BankMarketing

Krishna Teja

4/21/2020

Executive Summary

Marketing strategies of banks are of various types and cellular is one among them. Bank marketing data set have 16 dependent variables for determining the result of a campaign call. Each call can end with customer subscribing to the new scheme or rejecting. Aim is to build a model that efficiently predicts the positive result of a campaign.Building such model helps prioritising the customer records. As a result efficieny of the campaigns shoot up as bank staff would target the records with more probability of subscribing.

Gist of DataSet

Dataset consists of 45k rows and 17 variables. Out of the 17 variables, one variable comprises of output and the rest contribute to input variables. Each row represents one customer and Output variable takes the value of 0 or 1 indicating failure or success, respectively, of the campaign call with customer.

```
##       age                job          marital          education
##  Min.   :18.00   blue-collar:9732   divorced: 5207   primary  : 6851
##  1st Qu.:33.00   management :9458   married :27214   secondary:23202
##  Median :39.00   technician :7597   single  :12790   tertiary :13301
##  Mean   :40.94   admin.     :5171                    unknown  : 1857
##  3rd Qu.:48.00   services   :4154
##  Max.   :95.00   retired    :2264
##                  (Other)    :6835
##  default        balance       housing      loan            contact
##  no :44396   Min.   : -8019   no :20081   no :37967   cellular :29285
##  yes:  815   1st Qu.:    72   yes:25130   yes: 7244   telephone: 2906
##              Median :   448                           unknown  :13020
##              Mean   :  1362
##              3rd Qu.:  1428
##              Max.   :102127
##
##       day            month          duration         campaign
##  Min.   : 1.00   may    :13766   Min.   :   0.0   Min.   : 1.000
##  1st Qu.: 8.00   jul    : 6895   1st Qu.: 103.0   1st Qu.: 1.000
##  Median :16.00   aug    : 6247   Median : 180.0   Median : 2.000
##  Mean   :15.81   jun    : 5341   Mean   : 258.2   Mean   : 2.764
##  3rd Qu.:21.00   nov    : 3970   3rd Qu.: 319.0   3rd Qu.: 3.000
##  Max.   :31.00   apr    : 2932   Max.   :4918.0   Max.   :63.000
##                  (Other): 6060
##      pdays          previous           poutcome       y
##  Min.   : -1.0   Min.   :  0.0000   failure: 4901   no :39922
```

```
##  1st Qu.: -1.0   1st Qu.:  0.0000   other  : 1840   yes: 5289
##  Median : -1.0   Median :  0.0000   success: 1511
##  Mean   : 40.2   Mean   :  0.5803   unknown:36959
##  3rd Qu.: -1.0   3rd Qu.:  0.0000
##  Max.   :871.0   Max.   :275.0000
##
```

