Evaluating Machine Learning Predictions for Budget Overruns in U.S. Mass Transit Projects: A Correlational Study

Dissertation Manuscript

Submitted to National University

School of Business and Economics

in Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF BUSINESS ADMINISTRATION

by

PAUL ARTHUR BOUDREAU

San Diego, California

October 2025

Abstract

This study examined supervised and unsupervised machine learning methods to determine if statistical correlations could be established to validate mass transit project budgets. Capital budget forecasts for new mass transit projects in the United States have consistently been underestimated. These budget overruns resulted in funding allocation challenges for the Federal Transit Administration and obligated local transit authorities to secure additional financing. Developing accurate budget estimates is a significant challenge for project managers, particularly given the influence of cognitive biases such as optimism bias. Data from 108 projects were collected from publicly available U.S. government documents to serve as input for the machine learning models. Using a neural network and k-means clustering in SPSS software, the study created models that explained 94.6% (R^2 = 0.946) and 88.2% (R^2 = 0.882) of the variance in project budget outcomes. The findings highlighted the importance for organizations to apply AI-based machine learning technology to validate budget forecasts. This research built on prior research by applying machine learning methods that had been explored in other project contexts and problem domains. By employing SPSS software with configurable settings, rather than relying on customized Python code, the study enhanced the transparency and accessibility of machine learning regression analysis. The results demonstrated that a budget validation process could be developed using reproducible, data-driven predictive models that incorporate both project-specific and environmental variables.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

The ongoing trend toward urbanization has led to a growing demand for mass transit systems, resulting in an increasing number of projects initiated to meet the requirements (Statista, 2023). Mass transit system projects are part of global infrastructure spending, which is estimated at 3.9 trillion U.S. dollars per year (Woetzel et al., 2017). While mass transit systems have common components, such as transit stations, rail tracks, and vehicles, these projects can be very complex, and project success rates for complex infrastructure projects are estimated at 35% (Nieto-Rodriguez, 2020).

Project management is a discipline that consists of various tools, methods, and standards that help project managers deliver successful projects based on three significant objectives: project scope, budget, and schedule (Project Management Institute [PMI], 2021). Managing a project involves substantial challenges (Ammar et al., 2023; Dwi Ramadhan, 2023; Vinolia & Mudau, 2024), and one of the main responsibilities is ensuring an accurate project budget (Ackerman et al., 2024). Commonly used project management practices to create accurate budget estimates include analogous, parametric, template, and function points (PMI, 2021). According to the Project Management Body of Knowledge (PMBOK), these are cost estimating methods used under uncertainty (PMI, 2021). Two methods to validate budget accuracy are Monte Carlo simulation and reference class forecasting (Park, 2021). Project managers select the budget process based on the requirements of their project and the historical methods used by an organization (PMI, 2021).

Projects consist of organized processes for the deployment of solutions that support societal needs. Investment in projects is based on defining clear requirements and securing adequate funding to complete the scope of work (PMI, 2021). For mass transit projects in the

United States, funding from the U.S. federal government is based on comprehensive project plans, including budgets created for each project (Federal Transit Administration [FTA], n.d.). Mass transit funding is based on business cases and forms a long-term vision of capital spending (FTA, n.d.). An accurate project budget is an essential element for ensuring the overall plans are achievable within the funding requested (Gao & Touran, 2020).

Artificial intelligence (AI) has emerged as a transformative technology that can be applied to many areas of project management (Nenni et al., 2024). In a literature review, Nenni et al. (2024) highlighted the growing body of research on applying AI to project management, driven by low success rates in achieving project objectives. A component of AI is machine learning (Kelleher et al., 2020). Machine learning models use historical datasets to make predictions or perform classification (Murphy, 2012). These machine learning algorithms offer a new paradigm to predict the accuracy of a project budget (Pang et al., 2022) and predict the occurrence of project overspending based on historical datasets (Bakhshi et al., 2022). Two common categories of machine learning are supervised learning and unsupervised learning. Supervised learning uses labeled datasets to perform prediction, and unsupervised learning classifies the datasets into clusters based on common characteristics (Murphy, 2012).

Acquiring data and data management are integral components of machine learning (Boehm et al., 2022). Historical data must be sufficient and relevant to make accurate predictions and perform classification (Kelleher et al., 2020). Human bias has the potential to affect project budgets (Flyvbjerg, 2022). Machine learning algorithms also have the potential to be affected by and amplify bias (Dimitrakopoulou et al., 2024). Analysis using AI-based methods is a potentially innovative and less biased method to validate budget accuracy (Bakhshi et al., 2022;

Chandanshive & Kambekar, 2019; Ghimire et al., 2023), but needs further study to evaluate its effectiveness when applied to mass transit system capital budgets.

Statement of the Problem

The problem addressed in this study was the underestimation of budgets for mass transit system projects in the United States. There is ample evidence that traditional methods for project budgeting have resulted in budget overruns (Chen et al., 2023; Gao & Touran, 2020; Locatelli et al., 2017; Odeck, 2019; Park, 2021). Gao and Touran (2020) studied 81 U.S. transit projects completed between 1987 and 2018 and found that 77% exceeded their original budgets. A study of megaprojects across infrastructure industries, including mass transit, revealed that 90% had cost overruns of at least 50% (Flyvbjerg, 2014). Smaller projects are also susceptible to budget overruns (Mæhlen et al., 2024). Inaccurate project budgets have several negative implications, including reduced credibility in obtaining funding, financial loss, damage to customer trust, adverse publicity, deterioration of competitive advantage, and disappointed project stakeholders (Yim et al., 2015). Budget credibility issues can also result in hesitation to begin new projects (Thore Olsson et al., 2018). Projects that exceed the budget may be terminated due to declining confidence in achieving the project's value (Nieto-Rodriguez, 2021).

Many academic studies highlight the difficulty of identifying effective practices for the project budgeting process (Ammar et al., 2023; Denicol et al., 2020; Gómez-Cabrera et al., 2024; Moura & Ribeiro, 2021; Thore Olsson et al., 2018; Chen et al., 2023). A literature review conducted by Nenni et al. (2024) investigating AI in project management revealed a gap in addressing the accuracy of project budgets for mass transit. What is not known is the extent to which different AI methodologies, such as supervised and unsupervised learning, can reliably and effectively predict budget overruns for mass transit infrastructure projects and how these

predictions might reduce the frequency and scale of cost overruns compared to traditional forecasting methods.

Purpose of the Study

The purpose of this quantitative correlational study was to determine the effectiveness of machine learning algorithms using supervised and unsupervised learning to validate a budget for mass transit projects. The research assessed if machine learning could provide an objective method to validate a project budget. An accurate prediction of a project budget can reduce overruns by ensuring stakeholders and project funding providers are aware of a project's cost before the project begins, allowing time for action to address the situation (Al Jabouri, 2021; Doloi, 2013).

Project characteristics from historical datasets were used for input to create machine learning models. The dependent variable was the inflation-adjusted actual cost for the projects. The independent variable was the machine learning prediction of the project cost. The study analyzed data using SPSS software modules that contain supervised learning and unsupervised learning algorithms.

The target population for secondary data was mass transit projects in the United States. The publicly available data were mainly acquired from the U.S. Federal Transit Administration (FTA) website, which maintains the National Transit Database (NTD). Additional data relevant to mass transit systems were collected from publicly available sources, including the U.S. Census Bureau, the U.S. Bureau of Economic Analysis, and relevant city government documents. The additional data supplemented the datasets to expand the factors used by the machine learning algorithms. The potential impact of this research is significant, as it offers insights to promote changes in budget validation. The results can provide decision-makers with evidence-based

insights into the effectiveness of using AI-based methods for budget predictions. The study contributed to a greater understanding of how AI-driven analysis can be effectively applied to project management, particularly in the context of mass transit project budgets.

Introduction to the Conceptual Framework

The conceptual framework for this research was cognitive bias, which is the human tendency to make decisions that are not based on logic (Kahneman & Tversky, 1979). In project management, bias is a significant problem when determining the accuracy of a project budget (Prater et al., 2017). Kahneman's seminal research challenged the belief that humans make rational economic decisions (Kahneman et al., 2016). Kahneman and Tversky (1979) explained how people deviate from rational decision-making, have a tendency to overestimate positive outcomes, and frame project goals in overly positive terms. The term planning fallacy was proposed to describe underestimating task completion time and costs, regardless of available historical estimates for those tasks (Kahneman & Tversky, 1977). Flyvbjerg (2022) described 10 behavioral biases common to project managers. He categorized the two leading causes of inaccurate budgets as optimism bias and strategic misrepresentation (Flyvbjerg, 2013). Optimism bias is an innocent form of misrepresentation, whereas strategic misrepresentation is a deliberate strategy based on political pressures to acquire project approval (Flyvbjerg, 2013). This description of cognitive bias aligns with the concept of implicit bias being unintentional and explicit bias being deliberate (Greenwald & Banaji, 1995).

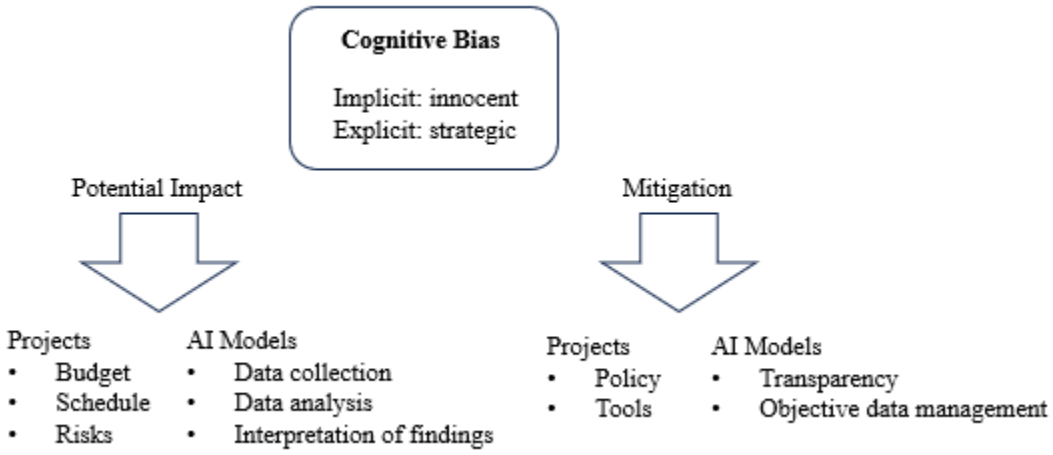
From a project perspective, the project scope, budget, schedule, and risks can all be affected by cognitive bias. Observable evidence of bias is the repeated occurrence of project budget overruns across successive projects. A mitigation approach for reducing optimism in creating a budget is to implement a process that changes the budgeting method. This intervention

can be a tool to reduce or eliminate bias from project estimates (De Reyck et al., 2017; Vela et al., 2022). A prediction of the project budget using a machine learning algorithm offers a new method to validate the budget, regardless of reasons for cognitive bias in creating the project scope, schedule, and budget.

Similar to the challenges in creating an accurate project budget, the research minimized or eliminated cognitive bias in all aspects of the dissertation work. Data collection, analysis, and presentation of the findings should be devoid of personal bias. The research included practical considerations for objectively implementing the research design and the adjustments that may occur during the study. Cognitive bias can influence the variables collected by a project manager that provide input to estimating methods. If machine learning algorithms attempt to reduce or eliminate human bias, caution must be exercised to avoid bias in data collection and selection (Hagendorff et al., 2023). A proposal by Min (2023) to address bias in data collection is to increase the range of data sources, add data from different contexts, and create synthetic data, if required, to balance the datasets. In addition, data should only be collected from reputable sources (Gudis et al., 2023). As a conceptual framework, bias needs to be addressed as part of the process of collecting the data, selecting the relevant data fields, completing the analysis, and describing the findings (Aquino, 2023). Transparency in the selection of analysis software and machine learning model configuration of the algorithms was an essential aspect of the study's methodology (Hagendorff et al., 2023). The findings were based on predetermined measures of statistical significance that promote an unbiased approach to the study results, as summarized in Figure 1.

Figure 1

Cognitive Bias: Implications and Mitigation



Note. Based on Flyvbjerg, 2022; Gudis et al., 2023; Vela et al., 2022.

Introduction to the Research Methodology and Design

This research was a quantitative study to evaluate the effectiveness of machine learning in predicting budget accuracy for mass transit system projects. A correlational research design was selected to determine if a relationship exists between two groups of variables (Cohen, 2010). This study used two types of machine learning models to perform the correlation analysis. A neural network algorithm was used for prediction based on supervised learning, and the k-means method was used for classification based on unsupervised learning. Pang et al. (2022) used three distinct datasets to enhance machine learning capability to predict information technology (IT) project cost and duration. These studies provided an opportunity to emulate parts of the research approaches. The sequence for completing the study was data collection, data

analysis, and interpretation of findings. These steps were guided by the conceptual framework of cognitive bias.

Secondary data were collected, consisting of a list of projects and project characteristics. The research included standards for collecting and using the data for input to the AI-based algorithms. Data fields containing project characteristics are called features in machine learning language and are based on the variables collected (Ewees et al., 2024). The main variables for the datasets were collected from the publicly available FTA website and the NTD, the primary source for transit data in the United States (National Transit Database, 2024). The NTD was mandated by Congress in 1974 (FTA, n.d.). The data were supplemented based on reports by the Government Accountability Office, which delivers reports to the U.S. Congress, and reports from city websites such as the Chicago Transit Authority.

Gao and Touran (2020) identified 81 mass transit projects and performed a paired *t*-test using seven data fields to evaluate if the implementation of a policy by the FTA in 2003 was effective in reducing budget overruns. Zhou et al. (2022) identified 14 data fields and used machine learning to predict mass transit operating expenses. The data fields for this research were based on these studies and supplemented by additional data fields relevant to predicting an accurate project budget. The main instrument for the study was a data collection template, which was maintained in a spreadsheet format.

The research design for the study was based on SPSS software capability. The data were categorized as nominal and scale according to the levels of measurement required by SPSS. The machine learning analysis function selected in SPSS applies preconfigured default options when creating models, thereby minimizing researcher bias in the configuration. The analysis was based on supervised learning using a neural network and unsupervised learning using k-means

classification. For the supervised learning results, the predicted value of the budgets was compared to the actual value for each project to determine a coefficient of determination (R^2) and the level of significance (p-value). For unsupervised learning, the datasets were grouped by project similarity. The mean of the project clusters was compared to the actual cost for each project to determine the R^2 value. A paired t-test was used to evaluate significance, The R^2 values generated from the supervised and unsupervised learning were also compared to assess the effectiveness of each method. The same project datasets were used for supervised and unsupervised methods. Any adjustments to the data or algorithms were explained and based on practices from previous research studies.

Research Questions

To address the problem of budget underestimation for mass transit projects, research was performed to determine the effectiveness of machine learning algorithms in predicting project budgets.

RQ1

To what extent can a machine learning algorithm using a supervised learning model accurately predict the budget for mass transit projects in the United States?

RQ2

To what extent can a machine learning algorithm using unsupervised learning accurately predict the budget for mass transit projects in the United States?

RQ3

Is there a statistically significant difference in predicted budget between supervised learning and unsupervised learning algorithms for mass transit projects in the United States?

Hypotheses

$H1_0$

A machine learning algorithm using supervised learning cannot achieve a statistically significant prediction for the budget of mass transit projects in the United States.

$H1_a$

A machine learning algorithm using supervised learning can achieve a statistically significant prediction for the budget of mass transit projects in the United States.

$H2_0$

A machine learning algorithm using unsupervised learning cannot achieve a statistically significant prediction for the budget of mass transit projects in the United States.

$H2_a$

A machine learning algorithm using unsupervised learning can achieve a statistically significant prediction for the budget of mass transit projects in the United States.

$H3_0$

Supervised learning does not perform significantly better than unsupervised learning in predicting the budget of mass transit projects in the United States.

$H3_a$

Supervised learning does perform significantly better than unsupervised learning in predicting the budget of mass transit projects in the United States.

Significance of the Study

The study advanced the understanding of how machine learning models, employing supervised and unsupervised learning algorithms, could be leveraged to predict budget overruns in mass transit infrastructure projects. The limitations of traditional forecasting methods were discussed in the academic literature review, leading to a more accurate and data-driven approach

to budget validation using machine learning. The contribution of this research to the field of project management and infrastructure planning could be substantial, as the literature consistently identified the prevalence of cost overruns and the inadequacies of current techniques, which have the potential to be influenced by cognitive biases such as optimism bias. By integrating AI, the study described new practices and approaches in project management, leading to more efficient capital expense planning, particularly in response to the increasing demand for effective public transit systems.

Definition of Key Terms

Project Budget

A project budget is a documented financial requirement based on completing all project-related activities to achieve the project's objectives (Project Management Institute [PMI], 2021).

Machine Learning

Machine learning is a subset of artificial intelligence that uses algorithms to identify patterns in data, which are used to perform prediction or classification (Manning, 2020).

Mass Transit

Mass transit is a system of transportation designed to move large numbers of people efficiently within an urban area. Typical characteristics of mass transit are high capacity, fixed routes, and public access (U.S. Department of Transportation, 2015).

Project Management

According to PMI (2021), project management is the application of knowledge, skills, tools, and techniques to activities to achieve the project objectives.

Supervised Learning

In supervised learning, the datasets have a label representing the target variable. The supervised learning algorithm uses correlational statistics to create a software model that best represents the label (Das et al., 2024).

Unsupervised Learning

Unsupervised learning uses unlabeled datasets to correlate the datasets into clusters of similar features (Das et al., 2024).

Summary

The purpose of this study was to explore the potential of machine learning models, utilizing both supervised and unsupervised learning algorithms, to enhance the accuracy of budget predictions for mass transit projects in the United States. The conceptual framework considered the impact of cognitive biases, particularly optimism bias, in budget planning and highlighted the potential of AI-driven methodologies to mitigate these biases. The research used a quantitative design with secondary data sourced from reliable public records mainly provided by the Federal Transit Administration and Government Accountability Office, ensuring the study's reliability and potential for replication by other researchers. The research approach used commercially available SPSS statistical analysis software to create and analyze correlations based on the machine learning models. The study's results could inform other research on innovation for project management practices, especially project budgeting, resulting in significant implications for the allocation of funding for infrastructure projects.

Chapter 2: Literature Review

The purpose of this literature review was to critically examine existing research on the intersection of project budgeting practices, cognitive bias, and machine learning, with a particular focus on infrastructure projects such as mass transit systems. Traditional budgeting methods were contrasted with research that evaluated using machine learning as a method to improve the accuracy of project budgets. By investigating the application of machine learning algorithms in project management, the review highlighted the potential for a more innovative approach that offers objective, data-driven methods for budget validation.

The literature review examined several major themes that established the conceptual foundation for this study. The themes reviewed included project management budgeting practices, an overview of the U.S. mass transit system, machine learning approaches in research, proper data management, the application of machine learning to projects, the use of SPSS for machine learning analysis, and addressing inherent biases in machine learning studies. The conceptual framework of cognitive bias influences all the themes, which were key areas of research addressed in the dissertation. The review explored project management practices for creating a budget with uncertain estimates, how they were applied, and if they were effective. This foundation provided context for understanding the historical and methodological challenges that affect cost-estimation accuracy in U.S. mass-transit projects.

A central focus of the review was the role of machine learning in addressing budgeting uncertainty within mass-transit projects. Studies were examined that applied machine learning to project-management problems, particularly for validating the accuracy of project budgets. A vital component of machine learning is data management, which involves data collection and analysis. Academic studies in this area were evaluated and critiqued in relation to the

conceptual framework. Various types of machine learning algorithms exist, but this study focused on two common methods, a neural network, and k-means clustering, to deliver statistical results that respond to the research questions. All themes were evaluated from a perspective of understanding bias that may be introduced and have a consequential impact on the project outcomes and the research process. Bias mitigation provided the rationale for using SPSS software to perform the machine learning analysis.

Pinto (2022) reflected on 40 years of project research, postulating the potential to reach a stagnation point and the risk of diminishing returns in academic productivity due to “safe" research topics. The article emphasized the need for a new path forward with a greater focus on an innovative approach and creative methodologies to address real-world project management issues. The literature review for this dissertation supports the quest for a more innovative structure for predicting budget variances and responding to the need for improved accuracy in project budgets. The adequacy of solutions for project budget overruns and the lack of a standard method for applying machine learning are scrutinized in the literature. While traditional methods such as expert judgment and parametric estimating are commonly used, they are susceptible to cognitive bias (Flyvbjerg, 2022). Researchers have explored a more objective assessment process using machine learning to improve project performance. However, a structured method is needed for how to apply these concepts to enhance the validation of a project budget.

Literature Review Strategy

The literature review identified, analyzed, and synthesized existing research on the use of machine learning to validate and predict project budgets. This included identifying gaps in the current literature, exploring the efficacy of machine learning algorithms, and understanding their application in project management, specifically within infrastructure and mass transit projects.

The primary search databases were accessed through the National University library, and included ProQuest, EBSCOhost, IEEE Xplore, ScienceDirect, SAGE, and Emerald Insight. An additional source was Google Scholar, which identified highly cited papers. Reference mining was also performed from relevant project management dissertations. For statistical data, the Statista database was accessed, and the retrieval of mass transit data variables was mainly from the Federal Transit Administration (FTA) and Government Accountability Office (GAO) websites.

The search process involved careful keyword selection and Boolean operators (AND, OR) to refine results. Keywords were categorized into three main areas: machine learning, budget variances, and project management. Under machine learning, terms such as "artificial intelligence," "machine learning algorithms," "supervised learning," and "unsupervised learning" were used. For budget variances, terms included were "project budgets," "budget validation," "cost prediction," "budget overruns," and "financial variance analysis." To ensure relevance to project management, keywords like "infrastructure projects," "construction projects," and "mass transit projects" were incorporated. Combined searches included phrases such as "machine learning AND budget variances," "AI OR machine learning AND cost prediction," and "budget overruns AND infrastructure projects AND artificial intelligence." The words "cognitive bias" and "bias" were appended to initial searches to include articles relating to the conceptual framework. Additional research was performed regarding SPSS software to inform decisions regarding configuration for the supervised and unsupervised learning functionality and statistical results.

The search process was iterative. Preliminary searches identified key papers, and the search terms were refined based on initial findings. Filters such as publication year focused on

recent advancements, specifically within the past five years. Citation chaining was employed to review references in foundational articles and track forward citations, ensuring the inclusion of seminal and emerging studies. Inclusion criteria emphasized peer-reviewed articles and reputable project management industry reports published from 2019 to 2025 that focused on traditional and innovative ways to understand budget accuracy. Exclusion criteria eliminated articles related to information technology (IT) projects or software budgets unless they included a machine learning analysis that could be used for infrastructure projects.

This strategy enabled a thorough understanding of the field, highlighting budget processes, mass transit project issues, machine learning applications, and the probability of bias. As gaps in the literature, such as a limited focus on mass transit projects and infrequent studies on using machine learning to validate budget accuracy, became evident, an expanded search was performed into more general infrastructure areas. Articles were reviewed to gain insight into how machine learning was applied and the approach used by different fields of study to obtain statistically significant results. The review provided a strong foundation for decisions in data acquisition, analysis, and evaluation of how the results fit into the potential for further research.

Conceptual Framework

The framework for this study was rooted in the theory of cognitive bias, which explores the ways human judgment deviates from rational decision-making. The seminal work of Kahneman identified how heuristics and psychological tendencies influence judgments, often leading to flawed decisions despite access to historical data (Kahneman & Tversky, 1979). As mentioned, the term planning fallacy described underestimating time and costs, regardless of available historical estimates for those tasks (Kahneman & Tversky, 1977). In the context of project management, Flyvbjerg's (2022) extensive research built upon this foundation,

identifying biases such as optimism bias and strategic misrepresentation as critical contributors to inaccurate project budgets and persistent cost overruns. These biases distort objective analysis and undermine decision-making, particularly in complex infrastructure projects. This framework provided the lens through which budgeting practices, machine learning applications, and the potential to mitigate these biases were evaluated. The framework for mitigating bias emphasized the need for data-driven approaches to improve accuracy and consistency in project outcomes.

In a study of the impact of cognitive bias in project management, Flyvbjerg (2022) identified 10 fundamental biases that contributed to distorted decision-making and suboptimal outcomes. Table 1 provides an overview of these biases, their descriptions, and their relevance to project budgeting and performance.

Table 1

Bias in Project Management

Bias Type	Description
Optimism bias	Overestimation of positive events or results
Strategic Misrepresentation	Deliberate distortion for political or strategic reasons
Anchoring bias	Reliance on one piece of information, often the first received
Uniqueness bias	The project is seen as more unique than it actually is
Planning fallacy	The propensity to underestimate costs, schedule, and risks and overestimate benefits
Overconfidence bias	Excessive belief that a person's answers are accurate
Hindsight bias	Past inaccurate estimates and events were predictable
Availability bias	The ease of information access directs actions and decisions

Bias Type	Description
Base-rate fallacy	Relying on a small sample or a specific event rather than generally available information
Escalation of commitment	A sunk cost fallacy that results in an increased commitment to a decision to maintain personal credibility

Note. Based on Flyvbjerg (2022) and Kahneman (2016).

Optimism bias is an innocent manifestation, whereas strategic misrepresentation is a deliberate strategy, typically based on political pressures to acquire project approval (Flyvbjerg, 2022). Kahneman et al. (2016) suggested ways to overcome bias, including observing variability in the results of past decisions and taking a more structured approach that uses algorithms. This supports an investigation into applying machine learning as a more structured and less biased approach to project budgeting.

In a literature review to discover areas affected by optimism bias in projects, Prater et al. (2017) identified the growth within project management studies to address this concern. The study captured 312 articles published between 2004 and 2015. The authors recommended additional quantitative studies to examine how to mitigate optimism bias as it affects project baselines. Aquino (2023) suggested that optimism affects the budgeting process but can also influence the variables collected that provide input to estimating methods, lauding the power of AI but highlighting the requirement to mitigate bias to ensure fair outcomes based on objective analysis.

Cognitive bias is also prevalent in other fields. In health care diagnosis, social bias, clinician bias, and technology bias have contributed to erroneous outcomes (Aquino, 2023). Critical thinking and AI tool validation were suggested to mitigate bias (Aquino, 2023). Vela

et al. (2022) provided details of the negative implications of bias in healthcare, where bias affects marginalized groups in the population. Critical intervention suggested to reduce or eliminate bias, included increased practitioner awareness and skills training. These change management methods were considered insufficient unless accompanied by structural system changes (Vela, 2022).

Jordao et al. (2020) confirmed that cognitive bias exists in business decisions and focused on optimism bias, overconfidence, and anchoring bias. The study verified previous research that refuted rational decision-making. The study suggested ways to mitigate bias, such as having managers deliberately seek out alternate options and contradictory perspectives. Machine learning algorithms offer a different perspective that may provide the necessary process change to remove or reduce bias in creating project budgets (Bakhshi et al., 2022; Chandanshive & Kambekar, 2019; Ghimire et al., 2023).

Love et al. (2022) presented a contrarian perspective, claiming there is no empirical evidence that correlates project cost overruns to optimism bias. The study analyzed social infrastructure projects in Hong Kong, finding 43% under budget and 53% over budget. The author admitted there is evidence that the planning fallacy exists at an individual level, but there is a lack of empirical support at a group level. Time constraints to finalize the budget were also proposed as a factor for optimistic estimates. A recommendation was made that more empirical research is required to verify the planning fallacy as the cause of budget overruns.

If a structured approach is introduced using machine learning algorithms as an attempt to reduce or eliminate human bias, caution must be exercised to avoid bias in the data collection, algorithm selection, and analysis process (Ayub & Banday, 2023). Ntoutsi et al. (2020) confirmed the requirement to scrutinize data to avoid inaccuracy. A definition of fairness in AI

systems is nebulous, but the study described processes for the development of objective strategies for data management, algorithm review, and analysis of outcomes. In response to these findings, this research study used machine learning algorithms to validate budgets for mass transit projects based on a process that incorporates these principles.

Project Budget Methodology

Project management is a multifaceted discipline with a variety of approaches, methods, and tools to address diverse challenges across industries (PMI, 2021). Modern project management can be traced to the 1950s when the field became recognized as a separate discipline (Cleland & Gareis, 2006). In a review of articles spanning six decades, Padalkar and Gopinath (2016) found that research on project management trends initially focused on technical aspects, then increasingly shifted to new methodologies and project performance. The authors identified three distinct but overlapping eras: deterministic, explanatory, and non-deterministic. The deterministic era focused on methods for optimizing scheduling and cost management, while the explanatory era examined success factors, performance management, and human elements.

