National College of Ireland

Exploratory analysis of bank marketing campaign using machine learning; logistic regression, support vector machine and k-nearest neighbour.

MSc Research Project
Fintech

Jamiu Olalekan Oni
Student ID: x19111240

School of Computing
National College of Ireland

Supervisor: Victor Del Rosa

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Jamiu Olalekan Oni
Student ID:	X19111240
Programme:	Fintech
Year:	2020
Module:	MSc Research Project
Supervisor:	Victor Del Rosa
Submission Due Date:	17 august 2020
Project Title:	Exploratory analysis of bank marketing campaign using machine; logistic regression, support vector machine and k-nearest neighbour
Word Count:	6211
Page Count:	24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: .....

17 august 2020

Date: .....

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Exploratory Analysis of Bank Marketing Campaign Using Machine Learning: Logistic Regression, Support Vector Machine, and K-Nearest Neighbor.

Jamiu Olalekan Oni
X19111240

Abstract

Bank marketing campaigns can be described as a technique or procedure designed by financial bodies particularly the banks to help reach the targeted needs or specifications of customers. This campaign could be said to be carried out or launched in various ways either using the internet, rally, social media, leaflet, emails, short message services, digital signage, blogging, strategic partnership, and other mediums. The endpoint of the campaign be it in any form is to meet the targeted needs of the customers thereby satisfying the customers. In this research work, the resampling technique was used to deal with the imbalance dataset and three classifiers were applied; Logistic regression, Support vector machine, and K-nearest neighbor were used to achieve the set objective. Comparative analysis was performed using correlation heatmap to identify the main factors that can increase customer subscriptions to a term deposit. The outcome shows that 'Duration' is the main factor that can increase customer subscriptions in the bank. two experiments were performed in this study. Of all the algorithms used in this work, KNN has the best performance with the accuracy of 91.8% in the first experiment and 91.7% in the second experiment as compared to the Support vector machine and Logistic regression.

Keywords: Correlation heatmap, K-nearest neighbor (KNN), Support vector machine (SVM), Logistic regression, Marketing campaign, Term deposit.

1 Introduction

Deposits are the main source of revenue for banks. Many banks offer different types of accounts to attract customers willing to deposit their funds. The terms and conditions of depositing depend on the type of account. For instance, current accounts are held by customers willing to withdraw their funds at any time. On the other hand, fixed deposits are held by customers ready to lock their funds for a given period. The rate of interest is one of

the motivating factors which encourage individuals to open fixed deposit accounts. A bank can increase the number of subscribers to term deposits through effective marketing. Banks should have an effective marketing campaign strategy to reach their customers. Customer service is one of the marketing techniques that should be applied by banks. In this regard, the bank should ensure that the customers are treated fairly. The response team should assist customers within the shortest time possible. Video content campaigns are also used by various banks to attract customers. The primary objective of the video content campaigns is to ensure that the customers understand the products offered by the bank. If customers do not understand the terms under a fixed deposit account, they may not subscribe to it. Notably, customers are likely to subscribe to something that they know. The choice of a marketing strategy plays an important role in determining the level of subscription by banks. With the improvements in technology, banks can use Big Data to collect and analyze customer data. These data can be used to identify the likelihood of customers to subscribe to term deposits. If the bank realizes that many customers do not understand the terms and conditions of term deposits, the application of direct marketing strategies would be appropriate. The interaction between the bank officials and customers may increase the number of customers who are willing to subscribe to term deposits. Such communications improve customers' understanding of the bank's products.

Studies show that bank marketing campaigns focus on competitive strategies. Some of the strategies include demographic targeting, customer outreach, loyalty programs, and technology adoption. These strategies not only help the banks to reach many customers but to sell their products to the general public. By targeting a specific group of customers, banks can achieve their organizational objectives. One of the goals is an increase in the number of subscriptions to term deposits (Grzonka et al., 2016). The literature review tries to understand the findings of various researchers. The focus of the review is the factors that can increase customers' subscription to a term deposit.

1.1 Research question

What is the main marketing campaign factor that can increase the customer's decision to subscribe to a term deposit?

1.2 Research objective

To identify the main factor that can increase the customer's subscription to a term deposit.

2 Related Work

Asare-Frempong and Jayabalan (2017) reveal that direct marketing enables banks to focus on the customers who show the possibility of subscribing to their products and packages. In this case, the researchers intended to find out the determinants of customer response to a direct marketing campaign by applying various classifiers such as Logic Regression, Decision Tree, Random Forest, and Multiplayer Perception Neutral Network (MLPNN). The evaluation of the classifiers shows that there are specific elements that determine the likelihood of a customer to subscribe to the bank's products and packages. The study also revealed that there are various attributes of customers who are likely to respond to direct marketing campaigns. A cluster analysis revealed that customers with a better understanding of the products are expected to subscribe to the products. According to the study carried out by Asare-Frempong and Jayaban (2017), the Random Forest classifier is the most appropriate in terms of predictive ability. The research shows that the Random Forest represented 87% of the likelihood of customers to subscribe to the bank's products. However, the classification accuracy analysis determined that the possibility of the customers subscribe to a bank product depends on the strategies applied by the bank. In some cases, logic regression can represent the highest percentage of the likelihood of subscription. Various studies have shown that banks are likely to apply different marketing strategies and analytical approaches. These approaches play an important role in determining the rate of subscription. The study of Elsalamony (2014), analysis bank direct marketing with the use of data mining techniques. The study emphasized the fact that every marketing campaign is dependent on the customers' data. This simply implies that the information or customer data to be used in achieving every bank marketing campaign is large enough thereby making it impossible for analysts to reach good decision making. The study went on to introduce the use of a data mining technique to help in achieving its campaigns. Some of the techniques introduced were multilayer perception neural network (MLPNN), tree augmented Naïve Bayes (TAN), Nominal regression or logistic regression (LR), and Ross Quinlan's new decision tree model. Elsalamony (2014), the purpose of introducing these techniques was to check its performance on real-world data knowing that human data is large to handle and in the long run improve campaign effectiveness by highlighting its success characteristics.

