Data Science for Business DECISION 520Q

Section C Team 06

Bryan Huang, Weiqi Li, Yashuo Wang (Gloria), Van Xu, Jose Zuart

Term Project Report

Business Understanding

According to information presented by the World Bank, the percentage of nonperforming loans to total gross loans in India has been rising dramatically since 2011. Currently, 7.39% of the total gross loans were nonperforming loans, which is about 200% higher than the level in 2011.[1] Defaulting on a personal loan is generally considered as a civil dispute in India. Once an individual defaults on their loan, banks will have to hand them over to professional recovery agents. Eventually the banks will need to resort to either a “Hair Cut” or the SARFAESI Act to recovery at least part of the loan, a time- and money-consuming process.[2]

To minimize the chance of customers defaulting on their loans, commercial banks require applicants to go through a very detailed screening process. During the process, banks want to improve their ability in identifying individuals with higher risks of defaulting on loans. Once identified, these individuals can be passed on to further scrutiny/analysis. Our motivation for the project comes mainly from the belief that more accurately estimating defaults can considerably increase the profitability of banks.

Banks or financial institutions with more powerful data mining techniques and forecasting models will have an advantage in being able to identify the probability that a

[1] https://data.worldbank.org/indicator/FB.AST.NPER.ZS?end=2020&locations=IN&start=2008&view=chart
[2] https://getmoneyrich.com/what-can-be-done-if-you-are-not-able-to-pay-back-your-loan/

client defaults, then charge rates and fees or turn down loan applications more accurately (to different customers), thus maximizing profits while lowering risks.[3] According to Stein's research work on the relationship between default prediction and lending profits, a bank can increase its profitability by around 11 basis points (BPS) for every dollar, which translates into a possible .11% increase in the income for every dollar in loans. According to Standard & Poors the market size for housing loans in India is 280 billions of dollars. Therefore, there can be a significant increase in the profitability of the business.

Data Understanding

We obtained the public personal loan data from Kaggle.com.[4] The dataset consists of 252,000 samples from the Indian commercial banking industry, with each sample representing the individual's personal information as well as the status of the risk flag that is given by the bank. 13 features are available in the dataset, including Id, Income, Age, Experience, Married/Single, House Ownership, Car Ownership, Profession, City, State, Years on the current job, Years on the current house and Risk flag.

Among all these variables, Income, Age, Experience, Years on the current job and Years on the current house are numeric variables that take on consecutive values, while the following are categorical variables:

Married/Single (single, married), House Ownership (rented, norent_noown, owned), Car Ownership (no, yes), Profession (51 professions), City (317 cities), State

[3] https://www.sciencedirect.com/science/article/pii/S0378426604000895
[4] https://www.kaggle.com/subhamjain/loan-prediction-based-on-customer-behavior

(29 states). Risk Flag is the target variable that we will need to predict, and it is a binary variable that takes the value of either 0 or 1.

Data Preparation

For the preparation of the data, a statistical summary of the variables was made, as well as checking that there was no missing data in the columns (Graph A). However, the database contains data in all columns.

