Business Question

How can **NextGen Financial Solutions** increase the subscription rate for its newly launched term deposit product? This project seeks to determine which customer segments are most likely to subscribe to the term deposit and design targeted marketing campaigns to increase conversions.

Business Problem

NextGen Financial Solutions, a digital bank, recently introduced a term deposit product. Despite robust marketing efforts, subscription rates have been lower than anticipated. The goal is to leverage customer data to identify key characteristics of likely subscribers, segment the target audience, and create two tailored campaigns that efficiently utilize marketing resources.

Data Supporting the Addressing of This Question:

The analysis uses customer demographic and behavioral data, which includes variables like age, job type, marital status, education level, account balance, housing and personal loan status, contact details, and responses to previous marketing campaigns. The dataset enables us to assess each customer's likelihood of subscribing, based on personal and historical engagement attributes variables include:

1. Demographic Variables:

- Age: Customer's age.
- Job: Type of job held by the customer.
- Marital Status: Relationship status.
- Education: Highest level of education achieved.

2. Financial Health Indicators:

- Balance: Customer’s bank balance, indicating financial health.
- Loan: Indicates if the customer has a loan.
- Housing Loan Status: Indicates if the customer has a housing loan.
- Default: Indicates if the customer has credit in default ("yes" or "no").

3. Previous Campaign Insights:

- Campaign: Number of contacts made during this campaign.

- Pdays: Days since the customer was last contacted in a previous campaign (999 indicates no prior contact).
- Previous: Number of prior contacts before this campaign.
- Poutcome: Outcome of the last marketing campaign.

4. **Contact Information:**
    - Contact: Type of communication used ("unknown", "telephone", "cellular").
    - Day: Last contact day of the month.
    - Month: Last contact month of the year.
    - Duration: Duration of the last contact in seconds.
5. **Deposit (Target Variable):**
    - Deposit: Indicates whether the customer subscribed to the term deposit ("yes" or "no")

Data Source: https://www.kaggle.com/datasets/janiobachmann/bank-marketing-dataset

Data Visualization

Outliers:

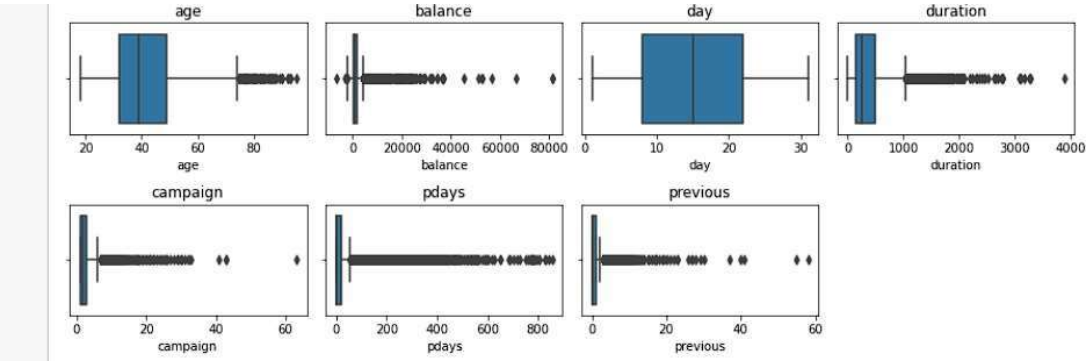


Figure 1

```
Outlier count and percentage by IQR method:
age: 171 outliers (1.53%)
balance: 1055 outliers (9.45%)
day: 0 outliers (0.00%)
duration: 636 outliers (5.70%)
campaign: 601 outliers (5.38%)
pdays: 2750 outliers (24.64%)
previous: 1258 outliers (11.27%)
```

Figure 2

Initial Findings from Outliers:

The outlier analysis shows that variables like **pdays (24.64%)** and **previous (11.27%)** have notable outlier counts, while others like **age (1.53%)** and **duration (5.70%)** have minimal outliers.

Although the outlier percentages are relatively low, we've retained them as they maycapture essential customer behaviors and financial states relevant to deposit subscription patterns. This retention aids in distinguishing customer segments with varied deposit probabilities and responses to previous campaigns, as shown in the clustering results.

Visualization for Relationships with target variable

1. Box Plots

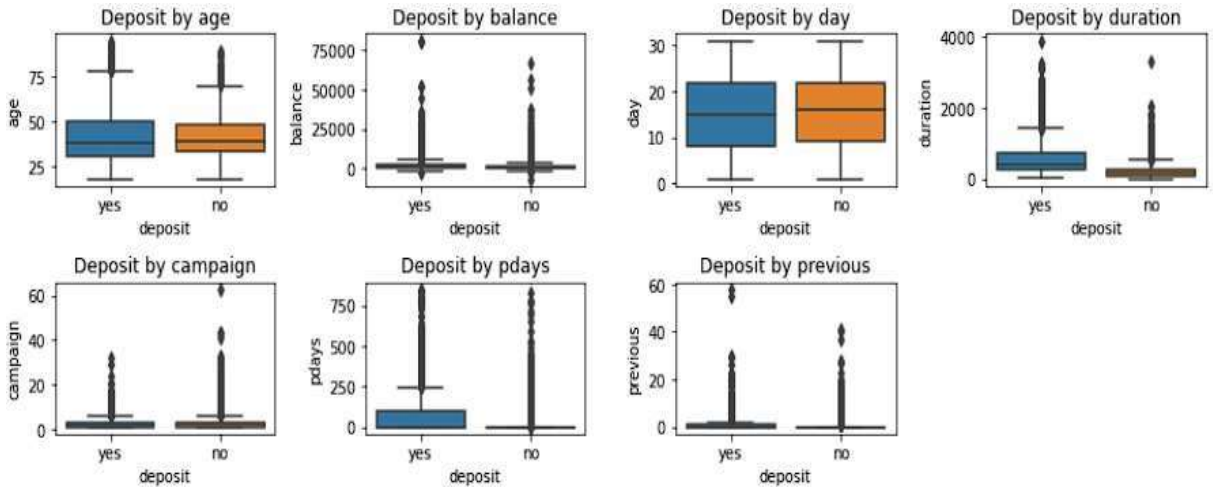


Figure 3

Initial Findings:

