A person walking in a city

Description automatically generated< Classification and Regression

Analysis of the Bank Marketing

Dataset >

< Name: Maykel Kalta  
Student Number: 501229519  
Date of Submission: 26/01/2024>

Table of Contents

Abstract[………………………………………………………………………………………….. 3](#_Toc452457337)

Literature Review…………………………………………...………………………………….... 4

Data preprocessing and EDA……………………………………………………………………. 9

Evaluation……………………………………………………………………………………… 24

Conclusion……………………………………………………………………………………….25

Limitation of the work …………………………………………………………………………. 26

GitHub link………………………………………………………………………………………27

Reference………………………………………………………………………………………. 28

Abstract

The abstract provided is for a project titled "Classification and Regression Analysis of the Bank Marketing Dataset," submitted by Maykel Kalta. The document outlines an initiative aimed at enhancing the effectiveness of telemarketing strategies for selling long-term deposit accounts in a Portuguese bank. The bank's telemarketing campaigns, which are the focus of the dataset analysis, necessitated multiple contacts with potential customers to assess their willingness to subscribe to term deposit accounts.

**Key Components of the Abstract:**

* **Objective:** The project's main goal is to assist the Portuguese bank in developing an effective telemarketing strategy to identify potential customers likely to subscribe to a deposit account. This involves analyzing factors influencing customer subscriptions to long-term deposit accounts and improving marketing campaigns based on dataset insights.
* **Data:** The analysis utilizes the Bank Marketing Dataset, encompassing data from telemarketing campaigns of a Portuguese bank. The dataset, with 17 attributes and 45,211 rows, includes details on bank client demographics, account information, and campaign interactions. Attributes include age, job, marital status, education level, default status, account balance, housing and personal loans, contact type, and outcomes of previous marketing campaigns.
* **Research Approach:** The project proposes using exploratory data analysis (EDA) for data visualization and analysis, addressing data quality issues like duplicates and outliers. Machine learning techniques such as logistic regression, decision trees, support vector machines (SVM), and K-Nearest Neighbors (KNN) will be employed to achieve the classification goal.
* **Tools:** The primary tools for the project are R Studio and Python, with various libraries to support data analysis and machine learning tasks.

Literature review

In banks, huge data records information about their customers. This data can be used to create and keep clear relationship and connection with the customers in order to target them individually for definite products or banking offers. Usually, the selected customers are contacted directly through personal contact, telephone cellular, mail, and email or any other contacts to advertise the new product/service or give an offer, this kind of marketing is called direct marketing. In fact, direct marketing is in the main a strategy of many of the banks and insurance companies for interacting with their customers [19].

Historically, the name and identification of the term direct marketing suggested first time in 1967 by Lester Wunderman, which he is the father of direct marketing [11].

In addition, some of the banks and financial-services companies may depend only on strategy of mass marketing for promoting a new service or product to their customers. In this strategy, a single communication message is broadcasted to all customers through media such as television, radio, or advertising firm, etc. [18]. In this approach, companies do not set up a direct relationship to their customers for new product offers. In fact, many of the customers are not interesting or respond to this kind of sales promotion [20].

Accordingly, banks, financial-services companies and other companies are shifting away from mass marketing strategy because its ineffectiveness, and they are now targeting most of their customers by direct marketing for specific product and service offers [19, 20]. Due to the positive results clearly measured, many marketers attractive to the direct marketing. For example, if a marketer sends out 1,000 offers by mail and 100 respond to the promotion, the marketer can say with confidence that the campaign led immediately to 10% direct responses. This metric is known as the 'Response Rate', and it is one of many clear quantifiable success metrics employed by direct marketers. In dissimilarity, general advertising uses indirect measurements, such as awareness or engagement since there is no direct response from a consumer [11]. From the literature, the direct marketing is becoming a very important application in data mining these days. The data mining has been used widely in direct marketing to identify prospective customers for new products, by using purchasing data, a predictive model to measure that a customer is going to respond to the promotion or an offer [7].

**DATA MINING OVERVIEW:**

* Data mining, or Knowledge Discovery in Databases (KDD), is a process that aims to unearth valuable information from large datasets. It seeks to reveal novel and significant patterns, hidden insights, or unrecognized relationships within the data by leveraging a combination of various techniques. This field has become fundamental across a broad spectrum of industries, underpinning many applications and businesses that benefit from data-driven insights.
* The concept of data mining was outlined by U. Fayyad and G. Shapiro in 1996 as a comprehensive, interactive, and iterative procedure. This process encompasses several critical steps: gaining an understanding of the application domain, selecting relevant data, performing data preprocessing and cleaning, integrating data from multiple sources, reducing, and transforming data to a more manageable form, choosing appropriate data mining algorithms, and finally interpreting and presenting the results in a meaningful way. The goal is to utilize the knowledge uncovered through this process effectively.
* Data mining techniques can be broadly categorized into two main types: descriptive and predictive. Descriptive data mining focuses on finding patterns that provide insight into the data without predicting future events, whereas predictive data mining uses historical data to predict future outcomes, trends, or behaviors. Together, these approaches enable businesses and researchers to extract profound knowledge from their data, leading to better decision-making and strategic planning.
* A diagram of a process

  Description automatically generatedData mining itself is a step-in knowledge discovery process. The steps involved in knowledge discovery are:

**Data Selection**: The data relevant to the analysis is decided and retrieved from the various data locations.

**Data Preprocessing**: This stage consists of:

Data Cleaning: this is the removing of noisy data and irrelevant data from the data collected.