Marinakos and Daskalaki (2017) found that sampling techniques play an important role in different algorithms and cluster-based methods of determining customer response. The researchers focused on a comparison of machine learning techniques and algorithms in

predicting customers' response to term deposits. Based on the results of the study, algorithms play an important role in determining the rate at which customers are likely to subscribe to the term deposits of the bank. However, the sampling technique determines the effectiveness of a mathematical or statistical approach. If the data collected for machine learning is more effective than that for an algorithm, the research may determine that machine learning is the best predictive approach. Therefore, banks should ensure that the sampling technique for the cluster-based procedure is established effectively. Information is the key determinant of the factors that can predict customers' response to term deposits. By using term deposits and the main point of consideration, Marinakos and Daskalaki (2017) revealed that direct marketing should be applied based on the available dataset. In most cases, banks tend to focus on customers' needs and preferences. Likewise, customers focus on the interest rate concerning term deposits. Therefore, banks should utilize distance-based and cluster-based sampling techniques to understand how customers respond to the products. Also, banks can use publicly available data for direct marketing. A study on the application of data mining in term deposit marketing by Zhuang, Yao, and Liu (2018) revealed that data mining techniques through SPSS Modeler predict customers' likelihood to subscribe to term deposits. The study acknowledged the fact that term deposits are experiencing challenges attributable to economic pressure and marketing competition. The results of the study showed that the application of the classification algorithm and clustering algorithm plays an important role in determining the rate of customer subscription to term deposits. The researchers recommended that banks should consider other underlying factors before implementing direct marketing techniques. Since the marketing technique determines the effectiveness of statistical and mathematical methods, banks should focus on the implementation of direct marketing (Zhuang et al., 2018). The researcher found that calling customers' phone numbers is an effective approach to direct marketing. A straightforward contract makes customers feel valued by the bank. Hence, the customers are more likely to subscribe to the term deposit.

Moro et.al (2011), also applied the use of CRISP-DM methodology in analyzing the bank marketing campaign. Studies have it that the CRISP-DM methodology is also a data mining technique used by the various analyst when dealing with real-life data that are seen and said to be large. Moro et.al (2011), indicated that the number of marketing campaigns carried out by banks is increasing daily and as such harming the public which includes the economic pressure that comes with it. Marketing managers according to the study tens to focus on target contacts which reduces their marketing performance. Due to this effect, the study introduces the use of Business intelligence (BI) and Data Mining (DA) techniques which are

simply the CRISP-DM. The use of CRISP-DM was applied to data collected from a Portuguese marketing campaign channeled towards deposits. Buhari and Elayidon (2015) applied the use of efficient CRM data mining in predicting customers' behavior towards a bank marketing campaign. The CRM was applied to determine the relationship that exists between customers and campaign carried by banks. The study made use of the classification models (Naïve Bayes and Neural Network) in specific to carry out its analysis. This study also applied the use of data collected from a Portuguese marketing campaign ranging from May 2008 to November 2010. Apampa (2016), Evaluates the techniques (Algorithm) used for Bank Customer Marketing Response Prediction. This study is seen to be similar and related to the study of Buhari and Elayidon (2015). The study attempts to improve the performance of classification algorithms used by other studies with the use of Random Forest (RF). Its aim is also to predict customer's behavior towards the bank marketing campaign. The study believed that RF stands as a better algorithm to be used for a bank marketing campaign as it stands to be better when dealing with real-life data. Sing'oei and Wang (2013), also emphasized on the framework for direct marketing by using a case study of bank marketing. The study identified the competitive market environment associated with a marketing campaign which has let to its increasing interest among academics and the public. The study further identified that its major increase was due to two (2) major reasons which are technology advancement and customer behavior. The study further noted that one of the challenges of this analysis was that customer's behavior is not easily predicted though some studies have tried predicting it using the various algorithm. For this purpose, the study initiated the use of a comprehensive framework guided by some systematic approach (mining approach) to promote the effectiveness of the marketing campaign and its research guide. Grzonka et.al (2016), weighed the performance and effectiveness of the bank marketing campaign with the use of selected specialized supervised classification. The study tries to solve the complex problems associated with a bank marketing campaign that could affect the larger economic society. The use of decision trees, bagging, boosting, and random forests were applied to assist decision making regarding the marketing campaign. The study of Nachev (2015), counters the previous study by trying to cover the gaps that the previous studies failed to underline with the use of a double testing procedure which combines cross-validation, multiple runs over a random selection of the folds and hyper-parameters, and multiple runs over a random selection of partitions. In general, Nachev (2015), tries to combine various techniques to access bank marketing campaign. Pan and Tang (2014) concentrated on the increasing cost of marketing campaigns particularly from the finance

sector and tries to devise ways of solving it. The study, therefore, decided to apply the use of the sophisticated technology to model behavior responses from the customers who the campaign is targeted at. The study compares bagging with boosting algorithms to measure the class imbalance behavior of the customers. Vajiramedhin and Suebsing (2014), also saw the need applied the use of data mining techniques to measure the performance of bank marketing campaigns with classification approaches.