1. **Age**: Customers who made deposits are generally younger, but age doesn't show astrong separation.
2. **Balance**: Higher balances are slightly more common among depositors, but there is a significant overlap.
3. **Duration**: Longer contact durations are more common among those who makedeposits.
4. **Campaign**: Fewer campaign contacts seem to correlate with deposits, while non-depositors experienced more repeated contacts.
5. **Pdays**: Depositors often have lower **pdays** values, suggesting recent contact may encourage deposits.
6. **Previous**: Non-depositors tend to have a higher number of previous contacts, possibly due to unsuccessful attempts.

2. Histogram

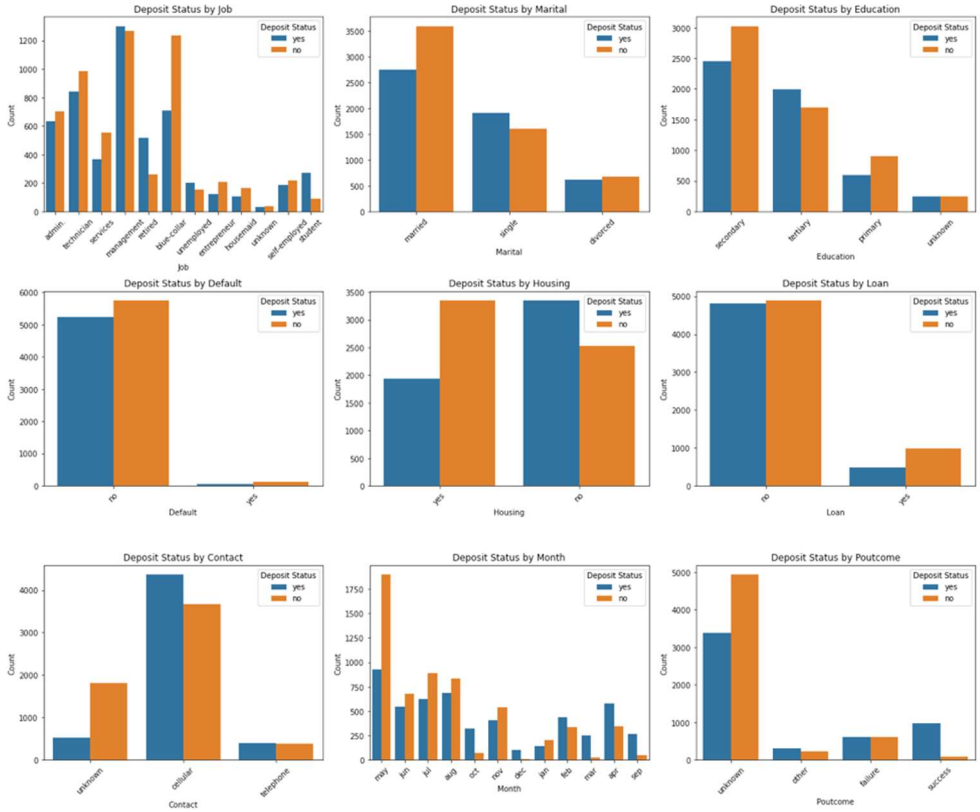


Figure 4

Initial Finding of Histogram:

1. **Contact**: More deposits occur with cellular contact, while telephone and unknown contacts see fewer deposits.
2. **Month**: Deposits are more common in May, August, and October.
3. **Poutcome**: A previous successful outcome leads to higher deposit rates, while most customers with an "unknown" outcome do not deposit.
4. **Marital Status**: Single individuals tend to deposit more than married or divorced ones.
5. **Education**: Higher deposit rates are seen among tertiary-educated individuals.
6. **Default**: Very few customers with credit in default make deposits.
7. **Housing Loan**: Customers without a housing loan are more likely to deposit.
8. **Personal Loan**: Those without personal loans have higher deposit rates.

Data Prediction Technique:

This code aims to analyze a bank marketing dataset to predict whether clients will subscribe to a term deposit. The approach involves preprocessing the data, training machine learning models, and evaluating their performance. Two different models, logistic regression and a decision tree classifier, are employed to compare predictive accuracy and gain insights into feature importance.

Logistic Regression:

The logistic regression model is used to explore the relationship between the features and the likelihood of a client subscribing to a term deposit. Before training, continuous variables are standardized to ensure consistent scaling, which enhances the model's performance.
The logistic regression coefficients are analyzed to identify the most significant features affecting the target outcome. Features with high absolute coefficient values are considered influential, providing insights into the factors driving client behavior.

The logistic regression model achieved an accuracy of approximately **73.49%**, indicating that it correctly classified about three-quarters of the test set data.

Important Variables

The most significant variables in the logistic regression model (based on coefficients) are as follows:

1. poutcome_success (1.63): The strongest positive predictor, indicating that a successful outcome from a previous campaign significantly increases the likelihood of a client subscribing.

2. duration (0.79): The length of the call is highly indicative of subscription likelihood, with longer durations correlating positively with success.

3. job_retired (0.59): Retired clients are less likely to subscribe, possibly reflecting financial instability or availability for investments.

4. job_student (0.51): Students also show a lower likelihood of subscribing, potentially due to their lack of funds and financial stability.

5. contact_cellular (0.47): Contacting clients via cellular phones positively impacts subscription rates, likely due to better engagement.

