

Data mining (Itm6285)

FINAL PROJECT



March 16, 2017

ANANYA AUDDY

NET ID: PY7552

Introduction: Almost every individual has a credit card and a correct usage of it does really help in building a good credit score. How important is Credit score? Ask anyone – specially here in US and all will know it. However, one default payment can cause the credit score to plummet. Not only credit score, it also has an adverse effect on the credit limit and future loans of any kind.

In this project, we would be dealing with a case of customer default payments in Taiwan. The dataset has 24 features and a class label and there are 30000 instances. We would be creating a predictive model to let the bank predict that whether their customer would be a defaulter for the next payment or not.

Objective: To build classifiers and use them to predict whether a credit card customer will be a defaulter in his next payment.

Dataset: **Data:** Bank Dataset from UCI Repository [https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients#](https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients)

Before we start our analysis, we need to get familiarized with the data set. The dataset is an imbalanced one i.e. majority of the class labels are non-defaulters. Out of 30000 instances, 78% are non-defaulters and remaining 22% are defaulters.

Following are the features present in the dataset. ID, Credit balance, Gender, Education, marital status and age are self-explanatory. Pay\_0, Pay\_2, Pay\_3, Pay\_4, Pay\_5, Pay\_6 are repayment status of months April till September respectively. Repayment status is defined as the delay in payment. Example: if the value of Pay\_0 is -1 then it means that the customer has duly paid, if the value is 2 then it means that the payment is overdue for two months. Bill\_Amt1 to Bill\_Amt6 are the credit card bill amounts for the month of April till September. Pay\_Amt1 to Pay\_Amt6 are the amount that the customer has paid against the credit card bill from the month of April till September.

N.B. The data belongs to year 2005.

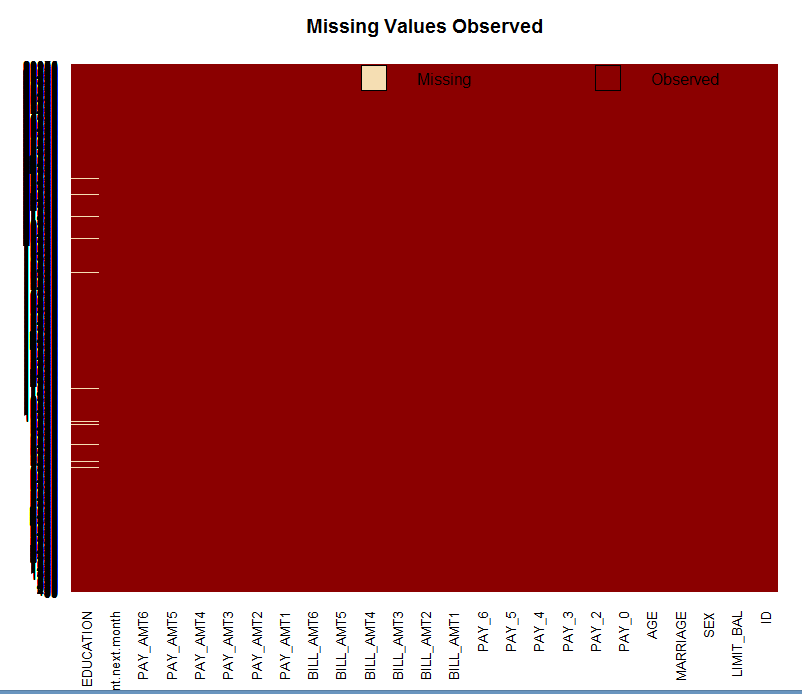
The features are listed down in a tabular form for a better understanding.

|  |  |
| --- | --- |
| **Attribute** | **Explanation** |
| ID | Unique Id of each record |
| Credit Balance | Credit amount on the given credit card |
| Gender | Gender of the customer i.e. male or female |
| Education | Education level of the customer i.e. high school, graduate, university, others |
| Marital Status | Marital status i.e. married, single, others |
| Age | Age of the customer |
| Pay\_0 | Repayment status in September |
| Pay\_2 | Repayment status in August |
| Pay\_3 | Repayment status in July |
| Pay\_4 | Repayment status in June |
| Pay\_5 | Repayment status in May |
| Pay\_6 | Repayment status in April |
| Bill\_Amt1 | Bill Amount in September |
| Bill\_Amt2 | Bill Amount in August |
| Bill\_Amt3 | Bill Amount in July |
| Bill\_Amt4 | Bill Amount in June |
| Bill\_Amt5 | Bill Amount in May |
| Bill\_Amt6 | Bill Amount in April |
| Pay\_Amt1 | Amount paid in September |
| Pay\_Amt2 | Amount paid in August |
| Pay\_Amt3 | Amount paid in July |
| Pay\_Amt4 | Amount paid in June |
| Pay\_Amt5 | Amount paid in May |
| Pay\_Amt6 | Amount paid in April |
| Default\_payment\_next\_month | Either 0 or 1 - 0 means for the next payment the customer is not a defaulter and 1 means the customer is a defaulter |