Data Transformation: This is where the selected data is transformed into forms appropriate for the mining procedure.

**Data Mining**: It is the crucial step in which clever techniques are applied to extract potentially useful patterns. The decision is made about the data mining technique to be used.

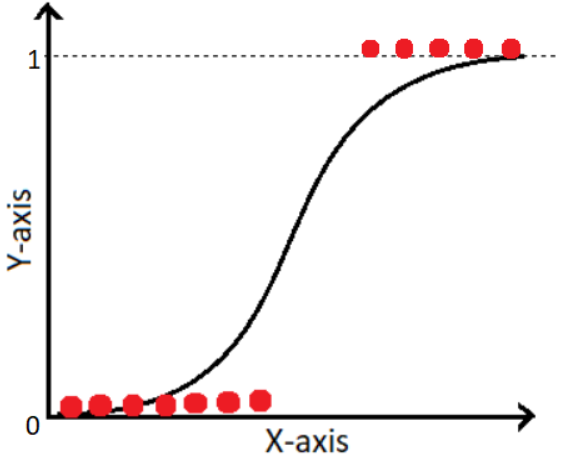
**Interpretation and Evaluation**: In this step, interesting patterns representing knowledge are identified based on given measures. The discovered knowledge is visually presented to the user. This essential step uses visualization techniques to help users understand.

**Classification in Data mining:**

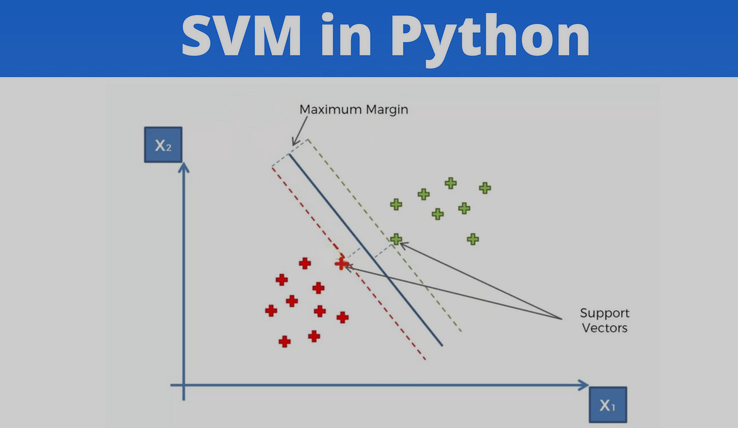
Classification in Data Mining

Classification is the most applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large [5]. Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. Basically, classification is used to classify each item in a set of data into one of predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network, and statistics [5]. A classification task begins with a data set in which the class assignments are known. In training process, a classification algorithm finds relationships between the values of the predictors and the values of the target. Different classification algorithms use different techniques for finding relationships. These relationships are summarized in a model, which can then be applied to a different data set in which the class assignments are unknown. Classification models are tested by comparing the predicted values to known target values in a set of test data. The historical data for a classification project is typically divided into two data sets: one for building the model; the other for testing the model.

**CLASSIFICATION ALGORITHM:**

1. **Logistic regression:** a process of modeling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome; something that can take two values such as true/false, yes/no, and so on.

**2)Support vector machine:** a kind of supervised learning algorithm that's employed in machine learning to tackle problems related to regression and classification; SVMs excel at handling binary classification problems, which call for dividing a data set's parts into two categories.

****

**3)KNN: a** non-parametric, supervised learning classifier which classifies data points based on proximity and forecasts how they will be grouped together.

The data is known as bank marketing from UC Irvine as shown in table1.

A table of attributes with text

Description automatically generated

*Data preprocessing and exploratory data analysis*

* **Data descriptive**

**A table with numbers and text

Description automatically generated**

The table displays summary statistics for numerical attributes of the bank marketing dataset. The 'count' row shows there are 45,211 entries for each attribute, indicating no missing values. The mean row provides the average value for each attribute, for example, the average age of clients in the dataset is approximately 40 years. The std (standard deviation) row measures the amount of variation or dispersion in the attributes; for instance, 'balance' has the highest variability with a standard deviation of over 3,044, indicating a widespread in the account balance amounts among the clients and previous has the lowest values indicating that the data points are clustered more closely around the mean. The min and max rows reveal the range of the data, from the lowest to the highest value, such as 'age' ranging from 18 to 95 years. The '25%', '50%' (median), and '75%' rows represent the quartile values, providing insights into the distribution of the data. For example, Age 25% of the individuals are younger than 33, the median age is 39, and 75% are younger than 48. the median day value is 16, suggesting that most last contacts occurred around the middle of the month. Campaign In 75% of the cases, individuals were contacted no more than 3 times

* **Some research questions are explicitly stated and appropriate for my project's objectives.**

**1 - How does the age distribution of clients subscribing to term deposits compare to those who do not?**

* The histogram indicates that consumers between the ages of 30 and 40 are more likely to subscribe to term deposits.

