George Washington University

School of Engineering and Applied Science

SEAS6401

Modeling of Predicting Loan Default with Machine Learning

Rui Zhang- Data Analytics

Ran Wei- Data Analytics

Dr. Benjamin Harvey

Submission Date: 12/07/2021

Abstract

In the lending industry, debtors can apply for loans from lenders, and investors can buy the loans from lenders in trade for the promise of compensation with interest. The investors will make income from the interest if the borrower repays the loan. However, failure to repay the loan will cause losses to lenders and investors. Therefore, lenders face the hassle of predicting the threat of a borrower being unable to repay a loan. In this project, the dataset from Lending Club is used to train Machine Learning models to understand the applicant's profile, to decide if the borrower has the ability to repay the loan, and to minimize the risk of future loan defaults.

Introduction

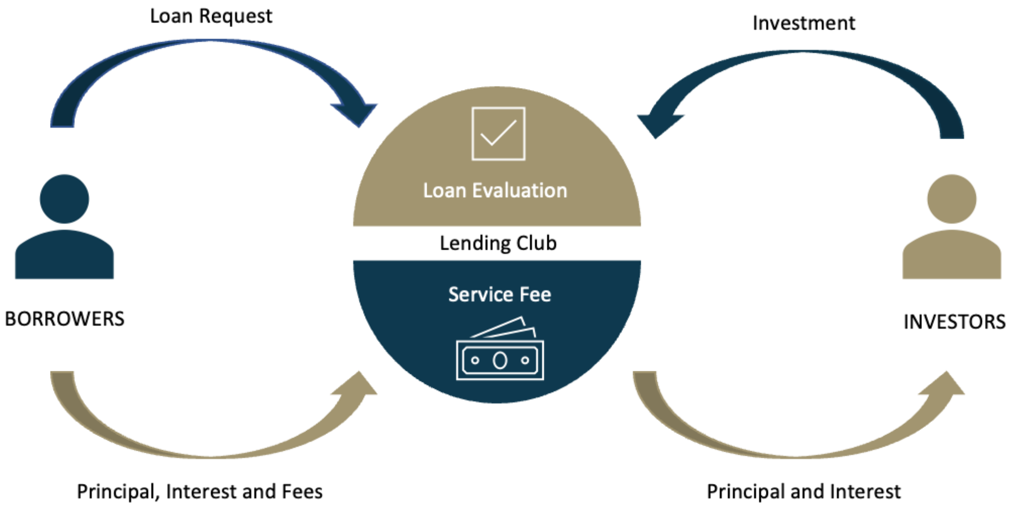
Nowadays, customers can make investments in client mortgage loans through Peer-to-Peer lending platforms. This kind of platform can offer loan opportunities outside of traditional lending institutions like banks and generally provides higher returns to the investors.

Lending Club, founded in 2007, is a pioneer P2P lending platform. Which directly connects borrowers and lenders without an intermediary. Potential investors would be able to view several borrowers' credit history, loan purpose, and other information provided by the borrowers when they applied for loans in order to make determinations as to which loans to fund. Once the transaction is completed, investors can get the principal and interest, and Lending Club would take a flat service fee. The schema of how it works is shown in Figure 1.

However, most loans on Lending Club are unsecured, which means investors will lose all their money when borrowers default. Therefore, the research goal of our project is to build a machine learning model that could predict the probability of loan default and could help investors to minimize the risk and maximize their investment returns.

Figure 1

Schema of How Lending Club Works



Methodology

Data Identification

The Lending Club dataset is no longer available to the public on the official website. We got the dataset from Kaggle, it covers all approved loan information and loan status (whether default or not) from 2007 to the third quarter of 2020. The description of each feature of this data set can be found in Appendix A.

Dependent Variable Acquisition

We use pandas in python to read and manage the dataset. The full dataset has 2.9 million loan records and 151 features for each loan. It is inefficient to read all the data, so we read the first 10000 records to make some feature selection. In order to make a prediction for a single column, we set 'loan_status' as the target value and the dependent variable. There are 10 unique values in this column, but we only need the 'Fully Paid' and 'Charged Off' data to demonstrate if the loan should have defaulted.

Independent Variable Validation & Filtering

To select appropriate predictors, firstly we chose features that would be available to an investor before deciding to fund the loan. We checked the correlation between other features and the target value 'loan_status'. After that, we got 21 selected features and read the whole data with the following features:

1) 'loan_amnt'
2) 'term'
3) 'int_rate'
4) 'sub_grade'
5) 'home_ownership'
6) 'annual_inc'
7) 'verification_status'
8) 'purpose'
9) 'addr_state'
10) 'dti'
11) 'open_acc'
12) 'pub_rec'
13) 'revol_bal'
14) 'revol_util'
15) 'initial_list_status'
16) 'application_type'
17) 'mort_acc'
18) 'pub_rec_bankruptcies'
19) 'loan_status_flag'
20) 'fico'
21) 'earliest_cr_line_y'

The dataset is filtered without duplicates.

