Final Project - Bank Marketing

Team Members:
Ulana Salyk
Jacqueline Chung
William Chinnery
Timothy Nguyen

1 Introduction

In this section you describe the project and the objectives and what you will present in this report.

The problem is related to direct marketing campaigns (phone calls) of a Portuguese banking institution. The **classification goal** is to predict whether the client subscribes a term deposit or not. The target class is the last attribute (*subscribed*) and has two values (*yes* and *no*).
The training set (trainset.csv) contains 3,196 subscribed and 26,076 unsubscribed records. The test set (testSet.csv) contains 1,444 subscribed and 1,047 unsubscribed records.
Attribute Information:
1. age (numeric)
2. **job**: type of job (categorical: 'admin.','blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired','self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')
3. **marital**: marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
4. **education**: (categorical: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')
5. **housing**: has housing loan? (categorical: 'no','yes','unknown')
6. **loan**: has personal loan? (categorical: 'no','yes','unknown')
7. **contact**: contact communication type (categorical: 'cellular','telephone')
8. month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
9. **day_of_week**: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
10. **duration**: last contact duration, in seconds (numeric).
11. **campaign**: number of contacts performed during this campaign and for this client (numeric, includes last contact)
12. **pdays**: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
13. **poutcome**: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')
14. **nr.employed**: number of employees - quarterly indicator (numeric)
15. **Target Attribute**: Subscribed - has the client subscribed a term deposit? (binary: 'yes','no')

Steps
The project involves the following steps:
1. **Data exploration**: try to know data and represents statistics for the important features among the features and the target attribute.
2. **Preprocessing the data**. The goal of this step is to extract features from records in the training set and use these features to test data sets. Note that the data have "**unknown**" values that need to be cleaned.
3. **Use a classification-learning method** provided by **R** to learn a model from the set of training examples. You can use any of the classification methods (decision tree, KNN, …) for this purpose.
4. **Test the learned model** on the test set and report the testing results.

The main objective of this project is to, as accurately as possible, predict whether or not clients will subscribe or not to a long-term deposit. A Portuguese banking institution is operating a direct marketing campaign through phone calls. Based on some attributes, the goal is to predict the likelihood of a client subscribing to the deposit.

This report will present our findings from data exploration (attributes with the most information gain), different

learning methods we have applied, and an evaluation of the performance of each learning method.

