Loan Default Classification Using Machine Learning

Miller (Yijia) Wu
Will (Yiyang) Zhang
Carey Business School @Johns Hopkins University

I. EXECUTIVE SUMMARY (BLUF)

- The study focuses on exploring the relationship between different factors and the probability of loan default to minimize the risks taken by financial institutions. In this case, we are using two classification models (Decision Tree and Light GBM), and find that interest rate spread, upfront charges are two significant drivers of increasing the probability of default on loans.
- Overall, the results sound very promising with an AUC approximately equal to 1 by using the Light BGM model. We assume that the models fit well in this particular dataset. If the explanations hold up on similar datasets, their use in the banking sector or other financial-related area could be explored.
- Some future efforts could be done to strengthen the model and reduce the model complexity to achieve a better result.

II. DATASET INTRODUCTION

Banks earn a major revenue from lending loans. But it is often associated with risk. The borrower may default on the loan. To mitigate this issue, the banks have decided to use Machine Learning to overcome this issue. They have collected past data on the loan borrowers & would like you to develop a strong ML Model to classify if any new borrower is likely to default or not.
The dataset is enormous & consists of multiple deterministic factors like borrower's income, gender, loan purpose, etc. The dataset is subject to strong multicollinearity & empty values. Our goal is to overcome these factors & build a strong classifier to predict defaulters.

III. EXPLORATORY DATA ANALYSIS

The original dataset we retrieve has 14860 entries with 32 variables. The variables “ID” and “year” are useless and we dropped them at the beginning. Then we checked missing values and brings up a heatmap (Exhibit 1) describing the missing value patterns. What we found interesting here is that the missing values mostly fall into the numeric variable category.

We also checked duplicate values (in total 0) and unique values (mostly numeric) and separated the variables into the quant, dummy, and label respectively.

Then, we assessed the skewness of quantitative variables, “loan amount”, “upfront charges”, “property value”, “income”, and “LTV” are the one's ned skewness corrections. We also found that due to the enormous numeric values in quantitative variables, many outliers need to be fixed.

We also created a pair plot (Exhibit 2) of the quantitative variables and a correlation heatmap (Exhibit 3) for these variables.

IV. DATA PREPROCESSING

The first thing we do in preprocessing is to impute the median of numeric missing values and drop the categorical missing values. Next, we encoded the categorical variables (dummy and label encoding), which enlarged the number of columns to 47. After the dataset is cleaned, it is time to address the skewness issue, and we transformed the skewness of the above-selected variables. We also used the Tukey rule to identify the outliers and winsorized those outliers with plots (Exhibit 4).

So far, we split data into Xs and y, and by plotting the target variable, we found out that it is highly unbalanced, only a quarter of data points is 1 while the rest are 0 (Exhibit 5). Then, we split the dataset into training and testing data with a proportion of 80%, and 20%.

V. DATA MODELING & EVALUATION

To build the model, we firstly SMOTE our dataset to get a balanced target variable (174,008 instances) and standardized the SMOTE training and testing data. Secondly, we come up with feature importance using the LGBM classifier to select the best predictors in our model. Initially, we selected nine variables, but it increases model complexity and eats RAM. Therefore, we finally determined the three best predictors. They are “rate of interest”, ”interest rate spread”, and ”upfront charges” (Exhibit 6).

When building the models, we utilized the cross-validation method of 5 folds in a pipeline to increase model performances and choose the best out of the others. The reason behind using cross-validation instead of holdout validation is to avoid uncertainties and randomness generated by different models.

We find out that nearly all the models generate very high scores. Thus, after considering the performance of each model, the LGBM classifier is selected as the best model with the highest scores. At this time, we grid searched learning rate and bagging fraction for the LGBM classifier, and found out that when bagging fraction = 0.1, and learning rate = 0.05, the model generates the highest AUC of 0.999977. We are using AUC/ROC as our primary metrics for evaluating the performance of the model. With such a high score, we assume that the model fits very well with the dataset with simplified variables. As a result, there is only 1 false-positive value given by the model. Then we created a bar chart plotting the propensity with a sample of 500 results (Exhibit 7).

After the model building and hyperparameter tuning, we used XAI tools for us to better understand the results. Both the permutation feature importance (Exhibit 8) and Shapley values (Exhibit 9) explain that “interest rate spread” and “upfront charges” contribute the most to our target variable status of default. In the Partial Dependence Plots, we find out that in “upfront charges”, only when the grid points are approximately between 0 and 0.4, showing a higher probability of default. The same logic applies to “interest rate spread” as well. However, when it comes to “rate of interest”, people are less likely to default when it exceeds 0 on the grid point.