According to the international journal of economics and financial issue (Parlar, 2017), the collection of customer information determines the effectiveness of the marketing campaign techniques. Concerning data mining techniques, customers' information is stored electronically and is used manually for analysis. The electronic data enables the banks to develop an effective direct marketing technique, which can attract many customers. By focusing on the algorithm, the research found that a supervised machine learning plays a vital role in determining the rate at which customers are likely to subscribe to a term deposit. The applicability of the machine learning technique depends on the availability of adequate data. Many banks keep electronic data about the customers; hence the application of supervised machine learning is appropriate. If the company does not have enough data about the customers, the machine learning approach would not be effective. The classification of information influences the results of the data mining techniques for detecting direct bank marketing. With the technological advancements in the banking sector, data can be stored in different patterns. Some of the most important information about customers is contact details. In terms of the application of direct bank marketing, sending direct messages or calling customers are appropriate techniques. The bank should have updated contact details about their customers. The address and location of customers should also be stored electronically (Parlar, 2017). The bank should note any change in the customer's contact details for a proper direct marketing strategy. The availability of the data will enable the bank to apply effective data mining techniques. The bank can use such methods to define and interpret the relevance of the bank campaign's effectiveness. According to a study by Migueis, Camanho, and Borges (2017), random forests influences customers' response to bank products. The study focused on target customers for banking campaigns. In this case, banks used target customer targeting strategies to determine the rate of response. However, the researchers found that imbalance is one of the problems experienced by banks when implementing a random forest approach. The risks of poor sampling techniques, banks can include demographic information, economic features, and contact details. Regardless of the sampling method that a bank uses, the availability of data influences the use of data mining techniques. By including

demographic information, banks will have adequate information about the customers. Such information can help in the random forest strategy. The study shows that random forest should be accompanied by an under-sampling algorithm to present a high prediction performance.

Class imbalance is a challenge to many banks with regards to data mining response. The account holders have varying attributes that determine their ability to respond to term deposits. The level of customers' income influences the likelihood of subscribing to bank products. Besides, a class imbalance is a great challenge in telemarketing; thus, it affects the authenticity of the data mining technique (Miguéis et al., 2017). The social class of the customers varies depending on the account holder. Some customers can manage multiple term accounts, while some may not manage one. The application of the data mining technique should include such factors. The inclusion of demographic information and social-economic features contributes to a discriminative performance. Telemarketing is one of the effective bank campaign strategies. A study carried out by Tekouabou, Cherif, and Silkan (2019) found that the effectiveness of five machine learning techniques depends on the application of effective bank telemarketing. The study focused on five machine learning techniques, including Logistic Regression, Support Vector Machines, Artificial Neutral Network, Decision Tree, and Naïve Bayes. By using accuracy and f-measure analysis, the study found that Artificial Neutral Network and Decision Tree techniques scored above 93% accuracy. The results showed that ANN and DT are important elements in determining customers' subscription to bank products. The application of machine learning techniques is influenced by the availability of the bank to interpret the results and availability of data. Tekouabou et al. (2019) developed a framework that represents a data modeling approach for classification challenges. Based on the structure, computing methodologies are the basis of a data mining strategy. Technology has enabled banks to collect adequate information from customers. The application of machine learning strategies depends on the availability of information and computing methodologies used by the bank. The framework also shows that the learning paradigm and supervised learning approached influence the application of direct marketing. A study on the application of selected supervised classification campaign by Daniel, Grayzna, and Barbara (2016) revealed that the complexity of the data used by banks could determine the application of data mining techniques. The form of direct marketing by banks influences the ability to get more information from the customers. The study found that more than 50% of the banks used direct marketing strategies for campaigns (Rubtcova and Pavenkov, 2019). These banks recorded an improvement in the use of the decision tree

approach in the determination of customer response. The likelihood of a customer to respond to term deposit depends on the sufficiency of the information provided by the bank through campaigns.

2.1 Limitations of the study

Lack of sample size for statistical measurement is one of the limitations of the study. Studies reveal that statistical measurement requires a large sample size (Uttley, 2019). In this case, the researcher would need to analyze many banks, which would be challenging. Following the restrictions on the banking industry, the researcher may not identify the most appropriate banks to study and the number of institutions to study. Lack of a large sample size is also problematic in terms of concluding. Research that focuses on a small sample size will not determine the actual phenomenon of the market. Therefore, it is important to focus on a large sample size.

Also, this research is subject to limited data. Concerning the restrictions in the banking industry, the research will have limited access to the data. Many banks would not allow researchers to get information on some essential aspects of the business. The nature of the banking business does not allow them to offer access to various information. Also, some banks are not willing to share information about their marketing strategies due to fear of competition. As a result, this research may not identify the actual campaign strategies used by the selected banks.

Time constraints are another limitation of the study. Like many organizations, banks work on a fixed schedule. The management of any bank ensures that time is utilized effectively. In this case, the researcher will write a letter to request for an appointment with various stakeholders. Due to the varying schedules of each bank, the research may not meet the guidelines. The limitation of time is also attributable to data collection and analysis.

3 Research Methodology

The methodology to be used for this research follows the cross-industry standard process for data mining (CRISP-DM).it provides a systematic approach to planning a data mining project. This approach is reliable and well-proven due to its step by step process and its general applicability (Gregory, 2018). This CRISP-DM includes five phases which are hierarchical and will be implemented during a data mining project (Rudiger and Jochen; 2000). These are shown in the diagram below;