6. poutcome_failure (-0.48): A failed outcome in a previous campaign negatively impacts the likelihood of subscription.

7. housing (-0.54): Homeownership slightly reduces the likelihood of subscription, possibly due to financial constraints.

8. loan (-0.55): Clients with existing loans are less likely to subscribe, suggesting financial limitations.

9. contact_unknown (-0.68): Lack of contact information is a significant negative predictor, emphasizing the importance of complete data.

10. poutcome_unknown (-0.79): An unknown outcome from a previous campaign strongly

decreases the likelihood of subscription.

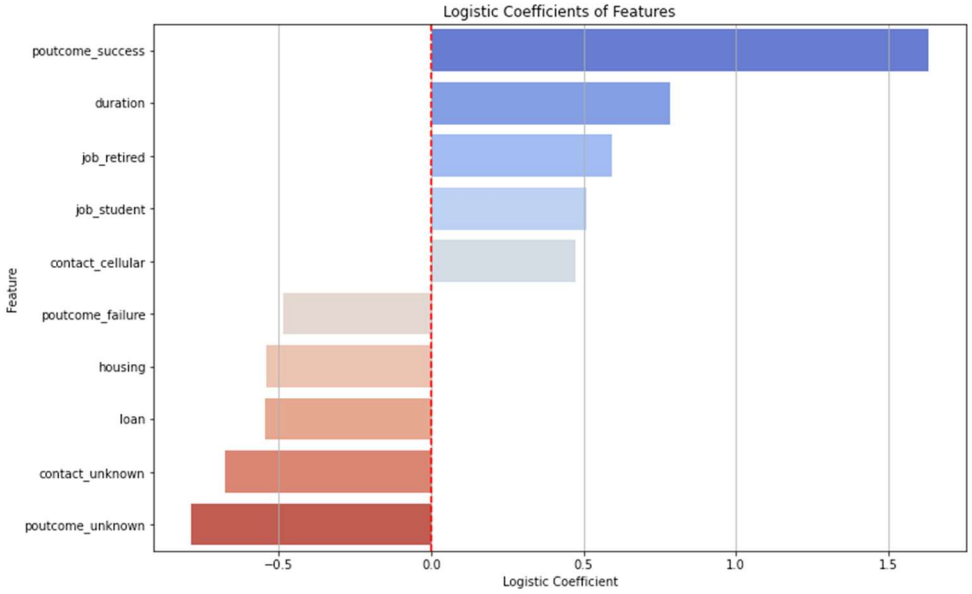


Figure 5

Decision tree:

The decision tree classifier, trained with a maximum depth of 6, achieved an accuracy of 72.50%. While slightly less accurate than logistic regression, the decision tree provides interpretability through feature importance, offering insights into the hierarchical decision-making process.

Important Variables

The decision tree's feature importance scores highlight the factors most influential in the classification process:

1. poutcome_success (0.30): The most critical feature, emphasizing that a successful outcome from a previous campaign is a significant determinant of client subscription.

2. duration (0.23): Call duration remains a key factor, reflecting its importance in determining client interest and engagement.

3. contact_unknown (0.18): The lack of contact information significantly impacts classification, indicating the importance of robust data collection.

4. month (0.10): The timing of the campaign affects subscription likelihood, suggesting

potential seasonality in client responses.

5. housing (0.08): Hosing loan status has a measurable impact, aligning with financial stability considerations.

6. poutcome_unknown (0.06): Uncertainty in the outcome of previous campaigns negatively influences predictions.

7. pdays (0.02): The recency of previous contact plays a minor role, contributing modestly to the decision-making process.

8. marital_single (0.01): Marital status has a minimal impact but still influences the predictions slightly.

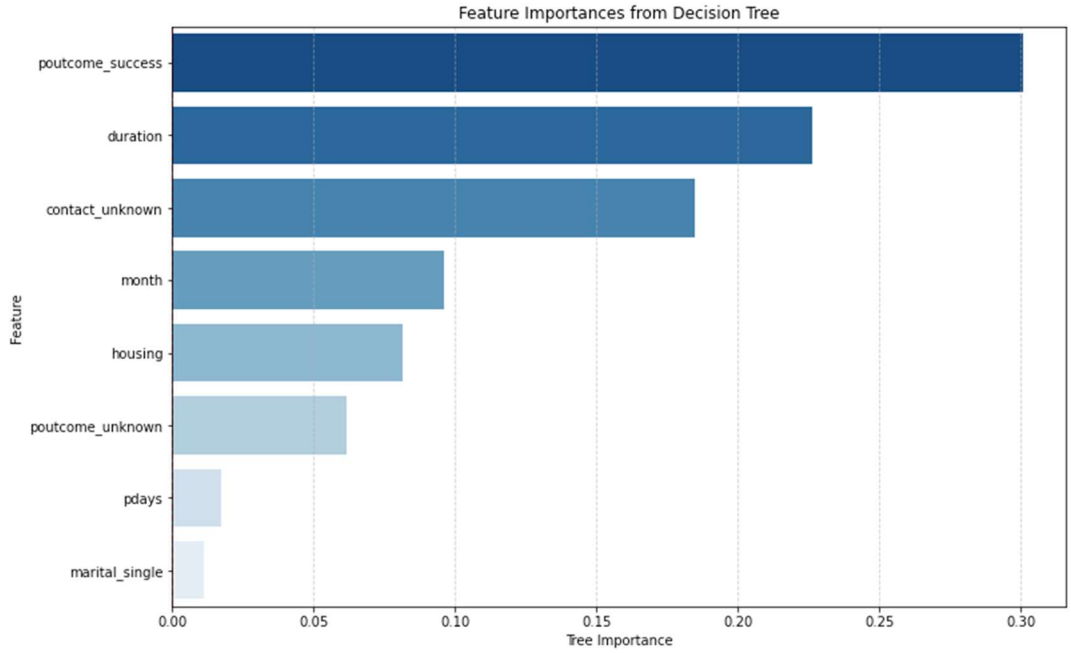


Figure 6

Having applied both the **Decision Tree** and **Logistic Regression** models, we have identified key features influencing customer deposit behavior, along with their respective importance scores and coefficients.