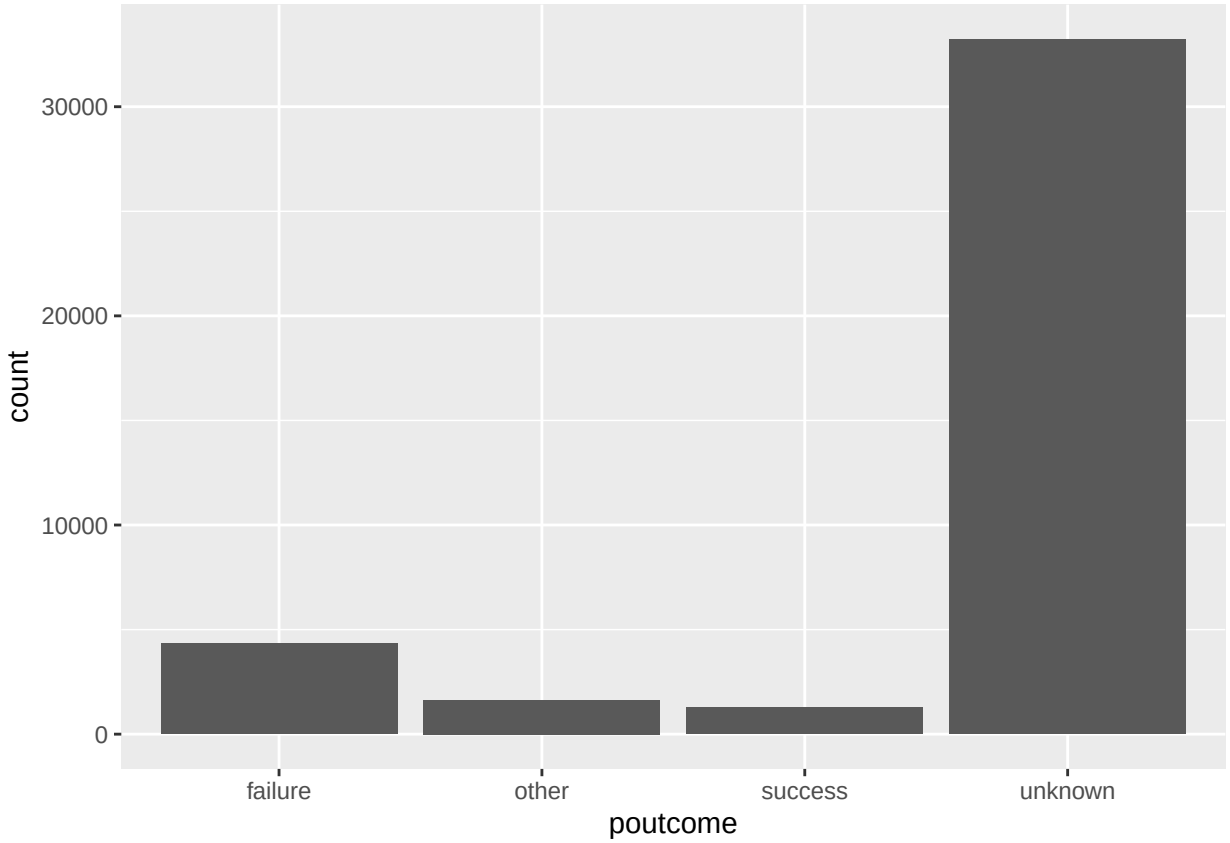
DataCleaning and Pre-processing

Check for the NA values. NA values make the prediction uncertain at various levels. To avoid all the conflicts and the ambiguity cleaning up NA values is important.

Remove outliers from all the numeric data. Balance is one main column which have wide range of values with outliers. Using interquartile range concept, outliers are found and respective rows are eliminated.

