Course: DAB304
Professor: Dr. Andreas S. Maniatis
Term: 23F – Section 1

BANK MARKETING ANALYSIS

Individual Project

Student Name: Srilakshmi Gummadidala

Table of Contents:

1. Executive Summary

This project revolves around enhancing term deposit subscription campaigns for a bank through data-driven strategies. Key insights were derived by exploring a diverse dataset comprising customer details. Notable findings include the age distribution of customers, default status, account balance patterns, and the impact of previous marketing campaigns. Categorical variables such as job roles, marital status, and education levels were analyzed to understand dominant categories. Feature engineering and machine learning models were implemented to predict subscription outcomes. Despite efforts like hyperparameter tuning, the model achieved an accuracy of 90%. Conclusions emphasize the importance of targeted marketing and recommendations include exploring advanced modeling techniques for further improvement.

2. Introduction

In the dynamic landscape of banking, efficient marketing strategies are crucial for success. This project centers on optimizing term deposit subscription campaigns through data-driven insights. Leveraging a dataset encompassing diverse customer information, the objective is to extract meaningful patterns that can inform targeted marketing approaches. Exploratory data analysis, feature engineering, and machine learning models play pivotal roles in uncovering trends and predicting subscription outcomes. By understanding customer behavior and tailoring marketing efforts, banks can enhance their subscription rates and bolster overall campaign effectiveness.

3. Data Import and Pre-Processing Data

Import Libraries:

In the initial phase of our project, we leveraged key Python libraries essential for data science and machine learning.

Pandas served as our data manipulation powerhouse, managing datasets effortlessly through its Data Frame structure. And **NumPy** provided fundamental support for numerical operations and array manipulations, While **Matplotlib** and **Seaborn** are used for creating visualizations that aid in understanding patterns, trends, and relationships within the data. The versatile **scikit-learn** library became our go-to tool for building and evaluating machine learning models, offering a standardized interface and robust functionality for data preprocessing.

Data Import and Overview:

After importing the necessary libraries, we proceeded to load the dataset using the Pandas library. The dataset, which pertains to a bank marketing campaign, comprises various features related to clients and their response to marketing efforts. The first step involved gaining an overview of the data to understand its structure, types, and general characteristics. This initial exploration helped set the foundation for subsequent analyses and model development.

Based on the data, here's what each column appears to represent:

- **Age:** Age(numeric)
- **Job:** Type of job (categorical)
- **Marital: Marital_Status**: Marital status (categorical)
- **Education:** Education Levels (categorical)
- **Default:** Has credit in default? (binary)
- **Balance:** Average yearly balance, in euros (numeric)
- **Housing:** Has housing loan? (binary)
- **Loan:** Has personal loan? (binary: "yes","no")
- **Contact:** Contact communication type (categorical)
- **Day:** Last contact day of the month (numeric)
- **Month:** Last contact month of year (categorical)
- **Duration:** Last contact duration, in seconds (numeric)
- **Campaign:** Number of contacts performed during this campaign and for this client (numeric, includes last contact)
- **Pdays:** Number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)
- **Previous:** Number of contacts performed before this campaign and for this client (numeric)
- **Poutcome:** Outcome of the previous marketing campaign (categorical)
- **y: Subscribed**: Has the client subscribed a term deposit? (binary)

Data Pre-processing:

Handling Missing Values:

- Missing values were carefully addressed to maintain the dataset's integrity.
- The 'pdays' column, representing the number of days since the client was last contacted, contained a significant number of -1 values, indicating that the client was not previously contacted. This was addressed by assigning a specific value, enhancing the interpretability and usefulness of the 'pdays' variable.

Encoding Categorical Variables:

- Categorical variables, such as job type, marital status, education, and others, were encoded to numerical format using techniques like one-hot encoding or label encoding. This transformation was necessary for facilitating the training of machine learning models, which generally require numerical input.