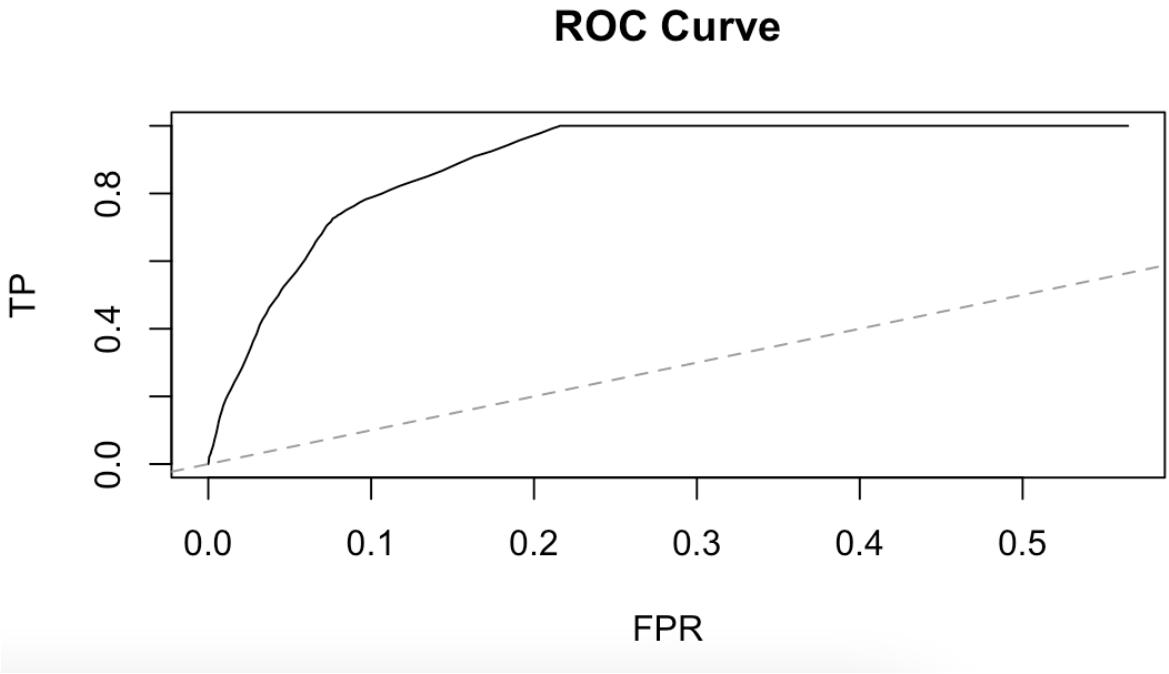
To understand the relationship of the variables between them. A correlation analysis was performed, with the table presented in the Appendix (Graph B). Which gives us an overview of the interaction of the variables.

We dropped the “ID” column from the original data since it is unrelated. The new dataset was then divided into a training set(202000 samples) and a holdout testing set(50000 samples).

Modeling

Our best model yet is the Random Forest model, with nodesize=5, ntree=500 and mtry=4. The AUC of the model is 0.9387 and the OOS Accuracy is 0.875104. We set the threshold to 0.1, which gives us the best result for our business problem--focusing on targeting high true positive rate. By setting the threshold to 0.1, the model eliminated 99.6 percent of possible loan defaults while keeping the mistakes at an 21.3% acceptable level. The pro for this model is that it gives very accurate predictions in comparison to other models we tested. The con is that this model is

computationally inefficient. If other commercial institutions want to modify our model or train it on a larger dataset, it would be challenging. (GRAPH K)



ROC for Random Forest

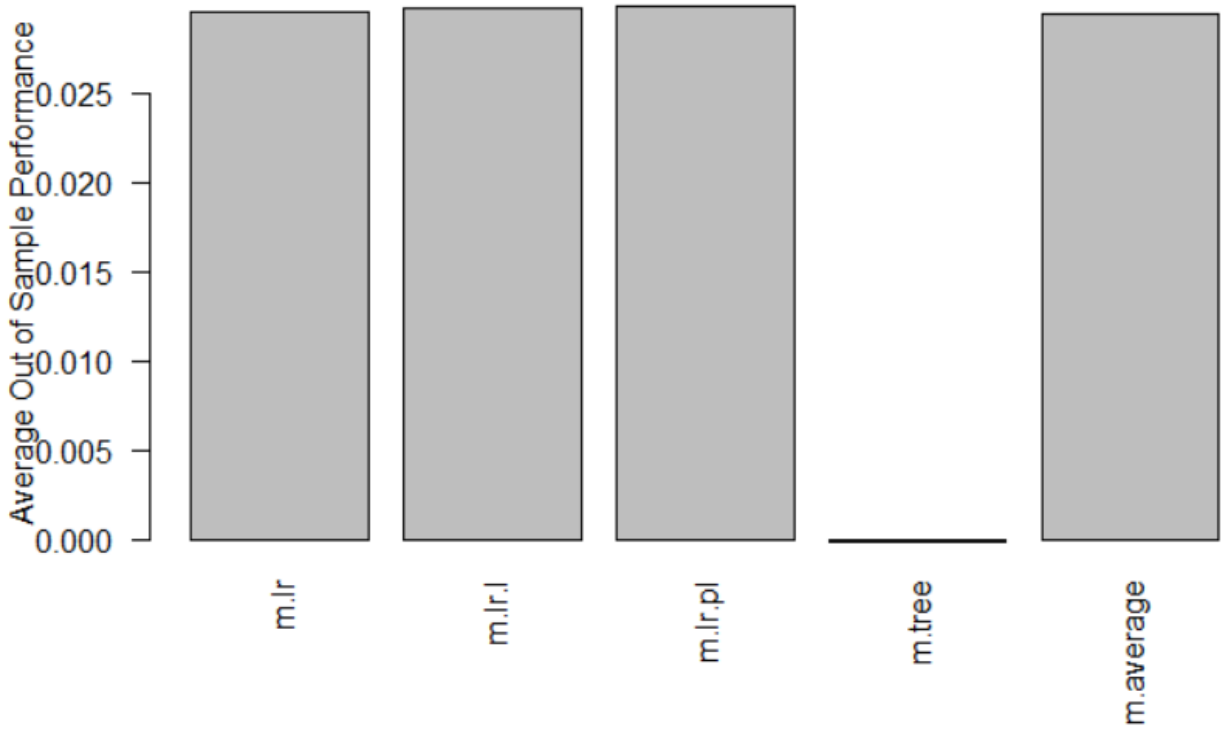
Our best alternative is the Logistic Regression model. The AUC of the model is 0.6518, out-of-sample accuracy is around 0.7897643. See the ROC curve in the Appendix (Graph G). In comparison to the Random Forest, the Logistic Regression takes a shorter time to run, but the prediction accuracy is also much lower.

