Bank Marketing Campaign

Kabwe Mpundu

27/08/2019

Introduction

A Portugues bank collected marketing information for its customers after a campaign to encourage the uptake of term deposits. This data is available to the public through Kaggle and was the centre of an academic paper related to data mining.

The dataset includes information on savers and non-savers all of whom were customers of the bank. In addition to this key information the following variables were included.

1. Age : The age of the customer
2. Job : A job category which may include:“admin.”,“unknown”,“unemployed”,“management”,“housemaid”,“entrepreneur”,“student”,“blue-collar”,“self-employed”,“retired”,“technician”,“services”)
3. Marital : The marital status of the customer; “married”,“divorced”,“single”; divorced also included widowed.
4. Education: The level of education of the customer: “unknown”,“secondary”,“primary”,“tertiary”)
5. Default:Did the client have any credit in default? : “yes”,“no”
6. Balance: The average yearly balance, in euros (numeric)
7. Housing: Did the client have a housing loan? : “yes”,“no”
8. Loan: Did the client have a personal loan? :“yes”,“no”

The following were related with the last contact of the marketing campaign:

1. Contact: The type of contact communication type: “unknown”,“telephone”,“cellular”
2. Day: The last contact day of the month (numeric)
3. Month: The last contact month of year “jan”,…, “dec”
4. Duration: last contact duration, in seconds (numeric)
5. Campaign: The number of contacts performed during this campaign and for this client (numeric, includes last contact)
6. Pdays: The number of days between contact from a previous campaigns (numeric, -1 means client was not previously contacted)
7. Previous: The number of contacts performed before the campaign and for this client.
8. Poutcome: The outcome of the previous marketing campaign:“unknown”,“other”,“failure”,“success”.

The target variable was the clients who subscribed for a term deposit as a result of the marketing campaign.

To solve this problem we shall begin by processing the data to ensure the it is clean and compatitable for Exploration, Visualisation, Training and Testing. The Methods Section details how the dataset was downloaded and how it was transformed to have gained insights into customer saving. It has also shown which Machine Learning tools were used.

Methods and Analysis

The dataset was first downloaded with the required packages, the following section of the r script demonstrates a part of the process:

```
dl <- tempfile()
download.file("https://archive.ics.uci.edu/ml/machine-learning-databases/00222/bank.zip", dl)


bank <- fread(text = gsub(";", "\t", readLines(unzip(dl, "bank.csv"))),
              header=TRUE)
set.seed(1)
ind <- createDataPartition(bank$y,p=0.1, list = FALSE)
bank_test <- bank[ind,]
bank_train <- bank[-ind,]
```

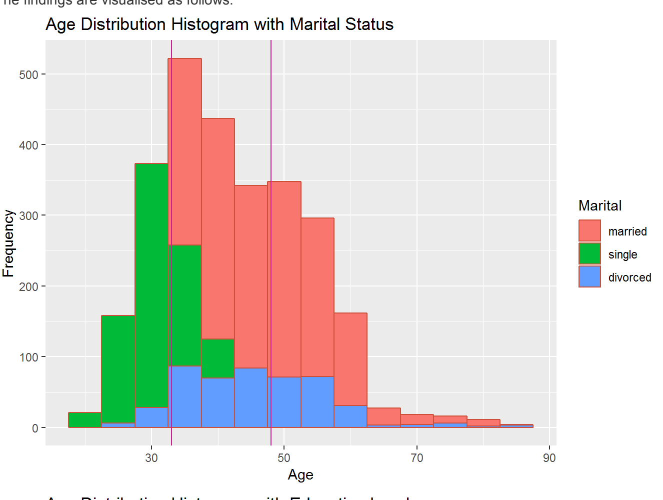
```
#We ensure the colnames are Capitalised.
colnames(bank_train) <- str_to_title(colnames(bank_train))
colnames(bank_test) <- str_to_title(colnames(bank_test))

#We then ensure the factor variables as stored as factors
bank_test <- bank_test %>%
  mutate_at(c('Job','Marital','Education','Default','Housing','Loan','Contact','Month','Poutcome','Y'),
          as_factor)
bank_train <- bank_train %>%
  mutate_at(c('Job','Marital','Education','Default','Housing','Loan','Contact','Month','Poutcome','Y'),
          as_factor)
```