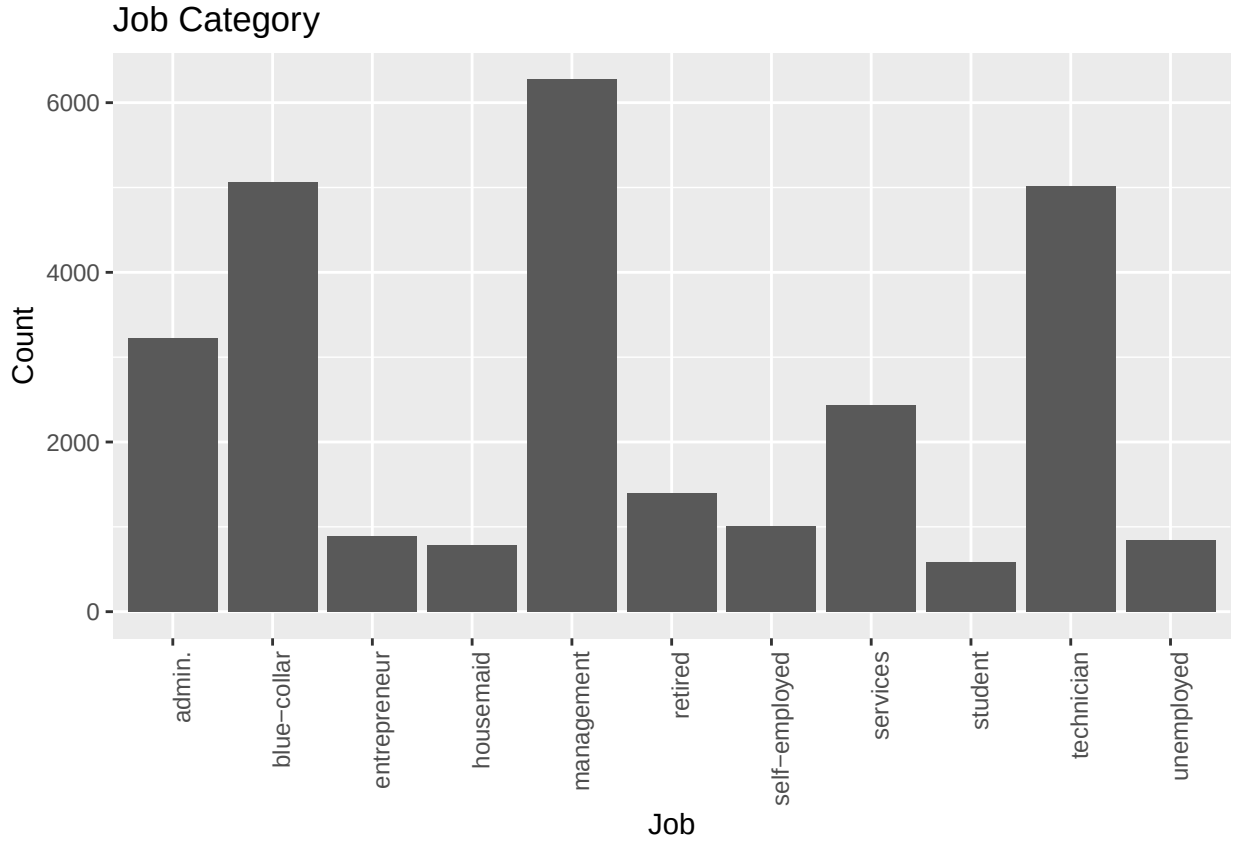
Within the categorical column data, check for the categories which reveal the least about the respective variable. For Instance, 'unknown' as a category reveal nothing in some cases. So, inclusion of such rows is equivalent to possessing NA values. Before considering to cleanup, verify the proportion of such categories in each variable.

