**SCHOOL OF ENGINEERING AND TECHNOLOGY**

**COURSEWORK FOR THE BSC (Hons) IS (BUSINESS ANALYTICS): YEAR 3**

**ACADEMIC SESSION: MARCH 2021; SEMESTER 8 & 9  
IST 3244: ADVANCED BUSINESS ANALYTICS DEADLINE: 16 JULY 2021**

**STUDENT NAMES: Chong Khai Yien, Chua Wen Soong, Khor Vi Vien, Ting Wei Jian, Yong Mun Yeng**

**NRIC/PASSPORT NO: 990319105441, 000228101829, 000526101552 , 000707140411, 000105141564**

**ASSIGNMENT:  
INSTRUCTIONS TO CANDIDATES**This is a group assignment that covers 70% of your final coursework marks.

| **IMPORTANT** The University requires students to adhere to submission deadlines for any form of assessment. Penalties are applied in relation to unauthorized late submission of work. |
| --- |
| - Coursework submitted after the deadline but within 1 week will be accepted for a maximum mark of 50%. - Work handed in following the extension of 1 week after the original deadline will be regarded as a non-submission and marked zero. |
| **Lecturer’s Remark (Use additional sheet if required)**  **I.............................. (Name) ...................std. ID received the assignment and read the comments....................................... (Signature/date)**  **Academic Honesty Acknowledgement**  **“I Chong Khai Yien, Chua Wen Soong, Khor Vi Vien, Ting Wei Jian, and Yong Mun Yeng. verify that this paper contains entirely my own work. I have not consulted with any outside person or materials other than what was specified (an interviewee, for example) in the assignment or the syllabus requirements. Further, I have not copied or inadvertently copied ideas, sentences, or paragraphs from another student. I realize the penalties (refer to page 16,**  ***Eriq, Peter, Vi Vien, Nathan, Skylar*  15/7/2021** |

**Table of Contents**

[**Introduction**](#_3veq3h76gh1e) **1**

[**Literature Review**](#_ezgllm88bovo) **2**

[**Research Methodology**](#_ibefghj9pfvo) **6**

[CRISP-DM](#_k5hr9txljck2) 6

[Methods of modelling](#_1ro5qlbdrdf7) 11

[Regression](#_pzxv7u86g43q) 12

[Decision Tree](#_m0m8u1pn4pii) 12

[kNN](#_gcy2uuapb1s5) 12

[Random Forest](#_ggpo8rhuzk4g) 13

[Evaluation Criteria](#_gqdirj8j6s0b) 13

[**Results**](#_eqoowss4wnlr) **15**

[Regression Results](#_u8tnou9zog14) 16

[Decision Tree Results](#_eo84emd5rg6) 16

[kNN Results](#_wfljqjqdjepc) 17

[Random Forest Results](#_45uvyny0niux) 17

[Overall](#_poj41ncqvggu) 17

[**Discussion**](#_9nyp025kq60t) **21**

[**Conclusion**](#_r5arphqxxeby) **25**

[**References**](#_vxpl3dou09p9) **26**

[**Appendix**](#_pssumk7hwmvy) **28**

**Predictive Analysis on Bank Deposit Subscription**

# Introduction

The banking industry, an important sector of the economy designated to holding financial assets and investing said financial assets as a leveraged way to create more wealth. Banks are also in charge of regulating the banking activities done by the government agencies, credit cards, insurance, mortgages and so on. Banks play a central role in the majority of any type of monetary transactions, with their primary role being an intermediary between depositors and borrowers. As banks play an important role in keeping a country's economy afloat, the target of this paper is to help direct the bank’s term deposit marketing campaign to future and current clients that have a higher likelihood to subscribe. The method selected is by implementing a prediction model using the clients’ characteristics to classify them.

Everyone can subscribe to bank term deposits. However, being able to determine the best possible candidates to target for a specific marketing campaign can be difficult. Without properly identifying potential clients, a bank would have to exhaust a substantial amount of resources which includes both financial and labour in their marketing campaigns. This is because the bank would have to contact all eligible adults in the region to promote their term deposit plan. This is highly inefficient as it can cause a bank to expend unnecessary resources on unwilling customers. As such, it is important to be able to identify the characteristics of potential clients so that a more precise market can be targeted. As there has been an increase in interest from banks to fine tune their campaigns so it can be directed at specific clients, the task at hand is to find a suitable model that is able to successfully predict which future clients are more likely to subscribe to the bank’s term deposit. In this paper, we are going to discuss the various techniques used like data exploration, data cleaning, variable selection, dealing with skewed data, and developing an effective machine learning model which will be used to predict which clients are more likely to subscribe to the term deposit.

The research questions which we are planning to study in this paper are as follows:

1. What are the characteristics of clients that are subscribed to the bank term deposits?
2. Can certain characteristics exhibited by an individual successfully predict their decision to subscribe?

# Literature Review

The first study we reviewed was by Odeyemi (2020), where predictive models were created to predict whether customers would subscribe to bank term deposits. In this study, they used Python to build the predictive models. Firstly, descriptive analyses were performed to have a rough idea of the dataset. Then, the data was preprocessed by encoding the categorical columns, rescaling numerical columns and splitting the dataset. The independent and dependent variables were also decided, where “subscribed” will be the target column. Dimensionality reduction was also performed on the dataset using Principal Component Analysis (PCA). Since it was found that the data was not balanced, Synthetic Minority Oversampling Technique (SMOTE) was performed to deal with the class imbalance in the training data set. Finally, the machine learning models that were created include logistic regression, XGBoost and Multi Layer Perceptron (MLP). The cross validation techniques used were K-Fold and Stratified K-Fold. These models were evaluated using accuracy, Area Under the ROC Curve (AUC), precision, recall and F1 score. It was found that the best performing model was XGBoost, with an accuracy of 87.23%, AUC of 94.55%, precision of 88.60%, recall of 85.43%, and F1 score of 86.88%.

The next study is similar in fashion, where Python was used to build the predictive models as well. In the data preprocessing stage, John (2020) used PyOD to clean the outliers. Label encoding and one-hot encoding were used to convert and transform the categorical variables. StandardScaler was also used to rescale all the numerical columns. Unlike the previous study, random oversampling was used to overcome the imbalanced data since 88.7% of the customers in the dataset declined and only 11.3% accepted. . In similar fashion, PCA was also used in this study to reduce dimensionality in the data. Then, the models created were XGBoost, MLP, logistic regression and Support Vector Machine (SVM). The cross validation techniques used were also K-Fold and Stratified K-Fold. The results of this study were aligned with those of Odeyemi (2020), where the XGBoost outshined the other models. It had an accuracy of 0.88, recall of 0.90, F1 score of 0.89 and ROC score of 0.88. This shows that John’s (2020) XGBoost model is better than that of Odeyemi (2020).

In this study, the author, Boateng (n.d), chooses to use R for their analysis on the same dataset. The data was first split into 70% train and 30% test. Then, recursive feature elimination (RFE) was used for feature selection. RFE is an efficient method in eliminating features from a training dataset for training. It works as follows: fitting a machine learning algorithm first, then it will rank the features by importance while discarding the least important features. After that, it will refit the model with the machine learning algorithm again. This process repeats until the desired number of features are taken. In the author’s case, 5 out of 16 features are selected, providing a 90% fit for their models. The models that the author uses are boosted log regression, extreme gradient boosting, random forest, support vector machine with radial kernel, generalized least squares, and Bayesian regularized neural network. The metrics that the author used are Area Under Lift Curve (ALIFT), accuracy and Receiver Operating Characteristic (ROC), however ROC is used to determine the best model. Their best model was found to be the Bayesian regularized neural network as it boasted the highest ROC.

The paper by Moro et al. (2014b) uses customer lifetime value and neural networks to improve prediction of bank deposit subscription in telemarketing campaigns. For feature selection, it is grouped into blocks for forward selection. Few features from LTV are then added into the logical blocks during forward selection. The features are discarded if AUC & ALIFT show no improvement. The models used in this paper are logistic regression, decision trees, support vector machine and neural network. The metrics used are AUC and ALIFT. The most accurate predictive model was then found to be the neural network model.

