MARKETING ANALYTICS
FOR
RETAIL BANKING

By Nikhil Sharma

Capstone Research Project

Presented to

Ryerson University, The G. Raymond Chang School of Continuing Education
Data Analytics, Big Data, and Predictive Analytics Certificate Program
Department of Engineering & Architectural Science

Toronto, Ontario, Canada, 2019

Supervised by
Dr. Ceni Babaoglu

ABSTRACT

Banks are always trying to find ways to improve their revenue models based on customer needs and behaviour by personalizing their products and services to drive greater levels of customer satisfaction and retention. As a result, a bank is searching for creative solutions to address their revenue decline, which may be linked to a decrease in the number of term deposit subscriptions. When customers purchase a term deposit account at a bank, it is considered a fixed-term investment where the principal investment is protected with predictable returns, however, the funds can't be withdrawn for a certain length of time. Term deposits allow banks to make a profit by lending the money out to clients in loans and earning interest payments in return. Term deposits are also considered a short- or long-term investment that can be used for retirement, buying a car, or a home. Telemarketing was the direct marketing strategy that was used by the bank to acquire investments, and big data analytics can help enhance this approach by exploring what customer demographic variables such as age, education level, and employment status are resulting in sales opportunities for the bank. In this paper, we will propose a marketing strategy by building a predictive model that will identify customers who have a higher probability of subscribing to a term deposit. The dataset is from the UCI Machine Learning Repository and consists of 45, 211 data points and 17 attributes. Logistic Regression, Random Forest, K-Nearest Neighbors, and Neural Networks are the supervised machine learning algorithms used to build a predictive model in R that can identify the correlation between independent variable x and the response variable y. Throughout this report, data cleaning, imputing, pre-processing, training, and validation will be demonstrated.

1. INTRODUCTION

With the bank looking to optimize its sales and marketing strategy based customer profiles such as age, education level, marital status, and employment status, this paper aims to develop a classification model that will help answer how many customers the bank labelled as non-subscribers for a term deposit account actually subscribed to a term deposit, and of all of the customers who are subscribers, how many of those did the bank correctly predict. For each of the supervised machine learning techniques (Logistic Regression, Random Forest, K-Nearest Neighbors, and Neural Networks) used to build a predictive model, accuracy, precision, recall, F1 score, and specificity were the performance metrics analyzed. Functions in the caret package (Classification And REgression Training) in R were primarily used to execute data cleaning, imputing, preprocessing, training, and feature selection.

2. LITERATURE REVIEW

Ejaz, S. (2016, May). *Predicting Demographic and Financial Attributes in a Bank Marketing Dataset*. Retrieved from https://repository.asu.edu/items/38651

Samira Ejaz conducted a similar study on the bank telemarketing dataset. Ejaz uses supervised machine learning classification techniques such as Support Vector Machine (SVM, Random Forest, and Logistic Regression to determine the highest predictability for seven features, including age, employment status, marital status, education level, housing loan, personal loan, and term deposit. Ejaz uses the F1 score as the evaluation metric for the algorithms which is the average of precision and recall to evaluate the performance of each model. Furthermore, Ejaz concludes that the Logistic Regression model achieved the best F1 score fo the majority of the classes tested and that the customer's decision to subscribe to a term deposit was most influenced by the relationship they had with the bank.

Elsalamony, H. A. (2014, January). *Bank Direct Marketing Analysis of Data Mining Techniques*. Retrieved from https://pdfs.semanticscholar.org/a911/cbe221347b400d1376330591973bb561ff3a.pdf

Hany A. Elsalamony also performed a study on the telemarketing dataset from a Portuguese bank. Elsalamony uses supervised machine learning classification techniques such as the Multi-layer Perception Artificial Neural Network (MLPNN), Tree Augmented Naive Bayes (TAN), Nominal regression or logistic regression (LR), and Ross Quinlan new decision tree model (C5.0) to identify which attributes are influencing a client's decision to subscribe for a term deposit. Elsalamony uses classification accuracy, sensitivity and specificity to evaluate the performance of each model. Furthermore, Elsalamony concludes that 'Duration' achieved the highest importance in C5.0, LR, and MLPNN. However, 'age' achieved the highest importance by the TAN model.

Choong, A. (2017, October). *Predictive Analytics in Marketing A Practical Example from Retail Banking*. Retrieved from https://actuaries.org.sg/files/library/other/WebsitePracticeArea/17CommitteesAndPracticeAreas/BigDataFolder/ResearchNoteNo01bySASBDC.pdf

Alvin Choong conducts a comparative study on the telemarketing dataset from a Portuguese bank. Choong uses Logistic Regression, Regression Trees, and Random Forest to determine the predictability of each feature. Choong also applies several boosting techniques to increase the accuracy of classification, including Boosted GLM and Gradient Boosted Trees. Choong concludes that the customer's job title and education are the most important feature in the Random Forest model's accuracy. Furthermore, Choong mentions that 'day of week' also had high predictive power. Choong also mentioned a few valid points regarding the dataset, including how the data does not offer the cost of each phone call in the marketing campaign, rather the problem focuses on increasing the efficiency of the strategy with the assumption that it will result in cost reduction of marketing the product. The author also mentioned that there are biased assumptions of there being a direct correlation between the attributes rather than considering other factors in the bank's revenue decline such as the interest rate environment which could be influencing the client not to invest their money.

3. DATA OVERVIEW

The values within the dataset are distinguished as two data types: categorical and continuous. In the table below you will notice that the dataset can be categorized into five different sub-categories, including customer demographics, personal banking, current and previous marketing campaign, and the response variable.

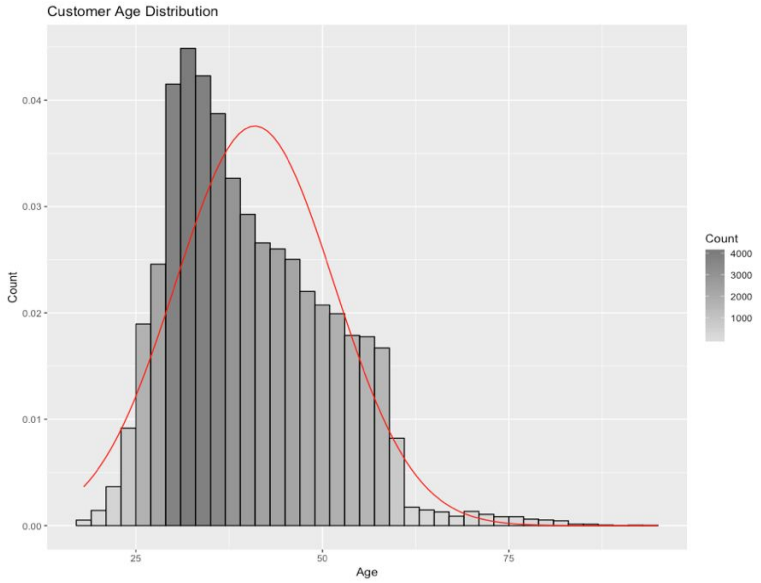
#	Attribute	Attribute Description	Data Type & Values
Customer Demographics			
1	age	client's age	numeric: [17, 98]
2	job	client's type of job	{categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown'}
3	marital	client's marital status	{categorical: 'divorced','married','single'}
4	education	client's education level	{categorical: 'primary','secondary','tertiary','unknown'}
Personal Banking			

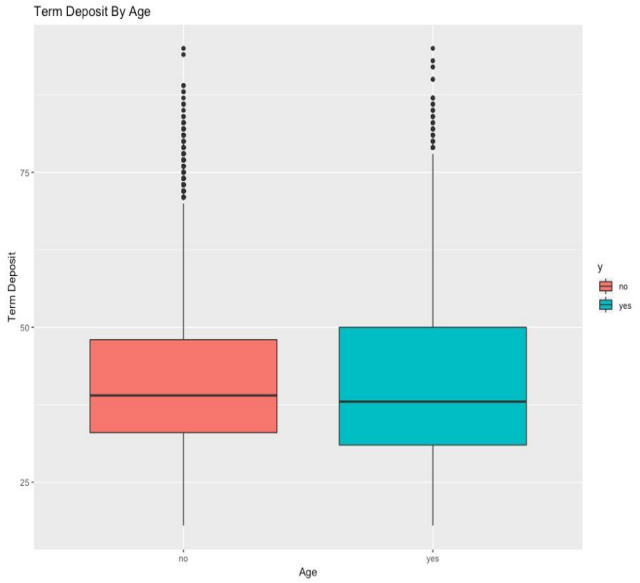
5	default	client has credit in default	{categorical: 'no','yes'}
6	balance	client's average yearly balance	numeric: [-8019, 102127]
7	housing	client has housing loan	{categorical: 'no','yes'}
8	personal	client has personal loan	{categorical: 'no','yes'}
Current Marketing Campaign			
9	contact	contact communication type	{categorical: 'cellular','telephone','unknown'}
10	day	last contact day of the month	numeric: [1, 31]
11	month	last contact month of the year	{categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec'}
12	duration	last contact duration (seconds)	numeric: [0, 4918]
13	campaign	number of contacts performed during this campaign and for this client including last contact	numeric: [1, 63]
Previous Marketing Campaign			
14	pdays	number of days after client was last contacted from previous campaign	numeric: [-1, 871], (-1 means client was not previously contacted)
15	previous	number of contacts performed to this client before this campaign	numeric: [0, 275]
16	poutcome	outcome of the previous marketing campaigns	{categorical: 'failure', 'other','success','unknown'}
Response Variable			
17	y	subscription to a term deposit	{binary: 'yes','no'}

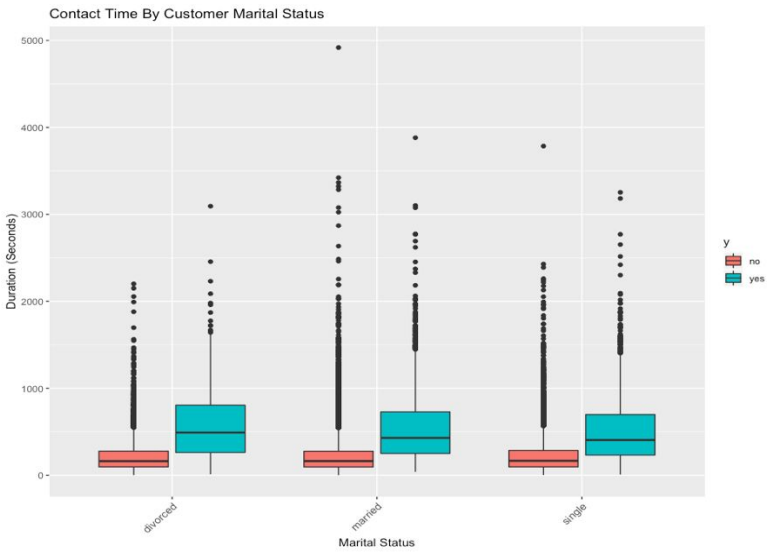
4. EXPLORATORY DATA ANALYSIS

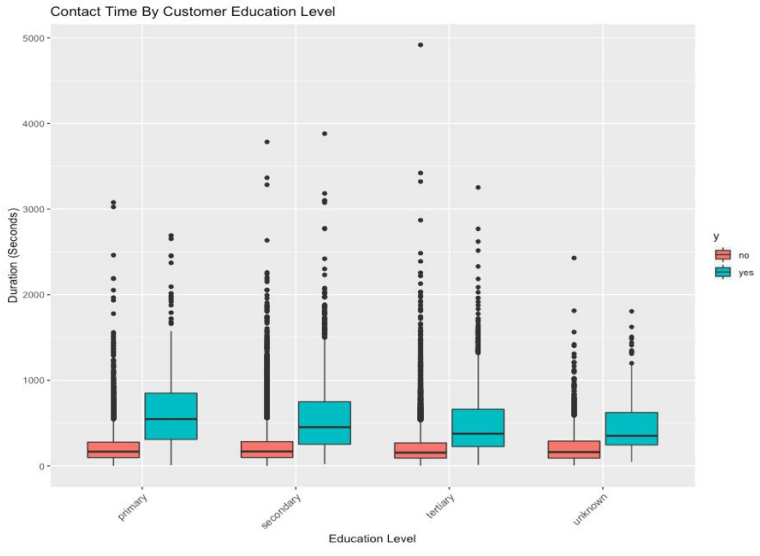
Exploratory analysis is completed to assess the quality of a dataset and to better understand how a dataset is structured. During exploratory analysis, categorical and continuous variables in the dataset will be evaluated to understand the distribution of each category and spread of each variable through univariate and bivariate analysis, respectively.

