Application of Data Mining techniques to analyze the success of bank telemarketing

Rijo Kuruvilla (14171011)

1. Abstract

The evolution of modern technology, awareness among consumers, and the rise in competition in financial technology are the few factors that have resulted in banks emphasizing heavily in marketing strategies. Term deposits are a crucial aspect of long-term savings and aid people to invest for a period of time and receive a fixed interest rate. The customer's money is locked away for the time the customer specifies (the term), typically between one month and five years [1]. A minimum amount is required to open a term deposit, and the minimum amount varies from bank to bank. The purpose of this report is to use data mining techniques like supervised classification to help banks gain insights on how to improve their customer reach to sell long-term deposits, as well as to use unsupervised clustering to identify potential natural clusters. The data used for the analysis were collected from Portuguese banking institutions and consisted of 17 features relating to the customers, their social and economic details, and the bank's records of contacting them. Logistic regression and Naive Byes classifier were the supervised classification techniques used to predict the success of the bank selling long-term deposits whereas Agglomerative Hierarchical Clustering (AHC) was the unsupervised technique utilized to cross-verify if the natural clusters align with the actual clusters. In predicting bank performance, Logistic regression coupled with the stepwise method generated the best accuracy results, while AHC with Ward's distance best separated the cluster resulting into two categories, thereby supporting the outcome of the dependent variable (whether a customer subscribes to a term deposit ("yes") or not ("no"). The analysis concludes that the data at hand can be used to determine the bank's success or failure in determining whether a customer subscribes to a term deposit. The upper leadership and bank managers can use the aforementioned models to improve their marketing strategies.

2. Introduction

The practice of attracting and acquiring new customers through traditional and digital media strategies is known as bank marketing [2]. The use of these media strategies aids in determining what type of customer is drawn to specific banking institutions. This also includes different banking institutions purposefully employing various strategies to attract the type of customer they wish to do business with. Bank marketing strategies have evolved. Due to the rapid advancement of technology and consumer demands, banks have become more agile in their efforts to provide excellent customer satisfaction. Customers nowadays prefer to rely on customer reviews and a personal touch from sellers when purchasing a product, rather than advertisements, which are slowly becoming an outdated form of product promotion.

Telemarketing is a subtype of bank marketing and is a common strategy used by businesses to sell products to current or potential customers, gather information from customer reviews, and maintain a positive relationship with customers via cellular phones or telephones. Telemarketing can be done from a call center, an office, or, increasingly, from the comfort of one's own home. Telemarketing can frequently involve a single call to assess interest or suitability, followed by follow-up calls to pursue a sale. Various data can be used to reduce large databases of names to a small number of high-probability customer prospects [3]. Although time-consuming, it helps in maintaining direct contact with the customers. This report focuses on using the data collected by a Portuguese bank using telemarketing and applying data mining techniques to predict whether the customer would subscribe to the bank's

long-term deposit. Unsupervised data mining is used to double-check the results of supervised techniques if or not the derived results align with the actual results. The accuracy derived from the model will aid the bank's marketing team to assess its telemarketing business model and boost its customer outreach.

3. Data

The dataset 'bank.csv' consists of 4521 observations with 17 features that were chosen at random from an older version of the dataset (bank-full.csv). Paulo Cortez (University of Minho) and Sérgio Moro (ISCTE-IUL) created this dataset in 2012. The dataset was made public for research in 2014 and can be found in the UCI Machine Learning Repository [4].
This data was obtained as part of an observational study and is related to a Portuguese banking institution's direct marketing campaign. The marketing campaigns relied on phone calls, and clients were frequently contacted more than once to determine whether a term deposit(response variable) subscription was successful. The dataset was evaluated for missing data and consisted of no missing values post-evaluation. The response variable (term deposit) is unbalanced, as only 11.5% of the observations are related to success as shown in FIGURE 1.

Out of seventeen variables, the ten "character" data type variables (job, marital, education, default, housing, loan, contact, month, outcome, and y(term deposit) were converted from character data type to factor data type after analyzing the original structure of the data before proceeding with the analysis. The variable 'pdays' was removed from the analysis because it contained many '-1' values (82.12%), possibly indicating that the customers had not been contacted in a long time. This can stymie the analysis because it is treated as missing data. As a result, this column has been removed. The dataset's final sample size is 4521 observations and 16 variables. The type and description of each variable are outlined in TABLE 1. There are no known interventions or pre-processing techniques that were applied to the data.