Moro et al. (2014a) suggests a data mining approach to predict whether customers will subscribe for a bank term deposit. Unlike the previous studies, this study used the rminer package in R Studio. Firstly, they identified that the dataset is unbalanced, where only 12.38% of the records had customers who were subscribed. The data was also time-ordered split where 4 years were placed into the training set and 1 year was placed in the test set. Next, the models Moro et al. (2014a) tested were logistic regression, decision tree, neural network and SVM. For feature selection however, the features were selected based on business intuitive knowledge, then an automated feature selection approach was used. This ended up with 22 relevant features. Then, the models were evaluated using ROC, AUC, and Area under the LIFT (ALIFT). In the end, the neural network model appeared to be the best resulting model, with an AUC score of 0.8 and ALIFT of 0.67.

The study by Muoki (2020) uses Python to seek a machine learning model that is able to predict which future client would subscribe to their term deposit. It begins with acquiring the same bank term deposit dataset from the UCI ML website. Then, it performs data pre-processing by first encoding categorical variables into numeric values. After that, outliers were handled by replacing them with the central measure of tendency, mean. Outliers are said to be removed as they would skew the data and affect the training process of a machine learning algorithm which results in an overall loss of accuracy in the models. The next step involves the scaling of all numerical columns using standardscaler in the SciKit Learn library. This step is important because most machine learning algorithms use Euclidean distance between two data points in their computations. Thus, standardisation scales the variables to a common magnitude, unit, and range which reduces the impact of one particular feature dominating other features in the dataset. Furthermore, dimensionality reduction techniques are used to reduce features based on their relative importance to the models. The author used four dimensionality reductions techniques, which are t-distributed stochastic neighbour embedding, autoencoders, and principal component analysis. Autoencoder's technique was chosen in the end as it can compress the information better into low dimensional latent space resulting in a higher accuracy for the prediction models later. The study also made use of the stratified k-fold cross-validation technique in selecting the best machine learning model. It is picked over the k-fold cross-validation technique as stratified k-fold generates test sets containing the same distribution of classes while preserving order dependencies in the dataset ordering. Before moving on to the predictive models, the author highlights the use of AUC score, F1 score, accuracy, precision, and recall. The predictive models used are logistic regression, XGBoost, multi-layer perceptron, support vector machine, decision trees, and random forest. In the end, the model was chosen based on its accuracy, interpretability, complexity, and scalability of the model. The model chosen was XGBoost. Its AUC score is 0.7927, F1 Score is 0.2621, accuracy is 0.8852, precision is 0.6176, and recall is 0.1663.

Studies have been previously done on this Bank Institution Term Deposit Predictive Model dataset that was curated by Moro et al. (2014a). These studies were performed with the goal of producing a predictive model that can accurately predict whether the client would subscribe to the term deposit based on 20 variables.

In one of the studies from Moro et al. (2014b), an additional construct, Lifetime Value (LTV) was built using Recency, Frequency, and Monetary (RFM) features that are used often in telemarketing research. RFM concept is widely known in literature, but its definition varies from research to research. In this study, the RFM features from the dataset were grouped into logical blocks and forward selection was performed on these blocks to find features that affected Area Under The Curve (AUC) and Area of the Lift Cumulative Curve (ALIFT curves).

In the preprocessing stage, John (2020) highlighted the imbalance dataset (88.7% decline, 11.3% accepted). Synthetic Minority Oversampling Technique (SMOTE) was used to overcome this imbalance (Odeymi, 2020). Furthermore, features were reduced using Principal Component Analysis (PCA) and then selected with Recursive Feature Elimination (RFE), Forward Selection, and Auto Encoder.

The studies have used many techniques such as Logistic Regression, XGBoost Classifier, Gradient Boosting, Random Forest, Decision Trees, Support Vector Machines, Multi Layer Perceptron, and Neural Networks. The metrics used to compare between models are accuracy, AUC, Precision, Recall, sensitivity, specificity, lift and ALIFT. It is shown that XGBoost and Neural Networks have the highest accuracy among all the models.

In our study, machine learning algorithms are used, namely, Logistic Regression, Decision Tree, Random Forest, and k-NN classifier. The main metrics used are accuracy, misclassification rate, precision, recall, ROC, and F1 Score.

# 

# Research Methodology

## CRISP-DM

The research methodology that will be used for this project is Cross Industry Standard Process for Data Mining (CRISP-DM). It is a standardised structured process for conducting data science projects. It involves 6 stages, business understanding, data understanding, data preparation, evaluation and deployment. The details of each stage are as follows.

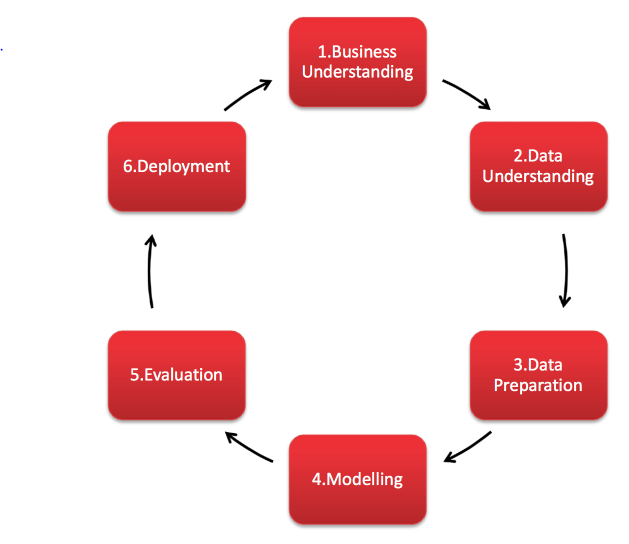


Figure 1: CRISP - DM Methodology

**CRISP-DM: Business Understanding Step**

The CRISP-DM framework begins with business understanding of the problem. This part focuses on understanding the project objectives and requirements from a business perspective so that it fulfills the client’s needs. The objective and requirements are then used to define the data mining problem and a project proposal is constructed.

**CRISP-DM: Data Understanding Step**

The second step of CRISP-DM is data understanding, which includes data retrieval and exploration to get a clear idea of the data to be analysed. The data is collected through primary or secondary data collection. Then, the data is explored by identifying key attributes, finding the distribution of important attributes, and looking for relationships between pairs of attributes.

**CRISP-DM: Data Preparation Step**

CRISP - DM’s third step is data preparation. This step consists of selecting the data, cleaning the data, constructing data, and integrating data. The data selected depends on the analyst. Some data features may be discarded because of low data quality or some other rationale for exclusion. Data cleaning is done to improve data quality for data analysis techniques that are selected. Next, required data such as derived attributes or generated records may be constructed to suit the analysis’ purposes. Integration of data is done by merging tables or aggregating records. It is done to combine similar information throughout multiple datasets.

**CRISP-DM: Modelling Step**

In the fourth step of the CRISP-DM methodology, the modelling phase, the modified data is put into a modelling tool to create multiple models. It is also at this phase when the modelling techniques to be used are decided. For example, a project could use support vector machines and random neural networks to create their model.

**CRISP-DM: Evaluation Step**

The evaluation step is important, as it assesses which model meets the business objectives and see whether it is useful to be used in deployment. The assessment metrics used are based on the modelling techniques used. For example, classification tasks use F1 score, accuracy, misclassification rate or AUC-ROC metrics to assess the model’s performance. After the models are evaluated and reviewed thoroughly, the next step, ‘deployment’ can begin.

**CRISP-DM: Deployment Step**

The final step of the CRISP-DM framework is deployment. In this step, the model created and evaluated is implemented into the business to help with its decision making. Besides that, a final report is to be written and the findings are presented.