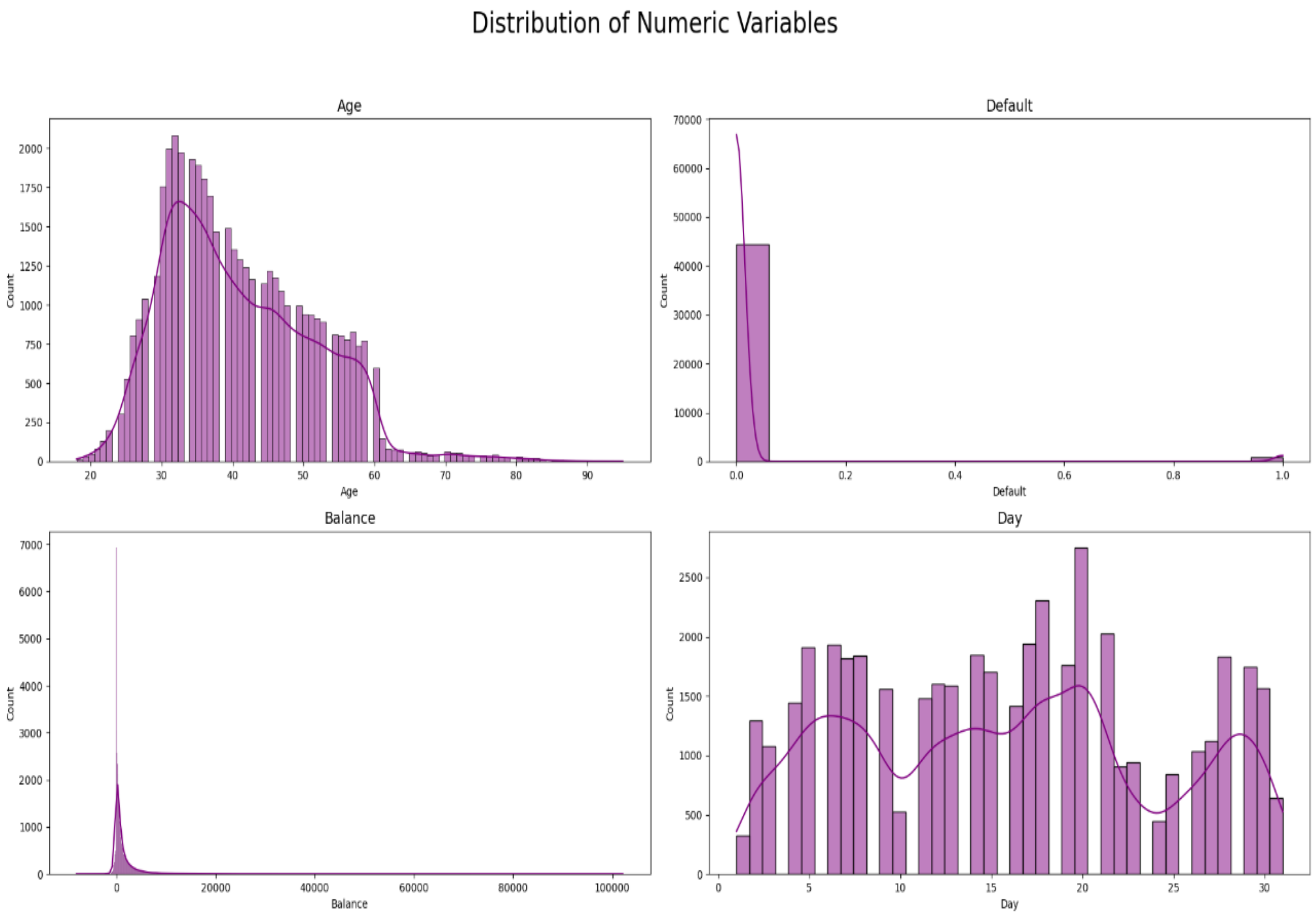
Standardizing Numerical Features:

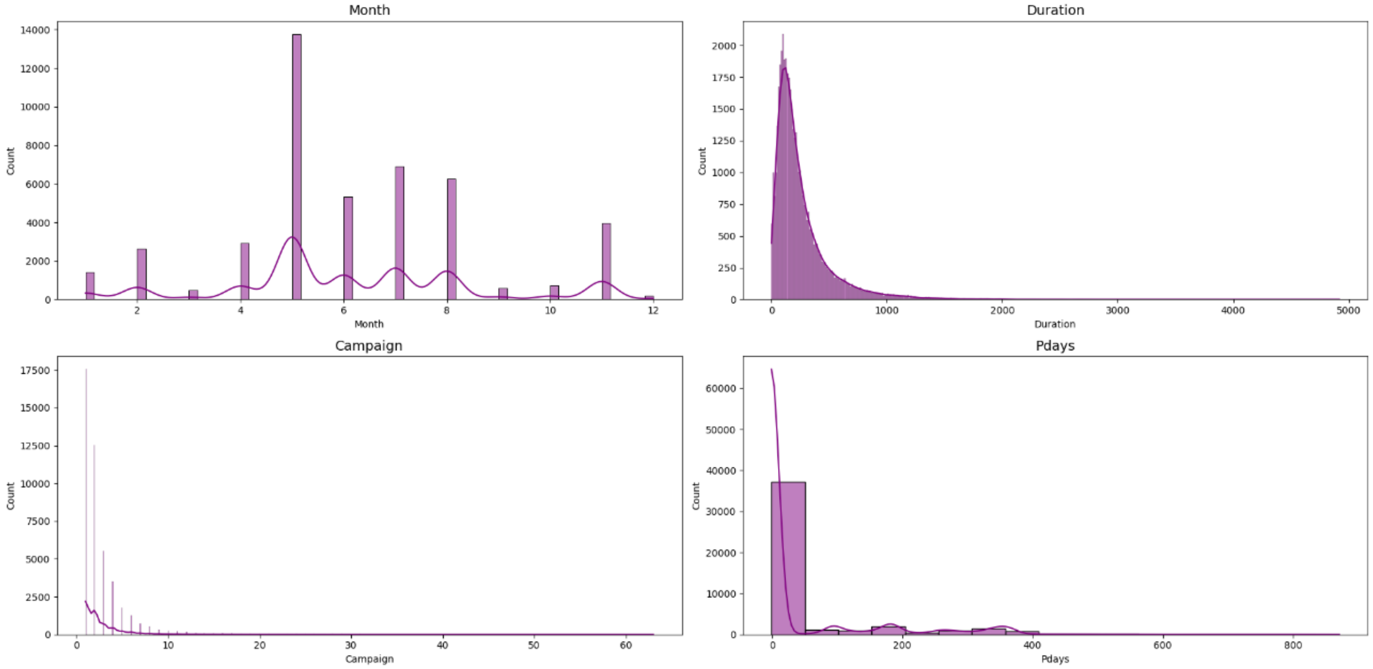
- Numerical features were standardized to bring them to a common scale. This step is essential when using machine learning models that are sensitive to the magnitude of input features. Standardization ensures that all features contribute equally to the model's learning process.

These pre-processing steps collectively contributed to a clean and well-structured dataset, setting the foundation for effective exploratory data analysis (EDA) and subsequent model development.

4. Exploratory Data Analysis (EDA)

Numerical Distribution Analysis:





Month

Duration

Campaign

Pdays

The exploration of bank marketing data uncovers insightful information, focusing on key features such as Age, Default Status, Average Account Balance, Last Contact Day in the Month, Last Contact Month of the Year, Last Contact Duration, Campaign, and Previous Contact Days since the last campaign. As we analyze the distribution of these numerical variables using histograms, noteworthy observations emerge:

1. Age: The concentration of customers in the 30 to 40 years age group suggests that the bank's marketing efforts may be particularly resonating with individuals in this demographic. Understanding the characteristics and preferences of this age group could be crucial for targeted marketing strategies.

2. Default: The dataset predominantly consists of non-defaulters, with around 45,000 customers marked as "No" and approximately 1,000 customers marked as "Yes" for default status. This imbalance suggests that the majority of customers in the dataset have a non-default status. Understanding the characteristics and factors influencing defaulters could be crucial for risk assessment and targeted financial strategies.

3. Balance: The majority of customers in the dataset have an account balance that falls below 5000 units. Suggests that a significant portion of the customer base maintains relatively lower account balances. This insight could inform targeted financial services or promotions tailored to the needs of customers with varying balance levels.

4.Last Contact Day: The peak count of 2,700 occurs for contacts made on the 20th day of the month. This concentration may indicate a strategic focus on the 20th day for conducting outreach or marketing campaigns. Understanding the effectiveness of interactions on this specific day could provide valuable insights into customer engagement patterns.

