Analysis of Portuguese Banking Institution's Direct Marketing Campaigns:

Predictive Modeling and Strategy Enhancement

CIND 820 - Big Data Analytics Project – D1H

Supervisor: Ceni Babaoglu, Ph.D.

Student: Isolda Veruska de Almirante Silva
Student ID: 501241365

Ryerson University

Table of Contents

Abstract

This project centers on analyzing a dataset from a Portuguese banking institution's direct marketing campaigns available in the UC Irvine website, primarily conducted via phone calls. The dataset can be found at the following link: http://archive.ics.uci.edu/dataset/222/bank+marketing and comprises two subsets: "bank-full.csv," including examples from May 2008 to November 2010, and "bank.csv," a 10% random sample. The goal is binary classification-predicting if clients will subscribe ('yes') or not ('no') to a term deposit.

In the realm of financial marketing, understanding customer behavior and campaign effectiveness is paramount. This dataset's unique aspect, involving repeated phone contacts and client engagement, makes it an intriguing subject. The main goal is to identify factors influencing term deposit subscriptions and construct predictive models for future marketing strategies.

The aim is to answer three fundamental questions: What drives clients to subscribe to a term deposit in this campaign, can we build accurate predictive models from the data, and how we can improve marketing campaigns?

The dataset contains 45,211 records in "bank-full.csv" and 4,521 in "bank.csv." It features 16 input attributes encompassing client data and one output attribute ('y') indicating term deposit subscription. Attributes range from numeric variables like age, balance, and contact duration to categorical descriptors like job type, marital status, education, and communication method. Binary indicators for credit default, housing loans, personal

loans, and campaign history, alongside numeric features detailing contact frequency and timing, are also included in the dataset.

The approach involves thorough data preprocessing, addressing missing values, encoding categorical data, and scaling numeric features. Various classification models can be explored for predicting term deposit subscriptions, including logistic regression, decision trees and random forests. Model performance is assessed using accuracy, precision, recall, and the F1-score. Python serves as the primary programming language for data manipulation, analysis, and visualization.

In conclusion, this project aims to unveil the key determinants of term deposit subscriptions in a banking marketing campaign. By comparing different modeling techniques, it seeks to provide actionable insights for refining future marketing strategies and enhancing campaign success

Keywords: direct marketing campaigns; banking; binary indicators; classification; predictive models; Python.

1. INTRODUCTION

Telemarketing, using phone calls to promote products and services, grew significantly in the 20th century. It serves various functions, including customer support and technical assistance, making it valuable for nurturing customer relationships. However, telemarketing often faces customer resistance. To address this, companies should carefully select and manage their telemarketing services.

In the banking sector, marketing plays a crucial role in exchanging financial products and services. Banks have extensive customer data. Effective marketing relies on understanding market dynamics and customer needs. Data analysis enhances telemarketing efficiency, allowing banks to craft personalized campaigns based on customer preferences and behaviors, strengthening customer relationships.

Data mining uses statistical, mathematical, AI, and machine learning techniques to extract valuable information from datasets. Classification, including methods like Naïve Bayes, Logistic Regression, and various neural networks, is a popular data mining task.

This researcher is trying to answer questions like what motivates clients to subscribe to a term deposit, whether accurate predictive models can be built from data, and how marketing campaigns can be improved.

The source code of this project is available in: https://github.com/isoldalmirante/Final-Project-Bank-Data-Analysis

2. LITERATURE REVIEW

Multiple publications were examined with a focus on use of data mining to analyze bank direct marketing.

Moro, Cortez, and Rita propose using data mining to predict telemarketing campaign success. A data-driven Decision Support System (DSS) was introduced, using data mining on a dataset spanning from 2008 to 2013. The study involved feature engineering and a comparison of four data mining models, with neural networks yielding the best results. The DSS significantly enhanced campaign efficiency, potentially achieving 79% of successful sales by contacting only half of the clients, and it created value for bank telemarketing managers in terms of campaign efficiency (Moro et al., 2014).

The article authored by Fereshteh Safarkhani and Fatemeh Safara compares five mining models, and Logistic Regression emerged as the standout performer with an ROC score of 0.93, indicating its superior ability to distinguish between positive and negative classes. With an accuracy rate of 91.21%, Logistic Regression appeared to be the most effective choice for bank telemarketing campaign managers, holding the potential to enhance marketing strategies and campaign success (Safarkhani & Safara, 2016).

Anas Nabeel Falih AL-Shawi and co-authors evaluate 11 customer, product, and socioeconomic attributes, and compare four data mining methods. The study emphasizes the significance of feature selection and highlights the superiority of a proposed hybrid classification method in achieving high accuracy (99%) and a strong area under the ROC

curve (97%). It underscores the critical role of data mining in decision support systems and decision-making processes (AL-Shawi et al., 2019).

The study of Elsalamony evaluates and compares four data mining models: Multilayer Perception Neural Network (MLPNN), Tree Augmented Naïve Bayes (TAN), Logistic Regression (LR), and C5.0 decision tree classification. These models are applied to a real-world dataset related to direct marketing campaigns by a bank. Results indicate that C5.0 outperforms the other models. The key attribute affecting success is "Duration," and this analysis contributes to improved customer targeting and campaign effectiveness in banking (Elsalamony, 2014).

Tuba Parlar and Songul Kakilli Acaravci employ two feature selection methods: Information Gain and Chi-square, to identify the most relevant features for enhancing the effectiveness of marketing campaigns. They conduct experiments using a Naive Bayes classifier on a bank marketing dataset, and the results demonstrate that a reduced set of features improves classification performance. Both IG and Chi-square methods assess feature importance and yield similar rankings for the top five features. Reducing the feature size enhances classification performance. The ten highest-ranked features comprise duration, outcome, month, pdays, contact, previous, age, job, housing, and balance (Parlar & Acaravci, 2015).

