Credit Risk Prediction using Classification Algorithms:

An Analysis of Lending Club Loan Data

May 14th, 2023

Team Members:

Yihe Chen

Haoxuan Huang

Ruitao Jiang

Sicheng Wang

Jingxi Yang

Table of Contents

Executive Summary

This project aims to build an effective machine learning model to predict loan customers that may potentially default. A dataset with 20K customer records and 13 features was cleaned and analyzed to understand how different features affect customers' potential defaults probability. Several machine learning models were fitted as part of the analysis. KNN, logistic regression, random forest, XgBoost, and Support Vector Machine, were trained to predict potential loan defaults. The parameters that give the best performance in each model were identified by performing grid searches. Based on the AUC as well as the recall metrics, the team identified logistic regression as the best-performing model that is able to spot two-thirds of defaulting customers.

Chapter 1: Introduction

In the wake of recent bank collapses, people have been paying more attention to establishing effective approaches to monitor and manage the credit risk of the banking system. The worldwide financial system has continuously adapted to incorporate cutting-edge technologies and methodologies to improve its effectiveness and security. A crucial challenge for financial institutions, especially banks, is determining credit risk and accurately forecasting loan defaults. It is essential for banks to have dependable and precise methods for evaluating borrowers' creditworthiness in order to reduce loan default risks, which contributes to the financial system's stability.

This project aims to tackle the issue of predicting loan defaults through the application of machine learning techniques and by examining the Lending Club Loan dataset. This dataset offers extensive information about credit applicants, such as loan amounts, interest rates, loan grades, credit scores, employment history, annual incomes, and other pertinent characteristics. Utilizing this data, we plan to apply machine learning techniques to enhance financial institutions' decision-making processes, decrease the risk of future loan defaults, enable fairer lending practices, and ultimately promote financial stability.

The significance of this project is its ability to enhance financial institutions' decision-making processes, empowering them to make more informed lending decisions. By reducing the risk of loan defaults, banks can increase their profitability and promote a more stable financial environment. Moreover, precise prediction models can result in more fair lending practices, as they can assist in eliminating biases and ensuring credit is granted to deserving applicants based on objective factors.

Our project seeks to investigate the following hypotheses:

1. Machine learning models can effectively predict loan defaults using the features present in the Lending Club Loan dataset.
2. Specific features, such as credit scores, loan grades, and employment histories, have a more significant impact on loan default predictions than others.
3. A refined machine learning model can outperform traditional credit evaluation methods in predicting loan defaults and reducing credit risk.

To accomplish our goals, the project will be organized into three primary phases:

1. Exploratory Data Analysis (EDA): In this phase, we conducted a thorough examination of the dataset to identify trends, patterns, and irregularities in the data. We performed numerous operations to prepare the data for analysis, including converting categorical variables to numerical forms using one-hot and binary encoding, handling missing values, removing outliers, splitting the dataset into training and testing sets, and so on.
2. Machine Learning: In this phase, we created predictive models using supervised binary classification methods to forecast loan defaults. The models we trained included Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Random Forest, and Extreme Gradient Boosting. A grid search was conducted to find the optimal parameters for each model. By optimizing our model, we aim to maximize its predictive potential and contribute to a more effective credit evaluation process. We then validated and tested the machine learning algorithms to determine the best-performing model according to the AUC metric, which is appropriate for our highly imbalanced dataset. We also evaluated

the recall rate of each model to compare their accuracy and effectiveness in predicting loan defaults.

3. Model Comparisons: Our analysis revealed that Logistic Regression outperformed other models with a comparable AUC, the highest recall rate, and a good cross-validation score. The findings of our study underscore the importance of recall rate in selecting efficient models. For effective loan default predictions, a high recall rate indicates a lower rate of false negatives, which is crucial for identifying loan default risks and improving profitability in a financial context.

In summary, this project's objective is to develop a machine learning model capable of accurately predicting loan defaults and assessing credit risk using the Lending Club Loan dataset. By uncovering patterns and trends in the data, we aim to improve financial institutions' decision-making processes and contribute to a more stable and equitable financial landscape. By using machine learning algorithms, we will strive to create a model that outperforms traditional credit assessment techniques and ultimately reduces the risk of future loan defaults.

Chapter 2: Background

For this project, we attempted to predict the risk of loan default using Lending Club's past loan data from 2007 to 2018. The data is publicly available on Kaggle. The dataset covers 20,000 records, where each record contains information regarding a borrower. The Lending Club's loan dataset can be downloaded from https://drive.google.com/file/d/1WFvu8dnVwZV5WuluHFS_eCMJv3qOaXr1/view [3].

Data Label:

The label of the dataset is whether or not a borrower ultimately defaulted or not. The original dataset has 7 loan statuses: charged off, current, defaulted, fully paid, in grace period, late (16-30 days), and late (31-120 days). To simplify the prediction into a binary classification problem, we labeled a record as 1 if the loan was "charged off" or "defaulted", and as 0 if the loan is "fully paid", ignoring all other categories such as "current", "late", or "in grace period".

For our labeled dataset, 80% of the records have a label of 0 (fully paid) and 20% of the records have a label of 1 (defaulted), which makes the dataset imbalanced. To address this problem, in the later section of this report, recall rate, also known as true positive rate, is adopted as one of the key performance metrics.