A graph of age distribution

Description automatically generated

**2 - How does the job type affect the likelihood of subscribing to a term deposit?**

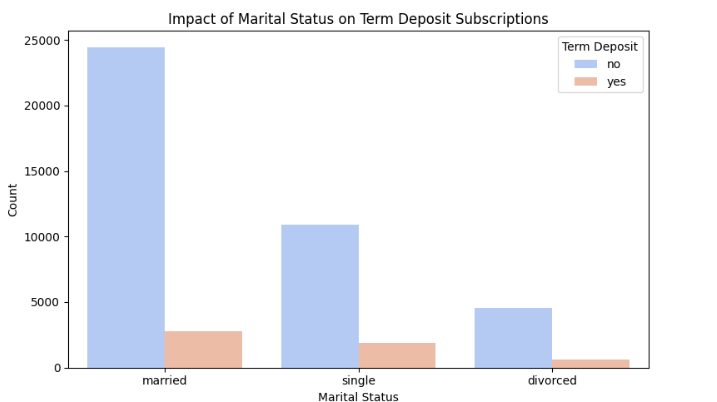
* The horizontal bar charts indicates that individuals in management position are more likely to subscribe to term deposits, but students are less in numbers they have a high subscription rate in relation to their group size.

**A screenshot of a computer

Description automatically generated**

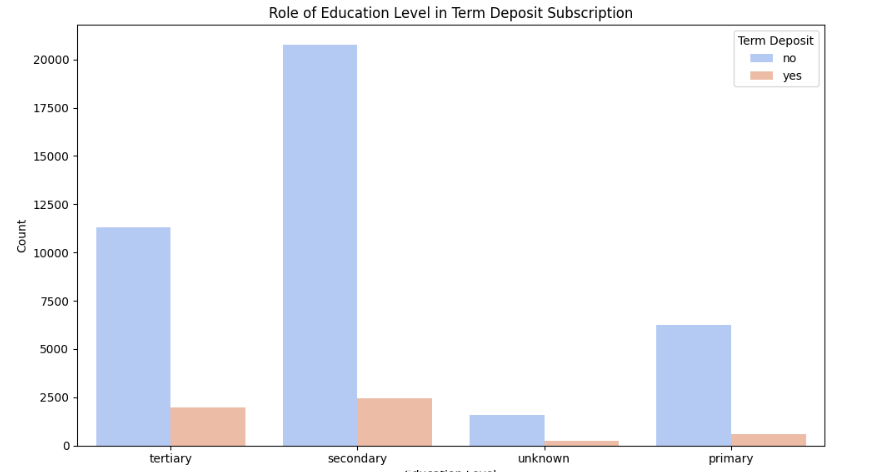
**3 - What is the impact of marital status on term deposit subscriptions?**

* Married clients have the highest count to not to subscribe to a term deposit as it shown in the bar chart.



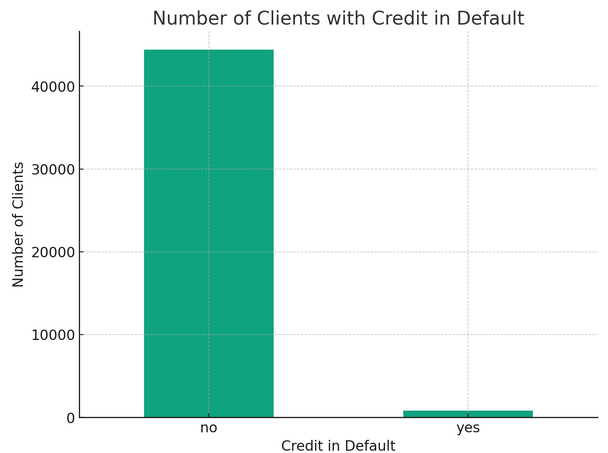
**4 – What role does education level play in the decision to subscribe to a term deposit?**

The bar chart indicates clients with secondary education had a lower proportion of term deposit subscriptions compared to those with tertiary education.

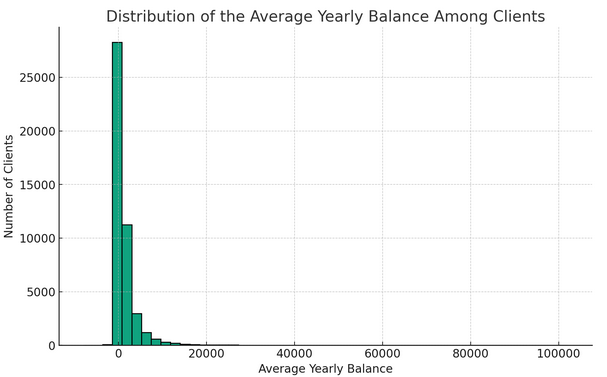


5 - How many clients have credit in default?

* The Credit in Default Distribution plot indicates the number of clients who have credit in default versus those who do not. It appears that a vast majority of clients do not have credit in default.