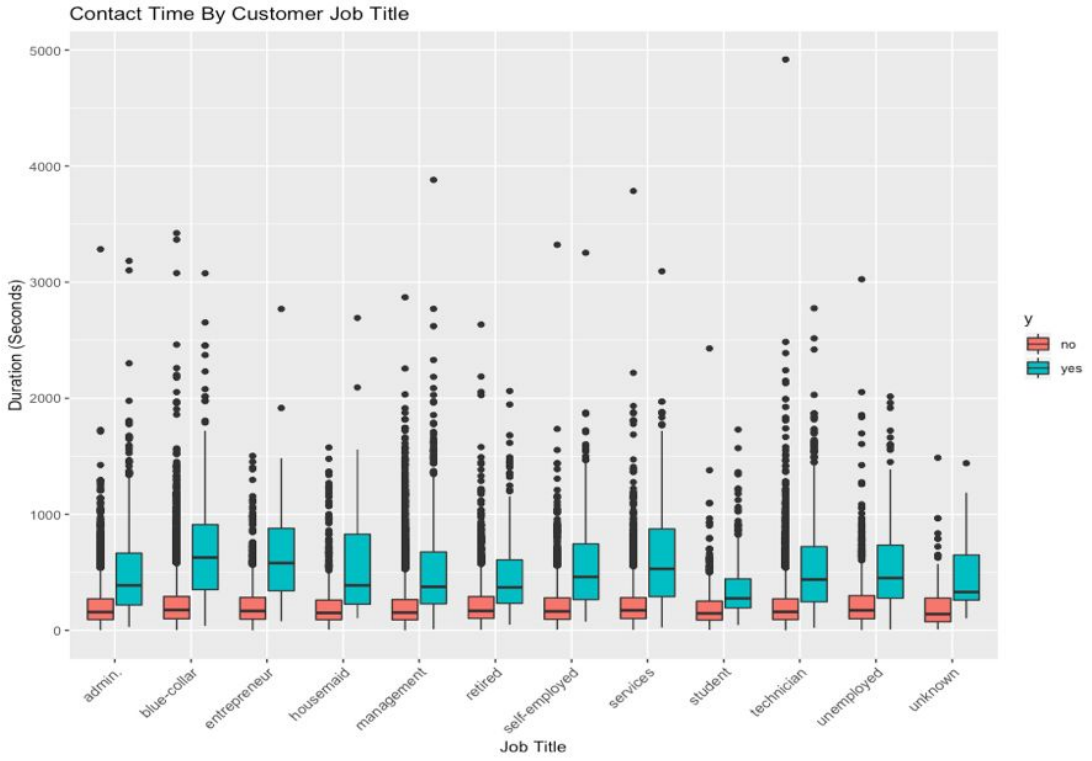
4.1 Customer Demographics

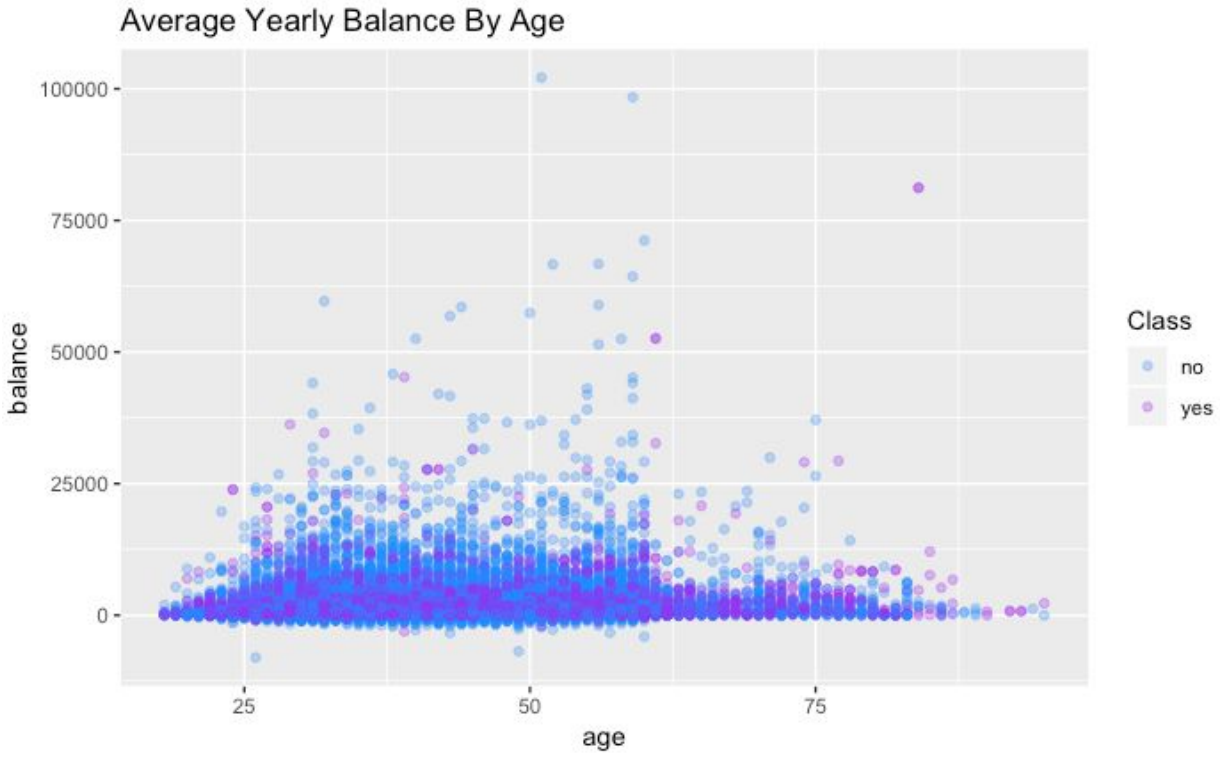


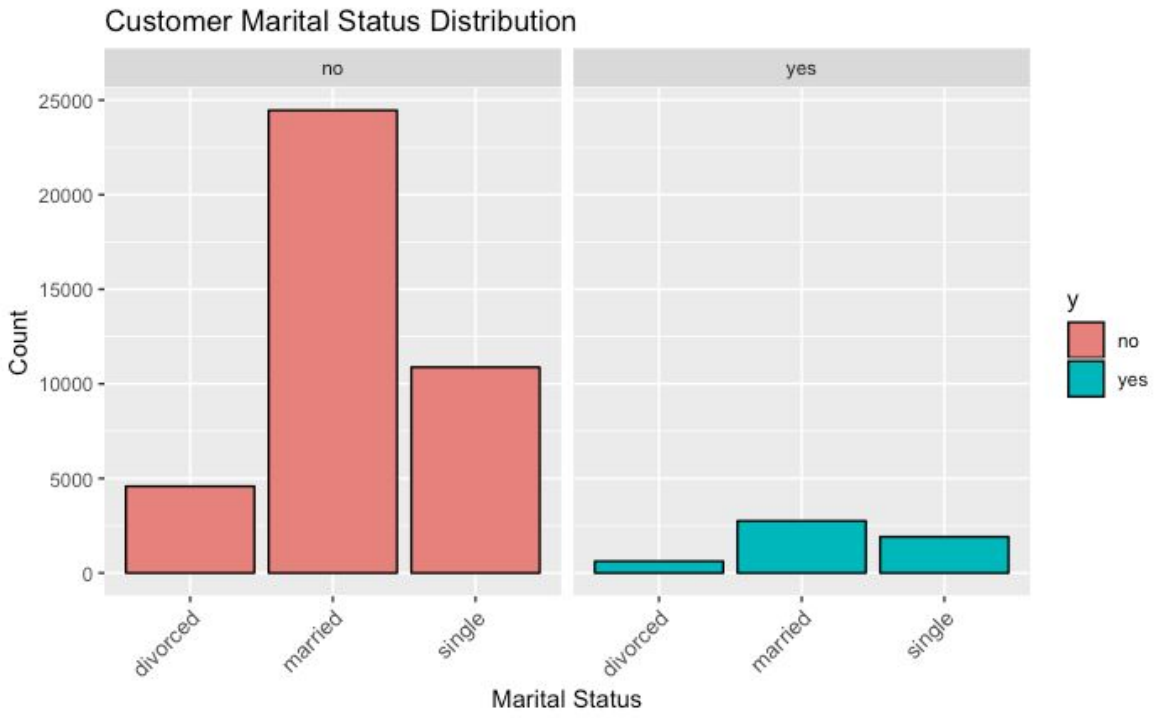


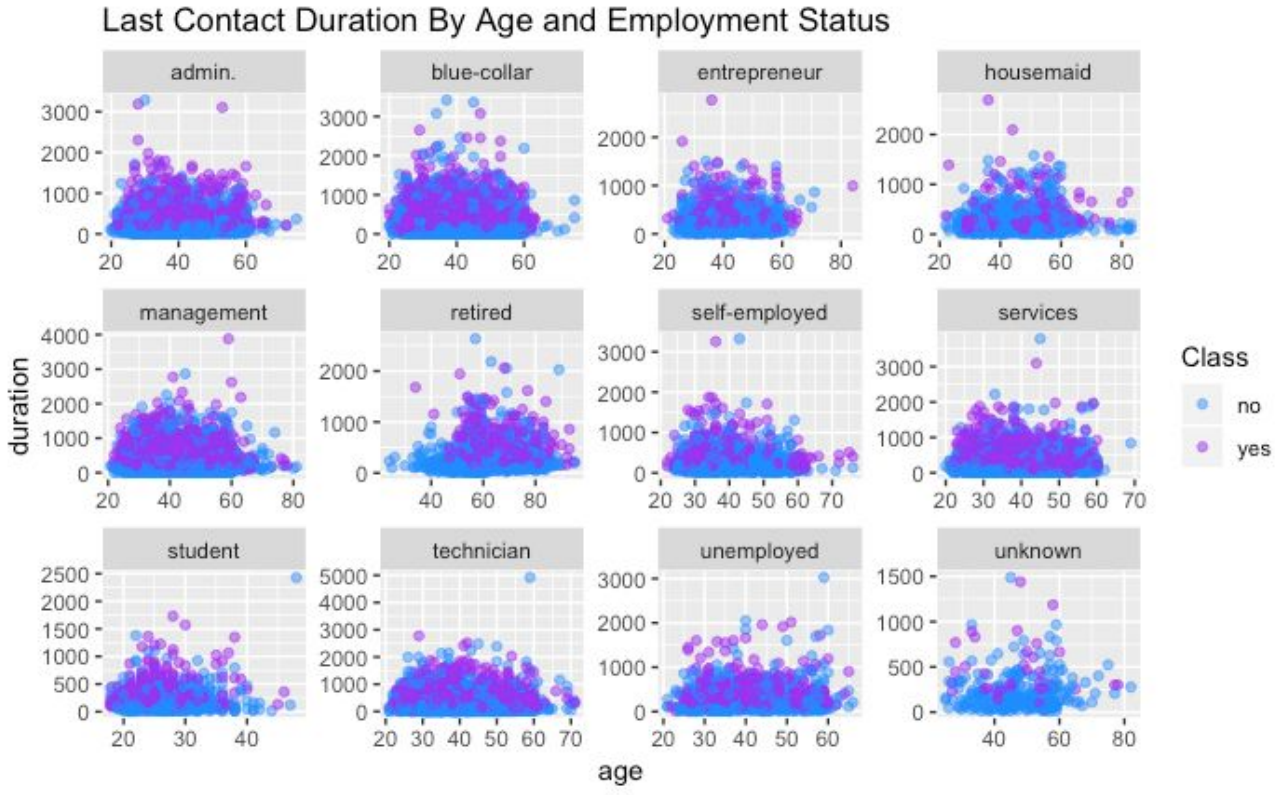


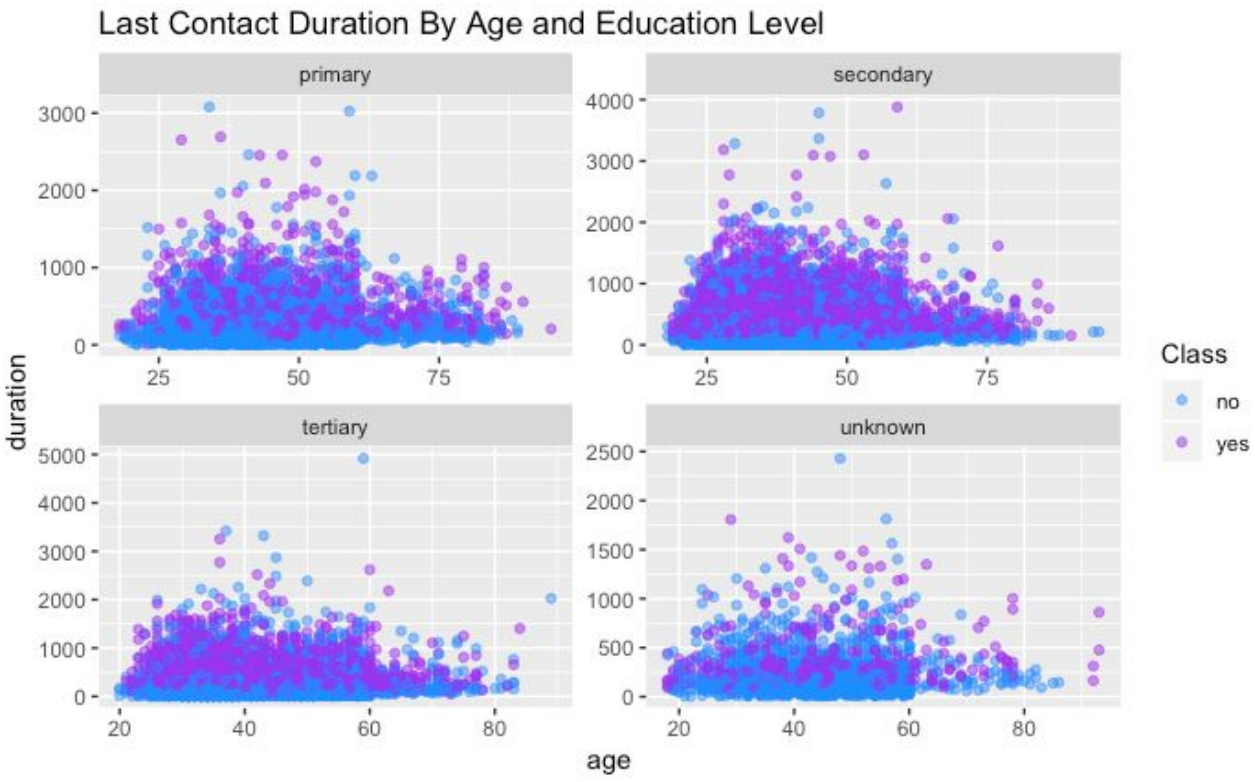



Contact Time By Customer Job Title
Duration (Seconds)
5000
4000
3000
2000
1000
0
admin.
blue-collar
entrepreneur
housemaid
management
retired
self-employed
services
student
technician
unemployed
unknown
Job Title
y
no
yes


Average Yearly Balance By Age
balance
100000
75000
50000
25000
0
25
50
75
age
Class
no
yes

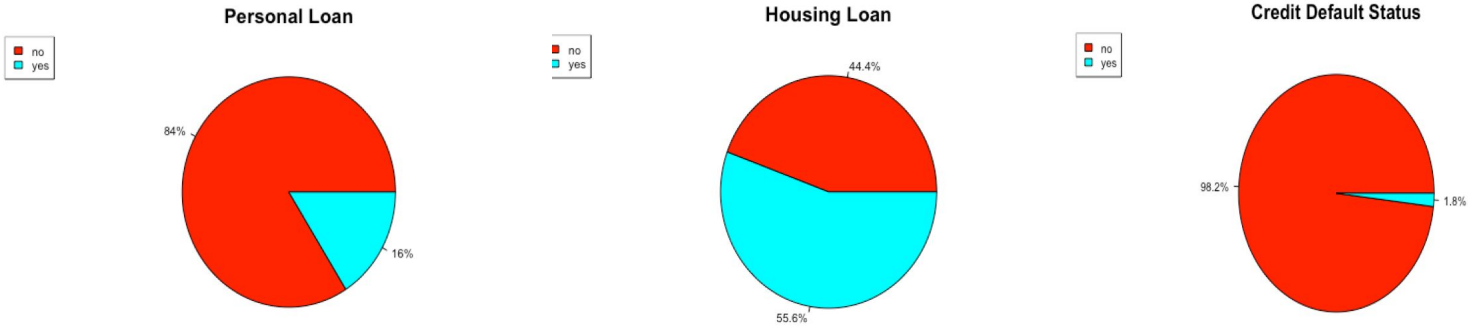

Customer Marital Status Distribution
no
yes
25000
20000
15000
10000
5000
0
Count
divorced
married
single
Marital Status
y
no
yes

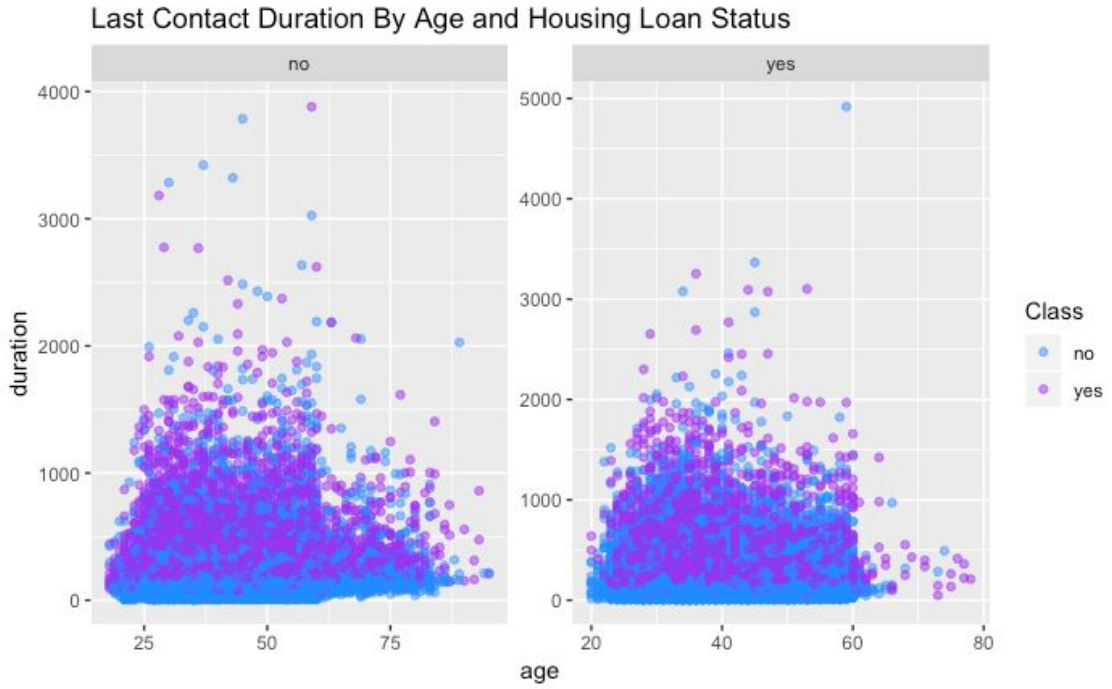

Last Contact Duration By Age and Employment Status
admin.
blue-collar
entrepreneur
housemaid
management
retired
self-employed
services
student
technician
unemployed
unknown
duration
age
Class
no
yes