The recent non-deterministic era considered complexity and uncertainty, emphasizing the connections between stakeholders and project dynamics. The research underscored the importance of addressing deterministic factors like cost and timeline adherence and non-deterministic factors like project complexity and uncertainty. Machine learning could bridge these paradigms by integrating historical data for deterministic predictions while accommodating dynamic changes in project environments. A model could incorporate stakeholder influences and external risks, addressing themes of complexity and interdependence highlighted in the article. This aligns with the article's call for blending empirical and conceptual methodologies. Machine learning systems could enhance project budget validation by creating adaptable models that

simulate real-world unpredictability, making them valuable for modern infrastructure projects. The inference of evolving eras in project management research aligns well with a data-driven, adaptive approach to using machine learning for mass transit budget validation.

Similar to evolving trends, success in a project has different interpretations. Albert et al. (2017) emphasized that the description of success has moved beyond the traditional baselines of delivering the scope on time and within budget. They highlighted the growing significance of stakeholder and customer satisfaction with the project results. There is no agreed standard for success criteria across different project types, suggesting that success metrics must be tailored to the specific context of each project (Iriarte & Bayona, 2020; Müller & Jugdev, 2012). Various factors can influence the measurement of a project's success. This is confirmed by Lamprou and Vagiona (2022), who claimed that traditional critical success factors remain valid, but stakeholder satisfaction and long-term effectiveness can also become relevant. Locatelli et al. (2017) cited factors that undermine project performance as complexity, weakness in organizational design, optimism bias, and strategic misinterpretation. They added another factor, the possibility of corruption in large infrastructure projects. Research has evolved from focusing solely on scope, budget, and schedule to embrace broader definitions of project success. According to Müller and Jugdev (2012), the field has developed more nuanced frameworks, advanced statistical analyses, and strategic perspectives that align project success with business objectives. Despite the advancements, they explained that challenges remain, including the subjective nature of success and the need for contextual definitions that vary across industries, project types, and stakeholders. Developing a customized set of metrics to define success in mass transit projects can be subjective. Success from the FTA’s perspective requires an accurate budget due to the strategic allocation of limited funding (FTA, n.d.).

The Project Management Institute (PMI) defined three domains in which project managers need to be knowledgeable: people, process, and business (PMI, 2021). The people domain is focused on interpersonal skills, the business domain is about strategic aspects of projects within an organization, and the process area emphasizes technical aspects such as tools and methodologies. The project budget is a critical baseline for assessing project performance, and budget processes are under scrutiny in the face of a dynamic, ever-changing environment.

A project budget consists of approved cost estimates based on scheduled activities and a contingency amount added for anticipated risks (PMI, 2021). Project budgeting practices rely on sources such as the Project Management Body of Knowledge (PMBOK) to provide guidance for creating accurate estimates when faced with uncertainty (PMI, 2021). Cost estimating methods are typically categorized as bottom-up or top-down. Bottom-up estimates require calculating costs based on the Work Breakdown Structure (WBS), which defines a detailed level of tasks (PMI, 2021). The costs are subsequently aggregated to create a total project estimate. Parametric and template methods are also considered bottom-up estimates. Parametric estimating uses historical data to calculate several items multiplied by an expected cost per item (Egboga et al., 2022). Parametric estimating can be combined with regression analysis to improve road and railway tunnel project cost estimates (Ahmed, 2021). An example of the template method is a list of the tasks required, the expected hours to complete each task, and an hourly rate (PMI, 2021). The calculations are made and summarized, and, if necessary, additional amounts are added for other factors, such as administration costs. Bottom-up processes for creating a budget are purportedly more credible (PMI, 2021). However, the estimates are still susceptible to optimism bias by the individual creating the estimate for each task (Flyvbjerg, 2022). Three-point estimating is a method that tries to balance the estimates based on using the optimistic,

pessimistic, and most likely estimate to calculate the cost for a task (PMI, 2021). This process aligns with the suggested solution by Jordao et al. (2020), which encourages individuals to look at all perspectives.

Top-down methods include analogous estimates, which are comparisons to previous projects, and new cost estimates that are prorated based on the perceived difference in proportional size or other tangible factors (PMI, 2021). The apportion method is another top-down method to create a budget. Based on historical data for a similar project, a summary budget is created, and the tasks are allocated a percentage of the total budget based on a subjective allocation (PMI, 2021). These top-down methods rely on historical information that may be biased, thereby propagating bias for new projects. A top-down or bottom-up method can benefit from a more objective, data-driven approach to identify the probability of budget overruns. Gómez-Cabrera et al. (2024) determined that cost overruns typically occur in the project execution stage or once the project has started. Their suggested improvement is greater consistency in using project management practices. If a method to validate the accuracy of project budgets is determined, additional work will be required to incorporate it as a standard practice. Al Juboori (2021) conducted a qualitative study to identify practices that increased the accuracy of project budgets. The best practices discovered were building stakeholder relationships, having experienced project managers, and using software and technology to reduce human error. This study suggested acceptance of a technical solution and aligns with using machine learning algorithms.

Budgets normally contain a contingency amount based on a historical standard or the calculation of the financial impact of risks that may occur on the project (PMI, 2021). Potential risks are identified with a probability of occurrence and a cost to mitigate or resolve the risk. The

probabilities are multiplied by the financial impact to determine a cost factor added to the budget as a contingency (PMI, 2021). Hollmann (2021) suggested that parametric estimating is a viable method to quantify risks in a project and predict the contingency amount required. The risk management process still has the potential to be affected by bias in the list of potential risks created, the probability assigned, and the budget impact to resolve them. Alternatively, optimism bias can be treated as a separate risk and added to the risk register.

Yim et al. (2015) highlighted the significant consequences of project failure and the need for improved risk management strategies in the face of project challenges. Project failure can result in financial losses, negative publicity, erosion of customer trust, and loss of competitive advantage. These apply to mass transit projects, where public perception and trust are critical (Rodrigue et al., 2024). The study by Yim et al. emphasized the need for a structured framework in managing risk to enhance the predictive and preventive measures applied to infrastructure projects. Developing a structured machine learning process supports this appeal for a more proactive approach to enhance project success.

Two common methods used to calculate contingency are the Monte Carlo method and reference class forecasting (RCF). The Monte Carlo method applies random probabilities to risk uncertainties and performs multiple simulations to determine different outcomes (McNeil et al., 2015). Monte Carlo simulation is used to predict the probability of an outcome based on the random probability of risks affecting the project (Paşcu, 2023). Deng and Jian (2022) used the Monte Carlo simulation method to determine a range of possible outcomes in cost deviations. The method required the project to be in progress and generate critical path method (CPM) and earned value management (EVM) status metrics. The approach did not quantify the variance at completion (VAC), only providing insight for the highest impact activities and range of probable

outcomes. Elghaish et al. (2021) suggested that Monte Carlo simulation is effective for determining construction project risks when faced with a paucity of information. This method may not be effective for mass transit projects where the activities are known, and costs can be based on historical information.

Ottaviani et al. (2024) described a framework for project contingency based on risk perception and highlighted how traditional methods that rely on subjective judgments by project managers lead to inefficiencies. Their proposed framework used Monte Carlo simulation to manage contingency reserves by anticipating short-term and long-term cost overruns. The intent was to optimize contingency budgets to efficiently address potential project cost overruns. Guan et al. (2024) used fault tree analysis to optimize the allocation of limited project contingency funds. Both of these approaches accept that contingency may be insufficient and does not address a validation process to determine the accuracy of the contingency amount. A study by Kwon and Kang (2018) introduced a method for project budget estimation that separates contingency for identified and unidentified risks to improve accuracy. They emphasized that traditional estimates often only cover identified or known risks, leading to a lack of comprehensive cost coverage. Analyzing 20 residential construction projects in South Korea, they found significant budget overruns primarily due to unidentified risks. The proposed method to address this used three-point estimation to determine the estimated mean risk cost and *R*-value analysis to determine cost variation. Their conclusion was a proposal that could reduce cost variances and enhance budget reliability when applied to projects already in progress. By distinguishing between these types of risks, the study applied different strategies to capture the total potential cost exposure, allowing for more accurate budget estimates.

Reference class forecasting is an approach based on selecting meaningful historical data, performing statistical analysis, and comparing the project data to the current project (Baerenbold, 2023). Flyvbjerg et al. (2016) applied this process to road construction projects in Hong Kong, and the results demonstrated an improvement in existing forecasting methods. Park (2021) analyzed results for 107 infrastructure projects in the United Kingdom that implemented RCF and identified that the average cost overrun declined from 38% to 5 %. The probability of completing a project within budget also increased by 12%. The results for the United States did not show the same level of improvement (Park, 2021).

Critics cite that optimism bias is used to justify reference class forecasting, and there are gaps in understanding how RCF interacts with the project budgeting process (Chen et al., 2023). For example, organizational factors such as political expediency may rely on reference class forecasting to justify unnecessarily large contingency reserves (Chen et al., 2023). In another study critical of reference class forecasting as a budget strategy, Baerenbold (2023) highlighted the lack of standards in using reference class data. The method also ignores the social and economic environment and has not been adequately tested across other industries.

The process of applying reference class forecasting is to obtain datasets of project characteristics that most closely resemble an upcoming project and then select the dataset that most closely matches the project conditions to validate the project budget. Finding similar datasets can be challenging, and selecting the closest match is left to subjective judgment. A less biased approach may be to use historical datasets as part of a machine learning classification process to match to the upcoming project. This is a gap in the literature and was the focus of this research which was to determine how machine learning can be applied to project budget validation.

Assuming optimism bias is not captured in the risk management contingency calculations, Chen et al. (2023) suggested imposing an administrative bias factor that adds funds to a budget to account for optimism bias. The concept described adding an overlay, called an optimism bias uplift, to transportation project budgets (De Reyck et al., 2017). A budget uplift is a contingency factor, not a method to validate a project budget. A study by Ika and Munro (2022) for World Bank projects suggested that optimism bias was responsible for 20% of the project underperformance, and the balance was due to factors such as complexity and uncertainty. Brandl et al. (2021) suggested that existing project management practices are inadequate for complex projects and that innovation is required for success. With no effective solution for project budget validation, performing research using machine learning methods is a viable alternative.

Another project methodology known as Projects IN Controlled Environments (PRINCE2) is more commonly used in Europe (Takagi et al., 2024). This differs from PMBOK, which is a guideline-based approach that contains best practices and knowledge areas for project management. The PRINCE2 framework emphasizes dividing projects into manageable stages, with clearly defined roles and responsibilities. The approach focuses on business justification and alignment with organizational goals (Office, 2017). As described in Table 2, PRINCE2 is based on seven key principles that provide structure for initiating and managing a project.

Table 2

Seven Principles of PRINCE2

Principle	Application to cost and budget
Continued business justification	Ongoing cost/benefit analysis and regular financial reviews
Defined roles and responsibilities	Clear accountability for decision-making
Learn from experience	Apply lessons learned from previous projects
Manage by exception	Use tolerances and escalate as required.
Manage by stages	Cost control at each stage
Focus on products	Clear deliverables to avoid scope creep
Tailor to suit the project environment	Flexible to size risk conditions

Note. Based on https://www.axelos.com/certifications/propath/prince2-project-management and the Association of Project Managers (APM) https://www.apm.org.uk/resources/find-a-resource/does-the-prince2-generalised-approach-suit-the-challenging-environment-faced-by-it-projects-particularly-in-relation-to-risk-management/

For projects in the public sector, Takagi et al. (2024) proposed integrating PRINCE2 with the concept of success management by adding structured activities to address the gaps in PRINCE2 for evaluating success and key performance indicators (KPIs). Cost and budget were not emphasized as part of the success criteria, which were focused more on efficiency and organizational benefits. In a comparison to PMBOK, Simonaitis et al. (2023) defined PRINCE2 as emphasizing structured progress monitoring and control to help manage project costs effectively. By establishing detailed timelines, clear quality control standards, and specific accountability, the PRINCE2 principles aim to reduce the risk of budget overruns by using accountability, quality standards, and timeline management to minimize deviations.

As the Oxford Chair of Project and Program Management, Flyvbjerg has been prolific in publishing research on project issues, focused mainly on persistently poor infrastructure performance. In a foundational study on cost escalation in transportation projects, Flyvbjerg et al. (2002) demonstrated that cost underestimation is a systemic issue. They identified an average cost overrun of 20% in road projects, 34% in fixed links like bridges and tunnels, and 45% in rail projects across 20 countries, indicating that inaccuracies in forecasting are widespread and consistent. Based on a global study of megaprojects, which were defined as projects with a budget of over $1 billion, Flyvbjerg devised a quotable expression for his concept known as the iron law of megaprojects: "over budget, over time, under benefits, over and over again" (Flyvbjerg, 2014, p. 2). This criticism of traditional project risk management and cost estimation methods was followed by research into optimism bias. Flyvbjerg proclaimed the need for a realistic approach to project planning with increased accountability and better data-driven decision-making (Flyvbjerg, 2009). He argued that planning and decision-making must incorporate empirical data from past projects to improve accuracy and outcomes (Flyvbjerg et al., 2014). This provides further support for research into using historical data and machine learning as a data-driven method for improving project performance.

As previously referenced, Flyvbjerg (2022) identified 10 behavior biases affecting project managers. In a study of 219 information technology (IT) projects, Flyvbjerg et al. (2024) explored uniqueness bias, the degree to which a project manager perceives their project as unique. A statistically significant correlation was reported between a higher perception of uniqueness and an increase in cost overrun ($p < 0.05$). The study did not find a correlation between uniqueness bias and schedule delays. The strategy recommended to

mitigate the effects of bias included using reference class forecasting based on historical datasets. Flyvbjerg's research was a call to instigate change from traditional practices by highlighting public infrastructure projects that resulted in excessive cost overruns (Flyvbjerg, 2022). An example was the Olympic Games hosting preparations from 1960 to 2016, which had an average of above 100% in cost overruns (Flyvbjerg, 2016). These studies infer a broader issue of cost overruns in project management that is not categorized by project type, size, or geographic location. The weakness in RCF is identifying contextual differentiation between successful and unsuccessful methods, particularly beyond the approach to project budgeting (Baerenbold, 2023). Flyberg (2022) concluded that the problem to be solved is addressing bias not the budgeting methodology.

Liao et al. (2022) conducted a literature review of 436 articles on intelligent risk management in construction projects and identified the potential of integrating AI and digital technologies, such as building information modeling (BIM) and Internet of Things (IoT), into risk management practices. They concluded that highly complex projects need to shift to technology-driven approaches. They suggested that intelligent risk management using AI to automate data analysis, risk identification, and decision-making can improve how construction risks are managed. For contingency budget calculations, these insights are significant as they promote a transition towards data-driven, AI-powered risk assessments that offer more comprehensive and accurate predictions of potential cost overruns and strengthen project finance resiliency. This is supported in a study by Khodabakhshian et al. (2023) where they highlighted the shift from traditional, manual methods, to AI-enhanced approaches such as using a neural network to deliver automation, improved decision-making, and the ability to handle complex data. They suggested that AI-driven risk management enhances the calculation of risk

contingency budgets by improving the prediction of potential cost overruns. Risk contingency can be validated separately, as suggested by this study, or as part of validating the accuracy of the overall project budget.

Another example of moving away from traditional methods is a study by Wang (2020) who constructed a comprehensive risk evaluation index system for the cost management of power grid projects using the WBS and a statistical evaluation of expert judgment ratings. The method resulted in key risk factors being identified. Establishing an efficient and practical risk evaluation system helps project managers identify and manage project costs using a more innovative and objective assessment than traditional methods. A structured approach or framework can also help with capital costs but needs to be based on unbiased methods.

Various factors may influence project budget results but they require judgment, which may be susceptible to cognitive bias. These factors include project complexity, amount of risk, the type of leadership skills, the effectiveness of resource allocation, and communication (Dinu, 2016; Robertson & Williams, 2006). For example, project complexity might be considered an independent variable that contributes to project budget overruns but is susceptible to subjective assessment as to what constitutes a differential level of a complex project. In an article by Dao et al. (2020), a methodology was outlined for assessing project complexity using a binary logistic regression model. Based on 44 construction projects, a set of 101 factors was reduced to 27 using univariate analysis and principal component analysis (PCA). Although statistical results were achieved, the input variables originated from a qualitative survey, which could be influenced by bias, requiring replication of the results at a broader level to be acknowledged as practical. Research by Bakhshi et al. (2016) viewed complexity as an important factor correlating to project failure, yet described a lack of understanding of what constitutes project

complexity. Through a review of 420 publications, they discovered a diversity of perspectives and lack of consensus on a definition of project complexity. An ambiguous definition of complexity means subjectiveness and bias would be incorporated into this variable if used as an independent variable.

Regarding project size, Denicol et al. (2020) and Flybjerg (2014) agreed with the definition of megaprojects as those exceeding US$1 billion and Flyvbjerg used the threshold of a budget over US$100 million to define large projects. In 2005, the U.S. Congress defined large projects in transportation as having capital costs above US$500 million (U.S. Congress, 2005). Other researchers (Jones & Lichtenstein, 2008; Merrow, 2011) defined project size using scope, complexity factors, stakeholder involvement, impact on the organization or external environment, and scale. These imprecise criteria can be quantified yet are subject to inherent human bias. Financial values are affected by inflation needing to be normalized using present value calculations to provide comparable outcomes. Using project size and complexity as independent variables risks subjective results. A data-driven approach would require a combination of different objective variables to represent these factors.

PMBOK and PRINCE2 contain practices and principles that suggest methods to create and maintain an accurate project budget. These approaches lack a process to validate that the budget will be accurate when the project is complete. A traditional method to validate a budget is obtaining expert judgment, which is using the knowledge of a project manager who has a lot of experience managing a similar project or projects. Research suggests this is inadequate due to cognitive bias and the inability to decipher and gain insights from previous experience that can be applied successfully to a complex project (Chen et al., 2023; Flyvbjerg et al., 2018; Kahneman et al., 2016). An objective and data-driven method using machine learning may be

more successful in improving the accuracy of project baselines. Research results suggest that using technology or technical solutions combined with historical project data can lead to more accurate project budgets.

U.S. Mass Transit Systems

The history of mass transit in the United States can be traced to a beginning over 200 years ago (Abdallah, 2023). It is now recognized as a critical element of sustainable urban transport, with opportunities to implement advanced technologies and materials that can significantly improve environmental outcomes (Abdallah, 2023). The global mass transit industry for new projects was valued at $110.53 billion in 2022 and is expected to expand at a compound annual growth rate (CAGR) of 13.0% from 2023 to 2030 (Grand View Research, n.d.). In 2022, the total funding spent in the public transit sector in the United States was approximately $84.2 billion, which includes projects at 32% and operational spending at 68% (Musick, 2022). In the United States, funding is available from the FTA. Since 1991, the FTA provided oversight for capital project implementation and, in 2003, implemented a risk assessment process as a requirement for funding (FTA, n.d.; Gao & Touran, 2020). Budget accuracy plays a critical role in capital planning as funding decisions are taken based on financial impact and prioritization of projects. Mass transit projects, especially the initial implementation in a city, tend to be large and complex. These projects typically use a combination of top-down and bottom-up cost estimating (Flyvbjerg et al., 2002; Kirk & Levinson, 2004; U.S. Department of Transportation, 2023). The projects also involve procurement contracts, require significant efforts in risk management, and may be subject to political issues (Federal Transit Administration, 2016; Flyvbjerg, 2022).

The first large-scale transit project in a city can be fraught with project issues, as evidenced in a report investigating project issues and budget overruns for the light rail deployment in Ottawa, Canada. The Ottawa Light Rail Transit Commission (2022) highlighted that both city officials and advisors were deficient in knowledge and experience, resulting in significant cost overruns and schedule delays. The report suggested that project managers need to find ways to "address the root causes of cognitive biases" (Ottawa Light Rail Transit Commission, 2022, p. 483). Uniqueness bias was also mentioned as a factor that resulted in poor project performance.

In the Gao and Touran (2020) study, data were collected data from 81 U.S. rail transit projects spanning 40 years to evaluate if a risk assessment mandated by the FTA in 2003 had an impact on reducing project budget overruns. The variables used were project size, project duration, and whether they were completed before or after 2003. The findings showed that budget overruns as a percentage improved over time and that larger projects, defined as costing over $500 million, had higher cost overruns. The higher cost overrun for larger projects was statistically significant ($p = .004$). The mean cost overrun for projects before 2003 was 6.7%, compared to 3.1% for projects after 2003. Using a *t*-test, to analyze the difference resulted in a *p*-value greater than 0.05, meaning there was no statistically significant difference due to the formal risk assessment implementation. While machine learning was not used, the study is important due to the sizable number of mass transit project datasets collected and analyzed for this type of study.

Sjögren and Norgren (2023) studied cost overruns for 114 Swedish infrastructure projects deployed between 2010 and 2022, identifying significant statistical relationships between cost overruns and factors of project size, type, and regional location. The researchers recognized the

difficulty of collecting factors that cause budget overruns. The regression model only explained 11% ($R^2 = 0.1128$) of the budget variance, indicating that numerous unmeasured or unpredictable factors also contributed. They suggested that cost overruns are partly driven by conditions not easily captured by standard project practices, resulting in an inherent unpredictability. The central and northern regions had higher cost overruns, making this a plausible independent variable. A paired *t*-test revealed that cost underestimation was more significant from 2018 to 2022 than in the earlier period, suggesting that changing project demands or external variables caused greater budget inaccuracy. The authors inferred without direct statistical correlation that optimism bias and strategic misrepresentation also contributed to cost underestimations. The study used only three independent variables to analyze the project budget overruns that ranged as high as 478% to projects that were under budget by as much as 92%. The mean cost overrun was 16.3%, with a standard deviation of 62.1%, signifying a skewed distribution toward extreme overruns. Additional independent variables could have contributed to a more comprehensive analysis. The authors may have been constrained by data availability or the desire for model simplicity, which led to results that were limited in their application for future projects.

Research by Mæhlen et al. (2024) involved collecting data from 489 projects in Norway and using regression analysis to determine that smaller projects experienced greater cost overruns. However, the factors used in regression analysis only explained 20% of the results. They concluded that project size and cost estimating practices are important predictors of cost overruns and suggested creating a standardized database for public project costs to improve prediction accuracy. The recommendation supports the findings of other studies that suggest using historical data as a method to improve budget accuracy and reduce budget overruns.

In a report from the California Institute of Transportation Studies, Elkind et al. (2022) used five case studies to analyze the high costs of transit projects. The study used the cost per kilometer as a measure to compare similar U.S. and international projects. The California projects had a range of 70% to 200% higher costs compared to the calculated U.S. standard, and 170% to 330% higher costs compared to the international metric. To curb higher costs, the authors identified the need for increased management expertise, greater adherence to project scope, coordination between contractors, and selecting more effective delivery methods. They noted that similar international projects had streamlined processes and used a more standardized approach. In addition to higher overall costs, they noted budget overruns, which they attributed to low-bid contracting, where underestimating costs for complex projects was identified as the primary reason.

A study by Odeck (2019) contributed insights to the collection and analysis of data for mass transit projects, particularly by demonstrating how mean percentage cost overruns (MPCOs) are influenced by specific characteristics such as project type and geographic region. The article highlighted the importance of accounting for variability in reported cost overruns rather than generalizing findings across all studies. For mass transit data collection, this approach underscored the need to gather diverse and detailed project data, including regional differences. The paper encouraged the inclusion of nuanced parameters in models for predicting and evaluating budget outcomes in mass transit projects. It strengthened the justification for applying innovative statistical methods and context-specific variables in planning infrastructure initiatives.

A review was conducted by Dempsey et al. (2022) to determine the most significant project management challenges. They selected eight articles from a comprehensive review of publications from 2000 to 2020. The main challenges were project complexity, control of project

baselines, and competence to manage technical, regulatory, and cultural issues. The inference is that projects are becoming more difficult to manage, and a project manager's ability to deliver successful projects will be based on increased knowledge and more effective processes. It is incumbent on researchers and practitioners to evaluate if machine learning technology offers viable process improvement solutions. Regardless of a myriad of factors that may contribute to project challenges, research suggests that the process needs to be anchored in an objective, data-driven method.

The use of data and statistical analysis is an underlying theme in the GAO Cost Estimating and Assessment Guide (U.S. GAO, 2020) which provided a detailed methodology for developing, managing, and evaluating program cost estimates within the U.S. government. Originally published in 2009. It highlights best-practices for ensuring accurate, reliable cost estimates. The challenges listed for producing accurate estimates include overly optimistic estimates, lack of high-quality data, and inadequate risk assessment. Without a reference to AI, the recommended 12-step process includes aspects aligned with using machine learning as part of the process. It describes collecting and validating data as a critical step. Under step seven, referring to point estimates, statistical methods are suggested such as *R-squared*, *t-statistic*, and *F-statistic*. This is similar to techniques applied in machine learning models to evaluate and validate predictions. A sensitivity analysis suggested in step eight encourages identification of the most significant variables in the cost estimate. The GAO report noted that data collected through subject matter expert interviews have the potential to be biased, stating that bias can originate from different sources, such as over-optimism, group think, dominating personalities, inexperience, or pressure from management. Motivational bias was cited as is a source of bias that arises when interviewees feel threatened if they give their true thoughts about a program.

Confidence bias is possible when the estimator is overly optimistic about the success of the program. Other forms of bias listed included a tendency to give more weight to recent events than earlier events, to assume patterns where none exist, and to assign connections to coincidences. When collecting data, the report emphasized that the estimator needed to be aware of any potential biases within the data.

A GAO report (GAO, 2019) highlighted deficiencies in the Federal Transit Administration's (FTA) cost estimating practices. Specifically, the FTA's budget creation guidance did not align with five of the 12 best practices outlined in GAO's cost estimation framework (GAO, 2009), particularly regarding sensitivity analyses, which assess the impact of changing assumptions on overall project costs. The report noted that the FTA's cost estimating information was scattered across 14 documents in a fragmented approach that undermined the ability to develop reliable estimates and effectively address cost overruns. To address these issues, the GAO report (2019) recommended two critical improvements. First, the FTA should ensure alignment with all 12 steps in GAO's Cost Estimating and Assessment Guide (U.S. GAO, 2009), Second, the FTA should consolidate all cost estimating information into a single, accessible resource. The objective should be to enhance the reliability of cost estimates, improve risk management, and minimize the occurrence of cost overruns in both public infrastructure and transit projects (U.S. GAO, 2019). Overall, the 12 steps align with the development of a structured machine learning approach to analyze historical data and perform budget validation based on statistical methods such as regression.

Based on the studies, higher costs and budget overruns appear endemic to mass transit projects in the United States and globally. Project management practices focus on creating an accurate budget and are not entirely successful. Regardless of how a project budget is created, a

method to validate the budget using historical data and an objective process can provide value to the field of project management.

Machine Learning

Machine learning is a component of AI that uses historical data to gain insights into a problem or situation using a form of regression analysis (Kelleher et al., 2020). There are different approaches to using machine learning for project management with two common methods being supervised and unsupervised learning (Sarker, 2021). Supervised learning performs predictions using labeled datasets to build a software model using project data features and a target variable or label. A common algorithm used for prediction is a neural network, which is a software technique that contains an input layer, hidden layers, and an output (Kelleher et al., 2020). The algorithm performs regression based on the features as independent variables and the label or target variable as the dependent variable to create a model. The algorithm employs forward propagation to determine the weights for each feature and employs backpropagation to adjust these weights to minimize the loss function, a statistical measure that indicates how well the model fits the features in each dataset (Hinton & Sejnowski, 1986). The input layer contains the features of each dataset, and the algorithm adjusts weights and balances until the optimum correlation is determined. When a new dataset is entered without a target variable, a comparison is made to the model, and an output in the form of a probability prediction is made (Ghori et al., 2020). The probability represents how closely the new dataset matches the model. The neural network algorithm uses hyperparameters which are predetermined settings (Hinton & Sejnowski, 1986). Typical settings are the number of hidden layers and the number of iterations the algorithm uses in the process to create the model. The advantage of neural network analysis is the ability to build a model that captures combined

contribution of each variable for the accurate prediction of a project budget variance (Hinton & Sejnowski, 1986).

Machine learning models, especially those using neural networks, can significantly enhance the diagnostic accuracy for mesothelioma (Kapila et al., 2023). The research underscored the potential of AI to aid medical professionals in early cancer diagnosis and improving patient outcomes. The statistical results had an $R^2 = 0.967$ and $F(5,9) = 52.31$, $p < .001$ indicating a significant relationship between model accuracy and the predictor variables. This unbiased approach in healthcare may seem unrelated to project budgets but the development of an early and accurate detection method for a diagnosis using machine learning suggests a similar approach can be developed at the start of a project to validate the accuracy of the project budget.

