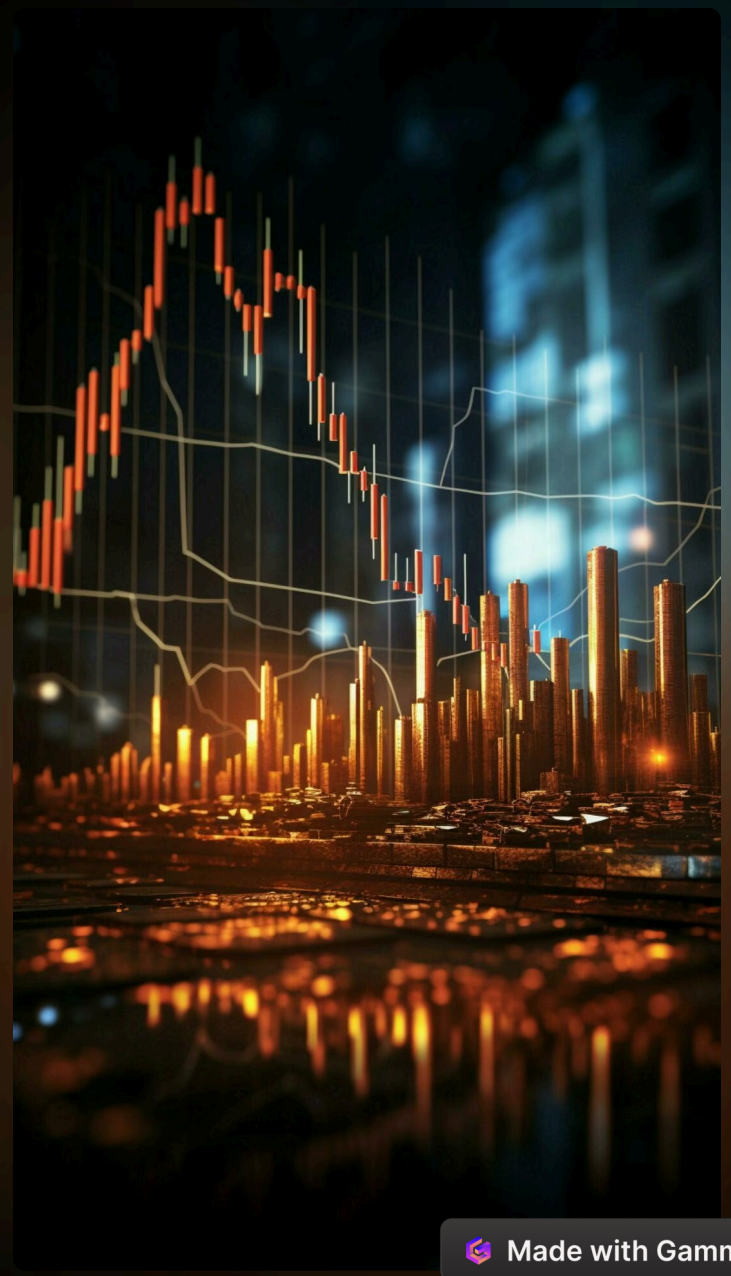
Analyzing Bank Marketing Data: Predicting Customer Responses

This project analyzes bank marketing data to develop predictive models capable of forecasting customer responses to future campaigns. The goal is to enhance marketing strategies by identifying key features for effective targeting.



Data Overview

Dataset Origin:

- This dataset is part of the **UCI Machine Learning Repository**, focused on financial institution marketing campaigns.

Business Context:

- The dataset tracks a bank's marketing efforts aimed at promoting **term deposit subscriptions**. The goal of this analysis is to uncover strategies that will improve future marketing campaigns by understanding customer behaviors and outcomes.

Key Details:

- **Total Records**: 11162
- **Features**: 17 columns, covering customer demographics, financial information, and interaction history.
- **Target Variable**: `deposit` (Did the client subscribe to a term deposit? `yes`/`no`).

Main Objective:

- Leverage this data to identify actionable insights for optimizing future marketing campaigns and increasing term deposit subscription rates.