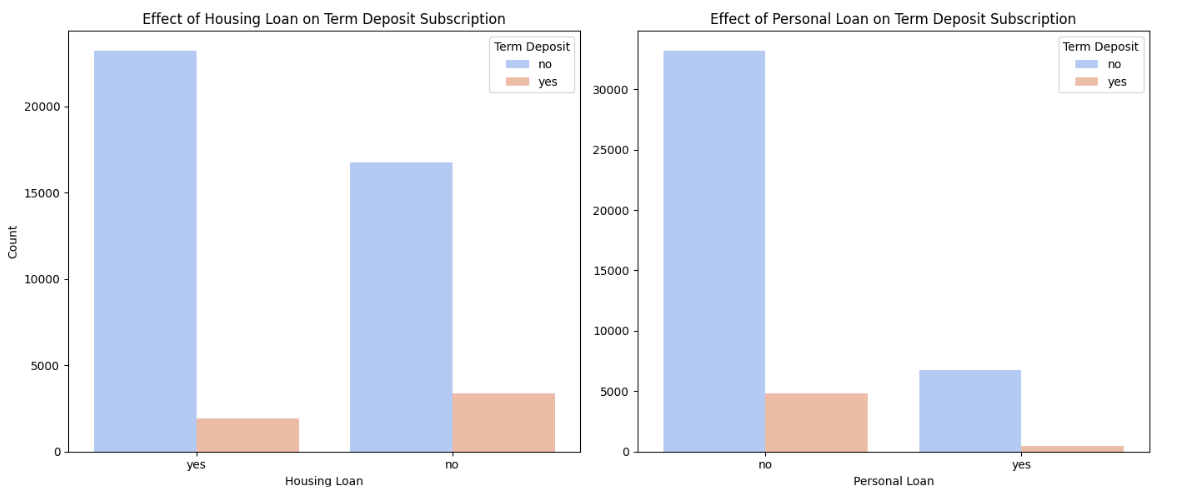


**6 - What is the distribution of the average yearly balance among the clients?**

* The Yearly Balance Distribution plot highlights the distribution of clients' average yearly balances. Due to the wide range of values, the plot is limited to showing balances between -5000 and 20000 for better visualization. The distribution is right-skewed, indicating that while most clients have relatively low balances, a few clients have significantly higher balances.

**7 -** How do housing and personal loans impact the likelihood of customers subscribing to a term deposit in a banking context?"

* Housing Loan Plot: The plot illustrates that customers without housing loans tend to subscribe to term deposits more frequently than those with housing loans, suggesting that financial commitments like housing loans may impact the propensity to invest in term deposits.
* Personal Loan Plot: Similarly, the analysis shows a higher subscription rate to term deposits among customers without personal loans, indicating that the absence of additional financial burdens like personal loans could correlate with an increased likelihood of term deposit subscriptions.



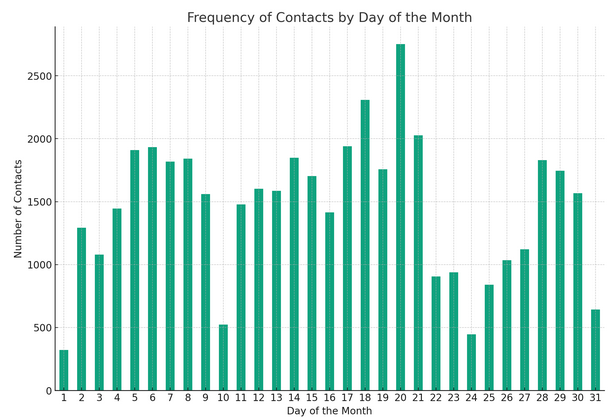
8 - What are the contact communication types used for the clients?

* A graph of different types of communication

  Description automatically generatedThe bar chart above displays the contact communication types used for clients. It shows the frequency of each type, revealing which methods are most used.

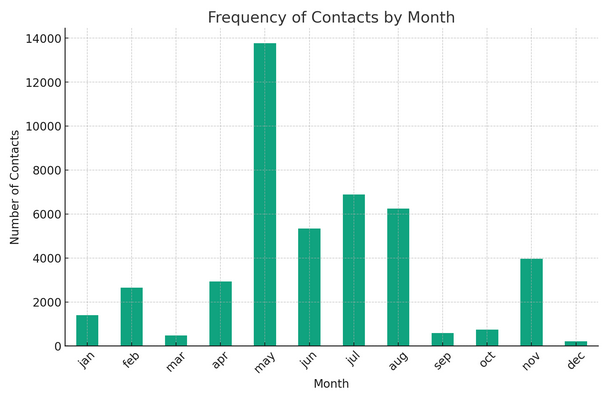
9 - On which days of the month are clients most frequently contacted?

* The bar chart above displays the frequency of contacts made on each day of the month to clients. It shows variation in the number of contacts, highlighting which days are most popular for client communication.



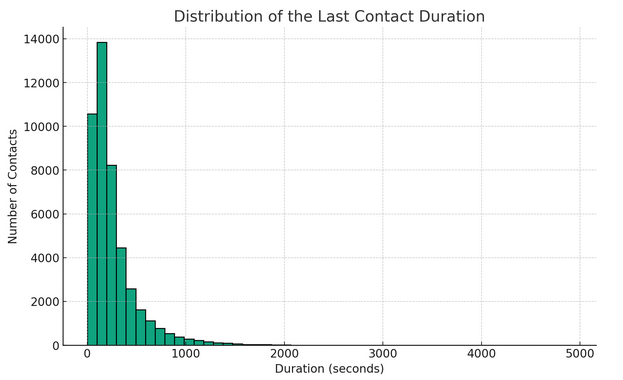
10- In which months are clients most frequently contacted?

* The Contact Month Distribution plot highlights the months in which clients are most frequently contacted. This reveals a clear pattern of contact frequency across different months, with certain months (like May) showing a higher number of contacts.



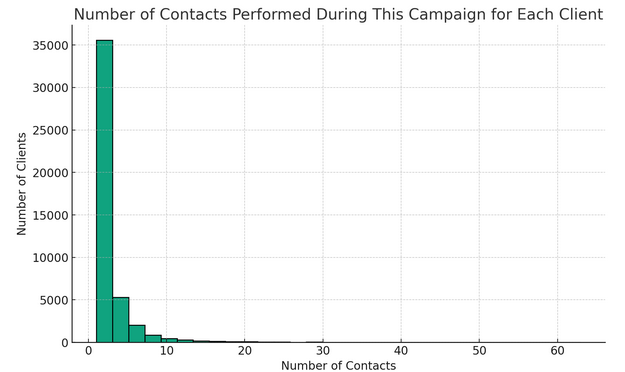
11 - What is the distribution of the last contact duration?

