## 

West Texas A&M University

Paul &Virginia Engler College of Business

Data Mining Final Project: Bank Term Deposit Prediction Models



# Executive Summary

Direct marketing, which promotes specific products to customers that have been identified as likely to buy those products, has proven to be a very successful marketing strategy in the banking sector due to the increasingly competitive nature of the industry (Parlar & Acaravci, 2017). Banks spend billions of dollars annually on marketing efforts and campaigns. For example, JP Morgan & Chase Co. reported over $3 billion in marketing expenses in 2021 (JP Morgan Chase & Co., 2021). Direct marketing can lower expenses and increase profits for banks by focusing marketing efforts on selling highly profitable products, such as term deposits, to the customers who have the highest probability of buying or signing up for those products. Term deposits are a significant avenue of revenue for banks because they are able to invest that money into products with a higher rate of return or loan it out to other customers at a higher interest rate, which makes them an ideal product for a direct marketing campaign (Chen, 2022). The objective of this project is to use classification data mining techniques to create a product that will help banks save money and increase profits by identifying and predicting which customers will most likely subscribe to a bank term deposit direct marketing campaign.

The data set used in this project was retrieved from www.kaggle.com (Moro, et al., 2011). It contains 45,211 records with 16 attributes, plus 1 output attribute or desired target, and no missing attributes. The data is based on phone calls made for direct marketing campaigns of a Portuguese banking institution to assess if the product (bank term deposit) would be (or not) subscribed.

To achieve the objective, we use several data mining techniques to identify the attributes of the customers most likely to subscribe to the product. Data preparation tasks were applied to check for missing values, anomalies, and outliers. RapidMiner, R, and Microsoft Excel were used to run descriptive statistics on the data set and check for correlation patterns. Using Rapid Miner and R’s splitting functionality, the data set was split to use 90% for training and 10% for prediction. Four classification techniques - decision tree, Naïve Bayes, logistic regression, and neural network - were applied to the data set. A confusion matrix was then used to evaluate the performance of each model.

The results of the data analysis indicate that the neural network model performed better than the other classification models in both accuracy and precision, making it the ideal classification model for predicting customers’ response to a direct marketing campaign for term deposits. The model will be a useful product for bank marketing teams.

### **I. Introduction**

Since the early 1990s when data mining began to proliferate due to the wide availability of databases, data warehousing, and commercial software products, financial institutions have been able to gather and store valuable demographic and fiscal information about their customers ([Linoff](https://www.wiley.com/en-us/search?pq=%7Crelevance%7Cauthor%3AGordon+S.+Linoff) & [Berry](https://www.wiley.com/en-us/search?pq=%7Crelevance%7Cauthor%3AMichael+J.+A.+Berry)*,* 2011). By using data mining techniques on their in-house data and by also adding to their data set by commercially available databases of demographic data - for instance, Axciom’s Infobase - banks can target the customers who are most likely to respond positively to their marketing campaigns (Infobase, 2022). Term deposits (or CDs) are desirable for the bank because the money is locked in at a set interest rate and the funds cannot be withdrawn without penalty. Banks primarily make money when they lend the money in their portfolio and charge interest to loan customers (Sang, 2022). The lending rates are higher than the CD deposit rates. This is why it is advantageous for banks to have deposits in the bank that have a high probability of not being withdrawn. There are fees for early withdrawal and some banks charge fees for maintaining the accounts as well (KochiesBiz, 2016). While the interest rate is higher than a traditional bank account, it is a product with few to no transactions that provides an expected amount of deposits for the bank.

There is a need for cost efficiency when running marketing campaigns to gain new customers of term deposit accounts because efficiency impacts all types of stakeholders, from transactional stakeholders such as customers and deposit-holders to cultural stakeholders such as regulators, within the banking organization. The bank is responsible for operating in a manner that preserves profits for stakeholders and delivers the most value to its customers. Any marketing campaign has costs involved, and the success of the marketing campaign is directly measured by how much money was spent on the campaign when looking at its profitability (Chou, et al., 2000). The significant labor hours involved in determining the customers most likely to open term deposits drives up the cost of customer acquisition. By evaluating a customer data set via data mining, the hours it takes to manually review volumes of customer data points can be reduced (Raj, 2015). This lowers costs by allowing marketers to focus efforts on the most likely individuals to open term deposits with the bank.

For the scope of this exercise and review, the goal is to use decision tree, Naïve Bayes, logistic regression, and neural network classification models to understand the customers within the data set that will be most likely to open a term deposit account with a particular bank. A successful evaluation will reduce the overall number of calls made by a bank’s marketing team, and customers that are contacted will find the product appealing. In turn the positive response rate will be higher over the course of the campaign.

### **II. Research Methodology and Modeling**

For research and study purposes, the team performed data analytics – data mining process to examine the customers who are likely to reject or accept bank term deposits by using different predictive models, like the decision tree, Naïve Bayes, logistic regression, and neural network to predict the probability of bank customers’ behavior on term deposit subscriptions.

###### *Data preparation and wrangling*

Preparation Goal: Ensure the variables and the attributes from the data sets were cleaned and ready for our models to use. Statistical values of the attributes are analyzed and decisions made for missing attributes, outliers, and other significant issues in the data. Taking the above essential and prerequisite steps of preparation allows the group to effectively analyze the data, decrease errors, and eliminate bias results from poor data quality and also inaccuracies that can occur to data during the modeling process (Talend, 2022).

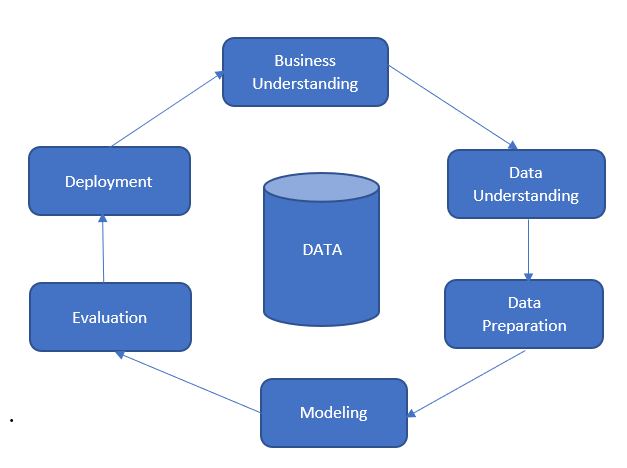


###### *Data Mining Techniques*

Data mining techniques provide different perceptions to business problems. Understanding the type of business problem will determine the type of data mining technique that will be suitable for that business. Understanding the bank term customers behavior and motives will draw more insights to data mining techniques for their solution and will result in more effective marketing strategy for the bank. Based on the bank business objectives and results from data preparation, the team will apply predictive modeling which will allow us to use different suitable approaches of data mining techniques to analyze the historical data records of the bank customers and how likely these customers will extend these behaviors into the future in regards to bank marketing response. The prediction modeling will combine other mining techniques like classification, probability, decision analysis, sequential patterns and clustering. We will use these techniques to analyze the relationship between the response or dependent (attributes) variables to predict the independent (target) variable with the data sets. This will help us better understand how the dependent variable's normal value changes when one of the independent variables is different. In contrast, the other independent variables remain unchanged (Sarkar, 2022).

The analyst will also incorporate the CRISP-DM, which stands for (Cross Industry Standard Process for Data Mining) Methodology, to help ensure that the DM life cycle is incorporated into our project and that we meet industry standards.

Figure 1: CRISP-DM Methodology

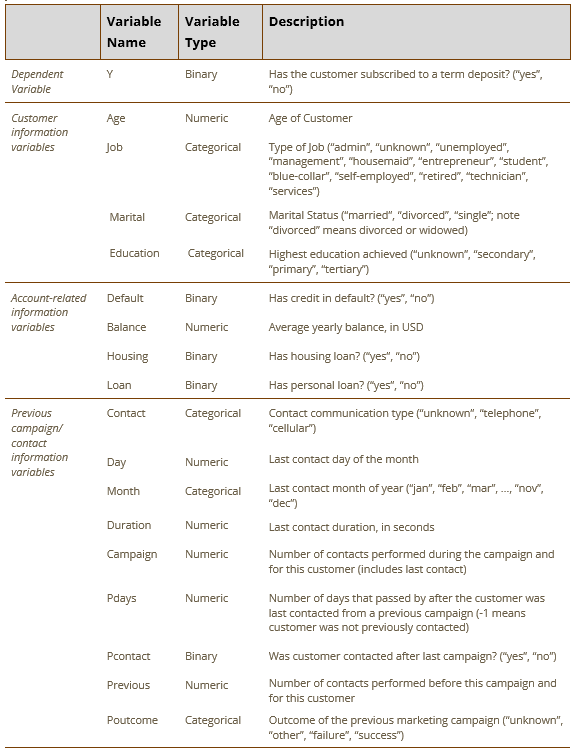


*Source: (Smart Vision Europe, 2020)*

### **III. Data Set Description**

The data set used for this project, Bank Direct Marketing, was retrieved from the www.kaggle.com website (Moro, et al., 2011). The original data is related to direct marketing campaigns of a Portuguese banking institution. The direct marketing campaign is used to make phone calls to bank customers. A customer can be contacted more than one time during the process to see if that customer will sign up for the bank product term deposit. The training data set has 45,211 examinations with 17 attributes (1 dependent attribute and 16 independent attributes). There were no missing attributes. The data set was split to use 90% for training and 10% for the prediction data set. The splitting of the data set was accomplished inside each model using Rapid Miner and R’s splitting functionality. The data set can be broken up into categories below. The categories include the dependent variable, customer information variables, account-related information variables, and previous campaign/contact information variables.

