**Final Project Report**

****

**Team**

**Sunil Raj Thota**

**Nalini Macharla**

**Date: 02/25/2021**

**Title: Final Project Report**

**ALY 6015 – Intermediate Analytics**

**Prof. Roseanna Hopper**

**Introduction**

In this assignment, we chose “Portuguese Bank Marketing” Dataset that was retrieved from the UCI Machine Learning Repository. Here, we will perform Exploratory Data Analysis, Inferential Statistics, and Regularization Techniques on the bank dataset and build a machine learning models like Logistic Regression to know the relationship between the variables, Support Vector Machines, Cross Validation, and Decision Tree Classifiers in order to predict the terms of a Deposit Subscription by the clients in the bank. It will be a supervised ML model which will try to solve the classification problem like whether the client will subscribe the term deposit in a bank or not (a simple yes or no).

Also, we will do some analysis on one-sample t-test, two-sample t-test, Paired t-test, the test of equal or given proportions, and F-tests. A one-sample t-test states whether an unidentified population means is dissimilar from a definite value. The two-sample t-test is also known as independent samples t-test to test whether the unknown population means of 2 groups are identical or not. Paired t-test also called the dependent sample t-test to discover whether the mean change between 2 sets is 0. They are validated two times, resulting for pairs of observations. The test of equal or given proportions will test whether or not a sample from a population represents the true proportion from the entire population. The last test, F-test signifies the linearity gives improved fit.

**Analysis & Methods**

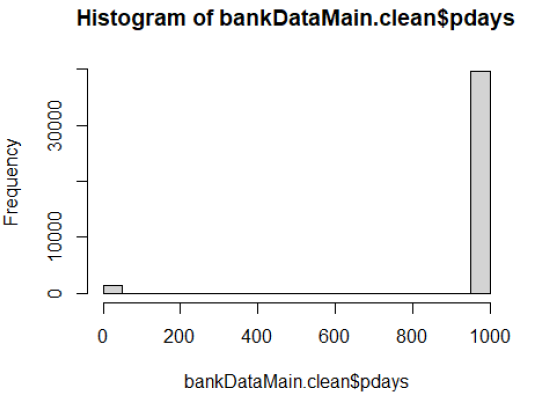
Let us now understand the dataset we have been given. What is there is in the data file and what are the columns present? This dataset contains 20 independent variables and 1 response variable, 41188 records of data. It has all the information belonging to a marketing campaign of a bank and we will have to analyze this data to come up with strategical planning to improve the marketing campaigns in the future for this financial institute. In the dataset given, the target column is ‘DEPOSIT’.

This dataset consists of the data such as client’s educational information, loans, balance, contact information of the client and so on. Based on these details, we can predict if the customer is interested in the term deposit and these values are shown in the form of YES or NO in our dataset.

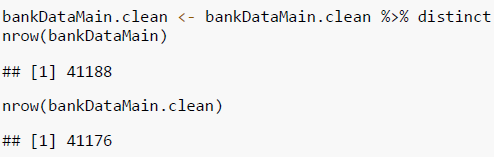
Now, let’s understand what exactly term deposit means, it is a deposit system which is a little different from the normal deposit account in a bank. This is specially designed for a specific period where one will receive back the amount after the certain period ends with a fixed rate of interest.

After deep analysis and close observations on the dataset which deals with numerical values like age, campaign, etc. We will further understand the categorical features and its unique values, like job, education, default, marital status, housing, loan, contact, month, poutcome and deposit. Deposit is described as if the client has subscribed to the term deposit or not and it is mentioned in the form of ‘yes’ or ‘no’.

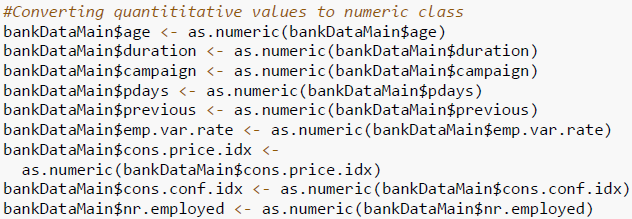
For this assignment we have used methods such as Logistic Regression, SVM, Decision Tree Algorithm, Random Forests. We also plotted various graphs such as histograms, bar plots, density plots using GGplot package.