These findings provide a solid foundation for understanding which variables most significantly

impact deposit likelihood, such as **poutcome_success**, **duration**, **contact_unknown, duration, job_retired and housing**. With an overall accuracy of around **73%** from both models, these insights demonstrate a moderate predictive ability but highlight specific areasfor improvement.

For the next step we will use the logistic regression features and perform **K-Means Clustering** to find two distinct clusters based on their relevance and impact on term deposit predictions. By comparing these clusters, we aim to identify the optimal grouping of features, which willinform our strategic approach to customer outreach and campaign targeting. This feature selection process will be essential in designing campaigns that maximize engagement and increase term deposit conversion rates.

Why Logistic Regression Was Chosen Over a Decision Tree

Logistic regression was chosen over the decision tree due to several factors. It provides clear and interpretable coefficients, making it ideal for explaining decision-making processes, whereas decision trees can become complex and sensitive to data changes. Logistic regression aligns well with datasets exhibiting linear relationships, while decision trees, though capable of capturing non-linear patterns, are more prone to overfitting on smaller or imbalanced datasets. Additionally, logistic regression generalizes better to unseen data, aided by regularization techniques, whereas decision trees may memorize patterns despite depth constraints. Performance metrics also favored logistic regression, as it showed balanced results across accuracy, precision, recall, and F1-score, outperforming the decision tree on test data. Finally, logistic regression is computationally efficient, training and predicting faster than decision trees, which require resource-intensive splitting and evaluations.

K-means Clustering:

In this part of the analysis, we applied K-Means clustering on the logistic regression features to identify patterns or segments within the dataset based on the most influential factors from the logistic regression model. The features selected for clustering were the ones identified as having the highest logistic regression coefficients, and we performed clustering to examine if there were distinct groups that exhibit different behaviors in terms of the target variable, "deposit."

K-Means Clustering Approach

We used **K-Means clustering** with 4 clusters, after standardizing the selected features (excluding the target variable "deposit") to ensure all features contribute equally to the clustering process. The clusters were then assigned based on the similarity of the feature values, and we analyzed the resulting groups based on their "mean deposit probability" and other relevant features.

Cluster Summary

We performed K-Means clustering with 4 clusters and analyzed the resulting groups based on the following metrics:

1. **Cluster 1 (Moderate Deposit Probability):**
    - Mean Deposit Probability: 0.671 – Clients in this cluster have a moderate probability of subscribing to the term deposit.
    - Cluster Size: 2817 clients.

2. **Cluster 2 (Moderate Deposit Probability, but lower success in prior campaigns):**
    - Mean Deposit Probability: 0.529 – These clients have a moderate probability of subscription, but slightly lower than Cluster 1.
    - Cluster Size: 3678 clients.
3. **Cluster 0 (Low Deposit Probability):**
    - Mean Deposit Probability: 0.396 – Clients in this cluster have a low probability of subscribing.
    - Cluster Size: 2321 clients.
4. **Cluster 3 (Very Low Deposit Probability):**
    - Mean Deposit Probability: 0.225 – This cluster represents clients with the lowest probability of subscribing to the term deposit.
    - Cluster Size: 2346 clients.

By identifying these patterns, we can create tailored marketing strategies to improve campaign effectiveness and optimize conversion rates. In this analysis, we will focus on **Clusters 2 and 3**, which have varying probabilities of deposit subscription. Cluster 2 represents customers with moderate interest in term deposits, while Cluster 3 includes those with a lower probability of subscribing. We will explore the unique features of each cluster and propose targeted marketing strategies to address their specific needs. **Cluster 2 (Moderate Deposit Probability - 0.529)**

Cluster Characteristics:

- **Deposit Probability**: 0.529 (moderate likelihood of subscribing).
- **Size**: 3678 clients.
- **Features**:
    - **poutcome**: No information of their last campeign.
    - **Duration**: Duration of the last contact is1-971 seconds.
    - **Job Retired (0):** people who are not retired.
    - **job_student(0):** People who are not students.
    - **contact_cellular**: People who contacted trough the cellular
    - **Hosuing (0):** People who do no have loan on a house.
    - **loan**: People who don't have another loan.

Marketing Strategy:

- **Targeted Communication**: Use **SMS or mobile app notifications** for personalized, immediate engagement.
- **Incentives**: Offer **special interest rates** or **limited time offers**.
- **Focus on Financial Flexibility**: Emphasize how a term deposit can enhance savings for those without housing loans.
- **Moderate Effort**: Moderate marketing campaigns to leverage mobile engagement and offer incentives, focusing on financial security and growth.

Cluster 3 (Low Deposit Probability - 0.226)

Cluster Characteristics:

- **Deposit Probability**: 0.226 (low likelihood of subscribing).
- **Size**: 2346 clients.
- **Features**:
    - **poutcome**: No information of their last campeign.
    - **Duration**: Duration of the last contact is1-971 seconds.
    - **Job Retired (0):** people who are not retired.
    - **job_student(0):** People who are not students.
    - **contact_cellular**: People didn't contact trough the cellular
    - **contact_unknown(1):** People tht the way of contact is unknown
    - **Hosuing (0):** People who have loan on a house.
    - **loan**: Don't have another loan.

