**Final Project: Preliminary Analysis**

****

**Team**

**Sunil Raj Thota**

**Nalini Macharla**

**Date: 02/10/2021**

**Title: Preliminary Analysis**

**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**

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.

Linear Regression model defines the relationship between two variables by setting a linear equation to the observed data. One variable is explanatory, and the other variable is dependent. Dependent is a variable whose values we want to explain or forecast. The explanatory variable is something that explains the other variables and to find if the values are independent.

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.

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

Now, let us talk about regularization techniques. It is a method for automatically penalizing the extra features that you use in your model. There is one type called as LASSO regression and here is the formula:

Minimize SSE +

**Conclusion**

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, factoextra, NbClust, party, caret, dplyr, tidyr, fpc, tidyverse, ggplot2, e1071, gmodels, gridExtra.

From the above observations, analysis, and model implementations we are looking to train and test the model and predict with an accuracy of ~85%. 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] Learntek (February 8, 2019) Data Mining Examples and Techniques was retrieved from <https://www.learntek.org/blog/data-mining-examples-and-techniques/>

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