# CIND 119 Class Project

Instructor - Dr. Syed Shariyar Murtaza
Name - Ashwin David
Section - DK0
Submission Date - 2021/04/20

# Table of Contents

## Members

Ashwin David, ashwin.david@ryerson.ca

## Project Description

The problem faced by the telemarketing section of the Portugese bank entails the lack of encouragement towards promoting long - term deposit accounts to their clients. The marketing campaigns used by the telemarketing team included phone calls from multiple contacts to guarantee the affirmation of said clients. The purpose of this report is to help construct an efficient and effective telemarketing strategy to successfully identify the intrinsic parameters that affect customers from subscribing for long term deposits. With the aid of different business and data analytic tools, the estimation towards the acceptance of a fit statistic will determine the levels of precision and accuracy of predicting future complications. It is important to note that the class label or feature of importance in this report is y, which has two categorical values, yes and no. The data preparation section will provide information on the basic statistics of the report and further eliminate necessary attributes that have no effect in the predictive outcome. This report will utilize classification techniques such as the regression tree and k- fold cross validation to determine the predictive value, whilst the value no will be used as the positive outcome in our case. The tools that will be used in this project will include weka explorer, statistical analysis system, R studio and python.

## Workload Distribution

| Member Name | Lists of Tasks Performed |
|---|---|
| Ashwin David | Project Description |
| Ashwin David | Data Preparation |
| Ashwin David | Predictive Modeling |
| Ashwin David | Conclusions and Recommendations |

## Data preparation

Using the dataset provided by the telemarketing group, the statistics of all attributes are displayed in tables 1 and 2 below. Table 1 depicts all statistics for the qualitative type - numerical while table 2 was constructed to display the categorical type - nominal. There are 17 attributes with 4521 observations or instances. The label or the class attribute, Y, will be used to determine whether subscriptions have been renewed, opted in or opted out of. After running the data through dynamic programming it can be concluded that there are no missing values or values denoted by NA.

| Attributes | Minimum | Maximum | Mean | Standard Deviation | Distinct Values |
|---|---|---|---|---|---|
| Age | 19 | 87 | 41.17 | 10.576 | 67 |
| Balance | -3313 | 71188 | 1422.658 | 3009.638 | 2353 |
| Day | 1 | 31 | 15.915 | 8.248 | 31 |
| Duration | 4 | 3025 | 263.961 | 259.857 | 875 |
| Campaign | 1 | 50 | 2.794 | 3.11 | 32 |
| Pdays | -1 | 871 | 39.767 | 100.121 | 292 |
| Previous | 0 | 25 | 0.543 | 1.694 | 24 |

*Table 1: Statistics for the Quantitative data provided by the Telemarketing group*

| Attributes | Type | Distinct values |
|---|---|---|
| Job | Nominal | 12 |
| Marital | Nominal | 3 |
| Education | Nominal | 4 |
| Default | Nominal | 2 |
| Housing | Nominal | 2 |
| Loan | Nominal | 2 |
| Contact | Nominal | 3 |
| Month | Nominal | 12 |
| Poutcome | Nominal | 4 |
| Y | Nominal | 2 |

*Table 2: Statistics for the Qualitative data provided by the Telemarketing group*

Figure 1 below provides detailed information on the five - number summary with the aid of the box plot function for representation. With the knowledge of explanatory data analysis it can be concluded that most attributes used in the prediction of whether or not customers are subscribed to long - term deposits are skewed to the right with multiple outliers giving us insight in potential unusual observations. After observing the boxplots it can be concluded that the attribute 'day' and 'age' have a well distributed shape with variability and a central value while the others have multiple outliers depicting highly influential values at extreme ends of the spectrum. It is also important to note that for the attribute Pdays which depict the number of days that have passed after contacting said client, has an assigned value of -1 which denotes that the client was not previously contacted causing the mean to be skewed to the right of the box plot representation. Figure 2 represents the numerical attributes after removing the outliers that posed potential fluctuations in statistical parameters. The number of observations removed is provided below each box plot assembled for the respective attribute. After the removal of outliers the disproportionate data has less effect on statistical significance providing better results and interpretations for the class label, y.

