Summer 2021 CS 687 Capstone Project Progress Report
Predicting Loan Default Likelihood Using Machine Learning

Patrick J Beck
Advisor: Dr. Sion Yoon
MS in Computer Science
School of Technology & Computing (STC)
City University of Seattle (CityU)
beckpatrick@cityuniversity.edu, yoonhee@cityu.edu

Abstract

In the financial industry, trying to determine an individual's credit worthiness can be a very difficult task. Traditional models that calculate credit worthiness are heavily dependent upon individual credit score or credit report information. This type of model makes it extremely difficult for individuals with little to no credit to obtain financial services, and if they are able to obtain these services, it is likely with very high interest rates and fees. To expand services to these individuals with more favorable rates while keeping the risk at acceptable levels, financial companies have been trying to figure out ways to implement non-traditional data into the credit risk process. Through the use of machine learning, non-traditional data is used to determine an individual's credit worthiness with an accuracy that is acceptable to financial businesses. This research project performed data processing and feature engineering on data provided by Home Credit Group. This data was then used to train a Light GBM machine learning model and was able to predict the likelihood of default with an AUC of 0.7759.

Keywords: Machine Learning, Credit, Default, Supervised Learning, Unbanked

1. INTRODUCTION

Problem Statement

In the United States, there are approximately 63 million Americans who are considered to be unbanked or underbanked (FDIC, 2017). These individuals have had difficulty opening bank accounts and gaining access to financial services such as credit cards and loans, mostly due to inadequate or a complete lack of credit history.

Through the use of machine learning, datasets will be analyzed to answer the question, can loan repayment be predicted in individuals with little to no credit history?

Motivation

It's quite a conundrum, individuals can't access financial services because of lack of credit history, but they can't build credit without these services. This forces many of these individuals to either remain unbanked or obtain credit with incredibly high interest rates while they try to build up their credit history. Either of these options results in the financial suppression of this subset of the population.

Approach

Prior research has been conducted to try to establish more accurate ways of measuring an individual's credit worthiness. Some of these methods involved creating completely new systems to determine credit worthiness, while other approaches built upon systems that are already in place. For the purpose of this paper, I will be trying to utilize machine learning methods that are already in use in the lending industry, as well as one unproved method, but will be focusing on applying these methods to more unconventional customer data rather than primarily focusing on an individual's credit score and report.

Conclusions

The expected outcome of this research is to find a combination of alternative data that can be used in a machine learning model to effectively predict an individual's likelihood of paying back a loan. To be considered a success, I will be targeting a default prediction area under the curve (AUC) better than 0.81724, which was the winning accuracy in the Kaggle public leaderboard for the dataset that I will be using.

2. BACKGROUND

The number of unbanked and underbanked has decreased 1.1% since 2009, but the number of adult individuals who fall into either of these categories was still at 63 million (FDIC, 2017). If individuals remain unbanked, they will likely never own a home, and they could be restricted from higher paying jobs, as more and more employers are using credit checks to try to gauge how responsible employees are. In order to establish credit or become banked, these individuals accept lines of credit with high interest rates or use alternative financial services. This results in paying a significantly higher amount to the lender compared to an individual with 'adequate' credit.

These financial practices disproportionately affect underserved communities, as 16.9% of African American, 14.0% of Hispanic households and 12.8% of households with other nationalities are unbanked, compared to just 3.0% of white households. To further elaborate on this data, 36.0% of black households and 31.5% of Hispanic households lacked any kind of mainstream credit, compared to 14.4% of white households.

To try to increase the financial opportunities for this subset of the population, Home Credit Group started a Kaggle competition in 2018. The purpose of the competition was to see if any team could figure out how to use alternative customer data to accurately predict an individual's credit worthiness. As part of the competition, Home Credit Group released a dataset to competitors that contained various information about customers other than just credit score or report information. The winning team for the private leaderboard was able to devise a machine learning algorithm that was able to predict a customer's ability to repay with an area under the curve (AUC) of 0.80570.

3. RELATED WORK

A business's determination of an individual's credit worthiness consists of many complex systems, the majority of which rely heavily on the individual's credit score. This heavy reliance on credit scores can limit business growth by only relying on customers with good credit. To attempt to grow the number of customers that a business can lend to, different approaches have been taken to try to determine customer credit worthiness. These alternate approaches have included redesigning credit risk scoring models and creating new risk scoring models that consider big data factors such as mobile phone data, geolocation data, and social media activity.