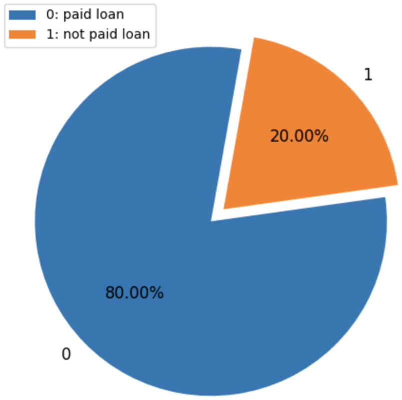

Figure 1. Ratio of two output classes

Features:

There are a total of 13 features in the dataset, 7 of them are categorical and 6 are numerical.

- *Grade:* This is an assigned credibility level for each loan applicant. The majority of the records have a grade of B, C, or D, with a smaller proportion having a grade of A or E and a tiny proportion having a grade of F or G. As seen below, better loan grades are associated with a lower likelihood of having unpaid loans.

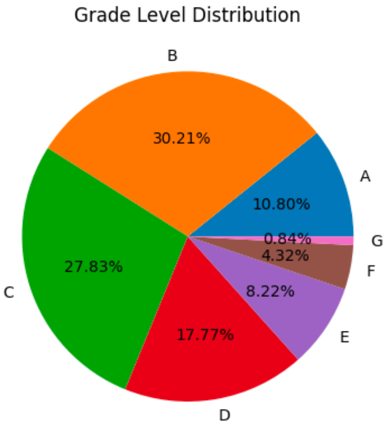

Figure 2. Distribution of grade levels

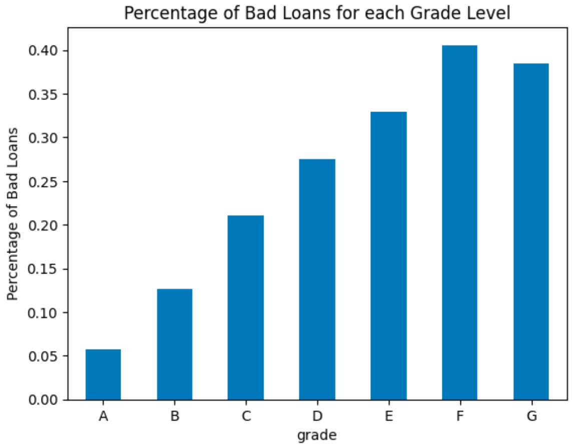

Figure 3. Percent of bad loan for each grade level

- *Annual_Inc:* This is the loan applicant's self-reported annual income level during the loan application process. The mean annual income for loan applicants who paid the loan is $10K higher than that for those who did not pay the loan.

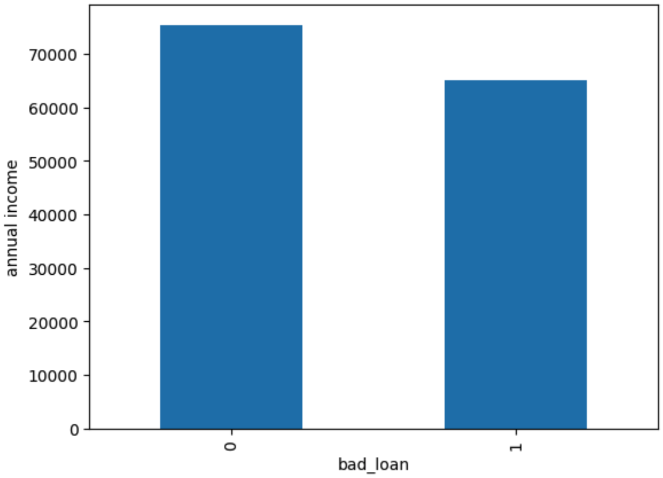

Figure 4. Mean annual income

- *Short_emp:* This feature has a value of 1 if the loan applicant is employed for 1 year or less (which is 88.75% of the records), and 0 otherwise (11.25% of the records). 24.18% of applicants who have been employed for 1 year or less ended up not paying the loan, and 19.47% of applicants who have been employed for more than 1 year ended up not paying the loan.

- *Emp_length_num:* This feature represents the number of full years the applicant has been employed up until the time of loan application. The maximum value for this feature is 11, so if an applicant has been employed for more than 11 years, the value for this feature is still 11. It can be seen below that those who are employed for less than 3 years are more likely to have an unpaid loan.

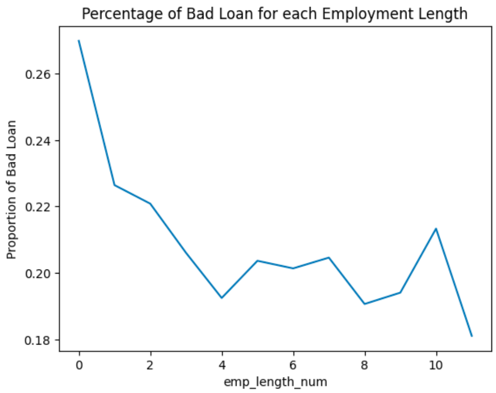

Figure 5. Percentage of bad loan for different employment lengths

- *Home_ownership:* Indicates the types of ownership that the applicant has, ex. rent, own, mortgage. Looking at the bar graph below, we can see that there is no significant proportional difference between loan defaults across different types of home ownership.