**Data Understanding**

The dataset is retrieved from the online data catalog data.world (Moro et al., 2014a). It includes information about clients of a bank and is used to determine if they will subscribe to a bank term deposit. There are a total of 41,188 records collected from May 2008 to November 2010 with 20 variables. The variables of the dataset are as follows:

| **Variable** | **Description** | **Possible values** |
| --- | --- | --- |
| age | The client’s age | Numeric |
| job | The client’s occupation (separated into several categories) | admin, blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed, unknown |
| marital | The client’s marital status | divorced, married, single, unknown |
| education | The client’s highest education level | basic.4y, basic.6y, basic.9y, high.school, illiterate, professional.course, university.degree, unknown |
| default | Whether the client has credit in default | no, yes, unknown |
| housing | Whether the client has a housing loan | no, yes, unknown |
| loan | Whether the client has a personal loan | no, yes, unknown |
| contact | The type of communication method | cellular, telephone |
| month | The month of the last contact with the client | jan, feb, mar, apr, may, jun, jul, aug, sep, oct, nov, dec |
| day\_of\_week | The day of the week of the last contact with the client | mon, tue, wed, thu, fri, sat, sun |
| duration | The duration of the last contact with the client in seconds | numeric |
| campaign | The number of contacts with the client during the current campaign | numeric |
| pdays | The number of days which passed since the last contact with the client from the previous campaign | numeric (999 means no previous contact) |
| previous | The number of contacts with the client before this campaign | numeric |
| poutcome | The outcome of the previous marketing campaign | failure, nonexistent, success |
| emp.var.rate | The employment variation rate | numeric |
| cons.price.idx | The consumer price index | numeric |
| cons.conf,idx | The consumer confidence index | numeric |
| euribor3m | The euribor 3 month rate | numeric |
| nr.employed | The number of employees that are employed | numeric |
| y | Whether the client subscribed to a bank term deposit | yes, no |

Table I: The name, description, and possible values of variables found in the dataset.

There are no null values in the dataset.

**Data Preparation**

Several modifications were done to the data as a preparation step before carrying out analysis on the dataset. One of the modifications that were done involves the variable “pdays”. The values for the variable “pdays” values were binned into two categories which are “Contacted” and “Not Contacted”. Values which are less than 999 were binned into “Contacted” while the values of 999 are binned into “Not Contacted”. This was done as all the data values are either 999, which represents a lack of contact during the previous campaign, or are in the range 0 - 27.

Besides that, certain values of the variable “job” were also recoded. The value “entrepreneur” was recoded as “self-employed” and “technician” was recoded as “blue-collar”. Both the values “admin.” and “management” were recoded as “white-collar”. Meanwhile the values “services” and “housemaid” were recoded as “pink-collar”.

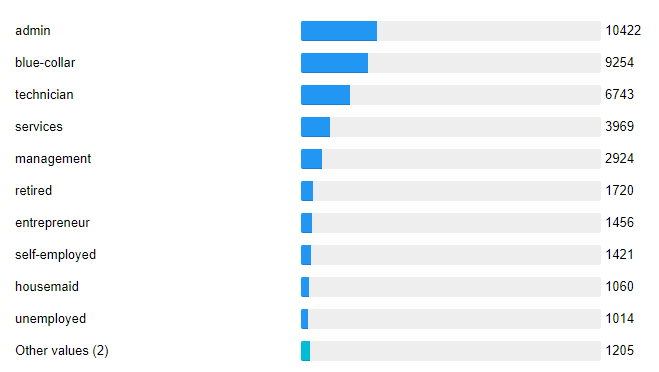


Figure 2: Before Recoding “job”

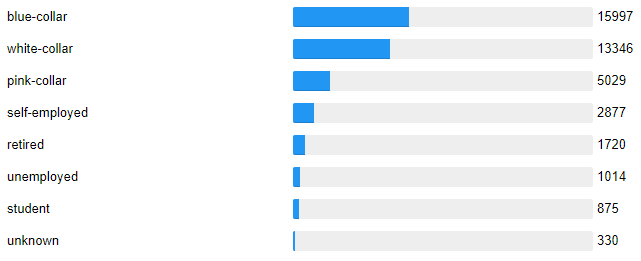


Figure 3: After Recoding “job”

There were also variables such as “duration” and “campaign” which were skewed. Data skewness refers to the presence of asymmetry in a statistical distribution, in which the distribution of data will appear distorted or skewed due to a few outliers. To overcome this issue, they were log-transformed to reduce skewness of their distributions.

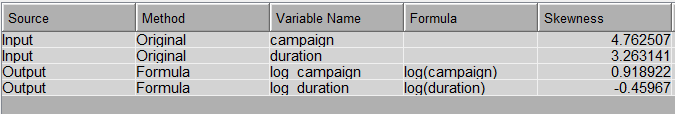
****

Figure 4: Skewness of “duration” and “campaign” Before and After the Transformation

## 

## Methods of modelling

The models generated in this research are classification models as they are used to classify a customer into groups of “yes” or “no” in the target variable “y”. SAS Enterprise Miner was used to carry out the modelling process. Several models were created using different methods which include regression, decision tree, k-nearest neighbours (kNN) and random forest. Before the generation of the models, the dataset was partitioned into 50:20:30 as well as 60:20:20 ratios for training, validating and testing the models to see whether different partitioning ratios can provide a better model.

### 

### Regression

Regression analysis is a basic method in statistics used to determine if a significant relationship exists between multiple variables (Chatterjee & Hadi, 2015). The regression model generally follows an equation, which is

Y = m0 + m1 X1 + m2 X2 + m3 X3 + … mn Xn + ϵ

where Y is the dependent variable, Xn are the independent variables, mn are the coefficients and ϵ is the random error of the model. In this paper, logistic regression will be used to generate the model as there are multiple qualitative data within the dataset.

### Decision Tree

The decision tree model is a hierarchical model that utilises decision rules to break down the dataset to be able to predict an occurrence (Song & Lu, 2015). These decision rules are usually in the form of inequality expressions relating to one or more predictor variables. There are several statistical values that can be considered when determining the best possible leaf node for a model, such as entropy, Gini index and gain ratio. However, this paper will be using chi-square p-values to evaluate the variables and determine the optimal leaf nodes.

### kNN

Another algorithm that will be used to carry out the predictive analysis is kNN. This algorithm makes use of the entire training dataset as it calculates the distance between a new data and the existing data and groups k amount of records closest with the data. After which, the value for the target variable with the highest frequency in the group is assigned as the value for the new data (Harrison, 2018). In the case of a binary target variable, it is recommended to use odd numbers for the value of k. In this paper, we will be analyzing the data with a model where k = 1, 3, 5 and 7.

### Random Forest

The fourth and final algorithm that we will be utilizing in our research is random forest. A random forest is an ensemble method of modelling which uses a set of hyperparameters before beginning the training of the model to create several decision trees. These hyperparameters include the size of the node, the amount of trees to be created and the amount of attributes to sample (IBM Cloud Education, 2020). The predicted class will be decided through the value of each decision tree, whereby the most common value of the target variable among all the trees created will be taken as the prediction. In this paper, we will be using the Gini importance value to rank the variables to determine the most important attributes in determining whether a user subscribes to a bank term deposit.

## 

## Evaluation Criteria

To measure the efficacy of the predictive models, an evaluation criteria is used. In literature, the evaluation criteria of predictive models normally consist of accuracy, misclassification rate, precision positive or precision negative, recall positive or recall negative, AUC-ROC curve, and F1-Score. They are calculated based on a confusion matrix, which shows the values of True Positive (TP), True Negative (TN), False Positive (FN), and False Negative (FN).

Below shows the formula of each evaluation metric.

Accuracy: Probability of the model predicting correctly for any one record. It is also the converse of the misclassification rate.

Misclassification Rate: Probability of the model predicting wrongly for any one record. It is also the converse of accuracy.

Precision Positive: Proportion of predicted actual positives over predicted positives

Precision Negative:Proportion of predicted actual negatives over predicted negatives.

Recall Positive: Proportion of predicted actual positives over total actual positives

Recall Negative: Proportion of predicted actual negatives over total actual negatives

ROC: Receiver Operator Characteristic (ROC) plots a graph of Sensitivity (recall positive) against Specificity (recall negative. A model with high discrimination, will have high Sensitivity at low Specificity at the same time. The area under the ROC curve (AUC) can also be interpreted as the accuracy of the model.