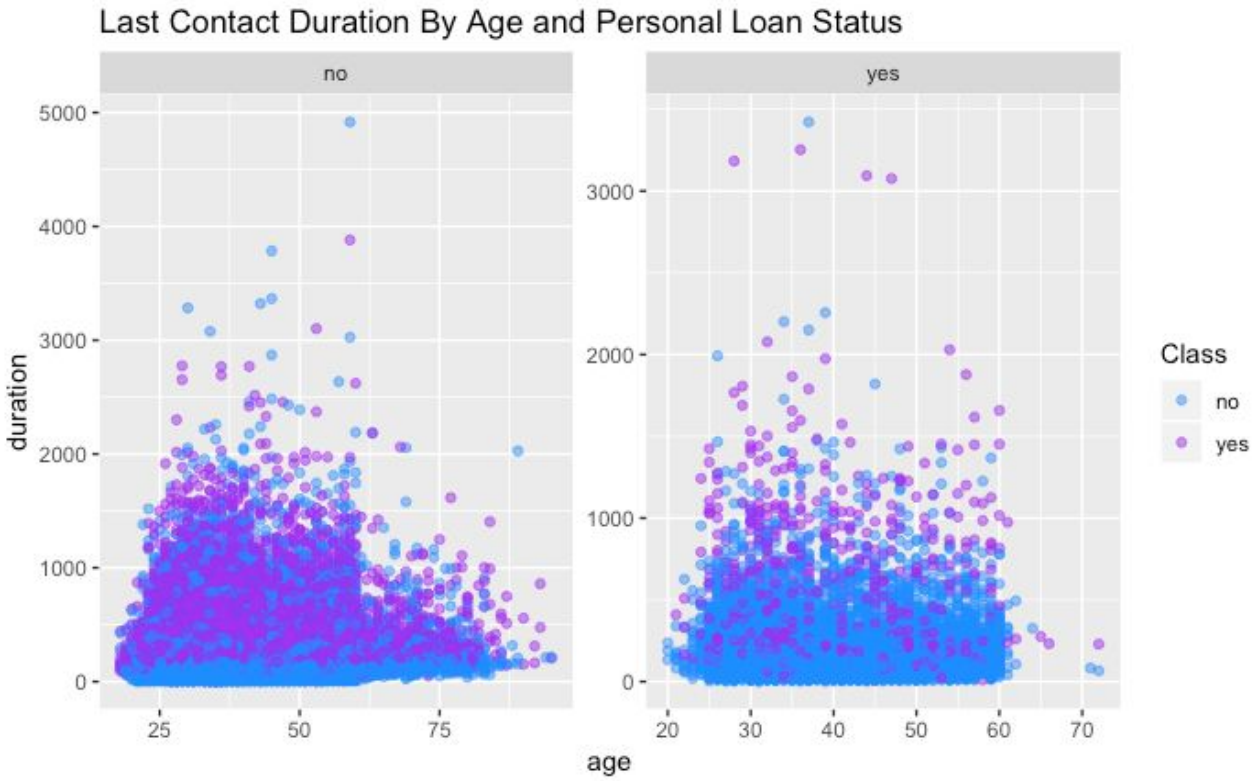


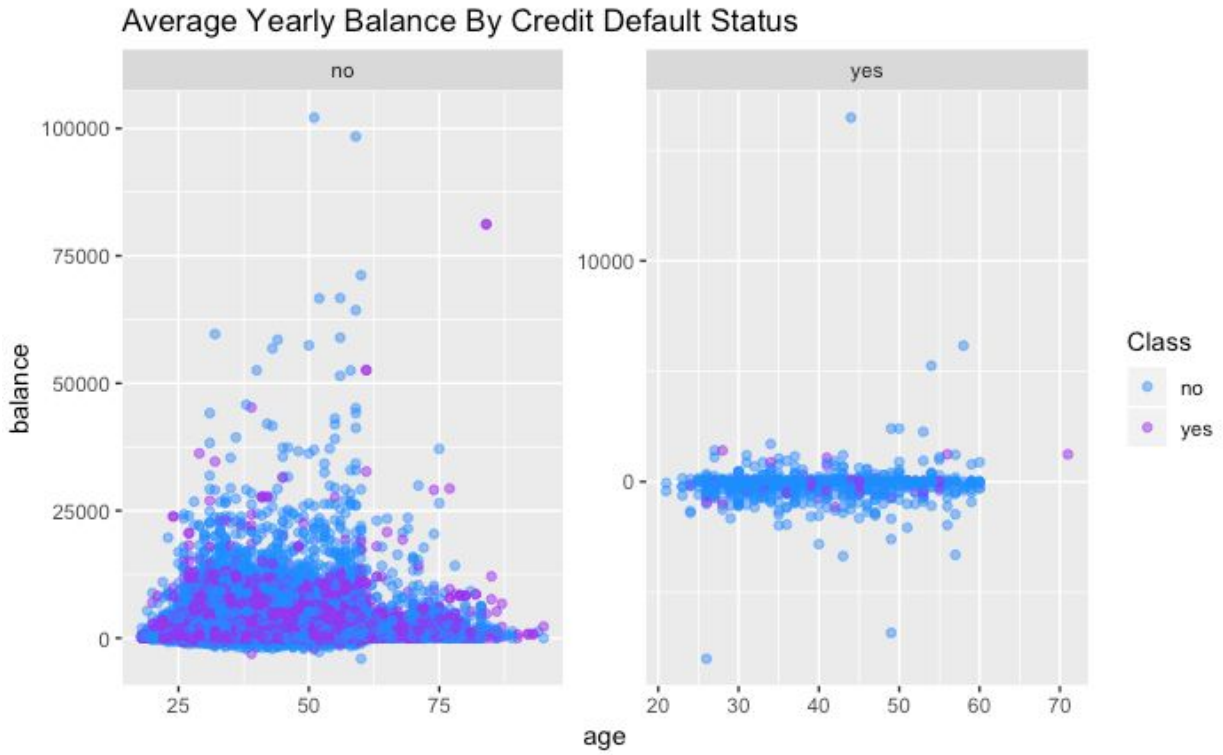
When interpreting the results from exploratory data analysis on customer demographics (age, job, education, and marital status, you will notice that the average age for a customer is 40-years-old. The youngest customer in the dataset is 18-years-old and the oldest is 95-years-old. Both data points can be referred to as outliers because both are inconsistent with the rest of the dataset. Furthermore, 50% of the customers are 39-years-old or younger, and 95% of the dataset consists of customers 59-years-old or younger are administrative workers and 22.5% are referred to as blue-collar workers. Out of 1,336 customers between 18 and 25-years-old, only 320 decided to subscribe to a term deposit at the bank. Only 2.1% of customers in the dataset are students; 60.5% of customers are married, and 28.1% are single. In the two scatter plots above, you will notice that most customers rarely spent more than 33 minutes (2000 seconds) on the phone with the bank regardless of education or employment level. You will also notice that the older the customer is, the less time they spent on the phone with the bank representative. There were a few outliers that did spend about 83 minutes on the phone but decided not to subscribe to a term deposit.

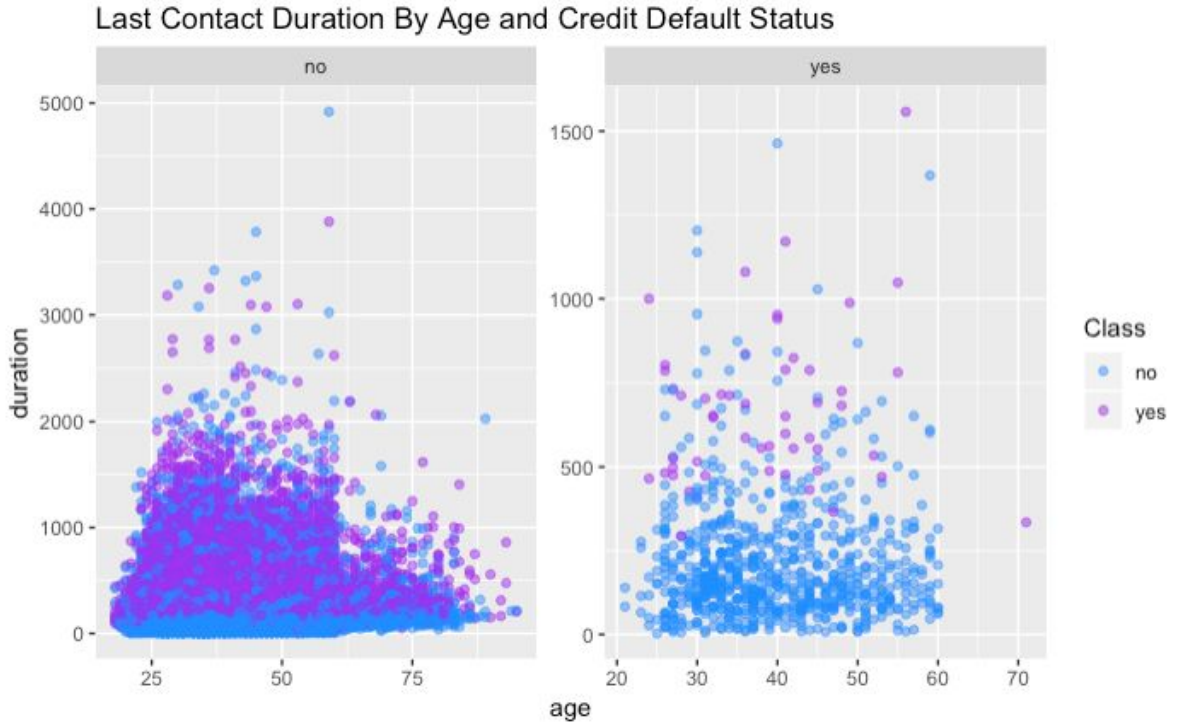
4.2 Customer Personal Banking




Last Contact Duration By Age and Housing Loan Status
no
yes
duration
age
Class
no
yes

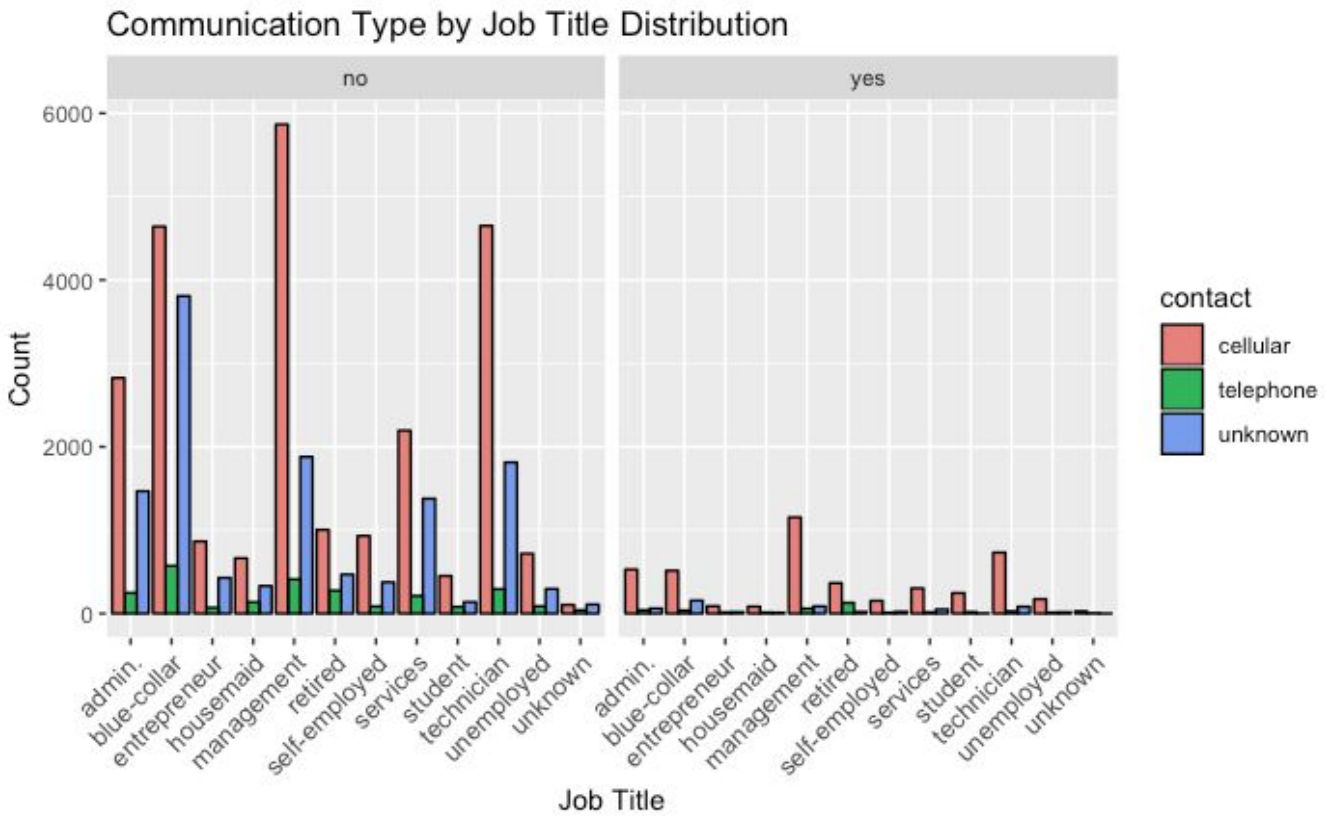

Last Contact Duration By Age and Personal Loan Status
no
yes
duration
age
Class
no
yes

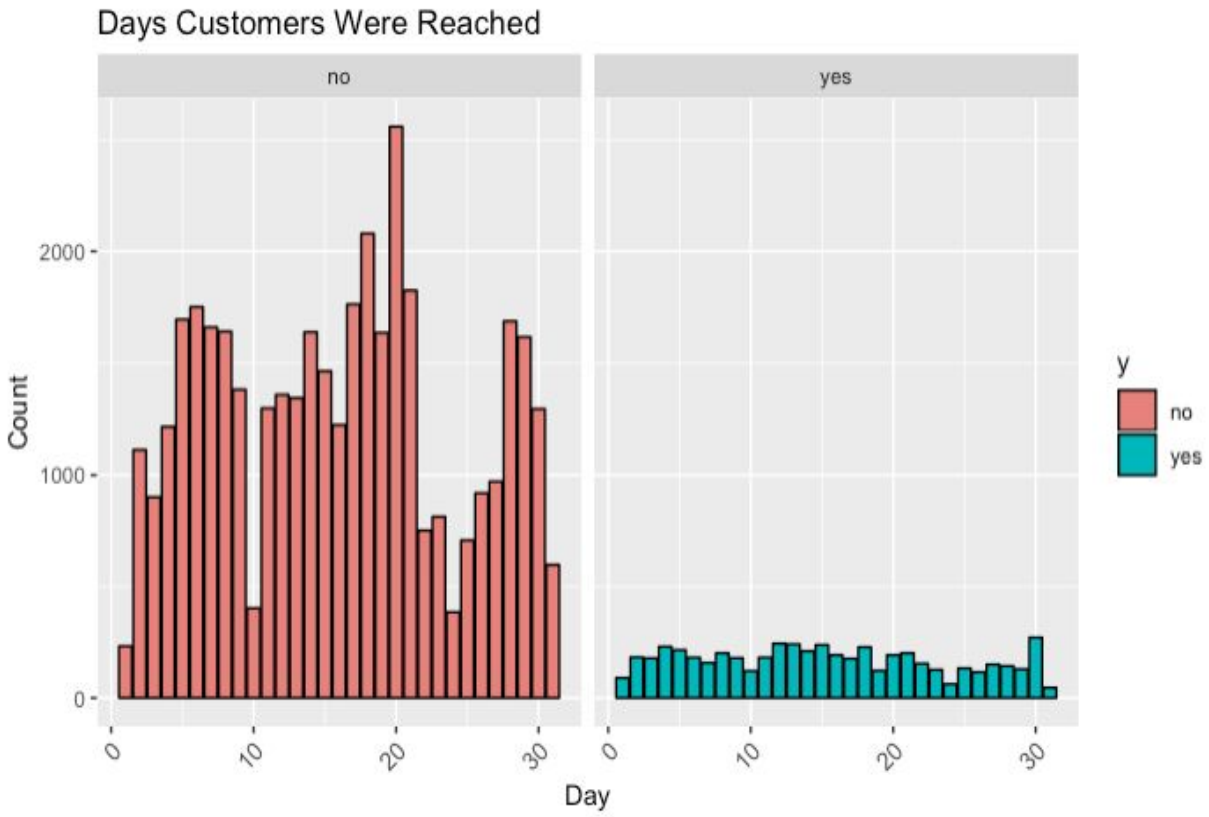

Average Yearly Balance By Credit Default Status
no
yes
balance
100000
75000
50000
25000
0
10000
0
25
50
75
20
30
40
50
60
70
age
Class
no
yes