If significant part of data is creating the confusion, getting rid of columns helps more than excluding huge number of rows. 'poutcome' is one such feature in this case. In case of 'poutcome' 70% of data is termed 'unknown' and excluding 70% of data is not the best thing to do. Rather exclude the 'poutcome'.

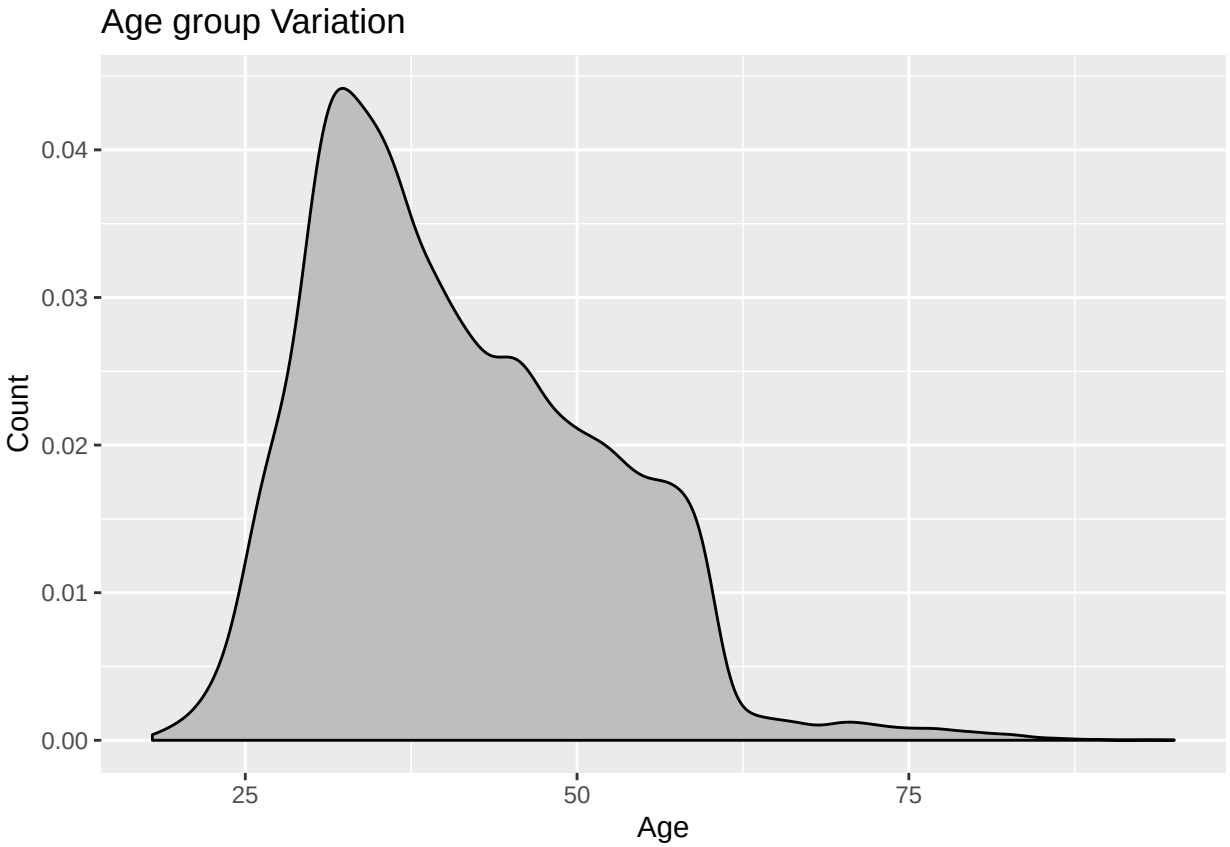


Exploratory Data Analysis

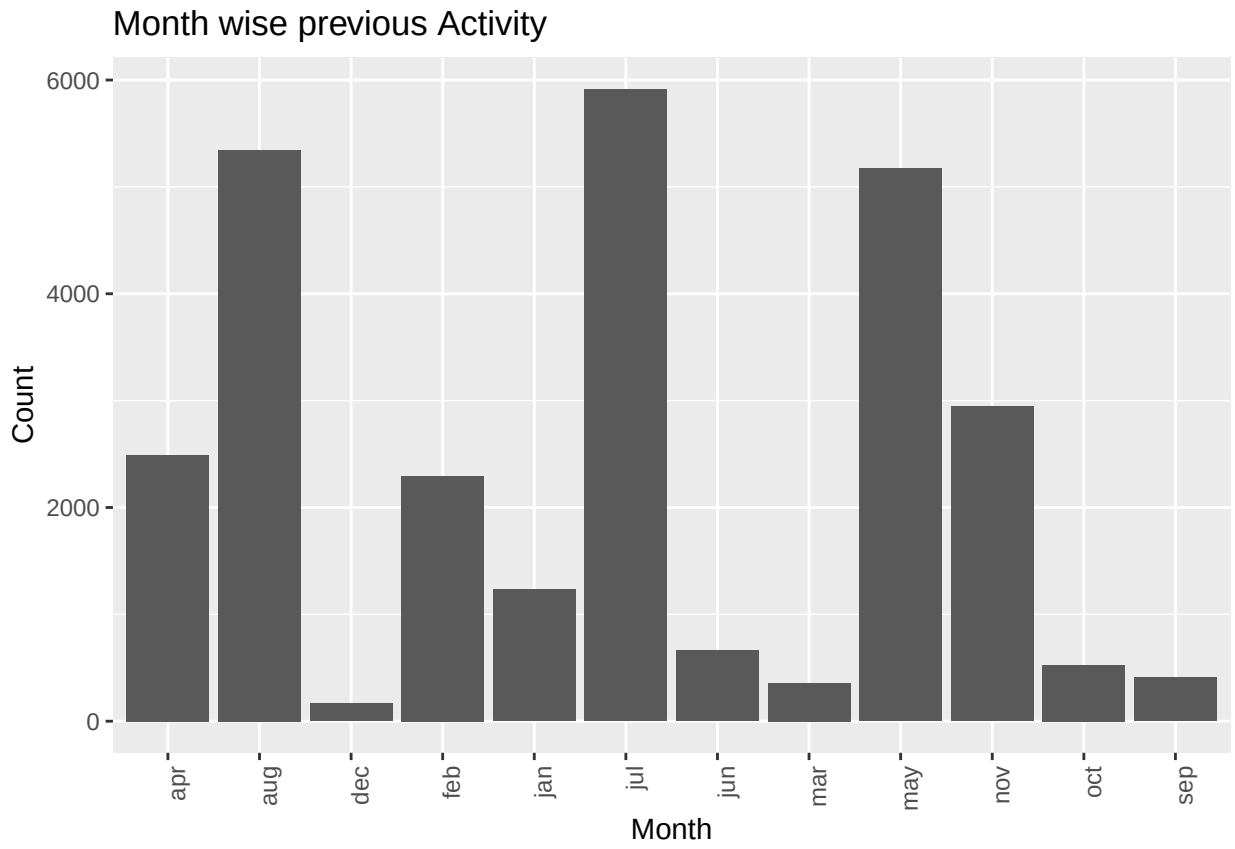
Types of job categories the customers possess are management, blue-collar, technician, services, admin, unemployed, entrepreneur, housemaid, retired, self-employed and student. Managment, technician and blue-collor jobs are more common professions.



