**Capstone Milestone Report**

**By: Shikher Singh**

**Mentor: Ankit Jain**

**Problem:**

This project aimed at the case of customers @ default payments in Taiwan and compares the predictive accuracy of probability of default using various machine learning methods. From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients. Because the real probability of default is unknown, we will train our model to see if we can predict probability of default

**Client:**

Consumer default payment data.

**Data:**

Will be using 30k customer data in which we will be predicting parameter (default payment next month)

**Approach:**

Looking @data initially, we found out that there are 30k consumers for whom 23 variables are available. We can employ various methods to predict default payment (Yes = 1, No = 0), as the response variable.

This project reviewed the literature and used the following 23 variables as explanatory variables:

Y: default Payment(which needs to be predicted)

X1: Amount of the 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 payment. We tracked the past monthly payment records (from April to September, 2005) 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:

-1 = pay duly;

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.

**DataSet is available in github:**

<https://github.com/singh0021/DataScienceMasters/blob/master/default_%20credit_clients.xls>

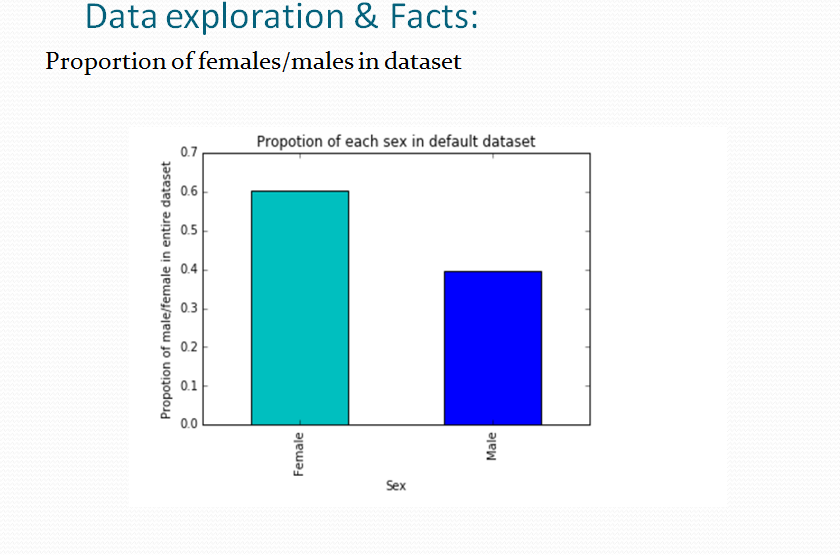
**Training DataSet:**

22k clients were picked to train the model. Cross-validation was also done on this training data set**.**

**Testing DataSet:**

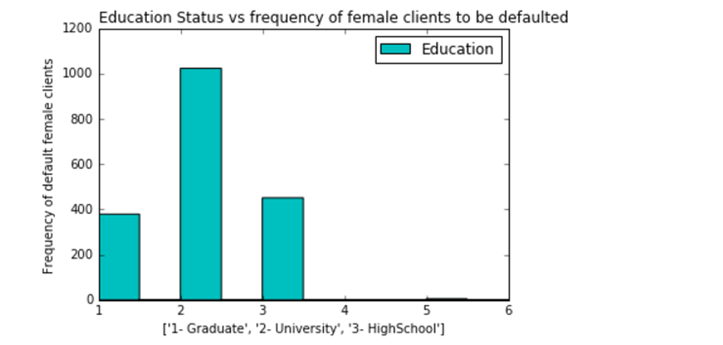
8k clients were chosen to test machine learning techniques.

**What to achieve:** Y to predict (0=will not default, 1= will default)

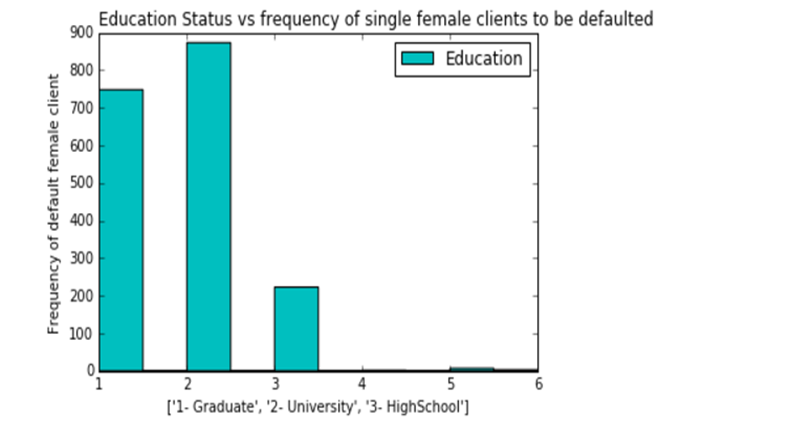


Just tried to build data stories doing data exploration and came to know that proportion of females are higher compared to males. It is evident from above diagram. Therefore, our focus would start from females first, then males. Rigorous exploration leads to below facts.

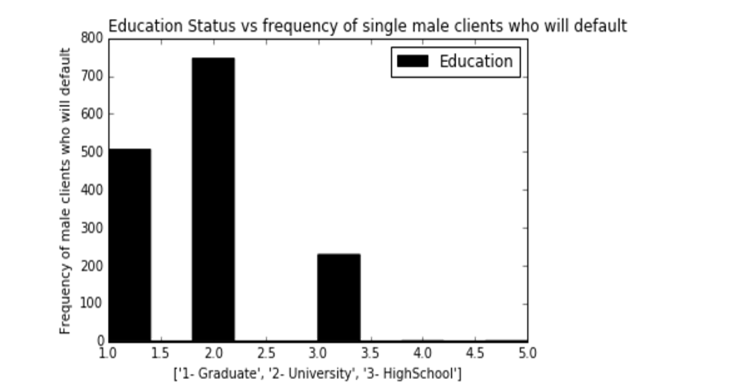
1. Married females who are university passed out are more likely to default than graduates.