Marketing Strategy:

- Aggressive Campaigns: Deploy large-scale marketing campaigns, focusing on re-engagement and incentivizing conversions.
- Educational Messaging: Use financial education to explain how term deposits could work in the long term for those with housing loans, focusing on security and future financial goals.
- Personalized Offers: Tailor offers based on their specific financial situation (e.g., flexible deposit options for individuals with housing loans).
- High Campaign Effort: Use intensive outreach, perhaps with multiple touchpoints and incentive-based campaigns that focus on overcoming past campaign failures and building trust.

Conclusion

The project effectively addresses the business problem by leveraging customer segmentation and predictive modelling to enhance the subscription rates for NextGen Financial Solutions' term deposit product. Using logistic regression, key features influencing deposit likelihood were identified, providing actionable insights into customer behaviours and characteristics. The model

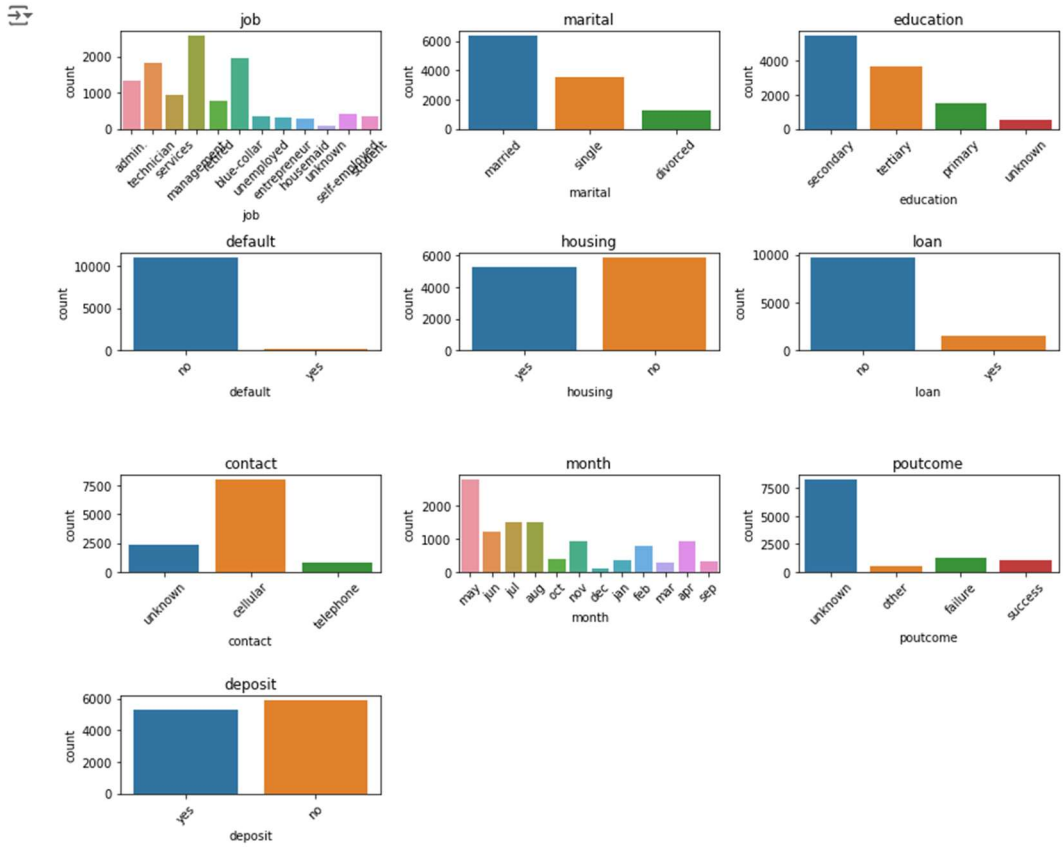
achieved a moderate performance, with an accuracy of approximately 73.5%, demonstrating a balanced ability to predict customer deposit probabilities.

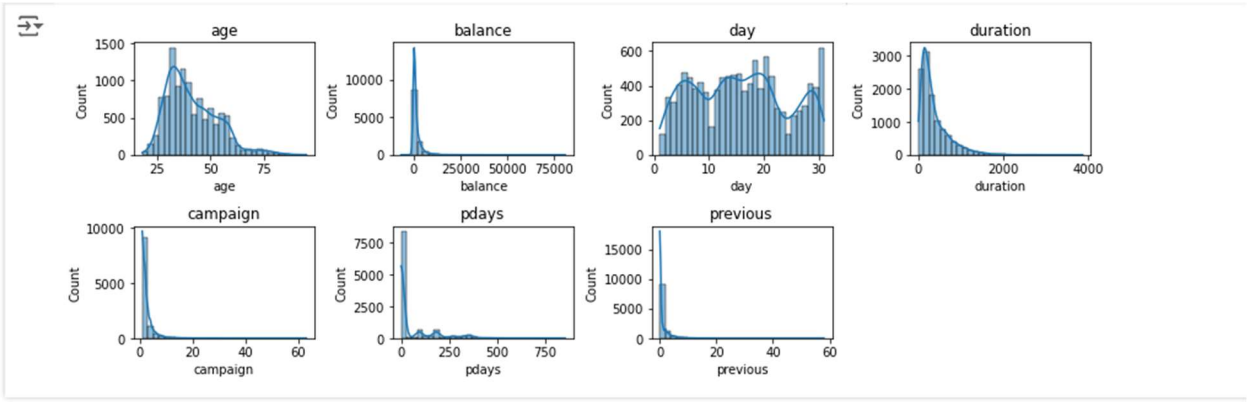
K-means clustering further refined customer segmentation, revealing distinct groups with varying probabilities of deposit participation. Two clusters were selected for targeted campaigns based on their characteristics and likelihood to buy. Cluster 2, with a higher likelihood of deposit participation, was addressed with a low-effort strategy focusing on automated notifications and streamlined processes. In contrast, Cluster 3, with a lower likelihood to buy, required a more intensive campaign that included educational initiatives, personalized consultations, and incentivized offers.