Age group of customers vary from 15 till 80 . Majority of the customers fall into the age group of 25-40. Spikes in the data below supports this inference.



July recorded the most activity followed by August and May. Least activity was recorded in December.

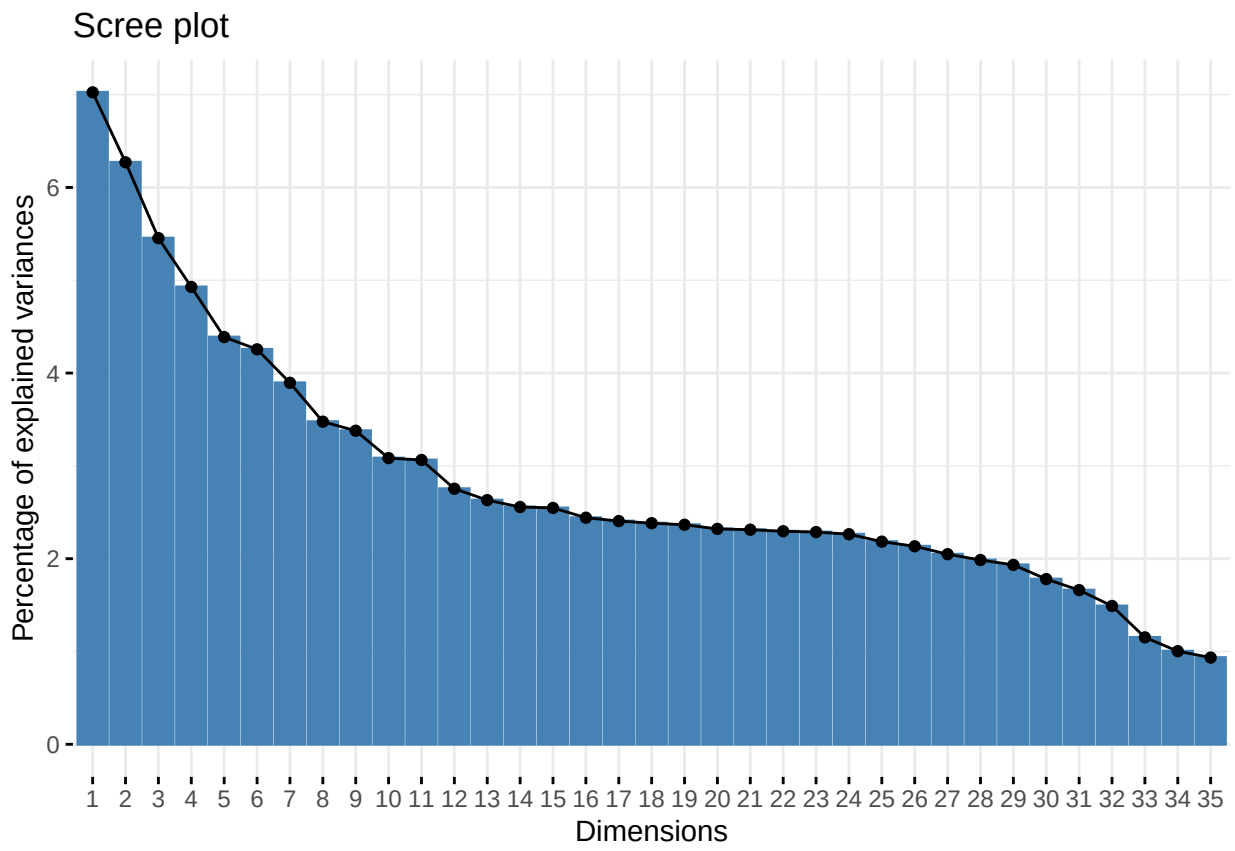


Data Preparation and Analysis

Features of type numeric require scaling and centering of data which will avoid the magnitude differences. For instance age will be ranging from 15-70 and account balance could go upto millions.All such different scales must be made unifrom.
Categorical features have to be handled before building models with them. Dummy variables come into picture and includes additional numeiric columns equivalent to categorical data.

