# **Yodlee** **Loan** **prediction** **challange** **2015**

## **Objective**

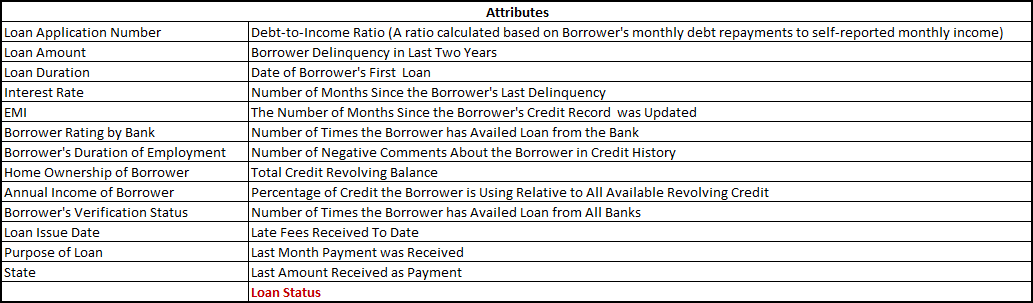
The challange is to model loan applications and predict if they will go bad or not.

## **Problem Statement**

In the datasets, the first row contains the labels of the attributes. In the training data set, there is an additional final column (named Loan Status) which has the information on whether the loans will turn bad (0) or not (1).

Our task here is to split the given training dataset into train and test (i.e., hold-out sample) and build a model on the train dataset and test the model predicted results against the Loan Status column in the test data set.

Labels of attributes is shown below.



## Setting up the working directory in R

## Load required R packages for this exercise

setwd("D:\\Kaggle\\Yodlee")  
  
library(tree)

## Warning: package 'tree' was built under R version 3.2.5

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(adabag)

## Warning: package 'adabag' was built under R version 3.2.5

## Loading required package: rpart

## Loading required package: mlbench

## Loading required package: caret

## Warning: package 'caret' was built under R version 3.2.5

## Loading required package: lattice

## Loading required package: ggplot2

library(glm2)

## Warning: package 'glm2' was built under R version 3.3.0

### Read and store the given training data set in R

Since we have 'percentage symbols' and 'months' attached with the values in EMI column and Loan Duration column respectively, it is desirable to remove them and convert the character type varibles into factor variables. '%>%' is dplyr's operator.

# Feature Engineering

* Number of loan applications per State
* Number of loan applications per Rating

# Missing values

Missing values are found and replaced by mean values of the respective columns.

* No\_of\_Loans (Number of times the borrower has availed the loan from the bank)
* Credit\_Balance (Total Credit Revolving balance)
* Percentage\_Used\_Credit (Percentage of Credit the Borrower is Using Relative to All Available Revolving Credit)
* No\_of\_Loans\_Other\_Banks (Number of Times the Borrower has Availed Loan from All Banks)

### Exclude the variables like Loan\_No (Loan Application Number) and other columns with date values

# Cross validation samples - Split the data set into two portions (70% - Train; 30% - Test)

# Logistic Regression Model

First two GLM models were studied and based on the results GLM3 (glm.fit3) is finalised.

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##   
## Call:  
## glm(formula = Loan\_Status ~ EMI + Late\_Fees + Last\_Amount + loans\_by\_state,   
## family = binomial, data = loan\_train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -6.2791 0.0014 0.2899 0.6666 3.5679   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.642e+00 5.201e-02 31.572 < 2e-16 \*\*\*  
## EMI -2.324e-04 1.197e-05 -19.414 < 2e-16 \*\*\*  
## Late\_Fees -3.370e-03 2.507e-04 -13.442 < 2e-16 \*\*\*  
## Last\_Amount 1.561e-04 7.755e-06 20.126 < 2e-16 \*\*\*  
## loans\_by\_state -5.245e-05 1.770e-05 -2.964 0.00303 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 13826 on 16799 degrees of freedom  
## Residual deviance: 11110 on 16795 degrees of freedom  
## AIC: 11120  
##   
## Number of Fisher Scoring iterations: 9

## GLM model was fit for test observations

# Confusion matrix and Model Accuracy

## loan\_Status\_test  
## glm.pred 0 1  
## 0 62 72  
## 1 1029 6037

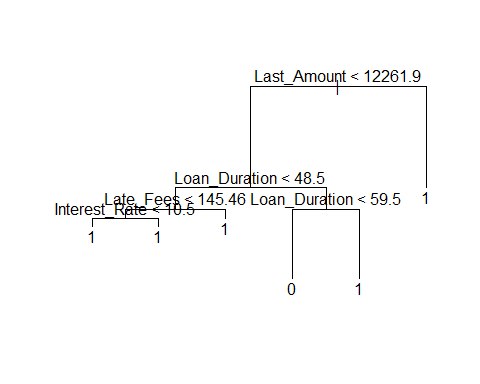
## [1] 0.8470833

# Decision Tree Model

## Decision tree is built using tree package in R

##   
## Classification tree:  
## tree(formula = Loan\_Status ~ ., data = loan\_train)  
## Variables actually used in tree construction:  
## [1] "Last\_Amount" "Loan\_Duration" "Late\_Fees" "Interest\_Rate"  
## Number of terminal nodes: 6   
## Residual mean deviance: 0.5655 = 9491 / 16780   
## Misclassification error rate: 0.1059 = 1778 / 16788