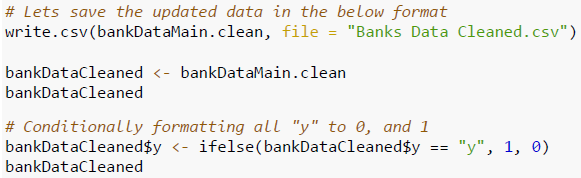
Plotted a histogram of pdays where the number of days completed after the client was last contacted.



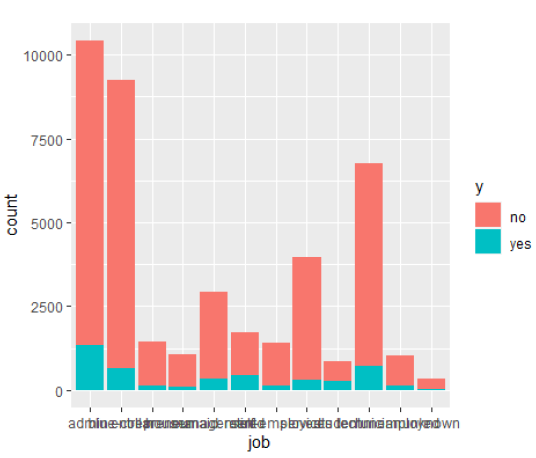
Originally, the dataset has 41188 observations but after cleaning the dataset it got reduced to 41176 distinct records. We have checked for the duplicate records and removed accordingly if any. And, also converted the quantitative values to numeric classes by using as.numeric()



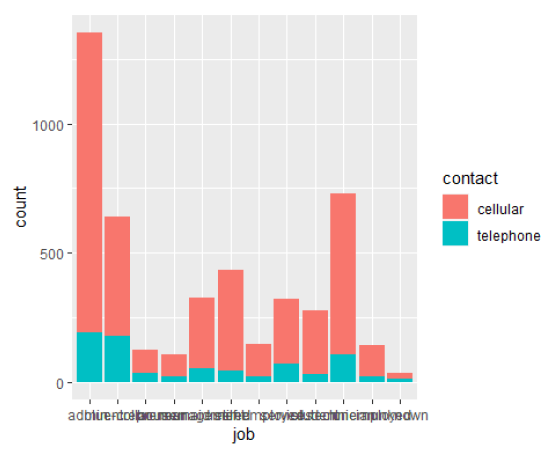
After cleaning the dataset, we have saved and updated the data in the below format



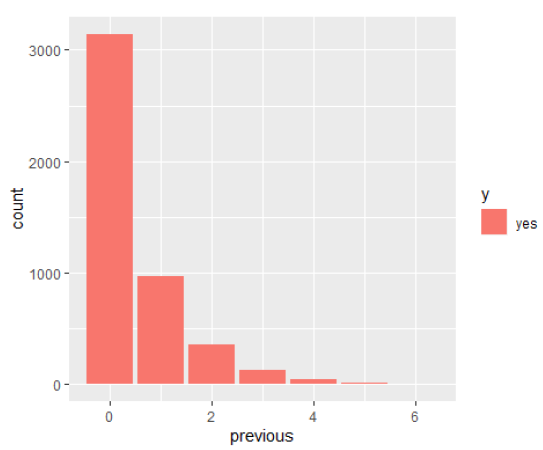
Also, conditionally formatted all the predictor variables that is “y” column to 0 and 1.



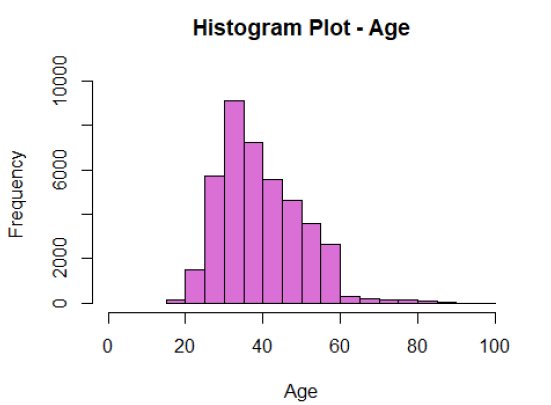
In the above plot, we can see the distribution of Age and Job with respect to y variable