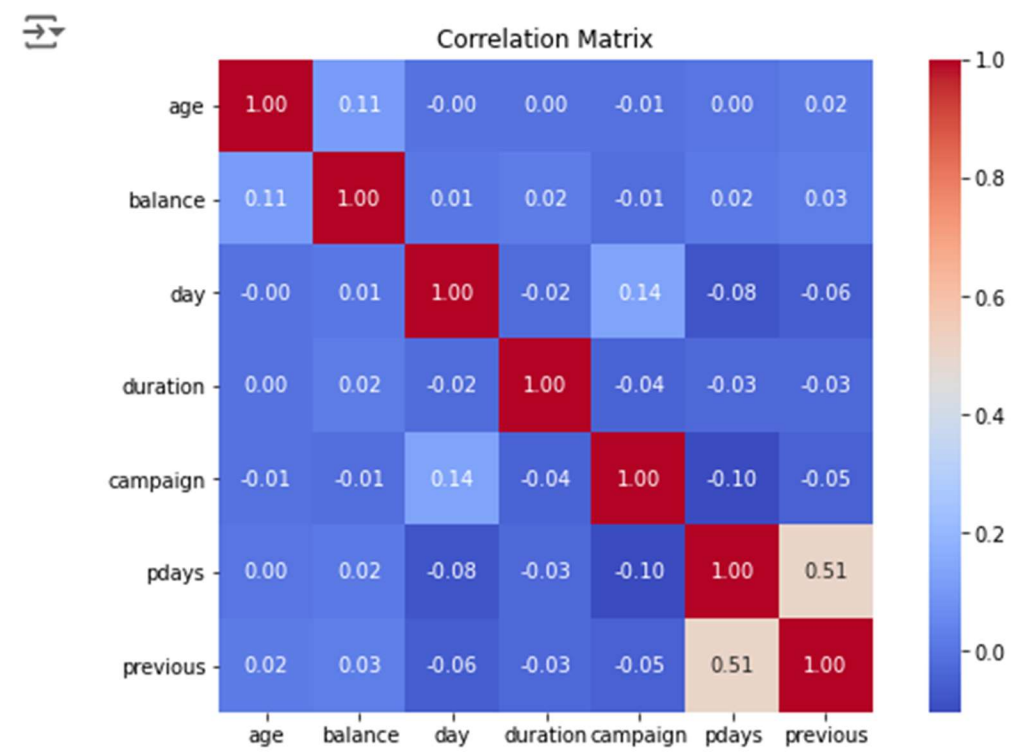
This project highlights the importance of segmenting customers based on their likelihood to subscribe to a term deposit. The insights from this analysis can significantly improve the effectiveness of marketing campaigns, leading to higher conversion rates and a better customer experience overall.

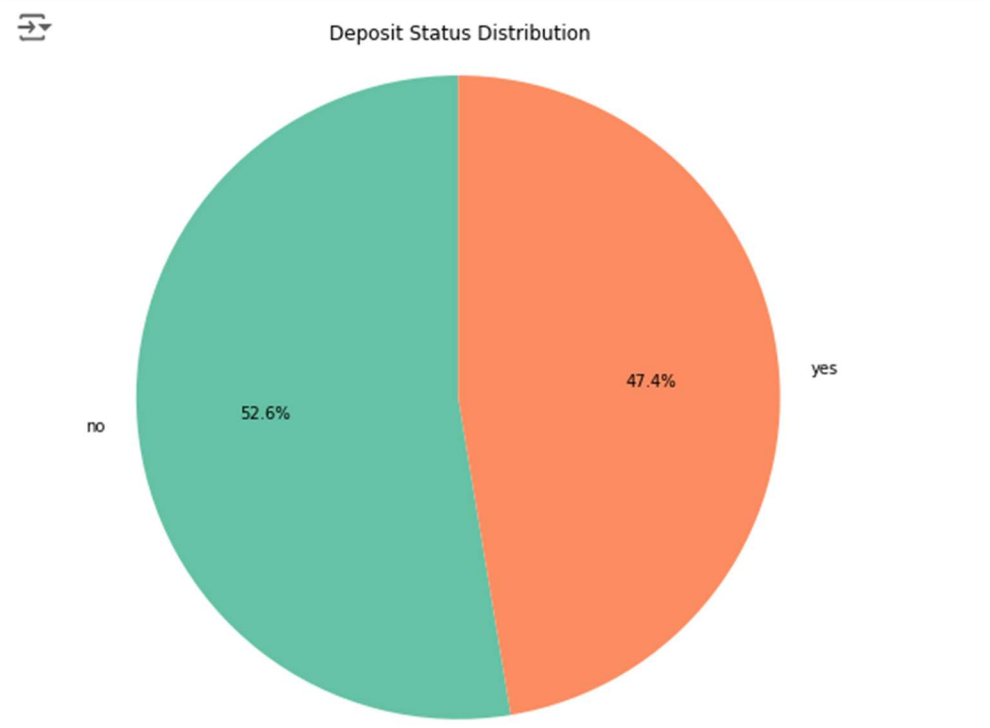
These data-driven strategies ensure optimal resource allocation, balancing short-term gains and long-term customer engagement. By integrating predictive insights with tailored marketing efforts, the project not only can increase subscription rates but also establishes a robust framework for customer relationship management and strategic decision-making. This approach aligns with NextGen Financial Solutions' goals of efficient resource utilization and sustained business growth.