Hashala and Andrews (2023) designed a quantitative study to compare the effectiveness of five different machine learning algorithms in determining project costs for 552 engineering automation projects in Africa and the Middle East. For this study, 70% of the datasets were used to create the model and 30% to test it. The findings showed that a neural network had the greatest prediction accuracy, and the accuracy varied from that of the other models. The study verified the use of machine learning as a viable method to improve prediction accuracy. Neither study revealed measures taken to minimize bias in the data or the creation of Python-based algorithms. Ghimire et al. (2023) compared machine learning to multiple linear regression (MLR) for construction projects in New York City between 1995 and 2003. The 16 variables collected per dataset from the public database were narrowed to 5 using feature engineering. The Python algorithm was split into 70% training and 30% test datasets, and the accuracy at predicting the project budget had an R^2 of 0.96 compared to MLR at R^2 of 0.28 (Ghimire et al.,

2023). This study offered more beneficial evidence for using machine learning to improve project methods.

Although supervised learning remains more popular for business applications due to its precision and direct applicability, unsupervised learning provides value when labeled datasets are limited (Sarker, 2021). Unsupervised learning, by contrast, does not rely on labeled datasets but identifies patterns and clusters within the data (Hinton & Sejnowski, 1999). Classification algorithms, such as k-means clustering, group datasets based on similarities, offering a different perspective for project management tasks like risk assessment and stakeholder segmentation (Mariani et al., 2023). Mulyono et al. (2023) demonstrated the effectiveness of k-means clustering in identifying customer segments. The study used 20 variables from a survey of 120 customers to successfully segment the supermarket's customers into four distinct groups using k-means clustering in SPSS. The analysis provided actionable insights for marketing strategies, including targeted promotions and product placement. They used ANOVA to validate the clustering process, ensuring that the segments were statistically sound and practically relevant.

Optimal clustering outcomes depend on the method and distance measure used (Löster, 2016). The study by Loster (2016) demonstrated that no method is universally superior and recommended using multiple coefficients to verify clustering validity. Mulyono et al. (2023) demonstrated the effectiveness of k-means clustering in SPSS by identifying customer segments to improve customer engagement. The k-means clustering method effectively segmented the datasets, providing actionable insights. The number of clusters was pre-specified as three, and the analysis focused on minimizing within-cluster variance while maximizing between-cluster distances.

Data Management

The process for performing prediction and classification includes data collection, selecting an algorithm, and interpreting the results. A concern for researchers is the amount of data required to make a valid prediction or classification. There is also a requirement to avoid or minimize bias in data collection since accurate algorithm predictions are based on obtaining reliable data (Sghir et al., 2023). A strategy must be selected to determine the data volume, sources, and collection methods. As a conceptual framework, bias needs to be addressed as part of the process of collecting the data, performing the research, completing the analysis, and describing the findings (Aquino, 2023). A proposal by Min (2023) to minimize bias in data collection is to increase the range of data sources, add data from different contexts, and create synthetic data if required to balance the datasets. The collection and application of data is an essential consideration for a quantitative analysis. This requires an investigation into methods used for similar studies, data sources, and ethical considerations. The data collection strategy should also consider commercial interests, and the proprietary nature of the data (Basystiuk et al., 2023).

Two elements of the machine learning process are data management and feature engineering. Data management, also known as data wrangling, is the process of collecting, organizing, and validating that the raw data is in an acceptable format for input to machine learning algorithms (Kim & Lee, 2020). This process is foundational to any data-driven study, as the quality of the raw data often dictates the reliability of the insights derived (Aggarwal, 2015). Effective data wrangling ensures that datasets are consistent, relevant, and free from errors, enabling more meaningful and interpretable results. Data wrangling can inadvertently introduce or perpetuate bias, impacting the validity of the outcomes. Mehrabi et al. (2021) emphasized the importance of fairness in data management to mitigate bias during data preprocessing. Bias can

infiltrate from various sources, such as incomplete datasets and decisions made during cleaning and transformation (Kelleher et al., 2020).

Feature engineering is a fundamental process in data science and machine learning that involves transforming raw data into meaningful and relevant input variables, or features, that enhance a model's predictive power (Kuhn & Johnson, 2019). It involves the manipulation of data collected during the data wrangling process to ensure a fair representation of the data. Feature engineering is the process of determining the most relevant features based on the data, the model, and the objective (Zheng & Casari, 2018). According to Bolón-Canedo et al., (2016) feature engineering encompasses several critical tasks, including feature construction, selection, and transformation. Feature construction requires the creation of new variables from the existing data, by combining, aggregating, or deriving metrics. Feature selection focuses on identifying the most relevant variables, reducing redundancy, and improving computational efficiency while avoiding the inclusion of unrelated outliers in the model which is known as overfitting. Feature transformation ensures the data is in a format suitable for machine learning algorithms, addressing issues like scale, distribution, and data type mismatches.

In the study of budget underestimation in mass transit projects, avoiding bias during feature engineering is essential to ensure that machine learning models produce accurate, fair, and actionable insights. Bias can arise from various sources, including unbalanced data representation, historical bias embedded in the dataset, or the inclusion of features that act as proxies for sensitive attributes (Kuhn & Johnson, 2019). Mass transit datasets might disproportionately represent well-funded projects while neglecting smaller initiatives. Historical data might reflect systemic underfunding in regions of the United States, which could unfairly influence the model's outputs. The research performed in this study addressed the strategy for

data management and feature engineering as well as steps to avoid or minimize bias in these areas that are critical inputs for machine learning.

Applying Machine Learning to Project Management

The appeal by Pinto (2022) for creative research methods aligns with leveraging machine learning to explore dynamic, interconnected datasets in mass transit systems, potentially identifying more effective processes. While a machine learning algorithm has the potential to become a new method for optimizing budget validation, the concept might have the ability to be generalized to other areas in project management. Investigating how to apply AI concepts to project management also supports Pinto's emphasis for the necessity to push boundaries in this area of research. Indicating the diverse applicability of neural networks, Selim et al. (2024) studied how to estimate seepage losses from irrigation canals with the objective to minimize water loss in lined canals. The research used SPSS to create neural network models. The neural network achieved a high correlation with the two models having $R^2 = 0.996$ and $R^2 = 0.965$, indicating a reliable result.

Machine learning capability is explored in several academic papers investigating project performance. In an expansive literature review of 215 academic articles from 1996 to 2023, Nenni et al. (2024) concluded that a structured AI framework is needed for proper integration into project management processes. Based on their study, the leading industries adopting AI technology were construction and IT, which was attributed to the complexity of the projects in these areas. The main applications of AI in project management were listed as advising, prediction, and classification, and the most critical project area for research was identified as risk management. The study concluded by suggesting that more research is needed in how to integrate AI-based methods into project management. The quantitative

literature review effectively captured a high-level perspective of the areas where AI technology was applied to project management. However, as a structured overview, there was no depth to the findings such as identification of benefits, finding the most practical solutions, or defining challenges to ensure an unbiased process. Alluding to the complexity of construction and IT projects portends the potential for a machine learning approach as AI tools surpass humans in analyzing complexity and large-scale data analysis (Brynjolfsson & McAfee. 2017).

Leu et al. (2023) constructed a supervised learning model using a Bayesian network, a Markov method, and an inference filter to proactively address project cost risks. The study used a mix of six building and mass transit projects in Taiwan to validate the model. The critical variables that influenced cost overruns were project size, duration, design, complexity, and external factors. The model error rate was consistently below 5%. A sensitivity analysis determined that timely intervention could reduce the overrun probability. The study used variables common to other studies, but the creation of a customized approach to developing a model with different algorithmic components may preclude general applicability and acceptance.

Supervised learning with a neural network algorithm was used in conjunction with four earned value management (EVM) metrics to predict increases in project costs (Bakhshi et al., 2022). The study evaluated different algorithms and determined the neural network as superior for the earliest prediction of cost overrun. Ghimire et al. (2023) provided further validation of these results, although their data were localized to a single city and government agency. Pang et al. (2022) studied the prediction of IT project cost overruns for 600 historical projects and determined that a machine learning algorithm with deep learning performed significantly

better than traditional methods and other machine learning models. The mean absolute percentage error (MAPE) for this model was 0.3%. The model was also verified in a real-world project in the telecommunications industry with a project that was 53% complete. The model predicted cost with a mean relative error (MRE) of 0.09, which outperformed traditional earned value management (EVM) metrics by 8.4% (Pang et al., 2022). Further study is required to determine if a similar approach is effective for transportation projects.

Applying machine learning to construction projects produced results more relevant to mass transit projects. Using 45 features from 139 project datasets, a neural network was trained using supervised learning to predict budget overruns in construction (Tajziyehchi et al., 2021). This study verified the ability of a machine learning algorithm to make accurate budget predictions for construction projects within the identified dataset limits. Chandanshive and Kambekar (2019) supported the findings of applying a neural network to accurately predict construction costs. The study created a neural network model based on 78 building projects in Mumbai, using the skeletal costs at input once the project was initiated to successfully predict the final construction cost. Ahiagu-Dagbui and Smith (2014) provided a background for moving from reference class forecasting to a neural network model. They used 1600 water infrastructure projects to create a neural network model using Statistica software that delivered an 87% budget validation score with an error range of ±5%. These studies suggest an opportunity to use a similar approach for mass transit projects.

A machine learning model was used to predict construction project budgets in Thailand (Kusonkhum et al., 2023). The data from 692 projects was collected from the government procurement system. An essential part of the process was feature engineering, which reduced the relevant characteristics to only four features to achieve the correlation (Kusonkhum et al., 2023).

The four categories of attributes were obtained for the mainly road and building projects. The datasets were split using 80% for model training and 20% to test the model accuracy. A Python software algorithm was created using a k-nearest neighbor (KNN) algorithm, which is a supervised learning method. The ability to accurately predict budget overruns was 86%. While the number of projects is large, the small number of data categories could be affected by bias or indicate a single factor determining the results. Feature engineering involves human decision-making to select or transform features for machine learning models and can result in inadvertently introducing biases into the process (Mehrabi et al., 2021).

Meng et al. (2024) studied 94 metro construction projects across China completed from 2002 to 2022 to develop a model to predict project cost. Seven variables were determined as having the most significant impact on cost: the length of the elevated line, the length of the ground-level line, the number of above-ground stations, the average station distance, inflation, infrastructure investment, and the gross domestic product of the region. A machine learning model using Light GBM, developed by Microsoft, obtained the most accurate results at a 13% error rate. A decision tree algorithm had a 20.44% error rate, and a backpropagation neural network had a 19% error rate. The type of machine learning algorithm was responsible for different error rates. The value of this study was in providing insight into the variables related to the project budget for metro construction.

Sandell (2020) examined how to improve cost estimating methods for the Finnish Transport Infrastructure Agency. The traditional method was a fixed percentage value based on the project phase and lacked precision. Multiple regression was compared to a neural network model for forecasting more accurately. The study used 60 projects completed between 2015 and 2018, consisting of road, waterway, and railway projects. For the design

stage, the neural network model had a mean absolute error (MAE) of 1.11, and the root mean square error (RMSE) was 1.57. Multiple linear regression produced similar results at an MAE of 1.19 and an RMSE of 1.59, producing more accurate estimates than the existing method. For the construction stage, multiple regression was more accurate, with an MAE of 0.89 and RMSE of 1.16, while the neural network had an MAE of 1.07 and RMSE of 1.52. Sandell concluded that the neural network was superior because it maintained consistent accuracy across training and test datasets, suggesting a better ability to adapt to different or more complex variables. This study demonstrated that a small sample size can be used to create an accurate neural network model, but it is limited geographically to Finland and does not directly address mass transit projects.

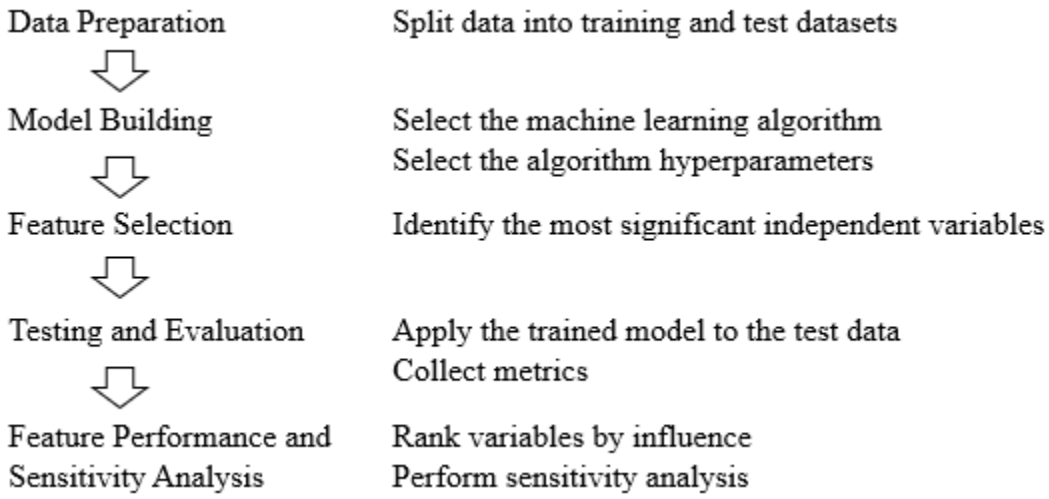
In a study for predicting operating costs for U.S. light rail projects, Zhou et al. (2022) collected 14 variables from 22 projects between 2008 and 2018. Traditional methods had resulted in up to 45% variances between actual and budgeted operating costs. Using supervised learning, the machine learning model calculated the relationship between the independent variables and the actual operating costs. The neural network model achieved an R^2 of 0.970, and multiple linear regression produced an R^2 of 0.887. The machine learning model was configured and evaluated using Minitab software (Zhou et al., 2022). The data were obtained from the National Transit Database (NTD), the U.S. Census Bureau, and the U.S. Bureau of Economic Analysis. While this study focused on transit operating expenses using statistical software for the analysis, many of these variables are also relevant for capital cost prediction.

Uddin et al. (2022) used a structured machine learning approach to enhance project analytics, focusing on construction project cost overruns. Using a publicly available dataset, they

obtained 139 construction projects in India with 44 independent variables and one dependent variable, which was cost overrun frequency. They employed six machine learning algorithm types, including a neural network, to predict the correct classification for projects divided into two categories: a project that had rare cost overruns as the project progressed or frequent cost overruns. The statistical value for a rare or frequent cost overrun was not defined and was already labeled as a categorical variable in the dataset. The neural network was 72.5% accurate in predicting the correct category. The study highlights the advantages of machine learning, such as improved handling of multi-dimensional data and its ability to optimize performance through continuous training. Based on their work, the authors created a framework for similar studies as shown in Figure 2.

Figure 2

Machine Learning-based Framework for Project Analytics



Note. Based on Uddin et al. (2022).

The framework provides a foundation for studies, but fails to provide algorithm details for model building. Their algorithms were created using Python utilities from the software library. Relying on statistical software such as Minitab or SPSS reduces subjective decision-

making as they provide default configurations in areas such as dataset splits and setting hyperparameters.

Table 3 displays the range of datasets and variables used in research for infrastructure project cost overruns. The researchers used available, relevant data and selected a method to achieve insights based on their research focus. No definitive minimum or optimal range of values was identified for the regression analysis.

Table 3

Number of Datasets and Variables Used for Project Management Studies

Academic Article	Projects/ Datasets	Variables	Method	Description
Mæhlen et al., 2024	489	5	Linear Regression	Cost overruns for small projects
Meng & Duan, 2024	94	7	Machine learning (GBM)	In progress transit construction cost estimation
Kusonkhum et al., 2023	692	4	Machine learning: KNN	Predict construction budgets
Capone & Narbaev, 2022	20	25	Machine learning: K-means	Risk grouping for contingency determination
Uddin et al., 2022	139	44	Machine learning: Neural network	Construction cost overruns
Zhou et al.,2022	22	14	Machine learning: Neural network	Predict mass transit operating costs
Tajziyehchi et al., 2021	139	45	Machine learning: Neural network	Predict construction budget overruns
Dao et al., 2020	44	27	Logistic regression	Assess project complexity
Chandanshive & Kambekar, 2019	78	11	Machine learning: Neural network	Predict construction costs

Academic Article	Projects/ Datasets	Variables	Method	Description
Ahiaga-Dagbui & Smith, 2014 Government Report	1600	6	Machine learning: Neural network	Predict causes of cost growth in projects
Sandell, 2020 Student Thesis	60	8	Machine learning: Neural network	Transportation project's indirect cost prediction
Sjögren & Norgren, 2023	114	3	Linear Regression	Infrastructure cost overruns

Talaei Khoei and Kaabouch (2023) compared supervised and unsupervised machine learning models in cybersecurity. Supervised learning models were more effective due to accuracy and efficiency. Unsupervised models were less reliable for this application but tended to be more flexible and used where labeled datasets were scarce. According to Sarker (2021), supervised learning is generally more prevalent in business applications because of its direct applicability to solving well-defined problems using labeled datasets. However, unsupervised learning provides a different perspective, and based on a study of AI adoption in project processes between 2017 and 2021, Noteboom et al. (2021) called for further exploration of unsupervised learning.

Classification was used in a study that successfully estimated the contingency budget required for project risks (Capone & Narbaev, 2022). Using k-means clustering to perform unsupervised learning, Capone and Narbaev (2022) grouped risks based on similarity as a method to determine a more accurate risk contingency amount for the project budget. The traditional expected monetary value (EMV) method to determine the contingency amount had a 43.94% probability of covering project risks, compared to the k-means classification, which had

a 90.97% probability. They used 20 risk registers containing 25 risks each. The researchers acknowledged the limitation of using random risk registers that cannot be generalized to all project types. The study highlighted the potential for machine learning as a more effective method for determining a contingency budget based on project risks. Mariani et al. (2023) used a partitioning method for unsupervised learning as an objective way to improve the classification of stakeholders compared to traditional methods, which were viewed as subjective. The study collected 18 variables from 124 stakeholders for an Italian IT project. The results classified stakeholders into three clusters, providing a more nuanced output than the conventional power-interest matrix used in project management (PMI, 2021). The contribution of this study is the effectiveness of using 18 variables to determine more objective results using a machine learning classification algorithm.

The k-means algorithm is an unsupervised learning technique that groups data into a distinct number of clusters (k) by minimizing intra-cluster variance, making it particularly suitable for segmenting project datasets into meaningful groups (Topaloglu, 2024). Its utility lies in the ability to reveal underlying patterns within the data, which is critical for analyzing different outcomes across projects. Pospieszny (2014) highlighted the need for advanced data mining techniques in project management to address the shortcomings of traditional estimation methods, such as expert judgment and parametric approaches. He emphasized the importance of using clean and preprocessed historical project data, including performance metrics like budgets, incurred costs, timelines, and resource utilization. Topaloglu (2024) applied k-means to a dataset with 10 critical risk assessment features, such as labor loss, risk scores, and hazard frequencies, across 26 sectors. The clustering process involved initializing the number of clusters and iteratively updating the centroids until the clusters stabilized. He noted that validation was

essential to assess the quality of clustering, often relying on statistical measures such as the Sum of Squared Errors (SSE) to ensure intra-cluster compactness and cohesion. Lower SSE values indicate better-defined clusters. The study illustrated the versatility of k-means in handling multivariate data and extracting meaningful clusters that reflect relationships among features. Topaloglu's hybrid model combined k-means and a Support Vector Machine (SVM) algorithm, to successfully divide datasets into subgroups. This clustering step allowed for subsequent classification tasks to yield a high accuracy rate of 99.6%.

Writing customized Python code or adding a sequence of components to create a machine learning algorithm to analyze a project issue limits the ability of other researchers to replicate the results unless the code, including hyperparameters and other software components, is openly available. An alternative is to apply statistical software such as MATLAB, Minitab, or SPSS to produce results. Statistical software programs request configuration options, which are more easily shared so researchers can obtain comparable results. There is also a possibility that custom Python code inherits more bias compared to statistical software.

Additional Algorithm Types

Machine learning algorithms using supervised or unsupervised learning may help validate a project budget, although other algorithms should be considered. Reinforcement learning algorithms use trial and error in the decision-making process. Datasets are captured and used to improve decision-making by avoiding previous poor decisions (Prigent et al., 2022). This requires the acquisition of datasets that reflect the project conditions. The algorithm assesses a historical result and rejects it if it does not provide the correct outcome. This type of algorithm is similar to reference class forecasting, where the reference class that most closely matches the project is selected. Jha et al. (2021) reviewed supervised learning, unsupervised learning, and

reinforcement learning, highlighting the strengths of each machine learning method. They concluded that no single algorithm produced an optimal result for all purposes. The effectiveness of a machine learning method is heavily dependent on the data, the objective of the study, and computational constraints.

Genetic algorithms are a form of machine learning based on simulating the theory of evolution (Wirsansky, 2020). When using a genetic algorithm, the result or outcome is known, and the algorithm iterates to find the best combination of features to achieve the result. Hashemi et al. (2019) integrated a neural network with a genetic algorithm to predict the accuracy of power plant budgets across Iran. Historical data from the plants was used to create the model, and the genetic algorithm was used to optimize the hyperparameters. The best model achieved an accuracy of 94.71% for predicting project costs based on the mean squared error (MSE). A sensitivity analysis revealed that the most influential factor was the type of power plant project, while the least significant factor was substation construction (Hashemi et al., 2019).

A software model using a genetic algorithm is faster at assessing risks than fuzzy logic or a logistic regression model (Chen et al., 2022). Genetic algorithms have been investigated for cost optimization where the cost is known, and the algorithm determines the optimal structural design decisions (Martins et al., 2023). In another study, Deb and Gupta (2023) identified an optimal allocation solution to minimize shipping costs. Genetic algorithms are also effective at determining a resilient schedule based on the required project end date (Milat et al., 2022). For a genetic algorithm to be successful, the dependent variable must be a known value, and the algorithm generates all possible combinations of features to achieve a correlation that best matches the known outcome. This type of machine learning is useful when the dependent variable is the same for each project.

While various machine learning algorithms exist, supervised and unsupervised learning methods are the most suitable for the study of mass transit budget overruns due to their ability to analyze historical datasets. Reinforcement learning relies on training to make decisions that maximize a reward (Prigent et al., 2022). Genetic algorithms are effective for optimizing known outcomes but are less useful when dependent variables already align with project results (Hashemi et al., 2019; Martins et al., 2023). Based on the research studies described, supervised and unsupervised learning methods offer the best balance of predictive power and adaptability for analyzing project budget performance.

Using SPSS for Machine Learning Research

Applying machine learning to improve project performance requires diligence in the face of a deficiency of uniform standards. The absence of standards for applying machine learning in project management promotes flexibility, but from a research perspective, it may impede the ability to replicate or enhance research findings. For supervised learning, a neural network can be created in software code using Python programming language, but a more structured approach is to use statistical software like SPSS to create a model and generate predictions. A process using SPSS capability for prediction and classification has occurred in other areas of research, and SPSS has emerged as a practical tool for applying machine learning in project management, offering structured and reproducible processes for supervised and unsupervised learning (Çevik & Tabaru-Örnek, 2020; Topaloglu, 2024).

Çevik, M., & Tabaru-Örnek (2020) used the neural network functionality in SPSS to predict academic achievement for 465 primary school students. The ability to accurately predict the success of students using SPSS was 81.5%. SPSS also correctly identified the contributing factors. Another study using SPSS demonstrated superior overall prediction accuracy of a neural

network, even as the datasets became more imbalanced (Harb & Jayousi, 2012). An imbalance in supervised learning means there is a disparity in the target variables for the datasets (Harb & Jayousi, 2012). In an article that compares a neural network and SPSS regression models for predicting customer churn in Iran's insurance industry, Pirmohammadi and Mast Zohouri (2020) achieved better prediction accuracy ($R = 0.7021$) with a neural network than using SPSS regression analysis ($R^2 = 0.286$). The study was performed using 138 customer datasets and 16 input features with customer status as the target variable.

Zhang and Xia (2017) considered SPSS an essential tool for evaluating prediction reliability and interpreting the statistical significance of forecasting results for short-term power loading. SPSS was particularly useful in the research for its statistical capabilities, especially in validating the model through error analysis and hypothesis testing. Table 4 shows records and features used in SPSS studies using a neural network and are not project-related.

Table 4

Data Details for Studies Using Neural Network Module

Study using SPSS	Records	Features	Rationale
Çevik & Tabaru-Örnek, 2020	465	10	Focused on educational and socio-economic predictors of student success
Pirmohammadi & Mast Zohouri (2020)	138	16	Captured variables influencing churn in the insurance industry
Harb & Jayousi, 2012	38,000	10	Captured socio-demographic variables for labor market predictions
Zhang & Xia, 2017	Not provided	720	Short-term power load predictions

K-means clustering in SPSS has proven to be a reliable tool for identifying homogeneous groups within datasets, making it particularly valuable for tasks such as classifying projects

(Ahamad & Bharti, 2021; Dauley et al., 2024; Mulyono et al., 2023). The k-means algorithm works by partitioning data into predefined clusters based on similarity measures, typically using Euclidean distances. This method ensures that objects within a cluster are as similar as possible while maximizing the differences between clusters. The adaptability of k-means to handle both continuous and categorical data has been demonstrated in various studies, making it a practical approach to project classification. These characteristics make k-means well-suited for analyzing projects with attributes such as cost overruns, project durations, and contractor selection.

K-means clustering in SPSS was used to analyze customer behavior at a shopping mall (Mulyono et al., 2023). The analysis segmented customers into three distinct clusters based on behavioral and demographic attributes, such as spending habits and social media usage. Similarly, Daulay et al. (2024) used k-means clustering to group coffee shop customers into three segments, successfully identifying target clusters based on preferences and purchasing behaviors. These studies emphasize how k-means can generate meaningful insights by uncovering hidden patterns within datasets.

A variety of datasets and variables can be used when applying k-means clustering in SPSS. For example, Daulay et al. (2024) worked with survey data that included seven attributes such as age, occupation, income levels, and purchasing habits to identify key customer groups. Ahamad and Bharti (2021) utilized datasets with 14 attributes. These examples highlight the flexibility of k-means to work with diverse and mixed datasets, which is highly relevant for project classification. By clustering projects based on identifiable factors, researchers can identify groups that share common budget performance traits.

Research into applying machine learning to improve project performance requires a structured approach. The variety of possible research designs poses a concern for reproducibility

and validation. Software like SPSS offers a statistical foundation and transparent process for supervised and unsupervised learning, enabling researchers to develop predictive models and validate results. Studies demonstrated the ability of SPSS's neural network and k-means capability to produce statistically significant results, making it a sensible method for project management research.

Addressing Bias in Machine Learning

Bias in machine learning is a critical concern that extends beyond technical errors to encompass broader ethical and societal implications. As machine learning systems are conceived, designed, and implemented by humans, they inevitably inherit biases embedded in the data and decision-making processes (Mehrabi et al., 2021). In the book *Weapons of Math Destruction*, O'Neil (2016) described how biased data and algorithms perpetuate and amplify social issues such as discrimination. This underscores the imperative for transparency and accountability in AI applications, especially in research contexts like machine learning for project management. A lack of transparency in how algorithms are configured or data is collected and analyzed can compromise research integrity, hinder reproducibility, and lead to flawed applications (Mehrabi et al., 2021).

There is a distinction between data privacy or security and AI bias. Data privacy and security focus on protecting sensitive information from unauthorized access, breaches, and misuse, ensuring that personal data is kept confidential and safe from malicious attacks (Data Privacy and Security, n.d.). This involves implementing measures such as encryption, secure data storage, and robust authentication protocols to safeguard data integrity and prevent exploitation (Kaur et al., 2023). Bias in a machine learning process pertains to systematic errors or unfair

outcomes resulting from unbalanced or prejudiced data, flawed algorithm designs, or societal inequities reflected in data (Mehrabi et al., 2021).

Various sources and types of bias in machine learning processes exist, including data, algorithmic, and user-induced biases (Mehrabi et al., 2021). Based on survey results, Mehrabi et al., (2021) concluded that a method to ensure fairness in AI is required. They defined fairness as demographic parity, which means equal positive outcomes across groups, equalized odds, equal true positive and false positive rates, and prediction consistency that matches actual outcomes. For societal implications, metrics such as measuring the sample diversity levels, checking error rate parity, and verifying probabilities are equally accurate across different groups are used to measure fairness and address biases in data and algorithms.

A study by Rana et al. (2023) presented a framework for identifying AI bias during software development and data acquisition. The authors categorized biases into pre-existing, technical, and emergent forms and proposed strategies for mitigating each type using conceptual, empirical, and technical approaches. The study emphasized the critical need for fairness management, advocating fairness sampling and training. Bias needs to be addressed so mitigation plans can be created. Dimitrakopoulou et al. (2024) reviewed sources, types, and methods to address bias in a machine learning process. Bias can emerge from data collection, algorithm design, model training, or deployment, posing significant risks in critical areas like healthcare and finance. Detection and correction strategies vary and may require a trade-off between fairness and accuracy. They described seven methods, as identified in Table 5, to detect bias in a machine learning system and identified methods to correct or minimize potential bias.