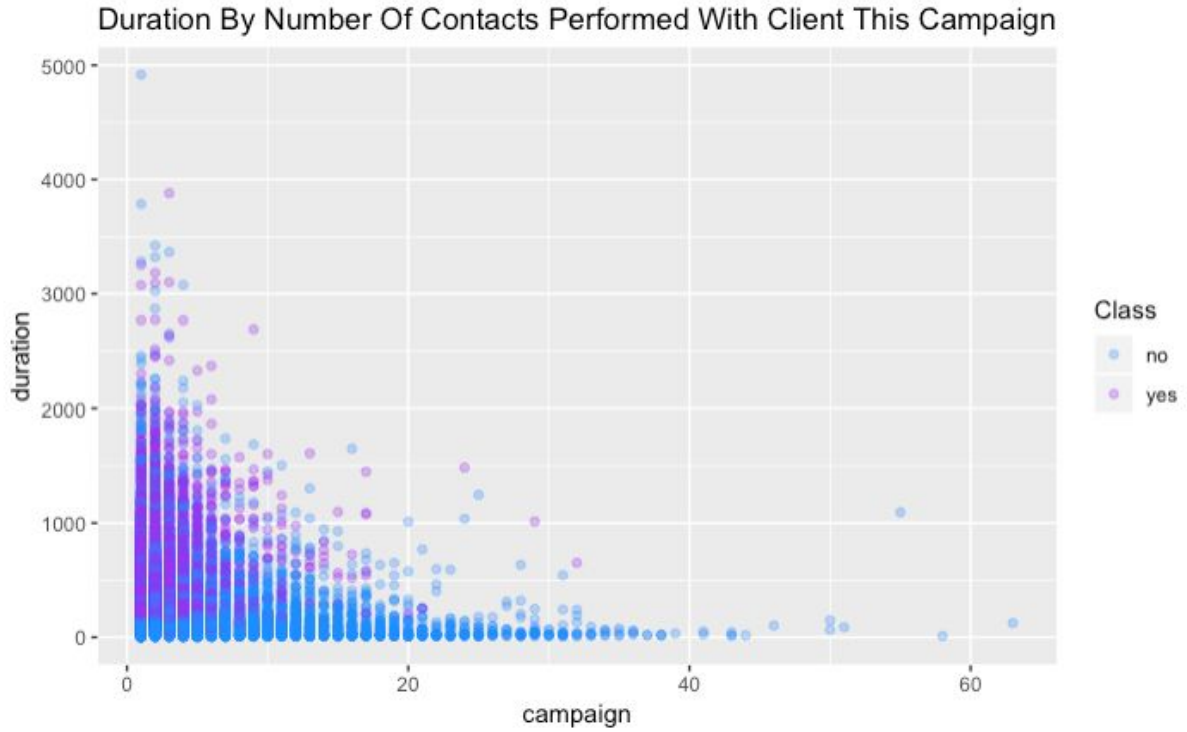

Last Contact Duration By Age and Credit Default Status
no
yes
duration
5000
4000
3000
2000
1000
0
1500
1000
500
0
25
50
75
20
30
40
50
60
70
age
Class
no
yes

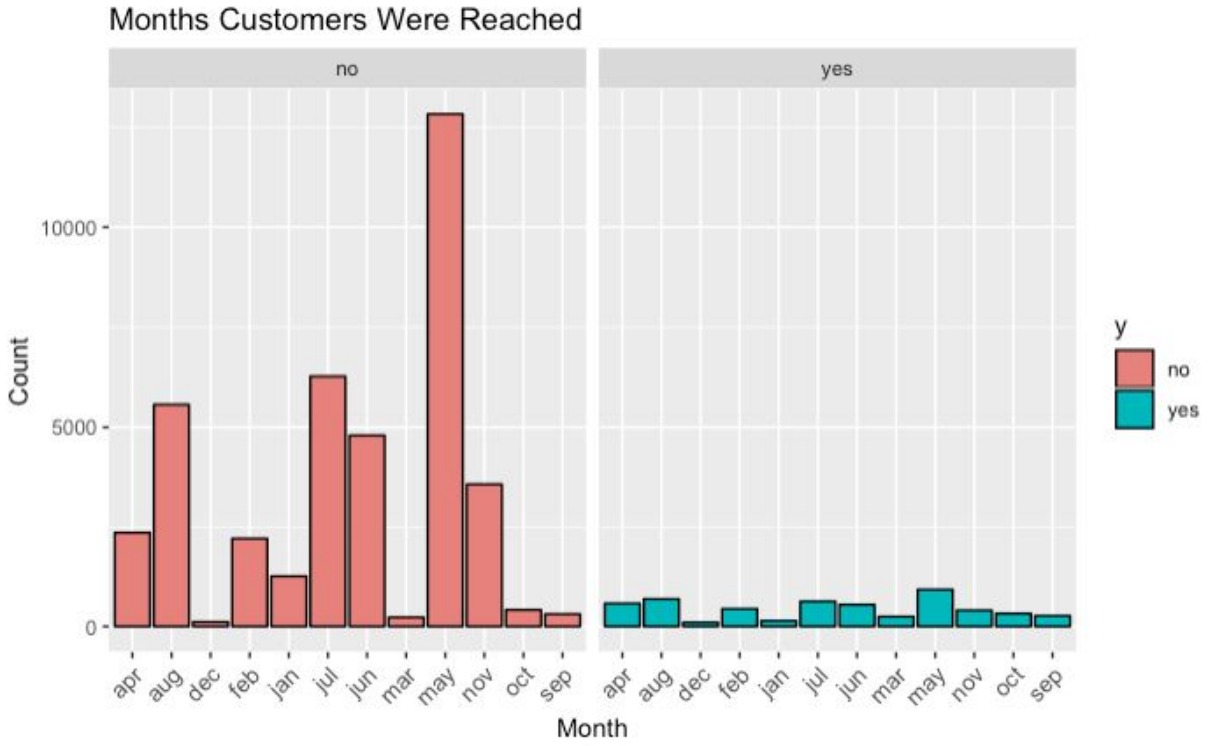
When interpreting the results from exploratory data analysis on customer personal banking, 92% of customers had no credit in default; 55.6 of customers had a housing loan, and 84% of customers did not have a housing loan. The lowest average yearly balance a customer had was -8019 and the highest was 102127. The average yearly balance a customer had was 1362, with 95% of customers having less than 5768 in their account. Only 5,237 of 44,396 customers that did not have credit in default subscribed to a term deposit. Furthermore, 3,354 of 20,081 customers that did not have a housing loan subscribed to a term deposit. Likewise, 4,805 of 37,967 customers that did not have a personal loan subscribed to a term deposit. If a customer had a personal or housing loan, they spent less time on the phone with a bank representative, and if they did have a lengthy conversation, they ended up subscribing to a term deposit. Customers that had credit in default also had a negative average yearly balance and most of them were not willing to subscribe to a term deposit.

4.3 Current Marketing Campaign



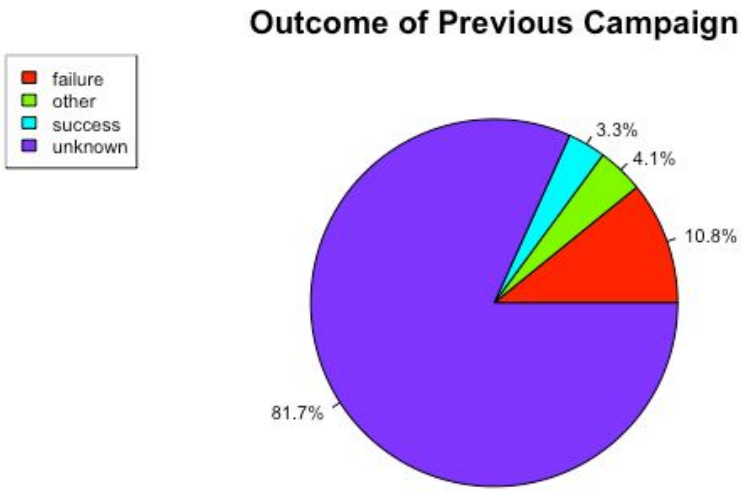

Days Customers Were Reached
no
yes
Count
2000
1000
0
0
10
20
30
Day
y
no
yes

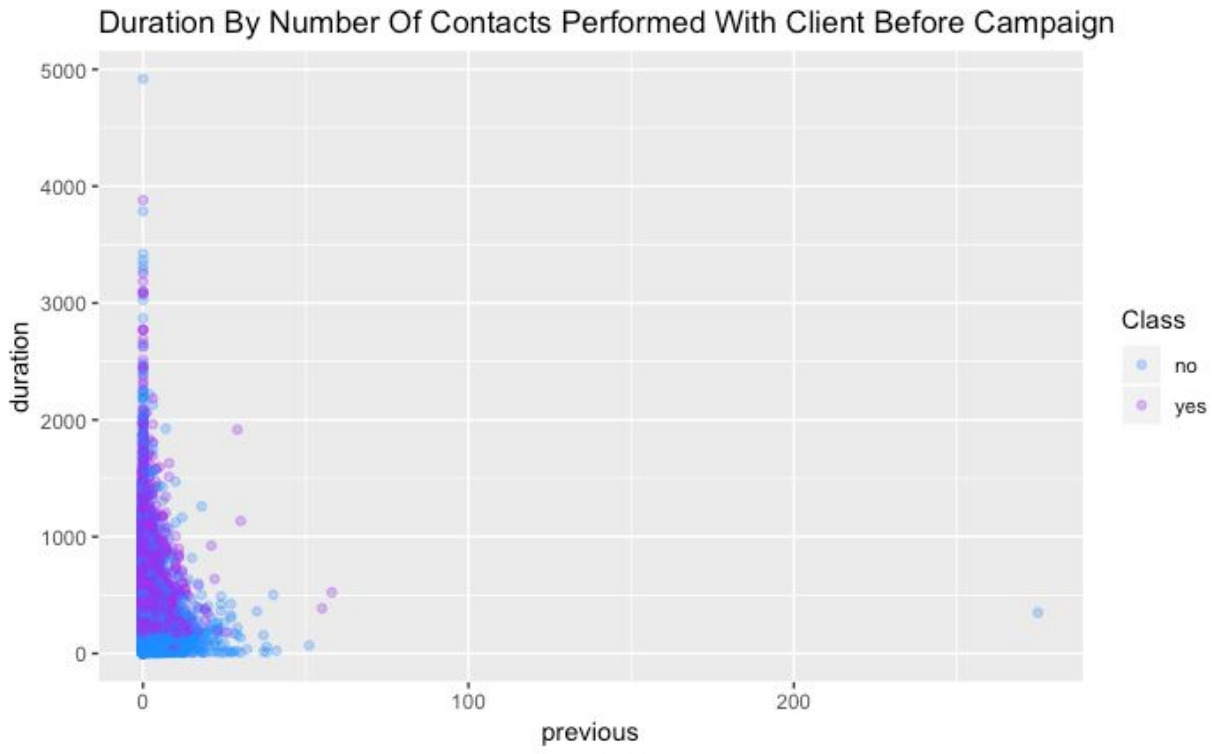

Duration By Number Of Contacts Performed With Client This Campaign
duration
5000
4000
3000
2000
1000
0
0
20
40
60
campaign
Class
no
yes

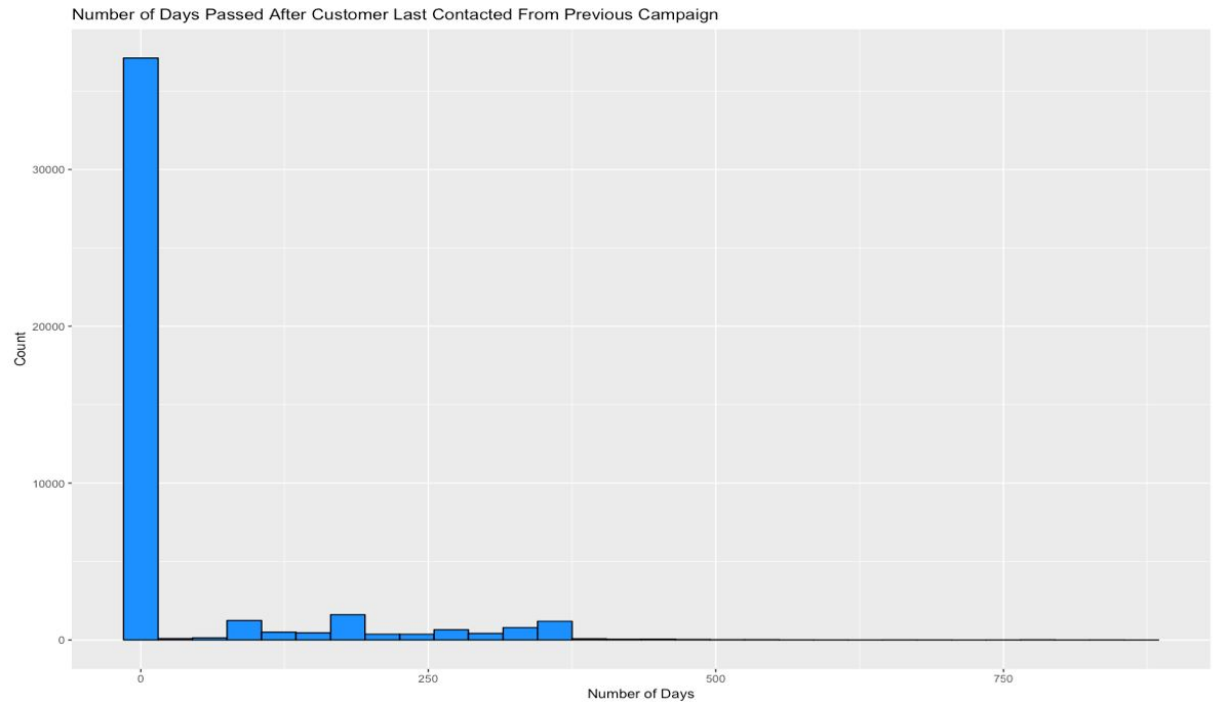


When interpreting the results from exploratory data analysis from the current marketing campaign, 64% of customers were contacted by a bank representative on their cellular device and 6% were reached by telephone. Most customers were reachable in the middle of the month or towards the end of the month. 30% of customers were last contacted in May. Moreover, 214 customers were last contacted by the bank in December and 100 of the customers subscribed to a term deposit. The average duration of communication was about 4 minutes (258.16 seconds). The highest was 81.96 minutes (4918 seconds). Customers were contacted twice on average during the current marketing campaign. Most of the customers were contacted less than 10 times during this campaign and most did not purchase a term deposit account.