In the article “An Efficient CRM-Data Mining Framework for the Prediction of Customer Behaviour”, two classification models were used: Multilayer Perception Neural Network (MLPNN) and Naïve Bayes (NB). The results show that MLPNN achieved a high accuracy rate of 88.63%, while NB performed well in terms of true positive rate and ROC area.

These findings provide valuable insights for customer behavior prediction in CRM (Bahari & Elayidom, 2015).

3. DATA SET

This project's dataset originates from a Portuguese bank's direct marketing campaigns, accessible at: http://archive.ics.uci.edu/dataset/222/bank+marketing. The dataset is divided into two subsets: "bank-full.csv" covering records from May 2008 to November 2010, and "bank.csv," representing a random 10% sample. This work uses the “bank-full.csv” dataset. The primary aim is binary classification, involving predicting whether clients will subscribe ('yes') or not ('no') to a term deposit, a type of fixed-income investment in which individuals invest a specific amount in a bank for a predetermined period, usually ranging from a few months to several years. To assess whether a client would subscribe to the product (a bank term deposit), multiple contacts with the same client were often necessary.

The full dataset comprises 45,211 records, with 16 input attributes related to client information and one output attribute ('y') indicating the subscription to a term deposit. These attributes include various data types such as numerical variables, categorical descriptors, and binary indicators. For a detailed attribute description, please refer to the table below:

Table 1. DATA DESCRIPTION

VARIABLE	VARIABLE DESCRIPTION	KIND
	# Personal information	
Age	Age of the client	Numeric
Job	Type of job	Categorical
Marital	Marital Status	Categorical
Education	("unknown", "secondary", "primary", "tertiary")	Categorical
Default	Has credit in default? (yes or no)	Binary
Balance	Average yearly balance, in euros	Numeric
Housing	Has housing loan? (yes or no)	Binary
Loan	Has personal loan? (yes or no)	Binary
	# Related with the last contact of the current campaign	
Contact	Contact communication type ("unknown","telephone","cellular")	Categorical
Day	Last contact day of the month	Numeric
Month	Last contact month of year	Categorical
Duration	Last contact duration, in seconds	Numeric
	# Other attributes	
Campaign	Number of contacts performed during this campaign and for this client (-1 means client was not previously contacted)	Numeric

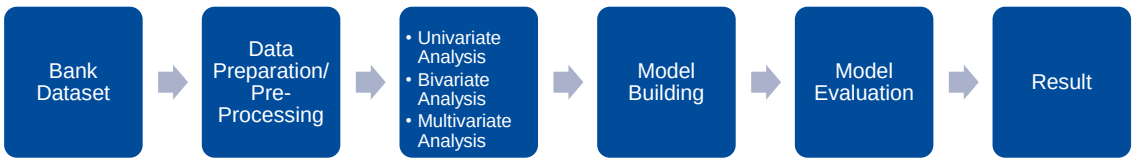
Pdays	Number of days that passed by after the client was last contacted from a previous campaign	Numeric
Previous	Number of contacts performed before this campaign and for this client	Numeric
Poutcome	Outcome of the previous marketing campaign	Categorical
	# Output variable (desired target):	
Y	Has the client subscribed a term deposit? (yes or no)	Binary

4. APPROACH

4.1 Data Preparation

The data preprocessing involved several key steps. First, a thorough check revealed no missing or duplicate data, ensuring data integrity. Next, binary 'yes/no' data was transformed into Boolean values. The dataset consisted of both numerical and categorical features. Descriptive statistical analysis was performed on the numerical variables to understand their distribution and tendencies. Notably, it became evident that the target variable "y" exhibited a significant class imbalance. Additionally, graphical analysis identified outliers in several variables, including 'age,' 'balance,' 'duration,' 'campaign,' 'pdays,' and 'previous.' To address these outliers, a grouping approach was chosen to manage their values, rather than removing them. As for the categorical variables, the 'job' variable was restructured into six distinct categories. Lastly, all categorical variables were converted into dummy variables (one-hot encoding) to make them suitable for machine learning models. The dataset now consists of a total of 66 numerical variables. These preprocessing steps have set the stage for more effective and robust predictive modeling.

Fig. 1. Overall methodology



4.2 Univariate Analysis

During the univariate analysis, a descriptive statistical analysis of the data was carried out, at first only on numerical data. In this dataset, the mean age of customers is approximately 40.94 years, with ages ranging from 18 to 95. Most customers have no default history (mean ≈ 0.018) and maintain a balance with a wide range from negative €8k to over €100k (mean ≈ €1,362.27). About 55.6% have housing loans, while around 16% have personal loans. Days of the month are represented (1 to 31), and call durations vary from 0 to 4,918 seconds (mean ≈ 258.16). The campaign contact count ranges from 1 to 63 (mean ≈ 2.76). Days since the last contact vary from -1 to 871 days (mean ≈ 40.20), but because -1 is a placeholder for customers who have not been contacted previously, the range of customers who have had contact in the past is between 1 and 871 days (mean ≈ 224.60). The number of previous contacts ranges from 0 to 275 (mean ≈ 0.58). The target variable "y" is binary, with "0" indicating "no" and "1" indicating "yes" responses.

Fig. 2. Statistics of numeric variables

	age	default	balance	housing	loan
count	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000
mean	40.936210	0.018027	1362.272058	0.555838	0.160226
std	10.618762	0.133049	3044.765829	0.496878	0.366820
min	18.000000	0.000000	-8019.000000	0.000000	0.000000
25%	33.000000	0.000000	72.000000	0.000000	0.000000
50%	39.000000	0.000000	448.000000	1.000000	0.000000
75%	48.000000	0.000000	1428.000000	1.000000	0.000000
max	95.000000	1.000000	102127.000000	1.000000	1.000000

	day	duration	campaign	pdays	previous
count	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000
mean	15.806419	258.163080	2.763841	40.197828	0.580323
std	8.322476	257.527812	3.098021	100.128746	2.303441
min	1.000000	0.000000	1.000000	-1.000000	0.000000
25%	8.000000	103.000000	1.000000	-1.000000	0.000000
50%	16.000000	180.000000	2.000000	-1.000000	0.000000
75%	21.000000	319.000000	3.000000	-1.000000	0.000000
max	31.000000	4918.000000	63.000000	871.000000	275.000000