Data Cleaning and Transformation

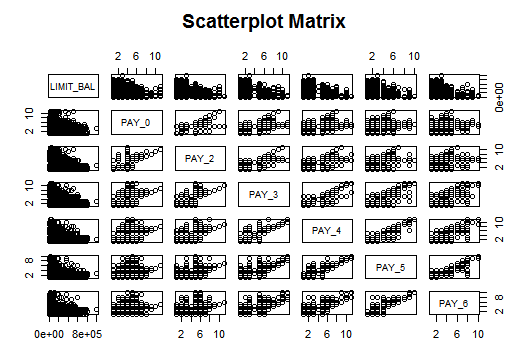
All the attributes of the dataset are integer, so we converted - gender, education, marital status, age and repayment status into factor. This would help in classifying the customers. We converted the class label i.e. **Default\_payment\_next\_month** to factor as well. The bill amount (Bill\_Amt) and the payment amount (Pay\_Amt) remains integer.

We delete the ID column as it is of no use in our analysis, it is just a sequential number identifying each instance uniquely. We also delete those rows which had missing values. The data had minimal missing values as shown in the following plot, thus removing such few rows will not vary the result.



Data Analysis

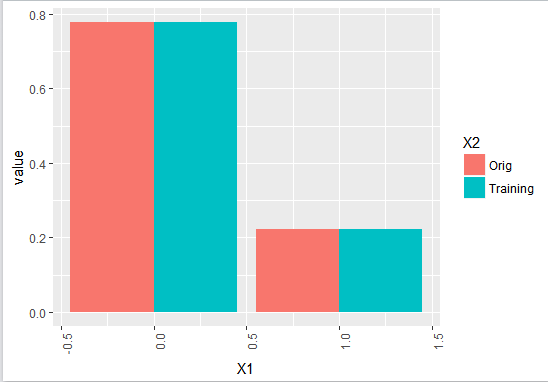
Below scatter plot shows there is a strong negative correlation between credit balance and repayment status (from April till September). Repayment status represents the delay in payment as mentioned before, thus as it increases the credit balance starts dropping.

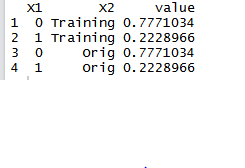


Next, we split the dataset into training dataset and test dataset. To build our model, we use stratified sampling method and use 10-fold cross validation to train and evaluate our model.

We build our training dataset with 45% of the data to process our classifier faster.

The proportion of classes between original and training dataset are depicted in the plot:





Building Models

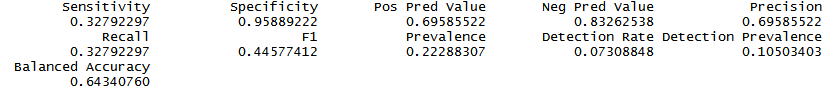
We would be using 4 algorithms C5.0, C5.0Cost, rpart and random forest and do a comparative study to determine the best predictive model.

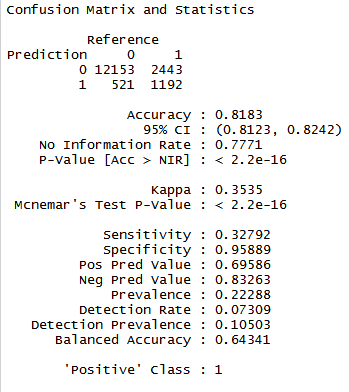
## Decision Tree:

We would be using the decision tree algorithm(C5.0) to classify the customers.

Following is the accuracy plot of decision tree algorithm, which depicts a higher accuracy of rules without winnowing, whereas in the adjacent plot, the accuracy is higher for the tree with winnowing. 

The following shows all the metrics against which the algorithm would be judged. The model seem to have almost 70% precision i.e.if it predicts 100 customers, 70 of them would be correct. Though the accuracy is 81%, but we cannot solely rely on it because we are dealing with an imbalanced dataset. Thus precision and recall should be the driving factor.