4.4 Previous Marketing Campaign



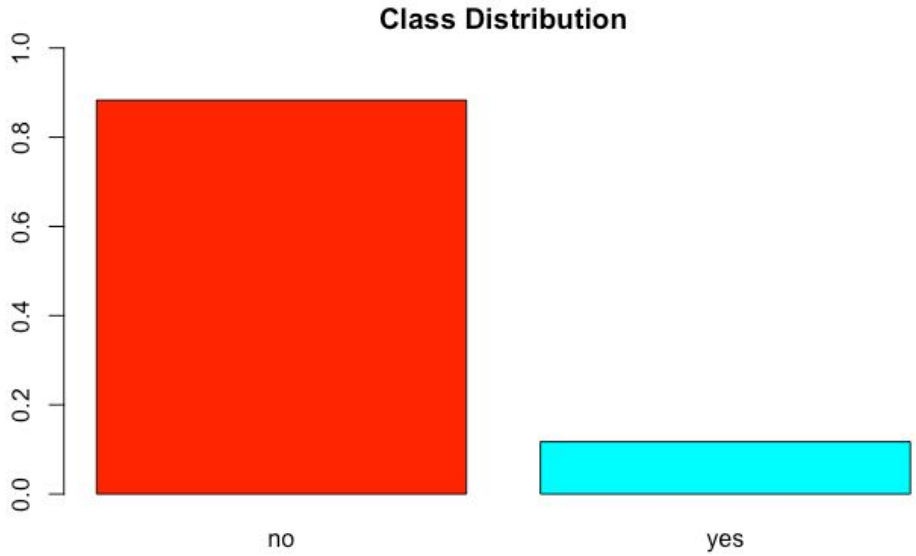




When interpreting the results from exploratory data analysis from a previous marketing campaign by the bank, you will notice that 75% of customers were not contacted by the bank since the previous campaign and 95% of customers were only contacted twice prior to the current campaign. The average number of days after the client

was last contacted from the previous campaign was about 40 days, and the client was rarely contacted by the bank before the current marketing campaign.

4.5 Class Distribution



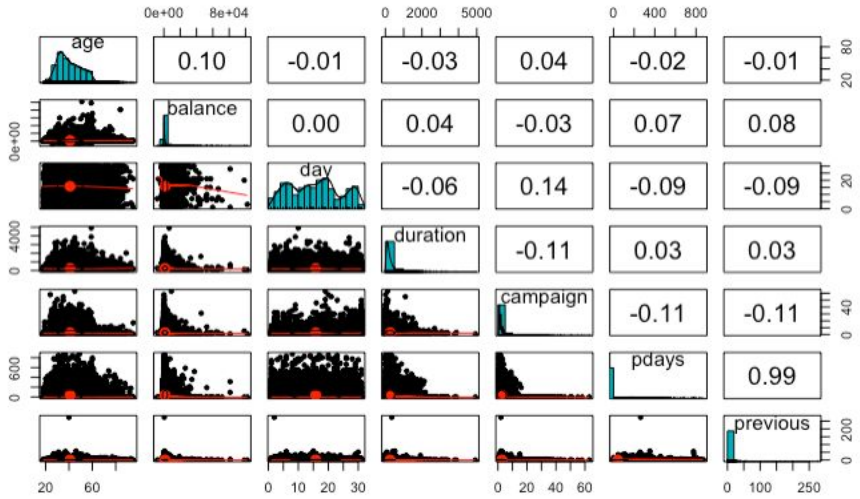
The dataset is highly imbalanced, which is a common problem in classification problems. There are multiple ways in order to handle the disproportionate ratio of observations in a dataset, including oversampling and under-sampling. Oversampling is when the minority class is duplicated until the class balance is achieved. Under-sampling removes the random instances of the majority class. Although the results may be strong, the model suffers from a significant loss of data points and ends up not generalizing well enough. For this dataset, the Synthetic Minority Over-Sampling Technique (SMOTE) package in R was used to balance the skewed class distribution of positives and negatives. If imbalanced data is trained on the classifier, the results may be biased in favour of the majority class simply because there weren't enough data points to learn about the population of the minority class. You could end up with test samples belonging to the minority being misclassified as the majority class. According to the report "SMOTE for Learning from Imbalanced Data: progress and Challenges" from the *Journal of Artificial Intelligence*, "instead of applying a simple replication of the minority class instances, the key idea of SMOTE is to introduce synthetic example," (Fernandez, Garcıa, Herrera, & Chawla, 2018, p. 866). The report goes on to explain how the new synthetic data points are created: "Based on a distance metric, several nearest neighbours of the same class (points xi1 to xi4) are chosen from the training set. Finally, a randomized interpolation is carried out in order to obtain new instances r1 to r4," (Fernandez, Garcıa, Herrera, & Chawla, 2018, p. 866). While it is a powerful balancing technique, there are some disadvantages to applying the SMOTE algorithm, including model overfitting due to duplicate data points.

5. LITERATURE REVIEW

Prior to selecting what type of statistical test to use for a specific dataset, we need to know how large the dataset is; if there are any linear dependencies between the features, and if the variables normally distributed.

5.1 Spearman Rank Correlation

Spearman's correlation coefficient is a nonparametric test that quantifies if there is a monotonic relationship between two variables that are either continuous or ordinal using rank values. Spearman's correlation coefficient does assumes that the variables are not normally distributed and do not have a linear relationship. According to the Handbook of Biological Statistics by John H. McDonald, "if you have a non-monotonic relationship (as X gets larger, Y gets larger and then gets smaller, or Y gets smaller and then gets larger, or something more complicated), you shouldn't use Spearman rank correlation."



5.2 Wilcoxon's Rank Sum Test

Wilcoxon rank-sum test (also known as the Mann-Whitney U Test) is a nonparametric test that assumes that the two samples come from continuous distributions equal medians and represent random variables x and y. The null hypothesis is that there is no shift in the shape of distribution. With the p-values of 7 continuous variables is $< 0.05 = a$, we reject the null hypothesis and conclude that there is enough evidence to suggest a statistically significant difference between x and y.

5.3 Chi-Square Test

The Chi-Square test of independence was applied on 9 categorical variables in the dataset. The Chi-Square test is used to measure if there is a relationship between two categorical variables. The null hypothesis is that there is no relationship between the two variables and that x and y variables are independent. With the p-values of 7 categorical variables is $< 0.05 = a$, we reject the null hypothesis and can conclude there enough evidence to suggest that the variables are dependent.

Attribute	P-Value	Statistical Test	Result
Age	< 2.2e-16	Wilcoxon Rank-Sum Test	Significant
Balance	< 2.2e-16	Wilcoxon Rank-Sum Test	Significant
Day	< 2.2e-16	Wilcoxon Rank-Sum Test	Significant
Duration	< 2.2e-16	Wilcoxon Rank-Sum Test	Significant
Campaign	< 2.2e-16	Wilcoxon Rank-Sum Test	Significant
Pdays	< 2.2e-16	Wilcoxon Rank-Sum Test	Significant
Previous	< 2.2e-16	Wilcoxon Rank-Sum Test	Significant
Job	< 2.2e-16	Chi-Square Test	Significant
Marital	< 2.2e-16	Chi-Square Test	Significant
Education	< 2.2e-16	Chi-Square Test	Significant
Default	2.45E-06	Chi-Square Test	Significant
Housing	< 2.2e-16	Chi-Square Test	Significant
Loan	< 2.2e-16	Chi-Square Test	Significant
Contact	0.0321	Chi-Square Test	Significant
Month	< 2.2e-16	Chi-Square Test	Significant
Poutcome	< 2.2e-16	Chi-Square Test	Significant

Moreover, the ‘findCorrelation’ function was used in R to reduce pairwise correlations based on absolute values with a cutoff of 0.75. Oftentimes, variables can be highly correlated due to missing values. Likewise, variables with a small or near-zero variance were analyzed using the nearZeroVar() function. Predictors with zero- or near zero-variance could result in a biased or incorrect model. Measures such as variance are important because it provides some insight on how well a model could perform on a dataset that it is not trained on such as a test dataset. For example, the variance that is high could lead to overfitting and incorrect predictions. There are a few ways in which high variance can be identified during the model building phase, including a training error that is the low and high test set error. However, if the variance is zero- or near zero, then that means that a few observations are different from the variable’s constant value and suggests that the feature has less predictive power. In the table below, you will notice that ‘default’ and ‘pdays’ have a near-zero variance. However, these variables will not be removed.

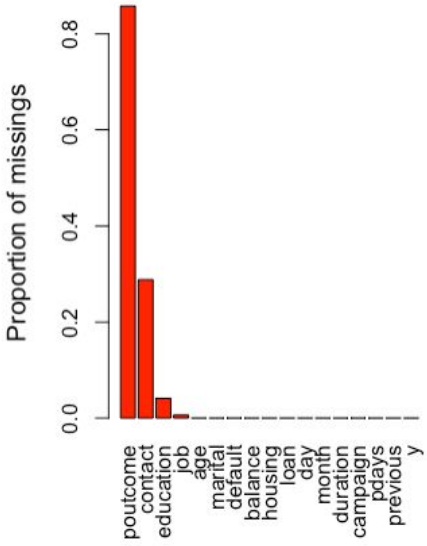
Attribute	freqRatio	percentUnique	zeroVar	nzv
Age	1.044589	0.170312535	FALSE	FALSE
Job	1.028346	0.024330362	FALSE	FALSE
Marital	2.127756	0.006635553	FALSE	FALSE
Education	1.764616	0.006635553	FALSE	FALSE
Default	54.47362	0.004423702	FALSE	TRUE
Balance	18.020513	15.85454867	FALSE	FALSE
Housing	1.251432	0.004423702	FALSE	FALSE
Loan	5.241165	0.004423702	FALSE	FALSE
Contact	9.588056	0.004423702	FALSE	FALSE
Day	1.192374	0.068567384	FALSE	FALSE
Month	1.996519	0.026542213	FALSE	FALSE
Duration	1.021739	3.479241777	FALSE	FALSE
Campaign	1.402959	0.106168853	FALSE	FALSE
Pdays	221.281437	1.236424764	FALSE	TRUE
Previous	13.331169	0.090685895	FALSE	FALSE
Poutcome	1.824099	0.004423702	FALSE	FALSE
Term Deposit	7.548119	0.004423702	FALSE	FALSE

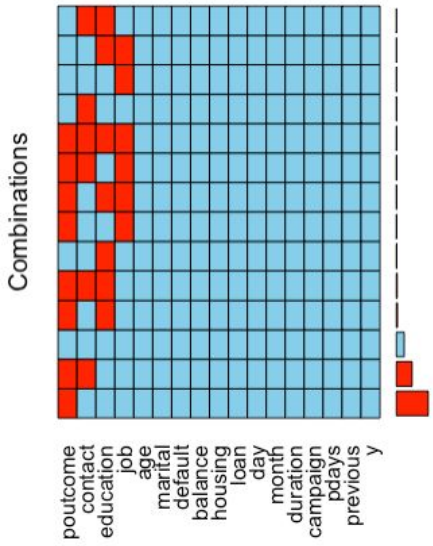
6. DATA PREPROCESSING

6.1 Data Cleaning

'Unknown' and "other" values were replaced with 'NA' to be recognized as missing values. Removing or imputing missing values can speed up the training time for each model and could increase the overall accuracy. However, the size of the dataset and the number of missing values should always be considered prior to decide to remove the select rows with missing values or replacing them with the mean of the variable of interest. If missing values are not treated, it could result in a biased model and incorrect classification. According to the book, *Flexible Imputation of Missing Data by Stef van Burren*, there are three types of missing values, including MCAR, MAR, and NMAR. Burren states: "if the probability of being missing is the same for all cases, then the data are said to be missing completely at random (MCAR)," (Buuren, 2018). Meaning, that there is no correlation to the causation of missing values in the data set. However, Burren goes onto state: "if the probability of being missing is the same only within groups defined by the observed data, then the data are missing at random (MAR)," (Buuren, 2018). Meaning. that there is a relationship between the missing values and observed data. But there are cases when both MCAR and MAR don't have enough conceptual evidence, which is when Missing Not At Random (MNAR) is considered. Burren states: "MNAR means that the probability of being missing varies for reasons that are unknown to us," (Buuren, 2018). Meaning, that the missing value may be intentionally unknown due to the participant. Among the variables in the dataset, 'Poutcome' had the highest amount of missing values with 38,799. Other variables with missing values were

'job,' 'education,' 'contact'. For this dataset, there were a significant amount of missing values that made the list-wise deletion and pairwise deletion, not the best strategy to treat missing values. Instead, the Multivariate Imputation by Chained Equations (MICE) package developed by Stef van Burren was used in R to impute missing values in this dataset which has multivariate data. According to the book, *Applied Missing Data Analysis with SPSS and (R)Studio* written by Martijn W. Heymans and Iris Eekhout, "a chain of regression equations is used to obtain imputations, which means that variables with missing data are imputed one by one. The regression models use information from all other variables in the model, i.e. conditional imputation models. In order to add sampling variability to the imputations, the residual error is added to create the imputed values," (Heymans & Eekhout, 2019). There were also zero duplicates in the dataset.





Variable	Missing Values
Job	288
Education	1857
Contact	13020
Poutcome	38799

6.2 Data Transformation

During the data transformation stage, the categorical values for the variable 'job' were bucketed into 'unemployed' and 'employed' to better identify records in a shared group to enhance algorithmic efficiency. The data set was then preprocessed using the 'preProcess' function using the caret package in R with method = "center" which subtracts the mean of the predictor's data from the predictor values, and method = "scale" which then divides by the standard deviation. Once the data processing was completed, the caret package has a function called "dummyVars," which transforms categorical data to dummy variables or a binary predictor (0, 1) using full-rank for modelling purposes. The full-rank parameterization generates n-1 columns to avoid collinearity.

7. MODEL DEVELOPMENT

7.1 Training-Test Split

The 'createDataPartition' function in the Caret package is used to divide the data set into balanced proportions. For example, a dataset is conventionally split into training, validation, and test set. However, for this dataset, the train-validation splitting was not used. Instead, Cross-validation (CV) was used through the caret package. According to the report, *Introduction to Data Science by Ron Sarafian*, "the data used for training (all except the test set) is randomly partitioned into k subsamples. Of the k subsamples, a single subsample is retained as the validation set, and the remaining k - 1 subsample are used for training. The process is then repeated k time with each of the k subsamples used exactly once as the validation set. All k results are then aggregated," (Sarafian, 2019).

Furthermore, 70% of the dataset was used for training and to build the model. Meanwhile, 30% was used for testing purposes and to evaluate the performance metrics such as precision and recall for each predictive model. The model's parameters were optimized with the best ROC using cross-validation. The main idea is to never use the test set to evaluate a model's performance using the 'actual' data. For each retraining, a distinct set of hyperparameters were evaluated, but still using the training set. Each retraining, we're also checking how strong the model generalizes by checking the performance of cross-validation. If the performance of the test set is compatible or does better than the results of cross-validation, then that is the final model.

8. CROSS-VALIDATION EVALUATION & RESULTS

For this dataset, accuracy is not the correct performance metric to use because there are more negatives than positives. Therefore predicting negatives would not be a correct labelling. Instead, the ROC (Receiver Operating Characteristics) curve will be used to evaluate how well a model can distinguish between classes when predicting. The higher the AUC (Area Under The Curve) or closer to 1, the stronger the model is at predicting the binary classes correctly.