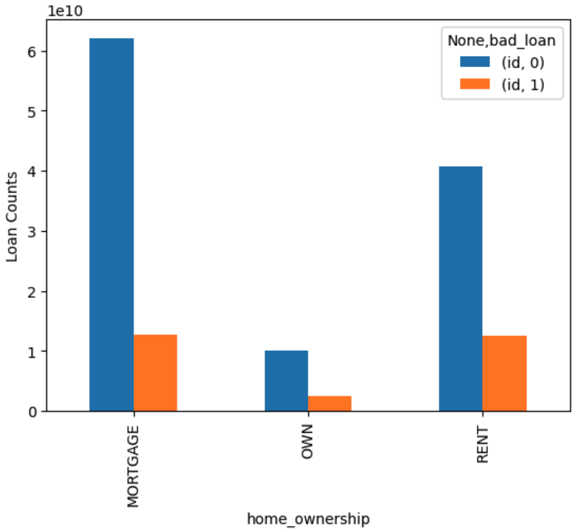

Figure 6. Frequency distribution for 3 types of home ownership

- *Dti:* Debt-to-Income Ratio, calculated using the borrower's total monthly debt payments on total debt obligations, excluding mortgage and the requested Lending Club loan, divided by the borrower's self-reported income [2]. If we compare the distribution of debt-to-income ratio of Defaulted loan holders to that of non-defaulted loan holders, there is a clear trend that, on average, Defaulted loan holders tend to have a higher debt-to-income ratio than non-defaulted loan holders. Thus, we can say that there is a positive association between Debt to income ratio and the probability of loan default.

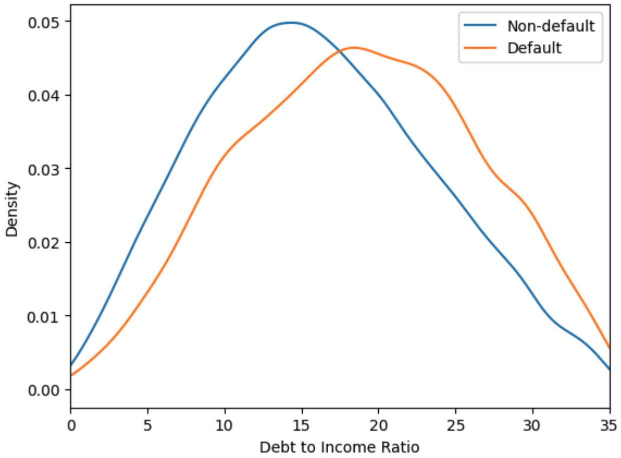

Figure 7. Distribution of debt to income ratio

- *Purpose:* Indicates the purpose of the loan, which is provided by the borrowers upon submitting the loan application. From the bar graph below, it is unequivocally clear that "debt consolidation" and "credit cards" are the main purposes for the loan that is defaulted on. This somehow indicates a positive relationship between the existing level of debt/credit card bill and the probability of future loan default. The graph can also further suggest that individuals who have already held significant debt and credit card bills tend to fall into the vicious cycle of accumulating more debt.

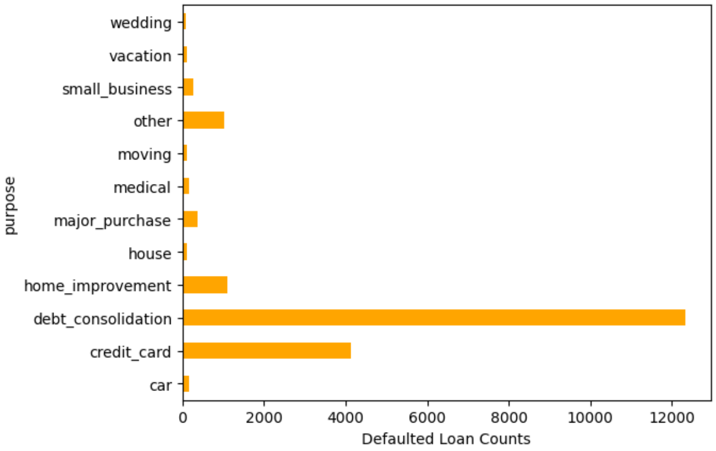

Figure 8. Distribution of loan purpose

- *Term:* Indicates the number of monthly payments on the loan. Values can be either 36 months or 60 months. As indicated in the pie chart below, 75% of the total loans are for 36 months, and 25% are for 60 months. Also as we can see in the bar chart below, 60 months loans, in proportion, default much more frequently than 36 months loans. The bar chart results suggest that there is a positive association between the number of monthly payments on a loan and its default probability.

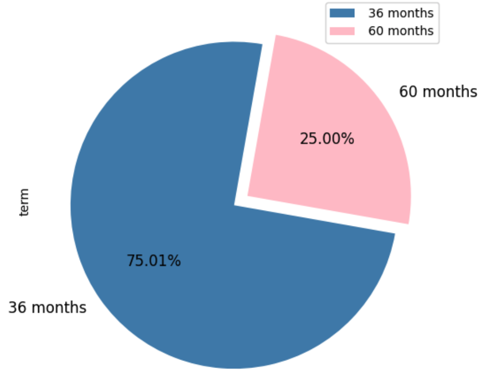

Figure 9. Ratio of two types of loan term

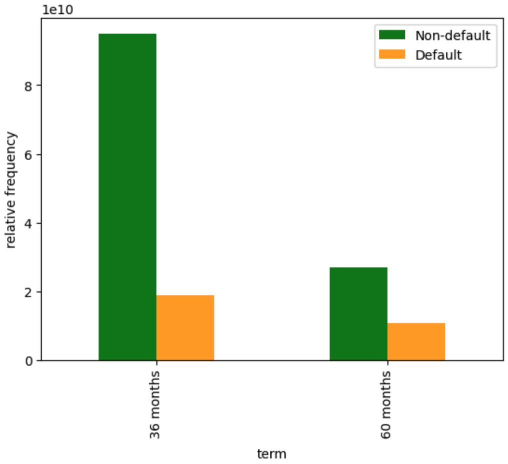

Figure 10. Relative frequency of two classes for 2 loan terms

- *Last_delinq_none:* A binary variable where 0 indicates the borrower has never had a delinquent payment on a loan and 1 indicates the borrower has had at least one incident of delinquency on loan payment [2]. Roughly 55% of the total loan holders had no delinquency before, and 45% had at least once. From the bar graph below, we can also see that a loan holder with a delinquency history is, on average, more likely to default on both a 36 months and 60 months loans than a loan holder without a delinquency history.

