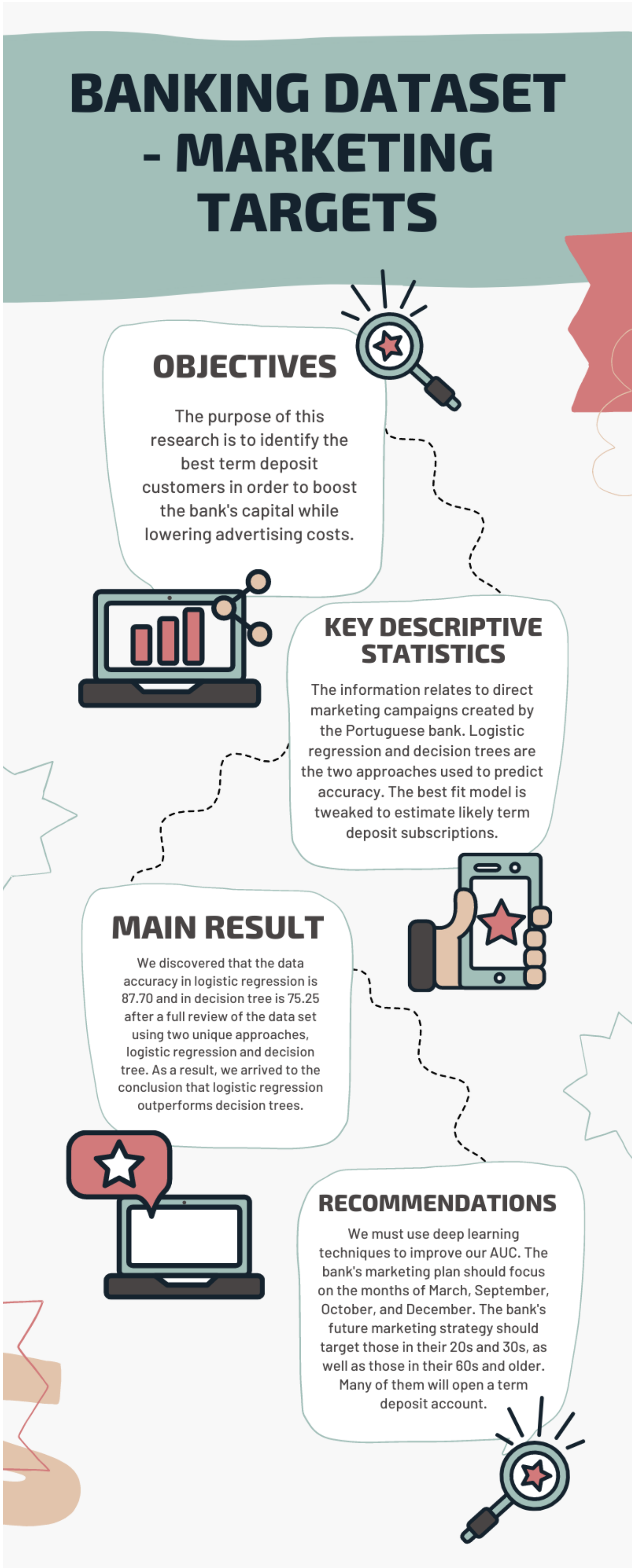
SUPERVISED LEARNING USING LOGISTIC AND DECISION TREE REGRESSION

Anusha Girish

(23638582)

Executive summary



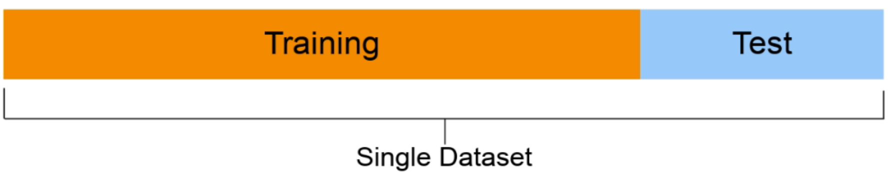
Introduction

The Ideology of the business is based on Bank marketing related to marketing, In which, a term deposit is a long-term investment that involves depositing funds into a financial institution's account. Term deposits are short-term investments with maturities ranging from one month to a few years with varied minimum deposit requirements. With the data recorded by the Portuguese team on the banking data, we are doing an analysis. Here is an example of what we would be predicted, when a bank customer deposits money, the bank can use it to lend to other consumers or businesses. In exchange for the right to use these funds for lending, they will provide the depositor compensation in the form of interest on the account balance. Most deposit accounts of this sort allow the owner to withdraw money at any time. This makes it difficult for the bank to forecast the amount of money it will be able to lend at any one time. Now Exploratory analysis is used to determine what factors have led to selecting which customer is able to do a term deposit. The marketing strategy used to collect the data by the team is by making calls.