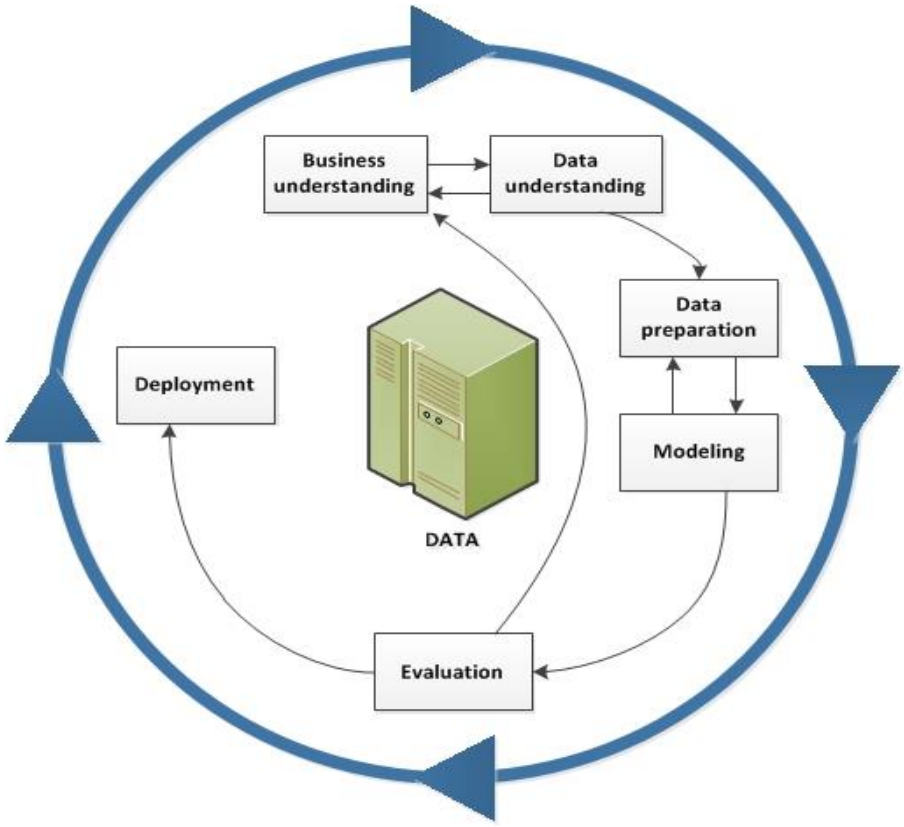


Figure. 1 CRISP-DM Approach

3.1 Business understanding

The first phase of the approach chosen for this data mining project is market segment awareness. For the main objective of this project to be achieved, an appropriate dataset needs to be put into consideration. The historical dataset used for this project was derived from the data.world. Containing 41188 rows with 21 columns.

3.2 Data preparation

The data for this analysis will be designed finally after having understood the data. The data is expected to be used for their potential modeling for this analysis. this analysis will start by gathering data from the data.world. The data will be obtained by downloading the bank marketing dataset which will be followed by the selected data exploration. The exploration will help to provide a clearer understanding of the data characteristics, size, and structure. Exploration will also assist in choosing the variables to be used while keeping in mind the study's question and purpose Better analysis of the data for this research will expose data quality problems and observations. It will eventually help in defining the type of data mining technique. It will ultimately help to decide what type of data mining technique to use for the research to better achieve a reasonable outcome.

3.3 Data modeling

The data will Present its independent variables and target variable after the data preparation has been completed to a certain level which will be further divided by a certain percentage for validation into test and training data. A certain percentage will be allocated to the training, while its percentage will be allocated to the test though smaller. The training is assigned higher so that the classifier can learn from the training larger part of the Dataset. This breaking will make it possible to apply the chosen modeling technique. This research will be limited to three chosen modeling techniques from other modeling approaches used for a data mining problem. Linear regression, support vector machine, k-nearest neighbor are the models to be used for this research work.

3.4 Model evaluation

These model(s) shall be evaluated for every analytical function. The models to be implemented for this analysis will be tested, as this will shape as necessary an important part of the research process. The models that will be used will be reviewed to determine their accuracy, and based on their accuracy, this will be done using the confusion matrix.

3.5 Deployment

This is the last stage of the study. The information gained from the results obtained will be reported and provided for use after thorough modeling with its assessment. This can also imply that the models used will be contrasted and therefore the one considered to have performed better will be suggested to the finance sector, especially the financial service providers to serve as means of improving their services. This will allow them to use this research to identify the main problems and focus on ways to develop innovative approaches to fix the existing problems, thereby providing customers with a better financial service that could reduce grievances and disputes.

4 Design Specification

The research would include a qualitative comparison of three machine learning models which will be considered as the strategies to be used in this research for the proper implementation of the models outlined for this report. The machine learning models can also be referred to as machine learning classifiers or classification approaches in the study for greater understanding and clarification since the study deals with a classification problem. It could also be said that the research work deals with supervised learning as the dataset to be used as a label for each example and that label is of a categorical type. The algorithms are;

4.1 Logistic regression

This algorithm is frequently used for analytical function. Empirically, this method increases the probability of logging to allow the data required to perform well (Scott et al 2006). It also indicates that it functions much like the linear regression only because it deals with variables that are categorical like Yes or No, 1 and 0, and others as can be seen in this analysis.

4.2 K-nearest neighbor

KNN can be used for statistical problems both in classification and regression. Nevertheless, it's used more broadly in business classification issues. This is also applied for solving problems of classification as well as regression. KNN is simpler to use, has quick execution time as the data quality determines the accuracy of the model, and the k value (nearest neighbor) must be adequate.

4.3 Support vector machine

A support vector machine is a supervised learning model with related learning algorithms analyzing the data used for classification and regression analysis. The Support Vector Machine (SVM) algorithm is a common machine learning tool that offers solutions for problems of classification and regression. In addition to performing linear classification, SVMs can perform a non-linear classification effectively using what is called the kernel trick, mapping their inputs into high-dimensional feature spaces implicitly.

5 Implementation

The figure below shows the mechanism owing by which the execution of this research is directed. the python programming language has been used to perform all the experiments.