Index	Variable Name	Variable Type	Variable Levels	Description
1	Age	Numeric	-	Age of the targeted customers
2	Job	Categorical	"admin"," unknown"," unemployed", "management"," housemaid"," entrepreneur", "student"," blue-collar"," self-employed", "retired"," technician", "services"	Type of jobs held by the customers
3	Marital	Categorical	"married", "divorced", "single" note: "divorced" means divorced or widowed	Marital status

4	Education	Categorical	"unknown", "secondary", "primary", "tertiary"	Education level
5	Default	Categorical – Binary	“yes”, “no”	Credit card default status
6	Balance	Numeric	-	Average yearly balance (euros)
7	Housing	Categorical – Binary	“yes” , “no”	Has a housing loan?
8	Loan	Categorical – Binary	“yes” , “no”	Has a personal loan?
9	Contact	Categorical	"unknown", "telephone", "cellular"	Contact communication type
10	Day	Numeric	-	Last contact day of the month
11	Month:	Categorical	"jan", "feb", "mar", ..., "nov", "dec"	Last contact month of year
12	Duration	Numeric	-	Last contact duration, in seconds
13	Campaign	Numeric	-	Number of contacts performed during this campaign and for this client
14	previous	Numeric	-	Number of contacts performed before this campaign and for this client
15	poutcome	Categorical	"unknown", "other", "failure", "success"	The outcome of the previous marketing campaign
16	y(term_depoist)	Categorical – Binary	“yes”, “no”	Has the client subscribed a term deposit?

TABLE 1: Data Summary Table

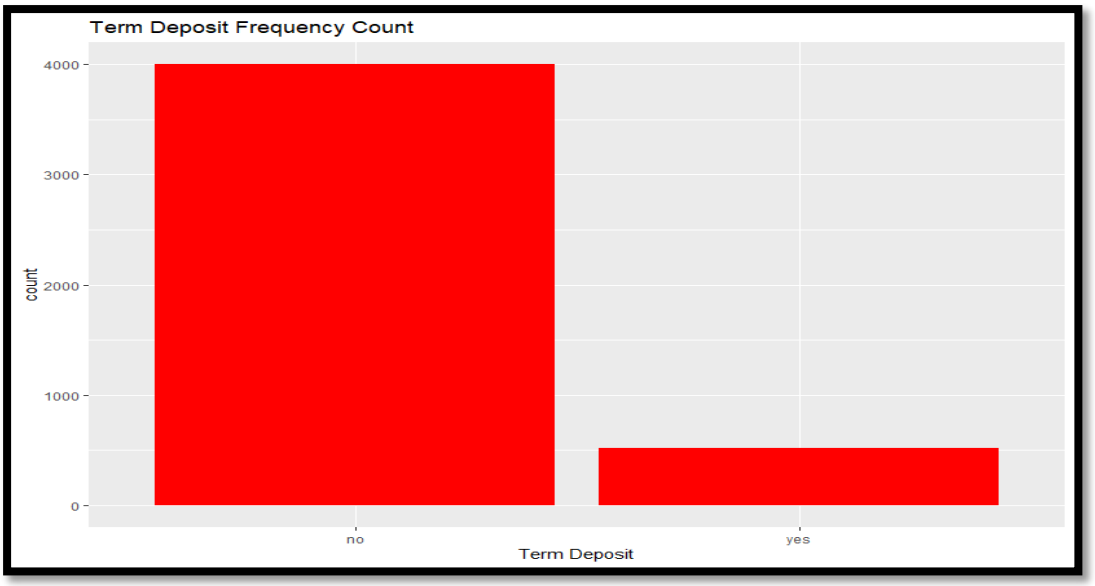

FIGURE 1: Frequency Count of the Response variable

4. Methods

The analysis was carried out with R programming in RStudio studio – version '2022.02.3+492' (release name – Prairie Trillium).