F1-Score: Measure of the accuracy of the model using precision positive and recall positive

Accuracy and misclassification rate are not as useful because this is an imbalanced dataset. To illustrate why accuracy and misclassification rate is a problem here, the dataset’s class imbalance is about 88.7% of people who subscribe to the term deposit to 11.3% of people who do not subscribe to the term deposit. Thus, if the model predicts everybody to decline the deposit, it will be trivial to reach 88.7% accuracy.

The most important evaluation metrics to measure the model performance in this case would be ROC and recall positive. Recall is taken as it is more important to reach more potential customers (people who accept the term deposit) as compared to precision which identifies the probability of the model predicting correctly whether a person will accept the term deposit. ROC is a graph that shows recall positive against recall negative. The larger the AUC of the graph, the better the model.

However, all metrics will be shown in the results section for clarity and comparison between models. All metric calculated and obtained will be based on the testing data set.

# Results

A total of 16 models are created, differing in the data partition ratio, model algorithm and algorithm specific features. Table II lists the model specifications as well as the name for the models which we will be using to refer to the individual models throughout this paper.

| **Model Name** | **Model Description** |
| --- | --- |
| 50 kNN, k=1 | kNN model where the value of k is 1 and the partition ratio is 50:20:30 |
| 50 kNN, k=3 | kNN model where the value of k is 3 and the partition ratio is 50:20:30 |
| 50 kNN, k=5 | kNN model where the value of k is 5 and the partition ratio is 50:20:30 |
| 50 kNN, k=7 | kNN model where the value of k is 7 and the partition ratio is 50:20:30 |
| 50 Stepwise Regression | Regression model with stepwise selection and 50:20:30 partition ratio |
| 50 Regression | Regression model with no selection and 50:20:30 partition ratio |
| 50 Decision Tree | Decision tree generated using 50% of the dataset for training |
| 50 Random Forest | Random forest generated using 50% of the dataset for training |
| 60 kNN, k=1 | kNN model where the value of k is 1 and the partition ratio is 60:20:20 |
| 60 kNN, k=3 | kNN model where the value of k is 3 and the partition ratio is 60:20:20 |
| 60 kNN, k=5 | kNN model where the value of k is 5 and the partition ratio is 60:20:20 |
| 60 kNN, k=7 | kNN model where the value of k is 7 and the partition ratio is 60:20:20 |
| 60 Stepwise Regression | Regression model with stepwise selection and 60:20:20 partition ratio |
| 60 Regression | Regression model with no selection and 60:20:20 partition ratio |
| 60 Decision Tree | Decision tree generated using 60% of the dataset for training |
| 60 Random Forest | Random forest generated using 60% of the dataset for training |

Table II: Model Name and Descriptions

## Regression Results

In this analysis, two logistic regressions were carried out for both the partitions. The difference between the two is that one does not have any variable selection method while the other employs a stepwise selection. As a result, one of the regression models takes into consideration every variable whereas the other only includes a portion of the variables. The variables included in the stepwise logistic regression are cons\_price\_idx, contact, day\_of\_week, default, emp\_var\_rate, euribor3m, log\_duration, month, pdays, poutcome, and previous when the partition is set to 50%. However, when the 60% of the dataset is used for training, it is noted that previous is no longer included in the stepwise logistic regression. From the results of the model, it can be seen that the regression model without a selection method has an accuracy of 91.26% while the model achieved through stepwise selection has an accuracy of 91.24% when 50% of the dataset is used for the training of the model. On the other hand, when the partition is set to 60%, the stepwise regression has an accuracy of 91.28% whereas the simple logistic regression model has an accuracy of 91.41%.

## Decision Tree Results

Through SAS Enterprise Miner’s decision tree node, a decision tree with a leaf size of 5 has been created for each of the partition values . The maximum depth of the tree was set to 20 to prevent underfitting of the data. As a result, two decision trees, both with a depth of 7 and a total of 18 leaf nodes, were generated. It was found that the decision tree model created using 50% of the dataset as the training dataset was the best model among all the models based on its misclassification rate of 8.50% on the testing dataset. The decision tree is attached in Appendix O.

## 

## kNN Results

To discover the best value for k or the amount of neighbours for the kNN model, 4 kNN models were generated using 1, 3, 5 and 7 as the values of k for each of the partition ratios. From the models that were generated, it can be observed that the model where k=7 has the best accuracy for both the partitions at 90.07% for 50% and 90.22% for 60%. However, all of the models created are still looked into as there may be different interpretations as to which model is the best when other values such as recall positive and precision negative.

## Random Forest Results

In this project, we generated two random forests with a difference in the amount of data used to train the model, whereby one uses 50% of the dataset and the other uses 60%. The hyperparameter set for the random forest models is 100 trees with a maximum depth of 50 and 25 variables. It can be seen that while both of these models have a lower misclassification rate (9.14% and 9.18%) when compared to the other models, they have a higher ROC index (0.939 and 0.941).

## Overall

After looking at the models individually, the ROC charts were also generated alongside calculating the different values based on the confusion matrix for each of the models. The ROC chart can be seen in Figure 5 whereas the compilation of values are recorded in Table III. The misclassification rate and ROC index of the different models are also obtained with the help of SAS Enterprise Miner and are shown in Figure 6. The confusion matrix evaluation metrics for the various models are recorded in Table I whereas the confusion matrices for each individual model are attached in the appendix as Appendix A-P. Moreover, the variable importance for our best models based on the misclassification rate and ROC index are shown in Figure 7 and Figure 8.