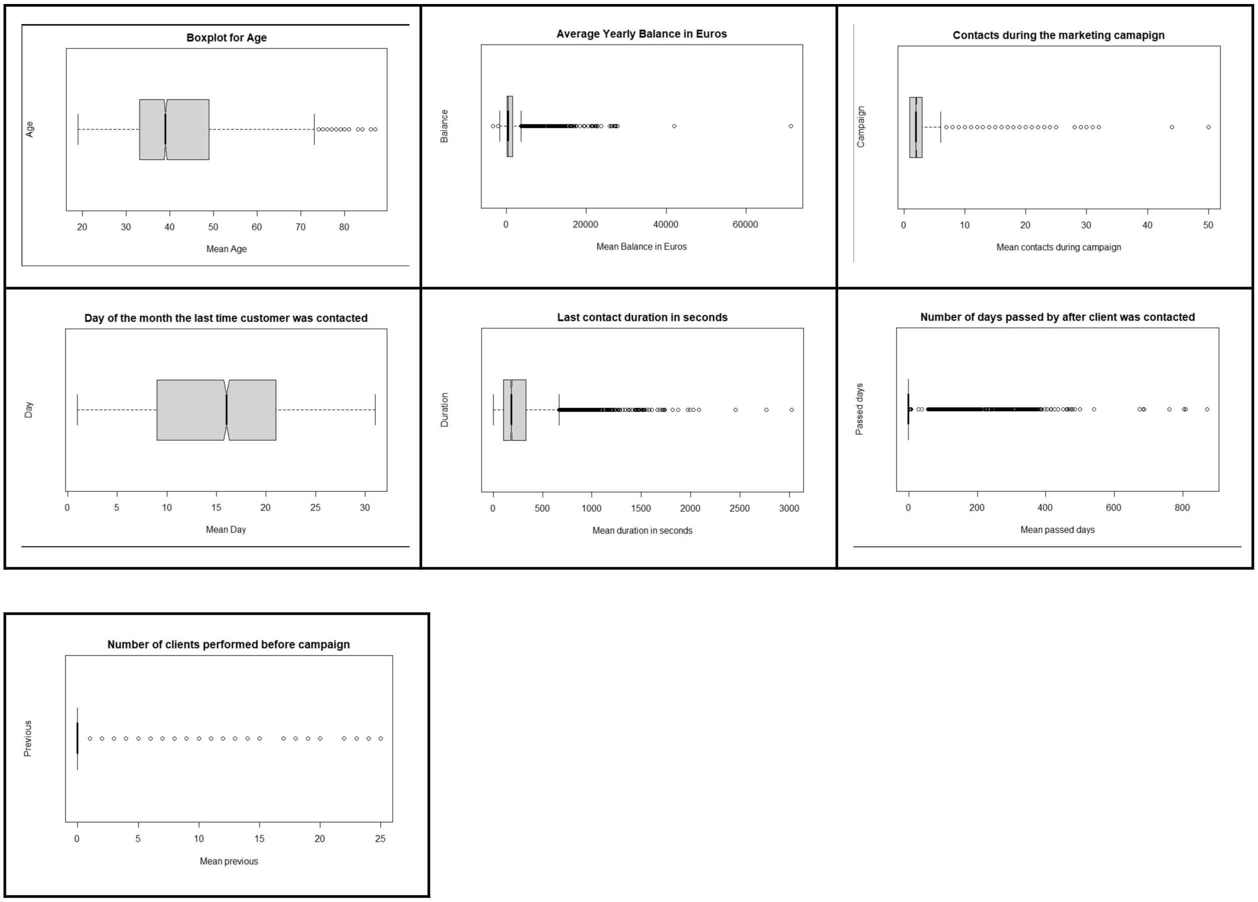
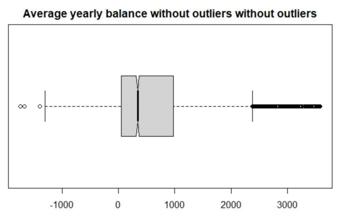
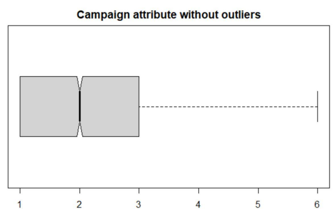


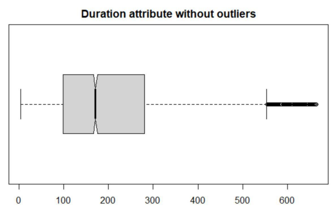
*Figure 1: Box Plots for all Numerical attributes from the telemarketing group*