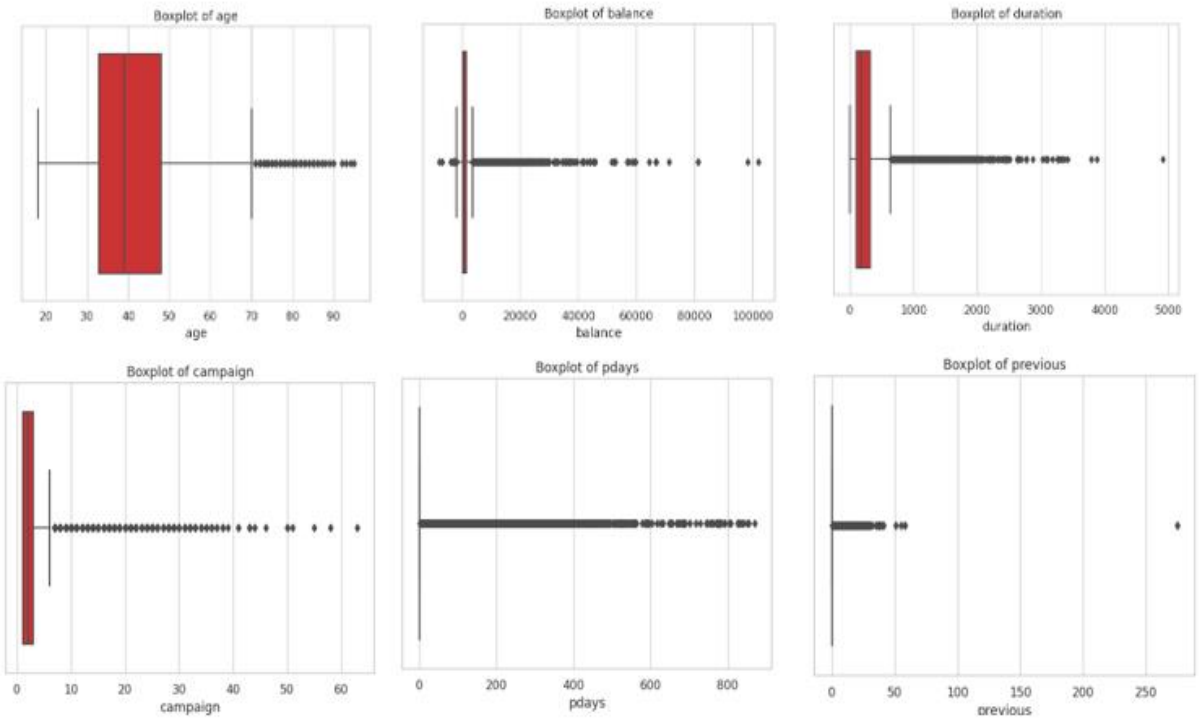
	y
count	45211.000000
mean	0.116985
std	0.321406
min	0.000000
25%	0.000000
50%	0.000000
75%	0.000000
max	1.000000

count	8257.000000
mean	224.577692
std	115.344035
min	1.000000
25%	133.000000
50%	194.000000
75%	327.000000
max	871.000000

Name: pdays, dtype: float64

As we can see in the graphs below, six numerical variables presented a large number of outliers:

Fig. 3. Numeric variables with outliers



Bar charts and column charts were generated to better understand the distribution of the values of the categorical and Boolean variables. Notably, the majority of clients exhibit some level of educational attainment, with secondary education being the most prevalent. Additionally, a significant portion of campaign contacts was conducted via clients' mobile phones. The month of May stands out as the peak for campaign outreach, with approximately 14,000 contacts, nearly double the count of July, which ranks second. However, information regarding the outcomes of the previous campaign is limited, with the vast majority of data marked as "unknown." As observed in the graphs, the variables ‘job’, ‘education’ and ‘contact’ also contain "unknown" values, which are retained. In the case of categorical variables, the utilization of an "unknown" value is warranted when a

category is undefined or ambiguous. This approach prevents the introduction of a fictitious category that could distort the analysis or context.

The dataset indicates a majority of married clients, with twice the number of unmarried clients and 5,000 individuals classified as divorced. In relation to Boolean variables, three out of the four variables exhibit significant imbalances, including the target variable "y," which demonstrates low campaign subscription, with only 5,000 adherents compared to 40,000 non-subscribers. Approximately 55.58% of clients have housing loans, and around 16.02% have personal loans. A small fraction of clients (about 1.8%) have a credit default.