The determination of if a customer will repay a debt is foundational to a financial business to succeed. It's for this reason that there has already been extensive research into developing the optimal credit risk assessment models that are appropriate for each individual business. While these models are accurate, the emergence of big data is allowing businesses to alter these models, allowing them to lend to more individuals by taking into consideration more factors in their credit risk models rather than just the traditional income and credit score.

Researchers are finding that by using big data and machine learning, an individual's credit worthiness can be determined with accuracy that rivals that of traditional models and sometimes exceeds their prediction accuracy.

Literature Review

In an attempt to create a new credit risk system, researchers at Shanghai Jiao Tong University in China implemented a two-stage dynamic credit risk assessment. This assessment involved using two different layers. The first layer, the aggregation layer, involved a static layer that represented individuals at a specific time frame, while the second layer, the RNN layer, was a dynamic layer that was representative of the dynamic attributes of the individual (Li et al., 2020). This two-stage approach machine learning model resulted in an accuracy improvement of 0.003 over an SVM-RBF machine learning model. The disadvantage of this approach is it is significantly more work for a relatively small gain in accuracy.

Rather than creating a new risk assessment system, a more popular research approach is using emerging technologies on current data to see how accurately credit risk can be assessed.

In one study, the deep learning framework DeepGBM is used to assess credit risk. As compared to LightGBM, PNN, and other machine learning methods, the DeepGBM outperformed

them all with an improvement of at least 0.04 increase in AUC compared to the next nearest model (Chen et al., 2019). Another model involved the use of a support vector machine (SVM) that was optimized using an adaptive mutation partial swarm algorithm (AMPSA). The data used in this research only consisted of 300 samples but resulted in an accuracy prediction of 88.141, roughly 0.8 higher than the next highest model (Fan et al., 2018). A simpler machine learning approach is using a decision tree. A researcher at the University of Kelaniya in Sri Lanka used a decision tree to attempt to predict the credit risk of leasing customers. The data was categorized, and binomial/binary logistic regression was utilized. A decision tree algorithm was then used, which resulted in a 92.34% accuracy on the training dataset (Perera, 2019). The disadvantage of this study is there wasn't a test dataset, so the accuracy outside the training set is unknown.

In another study, researchers at Fuzhou University in China used a support vector machine learning model to try to predict credit card defaults while using differential privacy (Cai et al., 2020). This was a more complex model that also took into account the privacy factors that must be maintained within the financial industry. Through the use of this model, the researchers were able to obtain an AUC similar to other machine learning models while maintaining data privacy. This model, if implemented, could help to make data more secure while also delivering similar results to current prediction models.

Researchers at Kennesaw State University performed a study to evaluate current machine learning methods to see how they could be optimized to improve the accuracy of risk models (Sherry Ni & Wang, 2019). Through this research tried to determine if there were optimal parameter settings that should be used for Logistic Regression (LG), K-Nearest Neighbors (KNN), Decision Tree (DT), and Artificial Neural Network (ANN) models. Through this research, it was determined that there isn't a one-size fits best parament setting for each model, although the use of KNN with bagging resulted in an accuracy, recall, and F1 score, while not having a negative impact on AUC, like boosting did.

The alteration of scoring models using big data is another area of research. Big data allows companies to have a much more thorough look at all aspects of a customer's life. The proper integration of this data within scoring models could significantly increase accuracy and increase the ability to lend to people who wouldn't otherwise qualify.

One such research project involved the use of mobile phone data to determine credit risk. The researcher used a data set from a mobile phone operator in Central Africa, which included data from individual phones, as well as airtime lending data. Using this data to train three random forest models, one for each feature, the model with the recharge and loan features resulted in a 0.80 area under the receiver operating characteristic (AUROC) (Shema, 2019). The limitation with this research is it only focused on airtime lending rather than other financial lending. A different research study used social media data to determine a personal credit score for individuals. The data used for this research was from Douban's social media data. During the data cleaning process, users were classified as either social media stars, "water army", or abnormal activity (Yu et al., 2020). These user's behavior is then used to calculate a personal credit score. The methods of the researchers resulted in abnormal credit score changes comparing before and after data cleaning. The researcher's conclusion was the user's personal activity data, as well as the social network structure, was needed to use this approach for calculating a personal credit score.

