SCMA 648 Business Data Analytics
Assignment 3 – Classification

The working directory was set according to the source file location. The saved rda file that was created for quarter 3 was loaded to the R environment along with the required packages.

Pre-processing steps:

- The random number generator seed has been used to get the same results each time.
- The rda file was loaded which contained the data frame called df.
- For the questions 2 and 3, two separate data frames have been created and some variables have been selected.
- The variables like annual_inc, dti, tot_cur_bal, and total_bc_limit had outliers. These outliers have been filtered from the dataset.
- The variables like home_ownership had some levels with less observations. These have been dropped.
- The variable emp_length was alpha-numeric. This has been converted to numeric values.

Splitting the data and Modeling:

- The data has been divided into training and testing dataset. 10% of the data has been used for training the model using createDataPartition.
- The missing values in certain features were imputed using preProcess function.
- There was a class imbalance for this dataset. The weights had been assigned to the classes of the target variables.
- The models Logistic Regression and Classification Tree have been built. Predictions were done using both the models. Confusion matrix has been constructed and misclassification rate was identified for all the models.
- The variables important for prediction were identified.

Question 2

The variables considered at the time at which a loan is awarded are loan_amnt, term, int_rate, home_ownership, annual_inc, fico_range_high, tot_cur_bal, dti, application_type, total_bc_limit.

These features were filtered from the dataset using select function. The summary for the dataset was determined and the outliers for the columns annual_inc, dti, total_cur_bal and total_bc_limit were removed using filter function. The dataset was then divided into training and testing dataset. Class imbalance was rectified using equivalent weights assigned to classes.

As mentioned above, the preprocessing for this dataset was done and model was built. The confusion matrix and misclassification rate are as follows:

- Confusion Matrix of Logistic Regression

	Charged Off	Fully Paid
Charged Off	7902	4429
Fully Paid	16467	28039

For this model, approximately 36000 observations were classified correctly. And the rest of them were misclassified.

- Misclassification rate for Logistic Regression
  0.3676478

From this model, the variables like term, int_rate, annual_inc and application_type are found to be statistically significant. This are important features for the prediction of the final status of the loan.

On the same dataset, a classification tree model has been built. The confusion matrix and misclassification rate are as follows:

- Confusion Matrix of Classification Tree

	Charged Off	Fully Paid
Charged Off	5511	6820
Fully Paid	11371	33135

For this model, approximately 38500 observations were classified correctly. And the rest of them were misclassified.

- Misclassification rate for Classification Tree
  0.3200556

Using variable importance, it was found that the variables int_rate, total_bc_limit, annual_inc are of importance for predicting the outcome of loan.

Question 3

Here, there were many features considered for the building the model including total_payment and last_payment_date. These variables are not generally available at the time at which a loan is awarded.

The model was built after required preprocessing. The confusion matrix and misclassification rate are as follows:

- Confusion Matrix of Logistic Regression

	Charged Off	Fully Paid
Charged Off	11617	787
Fully Paid	251	44733

For this model, approximately 56000 observations were classified correctly. And the rest of them were misclassified.

- Misclassification rate for Logistic Regression
0.01808741

On the same dataset, a classification tree model has been constructed. The confusion matrix and misclassification rate are as follows:

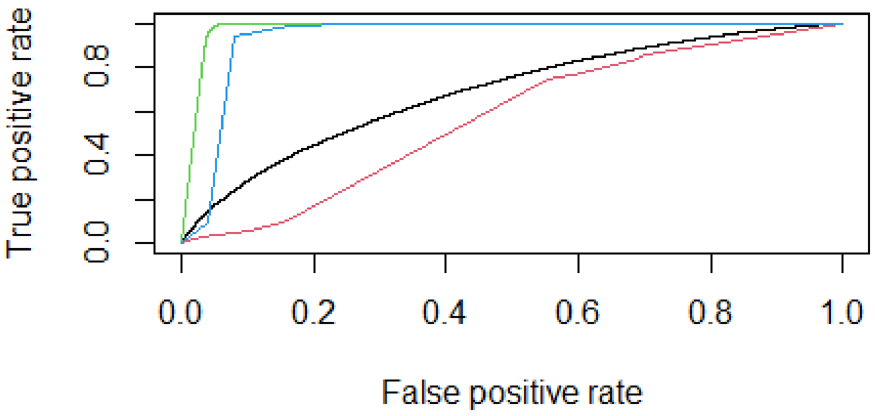
- Confusion Matrix of Classification Tree

	Charged Off	Fully Paid
Charged Off	11427	977
Fully Paid	2568	42416

For this model, approximately 53500 observations were classified correctly. And the rest of them were misclassified.

- Misclassification rate for Classification Tree
0.0617725

Question 4



The ROC curves were plotted for the models. If confusion matrix and misclassification rate are considered the third and fourth models perform the best. However, this is because of the attributes present in the second dataset. There is a data leakage. Total_payment and last_payment_date are variables which can not be considered for the prediction of loan outcome because it won't be known. Because of these attributes #3 models are able to perform better as compared to that of #2.

Question 5

From an investor's perspective, the first and second models are to be considered. The third and fourth model perform way better than first and second one because the attributes present in the dataset are not valid at the time of application. This is like revealing the answer to the model directly. However, in the first and second model, the confusion matrix and misclassification rate indicate that the models are fairly good considering the attributes given to them.