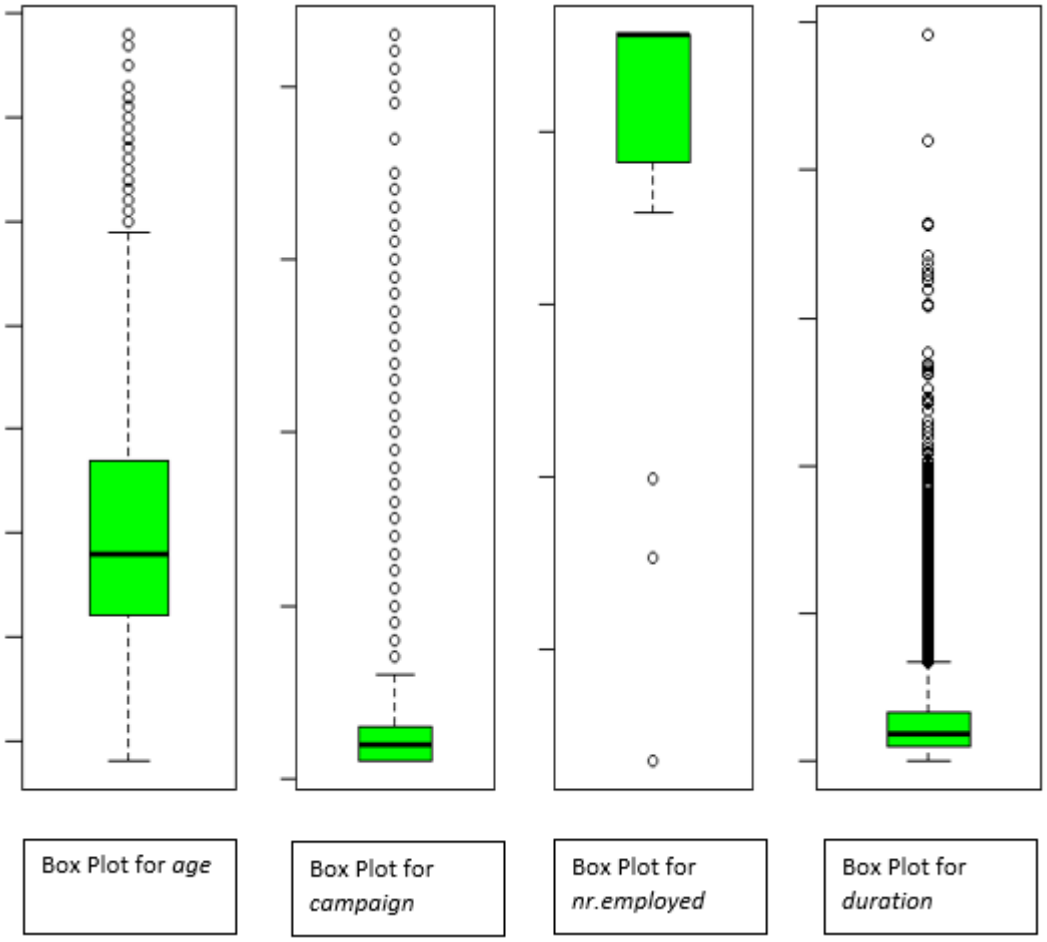
2 Data Exploration

Describe different information you found during data exploring (e.g., which attributes had higher information gain)

During our preliminary data exploration, we quickly found that the attribute “pdays” and “poutcome” do not provide us with any useful information. This is because pdays shows a majority of “999”, which means the client was not previously contacted, whereas poutcome shows a majority of “nonexistent”, which is considered irrelevant information.

During our data cleaning phase, we were able to remove about 7% of data. The rows that were removed contained values that were “unknown”. After the removal of these rows, we notice that there was not a lot of change that happened in terms of the accuracy.

We later performed a data preprocessing, including data cleaning and reduction. Using a series of Boxplots, we were able to identify the major outliers of numerical attributes. So, we removed outliers in the data set by looking for them in the boxplot. The image below shows the box plot for each of the numeric attributes:



Box Plot for *age*

Box Plot for *campaign*

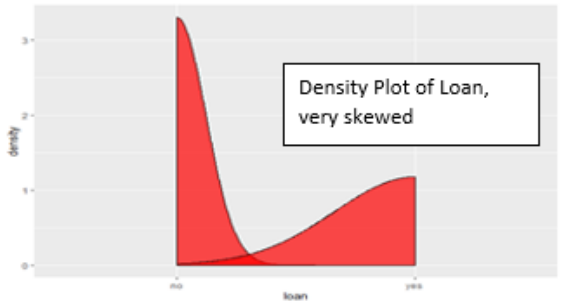
Box Plot for *nr.employed*

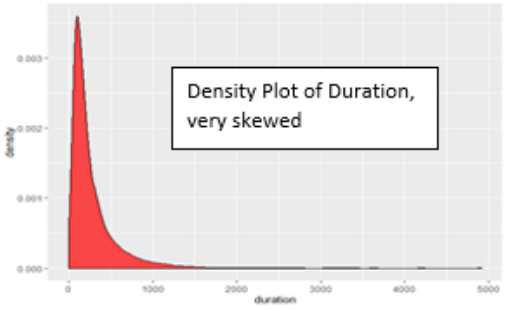
Box Plot for *duration*

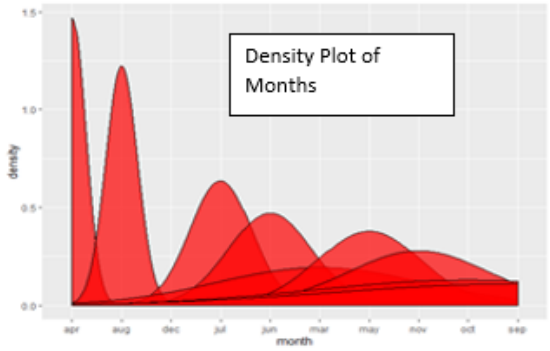
Using a density distribution chart, we looked at the density of each value for each attribute to see if data normalization is required. We can observe that many of the values are askew on the distributed charts, so we will normalize that data later on. From each of the attributes we were able to remove: values greater than “4000” from *duratio*n, values greater than “56” from *campaign*, values less than “5050” from *nr.employed*, delete “illiterate” from *education*, and removed “dec” and "sep" from *month.*