The data mining techniques used to process and analyze the data were Naive Bayes Classifier, Logistic Regression, and Hierarchical Clustering. The dataset was divided into training data (80%) and test data (20%) before applying the Naive Bayes and Logistic Regression classifiers. The data was split into 80% (training data) - 20% (test data) with a seed value of 1. Although the seed does not play a significant role in the model selection process, it ensures result reproducibility, which is critical in drawing conclusions and interpreting results [5]. The training data is used for feature analysis and model construction, while the test data is used for prediction and determining the model's accuracy.

i. Logistic Regression –

Logistic regression is one of the most common classification techniques where the dependent variable is binary (i.e., there are two classes only), whereas the predictor variables can be both numeric as well categorical. Logistic regression estimates the probability of an event occurring based on the independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1 [6]. The coefficient in this model is commonly estimated using maximum likelihood estimation (MLE). The log-likelihood function is generated by testing different coefficient values iteratively, and logistic regression seeks to maximize this function to determine the best parameter estimate. After determining the optimal coefficient(s), the conditional probabilities for each observation can be computed to produce a predicted probability.

Logistic regression was first applied to the training data to build the model. The model summary was used to extract significant features from the original set of features to make data predictions. Stepwise logistic regression with the "backward" method was used for feature selection. The "backward" stepwise method starts with a model with all features. It then iteratively begins eliminating the least significant variables one by one, until a predefined stopping rule is reached, or no variables remain to be examined

[13]. The assumptions for logistic regression were analyzed before proceeding with the analysis.

Validating the assumptions for Logistic Regression

a. The response variable is binary.

Logistic regression assumes that the response variable only takes two outcomes. The structure of the response variable (term deposit) confirmed that it is a binary categorical variable ("yes" or "no").

b. There is little or no multicollinearity between the variables.

When two or more predictor variables are highly correlated with one another, the regression model does not provide unique or independent information. The variance inflation factor (VIF) was used to detect multicollinearity, which measures the correlation and strength of correlation between predictor variables in a regression model. A VIF greater than 10 indicates high collinearity, whereas a VIF between 5 and 10 indicates the need for further investigation. All predictor variables had a VIF of less than 10, implying that there is little or no multi-collinearity between them.

c. The observations should be independent.

This assumption can be validated by looking at how the data is collected. Although each client was contacted multiple times in case they did not respond at first, all clients, if connected within the first try had their data stored by the bank. Only those customers who did not respond to the bank's first contact attempt were subjected to repeated calls in an attempt to establish contact. As a result, we can safely conclude that there were no repeated measurements resulting in this assumption being met.

d. Prefers large sample size

Logistic regression analysis yields reliable, robust, and valid results when the sample size of the dataset is large enough. As a rule of thumb, you should have a minimum of 10 cases with the least frequent outcome for each explanatory variable. This assumption was met since the sample size of this data is 4521 observations and 16 variables. To determine the significant features, the p-values of each variable generated from the logistic regression summary were examined. As significant features, all variables with p-values less than the assumed significance level of 0.05 were chosen.

The Akaike Information Criterion (AIC) was also used to see if the backward stepwise logistic regression produced the correct feature extraction. AIC scores with lower values are better, and AIC penalizes models with more parameters. So, if two models explain the same amount of variation, the one with fewer parameters has a lower AIC score and is a better fit [7]. The model summary demonstrated that there was a decrease in AIC values from 1807 to 1796.8, demonstrating that the stepwise regression model was better and correctly extracted the essential features. Forward logistic regression was used to build the model and make predictions using the significant features obtained through the backward stepwise method. Section 5 summarises the logistic regression model's results.

ii. Naive Bayes Classifier –

The Naive Bayes classifier was modelled using the features extracted from stepwise logistic regression using the backward method. A naive Bayes classifier is a type of simple "probabilistic classifier" that uses Bayes' theorem with strong (naive) independence assumptions between the features. Naive Bayes is a straightforward technique for building classifiers: models that assign class labels to problem instances represented as vectors of feature values, with the class labels drawn from a finite set [8]. The Naive Bayes classifier can work with both categorical and mixed (numerical and categorical) predictors, and it uses a frequentist approach to estimate the required probabilities from the data.