Decision Tree:



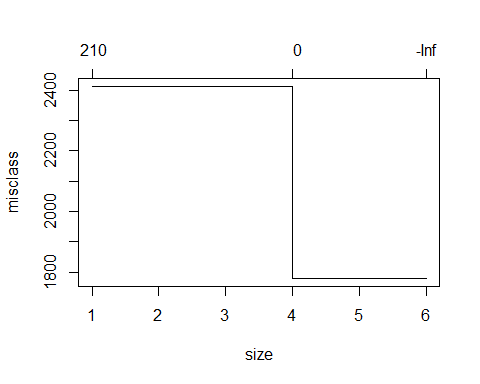
## Decision tree model is fit fot test observations and computed Confusion matrix.

## loan\_Status\_test  
## tree.pred 0 1  
## 0 335 68  
## 1 756 6041

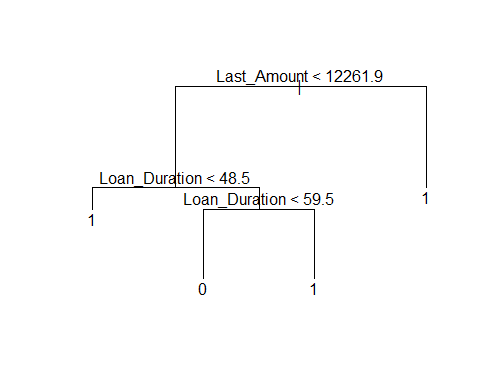
## [1] 0.8855556

# Tree Pruning using Cross Validation method

### Prune the tree to select number of trees to end up with.



## Application of Repruned tree



## loan\_Status\_test  
## tree.pred 0 1  
## 0 335 68  
## 1 756 6041

## [1] 0.8855556

# ADABOOST.M1 Algorithm

## Loan\_Status ~ pesos + Loan\_Duration + Interest\_Rate + EMI + Employment +   
## Home\_Ownership + Verification\_Status + Loan\_Purpose + Debt\_to\_Income\_Ratio +   
## Delinquency + No\_of\_Loans + Negative\_Comments + Credit\_Balance +   
## Percentage\_Used\_Credit + No\_of\_Loans\_Other\_banks + Late\_Fees +   
## Last\_Amount + loans\_by\_state + loans\_by\_Rating + loan\_to\_income  
## attr(,"variables")  
## list(Loan\_Status, pesos, Loan\_Duration, Interest\_Rate, EMI, Employment,   
## Home\_Ownership, Verification\_Status, Loan\_Purpose, Debt\_to\_Income\_Ratio,   
## Delinquency, No\_of\_Loans, Negative\_Comments, Credit\_Balance,   
## Percentage\_Used\_Credit, No\_of\_Loans\_Other\_banks, Late\_Fees,   
## Last\_Amount, loans\_by\_state, loans\_by\_Rating, loan\_to\_income)  
## attr(,"factors")  
## pesos Loan\_Duration Interest\_Rate EMI Employment  
## Loan\_Status 0 0 0 0 0  
## pesos 1 0 0 0 0  
## Loan\_Duration 0 1 0 0 0  
## Interest\_Rate 0 0 1 0 0  
## EMI 0 0 0 1 0  
## Employment 0 0 0 0 1  
## Home\_Ownership 0 0 0 0 0  
## Verification\_Status 0 0 0 0 0  
## Loan\_Purpose 0 0 0 0 0  
## Debt\_to\_Income\_Ratio 0 0 0 0 0  
## Delinquency 0 0 0 0 0  
## No\_of\_Loans 0 0 0 0 0  
## Negative\_Comments 0 0 0 0 0  
## Credit\_Balance 0 0 0 0 0  
## Percentage\_Used\_Credit 0 0 0 0 0  
## No\_of\_Loans\_Other\_banks 0 0 0 0 0  
## Late\_Fees 0 0 0 0 0  
## Last\_Amount 0 0 0 0 0  
## loans\_by\_state 0 0 0 0 0  
## loans\_by\_Rating 0 0 0 0 0  
## loan\_to\_income 0 0 0 0 0  
## Home\_Ownership Verification\_Status Loan\_Purpose  
## Loan\_Status 0 0 0  
## pesos 0 0 0  
## Loan\_Duration 0 0 0  
## Interest\_Rate 0 0 0  
## EMI 0 0 0  
## Employment 0 0 0  
## Home\_Ownership 1 0 0  
## Verification\_Status 0 1 0  
## Loan\_Purpose 0 0 1  
## Debt\_to\_Income\_Ratio 0 0 0  
## Delinquency 0 0 0  
## No\_of\_Loans 0 0 0  
## Negative\_Comments 0 0 0  
## Credit\_Balance 0 0 0  
## Percentage\_Used\_Credit 0 0 0  
## No\_of\_Loans\_Other\_banks 0 0 0  
## Late\_Fees 0 0 0  
## Last\_Amount 0 0 0  
## loans\_by\_state 0 0 0  
## loans\_by\_Rating 0 0 0  
## loan\_to\_income 0 0 0  
## Debt\_to\_Income\_Ratio Delinquency No\_of\_Loans  
## Loan\_Status 0 0 0  
## pesos 0 0 0  
## Loan\_Duration 0 0 0  
## Interest\_Rate 0 0 0  
## EMI 0 0 0  
## Employment 0 0 0  
## Home\_Ownership 0 0 0  
## Verification\_Status 0 0 0  
## Loan\_Purpose 0 0 0  
## Debt\_to\_Income\_Ratio 1 0 0  
## Delinquency 0 1 0  
## No\_of\_Loans 0 0 1  
## Negative\_Comments 0 0 0  
## Credit\_Balance 0 0 0  
## Percentage\_Used\_Credit 0 0 0  
## No\_of\_Loans\_Other\_banks 0 0 0  
## Late\_Fees 0 0 0  
## Last\_Amount 0 0 0  
## loans\_by\_state 0 0 0  
## loans\_by\_Rating 0 0 0  
## loan\_to\_income 0 0 0  
## Negative\_Comments Credit\_Balance  
## Loan\_Status 0 0  
## pesos 0 0  
## Loan\_Duration 0 0  
## Interest\_Rate 0 0  
## EMI 0 0  
## Employment 0 0  
## Home\_Ownership 0 0  
## Verification\_Status 0 0  
## Loan\_Purpose 0 0  
## Debt\_to\_Income\_Ratio 0 0  
## Delinquency 0 0  
## No\_of\_Loans 0 0  
## Negative\_Comments 1 0  
## Credit\_Balance 0 1  