Table 1: Attribute Names and Description



### **IV. Data Preparation**

### To further explore and understand the data, descriptive analysis using Rapid Miner and R were used. This gave a better understanding of the min, max, mean, and standard deviation of the numerical attributes. Table 2 provides details about the numerical attributes followed by categorical and binary attributes. The number of observations in the same groups are displayed for the categorical and binary attributes. Box plots of the numerical attributes provided further understanding.[[1]](#footnote-1)

Table 2: Descriptive Statistics Results

| **Variable** | **Number of records in each group** | **Min** | **Max** | **Mean** | **Median** |
| --- | --- | --- | --- | --- | --- |
| Balance |  | -8,019 | 102,127 | 1,362.27 | 448 |
| Age |  | 18 | 95 | 39 | 40.94 |
| day |  | 1 | 31 | 15.8 | 16 |
| duration |  | 0 | 4918 | 258.2 | 180 |
| campaign |  | 1 | 63 | 2.76 | 2 |
| pdays |  | -1 | 871 | 40.2 | -1 |
| previous |  | 0 | 275 | .58 | .58 |
| job | Blue-collar: 9732  Management: 9458  Technician: 7597  Admin: 5171  Services: 4154  Retired: 2264  Self-employed: 1579  Entrepreneur: 1487  Unemployed: 1303  Housemaid: 1240  Student: 938  Unknown: 288 |  |  |  |  |
| marital | Married: 27214  Single: 12790  Divorced: 5207 |  |  |  |  |
| education | Secondary: 23202  Tertiary: 13301  Primary: 6851  Unknown: 1857 |  |  |  |  |
| contact | Cellular: 29285  Unknown: 13020  Telephone: 2906 |  |  |  |  |
| month | May: 13766 Feb: 2649  Jul: 6895 Jan: 1403  Aug: 6247 Oct: 738  Jun: 5341 Sep: 579  Nov: 3970 Mar: 477  Apr: 2932 Dec: 214 |  |  |  |  |
| poutcome | Unknown: 36959  Failure: 4901  Other: 1840  Success: 1511 |  |  |  |  |
| default | Yes: 815 No: 44396 |  |  |  |  |
| housing | Yes: 25130 No: 20081 |  |  |  |  |
| loan | Yes: 7244 No: 37967 |  |  |  |  |
| pcontact | Yes:7951 No: 35791 |  |  |  |  |
| y | Yes: 5289 No: 39922 |  |  |  |  |

The following attributes were changed in the data set: the original data type of Y, which is now our target attribute, was changed from polynomial to binomial, since logistic regression measures probability based on binomial. The following tools were used to run the descriptive statistics analysis of our data set and check for missing values, anomalies, and outliers detection within the attributes: RapidMiner, R, and Microsoft Excel.

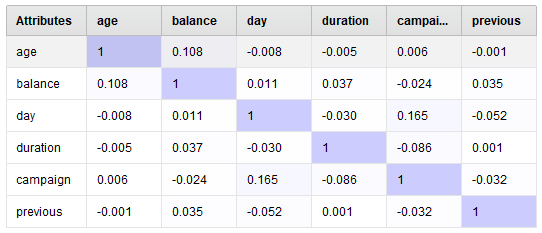
Further analysis of the standard deviation for the numerical variables indicated that outliers were present and needed attention.

The standard deviation for balance was 3,044.77, adjusting the range by 2 standard deviation units from the original min and max yielded a new range of -4,727.26 to 7,451.80. Removing the outliers decreased the data set by 1,469 records or 3.25%. The new data set contains 43,742 examples.

The standard deviation for age was 10.619, adjusting the range by 2 standard deviation units from the original min and max would yield a new range of 19.698 and 62.172; however it was decided to leave the age attribute values alone. The classification problem is centered around term deposits, which by definition “are an extremely safe investment and are therefore very appealing to conservative, low-risk investors” (Chen, 2022). It has been shown that as people age, they become risk averse, therefore outliers in age were not removed (Dohmen, et al., 2018).

The correlation matrix displayed no highly correlated items at or above 0.8, thus no further data was removed.

Figure 2: Correlation Matrix Between Attributes



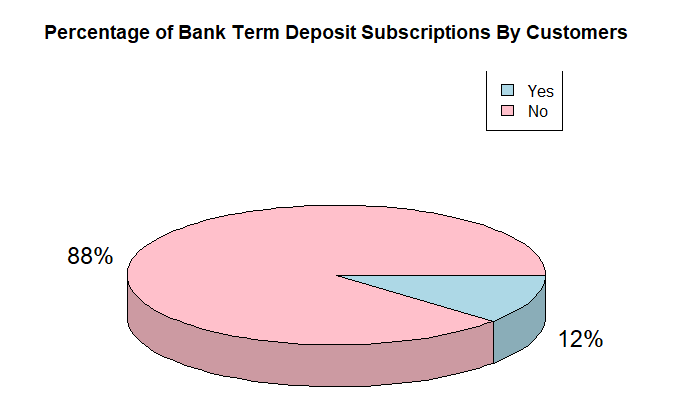
*Source: Analysis conducted in RM*

The attribute pdays is explained above as being the number of days that passed after the customer was last contacted from the previous campaign. The value is assigned -1 if the customer was not contacted. Over 50% of the customers were not contacted; therefore, pdays was transformed to a binary variable, pcontact. Pcontact is simply “yes” or “no” for contact or not. The transformation was achieved in Excel by adding a column, pcontact and putting a simple if statement to produce "yes" or "no" depending on the value for pdays. The if statement checked for values = -1 and assigned “no”, otherwise “yes”. The original column, pdays was then removed from the data set.

Once inside Rapid Miner, all previously defined binary attributes were updated to have a type of binomial instead of polynomial. This was consistent among all models and was especially needed for the logistic regression model which measures probability based on binomials. Also, inside R, all attributes imported as char were converted to factors for all models.

Running a general analysis of the data, the observation from the pie chart graph below indicates that for most of the bank customers contacted for term deposit subscriptions, about 88% said no, and 12% said yes.

Figure 3: Bank Term Deposits by Customers

****

### **V. Data Mining Modeling Results**

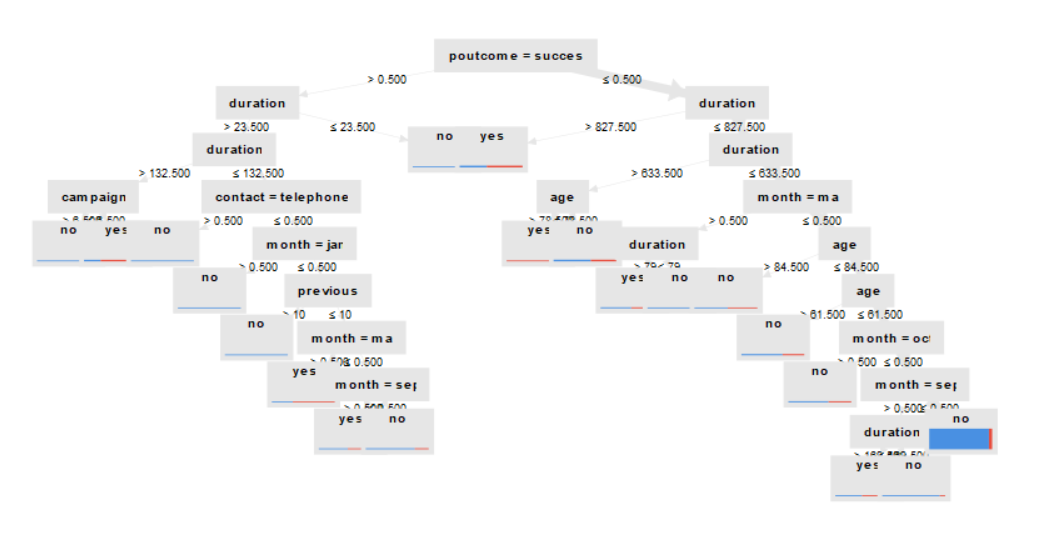
***Decision Tree Model***

A decision tree is a supervised learning model that will predict and classify the data with respect to the target attribute. The decision tree takes a large amount of data and divides it into smaller groups. Each time the data is split into a group, the more similar the data becomes. The data is presented in a tree format to show every possible output.

The decision tree is used in the problem to find how many customers will say yes to buy a term deposit. The decision tree predicted in RapidMiner 3,868 customers that replied – no – and 506 customers that would be willing to a term deposit - yes. It is worth noting that certain attributes affect the reply (e.g., duration).[[2]](#footnote-2) The decision tree predicted in R that 3,963 customers did not want a term deposit replied - no - and 412 customers wanted a term deposit replied - yes.

There is a difference in the predicted results because decision trees are sensitive to small variations of data. Decision trees have a high variance when it comes to the data being split. The data is being overfit due to it being split into different attribute combinations. As you can see in Figure 4, this proved to be true.

Figure 4: Decision Tree Visualization in RapidMiner



*Source: Rapid Miner*

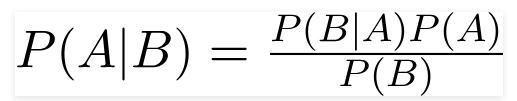
Based on the number of yes and no predictions we can further break down the data to form a confusion matrix. The confusion matrix tells us what the accuracy, sensitivity, and specificity is of the model. The accuracy is 90%, sensitivity is 62.12%, and the specificity is 92%. The specificity shows the percentage of true negatives. In relation to the problem 92% of customers **will** say no.

***Naive Bayes***

*“Naïve Bayes, is a wonderfully simple approach that often returns very accurate and stable models with very small sample sizes. The reason that Naïve Bayes often works so well is that it simplifies predictive modeling problems to avoid the curse of dimensionality” (Rice, 2014).*

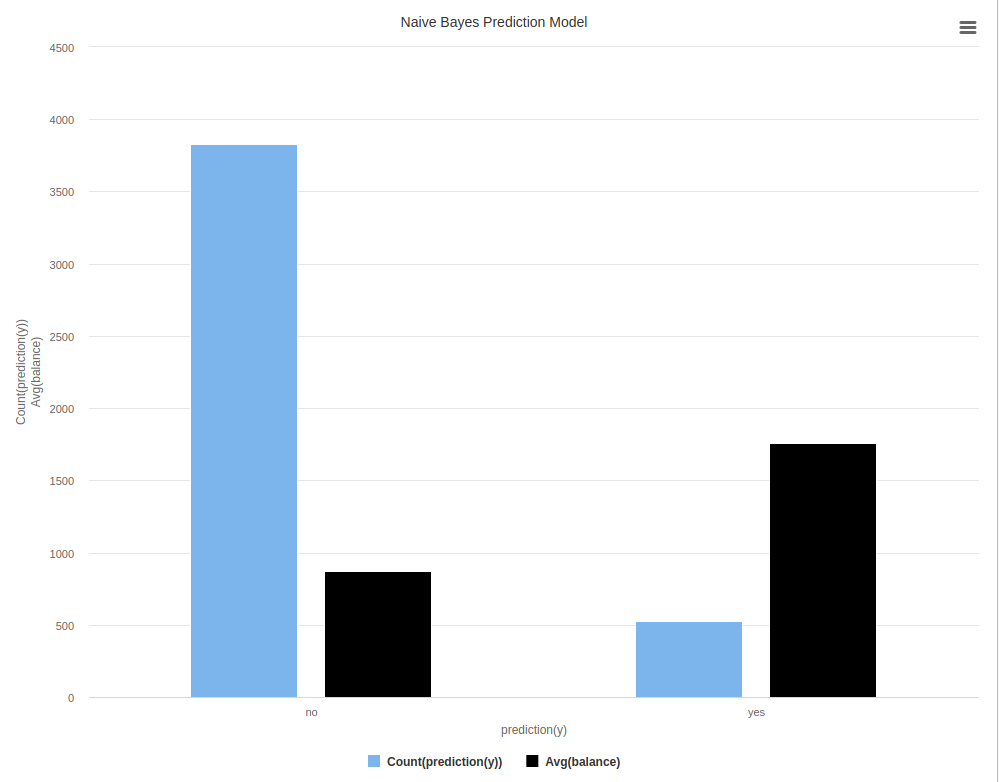
Rev. Thomas Bayes’ theories of probability were published in 1763. His primary interest was the probability of A given B, which is considered a conditional probability ([Linoff](https://www.wiley.com/en-us/search?pq=%7Crelevance%7Cauthor%3AGordon+S.+Linoff) & [Berry](https://www.wiley.com/en-us/search?pq=%7Crelevance%7Cauthor%3AMichael+J.+A.+Berry)*,* 2011, p. 211). The mathematical representation is:

Figure 5: Bayes Theorem



*Source: Gandhi, 2018*

When looking at the impact of various factors that influenced a Yes response, an attribute that influences the purchase is the Balance that a customer already holds:

Figure 6: Term Deposit Prediction and Influence of Yearly Bank Balance

Because term deposits or CDs require a minimum purchase of $1000 in most cases it is reasonable to understand that a customer with a higher yearly balance would be in a better position to make a purchase (Tumin, 2022).

Despite the independent nature of the attributes using the Naive Bayes Classifier, it was apparent that a higher yearly bank balance DID have an influence on the prediction of those who would purchase term deposits (Kumar, 2022). It is possible to review the results to find correlations and indicators that would be useful in understanding if the customer is a good target for marketing, but this is also a drawback of the model. It is two dimensional, and requires that a thorough comparison of each attribute be done against the prediction outcome. While this is possible, it increases the need for manual oversight, which defeats the purpose of using a data-mining model.

***Logistic Regression***

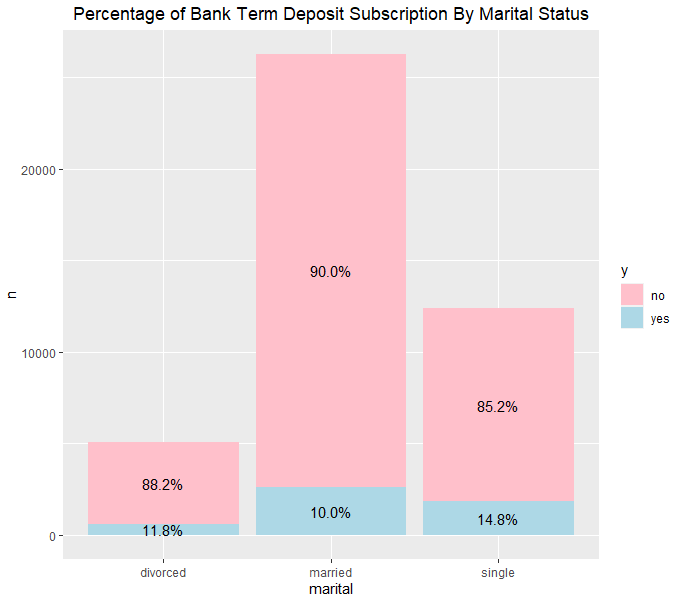
One of the methodologies used in our project is the logistic regression model because the problem we are trying to solve is related to the classification of Yes and No variables. Logistic regression is the classic model in terms of classification.

*“A logistic regression is a statistical method that relates a dependent variable to one or more independent (descriptive) variables. A logistic regression model is used in presenting changes observed in the dependent variable are associated with changes in one or more of the descriptive variables.”* (Beers, 2022).

The group performed a data analytics – data mining process to examine the customers who are likely to reject or accept bank term deposits by using logistic regression as one of the models to predict the probability of bank customers’ behavior on term deposit subscriptions.

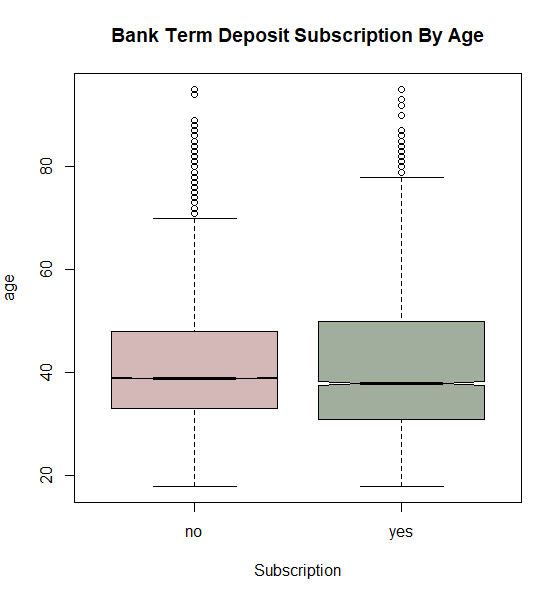
To build the model for a better prediction result and also to reduce the possibility of overfitting our data, the data was split into two subsets, one for training the model and the second for testing the model’s performance. The training data consists of 39,367 observations, and the testing data consists of 4,375 observations. The classification’s main objective is to predict which bank customers will likely sign up for a term deposit subscription. The target variable used is the attribute (“Y”) with a yes and no value. We choose other attributes like education, martial, age and further analyze and run them for individual evaluations to see how they responded to the bank term deposit subscription.

Figure 7: Percentage of Marital Status Distribution By Subscription



*Source: (RStudio)*

Based on the above bar chart, customers who are married are contacted for bank term deposit subscription and 90% of them will likely say no and the rest of the 10% from that group are likely to say yes. The unknown marital status number was less than 0.2% and was excluded from the bar chart.

Figure 8: Age Distribution By Subscription

*Source: (RStudio)*

After running different analysis on the attributes, we now fit the models using the training data, we set the random seed to 123 to make the sampling reproducible and check for the ratio of the training set. We now build the logistic regression model using the train set and apply the model to our test set for the probabilities result by using the confusion matrix function. The overall accuracy at 90%, with sensitivity at 98% and specificity at 33.4%. Based on the model result the accuracy rate is high, and the sensitivity is high meaning that it predicts fewer false positives.

In conclusion, using the logistic regression model to predict the probability of bank customers likely to subscribe to a long-term bank deposit. The model produced a higher sensitivity rate that should focus on because of its positive effects. This should be the primary significance of the model, not to compare the specificity rate, which is low and also a true negative rate.

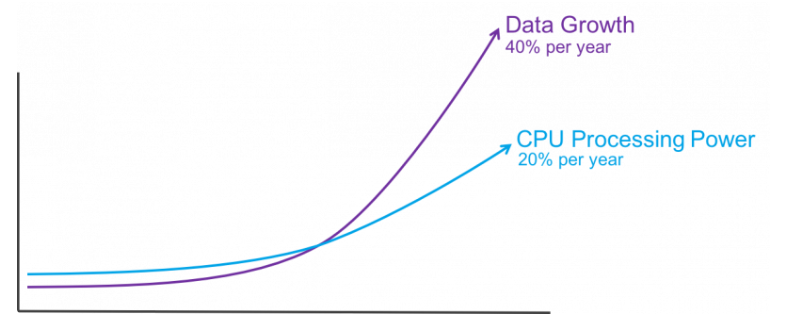
***Neural Network:***

The Neural Network model is one of the powerful, flexible models used for prediction and classification. Neural networks attempt to mimic the workings of the very powerful and complex human brain and are often called Artificial Neural Networks. The building blocks of the human brain are biological neurons. Neural networks utilize artificial neurons to compare attributes in order to discover strong connections among them. Before neural networks, computers were simply used to execute explicit instructions over and over. Neural networks have given computers the ability to learn. Neural networks indeed mimic the human brain’s ability to generalize and learn when they are fed well-defined domains or input. They are very useful in data mining and several successful examples use this model. ([Linoff](https://www.wiley.com/en-us/search?pq=%7Crelevance%7Cauthor%3AGordon+S.+Linoff) & [Berry](https://www.wiley.com/en-us/search?pq=%7Crelevance%7Cauthor%3AMichael+J.+A.+Berry)*,* 2011, pp. 281-320)

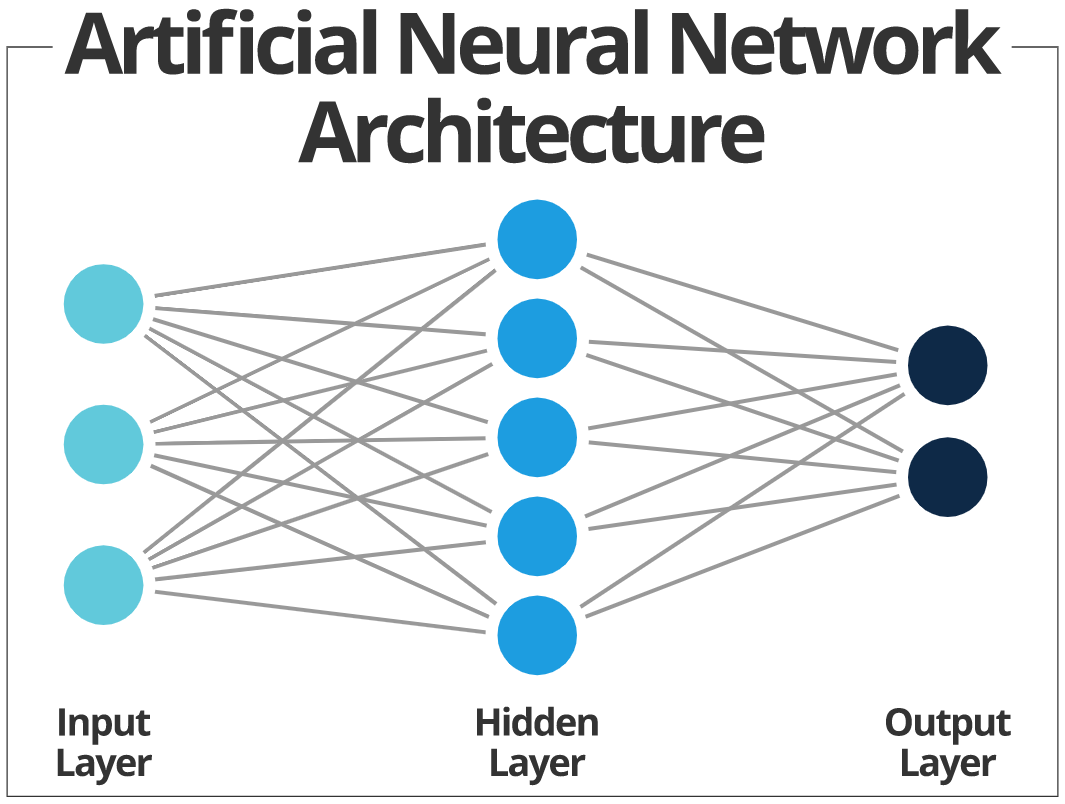
Neural networks come with their share of advantages and disadvantages. A major advantage is that neural network models generate better results most of the time. A key disadvantage is that the internal workings of the model are very complex. Inside the neural network training lies a complex process of assigning weights to all of the units in the network. After the weights are determined, they are compared back to the actual results. If they are different, adjustments are made and the process starts again. ([Linoff](https://www.wiley.com/en-us/search?pq=%7Crelevance%7Cauthor%3AGordon+S.+Linoff) & [Berry](https://www.wiley.com/en-us/search?pq=%7Crelevance%7Cauthor%3AMichael+J.+A.+Berry)*,* 2011) Because of the internal complexity of neural networks, the focus is centered on the inputs and outputs. The internal, complex workings of the neural network is treated as a black box. Neural networks are also more computationally expensive compared to the other models listed in this project. ”State of the art deep learning algorithms, which realize successful training of really deep neural networks, can take several weeks to train completely from scratch. By contrast, most traditional machine learning algorithms take much less time to train, ranging from a few minutes to a few hours or days.” (Donges, 2019). Given our data set size and number of attributes, execution time was around just over a minute. This seems like a reasonable time, but compared to only a couple seconds for the other models, there is a substantial difference. Another disadvantage of neural networks is the amount of data required to properly train, which is usually more than the other methods listed in this project. (Donges, 2019).