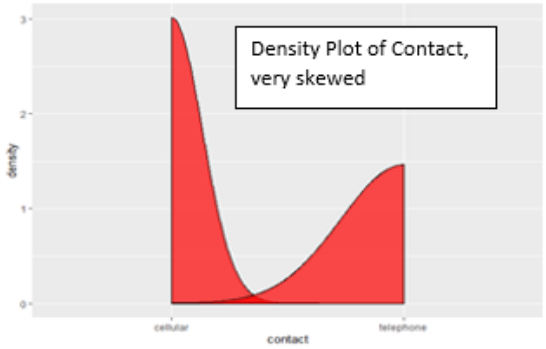
The images below show the values of each attribute and is represented by their density plots. We can see from density plots that most of the data is skewed. We used the density plot to check if the response variable is close to normal and to see sentimental values.

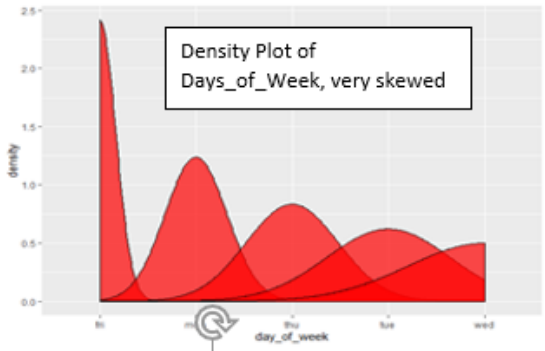
```
> summary(trainData$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   32.00   38.00   39.94   47.00   88.00
> summary(trainData$duration) # Remove >4000 (max), then 3000, 2000, 1000, 500
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0   103.0   179.0   266.9   328.0  4918.0
> summary(trainData$campaign) # Remove 56 (max), then >40, 30, 20, 10
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   2.000   2.735   3.000  43.000
> summary(trainData$nr.employed) # Remove < 5050, 5100, 5150
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5018    5191    5228    5205    5228    5228
> summary(trainData$job)
       admin.   blue-collar  entrepreneur     housemaid    management       retired
         6968          6157          1041           796          1974           866
self-employed      services       student    technician    unemployed       unknown
          983          2645           256          4827           665             0
> summary(trainData$marital)
divorced  married   single  unknown
    3142    17067     6969        0
> summary(trainData$education) # delete illiterate
           basic.4y            basic.6y            basic.9y         high.school
               2941                1614                4175                6365
         illiterate professional.course   university.degree             unknown
                 15                3724                8344                   0
> summary(trainData$housing)
     no unknown     yes
  12997       0   14181
> summary(trainData$loan)
     no unknown     yes
  22974       0    4204
> summary(trainData$contact)
 cellular telephone
    14558     12620
> summary(trainData$month) # delete dec and sep
 apr  aug  dec  jul  jun  mar  may  nov  oct  sep
 401 5163    1 6249 4227  122 7604 3198  121   92
> summary(trainData$day_of_week)
 fri  mon  thu  tue  wed
4828 5430 5705 5556 5659
```