Another approach to integrating big data and machine learning with credit-scoring systems was attempted by researchers from the University of Casablanca in Morocco. The researchers developed a multi-agent credit scoring system named CSMAS (Tounsi et al., 2020). This system pulls information from banking systems, payment systems, credit bureaus, and external databases and data sources. After processing, all of this data is then used in a machine learning model that utilizes Gradient Boosting Algorithms (GBA) to make predictions on customers, such as loan default. For this study, the dataset used was the Home Credit Group Kaggle competition dataset. Through the use of CSMAS, the researchers were able to obtain a 92% prediction accuracy using the CatBoost GBA, but the training time took almost an hour to complete. Using LightGBM, the researchers were able to obtain a 91.98% accuracy, with a much more favorable 5 minute 49 second training time.

Another focus of research in terms of machine learning is the bias of machine learning models and the inequalities that may result from these models. The argument is made that historical and societal discrimination can result in this discrimination being unconsciously built into machine learning models (Lee, 2019). Current

machine learning models try to remove bias from data through preprocessing and post processing of the data, in which data is 'corrected' to create a more accurate model. The argument is made that this one-size-fits-all approach currently used to process data could further amplify the bias or not remove enough of the bias.

Review Conclusions

The development of new credit scoring systems using overly complex algorithms has shown increased accuracy as compared to traditional models. The development of one such system, the two-stage dynamic risk assessment, ultimately showed increased accuracy compared to other models but only by 0.003. This system could possibly be improved upon, but with such a small increase in accuracy, it might not be worth the algorithm complexity in actual business use.

The use of new and emerging technologies to increase the accuracy of current models seems to be a more viable option. Through the use of DeepGBM, researchers were able to achieve a 0.04 increase in the AUC over the next best model. While the use of SVM with AMPSA resulted in a 0.8 increase in accuracy when compared to the next best performing model. In the case of SVM with AMPSA, there was a drawback to the study, being that there were only 300 samples upon which the research is based. Even the use of a decision tree had decent results. Researchers using binomial/binary logistic regression with a decision tree model achieved a 92% prediction accuracy in their training model. This research had a flaw as well, in that there wasn't a test model. Without a test model, the actual accuracy of this model is unknown.

The use of big data to determine an individual's credit score is actively being researched to determine the best way to use the data. Researchers using mobile phone data from Central Africa had some success determining credit worthiness for the mobile phone users, returning a 0.80 AUROC for their recharge and loan feature random forest model. The limitations of this research are they were only able to examine airtime lending and not any other financial data. The researchers who attempted to use social media activity to establish a personal credit score found that their data became abnormally altered after cleaning. They concluded that without user personal activity data and social network structure data, they couldn't accurately establish a personal credit score.

The ability to ensure that machine learning models are fair goes to the root of the research problem being addressed in this paper. It's for this reason that the research regarding 'fairness' in machine learning models needs to be given serious consideration. With a lot of machine learning models producing a binary categorical result, it could be extremely easy to unknowingly integrate data that could disproportionately affect one segment of the population. Taking this into consideration, caution needs to be taken about what type of data is used for machine learning models. Caution should also be used when trying to use historical data, patterns, and trends to teach machine learning models, as this could perpetuate the discrimination that occurred during that specific time period.

4. APPROACH

Requirements

To obtain the project objectives, I have several requirements that need to be met. One of the most important requirements for the project is that the data is configured in a way that can be best utilized by the machine learning model. To ensure this requirement is met, a method must be established to handle missing data, as well as determine the appropriate datatype for each feature.

The next requirement that must be met is, the results of the machine learning model must be easily explainable. This model is developed for use in the financial industry, so regulations require an explanation of why decisions are made regarding loans. For purposes of this project, I will be utilizing the DeepGBM framework, which uses information from decision trees and inputs it into neural networks to generate predictions. This approach has been previously researched on this dataset, as discussed in the literature review, but taking a different approach to data preprocessing the goal is to increase accuracy further for this approach while ensuring that results can still be easily explained.

The last requirement, and most important, is to attempt to reduce or, ideally, eliminate bias. As the field of artificial intelligence takes a larger role within businesses and society, it is imperative that models are designed to not perpetuate the discriminatory practices that are so ingrained within our society. While the goal of this model is to find alternative means to determine credit worthiness, the results need to be closely

analyzed to see what biases result from this model.