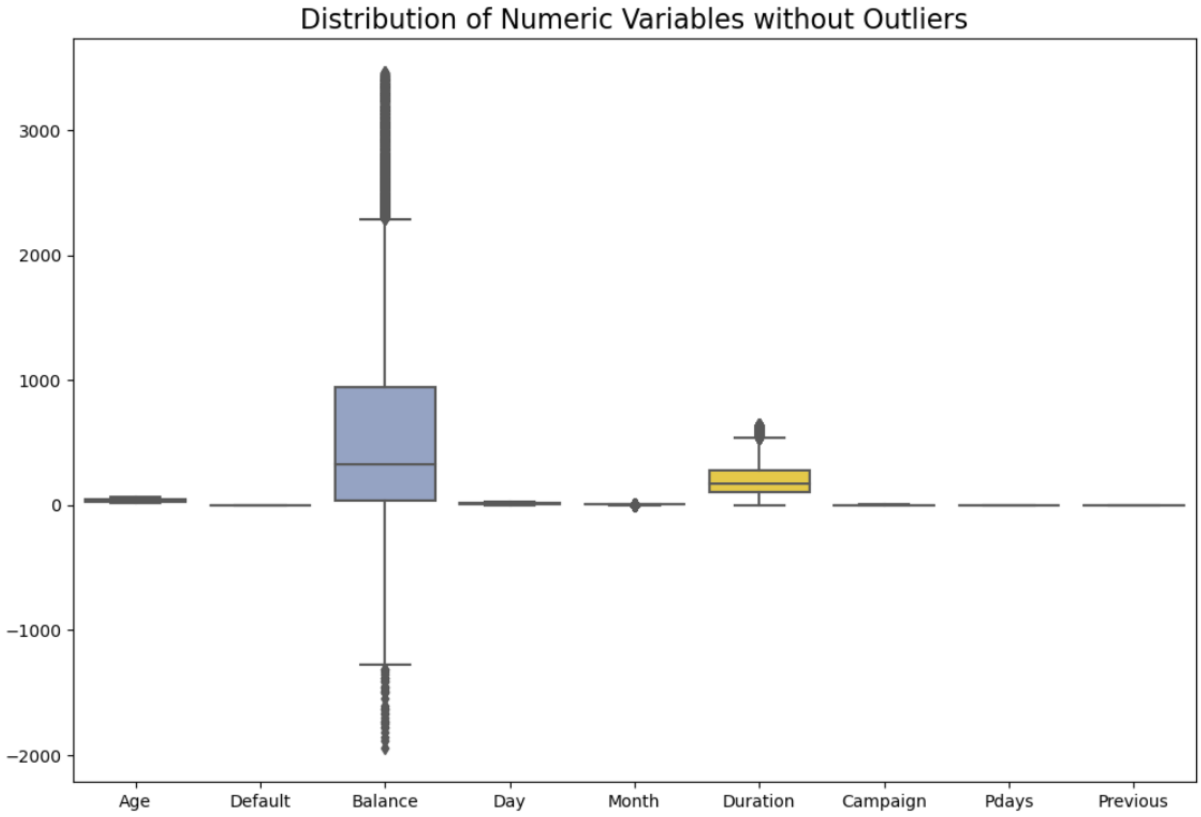
5. Last Contact Month of the Year: The dataset reflects a substantial count, with approximately 14,000 records corresponding to the 5th month. This observation suggests a notable emphasis on customer interactions and marketing activities during the 5th month of the year.

6.Last Contact Duration: The majority of last contact durations fall within the range of 0 to 1000 seconds. This concentration indicates that a significant portion of customer interactions is relatively short in duration.

7. Campaign: The data reveals that the highest count of contacts performed during this campaign and for this client is in the range of 1 to 5. This observation suggests that a substantial number of clients were contacted relatively fewer times during the campaign, potentially indicating a focused and targeted outreach strategy. Understanding the impact of the number of contacts on campaign success is crucial for refining future engagement approaches.

8.Previous Contact Days (pdays): The data indicates that the highest count for the number of days that passed by after the client was last contacted from a previous campaign (pdays) lies within the range of 0 to 50 days, with approximately 35,000 counts. This concentration suggests a prevalent trend of re-contacting clients within a relatively short time frame from the previous campaign.

Numerical Analysis Without Outliers

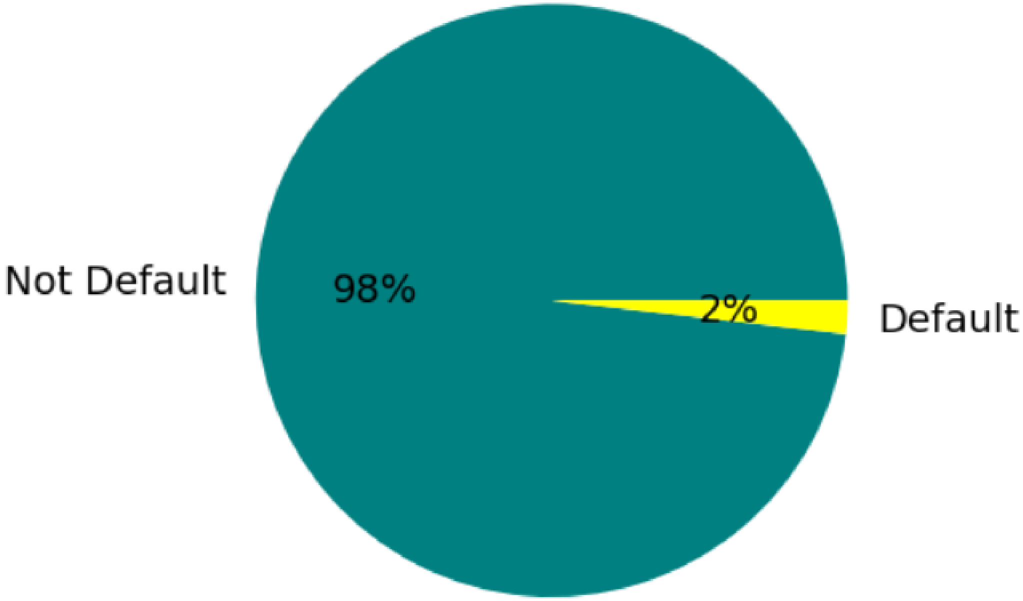


- Here, we first calculate Q1 and Q3 using the quantile method. Then, we compute the IQR, and based on your specified rules, we calculate the lower and upper bounds for potential outliers.