Table 5

Detecting Bias in Machine Learning

Detection Method	Description
Data analysis	Identify imbalances, outliers, or anomalies that may indicate bias. Check for underrepresented features.
Data visualization	Use charts and graphs to reveal patterns or trends that suggest bias.
Data annotation	Add headings and labels such as each category's data sources and quality level.
Algorithm analysis	Review logic, processes, and constraints.
Algorithm evaluation	Test the algorithm for accuracy, robustness, and generalization.
Algorithm feedback	Request input from experts and users to assess the algorithm's accuracy and potential bias.
Bias auditing	Test for bias and fairness in model outcomes.

Note. Based on Dimitrakopoulou et al., 2024.

Using SPSS or other statistical software may avoid algorithm design bias issues because a standard algorithm for machine learning is already incorporated. Configuration decisions and the overall process are still susceptible areas to mitigate bias. The actions in Table 6 are addressed in the research design.

Table 6

Correcting Bias in Machine Learning

Correction Method	Action
Data preprocessing	Clean, transform, and augment data. Ensure balanced data and datasets.
Fair algorithm design	Incorporate explicit fairness.
Model training adjustments	Improve algorithm training to reduce overfitting.
Model evaluation and fairness	Measure model fairness using metrics.
Transparency and explainability	Confirm that algorithms and processes are understandable and transparent by explaining how they work, the data used, and the outcome.

Note. Based on Dimitrakopoulou et al., 2024.

When using machine learning for project budget validation, bias in data or algorithms could skew predictions, leading to misleading results and, in a practical application, a misaligned capital planning strategy in infrastructure projects. A foundational understanding of fairness principles and practical mitigation strategies were incorporated into the research design. Fairness checks were considered in model configuration for the neural network and classification process by elucidating potential variation of results based on different configuration options. A similar approach was taken to critically scrutinize the processes for data management and analysis interpretation.

Summary

This literature review critically examined project budgeting practices, cognitive biases, and machine learning, focusing on infrastructure projects, especially mass transit systems. It explored the limitations of traditional budgeting methods, the unique challenges associated with mass transit projects, and the potential of machine learning to improve budget validation. Cognitive biases, such as optimism bias and strategic misrepresentation, significantly distort budget accuracy. The research highlighted the pervasive impact of these biases and emphasized the need for objective, data-driven approaches like machine learning to mitigate their effects.

Mass transit projects in the United States frequently experience budget overruns. While established project management methodologies, such as PMBOK and PRINCE2, provide approaches to improve forecasting accuracy, significant challenges persist in achieving reliable budget predictions. Existing research underscored the importance of leveraging historical data and objective methods. However, the application of machine learning for budget validation remains largely underexplored. Machine learning techniques, including neural networks and k-means clustering, have shown superior accuracy compared to traditional methods in related infrastructure and construction projects.

SPSS has emerged as a valuable tool for researchers performing prediction and classification tasks. Its robust statistical capabilities enable clear interpretation of results, identification of contributing factors, and validation of models through error analysis. Studies demonstrated that supervised learning, such as neural networks in SPSS, delivered accurate predictions and insights into key variables influencing outcomes. Additionally, unsupervised learning in SPSS provided an ideal method for exploratory analysis by uncovering hidden patterns within datasets and enabling meaningful segmentation, such as grouping projects based on shared attributes.

The most significant gaps identified in the literature are the limited application of machine learning techniques specifically dedicated to mass transit project budgeting and the lack of a standardized framework for integrating machine learning into budget validation processes. Addressing these gaps is essential for advancing research focused on improving project performance, particularly in enhancing the accuracy and reliability of project budgeting practices.

Chapter 3: Research Method

The problem addressed in this study was the underestimation of budgets for mass transit systems in the United States. The purpose of this quantitative correlational research was to determine the effectiveness of machine learning algorithms using supervised and unsupervised learning to validate a budget for U.S. mass transit projects. The research focused on machine learning as an objective method to validate a project budget. The growing trend of urbanization has driven a surge in demand for mass transit systems, which are implemented using project management processes. For U.S. mass transit projects, the allocation strategy for federal funding relies on precise budgets and long-term capital planning. However, there is a high rate of budget overspending on these projects (Gao & Touran, 2020). Artificial intelligence (AI), particularly machine learning, offers an innovative approach to improving budget accuracy and predicting overspending by leveraging historical data (Akanni, 2024).

This chapter provides an overview of the research methodology that was employed to investigate the application of machine learning algorithms in SPSS software to validate the accuracy of mass transit project budgets. The conceptual framework for the study is cognitive bias, which was used to address the aspects of data collection, configuration of the system, the analysis process, and interpreting the statistical outcome. The two areas of potential bias that were addressed are bias in the data and researcher bias (Ntoutsi et al., 2020). Data acquisition methods are described, detailing how project characteristics and key variables related to budget outcomes were identified, measured, and prepared for analysis. The process for employing supervised learning using a neural network and unsupervised learning using k-means is explained, emphasizing the suitability for predicting budget variances. The statistical procedures used to assess the effectiveness of each method are described, and an explanation of the selected

statistical outputs is presented to demonstrate how the analysis was performed. These components establish a foundation for evaluating the research questions and contribute to advancing the application of AI in addressing bias-related challenges in mass transit project budgeting.

Research Methodology and Design

Research was performed to determine the effectiveness of machine learning algorithms in predicting budget variances. There were three research questions. The first was to determine how accurately a machine learning algorithm using a supervised learning model can predict the budget for mass transit projects in the United States. The second research question was to determine how accurately a machine learning algorithm using unsupervised learning can predict the budget for mass transit projects in the United States. The third question was to evaluate if there is a statistically significant difference in the predicted budget between supervised learning and unsupervised learning algorithms for mass transit projects in the United States.

The approach for this study was a quantitative design using publicly available historical data, with the most critical data collected from U.S. government websites. The study employed SPSS software to perform analysis using preconfigured machine learning algorithms. A quantitative research method was a good fit for this study as the objective is the development of mathematical models to predict project budget overruns (Gajera, 2024; Tabassi & Bakar, 2018). A qualitative approach was considered less effective as subjective judgments on the reasons for budget overruns can be varied based on participant experience and personal bias (Flyvbjerg, 2022).

As noted in the literature review, there is encouragement for studies that employ innovative methods to solve project problems (Pinto, 2022). A data-driven approach using

machine learning algorithms is an opportunity to present new research methods in this area. Investigating budget validation from a social perspective by analyzing stakeholders and project communication is a traditional method that is open to bias in the data collected and by the researcher designing the study (McSweeney, 2021). It may be possible to optimize traditional methods, such as parametric estimating or reference class forecasting, by applying other statistical analyses. However, a different approach based on creating machine learning models using supervised and unsupervised learning to improve project performance supports a more innovative approach (Pinto, 2022).

Population

The population selected for this study was all metropolitan-based mass transit rail systems projects in the United States completed up between 1986 and 2024. The data were collected from reports on publicly available websites, with the main focus on the Federal Transit Administration (FTA). Due to the historical overspending of mass transit capital budgets, the FTA increased its role in technical assistance and oversight of these projects (Gao & Touran, 2020). The FTA data is considered reliable, as evidenced by the following statement.

> The data submitted to the Federal Transit Administration (FTA) annually by transit agencies are subjected to analysis and validation, both manual and automated. The process entails a detailed examination of each transit system's report, the identification of invalid entries based on the defined data types that appear in the database files, and the direct resolution of these problems in conjunction with the reporting transit system. FTA's role in this process is to identify and resolve questions of data completeness and accuracy. The CEO of each agency certifies the accuracy of the data contained in the report. FTA may reject a transit agency's report if this report is not in full compliance

> with reporting requirements, including decennial auditor's statements for financial data or annual statements for Federal Funding Allocation data. FTA employs a closeout process to ensure that the agency has addressed all data validation and that all validation measures have been met uniformly. (National Transit Database, 2024, p. 3)

The FTA provides reports to the U.S. Congress and is subject to reviews from the U.S. Government Accountability Office (GAO). The FTA instituted a risk assessment process in 2003, which required more structure for reporting project variables (Gao & Touran, 2020). For projects completed before 2003, it was more challenging to acquire data for all variables from the FTA (Voulgaris, 2017). In the study by Gao and Touran (2020), 81 project datasets were initially identified but reduced to 68 due to missing data. For this research, additional data were captured from the GAO, the U.S. Bureau of Economic Analysis (BEA), the U.S. Census Bureau, and the National Weather Service. Where FTA data were incomplete, data were captured from local government transit authority websites, the American Public Transit Association (APTA), and contractor websites publicizing the project. The multiple sources allowed triangulation of data (see Appendix A).

The variables collected were based on the unbiased nature of the project characteristics. Most variables are numeric values, such as the number of vehicles purchased for a project, or derived from simple equations, such as calculating a temperature range from minimum and maximum temperatures. Categorical variables collected were immutable text fields, such as a vehicle brand, that were encoded as nominal data using SPSS. As discussed earlier, including variables such as project complexity and size requires a determination that employs subjectivity. This study focused on variables that were common to similar studies on mass transit project costs, and the range of values in the variables objectively reflects project size and complexity.

Instrumentation

The data were extracted and manually entered into the data instrument, which is an Excel spreadsheet (see Appendix B). The spreadsheet was imported to SPSS, where variables were validated and converted to SPSS data types. Descriptive statistics and sample checking were performed to verify data were correctly entered. Poor preprocessing can lead to inaccurate analytics, whereas systematic preprocessing ensures robust, reliable insights (Joshi & Patel, 2020). By improving data quality, preprocessing minimizes errors and enhances the predictive power of machine learning models in validating and forecasting budget outcomes (Aggarwal, 2015). Data wrangling was used to collect, organize, and validate the input data into a format for machine learning algorithms (Kim & Lee, 2020). This process is essential for any data-driven study, as the quality of the raw data often dictates the reliability of the insights derived from the study (Aggarwal, 2015).

Measures were taken to avoid bias during data wrangling. The majority of the variables were raw numeric data. Nominal data were encoded for input to SPSS. Data that has a wide range may unduly influence regression analysis (Field, 2018) and were transformed to z-scores to allow for direct comparison with all variables despite their different units and magnitudes. Standardizing is important because it ensures variables contribute equally to the analysis rather than disproportionately influencing results due to differences in scale (Field, 2018). Software default configurations were used to standardize the analysis, and ANOVA results were used to verify model fit. The sample data were reviewed to ensure it represented the study focus, missing data were addressed transparently, and caution was exercised to avoid arbitrary removal of data points such as outliers. The emphasis on unbiased data management provided integrity for the study outcomes.

Operational Definition of Variables

Operationalizing the variables explains the definition and quantification of the variables, why the data fields are collected, and the data acquisition process. The study focused on project characteristics that are common to similar studies on mass transit project costs. The project data consisted of internal and external dependencies (PMI, 2021), as identified in Table 7. Internal factors were part of the project, including duration and physical items procured or built as part of the project scope. External factors refer to the physical environment, such as location or climate conditions, which may influence the ability to deliver the project (PMI, 2021). For machine learning, these project characteristics were considered features for analysis.

Table 7

Internal and External Project Variables

Ratio and Interval (Scale in SPSS)	Nominal (Nominal in SPSS)
Internal to the Project	
Project completion year	Vehicle Brand
Project duration	Initial or extension of the system
Track distance	
Underground track distance	
Number of stations	
Number of underground stations	
Number of vehicles purchased	
Number of maintenance facilities	
External to the Project (Environmental)	
Metropolitan density	Region
Minimum temperature	
Maximum temperature	
Temperature range	
State GDP	

Project Actual Cost, Inflation-adjusted (Dependent Variable). The inflation-adjusted actual cost is the project's actual cost in 2024 adjusted for inflation based on the year of project completion. The inflation-adjusted actual cost is a numeric variable expressed in U.S. dollars. In this study, the inflation-adjusted actual cost is the dependent variable and was correlated to the supervised and unsupervised learning predicted actual cost.

SPSS Neural Network Predicted Value (Independent variable). For supervised learning, the predicted value is the neural network model's calculation based on the target variables and derived from the learned relationships during training (IBM, 2024). The algorithm produced a predicted actual project cost for each dataset. As an independent variable, the predicted value was used to determine if there was a correlation to the actual project cost.

SPSS K-means Cluster Mean Value (Independent variable). For unsupervised learning, the mean values of the clusters were an independent variable. The cluster mean is a numeric value derived from calculating the mean of the actual project cost for each defined cluster. As an independent variable, the cluster mean was used to determine if there was a correlation to the actual project cost for each dataset.

Project Name. A project name was used to identify each dataset in the collection instrument. Mass transit projects can acquire different names over time, such as the original project name, a short form name, or a name provided by the metropolitan transit authority when the system is deployed. For this study, the project name was the one used by the FTA.

Project Budget. The project budget is the approved funds to complete the project work (Electronic Code of Federal Regulations, n.d.). The project budget is a numeric variable expressed in U.S. dollars and represents the value of the allocated budget at project

completion. The level of measurement is ratio, and the range depends on the project, but mass transit budgets typically range from millions to billions of dollars. The original project budget for mass transit projects is essential for research into project budget variances (Ashiaga-Dagbui & Smith, 2014; Elkind et al.,2022; Gao & Touran, 2020). The data were sourced from the FTA (FTA, n.d.).

Project Actual Cost. The project actual cost is the financial spending required to complete the project work (PMI, 2021). The project actual cost is a numeric value expressed in U.S. dollars. It represents the Year of Expenditure (YOE) value, reflecting the actual total cost in the year it is reported. The level of measurement is ratio, and the score range depends on the project scope, but mass transit budgets typically range from millions to billions of dollars. The original project actual cost for mass transit projects is essential for research studies into project budget overspending (Ashiaga-Dagbui & Smith, 2014; Elkind et al.,2022; Gao & Touran, 2020). The data were sourced from the FTA.

Project Start Year. The project start year is the year the first recorded financial allocation occurred for the project (FTA, n.d.). The project start year is an interval variable, and the range is based on projects completed before 2024. The project start year reflects the influence of temporal and economic factors, such as inflation, market conditions, and technological advancements, on project budgeting (PMI, 2021; Van Wee, 2007). Projects initiated in different years may experience varying degrees of budget accuracy due to shifting industry standards, regulatory environments, or resource availability. Incorporating the start year may allow for time-dependent factors, providing a more nuanced understanding of budget performance (Gao & Touran, 2022). The data were sourced from the FTA.

Project Completion Year. The project completion year is the year the transportation project was first opened to the public (FTA. n.d.). The project completion year is an interval variable, and the range for project completion year is from the first mass transit project completion in the 1980s to 2024. The project completion year may capture the impact of evolving economic conditions, technological changes, and policy shifts during a project's lifecycle (PMI, 2021; Van Wee, 2007). Projects completed in different years may face varying cost pressures, such as material price fluctuations, labor market dynamics, or changes in regulatory requirements. Including the completion year helps account for these temporal influences, offering insights into how external affect budget outcomes (Gao & Touran, 2020; Sjögren & Norgren, 2023). The source was the FTA.

Project Duration. Project duration was operationalized as the time elapsed from the project start year to project completion. The project duration is a ratio variable and was represented in full-year increments without allowance for partial years. Project duration is a commonly used factor in project studies (Ashiaga-Dagbui & Smith,2014; Gao & Touran, 2020; Siemiatycki, 2010). The project duration was calculated as the difference between the project completion year and the project start year or as reported in FTA documents.

Project Type. The project type is nominal data and was classified into two categories. The first mass transit project in a metropolitan area was a assigned 1, indicating a project to implement an initial or new mass transit project (FTA, n.d.). An extension project was assigned 0, indicating a project that adds to an existing mass transit system.

Project type affects mass transit project budgets because it may significantly impact cost structures and risks. The first deployment of mass transit systems typically involves higher initial costs for land acquisition, infrastructure development, and new systems. In

contrast, extensions or additions leverage existing assets and experience but may face integration challenges (Eno Center for Transportation, 2021). Differentiating between these project types can have an impact on the analysis of budget performance (Gao & Touran, 2020). The data were collected from the FTA.

Track Distance. Track distance is defined as the length of a physical transportation route measured in miles along its alignment and reported in track miles (FTA, n.d.). Track distance was measured in miles and is a ratio variable. For double or multi-track systems, the measurement applies to the distance of one track from the origin point to the end. Track distance reflects the scale and physical scope of a transit project, significantly impacting costs, materials, labor, and construction timelines. Measuring track distance enables analysis of its relationship with budget variances (Zhou et al. (2022). The source was the FTA and physical descriptions from local metropolitan transit authorities.

Underground Track Distance. Underground track distance is the length of the route below grade. Underground track distance was measured in miles, is a ratio variable, and is a subset of the total track distance. This variable reflects the physical scope and may significantly impact costs, materials, labor, and construction timelines (International Tunnelling Association Working Group 13, 2003). The data were sourced from the FTA or, if not documented, sourced from the local metropolitan transit authority.

Number of Stations. The number of new transit stations refers to facilities to serve passengers, constructed as part of the project and not previously existing before the project (FTA, n.d.). The number of stations was quantified as the total count of stations completed and operational by the end of the project. The number of stations is a ratio variable. The

number of stations is a tangible project requirement and a factor that has an impact on project budgets and expenses (Zhou et al., 2022). The data were sourced from the FTA.

Number of Underground Stations. Stations created below grade require tunneling for access and imply increased complexity due to construction methods (FTA, n.d.). The number of underground stations was the total count of stations completed and operational by the end of the project. The number of underground stations is a ratio variable and is a subset of the total number of stations. This tangible project requirement requires increased costs compared to ground-level stations based on susceptibility to seismic or sinkhole risk (International Tunnelling Association Working Group 13, 2003). The data were sourced from the FTA or, if not documented, sourced from the local metropolitan transit authority.

Vehicles Purchased. Vehicles purchased is the number of passenger-carrying vehicles purchased for the transportation project (FTA, n.d.). It is the count of transit vehicles acquired during the project explicitly as part of the project scope. Each vehicle was counted once, irrespective of size, type, or capacity. The number of vehicles purchased is a ratio variable. The number of vehicles is a tangible project requirement and a factor that has an impact on project budgets and expenses (Zhou et al., 2022). The data were sourced from the FTA.

Vehicle Brand. Vehicle brand is defined by the FTA as the name of the transit vehicle manufacturer responsible for producing the vehicles purchased by the project. The vehicle brand is a nominal variable. Brands that only have one entry were categorized as “other.” Separate categories for singular brands can compromise the reliability of statistical estimates by increasing the risk of overfitting in predictive models (Agresti, 2018). The model may capture random noise instead of meaningful trends. Having excess categories with a single entry can reduce the clarity of results, especially when the focus is on identifying overall patterns rather

than examining individual outliers (Agresti, 2018). Vehicle brand provides insights into procurement preferences and potential quality or performance differences across projects, which may have an impact on budgets (Amron, 2018). Vehicle brand is an objective, incontrovertible data field. The vehicle brand was acquired from the American Public Transportation Association (APTA) and verified by data from the local metropolitan transit authority.

Number of New Maintenance Facilities. Maintenance facilities are those constructed as part of the project requirements and did not exist before the project. A maintenance facility is where mechanics, machinists, and other maintenance personnel perform preventative maintenance, daily service and inspection, and other maintenance activities (FTA, n.d.). The number of new maintenance facilities was quantified as the total number of facilities completed in the project and is a ratio variable. The number of maintenance facilities is a tangible project requirement and a factor that has an impact on project budgets and expenses (Zhou et al., 2022). The data were sourced from the FTA.

Economic Region. The economic region is a physical location that was measured as a categorical variable, with each region assigned a unique identifier corresponding to the FTA's 10 regional designations (FTA, n.d.). The classification was based on the project's physical location and corresponds to the location of the Federal Transit Authority office. Each economic region is recorded as one of the 10 predefined FTA categories.

1. Region 1: Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont.
2. Region 2: New York, New Jersey, U.S. Virgin Islands.
3. Region 3: Delaware, Maryland, Pennsylvania, Virginia, West Virginia, Washington, D.C.

4. Region 4: Alabama, Florida, Georgia, Kentucky, Mississippi, North Carolina, South Carolina, Tennessee, and Puerto Rico.

5. Region 5: Illinois, Indiana, Michigan, Minnesota, Ohio, Wisconsin.

6. Region 6: Arkansas, Louisiana, New Mexico, Oklahoma, Texas.

7. Region 7: Iowa, Kansas, Missouri, Nebraska.

8. Region 8: Colorado, Montana, North Dakota, South Dakota, Utah, Wyoming.

9. Region 9: Arizona, California, Hawaii, Nevada, Guam, American Samoa, Northern Mariana Islands.

10. Region 10: Alaska, Idaho, Oregon, Washington.

The regions are nominal data. There is a fixed number of 10 categories corresponding to the defined regions 1 through 10. Economic regions may identify geographic patterns or disparities in transit project budgets, performance, and funding efficiencies. The FTA's standardized regional definitions ensure consistency and comparability across project locations, enhancing the validity of insights derived from the data. Economic region is a field used by other researchers for mass transit projects (Ashiaga-Dagbui & Smith, 2014; Elkind et al., 2022; Siemiatycki,2010; Sjögren & Norgren, 2023).

Metro Density. Metropolitan density is measured as the total population of the metropolitan area divided by its total land area, expressed as people per square mile (U.S. Census Bureau, n.d.). The variable is a numeric value and a ratio variable. The value of the variable depends on the size and urbanization of the metro area. Current year density was used. This provides a relative measure, and the highest growth for any major city has been 2.5 % over the past 5 years (U.S. Census Bureau, n.d.).

Metro density influences the scale, design, and cost of transit infrastructure required to meet demand. High-density areas may require more compact, complex, and expensive systems, such as underground tunnels or elevated tracks, but benefit from higher revenue potential. Low-density areas may need more extensive networks with higher per-user costs, impacting land acquisition, operating efficiency, and funding allocation (Zhou et al., 2022). The source of the data were from the U.S. Census Bureau. Metro density variables were standardized by transforming the variables to z-scores to avoid disproportionately influencing results due to differences in scale (Field, 2018).

Minimum Temperature at Location. The minimum temperature is the average lowest recorded daily temperature at the deployment location of the mass transit system over the past 5 years, as reported by the U.S. Weather Bureau. The variable is a numeric value represented in degrees Fahrenheit. Temperature is interval data. Climate and weather conditions, such as temperature, may have an impact on construction practices, materials, and costs, particularly in regions requiring cold-weather adaptations for transit infrastructure. Including this variable allows for the analysis of environmental factors that may influence project budget overruns (Amadi, A. I., 2022; Marzoughi et al., 2018; PMI, 2021). The data were sourced from the U.S. Department of Commerce, National Oceanic and Atmospheric Administration (NOAA), National Center for Environmental Information (NCEI).

Maximum Temperature at Location. The maximum temperature is the highest average recorded daily temperature at the deployment location of the mass transit system over the past 5 years, as reported by the U.S. Weather Bureau. The variable is a numeric variable represented in degrees Fahrenheit. Temperature is interval data. Climate and weather conditions, such as temperature, may have an impact on construction practices, materials, and

costs, particularly in regions requiring hot-weather adaptations for transit infrastructure. Including this variable allows for the analysis of environmental factors that may influence project budget overruns (Amadi, A. I., 2022; Marzoughi et al., 2018; PMI, 2021). The data were sourced from the U.S. Department of Commerce, National Oceanic and Atmospheric Administration (NOAA), National Center for Environmental Information (NCEI).

Temperature Range at Location. The temperature range is derived from the minimum and maximum temperature values. The range is a numeric value represented in degrees Fahrenheit. Temperature is interval data. As previously described, climate and weather conditions over a period of time may have an impact on construction practices, materials, and costs. Including this variable allows for the analysis of environmental factors that may influence project budgets (Amadi, 2022; Marzoughi et al., 2018; PMI, 2021).

State GDP. State GDP is measured as the annual Gross Domestic Product (GDP) of the state where the project is located, reported in billions or trillions of U.S. dollars (U.S. Bureau of Economics, n.d.). The State GDP was collected as of the latest year available. The variable is a ratio variable. State GDP reflects the economic capacity of a state to fund, implement, and support transit projects. Including this variable may provide an analysis of how economic strength correlates with the likelihood of budget overruns (Zhou et al., 2022). The data were sourced from the U.S. Bureau of Economics. State GDP variables were standardized by transforming the variables to z-scores to avoid disproportionately influencing results due to differences in scale (Field, 2018).

Study Procedures

The data for this study were collected from publicly available sources, beginning with the FTA and GAO websites. The focus was on obtaining data to complete the data fields in the study

instrument. If a specific data item was reported in several sources, the FTA-reported data were used as the source. The collection process followed a systematic approach to ensure transparency and replicability (FTA, n.d.; Tabassi & Bakar, 2018; Taherdoost, 2021). These steps ensured a systematic approach to data collection, enabling future researchers to replicate the process.

Step 1. Identify Relevant Government and Public Websites and Reports. The primary sites were the FTA, GAO, and the U.S. Department of Transportation. The U.S. Census Bureau, the U.S. Weather Bureau, and the U.S. Bureau of Economic Analysis were used for additional factors. Public sites, such as a metropolitan transit authority and the American Public Transport Authority, were used to fill in any gaps in the data. They were cross-checked with other sources to ensure accuracy.

Step 2. Identify Search Criteria. The search was based on the data fields identified in the data collection instrument within the parameters of the defined projects. The projects must be completed and meet the definition of a mass transit project.

Step 3. Collect Data and Enter it into the Data Collection Instrument. All sources were properly documented and referenced. Exceptions and data collection issues were reported, along with decisions regarding the data.

Step 4. Verify the Data. Where there were multiple sources, the data were cross-referenced. For accuracy, the FTA and GAO reports were given preference over other sources.

Step 5. Document the Data Collection Process. Discrepancies were identified and resolved with documented explanations for the decisions.

Data Analysis

Machine learning algorithms dynamically learn which features influence the results based on the variables in the datasets (Hinton & Sejnowski, 1986). The same datasets were used for supervised and unsupervised learning. Using the supervised learning process, the Pearson correlation coefficient from linear regression of the predicted and actual project costs was used to evaluate Research Question 1. The correlation coefficient from linear regression of the actual project cost and cluster means determined from the unsupervised learning process were used to evaluate Research Question 2. The two correlation coefficients were compared, and a paired t-test was used to determine the result for Research Question 3. The inflation-adjusted actual project cost for each dataset was compared to the predicted variances to determine the accuracy and efficacy of the budget forecasting models used. Decisions taken in the analysis process that may have affected the results were included in the final analysis.

Machine Learning Using Supervised Learning

The SPSS function for performing supervised learning utilized a neural network. Once the data were imported and properly categorized, the neural network options were configured. A target variable was used to perform the regression analysis and calculate predicted values. In this study, the target variable was the inflation-adjusted actual cost for each dataset. SPSS randomly partitions the datasets with an allocation of 70% to the training dataset and 30% to the testing dataset. This presupposes a balanced target variable. When the datasets are unbalanced, the default partitioning is adjusted to properly represent a balanced target variable in the training dataset. This process is handled through random sampling, which ensured that the subsets are representative of the overall datasets. The default settings and options selected are listed in Table 8.

Table 8

Default Settings Used for MLP in SPSS

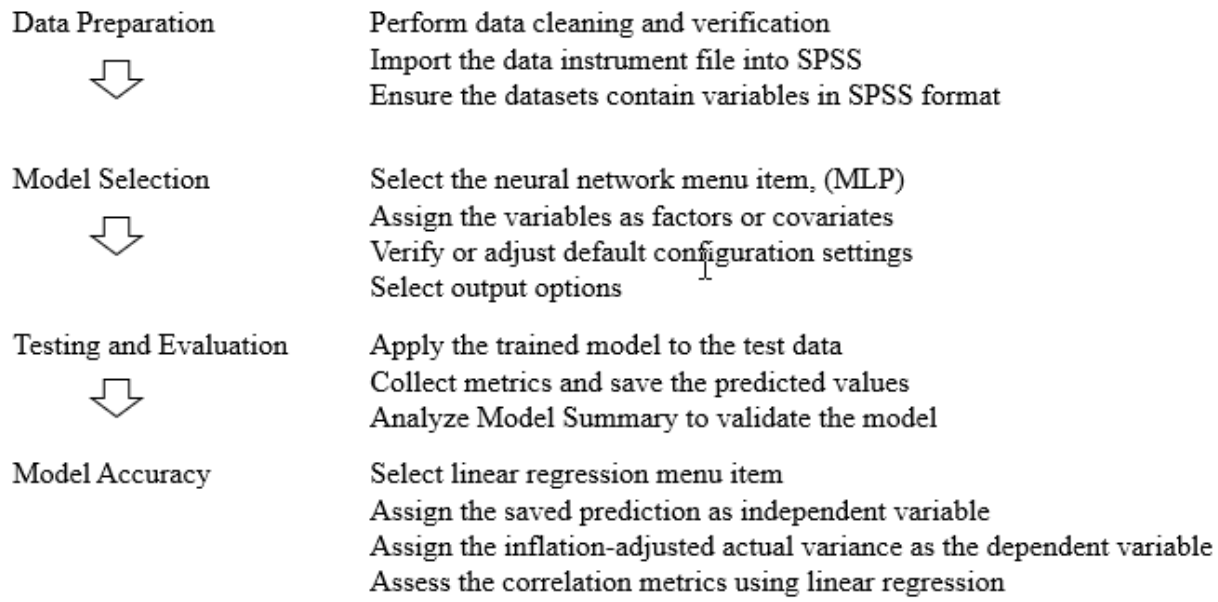
Option	Description	Selected
Variables	With scaling of covariates, standardized is the default, and normalized is used if extreme outliers exist in the data.	Default = standardized
Partition	The percentage split of datasets is determined for training and testing the model.	Default = randomly assign
Architecture	The number of hidden layers is defined.	Default = one layer
Training	The options are scaled conjugate gradient (SCG) or gradient descent. SCG is an advanced and efficient optimization method compared to gradient descent, requiring less manual parameter tuning.	Default = scaled conjugate gradient

Note. Based on SPSS User Guide (IBM, 2024).