## Percentage\_Used\_Credit 0 0  
## No\_of\_Loans\_Other\_banks 0 0  
## Late\_Fees 0 0  
## Last\_Amount 0 0  
## loans\_by\_state 0 0  
## loans\_by\_Rating 0 0  
## loan\_to\_income 0 0  
## Percentage\_Used\_Credit No\_of\_Loans\_Other\_banks  
## Loan\_Status 0 0  
## pesos 0 0  
## Loan\_Duration 0 0  
## Interest\_Rate 0 0  
## EMI 0 0  
## Employment 0 0  
## Home\_Ownership 0 0  
## Verification\_Status 0 0  
## Loan\_Purpose 0 0  
## Debt\_to\_Income\_Ratio 0 0  
## Delinquency 0 0  
## No\_of\_Loans 0 0  
## Negative\_Comments 0 0  
## Credit\_Balance 0 0  
## Percentage\_Used\_Credit 1 0  
## No\_of\_Loans\_Other\_banks 0 1  
## Late\_Fees 0 0  
## Last\_Amount 0 0  
## loans\_by\_state 0 0  
## loans\_by\_Rating 0 0  
## loan\_to\_income 0 0  
## Late\_Fees Last\_Amount loans\_by\_state  
## Loan\_Status 0 0 0  
## pesos 0 0 0  
## Loan\_Duration 0 0 0  
## Interest\_Rate 0 0 0  
## EMI 0 0 0  
## Employment 0 0 0  
## Home\_Ownership 0 0 0  
## Verification\_Status 0 0 0  
## Loan\_Purpose 0 0 0  
## Debt\_to\_Income\_Ratio 0 0 0  
## Delinquency 0 0 0  
## No\_of\_Loans 0 0 0  
## Negative\_Comments 0 0 0  
## Credit\_Balance 0 0 0  
## Percentage\_Used\_Credit 0 0 0  
## No\_of\_Loans\_Other\_banks 0 0 0  
## Late\_Fees 1 0 0  
## Last\_Amount 0 1 0  
## loans\_by\_state 0 0 1  
## loans\_by\_Rating 0 0 0  
## loan\_to\_income 0 0 0  
## loans\_by\_Rating loan\_to\_income  
## Loan\_Status 0 0  
## pesos 0 0  
## Loan\_Duration 0 0  
## Interest\_Rate 0 0  
## EMI 0 0  
## Employment 0 0  
## Home\_Ownership 0 0  
## Verification\_Status 0 0  
## Loan\_Purpose 0 0  
## Debt\_to\_Income\_Ratio 0 0  
## Delinquency 0 0  
## No\_of\_Loans 0 0  
## Negative\_Comments 0 0  
## Credit\_Balance 0 0  
## Percentage\_Used\_Credit 0 0  
## No\_of\_Loans\_Other\_banks 0 0  
## Late\_Fees 0 0  
## Last\_Amount 0 0  
## loans\_by\_state 0 0  
## loans\_by\_Rating 1 0  
## loan\_to\_income 0 1  
## attr(,"term.labels")  
## [1] "pesos" "Loan\_Duration"   
## [3] "Interest\_Rate" "EMI"   
## [5] "Employment" "Home\_Ownership"   
## [7] "Verification\_Status" "Loan\_Purpose"   
## [9] "Debt\_to\_Income\_Ratio" "Delinquency"   
## [11] "No\_of\_Loans" "Negative\_Comments"   
## [13] "Credit\_Balance" "Percentage\_Used\_Credit"   
## [15] "No\_of\_Loans\_Other\_banks" "Late\_Fees"   
## [17] "Last\_Amount" "loans\_by\_state"   
## [19] "loans\_by\_Rating" "loan\_to\_income"   
## attr(,"order")  
## [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## attr(,"intercept")  
## [1] 1  
## attr(,"response")  
## [1] 1  
## attr(,".Environment")  
## <environment: R\_GlobalEnv>  
## attr(,"predvars")  
## list(Loan\_Status, pesos, Loan\_Duration, Interest\_Rate, EMI, Employment,   
## Home\_Ownership, Verification\_Status, Loan\_Purpose, Debt\_to\_Income\_Ratio,   
## Delinquency, No\_of\_Loans, Negative\_Comments, Credit\_Balance,   
## Percentage\_Used\_Credit, No\_of\_Loans\_Other\_banks, Late\_Fees,   
## Last\_Amount, loans\_by\_state, loans\_by\_Rating, loan\_to\_income)  
## attr(,"dataClasses")  
## Loan\_Status pesos Loan\_Duration   
## "factor" "numeric" "numeric"   
## Interest\_Rate EMI Employment   
## "numeric" "numeric" "factor"   
## Home\_Ownership Verification\_Status Loan\_Purpose   
## "factor" "factor" "factor"   
## Debt\_to\_Income\_Ratio Delinquency No\_of\_Loans   
## "numeric" "numeric" "numeric"   
## Negative\_Comments Credit\_Balance Percentage\_Used\_Credit   
## "numeric" "numeric" "numeric"   
## No\_of\_Loans\_Other\_banks Late\_Fees Last\_Amount   
## "numeric" "numeric" "numeric"   
## loans\_by\_state loans\_by\_Rating loan\_to\_income   
## "numeric" "numeric" "numeric"