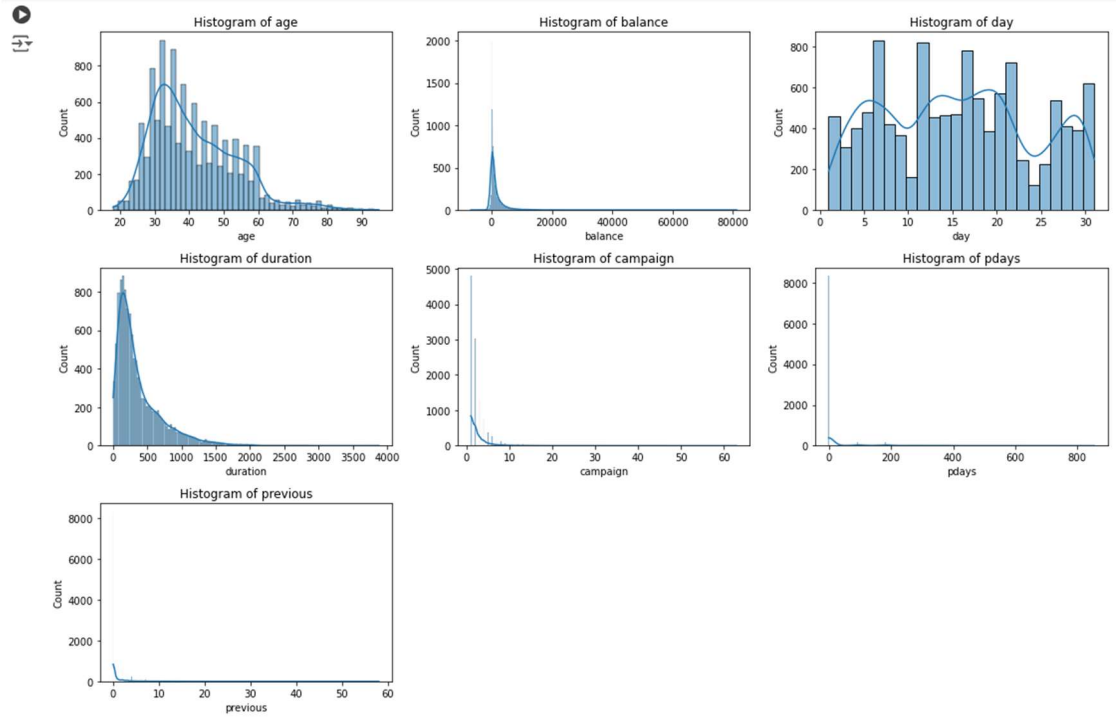
Appendix

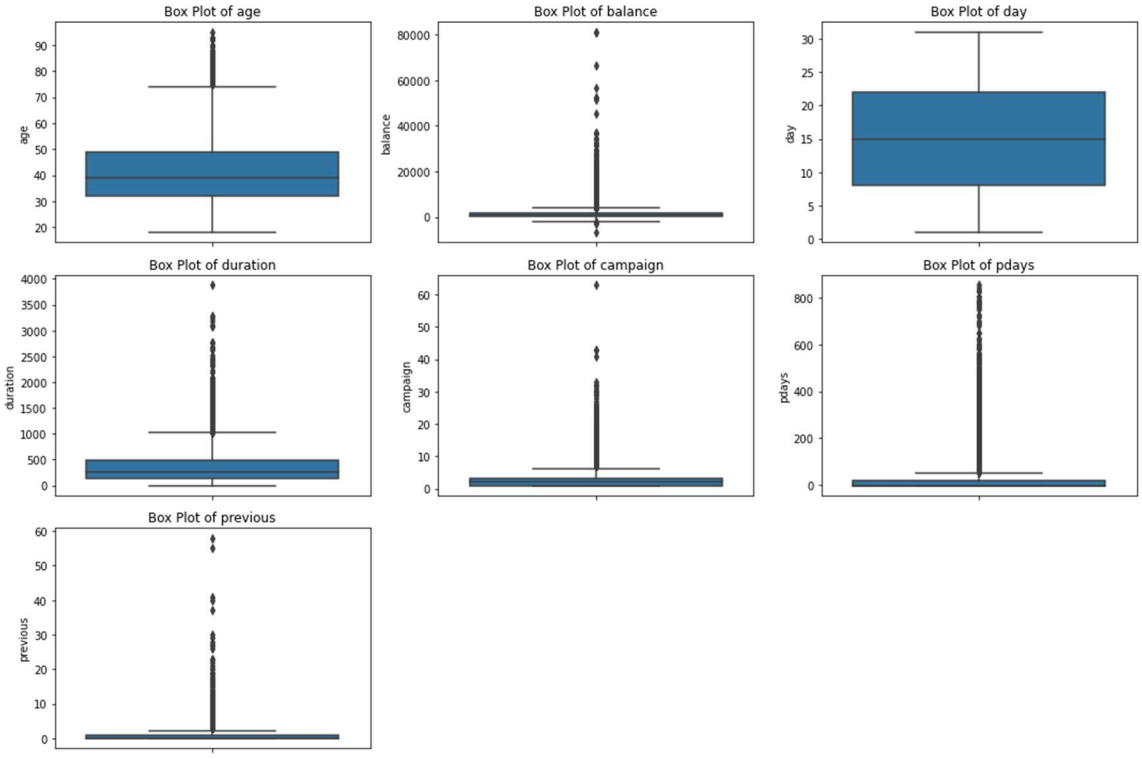

job
marital
education
default
housing
loan
contact
month
poutcome
deposit
count
admin.
technician
services
management
retired
blue-collar
unemployed
entrepreneur
housemaid
unknown
self-employed
student
married
single
divorced
secondary
tertiary
primary
unknown
no
yes
may
jun
jul
aug
oct
nov
dec
jan
feb
mar
apr
sep
cellular
telephone
other
failure
success