Furthermore, precision, recall, and F1 Score were also used to evaluate classification algorithms. Through a confusion matrix, recall indicates the percentage of correct positive classifications (true positives) from respondents that are actually positive. In other words, how many customers were correctly classified as term deposit subscribers by the model.

Meanwhile, precision indicates the percentage of correct positive classifications (true positives) from respondents that were predicted as positive. In other words, how many customers were correctly classified as non-subscribers and subscribers, respectively.

Ensuring a high precision percentage means that the bank would save time and money spent on calling customers who are likely to subscribe to a term deposit. However, ensuring a high recall percentage would be mean that the bank wouldn't miss out on subscribers who would actually subscribe to a term deposit regardless of calling the occasional customer who chooses not to subscribe to a term deposit. We need to be able to answer what is the cost of incorrectly classifying a customer who is a non-subscriber as a subscriber (false positive or Type I Error cost = marketing) versus what is the cost of failing to identify a customer who is a subscriber as a non-subscriber (false negative = revenue). When it comes to bank marketing, false positives are inexpensive compared to false negatives or Type II, which could lead to sales opportunities.

CONFUSION MATRIX	**ACTUAL VALUES**	
PREDICTED VALUES	**YES**	**NO**
YES	TRUE POSITIVE	FALSE POSITIVE
NO	FALSE NEGATIVE	TRUE NEGATIVE

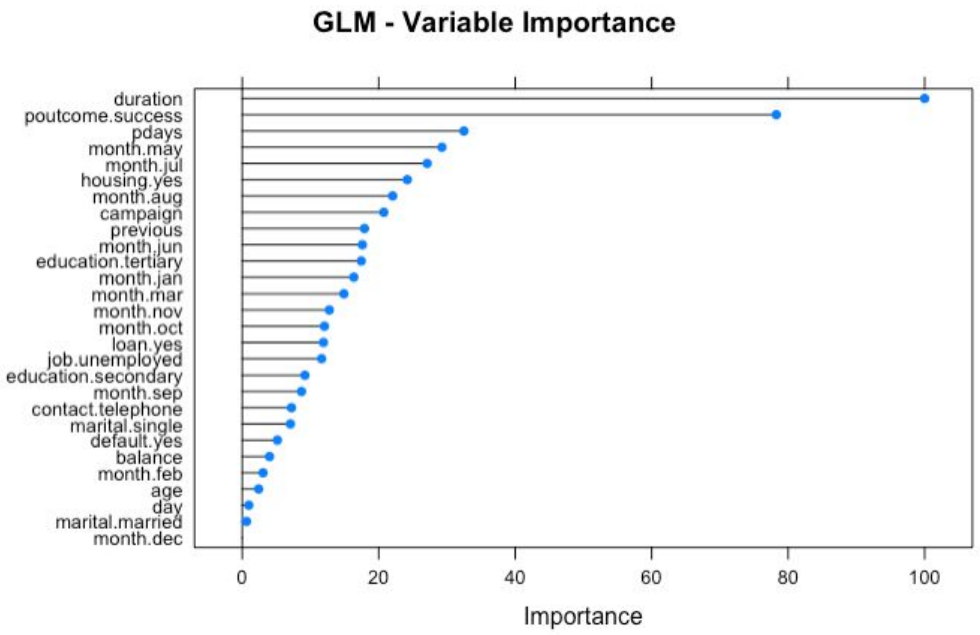
CONFUSION MATRIX METRICS
True Positive: Correctly predicting subscriber as a subscriber.
True Negative: Correctly predicting non-subscriber as a non-subscriber.
False Positive: Incorrectly predicting non-subscriber as a subscriber.
False Negative: Missing (Incorrectly predicting subscriber as a non-subscriber.

8.1 Logistic Regression

Logistic Regression is a generalized linear model used to predict the probability of binary classification of y, given x. Logistic Regression returns predicts $Y = 1$ when $p > 0.05$ and $Y = 0$ when $p < 0.5$. The probabilities are converted into binary values of 0 and 1 by the logistic function (sigmoid function). Logistic Regression does not require any parameter tuning and normalization of the data prior to being implemented. However, feature

engineering would be necessary for the model to perform well because the algorithm does not calibrate reliable predicted probabilities when it is fed a large number of predictors that do not have a high predictive power to the discrete outcome variable.

```
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 15235, 15236, 15236, 15234, 15235, 15234, ...
Resampling results:

  ROC        Sens       Spec
  0.9210069  0.8546408  0.849164
```



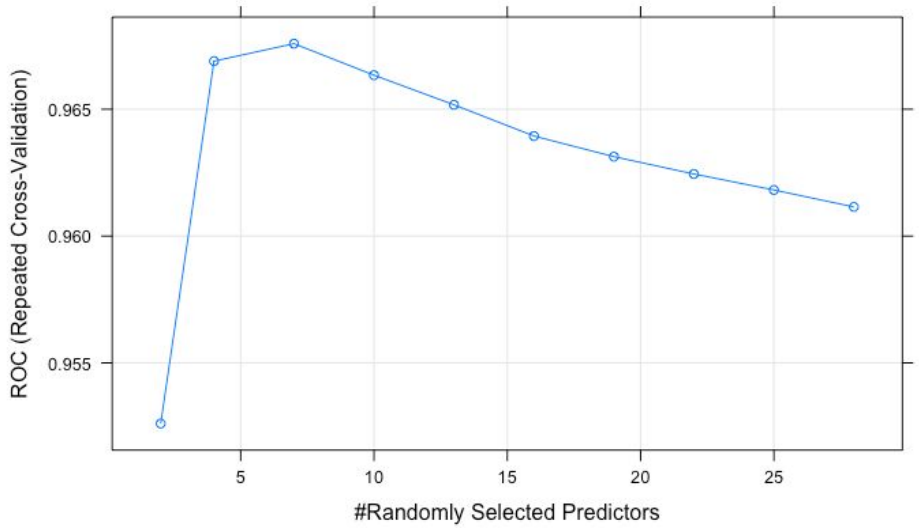

```
Cross-Validated (10 fold, repeated 3 times) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction  yes   no
       yes 42.7  7.5
       no   7.3 42.5

 Accuracy (average) : 0.8519
```

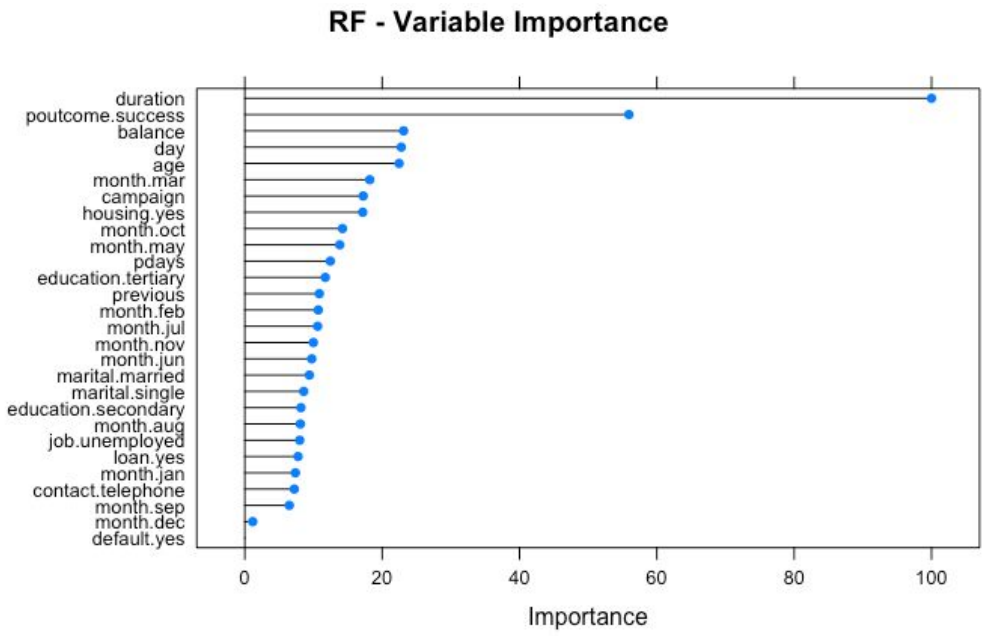
After applying 10-fold cross-validation for the logistic classifier, we can see that “duration” has the highest importance in the model.

8.2 Random Forest

Random Forest generates multiple decision trees, combines the results from each tree, and then splits nodes or subsets based on the input variables on a majority vote, and delivers a final output of class prediction. There are two parameters *mtry* and ntree that were tuned to enhance the model. *Mtry* is the number of variables random samples at each split and ntree is the number of trees to group. Random forest typically runs well and efficiently on large datasets and perform feature selection through the GINI Index and Mean Decrease Accuracy.

```
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 15235, 15236, 15236, 15234, 15235, 15234, ...
Resampling results across tuning parameters:

  mtry  ROC        Sens       Spec
   2    0.9526078  0.8863828  0.8856318
   4    0.9668973  0.9243084  0.8914618
   7    0.9675858  0.9251351  0.8955575
  10    0.9663433  0.9230482  0.8940213
  13    0.9651754  0.9216696  0.8942188
  16    0.9639437  0.9202912  0.8935104
  19    0.9631304  0.9195429  0.8935105
  22    0.9624471  0.9187160  0.8930769
  25    0.9618174  0.9182430  0.8922498
  28    0.9611491  0.9176136  0.8918562

ROC was used to select the optimal model using the largest value.
The final value used for the model was mtry = 7.
```





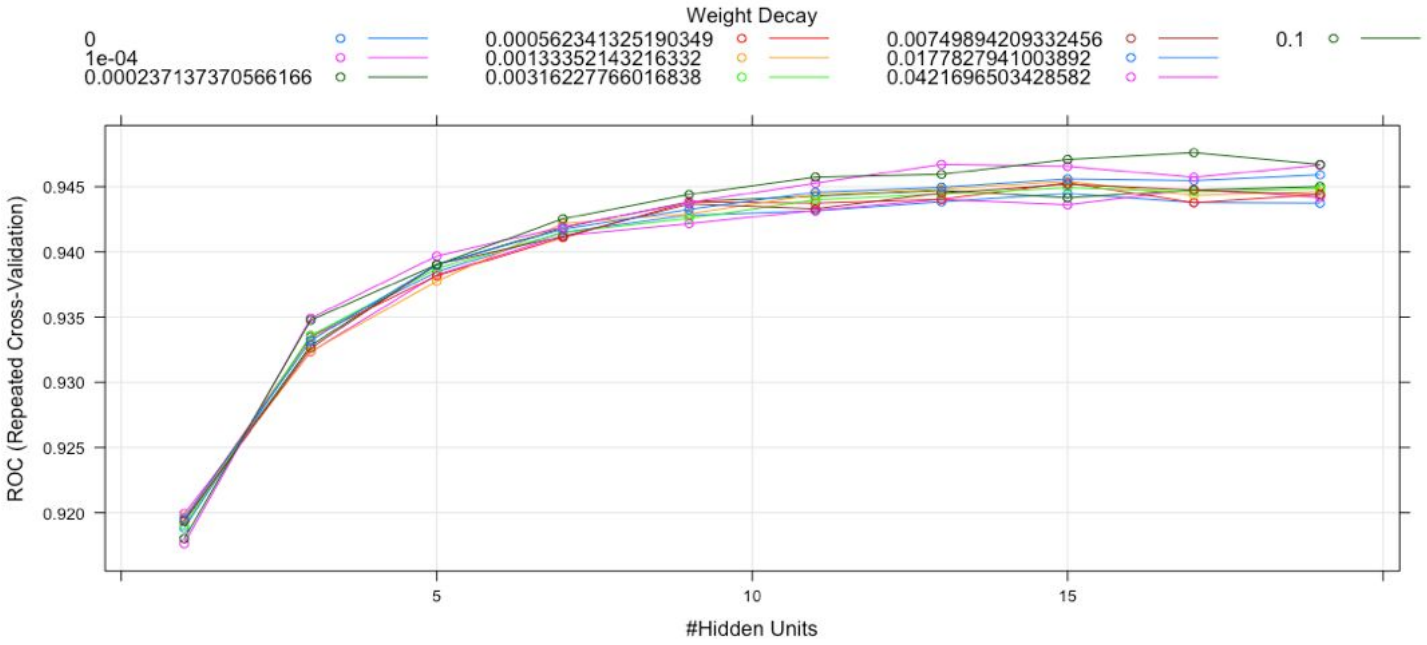

```
Cross-Validated (10 fold, repeated 3 times) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction  yes   no
       yes 46.3  5.2
       no   3.7 44.8

 Accuracy (average) : 0.9103
```

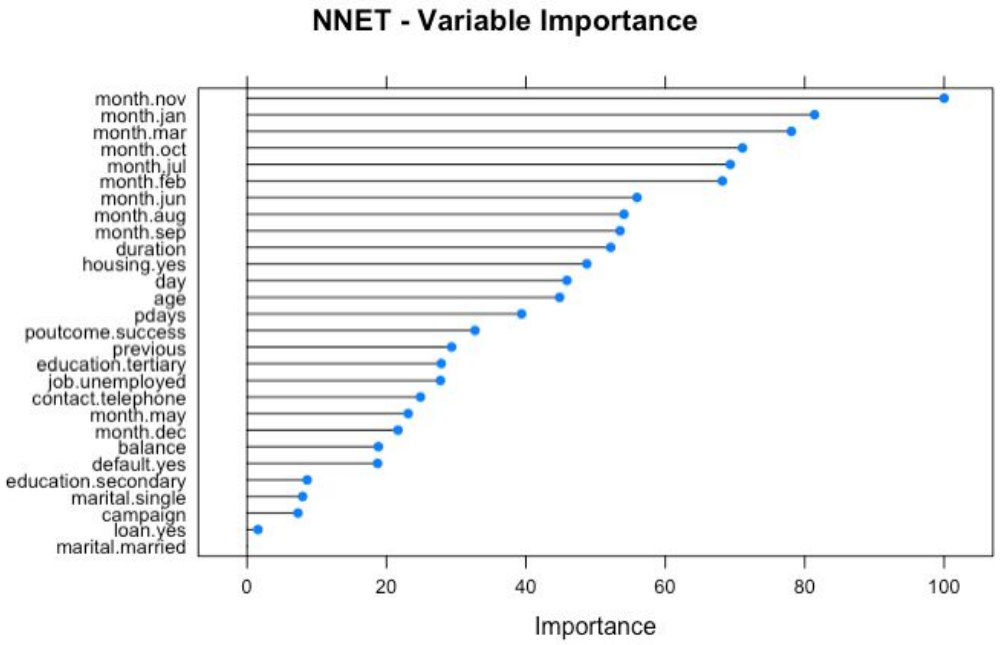
After applying 10-fold cross-validation for the random forest classifier, we can see that "duration" also has the highest importance in the model. In the grid of results are the average resamples estimates of performance where you will notice that when *mtry* at 7 was optimal.

8.3 Neural Networks

Neural network can be used for regression and classification problems with a dataset is numeric. The neural Network architecture consists of neurons and units, weights and parameters, and biases. According to Neural Network Toolbox by Howars Demuth and mark Beale, "Neural networks are composed of simple elements operating in parallel. These elements are inspired by biological nervous systems. As in nature, the network function is determined largely by the connections between elements," (Demuth, 2000). Furthermore, "batche training of a network proceeds by making weight and bias changes based on an entire sett (batch) of input vectors. Incremental training changes the weights and biases of a network as needs after the presentation of each individual input vector," (Demuth, 2000).

The algorithm performs well with nonlinear data and a large number of predictors and data points. However, Neural networks are “black boxes” and it is challenging to figure out how much the independent variable is impacting the dependent variable. It is also computationally heavy and training data can be a lengthy process, which could result in overfitting.