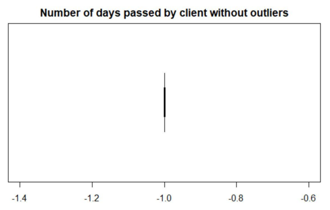
Obs removes - 506



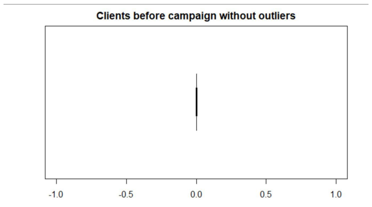
Obs removed - 318



Obs removed - 330



Obs removed - 816



Obs removed - 816

*Figure 2: Boxplots of Numerical attributes after removing outliers*

In order to understand the correlation or the joint probability distribution between two attributes and to analyze how similar they are a visual representation of each plot in the form of an histogram is provided below in figure 3. From the histograms provided, age represents an approximate evenly distributed shape when compared to other numeric attributes. All other attributes display a right skewed shape providing information on where statistical parameters such as mean, median and mode lie on. Figure 4 provides us with an insight towards the actual covariance and correlation between all numerical attributes and this is presented in the form of a matrix. When analyzing the data utilizing the correlation matrix most values tend to fall close to 0 resulting in a less linear relationship and proves an inversely proportional relationship between attributes.

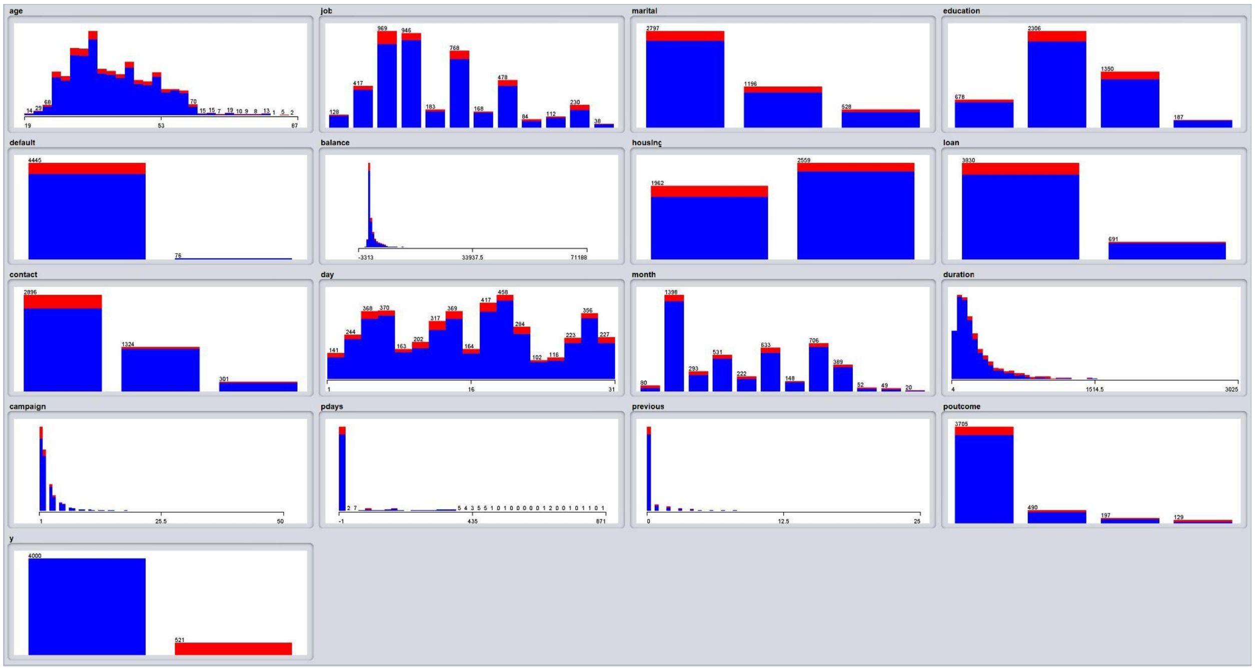
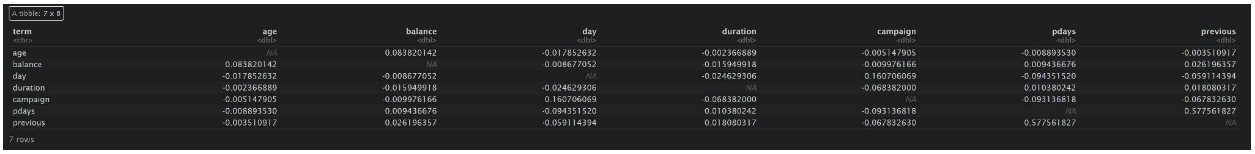