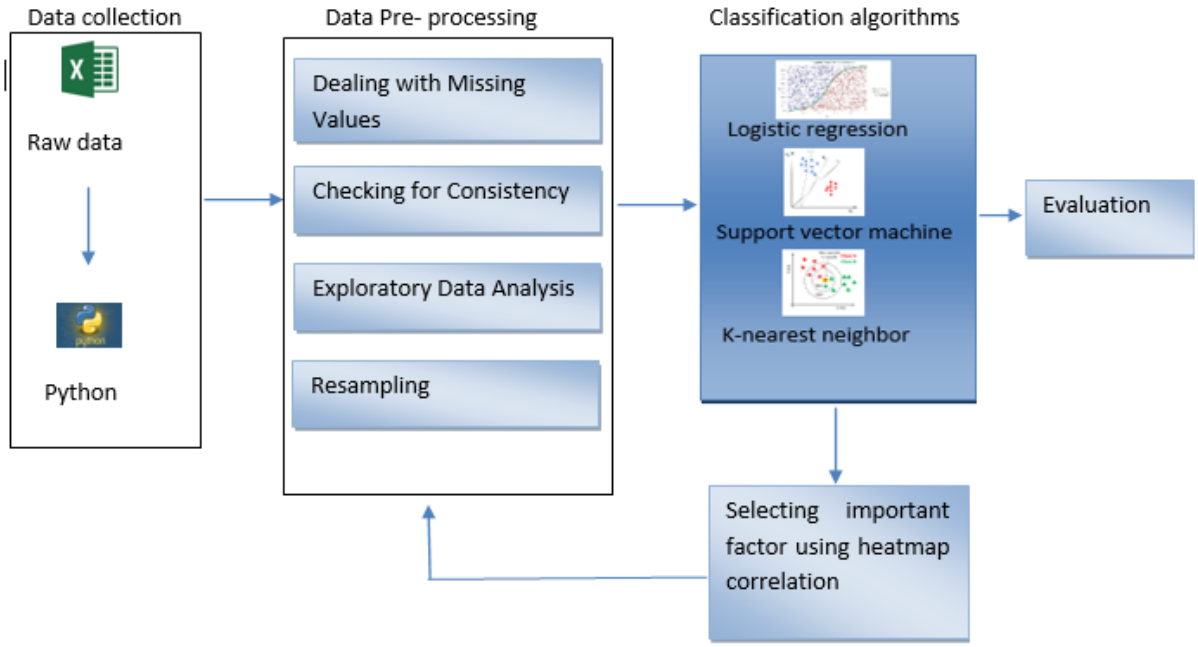


Figure. 2 flow diagram

The flow diagram above shows how the experiment on the dataset downloaded from the data world will be carried out using the python code.

5.1 Data collection

For this research work, the bank marketing dataset was downloaded from the Data world[1].it comprises 41188 rows and 21 columns.

5.2 Data preprocessing

After the data was extracted to achieve the specified goals, the dataset was read into a jupyter notebook using python code because it is flexible in handling large datasets. Anaconda which was also the open-source distribution of python was also used. A python feature was being called to replace strings that have space with an underscore, the isnull().sum() function was utilized to search for missing values in the imported data. The result shows that the data has no missing values, the target variable (y) was renamed to 'signed' for better understanding. with the use of boxplot, outliers were detected from the dataset and this will be properly addressed in the modeling phase. Also, the dataset used for this research work is consistent. The target variable(signed) is a binomial classification problem that was categorized into yes or no from the original dataset. As the chosen algorithms work better with numerical values, the yes was encoded with 1 and the no was encoded with 0, and pd.factorized function was

[1] Dataset, https://data.world/data-society/bank-marketing-data

used to convert categorical to numerical variables .after carrying out all these functions, resampling technique was carried out due to fact that the dataset is imbalanced.

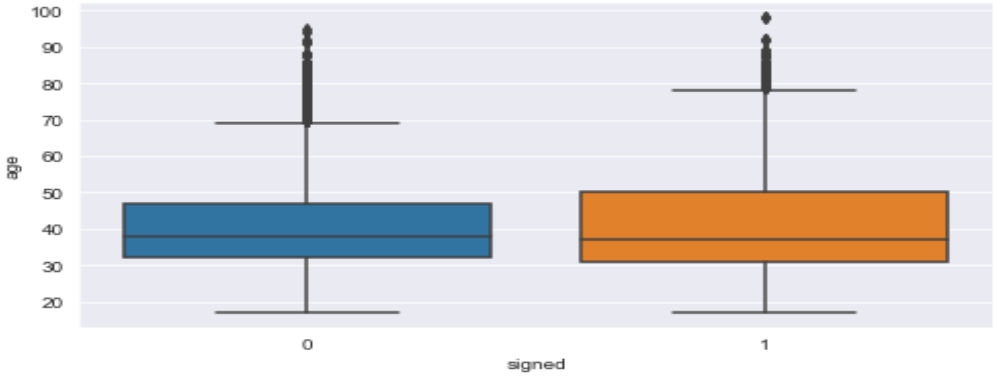


Figure .3 Boxplot showing outliers

```
#Checking for null values
BankAnal.isnull().sum()
```

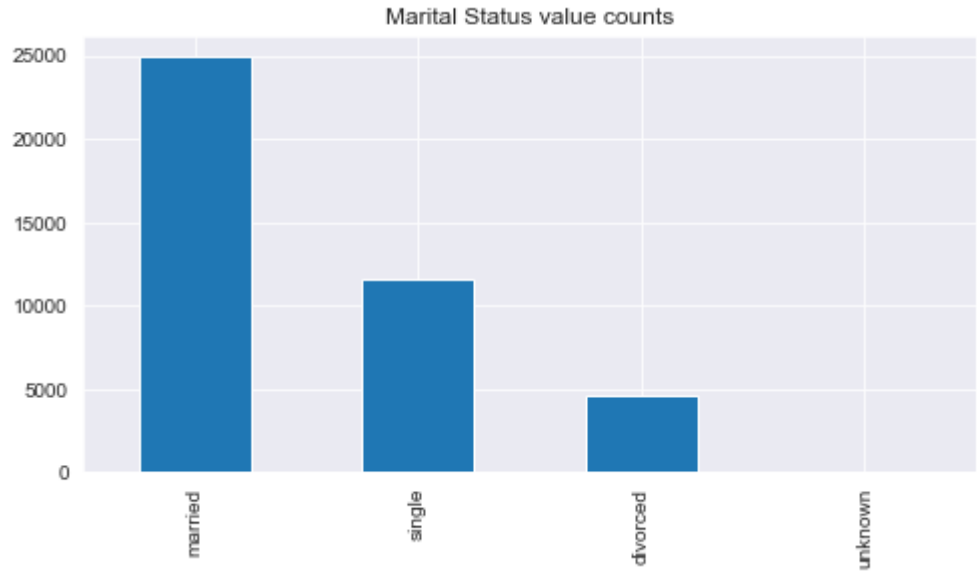
Figure. 4 Checking for missing values

```
from sklearn.utils import resample
df_0 =BankAnal[BankAnal.signed ==0]
df_1 =BankAnal[BankAnal.signed ==1]
df_Target=resample(df_1, replace = True, n_samples= 36548, random_state=0)
df_Target=pd.concat([df_Target,df_0])
df_Target.signed.value_counts()
```