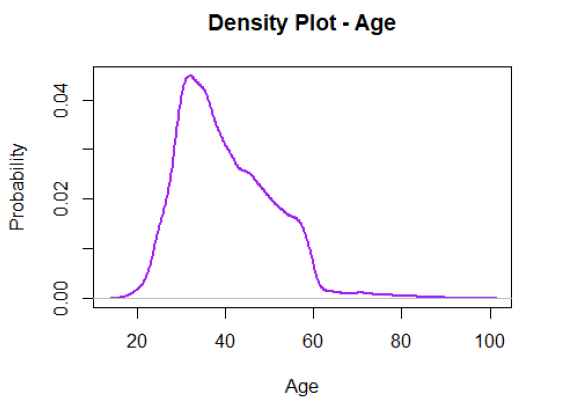
In the above plot, we can depict that the relation between job and contact. Management and Technician are most clients with job description that subscribed a term deposit.

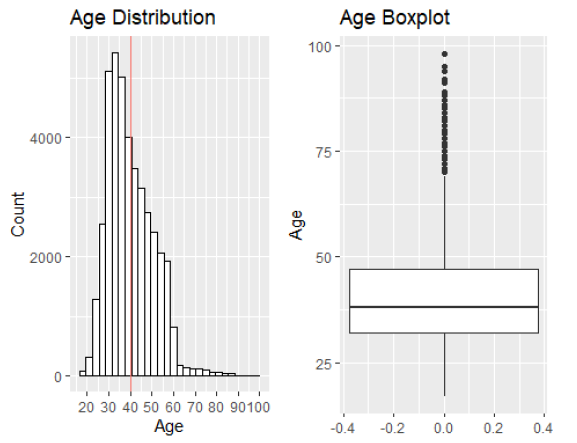


Previous number of contacts performed before this campaign and for this client has significantly less effect on subscription of product compared to clients who has no previous number of contacts performed.

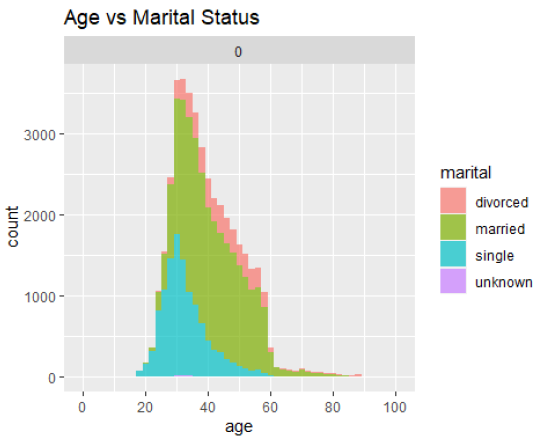


The bulk of clients are between the ages of 30 and 50 with mean lying around 40. Boxplot of Age tells the same statistics but we can see outliers > Age of 64

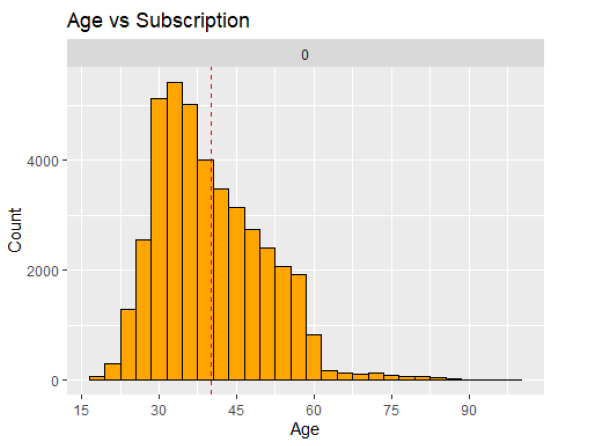




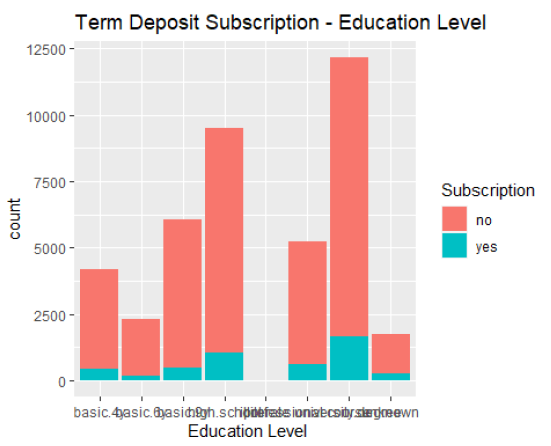
Now, lets deep dive into the Age distribution and marital status that subscribes Term Deposit.



