PROJECT ARTIFACT on

Underwriting Loan Applications

Advance Data Mining and Analytics

Project by,

Aditya Putta

Gordon Wall

Vijay Kumar Bollina

Mallikarjun Sasnur

Team Member Contributions	
Data Cleaning and Exploration	Aditya and Vijay
LGD Model	Aditya and Gordon
PD Model	Mallikarjun and Vijay
PPT and Report	Gordon and Mallikarjun

Table of Contents

1. Objective

Our main objective in this project is to build a Machine Learning Model to check whether we should lend money to a customer. Most of the profits made by banks are by investing money in people or companies. It is crucial to know the customer before we provide him/her with a loan.

We need to answer few questions before approving a loan

a. Can the customer repay the loan amount within stipulated time?
b. Is the customer going to miss any installments?
c. After analyzing these risks, do we approve him/her for a loan?

As underwriting teams for a bank is crucial to detail whether the customer would defaulter or not. This project will help the bank to better understand its customers in a better way. So, they can predict accurately and efficiently how risky is the customer.

2. Data

2.1 Overview

Data provided for modeling consists of 763 attributes with 80,000 observations. Few attributes of the data are significantly accounting for the prediction whether it's a good loan or bad loan.

In the given dataset, features are of two types. Some of parameters like rate of interest, principle amount, transaction ID, and percentage of default are related to banking. Parameters like age, annual income, credit score, age of credit history, payment history, credit card utilization and so on.

All attributes of the data given are of numeric data type. These contain both integer and double type.

2.2 Preprocessing

a. Initially, we omit all the duplicate columns and impute the missing values by utilizing a few prevalent imputation techniques like
   - kNN imputation
   - median imputation
   - mean imputation

We can use any one of the above imputation techniques and end up with approximately similar results. So, we have chosen to use **median imputation technique**

b. All the attributes are on different measures and it's difficult to explain the variability in the data. Hence, we need to generalize the features on to a similar scale. This process is known as standardization. There two types of standardization
   i. Normalization
   ii. Z Score

. We have chosen normalization process for this data preprocessing.

3. Modelling

3.1 Strategy

We are building two models based on regression and classification. In the first model, our task was to find out percentage of default. Since our target variable is a continuous, we have sought linear regression technique to predict. After prediction we use regularized linear regression and ensemble methods like boosting techniques, to check if accuracy can be improved than regression model.

In second model, we will classify whether the customer will default or not. Since our target variable is binary class (1 or 0), we have sought to use logistic regression and to enhance prediction accuracy we use ensemble methods like random forest.

In third model, we will calculate the loss by plugging results obtained from both the previous models in the following mathematical function,

Loss = PD * Loan Value * LGD

Gain = Loan Value * Interest Rate * (1 – PD) * Number of Years

PD – Probability of Default; **LGD** – Loss Given Default

From the above formula, if loss is greater than gain then it's a bad loan.

3.2 Technique

Linear Regression: We have used this technique to predict the target variable (Loss) which is continuous.

Regularized Linear Regression (Lasso/L1):

It is used to penalize and minimize the error rates of the predicted variable, which will improve the prediction accuracy.

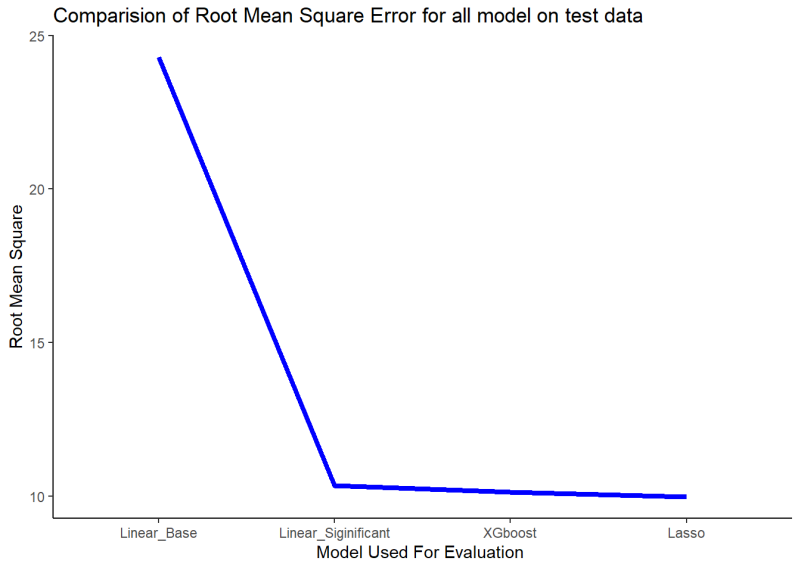
Ensemble Methods:

Random Forest/XGBoost: These are used for classification and regression tasks. In our model, we have used them to improve accuracy of logistic and linear regression models.

