Test Results:

Initially original training Data is modeled using SVM and Random Forest Classifiers. Though the accuracies are good enough, the models were not able to predict the positive 'yes' samples correctly as the original data is biased.(93% are negative 'no' while only 7 % are positive 'yes' Samples). So to overcome that training set is oversampled using ROSE library of R. The new balanced training set has got 49 percent training samples and 51 percent of negative samples. Models are then trained using the new balanced training set and test set accuracies and measures are computed. Given below are the test set accuracies and measures for various models.

	Measure				
Model	*Accuracy*	*Sensitivity*	*Specificity*	*Pos Pred Value*	*Neg Pred Value*
SVM UnBalanced	0.7374	0.08183	0.96815	0.47484	0.74977
SVM Balanced	0.7177	0.5483	0.7773	0.4642	0.8302
RF UnBalanced	0.7585	0.13952	0.97638	0.67518	0.76328
RF Balanced	0.6941	0.595	0.7289	0.4358	0.8365

Table showing different test set accuracy measures for models built on Original and balanced training data.

As you can see though the accuracies are higher for SVM and RF for original or unbalanced sets , Sensitivity is significantly low. i.e. these models cannot predict much of the positive samples perfectly. In the dataset taken in the project, there is not much cost involved in making phone calls, which signifies that all the people who are willing to subscribe for term deposit should be targeted even when more people who usually don't subscribe for term deposits also targeted with wrong prediction i.e. true positives should be considerably high and is the driving factor in these kind of campaign analysis, which makes sensitivity or Recall the dominating factor for test measure.

Given below the graphical representation of variations of test measures before and after balancing training data.

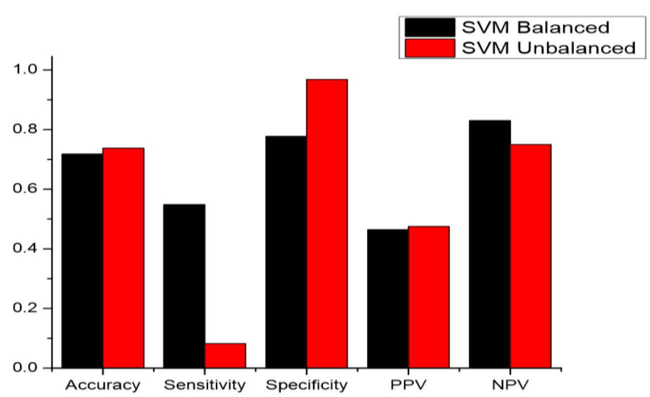

Fig1: Test measures for SVM Modeled on Balanced and Original Datasets

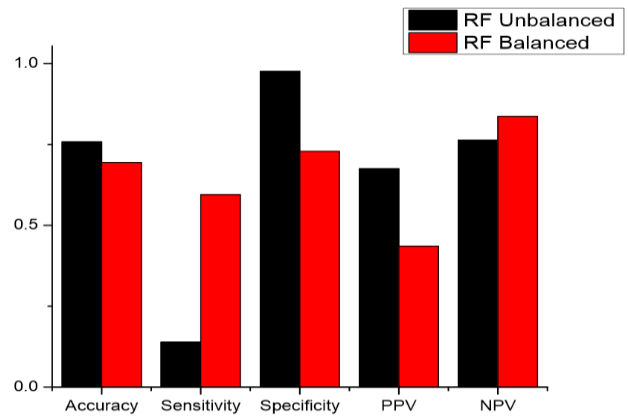

Fig2: Test measures for RF Modeled on Balanced and Original Datasets

As one can see sensitivity (recall) of test data significantly improved after balancing the training data.

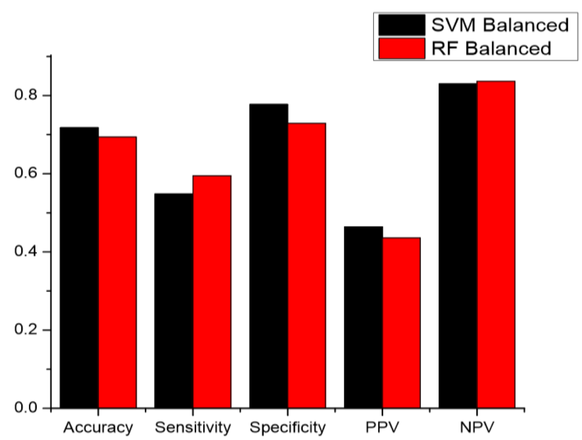

Fig3: Test measures for SVM and RF models built on balanced data.

