**Predictive Modelling for Customer’s Credit worthiness**

**1. Data Exploration Insights:**

The overall data set is divided into 3 segments where the account and enquiry segment are based on the customer id with each weight. The account segment has the features based on each customer's multiple transaction histories. The enquiry segment has the features based on the enquiry made by the customer with each date and amount. The overall data set has the target variable called 'Bad\_label' and the variable indicates the customer creditworthiness where 0 value represents the positive creditworthiness and 1 represents the negative creditworthiness. The data is split into two data sets by 70% and 30% for train and test data respectively. The data is in the form of categorical, numerical and Alpha numerical variables.

**Enquiry segment:** The enquiry segment has the total number of 6 features where it has the key index of customer id. Each customer id has multiple enquiries based on the date, amount and enquiry purpose. Enquiry purpose is a categorical variable which has the numerical values represent the enquiry purpose for each amount.

**Account segment:** The account segment is based on transaction history which has 21 features. The payment history is the main key for the rest of the feature where each customer has multiple payment history with payment date, credit limit, cash limit, current balance…etc. The payment history contains the values of Days Past Due for 18 months.

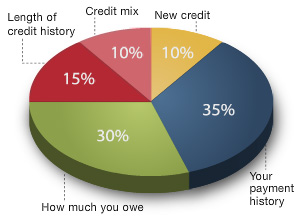
**Data Segment:** The data segment has the large number features based on customer bank account. The target variable is 'Bad\_label' and it is based on customer id as key index. The features contain the customer account type, credit, age, location…etc. The total number of features in the data segment is 83. The overall data has plenty of missing values and NaN values.

The target variable is Customer creditworthiness either with positive or negative flag. The Credit score is denoted by the Customer creditworthiness variable which has the general weights for customer's features such as Length of the credit history, credit amount, payment history, credit mix, new credit...etc. Each feature has their own weight which has more gain in target variable. The suggested features such as total\_diff\_lastpaymt\_opened\_dt, utilisation\_trend, count\_enquiry\_recency\_365, Ratio\_currbalance\_creditlimit are having more weight in credit score which has higher gain value comparing to the other features.

* The overall customers falling under the negative bad label is 4.29% which means 95.71% of the customers having the positive creditworthiness.
* From the correlation heat map, the feature 35 and 69 has the correlation value of 1 which means that the features are same and repeated features. The feature 34 and 68 have the same repeated feature which has the correlation value of 1.
* The feature 29, 66 and 44 has second highest correlation between them where these two features are representing the area postcode of the customer address or bank address.
* The feature 3 and 7 has the third highest correlation between them.
* The actual payment amount and high credit amount has the highest correlation of 0.769 in account segment
* The cash limit and credit limit has the correlation value of 0.758
* The current balance amount and high credit amount has the next highest correlation value of 0.755

**2. Feature matrix:**

The 3 segment has the data which is based on the customer id as the key index to connect between them. The features need to be created according to the important factor for credit score such as payment date, credit limit, cash limit, current balance…etc. These features needed to be created with customer id index and it should be blended or added to the data segment which has the all the features include target variable. The new features should reflect the payment history, enquires, credit…etc. The suggested features were the features which have the more weights for the bad label.



The above pie chart shows the credit score weight for each future. The features need to be created from the account and enquiry segment should be focused on this weight which helps the overall model to achieve more accuracy with higher gain. The payment history has the more weight in the credit score were the features belongs to payment history will be having the higher gain comparing to the rest of the features in other parts. The list of features were selected on the basis of feature importance method using RF classifier.

|  |  |  |
| --- | --- | --- |
| **S.No** | **Feature** | **Gain** |
| **1** | **mean\_diff\_lastpaymt\_opened\_dt** | **0.0273534** |
| **2** | **total\_diff\_lastpaymt\_opened\_dt** | **0.0267480** |
| **3** | **count\_enquiry\_recency\_365** | **0.0233670** |
| **4** | **mean\_diff\_open\_enquiry\_dt** | **0.0231488** |
| **5** | **payment\_history\_mean\_length** | **0.0228567** |
| **6** | **avg\_enq\_amount** | **0.0172696** |
| **7** | **Ratio\_currbalance\_creditlimit** | **0.0146028** |
| **8** | **utilisation\_trend** | **0.0104339** |
| **9** | **max\_freq\_enquiry** | **0.0096361** |

**3.** **Model evaluation**

**Machine Learning Model:** The algorithm which is used in this project is Random Forest and it will work better, if the data set is categorical and labelled data. The final data of this project is well labelled and categorized by data cleaning methods. The train and test set are previously divided by the provider itself. The test and train data is joined to create the categorical dummies using the one hot encoding method. If one hot encoding is used in the split data, there will be a lot of dummies change in the train and test data features due to the lack of categorical variable present in the test data comparing to train data. The train and test data fitted into the model and the target variable were predicted. The default method of random forest classifier is based on Gini index.

**Model Accuracy = 95.5%**

|  |  |  |
| --- | --- | --- |
| **Rank** | **Feature** | **Feature Gain** |
| **1** | **feature\_7** | **0.0375878** |
| **2** | **mean\_diff\_lastpaymt\_opened\_dt** | **0.0273534** |
| **3** | **total\_diff\_lastpaymt\_opened\_dt** | **0.0267480** |
| **4** | **count\_enquiry\_recency\_365** | **0.0233670** |
| **5** | **mean\_diff\_open\_enquiry\_dt** | **0.0231488** |
| **6** | **payment\_history\_mean\_length** | **0.0228567** |
| **7** | **feature\_3** | **0.0221267** |
| **8** | **feature\_21** | **0.0216086** |
| **9** | **feature\_52** | **0.0214672** |
| **10** | **count\_enquiry\_recency\_90** | **0.0192718** |