*Figure 3: Histograms of all attributes with respect to the class label Y*



A tibble: 7 x 8

| term | age | balance | day | duration | campaign | pdays | previous |
|---|---|---|---|---|---|---|---|
| <chr> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> |
| age | NA | 0.083820142 | -0.017852632 | -0.002366889 | -0.005147905 | -0.008893530 | -0.003510917 |
| balance | 0.083820142 | NA | -0.008677052 | -0.015949918 | -0.009976166 | 0.009436676 | 0.026196357 |
| day | -0.017852632 | -0.008677052 | NA | -0.024629306 | 0.160706069 | -0.094351520 | -0.059114394 |
| duration | -0.002366889 | -0.015949918 | -0.024629306 | NA | -0.068382000 | 0.010380242 | 0.018080317 |
| campaign | -0.005147905 | -0.009976166 | 0.160706069 | -0.068382000 | NA | -0.093136818 | -0.067832630 |
| pdays | -0.008893530 | 0.009436676 | -0.094351520 | 0.010380242 | -0.093136818 | NA | 0.577561827 |
| previous | -0.003510917 | 0.026196357 | -0.059114394 | 0.018080317 | -0.067832630 | 0.577561827 | NA |

7 rows

*Figure 4: Correlation matrix for all Numerical attributes*

From the matrix provided in figure 4, the correlation coefficients for each numerical attribute was calculated. This provides information about the linear relationship and the proportionality of how one attribute depends on another. With correlation coefficients greater than |0.7|, suggests that there is a strong correlation relationship between the two variables. An important factor to note is that the terminals for positive and negative propose the degree of correlation and the direction. From the data observed all attributes have lower correlation coefficients and thus this implies the relative degree of correlation and it could be concluded that they do not share similar characteristics and levels of dependency.

The next process that will be discussed is the elimination of certain attributes that are deemed impractical in determining the subscription rate for long term deposits. Using the feature selection from Weka, the raw data according to each attribute was evaluated with respect to the class label and certain attributes were found to be optimal in the machine learning process.

```
Evaluator:    weka.attributeSelection.CfsSubsetEval -P 1 -E 1
Search:       weka.attributeSelection.BestFirst -D 1 -N 5

Selected attributes: 1,12,16 : 3
                     age
                     duration
                     poutcome
```

Using the SUbset evaluator and search method of Best first, the conclusion can be made that the attributes age, duration and poutcome and are significant in determining the class label and have a greater impact.

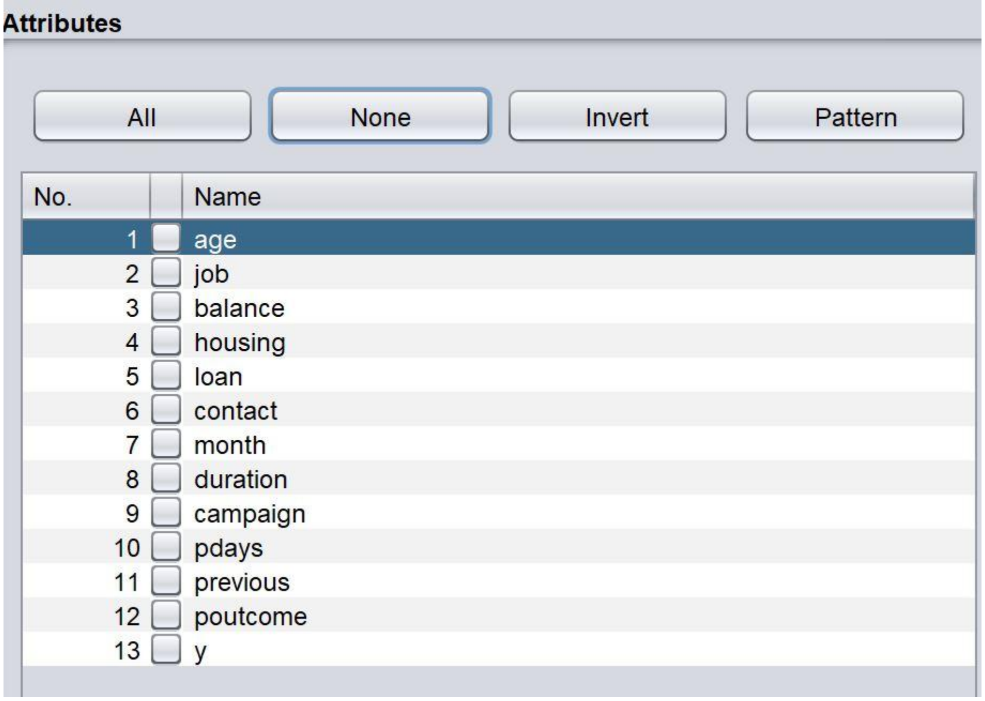
```
Search Method:
     Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 17 y):
     Information Gain Ranking Filter

Ranked attributes:
 0.10811967  12 duration
```

```
0.03758116  16 poutcome
0.03553361  14 pdays
0.0299014   11 month
0.01633501   9 contact
0.01622639  15 previous
0.00999086   2 job
0.00971603   1 age
0.00782731   7 housing
0.00533738   6 balance
0.0041129    8 loan
0.00304631  13 campaign
0.00297254   3 marital
0.00236554   4 education
0.00000121   5 default
0           10 day
```