Method

The Bank- marketing Dataset selected has 45211 observations and 16 variables. The Variable which must be predicted is y (i.e., has the client subscribed to a term deposit?), So the dependent variable in this Dataset is y and all the other variables in the data set are independent variables. I have used two algorithms to predict the dependent variables those are Logistic regression and Decision tree. Here is a brief description Logistic regression is a method which comes under the category of supervised machine learning techniques and is used for classification purposes. Supervised learning implies that the machine learns from labelled data (Lao, 2018). A decision tree is also explored and included in this analysis. AUC represents the degree or measure of separability, whereas ROC is a probability curve. It indicates how well the model can distinguish between classes. The AUC indicates how well the model predicts 0 courses as 0 and 1 classes as 1.

Before the implementation of the algorithms (logistic/Decision Tree), Data processing will be carried out to the original bank data set which is hereby classified into 2 types Training and testing data sets.



- Training Data: The first data needed to train machine learning models is known as training data (or a training dataset). All Machine learning algorithms are taught how to make predictions or perform a task using training datasets.
- Testing Data: Data that has been explicitly identified for use in testing, usually of a computer programme, is known as test data. Some data can be utilised in a confirmatory manner, for example, to ensure that a particular set of inputs to a function provides the desired output.

The Ratio of split for the Data sets I have chosen is 80:20 as it's the globally preferred split ratio. As the Original data is 45211, the train will have 36170 records whereas test data set has 9041 records. The Data set selected for the analysis is from Kaggle https://www.kaggle.com/datasets/prakharrathi25/banking-dataset-marketing-targets?datasetId=223954&searchQuery=R

Descriptive Statistics and Preliminary Correlation Analysis

View of the Data Set: The command used to view the data is str(dataset_name)

```
> str(bank_data)
'data.frame':	45211 obs. of  17 variables:
 $ age      : int  58 44 33 47 33 35 28 42 58 43 ...
 $ job      : chr  "management" "technician" "entrepreneur" "blue-collar" ...
 $ marital  : chr  "married" "single" "married" "married" ...
 $ education: chr  "tertiary" "secondary" "secondary" "unknown" ...
 $ default  : chr  "no" "no" "no" "no" ...
 $ balance  : int  2143 29 2 1506 1 231 447 2 121 593 ...
 $ housing  : chr  "yes" "yes" "yes" "yes" ...
 $ loan     : chr  "no" "no" "yes" "no" ...
 $ contact  : chr  "unknown" "unknown" "unknown" "unknown" ...
 $ day      : int  5 5 5 5 5 5 5 5 5 5 ...
 $ month    : chr  "may" "may" "may" "may" ...
 $ duration : int  261 151 76 92 198 139 217 380 50 55 ...
 $ campaign : int  1 1 1 1 1 1 1 1 1 1 ...
 $ pdays    : int  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
 $ previous : int  0 0 0 0 0 0 0 0 0 0 ...
 $ poutcome : chr  "unknown" "unknown" "unknown" "unknown" ...
 $ y        : chr  "no" "no" "no" "no" ...
> 
```

With the variables in the dataset, the variable "previous" doesn't make sense for the analysis so we have dropped the variable using the below command.