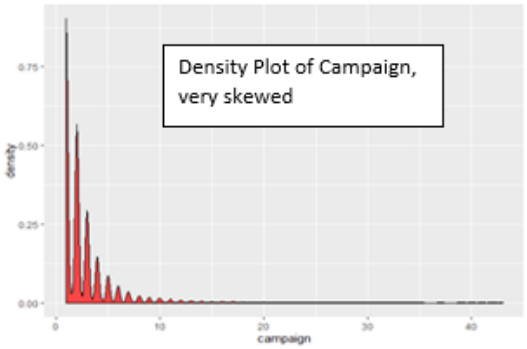

Density Plot of Loan, very skewed
density
loan
no
yes

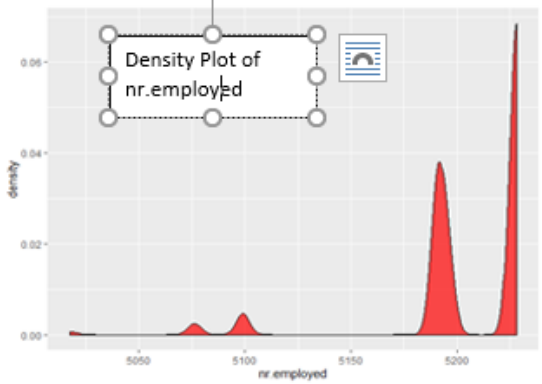

Density Plot of Duration, very skewed
density
duration


Density Plot of Months
density
month
apr
aug
dec
jul
jun
mar
may
nov
oct
sep

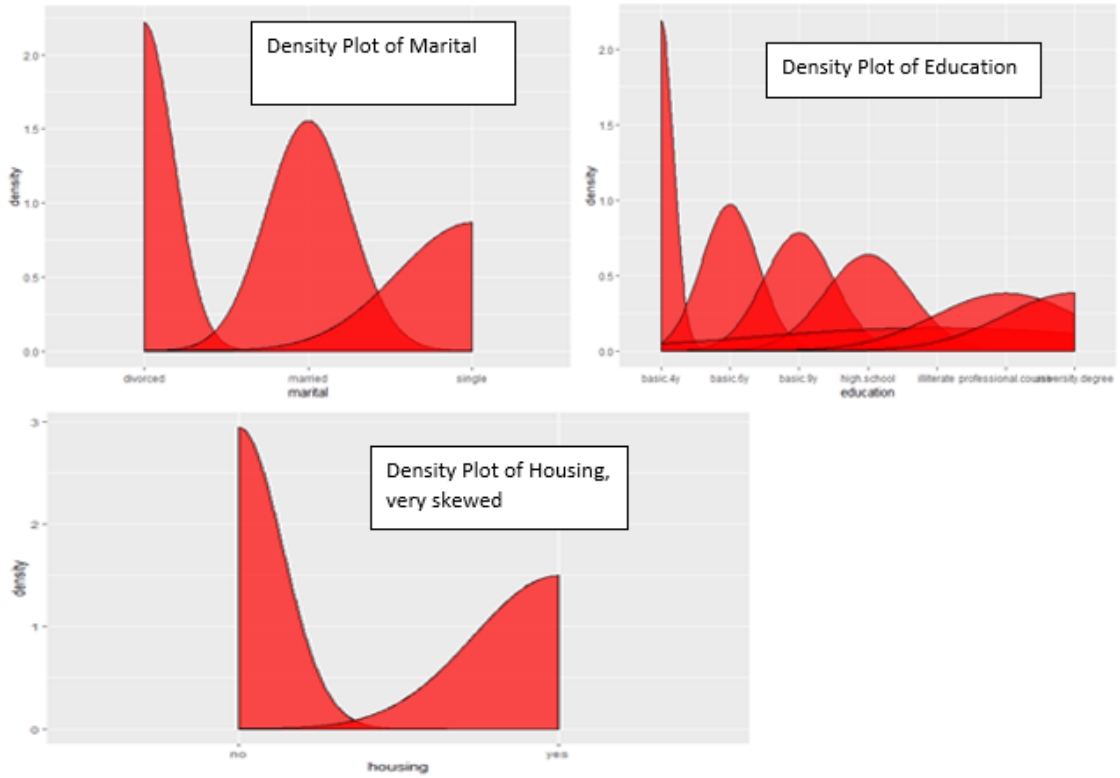

Density Plot of Contact, very skewed
density
contact
cellular
telephone


Density Plot of Days_of_Week, very skewed
density
day_of_week
thu
tue
wed


Density Plot of Campaign, very skewed
density
campaign


Density Plot of nr.employed
density
nr.employed
5050
5100
5150
5200


Density Plot of Job
density
job
admin.
student
unemployed



After cleaning the data, we calculated the information gain for each attribute. Below is a table showing the attributes with the highest information gain, top 3, 5, and 10 respectively:

Top # Attributes	**Attributes (including *nr.employed)***
Top 3	*nr.employed, duration, month*
Top 5	*nr.employed, duration, month, contact, age*
Top 10	*nr.employed, duration, month, contact, age, job, campaign, marital, education, housing*