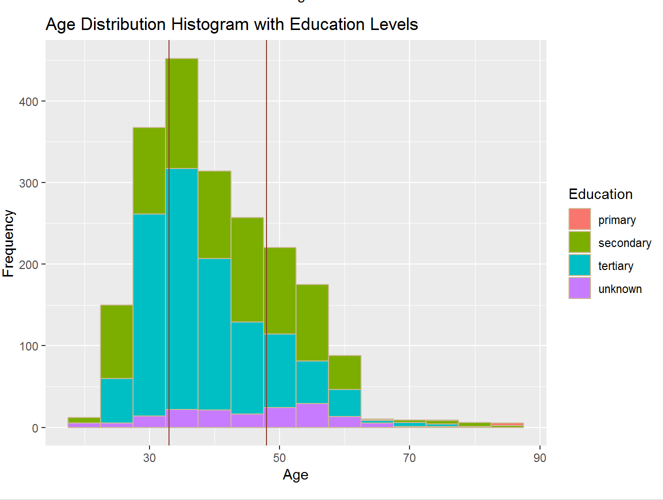
The data has 17 variables and over a 45, 211 observations for the train set and 4, 521 for the test set. We then ensure then code the following variables as categorical. A summary of the data shows the distributions of the variables we find the some variables will be more useful than others at informing our decisions due to the qualities they present such as spread and distribution.

```
##       Age                    Job           Marital          Education
##  Min.   :19.00   management :883   married :2529   primary  : 606
##  1st Qu.:33.00   blue-collar:845   single  :1072   secondary:2069
##  Median :39.00   technician :694   divorced: 467   tertiary :1236
##  Mean   :41.24   admin.     :430                   unknown  : 157
##  3rd Qu.:49.00   services   :365
##  Max.   :87.00   retired    :209
##                  (Other)    :642
##  Default        Balance       Housing      Loan           Contact
##  no :3999   Min.   :-3313   no :1762   no :3444   cellular :2608
##  yes:  69   1st Qu.:   68   yes:2306   yes: 624   unknown  :1185
##             Median :  440                         telephone: 275
##             Mean   : 1416
##             3rd Qu.: 1464
##             Max.   :71188
##
##       Day             Month          Duration          Campaign
##  Min.   : 1.00   may    :1271   Min.   :   4.0   Min.   : 1.000
##  1st Qu.: 9.00   jul    : 642   1st Qu.: 104.0   1st Qu.: 1.000
##  Median :16.00   aug    : 574   Median : 186.0   Median : 2.000
##  Mean   :15.97   jun    : 464   Mean   : 263.2   Mean   : 2.773
##  3rd Qu.:21.00   nov    : 336   3rd Qu.: 330.0   3rd Qu.: 3.000
##  Max.   :31.00   apr    : 268   Max.   :2769.0   Max.   :50.000
##                  (Other): 513
##      Pdays           Previous           Poutcome       Y
##  Min.   : -1.00   Min.   : 0.0000   unknown:3330   no :3600
##  1st Qu.: -1.00   1st Qu.: 0.0000   failure: 436   yes: 468
##  Median : -1.00   Median : 0.0000   other  : 180
##  Mean   : 39.79   Mean   : 0.5455   success: 122
##  3rd Qu.: -1.00   3rd Qu.: 0.0000
##  Max.   :808.00   Max.   :25.0000
##
```

The analysis above shows most customers were contacted in May, they had low average balances, were married and between the ages 33 and 48. It also showed most clients had taken a pay_day loan, It appears the marketing campaign was mostly targeted at struglling low income families. This was supported by the fact that the majority of our population had no tertiary education. It was also noted most customers were contacted by phone and the quality of the information was poor given, with many blanks or irrelvant varibles.

The findings are visualised as follows:



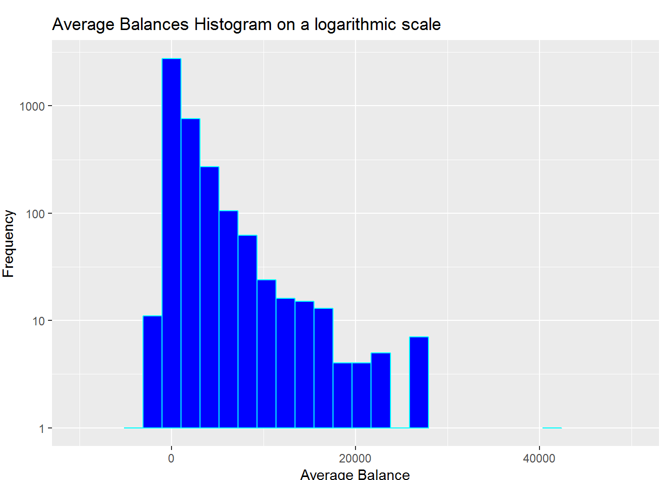



```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 1 rows containing non-finite values (stat_bin).
```