Fig. 4. Bar charts of categorical variables

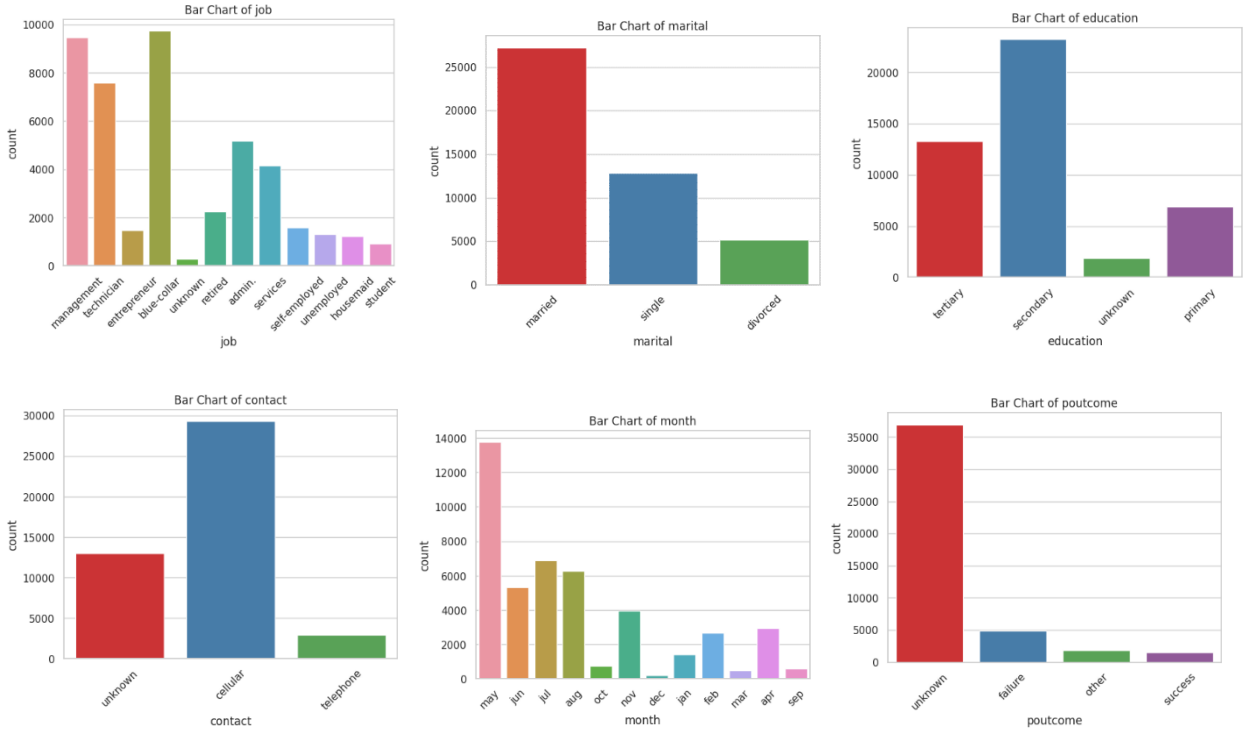
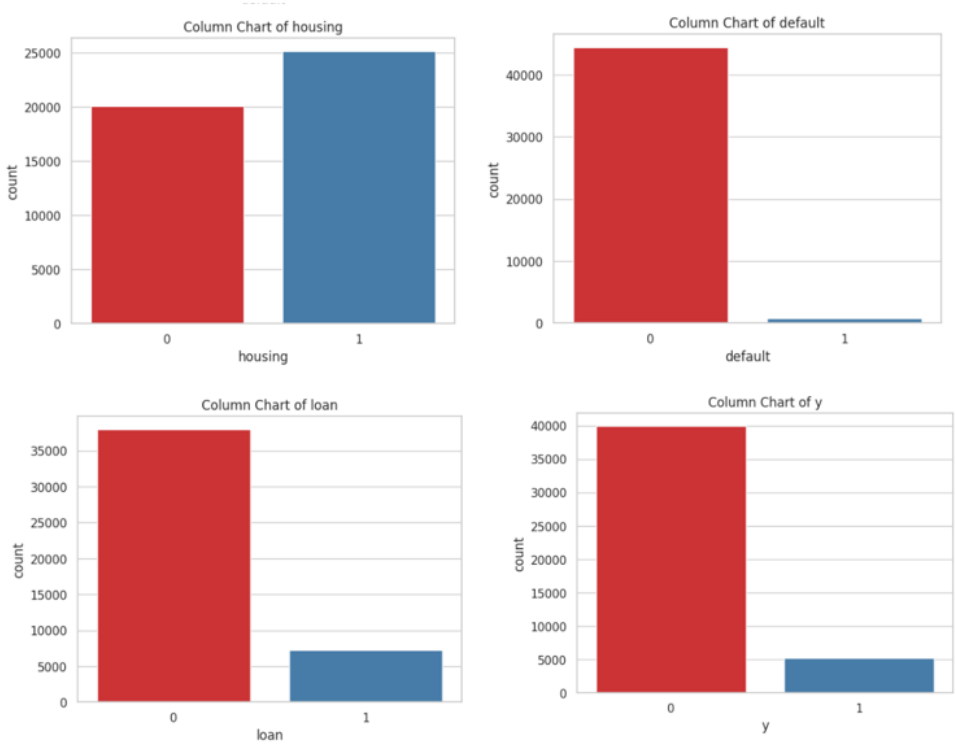


Fig. 5. Column charts of Boolean variables



Data was transformed by bucketing followed by one-hot encoding to address the outliers in the data and convert categorical variables into numeric variables in preparation for bivariate analysis. In total the prepared data now contains 66 variables, including the target variable.