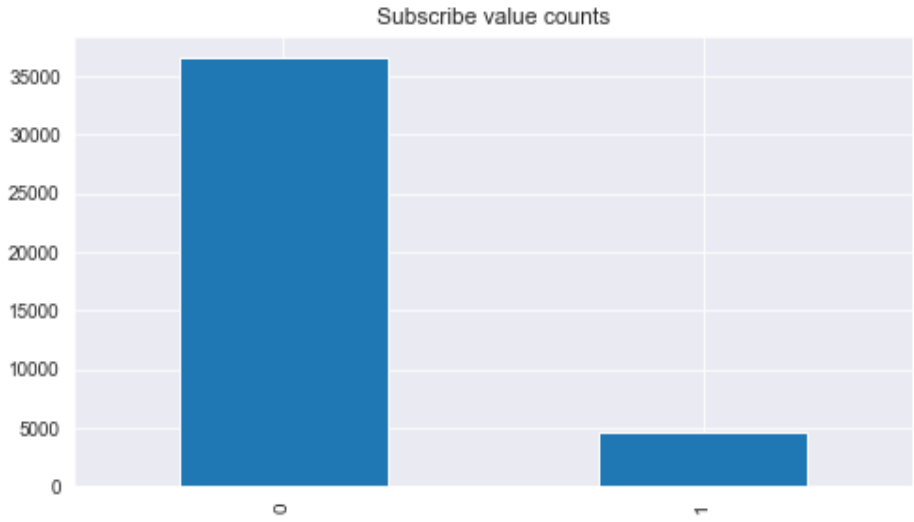
Figure .5 Resampling of dataset

5.3 Exploratory analysis

From the bank marketing dataset, the bar chart was plotted against the marital status and the result depicts that the highest number of customers came from the married people and took exactly 60.5% of the entire population. while the singles took just 28% of the population. The divorce was only 11.19%. and 0.19% was unknown.



From the figure below, the output indicates that over 35000 customers refuse to subscribe to the bank. On the other hand, the rate of customers subscription is less than 5000.



5.4 Feature engineering

From the exploratory analysis carried out, the target variable (signed) was an imbalance, and the resampling technique which is a function in python was applied to the balance of the dataset for better analysis. Before the balancing of the data, 89% of the customers did not subscribe to a term deposit of the bank while the remaining 11% of customers manage to subscribe.

5.5 Modeling

In other to avoid overfitting, the model_selection function was imported for splitting of the data into train and testing using ratio 70:30 before training the model on the training set and testing on the test set. sklearn.linear_model was used in training the linear regression

algorithm to classify the number of subscribers. Support vector machine and k nearest neighbor was also trained to classify the number of subscribers.

For the research question and objective to be met, the second experiment was carried out by bringing out the important factors that contributed to the customer's decision to subscribe to a term deposit. five (5) important factors identified with correlation heatmap and were used to retrain the three algorithms as used for the first experiment. The six identified important factors are contact, month, duration, previous, Euro interbank offer rate 3 months (euribor3m), and the number of employees (nr.employed). the new important factor was then stored in a new data frame.

```
  #Standardizing the data for them to be on the same scale

from sklearn.preprocessing import StandardScaler
sc_X = StandardScaler()
X_train_scaled = sc_X.fit_transform(X_train)
X_test_scaled= sc_X.transform(X_test)
```

Figure. 6 Standardization of the data

```
 #Splitting the data into training and test set
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(df_Target.drop(['signed'] ,axis=1),
                                df_Target['signed'], test_size=0.3, random_state=0)
```

Figure. 7 Splitting of data into train and testing

6 Evaluation

Following the implementation of the data mining methodology used for this analysis, as stated earlier, three algorithms are extensively used to help recognize patterns from the collected data. These methods used in this study utilize multiple forms to deliver performance or reports. The results obtained from the use of the algorithms will then be presented and analyzed at this point in the analysis but not clearly explained as this will be discussed in the following paragraphs. Classification models are described in a variety of ways. This implies that it is possible to use multiple matrices to test findings resulting from the use of models of classification. For this research, the findings for this analysis are calculated based on the confusion matrix for calculating the errors that were misclassified for each algorithm or method from which each value was measured for accuracy. In this analysis, accuracy is used because its outcome is on average (equal value) with the appropriate quality. It is pertinent to realize that this model was originally divided into training and testing. Therefore, the analysis and outcomes of the three models will be listed below.

6.1 Logistic regression

A distinction will be made between the testing and training data for this method, keeping in mind that forecasting was made with the test dataset. The confusion matrix for this algorithm's test dataset is shown in table 1 below.

Prediction using the test dataset:

N = 21929	Predicted: 0	Predicted: 1
Actual 0:	TN= 9465	FP= 1619
Actual 1:	FN= 1425	TP= 9420

Table 1 Confusion matrix for the test dataset

From the table 1 above, there are two possible predicted classes: '0'(no) and '1'(yes).as the number of subscribers is predicted, 0 means customers will not subscribe and 1 means customers will subscribe. The above table shows customer's controversial responses using a logistic regression algorithm for the test dataset. According to the model, the accuracy is 86.12% whereas the misclassification rate is 13.88%. from the total observation of 21,929, the model predicted 9,465 No and 9,420 as Yes. also, it incorrectly predicted 1,425 negatives as Yes and 1,619 positives as No.

6.2 Support vector machine

The demand was reached for this model, too. The call provides a summary that shows how many variations there are. The confusion matrix for this algorithm's training data set is shown in table 2 below.