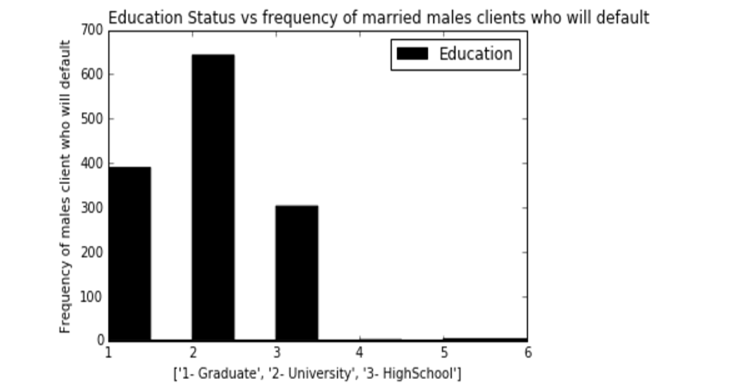
2. Single female university students are also likely to default more in comparison to graduate student.



3. Single males who are university passed out also likely to default as compared to graduates.



4. Married university passed out males are also more likely to default than graduates are



Based on above findings, we came to conclude that there is high rate of default in females as compared to males esp. University passed out students. Therefore, we excluded other features such as BILL\_PAYMENT in consecutive months, BILL\_AMOUNT in consecutive months, PAID\_AMOUNT in consecutive months and LIMIT BALANCE in clients’ credit card.

Considering above points in mind, shall try to find probability of default among females, males esp. In university passed out. To achieve these results, we will implement machine learning algorithms KNN, SVM, RandomForest, Logistics Regression and NaiveBayes.

**Logistics Regression:**

It is a classification problem, we started with Logistics Regression to train our 22k training example to see if credit card default rate can be predicted effectively. Cross-validation is done in five folds based on training data-set. Hyperparameter or regularized parameter(C=1) was achieved to get the best predictor. However, testing and training accuracy can only be maximized up to 79.53 %. Hence we will try our next model.

**K Nearest Neighbors:**

Second classification model was built using optimized number of neighbors(30) so that default clients can be predicted accurately. We got only 78.65 % testing as well as training accuracy.

**Support Vector Machine**:

Support vector machine are considered best models since we only care about support vectors to train the model.

We normalized the training and testing data(based on training data only).

However, result have not improved much even after applying Cross-validation to get best Slack (1) , Gamma(0.001) and Kernel(RBF). Accuracy moves around 78-79%.

This technique took much time to train 22k records.

**NaiveBayes:**

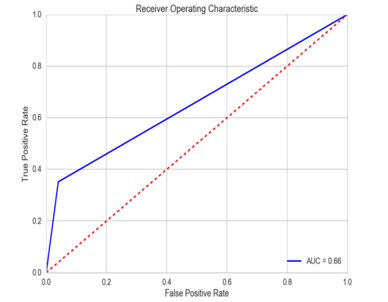
NaiveBayes is based on prior probability and then posterior probability is predicted. Algorithm need features to be independent. After dataset independent features are used in NaiveBayes, accuracy improved to 80 %.

**Random Forest :**

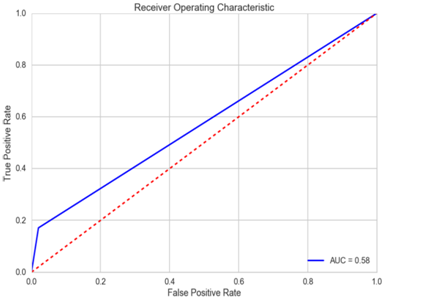
Random Forest is an ensemble model based on decision trees. We don’t need to normalize the data even though cross-validation was used to pick best parameters(Number of trees-20, maximum depth-7).

Results were surprising to see 82 % training as well as testing accuracy.

Also ROC curves were plotted for each model. ROC curve represents how efficiently, effectively test data can be plotted on the graph. The accuracy of test depends upon how well test data separates the group being tested into those who defaulted and who did not default. RandomForest build the best ROC curve as evident from below graph.



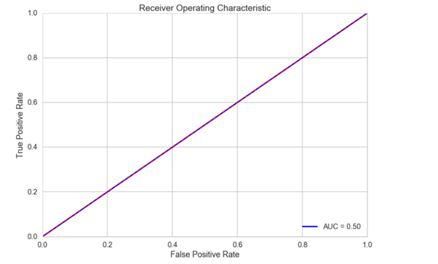
**RandomForest**



**NaiveBayes**

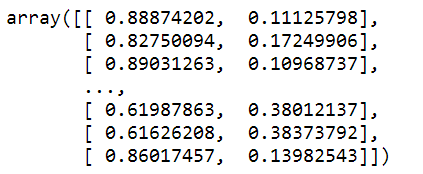


**KNN**



**SVM**

Using Random Forest to test on independent test data for females who are university passed out, probability of default were found and below is result.



Results looks pretty convincing to say that bank should implement below recommendation when targeting customer to avoid loss due to default payment.

1. Try to provide credit cards to males more as compared to females.
2. Always do education status check before providing cards to the client as proportion of default is very high in university passed out.
3. Plan to give lower limit schemes for university passed out males/females so that they are less likely to default.