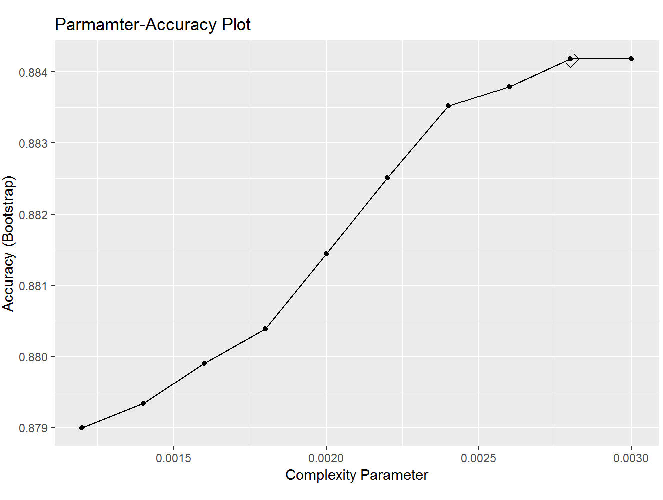
```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Removed 13 rows containing missing values (geom_bar).
```



The data required further analysis for modeling, a decision tree was trained on to the data to highlight the key determinants for saving. # Decision Tree

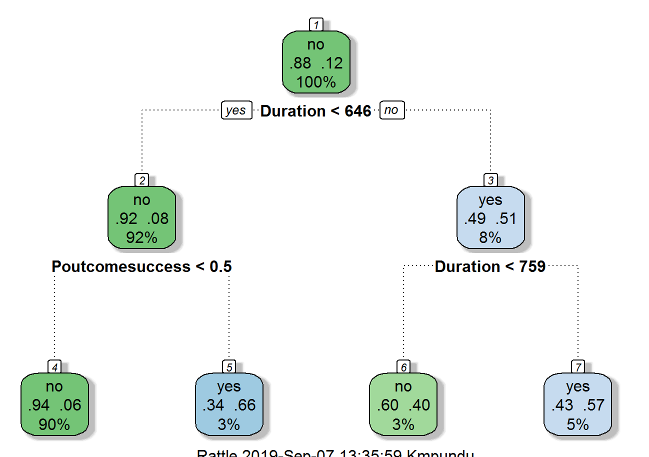
```
tree_fit <- train(Y~.,
                  data=bank_train,
                  method='rpart',
                  tuneGrid=expand.grid(cp=seq(0.0012,0.003,0.0002)))
tree_Y <- predict(tree_fit, bank_test)
ggplot(tree_fit, highlight = TRUE)+
  labs(title='Parmamter-Accuracy Plot')
```




```
#The decision tree analysis
#fancyRpartPlot(tree_fit$finalModel)
```

The decision tree had many branches, but would acheived an accruacy over 90%, for the purposes of illustration the less significant branches were chopped off. The tree below illustrated the most important discriminants.

```
#Showing a simpler tree
fancyRpartPlot(prune(tree_fit$finalModel, cp=0.02))
```

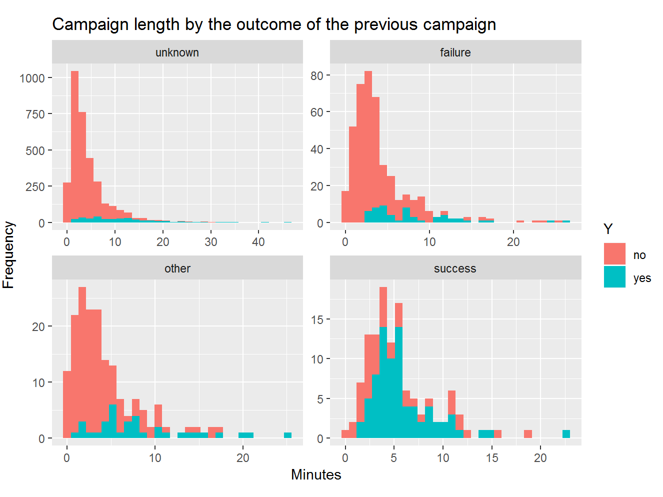


Rattle 2019-Sep-07 13:35:59 Kmpundu

The decision tree showed the key discriminants were the outcomes of previous marketing campaigns and duration of the current marketing campaign.