Finally, we built a surrogate model (Exhibit 10) as a proxy to walk through the story of whether a person will be likely to default on the loan. We chose a depth = 3 to increase the accuracy of prediction. In the first story, when the interest rate spread is less than 0.39, no one will default. In the second story, if the interest rate spread is larger than 0.39 and less than 0.39 in decimal points with an upfront charge smaller than 46.692, all people will default. Therefore, to mitigate the risk for the financial institution, it is better to let the applicant of the loan applications for an asset-backed loan. As a result, the correlation coefficient of LGBM prediction and surrogate model prediction is equal to 1.

VI. DISCUSSION & LIMITATIONS

A. Recommendation on Model

We would highly recommend financial institutions use our model as a starter to build their machine learning models because our results are good enough for now. Although the model is highly accurate, we still need more datasets to train the model to stabilize the performance. In the modeling part, we see not only the LGBM classifier runs a good result, but also XGBoost, Decision Tree. Therefore, if we can apply the models we generated with a similar dataset, it can be used widely in the realm of finance and banking.

B. Potential Risks

Although our model runs good results, there are still limitations and defects in the way we process the dataset. We only impute median values for the missing values, but there are better ways of doing it at the hands of financial technology professionals. They understand those data better than we do. Also, even though we are deleting almost all the variables in the construction of models, there could be risks of biased models or potential overfitting. Therefore, if people are specializing in the evaluation of the model to find and fix such problems, we are better off then. Lastly, the model we built takes time to run and evaluate, and there are automated AIs on the market to do the jobs for us. They could lessen the limitations we have currently and perhaps do a better job.