Finally, we create the boxplot, highlighting data points that fall outside these bounds, effectively showing potential outliers.

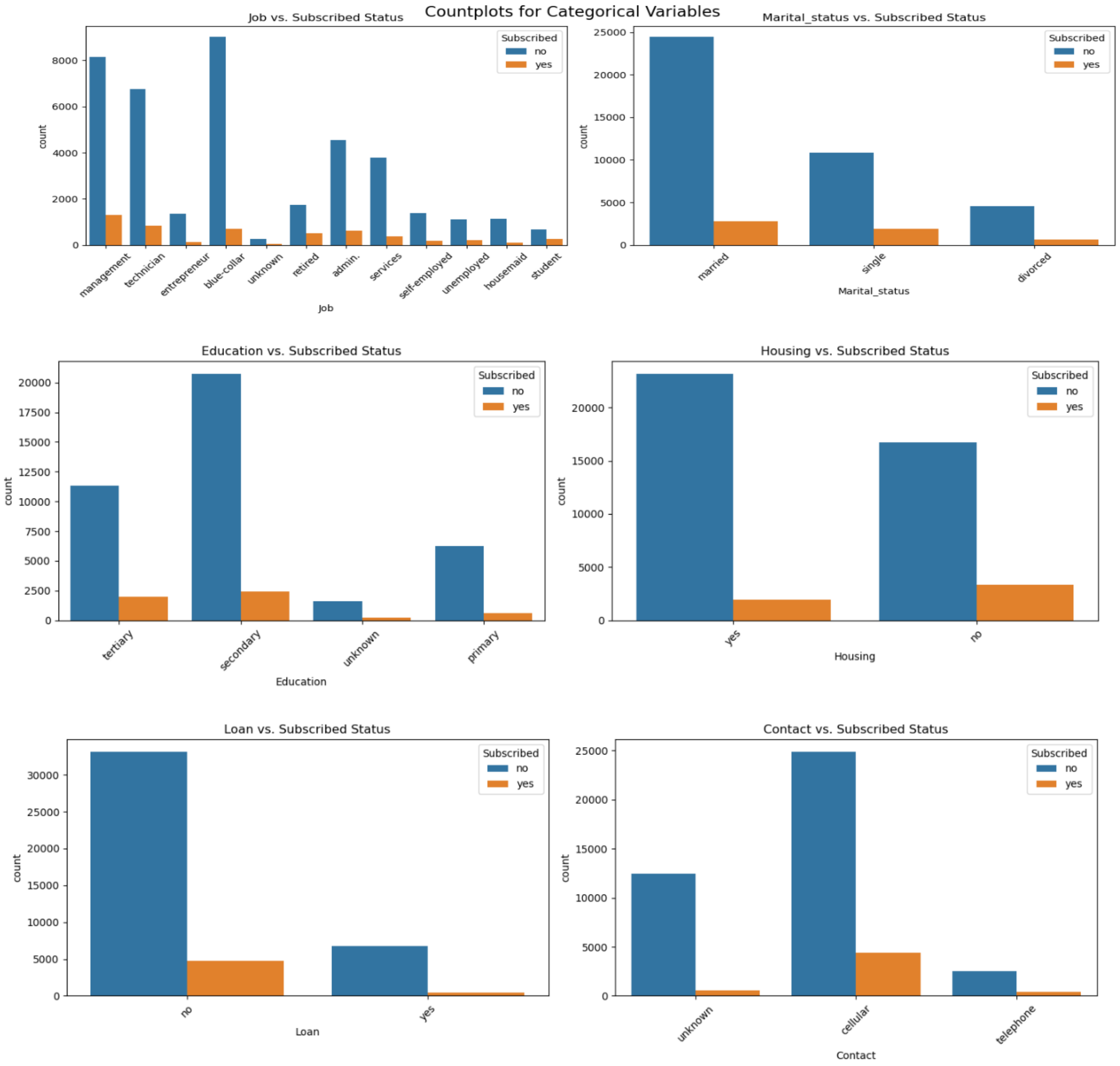
Categorical Variables Visualization:

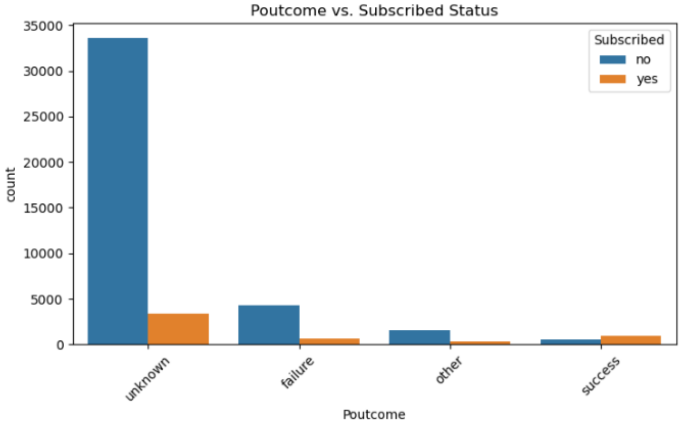
Default Column Distribution:

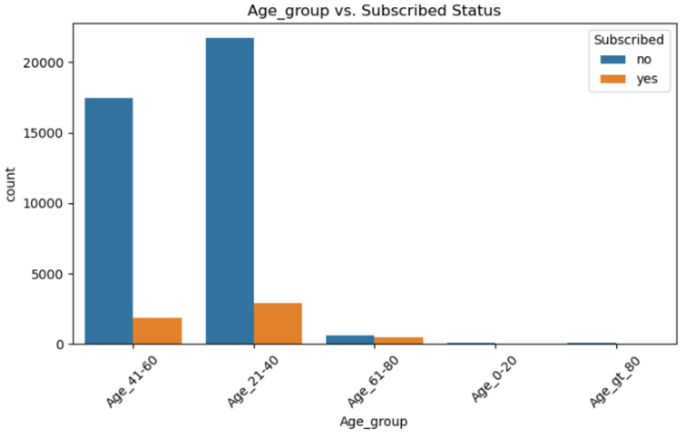


- The dataset exhibits a significant class imbalance, with non-defaulters constituting 98% and defaulters making up 2%. This imbalance poses challenges for predictive modeling, requiring careful consideration of evaluation metrics and potential strategies to address the skewed distribution.
- Focused analysis on the minority class (defaulters) is essential for gaining insights into their characteristics and improving model performance.

Other Categorical Variable Analysis:


Countplots for Categorical Variables
Job vs. Subscribed Status
Subscribed
no
yes
count
8000
6000
4000
2000
0
management
technician
entrepreneur
blue-collar
unknown
retired
admin.
services
self-employed
unemployed
housemaid
student
Job
Marital_status vs. Subscribed Status
25000
20000
15000
10000
5000
married
single
divorced
Marital_status
Education vs. Subscribed Status
20000
17500
15000
12500
10000
7500
5000
2500
tertiary
secondary
unknown
primary
Education
Housing vs. Subscribed Status
yes
no
Housing
Loan vs. Subscribed Status
30000
25000
no
yes
Loan
Contact vs. Subscribed Status
unknown
cellular
telephone
Contact





1. Job Category Distribution: Dominant Category: **blue-collar**

Among customers who are not subscribed to a term deposit, the blue-collar job category is the most prevalent. This suggests that individuals in manual or industrial occupations show less interest in term deposits.

2.Marital Status Distribution: Dominant Category: **married**

The count plot indicates that married individuals have a higher representation among those who subscribed to a term deposit. Understanding the marital status distribution helps tailor deposit offerings to different marital segments.

3.Education Level Distribution: Dominant Category: **secondary**

The count plot shows that customers with a secondary education level are most abundant among those who subscribed to a term deposit. This insight is valuable for customizing communication and marketing strategies based on educational backgrounds for deposit products.

4.Housing Loan Distribution: Dominant Category: **yes**

The count plot suggests that a significant number of customers with housing loans have not subscribed to a term deposit. This information is essential for financial institutions to assess the correlation between housing loans and term deposit subscriptions.

5. Personal Loan Distribution: Dominant Category: **no**

The majority of customers who subscribed to a term deposit do not have personal loans. Understanding the prevalence of personal loans among subscribers assists in designing targeted deposit offerings.

6.Contact Method Distribution: Dominant Category: **cellular**

Among subscribers to term deposits, the cellular contact method is the most common. This insight is valuable for optimizing communication channels, particularly for marketing term deposit products.