As Naive Bayes assumes that all predictors are statistically independent within each class, a Chi-squared test of independence was used to determine the association between the categorical variables. The null and the alternate hypothesis were stated as follows:

H0: Significant relationship does not exist between the predictors and response
Ha: Significant relationship exists between the predictors and response

The test yielded a p-value of 0.0004998 for all categorical predictors, which is less than the assumed level of significance (0.05), indicating that the alternate hypothesis (Ha) is correct, and a significant relationship exists between the predictors and the response variable. Even though the classifier's assumption was violated, the naive Bayes classifier usually performs well even when the independence assumption is violated. The reason for this is that the classifier does not require exact probability values, which may be inaccurate due to a violation of the independence assumption, but such inaccuracies are harmless if the largest probability can still be determined correctly. For numerical variables, the kernel density function was used to handle the assumption that the variables follow a normal distribution and to estimate their class-conditional distribution. The Naive Bayes summary defines two classes, "yes" and "no," with seven features each. The prior probabilities of no(0.8847) and yes(0.1153) justify the data imbalance mentioned in section 3.

FIGURE 2 displays the visual summary of the Naive Bayes model. A customer is more likely to sign up for a term deposit if he is married, has a stable income, and has no outstanding loans. A customer with a cell phone is more likely to sign up for a term deposit than one with a telephone. The highest number of subscriptions was recorded in the months of May and August, while December had the lowest number of subscriptions which could be attributed to the Christmas and New Year holidays. The variable duration shows a poor separation of classes indicating poor subscription results. The model performance and accuracy are discussed in section 5.

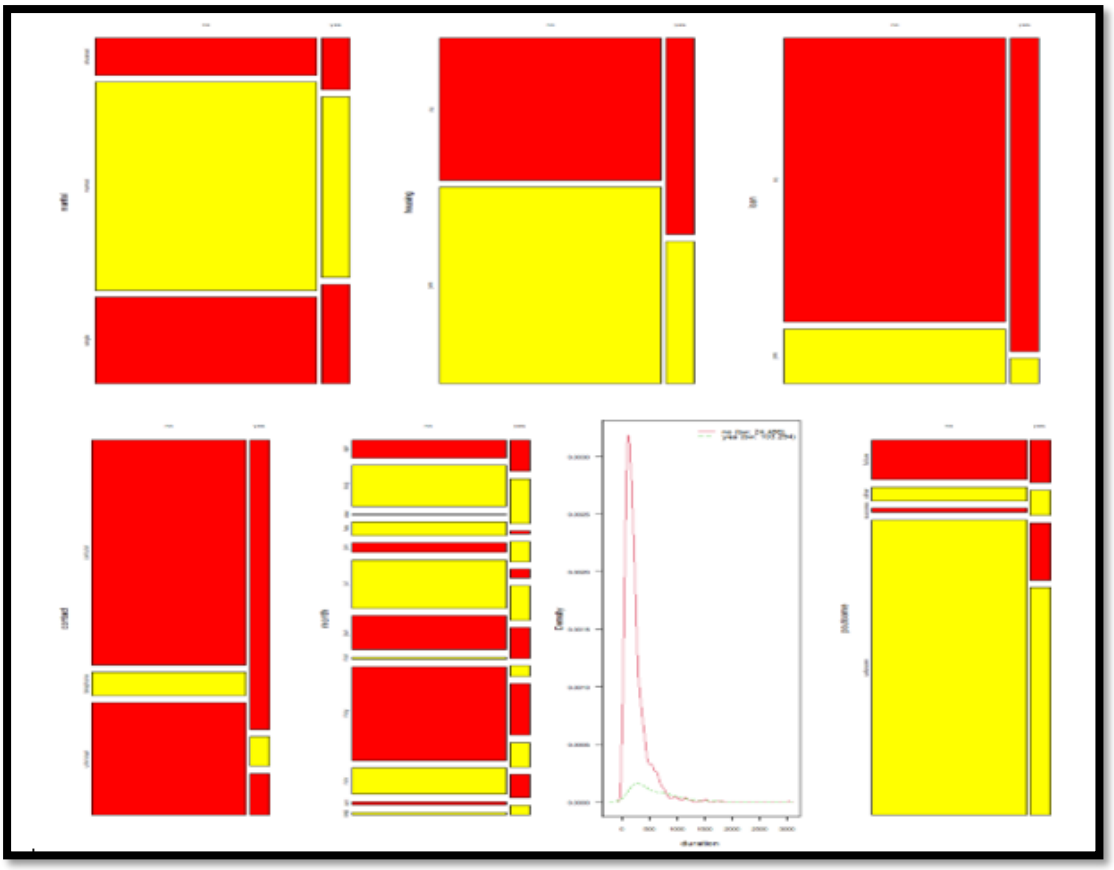

FIGURE 2 - Naive Bayes Summary

iii. Hierarchical Clustering –

As the dataset consists of a mixture of categorical and numerical data, hierarchical clustering is the technique used to validate the best separation of classes. The Agglomerative Clustering algorithm begins by considering each object to be a singleton cluster. Following that, pairs of clusters are merged one by one until all clusters have been merged into one large cluster containing all objects. The result of the AHC is a visually appealing tree-based representation of the objects, called the dendrogram which makes the clustering model performance easy to interpret and analyze.