Design

The design of this machine learning model will be relatively simple. After the dataset is processed, the DeepGBM model will be trained using our training set. The test data will then be run through our model to see our results. The results will then be analyzed, and adjustments will then be made on the training dataset and machine learning model to see if accuracy can be further increased. The most complex part of our design is determining the best approach to take with our dataset. This includes the handling of missing data, changing datatypes, and feature engineering. During the analysis of the data, new features will be created from the existing data to attempt to find hidden correlations within the dataset.

Implementation

After analysis of the project goals, the datasets available, and the research already completed on this project, the decision was made to use the LightGBM and the DeepGBM frameworks.

The Light GBM framework is a gradient boosting decision tree (GBDT) that utilizes gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB) that was developed by Microsoft in 2016 (Ke et al., 2017).

The DeepGBM framework was developed by researchers at the University of Posts and Telecommunications and the Chongqing Housing Provident Fund Management Center, located in China. This framework combines the use of a categorical neural network (CatNN) for sparse categorical data and a gradient boosting decision tree 2 neural network (GBDT2NN) for numerical features.

5. THE DATA

Data Collection

The data utilized in this project was obtained from the Kaggle competition, Home Credit Default Risk. The dataset was preprocessed by the competition sponsor, Home Credit Group. With the data already being collected, the focus was instead given on further processing of the datasets and feature engineering.

Data Cleaning

The data for this machine learning project, as explained previously, was supplied from Home Credit Group. They already performed a lot of cleaning on the data so that it could be easier to utilize for machine learning tasks, but this precleaning is not quite sufficient for our purposes, so further cleaning was performed.

Taking a brief look at the training dataset, there are 307,511 rows of data and 122 columns, resulting in 37,516,342 instances. Taking a closer look at our features, there are several with a significant amount of missing data. The OWN_CAR_AGE feature is missing 202,929 instances, which is 66% of the data for this feature. There are many more features that are like this, especially for features describing the properties of the client's residence. For features that are categorical, a category label of 'MISSING' will be used for missing data. For features that contain numerical data the median value of the feature dataset will be used for missing data.

In looking at data to determine if a loan applicant is likely to default, we want to ensure we look at the complete dataset. For applicants who default on loans, there are likely data at the extremes of the data ranges. To ensure that this data is not excluded from our training data, outliers are not eliminated from the dataset.

Data Analysis

The first feature analyzed was our 'Target' feature, which gives information regarding whether loans issued are repaid or not. The number of loans that are categorized as 'Default' only accounts for 8.07% of our dataset, as depicted in Figure 1. This is a very small proportion of our dataset, and since this is what we are trying to predict, this low number of instances could result in low model accuracy.

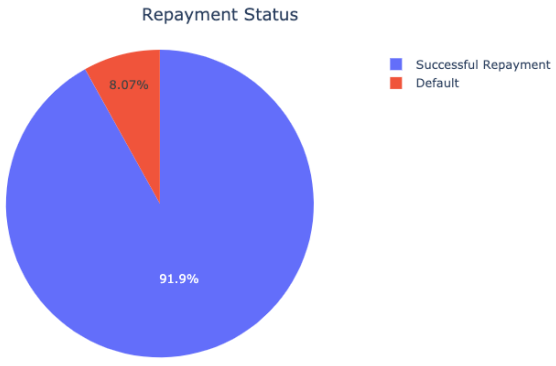

Fig 1. Loan Repayment Status

The next feature analyzed was the types of loans issued. Analysis of this feature's data shows that the majority of loans issued are revolving loans, with this category making up 90.5% of all loans issued. While the cash loans category makes up

only 9.5% of the data, as depicted in Figure 2 below. This distribution of data is very similar to the distribution of our 'Target' feature.