4.3 Bivariate Analysis

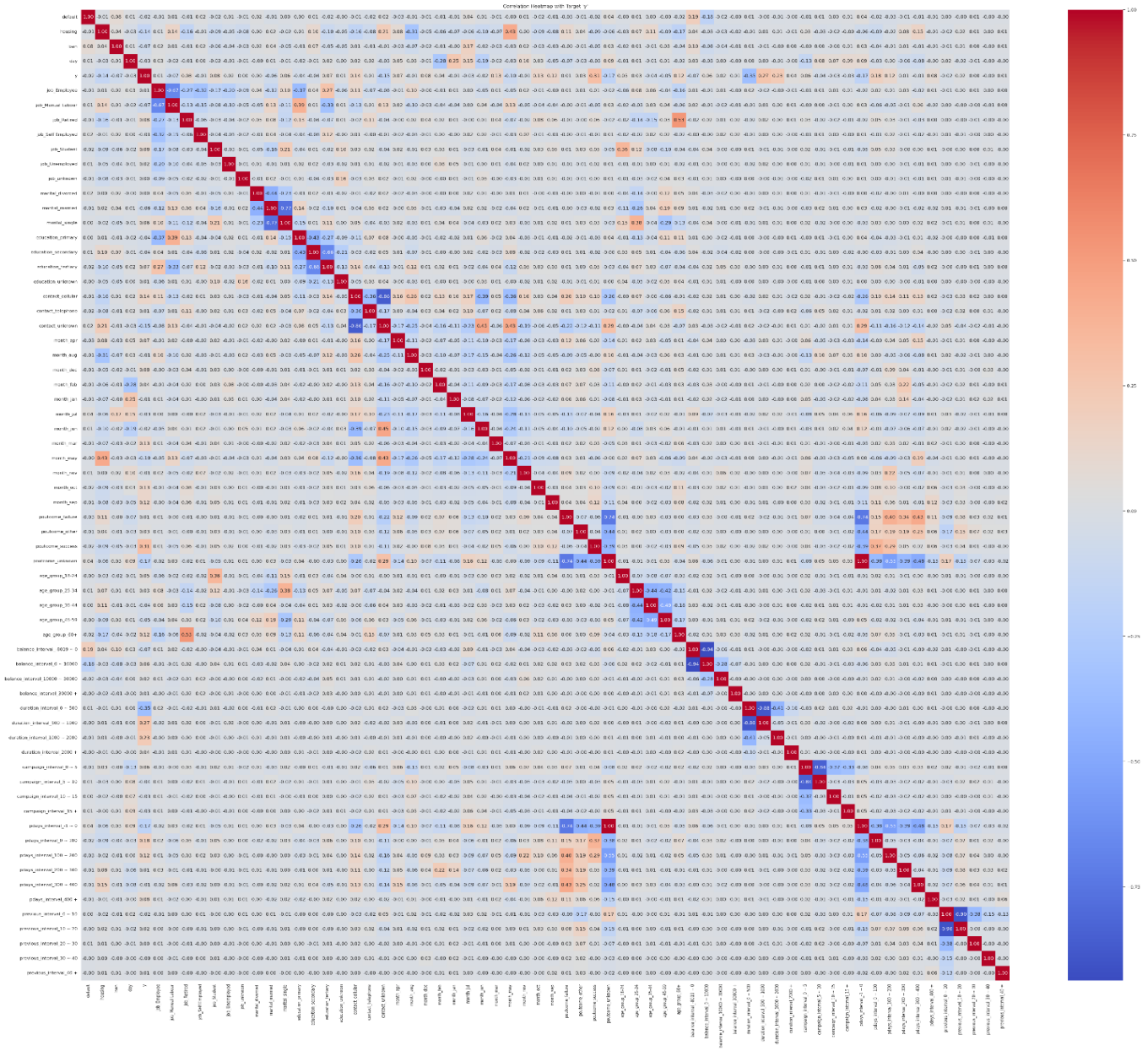
Following data transformation, a correlation matrix was generated to explore potential associations between variable pairs. To prevent correlations across variables stemming from the same category (e.g., employee and manual labor), we can consider preserving only one of the variables and dropping the other. Primary education exhibited a negative correlation with employee labor and a positive correlation with manual labor, while tertiary education showed a negative correlation with manual labor. Due to the significant correlation between education level and job types, it might be advantageous to consider dropping one category.

We anticipated a correlation between the 18-24 age group and student status. A clear correlation was also observed between the oldest age group and retirement. In addition, age groups also displayed an expected relationship with marital status, with younger age groups positively correlated with the "single" status and older age groups displaying a negative correlation. These instances suggest the possibility of excluding one or more of these variables with strong association.

Peculiarly, there was a positive correlation between housing and the month of May and a negative correlation with the month of August. A similar situation was noted with the negative correlation between mobile contact and the months of May and June versus the positive correlation between the unknown mode of contact and the same months. However, these correlations may potentially be spurious and may require further analysis.

Notwithstanding, a correlation was observed between the success of the previous campaign and the current one, as well as age groups 18-24 and 60+. The duration of the call also exhibited correlations with campaign success, with shorter calls displaying a negative correlation, and medium-duration calls showing a positive correlation. Strikingly, there was no apparent correlation between the longest call duration interval and the campaign's outcome. Lastly, clients contacted more recently tended to achieve more successful outcomes in the previous campaign than those contacted less recently. It is important to note, however, that the previous call interval of -1 did not apply to any of the previous contact outcome variables.

Fig.6. Heatmap of the correlation matrix



4.4 Variable Selection Following Bivariate Analysis

Based on the analysis of the correlation matrix, it was decided to drop specific variable groups, resulting in the removal of 'job type,' 'marital status,' and 'months' from the dataset in order to simplify it by eliminating redundant or less informative variables, which could potentially reduce multicollinearity and enhance the interpretability and performance of the model.

The decision to drop 'job type' stemmed from the significant correlations between job types and other variables. Specifically, 'age' already effectively captured 'retired' and 'student' job types, rendering a separate 'job type' variable redundant. Additionally, 'education' already represented ‘manual labor' in the case of primary education and 'management' in the case of tertiary education.