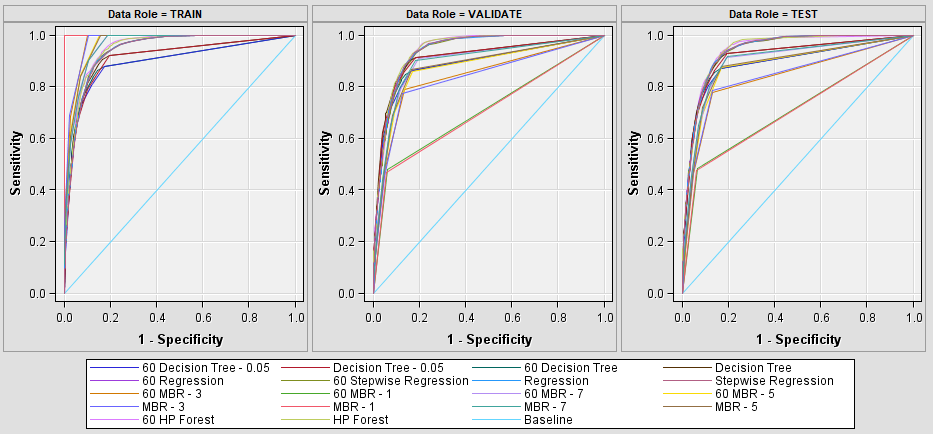


Figure 5: ROC Chart

|  | Accuracy | Misclassifi-cation Rate | Precision Positive | Precision Negative | Recall Positive | Recall Negative | F1 - Score |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 50 kNN, k=1 | 88.36% | 11.64% | 48.32% | 93.34% | 47.41% | 93.56% | 47.86% |
| 50 kNN, k=3 | 89.65% | 10.35% | 54.74% | 93.38% | 46.91% | 95.08% | 50.52% |
| 50 kNN, k=5 | 89.96% | 10.04% | 56.80% | 93.23% | 45.33% | 95.62% | 50.42% |
| 50 kNN, k=7 | 90.07% | 9.93% | 57.58% | 93.21% | 45.04% | 95.79% | 50.54% |
| 50 Stepwise Regression | 91.24% | 8.76% | 66.56% | 93.27% | 44.76% | 97.15% | 53.53% |
| 50 Regression | 91.26% | 8.74% | 66.46% | 93.23% | 45.26% | 97.10% | 53.85% |
| 50 Decision Tree | 91.50% | 8.50% | 62.82% | 94.97% | 60.20% | 95.48% | 61.48% |
| 50 Random Forest | 90.86% | 9.14% | 75.89% | 91.49% | 27.59% | 98.89% | 40.47% |
| 60 kNN, k=1 | 88.37% | 11.63% | 48.38% | 93.44% | 48.28% | 93.46% | 48.33% |
| 60 kNN, k=3 | 89.48% | 10.52% | 53.82% | 93.31% | 46.34% | 94.95% | 49.80% |
| 60 kNN, k=5 | 89.86% | 10.14% | 56.33% | 93.15% | 44.61% | 95.61% | 49.79% |
| 60 kNN, k=7 | 90.22% | 9.78% | 58.47% | 93.26% | 45.37% | 95.91% | 51.09% |
| 60 Stepwise Regression | 91.28% | 8.72% | 67.50% | 93.15% | 43.64% | 97.33% | 53.01% |
| 60 Regression | 91.41% | 8.59% | 67.80% | 93.32% | 45.15% | 97.28% | 54.20% |
| 60 Decision Tree | 91.39% | 8.61% | 64.66% | 94.06% | 52.10% | 96.39% | 57.70% |
| 60 Random Forest | 90.82% | 9.18% | 75.90% | 91.45% | 27.16% | 98.91% | 40.00% |

Table III: Values of Evaluation Metrics Obtained Through Confusion Matrix

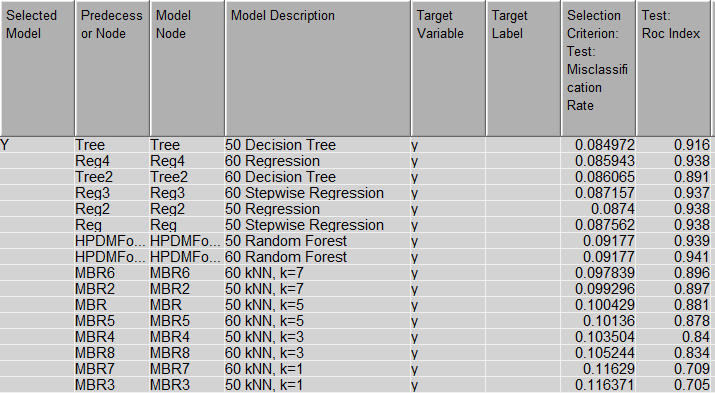


Figure 6: Misclassification Rate and ROC Index generated by SAS Enterprise Miner

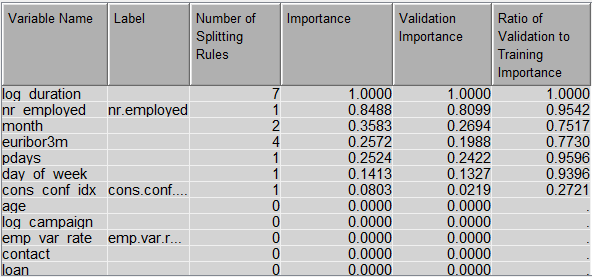


Figure 7: Variable Importance for 50 Decision Tree

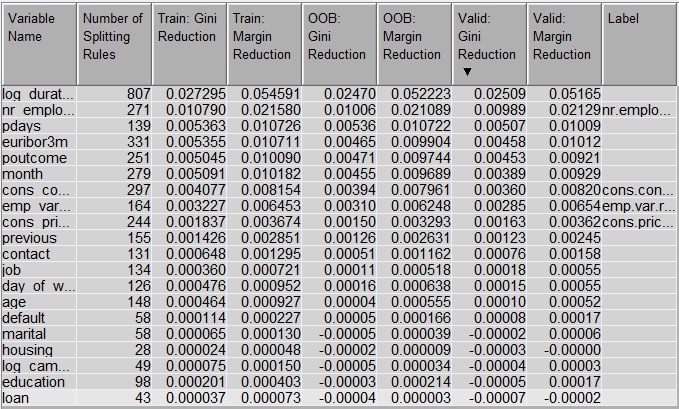
****

Figure 8: Variable Importance for 60 Random Forest

# Discussion

In the following scenarios, the performance of each proposed model of kNN, random forest, regression, and decision tree is compared. Following the values shown in Table I, the decision tree with 50% data in the training dataset, leaf size of 5 and maximum depth of 20 scored the highest among the models in terms of accuracy at 91.50%. In addition, the decision tree model scored the lowest misclassification rate among the models at 0.0849, which means that the decision tree model is likely to only misclassify 8.5% of the results. Furthermore, when comparing the recall value of the models, the decision tree scored the highest at returning 60.20% of the relevant instances in ratio to the total relevant instances. However, when comparing the precision achieved by each model, the random forest model with 60% data partitioned into the training dataset scored the highest at returning 75.90% of relevant instances in ratio to the total instances. This could mean that while the random forest is able to classify and rank the patterns well, the selection of threshold is bad, thus leading to a high ROC index but lower overall accuracy.

Figure 5 plots the ROC curves for all the models tested. A good model should have high sensitivity and low specificity. According to the ROC chart in Figure 5, the kNN model with k value of 1 outperformed every other model in the training set. However, it performed the worst in the validation and testing set. The random forest models with 50% and 60% data partitioned in the training sets showed the most consistent performance throughout the training, validation and testing sets. Although it was outperformed by all the kNN models in the training sets, both random forest models outperformed every other model in both validation and testing sets as shown in the ROC chart. This is further supported by the ROC index achieved by the models, with the random forest model of 50% data partitioned in the training dataset scoring 0.941 which is higher than the rest of the models.

Comparing the F1-Score of each model which is tabulated in Table III, the decision tree model with 50% data partitioned into the training set scored the highest at 0.6148, followed by a F1-score of 0.5750 by the decision tree model with 60% data partitioned into the training dataset. Then, the regression model with 60% data partitioned into the training set scored 0.5420. The lowest scoring model is the random forest model with 60% data partitioned into the training dataset with a F1-Score of 0.40. This indicates that the actual classification accuracy of our models are not optimal. This may be due to the class distribution varying from a large range of values thus reflecting the low overall F-1 score across all the models.

In comparison to the models done by Golecha (2017), the model found to have the highest accuracy of 93.37% was an Adaboost model. Following the Adaboost model, the decision tree model outperformed both support vector machine and logistic regression models in terms of accuracy. The same can be said for the output achieved by our models, where the decision tree model scored the highest in accuracy when compared to our logistic regression and kNN models. According to a research done by Muoki (2020), the XGBoost model was deemed the best performing model with a 88.52% accuracy, followed by decision tree and logistic regression models with 88.44% accuracy each, followed closely by the random forest model at 88.40% Similarly, our results are more or less the same as Muoki’s (2020) findings, where decision tree was found to perform better than the logistic regression, followed by the random forest models. This is supported by the accuracy of the decision tree model scoring higher than the logistic regression model. Though, it should be noted that Muoki (2020) used an auto-encoder to select the features. Hence, there might be potential research gaps in feature selection when it comes to building predictive models for this dataset.

When compared to a paper done by John (2020), the models generated have a much higher accuracy with the XGBoost classifier achieving a F1-score of 0.88, which is significantly higher than our highest F1-score of 0.6148. This could be due to John using a dimensionality reduction method using principal component analysis. By doing this step, the data will be linearly mapped to achieve a lower dimensional space in such a way that data in low dimensions can have similar impact as high dimensional data. This could prove to be useful as the dataset we used has data ranging from very low dimensionality to very high dimensionality. However, when comparing classification accuracy, the XGBoost model scored 88% which is lower than our best model with 91.50% accuracy. This can be interpreted as our model is better at ranking the test data, with most of the positive results on one end and negative cases on the other end. However, since the data used is considered highly skewed, the F1-score should be used to select the best model.