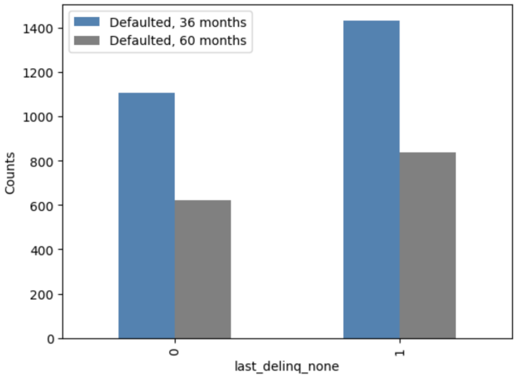

Figure 11. Relative frequency for defaulted/non-defaulted in terms of last delinquent payment and length of loan

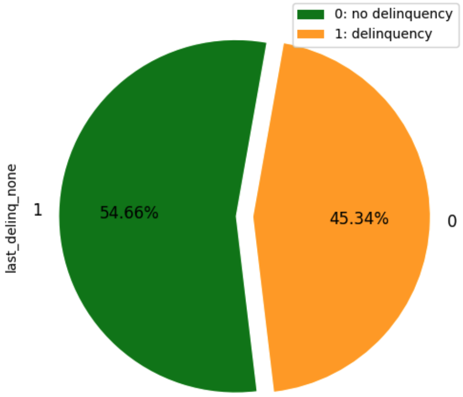

Figure 12. Proportion of last delinquent accounts in the total

- *Last_major_derog_none:* This binary feature has a value of 1 if the borrower has at least 90 days of bad rating in credibility and 0 otherwise [2]. 436 out of 20000 samples have a value of 1 and 138 samples have 0 for this feature, and the rest of the records are empty (nan).

- *Revol_util:* Revolving line utilization rate [2]*,* this feature represents the amount of loan the applicant is borrowing relative to the applicant's total credit limit. The distribution of the majority revolving line utilization rate ranges from 0-128, and there is an outlier of 5010. The graph below represents the distribution excluding the outlier.

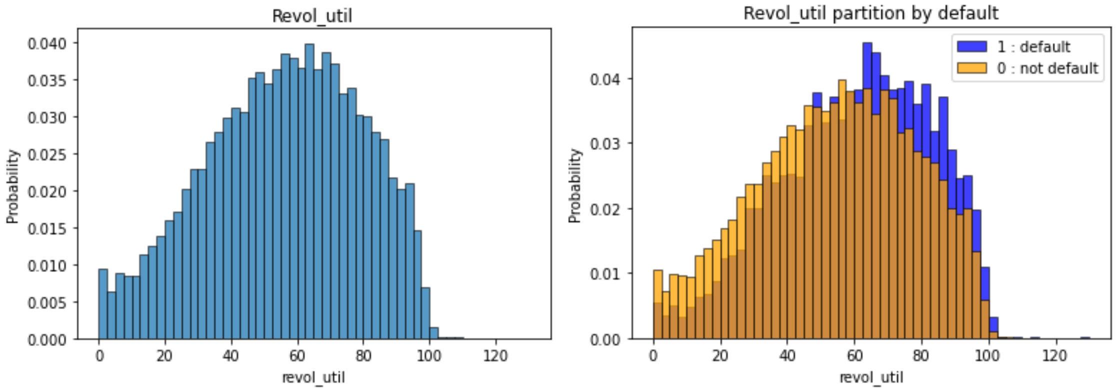

Figure 13. Histograms for Revolving Line Utilization Rate

- *Total_rec_late_fee:* This feature represents the total amount of late fees incurred by the applicant up to the application date [2]. The majority (19769 out of 20000) of the applicants have 0 late fees. Applicants with non-zero late fees are more likely to have bad loans (72.2% of the applicants with late fees turn out to be bad loan holders, compared to the overall rate of 20%). The histogram of the feature with non-zero values is shown below.

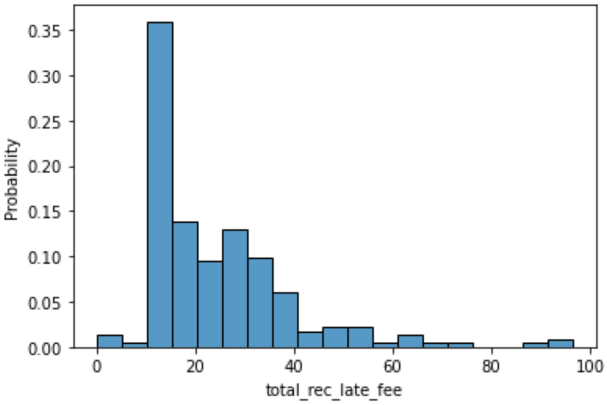

Figure 14. Histogram of Late Fees incurred by customers

- *Od_ratio:* This feature represents the overdraft ratio (the amount of money the applicant is borrowing relative to the amount of wealth they have) [2], which looks uniformly distributed. No significant trend can be identified from the scatter plot below (in the scatter plot, the y-axis is bad_loan, and the x-axis is od_ratio).