bank_data <- subset(bank_data, select = -c(previous))

```
> summary(bank_data)
      age                 job           marital          education     default        balance       housing    
 Min.   :18.00   blue-collar:9732   divorced: 5207   primary  : 6851   no :44396   Min.   : -8019   no :20081  
 1st Qu.:33.00   management :9458   married :27214   secondary:23202   yes:  815   1st Qu.:    72   yes:25130  
 Median :39.00   technician :7597   single  :12790   tertiary :13301               Median :   448              
 Mean   :40.94   admin.     :5171                    unknown  : 1857               Mean   :  1362              
 3rd Qu.:48.00   services   :4154                                                  3rd Qu.:  1428              
 Max.   :95.00   retired    :2264                                                  Max.   :102127              
                 (Other)    :6835                                                                              
  loan           contact           day            month              duration         campaign          pdays      
 no :37967   cellular :29285   Min.   : 1.00   Length:45211       Min.   :   0.0   Min.   : 1.000   Min.   : -1.0  
 yes: 7244   telephone: 2906   1st Qu.: 8.00   Class :character   1st Qu.: 103.0   1st Qu.: 1.000   1st Qu.: -1.0  
             unknown  :13020   Median :16.00   Mode  :character   Median : 180.0   Median : 2.000   Median : -1.0  
                               Mean   :15.81                      Mean   : 258.2   Mean   : 2.764   Mean   : 40.2  
                               3rd Qu.:21.00                      3rd Qu.: 319.0   3rd Qu.: 3.000   3rd Qu.: -1.0  
                               Max.   :31.00                      Max.   :4918.0   Max.   :63.000   Max.   :871.0  

    poutcome        y        
 failure: 4901   no :39922  
 other  : 1840   yes: 5289  
 success: 1511              
 unknown:36959              
```

Summary of the Dataset:

In the above image, we can have a view of the mean, median, quartiles and range for continuous variables and the total number of values and categories for each nominal value.

A crosstab is a table which shows the detailed relationship between two or more variables (i.e., for categorical variables here)

```
> ftable(table_1)
    housing    no   yes
y
no          16727 23195
yes          3354  1935
```

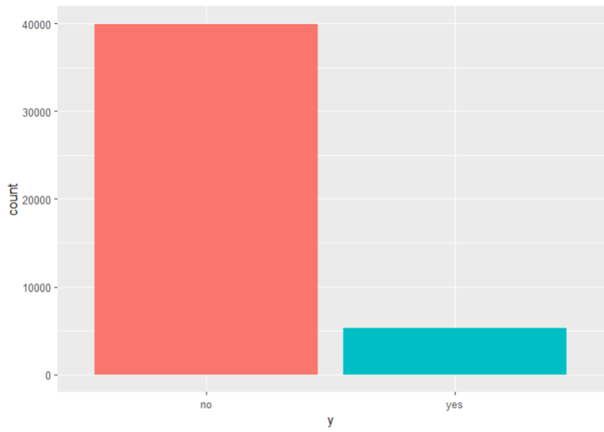
```
> ftable(table_2)
    education primary secondary tertiary unknown
y
no               6260     20752    11305    1605
yes               591      2450     1996     252
>
```

```
> ftable(table_3)
    job admin. blue-collar entrepreneur housemaid management retired self-employed services student technician unemployed unknown
y
no        4540        9024         1364      1131       8157    1748          1392     3785     669       6757       1101     254
yes        631         708          123       109       1301     516           187      369     269        840        202      34
```

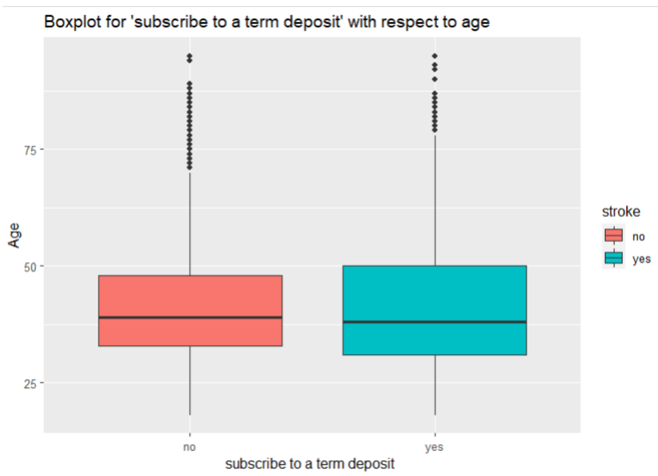
```
> ftable(table_4)
    marital divorced married single
y
no              4585   24459  10878
yes              622    2755   1912
>
```

```
> ftable(table_5)
    loan    no   yes
y
no       33162  6760
yes       4805   484
```

Charts: The charts which we have used for the analysis are histogram, box and whisker plots to analyse the continuous variables data in the dataset.