Above is the results of evaluating the raw data the bank telemarketing team provided with the InfoGainAttributeEval provided as one of the classification techniques that provides data on the degree of preserving the validity of the class label. In conclusion, the attributes day, default, education and marital displayed a lower entry value and therefore will be neglected from the predictive modeling of the dataset. After removing the 4 attributes, the remaining variables include 12 attributes and one class label, y, which is displayed below in figure 5.



| No. |  | Name |
|---|---|---|
| 1 | □ | age |
| 2 | □ | job |
| 3 | □ | balance |
| 4 | □ | housing |
| 5 | □ | loan |
| 6 | □ | contact |
| 7 | □ | month |
| 8 | □ | duration |
| 9 | □ | campaign |
| 10 | □ | pdays |
| 11 | □ | previous |
| 12 | □ | poutcome |
| 13 | □ | y |

*Figure 5: Remaining attributes that will be used in predictive modeling*

## Predictive Modeling (Classification)

The first classification algorithm that will be used is the 10 - fold split. The data split method used in this report is the k - fold cross validation. The k - fold cross validation method is used to estimate the skill of the given data set with  a lower bias by dividing the data set into k number of folds or commonly known as the stratified sampling technique. This process is repeated a number of times with the training data set to increase accuracy and eliminate preliminary errors such as overfitting and underfitting of data points. Given certain attributes are eliminated,the 10 - fold split is used in this data set as the instances provided are densely distributed and thus optimistic estimate of the model skill is obtained.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        4000               88.476  %
Incorrectly Classified Instances       521               11.524  %
Kappa statistic                          0
Mean absolute error                      0.2041
Root mean squared error                  0.3193
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances             4521

