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.

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

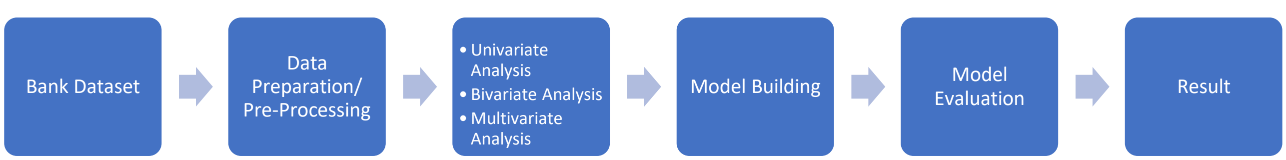
	# Related with the last contact of the current campaign	
Contact	Contact communication type ("unknown","telephone","cellular")	Categorical
Day	Last contact day of the month	Numeric
Month	Last contact month of year	Categorical
Duration	Last contact duration, in seconds	Numeric
	# Other attributes	
Campaign	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	

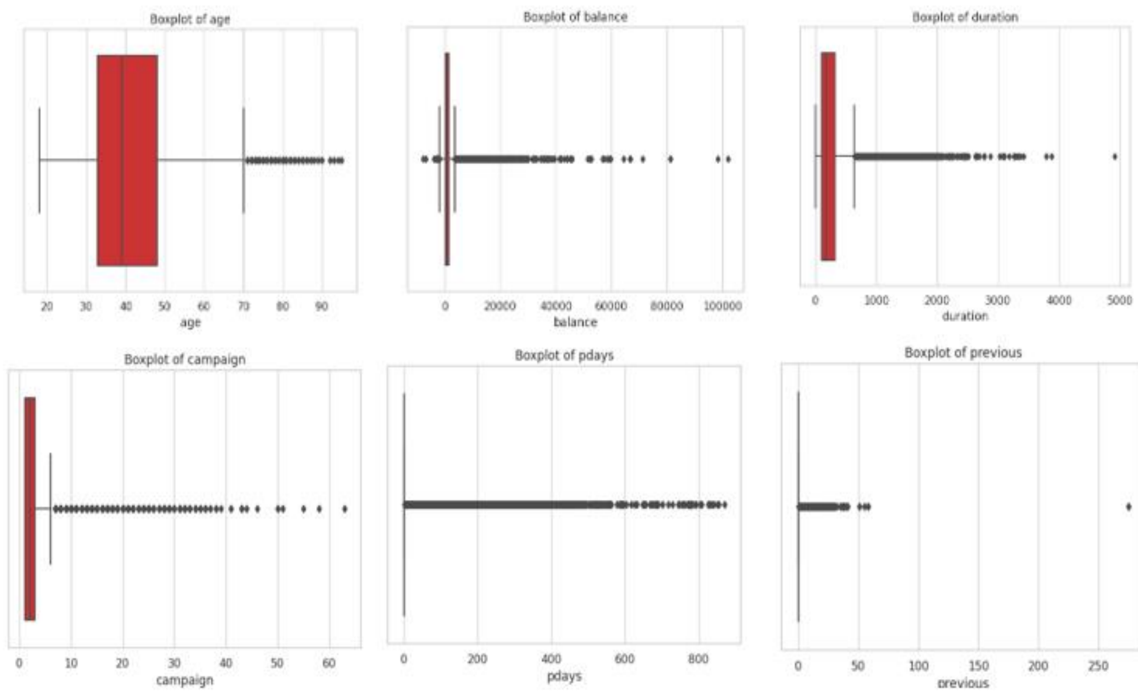
	y
count	45211.000000
mean	0.116985
std	0.321406
min	0.000000
25%	0.000000
50%	0.000000
75%	0.000000
max	1.000000