VII. CITATION

M. Yasser H, Loan Default Dataset, *Kaggle,*<https://www.kaggle.com/datasets/yasserh/loan-default-dataset>

VIII. Appendix

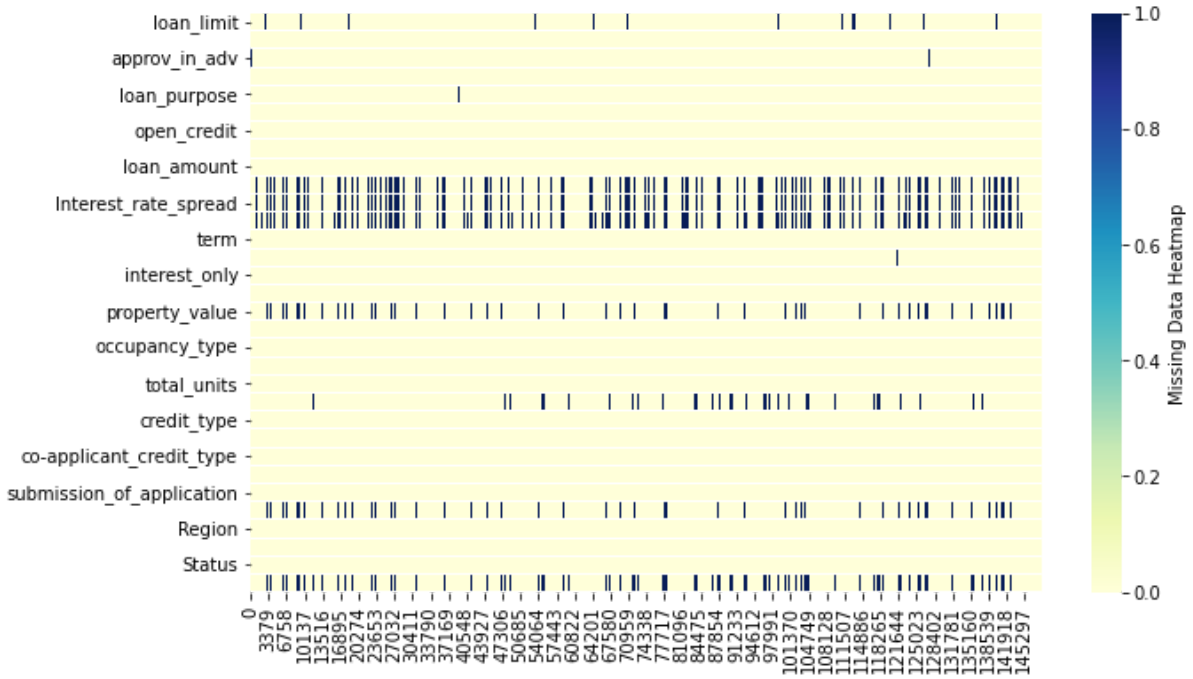

Exhibit 1

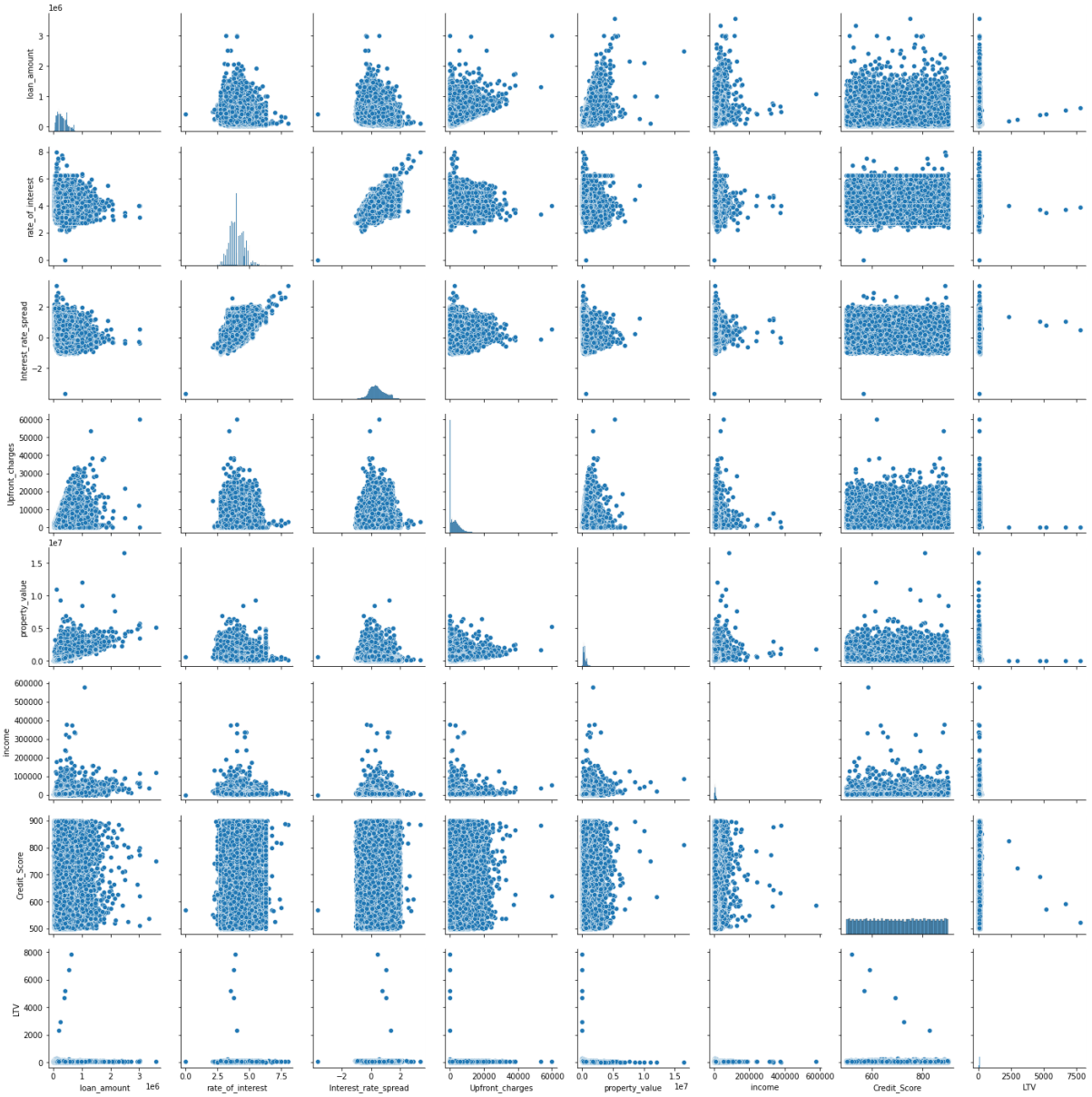
loan_amount
rate_of_interest
Interest_rate_spread
Upfront_charges
property_value
income
Credit_Score
LTV
1e6
1e7