Further explorations revealed the following about succesful outcomes;

1. Customers who subscribed previously were more likely to subscribe again.
2. Short campaings were not effective, escpecially under three minutes.
3. Campaigns under 15 minutes long were most effective



Besides the decision tree the Generalised Linear Model (GLM) was used to predict the marketing outcomes. It was found to have a high accuracy and specificity, unfortunately the sensitivity of the model was below 50%. The similar obsevations were noted with KNN neighbours and the decision tree analysis

```
#The GLM method is used for training
glm_fit <- train(Y~.,
                 method='glm',
                 data=bank_train)

glm_Y <- predict(glm_fit,bank_test)
```

```
#A knn model was also trained and fit
knn_fit <- train(Y~.,
                method='knn',
                data=bank_train)

knn_Y <- predict(knn_fit,bank_test)
```

```
####After the chosen models were fitted we analysed theire confusion matices.
tree_cm_test <- confusionMatrix(data = tree_Y, reference= bank_test$Y)
```

```
## Warning in confusionMatrix.default(data = tree_Y, reference = bank_test$Y):
## Levels are not in the same order for reference and data. Refactoring data
## to match.
```

```
tree_cm_test
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction yes  no
##        yes  20  18
##        no   33 382
##
##                Accuracy : 0.8874
##                  95% CI : (0.8546, 0.915)
##     No Information Rate : 0.883
##     P-Value [Acc > NIR] : 0.42021
##
##                   Kappa : 0.3789
##
##  Mcnemar's Test P-Value : 0.04995
##
##             Sensitivity : 0.37736
##             Specificity : 0.95500
##          Pos Pred Value : 0.52632
##          Neg Pred Value : 0.92048
##              Prevalence : 0.11700
##          Detection Rate : 0.04415
##    Detection Prevalence : 0.08389
##       Balanced Accuracy : 0.66618
##
##        'Positive' Class : yes
##
```

```
glm_cm_test <- confusionMatrix(data = glm_Y, reference= bank_test$Y)
```

```
## Warning in confusionMatrix.default(data = glm_Y, reference = bank_test$Y):
## Levels are not in the same order for reference and data. Refactoring data
## to match.
```

```
glm_cm_test
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction yes  no
##        yes  13  11
##        no   40 389
##
##                Accuracy : 0.8874
##                  95% CI : (0.8546, 0.915)
##     No Information Rate : 0.883
##     P-Value [Acc > NIR] : 0.4202
##
##                   Kappa : 0.2856
##
##  Mcnemar's Test P-Value : 8.826e-05
##
##             Sensitivity : 0.24528
##             Specificity : 0.97250
##          Pos Pred Value : 0.54167
##          Neg Pred Value : 0.90676
##              Prevalence : 0.11700
##          Detection Rate : 0.02870
##    Detection Prevalence : 0.05298
##       Balanced Accuracy : 0.60889
##
##        'Positive' Class : yes
##
```

```
knn_cm_test <- confusionMatrix(data = knn_Y, reference= bank_test$Y)
```

```
## Warning in confusionMatrix.default(data = knn_Y, reference = bank_test$Y):
## Levels are not in the same order for reference and data. Refactoring data
## to match.
```

```
knn_cm_test
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction yes  no
##        yes  10  18
##        no   43 382
##
##                Accuracy : 0.8653
##                  95% CI : (0.8304, 0.8954)
##     No Information Rate : 0.883
##     P-Value [Acc > NIR] : 0.89133
##
##                   Kappa : 0.1806
##
##  Mcnemar's Test P-Value : 0.00212
##
##             Sensitivity : 0.18868
##             Specificity : 0.95500
##          Pos Pred Value : 0.35714
##          Neg Pred Value : 0.89882
##              Prevalence : 0.11700
##          Detection Rate : 0.02208
##    Detection Prevalence : 0.06181
##       Balanced Accuracy : 0.57184
##
##        'Positive' Class : yes
##
```

```
# Reviews of these models showed low specificity, which in  the banks case is not helpful to anyone.
```

The models above add little value to the bank as they would turn away most prospective valuable candidates, and also a significant proportion of people predicted to save had been wronlgy classified. That is there are many true negatives and false positivies.

Results

The sensitivity of the models varied ranging from 15% to 50%, thus the models above add little value to the bank as they would turn away most prospective valuable candidates, and also a significant proportion of people predicted to save had been wrongly classified. That is there are many true negatives and false positives.

Conclusion

Because of the low prevalence about 10%, the accuracy was high at around 90%, but the sensitivity was low, ranging between 10% and 60%. The predictions from K-Nearest Neighbours had the lowest sensitivity at around 15% ,the Generalised Linear Model followed around 24% percent,and the most helpful is the decision tree with a sensitivity around 38%.

Nonetheless, The test still showed Previous Customers were very likely to subscribe, it also showed the time spent on a customer had a non-linear relationship with savings. The customers age also played a factor with young and elderly people saving more than their middle aged counterparts.

Recommendations

1. To target younger and older people in future campaigns.
2. To build a polynomial regression model to predict marketing outcomes. 3, Collect more statistics, as many variables reported blanks.
3. Conducting the campaigns in March, September and December as there was a higher uptake during these months, and the current campaign was mostly conducted in May.
4. Given the low prevalence, and the lack of relevant variables the Decision tree should not be used, unless corrected for prevalence.