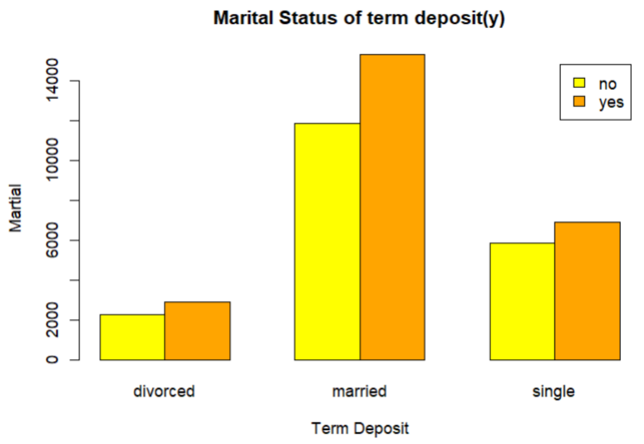
This is the graph which displays the count of yes and no variables for the dependent variable that is y in the dataset.

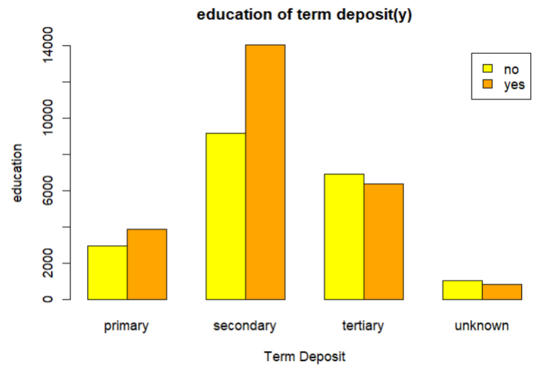


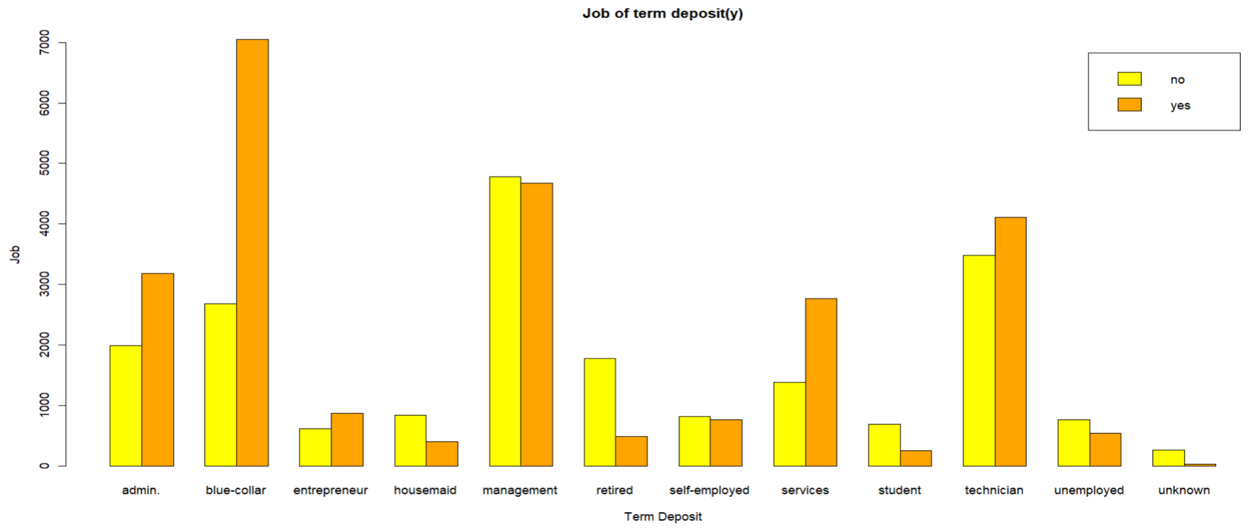
This is a gg plot that describes the relationship between the dependent variable and Age.