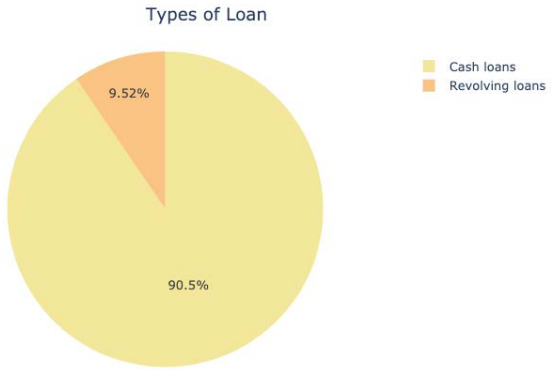

Fig 2. Types of loans issued

The distribution of the 'AMT_INCOME_TOTAL' and 'AMT_CREDIT' features was then analyzed. Both of these features had a similar right skewed distribution, as shown in Figures 3 and 4.

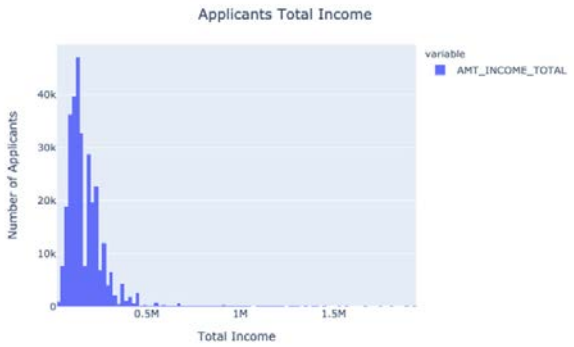

Fig 3. Distribution of Applicants Income

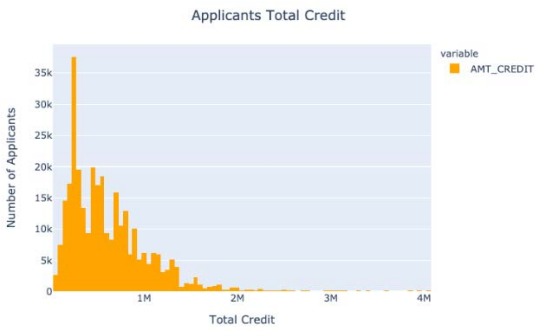

Fig 4. Distribution of Applicants Credit

After seeing the distribution between these two features, an analysis was then conducted to determine if there is a relationship between income level and loan repayment. Applicants with an income level of greater than $100,000 had a repayment rate of 92%, as compared to 91.8% for applicants with an income below $100,000. The small difference between these two groups could be caused by the limitation of our dataset resulting from the small number of total loan defaults.

With the difference between default rates of applicants with incomes above and below $100,000 so small, the next focus of the analysis was a comparison between income sources and repayment status.

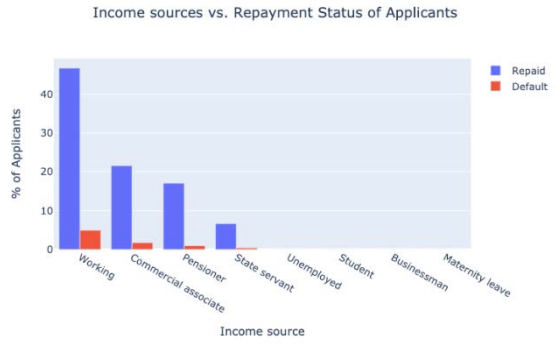

Fig 5. Comparison of Applicant Income Source and Repayment Status

From the graph depicted in Figure 5, it can be seen that a higher percentage of applicants that have an income classification as 'Working' make up a larger percentage of loan defaults than other categories. This category also makes up the largest percentage of all loan applicants, so the higher percentage default rate is not surprising. To gain more insight here, future research could investigate default percentages within each category rather than compared to all applicants, but for this project's purpose this will not be done.

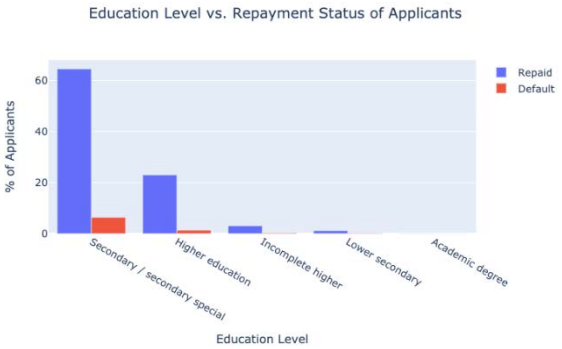

Fig 6. Comparison of Applicant Education Level and Repayment Status

Taking a closer look at loan customer education level and their repayment status shows that customers with a secondary education make up 71% of all loan customers, as shown in Figure 6. Approximately 6% of all loan customers are

those with secondary education and have defaulted.