To address the research questions, the factors that were most prominent in predicting whether clients would subscribe to the bank’s term deposit were number of employees, duration of last contact and the last contact month of the year. To be specific, 5087.65 employees, 6.27 seconds of last contact, and contact during May, June, July, August and November. Other than that, none of the client characteristics such as age, job, marital status and education level were significant factors in predicting their decision to subscribe to a bank term deposit. Client purchase subscription history such as a housing loan or personal loan did not have a significant effect either. According to the variable importance of 50 decision tree, there are also 4 variables that are more important than the rest. They are duration of last contact, number of employees, number of days passed since last contact, the day of the week, month of contact and the euribor3 month rate. These features were pointed out because the ratio of validation to test of more than 0.50. However, there is only one variable that is clearly more important than the rest in the 60 random forest, which is duration of last contact.

After completing the entire process from the research methodology to the analysis, we have observed that there are several limitations which may be affecting the outcome of the analysis carried out in this study. This may cause the outcome to be not the most accurate at reflecting the desired output. One of the limitations that we determined is the timeliness of the data. The data used in this analysis was collected around a decade ago from 2008 to 2010. There have been many technological as well as cultural changes which have occurred between the time the data was collected 10 years ago to the present time where the data is analysed. This means that the data does not accurately portray the characteristics of the society of present times. Moreover, if the analysis is carried out using the data collected today, we may have drastically different results as the thought process and social norms have evolved. Additionally, the distribution of the data has also led to several variables being highly skewed, which in turn, affected the evaluation metrics.

Besides that, after inputting the data into Python, we also noticed that the values of the targeted variable in the dataset are not balanced as shown in Figure 9 below. A total of 88.7% of the clients had declined to subscribe to the term deposit, leaving only 11.3% of the clients who had subscribed to it. Having such a large difference in proportion between the values of the target variable may cause the model to be biased. It would also become not suitable for predictive analysis and may possibly incorrectly predict the factors which contribute to the target variable. This means that the characteristics of the clients who subscribed to the term deposit may not be accurate simply due to the considerably larger proportion of clients who declined, which would influence the models. Furthermore, it is also found that there are many outliers within the dataset. This has caused the data to be significantly skewed and not well balanced, which in turn, affects the results of the analysis. The skewness may influence the individual characteristic to appear to have an impact on the target variable more than it actually does.

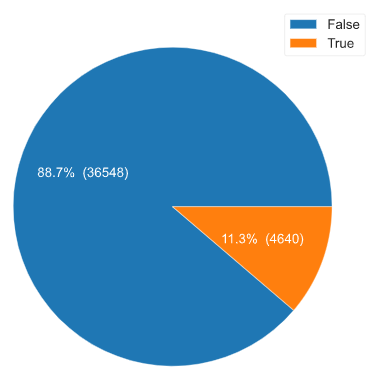


Figure 9: Pie chart for variable “y” showing data imbalance.

# 

# Conclusion

In conclusion, the factors that affect a clients’ decision to subscribe to the bank term deposit are number of employees, duration of last contact, and the month of the last contact with the client. However, none of the individual characteristics were able to be used to reliably predict the decision to subscribe. The results of this analysis can be used to segment the potential clients in marketing campaigns for bank term deposits. Companies such as banks, can use this information to target the correct group of customers in order to achieve higher success rates and avoid expending unnecessary resources.

Based on our results, it can be seen that the decision tree model that was created using 50% of the dataset as the training dataset was the best model among all the models. It has the highest accuracy of 91.50% and the lowest misclassification rate of 8.50% on the testing dataset. Looking at the ROC chart, the random forest models do not have the best performance. However, they were only outperformed by the kNN models in the training sets, and both the random forest models managed to outperform every other model in both validation and testing sets. Meanwhile, the decision tree created with 50% partition for training had the best F1-Score with the value of 0.6148.

There are several methods to further improve the analysis carried in this paper. Future researchers should first collect new data. As mentioned earlier in the paper, one of the limitations of this analysis is the timeliness of the data. Thus, collecting new data is crucial to accurately reflect the behaviour of the present society. Other than the quantity of data collected, its quality should also be taken into account. Future researchers should try to ensure that the data is well balanced to avoid the possibility of a biased result. Besides that, future researchers can also investigate and eliminate outliers from the dataset to reduce the skewness of the data. However, this may cause some data loss, thus precautionary measures should be taken. Following this, future research should experiment with feature selection. This is said as the decision tree model we created outperformed several other researches (Golecha, 2017; Muoki, 2020) through the use of SAS Enterprise Miner’s Chi-Square variable selection criterion.

# References

Boateng, N. (n.d.). *Predicting bank loan term deposit subscription*. <https://rstudio-pubs-static.s3.amazonaws.com/313414_c2eeb193f3a140f68c2a73a4cbc8ae91.html>

Golecha, Y. S. (2017). *Analyzing term deposits in banking sector by performing predictive analysis using multiple machine learning techniques*. School of Computing National College of Ireland.  
<http://norma.ncirl.ie/3100/1/yogeshsanjaygolecha.pdf>

Harrison, O. (2019, July 14). *Machine learning basics with the k-nearest neighbors algorithm*. <https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761>

John, L. (2020, August 31). *Bank institution term deposit predictive model*. <https://towardsdatascience.com/bank-institution-term-deposit-predictive-model-9f0b7c2fd38f>

Moro, S., Cortez, P., & Rita, P. (2014a). *A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62, 22–31.* [doi:10.1016/j.dss.2014.03.001](https://www.sciencedirect.com/science/article/abs/pii/S016792361400061X)

Moro, S., Cortez, P., & Rita, P. (2014b). *Using customer lifetime value and neural networks to improve the prediction of bank deposit subscription in telemarketing campaigns. Neural Computing and Applications, 26(1), 131–139.* [doi:10.1007/s00521-014-1703-0](https://link.springer.com/article/10.1007/s00521-014-1703-0)

Muoki, M. (2020). *Bank institution term deposit predictive model.*  
<https://medium.com/analytics-vidhya/bank-institution-term-deposit-predictive-model-14af2bbba70e>

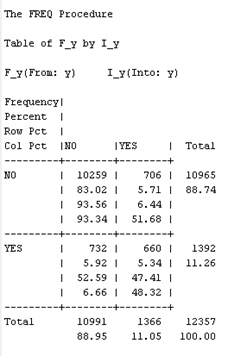
Odeyemi, G. (2020). *Bank institution term deposit predictive model.*  
<https://towardsdatascience.com/bank-institution-term-deposit-predictive-model-83afe1d2b08c>

Chatterjee, S., & Hadi, A. S. (2015). *Regression analysis by example*. <https://books.google.com.my/books?hl=en&lr=&id=zyjWBgAAQBAJ&oi=fnd&pg=PP1&dq=regression+analysis&ots=O_cCcN0MoX&sig=jvgjG5yHfmYgCpMeYsVeZfrx22E#v=onepage&q&f=false>

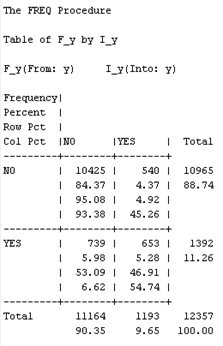
IBM Cloud Education. (2020). *Random forest*. <https://www.ibm.com/cloud/learn/random-forest>

Song, Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry, 27(2), 130-135. doi: 10.11919/j.issn.1002-0829.215044  
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4466856/>

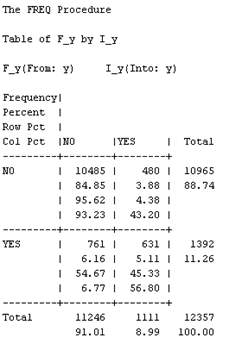
# Appendix



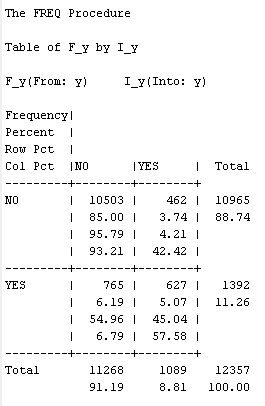
Appendix A: Confusion Matrix of 50 kNN, k=1