From fig3 when both models are compared though accuracy of SVM is a bit more Random Forest has got more sensitivity(Recall) and is able to predict more number of true positives than SVM(Evident from confusion matrices listed in the later part of this document). So Random Forest is taken as base model for classifying the input data in this project.

A Comparative analysis is made on different categories of features (categories are taken based on the domain and the way they have been collected) and test measures for those subsets are computed and listed in the table below.

	Measure				
Model	*Accuracy*	*Sensitivity*	*Specificity*	*Pos Pred Value*	*Neg Pred Value*
RF_BankClientData	0.5484	0.5867	0.5349	0.3074	0.7862
RF_PreviousContactInfo	0.6984	0.4265	0.7941	0.4215	0.7973
RF_SocioEconomic	0.4738	0.9167	0.3179	0.3211	0.9156

Table listing test accuracy measures for Categories BankClientAttributes,PreviousContactInfo attributes and Social Economic Attibutes trained using Random Forest Classifier.

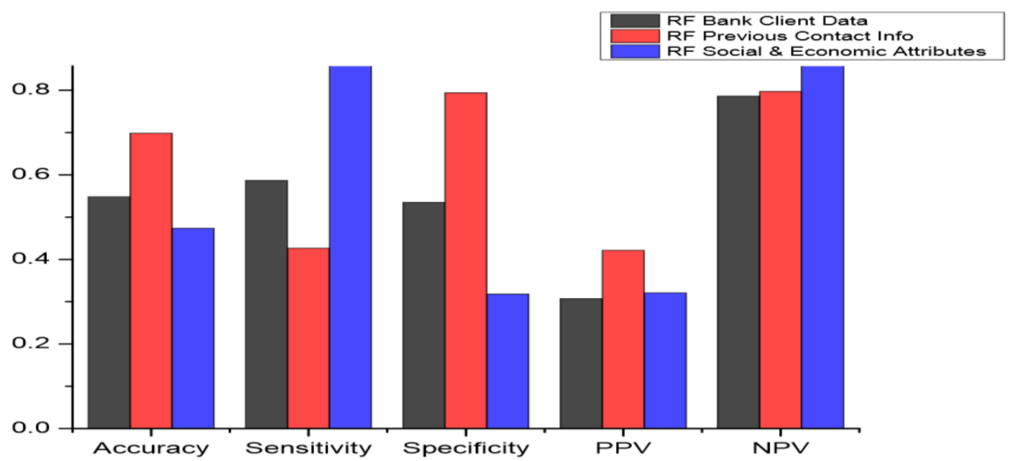

Fig4: Test Measures for different categorical attributes trained using Random Forests.

Categorically Bank Client information attributes and Social Economic Attributes contribute more to sensitivity i.e. prediction of more number of true positives.

Individual feature contribution to classification is computed using gini measure for both balanced and original datasets and the results are shown below:

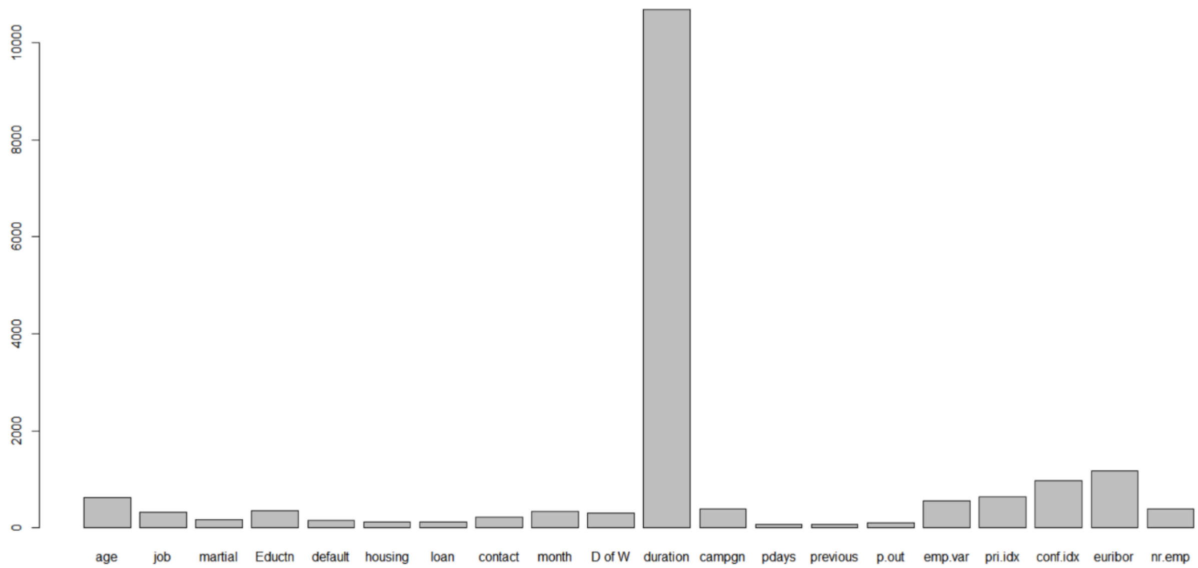

Fig5: Gini measures for all attributes for balanced dataset trained using RF classifier

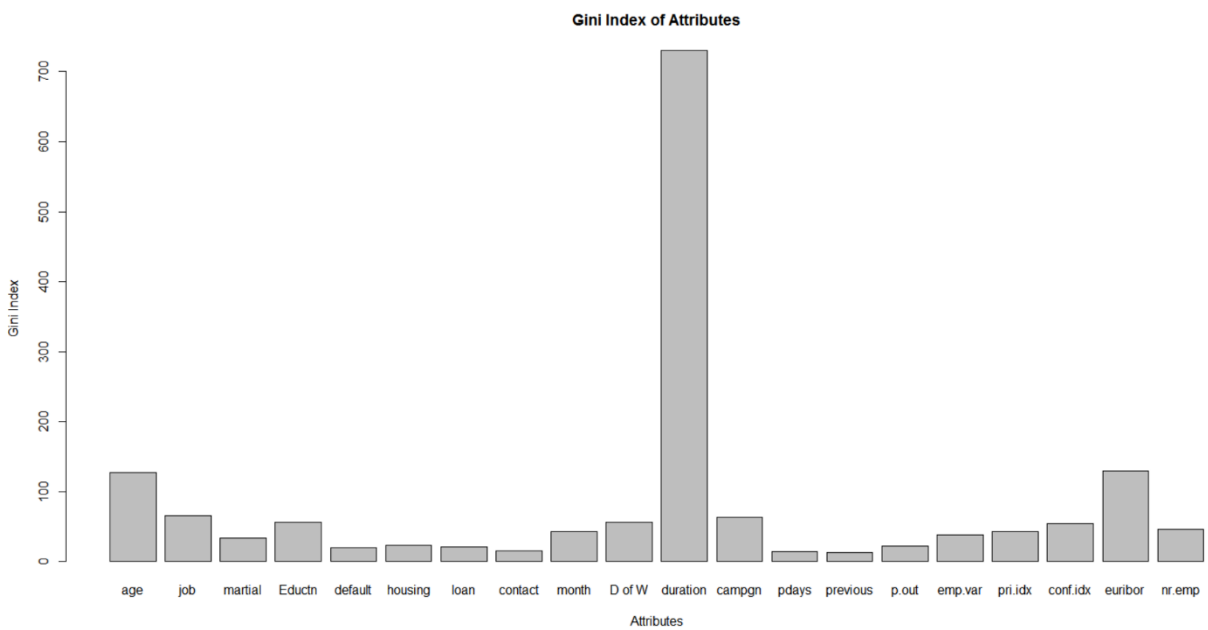

Fig6: Gini measures for all attributes for Original dataset trained using RF classifier

From the above graphs it is clearly evident that call duration is the dominating predictor for term deposit subscription, i.e. longer call duration signify that the targeted customer more likely to subscribe for term deposit while 0 call duration signifies sure shot denial of term deposit.