Once the data is preprocessed, Challenge is to choose part of dataset that defines maximum part of output variable.Correlation metrics proved that there is not much correlation between the 6 numerical features.

By principle component analysis we define new variables with transformed columns. By performing PCA, we lose the structure of the data but the variance between the variables is preserved.Output of PCA will be the multiple principle componenets with a count equal to number of features post preprocessing. Out of the 48 variables, top 30 explains more than 90 % of the variance. So choosing 30 variables is the best optino considering computation and efficieny.



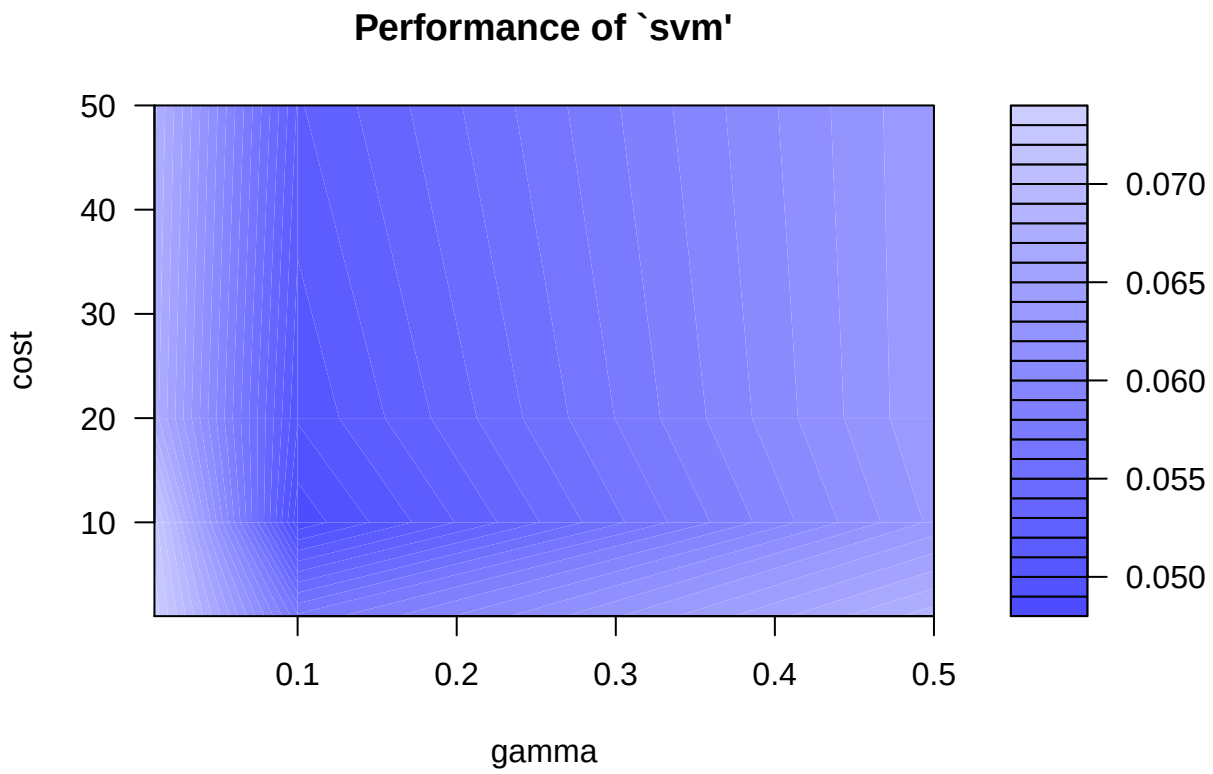
Machine Learning Models

Before applying machine learning, Data is to be partitioned into test and train Datasets. Any algorithm have a set of parmaeters which define the models. Hyperparameter tuning gives the best possible values for the parameters.Each model is built with parmeter values from hyperparameter tuning and is trained with train dataset. Predictions are compared by implemeting using testData.

Support vector machine(SVM)

Hyper Parameter tuning

Choosing the kernel for SVM gives the set of parameters. For radial kernel, cost and gamma defines the model output. Post hyperparameter tuning the values for the parameters like cost and gamma are 10 and 0.1 respectively implies that a model built with these values gives the best possible SVM results for the data.