Agglomerative clustering works from the bottom up. In other words, each object begins as a single-element cluster (leaf). The algorithm is executed step by step.
Two of the most similar clusters are combined to form a larger cluster (nodes)[10]. This process is repeated until all of the points are part of a single large cluster (root).
Both K-means clustering and DBSCAN clustering are not suitable for this dataset as K-means and DBSCAN both require numerical predictors and cannot be performed with mixed variables.

Since clustering is an unsupervised learning technique, the response variable was first removed from the data before commencing the analysis. The numerical variables were scaled (standardized) so that they all had the same scale. Scaling prevents larger-scale variables from dictating how clusters are defined [9]. It enables the algorithm to consider all variables equally important. The Gowers distance is used to compute the distance matrix for data with a mixed set of variables. For each variable,

a similarity measure is computed within the range [0,1]. The overall similarity is calculated by taking the average of these values.

After computing the Gowers distance, the following types of hierarchical clustering were then implemented on the data:

a. single-linkage(SL) – uses the smallest pairwise distance as the measure of dissimilarity between the clusters.
b. complete linkage(CL) - uses the largest pairwise distance as the measure of dissimilarity between the clusters.
c. average-linkage(AL) - uses the average pairwise distance between as the measure of dissimilarity between the clusters.
d. Wards distance – It minimizes the total within-cluster variance. At each step, the pair of clusters with minimum between-cluster distance are merged.

The clustering results are discussed in section 5.

5. Results and Discussion

Logistic Regression and Naive Bayes classifiers both produce high-accuracy results. As shown in TABLE 2**,** Logistic Regression had the highest accuracy score of 89.16%, while Nave Bayes had the lowest at 84.29%. Both classifiers were replicated ten times with a for loop, and their accuracy means, and standard deviation means were calculated across ten test sets. The mean accuracy scores for Logistic Regression and Naive Bayes classifiers were 90.15% and 89.98%, with mean standard deviations of 0. 0.00845 and 0.00704, respectively, indicating low variance in the data. TABLE 3 displays the confusion matrix, which aids in determining the model's accuracy results. The rows represent what the algorithm predicted, while the columns represent the actual results.

Logistic Regression correctly classifies 782 out of 800 customers who did not ("no") take subscriptions while incorrectly classifying 80 out of 104 "yes" subscriptions. The Naive Bayes model correctly categorizes 756 out of 800 "no" subscriptions while incorrectly categorizing 98 out of 104 "yes" subscriptions.

Classifier	**Accuracy Score (%)**	**Mean Accuracy Score (%) – 10 iterations**	**Mean Standard Deviation – 10 iterations**
Logistic Regression	89.16	90.15	0.00845
Naive Bayes	84.29	89.98	0.00704

TABLE 2: Accuracy summary for the classifiers

Logistic Regression	Prediction Logistic Regression	no	yes
	no	782	80
	yes	18	24
Naive Bayes	Prediction Naive Bayes	no	yes
	no	756	98
	yes	44	6