4. Model Performance

4.1 Model 1

- When we use Linear Regression, model overfits on the training data and does not perform well on test data
- We built another regression model considering only significant variables from the base model. When we use this model on test data, it was able to explain the similar percentage of variability.
- There are no penalty metrics added to the linear regression model. This we were able to achieve by using Lasso (L1) regularization, since it zeros out the coefficient of the variable. In lasso regularization, we obtain closer percentage of variability on both train and test data.
- We further tried to improve the performance by using ensemble method
- Using XGboost, we construct cross validation model to find the hyper tuning parameters but the model was overfitting.

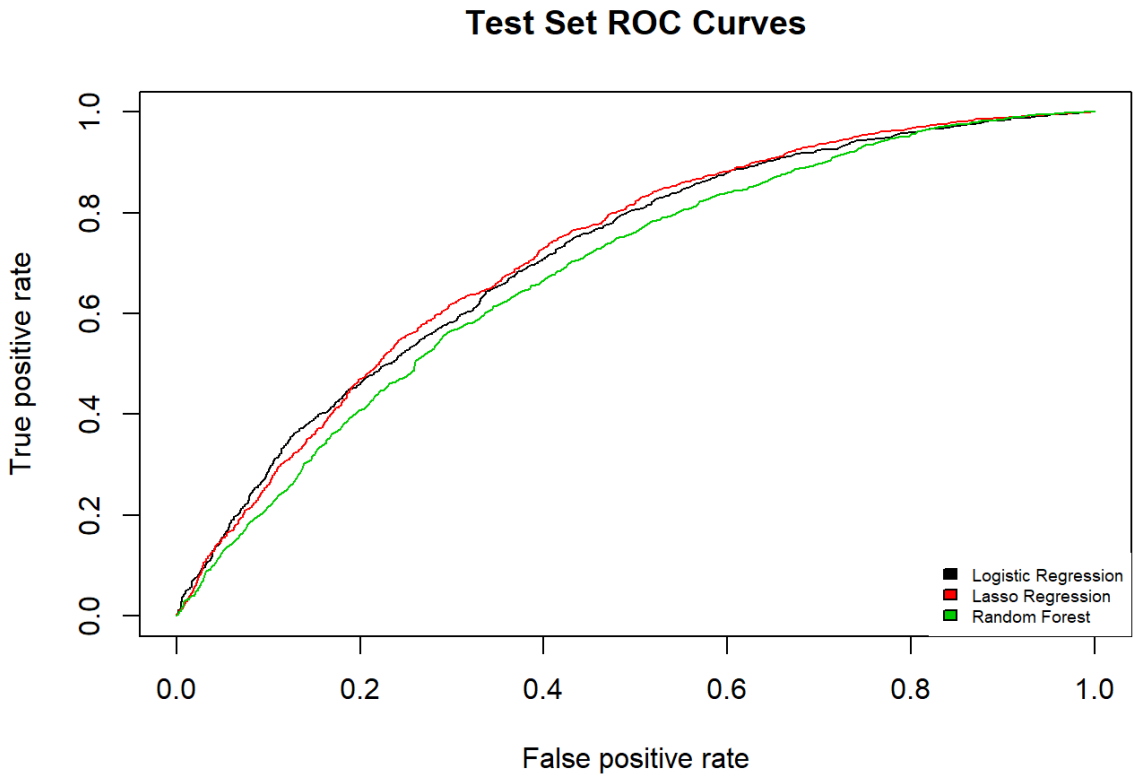


From the above graph, we can observe that Lasso model has less error rate than other models. Hence, we have chosen lasso regularization.