Data Aggregation & Visualization

We apply seaborn and matplotlib to display our analysis results. To dig into each feature, we conduct analysis for each selected variable, including target value and other values. For some continuous variables, we need to group these variables so that we could figure out the charged-off rate for each group. Some notable conclusions, which might be useful for business evaluation, will be shown in the next section of this article. All of the results can be found in Appendix C.

Data Representation

As for modeling, we need to digitize each categorical feature and process missing data at first:

1) Created a new column named 'loan_status_flag' based on the dependent variable 'loan_state', and label 'Charged Off' as 1 and 'Fully Paid' as 0.
2) Grabbed the year in which the loan was funded from 'issue_d' in order to analyze the data by time.
3) There are four features that have missing data: 'mort_acc', 'revol_util' ,'dti' and ‘pub_rec_bankruptcies’, the missing percentages are 2.54%, 0.08% , 0.06% and 0.004%. For the features with low missing percentage like 'revol_util' and 'dti', we just simply drop the missing values. 'mort_acc' has the highest correlation coefficient with 'earliest_cr_line_y', as the years of earliest credit line opening increase, the number of mortgage accounts increase. So we grouped the data by 'earliest_cr_line_y' and calculated the average number of mortgage accounts for each group, then filled in the 'mort_acc' variable. After dealing with the missing value, we got 1,857,132 records.
4) Created dummy variables by ‘get_dummies’ function.

Predictive Model

Besides analyzing the used car market, another goal of this study is to build a predictive model to predict the probability of default given the loan origination information. To achieve this goal, we built three different types of models, the logistic regression model, the random forest model, and the XGBoost model. This section will present the modeling strategy and report the performance.

Before making the decision of funding a loan, the top question that an investor is interested in is if the loan will be repaid. To address this question, we restrict the prediction target to the loan status when it is closed--fully paid or charged-off. The dichotomous nature of the response variable directed us to binary classification models. We considered three popular classification models for the tabular data: logistic regression, random forest, and XGBoost.

The logistic regression model is the most classic method having an intrinsic linear structure of the features. The logistic regression model assumes the response variable, the loan status at the closing time follows the Bernoulli distribution with a probability of charged-off being *p*. Then it models the log odds ratio log(*p*/(1-*p*)) with a linear combination of features. This linear structure is easy to understand and the interpretation of the model is relatively intuitive.

The random forest is an ensemble method that gathers many weak learners' results to achieve collective intelligence. Each learner is a decision tree fitted by a random sample of data and a random subset of features. This ensemble fashion can help reduce the variance of the tree collection while maintaining the low bias of each tree.

The XGBoost is a computationally optimized implementation of gradient boosting trees. Although XGBoost is also an ensemble method, unlike random forest, each tree in the XGBoost fits the error and boosts the performance from the last tree in a sequential manner. The XGBoost soon became the industry standard after its initial launch in 2014 due to its strong prediction power and computation efficiency.

Feature Selection

The purpose of this model is to predict the loan status at the end of its life cycle given the information at the start of the cycle. The first concern is the data availability at the origination time. For example, no payment history of the loan is available before the loan is funded, so this type of information has to be excluded from the feature set from a practical point of view. Some features are excluded because they have a large portion of missing values.

Data Split

While ‘loan_status_flag’ was set as the dependent variable, the other 21 selected features were set as the independent variable. The positive label is “charge off”, and the negative label is “fully paid”. The ratio between the number of positive labels and negative labels is about 1:4. The imbalance may cause the model to favor the negative labels. We did downsampling on the negative samples to make them have the same size as the positive samples.

We applied the ‘train_test_split’ function from sklearn to divide the dataset into train and test groups for prediction. Meanwhile, we set the proportion of the training group to 20% and fixed the split data by setting the ‘random_state’ parameter.

Model training

After splitting the data set, we tried to train the data with different methods:

1) Logistic Regression: which is used the classification problems with discrete output;
2) Random Forest Classifier: which utilizes ensemble learning and consists of many decision trees, could get higher accuracy for a huge sample dataset;
3) XGBoost: which is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework.

The XGBoost models yield better performance on our dataset in most of our initial experiments. We decided to use the XGBoost model and further tune it to achieve better performance.

The hyperparameter tuning used randomized grid search based on the 3-fold cross-validation. Some important parameters of the model are listed below.

learning_rate = 0.69472,
max_depth = 2,
n_estimators = 400,
subsample = 0.996.