3 Learning Methods

In this section you present the method(s) you used and why you used them.

a) Decision Tree

Decision Trees are easy to implement, use and interpret. Since this data set has many attributes, using a decision tree can help identify non-linear relationships. As a result, we made several decision trees, including rpart, CTree, and J48Tree.

b) kNN

The kNN (K-Nearest Neighbour) algorithm identifies the 'k' nearest neighbors by operating on the principle that similar things are close together. The algorithm uses this idea of proximity or closeness and calculates the distance between data points. In our use of kNN, after the package RWeka is installed, we first indicate the variables that we will be predicting, starting with the top 3 attributes with highest information gain (*nr.employed, duration,* and *month),* we also set our classifications using the parameter "CL" which is used to contain the categories of our response variables that belong to the training data set (trainset.csv).

Between the decision tree method and kNN method, we found that decision tree gave us a better result.

4 Evaluation

All the performance measures you used should be described first and then draw figures that represent results and discuss the results and compare the methods.

Most of our performance measures were based off of precision, recall and accuracy tests.

Below shows the results for our best decision tree method (which includes month and duration):

```
> cmRPart1 <- confusionMatrix(table_mat)
> cmRPart1$overall['Accuracy'] # Accuracy: 66.74%
 Accuracy
0.6822187
> # Calculate Precision, Recall, and F-Score
> # It is reversed for this dataset
> # Precision: tn/(tn + fn):
> prec <- table_mat[1,1]/sum(table_mat[1,1:2])
> prec
[1] 0.7267259
> # Recall: tn/(tn + fp):
> recall <- table_mat[1,1]/sum(table_mat)
> recall
[1] 0.6386674
> # F-Score: 2 * precision * recall /(precision + recall):
> f_score <- 2 * prec * recall / (prec + recall)
> f_score
[1] 0.6798571
```

Compared to KNN, we can see that Decision Tree gives us the best accuracy.

5 Discussion

Any discussion based on your evaluation and methods should be presented here.

After running through the different learning methods, we believe that the decision tree has the accuracy compared to the other method we ran. we notice that as we increase the amount of attributes from Top 3 to Top 5 and then from Top 5 to Top 10, the accuracy of the decision tree for training data continued to increase slightly however the accuracy of the test data began to decrease.

After our preliminary data processing we wanted to test the preliminary accuracy of different trees. We calculated the accuracy of trees with attributes of: *duration* + *month* + *contact* + *age* + *job* + *campaign* + *marital* + *education* + *housing* (1st tree); *duration* + *month* + *contact* + *age* + *job* + *campaign* + *marital* + *education* (2nd tree); *duration* + *month* + *contact* + *age* + *job* + *campaign* + *marital* (3rd tree); *duration* + *month* + *contact* + *age* + *job* + *campaign* (4th tree); *duration* + *month* + *contact* + *age* + *job* (5th); *duration* + *month* + *contact* + *age* (6th tree); *duration* + *month* + *contact* (7th tree); *duration* + *month* + *age* (8th tree)*; nr.employed* + *duration* + *month* + *age* + *job* + *campaign* + *marital* + *education* + *housing* (9th tree); *duration* + *month* + *age* + *job* + *campaign* + *marital* + *education* + *housing* (10th tree); *duration* + *month* + *age* + *job* + *campaign* + *marital* + *education* (11th tree); *duration* + *month* + *age* + *job* + *campaign* (12th tree)*; duration* + *month* + *age* + *job* (13th tree); *duration* + *month* + *job* (14th tree); *duration* + *month* + *job* (15th tree)

From the 1st to the 6th tree, the accuracy remained the same at 23.27%. The 7th tree has an increase of accuracy to 23.87%. The 8th tree had an increase to 65.98%. With the addition of *nr.employed* in the 9th tree, the accuracy significantly dropped to 14.28%. From the 10th tree to the 13th tree, the accuracy stayed at 65.98%. With the removal of *age* in the 14th tree, the accuracy with up slightly to 66.74%. Although the increase isn't much, it tells us that *housing, campaign, age, marital*, and *education* is not as important.

We then decided to create rpart trees and pruned the branches. The 1st pruned tree gave an accuracy of 87.88%, the 2nd pruned tree was 67.28%. After extra pruning, the 3rd and 4th tree gave a 14.28% accuracy. Although the first pruned tree gave us the highest accuracy of 87.88%, it was not an effective tree because there is only the root node remaining on that tree.