We first started with K-means clustering. When we tried to select k for the K-means, we first had the regularization idea via information criteria (IC). Based on the HDIC result, our most optimal k value is 17. However, both BIC and AIC have optimal k values over 50, which may lead to severe overfitting since our dataset contains 252000 observations.

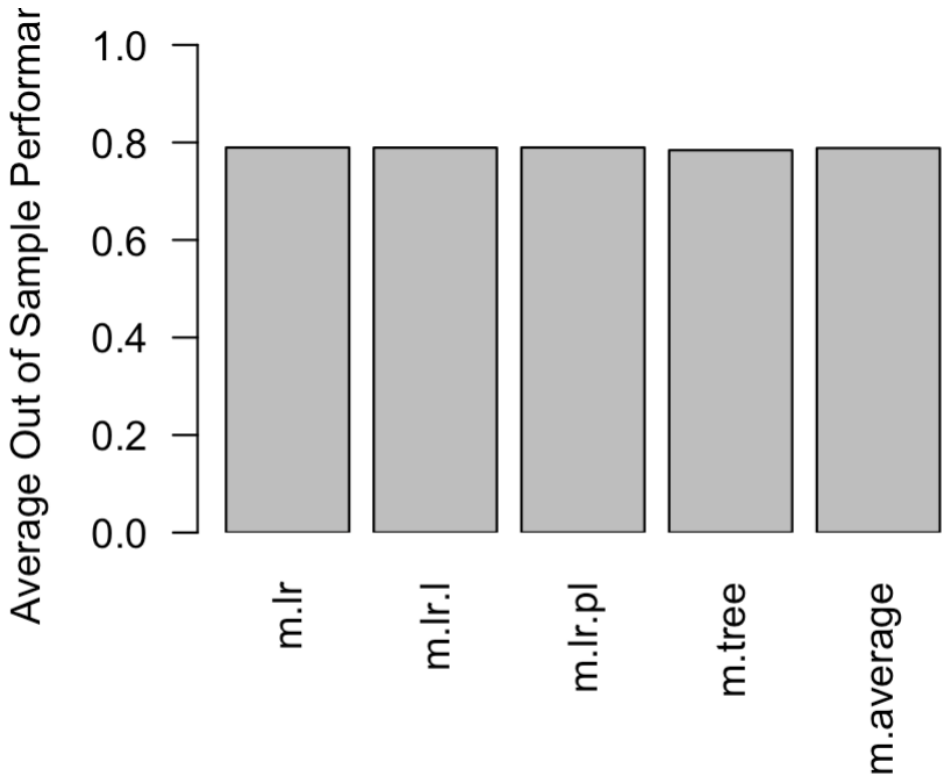
Since the dataset we are using is supervised. K-means clustering does not work for our dataset, therefore we abandon such methods in our modeling.

We then run the PCA. However, since the data does not contain any unlabeled data sets. (GRAPH D) PCA does not work effectively. What is more, if we take a close look at all PCA, they are all complicated combinations of large numbers of features which certainly do not help us to build the model.