Project features were assigned as inputs. In SPSS, categorical variables are considered factors, and ratio variables are entered as covariates. Once data fields and settings were entered, the analysis was executed, prompting SPSS to process the data. An important step in interpreting the output was to assess variable importance to understand the relative contribution of features while considering any systematic biases that may skew these contributions. For instance, if some variables dominated the correlation due to unbalanced target variables, it could result in misleading insights about their relevance. The process is described in Figure 3.

Figure 3

Machine Learning Process in SPSS for Supervised Learning Using a Neural Network



Note. Based on SPSS User Guide (IBM, 2024).

For supervised learning, a model summary table was used to display the results of the model. The sum of squared errors (SSE) measures the total squared differences between the predicted and actual values, indicating the overall model error. Differences between the training SSE and testing SSE determined if the model performs better on the test or training dataset (Zhang et al., 2019). Relative error measured the average deviation of predicted values from actual values, normalized by the actual values, providing a scale-independent accuracy metric. A low relative error indicates that predictions are close to actual values, while a high relative error suggests potential issues with accuracy (Tofallis, 2015). SSE can also signal overfitting if the training error is much lower than the test error or indicates underfitting if both errors are high. However, there is a lack of exact guidelines, and interpretation depends on context when assessing the proximity of SSE values (Tofallis, 2015). Since neural networks involve

randomness in dataset selection for training and testing, an acceptable model has a 70/30 dataset split and demonstrate minimal model overfitting.

This structured approach to using SPSS's neural network module, with explicit attention to bias considerations such as data imbalance, sampling errors, and optimism bias, ensured a robust and transparent analysis. Evidence of model fitness and how well it performs on the training and test datasets was captured by comparing relative error metrics. The model produced predicted actual costs for each dataset. To answer Research Question 1, the neural network predicted actual cost was compared to the inflation-adjusted actual cost for each project using linear regression, which produced a Pearson correlation coefficient and significance level.

Machine Learning Using Unsupervised Learning

The SPSS capability for unsupervised learning used a k-means algorithm. Once the data were imported and properly categorized, the k-means options were configured as identified in Table 9. The optimal number of clusters was determined using Ward's method and evaluated with the elbow criterion. This method was chosen because it minimizes intra-cluster variance, ensuring that the regions within each cluster are as similar as possible while maximizing the differences between clusters (Majerova & Nevima, 2017; Serin et al., 2023). Ward's method identifies statistically significant differences between clusters, which is important in the context of mass transit projects. In a hierarchical clustering approach, Ward's method determined the optimal number of clusters that best represent differences in the datasets. In hierarchical clustering in SPSS, an agglomeration schedule table indicated the order in which clusters are merged and the distance or coefficients at each step. A significant increase in these coefficients produces an elbow effect that indicates two relatively dissimilar clusters have been merged, and the clustering process has reached a point where combining more clusters may lead to less

meaningful solutions. Once the largest incremental increase is identified, the optimal number of clusters is the number before the most significant increase. This number is used to initialize the k-means algorithm, ensuring that the cluster centers are optimally placed and the classification process is both robust and statistically sound.

Table 9

Settings Used for K-Means in SPSS

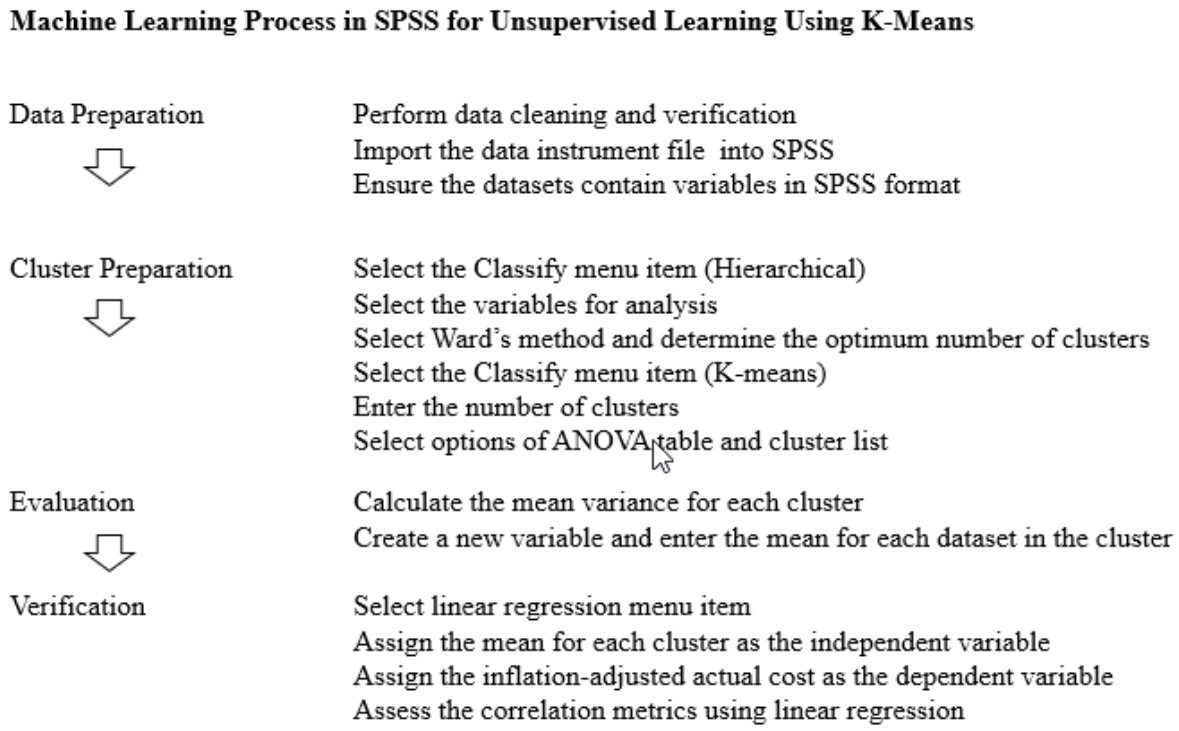
Option	Description	Selected
Number of Clusters	The number of desired groupings in the classification	Calculated using Ward's method
Iterate	Maximum number of iterations	Default = 10

Note. Based on SPSS User Guide (IBM, 2024).

Project features were assigned as variables, including nominal data, which was encoded. Once data fields and settings were entered, the k-means function was employed, prompting SPSS to process the data. The results produced a table displaying the cluster centers, iteration history, cluster membership, final cluster centers, an ANOVA table identifying the significance of each feature, and the number of datasets in each cluster. The process is described in Figure 4.

Figure 4

Machine Learning Process in SPSS for Unsupervised Learning Using K-Means



Note. Based on SPSS User Guide (IBM, 2024).

The mean inflation-adjusted actual cost was calculated for each cluster and compared to the inflation-adjusted actual cost for each dataset in the cluster. Linear regression was used, assigning the inflation-adjusted actual cost of each dataset in the cluster as the dependent variable and the cluster mean as the independent variable. The regression produced a Pearson correlation coefficient and significance level to evaluate Research Question 2. The Pearson correlation coefficient and a paired *t*-test was used to compare the supervised and unsupervised learning methods to respond to Research Question 3.

Assumptions

The study was based on several assumptions. First, the data collected by the FTA was assumed to be accurate and reported properly. Next, after data collection, it was assumed that

there were sufficient datasets to perform the neural network correlation without underfitting or overfitting. Another assumption was that the SPSS configuration settings used for the neural network and k-means analysis created representative models. Finally, it was assumed that by using SPSS, the research design and analysis minimized any cognitive bias in the study.

Limitations

The study was to determine if there is a relationship between the dependent and independent variables using machine learning methods. There were no recommended actions to resolve specific variances. The study was limited by the project type, the availability of data, and the historical nature of the data available for completed projects. While the dataset was an attempt to represent the entire population of mass transit projects, the population size could limit the complexity of the models or reduce the robustness of predictive algorithms. The study objective was to build machine learning models based on available data, making them directly relevant to this set of projects. The reliability of the data affects model accuracy, as data may contain inconsistencies or incomplete entries. Additional objective data, such as double tracking, light signals, and right-of-way land acquisition, was excluded due to the difficulty in obtaining accurate data from reliable sources.

Delimitations

The study did not seek to use different machine learning algorithms such as random forest, decision trees, or SVM, opting instead for the neural network capability in SPSS. Based on the literature review indicating they were ineffective for this type of analysis, additional machine learning methods, such as genetic algorithms and reinforcement learning, were not used. The study focused on U.S. mass transit projects due to the availability of data and previous

studies that highlighted consistent budget overspending (Gao & Touran, 2020; Zhou et al., 2022).

Ethical Assurances

Prior to data collection, this study obtained approval from the Institutional Review Board (IRB) of the National University. The study used publicly available historical data and did not involve human subjects. The research used the SPSS premium edition, which contains machine learning functions and is licensed by the National University. The researcher was responsible for ensuring integrity, objectivity, and transparency exist in conducting the study. This included taking steps to avoid bias and accurately representing findings without misrepresentation. The study conclusions were properly handled in the event they resulted in privacy issues. For example, where results indicated that a specific vehicle manufacturer correlated to consistently poor project performance, the study revealed the findings based on encoded values rather than indicating specific named entities. The researcher does not work for or produce publicly available comments about mass transit and has no conflict of interest in this study.

Summary

This study was a quantitative research design to evaluate the efficacy of machine learning algorithms in predicting budget overspending in mass transit projects across the United States. The research employed an objective data collection approach, ensuring an unbiased methodology, with statistical analyses conducted using SPSS software. Data were sourced from reputable government repositories, systematically organized into a structured dataset, and verified for accuracy. Both supervised and unsupervised machine learning models were analyzed using the same dataset, allowing for a comparative analysis of their predictive capabilities. The research design followed a clearly defined and replicable process, ensuring methodological rigor

and reproducibility. Ethical considerations were minimal as the study relied exclusively on historical, publicly available data. The findings contribute to the growing body of research on artificial intelligence in project management, offering insights into how machine learning can enhance budget forecasting and decision-making in large-scale infrastructure projects.

Chapter 4: Findings

Mass transit system projects frequently overspend their budget estimates. A study by Gao & Touran (2020) reported that 77% of mass transit projects in the U.S. experienced cost overruns. This has a significant impact on strategic planning and grant allocations from funding agencies such as the U.S. Federal Transit Administration (FTA). Machine learning algorithms offer an innovative approach to validate budget estimates. The problem addressed in this study is the underestimation of budgets for infrastructure projects, specifically those related to mass transit systems. The purpose of this research was to determine the effectiveness of machine learning algorithms, utilizing both supervised and unsupervised learning, in validating a budget for mass transit projects. The research analysis utilized SPSS software to create supervised machine learning models and to evaluate the correlation between the model results and the inflation-adjusted actual costs of completed projects. The findings are organized by the data's validity and reliability, the research questions, the hypotheses addressed, an evaluation of the findings, and a summary.

Validity and Reliability of the Data

The focus of data collection was on reliable government sources, with cross-referencing to confirm data accuracy. The data were manually entered into the data instrument, a spreadsheet. The study utilized secondary data that identified the characteristics of U.S. mass transit infrastructure projects. The data collected from the FTA were submitted by transit authorities using standardized federally mandated reporting protocols, ensuring consistency and reliability across reporting agencies (GAO, 2010). The data collected from local transit authorities is from public websites and supported by maps and images of the physical aspects of the transit system. The use of administrative data for empirical research is shown to yield valid

results when properly verified and interpreted (Johnston, 2014; Vartanian, 2010). To ensure data integrity, manually extracted data entries were independently verified by a second reviewer (see Appendix C). For the external project variables, the auditor checked 100% of the data fields. The internal project characteristics were subjected to audit review for 30% of the project datasets. Any discrepancies identified were cross-referenced and resolved using the source documents. This verification process helped ensure the accuracy and consistency of key data variables, including cost, duration, number of stations, and vehicle procurement. Given the official nature of the data sources and the validation steps taken, the datasets are deemed suitable for quantitative analysis of mass transit project characteristics.

Data Collection

The data encompassed a diverse range of projects, varying in terms of size, complexity, and budget. Data were collected for 108 U.S. mass transit projects, spanning a range of completion dates from 1986 to 2024, and inflation-adjusted costs from $52 million to $22 billion. The inflation-adjusted actual cost was calculated based on the year of completion converted to 2024 constant dollars. For each project, 17 variables were captured and verified. Excluded from the projects were bus transit systems, streetcar systems, and projects dedicated to the reconstruction and renovation of existing transit systems. The completed projects did not include projects completed prior to 1986 due to the availability and reliability of the data. Three projects completed after 1986 were initially considered but excluded due to incomplete data, including one where the city failed to disclose the actual project cost.

The variables defined and the method of data collection minimized bias in calculating the predicted project cost based on project characteristics. The project variables collected are from physical project and environmental factors as reported by reputable sources. An independent

audit process verified the integrity of the data captured. The data, as described in Table 10, formed the foundation for unbiased findings based on statistical analysis that evaluated the research questions.

Table 10

Descriptive Statistics for U.S. Mass Transit Project Data

Variable	*N*	Minimum	Maximum	Mean	Std. Deviation
Inflation Adj Actual Cost ($M, 2024)	108	52.23	22,757.83	1,406.68	2,353.28
Project Duration (years)	108	2.00	17.00	6.56	3.148
Track Distance (miles)	108	1.00	44.00	9.55	7.51
Underground Miles	108	0.00	13.00	0.78	1.90
Number of New Stations	108	1.00	29.00	8.83	6.40
Number of Underground Stations	108	0.00	13.00	0.81	2.04
Number of New Maintenance Facilities (count)	108	0.00	1.00	0.49	0.50
Initial (1) or Extension (0)	108	0.00	1.00	0.30	0.46
Vehicles Purchased	108	0.00	294.00	24.04	33.82
Vehicle Brand (coded numerically)	108	1.00	7.00	4.01	1.48
FTA Region (numbered)	108	1.00	10.00	6.95	2.62
State GDP ($M)	108	115,627.00	4,103,124.00	1,781,130.03	1,575,534.51
Metro Density (per mi^2)	108	1263.00	29,303.20	7,314.85	4,918.22
Min. Avg. Temp (^{0}F)	108	12.20	72.30	36.405	12.51
Max Avg. Temp (^{0}F)	108	72.50	105.20	83.98	7.70
Temperature Range (^{0}F)	108	16.00	68.80	47.57	15.23

Results

The study employed both supervised and unsupervised machine learning techniques in SPSS to examine correlations between project characteristics and budget outcomes. Neural networks and k-means clustering were used to analyze patterns and assess the predictive

accuracy of cost estimations for U.S. mass transit projects. The intent was to provide insights into the application of machine learning models to validate project budgets.

The supervised learning method used SPSS to create a neural network model. To address the inherent variability in neural network training caused by random dataset partitioning and weight initialization, the model was trained and evaluated 10 times, with SPSS randomly assigning data to training and testing sets in each run. For each iteration, the relative error was calculated, and the model with the lowest relative error was selected as the optimum model. This process reflects a repeated cycle of training and validation using random sub-sampling, which improves model robustness by accounting for randomness in both data splits and initial conditions (Dietterich, 1998). The use of relative error as the selection criterion ensured that the chosen model minimized the average deviation between predicted and actual values, aligning model performance more closely with the underlying data patterns (Witten et al., 2011). Repeating the training process helped mitigate the effects of overfitting and improved the robustness of the final model by reducing sensitivity to initial conditions and data split randomness (Goodfellow et al., 2016).

Table 11 presents the metrics of 10 neural network models with the inflation-adjusted project cost as the target variable. The datasets were randomly partitioned into training and test sets with a dataset split of 70% for training and 30% for testing. The lowest relative error in training was achieved by Model 2. The relative error of .032, which is close to zero, implies high accuracy in training. The higher relative error for testing at 0.463 suggests potential for overfitting, which is consistent with the variation expected when models are applied to diverse datasets, but is still within the lower range of relative error for model testing (Hastie et al., 2009; Tofallis, 2015).

Table 11

Training and Testing Dataset Splits with Relative Error for 10 Models

Model	Training Datasets	Testing Datasets	Relative error: Model Training	Relative error: Model Testing
1	74 (69.2%)	33 (30.8%)	.585	.401
2	78 (72.2%)	30 (27.8%)	.032	.463
3	78 (72.2%)	30 (27.8%)	.081	.486
4	68 (63.6%)	39 (36.4%)	.070	.561
5	76 (70.4%)	32 (29.6%)	.140	.734
6	84 (77.8%)	24 (22.2%)	.723	.322
7	80 (74.1%)	28 (25.9%)	.538	.735
8	80 (74.1%)	28 (25.9%)	.052	.667
9	79 (73.1%)	29 (26.9%)	.063	.766
10	79 (73.1%)	29 (26.9%)	.059	.420

The predicted project cost values generated for each dataset, as determined by Model 2, were compared to the inflation-adjusted actual cost for each project using linear regression. SPSS reported that the model accounted for 94.6% of the variance in inflation-adjusted actual cost as displayed in Table 12. As the best-performing neural network model, Model 2 can be saved and used to generate predictions on similar datasets.

Table 12

Supervised Learning Model 2 Summary for Linear Regression from SPSS

Model	R	R^2	Adjusted R^2	Std. Error of the Estimate
2	.984	.946	.946	421.09

The regression model significantly predicted inflation-adjusted actual cost, $F(1, 106) = 1872.91$, $p < .001$, indicating a statistically significant relationship between the predicted and inflation-adjusted actual project cost, as observed in Table 13. The p-value falls well below the

conventional threshold of 0.05, providing strong statistical evidence that the model's predictions are closely associated with the observed actual inflation-adjusted cost.

Table 13

Supervised Learning ANOVA Summary for Predicting Inflation-Adjusted Actual Cost

Model		Sum of Squares	df	Mean Square	*F*	*p*
2	Regression	560,821,233.66	1	560,821,233.66	1872.91	< .001
	Residual	31,740,501.18	106	299,438.59		
	Total	592,561,734.83	107			

Research Question 1

Research Question 1 was "To what extent can a machine learning algorithm using a supervised learning model accurately predict the budget for mass transit projects in the United States?" The associated null and alternative hypotheses are as follows:

$H1_0$ A machine learning algorithm using supervised learning cannot achieve a statistically significant prediction for the budget of mass transit projects in the United States.

$H1_a$ A machine learning algorithm using supervised learning can achieve a statistically significant prediction for the budget of mass transit projects in the United States.

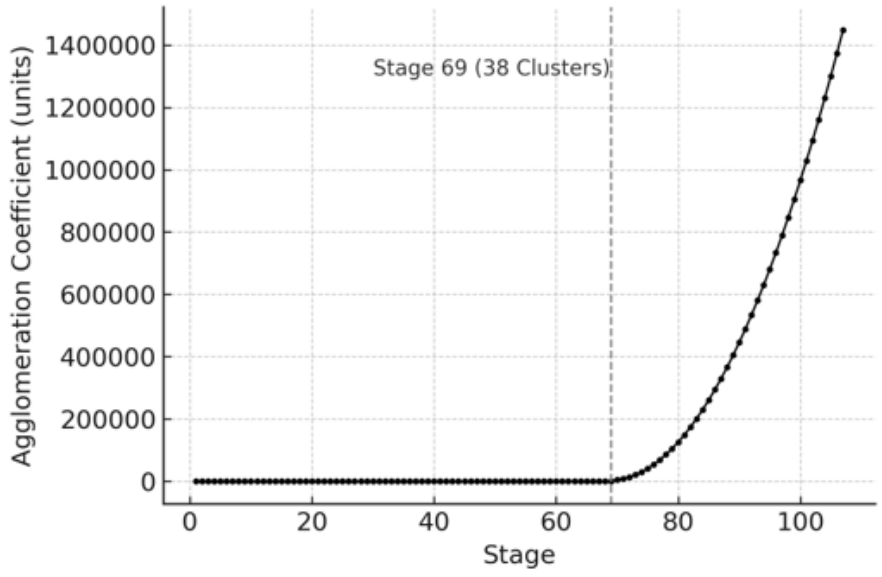
A statistically significant relationship was observed between the predicted and actual costs of U.S. mass transit projects. Based on project characteristics, a neural network model was developed to generate cost predictions, which were compared to the inflation-adjusted actual cost using linear regression. The ANOVA results indicated that the regression model significantly predicted the inflation-adjusted actual cost, accounting for more than 94% of the variance, with a

highly significant *p*-value ($p < .001$). These results provide sufficient evidence to reject the null hypothesis.

The unsupervised learning method applied the clustering functionality in SPSS on the same datasets used for supervised learning. Hierarchical cluster analysis using Ward's method was employed to identify a statistically appropriate number of clusters, which was determined to be 38. The agglomeration table produced by SPSS details the stepwise merging of clusters as displayed in Figure 5. Each merge is represented by the increase in within-cluster variance and is indicated by the agglomeration coefficient. The early stages display minimal increases in variance, indicating that similar data points are being grouped. As the clustering process progresses, the agglomeration coefficients begin to rise more sharply, reflecting the merging of increasingly dissimilar clusters.

Figure 5

Agglomeration Coefficients by Clustering Stage



Note. Output generated by IBM SPSS Statistics, Version 30.0 (IBM Corp.).

The elbow method was applied by plotting the agglomeration coefficients against the clustering stages. The elbow is the point at which the increase in variance accelerates noticeably and marks a natural breakpoint beyond which additional merging leads to a significant loss of cohesion within clusters. Stage 69 was the clustering step identified as the elbow point, indicating that a 38-cluster solution effectively captures the underlying structure of the data while preventing excessive merging of distinct clusters. The cluster value of 38 is calculated by subtracting the number of datasets, 108, from the next stage, 70.

The mean for each cluster was calculated and entered as the predicted value for each dataset. Linear regression was used to compare each dataset mean, the independent variable, to the inflation-adjected actual cost, the dependent variable. For the unsupervised learning method, the correlation between the cluster means and the inflation-adjusted actual cost accounted for over 88% of the variance in inflation-adjusted actual cost, as displayed in Table 14.

Table 14

Unsupervised Learning Linear Regression from SPSS

Model	R	R^2	Adjusted R^2	Std. Error of the Estimate
1	.939	.882	.881	810.72

The cluster method significantly predicted inflation-adjusted actual costs, $F(1, 106) = 795.564$, $p < .001$, indicating a statistically significant relationship between the predicted and actual project cost values, as observed in Table 15. The *p*-value falls well below the conventional threshold of 0.05, providing strong statistical evidence that the model's predictions are closely associated with the observed actual inflation-adjusted cost.

Table 15

Unsupervised Learning ANOVA Summary for Predicting Inflation-Adjusted Actual Cost

Model		Sum of Squares	df	Mean Square	*F*	*p*
1	Regression	522,892,193.79	1	522,892,193.79	795.56	< .001
	Residual	69,669,541.04	106	657,259.82		
	Total	592,561,734.83	107			

Research Question 2

Research Question 2 was "To what extent can a machine learning algorithm using an unsupervised learning model accurately predict the budget for mass transit projects in the United States?" The associated null and alternative hypotheses are as follows:

$H1_0$ A machine learning algorithm using unsupervised learning cannot achieve a statistically significant prediction for the budget of mass transit projects in the United States.

$H1_a$ A machine learning algorithm using unsupervised learning can achieve a statistically significant prediction for the budget of mass transit projects in the United States.

The results showed that a statistically significant correlation can be obtained using unsupervised learning to classify the datasets. Based on the project characteristics, the datasets were classified using the k-means clustering algorithm. Using linear regression, the ANOVA table indicated that the regression model significantly predicted the inflation-adjusted actual cost, explaining 88.2% of the variance, with a highly significant p-value ($p < .001$). The null hypothesis is rejected.

Research Question 3

Research Question 3 compares the results of supervised and unsupervised learning. This question was stated as “Is there a statistically significant difference in predicted budget between supervised learning and unsupervised learning algorithms for mass transit projects in the United States?” The associated null and alternative hypotheses are as follows:

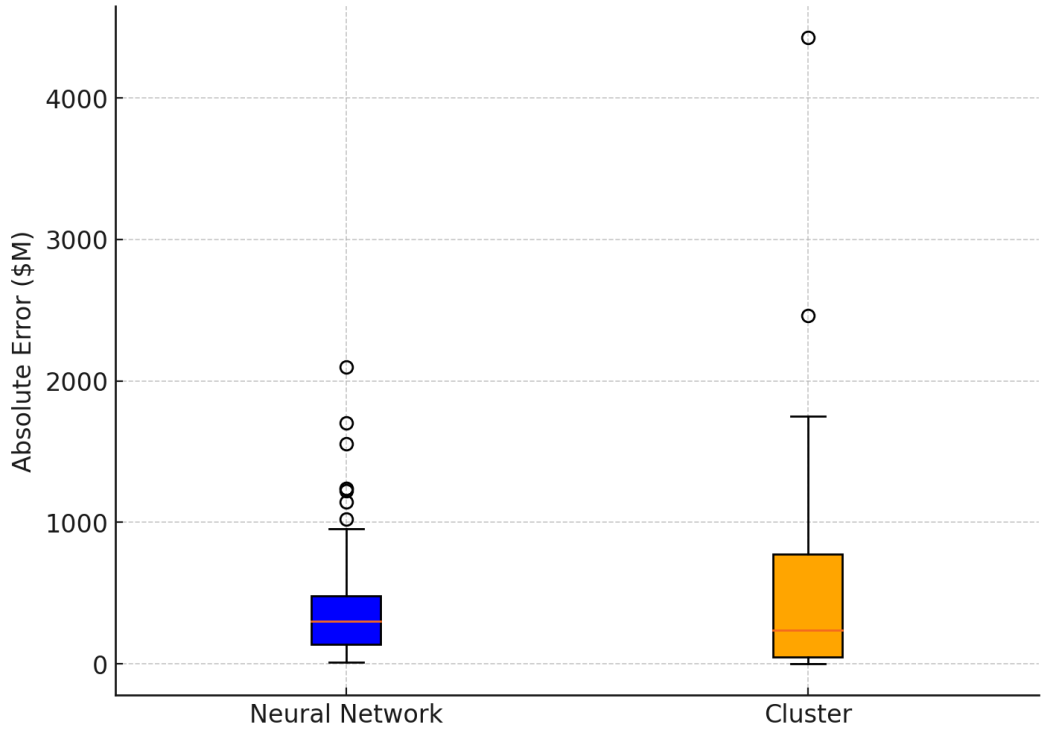
$H3_0$ Supervised learning does not perform significantly better than unsupervised learning in predicting the budget of mass transit projects in the United States.

$H3_a$ Supervised learning does perform significantly better than unsupervised learning in predicting the budget of mass transit projects in the United States.

A paired-samples *t*-test was conducted to compare the absolute prediction error between the supervised learning model (neural network) and the unsupervised clustering model across 108 transit projects. The supervised model (M = 393.45, SD = 377.41) had a lower mean absolute error (MAE) than the unsupervised model (M = 488.49, SD = 643.11). However, the difference was not statistically significant, $t(107) = 1.41$, $p = .160$. The mean difference in error was 95.04 (95% CI [−38.06, 228.41]), suggesting that although the supervised model showed lower prediction error on average, the evidence was insufficient to conclude a statistically significant performance advantage. The effect size was small (Cohen’s $d = .136$), indicating a modest practical difference between the two methods. A visual representation is presented in Figure 6, which compares the absolute prediction errors between the neural network and cluster-based models. Outliers are displayed as individual points.

Figure 6

Comparison of Prediction Errors



Note. All data points, including the highest outlier in the cluster-based model with an absolute error of $4,424.26 million, were included and displayed in the chart.

The mean and standard deviation data suggest that supervised learning provides more accurate budget predictions than unsupervised learning for U.S. mass transit projects, but the results are not significant. The boxplot reveals a lower median for unsupervised learning, but a higher mean, which is influenced by extreme outliers in the budget predictions. A paired-samples *t*-test revealed that the neural network model did not produce a statistically significant lower MAE than the clustering model. Based on this evidence, the null hypothesis was not rejected, indicating that neither method was statistically superior in predictive accuracy.