* Last Contact Duration Distribution plot shows the spread of the last contact duration in seconds. Most contacts are relatively short, but the distribution has a long tail, indicating that some conversations are much longer.

.

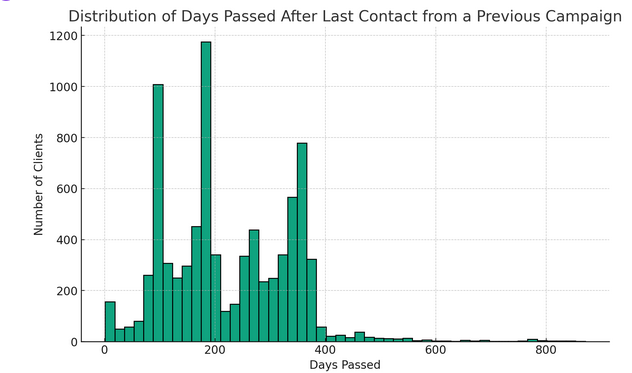
**12 -How many contacts are performed during this campaign for each client?**

* The Campaign Contacts Distribution plot provides insight into the number of contacts made to each client during the current campaign. It shows that most clients are contacted a few times, with the frequency of contacts quickly decreasing as the number of contacts increases.

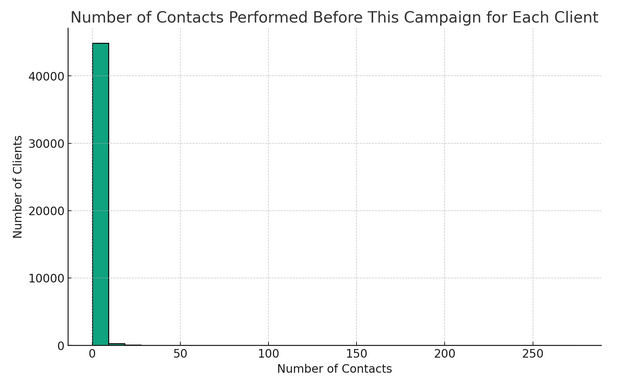


13 -What is the distribution of days that passed by after the client was last contacted from a previous campaign?

* The Days Since Last Contact Distribution plot shows the distribution of days that have passed since clients were last contacted from a previous campaign, excluding cases where pdays is -1 (indicating no previous contact). The distribution suggests that a significant number of contacts are made shortly after the previous contact, but there's a wide range overall.

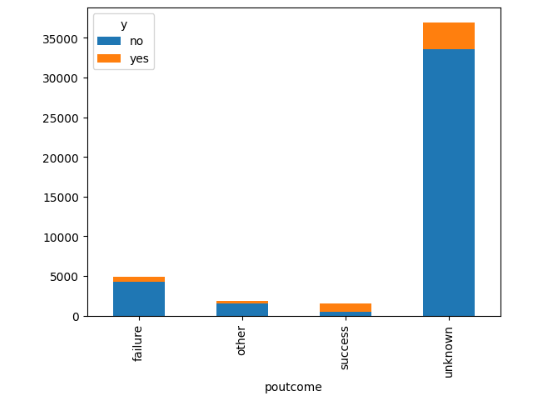


14 - How many contacts were performed before this campaign for each client?

* The Previous Contacts Distribution plot illustrates the number of contacts performed before this campaign. Most clients have not been contacted or have been contacted very few times before this campaign, as indicated by the peak at zero and the rapid drop-off

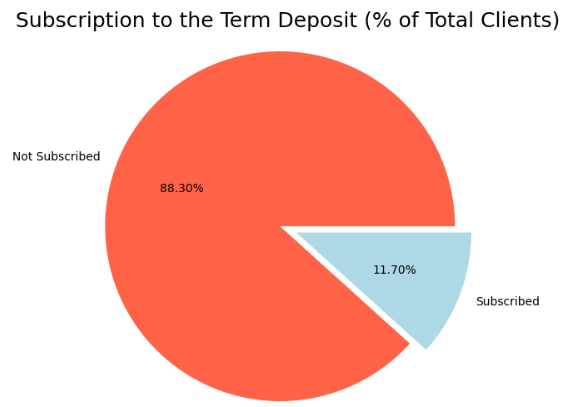
15 - **How does the outcome of previous marketing campaigns affect the likelihood of a client subscribing to a term deposit in the current campaign?**

* it's evident that clients with a successful outcome in the past campaign are significantly more likely to subscribe to a term deposit in the current campaign, indicating that past success is a strong predictor of current campaign success. Clients with an unknown past campaign outcome are the least likely to subscribe, which could suggest the importance of having a clear and positive history with clients to increase the likelihood of a subscription.



**16 - y - Research Question**: What is the proportion of customers subscribing to a term deposit compared to those who do not, within a banking context?

* The pie chart distinctly highlights the disparity between customers who have subscribed to a term deposit and those who have not, emphasizing the challenge banks face in converting a larger segment of their clientele into term deposit subscribers.



* **Find any missing values.**

As per the initial analysis with the Python code, there were no missing values reported in any of the columns of the dataset. This means that you won't need to perform any imputation or cleaning related to missing data.