## [1] 0.997343394 0.662805515 0.498419144 0.318935157 0.313220523  
## [6] 0.309432875 0.263115337 0.216799296 0.204354741 0.212483411  
## [11] 0.219346518 0.153511397 0.140857101 0.154611027 0.137422235  
## [16] 0.092337512 0.110334353 0.089004496 0.126749405 0.151621238  
## [21] 0.087366563 0.116284103 0.053679844 0.075040912 0.060072889  
## [26] 0.066208173 0.049564478 0.085467024 0.045136327 0.044583957  
## [31] 0.005193925 0.002612367 0.059787372 0.078752990 0.099646231  
## [36] 0.040909022 0.059277159 0.099194773 0.038336781 0.058484085  
## [41] 0.053463350 0.066792348 0.038094402 0.021848262 0.029486298  
## [46] 0.049693170 0.007357189 0.032444969 0.022764077 0.021966101  
## [51] 0.066933337 0.072118731 0.047155620 0.035918676 0.000614611  
## [56] 0.017379091 0.022392754 0.055144097 0.034406441 0.025508795  
## [61] 0.027265888 0.080904135 0.026337747 0.010464666 0.017398868  
## [66] 0.057620205 0.051009160 0.063505575 0.043339513 0.022679070  
## [71] 0.009247068 0.006276462 0.044093071 0.028729162 0.017570037  
## [76] 0.007681528 0.026768829 0.080728170 0.040160595 0.019109344  
## [81] 0.023630225 0.029069256 0.007031071 0.015103014 0.042608746  
## [86] 0.002336451 0.041151311 0.021971276 0.024145205 0.011073581  
## [91] 0.008139204 0.033682992 0.017921543 0.008761952 0.010761719  
## [96] 0.005833927 0.029838029 0.069940207 0.023403170 0.011507790

## Predict the boosting model on test observations and test errors are evaluated.