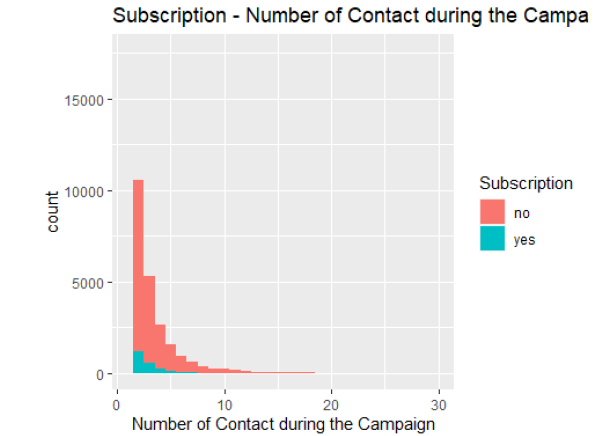
The bulk of clients are married or divorced. Sharp drop of clients above age 60 with marital status ‘divorced’ and ‘married’. \*Single clients drop in numbers above age 40.



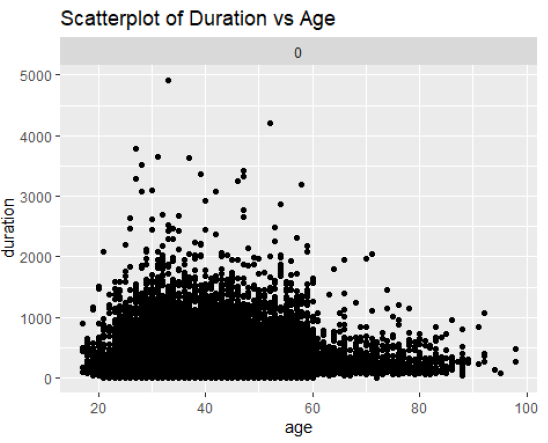
The above plot showcases with the most clients that subscribe are between age 24 to 46. Mean age for all clients is > 40 years of age.



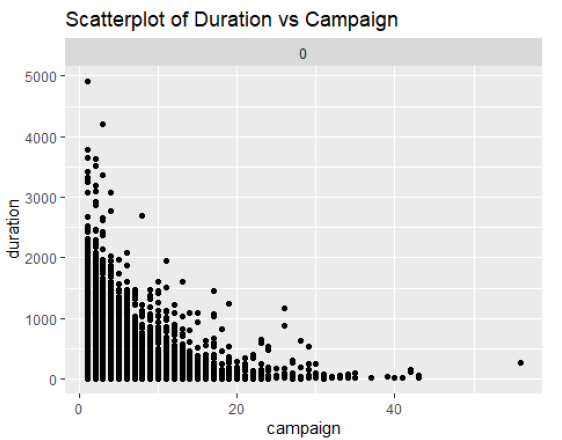
Now, let see the relation between higher education and term deposit. Most clients who subscribe are from ‘secondary’ and ‘tertiary’ education levels. Tertiary educated clients have higher rate of subscription (16%)



From the above plot, It can be observed from bar chart that there will be no subscription > 7 contact during the campaign. Future campaigns would improve resource utilization by taking proper thresholds to contacts during a campaign. Future campaigns can focus on 1st 3 contacts as it will have greater rate of subscription.



Scatterplot between Duration and Age. We can see that there are less clients after age of 61.

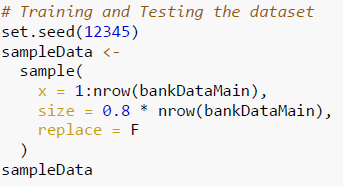


This plot shows that the duration on call similar for 1st ten contacts during campaign. A good subscription occur within first 10 contacts.

We have done some Data Correlation Analysis and got to analyze these things: We wanted to determine the effectiveness, accuracy, and precision of Age, Marital, Loan, Job, and Housing attributes. We conducted these tests by using the correlation test.