If a sufficient data set is available for the neural network today and it is computationally feasible today, will it be in the future? If the company’s data set grows faster than the speed of their cpu processing power, this would result in a model that would cost more to run as the data set grew. Figure 9 is a representation of data growing faster than cpu processing power.

Figure 9: Data Growth vs CPU Processing Power



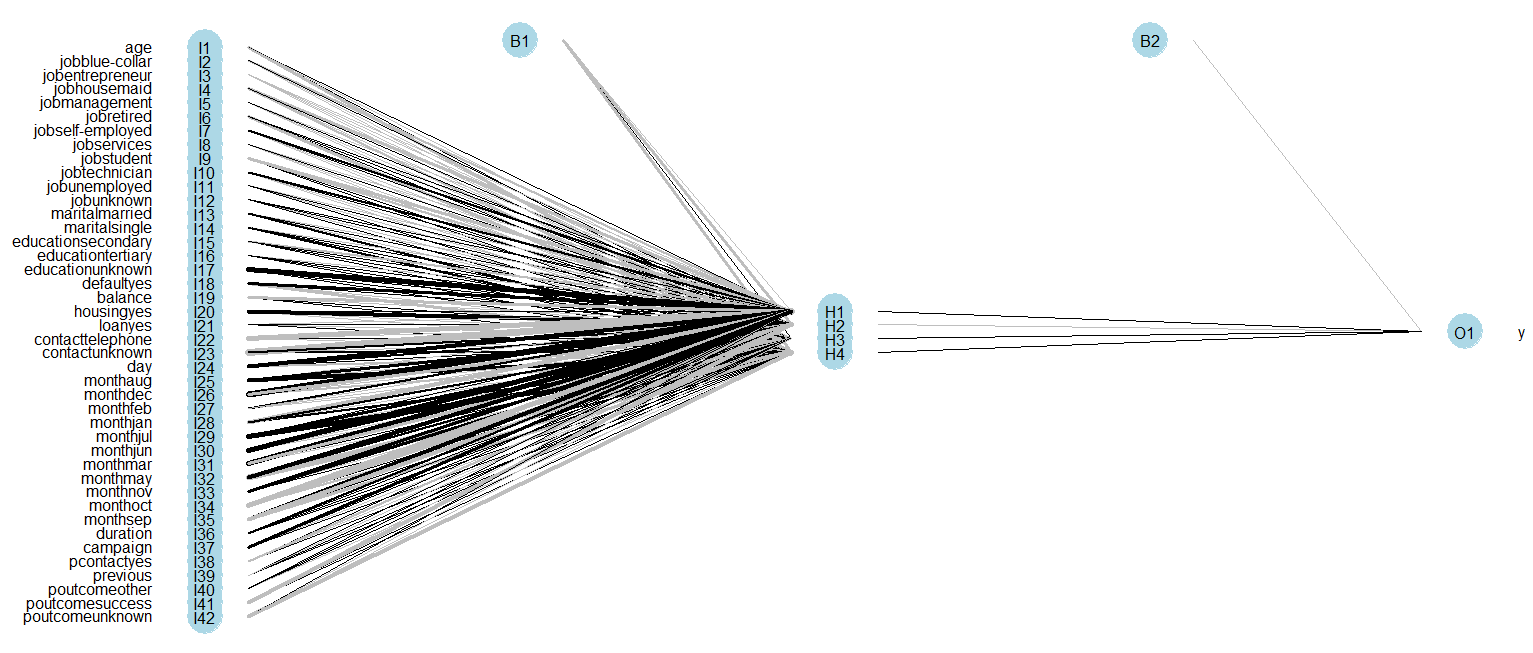
*Source: (Mostak, 2019)*

Figure 10: Neural Network Architecture

*Source: (Smartsheet, n.d. )*

For the neural network model, the seed was set to 1,000 to remain constant with each run, with y as the target attribute and other attributes as predictor attributes. The size parameter indicates the number of nodes to use in the hidden layer. Size was set to 1 in RM and 4 in R. Maximum iterations was set to 10000. The prediction results were classified into “yes” and “no” based on a probability of greater than.5.   
 The plot of the neural network generated in R is below in Figure 11. Notice that there are 4 hidden layers and that the bias is also shown. Also notice the thicker and darker neurons, lines between the nodes. The thicker and darker the neurons between nodes represent the strong affinity between the nodes (North, 2018). This model did not use all 10,000 iterations as it converged just after iteration 860 in R.

Figure 11: Neural Network Visualization in R



*Source: R Studio*

The neural network models produced an accuracy of 90.95% in RapidMiner and 91.02% in R. Neural networks typically are better at finding the strength of relationships between attributes (North, 2018). Data preparation resulted in a data set that contains desired attributes and values to use in a neural network to find such strengths of relationships. This model produced 506 Yes and 3,868 No in Rapid Miner and 509 Yes and 3,866 No in R.

With the emphasis on neural networks being the input and output, great care must be taken when selecting the input and the preparation of the input. Neural networks “are most easily trained when input fields have been mapped to a small range close to zero.”([Linoff](https://www.wiley.com/en-us/search?pq=%7Crelevance%7Cauthor%3AGordon+S.+Linoff) & [Berry](https://www.wiley.com/en-us/search?pq=%7Crelevance%7Cauthor%3AMichael+J.+A.+Berry)*,* 2011, p. 319) Reflecting on this and looking at initial runs of the models yielding lower than expected accuracy rates, all numerical values were normalized to a range between 0 and 1. Normalizing the data and altering the size of the hidden layers produced better results than initial model runs. Rapid Miner’s best accuracy was produced with a hidden layer with size 1 at 90.95% accuracy. R’s accuracy increased as the hidden layers increased. The size of 1 yielded an accuracy of 88.37%, size 2 yielded 90.83%, size 3 yielded 90.99%, and finally size 4 yielded 91.02%. Increasing to size 5 resulted in a lower accuracy of 90.65%. As the size increased, so did the number of iterations to convergence. It was noted that the higher the size, the more iterations and processing of the neural network for convergence.

### **VI. Evaluation**

###### *Modeling benchmark*

In evaluating the models’ performance, a confusion matrix will be used. The matrix will help identify the customers that will respond to getting a term deposit. Accuracy, precision, recall, and F-measure are the indicators that will be used to find the most effective model at predicting the responses. Values in the tables below were calculated and produced in Rapid Miner, R, and Excel. The confusion matrix indicates that the neural network is the optimum model choice when compared to the logistic regression, decision tree, and Naïve Bayes models, in both RM and R.

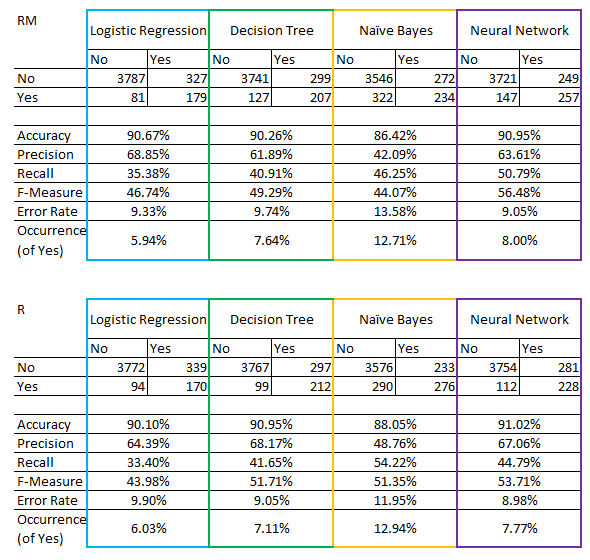
In regard to the three indicators, accuracy, precision and recall, the neural network is the most appealing model based on the RM confusion matrix. Accuracy and precision are higher than the other models with 90.95% and 63.61% respectively.[[3]](#footnote-3) Following closely behind the neural network are the logistic regression and decision tree models. These models performed at 90.67% and 90.26% accuracy respectively. Of note, the logistic regression model had the highest performance in precision with 68.85%. Having a high precision value with a low recall value creates a bit of tension. Precision identifies how many (yes) answers there are, including false positives, whereas recall identifies how many actual (yes) answers there are, excluding false positives. The logistic regression model performed the best in the aspect of recall with a mark of 35.38% Another indicator is needed to evaluate both precision and recall, this is where F-measure comes in. F-measure gives all the models the same weight for precision and recall, with this indicator the model with the highest value is the neural network, with 56.48%(RM).