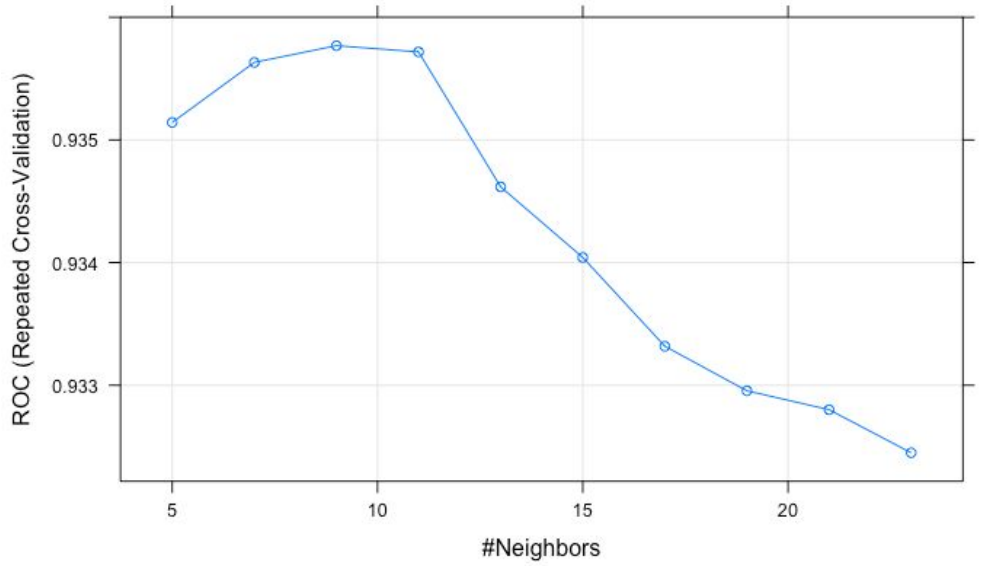
```
Cross-Validated (10 fold, repeated 3 times) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction  yes   no
       yes 44.8  6.8
       no   5.2 43.2

 Accuracy (average) : 0.8805
```

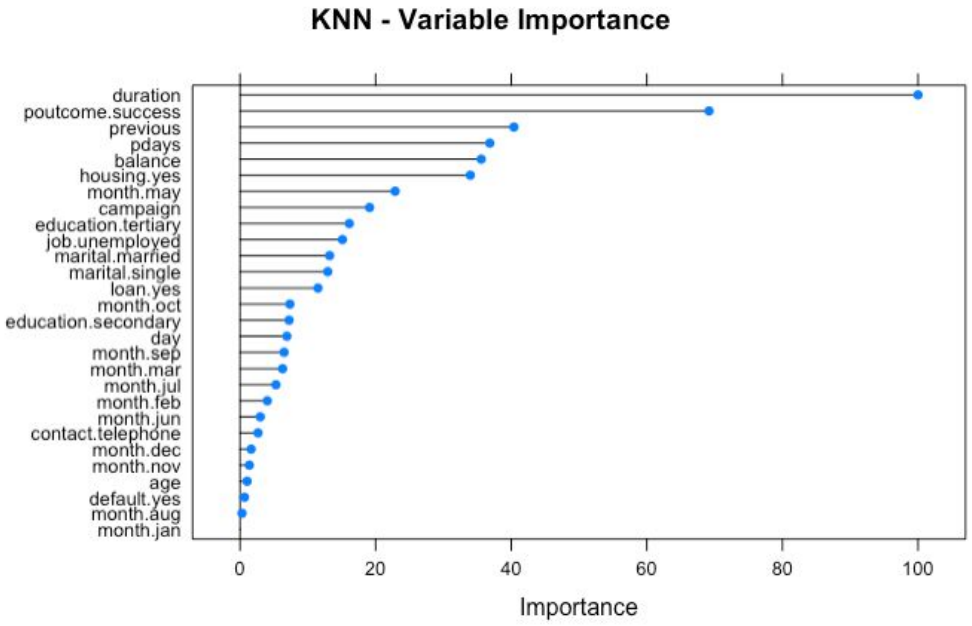
After applying 10-fold cross-validation for Neural Networks, we can see that "month" had the highest importance in the model.

8.4 k-NN

With the K Nearest Neighbour classifier, the similarity between instances or observations is measured. Each instance is then stored in the training set and the distance between all instances is determined using measures such as Euclidean Distance. The model predicts the target label by finding the nearest neighbour class. K-NN is for classification and regression. The k-NN model is sensitive to outliers and works better with a small subset of features.

```
Cross-Validated (10 fold, repeated 3 times) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction  yes   no
       yes 44.8  7.5
       no   5.2 42.5

 Accuracy (average) : 0.8723

k-Nearest Neighbors

16928 samples
   28 predictor
    2 classes: 'yes', 'no'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 15235, 15236, 15236, 15234, 15235, 15234, ...
Resampling results across tuning parameters:

  k   ROC        Sens       Spec
   5  0.9351426  0.9015826  0.8504630
   7  0.9356322  0.8965813  0.8498336
   9  0.9357684  0.8952028  0.8493605
  11  0.9357175  0.8918161  0.8499520
  13  0.9346172  0.8890193  0.8487713
  15  0.9340419  0.8870896  0.8492437
  17  0.9333177  0.8852775  0.8490071
  19  0.9329547  0.8831908  0.8481401
  21  0.9328007  0.8818913  0.8470371
  23  0.9324494  0.8822066  0.8480216

ROC was used to select the optimal model using the largest value.
The final value used for the model was k = 9.
```



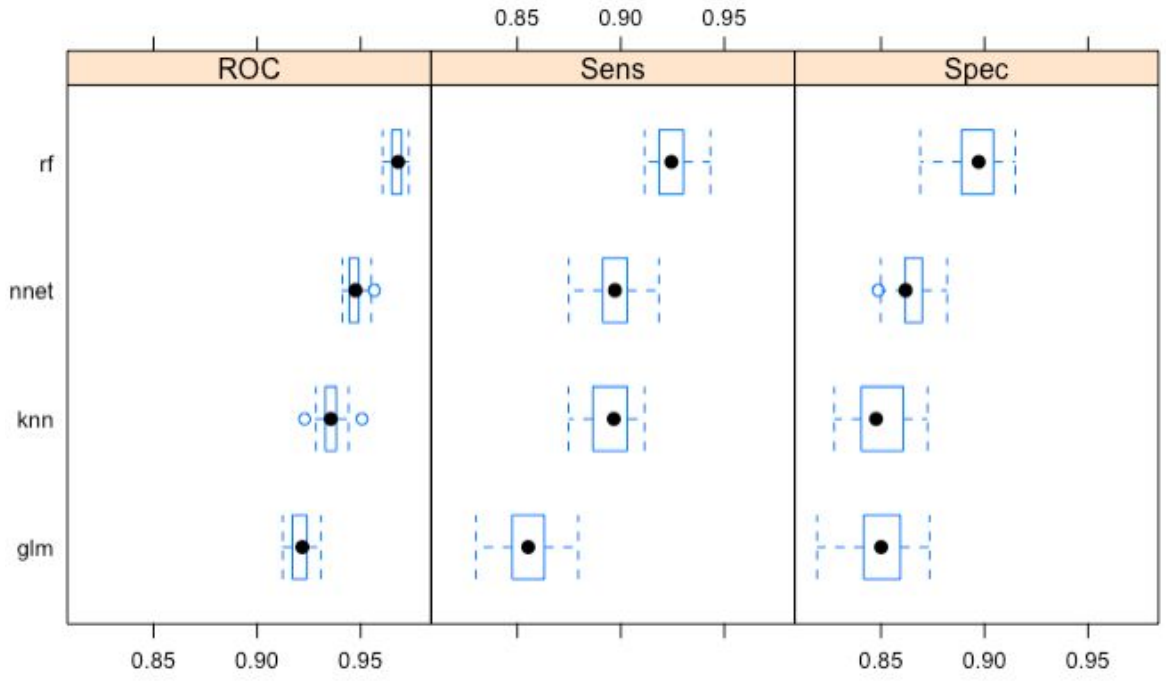


After applying 10-fold cross-validation for k-NN, we can see that “duration” had the highest importance in the model.

8.5 Model Comparison

```
Call:
summary.resamples(object = multi_models)

Models: glm, rf, knn, nnet
Number of resamples: 30

ROC
          Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
glm  0.9124234 0.9171596 0.9218282 0.9210069 0.9239618 0.9307904    0
rf   0.9606333 0.9650433 0.9681141 0.9675858 0.9696498 0.9732321    0
knn  0.9229658 0.9331884 0.9355823 0.9357684 0.9382116 0.9507045    0
nnet 0.9412082 0.9447174 0.9475269 0.9476014 0.9490062 0.9566308    0

Sens
          Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
glm  0.8299882 0.8475177 0.8552868 0.8546408 0.8621606 0.8794326    0
rf   0.9114522 0.9191030 0.9244392 0.9251351 0.9299855 0.9432624    0
knn  0.8747045 0.8874438 0.8966335 0.8952028 0.9027778 0.9114522    0
nnet 0.8748524 0.8915485 0.8971631 0.8966223 0.9028067 0.9184397    0

Spec
          Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
glm  0.8191489 0.8416076 0.8500590 0.8491640 0.8587893 0.8735225    0
rf   0.8689492 0.8888889 0.8971631 0.8955575 0.9040446 0.9148936    0
knn  0.8274232 0.8407210 0.8476077 0.8493605 0.8597993 0.8724911    0
nnet 0.8486998 0.8617021 0.8618654 0.8643267 0.8691293 0.8819362    0
```

In the output above, we can see that the random forest model had the highest ROC average over 30 resamples. We can also see that the model also had the highest specificity and sensitivity compared to the other algorithms.



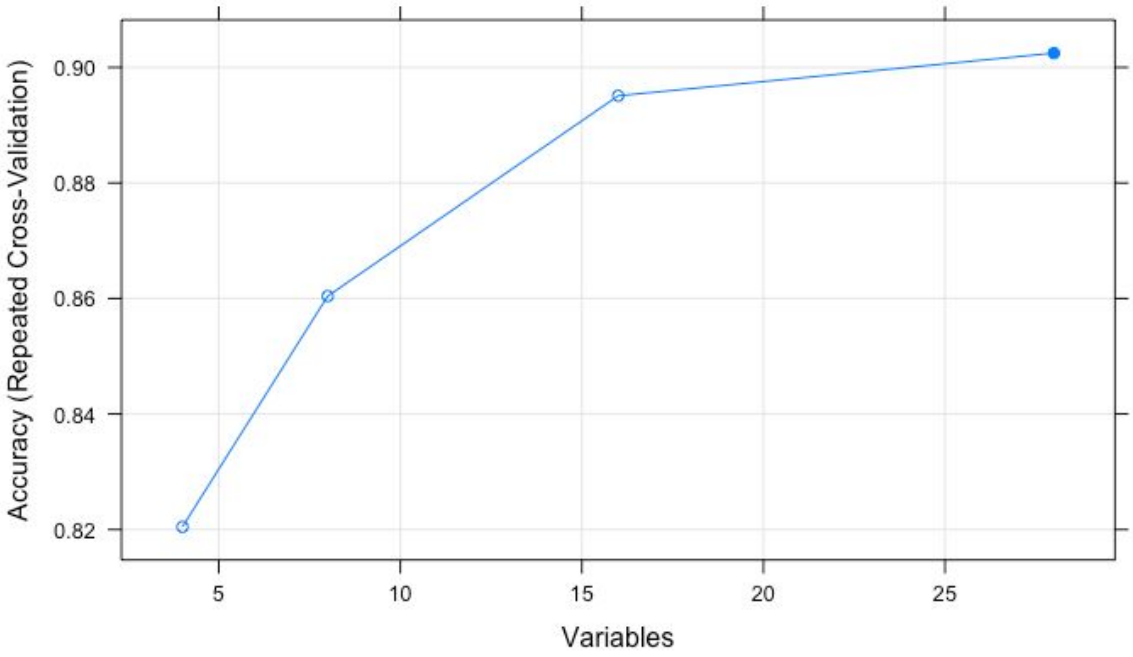
9. RFE: RECURSIVE FEATURE ELIMINATION

There are two feature selection methods, including the wrapper method and filter method. According to the report The caret Package by Max Kuhn, “Wrapper methods evaluate multiple models using procedures that add and/or remove predictors to find the optimal combination that maximizes model performance,” (Kuhn, 2019). Meanwhile, Kuhn goes onto explain filter methods. Kuhn states: “Filter methods evaluate the relevance of the predictors outside of the predictive models and subsequently model only the predictors that pass some criterion,” (Kuhn, 2019).

There are disadvantages to both approaches. Filter methods are more efficient, but the features are not selected based on model performance and highly-correlated features may be selected before each variable is evaluated individually. Meanwhile, the wrapper method takes longer computationally because parameter tuning may be necessary in order to select the variables with the highest predictive power.

For this dataset, the wrapper method: recursive feature elimination was used. Recursive feature elimination fits a model using all variables and calculates the importance or rank for each variable. The worst performing variables are eliminated at each iteration and the model is retrained to compute the importance for each variable. The feature selection process repeated until the subset-size chosen is satisfied. The best subset of predictors is selected based on the highest variable importance or rank with 10-fold cross-validation and is used in the final model. “Duration,” “poutcome.success,” “age,” “day,” “balance.”

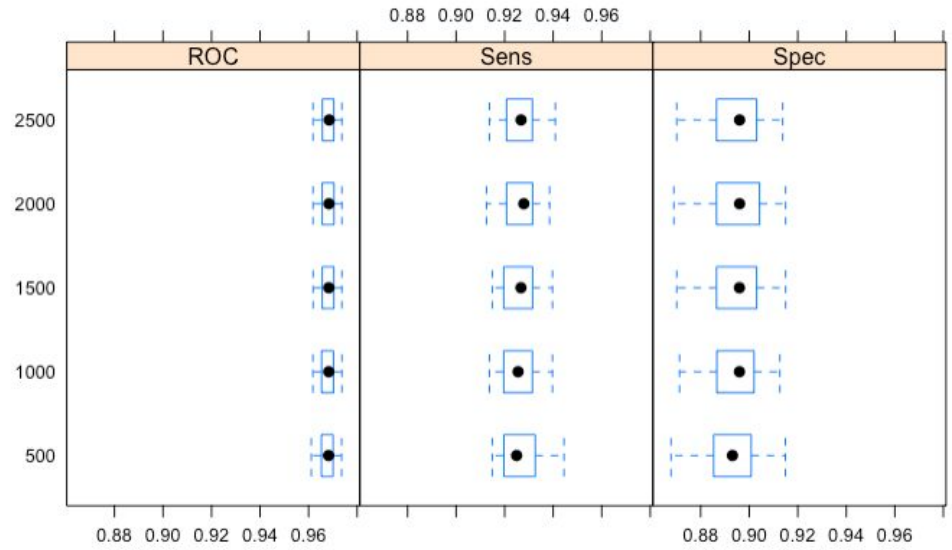
According to the book, Feature Engineering and Selection: A Practical Approach for Predictive Models by Max Kuhn and Kjell Johnson, “not all models can be paired with the RFE method, and some models benefit more from RFE than others. Because RFE requires that the initial model uses the full predictor set, then some models cannot be used when the number of predictors exceeds the number of samples,” (Kuhn & Johnson, 2019). Kuhn and Johnson go onto mention that backwards selection is often used with Random Forest models. As stated: “Increased performance in ensembles is related to the diversity in the constituent models; averaging models that are effectively the same,e does not drive down the variation in the model predictions. For this reason, random forest coerces the trees to contain sub-optimal splits of the predictors using a random sample of predictors. The act of restricting the number of predictors that could possibly be used in a split increases the likelihood that an irrelevant predictor will be used in the split,” (Kuhn & Johnson, 2019).