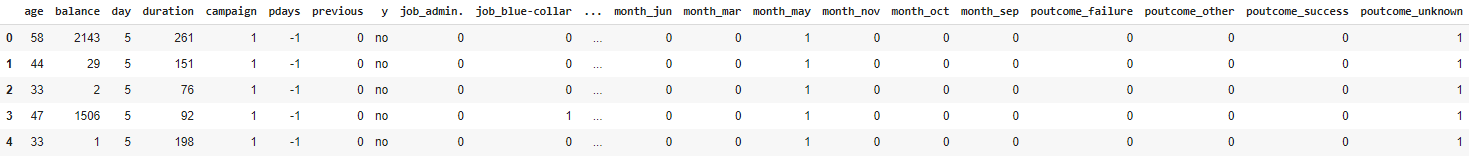
A list of words on a white background

Description automatically generated

* **One hot encoding**

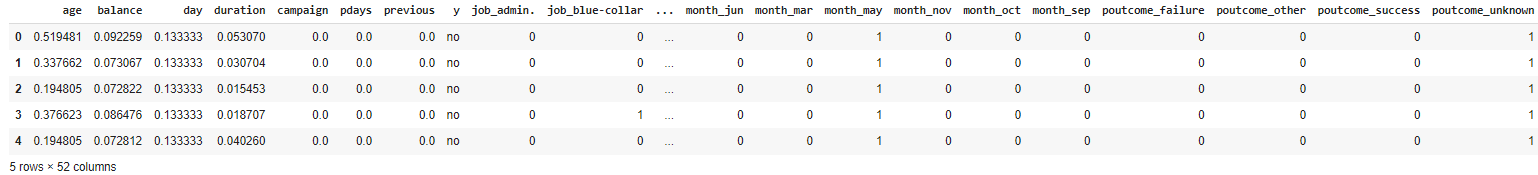
This technique converts categorical values into a format that can be provided to ML algorithms to do a better job in prediction. It works by creating a binary column for each category and returns a matrix with the results. Using the same color example, one-hot encoding would create three columns, one for each color. If the color is 'Red', for example, the column for 'Red' would have a 1, and the others 0.

Applying one-hot encoding to categorical variables indeed increases the number of features in your dataset. as you can see 52 columns instead of 17. so, I apply Feature Selection to keep only the most relevant features, reducing the dimensionality of my dataset.



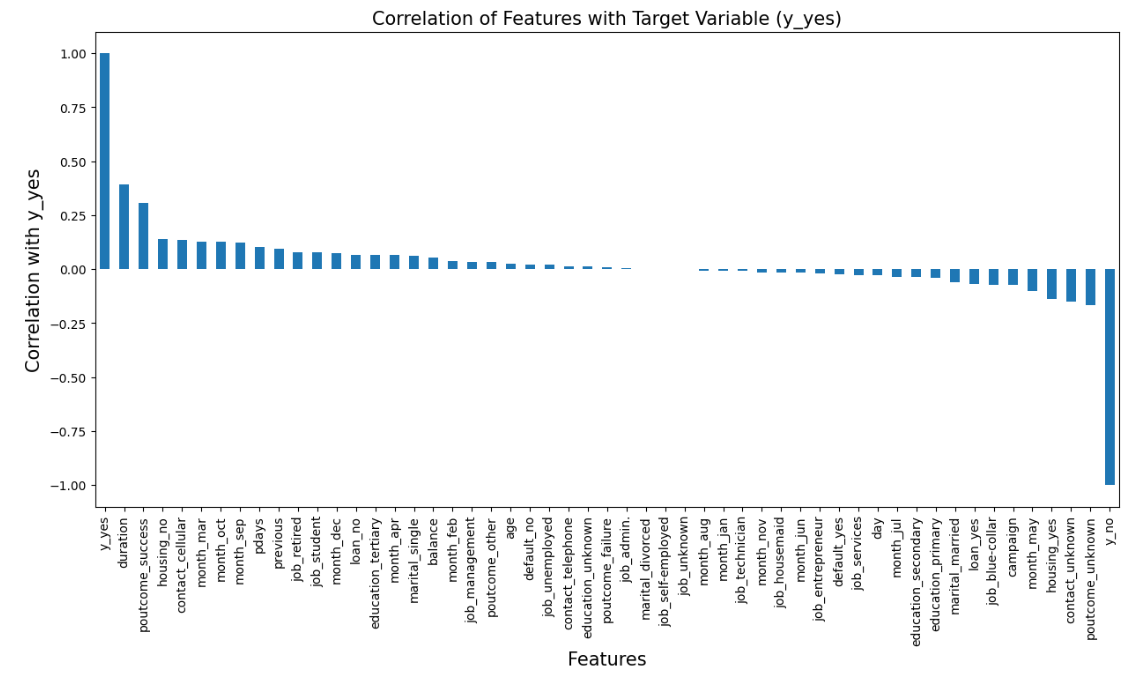
* **Normalize the dataset.**

The dataset has been normalized using MinMax scaling, which scales and translates each feature individually such that it is in the range [0,1]



* **Correlations between the variables**

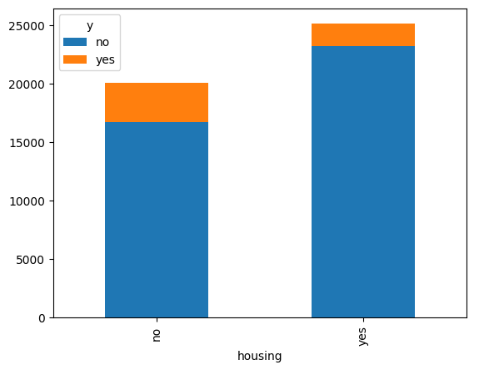
Categorical variables are converted to numerical format using techniques like one-hot encoding to enable correlation analysis with other numerical variables, as correlation coefficients require quantitative data for calculation. This transformation allows for the assessment of relationships and dependencies between categorical and continuous variables within a dataset.