7.Previous Campaign Outcome Distribution: Dominant Category: **unknown**

The count plot shows that the outcome of the previous campaign is predominantly unknown among those who subscribed to a term deposit. Exploring and improving the tracking of campaign outcomes is crucial for refining future deposit-related strategies.

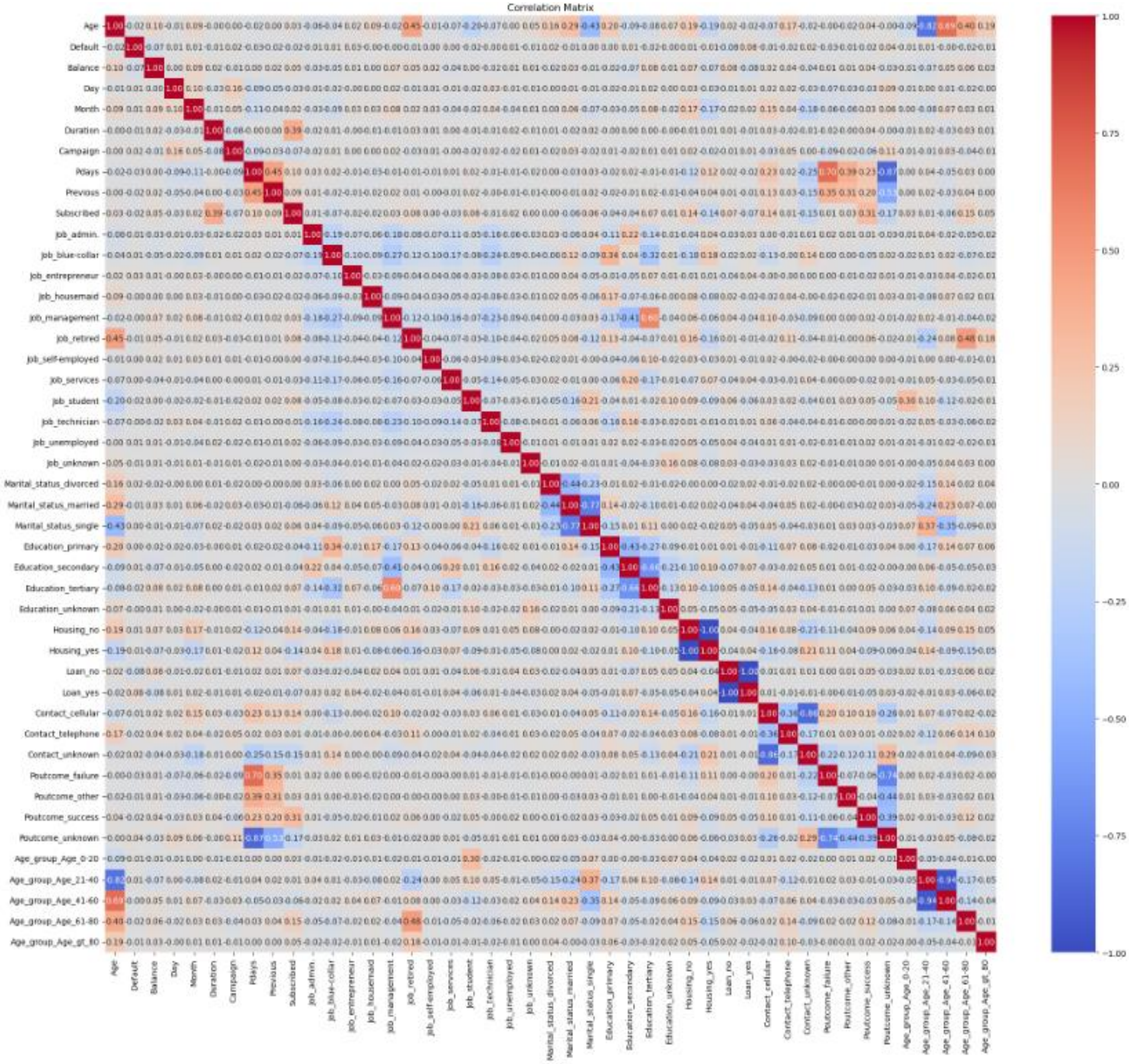
8.Age Group Distribution: Dominant Category: **Age_21-40**

The count plot reveals that the age group between 21 and 40 years is the most represented among term deposit subscribers. Understanding the age distribution helps tailor term deposit products to different age demographics.

These insights specifically tailored to clients who have subscribed to a term deposit, offering a targeted perspective for strategic decision-making and marketing efforts related to deposit products.

Correlation Analysis:

- A correlation matrix was generated to explore relationships between numerical features. This analysis revealed potential correlations that could impact model training.