SVM Results

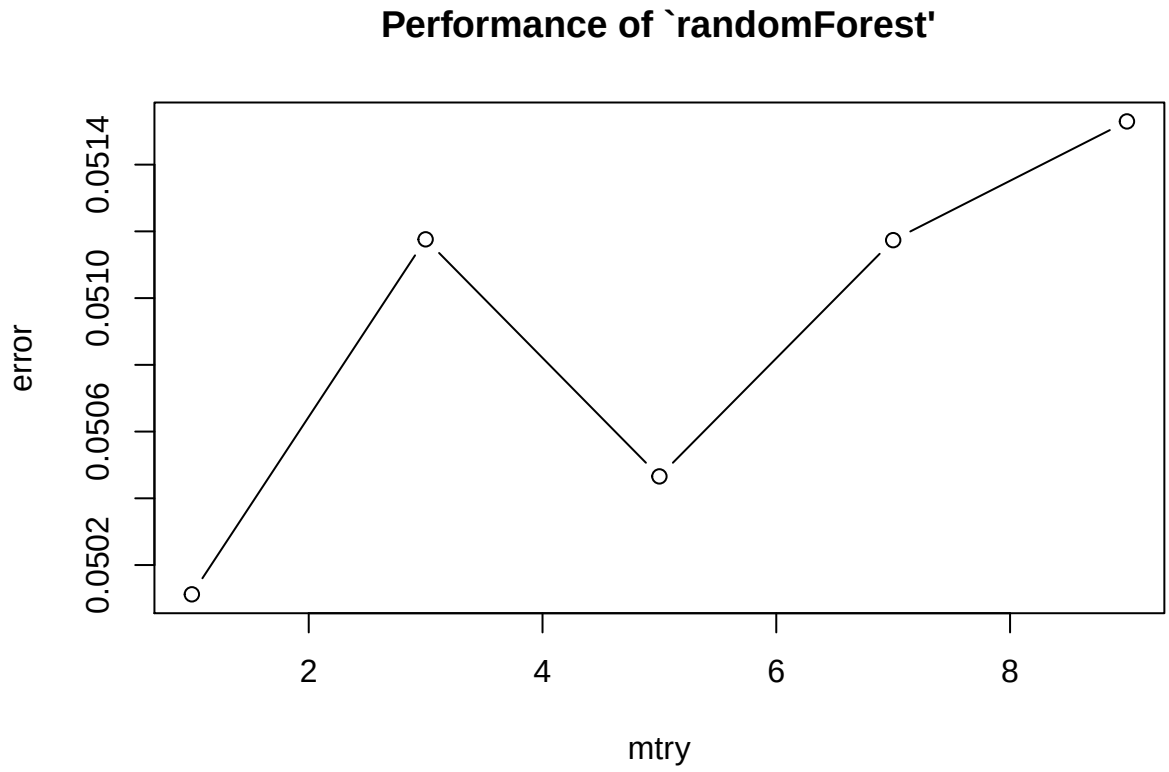
Prediction results of the model are as below. Model is giving an accuracy of 83.9% and recall score of positive target variable is 61%.

```
##       Predicted
## Actual    0    1
##      0 6229  874
##      1  449  703
```

```
## [1] "Accuracy:"
## [1] 0.8397335
## [1] "Precision:"
##         0         1
## 0.9327643 0.4457831
## [1] "Recall:"
##         0         1
## 0.8769534 0.6102431
## [1] "F1 Score:"
##         0         1
## 0.9039983 0.5152070
## [1] "Error Rates"
## [1] 0.1602665
```

Random Forest

Hyper parameter tuning included finding best balue for mtry and max.depth parameters. In this case the best parameters is 5 for mtry. Though value of 1 seems to give the best model, there is no marginal difference when compared with that of 5.



Random Forest Results

Prediction results of the model are as below. Model is giving an accuracy of 84.8% and recall score of positive target variable is 72%.