The second confusion matrix in R determined that the neural network is also the best model. Given that the neural network and accuracy is valued at 91.02% and decision tree is 90.95%, the two make them the top models in R. The precision indicator is second behind decision trees at 67.06%, and recall is also ranked second behind decision trees with 44.79%. Neural network recovers the top rank with the lowest Error rate of 8.98% and F-measure also indicates that neural network is the best model with a value of 53.71%.

In both of the matrices, Naïve Bayes failed to be the most valuable in any of the indicators. If Naïve Bayes were to be chosen there would be a low accuracy, would be expensive and perform poorly in the direct marketing campaign. Thus, Naïve Bayes must be eliminated as a possibility for deployment. This is also made evident in the graphical representations of accuracy and error rates.[[4]](#footnote-4)

In comparing both of the matrices the neural network model has been a constant in the top two best models making the neural network model the best model for deployment.

**Confusion Matrix and Descriptive Statistics Based on Testing in RapidMiner and R**

****

When reviewing the Lift Charts across each of the 4 DM models[[5]](#footnote-5), we review the minimum confidence and the true case matches across the bins that the lift chart produces. To understand the lift chart confidence we must understand that a random guess is valued at 0.1. Therefore, any value greater than 0.1 is considered significant.

The neural network confidence is .473 in the first bin with true case matches of 269. This bin provides the greatest lift. Bins 2 and 3 add 108 and 68 true cases, respectively, though the confidence in these bins falls below the random guess value for both bins.

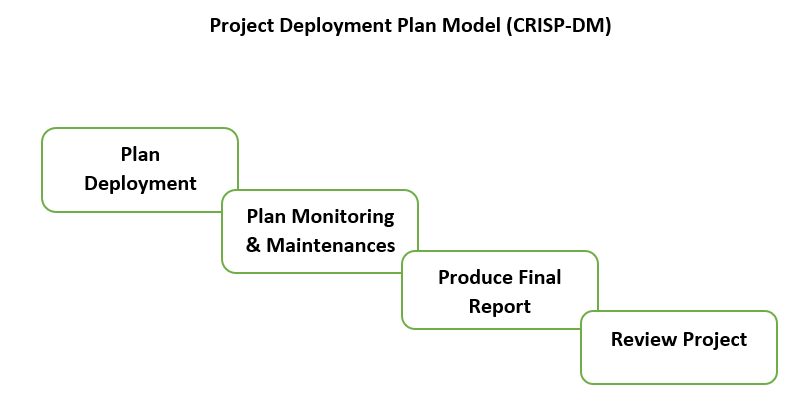
As we look over the lift charts for logistic regression and Naive Bayes, we find that both of those models produced significant lift in the first and second bins, and both have an extremely high confidence level in the first bin, but then drop significantly in confidence thereafter. We can see that true cases for the LR model are the highest at 378 for bins above the 0.1 mark, but for Naive Bayes, true cases are only 199 from the first bin, which is the only one above the random guess level.

The lift chart for the decision tree model did not produce a similar 10 bin output despite the same settings being used in the modeling in RapidMiner. This is because the lift chart identifies the performance of the leaves which can vary greatly in size. If the model was used in this data mining exercise, the lift chart would serve to identify the leaves continuing the most likely positive outcomes. The confidence level of the first 12 bins greatly outperforms the 0.1 random guess mark with the lowest confidence at .221 and the greatest at 1.

###### VII. Discussion

In conclusion, because cost efficiency and profitability are priorities in the banking industry, the ability to predict customers’ responses to a direct marketing campaign for term deposits is an invaluable tool for developing a successful campaign. By applying the classification data mining techniques to the Portuguese bank data set, we were able to produce a neural network model that can predict the likelihood of a customer responding “yes” to purchasing a term deposit through direct marketing with a high level of accuracy and precision. In order for banks to use this model to improve their direct marketing campaigns, the next step must include developing a strategic deployment plan for this model (see Figure 13).

Figure 13: Deployment Plan Model



First, changes to the RM process and R code should be implemented, including changing RM to be connected to a file that is not static. Both RM and R models already have logic to check for highly correlated attributes and remove them if they are above +-.80. This portion might be executed with refreshed and newer data sets. Next, several pilot tests, including focus groups, must be conducted, and feedback from subject matter experts (SMEs) must be reviewed. The model must then be adjusted based on the feedback received. Finally, the model must be reviewed to ensure it follows the organization’s Standard Operating Procedure.

Once deployment is planned and documented, the model must be monitored. Maintenance must be performed regularly to ensure that the model is operating properly and to deal with any issues that may arise. After this step, a final report must be produced for the bank to review, and a meeting should be scheduled to discuss the model’s overall performance, including what went well and what could be improved upon.

Limitations of this study include the lack of information regarding the specific term deposit rates, as well as the status of the economy during the date range that the data collection occurred. There is also no research provided about whether or not regional or country differentiation should be considered when determining the importance of each attribute. For future improvements, these factors should be considered.[[6]](#footnote-6) Additionally, during the analysis, while all the attributes in the data set were used, some attributes were found to be irrelevant in predicting how many people would say yes to getting a term deposit. To determine the weight each attribute had on “y,” the Explain Predictions operator in RM was applied. When used in all the models, the operator determined that “Day” and “Loan” were irrelevant to predicting “y.”[[7]](#footnote-7)

### 

### **References**

Beers, B. (2022). *What is Regression? Definition, Calculation, and Example*.

Investopedia. Retrieved from: https://www.investopedia.com/

terms/r/regression.asp

Chen, J. (2022). *Time Deposit: Definition, How It's Used, Rates, and How to Invest*.

Investopedia. Retrieved from: https://www.investopedia.com/terms/t/

termdeposit.asp#:~:text=Term%20deposits%20are%20an%20extremely,Deposit%20Insurance%20Corporation%20(FDIC).

Chou, P., Grossman, E., Gunopulos, D., & Kamesam, P. (2000). Identifying

Prospective Customers. [Paper presentation]. Sixth International Conference on Knowledge Discovery and Data Mining, Boston, MA, USA. doi:10.1145/347090.347183.

Dohmen, T., Falk, A., Golsteyn, B., Huffman, D., & Sunde, U. (2018). *Identifying the*

*Effect of Age on Willingness to Take Risks*. CEPR. Retrieved from: https://cepr.org/voxeu/columns/identifying-effect-age-willingness-take-risks

Donges, N. (2019). *4 Reasons Why Deep Learning and Neural Networks Aren't*

*Always the Right Choice*. Built In. Retrieved from: <https://builtin.com/data->

science/disadvantages-neural-networks

Gandhi, R. (2018). *Naive Bayes Classifier*. Towards Data Science. Retrieved from:

https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c

Infobase. (2022). Acxiom. Retrieved from: https://www.acxiom.com/customer-data/

infobase/

JPMorgan Chase & Co. . (2021). *JPMorgan Chase & Co. Annual Report 2021*. JP

Morgan Chase & Co. Retrieved from: https://www.jpmorganchase.com

/content/dam/jpmc/jpmorgan-chase-and-co/investor-

relations/documents/1q21-10-q.pdf

KochiesBiz. (2016). *Term Deposits: Simple, Safe & Secure Way To Invest* [Video].

Youtube. https://www.youtube.com/watch?v=uKXUbMGLrYc

Kumar, N. (2022). *Naive Bayes Classifiers*. GeeksforGeeks. Retrieved from:

https://www.geeksforgeeks.org/naive-bayes-classifiers/

# [Linoff](https://www.wiley.com/en-us/search?pq=%7Crelevance%7Cauthor%3AGordon+S.+Linoff), G.S., & [Berry](https://www.wiley.com/en-us/search?pq=%7Crelevance%7Cauthor%3AMichael+J.+A.+Berry)*,* M.J.A*.* (2011). *Data Mining Techniques: For Marketing, Sales,*

# *and* *Customer Relationship Management, 3rd Edition.* Wiley Publishing, Inc.

Moro, S., Laureano, R., & Cortez, P. (2011). Using Data Mining for Bank Direct

Marketing: An Application of the CRISP-DM Methodology. *Proceedings of the European Simulation and Modelling Conference - ESM'2011*, pp. 117-121, Guimarães, Portugal, 2011. Retrieved from: https://www.kaggle.com/

datasets/psvishnu/bank-direct-marketing?group=bookmarked

Mostak, T. (2019). *What is a GPU and why should Telcos Care?* Design

World. Retrieved from: https://www.designworldonline.com/what-is-a-gpu-

and-why-should-telcos-care/

North, M. (2018). *Data Mining for the Masses, Third Edition: With Implementations*

*in RapidMiner and R*. Infinite.

Parlar, T., & Acaravci, L. (2017). Using Data Mining Techniques for Detecting the

Important Features of the Bank Direct Marketing Data. I*nternational Journal of Economics and Financial Issues, 7(2)*, 692-696.

Raj, S. (2015). A Review of Data Mining Applications in Banking. [Paper

presentation]. International Conference on Recent Advances in Electronics, Computer Science and Information Technology, Chennai, India. doi:10.4236/ce.2015.615165.

Rice, D. (2014) Calculus of Thought. Academic Press.

https://doi.org/10.1016/B978-0-12-410407-5.00004-0.

Sang, A. (2022). *How Do Banks Make Money?* Clever Girl Finance. Retrieved from:

https://www.clevergirlfinance.com/blog/how-do-banks-make-money/

Sarkar, P. (2022). *What is Regression Analysis? Types, Techniques, Examples*.

KnowledgeHut. Retrieved from: https://www.knowledgehut.com/blog/

data-science/regression-analysis-and-its-techniques-in-data-science

Smartsheet. (n.d.) *Real-life Applications of Neural Networks*. Smartsheet. Retrieved

from: https://www.smartsheet.com/neural-network-applications

Smart Vision Europe. (2020). *CRISP DM methodology*. Smart Vision Europe.

Retrieved from: https://www.sv-europe.com/crisp-dm-methodology/

Talend. (2022). *What is Data Preparation? Processes and Example*. Talend.