=== Detailed Accuracy By Class ===
```

| **Model Based Fit Statistic** | **10 - Fold Validation** |
|---|---|
| Accuracy | 88.476 |
| True Positive | 1.00 |
| False Positive | 1.00 |
| Precision | 0.885 |
| Recall | 1.00 |
| ROC area | 0.499 |

*Table 3: Model based Fit statistics for the 10 - fold validation*

The next technique used to classify the data set is the decision tree. This utilizes the method of classification and regression in analyzing the data and explicitly providing decisions that follow the guidelines for machine learning. When running the data through a statistical analytic program, only the numeric attributes were used in constructing the decision tree while the class label y, which is known as the feature of interest was used in determining the direction of flow. The criteria based on impurity used for the classification process was entropy to split the data based on categorical responses.

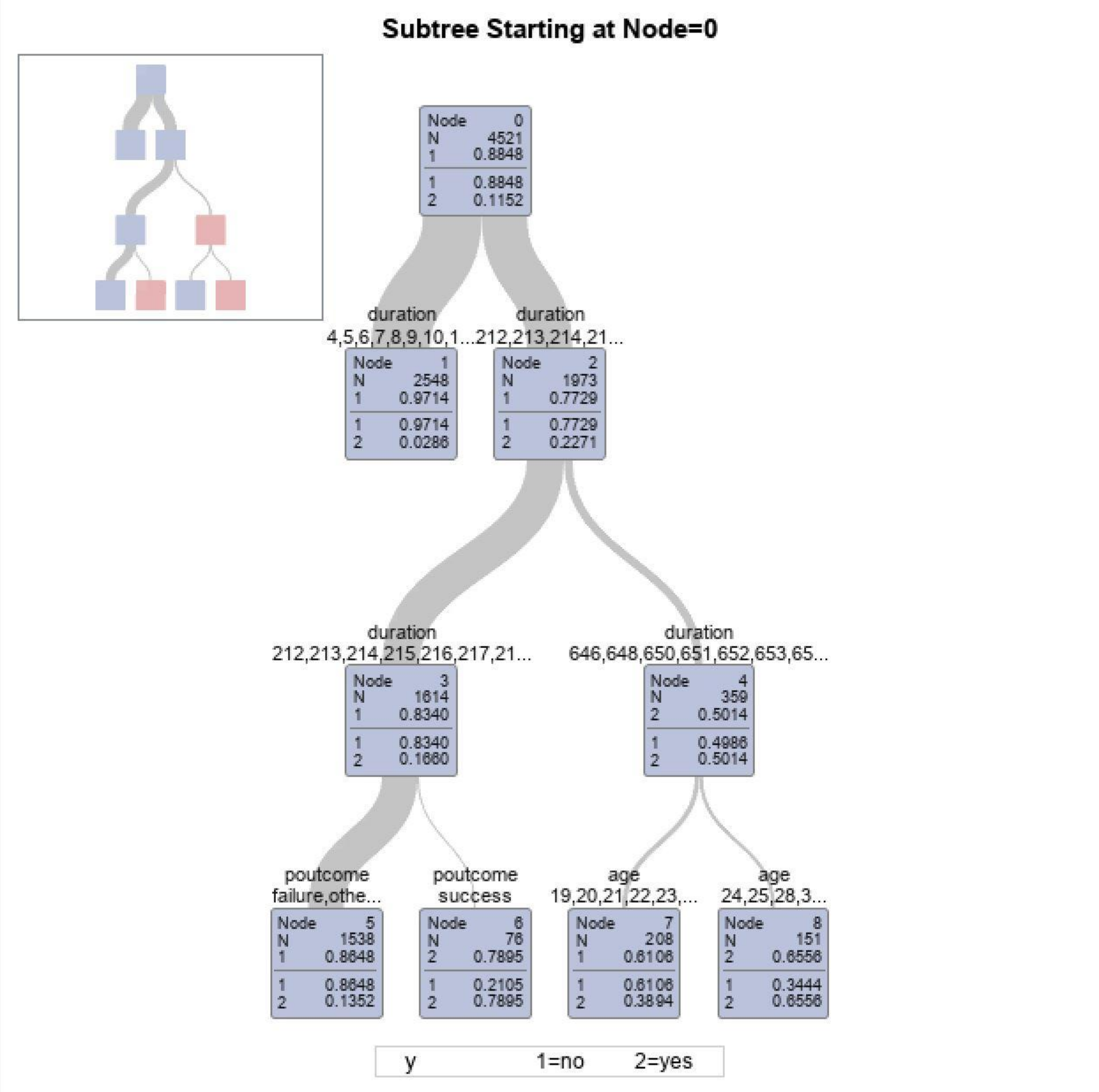


*Figure 6: Decision Tree*

| Model based Fit Statistic | Decision Tree |
|---|---|
| Accuracy | 0.905 |
| True Positive | 3932 |
| False Positive | 362 |
| Precision | 0.916 |
| Recall | 0.983 |
| ROC area | 0.81 |

*Table 4: Model Based Statistic for the Decision Tree*

## Conclusion and Recommendations

The data set provided to us consisted of 16 attributes containing both numerical and categorical values and one class label which was used to distinguish between whether or not a client would subscribe for long term deposits. During the data preparation phase a number of steps were taken to normalize and help visualize the data. Tables 1 and 2 provided important statistical information on each attribute and a hidden process to determine whether any NULL or NA values were available. The next step included determining the outliers of each attribute with the aid of box plots using R studio and then further removing these outliers and cleaning the data set from potential fluctuations and data points that indulge in errors.Moreover, histograms were used to visualize the remaining data and information regarding the normalization of each plot was considered. Finally, unwanted or attributes that posed the least significant changes towards determining the outcome were removed from the dataset with the help of attribute evaluation available in weka. The next major step used was the Predictive modeling phase which included two classification techniques such as the k - fold cross validation technique and the decision tree. From the values provided in tables 3 and 4, it can be concluded that the decision tree provided better results than the 10 - fold cross validation technique. When comparing each model based statistic the accuracy, precision and recall statistics were close to 0.9 and 1.0 resulting in a better outcome of results. Finally, we would like to advise the telemarketing team of the Portugese bank that most customers are inclined not to subscribe for long term deposits and this is provided in the categorical attributes such as housing and loan that provide higher percentages for No rather than Yes. One last recommendation that our team would like to advise is to build a social media marketing campaign to indulge and interact with customers virtually at a higher rate rather than awaiting confirmation via phone calls.