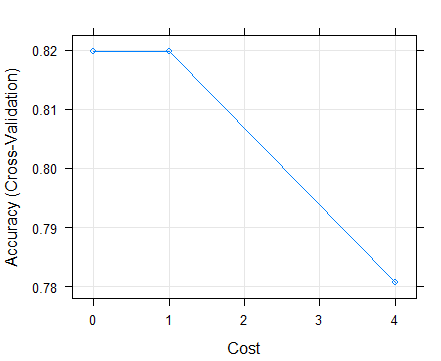
Adding Cost to Decision Tree:

In our earlier classifier, the cost of predicting a false positive and false negative is same. However, that shouldn’t be the case. The cost of predicting a credit card default customer as a non-defaulter would be more. There we have used another classifier C5.0Cost and added a cost matrix. We have added the following cost matrix:

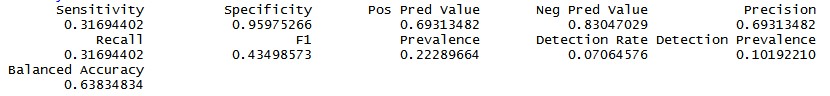
COST MATRIX:

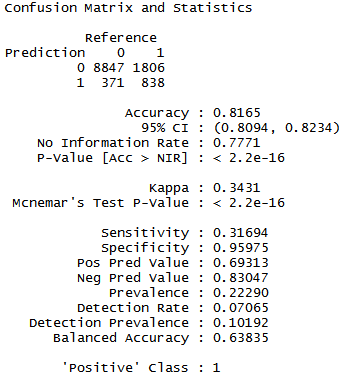
|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | NA | 1 |
| 1 | 4 | NA |

The accuracy of the model decreases as the cost increases.



Below are the metrics based on which the algorithm would be judged.

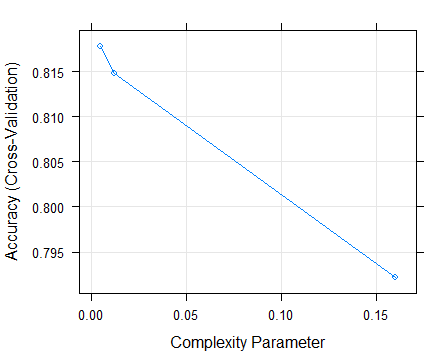




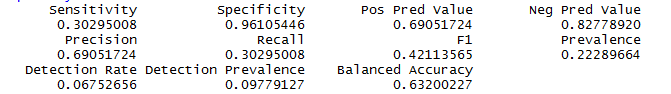
Decision Tree using Recursive Partitioning and Regression Trees:

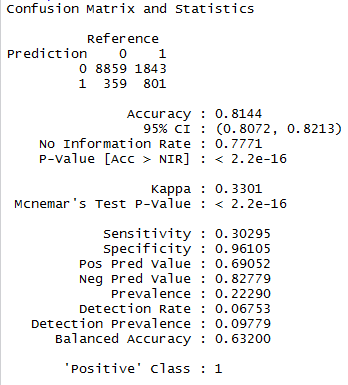
We would be using the decision tree algorithm(rpart) to classify the customers.

The below plot shows how the accuracy of the model decreases as the complexity increases.



Below are the metrics based on which the algorithm would be judged.