Exhibit2

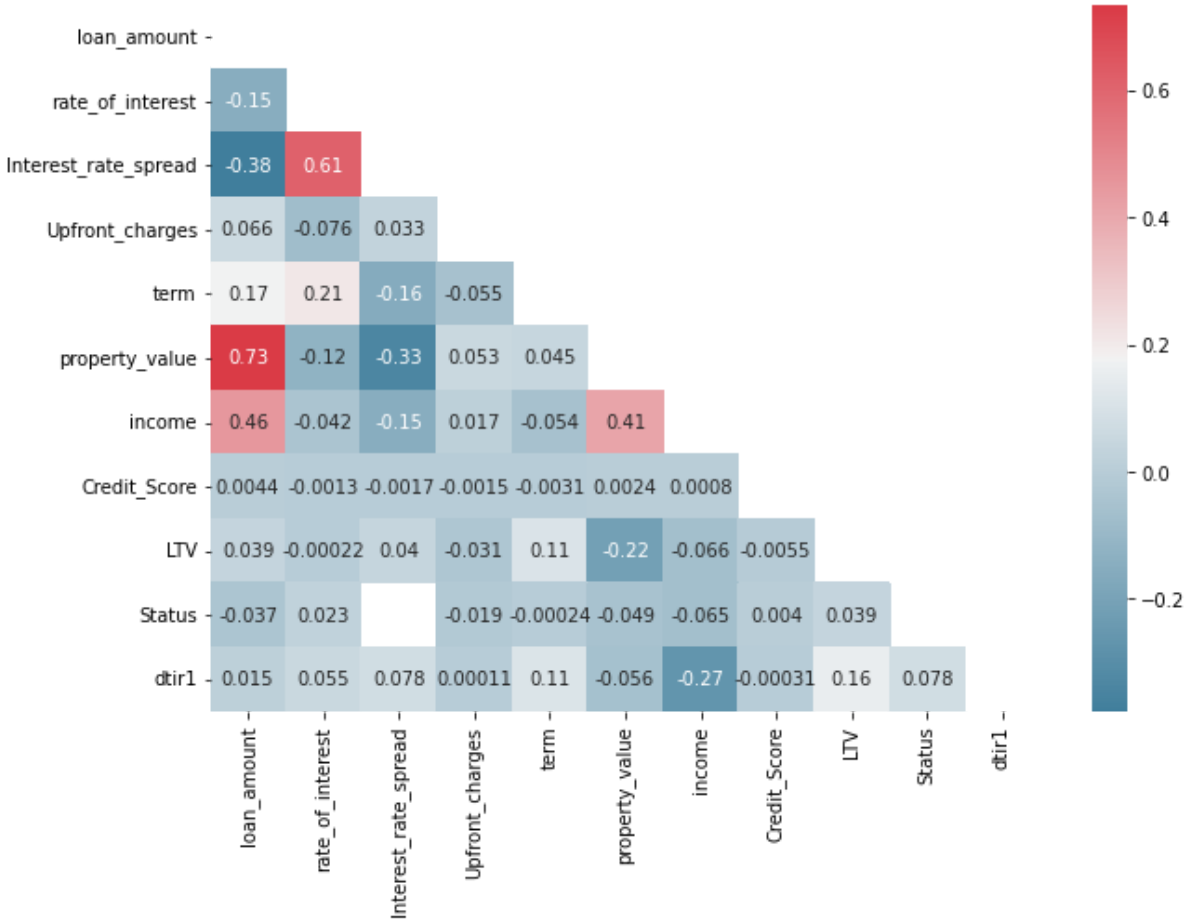

Exhibit3

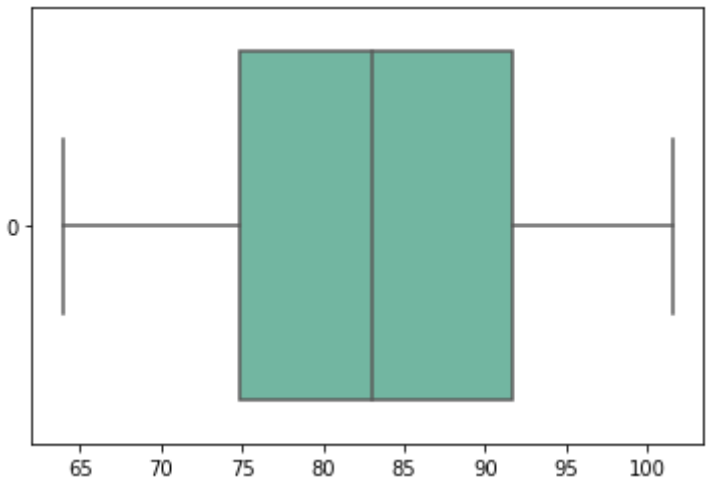

Exhibit 4.a(loan_amount)

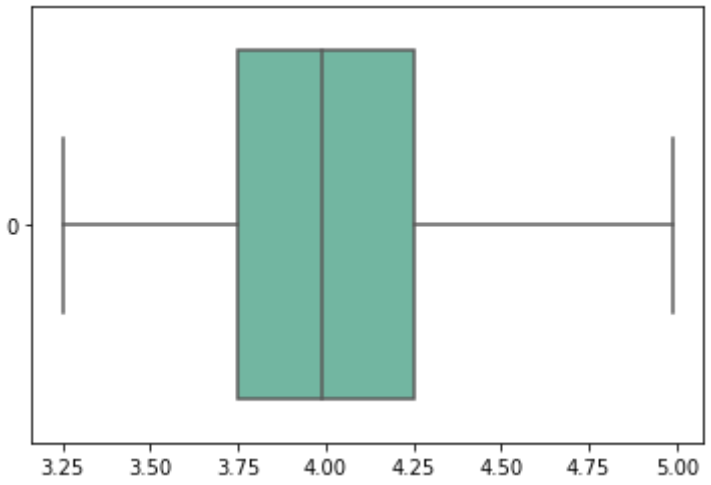

Exhibit 4.b(rate_of_interest)