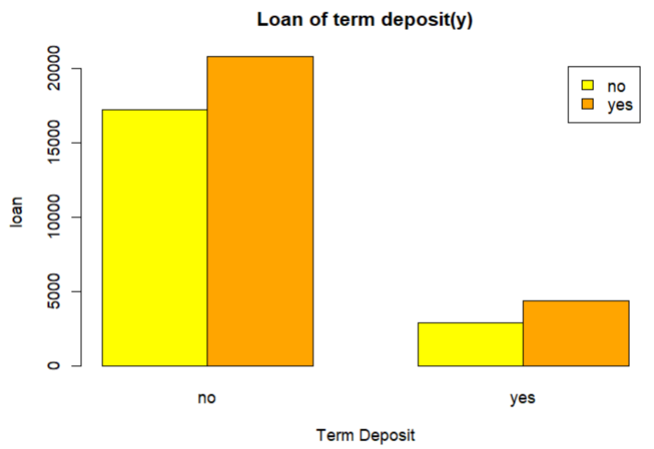
When observed it gives clarity that there are many outliners when compared with age.

We shall now compare the dependent variable with a few of the Independent variables and below are the graphical comparison.



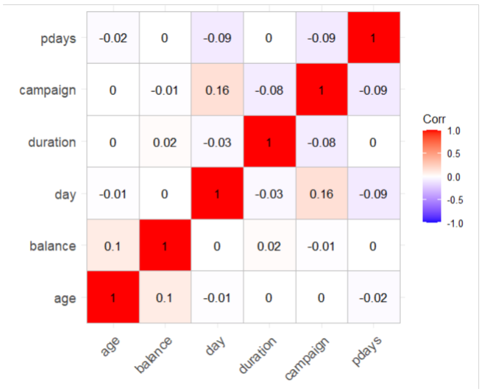




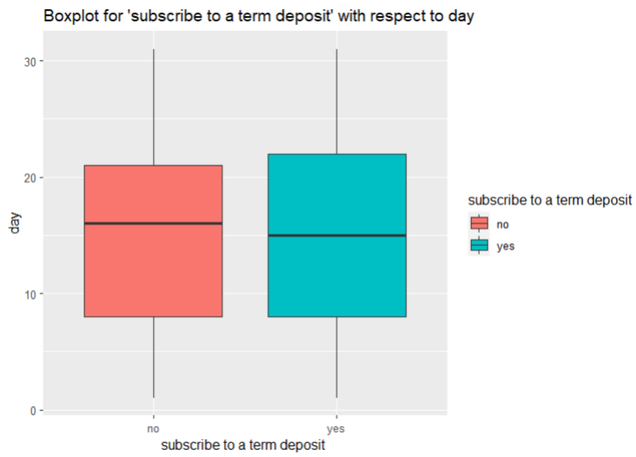


The correlation matrix includes p-values and confidence ranges to assist users in determining the relationships' statistical significance.

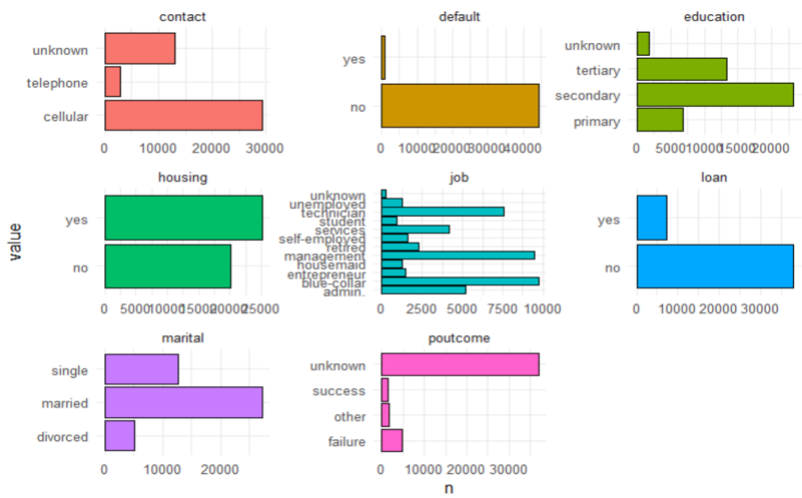
The below graph gives a view of the correlation matrix for the bank data set.



With respect to the boxplot, the data is being populated based on Term deposit and day and in the analysis of this, we see that there are no outliers which means that the data is being manipulated.



A histography view of all the categorical variables.



Analytics

To Start the analytics first we need to convert data to training and testing with the ratio selected 80:20 by setting up the seed.

Seed: The objective is to ensure that we use the same training and validation data set when evaluating the performance of multiple models same hyperparameters or machine learning techniques.