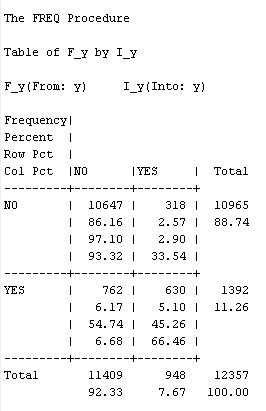
Appendix B: Confusion Matrix of 50 kNN, k=3



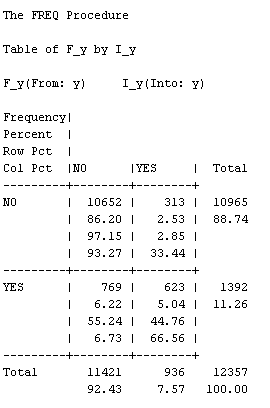
Appendix C: Confusion Matrix of 50 kNN, k=5



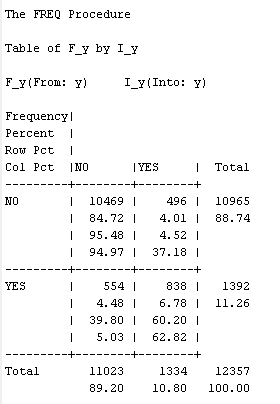
Appendix D: Confusion Matrix of 50 kNN, k=7



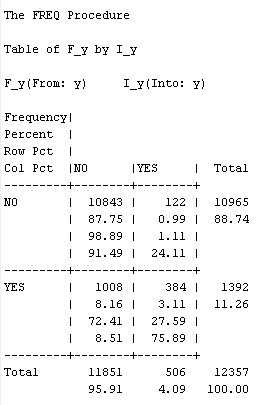
Appendix E: Confusion Matrix of 50 Regression



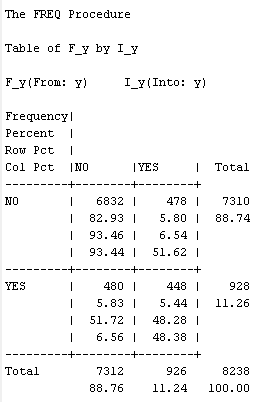
Appendix F: Confusion Matrix of 50 Stepwise Regression



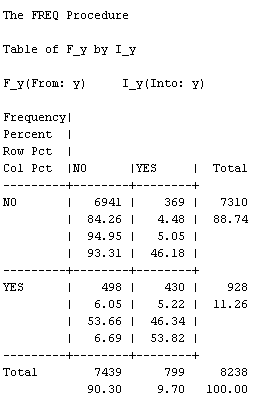
Appendix G: Confusion Matrix of 50 Decision Tree



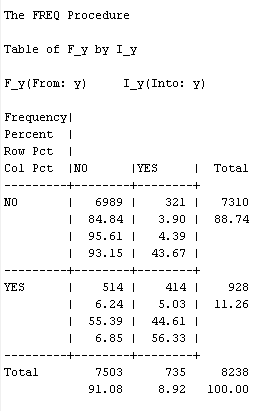
Appendix H: Confusion Matrix of 50 Random Forest



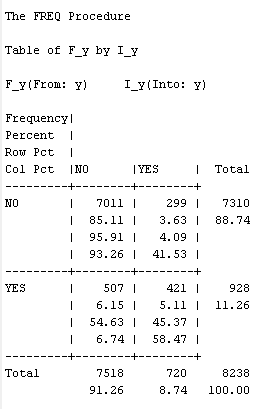
Appendix I: Confusion Matrix of 60 kNN, k=1



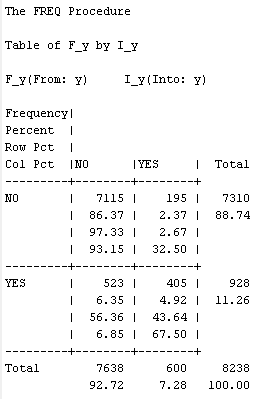
Appendix J: Confusion Matrix of 60 kNN, k=3



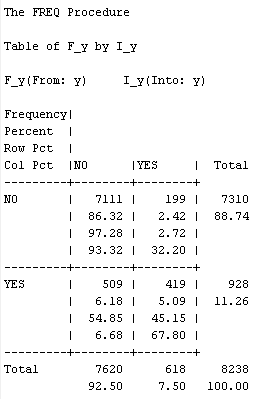
Appendix K: Confusion Matrix of 60 kNN, k = 5



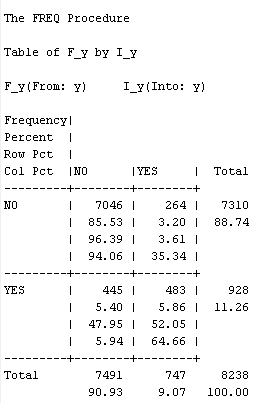
Appendix L: Confusion Matrix of 60 kNN, k-7



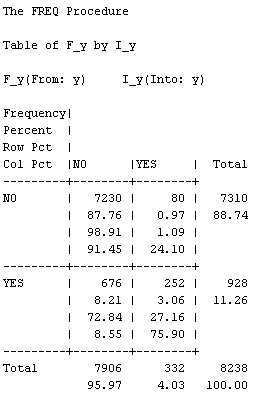
Appendix M: Confusion Matrix of 60 Stepwise Regression



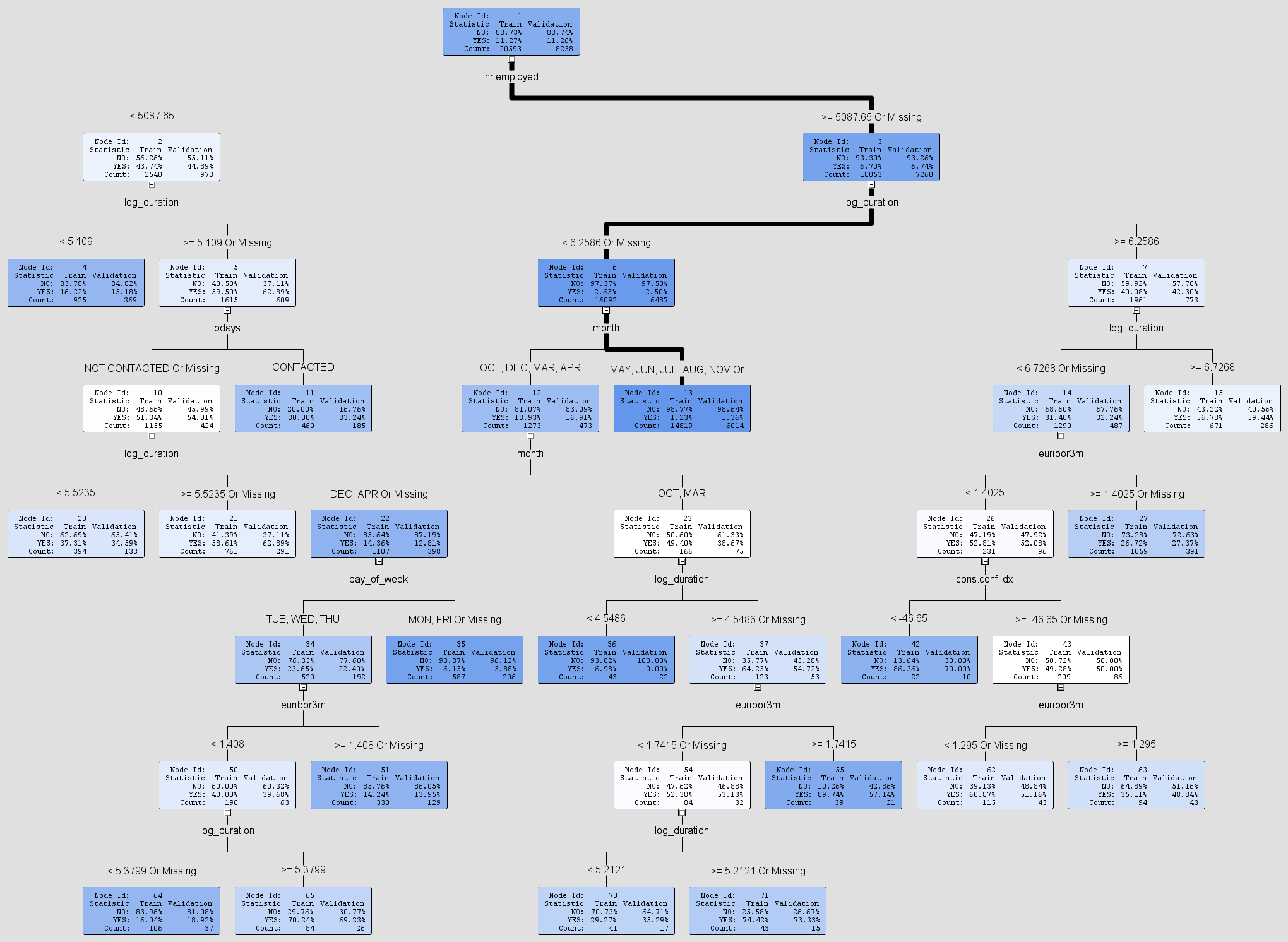
Appendix N: Confusion Matrix of 60 Regression