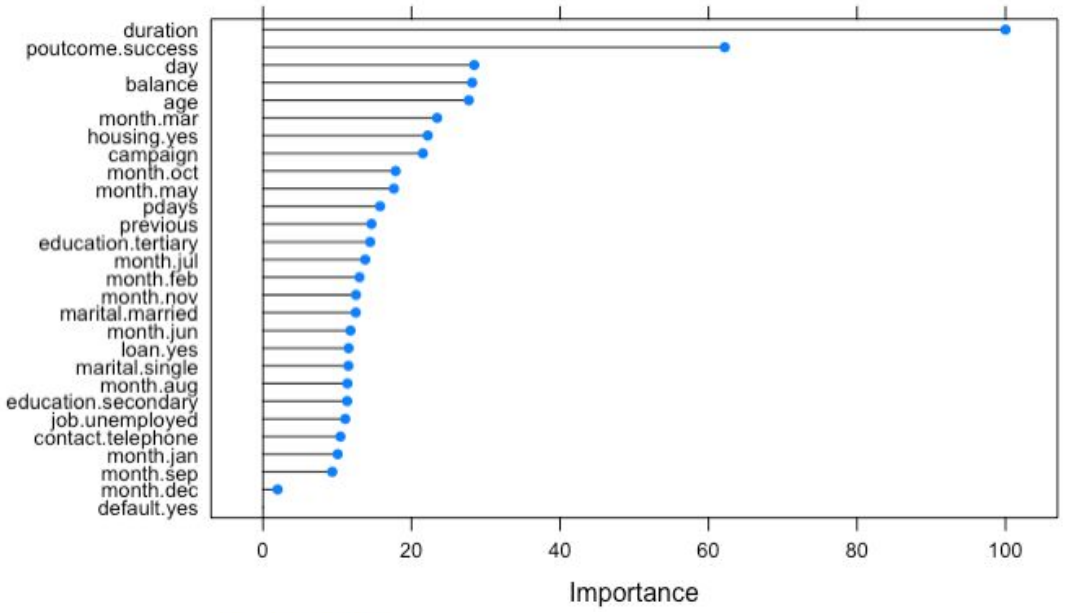
After running the recursive feature elimination we can see that the accuracy of the model does increase with all of the predictors used as opposed to the subset.

10. MODEL TUNING

```
Call:
summary.resamples(object = results)

Models: 500, 1000, 1500, 2000, 2500
Number of resamples: 30

ROC
          Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
500  0.9611107 0.9653746 0.9682251 0.9678324 0.9699916 0.9735883    0
1000 0.9617422 0.9654580 0.9682935 0.9680098 0.9702259 0.9737068    0
1500 0.9618349 0.9656535 0.9683237 0.9681026 0.9703873 0.9736900    0
2000 0.9618077 0.9656223 0.9683843 0.9681054 0.9704529 0.9737103    0
2500 0.9617965 0.9657090 0.9684764 0.9681238 0.9703814 0.9736998    0

Sens
          Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
500  0.9148936 0.9200118 0.9249409 0.9265923 0.9323286 0.9444444    0
1000 0.9137116 0.9196455 0.9255319 0.9260803 0.9314421 0.9397163    0
1500 0.9148936 0.9196455 0.9267139 0.9263952 0.9315230 0.9397163    0
2000 0.9125296 0.9208272 0.9278960 0.9263165 0.9315230 0.9385343    0
2500 0.9137116 0.9211223 0.9267572 0.9265924 0.9314421 0.9408983    0

Spec
          Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
500  0.8677686 0.8859670 0.8930888 0.8938637 0.9008264 0.9148936    0
1000 0.8713105 0.8869208 0.8959811 0.8944546 0.9019781 0.9126328    0
1500 0.8701299 0.8865583 0.8959811 0.8948093 0.9022164 0.9149941    0
2000 0.8689492 0.8868535 0.8960425 0.8952030 0.9033981 0.9149941    0
2500 0.8701299 0.8865583 0.8960425 0.8948487 0.9025116 0.9138135    0
```



RF Tuned Model - Variable Importance




```
Random Forest

16928 samples
   28 predictor
    2 classes: 'yes', 'no'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 15236, 15236, 15234, 15234, 15236, 15236, ...
Resampling results:

  ROC        Sens       Spec
  0.9680981  0.9265514  0.8937459

Tuning parameter 'mtry' was held constant at a value of 5.385165
```

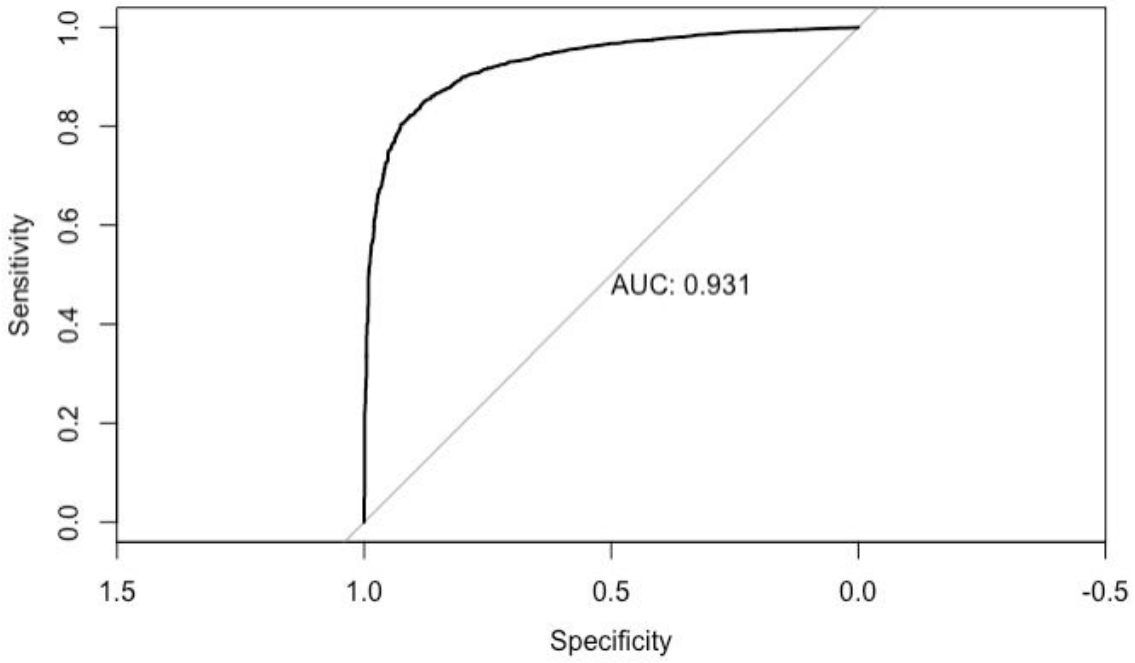
```
Cross-Validated (10 fold, repeated 3 times) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction  yes   no
       yes 46.3  5.3
       no   3.7 44.7

 Accuracy (average) : 0.9101
```

In order to find the optimal values for the random forest model, different values for *ntree* while holding *mtry* constant default value was done. You can see that the highest ROC value for *ntree* was 2500 (an increase over the first experiment using a random search for mtry). This experiment could be repeated to try different combinations of *ntree* and *mtry*.

11. TEST RESULTS

CONFUSION MATRIX	**ACTUAL VALUES**	
PREDICTED VALUES **N = 9041**	**YES**	**NO**
YES	888 TRUE POSITIVE	1023 FALSE POSITIVE
NO	169 FALSE NEGATIVE	6961 TRUE NEGATIVE

PERFORMANCE METRICS					
Accuracy	**Precision**	**F1 Score**	**Specificity**	**Sensitivity**	**Recall**
0.8682	0.4647	0.5984	0.8719	0.8401	0.8401



As mentioned earlier, false positives are less costly in comparison to false negatives because the bank would rather pay the marketing cost it would take to call the occasional customer who chooses not to subscribe to a term deposit rather than missing out on high-value customers and lose out on sales opportunities.

We can see that the model has an AUC of 0.931 when applied on the test set which means it's strong at predicting the binary classes correctly. When the Random Forest model is tuned and cross-validated, the ROC does increase from .9675858 to .9681238. Likewise, sensitivity, also known as the true positive rate or recall, increases from .9251351 to .9263165. Meanwhile, specificity, also known as the false positive rate, decreases from .8955575 to 8948487. A high sensitivity means few false negatives (incorrectly predicting subscriber as a non-subscriber) and a low specificity means many false positives (incorrectly predicting non-subscriber as a subscriber). From the graph above, we can see that the closer the ROC curve is to the upper left corner the stronger the accuracy of the test will be. As mentioned, ensuring a high recall percentage would mean that the bank wouldn't miss out on subscribers who would actually subscribe to a term deposit.

12. RECOMMENDATIONS

While the dataset does not provide the cost of each phone call during the bank's marketing campaign, however, the efficiency of the bank's strategy was increased with the assumption that it will result in cost reduction of marketing the product. Also, the length of the call was the most important feature in the dataset followed by whether the client was known to the bank; the day the call was made by the sales representative; how much money was in the client's account and the age of the client.

Works Cited

Buuren, S. V. (2018). *Flexible Imputation of Missing Data, Second Edition*. Retrieved from https://stefvanbuuren.name/fimd/sec-MCAR.html

Choong, A. (2017, October). *Predictive Analytics in Marketing A Practical Example from Retail Banking*. Retrieved from https://actuaries.org.sg/files/library/other/WebsitePracticeArea/17CommitteesAndPracticeAreas/BigDataFolder/ResearchNoteNo01bySASBDC.pdf

Ciaburro, G., & Venkateswaran, B. (2017). *Neural Networks with R: Smart models using CNN, RNN, deep learning, and artificial intelligence principles*. Retrieved from https://edu.kpfu.ru/pluginfile.php/419285/mod_resource/content/2/neural-networks-r.pdf

Demuth, H. B. (2000). *Neural Network Toolbox; for Use with MATLAB; Computation, Visualization, Programming; User's Guide, Version 4*. Retrieved from http://cda.psych.uiuc.edu/matlab_pdf/nnet.pdf

Elsalamony, H. A. (2014, January). *Bank Direct Marketing Analysis of Data Mining Techniques*. Retrieved from https://pdfs.semanticscholar.org/a911/cbe221347b400d1376330591973bb561ff3a.pdf

Ejaz, S. (2016, May). *Predicting Demographic and Financial Attributes in a Bank Marketing Dataset*. Retrieved from https://repository.asu.edu/items/38651

Fernandez, A., Garcıa, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research*, 863-905. Retrieved from https://pdfs.semanticscholar.org/fbcc/f994f4eb5b551ccc2376e8d0fe35d845a6e3.pdf

Heymans, M. W., & Eekhout, I. (2019). *Applied Missing Data Analysis With SPSS and (R)Studio*. Retrieved from https://bookdown.org/mwheymans/Book_MI/

Heymans, M. W., & Eekhout, I. (2019). *Applied Missing Data Analysis With SPSS and (R)Studio*. Retrieved from https://bookdown.org/mwheymans/Book_MI/

IBM. (n.d.). IBM Knowledge Center - Handling Missing Values. Retrieved from https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.modeler.help/missingvalues_treating.htm

IBM. (n.d.). IBM Knowledge Center - Chi-Square Test. Retrieved from https://www.ibm.com/support/knowledgecenter/en/SSLVMB_23.0.0/spss/base/idh_ntch.html

IBM. (n.d.). IBM Knowledge Center - Checking for Duplicates. Retrieved from https://www.ibm.com/support/knowledgecenter/en/SS3J58_9.1.1/com.ibm.i2.ibase.doc/checking_duplicates.html

IBM. (n.d.). IBM Knowledge Center - Handling Missing Values. Retrieved from https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.modeler.help/missingvalues_treating.htm

IBM. (n.d.). IBM Knowledge Center - Logistic Regression. Retrieved from https://www.ibm.com/support/knowledgecenter/zh/SSLVMB_24.0.0/spss/regression/idh_lreg.html

IBM. (n.d.). IBM Knowledge Center - Logistic Regression. Retrieved from https://www.ibm.com/support/knowledgecenter/zh/SSLVMB_24.0.0/spss/regression/idh_lreg.html

IBM. (n.d.). IBM Knowledge Center - C5.0 Node. Retrieved from https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.modeler.help/c50node_general.html

IBM. (n.d.). IBM Knowledge Center - Understanding Metrics. Retrieved from https://www.ibm.com/support/knowledgecenter/en/SSRU69_1.1.1/base/vision_metrics.html

IBM. (n.d.). IBM Knowledge Center - Example: Using a Two-way Chi-Square Test to Compare Default Rates and Level of Education. Retrieved from https://www.ibm.com/support/knowledgecenter/en/SSEP7J_10.1.1/com.ibm.swg.ba.cognos.ug_cr_rptstd.10.1.1.doc/t_id_example_2waycs.html

IBM. (n.d.). IBM Knowledge Center - Random Forest Node. Retrieved from https://www.ibm.com/support/knowledgecenter/en/SS3RA7_sub/modeler_mainhelp_client_ddita/clementine/python_nodes_rf.html

IBM. (n.d.). IBM Knowledge Center - Naive Bayes. Retrieved from https://www.ibm.com/support/knowledgecenter/en/SS6NHC/com.ibm.swg.im.dashdb.analytics.doc/doc/r_naive_bayes.html

Jones, S., Ye, Z., Xie, Z., Root, C., Prasuctchai, T., Anderson, J., … Roggenburg, M. (2018). *A Proposed Data Analytics Workflow and Example Using the R Caret Package*. Retrieved from http://matthewalanham.com/Students/2018_MWDSI_R%20caret%20paper.pdf

Kenn State University. (August 12). *SPSS Tutorials: Chi-Square Test of Independence*. Retrieved from https://libguides.library.kent.edu/SPSS/ChiSquare

Kuhn, M. (2019). *The caret Package*. Retrieved from https://topepo.github.io/caret/index.html
McDonald, J. H. (n.d.). Handbook of Biological Statistics. Retrieved from http://www.biostathandbook.com/spearman.html

McKinsey&Company. (2015, March). *Marketing & Sales: Big Data, Analytics, and the Future of Marketing & Sales*. Retrieved from https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/Marketing%20and%20Sales/Our%20Insights/EBook%20Big%20data%20analytics%20and%20the%20future%20of%20marketing%20sales/Big-Data-eBook.ashx

Moro, S., Cortez, P., & Rita, P. (2014, March). *A data-driven approach to predict the success of bank telemarketing*. Retrieved from http://media.salford-systems.com/video/tutorial/2015/targeted_marketing.pdf2015/targeted_marketing.pdf

Narkhede, S. (2018, June 26). Understanding AUC - ROC Curve. Retrieved from https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

UCI Machine Learning Repository. (2014). *UCI Machine Learning Repository: Bank Marketing Data Set*. Retrieved from https://archive.ics.uci.edu/ml/datasets/Bank+Marketing#

Research Data Services + Sciences, University of Virginia Library. (n.d.). The Wilcoxon Rank Sum Test. Retrieved from https://data.library.virginia.edu/the-wilcoxon-rank-sum-test/

Sarafian, R. (2019). *Introduction to Data Science*. Retrieved from https://bookdown.org/ronsarafian/IntrotoDS/

Sullivan, L. (n.d.). Nonparametric Tests. Retrieved from http://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_nonparametric/BS704_Nonparametric_print.html

findCorrelation: Determine highly correlated variables in caret: Classification and Regression Training. (2019, April 27). Retrieved from https://rdrr.io/cran/caret/man/findCorrelation.html

Two-Way Tables and the Chi-Square Test. (n.d.). Retrieved from http://www.stat.yale.edu/Courses/1997-98/101/chisq.html

NCSS Statistical Software. (n.d.). Correlation Matrix. Retrieved from https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Correlation_Matrix.pdf