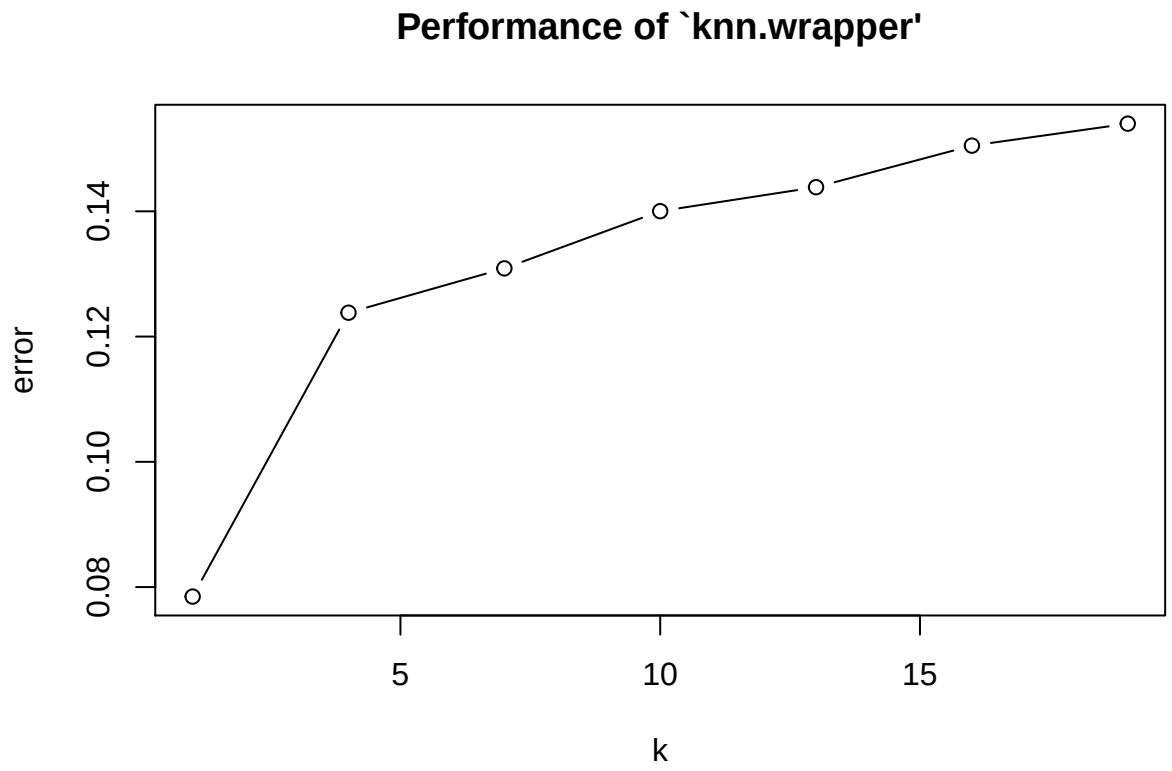
```
##       Predicted
## Actual    0    1
##      0 6164  939
##      1  335  817
```

```
## [1] "Accuracy:"
## [1] 0.8456693
## [1] "Precision:"
##         0         1
## 0.9484536 0.4652620
## [1] "Recall:"
##         0         1
## 0.8678023 0.7092014
## [1] "F1 Score:"
```

```
##         0         1
## 0.9063373 0.5618982
## [1] "Error Rates"
## [1] 0.1543307
```

K-nearest neighbours

Hyperparameter tuning of KNN includes finding best K value for the dataset. K value of 7 gives the best re-



sults in this case.

KNN Results.

Prediction results of the model are as below. Model is giving an accuracy of 85.13% and recall score of positive target variable is 74.4%.

```
## [1] 85.16051
```

```
##       Predicted
## Actual    0    1
##      0 6171  932
##      1  293  859
```

```
## [1] "Accuracy:"
## [1] 0.8516051
## [1] "Precision:"
```

```
##         0         1
## 0.9546720 0.4796203
## [1] "Recall:"
##         0         1
## 0.8687878 0.7456597
## [1] "F1 Score:"
##         0         1
## 0.9097074 0.5837581
## [1] "Error Rates"
## [1] 0.1483949
```

Conclusion:

After preprocessing, EDA and PCA of data, smote sampling is done to handle the bias in the target variable. Three classification algorithms(SVM, randomForest and KNN) are implemented on the data set. Main motto of the campaign is to find the customer records with good chance of converision. So more than accuracy, recall score determines the model in this case.Out of the 3 algorithms, KNN with a recall score of 74.4 and an error rate of 0.14 gives the best model.