4.2 Model 2

- Initially we constructed Logistic regression model to predict if a customer is defaulter or not.
- In this model, we found that accuracy is high for the test data, but no information rate value implies that largest proportion of the observed classes here are non-defaulters. We computed hypothesis test to evaluate whether the overall accuracy rate is greater than the rate of the largest class.
- We found predictions of the model are biased towards non-default class as they are more prevalent across the dataset. Hence, we evenly balanced the data.
- We reconstruct logistic regression model for balanced data. Here we obtained lesser accuracy from model, but it was **not biased**.
- We rebuild Logistic regression using Lasso (L1) regularization. It only considers significant variables to train the model.

- In classification model, Area Under Curve (AUC) is the best metric to measure the performance of the model. So, lasso gives better AUC value than others.
- We build ensemble (random forest) model to improve the performance lasso regression model. However, there are less number of significant variables and random forest could not perform value.



Based on above plot we see lasso model is performing better than logistic or random forest model. Thus, we will be using lasso model to predict target variable for test scenario.

5. Insights and Results

Here, we are defining new parameters called Risk, Gain, and Delta.

- Risk is given by, Risk = Loan value * LGD * PD
- Gain is given by, Gain = Loan value * Interest Rate * (1-PD) * Number of years
- Delta is given by, Delta = Gain - Risk

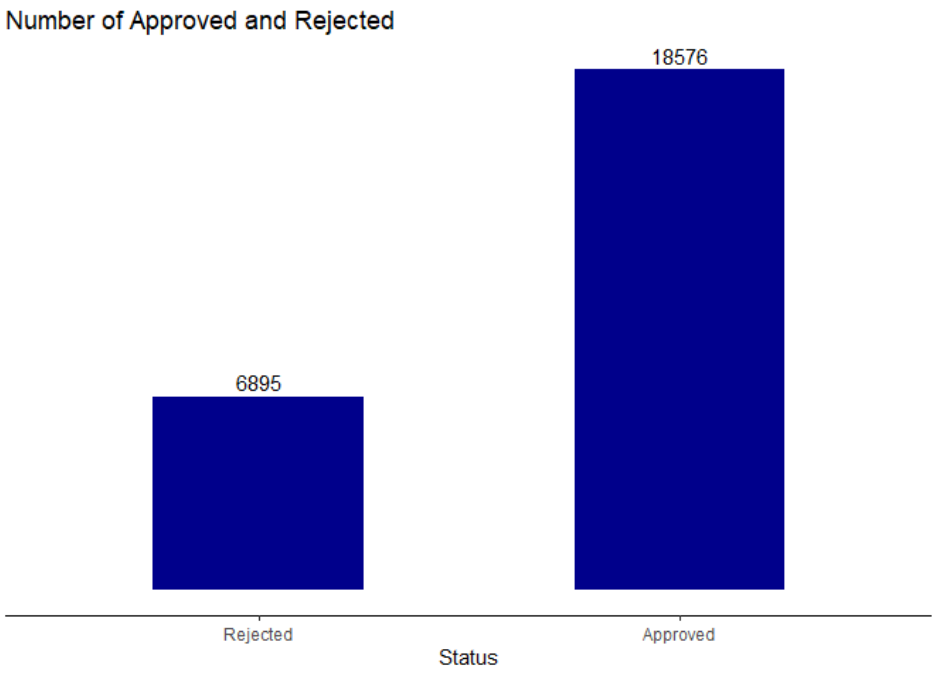
Loan value is defined as requested loan amount by each customer for 5 years.

LGD is *Loan Given Default*, it is percentage predicted by the model on test scenario.

PD is *Probability of Default*, where 1 (defaulter) & 0 (non-defaulter) are predicted by the model on test scenario.