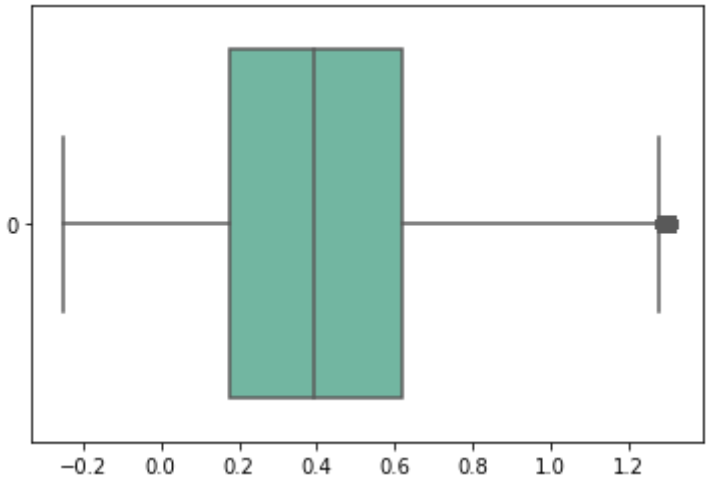

Exhibit 4.c(Interest_rate_spread)

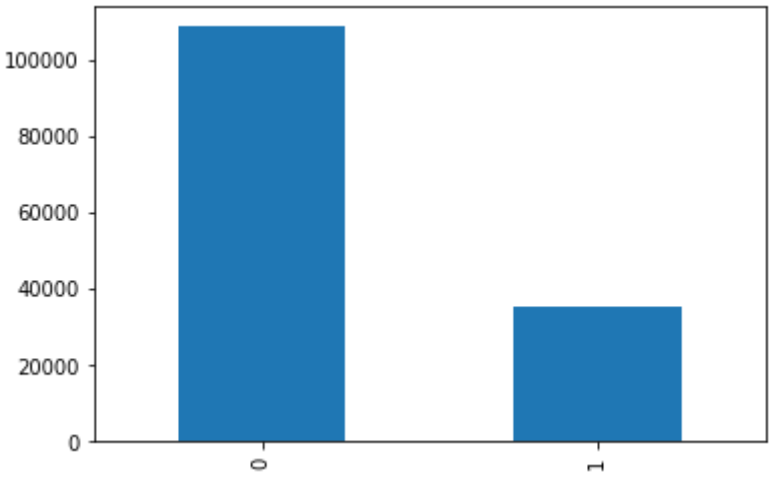

Exhibit 5

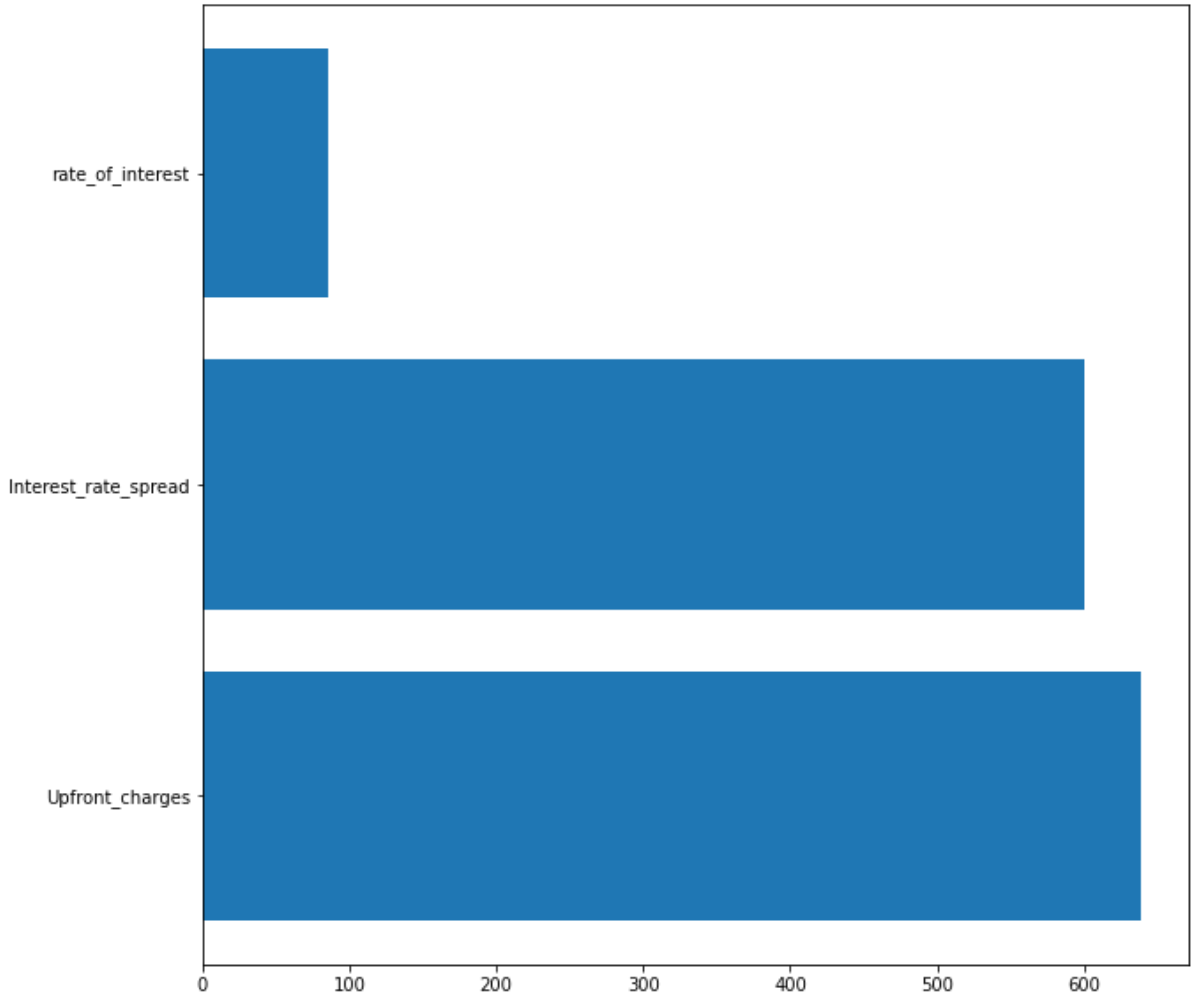
rate_of_interest