Our attention turned towards Classification Tree and Regressions. We first ran Cross Validation for OOS R^2 and OOS Accuracy on Classification Tree, Lasso, Post Lasso, and Logistic Regression (GRAPH E & F). Random Forest was deemed time-consuming and not realistic to include in this step due to our hardware limitations. All models performed well in the OOS Accuracy test. However, in the OOS R^2 test, the numbers are relatively low. We kept that in mind and continued to test the individual performance of models from the ROC perspective.



OOS R^2



OOS Accuracy

For the Classification Tree, the initial model with all variables failed to deliver a satisfactory result. The AUC of the model is 0.5, which indicates that it is no better than the null model. We tried to use different combinations of independent variables but nothing seemed to work to improve the performance of the model. See the ROC curve in Appendix (Graph I).

We also tested Lasso and Post-Lasso. The OOS Accuracies are 0.7897643 and 0.7901111 respectively. The AUCs are 0.6363 and 0.6367 respectively. These results are not very ideal so our team moved these models to the bottom of the list.

By far, the best candidate is the Logistic Regression. The OOS Accuracy is 0.790094 and the AUC is 0.6367. Since it is only slightly better than Post-Lasso, our team decided to make a final attempt using Random Forest.

For the Random Forest, we used nodesize=5, ntree=500, and mtry=4. The result was encouraging. The model will efficiently identify those who are more likely to default on their loans and flag them, giving the banks another line of defense when it comes to personal loan applications.

Evaluation

The result of the data mining should be evaluated by TPR and FPR. A high TPR and low FPR would suggest that the predictive model is sufficient.

Because of our high TPR rate (0.996), nearly all risky customers will be given the risk flag. Serving with risky customers is costly since defaults will cause huge loss to the bank. Dealing with default is also time-consuming for the bank. Therefore reducing the risk will definitely help the bank on not only cost but also time. With the precise prediction on that, banks can reduce a lot of cost on the default customers since they all have risk flags.

On the other hand, when we consider FPR, the rate is relatively low, about 0.2. It means only a 20 percent chance that a person with a flag is not risky to have that flag. However, in the bank's view,being conservative in a mortgage market is not a bad thing to do. 0.2 is also not so bad that will make the bank lose a large amount of customers because of the risk flag given.

At the current stage, it is very difficult for us to project expected profit that the model will bring about. We need more data on the interest rates/fees that banks charge on their “good” and “bad” customers and the current screening procedure that banks are utilizing. We can, however, monitor the monthly loan default rate as well as the cost related to default loan recovery before and after the model is deployed into the screening system, or observe two similar banks, one with and one without the help of the model. With more data on loan default rate and cost, we will be able to develop strategies to project the expected benefit of our model. A/B testing can also be conducted to see how much improvement our model can bring to the current system.

Deployment

This model can serve as the first defense for personal loan application in Indian commercial banks. The banks can put new applicants’ information into the model and identify those that are “flagged”. They can then pass these individuals on to further investigation to see if the flags are justified or not. Those whose flags are “justified” will be handed over to the next department to determine if their applications should be rejected or if they should be given higher interest rates/fees considering their higher chances of defaulting; those who are mistakenly flagged (false positive) will be put back into the general loan application screening process.

Since we sacrifice 20 percent in FPR, that may be reasonable for the bank to be conservative because we use 10 percent increases in FPR and then get a 20 percent increase in TPR (GRAPH K). It should be economically beneficial for most of the banks since the cost of getting a default would overweight revenues generated by a new

customer, but some banks may want to change based on their situation. Their business strategies change with respect to the loan amount they hold and their financial situation. If a bank wants to experience some risk in order to issue more loans, one can just change the threshold and use a new predicted model. Based on our attempts, Random Forest is the best model.

The risk flags will potentially bring biases onto the customers' profiles, and they might be judged unfairly when interacting with the banks for other services that are not related to loans.

The potential risks associated with our proposed plan is that the firm might be too dependent on the predictive model when evaluating customers' risk flag. Though our predictive model can predict whether the client will default on a loan or not, there are certainly other factors that might affect such outcomes. In order to mitigate such risk, we can collect more information about each individual client to have a better understanding about the client's behavior.

Appendix 1

Contribution:
Report: everyone
R Code: everyone
Slides: everyone
Presentation: Gloria and Bryan