We also defined a function to output accuracy score, classification report, and confusion matrix.

Utilization of Analysis Results

Once the model is tuned and trained based on the training data, we will evaluate the model performance based on the testing data set. The final model is then deployed as a web application using the streamlit python library. The application takes loan information from the user, then it predicts the default probability of the loan. This application will help the investors to evaluate the credit risk associated with specific loans at the origination time.

Data Analysis

Dependent Variable

Due to the reason that we will focus on a prediction of loan status in this research, as a dependent variable, the distribution of loan status should be analyzed at first. Our general exploratory data analysis reveals the rise in popularity of Lending Club loans. The number of loans is rising year by year, except in 2020, since the dataset was collected by the end of the third quarter of 2020. As shown in Figure 2.

Figure 2

Number of Loans by Year from 2007 to the 3rd quarter of 2020

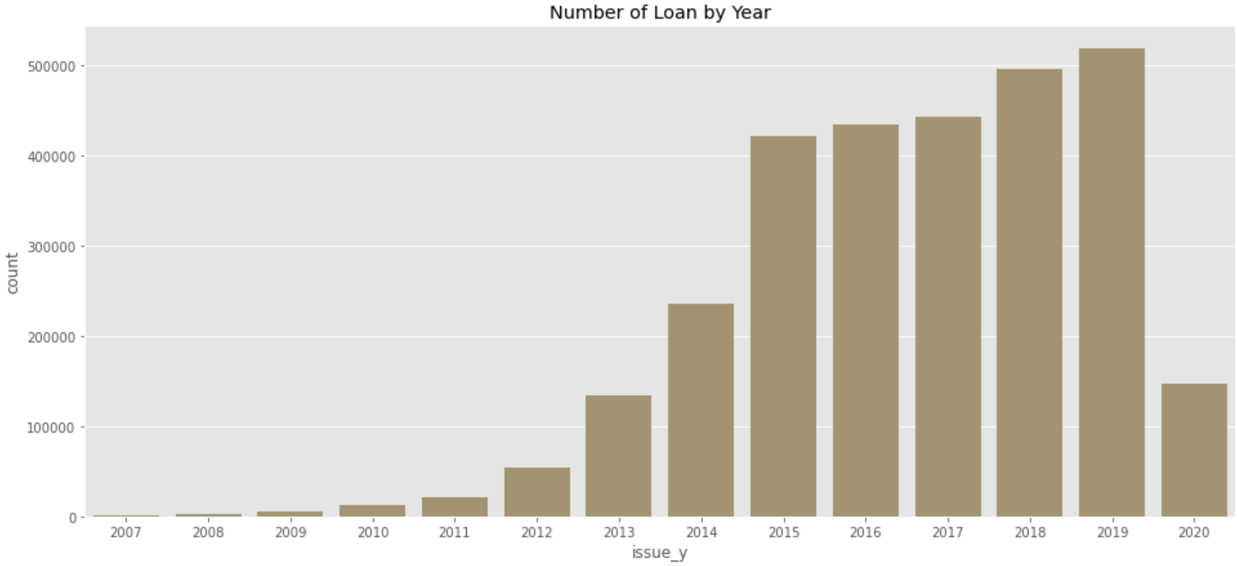
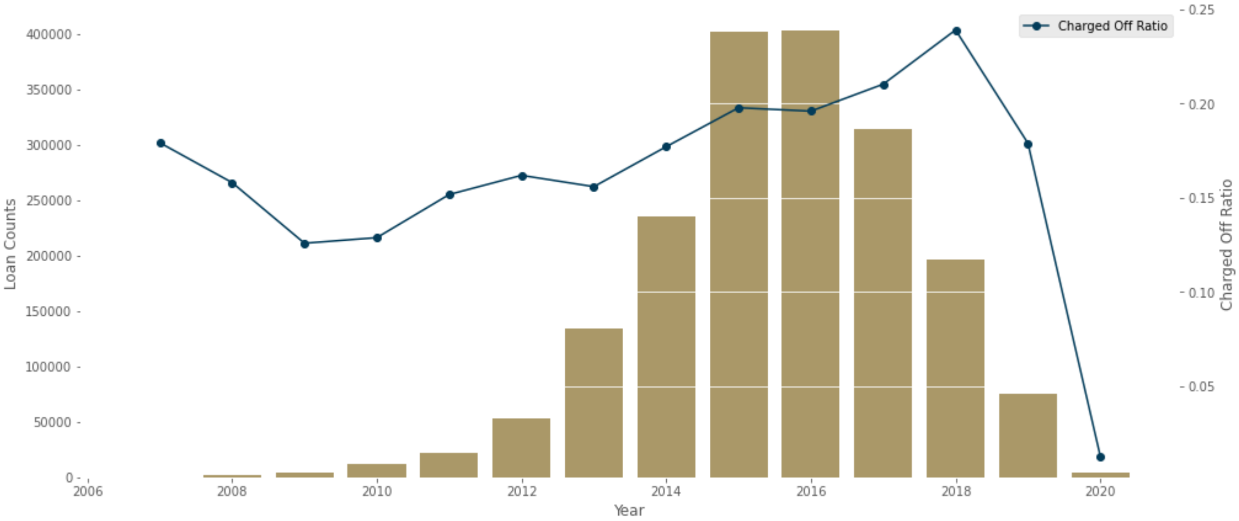