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

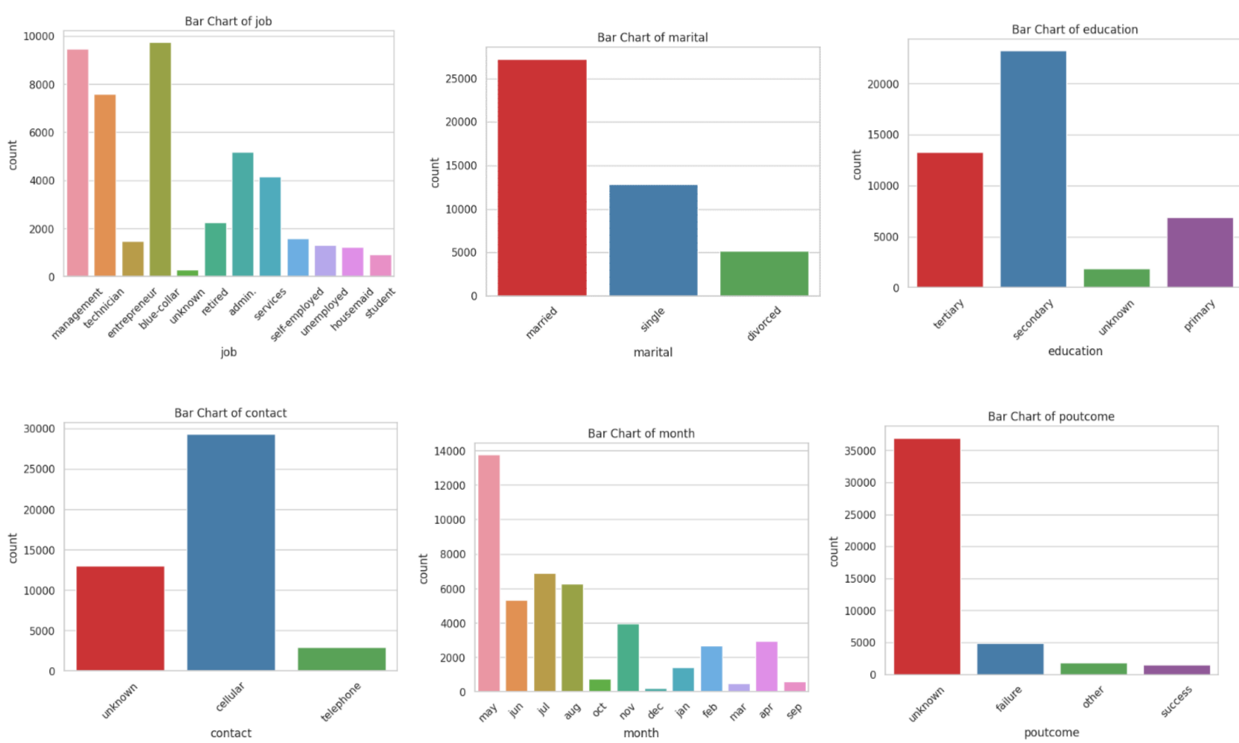
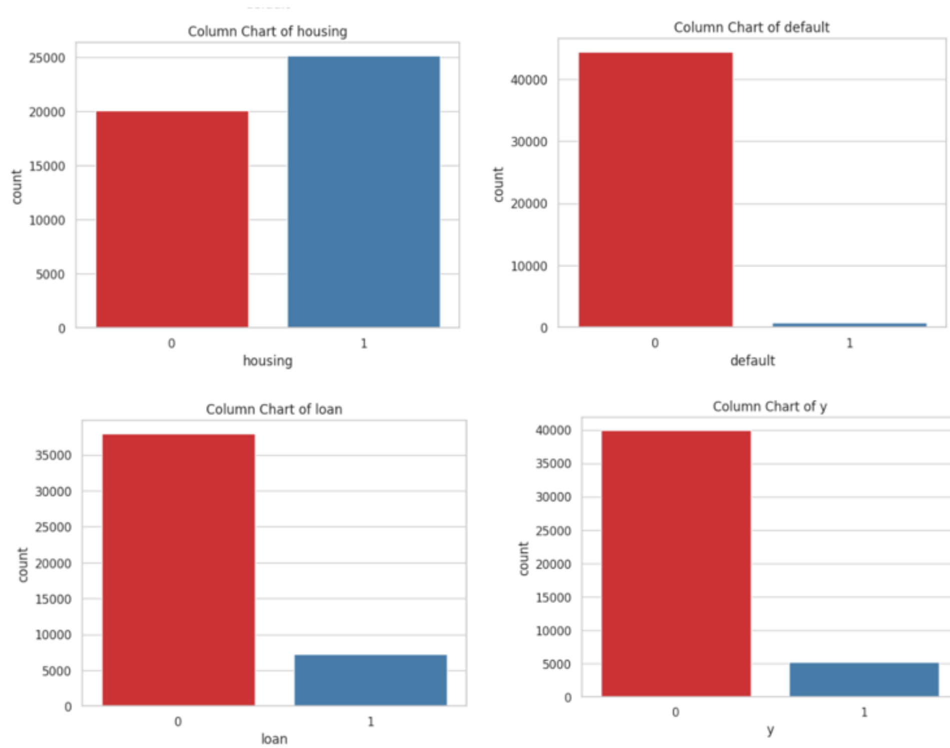