Evaluation of the Findings

The study incorporated multiple bias-mitigation strategies as outlined in Chapter 3, drawing on the framework presented by Dimitrakopoulou et al. (2024). A data audit was conducted to validate the integrity and reliability of the data. The data included a wide range of budget values, thereby addressing potential bias related to limited variability in the target variable. To reduce the risk of overfitting, a repeated random sub-sampling validation approach was employed to identify the optimal model configuration. Model performance was evaluated using relative error metrics for both training and test datasets. In addition, an ANOVA table was generated to assess statistical significance in both supervised and unsupervised learning scenarios. All procedures were executed using commercially available SPSS software, with sufficient explanations provided to ensure the replicability and transparency of the analysis.

The results of this study can be interpreted through the lens of cognitive bias, particularly optimism bias, which often leads to systematic underestimation of project costs. Both supervised and unsupervised machine learning models demonstrated high predictive accuracy, relying on empirical data rather than subjective estimates. The use of repeated random sub-sampling validation for neural network training reduced the influence of randomness and overfitting, while the clustering method grouped similar projects to smooth out anomalies. These techniques counter optimism bias by grounding predictions in observed cost patterns.

The research analysis is consistent with other research findings. The study by Zhou et al. (2022) aligns closely with this study in that it used neural networks to predict project operating costs for U.S. light rail systems, employing machine learning methods. They found that machine learning models substantially outperformed traditional multiple linear regression (MLR) in predictive accuracy. The methodology is also aligned, as both studies employed normalization of variables, used relative error metrics, created a model with a 70/30 data split, and minimized bias

by selecting diverse predictors. Research by Pang et al. (2022) also complements this research. They utilized neural networks to predict project capital costs, with a focus on IT projects. The use of diverse datasets and data verification minimized bias and confirmed the generalizability of their model.

Research by Mariani et al. (2025) supported the use of unsupervised learning to enhance project cost prediction. The study applied k-means clustering to segment project datasets based on shared characteristics. They conducted further analysis, using cluster means in regression modeling. The study emphasized how clustering reduces unsystematic variation and increases homogeneity within groups, thereby improving the reliability of predictions. Their research also used the elbow method to determine the optimal number of clusters, balancing model complexity with interpretability. Their work demonstrated that k-means clustering effectively reveals underlying patterns in project data, which enhances the accuracy of performance benchmarking and forecasting. The study affirmed the validity of combining clustering with regression in machine learning pipelines for project cost analysis.

Additional Findings

A statistical analysis was conducted to assess the relationship between inflation-adjusted project cost and absolute error, defined as the difference between the predicted value and the actual inflation-adjusted cost. Linear regression analyses were applied to the results of both the supervised and unsupervised models. In both cases, no statistically significant correlation was identified. These findings suggest that project size was not associated with higher prediction errors or cost overruns, indicating that larger projects did not exhibit disproportionately greater deviations in cost compared to smaller ones.

The correlating factors were similar for both studies. For the neural network model, SPSS produced an independent variable importance table that quantified the contribution of each input variable to the model's predictive accuracy. The number of vehicles purchased had the highest normalized importance at 100%, followed by the number of underground stations at 70.4% and the temperature range at 59.4%. Normalized importance expresses each variable's relative influence on the model's prediction, with the most impactful variable scaled to 100%. Other variables are proportionally ranked to show how their contributions compare. The unsupervised learning process yielded an ANOVA table indicating statistical significance in the relevance of specific variables in differentiating the 38 clusters. The most significant factors for clustering were the number of vehicles purchased, the temperature range, and track distance, as evidenced by high F-values and strong significance levels ($p < .001$).

Summary

The research provided evidence that machine learning can be used to generate statistically significant results for validating project budgets. A structured data validation process was implemented, including independent audit checks and cross-referencing with source documents. This process helped ensure the accuracy of the data instrument. The supervised learning model, utilizing a neural network with repeated random sub-sampling validation, yielded a coefficient of determination (R^2) of 0.946, indicating that the model explained 94.6% of the variance in actual costs. The unsupervised learning model, based on hierarchical clustering, identified 38 clusters and yielded an R^2 of 0.882. Both models produced statistically significant results ($p < .001$). The data validation and statistical approach employed in this study align with other studies that have attempted to minimize cognitive bias in budget predictions. The analysis demonstrated a machine learning approach that delivered statistically significant findings.

Chapter 5: Implications, Recommendations, and Conclusions

Infrastructure projects and more specifically, mass transit systems, lack an effective method to validate the accuracy of a project budget. Consequently, organizations are faced with unexpected cost overruns that impact strategic funding allocation and project delivery (Gao & Touran, 2020). The purpose of this correlational study was to evaluate whether machine learning algorithms were effective in predicting a project budget based on unbiased project characteristics. The study assessed the application of a supervised learning neural network model and an unsupervised learning approach, specifically k-means clustering, to predict the inflation-adjusted actual costs of mass transit projects in the United States. Using a dataset collected from publicly available government sources, the study modeled project outcomes based on project and environmental characteristics. A finding during data collection was that the scope was reduced on at least one project as a way to deliver the project within the budget (Federal Transit Administration, 2013), providing evidence that the underestimation of project budgets for mass transit projects is likely understated.

Factors Influencing Interpretation

Based on the limitations identified for the study, the results were influenced by several factors. These factors highlight the practical challenges of working with historical datasets collected from government sources. They also demonstrate how data collection limitations can affect the accuracy and performance of machine learning models.

Data Source Reliability

Government data may contain inconsistencies or incomplete entries, which can potentially affect model accuracy (GAO, 2010). This issue was evidenced in the data collection phase, where FTA reports from different time periods had different formats and occasionally

omitted specific project values. Remediation consisted of identifying a second source, such as the local transit authority, to verify the physical characteristics of the transit system. Variability in reporting standards may still have introduced measurement error, which was addressed in SPSS through model cross-validation.

Dataset Limitations

The analysis was limited to available U.S. mass transit projects, which may reduce the generalizability of the findings (Witten et al., 2011). A larger and more diverse dataset would improve model training and robustness (Althnian et al., 2022). In addition, the interpretation of model results depends on the consistency of input variables across projects. Variations in how variables are defined or categorized can introduce structural bias (Witten et al., 2011). Specific physical project characteristics, such as double tracking and signaling infrastructure, were excluded due to inconsistent or unavailable data, potentially omitting relevant predictors.

Potential Confounders

Contextual variables, such as local procurement practices and regulations, organizational preferences, project manager experience, and less measurable project factors, may influence cost outcomes in real-world settings (GAO, 2010). Any subjectiveness in the approach taken by the original budget estimators was not captured but is expected to exist (Flyvbjerg, 2022).

Implications

The significant correlation between predicted and actual costs indicated that machine learning provided a reliable and bias-resistant alternative to traditional forecasting methods, even though it does not eliminate the risk of project budget overruns. The results are relevant to project managers, organizational executives, and funding agencies. The value of this study lies in the practical application of machine learning to a complex public-sector problem. By using

supervised and unsupervised learning approaches, the research outlined a data-driven framework for evaluating project budgets. The insights can help planners, policymakers, and project managers assess budgets more effectively during the early stages of the planning process.

The findings demonstrated that machine learning methods can be applied to budget validation when used with appropriate evaluation and data verification processes. As a funding organization for U.S. transit projects, the FTA relies on accurate and consistent data collection. Templates can be defined and implemented that provide the critical data fields for the development of effective machine learning models. Reliable, accessible, and consistent data captured over time facilitates the application of innovative technologies such as machine learning to enhance predictive analysis and decision making.

The study also demonstrated that machine learning could be effectively implemented in SPSS using configurable settings. This approach avoids reliance on a software development environment to create code. The research highlighted the importance of data and process over algorithm creation.

Research Question 1

Research Question 1 results demonstrated that a neural network model accurately predicted project budgets for U.S. transit system projects. Although a significant correlation was observed, the mean error indicated variation in prediction accuracy. As a budget validation method, the process of data collection, verification, and using SPSS can still deliver an unbiased evaluation of budget accuracy. The results challenge institutional norms regarding the efficacy of traditional budget planning and highlight the need to integrate machine learning technology to counteract cognitive bias. Organizations such as the FTA can use a machine learning model to

evaluate budget accuracy across a portfolio of funded projects. The model can be used as developed in this study or improved with additional objective project factors.

Organizations and project managers can find value in the straightforward process of creating and evaluating a supervised learning model using SPSS. The reproducibility of the supervised learning process is grounded in the transparency of SPSS configurable settings as opposed to the nuances and complexity of creating software code. The implication is that machine learning becomes accessible to anyone with a basic knowledge of SPSS and supervised learning, enabling them to perform correlations for a budget validation process. SPSS is scalable and can technically handle over 2 billion cases and more than 32,000 variables (IBM Corp., n.d.). A realistic limit is 10 million projects, assuming sufficient computer memory and processing power (IBM, 2024). The ability of project managers who have statistical knowledge and can use SPSS democratizes machine learning by making the process accessible beyond data science experts and software developers.

Research Question 2

The unsupervised learning approach yielded clusters of projects based on attributes, grouping similar projects and segregating dissimilar projects into separate groups. Unsupervised learning required a series of steps to deliver a prediction. The value in clustering is the ability to statistically group similar projects for project budget validation and identify outliers. The correlation was significant, and this method may outperform supervised learning, depending on the data and the analysis objective.

Clustering projects enables the investigation of scope details that may explain differences or divergent methods in budget development. For example, in one project in the study, all the railcars were purchased through a separately funded procurement process (FTA, 1997). The

number of railcars emerged as a significant predictor in project costs, suggesting procurement strategies influence budget estimates. Although unsupervised learning did not achieve the same level of predictive accuracy as supervised learning, it demonstrated the potential to be applied to a different set of project problems (Mariani et al., 2025).

Based on the research process, clustering can be applied to reference class forecasting (RCF). In RCF, a key challenge is identifying the most appropriate dataset to use as a reference when deciding which historical project is comparable to the current one (Baerenbold, 2023). The analysis in this study demonstrated that unsupervised learning can address this limitation by classifying a project into the most similar group based on project characteristics. By using clustering techniques, the model objectively assigns the project to a reference class, reducing subjectivity and improving the relevance of cost comparisons. This approach enhances the traditional RCF method by introducing data-driven classification, which strengthens the foundation for more accurate forecasts.

Research Question 3

The study was not intended to provide a method for creating a project budget, but to investigate two methods to validate the level of budget accuracy. The results for Research Question 3 did not indicate a significantly better performing method. Unsupervised learning produced more extreme outliers, whereas supervised learning exhibited a narrower standard deviation around the mean prediction. The study did not indicate how an organization should select either method as part of a budget validation process.

Consistency with Existing Research and Theory

The findings of this study were consistent with existing research and theories on the persistent challenges of accurately forecasting transit project budgets. Traditional cost estimates

for infrastructure projects typically include bottom-up, parametric, or analogous methods (PMI, 2021). Commonly used budget validation methods include analogous, Monte Carlo simulation, and reference class forecasting (Park, 2021). The availability of these approaches has not prevented underestimation and cost overruns in mass transit projects, suggesting the prevalence of optimism bias described by Flyvbjerg (2022). The results of this study reinforce these historical findings, extending prior research that used machine learning and SPSS to examine project-related issues. Zhou et al. (2022) utilized a dataset that included seven of the 14 physical characteristics examined in this study to predict mass transit operating costs. Traditional methods resulted in a variance of up to 45% between actual and budgeted operating costs, whereas a supervised learning machine learning model achieved a prediction accuracy of 97% ($R^2 = 0.970$). Unsupervised learning, utilizing a k-means algorithm, was employed by Capone & Narbaev (2022) to estimate a contingency budget. The traditional expected monetary value method yielded a 43.94% probability of contingency budget accuracy, while the k-means approach demonstrated a 90.97% probability of the budget being sufficient.

The results in this study align with existing research and theoretical frameworks that emphasize the value of empirical, data-driven approaches in reducing error and bias in project forecasting. Prior studies in project management and behavioral economics highlighted the limitations of expert judgment and conventional estimation methods, particularly when affected by optimism bias and strategic misrepresentation (Flyvbjerg, 2009; Lovallo & Kahneman, 2003). The use of supervised and unsupervised machine learning aligns with the approach proposed by Kahneman and Tversky (1979), which advocates basing forecasts on actual outcomes from similar past projects rather than on subjective assumptions. The clustering and predictive capabilities demonstrated in this study reflect core principles from systems theory and decision

science, where complexity is best addressed through adaptive, evidence-based modeling (Sterman, 2000). The machine learning methods in this study used SPSS and comply with the machine learning reproducibility checklist developed by Pineau et al. (2021) to promote transparency, rigor, and reproducibility in machine learning research. By validating cost estimates using machine learning, this study supports the premise that integrating objective data analysis into planning processes enhances the reliability of project budgets.

Contribution to Existing Literature and Framework

The contribution to the conceptual framework of cognitive bias is demonstrated by using objective factors from historical data to determine a significant correlation. These findings support the importance of incorporating environmental and physical variables into budget forecasting models for transit projects to yield reliable results. The study contributes to the existing literature by offering a data-driven machine learning alternative to traditional project forecasting methods, which are often susceptible to cognitive biases such as optimism bias and strategic misrepresentation (Flyvbjerg, 2013). These biases, well-documented in infrastructure planning, frequently result in underestimated costs and overestimated benefits (Flyvbjerg, 2009).

By applying supervised and unsupervised machine learning techniques to predict project costs, the research demonstrated how algorithmic models could complement or even challenge human judgment. Within the framework of cognitive bias, the use of machine learning introduces objectivity and consistency, reducing the influence of subjective assumptions that commonly distort budget forecasts (Flyvbjerg, 2002). This aligns with research that encourages outside-view approaches and empirical modeling to offset human error in complex project environments (Flyvbjerg, 2013; Kahneman & Tversky, 1979; Mariani et al., 2025). The study contributes to the growing body of research at the intersection of artificial intelligence, project

management, and behavioral economics, demonstrating how AI tools can mitigate known biases and improve the reliability of decision-making in capital project planning.

Recommendations for Practice

Project management offices and agencies such as the FTA should begin using machine learning models based on historical data to validate that project budgets align with funding allocation assumptions. The findings highlight that reliable data forms the foundation for the effective application of AI-based analysis. The availability and quality of data underscore the essential role of consistent data capture and retention by the FTA. Calculating a cost per mile has been used as a form of budget comparison (California State Transportation Agency, 2025). However, this method can be inadequate, as it fails to capture the complexity and variability of infrastructure projects. Machine learning models offer a more data-driven approach, incorporating a wider range of project variables to identify complex, non-linear relationships and interactions among factors that traditional cost-per-mile calculations overlook. The study highlighted the importance of data as machine learning algorithms become more accessible. With a reliable model, correlational analysis can be performed by changing the data, not the algorithm.

The next logical step in this line of study is to expand the dataset by incorporating additional variables that may influence project outcomes to improve model performance. While the current study focused on core project characteristics, such as cost, physical characteristics, and project environmental conditions, future iterations could include additional factors. Adding more granular data, such as contract type, procurement method, or risk mitigation strategies, may enhance the model's predictive accuracy. By enriching the dataset, machine learning models can be used by organizations such as the FTA to capture the complexity of real-world transit projects

and produce more reliable forecasts, ultimately supporting more informed decision-making in project planning and execution.

Recommendations for Future Research

Future research should test the model's generalizability. The models can be tested across geographies, different project types, and other project issues. The dataset can be reconfigured to correlate features for international projects, diverse infrastructure projects, or issues such as project duration and risk. Predictive results provide enhanced efficiency in planning, a factor recognized as critical to successful project delivery (Flyvbjerg & Gardner, 2023).

Geographic Regions and Project Types

The methodology of this study can be extended beyond the United States. Both the supervised and unsupervised learning approaches can be applied to assess whether the models generalize effectively to transit projects in international regions. Additionally, the approach can be adapted for use in other types of infrastructure projects where cost overruns are common. Expanding the dataset to include a broader range of projects or more detailed project characteristics may enhance model accuracy and predictive power. In addition to validating budgets, future research can focus on estimating specific budget components, such as risk contingencies, based on a project's internal and external risk factors.

Project Issues

The methodology developed in this study can also be applied to other critical project management issues, such as predicting project duration. A study of 180 U.S. and international transit projects found that construction durations in the U.S. were significantly longer, with tunneled projects taking nearly 18 months longer, and at-grade projects taking six months longer on average than similar international counterparts (Eno Center for Transportation, 2023).

Supervised learning models trained on historical data that include variables such as project scope, delivery method, number of stations, and details of underground construction are a potential method to validate the duration of a new transit project. Unsupervised learning can be used to group projects by similarity based on duration-related characteristics, providing insights into common delay factors. This approach can help project planners set more realistic timelines and proactively identify issues that contribute to schedule overruns.

Optimal Machine Learning Method

An important avenue for future research is determining which project issues are more conducive to accurate results from supervised versus unsupervised machine learning methods. Certain types of projects, such as those with well-documented historical data and clearly defined outcome variables, may benefit more from supervised learning, where prediction is the primary goal. In contrast, projects with less structured data or lacking predefined output variables may be better suited for unsupervised learning, which can reveal hidden patterns or groupings (Naeem et al., 2023). Establishing clear criteria or a decision framework for selecting the appropriate machine learning method based on project characteristics, data quality, and analytical objectives would enhance the strategic application of machine learning approaches in project planning and management.

Range of Data Volume for Effective Project Models

Researchers who acquire secondary data based on availability may be limited by accessible datasets. Studies in project management need to determine the optimal volume of data, in the form of the number of datasets and the number of project characteristics, that are more likely to deliver a significant result. The effective range in the amount of data required may

vary depending on the research objective or the selection of supervised and unsupervised learning methods.

Conclusions

This dissertation explored the application of supervised and unsupervised machine learning techniques to validate budget forecasts for U.S. mass transit projects, addressing a persistent challenge in infrastructure planning. Cost overruns are driven by inaccurate early estimates and cognitive biases. The study demonstrated that machine learning provides a data-driven alternative to traditional forecasting methods, supporting more objective and consistent analysis. Using a structured dataset of completed transit projects, a supervised learning model was trained to predict actual project costs. An unsupervised learning process was applied to classify projects into relevant reference groups to determine a cost prediction. The results showed that both approaches generated significantly predicted project costs. These findings align with and reinforce existing theories on optimism bias and strategic misrepresentation, demonstrating how machine learning can improve results by shifting decision-making toward empirical evidence.

The study also revealed practical challenges in data collection, underscoring the importance of consistent data capture and retention practices at the FTA and metropolitan transit authorities to support advanced analytics and innovative technologies, such as artificial intelligence. Beyond cost prediction, the developed methodology has broader applicability, as it can be adapted to other geographies, infrastructure types, and forecasting goals, including schedule prediction or contingency allocation. An important recommendation for future research is to expand the dataset with additional project characteristics and contextual variables to

improve model accuracy. Establishing decision criteria for when to apply supervised versus unsupervised methods also presents a valuable research opportunity.

Using SPSS instead of custom Python code in this study enhanced transparency and accessibility in both data handling and model development. The structured approach ensured that analytical results were traceable, reproducible, and easily reviewed by others. The research demonstrated that advanced techniques, such as neural networks and k-means clustering, are more accessible to non-programmers, enabling broader participation in machine learning analysis and decision-making.

This dissertation makes significant contributions both theoretically and practically to how machine learning applies to project management and behavioral economics. As the complexity and scale of public investments continue to grow, the integration of AI-based tools into project planning and oversight becomes increasingly important. By demonstrating that machine learning can mitigate forecasting bias and improve the reliability of cost estimates, the research contributes to the advancement of evidence-based decision-making and promotes greater accountability in the planning and delivery of large-scale infrastructure projects.

References

Abdallah, T. (2017). *Sustainable mass transit*. Elsevier.

Ackermann, F., Maytorena, E., Gavin, C., & Forsyth, S. (2024). Developing enduring leadership competencies in complex project management: Charting a course, embarking on a journey. *Journal of Management Development*, *43*(1), 35–48. https://doi.org/10.1108/JMD-09-2022-0218

Aggarwal, C. C. (2015). *Data mining: The textbook.* Springer.

Agresti, A. (2018). *Statistical Methods for the Social Sciences* (5th ed.). Pearson Education.

Ahamad, M. K., & Bharti, A. K. (2021). Analysis of cluster performance of real dataset using SPSS tool with K-means approach via PCA. *Advances in Mathematics: Scientific Journal, 10*(1), 535–542. https://doi.org/10.37418/amsj.10.1.53

Ahiaga-Dagbui, D. D., & Smith, S. D. (2014). Rethinking construction cost overruns: Cognition, learning and estimation. *Journal of Financial Management of Property and Construction, 19*(1), 38–54. https://doi.org/10.1108/JFMPC-06-2013-0027

Ahmed, C. (2021). Early cost estimation models based on multiple regression analysis for road and railway tunnel projects. *Arabian Journal of Geosciences*, *14*(11). https://doi.org/10.1007/s12517-021-07359-x

Akanni, A. O. (2024). AI in dynamic budgeting and forecasting. *European Journal of Business and Innovation Research*, 12(6), 86–99. https://doi.org/10.37745/ejbir.2013/vol12n68699

Al Juboori, A. A. (2021). Practices that increase the budget estimate's accuracy during the initial phase of the construction project's life cycle. *International Journal of Construction Project Management*, *13*(2), 129–159.

Albert, M., Balve, P., & Spang, K. (2017). Evaluation of project success: A structured literature review. *International Journal of Managing Projects in Business*, *10*(4), 796–821. https://doi.org/10.1108/ijmpb-01-2017-0004

Althnian, A., AlSaeed, D., Al-Baity, H., Samha, A., Dris, A. B., Alzakari, N., Abou Elwafa, A., & Kurdi, H. A. (2022). Impact of dataset size on classification performance: An empirical evaluation in the medical domain. *Applied Sciences, 12*(5), 2497. https://doi.org/10.3390/app12052497

Amadi, A. I. (2022). The financial implication of weather condition as a determinant of on-site productivity and contract duration. *International Journal of Construction Engineering and Planning, 7*(2), 6–14. https://doi.org/10.37628/IJCEP

Ammar, T., Abdel-Monem, M., & El-Dash, K. (2023). Appropriate budget contingency determination for construction projects: State-of-the-art. *Alexandria Engineering Journal, 78*, 88–103. https://doi.org/10.1016/j.aej.2023.07.035

Amron, A. (2018). Effects of product quality, price, and brand image on the buying decision of city car product. *Archives of Business Research*, *6*(4), 1–8. https://doi.org/10.14738/abr.64.4374

Aquino, Y. (2023). Making decisions: Bias in artificial intelligence and data-driven diagnostic tools. *Australian Journal of General Practice, 52*(7), 439–442. https://doi.org/10.31128/AJGP-12-22-6630

Ayub, Z., & Banday, M. T. (2023). Ethics in artificial intelligence: An analysis of ethical issues and possible solutions. In *Proceedings of the 2023 Third International Conference on Smart Technologies, Communication and Robotics (STCR)* (pp. 1–6). https://doi.org/10.1109/STCR59085.2023.10396966

Baerenbold, R. (2023). Reducing risks in megaprojects: The potential of reference class forecasting. *Project Leadership and Society, 4*(100103-). https://doi.org/10.1016/j.plas.2023.100103

Bakhshi, J., Ireland, V., & Gorod, A. (2016). Clarifying the project complexity construct: Past, present and future. *International Journal of Project Management, 34*(7), 1199–1213. https://doi.org/10.1016/j.ijproman.2016.06.002

Bakhshi, R., Moradinia, S. F., Jani, R., & Poor, R. V. (2022). Presenting a hybrid scheme of machine learning combined with metaheuristic optimizers for predicting the final cost and time of project. *KSCE Journal of Civil Engineering, 26*(8), 3188–3203. https://doi.org/10.1007/s12205-022-1424-3

Basystiuk, O., Melnykova, N., & Rybchak, Z. (2023). Multimodal learning analytics: An overview of the data collection methodology. *2023 IEEE 18th International Conference on Computer Science and Information Technologies (CSIT), Computer Science and Information Technologies (CSIT), 2023 IEEE 18th International Conference On*, 1–4. https://doi.org/10.1109/CSIT61576.2023.10324177

Boehm, M., Kumar, A., & Yang, J. (2022). *Data management in machine learning systems*. Springer Nature.

Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2016). Feature selection for high-dimensional data. *Progress in Artificial Intelligence*, *5*, 65–75.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Brandl, F. J., Roider, N., Hehl, M., & Reinhart, G. (2021). Selecting practices in complex technical planning projects: A pathway for tailoring agile project management into the

manufacturing industry. *CIRP Journal of Manufacturing Science and Technology, 33,* 293–305. https://www.sciencedirect.com/science/article/pii/S1755581721000559

Brynjolfsson, E., & McAfee, A. (2017). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. W.W. Norton & Company.

California State Transportation Agency. (2025, May). *Transit Transformation Task Force meeting 10: Staff report for item 4 – Recommendations* [PDF]. California State Transportation Agency. https://calsta.ca.gov/-/media/calsta-media/documents/tttf_10_staff_report_for_item_4-recommendations-a11y.pdf

Capone, C., & Narbaev, T. (2022). Estimation of risk contingency budget in projects using machine learning. *IFAC PapersOnLine, 55*(10), 3238–3243. https://doi.org/10.1016/j.ifacol.2022.10.140

Çevik, M., & Tabaru-Örnek, G. (2020). Comparison of MATLAB and SPSS software in the prediction of academic achievement with artificial neural networks: Modeling for elementary school students. *International Online Journal of Education and Teaching (IOJET)*, *7*(4), 1689–1707. Retrieved from http://iojet.org/index.php/IOJET/article/view/999

Chandanshive, V. & Kambekar, A. (2019). Estimation of building construction cost using artificial neural networks. *Journal of Soft Computing in Civil Engineering, 3*(1), 91–107. https://doi.org/10.22115/scce.2019.173862.1098

Chandrasekaran, B., & Ravindran, B. (2023). *Machine Learning for High-Dimensional Data: Feature Selection and Feature Engineering Techniques.* Springer.

Chen, Y., Ahiaga-Dagbui, D. D., Thaheem, M. J., & Shrestha, A. (2023). Toward a deeper understanding of optimism bias and transport project cost overrun. *Project Management Journal, 54*(5), 561–578. https://doi/10.1177/87569728231180268

Chen, Z.-J., Hsieh, T.-S., & Davoudi, S. M. M. (2022). Analysis of factors affecting the success of sustainable development projects with the help of machine learning tools. *Discrete Dynamics in Nature and Society, 2022,* Article ID 1956879. https://doi.org/10.1155/2022/1956879

Cleland, D., & Gareis, R. (2006). *Global project management handbook: Planning, organizing and controlling international projects*. McGraw-Hill Education. ISBN 0071460454.

Cohen J. (2010). *Statistical power analysis for the behavioral sciences*. Routledge.

Dao, B., Kermanshachi, S., Shane, J., Anderson, S., & Damnjanovic, I. (2020). Developing a logistic regression model to measure project complexity. *Architectural Engineering and Design Management*. https://doi.org/10.1080/17452007.2020.1851166

Das, S., Ashrafuzzaman, M., Sheldon, F. T., & Shiva, S. (2024). Ensembling supervised and unsupervised machine learning algorithms for detecting distributed denial of service attacks. *Algorithms*, *17*(3), 99. https://doi.org/10.3390/a17030099

Data Privacy and Security. (n.d.). *IBM*. https://www.ibm.com/security/data-security

Daulay, R. Y., Passalaras, R. A., & Heikal, J. (2024). Customer segmentation using K-means clustering with SPSS program in a case study of consumer interest in current coffee shop budgeting. *Journal of Business, Management and Accounting, 5*(2), 721-740. https://doi.org/10.31539/budgeting.v5i2.9288

Deb, I., & Gupta, R. K. (2023). A genetic algorithm based heuristic optimization technique for solving balanced allocation problem involving overall shipping cost minimization with

restriction to the number of serving units as well as customer hubs. *Results in Control and Optimization, 11*. https://doi.org/10.1016/j.rico.2023.100227

Dempsey, M., Brennan, A., Holzberger, A., & McAvoy, J. (2022). A review of the most significant challenges impacting conventional project management success. *IEEE Engineering Management Review*, *50*(3), 1–5.

