Name: Fahad Idrees

Roll No : MSCS15029

Data Mining

Submitted to : Faisal Kamiran

**CRISP-DM** on Loan Applications Approval Application (Data Understanding and Prediction)

**Business Understanding**:

Bank / Financial institute wish to have a decision engine of loan approval in which they can decide based on the record, whether it is beneficial for a bank to approve/ decline the loan of applicant(s) and predict the defaulters and normal customers based on the applicant’s profile and previous record. How much risk is involved in approving loan decision? what are some strategies in which if applicant is intended to pay the loan so bank should approve the applicant so to enhance the profitability of the bank and refrain the approval of that applicant(s) in which bank can face the financial loss due to fraudulent cases and in result can lessen the man effort and time of the bank resources.

**Data Understanding:**

In the problem we have online data available in Kaggle competitions website free includes both test and train set but I have just use train dataset. In Train dataset we have 105471 records (observations) and 770 features(attributes). This Data also have 2 columns i.e. IDs and loss, IDs are unique and represent one observation of the data but loss are varying from 0-100 which means this some percentage, loss is the percentage of loan that was not returned, 0 means 100% loan was reimbursed and 50 means 50% returned and 50% is not.The data set also have missing values in more than 30% columns in the form of NAs in various attributes.

**Data Preparation:**

For Data Pre-Processing I have dropped the columns with NAs attributes more than 1500, for this I have calculated the SUM of each column and took only those columns where NAs are more or equal to 1500 and found 86 columns above this threshold, for more refine result I than apply variance on each column and get 4 different columns where variance is equal to zero almost. So, after this left with 675 features column, After Data cleaning, I have applied the median to left NAs attributes based on their column value and replace NAs values with mean of that column.

**Modeling**:

Before modeling I remove outliers from the data and clean for processing, Outliers remove only numeric data attributes and update each attributes after each attributes. Afterward I use Naïve Bayes Classifier for 70% training set and 30% testing set, predict the loss on training set by using Naïve available library in R (e1071). For me this classifier fails miserable and don’t give me correct expected result. I tried all possibilities to run and evaluate the result but could not lucky though.

For modeling I use decision tree in which we cluster the data into two parts one is 75% which is use for training and 25% in use for testing, Unfortunately I could not get the expected result as required but I have learned many things varies from tools learning to play with big data.

**Conclusion:**

I could not get the required result and expected predictor engine which can help the bank to deal with applicant’s loan applications, From the beginning I have faced hardware issues with computer and shifted Core i3 to Core i5 and than data set not loaded on that machine, and took too much time but on MAC with Core i5 and RAM 16 GB I face some comfort, as loading the data.

I tried sampling the data using brute-force so that I can process the data efficiently but afterward I have remove that part as MAC is handled the processing of this data.

In the end I realized that can use many approach to work on this project, For Data processing we can use median, neighborhood classifiers, average values or entirely remove the attributes to remove the missing values, NAs attributes. This was a tough problem and probably time was not good manage to try couple of different algorithms. Even for Data preparation I use first mean values and afterward zero variance helped better to clean the data.

For best 5 correlations, I have used the correlation of loss column with each of 770 columns and than ordered the column and fetch the values from original csv of that respective columns.

To conclude the hardest part and time consuming part was to deal with large quantity of data and remove the unused or fault values with some appropriate values, I have spent almost 50 hours on this project in 1 week but unfortunately could get additional time to try different approaches to get the better result. But in future this will help me to understand the problem with data and time management.

**Thank you so much for this project, Learn a lot !!**