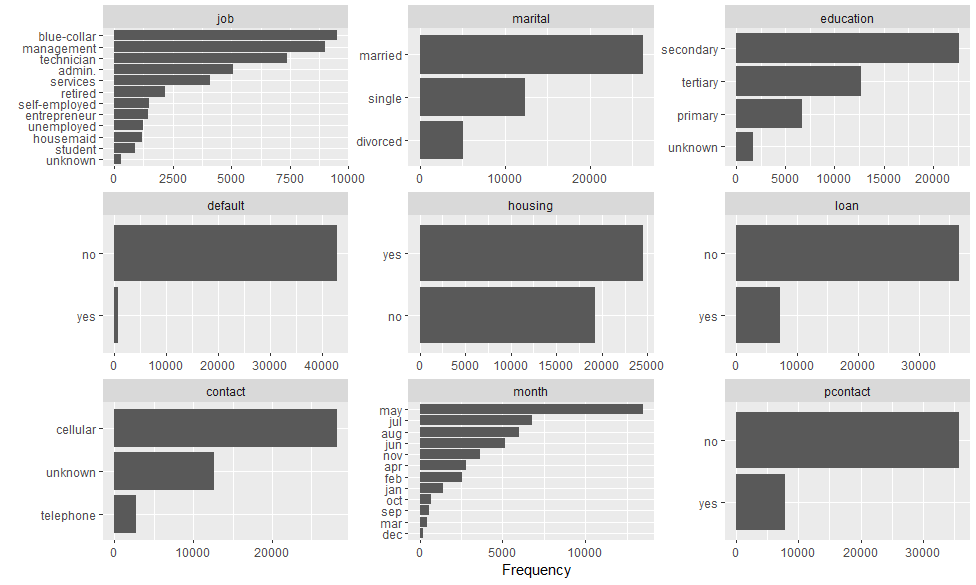
Retrieved from: https://www.talend.com/resources/what-is-data-preparation/

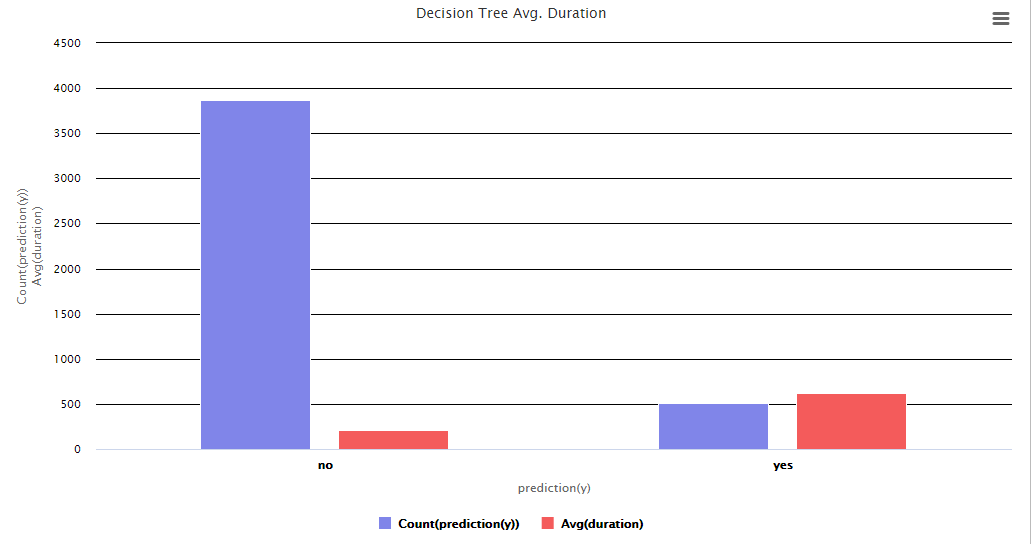
Tumin, K. (2022). *Best CD Rates* . Deposit Accounts by LendingTree. Retrieved

from: https://www.depositaccounts.com/cd/

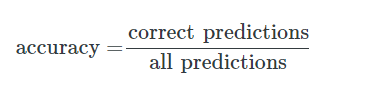
**Appendices**

Appendix 1: Boxplots of Numerical Attributes



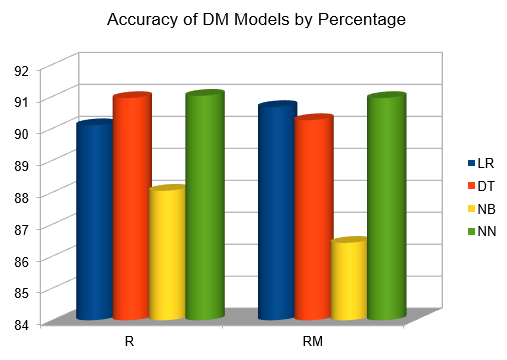
Appendix 2: Decision Tree Model Bar Chart in RapidMiner

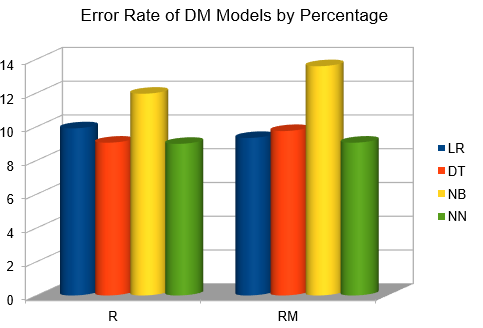
Appendix 3: Accuracy Formula and Calculation