Fig. 5. Column charts of booleans variables



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

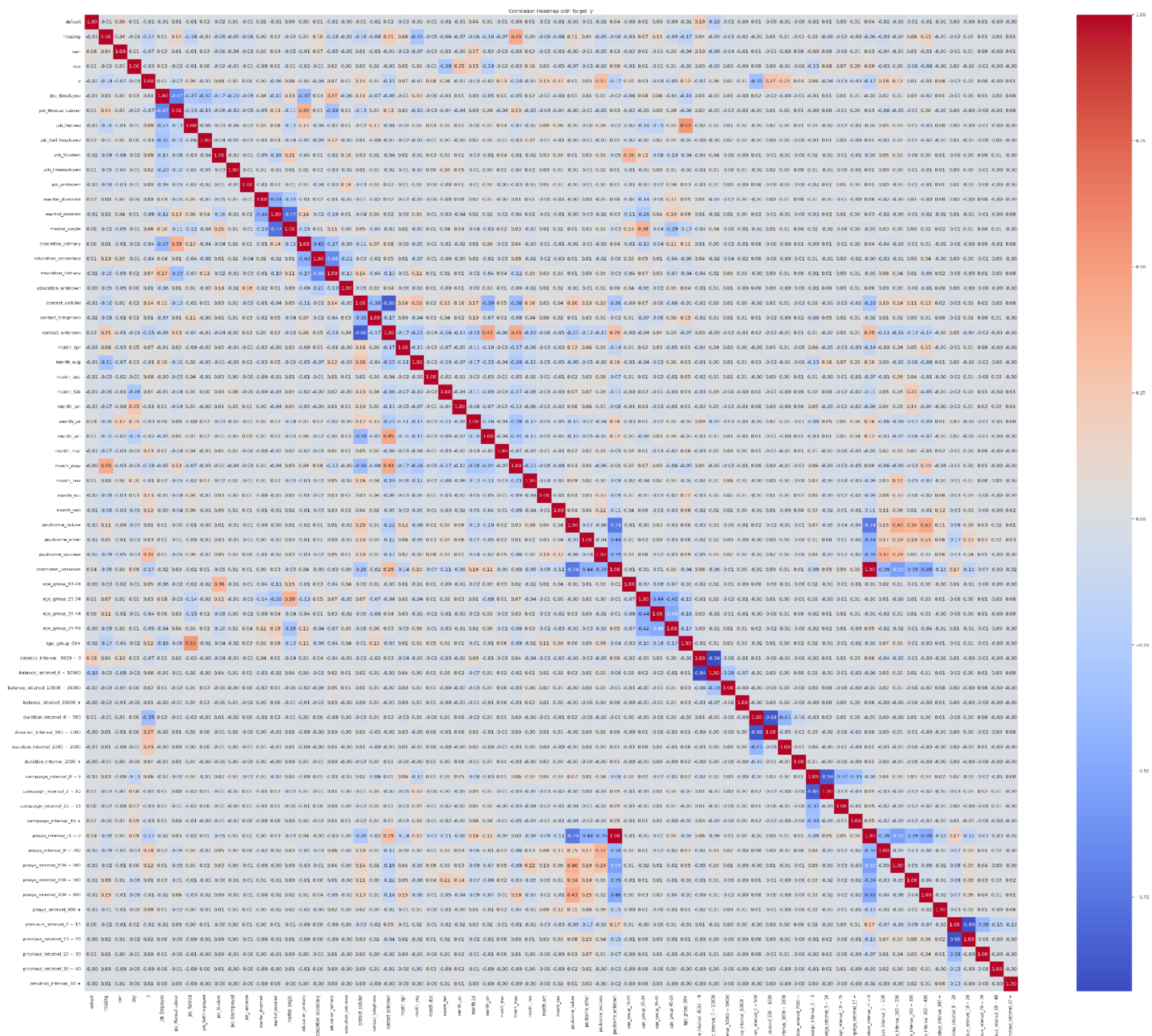
Following data transformation, a correlation matrix was generated to explore potential associations between variable pairs. To prevent correlations across variables stemming from the same category (e.g., employee and manual labor), we can consider preserving only one of the variables and dropping the 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. 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



References

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)

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

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

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

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

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