Next, we examined CTrees and J48 Trees. From the 1st Ctree and J48 Tree we created from the model, we got accuracy of 95.97% and 97.06% respectively from testing the training data set. However, when testing these trees from the test data set, the accuracy significantly dropped to 14.36% and 14.31% respectively. This tells us that there is definitely overfitting in the models. This means that these decision trees are not appropriate to use. The table below shows the accuracy of each

type of tree with each amount of the Top number of Attributes.

# of Top Attributes	CTrees Accuracy %	J48 Trees Accuracy %
Top 3	14.36%	14.28%
Top 5	14.37%	14.28%
Top 10	14.36%	14.29%

After our secondary data cleaning phase, we recreated the trees and noticed an overall increase in the accuracy. The rpart tree that contained *duration, month,* and *job* had an accuracy of 68.22%. To demonstrate the cleaning the data created an increase in accuracy, we create a tree that included "sep" from *month* and noticed it dropped to 66.79%. Using the Top 3 attributes, we built a CTree and J48 Tree and received an increase of accuracy of 68.10% and 68.25% respectively.

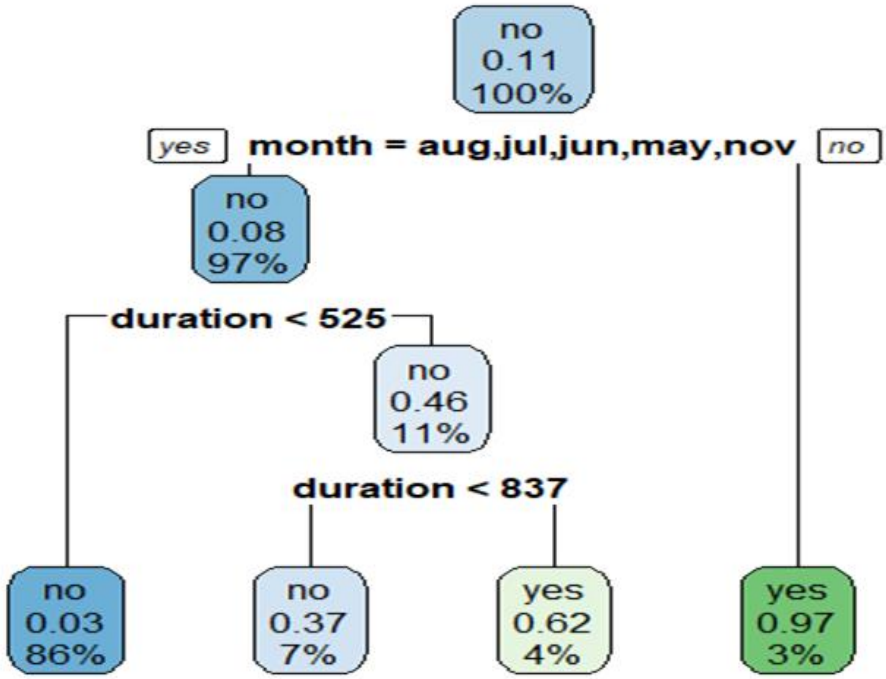
We then continued to make rpart trees. To enhance the accuracy of the tree, we had to prune the branches of the trees. As expected, the 9th rpart tree that was created provided us with the highest accuracy.

Rtree	Attributes	Accuracy
1st tree	*duration + month + age + campaign + job + marital + education + loan + contact + day_of_week*	24.45%
2nd tree	*duration + month + age + campaign + job + marital + education + loan + contact*	24.45%
3rd tree	*duration + month + age + campaign + job + marital + education + loan*	67.09%
4th tree	*duration + month + age + campaign + job + marital + education*	67.09%
5th tree	*duration + month + age + campaign + job + marital*	67.09%
6th tree	*duration + month + age + campaign + job*	67.09%
7th tree	*duration + month + age + campaign*	67.09%
8th tree	*duration + month + age*	67.09%
9th tree	*duration + month*	68.22%

In addition, we conducted some preprocessing techniques, such as standardization, normalization, range, etc. this helped boost up our accuracy for our decision trees.