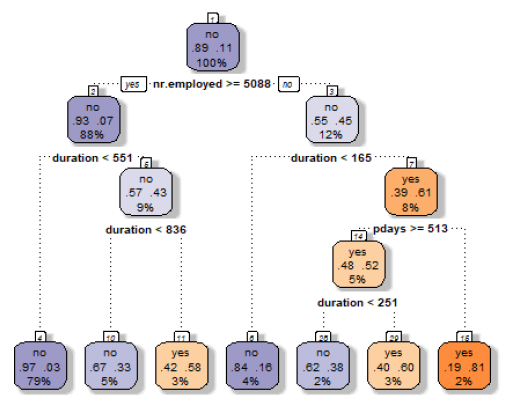
* Correlation Test: Age ~ Term Deposit - cor = 0.03033926
* Correlation Test: Job ~ Term Deposit - cor = 0.02512217
* Correlation Test: Marital ~ Term Deposit - cor = 0.04620261
* Correlation Test: Education ~ Term Deposit - cor = 0.05779889
* Correlation Test: Housing ~ Term Deposit - cor = 0.01155169
* Correlation Test: Loan ~ Term Deposit - cor = -0.004908593
* Correlation Test: Housing + Loan ~ Term Deposit - cor = 0.006273869

Therefore, 95% confidence in the correlation is proven.

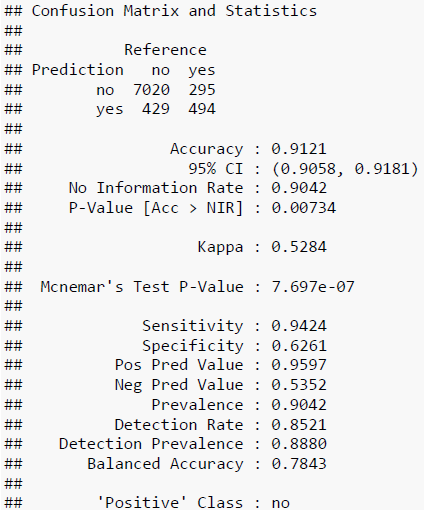


Now, let us split the dataset into training and testing dataset each holds around 80: 20 of the original data.

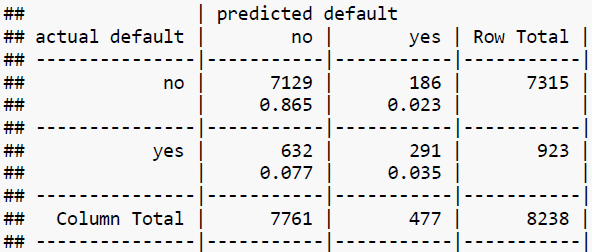
Decision Tree is one of the very simplest and useful ML algorithm. It can be broadly used for two purposes. One is for classification and the other is for regression. Regression is really about determining the set of continuous numbers, so if you wanted to predict a number or something numerical that’s when we use regression. Classification is when you want to predict a class as into which class of the given data in the dataset does it actually belong to. The combination of both these terms is known to us and is referred to as CART. Having a good representative set for the training data, the data that you feed into your training clearly influences the results that you get off your predictive models.



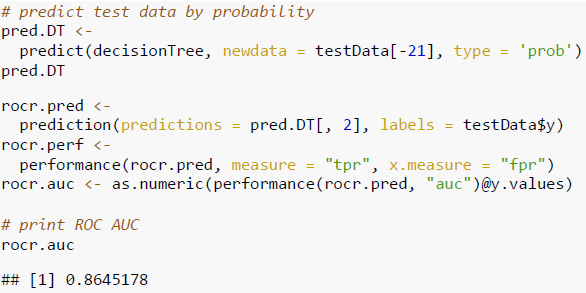
In this assignment, we used Decision Tree model to build and plot it. We got an accuracy of **91.21%**



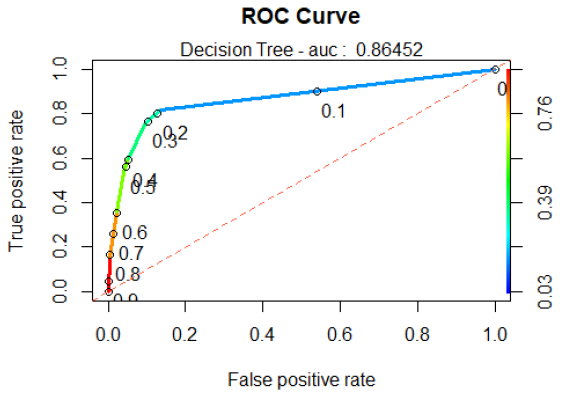
And, tabulated the KNN predicted values in the below table