Logistic and Decision tree models:

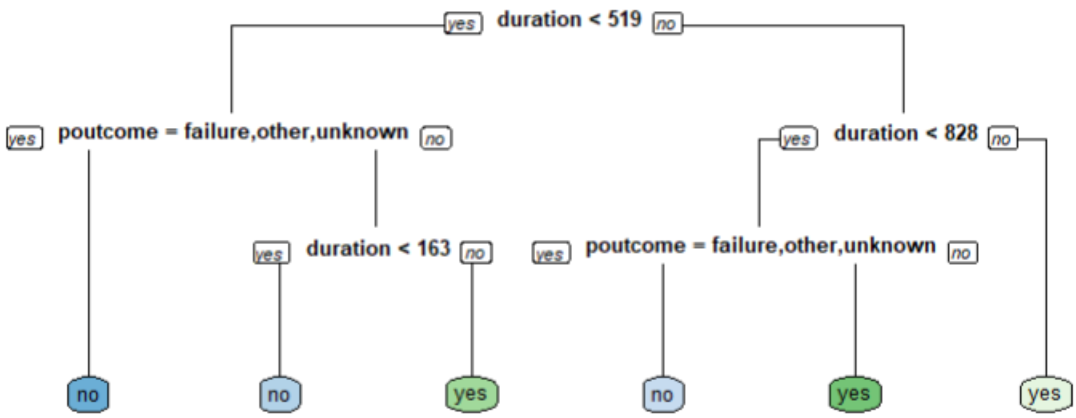
There are three models which are implemented for logistic regression and one for decision tree. With the 3 models applied for the same dataset with changes in variables, The training data is used to build both logistic and decision tree models, which are subsequently used to make predictions on the test data. Confusion matrices may be used to determine the accuracy of each model. The model with the highest accuracy and lowest AIC will be chosen and deemed as the best model for logistic regression. For logistic regression and decision trees, the AUC-ROC curve is also utilised. The values of the confusion matrix are observed and anticipated.

```
> model_1[["aic"]]
[1] 21827.92
> anova(model_1, test = "LRT")
Analysis of Deviance Table

Model: binomial, link: logit

Response: y

Terms added sequentially (first to last)


             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                         36169      26108
job          11   620.89     36158      25487 < 2.2e-16 ***
marital       2   101.27     36156      25386 < 2.2e-16 ***
education     3    74.92     36153      25311 3.771e-16 ***
age           1    39.46     36152      25272 3.349e-10 ***
housing       1   415.23     36151      24857 < 2.2e-16 ***
contact       2   642.06     36149      24215 < 2.2e-16 ***
default       1    25.81     36148      24189 3.760e-07 ***
loan          1   119.22     36147      24069 < 2.2e-16 ***
month        11  1107.28     36136      22962 < 2.2e-16 ***
poutcome      3  1116.37     36133      21846 < 2.2e-16 ***
day           1     0.35     36132      21845    0.5535
campaign      1    95.47     36131      21750 < 2.2e-16 ***
day:campaign  1     2.03     36130      21748    0.1540
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> logLik(model_1)
'log Lik.' -10873.96 (df=40)
> deviance(model_1)
[1] 21747.92
> |
```

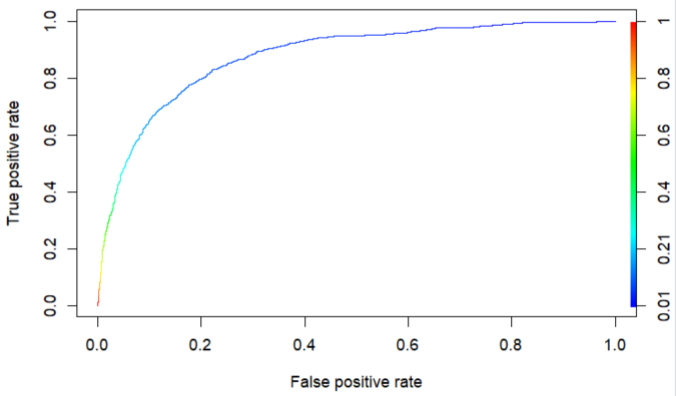
```
> model_2[["aic"]]
[1] 25215.76
> anova(model_2, test = "LRT")
Analysis of Deviance Table

Model: binomial, link: logit

Response: y

Terms added sequentially (first to last)


          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                      36169      26108
job       11   620.89     36158      25487 < 2.2e-16 ***
marital    2   101.27     36156      25386 < 2.2e-16 ***
education  3    74.92     36153      25311 3.771e-16 ***
loan       1   131.45     36152      25180 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> logLik(model_2)
'log Lik.' -12589.88 (df=18)
> deviance(model_2)
[1] 25179.76
> |
```

```
> model_3[["aic"]]
[1] 19003.27
> anova(model_3, test = "LRT")
Analysis of Deviance Table

Model: binomial, link: logit

Response: y

Terms added sequentially (first to last)


         Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                     36169      26108
poutcome  3   2123.9     36166      23984   <2e-16 ***
loan      1    116.6     36165      23868   <2e-16 ***
housing   1    538.2     36164      23330   <2e-16 ***
pdays     1      0.3     36163      23329   0.5594
duration  1   4342.1     36162      18987   <2e-16 ***
age       1      1.9     36161      18985   0.1678
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> logLik(model_3)
'log Lik.' -9492.635 (df=9)
> deviance(model_3)
[1] 18985.27
> |
```