After our preliminary conclusion that the model of the 9th tree gave us the highest accuracy, we tested it by using the undersampling method. We later calculated the precision, recall and F-score. Below shows the results for the evaluation.

```
> cmRPart1 <- confusionMatrix(table_mat)
> cmRPart1$overall['Accuracy'] # Accuracy: 66.74%
 Accuracy
0.6822187
> # Calculate Precision, Recall, and F-Score
> # It is reversed for this dataset
> # Precision: tn/(tn + fn):
> prec <- table_mat[1,1]/sum(table_mat[1,1:2])
> prec
[1] 0.7267259
> # Recall: tn/(tn + fp):
> recall <- table_mat[1,1]/sum(table_mat)
> recall
[1] 0.6386674
> # F-Score: 2 * precision * recall /(precision + recall):
> f_score <- 2 * prec * recall / (prec + recall)
> f_score
[1] 0.6798571
```



For our kNN method, we set our k = 1 and k = 5. We were not able to find the accuracy for this method. Below are the results for our kNN method:

```
> summary(trainDataLessOut2)
      age                  job             marital                   education        housing          loan           contact            month         day_of_week     duration         campaign
 Min.   :18.00   admin.       :6932   divorced: 3133   university.degree  :8305   no     :12950   no     :22883   cellular :14463   may    :7602   fri:4804    Min.   :   0.0   Min.   : 1.000
 1st Qu.:32.00   blue-collar  :6148   married :16990   high.school        :6345   unknown:    0   unknown:    0   telephone:12607   jul    :6244   mon:5418    1st Qu.: 103.0   1st Qu.: 1.000
 Median :38.00   technician   :4815   single  : 6947   basic.9y           :4173   yes    :14120   yes    : 4187                     aug    :5159   thu:5691    Median : 179.0   Median : 2.000
 Mean   :39.92   services     :2640   unknown :    0   professional.course:3710                                                     jun    :4227   tue:5516    Mean   : 266.6   Mean   : 2.739
 3rd Qu.:47.00   management   :1968                    basic.4y           :2928                                                     nov    :3195   wed:5641    3rd Qu.: 327.0   3rd Qu.: 3.000
 Max.   :88.00   entrepreneur:1036                     basic.6y           :1609                                                     apr    : 400               Max.   :4918.0   Max.   :43.000
                 (Other)      :3531                    (Other)            :   0                                                     (Other): 243
     pdays            poutcome       nr.employed    Subscribed   age_imp          job_imp          marital_imp      education_imp    housing_imp      loan_imp         contact_imp      month_imp
 Min.   :  0.0   failure    :  885   Min.   :5018   no :24197   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 1st Qu.:999.0   nonexistent:25948   1st Qu.:5191   yes: 2873   FALSE:27070     FALSE:27070     FALSE:27070     FALSE:27070     FALSE:27070     FALSE:27070     FALSE:27070     FALSE:27070
 Median :999.0   success    :  237   Median :5228
 Mean   :989.7                       Mean   :5206
 3rd Qu.:999.0                       3rd Qu.:5228
 Max.   :999.0                       Max.   :5228

 day_of_week_imp duration_imp    campaign_imp    pdays_imp       poutcome_imp    nr.employed_imp Subscribed_imp
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
 FALSE:27070     FALSE:27070     FALSE:27070     FALSE:27070     FALSE:27070     FALSE:27070     FALSE:27070
```

```
> head(trainDataLessOut2)
  age          job  marital           education housing loan   contact month day_of_week duration campaign pdays    poutcome nr.employed Subscribed age_imp job_imp marital_imp education_imp
1  41  blue-collar divorced            basic.4y     yes   no telephone   may         mon     1575        1   999 nonexistent        5191        yes   FALSE   FALSE       FALSE         FALSE
2  49 entrepreneur  married   university.degree     yes   no telephone   may         mon     1042        1   999 nonexistent        5191        yes   FALSE   FALSE       FALSE         FALSE
3  49   technician  married            basic.9y      no   no telephone   may         mon     1467        1   999 nonexistent        5191        yes   FALSE   FALSE       FALSE         FALSE
4  41   technician  married professional.course     yes   no telephone   may         mon      579        1   999 nonexistent        5191        yes   FALSE   FALSE       FALSE         FALSE
5  45  blue-collar  married            basic.9y     yes   no telephone   may         mon      461        1   999 nonexistent        5191        yes   FALSE   FALSE       FALSE         FALSE
6  42  blue-collar  married            basic.9y     yes  yes telephone   may         mon      673        2   999 nonexistent        5191        yes   FALSE   FALSE       FALSE         FALSE
  housing_imp loan_imp contact_imp month_imp day_of_week_imp duration_imp campaign_imp pdays_imp poutcome_imp nr.employed_imp Subscribed_imp
1       FALSE    FALSE       FALSE     FALSE           FALSE        FALSE        FALSE     FALSE        FALSE           FALSE          FALSE
2       FALSE    FALSE       FALSE     FALSE           FALSE        FALSE        FALSE     FALSE        FALSE           FALSE          FALSE
3       FALSE    FALSE       FALSE     FALSE           FALSE        FALSE        FALSE     FALSE        FALSE           FALSE          FALSE
4       FALSE    FALSE       FALSE     FALSE           FALSE        FALSE        FALSE     FALSE        FALSE           FALSE          FALSE
5       FALSE    FALSE       FALSE     FALSE           FALSE        FALSE        FALSE     FALSE        FALSE           FALSE          FALSE
6       FALSE    FALSE       FALSE     FALSE           FALSE        FALSE        FALSE     FALSE        FALSE           FALSE          FALSE
> trainDataLessOut2 <- subset(trainDataLessOut2, select = duration:age)
> head(trainDataLessOut2)
  duration day_of_week month   contact loan housing           education  marital          job age
1     1575         mon   may telephone   no     yes            basic.4y divorced  blue-collar  41
2     1042         mon   may telephone   no     yes   university.degree  married entrepreneur  49
3     1467         mon   may telephone   no      no            basic.9y  married   technician  49
4      579         mon   may telephone   no     yes professional.course  married   technician  41
5      461         mon   may telephone   no     yes            basic.9y  married  blue-collar  45
6      673         mon   may telephone  yes     yes            basic.9y  married  blue-collar  42
```