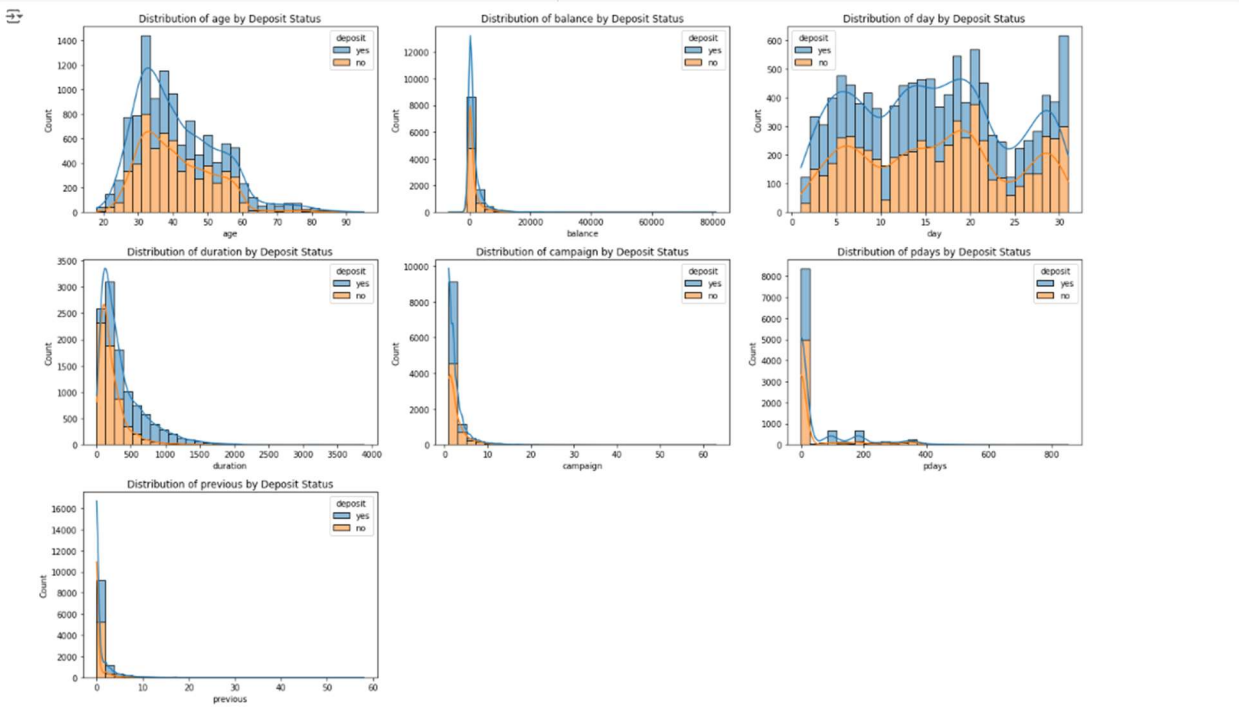

age
balance
day
duration
campaign
pdays
previous
Count


Correlation Matrix
age
balance
day
duration
campaign
pdays
previous
1.00 0.11 -0.00 0.00 -0.01 0.00 0.02
0.11 1.00 0.01 0.02 -0.01 0.02 0.03
-0.00 0.01 1.00 -0.02 0.14 -0.08 -0.06
0.00 0.02 -0.02 1.00 -0.04 -0.03 -0.03
-0.01 -0.01 0.14 -0.04 1.00 -0.10 -0.05
0.00 0.02 -0.08 -0.03 -0.10 1.00 0.51
0.02 0.03 -0.06 -0.03 -0.05 0.51 1.00
1.0
0.8
0.6
0.4
0.2
0.0


Deposit Status Distribution
47.4%
yes
52.6%
no


Histogram of age
Count
age
Histogram of balance
Count
balance
Histogram of day
Count
day
Histogram of duration
Count
duration
Histogram of campaign
Count
campaign
Histogram of pdays
Count
pdays
Histogram of previous
Count
previous


Box Plot of age
age
Box Plot of balance
balance
Box Plot of day
day
Box Plot of duration
duration
Box Plot of campaign
campaign
Box Plot of pdays
pdays
Box Plot of previous
previous



	Cluster	Mean Deposit Probability	Count	Mean poutcome_success	Mean duration	Mean job_retired	Mean job_student	Mean contact_cellular	Mean poutcome_failure	Mean housing	Mean loan	Mea contact_unknow
2	2	0.913165	1071	1	0.029879	0.126050	0.062558	0.915033	0	0.264239	0.047619	0.00560
3	3	0.503257	1228	0	0.043160	0.059446	0.035016	0.930782	1	0.597720	0.136808	0.00488
1	1	0.485290	6526	0	0.078302	0.073552	0.035550	0.906987	0	0.395035	0.140362	0.00000
0	0	0.225075	2337	0	0.091570	0.038511	0.007702	0.000000	0	0.721438	0.139067	0.99871

Choosen cluster for first campaign (Less marketing effort higher likelihood to buy)

	Cluster	Mean Deposit Probability	Count	Mean poutcome_success	Mean duration	Mean job_retired	Mean job_student	Mean contact_cellular	Mean poutcome_failure	Mean housing	Mean loan	Mean contact_unknown
0	0	0.225075	2337	0	0.09157	0.038511	0.007702	0.0	0	0.721438	0.139067	0.998716

Chosen cluster for second campeign (More marketing effort lower likelihood to buy)

	Cluster	Mean Deposit Probability	Count	Mean poutcome_success	Mean duration	Mean job_retired	Mean job_student	Mean contact_cellular	Mean poutcome_failure	Mean housing	Mean loan	Mea contact_unknow
1	1	0.48529	6526	0	0.078302	0.073552	0.03555	0.906987	0	0.395035	0.140362	0.(