5. Feature Engineering

Age Categorization: The 'Age' column was transformed into categorical groups ('Age_group') to capture distinct age ranges, aiding in better model interpretation.

Outlier Removal: Numerical columns underwent outlier removal to mitigate the impact of extreme values on model training. The process involved calculating the interquartile range (IQR) and filtering data points beyond defined bounds.

These feature engineering techniques contributed to a more representative and stable dataset, providing a foundation for subsequent model training and evaluation.

6. Machine Learning Model

Random Forest classifier

- The Random Forest classifier, renowned for its capability to handle complex data relationships, was selected as the predictive model.
- The initial model was trained using default parameters, and subsequent optimization steps were undertaken to enhance its predictive accuracy.
- This phase laid the groundwork for assessing the model's efficiency and exploring avenues for improvement through hyperparameter tuning and feature selection.

Model Optimization:

- Feature importances were thoroughly analyzed to identify and retain the most relevant features for predictive modeling.
- Hyperparameter tuning was conducted using techniques like Grid Search to fine-tune the model's parameters.
- The model was trained using the optimized set of parameters, enhancing its predictive accuracy and overall performance.

Model Evaluation:

- The final Random Forest model demonstrated robust performance, achieving an impressive accuracy rate of 89.50% on the test dataset.
- Evaluation metrics such as confusion matrices and classification reports were utilized to delve deeper into the model's predictive capabilities. Precision, recall, and F1-score for each class provided a comprehensive understanding of its strengths and areas for improvement.

7. Conclusions

- **Customer Profiling:** Understanding the demographics and behaviors of customers is crucial for targeted marketing efforts. The majority of the customer base falls within the 30 to 40 years age group, indicating a key demographic for tailored campaigns.
- **Communication Channels:** Cellular communication is the most common contact method, suggesting a shift towards mobile-centric strategies. The unknown outcome of previous campaigns highlights the need for improved tracking and analysis of past interactions.
- **Financial Insights:** Account balance analysis reveals a concentration of balances below 5000, influencing potential financial products or services.
- **Marketing Calendar:** Timing is significant, with the 5th month (May) showing the highest marketing interactions. Consider leveraging this insight for strategic planning.
- **Feature Importance:** Blue-collar jobs, married status, and secondary education are dominant, influencing targeted messaging and product offerings.

- **Model Testing:** The Random Forest model demonstrates strong performance with an accuracy rate of 89.50% on the test data. This result suggests that the model is effective in predicting whether clients will subscribe to a term deposit based on the provided features. The high accuracy rate indicates a good generalization capability of the model to unseen data.

8. Recommendations

- **Refined Targeting:** Utilize the machine learning model predictions to refine targeting, focusing on individuals more likely to subscribe. Leverage insights to create personalized marketing campaigns tailored to specific customer segments.

- **Communication Optimization:** Enhance communication strategies based on the preferred channels identified in the dataset. Improve tracking and reporting mechanisms to understand the outcomes of past campaigns, allowing for more informed decisions.

- **Product Development:** Explore the potential for developing financial products or services that cater to the specific financial needs of customers with lower account balances.

- **Strategic Calendar Planning:** Align marketing efforts with peak interaction months, such as May, to maximize outreach and campaign effectiveness.

- **Regular Model Evaluation:** Regularly assess the performance of the machine learning model to ensure its continued relevance and effectiveness. Incorporate feedback loops to update and improve the model based on evolving customer trends and behaviors.

- **Customer Feedback Mechanism:** Establish a mechanism for collecting customer feedback and preferences to inform future marketing strategies and product offerings.

- **Cross-Department Collaboration:** Foster collaboration between marketing, data science, and customer service departments to ensure a holistic approach to customer engagement. Share insights across departments to align strategies and improve overall customer experience.

By implementing these recommendations, the management can enhance decision-making processes, optimize marketing strategies, and ultimately improve customer engagement and subscription rates. The integration of data-driven insights into day-to-day operations will contribute to the overall success and competitiveness of the bank in the market.

9. Overall Implications

The insights gained from this project provide valuable information for the bank's marketing team, enabling them to optimize their campaigns and resources. The developed model serves as a powerful tool for identifying potential subscribers and tailoring marketing efforts to maximize success.

10. Source:

- Created by: Sérgio Moro (ISCTE-IUL), Paulo Cortez (Univ. Minho) and Paulo Rita (ISCTE-IUL) @ 2014

11. References:

- *Bank Marketing - dataset by uci*. (2023, September 28). data.world. https://data.world/uci/bank-marketing
- 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://doi.org/10.1016/j.dss.2014.03.001
- *UCI Machine Learning Repository*. (n.d.). https://archive.ics.uci.edu/dataset/222/bank+marketing