Figure 3 displays the number of completed loans by year. It's obvious that the number of loans reached its peak in 2015-2016, followed by a decrease from 2017. That's maybe because some loans are not completed yet. The charged-off ratio reached its highest in 2018 and dropped sharply.

Figure 3

Number of Completed Loan by Year



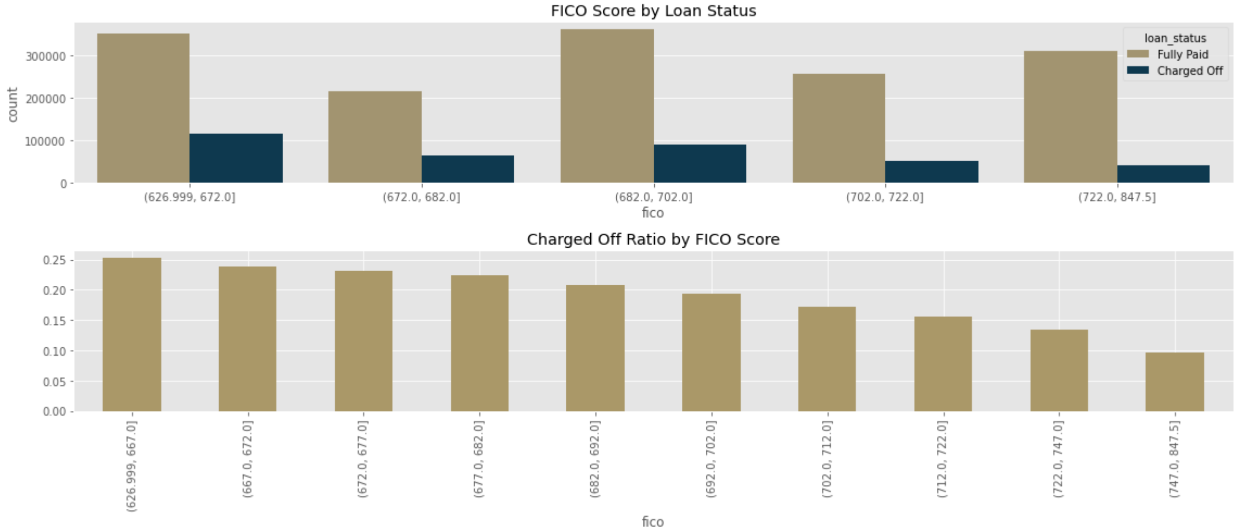
Predictors

In order to provide more information about the dataset, we also conduct the analysis of all independent variables, which will be applied to the prediction model. All the results are shown in the 'Image' file. Some observations are remarkable as below:

1) FICO Score is a credible indicator in general since it is calculated according to consumer behavior. It's obvious, as shown in Figure 4, that as the FICO score increases, the charged-off percent decreases. Charged-off loans have a FICO score 10 points lower on average.