Prediction using test dataset:

N = 21929	Predicted: 0	Predicted: 1
Actual 0:	TN= 9137	FP= 1947
Actual 1:	FN= 1205	TP= 9640

Table. 2 Confusion matrices for the test dataset

Table 2 above shows customer's controversial responses using Support vector machine algorithm for the test dataset. Based on the model, the accuracy is 85.63% whereas the misclassification rate is 14.37%. from the total observation of 21,929, the model predicted 9,137 No and 9,640 as Yes.

6.3 K-nearest neighbor

A distinction was made between the testing and training data for this method, keeping in mind that forecasting was made with the test dataset.

Prediction for test dataset:

N = 21929	Predicted: 0	Predicted: 1
Actual 0:	TN= 9439	FP= 1645
Actual 1:	FN= 170	TP= 10675

Table .3 Confusion matrix for the test dataset

Using the K-nearest neighbor algorithm for the test dataset, shows the customers' controversial responses to subscribe to a term deposit. The accuracy of this model is 91.7% which happens to be the highest of the models while the misclassification rate is 8.28%

Experiment 1: Using all factors gotten from the original dataset

Models	Accuracy	Precision	Recall	AUC
Logistic Regression	0.861	0.853	0.869	0.861
K Nearest Neighbor	0.918	0.869	0.982	0.918
Support Vector Machine	0.868	0.844	0.898	0.868

Table 4 result of the models from the original dataset

Experiment 2: Using the six important factors extracted from the original dataset

Models	Accuracy	Precision	Recall	AUC
Logistic Regression	0.848	0.841	0.854	0.848
K Nearest Neighbor	0.917	0.866	0.984	0.918
Support Vector Machine	0.856	0.832	0.889	0.857

Table 5 result of the models from the 6 important variables extracted from the original dataset

6.4 Discussion

This thesis started as a guideline for researching by setting out a research question and its specified objectives. On this basis, a methodology of data mining was chosen and implemented with the analysis of three different algorithms to facilitate the specified objectives, address the question, and accomplish a meaningful analysis. This research adopted three classifiers approach to accommodate its specified goals based on performance conducted. on this basis, different outcomes or findings were generated using the three chosen algorithms adopted for this research: Logistic Regression, Support vector machine, and K nearest neighbor. the result generated from the analysis is to help the banks in identifying the main factor that can increase customer subscriptions to a term deposit. based on the result from the models carried out, KNN has the highest accuracy rate of 91.7% in the test data. This indicates that KNN will be a good model for predicting if customers will subscribe to a term deposit or not. however, the other two models also perform well but KNN performs exceptionally as the best model. Besides, two experiments were carried out in this study, the first experiment was done using all the variables present in the dataset after pre-processing, while the second experiment was performed using 6 important variables to improve the output of the model using lesser variables.

In experiment one, K nearest neighbor had 91.8% for both Accuracy and AUC. Looking at the model that correctly classifies the number of subscribers i.e. Recall, KNN had 98.2%, SVM and Logistic regression had 89.8% and 86.9% respectively.

The second experiment was done to achieve the research objectives and research question as highlighted in sections 1.1 and 1.2 by identifying the main factors that can influence the customer's decision in signing up for a term deposit in the bank. The results have gotten shows that K nearest neighbor has the highest Accuracy with 91.7% which is the same as that of test data and AUC score of (91.8%) which has the same values as that of experiment one. Looking at the model that correctly classifies the number of subscribers i.e. Recall, KNN had 98.4%, SVM and Logistic regression had reduced scores of 88.9% and 85.4% respectively. for Precision, KNN had a reduced score of 86.6%, Logistic regression, and SVM had 84.1% and 83.2% respectively as shown in table 4 and 5 above.

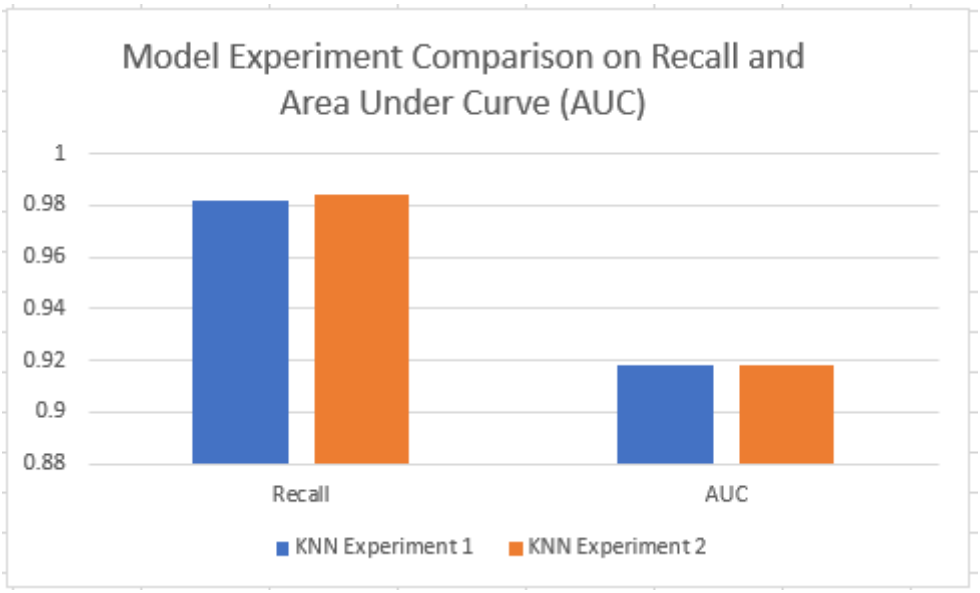


Figure 8: Model Comparison Based on Recall and AUC

The figure above shows that KNN had the same AUC score of 91.8% for both experiment one and experiment two, with an improved Recall score of 98.4% in experiment two compared to experiment one with a score of 98.2%, which implies that the fewer factors obtained from the second experiment will help the bank to identify the main factor that can increase the number of customers that are likely to subscribe to a term deposit. The financial institution and even the fintech community, whose goal is to provide customers with better financial services, can leverage this report. It can also help them make better service strategic decisions they prevent potential issues with their services. This research can also be used to explore the probability that respective customers will refute their reactions based on the existing attributes that support their predicted issue.