Deng, J. & Jian, W. (2022). Estimating construction project duration and costs upon completion using Monte Carlo simulations and improved earned value management. *Buildings*, *12*(12), 2173. https://doi.org/10.3390/buildings12122173

Denicol, J., Davies, A., & Krystallis, I. (2020). What are the causes and cures of poor megaproject performance? A systematic literature review and research agenda. *Project Management Journal, 51*(3), 328–345. https://doi.org/10.1177/8756972819896113

De Reyck, B., Grushka-Cockayne, Y., Fragkos, I., Harrison, J., & Read, D. (2017). Optimism Bias Study: Recommended adjustments to optimism bias uplifts. *UK Department for Transport, UK, Horseferry Road, London*.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation, 10*(7), 1895-1923. https://doi.org/10.1162/089976698300017197

Dimitrakopoulou, G., Kapsalis, N., & Kokkinis, G. (2024). Bias detection and correction methods for machine learning algorithms. In *Proceedings of the 2024 5th International Conference on Electronic Engineering, Information Technology & Education (EEITE)* (pp. 1–6). https://doi.org/10.1109/EEITE61750.2024.10654404

Dinu, A.-M. (2016). Project risk management - Reasons why projects fail. *Quality - Access to Success*, *17*, 208–213.

Doloi, H. (2013). Cost overruns and failure in project management: Understanding the roles of key stakeholders in construction projects. *Journal of Construction Engineering and Management*, *139*(3), 267–279.

Dwi Ramadhan, G. C. (2023). Bridging the time gap: Exploring escalation method for accurate, precise, reliable, and valid project budgeting. *PM World Journal*, *12*(10), 1–29.

Egboga, I., Daniel, C. O., & Abubakar, H. L. (2022). Effect of cost estimation on project performance in construction firms in Abuja. *Turkish Online Journal of Qualitative Inquiry*, *13*(1), 1052–1063.

Electronic Code of Federal Regulations. (n.d.). *Project management oversight*. Title 49, Subtitle B, Chapter VI, Part 633. U.S. Government Publishing Office. Retrieved January 22, 2025, from https://www.ecfr.gov/current/title-49/subtitle-B/chapter-VI/part-633

Elghaish, F., Abrishami, S., Hosseini, M. R., & Abu-Samra, S. (2021). Revolutionising cost structure for integrated project delivery: a BIM-based solution. *Engineering Construction & Architectural Management (09699988)*, *28*(4), 1214–1240. https://doi.org/10.1108/ECAM-04-2019-0222

Elkind, E. N., Segal, K., & Lamm, T. (2022). *Getting back on track: Policy solutions to improve California rail transit projects.* University of California Institute of Transportation Studies. https://doi.org/10.7922/G2V986CM

Eno Center for Transportation. (2021, July 30). *Eno releases major report on U.S. transit costs and project delivery*. Retrieved January 15, 2025, from https://enotrans.org/article/eno-releases-major-report-on-u-s-transit-costs-and-project-delivery/

Eno Center for Transportation. (2023). *Eno releases major report on U.S. transit costs and project delivery*. https://enotrans.org/article/eno-releases-major-report-on-u-s-transit-costs-and-project-delivery/

Ewees, A. A., Gaheen, M. A., Alshahrani, M. M., Anter, A. M., & Ismail, F. H. (2024). Improved machine learning technique for feature reduction and its application in spam email detection. *Journal of Intelligent Information Systems: Integrating Artificial Intelligence and Database Technologies*, 1–23. https://doi.org/10.1007/s10844-024-00870-z

Federal Transit Administration (n.d.), *The National Transit Database (NTD),* Federal Transit Administration. Retrieved September 5, 2024, from https://www.transit.dot.gov/ntd

Federal Transit Administration. (1997). *Annual report on funding recommendations: Fiscal year 1998, Capital Investment Grant Program*. U.S. Department of Transportation. https://www.transit.dot.gov/sites/fta.dot.gov/files/docs/FY98_Annual_Report_on_Funding_Recommendations.pdf

Federal Transit Administration. (2013). *Metro Gold Line Extension project: Before-and-after study, Los Angeles, California*. U.S. Department of Transportation. https://www.transit.dot.gov/before-and-after-studies

Federal Transit Administration. (2016). *Project and construction management guidelines: March 2016.* U.S. Department of Transportation. Available at http://www.fta.dot.gov

Field, A. (2018). *Discovering Statistics Using IBM SPSS Statistics* (5th ed.). SAGE Publications.

Flyvbjerg, B., Holm, M. K. S., & Buhl, S. L. (2002). Underestimating costs in public works Projects: Error or lie? *Journal of the American Planning Association, 68*(3), 279–295.

Flyvbjerg, B. (2009). Survival of the un-fittest: Why the worst infrastructure gets built—and what we can do about it. *Oxford Review of Economic Policy, 25*(3), 344–367. https://doi.org/10.1093/oxrep/grp024

Flyvbjerg, B. (2013). Delusions of success: Comment on Dan Lovallo and Daniel Kahneman. *Harvard Business Review, 81*(12).

Flyvbjerg, B. (2014). What you should know about megaprojects and why: An overview. *Project Management Journal, 45*(2), 6–19. https://doi.org/10.1002/pmj.21409

Flyvbjerg, B., Stewart, A., & Budzier, A. (2016). *The Oxford Olympics study 2016: Cost and cost overrun at the games*. Saïd Business School, University of Oxford.

Flyvbjerg, B., Ansar, A., Budzier, A., Buhl, S., Cantarelli, C., Garbuio, M., Glenting, C., Holm, M. S., Lovallo, D., Lunn, D., Molin, E., Rønnest, A., Stewart, A., & van Wee, B. (2018). Five things you should know about cost overrun. *Transportation Research Part A*, *118*, 174–190. https://doi.org/10.1016/j.tra.2018.07.013

Flyvbjerg, B. (2022). Top ten behavioral biases in project management: An overview. *Project Management Journal 52*(6). doi.org/10.1177/87569728211049046

Flyvbjerg, B., Budzier, A., Christodoulou, M. D., & Zottoli, M. (2024). *Uniqueness bias: Why it matters, how to curb it*. Saïd Business School Working Paper, University of Oxford.

Flyvbjerg, B., & Gardner, D. (2023). *How big things get done: The surprising factors behind every successful project—from home renovations to space exploration*. Currency.

Gajera, R. (2024). The role of machine learning in enhancing cost estimation accuracy: A study using historical data from project control software. *Letters in High Energy Physics, 2024*(495), 495–500.

Gao, N., & Touran, A. (2020). Cost overruns and formal risk assessment program in U.S. rail transit projects. *Journal of Construction Engineering and Management*, *146*(5), 05020004.

Ghimire, P., Pokharel, S., Kim, K., & Barutha, P. (2023, June). Machine learning-based prediction models for budget forecast in capital construction. In *Proceedings of the 2nd International Conference on Construction, Energy, Environment & Sustainability, Funchal, Portugal* (pp. 27–30).

Ghori, K., Abbasi, R., Awais, M., Imran, M., Ullah, A.& Szathmary, L. (2020). Performance analysis of different types of machine learning classifiers for non-technical loss detection. *IEEE Access, 8,* 16033–16048. https://doi.org/10.1109/ACCESS.2019.2962510

Gómez-Cabrera, A., Gutierrez-Bucheli, L., & Muñoz, S. (2024). Causes of time and cost overruns in construction projects: a scoping review. *International Journal of Construction Management, 24(*10), 1107–1125. https://doi.org/10.1080/15623599.2023.2252288

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Gosling, G. D., & Freeman, D. (2012). *Case study report: Portland MAX Airport Extension (MTI Report 10-26)*. Mineta Transportation Institute, College of Business, San José State University. https://transweb.sjsu.edu/research/portland-max-airport-extension-case-study-report

Grand View Research. (2023). *Public transportation market size, share & trends analysis report by mode (bus, light rail transit, metro rail, regional taxi), by application (city, rural), by region, and segment forecasts, 2023–2030*. Grand View Research. Retrieved from

https://www.grandviewresearch.com/industry-analysis/public-transportation-market-report

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*(1), 4-27. https://doi.org/10.1037/0033-295X.102.1.4

Guan, X., Servranckx, T., & Vanhoucke, M. (2024). Risk response budget allocation based on fault tree analysis and optimization. *Annals of Operations Research, 337*(1), 523–564. https://doi.org/10.1007/s10479-022-05155-8

Gudis, D. A., McCoul, E. D., Marino, M. J., & Patel, Z. M. (2023). Avoiding bias in artificial intelligence. *International Forum of Allergy & Rhinology*, *13*(3), 193–195. https://doi.org/10.1002/alr.23129

Hagendorff, T., Bossert, L. N., Tse, Y. F., & Singer, P. (2023). Speciesist bias in AI: how AI applications perpetuate discrimination and unfair outcomes against animals. *AI and Ethics, 3*(3), 717–734. https://doi.org/10.1007/s43681-022-00199-9

Harb, A., & Jayousi, R. (2012). Comparing neural network algorithm performance using SPSS and NeuroSolutions. *The 13th International Arab Conference on Information Technology (ACIT)*, Dec. 10–13. Retrieved from http://acit2k.org

Hashala, A. M., & Andrews, K. S. (2023). Improving project budgeting systems by developing machine learning models. *Journal of Applied Research, 19*.

Hashemi, S. T., Ebadati, O. M., & Kaur, H. (2020). Cost estimation and prediction in construction projects: A systematic review on machine learning techniques. *SN Applied Sciences, 2*(1703). https://doi.org/10.1007/s42452-020-03497-1

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction (2nd ed.).* Springer.

Hinton, G., & Sejnowski, T. J. (1986). Learning and relearning in Boltzmann machines. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations* (pp. 282–317). MIT Press.

Hinton, G., & Sejnowski, T. J. (Eds.). (1999). *Unsupervised learning: Foundations of neural computation*. MIT Press.

Hollmann, J. K. (2021). The case for parametric quantification of systemic risks for transportation projects. *International Technical Paper*. *Association for the Advancement of Cost Engineering (AACE).*

IBM Corp. (n.d.). *IBM MySupport*. IBM Support. Retrieved July 4, 2025, from https://www.ibm.com/mysupport/s/?language=en_US

IBM. (2024). *IBM SPSS Statistics brief guide (Version 30.0)*. Retrieved from https://www.ibm.com/support/pages/ibm-spss-statistics-30-documentation

Ika, L. A., & Munro, L. T. (2022). Tackling grand challenges with projects: Five insights and a research agenda for project management theory and practice. *International Journal of Project Management*, *40*(6), 601–607. https://doi.org/10.1016/j.ijproman.2022.05.008

International Tunnelling Association Working Group 13. (2003). *Underground or aboveground? Making the choice for urban mass transit systems*. International Tunnelling Association. Retrieved from https://tunnel.ita-aites.org

Iriarte, C., & Bayona, S. (2020). IT projects success factors: a literature review. *International Journal of Information Systems and Project Management*, *8*(2), 49-78.

Jha, K. K., Jha, R., Jha, A. K., Hassan, M. A. M., Yadav, S. K., & Mahesh, T. (2021). A brief comparison on machine learning algorithms based on various applications: A comprehensive survey. *2021 IEEE International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, 1–5, https://doi: 10.1109/CSITSS54238.2021.9683524

Johnston, M. P. (2014). Secondary data analysis: A method of which the time has come. *Qualitative and Quantitative Methods in Libraries, 3* (3), 619–626.

Jones, C., & Lichtenstein, B. (2008). Temporary inter-organizational projects: How temporal and social embeddedness enhance coordination and manage uncertainty. In S. Cropper, M. Ebers, C. Huxman, & P. S. Ring (Eds.), *The Oxford handbook of inter-organizational relations* (pp. 231–255). Oxford University Press.

Jordao, A. R., Costa, R., Dias, A. L., Pereira, L., & Santos, J. P. (2020). Bounded rationality in decision making: An analysis of the decision-making biases. *Business: Theory and Practice*, *21*(2), 654. https://doi.org/10.3846/btp.2020.11154

Joshi, A. P., & Patel, B. V. (2020). Data preprocessing: The techniques for preparing clean and quality data for data analytics process. *Oriental Journal of Computer Science and Technology, 13*(2-3), 78–81. https://doi.org/10.13005/ojcst13.0203.03

Kahneman, D., & Tversky, A. (1977*).* Intuitive prediction: Biases and corrective procedures. *Judgment Under Uncertainty: Heuristics and Biases. Vol. 185.* Cambridge University Press.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291. https://doi.org/10.2307/1914185

Kahneman, D., & Tversky, A. (1982). Intuitive prediction: Biases and corrective procedures. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 414–421). Cambridge University Press.

Kahneman, D., Rosenfield, A. M., Gandhi, L., & Blaser, T. (2016). *Reducing noise in decision making. Harvard Business Review, 94*(10), 38–46.

Kapila, D., Panwar, S., Maruga Raja, M. K. M., Mondal, T., Rafi, S. M., Singh, S. P., & Kumar, B. (2023). Applications of neural network-based plan cancer method for primary diagnosis of mesothelioma cancer. *BioMed Research International*, 2023, Article ID 3164166. https://doi.org/10.1155/2023/3164166

Kaur, A., Dhiman, A., & Singh, M. (2023). Comprehensive review: Security challenges and countermeasures for big data security in cloud computing. *2023 7th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech), Electronics, Materials Engineering & Nano-Technology (IEMENTech), 2023 7th International Conference On*, 1–6. https://doi.org/10.1109/IEMENTech60402.2023.10423449

Kelleher, J. D., Mac Namee, B., & D'arcy, A. (2020). *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*. MIT Press.

Khodabakhshian, A., Puolitaival, T., & Kestle, L. (2023). Deterministic and probabilistic risk management approaches in construction projects: A systematic literature review and comparative analysis. *Buildings, 13*(5), 1312. https://doi.org/10.3390/buildings13051312

Kim, H., & Lee, K. J. (2020). Data cleaning and preprocessing: A practical guide to data management in applied research. *Journal of Big Data, 7*(1), 1–21.

Kirk, R. S., & Levinson, H. S. (2004). Light rail transit systems: An overview. *Transportation Research Board, National Research Council*.

Kuhn, M., & Johnson, K. (2019). *Feature engineering and selection: A practical approach for predictive models.* CRC Press.

Kusonkhum, W., Srinavin, K., & Chaitongrat, T. (2023). The adoption of a big data approach using machine learning to predict bidding behavior in procurement management for a construction project. *Sustainability (2071-1050), 15*(17), 12836. https://doi.org/10.3390/su151712836

Kwon, H., & Kang, C. W. (2018). Improving project budget estimation accuracy and precision by analyzing reserves for both identified and unidentified risks. *Project Management Journal, 50*(1), 86–100. https://doi.org/10.1177/8756972818810963

Lamprou, A. & Vagiona, D. G. (2022). Identification and evaluation of success criteria and critical success factors in project success. *Global Journal of Flexible Systems Management, 23*(2), 237–253. https://doi.org/10.1007/s40171-022-00302-3

Leu, S.-S., Lu, C.-Y., & Wu, P.-L. (2023). Dynamic-Bayesian-network-based project cost overrun prediction model. *Sustainability, 15*(4570). https://doi.org/10.3390/su15054570

Liao, C., Aminudin, E., Mohd, S., & Yap, L. S. (2022). Intelligent risk management in construction projects: Systematic literature review. *IEEE Access, 10*, 72936–72952. https://doi.org/10.1109/ACCESS.2022.3189157

Locatelli, G., Invernizzi, D. C., & Brookes, N. J. (2017). Project characteristics and performance in Europe: An empirical analysis for large transport infrastructure projects. *Transportation Research Part A, 98*, 108–122. https://doi.org/10.1016/j.tra.2017.01.024

Löster, T. (2016). Determining the optimal number of clusters in cluster analysis. *The 10th International Days of Statistics and Economics, Prague, September 8-10, 2016,* 1078–1090. Retrieved from http://msed.vse.cz/

Lovallo, D., & Kahneman, D. (2003). Delusions of success: How optimism undermines executives' decisions. *Harvard Business Review, 81*(7), 56–63.

Love, P. E. D., Ika, L. A., & Sing, M. C. P. (2022). Does the planning fallacy prevail in social infrastructure projects? Empirical evidence and competing explanations. *IEEE Transactions on Engineering Management, 69*(6), 2588–2602. https://doi.org/10.1109/TEM.2019.2944161

Mæhlen, J., Bekkevold, J. P., Welde, M., & Olsson, N. O. E. (2024). Reducing cost overruns through data-driven methods used in uncertainty analyses. *Procedia Computer Science*, *239*, 209–216. https://doi.org/10.1016/j.procs.2024.06.164

Majerova, I., & Nevima, J. (2017). The measurement of human development using the Ward method of cluster analysis. *Journal of International Studies, 10*(2), 239–257. https://doi.org/10.14254/2071-8330.2017/10-2/17

Manning, C., (2020). Artificial Intelligence definitions, *Stanford University*. https://hai.stanford.edu/sites/default/files/2020-09/AI-Definitions-HAI.pdf

Mariani, C., Navrotska, Y., & Mancini, M. (2023). Unsupervised machine learning for project stakeholder classification: Benefits and limitations. *Project Leadership and Society, 4*, 100093. https://doi.org/10.1016/j.plas.2023.100093

Mariani, C., Cellerino, F., Araya Aliaga, E. H., Atencio, E., & Mancini, M. (2025). The C-BA method: Enhancing megaproject forecasting through the "Fifth Hand" principle.

International Journal of Managing Projects in Business, 18(8), 50–78. https://doi.org/10.1108/IJMPB-11-2024-0281

Martins, J. P., Correia, J., Ljubinković, F., & Simões da Silva, L. (2023). Cost optimisation of steel I-girder cross-sections using genetic algorithms. *Structures, 55*, 379–388. https://doi.org/10.1016/j.istruc.2023.06.030

Marzoughi, F., Arthanari, T., & Askarany, D. (2018). A decision support framework for estimating project duration under the impact of weather. *Automation in Construction*, *87*, 287-296

McNeil, A. J., Frey, R., & Embrechts, P. (2015). *Quantitative risk management: Concepts, techniques, and tools.* Princeton University Press.

McSweeney, B. (2021). Fooling ourselves and others: Confirmation bias and the trustworthiness of qualitative research – Part 1 (the threats). *Journal of Organizational Change Management*, *34*(5), 1063–1075. https://doi.org/10.1108/JOCM-04-2021-0117

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), 1–35. https://doi.org/10.1145/3457607

Meng, C., Qu, D., & Duan, X. (2024). Cost estimation of metro construction projects using interpretable machine learning. *Journal of Computing in Civil Engineering, 38*(6), 04024038. https://doi.org/10.1061/JCCEE5.CPENG-6018

Merrow, E. W. (2011). *Industrial megaprojects: Concepts, strategies, and practices for success*. John Wiley & Sons.

Milat, M., Knezić, S., & Sedlar, J. (2022). Application of a genetic algorithm for proactive resilient scheduling in construction projects. *Designs, 6*(1), 16. https://doi.org/10.3390/designs6010016

Min, A. (2023). Artificial intelligence and bias: Challenges, implications, and remedies. *Journal of Social Research, 2*(11), 3808–3817. https://doi.org/10.55324/josr.v2i11.1477

Mohammadzadeh Gonabadi, A., Fallahtafti, F., Heselton, J., Myers, S. A., Siu, K.-C., & Boron, J. B. (2025). Modeling dual-task performance: Identifying key predictors using artificial neural networks. *Biomimetics*, *10*(6), Article 351. https://doi.org/10.3390/biomimetics10060351

Morrow, R., & Brown, D. (1994). *Critical theory and methodology*. SAGE Publications, Inc.

Moura, H. P., & Ribeiro, J. L. D. (2021). Project managers' change readiness and its impact on project success. *Journal of Project Management, 36*(4), 12-25. https://doi.org/10.1234/jpm.v36i4.5678

Müller, R., & Jugdev, K. (2012). Critical success factors in projects: Pinto, Slevin, and Prescott – the elucidation of project success. *International Journal of Managing Projects in Business*, *5*(4), 757–775. https://doi.org/10.1108/17538371211269040

Mulyono, R., Ndini, A. S., Kharisma, G., & Heikal, J. (2023). Segmentation K-means clustering model with SPSS program: Case study customer the Park Mall Sawangan. *Syntax Literate: Journal Ilmiah Indonesia, 8*(2), 1301-1314. https://doi.org/10.36418/syntax-literate.v8i2.11429

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. The MIT Press.

Musick, N. (2022). *Government spending on public transportation and other infrastructure*. Congressional Budget Office. https://www.cbo.gov/publication/58086

National Transit Database. (2024, October). *2023 Annual data publications guide*. Office of Budget and Policy. U.S. Department of Transportation. Retrieved from https://www.transit.dot.gov/sites/fta.dot.gov/files/2024-10/2023%20Annual%20NTD%20Data%20Publications%20Guide_1.pdf

Naeem, S., Saeed, M. A., Khan, A., Mehmood, R., Khan, W., Ullah, F., & Jan, A. A. (2023). An unsupervised machine learning algorithm: Comprehensive review. *International Journal of Computing and Digital Systems, 12*(1), 291–303. https://doi.org/10.12785/ijcds/120125

Nenni, M. E., De Felice, F., De Luca, C., & Forcina, A. (2024). How artificial intelligence will transform project management in the age of digitization: a systematic literature review. *Management Review Quarterly: Systematic Literature Reviews, Meta-Analyses, and Replication Studies*, 1–48. https://doi.org/10.1007/s11301-024-00418-z

Nieto-Rodriguez, A. (2021). *Harvard business review project management handbook: how to launch, lead, and sponsor successful projects*. Harvard Business Press.

Noteboom, C., Ofori, M., & Shen, Z. (2021). *The applications of artificial intelligence in managing project processes and targets: A systematic analysis*. *Journal of International Technology and Information Management, 31*(3), 77-100.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M. E., ... & Staab, S. (2020). Bias in data-driven artificial intelligence systems - An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3), e1356.

Odeck, J. (2019). Variation in cost overruns of transportation projects: an econometric meta-regression analysis of studies reported in the literature. *Transportation: Planning - Policy - Research - Practice, 46*(4), 1345–1368. https://doi.org/10.1007/s11116-017-9836-5

Office, T. S. (2017). *Managing successful projects with PRINCE2*. The Stationery Office.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy.* Crown Publishing Group.

Ottaviani, F. M., De Marco, A., Rafele, C., & Castelblanco, G. (2024). Risk perception-based project contingency management framework. *Systems, 12*(3), 93. https://doi.org/10.3390/systems12030093

Ottawa Light Rail Transit Commission. (2022). *Report of the Ottawa Light Rail Transit Public Inquiry: Final report*. Retrieved from https://www.ottawalrtpublicinquiry.ca

Padalkar, M., & Gopinath, S. (2016). Six decades of project management research: Thematic trends and future opportunities. *International Journal of Project Management*, *34*, 1305-1321. https://doi.org/10.1016/j.ijproman.2016.06.006

Pang, D.-J., Shavarebi, K., & Ng, S. (2022). Development of machine learning models for prediction of IT project cost and duration. *2022 IEEE 12th Symposium on Computer Applications & Industrial Electronics (ISCAIE), Computer Applications & Industrial Electronics (ISCAIE), 2022 IEEE 12th Symposium On*, 228–232. https://doi.org/10.1109/ISCAIE54458.2022.9794529

Park, J. (2021). Curbing cost overruns in infrastructure investment: Has reference class forecasting delivered its promised success? *European Journal of Transport & Infrastructure Research, 21*(2), 120–136. https://doi.org/10.18757/ejtir.2021.21.2.5504

Pașcu, P. (2023). Monte Carlo simulation in project risk management. *USV Annals of Economics & Public Administration, 23*(2). Ștefan cel Mare University of Suceava.

Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., & Larochelle, H. (2021). Improving reproducibility in machine learning research (a

report from the NeurIPS 2019 reproducibility program). *Journal of Machine Learning Research, 22*(164), 1–20.

Pinto, J. K. (2022). Avoiding the inflection point: Project management theory and research after 40 years. *International Journal of Project Management*, *40*(1). https://doi.org/10.1016/j.ijproman.2021.11.002

Pirmohammadi, G., & Mast Zohouri, M. (2020). Comparison of artificial neural network and SPSS model in predicting customers churn of Iran's insurance industry. *International Journal of Computer Applications, 176*(32), 14–21. https://doi.org/10.5120/ijca2020920345

PosPieszny, P. (2017). Application of data mining techniques in project management: An overview. *CEA Annals*, *43*, 199–220.

Prater, J., Kirytopoulos, K., & Ma, T. (2017). Optimism bias within the project management context. *International Journal of Managing Projects in Business*, *10*(2), 370–385. https://doi.org/10.1108/IJMPB-07-2016-0063

Prigent, C., Cudennec, L., Costan, A., & Antoniu, G. (2022). A methodology to build decision analysis tools applied to distributed reinforcement learning. *ScaDL 2022: Scalable Deep Learning over Parallel and Distributed Infrastructures - An IPDPS Workshop*, Lyon/Virtual, France, 1–10. https://hal.science/hal-03613558

Project Management Institute. (2021). *A guide to the project management body of knowledge (PMBOK guide) (7th ed.).* Project Management Institute.

Rana, S. A., Azizul, Z. H., & Awan, A. A. (2023). A step toward building a unified framework for managing AI bias. *PeerJ Computer Science, 9*(e1630). https://doi.org/10.7717/peerj-cs.1630

Robertson, S., & Williams, T. (2006). Understanding project failure: Using cognitive mapping in an insurance project. *Project Management Journal*, *37*(4), 55–71.

Rodrigue, L., Soliz, A., Manaugh, K., Kestens, Y., & El-Geneidy, A. (2024). Opinions matter: Contrasting perceptions of major public transit projects in Montréal, Canada. *Transport Policy*, *157*, 34–45. https://doi.org/10.1016/j.tranpol.2024.08.006

Sandell, H. (2020, March). Forecasting indirect costs in Finnish public transport infrastructure projects: Applications with machine learning models. *Finnish Transport Infrastructure Agency.* Retrieved from https://www.vayla.fi

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science, 2*(160). https://doi.org/10.1007/s42979-021-00592-x

Selim, T., Elshaarawy, M. K., Elkiki, M., & Eltarabily, M. G. (2024). Estimating seepage losses from lined irrigation canals using nonlinear regression and artificial neural network models. *Applied Water Science, 14*(90). https://doi.org/10.1007/s13201-024-02142-1

Serin, H., Körez, M. K., Tekin, M. E., & Siren, S. (2023). Classification of forest fires in European countries by clustering analysis techniques. *Sakarya University Journal of Science, 27*(5), 987–1001.

Sghir, N., Adadi, A., & Lahmer, M. (2023). Recent advances in predictive learning analytics: A decade systematic review (2012-2022). *Education and Information Technologies, 28*(7), 8299–8333. https://doi.org/10.1007/s10639-022-11536-0

Siemiatycki, M. (2010). Managing optimism biases in the delivery of large infrastructure projects: A corporate performance benchmarking approach. *European Journal of Transport and Infrastructure Research*, *10*(1). https://doi.org/10.18757/ejtir.2010.10.1.2866

Simonaitis, A., Daukšys, M., & Mockienė, J. (2023). A comparison of the project management methodologies PRINCE2 and PMBOK in managing repetitive construction projects. *Buildings, 13*(7), 1796. https://doi.org/10.3390/buildings13071796

Sjögren, E., & Norgren, J. (2023). *Cost overrun in Swedish infrastructure transport projects: An analysis of cost overrun in Swedish infrastructure transport projects between 2010–2022* (Bachelor thesis, Jönköping University).

Statista. (2023). *Urbanization*. Retrieved from https://www-statista-com.eu1.proxy.openathens.net/study/136786/urbanization/

Sterman, J. D. (2000). *Business dynamics: Systems thinking and modeling for a complex world.* McGraw-Hill Education.

Tabassi, A. A., & Bakar, A. H. A. (2018). *Quantitative research method. A complete guide to academic research* In Built Environment and Engineering (Penerbit USM).

Taherdoost, H. (2021). Data collection methods and tools for research; a step-by-step guide to choose data collection technique for academic and business research projects. *International Journal of Academic Research in Management (IJARM)*, *10*(1), 10-38.

