Assignment-1
Machine Learning

Submitted by: (Group 9)
Amrit Kochar(2014B4A70821P)
Siddharth Nagpal(2014B3A70743P)

1. Problem Statement:

We aim to classify people, by applying classification techniques on given set of attributes, as good or bad credit risks. In the world of finance, this issue is faced by each and every institution. We want to build a model which accurately predicts the probability of default payments by credit card customers. This project considers the case of customer default payments in Taiwan.

2. Data Description:

A binary variable, default payment (Yes = 1, No = 0) is the only response variable.

Here is the description of attributes in the selected dataset. ***(23 attributes)***

X1: Amount of given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
X2: Gender (1 = male; 2 = female).
X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
X4: Marital status (1 = married; 2 = single; 3 = others).
X5: Age (year).
X6 - X11: History of past payments. The monthly payment records (from April to September, 2005) have been tracked as follows:
X6 = the repayment status in September, 2005;
X7 = the repayment status in August, 2005; . . .; X11 = the repayment status in April, 2005.
The measurement scale for the repayment status is: -2 = No Consumption ; -1 = pay duly; 0 = Use of revolving credit; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.

X12-X17: Amount of bill statement (NT dollar).
X12 = amount of bill statement in September, 2005;
X13 = amount of bill statement in August, 2005; . . .;
X17 = amount of bill statement in April, 2005.

X18-X23: Amount of previous payment (NT dollar).
(X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .
X23 = amount paid in April, 2005.)

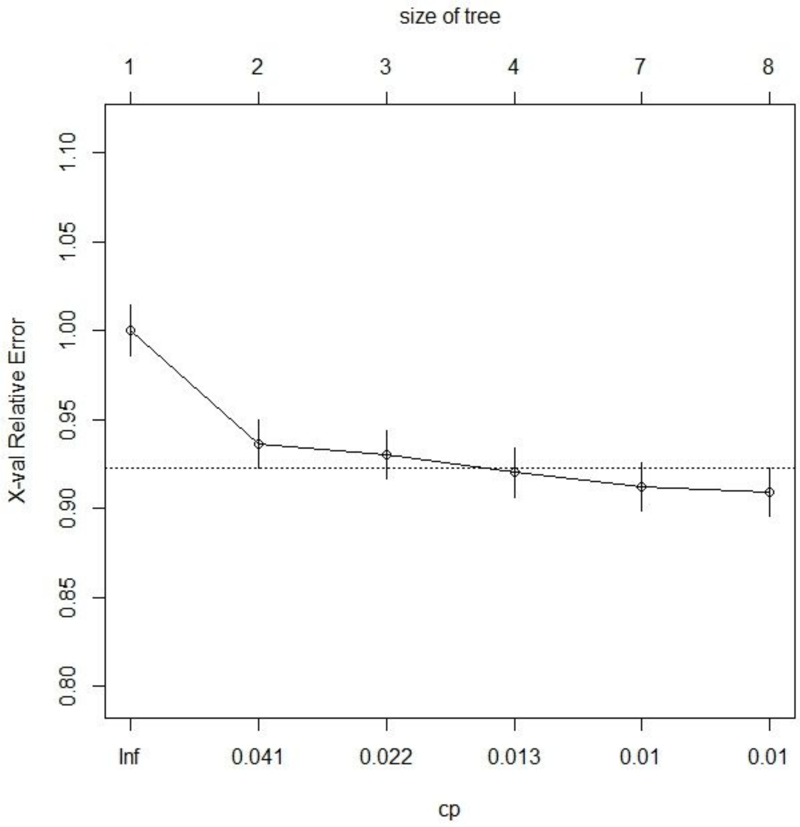
3. Methodology

Here are the steps we followed for this problem:

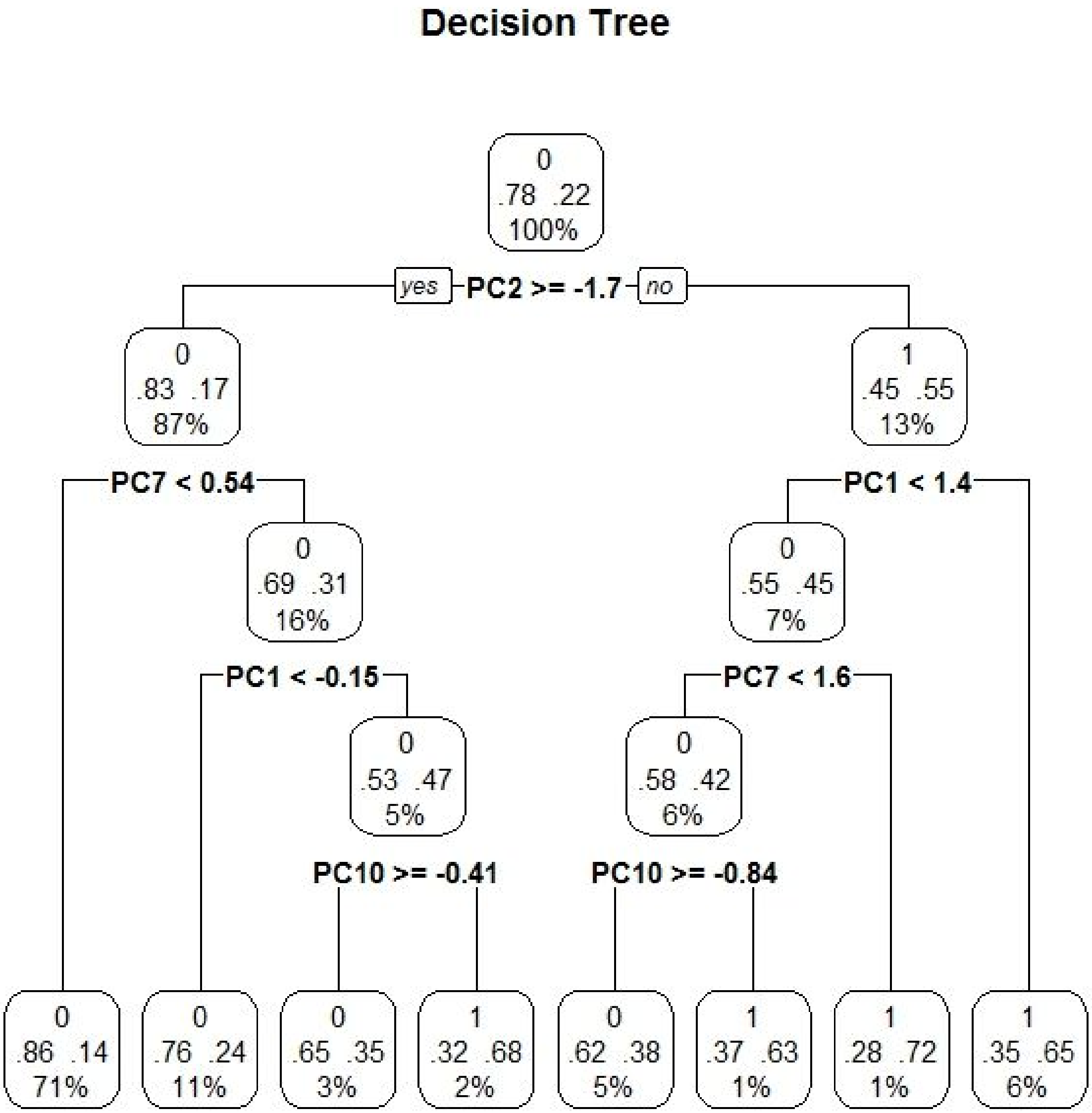
1. We use three libraries to for ease in preprocessing, to create the tree and plotting the tree, **caret and rpart and rpart.plot**.
2. First we save the response variable column separately before preprocessing the data columns. We remove this column from the dataset temporarily.
3. We **compute the difference in bill amount and amount paid for each of the six months** and store it as individual columns. We delete the individual columns of bill amount and paid amount from the dataset. Total deleted columns is 12 and new columns is 6. This is done because the net difference holds significance of the net position of each customer in each month. We don't need individual values.
4. We **normalize** the data points.
5. We apply **PCA**.
6. We append the response variable column back to the dataset.
7. We compute training and testing datasets. 60% of the data available is used for training the model.
8. **Modelling:** Create the Decision Tree over training dataset.
9. **Prune the tree** and plot the Decision Tree.
10. **Compute the performance** over testing dataset.

4. Results:

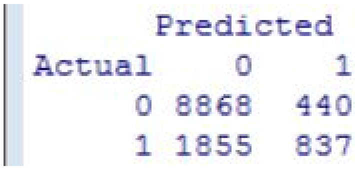
Training of the model.



This is the decision tree created to model the problem.



The final predictive accuracy we got is **0.80875**.



```
         Predicted
Actual      0    1
     0 8868  440
     1 1855  837
```

5. Applications and Scope:

- The credit to a customer by a business or a bank may be increased or the interest rate may be reduced in the case of improving creditworthiness of the individual or a business, while the opposite may occur if there is a decline in the customer's credit profile. Therefore, this a major application for any customer-credit business.
- In a well developed and stable financial system, risk prediction plays an important role. The major purpose of risk prediction is to use financial information, such as business financial statement, customer transaction and repayment records, etc., to predict business performance or individual customer's credit risk and to reduce the damage and uncertainty.

6. References:

1. https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients
2. https://www.researchgate.net/publication/311714926_Analyzing_Default_Payments_of_Credit_Card_Clients_in_Taiwan
3. http://ieeexplore.ieee.org/document/5713481/?reload=true
4. http://mitsloan.mit.edu/media/Lo_ConsumerCreditRiskModels.pdf