Figure 4

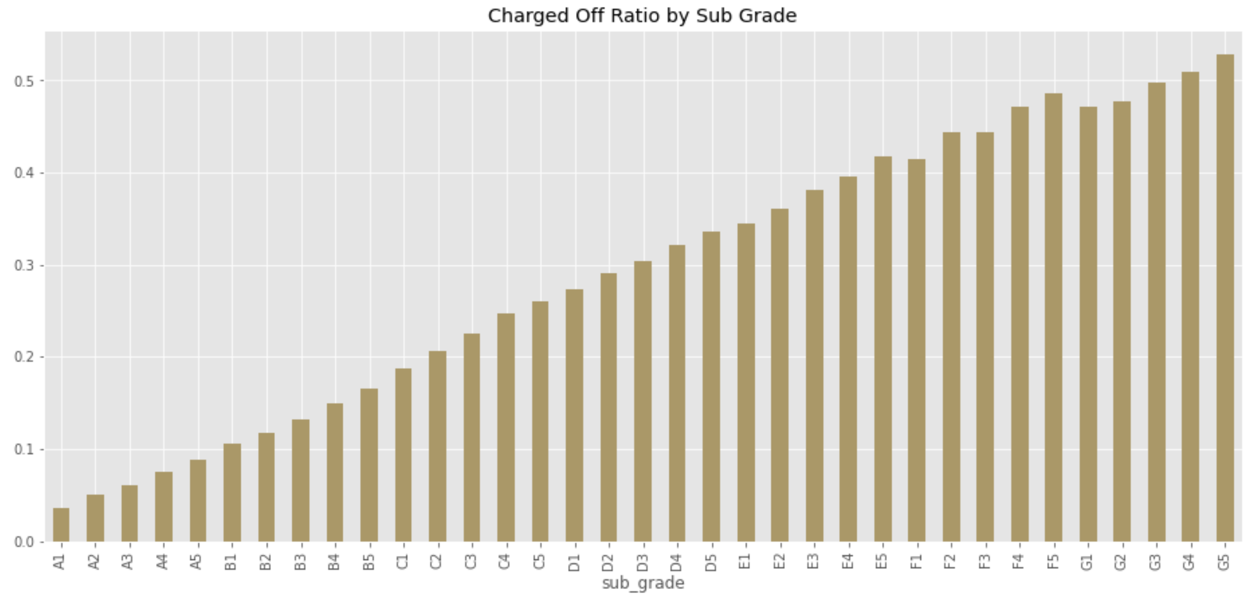
FICO Score Analysis



2) Every loan is ranked into grades from the best: A, to the worst: G. To be specific, each grade is divided into 5 levels from the best: 1, to the worst: 5. Figure 5 displays the analysis of sub-grade in terms of the loan status. As the subgrade gets worse, the charged-off rate increases gradually, and the worst level: G5 even shows a charged-off rate of more than 55%.

Figure 5

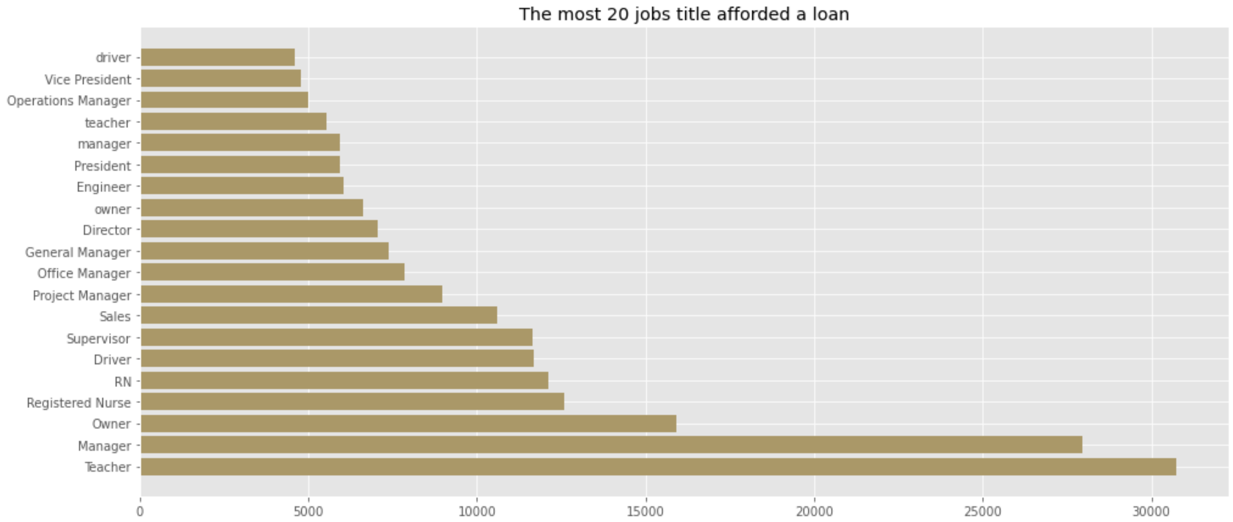
Sub-Grade Analysis



3) The top 20 occupations with the most loans are shown in Figure 6. Teachers surprisingly own the largest value of loans even before managers.