Interest_rate_spread
Upfront_charges
0
100
200
300
400
500
600

Exhibit 6

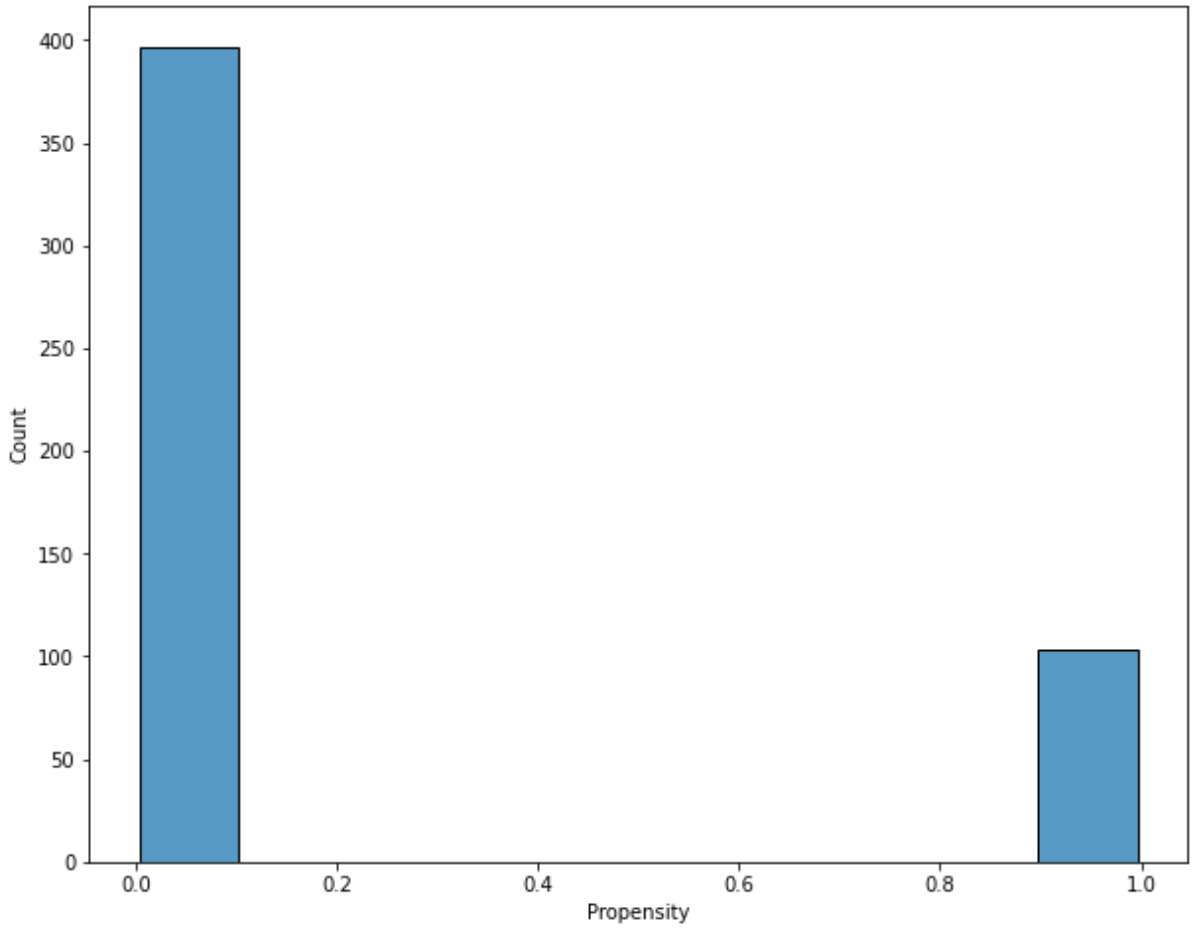
400
350
300
250
200
150
100
50
0
Count
0.0
0.2
0.4
0.6
0.8
1.0
Propensity