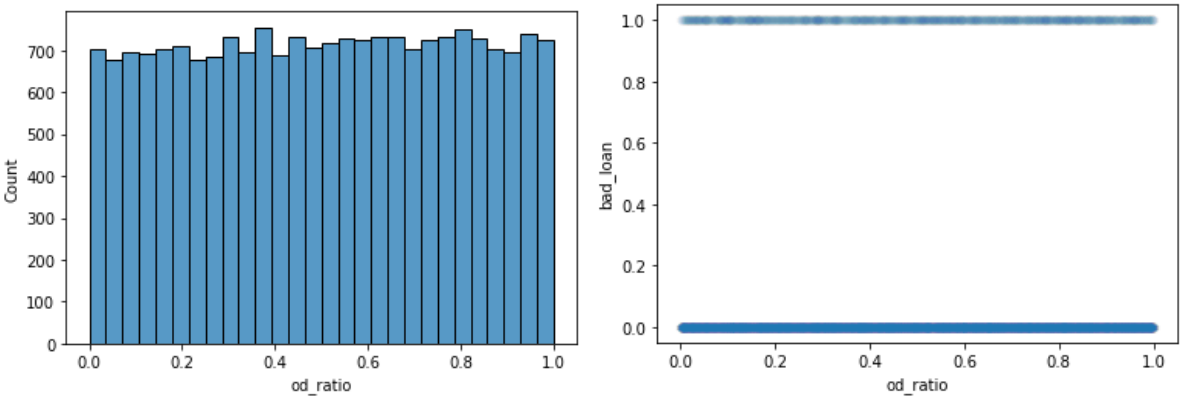

Figure 15. Histogram of Overdraft Ratio & Scatter Plot of Overdraft Ratio and Bad Loan

Chapter 3: Methodology

Data Cleaning

First of all, data-cleaning has been performed on several features of the dataset to make the dataset usable for building the model. Features that are cleaned include:

Grade*:* letter grades are converted to numerical values, where higher values denote a better grade as well as a better credibility.

Home_ownership: as 7.5% of values for this feature are NULL, we made NULL a separate category, so that there are four categories for this feature (mortgage, rent, own, NULL). Then, one-hot encoding was performed on this feature to convert the categorical features into numerical values, where the binary values of the resulting columns denote the category that each sample belongs to (1 if the sample belongs to the column's corresponding feature, and 0 otherwise).

Term: one-hot encoding was performed on this feature (same logic as above)

Purpose: There are 12 categories for this feature, so performing one-hot encoding will lead to 11 resulting columns, which will lead to a serious multicollinearity problem, making it hard to identify the effect of each category on the default probability (our label). To avoid multicollinearity while still capturing the different features, binary encoding was performed on this feature. In binary encoding, each category gets a binary code (each digit in the code must be 0 or 1), and one resulting feature will represent one individual digit in the binary code. In this way, 4 digits can capture up to 2^4=16 categories, so the 12 categories for this Purpose feature can be captured by 4 features, which will greatly reduce the multicollinearity problem.

Dti: 150 values out of the 20,000 records are NULL, so those 150 records were removed.

Revol_util: this feature denotes the utilization rate, which should be between 0 and 100, so a record with a value of 5010 was removed.

Last_major_derog_none: more than 95% of the records have NULL for this feature, meaning that this feature does not contain much useful information. Thus, this feature is dropped.

Model Preparation

Before starting to build the models, feature selection was performed via correlation analysis so that if certain features are strongly correlated with each other, then only one out of those features will be included in the model. The correlation analysis was performed by building a heatmap in the Seaborn package, and the result is below:

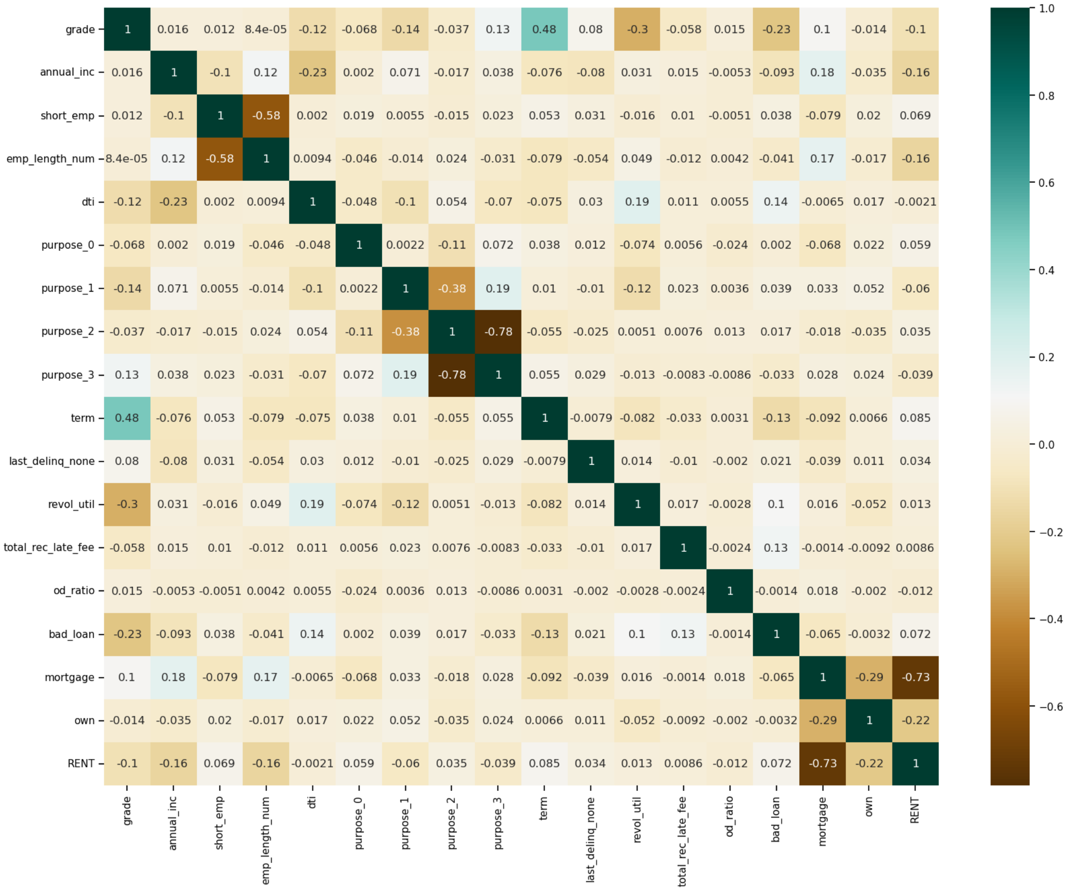

Figure 16. Correlation coefficients between different features

It can be seen that the only features that are strongly correlated with each other are those associated with one-hot encoding or binary encoding (features are strongly correlated with each other if the absolute value of the correlation coefficient is greater than 0.7), but strong correlation is the nature of one-hot encoding, which is important for capturing the categorical features. Thus, all features will be kept for the purpose of model building.

Then, the dataset was randomly split so that 70% of the data became training data and 30% became testing data. Then, normalization was performed by applying the standard scaler in the sk-learn package separately, to the training data and the testing data so that every feature had a mean of 0 and a standard deviation of 1.

Model Building

The learning problem we consider is supervised binary classification. The goal is to find the best classification model to predict whether a set of loan data is a bad loan. We plan to compare the following models: Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Random Forest, and Extreme Gradient Boosting.

Logistic Regression: It is a linear model used for binary classification tasks. It models the probability of an instance belonging to a particular class by fitting a logistic function to the input features. The coefficients of the logistic function are learned by minimizing the negative log-likelihood of the training data, often using techniques like gradient descent.

K-Nearest Neighbors (KNN): KNN is a non-parametric, instance-based learning algorithm. It classifies instances based on their similarity to training instances. Given a new instance, KNN finds the k training instances closest to it using a distance metric and assigns the majority class label among those neighbors.

Support Vector Machine (SVM): SVM is a maximum-margin classifier that seeks to find the best hyperplane that separates instances of different classes. It uses the concept of support vectors, which are instances closest to the decision boundary. The algorithm aims to maximize the margin between support vectors, which improves generalization performance.

Random Forest: This is an ensemble learning method that combines multiple decision trees. Each tree is built using a random subset of features and instances from the training set with replacement. The final prediction is obtained by majority voting from the predictions of the individual trees.

Extreme Gradient Boosting (XGBoost): XGBoost is a boosted tree algorithm that builds a sequence of decision trees, where each tree is trained to correct the errors made by the previous tree. The trees are combined in a weighted manner to form the final model. XGBoost uses gradient boosting to optimize the loss function and employs regularization techniques to prevent overfitting.

The grid search algorithm will be run on each model, out of many different values on several parameters, in order to identify the optimal parameters for each model. The optimal parameters for each model are listed below:

Model	Optimal Parameters
Logistic Regression	C=1, class_weight='balanced', solver='newton-cg'
K-Nearest Neighbors	n_neighbors=29
Support Vector Machine	C=100, class_weight='balanced', degree=2, probability=True, verbose=True
Random Forest	max_depth=9, min_samples_leaf=2
Extreme Gradient Boosting	Default

Table 1. Optimal Parameters for models determined by Grid Search

Chapter 4: Results & Discussion

Given that we have imbalanced data, we used ROC/AUC as the major metric to evaluate the performance of the models. The higher the value of the ROC/AUC metric, the better the model. While a random guess model would have an AUC value of 0.5 and a perfect model would have 1.0, we expect that our models' scores would fall somewhere between 0.5 and 1.

Additionally, we used recall rate (also known as true positive rate) as our secondary performance metric:

Recall = True Positive / (True Positive + False Negative)

It measures the proportion of positive samples that are actually being classified as positive by the model. In the context of this project, the recall rate is measuring the proportion of defaulting customers that are actually classified as default. We believe this is the most important metric of a credit risk control model as the primiary goal of the financial institutions is to spot the high risk borrowers that can potentially default. While low accuracy and low precision may be acceptable, a low recall rate means failure for a credit risk model.