TABLE 3: Confusion Matrix for the classifiers

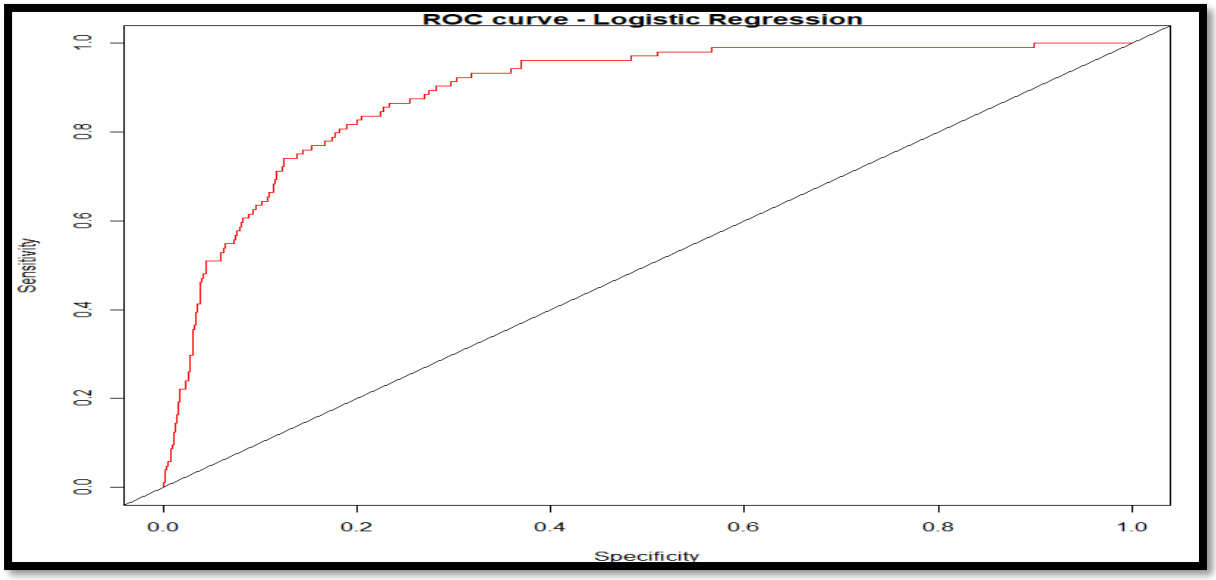

FIGURE 3: AUC-ROC – Logistic Regression.

The Area under the Receiver Operating Characteristic Curve (AUC-ROC) was also plotted to evaluate the classifier's performance, as shown in FIGURE 3, which displays the accuracy results of Logistic Regression. The ROC curve is formed by changing the classification probability and plotting the True Positive rate vs the False Positive Rate. The red diagonal line corresponds to a pure random allocation of class labels. Values close to 0.5 indicate performance comparable to random, while values less than 0.5 indicate performance worse than random. AUC values closer to one indicate that the classifier performs better. AUC-ROC also works best for imbalanced data and hence is ideal for this dataset. AUC-ROC for the logistic regression is 89.43% which indicates a good model performance in determining the response variable.

Clustering was used to cross-check the results of the naive Bayes and logistic regression classifier and to check if there are more natural clusters as compared to the original two classes. FIGURE 4 depicts the dendrograms for all types of linkages used to determine clustering results. Dendrograms are excellent visualizing structures, and the key to understanding them is to concentrate on the height at which any two objects are joined together. Each dendrogram leaf represents an observation. As we climb the tree, the observations of each mother merge into branches. As we climb the tree, the branches join together with leaves or other branches. This fusion's vertical axis height indicates how different the two observations are. Thus, observations fusing near the bottom of the tree are very similar, whereas observations fusing near the top of the tree are very different.