1. As can be seen from the plot duration is a very important feature. This is the duration of last call with client.
2. If the call duration is more , there are higher chances of getting a yes from the client.
3. It has been sorted in descending order.
4. poutcome\_sucess, housing\_no and contact\_cellular are highly correlated with y\_yes
5. poutcome\_unknown, contact\_unknown and housing\_yes are highly correlated with y\_n

* **Exploring Attribute Correlations with Target Variables**

A graph with blue and orange bars

Description automatically generatedA graph of a number of blue and orange bars

Description automatically generated

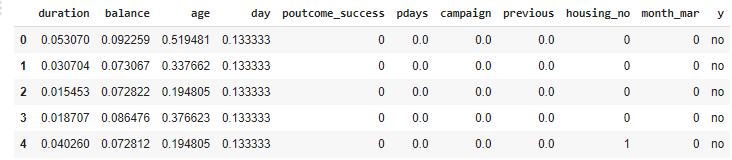
* **Splitting of data into training and testing set**

here we split the data into training and test set so that we can fit and evaluate a learning model.

****

* **Feature Selection**:

Selecting the most relevant features to use in model construction and reducing the dimensionality of my dataset.



A graph with blue bars

Description automatically generated

Here's a graph visualizing the top 10 most important features from the table, along with their importance scores. The graph clearly illustrates that "Duration" is the most significant predictor of the outcome, followed by "Balance, Age, and others, based on the analysis of the bank marketing dataset.

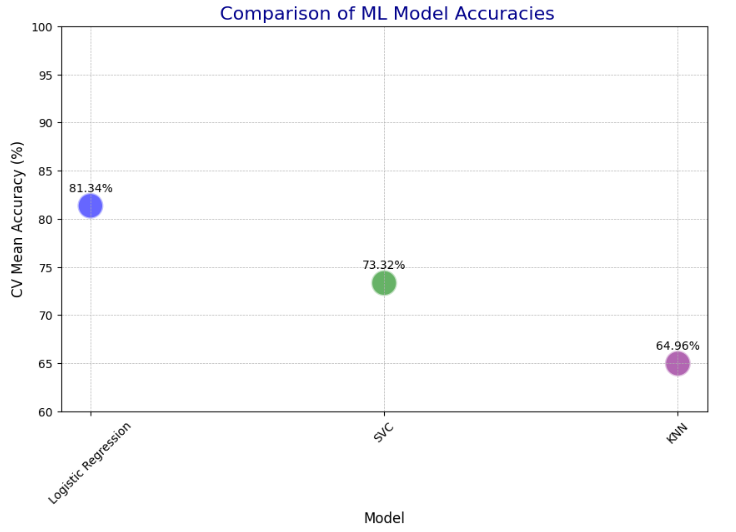
Evaluation

* Following the implementation of the data mining methodology used in this analysis, as previously stated, three algorithms have been extensively employed to identify patterns within the collected data. These methods have been utilized in a variety of forms to deliver performance insights and reports. The results derived from the application of these algorithms will be presented and analyzed at this juncture, with a more detailed explanation to follow in subsequent paragraphs. While classification models can be interpreted through multiple metrics, this research adopts a different approach for validation.
* Instead of using a confusion matrix to calculate misclassification errors, this study employs cross-validation techniques to assess the performance of the classification models. The use of cross-validation allows for a more robust verification of the model's effectiveness by testing it against multiple subsets of the data. This ensures that the accuracy reported is not based on a single train-test split but is rather an average performance across multiple folds of the data, providing a more balanced and generalized assessment of the model's predictive capabilities.
* The research thus relies on the comprehensive nature of cross-validation to calculate accuracy, which, in this context, is an aggregate measure reflecting the consistent quality of the model across various iterations of training and testing. The detailed analysis and outcomes of the three models, substantiated through cross-validation, will be elucidated in the sections that follow.

1. Logistic regression
2. SVC
3. KNN

Conclusion

Based on the results obtained from cross-validation, it appears that Logistic Regression achieved the highest mean accuracy score of 81.34%, followed by Support Vector Classifier (SVC) with a mean accuracy of 73.32% and K-Nearest Neighbors (KNN) models achieved mean accuracies of 64.96%, respectively. These results suggest that Logistic Regression might be more suitable for this classification task compared to SVC and KNN.



Limitation of the work

This study on the classification and regression analysis of the bank marketing dataset has provided valuable insights into the effectiveness of telemarketing strategies for selling long-term deposit accounts. However, like any research, it has its limitations:

1. **Dataset Constraints**: The dataset used is limited to a Portuguese bank's telemarketing campaigns, which may not represent the diversity of banking environments and customer behaviors worldwide. Additionally, the dataset's scope, covering only specific periods and types of communication, restricts the generalizability of the findings.
2. **Feature Selection and Engineering**: While significant effort was made to select relevant features and apply one-hot encoding for categorical variables, the process might have omitted potentially informative attributes or interactions between variables that could enhance the models' predictive capabilities.
3. **Model Selection and Performance**: The study focused on three algorithms: Logistic Regression, Support Vector Classifier (SVC), and K-Nearest Neighbors (KNN). Other advanced machine learning techniques or ensemble methods that might offer better performance were not explored.
4. **Imbalance in Target Variable**: The target variable's distribution, representing customers who subscribed to a term deposit versus those who did not, might exhibit imbalance. The study did not explicitly address strategies for dealing with imbalanced datasets, which could affect the models' ability to generalize to less represented classes.
5. **Evaluation Metrics**: The project primarily employed accuracy as the metric for model evaluation. Reliance on a single metric may overlook aspects of model performance, such as precision, recall, and the F1 score, especially important in imbalanced datasets.
6. **Temporal Dynamics**: The dataset captures a snapshot in time of customer interactions and outcomes. It does not account for temporal changes in customer behavior, economic conditions, or marketing strategies that could influence the effectiveness of telemarketing campaigns over time.
7. **External Validity**: The conclusions drawn are based on data from a specific banking context, limiting the applicability of the findings to other sectors or types of marketing campaigns.

