**BANK DATASET Report**

We’re provided with the Bank dataset and asked to predict the output of outcome variable y, yes or no, to predict whether a client will subscribe to a term deposit. Since the outcome variable, y, is a dichotomous variable we create a few different models by using Logistic regression, K-Nearest Neighbors and Classification trees, and then compare the accuracy of the models we create in order to determine the model that is best suited for use in this case.

We begin by checking for any missing values or duplicate values in the dataset before we proceed with our analysis. After we determine that the data is clean, we conduct some exploratory data analysis to see how different factors given in the model affect the outcome variable y. Below are some of the different bar plots that we have created to visualize the association, if any, between certain factors and the variable y.

We conduct correlation analysis and note that there isn’t anything of significance shown by it. There are weak positive correlations between Age & Balance, Day & Campaign, and a strong positive correlation between Previous & pdays.

Chart, treemap chart

Description automatically generated

We determine that individuals that work in management or individuals that are entrepreneurs are less likely to subscribe to the term deposit. However, it is more unlikely than it is likely, that a person does not subscribe to a term deposit regardless of their occupation. We can prove this by checking what the predominant output is for variable y in the dataset, of which 88.476% are No and 11.524% are yes.

We perform further exploratory data analysis to determine the predominant age group of individuals in the sample. A large portion of the individuals are 20 and 60 years old, and the highest frequency is within the age range of 30-45 years old. Most of the individuals are married, and a very small percentage of them are divorced, while there is a group of individuals that are single as well. Majority of the individuals have completed secondary or tertiary education, only a very small percentage of them have only obtained primary education.

Chart, bar chart, histogram

Description automatically generated

Most individuals do not have a personal loan and nearly 60% of people have a housing loan.

We convert the given attributes to either Numeric or Factor form to make it easier to incorporate them into models. We convert outcome variable y into a binary variable where the output is now 0 or 1, as opposed to yes or no,

**Logistic Regression:**

We create a logistic regression model, using the training and testing datasets that we defined, to help predict the output of y as the outcome variable is dichotomous.

A picture containing graphical user interface

Description automatically generated

We obtain an accuracy of 90.3% as shown above after forming a table containing the actual value and predicted value that we obtain through the application of the model. We create an ROC curve to further validate the strength of this model. If the Area under this curve is high, the model is highly significant and very useful. We obtained an AUC value of 0.886, indicating that the model is indeed strong.

Chart

Description automatically generated

**Classification Trees:**

We use the training and test dataset created to form a classification tree, which can also be used to derive a prediction for variable y, that we will use to create a model in order to predict whether an individual will subscribe to a term deposit or not.

Diagram, timeline

Description automatically generated

We produce a confusion matrix for the model created to assess the accuracy and viability of this model.

A screenshot of a computer

Description automatically generated with low confidence

The model has an accuracy of 89.37%. The sensitivity is high at 96.44% but the specificity is 34.19% which is subpar. A value closer to 1 for both, specificity and sensitivity, would indicate maximum accuracy of the model.

**K-Nearest Neighbors:**

We perform KNN where we use the training dataset as the input to train a model for prediction of the output variable using this method. A confusion matrix is created:

A picture containing text, receipt

Description automatically generated

The accuracy that we obtain from this model is 88.05%. The sensitivity once again is high; however, the specificity is even lower than that which we obtained with the use of classification tree. We use cross-fold validation to validate the above model.Text

Description automatically generated with medium confidence

**Conclusion:**

We create 3 models to predict whether a client will subscribe to the term deposit or not, one using logistic regression, another using Classification trees and the final one using K-nearest neighbors. After obtaining the three models and measuring for accuracy of each of them, we conclude that the best model for use is the Model 1, created using Logistic Regression. It has an accuracy of 90.3%, with an AUC value that supports the fact that it is an accurate and strong model. It has a higher accuracy rate than the other two models we have produced and will most accurately predict our outcome variable y based on different factors that are determined using reliable methods such as stepwise regression.