Loan Default Risk Project

Ram Chandra Tiwari

The work consists of importing required libraries, data loading, data cleaning, data conversion, data visualization, checking correlation, model fitting for training data using Logistic Regression, Decision Tree, Random Forest , Support Vector Machine, Radial Support Vector Machine and chose the best model to predict outcome of test model.

**Executive Summary**

Interest of loan funded is an important source of income for commercial institutions. However, delinquency can be hazardous for them. As a Data Scientist, I am trying to provide a pre-screening guidelines for issuing loans by observing First Payment Default. The goal of this project is to predict the chance of being able to pay loan in Loan Due Date based on applicants’ financial status, location, residency status and past credit activities. This project will have effort to increase accuracy in determining applicants’ capability of loan repayment.

**Introduction**

The most common process of approving a personal loan is pre-screening the applicant’s credit history, followed by careful determining the amount of the loan. My goal of the project is to create a pre-screening model for a company to quickly screen out applicants who are likely to default in First Payment without misclassifying applicants who should qualify for loans.

The rest of the report is organized as follows. In Section 3, I describe the data and its source, and we present you with a summary of our findings. Section 4 demonstrates the visualization and Analysis of the relationship of variables and interpret the results into plain English. A conclusion and possible improvement are given out in Section 5.

**Data**

### Data for the study was provided by Keely Wolford from Zookeeper. Train data set and test data set with total 2,000 anonymous loan borrowers with their SetID is provided. For the test data, First Payment result is given but it is not given (NA) for test data. The list of features for each applicant is described in the table below. Variables are renamed to make the study easier.

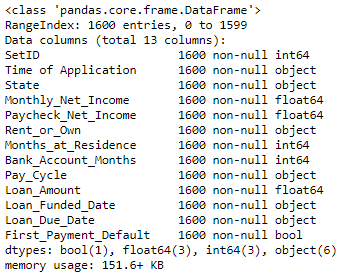


Table 1: Information of variables

From Table1, there are no non-null values in any of the columns. So, it is not necessary to play with missing values.

**Visualization and Analysis of Data**

Time of application is an object type variable so it has been converted into Month, Day and hour format to observe the effect of time on the Loan Payment. After conversion, Time of Application column can be deleted. The application time has been flagged for the deep study of its effect on the target variable. From the data preprocessing we did not find any duplicate IDs so it is the unique on given dataset. Moreover, there are no missing values so we don’t need to play with it.

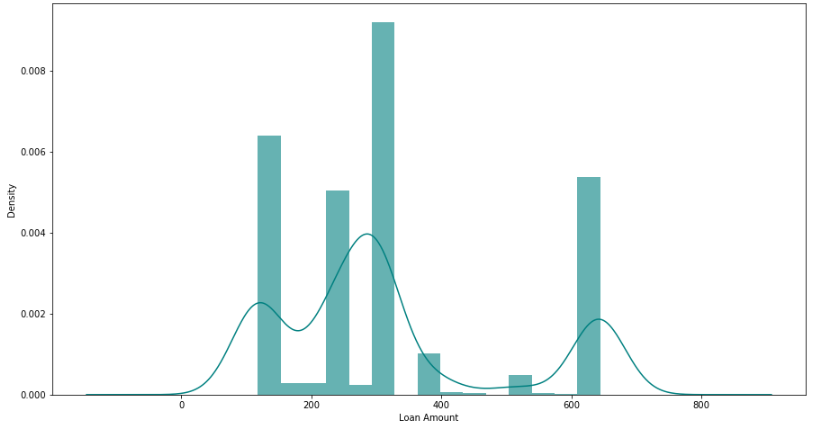


Fig.1: Density curve for Loan Amount

Density Curve for Loan Amount in Fig. 1 shows that the amount of loan of the customers is maximum for 150 to 350 and 600 to 650 but it is insignificant for 350 to 600.

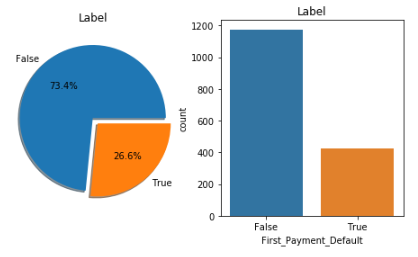


Fig.2: Dependent variable distribution

From the Fig. 2: of dependent variable distribution, it can be seen that the distribution of classes (False/True) is skewed so we need to do stratified sampling approach while training the model.

For the study of correlation between variables, we can see correlation plot as shown in Fig.3. From the figure, it can clearly be seen that Monthly Net Income and Paycheck Net Income are highly correlated.

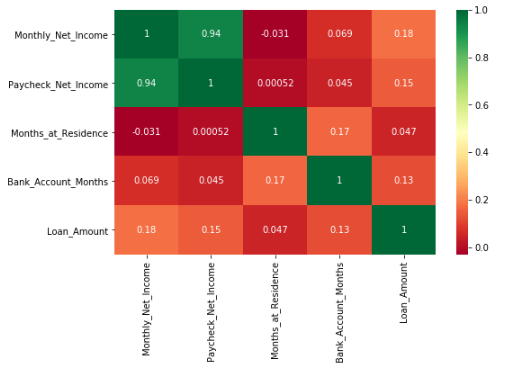


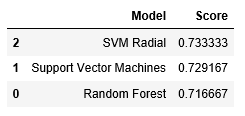
Fig.3: Correlation Plot

Column families has been created for different data types and normalized. After normalization of the data Label Encoding and One Hot Encoding is performed. Label encoding is the approach to encode categorical values. Label encoding simply converts each value in a column to a number. The disadvantage of label encoding is that numeric value can value misinterpreted by algorithms. For example, the value 0 is obviously less than 2 but does not really mean the same on the dataset. In order to mitigate this issue,we have done one hot encoding on the variables.

After finishing data processing, we performed Stratified Sampling (Train and Test Split), we have performed classifying modelings: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and SVM Radial.

**Analysis**

In this project, we would like to predict whether or not an applicant will pay loan on loan due date based on applicants’ financial status, location, residency status and past credit activities. For this analysis, logistic regression model, decision trees, random forests, support vector machine - linear, and support vector machine - radial models were used for modeling and prediction. The models were built on the train data, the accuracies for different models were obtained from the validation data (split test data), and the best model was used for making predictions in the test data. The accuracy of the different models on validation data are summarized on the following table:



The top three models based on the accuracy are tabulated on the above table. From the table, we can see that the support vector machine - radial (SVM-R) model yielded the efficiency of 73.33 %. So, SVM-R is the best model for making predictions for this project.

At first, logistic model was built on the train data, and the efficiency was checked on the validation test data. The area under the curve (AUC) for logistic model was 0.608 and average precision score was 0.316. Then, decision tree model was also built on the training data, and different parameters were calculated by running the model with validating data. The AUC for decision tree was 0.599.

|  |  |
| --- | --- |
|  |  |
| The AUC curve for Logistic Classification | The AUC curve for Decision Tree Classifier. |

**Conclusion**

In this assignment, data cleansing, data preprocessing, data visualization and data classification modeling was performed using Logistic Regression, Decision Tree, Random Forest, SVM and Radial SVM. Among them, SVM Radial model was found to have the highest accuracy. Values for the target variable of Test data was predicted and found 395 False and 5 True values.