Tajziyehchi, N., Moshirpour, M., Jergeas, G., & Sadeghpour, F. (2021). A methodology and tool for the predictive analysis of cost growth in construction projects. In *Proceedings of the 2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI)* (pp. 240–247). https://doi.org/10.1109/IRI51335.2021.00039

Takagi, N., Varajão, J., Ventura, T., Ubialli, D., & Silva, T. (2024). *Managing success criteria and success factors in a BPM project: An approach using PRINCE2 and Success*

Management on the public sector. *Cogent Business & Management, 11*(1), 2336273. https://doi.org/10.1080/23311975.2024.2336273

Talaei Khoei, T., & Kaabouch, N. (2023). A comparative analysis of supervised and unsupervised models for detecting attacks on the intrusion detection systems. *Information, 14*(2), 103. https://doi.org/10.3390/info14020103

Thore Olsson, A.-C., Johannesson, U., & Schweizer, R. (2018). Decision-making and cost deviation in new product development projects. *International Journal of Managing Projects in Business, 11*(4), 1066–1085. https://doi.org/10.1108/IJMPB-02-2018-0029

Tofallis, C. (2015). A better measure of relative prediction accuracy for model selection and model estimation. *Journal of the Operational Research Society, 66*(8), 1352-1362. https://www.jstor.org/stable/24505756

Topaloglu, F. (2024). A hybrid approach based on k-means and SVM algorithms in the selection of appropriate risk assessment methods for sectors. *PeerJ Computer Science, 10,* e2198. https://doi.org/10.7717/peerj-cs.2198

Uddin, S., Ong, S., & Lu, H. (2022). *Machine learning in project analytics: A data-driven framework and case study*. *Scientific Reports, 12*, 15252. https://doi.org/10.1038/s41598-022-19728-x

U.S. Bureau of Economic Analysis. (n.d.). *Homepage*. Retrieved from https://www.bea.gov

U.S. Census Bureau. (n.d.). *Homepage*. Retrieved from https://www.census.gov

U.S. Congress. (2005). *Safe, accountable, flexible, efficient transportation equity act: A legacy for users (SAFETEA-LU), Pub. L. No. 109-59*. U.S. Government Publishing Office. https://www.govinfo.gov/content/pkg/PLAW-109publ59/html/PLAW-109publ59.htm

U.S. Department of Commerce, National Oceanic and Atmospheric Administration, National Center for Environmental Information. (2023). *Homepage*. Retrieved from https://www.ncei.noaa.gov

U.S. Department of Transportation (2015). *Expand public transportation systems and offer incentives*. https://www.transportation.gov/mission/health/Expand-Public-Transportation-Systems-and-Offer-Incentives

U.S. Department of Transportation, Federal Transit Administration. (2023). *Public transportation funding and financing*. https://www.transit.dot.gov

U.S. Government Accountability Office (GAO). (2009). *GAO Cost Estimating and Assessment Guide: Best practices for developing and managing capital program costs* (GAO-09-3SP). Washington, DC: Author. https://www.gao.gov/products/GAO-09-3SP

U.S. Government Accountability Office. (2010). *Public transportation: Federal funding for the Capital Investment Program (GAO-10-630).* U.S. Government Accountability Office. https://www.gao.gov/products/gao-10-630

U.S. Government Accountability Office (GAO). (2019). *Rail transit: Federal Transit Administration could improve information on estimating project costs* (GAO-19-562). Washington, DC: Author. https://www.gao.gov/products/GAO-19-562

U.S. Government Accountability Office. (2020). *Cost estimating and assessment guide: Best practices for developing and managing program costs* (GAO-20-195G). https://www.gao.gov/products/gao-20-195g

Van Wee, B. (2007). Large infrastructure projects: a review of the quality of demand forecasts and cost estimations. *Environment and Planning B: Planning and Design*, *34*(4), 611–625.

Vartanian, T. P. (2010). *Secondary data analysis*. Oxford University Press.

Vela, M. B., Erondu, A. I., Smith, N. A., Peek, M. E., Woodruff, J. N., & Chin, M. H. (2022). Eliminating explicit and implicit biases in health care: Evidence and research needs. *Annual Review of Public Health*, *43*, 477–501. https://doi.org/10.1146/annurev-publhealth-052620-103528

Vinolia, M., & Mudau, J. (2024). Evaluating municipal budgeting and financial allocations for capital projects within the city of Tshwane municipality. *African Journal of Development Studies*, *14*(1), 135. https://doi.org/10.31920/2634-3649/2024/v14n1a7

Voulgaris, C. T. (2017). *Crystal balls and black boxes: Optimism bias in ridership and cost forecasts for new starts rapid transit projects*. University of California, Los Angeles.

Wang, J. (2020). Construction of risk evaluation index system for power grid engineering cost by applying WBS-RBS and membership degree methods. *Mathematical Problems in Engineering*, 1–9. https://doi.org/10.1155/2020/6217872

Wirsansky, E. (2020). *Hands-on genetic algorithms with Python: applying genetic algorithms to solve real-world deep learning and artificial* intelligence *problems*. Packt Publishing Ltd.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques* (3rd ed.). Morgan Kaufmann.

Woetzel, J., Garemo, N., Mischke, J., Hjerpe, M., & Palter, R. (2017). *Bridging infrastructure gaps: Has the world made progress?* McKinsey Global Institute.

Yim, R., Castaneda, J., Doolen, T., Tumer, I., & Malak, R. (2015). A study of the impact of project classification on project risk indicators. *International Journal of Project Management, 33*(4), 863–876. https://doi.org/10.1016/j.ijproman.2014.10.005

Zhang, M., & Xia, C. (2017). A loose wavelet nonlinear regression neural network load forecasting model and error analysis based on SPSS. *International Journal of Information Technology and Computer Science (IJITCS)*, *9*(4), 24–30. https://doi.org/10.5815/ijitcs.2017.04.04

Zhang, N., Shen, S.-L., Zhou, A., & Xu, Y.-S. (2019). Investigation on performance of neural networks using quadratic relative error cost function. *IEEE Access, 7*, 106642–106652. https://doi.org/10.1109/ACCESS.2019.2930520

Zheng, A., & Casari, A. (2018). *Feature engineering for* machine *learning: Principles and techniques for data scientists*. O'Reilly Media.

Zhou, G., Etemadi, A., & Mardon, A. (2022). Machine learning-based cost predictive model for better operating expenditure estimations of U.S. light rail transit projects. *Journal of Public Transportation*, *24*. https://doi.org/10.1016/j.jpubtr.2022.100031

Appendix A Data Sources

The data sources for acquiring data for the study used credible websites. Within each source, the search capability accessed relevant documents.

Project Variables	Sources
Project Name Project Budget Project Actual Cost Project Duration Project completion or opening date Track distance and underground track distance Number of new transit stations and number of underground stations Number of new vehicles purchased Project type: Initial or extension Number of new maintenance facilities	https://www.transit.dot.gov/ https://www.transit.dot.gov/funding/grant-programs/capital-investments/and-after-studies-new-starts-projects
Inflation Calculator	https://www.bls.gov/data/inflation_calculator.htm
Vehicle Brand	https://www.apta.com/research-technical-resources/transit-statistics/vehicle-database/
Economic Region	https://www.transit.dot.gov/about/regional-offices/regional-offices
Metro density	https://www.census.gov/
Location temperature	https://www.ncei.noaa.gov/cdo-web/
State GDP	https://www.bea.gov/data/gdp/gdp-state
Additional Sources and Triangulation	
Government Accountability Office	https://www.gao.gov/ https://www.gao.gov/reports-testimonies?f%5B0%5D=topic%3A156&f%5B1%5D=topic%3A161&f%5B2%5D=topic%3A286

	https://www.transit.dot.gov/regulations-and-programs/access/ccam/government-accountability-office-gao-reports

Metro Transit Authorities

Atlanta	https://atltransit.ga.gov/
Baltimore	https://www.mta.maryland.gov/
Buffalo	https://www.nfta.com/
Charlotte	https://www.charlottenc.gov/CATS/Home
Chicago	metra.com
Dallas	dart.org
Denver	rtd-denver.com
Detroit	https://detroitmi.gov/departments/detroit-department-transportation
Houston	ridemetro.org
Jacksonville	https://www.jtafla.com/
Los Angeles	https://www.metro.net/
Memphis	https://www.matatransit.com/
Miami	https://www.masstransitmag.com/home/company/11175417/miami-dade-transit-mdt
Minneapolis	https://www.metrotransit.org/
New Jersey	njtransit.com
New York	mta.info
Norfolk	gohrt.com
Orlando	sunrail.com
Phoenix	valleymetro.org
Pittsburgh	portauthority.org
Portland	trimet.org
Sacramento	https://www.sacrt.com/
Saint Louis	https://www.metrostlouis.org/
San Franscisco	https://www.bart.gov/
San Diego	sdmts.com
San Juan	dtop.gov.pr
San Jose	https://www.vta.org/
Santa Clara	https://www.vta.org/
Salt Lake City	rideuta.com
Seattle	soundtransit.org
Washington	wmata.com

Appendix B Data Instrument

The data instrument is a spreadsheet that captures the variables.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Project Name	Original budget	Actual cost	Variance %	CPI - adjusted actual cost	Project completion (year)	Project duration (years)	Initial or extension	Track distance (miles)	Underground track distance	Number of new stations	Number of underground stations	Vehicles purchased	Vehicle brand	Number of new maintenance faciltites	Economic region	Metropolitan density	Low temp	High temp	State GDP
2																				
3																				
4																				
5																				

Appendix C Archived Data Retrieval Audit Findings and Summary Form

Researcher's Name: Paul Boudreau

Title of the Study: Evaluating the Effectiveness of Machine Learning Models to Predict Budget Overruns for U.S. Mass Transit Projects: A Correlational Study

Dissertation Chair's Name: Dr. Sharon Kimmel

Instructions: Please answer all the questions below.

1a. How many reviewers participated in the data set audit? For each reviewer, please enter name, relationship with the researcher, and, as appropriate, related audit skills.

1. Spouse. Former Manager, Health Canada, Natural Health Products Directorate, responsible for overseeing license issuing. Currently, Office Manager for an IT Consulting firm, reviewing AI research and training development applied to project management.
2. Neighbor. Business Intelligence analyst and software developer.

1b. How many cases were audited by each expert? Also include the total number of cases within the dataset.

The dataset consists of 108 projects, featuring 16 raw variables and 2 calculated variables. Expert 1 performed a random audit on 33 of the 108 datasets (30%). The audit examined 11 data fields related to project characteristics.
Expert 2 audited 100% of the dataset for the 5 external variables and the two calculated variables (Inflation-adjusted cost and temperature range).

2. Please summarize audit findings and, if applicable, corrective actions implemented based on audit findings. If no changes or corrective actions were made, then indicate 'no changes.'

Expert 1. For the actual cost data field, one entry of $207.8 million was rounded to 208. This was corrected to 207.8. An entry of $402.32 million was corrected to 401.32 million. Based on the findings, an additional 30 actual cost data fields were audited, and no discrepancies were identified.

Expert 2. Found no discrepancies in 4 data fields. In the Metro density data field, the number 8109.1 was entered as 8901.1. The entry was corrected. There were no other discrepancies, and this field was subject to a 100% audit.

3. For the actual dissertation study, are you revising any of the data collection procedures, sources, or research procedures that you wrote in the Dissertation Proposal?
   ___ Yes __X_ No

   (***Please answer yes or no. If you answer yes, please summarize refinements made.***

No

4. What other changes (from the Dissertation Proposal proposed study plan) for the actual dissertation research study are you proposing to implement based on what you learned from the data set audit? If no changes, then indicate 'no changes.'

No changes

Signature of Dissertation Chair: Sharon R. Kimmel, Ph.D., ASCE

Date: 20 June 2025

Signature of the Dissertation Candidate: Paul Boudreau

Date: 23 June 2025

Appendix D Data Collection Results

Project	Project Completion (year)	Actual Cost ($Millions)	Inflation Adj Actual Cost (2024)	Duration	Distance	Underground miles	New Stations	Underground stations	New maintenance fac.	Initor Ext	Vehicles purchased	Vehicle Brand Nominal	REGION	State GDP 2024	Metro density	Min temp 1991-2020	Max temp	Range temp
Atlanta North Extension	1999	463.18	868.58	5	1.9	0	2	0	0	0	28	1	4	882535	3685.7	37.6	88.7	51.1
Baltimore Phase 1 Rapid Transit	1987	1289.00	3525.26	15	7.6	5	13	7	1	1	72	2	3	542766	7235.8	27.3	86.6	59.3
Baltimore Ext. to John Hopkins	1995	302.00	620.93	6	6	6	6	6	0	0	0	2	3	542766	7235.8	27.3	86.6	59.3
Baltimore BWI, Hunt Valley Penn Station Ext.	1997	106.00	207.40	4	7.5	0	8	0	0	0	18	2	3	542766	7235.8	27.3	86.6	59.3
Baltimore Central LRT Double Tracking	2006	151.60	237.09	5	9.4	0	4	0	0	0	0	2	3	542766	7235.8	27.3	86.6	59.3
Boston Green Line extension (GLX)	2022	2297.62	2443.22	10	4.3	3	7	4	1	1	24	7	1	780666	13976.7	25.6	79.6	54.0
Buffalo Minimum LR Rapid Transit	1986	722.00	1891.01	8	6.4	1.2	14	8	1	1	27	7	2	2297028	6893.6	21.4	78.3	56.9
Central Florida Commuter Rail Transit – Initial Operating Segment	2014	357.20	480.10	7	32	0	12	0	1	1	14	3	4	1705565	2780.4	51.6	91.5	39.9
Charlotte South Corridor Light Rail	2007	462.75	695.34	5	9.6	0	15	0	1	1	20	4	4	839122	2836.9	33.7	88.6	54.9
Charlotte LYNX Blue Line Northeast	2018	1160.08	1457.32	6	9.3	0	11	0	1	0	22	4	4	839122	2836.9	33.7	88.6	54.9
Chicago Southwest Extension (Orange Line)	1983	500.00	1577.77	5	9.2	0	8	0	1	1	54	2	5	1137244	12059.8	21.6	82.5	60.9
Chicago SW Transitway	1989	185.50	464.27	4	11	0	2	0	0	0	13	2	5	1137244	12059.8	21.6	82.5	60.9

Metra North Central	2006	216.80	339.06	6	18.6	0	5	0	0	0	0	5	5	1137244	12059.8	21.6	82.5	60.9
Metra SW Corridor	2006	185.30	289.80	5	11	0	3	0	1	0	0	5	5	1137244	12059.8	21.6	82.5	60.9
Metra UP West	2006	106.10	165.94	5	8.5	0	2	0	1	0	0	5	5	1137244	12059.8	21.6	82.5	60.9
Dallas South Oak Cliff	1996	295.60	588.23	17	20	0	13	0	1	1	48	5	6	2709393	3841.1	38.2	94.3	56.1
Dallas North Central LRT	2002	437.30	762.93	6	12.5	0	9	0	1	0	21	5	6	2709393	3841.1	38.2	94.3	56.1
Dallas Northwest-Southeast	2010	1406.20	2024.85	10	20.9	0.29	16	0	1	0	18	5	6	2709393	3841.1	38.2	94.3	56.1
Denver Central Corridor D	1994	116	244.56	4	5.3	0	14	0	1	0	11	4	8	553323	4674.3	18.9	86.9	68.0
Denver Southwest Line	2000	176.32	319.81	4	8.7	0	5	0	1	0	14	4	8	553323	4674.3	18.9	86.9	68.0
Denver Central Platte Valley	2002	47.8	83.39	2	1.8	0	4	0	0	0	0	4	8	553323	4674.3	18.9	86.9	68.0
Denver Southeast Corridor	2006	850.80	1330.61	8	19.1	0	13	0	1	0	34	4	8	553323	4674.3	18.9	86.9	68.0
Denver West Rail Line LRT	2013	709.80	961.24	9	12.1	0.5	12	0	0	0	34	4	8	553323	4674.3	18.9	86.9	68.0
Denver Eagle P3	2016	2043.14	2670.84	6	30.2	0	13	0	1	1	44	4	8	553323	4674.3	18.9	86.9	68.0
Denver Southeast extension	2019	233.1	286.28	3	2.3	0	3	0	0	0	8	4	8	553323	4674.3	18.9	86.9	68.0
Detroit Central Automated Transit	1987	215.00	588.00	15	2.9	0	12	0	1	1	12	7	5	706616	4606.8	22.1	81.8	59.7
Honolulu Skyline Segment 1	2023	1400.00	1440.43	13	10.75	0	9	0	1	1	12	7	9	115627	1694.2	67.7	87.9	20.2
Houston Red Line	2004	324.00	537.34	3	7.5	0	16	0	1	1	18	4	6	2709393	3594.4	46.7	89.9	43.2
Houston Red Line North extension	2013	756.00	1023.81	4	5.3	0	8	0	0	0	22	4	6	2709393	3594.4	46.7	89.9	43.2
Houston Purple Line (Southeast Line)	2015	822.91	1098.04	6	6.5	0	13	0	1	1	29	4	6	2709393	3594.4	46.7	89.9	43.2
Houston Green Line	2015	279	372.28	6	3.3	0	7	0	0	1	10	5	6	2709393	3594.4	46.7	89.9	43.2
Jacksonville Skyway Extension	1999	183.00	343.17	15	2.5	0	8	0	0	0	16	3	4	1705565	1270.7	46.7	89.9	43.2

New Jersey Hudson-Bergen MOS 1 & 2	2006	886.50	1386.44	6	6.1	0	7	0	0	0	23	5	2	846587	1263.0	28.0	84.5	56.5
Los Angeles Blue Line	1990	877.00	2532.35	5	22	0	22	0	1	1	54	4	9	4103124	8304.2	49.5	74.6	25.1
Los Angeles Red Line MOS 1	1993	1450.00	3138.73	6	4.4	4.4	5	5	1	1	30	1	9	4103124	8304.2	49.5	74.6	25.1
Los Angeles Red Line MOS 2	1999	1736.00	3255.44	6	6.7	6.7	8	8	0	0	42	1	9	4103124	8304.2	49.5	74.6	25.1
Los Angeles Red Line MOS 3	2000	1341.00	2432.34	6	6.3	6.3	3	3	0	0	32	1	9	4103124	8304.2	49.5	74.6	25.1
Los Angeles Metro Gold Line Eastside Extension	2009	899.10	1314.02	5	6	1.8	8	2	0	0	10	1	9	4103124	8304.2	49.5	74.6	25.1
LA to Perris Valley started 2013	2016	248.30	324.58	9	24	0	4	0	1	0	0	7	9	4103124	8304.2	49.5	74.6	25.1
Los Angeles Expo line now part of E to Santa Monica P	2012	930	1278.36	6	8.7	0	10	0	0	0	23	5	9	4103124	8304.2	49.5	74.6	25.1
Los Angeles Expo line Phase 2	2016	1500	1960.83	5	6.6	0	7	0	0	0	44	5	9	4103124	8304.2	49.5	74.6	25.1
Los Angeles Crenshaw (Metro K Line)	2022	2100.00	2233.08	8	8.5	1	8	3	1	0	28	5	9	4103124	8304.2	49.5	74.6	25.1
Los Angeles Regional Connector	2023	1795.00	1846.84	11	1.9	1.9	3	3	0	0	4	5	9	4103124	8304.2	49.5	74.6	25.1
Memphis Med Center LRT	2004	74.58	123.69	6	2	0	6	0	0	0	4	7	4	549709	2131.8	35.4	90.9	55.5
Miami Dade County HRT	1988	1341.00	3512.25	8	21	0	21	0	1	1	136	1	4	1705565	10774.7	62.8	90.2	27.4
Miami DPM Starter Line	1986	159.00	454.13	3	1.9	0	8	0	1	1	12	3	4	1705565	10774.7	62.8	90.2	27.4
Miami Omni &Brickell Ext	1994	248.00	522.85	6	2.5	0	12	0	0	0	17	3	4	1705565	10774.7	62.8	90.2	27.4
Minneapolis Hiawatha Line	2004	713.00	1182.48	7	12	1.5	17	1	1	1	27	3	5	500851	7962.1	12.2	81.0	68.8
Minneapolis Northstar Corridor Rail	2009	308.50	450.87	2	1	0	7	0	1	0	18	3	5	500851	7962.1	12.2	81.0	68.8
Minneapolis St Paul	2014	926.50	1245.29	4	9.7	0	18	0	1	0	47	4	5	500851	7962.1	12.2	81.0	68.8

New York Second Avenue Subway Phase 1	2017	4450.00	5696.98	10	2.3	2.3	3	3	0	0	0	6	2	2297028	29303.2	28.5	81.2	52.7
Newark Rail Link MOS-1	2006	207.70	324.83	4	1	0.15	5	0	0	0	4	5	2	846587	12903.8	28.0	84.5	56.5
Norfolk The Tide Light Rail Project	2011	314.60	439.97	9	7.3	0	11	0	1	1	9	4	3	764475	4467.5	34.6	88.7	54.1
Phoenix, Central Phoenix/East Valley	2008	1405.00	2109.26	5	19.7	0	28	0	1	1	36	5	9	552167	3104.5	44.3	102.6	58.3
Central Mesa Light Rail Extension	2016	196.70	257.13	6	3.1	0	4	0	0	0	0	4	9	552167	3635.5	46.8	105.2	58.4
Phoenix Northwest	2016	326.6	426.15	3	3.2	0	3	0	0	0	0	4	9	552167	3104.5	44.3	102.6	58.3
Phoenix ext Phase 2	2024	401.32	401.32	3	1.6	0	3	0	0	0	3	4	9	552167	3104.5	44.3	102.6	58.3
Phoenix Mesa Gilbert road ext	2019	184	225.98	3	1.9	0	2	0	0	0	11	4	9	552167	3635.5	46.8	105.2	58.4
Pittsburgh South Hills Reconstruction	1987	622.00	1701.09	9	17	0	17	0	1	1	55	4	3	1024206	5471.3	23.8	81.3	57.5
Pittsburgh North Shore	2012	510.40	701.59	9	1.2	0.52	2	1	0	0	0	4	3	1024206	5471.3	23.8	81.3	57.5
Pittsburgh Stage II reconstruction	2005	385.00	617.42	9	5.5	0	10	0	0	0	28	4	3	1024206	5471.3	23.8	81.3	57.5
Portland Banfield Corridor	1986	266.00	696.69	6	15.1	0	27	0	1	1	26	3	10	331029	4889.5	36.4	79.5	43.1
Portland Westside/Hillsboro MAX	1998	963.52	1855.35	6	17.7	2.9	20	1	1	0	36	4	10	331029	4889.5	36.4	79.5	43.1
Portland Airport MAX Extension	2001	128.80	230.05	3	5.5	0	4	0	0	0	6	4	10	331029	4889.5	36.4	79.5	43.1
Portland Interstate MAX	2004	350.00	580.46	4	5.8	0	10	0	0	0	17	4	10	331029	4889.5	36.4	79.5	43.1
Portland Oregon (Washington and Multinomah)	2009	162.00	236.76	3	14.7	0	5	0	1	1	4	4	10	331029	11280.7	27.0	85.7	58.7
Portland Green Line	2009	576.00	841.81	5	14.5	0.03	8	0	0	0	22	4	10	331029	4889.5	36.4	79.5	43.1
Portland-Milwaukee Rail	2015	1490.35	1988.64	4	7.3	0	10	0	0	0	18	4	10	331029	4889.5	36.4	79.5	43.1

Project (Orange Line)																	
Portland Airport MAX Extension	2024	215.00	215.00	3	10.6	0	10	0	1	0	4	4	10	331029	4889.5	36.4	79.5	43.1
Sacramento Starter Line (Stage I LRT)	1987	188.00	514.16	15	18.3	0	29	0	1	1	27	4	9	4103124	5323.4	38.5	91.5	53.0
Sacramento Gold Exr Mather Field	1998	37	71.25	2	2.3	0	2	0	0	0	0	5	9	4103124	5323.4	38.5	91.5	53.0
Sacramento South LRT (Phase 1)	2003	222.00	380.16	6	6.3	0	7	0	1	1	24	6	9	4103124	5323.4	38.5	91.5	53.0
Sacramento Gold Line Sunrise	2004	89	147.6	2	3.7	0	3	0	0	0	0	7	9	4103124	5323.4	38.5	91.5	53.0
Sacramento South LRT (Phase 2)	2015	270.00	360.27	8	4.9	0	4	0	0	0	0	4	9	4103124	5323.4	38.5	91.5	53.0
Salt Lake North-South Line	1999	312.00	585.08	9	15	0	16	0	1	0	23	4	8	300904	1810.1	26.0	89.9	63.9
Salt Lake City University line	2003	207.8	355.85	3	3.74	0	7	0	0	0	7	4	8	300904	1810.1	26.0	89.9	63.9
Salt Lake City Mid-Jordan Light Rail Project	2011	509.80	712.96	4	10.6	0	9	0	0	0	28	4	8	300904	1810.1	26.0	89.9	63.9
The Weber County to Salt Lake	2008	614.00	921.77	3	44	0	9	0	1	1	24	4	8	300904	1810.1	26.0	89.9	63.9
Salt Lake City Draper corridor	2013	212.21	287.38	3	3.8	0	3	0	0	0	5	4	8	300904	1810.1	26.0	89.9	63.9
Utah Trax Green	2011	370.00	517.45	3	5.1	0	4	0	0	0	0	4	8	300904	1810.1	26.0	89.9	63.9
San Diego, Calif., East Urban Corridor	1989	108.00	270.30	6	11.3	0	9	0	0	0	15	4	9	4103124	4255.9	50.6	74.8	24.2
Mission Valley East LRT Extension	2005	506.20	811.78	9	6.9	0.7	4	1	0	0	11	4	9	4103124	4255.9	50.6	74.8	24.2
San Diego Oceanside Sprinter	2008	477.63	717.04	8	22	0	15	0	1	1	12	4	9	4103124	4255.9	50.6	74.8	24.2
San Diego Mid Coast Corridor	2021	2171.2	2457.81	7	10.9	0.03	9	0	0	0	45	4	9	4103124	4255.9	50.6	74.8	24.2
San Fran North Concord	1995	86	176.82	4	2.5	0	1	0	0	0	0	3	9	4103124	18629.1	45.1	72.5	27.4

San Francisco, Colma BART	1996	165.70	329.73	9	1.6	0.1	1	1	0	0	0	1	9	4103124	18629.1	45.1	72.5	27.4
San Fran Dublin to Pleas	1997	517	1011.58	6	14.7	0.5	2	1	0	0	0	2	9	4103124	18629.1	45.1	72.5	27.4
BART extension to SFO	2003	1551.60	2657.04	7	8.7	6.1	4	0	1	0	0	1	9	4103124	18629.1	45.1	72.5	27.4
BART to Warm Springs	2018	2330.00	2927.00	7	10.15	0	2	0	0	0	40	4	9	4103124	18629.1	45.1	72.5	27.4
San Franscisco Central Subway	2022	1950.00	2073.57	12	1.7	1.3	4	3	0	0	4	4	9	4103124	18629.1	45.1	72.5	27.4
San Francisco SMART	2019	42.53	52.23	5	2.1	0	1	0	0	0	0	4	9	4103124	18629.1	45.1	72.5	27.4
San Francisco Antioch extension	2018	525.00	659.52	9	10	0	2	0	1	0	8	7	9	4103124	18629.1	45.1	72.5	27.4
San Jose, Calif., Guadalupe Corridor	1991	556.10	1272.72	6	20.3	0	28	0	1	1	50	5	9	4103124	5684.1	42.4	81.0	38.6
San Jose Tasman West	1999	328.00	615.08	8	7.6	0	12.2	0	0	0	0	5	9	4103124	5684.1	42.4	81.0	38.6
San Juan Tren Urbano	2005	2250.00	3608.29	6	10.6	1.1	16	2	1	1	74	4	4	117760	8109.1	72.3	88.9	16.6
St. Louis Metrolink	1993	465.00	1006.56	5	17.5	0	13	0	1	1	31	4	7	451201	4885	26.6	87.9	61.3
St Louis Metrolink extension	2001	339.20	605.85	5	17.4	0	8	0	1	0	20	4	7	451201	4885	26.6	87.9	61.3
St Louis Metrolink Cross County Ext	2006	676.788	1058.46	4	8.2	1.4	10	2	0	0	22	4	7	451201	4885	26.6	87.9	61.3
Santa Clara Tasman East Line	2004	434.90	721.26	8	8.2	0	8	0	0	0	70	5	9	4103124	6984	42.4	81.0	38.6
Seattle Central Link	2009	2558.00	3738.46	9	15.6	2.3	13	5	1	1	35	5	10	854683	8791.8	37.5	75.4	37.9
Seattle University Link	2016	1674.80	2189.33	8	3.15	3.15	2	2	0	0	27	5	10	854683	8791.8	37.5	75.4	37.9
Seattle Northgate Link	2021	1900.00	2150.81	9	4.3	3.5	3	2	0	0	41	4	10	854683	8791.8	37.5	75.4	37.9
Seattle Lynwood Link extension	2024	3260.40	3260.40	7	8.5	3.5	4	0	1	0	34	4	10	854683	8791.8	37.5	75.4	37.9
Seattle Eastside	2024	3677	3677	7	6.5	0.5	8	1	1	0	10	5	10	854683	8791.8	37.5	75.4	37.9
Washington METRO	1986	7968.00	22757.83	15	26.4	13	26	13	1	1	294	1	3	186165	11280.7	27.0	85.7	58.7

Washington Largo Extension	2004	426.40	707.17	8	3.1	2	2	0	1	0	14	1	3	186165	11280.7	27.0	85.7	58.7
Wash. Dulles Corridor	2014	3047.00	4095.40	9	11.7	0	5	0	0	0	64	6	3	186165	11280.7	27.0	85.7	58.7
Wash. Silver Line Ph 2	2024	2778.24	2778.24	8	11.4	0	6	0	1	0	64	6	3	186165	11280.7	27.0	85.7	58.7

Note. Project names in the first column were truncated to fit the column width.