Figure 6

Job Title Analysis

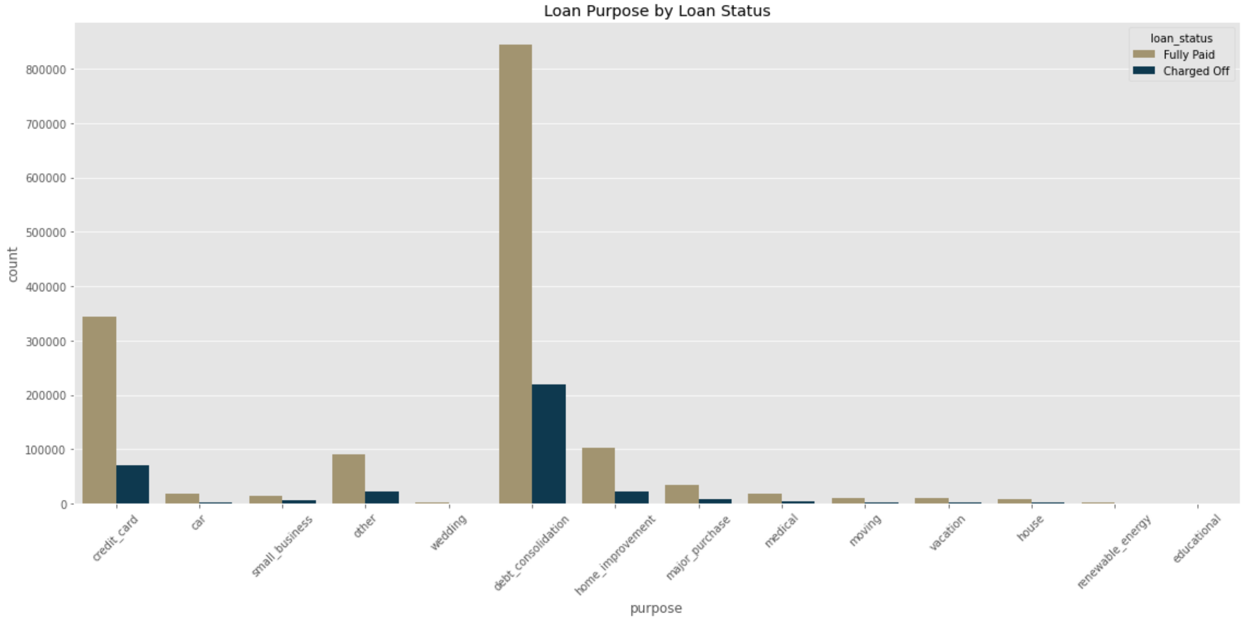


4) People take out loans mainly for debt consolidation or credit cards as shown in Figure 7. Loans for debt consolidation have the highest counts, also the highest counts of fully paid

and charged off. But the highest charged off percent belongs to loans for small businesses. And loans for weddings have the smallest charged-off percent.

Figure 7

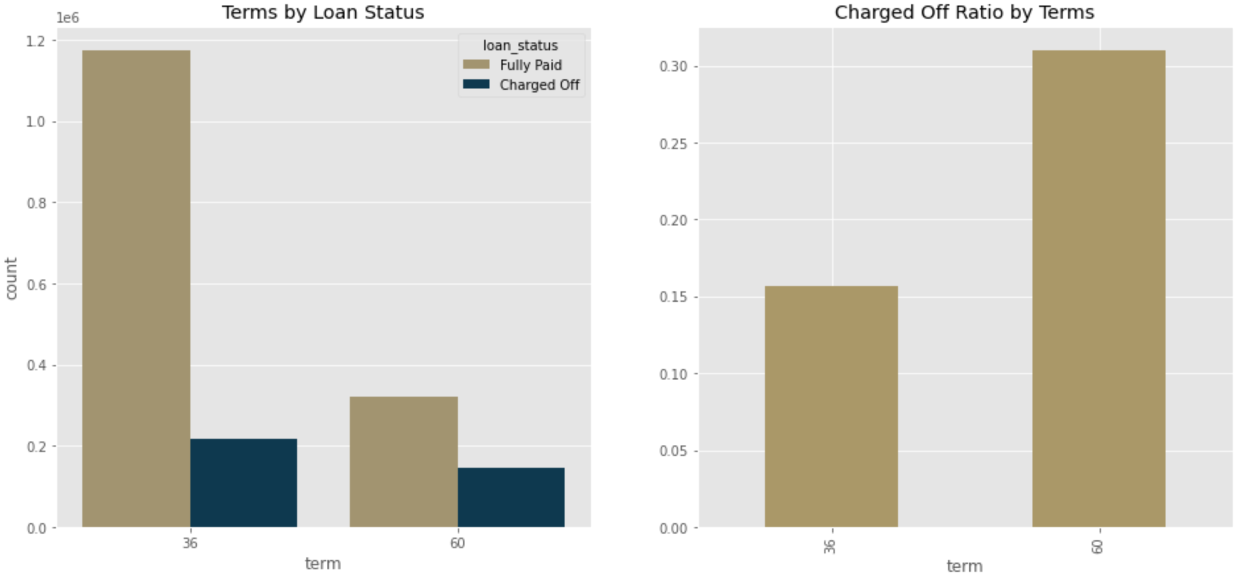
Loan Purpose Analysis



5) Lending Club provides two types of terms for the customers: 36-month and 60-month, as shown in Figure 8. The loan with 36-month payment times has the higher counts of either fully Paid or charged Off, but the lower charged off percent. The loan with 60-month payment times has the higher charged off percent. The trend shows that more people prefer short-term loans, and the charged-off rate is higher in long-term loans.