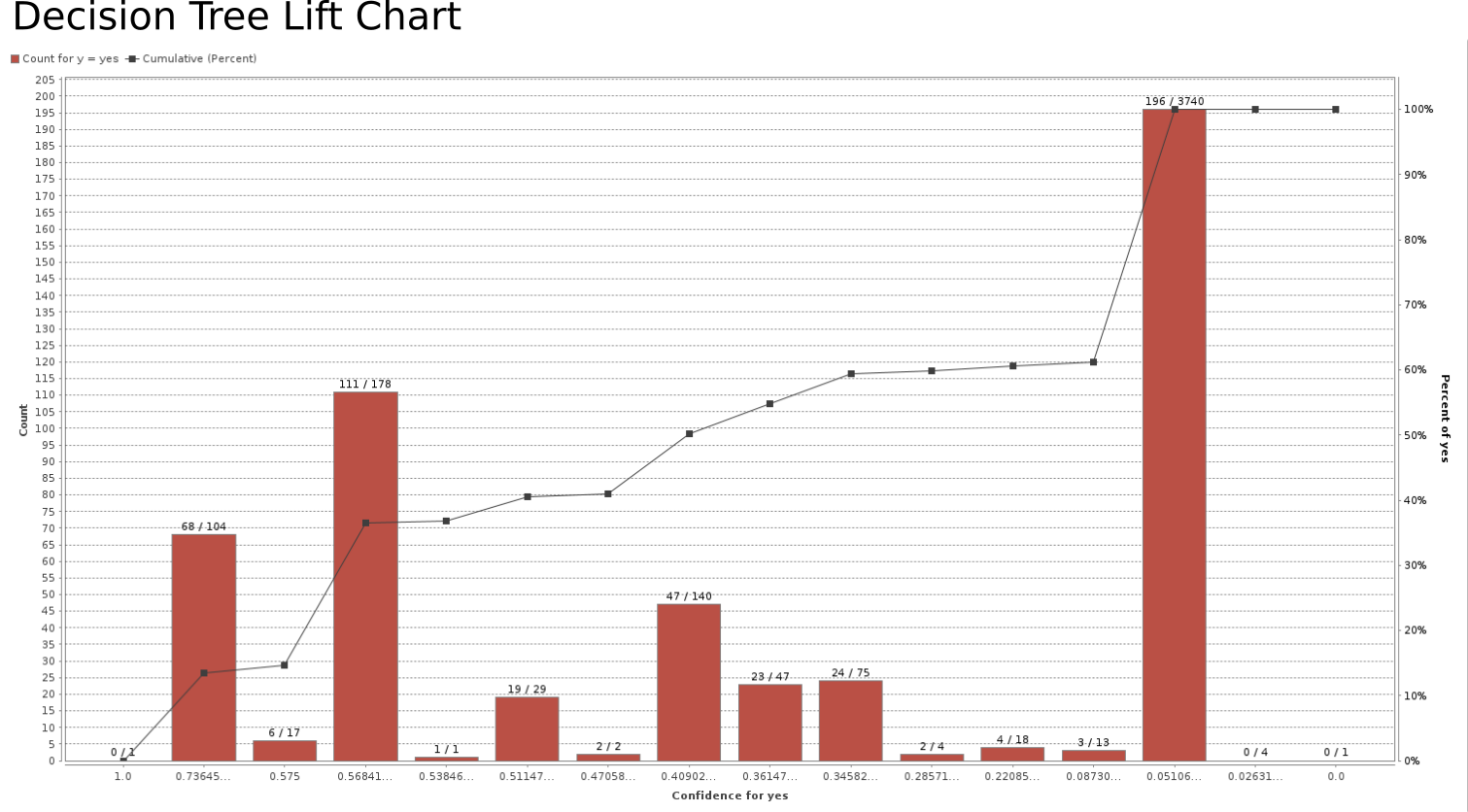
**A = (3772 + 170) /4375 = 90%**

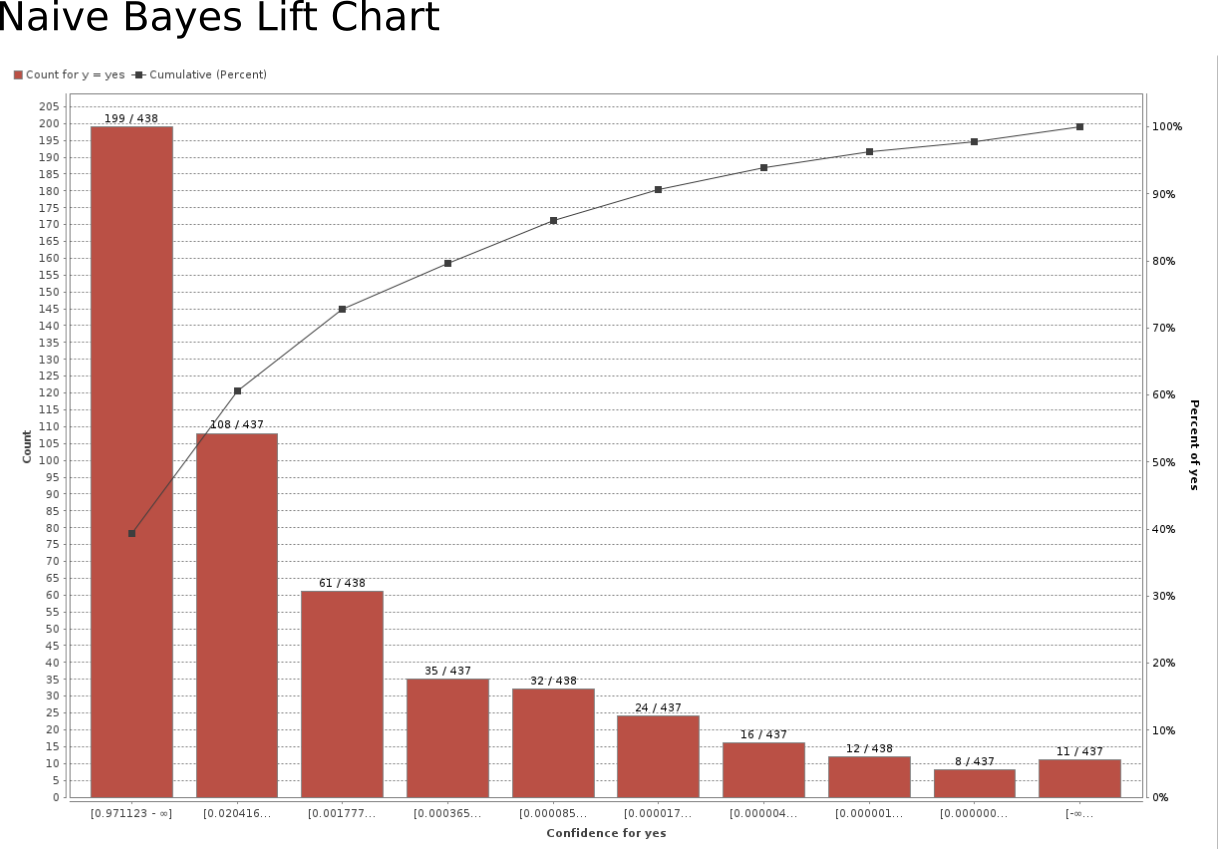
Appendix 4: DM Models by Percentages

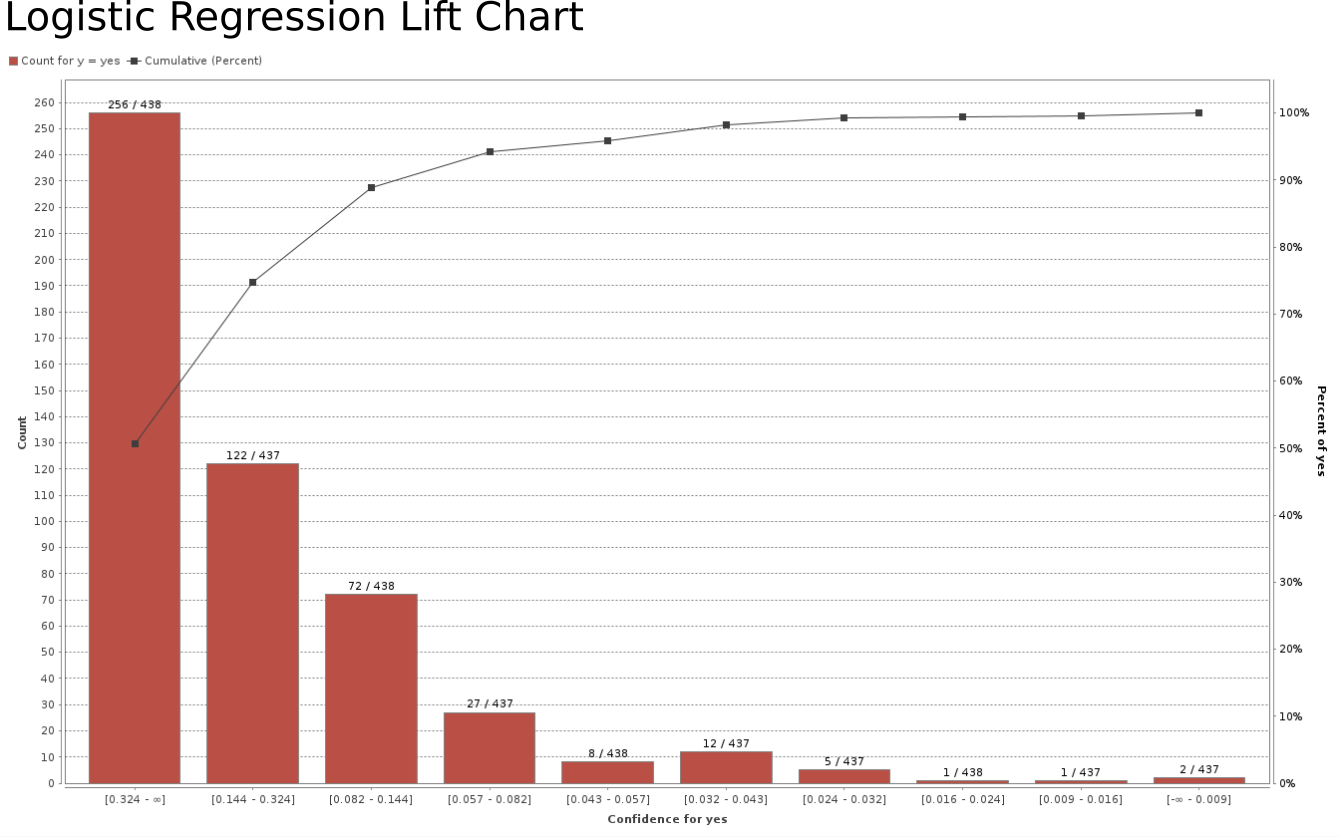
****

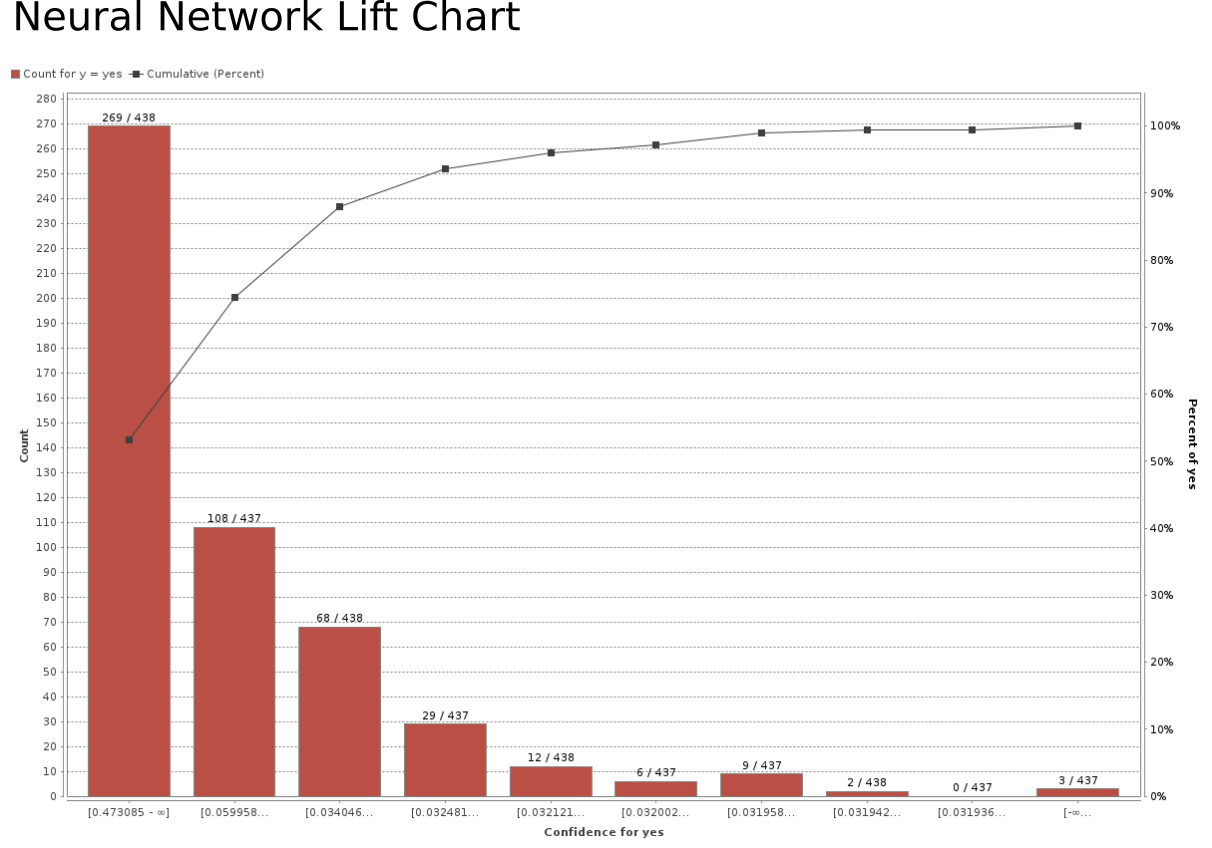
****

Appendix 5: Lift Charts for 4 DM Models

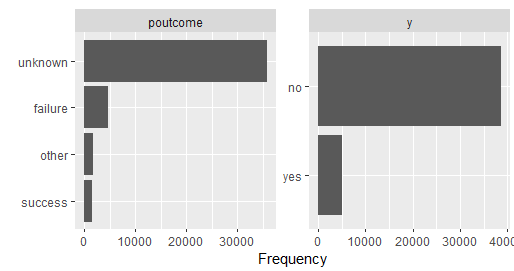




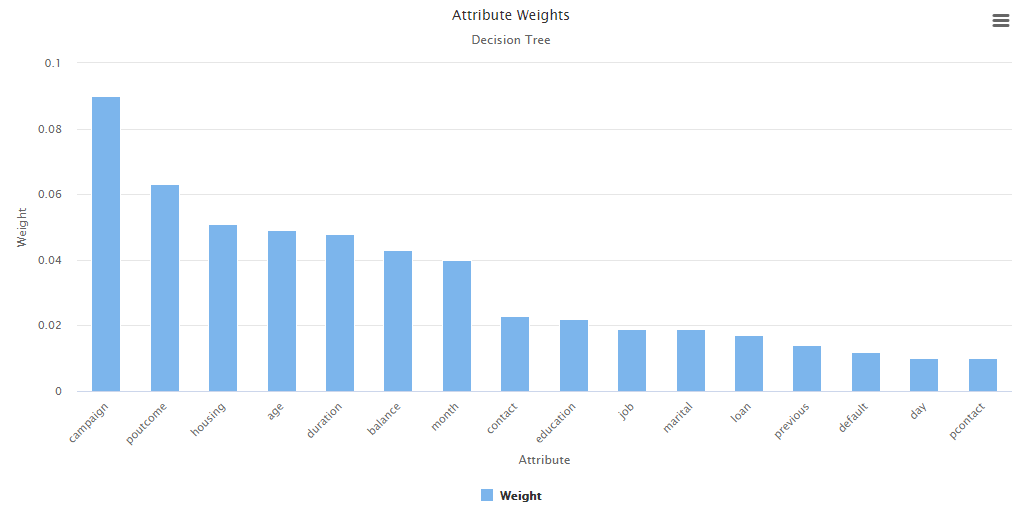




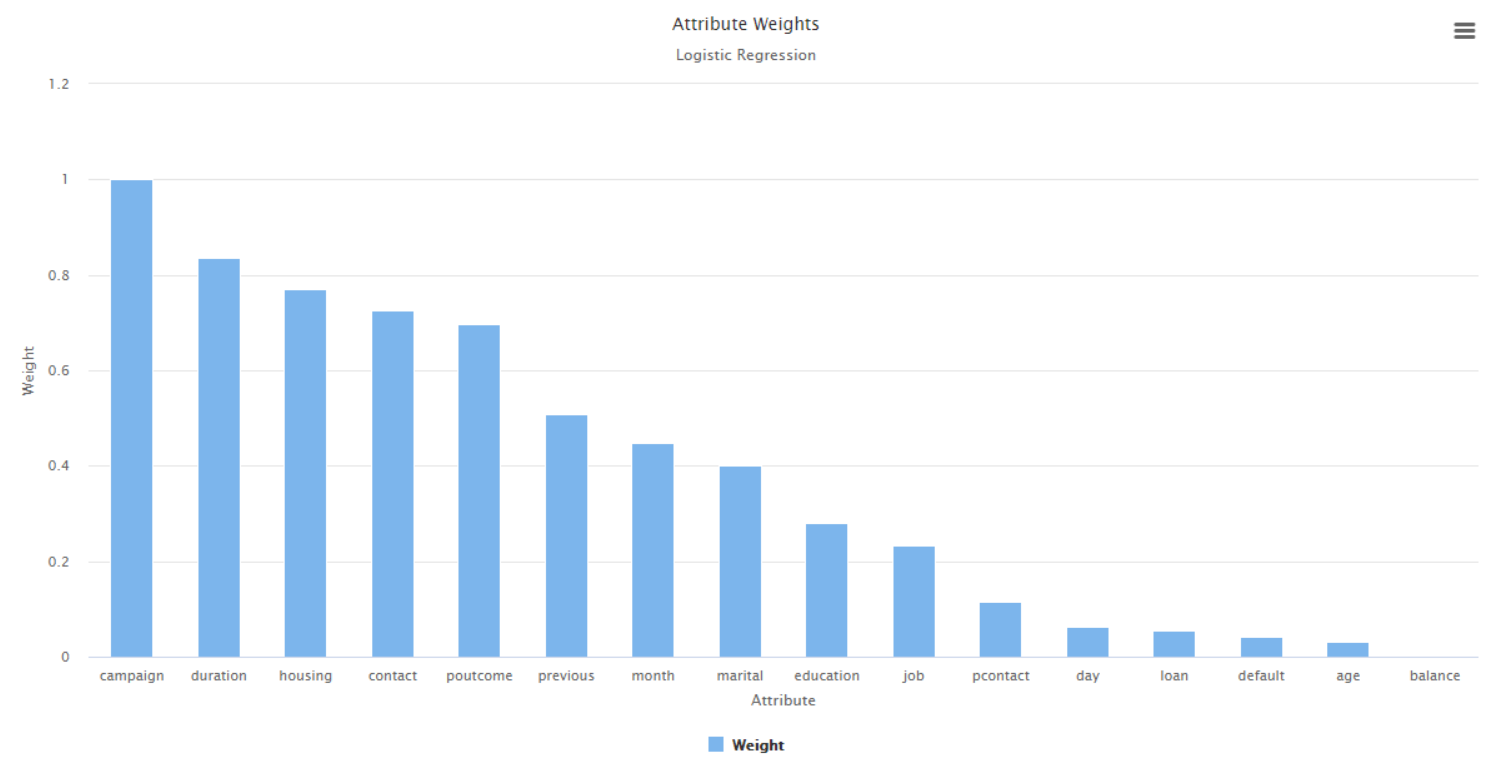
Appendix 6: (Title)



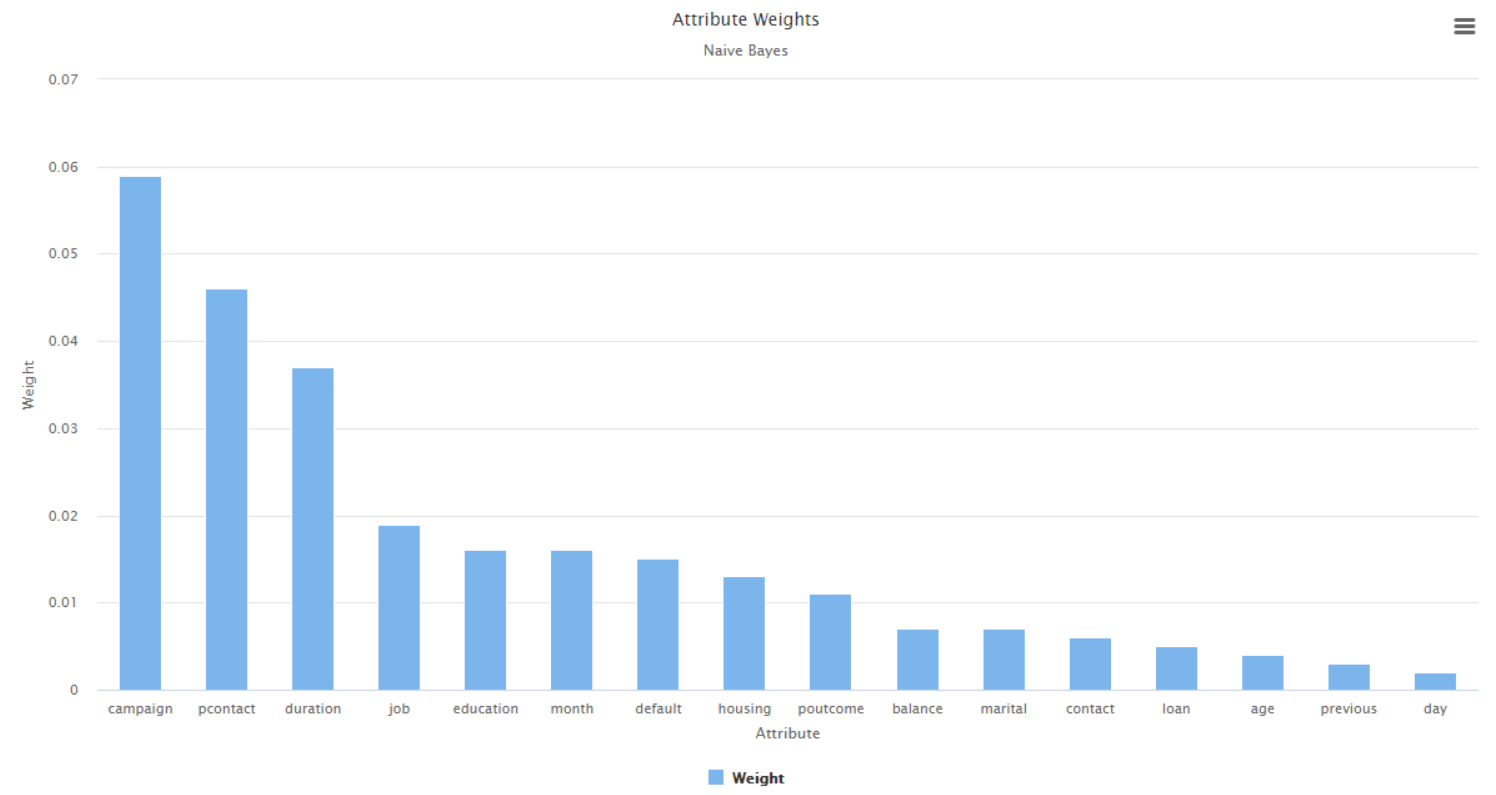
Appendix 7: Attribute Weights for Decision Tree

**

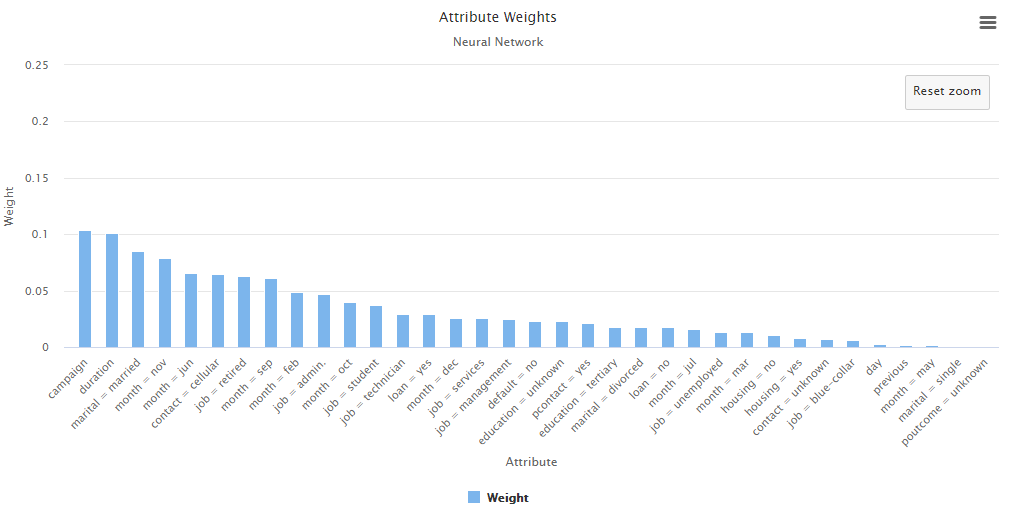
Appendix 8: Attribute Weight for Logistic Regression

**

Appendix 9: Attribute Weight for Naive Bayes

**

Appendix 10: Attribute Weight for Neural Network

**

**Link for Presentation**

<https://buffswtamu-my.sharepoint.com/:v:/g/personal/lsbracken1_buffs_wtamu_edu/ET1hcg7DkrxCittOC1kubEkBuUOou2NXKOXjAtqaR2gJoQ?e=PPl0Bb>

1. Refer to Appendix 1 for details [↑](#footnote-ref-1)
2. Refer to Appendix 2 for details [↑](#footnote-ref-2)
3. Refer to Appendix 3 for details [↑](#footnote-ref-3)
4. Refer to Appendix 4 for details [↑](#footnote-ref-4)
5. Refer to Appendix 5 for details [↑](#footnote-ref-5)
6. Refer to Appendix 6 for details [↑](#footnote-ref-6)
7. Refer to Apendices 7-10 for details [↑](#footnote-ref-7)