We trained every model using the optimal parameters identified above, and the models' performances are presented in the figure below:

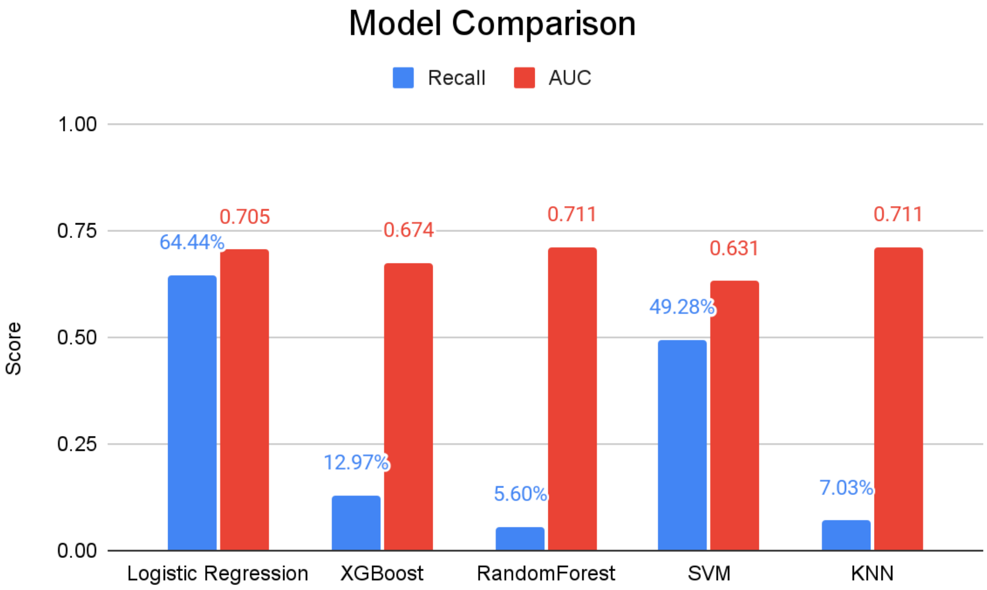

Figure 17. AUC and Recall of every model

From Figure 17 we can see that Random Forest and KNN have the best performance under the AUC metric. However, the problem with Random Forest and KNN is that their recall rates are too low, which is unacceptable for the purpose of this project. Evaluating both AUC and Recall rate, Logistic Regression, which performs reasonably well in both metrics, was choosen as our final model option. Though the AUC of Logistic Regression is slightly lower than that of Random Forest and KNN, but it is the only model that delivers a recall rate above 50% (indicating it is the only model with more prediction power than a random guess model) .

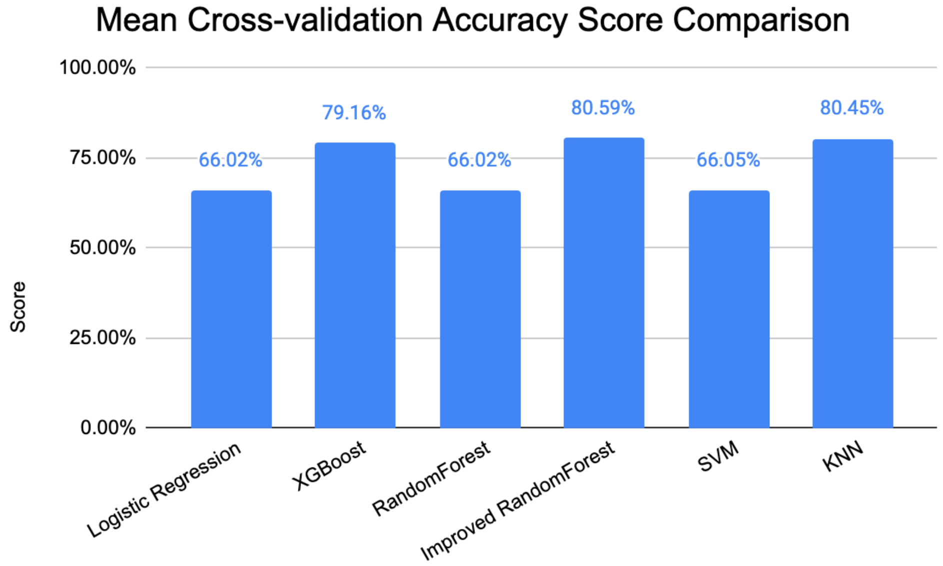

Figure 18. Accuracy of every model

Mean cross validation was used to get the accuracy score of the models. Even though XGBoost, Improved Random Forest (Random Forest trained by only the impotent features), and KNN have the highest accuracy, they suffer from the lowest recall rate. The 80% accuracy for the three models seem rather very suspicious. Given that 80% of the sample data are actually negative (non-defaulting), this means these three models basically predict almost all samples as negative, leading to a 80% accuracy score but terrible recall score. This finding is further illustrated in the confusion matrix below:

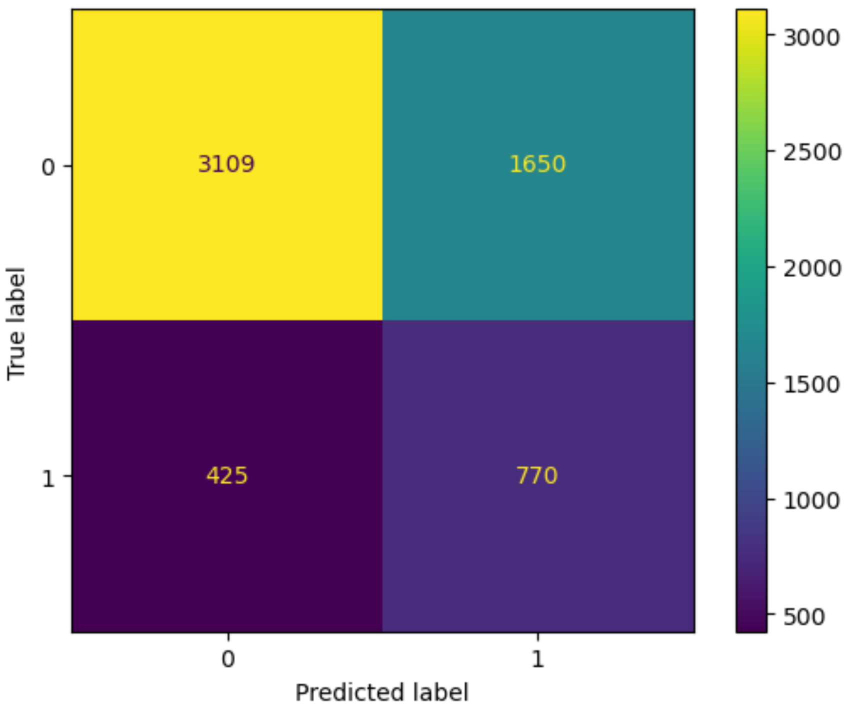

Figure 19. Confusion Matrix for Logistic Regression model

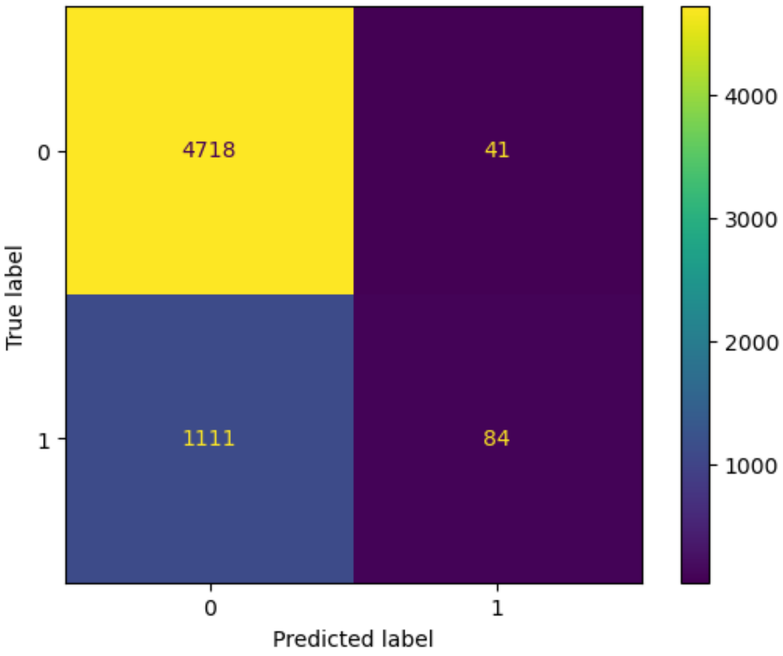

Figure 20. Confusion Matrix for the KNN model

Chapter 5: Conclusions & Recommendations

In summary, this project investigated a binary classification problem in credit risk prediction utilizing machine learning techniques. We first preprocessed and explored a total of 13 features in the dataset. After we found no significant correlation between features, we proceeded to train and test 5 different models with a 70-30 split. We found that, while all models have an AUC around 70%, logistic regression has the highest recall rate of 64% (correctly predicting 64% of the actual loan defaults). Further, with cross validation, we showed how accuracy is not a good metric with such imbalanced data. In the end, we came to the conclusion that, with the an objective to maximize recall rate, logistic regression is the best choice out of the five.

As a conclusion for this project, we showed that machine learning models can be effective at predicting loan defaults. More specifically, certain features do have a more significant impact on loan default predictions than the others. For instance, credit scores, loan grades, and employment histories tend to have more predictive power than the other information of a borrower. What's more, the findings of our study underscore the importance of recall rate in selecting an efficient credit risk model, since it is crucial to identifying loan default risks and improving decision-making in a financial context. Although the accuracy and recall rate of the models are not extremely high, the final result is robust enough for a highly complex problem of credit risk prediction, where the outcome would depend on countless factors, many of which are not captured by this dataset with only 13 features.

For future investigations in credit risk predictions, we have three recommendations. First, collecting more data can allow us to gain better insight into the problem as a whole, whether it is by increasing the number of records or the number of features in the dataset. Alternatively, we can choose to apply more advanced models, such as neural networks with multiple hidden layers, to see if that can strengthen the predictive capability. Finally, we could try to use more advanced techniques to deal with the problem of class imbalance. SMOTE is one of the options, which generates synthetic samples for the minority default class. After such manipulation, we would be able to train the model with a more balanced data set, allowing the model to better learn the differences between the two classes and make more accurate classifications.

References

1. George, N. (2019, April 10). *All lending club loan data*. Kaggle. Retrieved April 11, 2023, from https://www.kaggle.com/datasets/wordsforthewise/lending-club

2. Gomes, G. G. (2022, January 16). *Machine learning: Predicting bank loan defaults*. Medium. Retrieved April 11, 2023, from https://towardsdatascience.com/machine-learning-predicting-bank-loan-defaults-d48bffb9aee2

3. Gomes, G. G. (n.d.). *Lending_club_loan_dataset*. Google Drive. Retrieved April 11, 2023, from https://drive.google.com/file/d/1WFvu8dnVwZV5WuluHFS_eCMJv3qOaXr1/view?pli=1

4. *Predict LendingClub's loan data*. (n.d.). Retrieved April 11, 2023, from https://rstudio-pubs-static.s3.amazonaws.com/203258_d20c1a34bc094151a0a1e4f4180c5f6f.html

5. Mahoney, A. J. (2020, September 30). *Credit risk modeling with machine learning*. Medium. Retrieved April 11, 2023, from https://towardsdatascience.com/credit-risk-modeling-with-machine-learning-8c8a2657b4c4