Figure 8

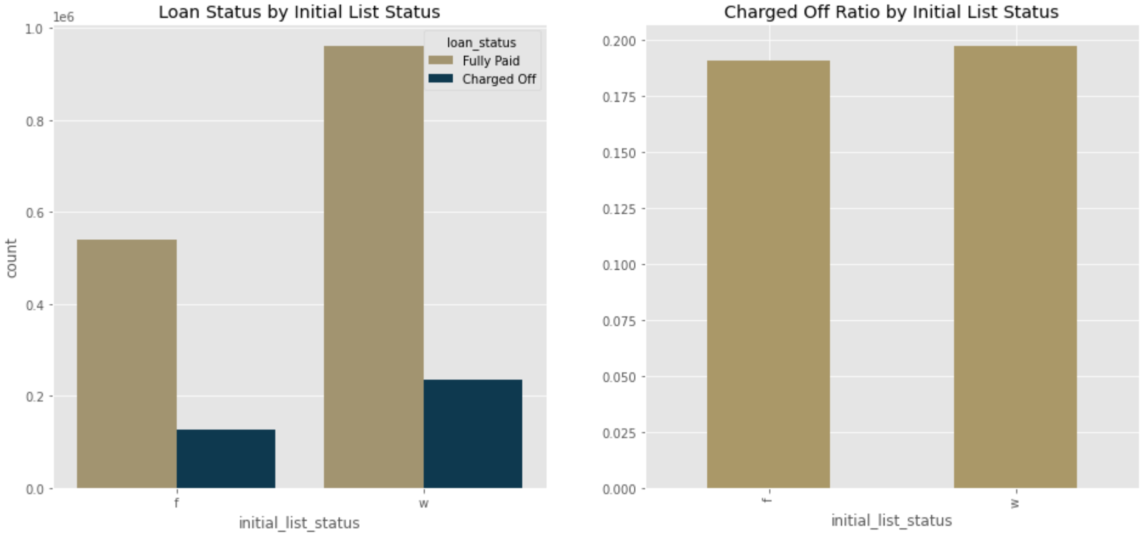
Loan Term Analysis



6) Figure 9 displays different types of loan applications, and 'W' means Whole, 'F' means Fraction. The number of individual applications is pretty higher than joint applications. But the charged-off percent of joint applications is higher. It seems Joint loans are slightly more likely to be charged off.

Figure 9

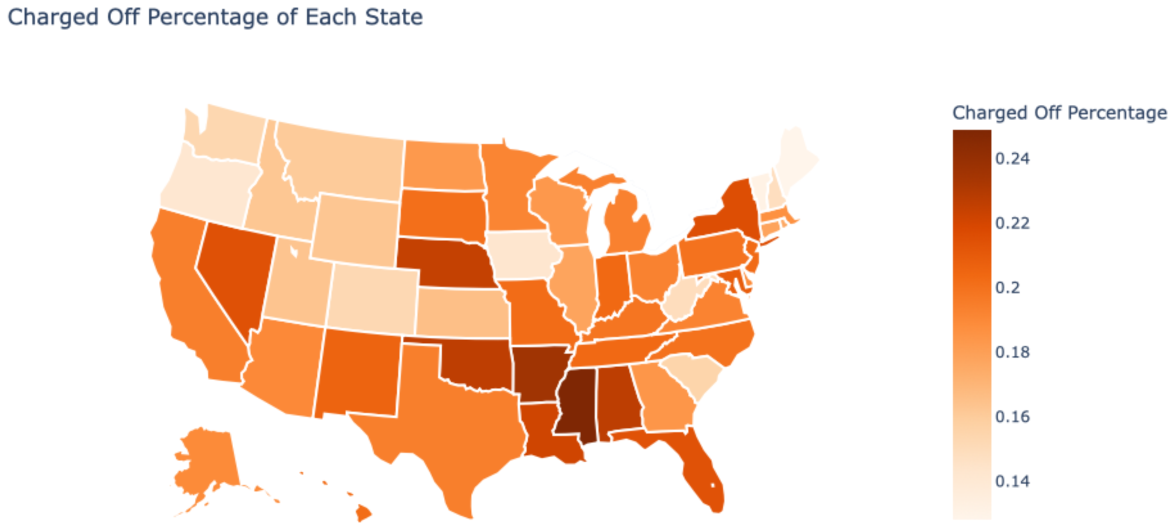
Application Type Analysis



7) We also visualized the charged-off rate into the map as shown in Figure 10. In the Northeast corner of America, Maine has the lowest charged-off ratio which is about 12.82%, but on the contrary, Mississippi has the highest ratio which is about 24.89% in southern America.