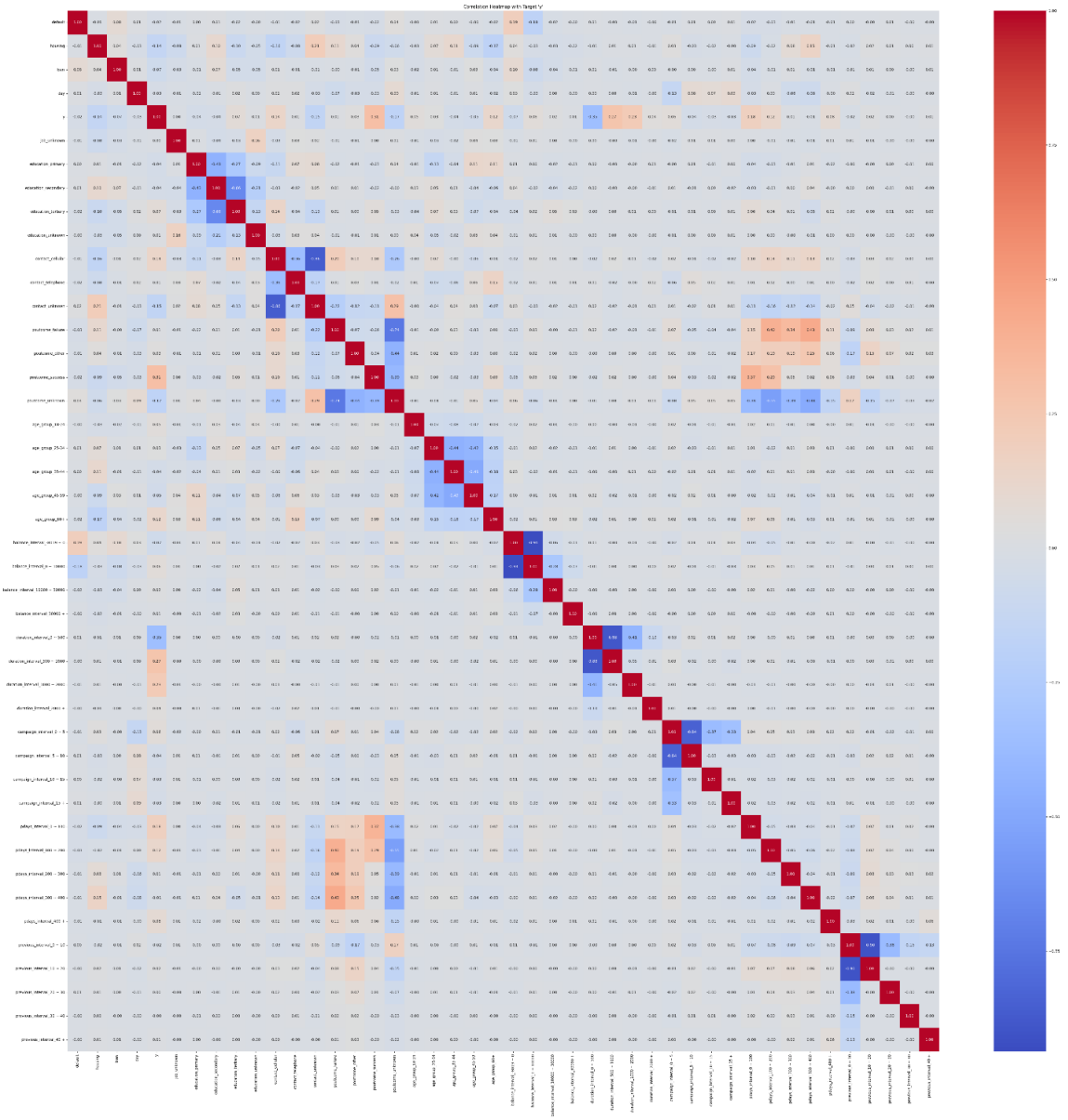
Similarly, 'marital status' was dropped as it exhibited strong correlations with 'age', with younger age groups positively correlated with the 'single' status, and older age groups displaying a negative correlation. Thus, it became apparent that this variable group had already captured the pertinent information, thus negating the need for a separate 'marital status' group of variables.

The decision to drop the 'months' variables was influenced by the strong correlation with 'housing' status. Notably, 'May' exhibited a robust positive correlation with 'housing' = 'yes,' while 'August' showed a strong negative correlation with 'housing' = 'no.' These correlations indicated that perhaps certain months may have been associated with

targeted campaigns based on the housing loan situation, while other months showed no apparent significance.

Lastly, the variable ‘pdays_interval_-1 ~ 0’ was dropped due to its perfect correlation with ‘poutcome_unknown’. A new correlation matrix was generated following this analysis.

Fig.7. Revised heatmap of the correlation matrix



5. INITIAL RESULTS

5.1 Baseline Models

A predictive model is a process that utilizes statistics and machine learning models to make predictions based on historical data. In our case, we are interested in finding the model that will most accurately predict the number of subscriptions to the marketing campaign. Initially, we will employ three baseline models: Logistic Regression, Random Forest, and Naïve Bayes. These models will serve as a starting point for refinement and evaluation. Cross-validation is a technique used to assess a model's performance by splitting the dataset into multiple subsets, training and testing the model on different partitions to obtain a more robust evaluation of its generalization capabilities. The performance of these three models can be observed in the following figures:

Fig.8 (a) Result of Logistic Regression Baseline Model

```
Cross-Validation Scores:
[0.8980833  0.8956682  0.89511521 0.89806452 0.8921659 ]
Classification Report (Validation Set):
              precision    recall  f1-score   support

           0       0.92      0.97      0.94      7994
           1       0.61      0.32      0.42      1048

    accuracy                           0.90      9042
   macro avg       0.76      0.65      0.68      9042
weighted avg       0.88      0.90      0.88      9042

Confusion Matrix (Validation Set):
[[7781  213]
 [ 711  337]]
Classification Report (Test Set):
              precision    recall  f1-score   support

           0       0.92      0.97      0.94      7990
           1       0.58      0.34      0.43      1053

    accuracy                           0.89      9043
   macro avg       0.75      0.65      0.68      9043
weighted avg       0.88      0.89      0.88      9043

Confusion Matrix (Test Set):
[[7736  254]
 [ 698  355]]
```

Fig.8 (b) Result of Random Forest Baseline Model

```
Cross-Validation Scores: [0.87873203 0.88442396 0.88       0.87612903 0.88202765]
Mean CV Accuracy: 0.880262535224665
Classification Report (Validation Set):
              precision    recall  f1-score   support

           0       0.92      0.95      0.93      7994
           1       0.49      0.37      0.42      1048

    accuracy                           0.88      9042
   macro avg       0.71      0.66      0.68      9042
weighted avg       0.87      0.88      0.88      9042

Confusion Matrix (Validation Set):
 [[7594  400]
 [ 660  388]]
Classification Report (Test Set):
              precision    recall  f1-score   support

           0       0.92      0.94      0.93      7990
           1       0.46      0.36      0.40      1053

    accuracy                           0.88      9043
   macro avg       0.69      0.65      0.67      9043
weighted avg       0.86      0.88      0.87      9043

Confusion Matrix (Test Set):
 [[7546  444]
 [ 676  377]]
```