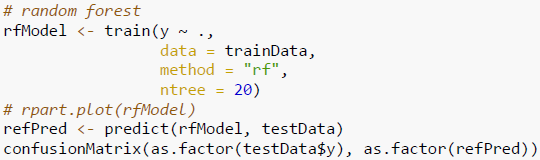
After some predictions and testing the data by probability, we got ROC AUC as 0.8645178. Let's predict our model with the help of testing data set which was initially splitted as testData with 20% of main data set. For this, we have to use predict() method on the testing data set. In this, lets mix and match various columns to implement several various decision tree models and compare the prediction results of each.



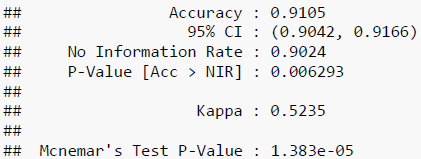
Plotted the ROC curve – Decision Tree



We have also evaluated and predicted the Logistic Regression model on testing data. It defines whether the variable produces and event like yes/ no or 0/ 1. It is used to define data and to enlighten the bond among 1 dependent binary (0/ 1) and 1 or more independent variables.



Used Bagging in the Random Forest model and predicted the accuracy with **91.05%**



Random Forest is the most powerful supervised ML Algorithm that is capable of performing both regression and classification tasks. As the name implies, this algorithm creates forest with a number of decision trees. In general, the more trees in the forest, the more robust is the prediction and thus gives us higher accuracy to model multiple decision trees to create the forest. We are going to construct the decision with the information gained amongst other algorithms. A regular linear regression would say that I want to just to minimize the sum of the squared errors in my fit

**Conclusion**

|  |  |  |
| --- | --- | --- |
| **MODEL** | **ACCURACY (%)** | **RANK** |
| Logistic Regression | **90.38** | **3** |
| k-Nearest Neighbors | **90.07** | **4** |
| **Decision Trees** | **91.21** | **1** |
| Random Forests | **91.05** | **2** |

The accuracy of the predictions are verified with a Probability Value of 0.00734 and a 95% confidence interval of 0.9058 to 0.9181 For better understanding we can go further by building XG Boost, Ada Boost, GBM, Light GBM, and Neural Network Models and figure out the best accurate predictor and use it in the Bank’s Marketing Campaign.

To perform multiple analytical operations, we use R on a wider range of data like various datasets. R is a very powerful tool to perform analysis which is mainly built by researchers, statisticians, and developers. In the beginning, we did Descriptive Statistics Analysis, Exploratory Data Analysis, and we will continue with the Logistic regression, k means clustering, density-based clustering, and hierarchical clustering Data Mining Techniques to get understandings from the provided data. Plotted relevant parameters, to helps us to understand the data in a clear way. We are going to use these packages for our dataset plyr, party, caret, tidyr, rattle, randomForest, ROCR, rpart, ggplot2, e1071, gmodels, gridExtra.

The results show that the Decision Tree model are fitted to evaluate train data considering that the accuracy in the test set is **91.21**%. From the above observations, analysis, and model implementations we have trained and tested various models and predicted with an accuracy of ~91.21%. To get a good performant model we need to implement a rigorous model and compare with other classification models.

**References**

[1] Larry Alton (December 22, 2017) The 7 Most Important Data Mining Techniques was retrieved from <https://www.datasciencecentral.com/profiles/blogs/the-7-most-important-data-mining-techniques>

[2] Hand, D. J., & Adams, N. M. (2014). Data mining. Wiley StatsRef: Statistics Reference Online, 1-7

[3] Phil Spector, using t-tests in R Originally for Statistics 133 was retrieved from https://statistics.berkeley.edu/computing/r-t-tests

[4] Learntek (February 8, 2019) Data Mining Examples and Techniques was retrieved from <https://www.learntek.org/blog/data-mining-examples-and-techniques/>

[5] Hand, D. J. (2007). Principles of data mining. Drug safety, 30(7), 621-622