The question that I was trying to answer was if the result of the loan application could be predicted given the set of inputs.

The question was important one to answer so that loan application system might could be automated as such the loan application with bad prediction from the model could right away be rejected with a specific reason or with minimal time frame. The system could also be used to enhance the decision making process of whether to issue loan for an individual given the circumstances.

Dataset of loan applications at a financial company during the year of 2015 was provided. Dataset consisted of 2 csv files of which one consisted of all the accepted loan application and the details of it and other contained rejected loan application.

The dataset was obtained at https://www.lendingclub.com/info/download-data.action. Lending club was so generous to provide the dataset of the loan applications.

To explore the dataset, first CSV files were opened using Microsoft Excel. The conclusion was that number of columns in both dataset is different. Each set of dataset had different naming conventions for each column. One of the first task that was done as part of the data exploration was to figure out the common column between both of the datasets and extract them. The common columns included loan amount, credit score, debt to income ratio, state, employment length, policy code and zip code. and Another problem with the dataset was the scale of credit score were different in both datasets. Apart from the employment history and debt to income ratio were in string format instead of numbered format. Zip code was provided in terms of first 3 digit of the code and last two digits were encoded as XX for the privacy purpose. Finally, common column named policy code was removed from the dataset as it could have introduced bias in the model. The reason that the column could have introduced bias in the model was that the policy code for all the accepted loan application was 1 whereas the policy code for all the rejected loan application was 0.

In order to transform the dataset in the proper format to feed into machine learning algorithm Apache Spark was used. Credit score were converted on same scale by determining the range belong to each interval of the credit score. String processing was performed to extract the proper debt to income ratio, first three digits of the zip code and employment history. One of column named the loanAccept was added to the dataset during the ETL process denoting the label 0 or 1 to the dataset, where 1 represents that the loan application was accepted. All the category columns were factorized before training a classifier on the dataset. As the dataset was reasonably small spark was run locally to perform the task of ETL.

I decided to use Decision tree to build model on the dataset. Main task to be performed on the dataset was classification into two classes namely loan accept or loan reject. Decision trees are very good at the task of classification with high accuracy and requires less computation to classify the future node. Hence, final decision was to proceed with the decision tree.

Error matrix for the decision tree was to determine the accuracy (% of the records classified correctly using the algorithm), precision and F1 Score.

The data were divided into 80 is to 20 using random selection in train and test. Then I decided to train model using all the available feature and evaluate the error. Initial accuracy that was obtained using all the feature was 90%. In order to simplify the model, I retrained the classifier using the individual features. The conclusion was the if just one feature named employment length was used to classify the loan records then it would be almost as accurate as if all the feature were used to classify the application. The evaluation matrix using the employment length as feature obtained as high as 89% accuracy on the test data.

Technology that I used to train classifier on the dataset was python and scikit-learn library. Apart from that Apache Spark (pySpark) was mainly used to perform ETL on the dataset.

Drawbacks of the Model

- Model contains high bias and hence will have high recall rate for the loan accept applications compared to the loan reject application. Such bias was introduced in the model due to the sheer difference in number of records for each type of the application namely accept and reject. There was lot more rejected application example in the dataset that was being used to train the model compared the loan accept application.

Future Improvements to the Model

- One improvement that could be brought to the model is eliminating the bias from the model. Another such work that can be done is to reduce the recall rate of the accept application depending on the requirement. Lastly, the ability of model to predict the rejected loan application could be enhanced as part of the future work depending on the need.
- Another improvement that could be bought is via using the cross-validation to train the model and select the best model for given feature set.
- Complexity of the model could be decreased by ranging the values of the feature dataset. For example, loan amount could be categorized at every 10,000 or 25000 so the decision tree would be simplified.
- Another way to bring noticeable improvement in the model would be to normalize the dataset and retrain the model.
- Lastly, Logistic regression of SVM could be used as another classifier to predict the outcome. All various models could be compared for the accuracy and best one could be selected.