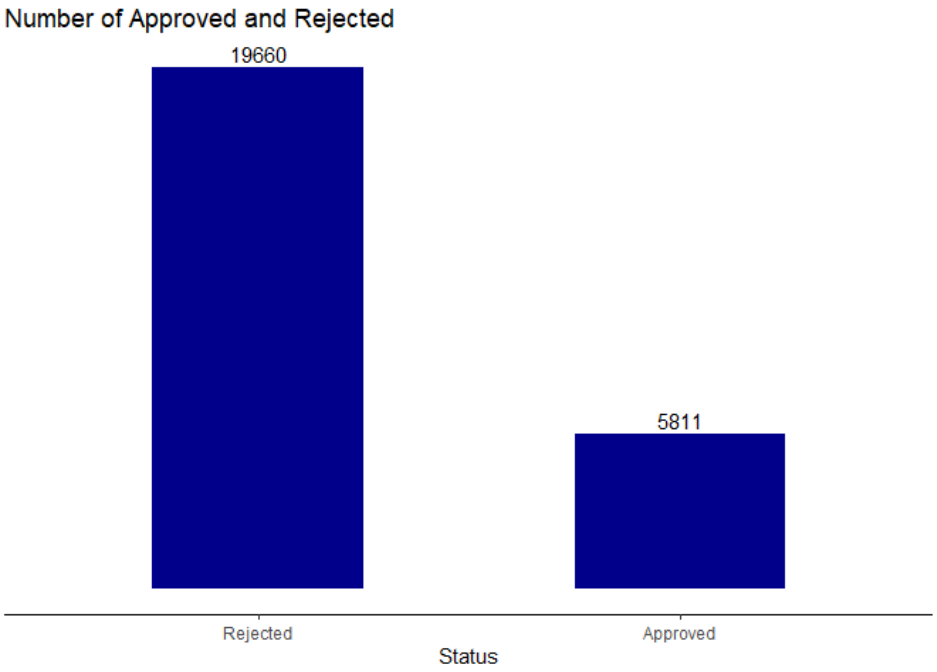
5.1 Scenario 1

Given, the total capital of 1.4 billion, we should consider the Risk = 0 which implies application is approved else rejected.



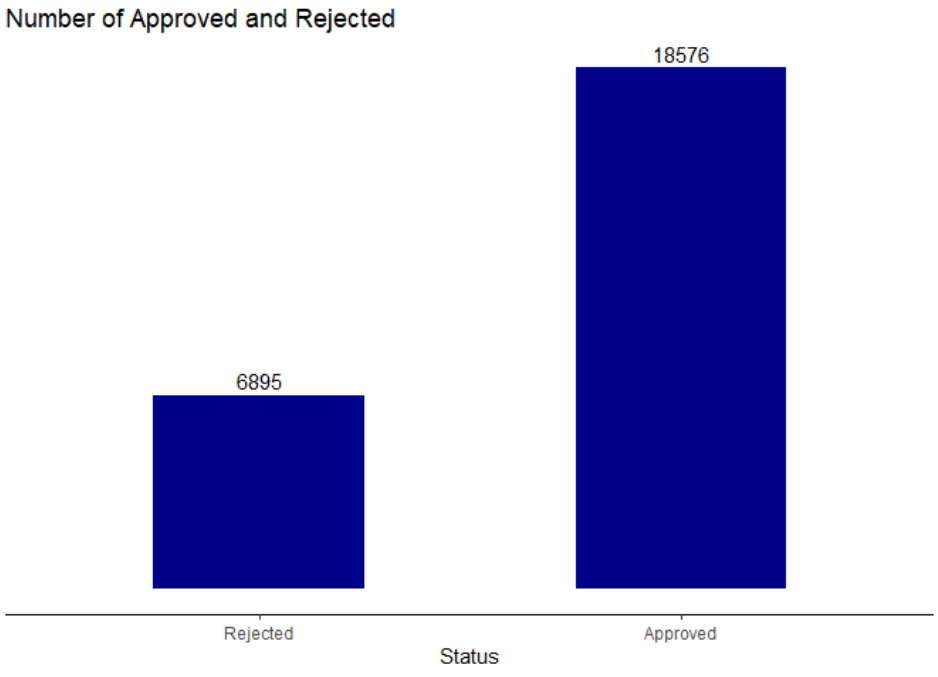
5.2 Scenario 2

Given the total capital of 450 million is allocated among customers based on ranking Delta > 0 which implies application is approved else rejected.



5.3 Scenario 3

Given the total capital of 1.4 billion, we should consider the Delta > 0 which implies application is approved else rejected.



6. Conclusion

From our 3-parted strategy, our underwriting team was able to successful predict both the probability of default and loss given default of each customer and leverage that information to classify for loan approval given the delta between gain and risk of each given customer. We achieved this successful model by utilizing concepts of imputation, Lasso penalizing, logistic and linear regression, and algebraic methods.

Given the continual addition of more customer loan data, this modeling process could be fine-tuned to serve as an accurate decision-maker at our bank. We were able to stay at-or-under budget while minimizing risk and maximizing gain for our bank.