Exhibit 7

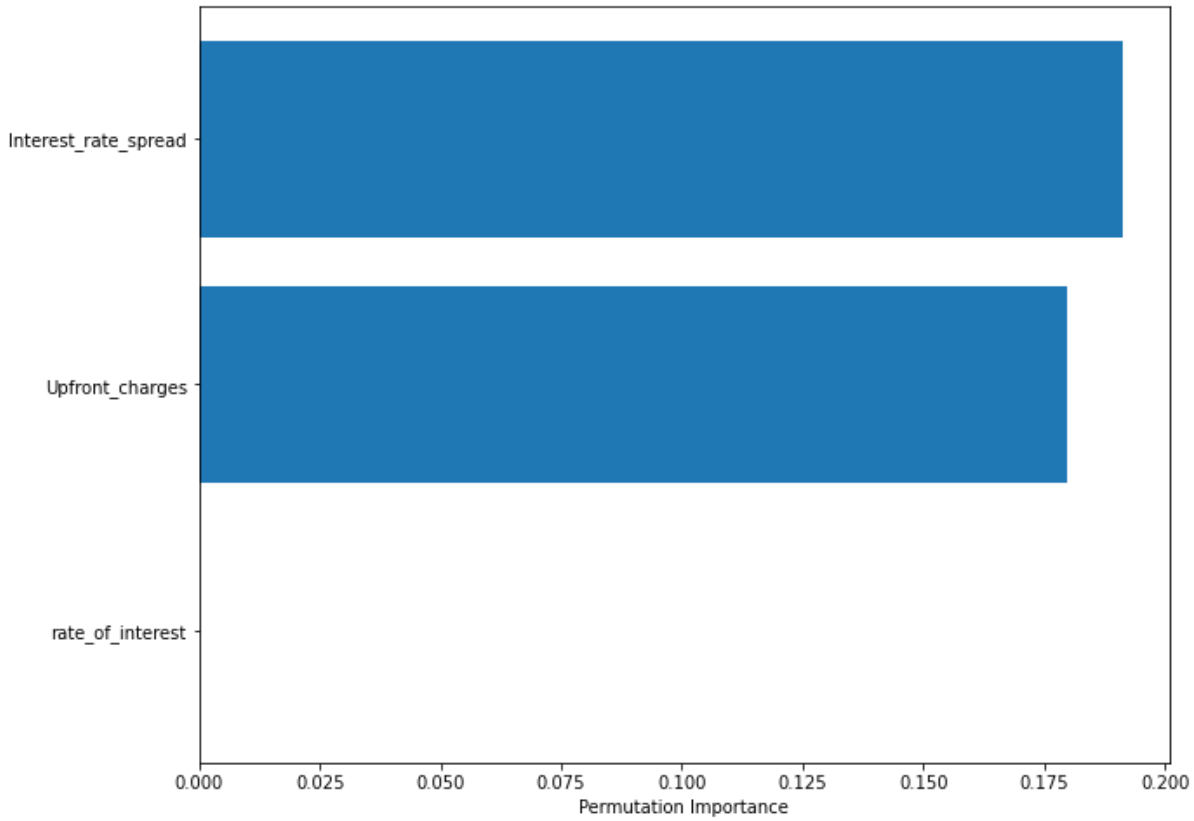
Interest_rate_spread
Upfront_charges
rate_of_interest
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
Permutation Importance