Fig.8 (c) Result of Naïve Bayes Baseline Model

```
Cross-Validation Scores: [0.83984519 0.80387097 0.85327189 0.84976959 0.85105991]
Mean CV Accuracy: 0.839563508011435
Classification Report (Validation Set):
              precision    recall  f1-score   support

           0       0.94      0.87      0.91      7994
           1       0.38      0.60      0.47      1048

    accuracy                           0.84      9042
   macro avg       0.66      0.74      0.69      9042
weighted avg       0.88      0.84      0.86      9042

Confusion Matrix (Validation Set):
 [[6991 1003]
 [ 422  626]]
Classification Report (Test Set):
              precision    recall  f1-score   support

           0       0.94      0.88      0.91      7990
           1       0.40      0.60      0.48      1053

    accuracy                           0.85      9043
   macro avg       0.67      0.74      0.69      9043
weighted avg       0.88      0.85      0.86      9043

Confusion Matrix (Test Set):
 [[7035  955]
 [ 423  630]]
```

The results of the three classification models using cross-validation were compared. In order to compare the performance of the models, we will be focusing on two metrics: accuracy and recall. While accuracy is used to measure the overall performance of the model, recall is used to measure our true positives rate (representing those individuals the model correctly predicted as having subscribed to the campaign). In this context of a marketing campaign where the accurate identification of campaign subscribers can be crucial in the efficient allocation of resources, recall should be our focus. The results show that Logistic Regression achieves an average accuracy of around 89%, but it has a relatively low recall for class 1, indicating a challenge in correctly identifying these cases. Random Forest, although having a similar average accuracy, slightly improves the recall of class 1 compared to Logistic Regression. On the other hand, Naïve Bayes has a higher recall for class 1, making it more effective in identifying people who subscribed to the campaign, although its accuracy is slightly lower.

5.2 Hyperparameter Tuning

Hyperparameter tuning aims to find the best configuration for a machine learning model, optimizing its accuracy. This is done by varying hyperparameters and evaluating performance on validation data. The goal is to maximize the model's performance to meet the problem's requirements.

Hyperparameter tuning was performed in all the models. In the Logistic Regression model, the hyperparameter 'C' governs regularization, and a grid search tested various values (0.001, 0.01, 0.1, 1, 10, 100) to find the best one. The optimal 'C' value was selected based on cross-validation, resulting in an improved Logistic Regression model.

Fig. 9 (a) Result of Logistic Regression model after Hyperparameter tuning

```
Cross-Validation Scores:
[0.83984519 0.80387097 0.85327189 0.84976959 0.85105991]
Classification Report (Validation Set):
              precision    recall  f1-score   support

           0       0.92      0.97      0.94      7994
           1       0.61      0.32      0.42      1048

    accuracy                           0.90      9042
   macro avg       0.76      0.65      0.68      9042
weighted avg       0.88      0.90      0.88      9042

Confusion Matrix (Validation Set):
[[7778  216]
 [ 708  340]]
Classification Report (Test Set):
              precision    recall  f1-score   support

           0       0.92      0.97      0.94      7990
           1       0.58      0.34      0.43      1053

    accuracy                           0.89      9043
   macro avg       0.75      0.65      0.68      9043
weighted avg       0.88      0.89      0.88      9043

Confusion Matrix (Test Set):
[[7735  255]
 [ 697  356]]
```

Comparing the two Logistic Regression models, we can see that after hyperparameter tuning, the model's performance remained the same of recall for class 1. In both models, the recall for class 1 was 32% on the validation set and 34% on the test set. The other metrics like precision, f1-score, and accuracy also remained consistent.

In the Random Forest model, a hyperparameter grid was defined, including the number of estimators, the maximum tree depth, the splitting criterion, and class weights.

Fig. 9 (b) Result of Radom Forest model after Hyperparameter tuning

```
Cross-Validation Scores: [0.84647991 0.81861751 0.83778802 0.82875576 0.83078341]
Mean CV Accuracy: 0.8324849223995747
Classification Report (Validation Set):
              precision    recall  f1-score   support

           0       0.96      0.84      0.90      7994
           1       0.38      0.74      0.50      1048

    accuracy                           0.83      9042
   macro avg       0.67      0.79      0.70      9042
weighted avg       0.89      0.83      0.85      9042

Confusion Matrix (Validation Set):
 [[6718 1276]
 [ 277  771]]
Classification Report (Test Set):
              precision    recall  f1-score   support

           0       0.96      0.84      0.90      7990
           1       0.38      0.74      0.50      1053

    accuracy                           0.83      9043
   macro avg       0.67      0.79      0.70      9043
weighted avg       0.89      0.83      0.85      9043

Confusion Matrix (Test Set):
 [[6749 1241]
 [ 279  774]]
```

In the Random Forest model before hyperparameter tuning, the average cross-validation accuracy is approximately 88%, with a recall of 37% for class 1 in the validation set and 35% in the test set. After hyperparameter tuning, the average cross-validation

accuracy dropped slightly to 83%, but with a recall of 74% for class 1 in the validation set and 74% in the test set. The results indicate that hyperparameter tuning significantly improved the recall of class 1, making the model more effective in identifying this class.

In the Naïve Bayes model, the adjusted hyperparameter was "priors," which represents the a priori probability of each class in the Naive Bayes model. Different sets of priors values were tested to find the combination that optimizes the model's performance.