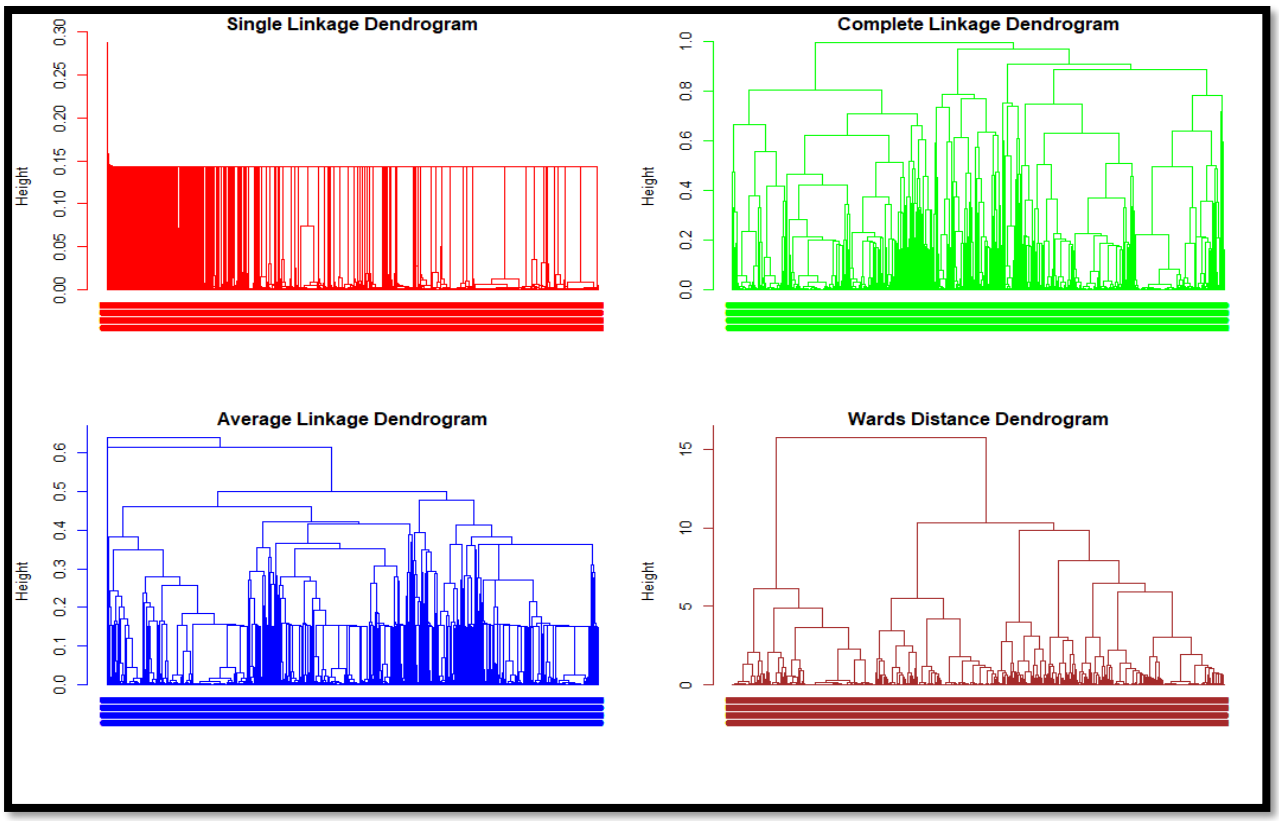

FIGURE 4: AHC Dendrograms for all linkages

When compared to other distances, the single linkage hierarchical clustering displayed the worst result with almost no separation between the classes. Both the average and complete dendrograms have poor linkages at the bottom but good separation at the top. However, the Wards distance, which reduces overall within-cluster variance and merges two clusters with the shortest between-cluster distance, provides the best class separation to determine "yes" and "no" term deposit subscriptions. The dendrogram shows good separation from the bottom to the top, and the hull plot in FIGURE 5 supports the dendrogram's results.