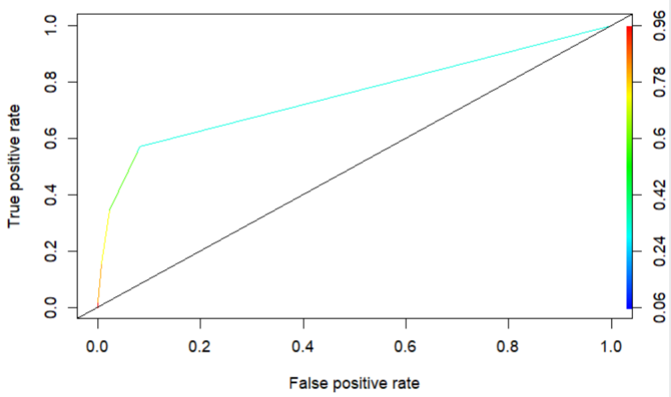
Now if we compare the three models created the data is model3 more accurate compared to the other two model's.



ROC-AUC



The accuracy of the data is 87.70 when applied model3 in logistic regression.



The Accuracy of the data is 75.25 when applied using a decision tree.

With the above analysis, we can conclude that Logistic regression is more accurate than a decision tree.

Recommendations and Conclusions:

We can impute p-days which have around 70% of values with 999 by using different imputation methods We have trained our models on the very limited data set. To improve our AUC score we can use Deep Learning techniques which require more datasets for training. (Pro-Dut2000)

1) Months of marketing activity: We discovered that the month of May had the highest degree of marketing activity. This was the month, however, when potential consumers were most likely to reject term deposit proposals (Lowest effective rate: -34.49 per cent). The bank should concentrate its marketing efforts on the months of March, September, October, and December for the next marketing campaign.
2) Seasonality: During the fall and winter seasons, potential clients choose to subscribe to term deposits. During these seasons, the next marketing campaign should concentrate its efforts.
3) Age Category: The bank's next marketing effort should target potential consumers in their 20s and younger, as well as those in their 60s and older. The youngest category had a 60% likelihood of signing up for a term deposit, while the oldest category had a 76% chance. It would be fantastic if the bank addressed these two categories in the upcoming campaign, increasing the possibility of more term deposit subscriptions.
4) Campaign Calls: To minimise time and effort in obtaining new potential clients, a policy should be created that says that no more than three calls should be made to the same possible client. Remember that the more times we call a potential client, the more likely he or she is to refuse to establish a term deposit.

It is likely that the bank's next marketing campaign will be more effective than the present one if all of these methods are combined and the target market for the next campaign is simplified.

References

BANKING MARKETING ANALYSIS. (n.d.). Retrieved from Jovian: https://jovian.ai/pro-dut2000/banking-market-analysis

Lao, R. (2018, July). *A beginners guide to machine learning*. Retrieved from medium.com: https://medium.com/@randylaosat/a-beginners-guide-to-machine-learning-5d87d1b06111

RATHI, P. (n.d.). *Banking Dataset - Marketing Targets*. Retrieved from Kaggle: https://www.kaggle.com/datasets/prakharrathi25/banking-dataset-marketing-targets?datasetId=223954&searchQuery=R