Exhibit 8

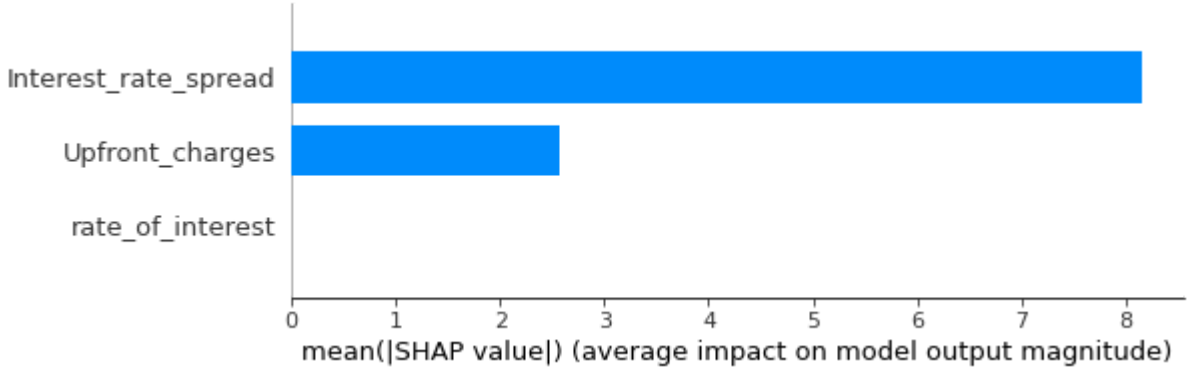

Exhibit 9

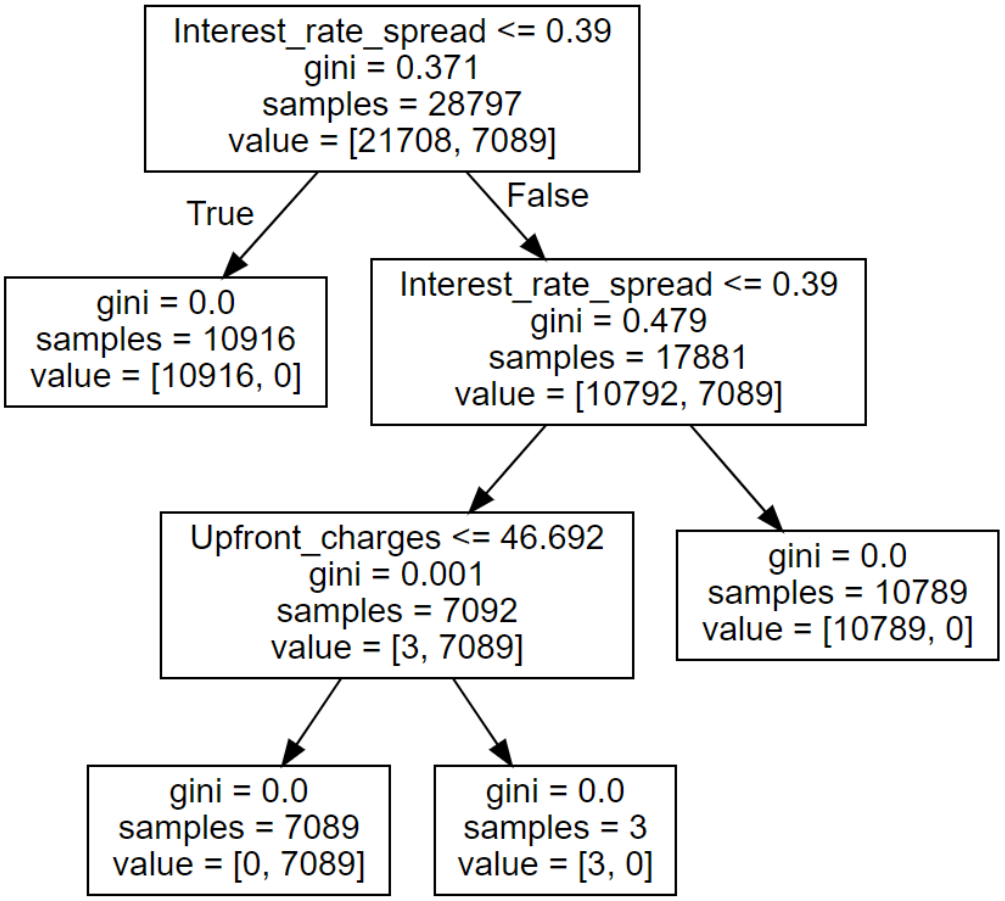

Exhibit 10