BANK MARKETING ANALYSIS REPORT

Nirali Bandaru

Introduction

This study uses the "Bank Marketing" data set from the UCI Machine Learning Repository. The goal of this investigation is to predict whether or not a potential customer chooses to subscribe to a term deposit, described by target variable "y." The problem statement has been categorized as a classification problem.

The machine learning model chosen for this application is random forest. The model fits the data well. However, the model initially produced biased class errors due to an imbalance in the dataset. In order to overcome this, the data set was balanced, and lower error rates were produced as desired.

Dataset and Observations from Exploratory Data Analysis

Number of observations = 45211

Number of predictors = 16

Target variable = y = whether client has subscribed a term deposit

Variable	Range/Levels	Description	Observation (EDA)
Age (N)	18-95	Age of client	Slightly skewed right, with mean at around 40 years old, most people at 30 years old.
Job (C)	7	Client's occupation	Most popular jobs were blue-collar, management, and technician.
Marital (C)	3	Client's marital status	Highest number of clients were married, followed by single clients, and lastly divorced clients.
Education (C)	4	Client's education level	Of known education levels of people, secondary education was highest.
Default (C)	2	Whether the client has ever defaulted	Less than 2% of people contacted defaulted.
Balance (N)	-8918 – 102127	Current bank balance	Distribution skewed right, with many outliers.
Housing (C)	2	Whether client has housing loan	Nearly 80% have housing loan, 20% do not.
Loan (C)	2	Whether client has personal loan	Only 20% have personal loan.
Contact (C)	3	Contact type	Cellular contact method was far more popular than telephone. Many people's contact method is unknown.
Month (C)	7	Last month of contact	Most people were last contacted in May.
Day (N)	1-31	Day of the month of last contact	Approx. the same number of people were contacted each day of the month, with dips in the distribution assumed to be weekend days and an unusual peak in the middle of the month.
Duration (N)	0.0 – 4918 (s)	Duration in seconds of last contact	Largely right-skewed distribution, with many outliers.
Campaign (N)	1-63	Number of times client has been contacted during this campaign, including last contact	Distribution skewed right, with many outliers.
Pdays	-1 – 871	Number of days passed since last contact	Most people were not previously contacted, mode = 0.
Previous	0-275	Number of times contacted before campaign	Most people were not previously contacted, with exceptions of a few extreme outliers.
Poutcome	4	Result of previous marketing campaign for this client	Most clients' previous outcome is unknown. Of the known outcomes, most of the previous campaigns failed.
y	2	Whether client has subscribed a term deposit (target variable)	Most people did not subscribe. Number of "no's" is approximately seven times the number of "yes's."

Model Selection and Validation

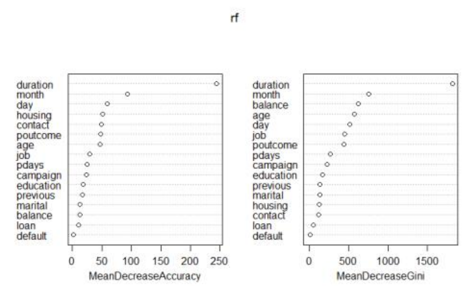
The response variable “y” is categorical and binary. This indicates that the most suitable model would be a classification model. The chosen model is **decision trees (random forest).** This is because decision trees are versatile and also provide information on which attributes have the highest predictive value. In this particular application, knowing what affects clients to subscribe to a plan at the bank would be the most useful insight that can be gained from analyzing a set of marketing data. Another suitable model is the *support vector classifier* which is conveniently intended to be used as a binary classifier. However, decision trees were chosen over SVC because of the higher interpretability of the former compared to the latter. Due to the large size of the data set, 70-30 training-testing ratio was used as opposed to k-fold validation. Since number of predictions = 16, mtry = sqrt(16) = 4.

Random Forest Model Summary:

```
Call:
 randomForest(formula = y ~ ., data = df, mtry = 4, importance = TRUE,      subset = training)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 9.41%
Confusion matrix:
       no  yes class.error
no  26916 1017  0.03640855
yes  1962 1752  0.52827141
```

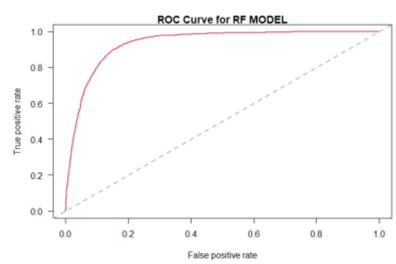
Importance Plot:



Importance Summary:

```
> importance(rf)
                 no         yes MeanDecreaseAccuracy MeanDecreaseGini
age        45.5750378   6.3938850            46.627024        571.83006
job        36.6956435  -8.6313791            29.146448        449.97318
marital     9.0657469   9.2184025            12.911625        129.85005
education  20.7809562  -0.7418906            18.236612        165.85247
default     0.1366984   5.1530924             2.464396         10.82605
balance     8.1860751   9.3222923            12.527659        618.69804
housing    45.8917223  19.6938153            51.790762        122.15869
loan        0.9796217  17.1012209            10.555471         52.39473
contact    47.9837381   6.4314322            49.631137        116.17471
day        59.3299592   3.0261672            59.122709        511.49294
month      90.1256866  25.0212682            93.797100        755.21478
duration  141.4805902 256.5719270           244.106438       1821.53806
campaign   22.4619399   8.8302175            24.187353        227.09100
pdays      22.7657733  20.6394849            25.243483        268.83332
previous   16.7728121  10.4286171            16.836818        134.28498
poutcome   33.2181352  16.5148282            48.170718        436.94353
```

ROC Curve:



Model Interpretation

The random forest model results indicated an overall error rate of 9.41% indicating the model’s a very good fit for the data. However, observing the class error rates, it can be seen that the majority class (where campaign outcome is “No,” a person did not subscribe to a plan at the bank) has a very low error rate at 3.6%, whereas the minority class (“Yes,” a person did subscribe to a plan at the bank) has an overwhelming error rate at 52.8%. This is due to the unbalanced nature of the data where a majority of the data points are for people who fall into the “No” category of the response variable y. Given that this analysis is for a marketing campaign, it is highly undesirable to have a high error rate for the “Yes” category since that is the desired outcome of the campaign. In order to overcome this unbalanced classification problem, the minority class was over-sampled, and the majority class was under-sampled to have an equal number of data points in each class. The new training data contains 7428 data points for each class (“yes” and “no” in response variable y). The following image shows the model summary for the new training data set, and the error rate for the testing data set:

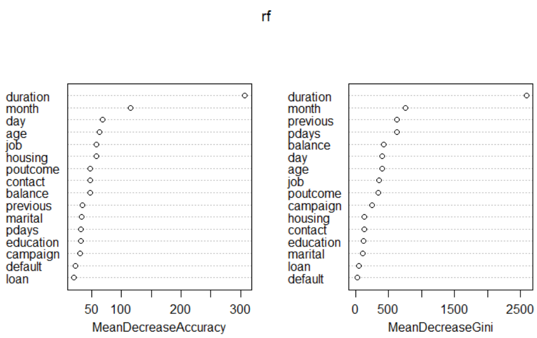
```
> print(rf)

Call:
 randomForest(formula = y ~ ., data = train_smote, mtry = 4, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 9.91%
Confusion matrix:
      no  yes class.error
no  6559  869  0.11698977
yes  603 6825  0.08117932
```

The new model now has an overall error rate of 9.91%, with class errors at 11.6% and 8.1% for classes "no," and "yes," respectively. The class error rates have improved significantly and the overall error rate is only slightly higher compared to the previous model. Overall, this model is a great fit for the data set.

Updated importance plot:



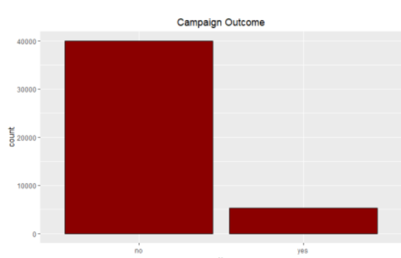
Updated importance summary:

```
> importance(rf)
                 no         yes MeanDecreaseAccuracy MeanDecreaseGini
age        62.85687  18.2477770             62.78337        402.92493
job        74.15384  -0.6968741             57.56401        354.86746
marital    37.00219   9.9652585             33.79819        100.71094
education  36.59165  -2.2824758             32.38476        120.24123
default    20.15575  12.5624657             23.25874         24.98953
balance    46.98819  20.3467273             47.04300        419.48082
housing    56.17297  11.6957091             57.42175        133.29982
loan       23.29063   7.2901424             20.80279         46.85907
contact    45.82120 -13.6873428             47.41227        128.35927
day        69.47105  16.4740567             68.31109        404.23501
month     105.05264  27.0828436            115.37815        759.96426
duration  291.74230 224.7466834            306.72011       2586.39307
campaign   28.08324  21.4354238             30.77766        250.99010
pdays      32.11528  21.1668150             32.45877        622.31080
previous   39.97299  10.1156783             34.35496        623.87866
poutcome   54.49316 -46.0377853             47.64552        346.67794
```

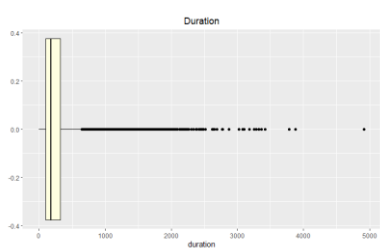
The predictors most significantly impacting the results are duration and month, the rest of them standing at similar values. In the bank marketing context, it makes logical sense to correlate the client's interest in the bank's plans with the duration of the contact during the campaign. It is interesting to note that month has a high value. A possible explanation could be that the greatest number of contacts were done in the month of May, which also could mean that most people who decided to subscribe were also contacted in the month of May. The importance plot also indicates that the predictors "age," "job," and "housing" also somewhat impacted the response variable. It is interesting to observe that the ranking provided by the importance plot changed after changing the sample training data.

Graphical representations of response variable and most important predictors:

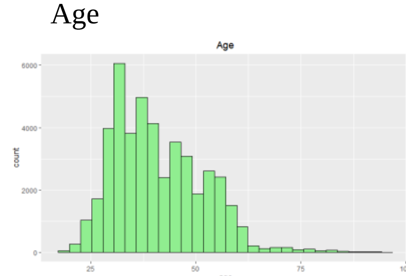
Response Variable



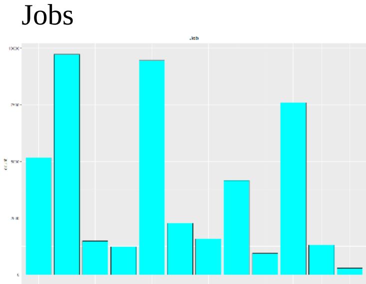
Duration



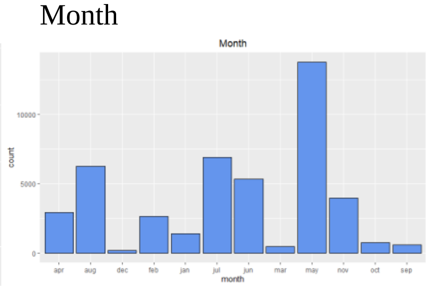
Age



Jobs



Month



Day