6 Conclusion

A summary of what you have done in this report and the main results you obtained.

This report has shown the process we have taken to clean the data, apply learning methods and provides a discussion based on the analysis on the models we used. We cleaned the data set by removing outliers with the use of box plots and normalized the data by using density plots. We applied two learning methods: decision tree and kNN. The accuracy of the results we retrieved from the best decision tree was from the rpart tree we created.

References

https://www.r-bloggers.com/cheat-sheet-for-prediction-and-classification-models-in-r
KNN: https://stat.ethz.ch/R-manual/R-devel/library/class/html/knn.html
Decision Tree: https://www.r-bloggers.com/using-decision-trees-to-predict-infant-birth-weights
Data preprocessing: https://www.r-bloggers.com/preparing-the-data-for-modelling-with-r
https://www.rdocumentation.org/packages/DMwR/versions/0.4.1/topics/kNN
https://cran.r-project.org/web/packages/FNN/FNN.pdf
https://cran.r-project.org/web/packages/jtools/vignettes/summ.html
https://www.datacamp.com/community/tutorials/linear-regression-R
https://www.dataquest.io/blog/statistical-learning-for-predictive-modeling-r/
https://rpubs.com/cyobero/regression-tree
https://tutorials.iq.harvard.edu/R/Rstatistics/Rstatistics.html
http://r-statistics.co/Linear-Regression.html
https://www.machinelearningplus.com/machine-learning/complete-introduction-linear-regression-r/
https://rpubs.com/sidTyson92/329310
https://topepo.github.io/caret/pre-processing.html
https://www.machinelearningplus.com/machine-learning/complete-introduction-linear-regression-r/
http://www.sthda.com/english/articles/38-regression-model-validation/158-regression-model-accuracy-metrics-r-square-aic-bic-cp-and-more/
https://medium.com/@rishabhjain_22692/decision-trees-it-begins-here-93ff54ef134
https://www.r-bloggers.com/classification-trees-using-the-rpart-function/
https://www.rdocumentation.org/packages/caret/versions/3.13/topics/preProcess

https://www.kaggle.com/jhuno137/classification-tree-using-rpart-100-accuracy