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:
We 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**.**

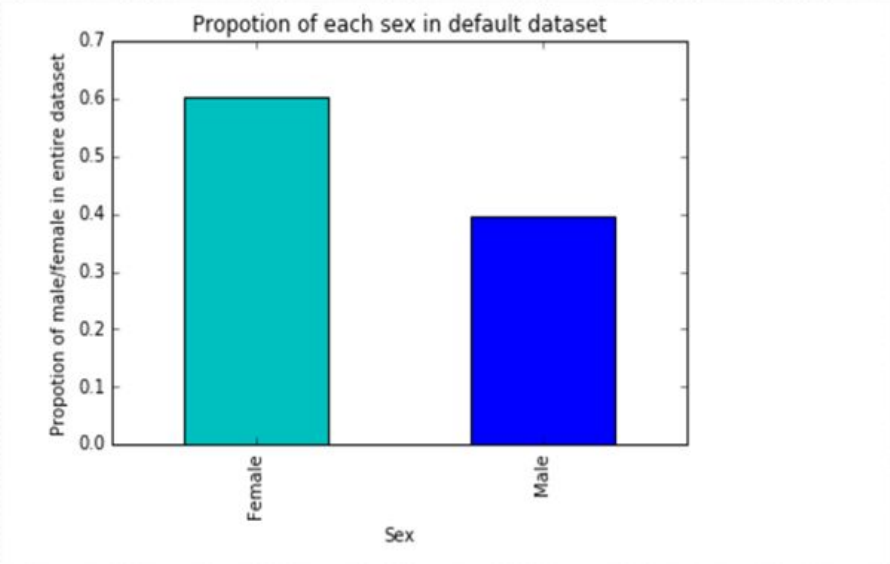
Testing DataSet:

8k clients were chosen to test machine learning techniques.

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

Data exploration & Facts:

Proportion of females/males in dataset



Recommendations:

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.

Requirement of machine learning:

Trying to find probability of defaults among females and males , hence using classification models

Logistics Regression:

- Features to be excluded: Initial amount of balance given, bill amount in last 3 months as well as amount of previous statements in last 5 months excluded
- Cross-validation : regularized parameter(C=1) was the best estimator
- Max accuracy obtained: 79.53 (better than before)

K Nearest Neighbour:

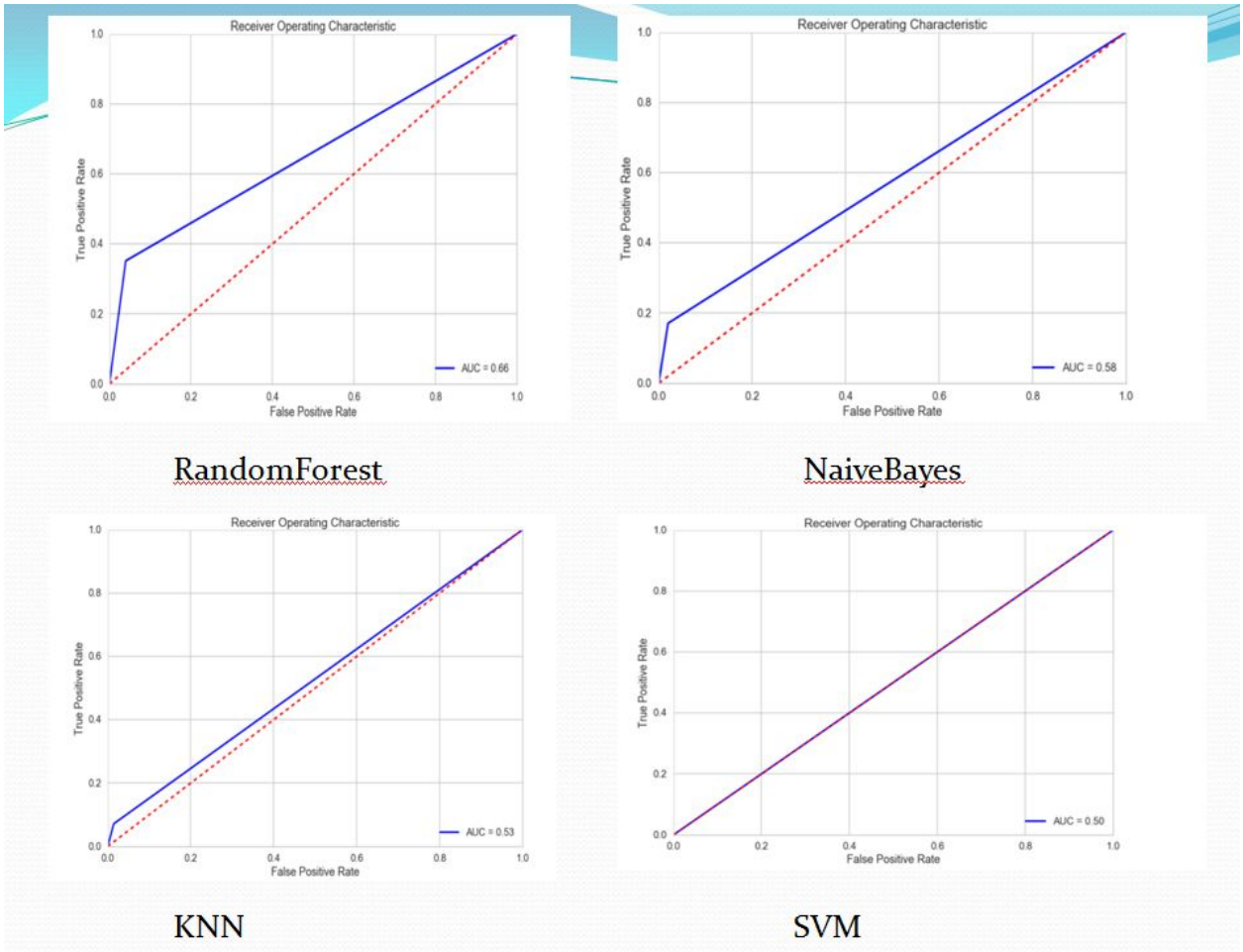
- Same features were excluded as was in Logistics Regression
- Cross-validation was used to pick optimized value of neighbours
- Neighbors-30
- Accuracy achieved: 78.65%

NaiveBayes

- Initial accuracy achieved was 36 percent and thought of reject naive-Bayes.
- Go for independent features
- Accuracy score achieved was 80.38% percent

Random Forest :

- No normalization needed
- More features were excluded such as history of payments for last 4 months
- Pretty fast
- Cross-validation were used to pick best features(Number of trees-20, maximum depth-7)
- Accuracy score achieved: 82.77 %
- So far the best estimator



Using the best model for this dataset probability of default was calculated between males and females.

```
array([[ 0.88874202,  0.11125798],
       [ 0.82750094,  0.17249906],
       [ 0.89031263,  0.10968737],
       ...,
       [ 0.61987863,  0.38012137],
       [ 0.61626208,  0.38373792],
       [ 0.86017457,  0.13982543]])
```