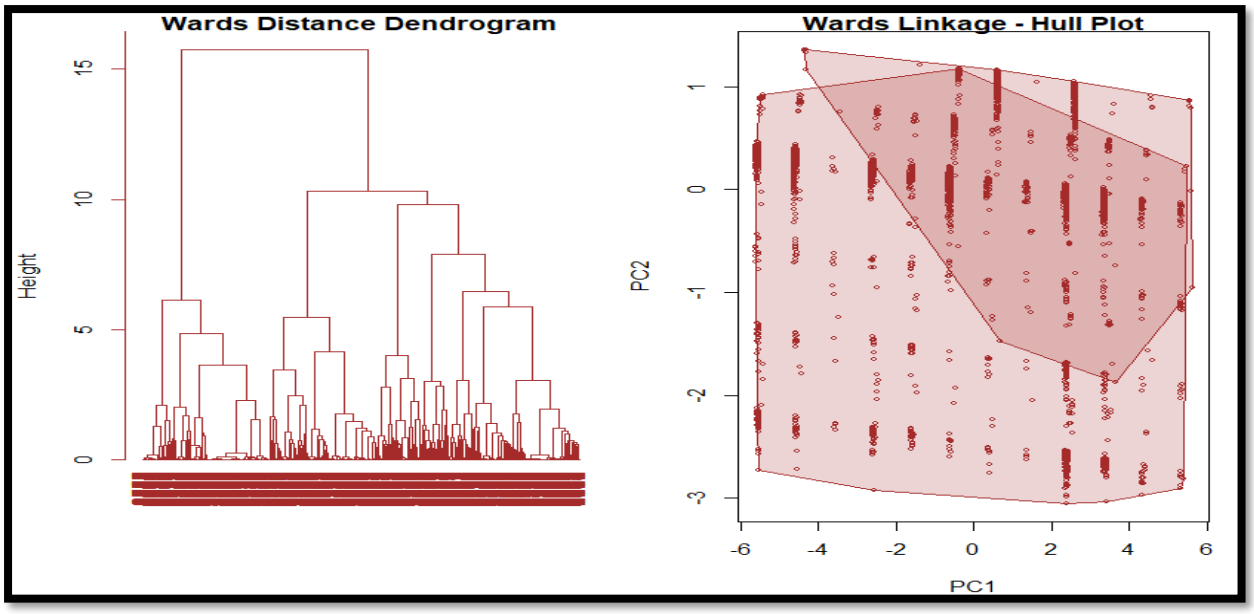

FIGURE 5: Dendrogram and Hull Plot for Wards distance

6. Conclusions

Direct and healthy customer relationships can be maintained using efficient telemarketing strategies. Such strategies offer customers attractive product promotions provided by banking institutions. The data was collected by Portuguese banking institutions, and cellular phones and telephones were the primary mediums through which the clients were contacted and details regarding their personal as well as socio-economic status were collected. Predictive data mining techniques were deployed to determine whether or not a long-term deposit subscription was taken by a customer. Clustering, an unsupervised learning technique, was used to determine if more insights could be determined concerning long-term deposits. Having more long-term deposits would massively boost the capital requirements of banking institutions.

The response variable is highly imbalanced and shows that more customers did not subscribe to a long-term deposit plan. Logistic Regression and Naive Bayes classification models were deployed, and predictions based on those models determined that Logistic Regression generated the best accuracy (90.15%) and AUC score (87.85%) in predicting the customer response. AHC using the Wards method was the unsupervised technique used to validate the natural cluster separation with the actual cluster separation. AHC with the Wards distance provided the best class separation clearing distinguishing between the “yes” or “no” long-term deposit subscriptions. Utilizing the results of these modelling techniques, the bank’s marketing team can improve marketing efficiency by reducing contact costs and employing selective customer focus. More data can be collected from customers in the future (number of family members, previous long-term deposit subscriptions held, etc.) to enrich the dataset, and advanced machine-learning techniques can be utilized to improve model performance.

References

[1] *What is a term deposit? (n.d.). Www.westpac.com.au.* https://www.westpac.com.au/personal-banking/bank-accounts/term-deposit/what-is-a-term-deposit/

[2] admin. (2019, July 12). Modern Bank Marketing - A Comprehensive Guide (2021). Mediaboom. https://mediaboom.com/news/bank-marketing/

[3] Kenton, W. (2020, June 21). *Telemarketing*. Investopedia. https://www.investopedia.com/terms/t/telemarketing.asp

[4] *UCI Machine Learning Repository: Bank Marketing Data Set*. (2012). Uci.edu. https://archive.ics.uci.edu/ml/datasets/bank+marketing

[5] *Random seed*. (2021, October 16). Wikipedia. https://en.wikipedia.org/wiki/Random_seed

[6] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R* (2nd ed., pp. 133–147) [Review of *An Introduction to Statistical Learning with Applications in R*]. Springer.

[7] Bevans, R. (2020, March 26). *Akaike Information Criterion | When & How to Use It*. Scribbr. https://www.scribbr.com/statistics/akaike-information-criterion/

[8] Wikipedia Contributors. (2019, June 17). *Naive Bayes classifier*. Wikipedia; Wikimedia Foundation. https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[9] Standardization in Cluster Analysis. (2018, September 24). Alteryx Community. https://community.alteryx.com/t5/Alteryx-Designer-Knowledge-Base/Standardization-in-Cluster-Analysis/ta-p/302296

[10] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R* (2nd ed., pp. 521–531) [Review of *An Introduction to Statistical Learning with Applications in R*]. Springer.

[11] *Interpreting ROC Curves, Precision-Recall Curves, and AUCs*. (2018, December 8). Www.datascienceblog.net. https://www.datascienceblog.net/post/machine-learning/interpreting-roc-curves-auc/

[12] [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

[13] *Understand Forward and Backward Stepwise Regression – Quantifying Health*. (n.d.). https://quantifyinghealth.com/stepwise-selection/