Appendix O: Confusion Matrix of 60 Decision Tree



Appendix P: Confusion Matrix of 60 Random Forest



Appendix Q: Decision Tree

Appendix R: Table of reference

| [Bank Institution Term Deposit Predictive Model | by Glory Odeyemi | Towards Data Science](https://towardsdatascience.com/bank-institution-term-deposit-predictive-model-83afe1d2b08c) | Python  EDA: No missing values, Highly imbalanced  target variable (lots of N few Y)  Preprocessing: Encoding categorical, Rescaling numerical, split dataset, PCA, SMOTE  Models: Log Regression, XGBoost, MLP  Cross Validate: K-Fold, Stratified K-Fold  Metrics: Accuracy, AUC, Precision, Recall, F1 Score (Weighted average of precision and recall)  Conc: XGBoost is the best model, Recommends: SVM, Random Forest, Decision Trees |
| --- | --- |
| [Bank Institution Term Deposit Predictive Model | by Lotome John | Towards Data Science](https://towardsdatascience.com/bank-institution-term-deposit-predictive-model-9f0b7c2fd38f) | Python  EDA: Imbalance data 88.7% decline 11.3% accepted.  Outliers: Use k-Nearest Neighbors Detectors  Preprocessing: One-hot encoding for categorical, StandardScaler to rescale numerical, Random oversampling for imbalance, PCA  Models: XGBoost Classifier, MLP, Log Regression, SVM  Cross Validate: k-fold, Stratified k-fold  Metrics: Accuracy, AUC, Precision, Recall,  Conc: XGBoost is best, stratified k-fold shows higher scores |
| [Predicting Bank loan Term Deposit Subscription (rstudio-pubs-static.s3.amazonaws.com)](https://rstudio-pubs-static.s3.amazonaws.com/313414_c2eeb193f3a140f68c2a73a4cbc8ae91.html) | R  Split: 70% train 30% test based on outcome  Feature Selection: RFE 16 features,takes top 5 features, providing 90% accuracy  Models: Boosted log regression, extreme gradient boosting, random forest, SVM with radial kernel, generalized least squares, bayesian regularized neural network  Metrics: ALIFT, sensitivity and specificity  Most accurate model: Neural net |
| [Sci-Hub | Using customer lifetime value and neural networks to improve the prediction of bank deposit subscription in telemarketing campaigns. Neural Computing and Applications, 26(1), 131–139 | 10.1007/s00521-014-1703-0](https://sci-hub.se/10.1007/s00521-014-1703-0) | Uses LTV tho  Split: 51651 old contacts and 1293 recent contacts.  Feature selection: Group into logical blocks, forward selection. Add few features from logical blocks into forward selection. Discarded if AUC & ALIFT no improvement.  Models: Logistic regression, Decision trees, support vector machines, neural network.  Metrics: AUC, ALIFT  Most accurate model: NN |
| [Sci-Hub | A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62, 22–31 | 10.1016/j.dss.2014.03.001](https://sci-hub.se/10.1016/j.dss.2014.03.001) | RMiner  Split: 51651 old contacts and 1293 recent contacts.  Selection: Semi-automatic (subjective + forward)  Models: LR, DT, NN, SVM  Metrics: AUC, ALIFT Best model: NN |
| [Bank Institution Term Deposit Predictive Model | by Mwikali Muoki | Analytics Vidhya | Medium](https://medium.com/analytics-vidhya/bank-institution-term-deposit-predictive-model-14af2bbba70e) | Selection: Auto encoder  Models: LR, XGBoost, MLP, SVM, DT, RF  Metrics: AUC, ALIFT  Best model: XGBoost |
| [Analyzing Term Deposits in Banking Sector by Performing Predictive Analysis Using Multiple Machine Learning Techniques (ncirl.ie)](http://norma.ncirl.ie/3100/1/yogeshsanjaygolecha.pdf) | Excel / RapidMiner / R  Method: CRISP-DM  Cross validation: K-fold cross validation Selection: Manually based on background understanding  Models: LR, Adaptive boosting, DT, SVM  Metric: Accuracy  Best model: Adaptive boosting 93.37% |

Appendix S: Proposal of the project

**Proposal - Dancing Queens**

**Domain Research:** Bank Marketing for Term DepositsIt was chosen because we were interested in whether the characteristics of an individual can help with predicting their decision to subscribe to bank term deposits through telemarketing.

**Problem Statement:**Everyone can subscribe to bank term deposits. However, being able to determine the best possible candidates to target for a specific marketing campaign can be difficult. Without properly identifying potential clients, a bank would have to exhaust a substantial amount of resources which includes both financial and labour in their marketing campaigns. This is because the bank would have to contact all eligible adults in the region to promote their term deposit plan. This is highly inefficient as it can cause a bank to expend unnecessary resources on unwilling customers. As such, it is important to be able to identify the characteristics of potential clients so that a more precise market can be targeted.

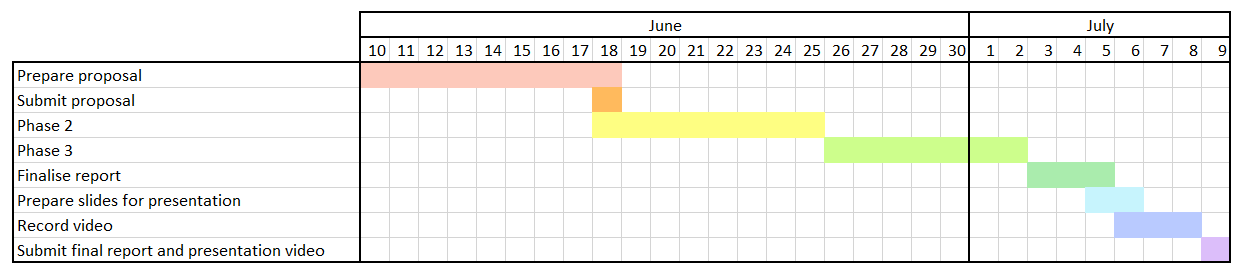
**Research Questions:**

1. What are the characteristics of clients that are subscribed to the bank term deposits?
2. Can certain characteristics exhibited by an individual successfully predict their decision to subscribe?

**Research Objectives:**

1. Determine the characteristics of clients with a higher likelihood to subscribe.
2. Classify whether a client would subscribe based on their characteristics.

**Research Plan:**

****

**Expected outcome:**

1. Successfully identify characteristics of a client that will influence their decision to subscribe to a bank term deposit.
2. A model that allows for the classification of clients into groups of will subscribe and will not subscribe is created.

**Dataset: Bank Marketing (**[**https://data.world/xprizeai-ai/bank-marketing**](https://data.world/xprizeai-ai/bank-marketing)**)**

**This dataset contains information of clients in a Portugese banking institution with over 41 thousand unique clients.**