Feature Engineering

During feature engineering, the results from our data analysis were used to try to create relevant new features from the existing dataset.

After analyzing the application_train dataset, a column was created for the amount of credit issued as compared to income, represented as a percentage, labeled as Cred_Inc_Per., and a flag for if income is higher than debt, labeled as Inc_Cred_Flag. Other features created in the application_train dataset were annuity to income percent labeled as Annu_Inc_Per and number of days employed labeled as Per_Days_Employ.

The original training dataset consisted of 307,511 rows and 122 columns. After cleaning, processing, and feature engineering, our final training dataset contained 307,511 rows and 370 columns.

6. MACHINE LEARNING

The problem this project is trying to address will require the classification of a loan customer as either 'successful repayment' or 'will default'. After analyzing machine learning models, it was decided that the DeepGBM framework would be used and compared against a Light GBM model. Deep GBM is a decision tree model that utilizes a categorical neural network (CatNN) for categorical data and a gradient boosting decision tree 2 neural network (GBDT2NN) for dense numerical data. This model was initially developed in 2019 and was tested on the same Home Credit Kaggle dataset that is being used for this project. The researchers who developed this model were able to obtain a 0.755832 AUC on this dataset. This result was better than all other models tested, with the next closest model, the LightGBM model, finishing with an AUC of 0.73466

Before training the DeepGBM model a LightGBM model was trained, so the two models could be compared. The LightGBM model used had a varying max_depth of 3, 5, 7, and 10 so the optimal depth could be found and used. Other parameters used were an nthread of 5, the num_leaves of 32, a max_bin of 512, and a learning rate of 0.05, along with several other parameters that can be further explored in the Github code repository for this project. I didn't have time to optimize the hyperparameters so many of the parameter settings I used were the settings most often used by the top scoring Kaggle competition teams.

7. FINDING

After splitting out data into the train, validation, and test sets a LightGBM model was trained to determine what the most important features in our dataset were. The most important features were those with a calculated value of greater than or equal to 50, which resulted in the selection of 181 features. These features were then used to train the LightGBM model to run against the test data.

```
The best max_depth is 3, resulting in a Train AUC score of 0.8119709901534501
The best max_depth is 3,resulting in a Cross validated AUC score of 0.7791598502880408
The best max_depth is 3, resulting in a Test AUC score of 0.7759957167640366
The test AUC score is : 0.7759957167640366
```

Fig 7. AUC Scores from Light GBM Model

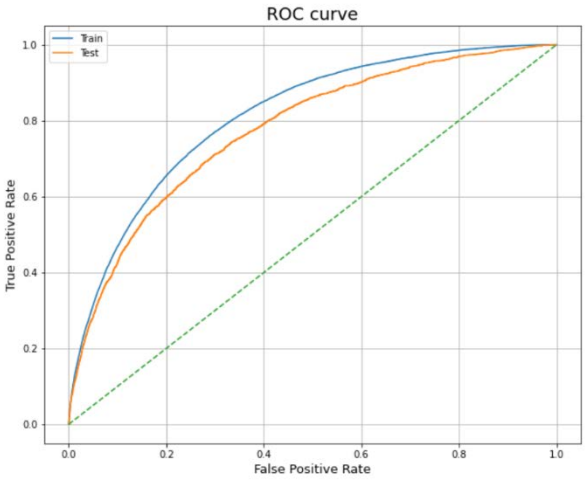

Fig 8. Light GBM AUC_ROC Curve

The initial training of this model resulted in an AUC of 0.81197, as shown in Figure 7. The AUC then increased further when our validation data was run through the model. Lastly, when the fully trained model was used with the test data, it resulted in an AUC of 0.77599, as depicted in the AUC_ROC graph in Figure 8.

Unfortunately, when trying to implement the Deep GBM model I was unable to get it to function properly, so there are currently no results for the Deep GBM model.

8. CONCLUSION

The DeepGBM researchers were able to achieve an AUC of 0.735 with LightGBM and 0.756 with their DeepGBM framework, which was a 2.78% increase. With proper data preparation and feature engineering I was able to obtain an AUC of 0.7759 with LightGBM, which is a 0.0199 AUC improvement as compared to the DeepGBM researchers. As I was unable to get the DeepGBM model to function properly it is unknown if I would have seen a similar increase in AUC between the LightGBM and DeepGBM models.