Given below the confusion matrices and gini indices that are computed in r and which are used for the a bove analysis.

```
> confusionMatrix(predSVMUnBalanced,test$y, positive = 'yes')
Confusion Matrix and Statistics
          Reference
Prediction   no  yes
       no  7296 2435
       yes  240  217
               Accuracy : 0.7374
                 95% CI : (0.7288, 0.746)
            Sensitivity : 0.08183
            Specificity : 0.96815
         Pos Pred Value : 0.47484
         Neg Pred Value : 0.74977
       'Positive' Class : yes
```

```
> confusionMatrix(predSVMBalanced,test$y, positive = 'yes')
Confusion Matrix and Statistics
          Reference
Prediction   no  yes
       no  5858 1198
       yes 1678 1454
               Accuracy : 0.7177
                 95% CI : (0.7089, 0.7264)
            Sensitivity : 0.5483
            Specificity : 0.7773
         Pos Pred Value : 0.4642
         Neg Pred Value : 0.8302
       'Positive' Class : yes
```

```
> confusionMatrix(pred_randomForest_UnBalanced,testConvert$y, positive = '2')
Confusion Matrix and Statistics
          Reference
Prediction    1    2
         1 7358 2282
         2  178  370
               Accuracy : 0.7585
                 95% CI : (0.7501, 0.7668)
           Sensitivity : 0.13952
            Specificity : 0.97638
         Pos Pred Value : 0.67518
         Neg Pred Value : 0.76328
       'Positive' Class : 2
```

```
> confusionMatrix(pred_randomForest_Balanced,testConvert$y, positive = '2')
Confusion Matrix and Statistics
          Reference
Prediction    1    2
         1 5493 1074
         2 2043 1578
               Accuracy : 0.6941
                 95% CI : (0.685, 0.703)
        Sensitivity : 0.5950
            Specificity : 0.7289
         Pos Pred Value : 0.4358
         Neg Pred Value : 0.8365
       'Positive' Class : 2
```

```
> confusionMatrix(pred_randomForest_Balanced_BankClientData,testConvert$y, pos
itive = '2')
Confusion Matrix and Statistics
          Reference
Prediction    1    2
         1 4031 1096
         2 3505 1556
               Accuracy : 0.5484
                 95% CI : (0.5387, 0.5581)
           Sensitivity : 0.5867
            Specificity : 0.5349
         Pos Pred Value : 0.3074
         Neg Pred Value : 0.7862
       'Positive' Class : 2
```

```
>  confusionMatrix(pred_randomForest_Balanced_PreviousContactInfo,testConvert$
y, positive = '2')
Confusion Matrix and Statistics
          Reference
Prediction    1    2
         1 5984 1521
         2 1552 1131
               Accuracy : 0.6984
                 95% CI : (0.6894, 0.7073)
            Sensitivity : 0.4265
            Specificity : 0.7941
         Pos Pred Value : 0.4215
         Neg Pred Value : 0.7973
       'Positive' Class : 2
```

```
 confusionMatrix(pred_randomForest_Balanced_SocioEconomic,testConvert$y,  posi
tive = '2')
Confusion Matrix and Statistics
          Reference
Prediction    1    2
         1 2396  221
         2 5140 2431
               Accuracy : 0.4738
                 95% CI : (0.4641, 0.4835)
            Sensitivity : 0.9167
            Specificity : 0.3179
         Pos Pred Value : 0.3211
         Neg Pred Value : 0.9156
       'Positive' Class : 2
```

```
> importance(model_randomForest_Balanced)
                 MeanDecreaseGini
age                     590.65537
job                     330.44504
marital                 163.22718
education               359.29253
default                 161.63397
housing                 119.22438
loan                    107.69972
contact                 236.94526
month                   433.02469
day_of_week             322.46548
duration              10469.33353
campaign                340.53106
pdays                    64.30814
previous                 66.19270
poutcome                103.42929
emp.var.rate            388.84701
cons.price.idx          506.40387
cons.conf.idx           904.05571
euribor3m              1172.03596
nr.employed            1042.82195
```

```
> importance(model_randomForest_UnBalanced)
                 MeanDecreaseGini
age                     127.47427
job                      65.54854
marital                  33.65610
education                56.76417
default                  19.47196
housing                  22.83553
loan                     21.18036
contact                  14.85211
month                    42.42188
day_of_week              56.39580
duration                729.94529
campaign                 63.48623
pdays                    13.37985
previous                 12.30144
poutcome                 22.10939
emp.var.rate             38.14709
cons.price.idx           42.19490
cons.conf.idx            54.09051
euribor3m               129.55397
nr.employed              45.77751
```