Fig. 9 (c) Result of Naïve Bayes model after Hyperparameter tuning

```
Cross-Validation Scores: [0.83984519 0.80387097 0.85327189 0.84976959 0.85105991]
Mean CV Accuracy: 0.839563508011435
Classification Report (Validation Set):
              precision    recall  f1-score   support

           0       0.94      0.87      0.91      7994
           1       0.38      0.60      0.47      1048

    accuracy                           0.84      9042
   macro avg       0.66      0.74      0.69      9042
weighted avg       0.88      0.84      0.86      9042

Confusion Matrix (Validation Set):
 [[6991 1003]
 [ 422  626]]
Classification Report (Test Set):
              precision    recall  f1-score   support

           0       0.94      0.88      0.91      7990
           1       0.40      0.60      0.48      1053

    accuracy                           0.85      9043
   macro avg       0.67      0.74      0.69      9043
weighted avg       0.88      0.85      0.86      9043

Confusion Matrix (Test Set):
 [[7035  955]
 [ 423  630]]
```

Before hyperparameter tuning in the Naïve Bayes model, the average cross-validation accuracy is approximately 84%, with a recall of 60% for class 1 in the validation set and 60% in the test set. After hyperparameter tuning, the average cross-validation accuracy remains around 84%, with a recall of 60% for class 1 in the validation set and 60% in the test set. The results indicate that hyperparameter tuning did not impact the model's performance in terms of recall for class 1.

Analyzing all the results after hyperparameter tuning, we can see that the Random Forest model achieved the best result with a recall of 74% for class 1.

6. Conclusion

The conducted research focused on predicting success in bank telemarketing campaigns using models such as Logistic Regression, Random Forest, and Naïve Bayes. The starting point was the assumption presented by Moro, Cortez, and Rita (2014) regarding the effectiveness of data mining in this context. Due to the imbalance in the target variable in the dataset, accuracy and recall metrics were prioritized. The dataset, originating from direct campaigns of a Portuguese bank, underwent a comprehensive preparation process, including data transformation, univariate and bivariate analysis, and variable selection. The study focused on binary classification to predict whether clients would subscribe to a term deposit or not.

In AL-Shawi et al. (2019), 11 attributes were evaluated, with emphasis placed on the importance of feature selection. This study, on the other hand, adopted a more comprehensive approach, considering the 16 attributes related to customer information. Elsalamony (2014) identified the C5.0 model as superior in their context, focusing on direct marketing campaigns of a bank. In contrast, this work focused on more common models, providing a comparative perspective on these widely used models in classification problems.

Naïve Bayes yielded better results in the initial test, with a class 1 recall reaching about 60% – almost double that of other models. In the work of Bahari & Elayidom Naïve Bayes also showed a higher class 1 recall of 47% compared to the other model, the Multilayer Perception Neural Network. However, following hyperparameter tuning, Random Forest significantly increased its class 1 recall to 74%, while Logistic Regression remained at around 34%, and Naïve Bayes maintained the 60%. Although Random Forest's accuracy was slightly lower (83%) compared to the other models (between 85% and 90%), the considerable increase in true positives for class 1 suggests that Random Forest is the most effective model among the three, making it possible to create a predictive model.

Conducting this project yielded an opportunity to explore the application of data mining, Python language, and machine learning models to a business problem. This work has shown that it is possible to obtain satisfactory accuracy and recall using a Random Forest model. However, an opportunity exists to further explore more machine learning models and apply them, as well as explore the use of a hybrid classification method, as was done by AL-Shawi et al. (2019).

6.1 Findings and Recommendations

Bivariate analysis revealed a significant uptake of long-term deposits among two specific demographic groups: young individuals aged 18 to 24, predominantly unmarried and students, and those over 60 years old, mostly married and retired. In this regard, we suggest targeting telemarketing campaigns towards these groups or even creating exclusive products that cater to their distinct needs. For instance, young individuals may be interested in utilizing long-term deposits to fund their studies or acquire their first car/house, while older individuals may view the product as an additional investment. A more in-depth analysis of these groups can provide valuable insights into their specific requirements.

The duration of the calls also showed correlation with campaign success. Calls of medium duration exhibited a positive correlation, indicating an increase in campaign success. In contrast, short and long calls showed negative correlations. It is important to note that the negative correlation in short calls may be a natural consequence, as many individuals disconnect telemarketing calls before even hearing the proposal. Regarding long calls, the negative correlation may suggest a potential lack of employee skills in explaining the product effectively. This aspect can be improved with more effective employee training, empowering them to present the product more clearly and persuasively.

Remarkably, clients contacted more recently since the previous campaigns tended to have more successful outcomes in the new campaign. Therefore, it is suggested that a scheduling or reminders system be put in place in order to avoid long periods of lack of contact with clients in between campaigns. However, it is crucial to emphasize that excessive call frequency in a short period of time can be perceived as inconvenient, leading the client to react negatively to future contacts from the new campaign. Therefore, careful management of contact frequency is essential to maintain client receptiveness.

References

[1] AL-Shawi, A. N. F., et al. (2019). Hybrid Datamining Approaches to Predict Success of Bank Telemarketing. *International Journal of Computer Science and Mobile Computing*. *Vol.8 Issue.3, March,* 49-60. Moro, S., Cortez, P. & Rita, P. (2014)

[2] Elsalamony. H. A. (2014). Bank Direct Marketing Analysis of Data Mining Techniques. *International Journal of Computer Applications (0975 – 8887) Volume 85 – No 7, January*. https://www.researchgate.net/publication/263054095_Bank_Direct_Marketing_Analysis_of_Data_Mining_Techniques

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

[4] Parlar, T., Acaravci, S. K. (2017). Using Data Mining Techniques for Detecting the Important Features of the Bank Direct Marketing Data. *International Journal of Economics and Financial Issues, 7(2), 692-696*. https://www.econjournals.com/index.php/ijefi/article/view/4580

[5] Safarkhani, F. & Safara, F. (2016). Using data mining techniques to predict the success of bank telemarketing. *1st International Conference on New Research Achievements in Electrical and Computer Engineering (ICNRAECE).* https://www.researchgate.net/publication/355108256_Using_data_mining_techniques_to_predict_the_success_of_bank_telemarketing

[6] T. Femina Bahari and M. Sudheep Elayidom. (2015). An Efficient CRM-Data Mining Framework for the Prediction of Customer Behaviour. *Procedia Computer Science 46, 725 – 731.* https://www.researchgate.net/publication/275721122_An_Efficient_CRM-Data_Mining_Framework_for_the_Prediction_of_Customer_Behaviour