Figure 10

Charged-Off Rate by State



Main Results

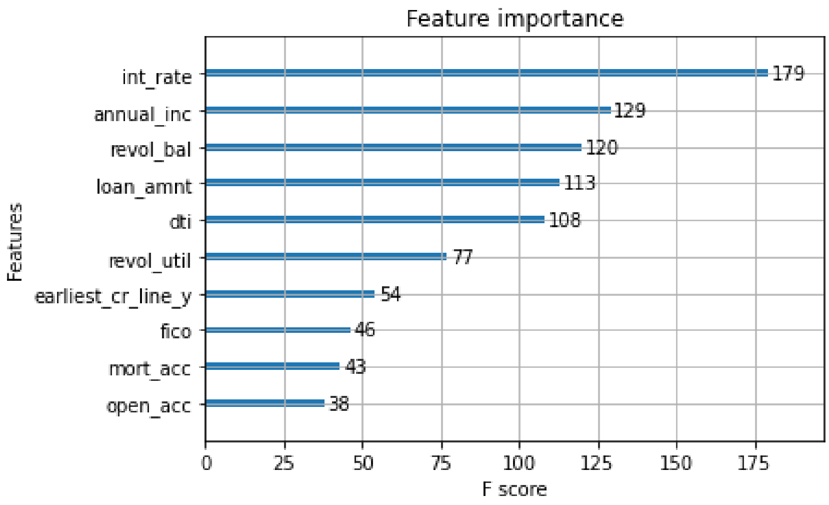
The results of different models are shown in Table 1. It shows five metrics: precision, recall, F1, ROC, and accuracy. The XGBoost achieves the best result on all metrics. The downsampling largely improves the recall. For example, the logistic regression model only achieves 0.06 recall when it was trained on data without downsampling. The downsampling sacrifices some accuracy, but the accuracy is relatively less important because correctly identifying default loans is more important in credit risk predictions.

Table 1 *Summary of model performance on the test set*

	Logistic Regression	Random Forest	XGBoost
Precision	0.65	0.65	0.65
Recall	0.53	0.66	0.68
F1	0.58	0.65	0.66
AUC_ROC	0.68	0.71	0.72
Accuracy	0.62	0.65	0.66

Note that the performance metrics are generally not very high. One important reason is that this dataset only includes the accepted loans. The loans that are obviously bad would not be accepted by the lending club. It increases the difficulty of predicting bad loans. Also, we didn't have access to the payment activities of the loan. The model can be further improved if we can combine the payment history over the life cycle.

Figure 11 *Top 10 important features and their importance scores of the XGBoost model*



The feature importances are shown in Figure 11 to help us understand the model mechanism. On the top of the list, there are interest rate, annual income, revolving balance, loan amount, and debt-to-income ratio.

The model is then deployed as a web application (link). Using this application, the user can specify loan information at the origination time and it will return the prediction for the loan status at the closing time.

Conclusion

This project aimed to build a predictive model to predict the loan status at the closing time given the loan origination information. We obtained the data from the Kaggle dataset service and conducted extensive visualization and data preprocessing. We built logistic regression, random forest, and XGBoost models to predict if the loan would default in the end. The most important features of the model include interest rate, annual income, revolving balance, loan amount, and debt-to-income ratio. The model was deployed as a web application to score the risk of a loan. The XGBoost model achieved 0.66 F1 and 0.72 ROC. The weak performance of the model is due to the fact that the dataset only included accepted loans and the model only used origination time information. In practice, we can choose a high threshold close to 1 to elevate the precision by sacrificing the recall and use this model as a preventive tool for the investors to avoid risky loans.

References

Chaitali Majumder, Thomas Kim. (2021, May 5). *Investment in Lending Club - Loan Default & Investor ROI Prediction*. NYCDATA SCIENCE ACADEMY. https://nycdatascience.com/blog/student-works/investment-in-lending-club-loan-default-investor-roi-prediction/

Gonçalo Guimarães Gomes. (2021, June 10). *Machine Learning: Predicting Bank Loan Defaults*. towards data science. https://towardsdatascience.com/machine-learning-predicting-bank-loan-defaults-d48bffb9aee2

Philippe Heitzmann. (2020, December 1). *Predicting Loan Defaults using Machine Learning Classification Models*. NYCDATA SCIENCE ACADEMY. https://nycdatascience.com/blog/student-works/predicting-loan-defaults-using-machine-learning-classification-models/#portfolio

Appendices

Appendix A

Lending Club Data Dictionary

Appendix B

Image of EDA

Appendix C

Code of EDA

Appendix D

Code of Model