The stated goal for this project was to obtain an AUC of 0.81724, which I was unable to obtain through the use of a LightGBM model with feature engineering. If I was able to get the DeepGBM model working and I saw the same 2.78% increase between models as the DeepGBM researchers, theoretically I could have been much closer to this goal with an AUC of 0.7974.

While this model isn't accurate enough yet for deployment for commercial use it does show that through proper data preparation and feature engineering, data other than credit scores can successfully be used to predict an individual's credit worthiness in a test setting.

9. FUTURE WORK

There is still a lot of work that could be done with this project. The first area of focus would be to continue work on trying to implement the DeepGBM framework. While research on the DeepGBM framework is limited, it has shown promise in increasing machine learning model AUC, as compared to other models.

Another area of work would be to try and bring in more data to train the model. The entire focus of this research was to try to use data other than the typical credit score data to determine credit worthiness. The dataset supplied by Home Credit Group was still very much focused on more traditional financial numbers. I would like to bring in other data such as mobile phone data and other payment history such as rental and utilities. It would be interesting to see how this would affect the model, as some of these factors are currently being used by some lending institutions already.

In addition to bringing in more data, I would also like to see more feature engineering, especially after bringing in more data. Being able to find hidden relationships in the data can have a significant impact on model performance, as seen in the LightGBM results I was able to obtain compared to the results obtained by other researchers.

Lastly, I didn't have time to optimize the hyperparameter settings of the LightGBM model. Just by tuning the hyperparameters it is likely that this model can increase the AUC without bringing in further data or further feature engineering.

10. REFERENCE

Cai, J., Liu, X., & Wu, Y. (2020, November). SVM Learning for Default Prediction of Credit Card under Differential Privacy. In Proceedings of the 2020 Workshop on Privacy-Preserving Machine Learning in Practice (pp. 51-53).

Chen, X., Liu, Z., Zhong, M., Liu, X., & Song, P. (2019, September). A deep learning approach using DeepGBM for credit assessment. In Proceedings of the 2019 International Conference on Robotics, Intelligent Control and Artificial Intelligence (pp. 774-779).

Fan, Q., Liu, X., Zhang, Y., Bao, F., & Li, S. (2018, October). Adaptive mutation PSO based SVM model for credit scoring. In Proceedings of the 2nd International Conference on Computer Science and Application Engineering (pp. 1-7).

FDIC. (2017). *2017 FDIC National Survey of Unbanked and Underbanked Households*. https://www.fdic.gov/householdsurvey/2017/2017report.pdf

Lee, M. S. A. (2019). Context-conscious fairness in using machine learning to make decisions. AI Matters, 5(2), 23-29.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*, 3146-3154.

Li, R., Deng, S., Zhang, J., He, H., Jin, Y., & Duan, J. (2020, July). A Two-Stage Dynamic Credit Risk Assessment System. In Proceedings of the 2020 4th International Conference on Deep Learning Technologies (ICDLT) (pp. 99-103).

Perera, P. (2019, September). Decision tree approach for predicting the credit risk of leasing customers in Sri Lanka. In Proceedings of the 3rd International Conference on Business and Information Management (pp. 65-68).

Shema, A. (2019, January). Effective credit scoring using limited mobile phone data. In Proceedings of the Tenth International Conference on Information and Communication Technologies and Development (pp. 1-11).

Tounsi, Y., Anoun, H., & Hassouni, L. (2020, March). CSMAS: Improving multi-agent credit scoring system by integrating big data and the new generation of gradient boosting algorithms. In Proceedings of the 3rd international conference on networking, information systems & security (pp. 1-7).

Wang, Y., & Ni, X. S. (2019, April). Developing and Improving Risk Models using Machine-

learning Based Algorithms. In Proceedings of the 2019 ACM Southeast Conference (pp. 281-282).

Yu, X., Yang, Q., Wang, R., Fang, R., & Deng, M. (2020). Data cleaning for personal credit scoring by utilizing social media data: An empirical study. IEEE Intelligent Systems, 35(2), 7-15

11. DEMO

https://youtu.be/CW1ZBKEOZ2Q

12. GITHUB

https://github.com/PJBeck84/Capstone.git