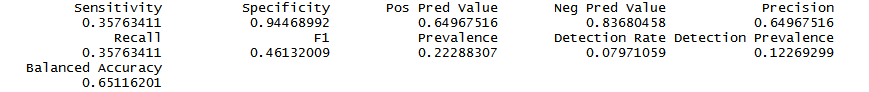


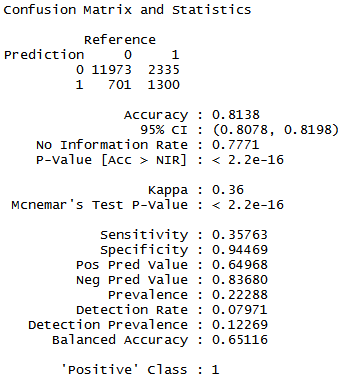


## Random Forest:

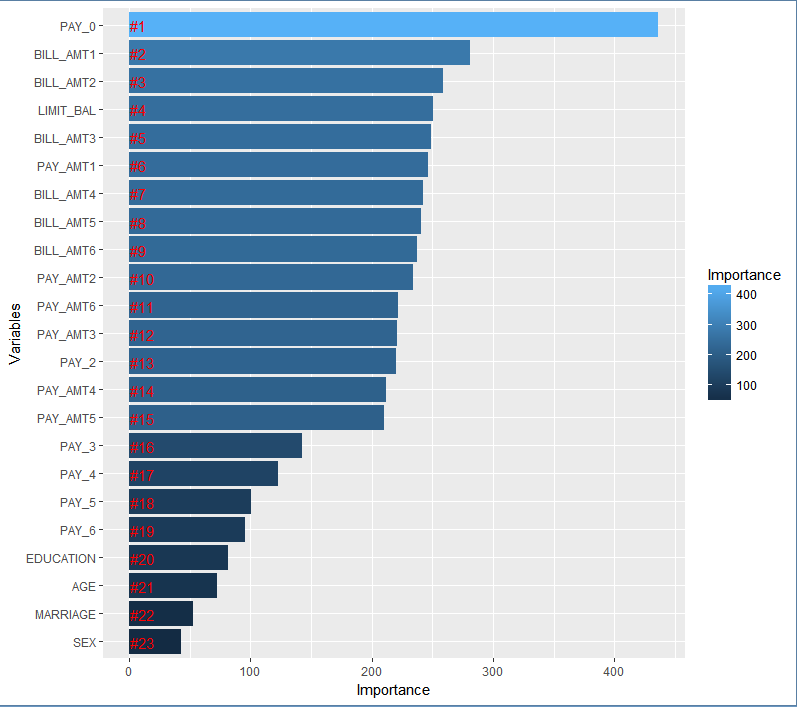
Confusion Matrix of Random Forest Algorithm

The following are the metrics of Random Forest classifier.

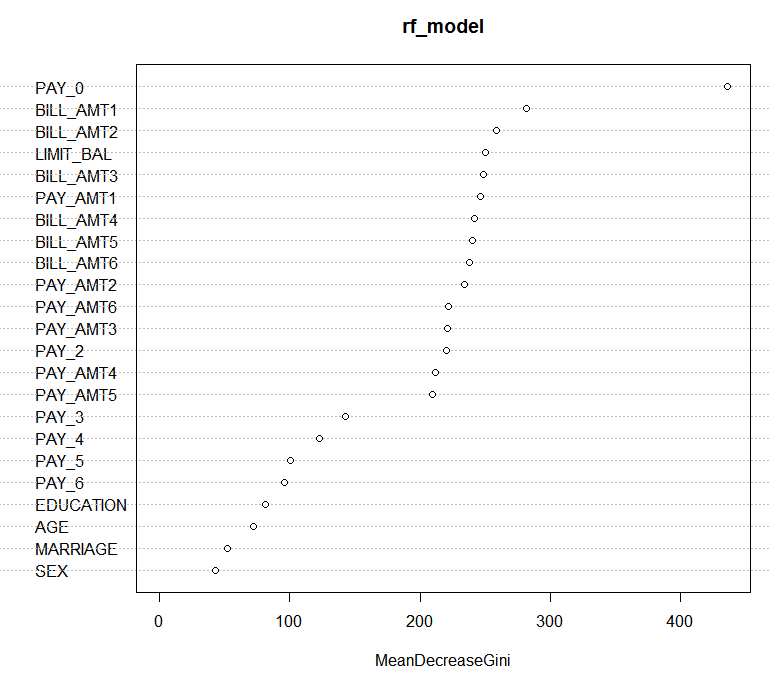




Ranking of all the atrributes:



Mean Decrease Gini: This coefficient helps in measuring how each variable contributes to the homogeneity of the nodes and leaves in the resulting random forest. In simple terms, it ranks the importance of the variables from top to bottom. Least important are the demographics data i.e. education, age, marital status and sex.



## Comparative study of all four Classifiers

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **C5.0** | **C5.0Cost** | **Rpart** | **Random Forest** |
| **Accuracy** | 0.8183 | 0.8165 | 0.8144 | 0.8138 |
| **Sensitivity** | 0.32792 | 0.31694 | 0.30295 | 0.35763 |
| **Specificity** | 0.95889 | 0.95975 | 0.96105 | 0.94469 |
| **Precision** | 0.69586 | 0.69313 | 0.69052 | 0.64968 |

As per this business problem, a classifier would be considered better if it has a higher recall value because the bank would be more interested in knowing the actual positives i.e. number of customers who would likely be a defaulter for the next month.

The F1 score (also F-score or F-measure) is a measure of a test's accuracy. It considers both the [precision](https://en.wikipedia.org/wiki/Precision_(information_retrieval)) p and the [recall](https://en.wikipedia.org/wiki/Recall_(information_retrieval)) r of the test to compute the score. There are two other measures namely F2, which weighs recall higher than precision (by placing more emphasis on false negatives), and F0.5, which weighs recall lower than precision (by attenuating the influence of false negatives).

Thus, the beta value would be 2.

Weighted F- Measure

|  |  |
| --- | --- |
| **Models** | **Weighted F-measure** |
| C5.0 | 0.366698825 |
| C5.0Cost | 0.355532394 |
| Rpart | 0.341257727 |
| Random Forest | 0.392959322 |

Recommendation: As per the weighted F-measure, Random Forest is the best classifier. This would help the bank to identify its potential risk customers, monitor proactively and take actions as required.