GitHub link

https://github.com/mkalta1994/ML-Model-Comparisons-CV

Reference

1. [**https://en.wikipedia.org/wiki/Category:Classification\_algorithms**](https://en.wikipedia.org/wiki/Category:Classification_algorithms)
2. **E.W.T. Ngai, L. Xiu, and D. C. K. Chau, “Application of Data Mining Techniques in Customer Relationship Management: A Literature Review on Classification”, Expert Systems with Applications, vol.36, no. 2, Part 2, 2009, pp.2592-2602.**
3. **S. Mitra, S.K Pal, and P.Mitra, “Data Mining in Soft Computing Framework: A Survey”, Neural Networks, IEEE Transactions, vol. 13, no, 1, 2002, pp. 3-14.**
4. [**https://www.ibm.com/topics/decision-trees**](https://www.ibm.com/topics/decision-trees)
5. [**https://en.wikipedia.org/wiki/Weka\_(machine\_learning)**](https://en.wikipedia.org/wiki/Weka_(machine_learning))
6. **M. Purnachary,B. Srinivasa S P Kumar, Humera Shaziya“Performance Analysis of Bayes Classification Algorithms in WEKA Tool using Bank Marketing Dataset” International Journal of Engineering Research**
7. [**http://en.wikipedia.org/wiki/Direct\_marketing**](http://en.wikipedia.org/wiki/Direct_marketing) **Wikipedia has a tool to generate citations for particular articles related to direct marketing.**
8. [**https://d1wqtxts1xzle7.cloudfront.net/59954293/IRJET-V6I30820190707-55305-1e47yyf-libre.pdf?1562564671=&response-content-disposition=inline%3B+filename%3DIRJET\_Evaluation\_of\_Classification\_Algor.pdf&Expires=1708381053&Signature=H9AbSILP9x07G6~9imCdFVw9cdgHTnzkeUUYnT9HDgi584Phc4MiAU4s~w9ABaNAG5sLBg2vMGO7uCdQEhxuzU1ufL7FUPTixvEoCRLV-OnSatz2Cd8A8oR2lbCAr0iHcIFku1jtZld0DN42lU7p8zWnGQ8BlrOGh6F48rX6bzLu7~G4RmzAVY6PNeLqJq3eTUY8vICQP-lteitYhAitszXXs-U9JecBnmBTpodjyiv9fgFvQQbWaA4a3Dgb10sAto8w27oco-aQv5Vz9AHCkyxANa5m6I0drveNQCDxpWqIl8LtTyJZ7Gw5ws5ktp3CUSjGp9HxznXMqiMKeE1pZA\_\_&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA**](https://d1wqtxts1xzle7.cloudfront.net/59954293/IRJET-V6I30820190707-55305-1e47yyf-libre.pdf?1562564671=&response-content-disposition=inline%3B+filename%3DIRJET_Evaluation_of_Classification_Algor.pdf&Expires=1708381053&Signature=H9AbSILP9x07G6~9imCdFVw9cdgHTnzkeUUYnT9HDgi584Phc4MiAU4s~w9ABaNAG5sLBg2vMGO7uCdQEhxuzU1ufL7FUPTixvEoCRLV-OnSatz2Cd8A8oR2lbCAr0iHcIFku1jtZld0DN42lU7p8zWnGQ8BlrOGh6F48rX6bzLu7~G4RmzAVY6PNeLqJq3eTUY8vICQP-lteitYhAitszXXs-U9JecBnmBTpodjyiv9fgFvQQbWaA4a3Dgb10sAto8w27oco-aQv5Vz9AHCkyxANa5m6I0drveNQCDxpWqIl8LtTyJZ7Gw5ws5ktp3CUSjGp9HxznXMqiMKeE1pZA__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA)
9. [**https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=3b09589fc06dc60c5594da1708da276d0e22287b**](https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=3b09589fc06dc60c5594da1708da276d0e22287b)
10. [**https://www.researchgate.net/profile/Hany-Elsalamony/publication/263054095\_Bank\_Direct\_Marketing\_Analysis\_of\_Data\_Mining\_Techniques/links/5519357f0cf2d241f355c920/Bank-Direct-Marketing-Analysis-of-Data-Mining-Techniques.pdf**](https://www.researchgate.net/profile/Hany-Elsalamony/publication/263054095_Bank_Direct_Marketing_Analysis_of_Data_Mining_Techniques/links/5519357f0cf2d241f355c920/Bank-Direct-Marketing-Analysis-of-Data-Mining-Techniques.pdf)
11. [**https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=0e85310bc60dfcf2cbe9c8d303129adb20c8ef23**](https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=0e85310bc60dfcf2cbe9c8d303129adb20c8ef23)
12. [**https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=3b09589fc06dc60c5594da1708da276d0e22287b**](https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=3b09589fc06dc60c5594da1708da276d0e22287b)
13. **https://norma.ncirl.ie/4574/1/jamiuolalekanoni.pdf**