7 Conclusion and future work

Many banks use direct marketing strategies to enable customers to access adequate information about the products. Researchers suggest that the applicability of data mining techniques depends on the availability of customers' information. Also, studies reveal that machine learning techniques determine customer response to bank products. The ability of the customers to subscribe to term deposits depends on the marketing campaign by the bank. In this research work, the resampling method was used in dealing with the problem of imbalanced data, and three machine learning algorithms (Logistic regression, KNN, and SVM) were deployed to find out the main factor that influences customers decision to subscribe to a term deposit in the bank. Correlation heatmap was then used to get the most important factors that influence customer decisions and was then retrained to perform the

second experiment. Both experiments had good accuracies ranging from 84% to 92% with KNN performing better with an improved recall score of 98.4% and AUC score of 91.8% compared to other algorithms used for this research. The correlation heatmap highlighted five factors that can influencing the customer's decision and ‘duration’ has the highest correlation coefficient with dependent variable (positive correlation) of 0.46. Which means that the longer the bank continue to advertise their product and service, the more customers can subscribe to a term deposit. Banks should focus on direct marketing techniques when applying statistical and mathematical approaches to determine customer response. This project can further be enhanced by using other techniques like univariate selection and feature importance on the dataset to identify more factors that can influence customer's decision to subscribe to a term deposit in the bank.

Acknowledgment

I will like to give thanks to God Almighty for the success of this research work. Also, I would like to thank my supervisor Victor Del Rosa for the guidance and encouragement throughout the research process. A big thanks to my family and friends who have contributed immensely to this great success.

References

Apampa, O. (2016) ‘Evaluation of classification and ensemble algorithms for bank customer marketing response prediction’, Journal of International Technology and Information Management, 25(4), p.6.

Asare-Frempong, J. and Jayabalan, M. (2017) ‘Predicting customer response to bank direct telemarketing campaign in 2017 International Conference on Engineering Technology and Technopreneurship’, (ICE2T). IEEE, pp. 1–4.

Bahari, T.F. and Elayidom, M.S. (2015) ‘An efficient CRM-data mining framework for the prediction of customer behavior’, Procedia computer science, 46, pp.725-731.

Elsalamony, H.A. (2014) ‘Bank direct marketing analysis of data mining techniques’, International Journal of Computer Applications, 85(7), pp.12-22.

Gregory, P. (2014) ‘CRISP-DM, still the top methodology for analytics, data mining, or data science projects’, [Online] Available: https://www.kdnuggets.com/2014/10/crisp-dm-topmethodology-analytics-data-mining-data-scienceprojects.html. Accessed on: Aug. 10 2018.

Grzonka, D., Suchacka, G. and Borowik, B. (2016) ‘Application of selected supervised classification methods to bank marketing campaign. Information Systems in Management’, 5(1), pp.36-48.

Marinakos, G. and Daskalaki, S. (2017) ‘Imbalanced customer classification for direct bank marketing’, J. Mark. Anal. 5, 14–30.

Miguéis, V.L., Camanho, A.S. and Borges, J. (2017) ‘Predicting direct marketing response in banking: comparison of class imbalance methods’, Serv. Bus. 11, 831–849.

Moro, S., Laureano, R. and Cortez, P. (2011) ‘Using data mining for bank direct marketing: An application of the crisp-dm methodology’, In Proceedings of European Simulation and Modelling Conference-ESM'2011 (pp. 117-121). EUROSIS-ETI.

Nachev, A. (2015) ‘Application of data mining techniques for direct marketing’, Computational Models for Business and Engineering Domains, pp.86-95.

Pan, Y. and Tang, Z. (2014) ‘June Ensemble methods in bank direct marketing’, In 2014 11th International Conference on Service Systems and Service Management (ICSSSM) (pp. 1-5). IEEE.

Parlar, T. (2017) ‘Using data mining techniques for detecting the important features of the bank direct marketing data’, Int. J. Econ. Finance. Issues 7, 692.

Rubtcova, M. and Pavenkov, O. (2019) ‘Features of Integrated Marketing Communications of the Russian Bank Sphere’, in RF-360th International Conference on Management, Economics & Social Science-ICMESS.

Rudiger, W. and Jochen, H. (2000) 'CRISP-DM: Towards a standard process model for data mining’, in 2000 Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining (ICPAKDDM). , Berlin, Germany,18-20 june, pp. 29-39., IEEE Xplore. doi: 10.1024/ICPAKDDM.2000.7942220.

Scott, N. A., Sunil, G., Wagner, K., Junxiang, L. and Mason, C. H. (2006) ‘Defection detection: Measuring and understanding the predictive accuracy of customer churn models, Journal of Marketing Research’, 43(7): (204)211.

Sing'oei, L. and Wang, J. (2013) ‘Data mining framework for direct marketing: A case study of bank marketing. International Journal of Computer Science Issues (IJCSI)’, 10(2 Part 2), p.198.

Tekouabou, S.C.K., Cherif, W., Silken, H. (2019) 'A data modeling approach for classification problems: application to bank telemarketing prediction, in Proceedings of the 2nd International Conference on Networking, Information Systems & Security', Pp. 1–7.

Uttley, J., 2019. Power analysis, sample size, and assessment of statistical assumptions—

Improving the evidential value of lighting research. Leukos 15, 143–162.

Vajiramedhin, C. and Suebsing, A. (2014) 'Feature selection with data balancing for prediction of bank telemarketing. Applied Mathematical Sciences', 8(114), pp.5667-5672.

Zhuang, Q.R., Yao, Y.W., Liu, O. (2018) 'Application of data mining in term deposit marketing, in Proceedings of the International MultiConference of Engineers and Computer Scientists'.