Brief Description of the Data Set

- **age**: int64 - Age in years
- **job**: object - Type of job (categorical values: 'admin.', 'technician', 'services', 'management', 'retired', 'blue-collar', 'unemployed', 'entrepreneur', 'housemaid', 'unknown', 'self-employed', 'student')
- **marital**: object - Marital status (categorical values: 'married', 'single', 'divorced')
- **education**: object - Education background (categorical values: 'secondary', 'tertiary', 'primary', 'unknown')
- **default**: object - Has credit in default? (categorical values: 'no', 'yes')
- **balance**: int64 - Balance of the individual
- **housing**: object - Has housing loan? (categorical values: 'yes', 'no')
- **loan**: object - Has personal loan? (categorical values: 'no', 'yes')
- **contact**: object - Contact communication type (categorical values: 'unknown', 'cellular', 'telephone')
- **day**: int64 - Last contact day of the week (categorical values: 'mon', 'tue', 'wed', 'thu', 'fri')
- **month**: object - Last contact month of year (categorical values: 'may', 'jun', 'jul', 'aug', 'oct', 'nov', 'dec', 'jan', 'feb', 'mar', 'apr', 'sep')
- **duration**: int64 - Last contact duration, in seconds
- **campaign**: int64 - Number of contacts performed during this campaign and for this client
- **pdays**: int64 - Number of days that passed after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
- **previous**: int64 - Number of contacts performed before this campaign and for this client
- **poutcome**: object - Outcome of the previous marketing campaign (categorical values: 'unknown', 'other', 'failure', 'success')

Data Sources and Preprocessing

Data Acquisition

We will use data from a Kaggle dataset, including customer demographics, financial information, and previous campaign interactions.

Data Cleaning

The data will be cleaned and preprocessed to address missing values, inconsistencies, and outliers. This ensures data quality and reliability for analysis.

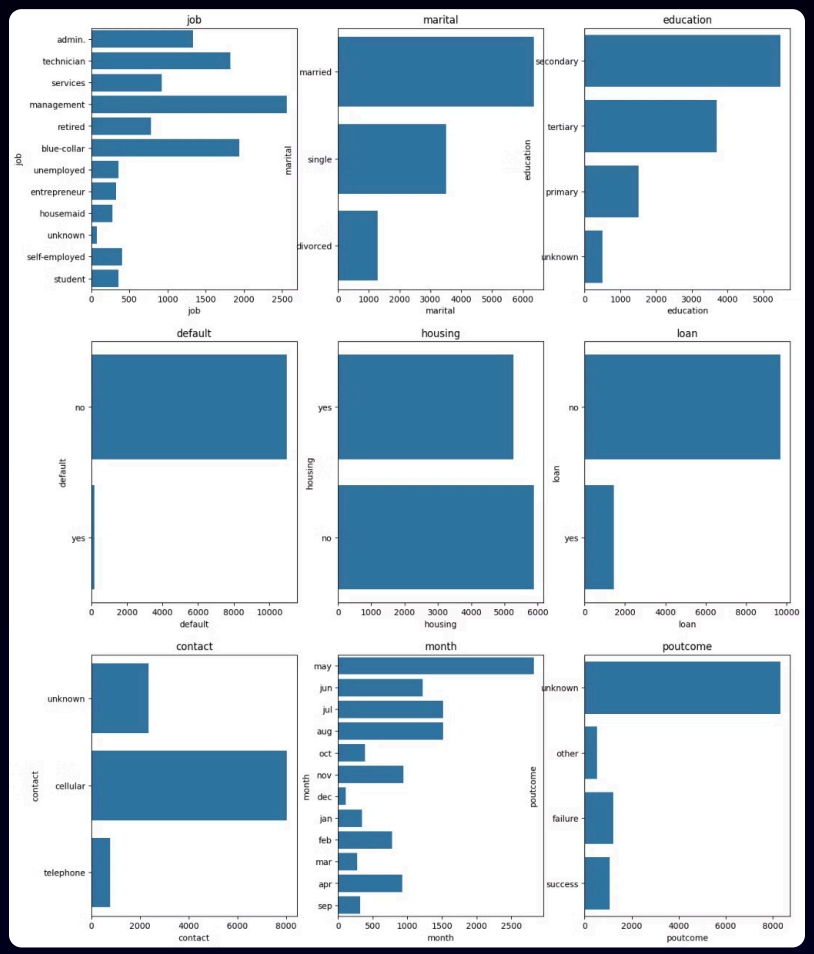
Feature Engineering

New features will be engineered from existing ones to enrich the dataset. This may involve creating interaction terms or transforming variables to better capture relationships.

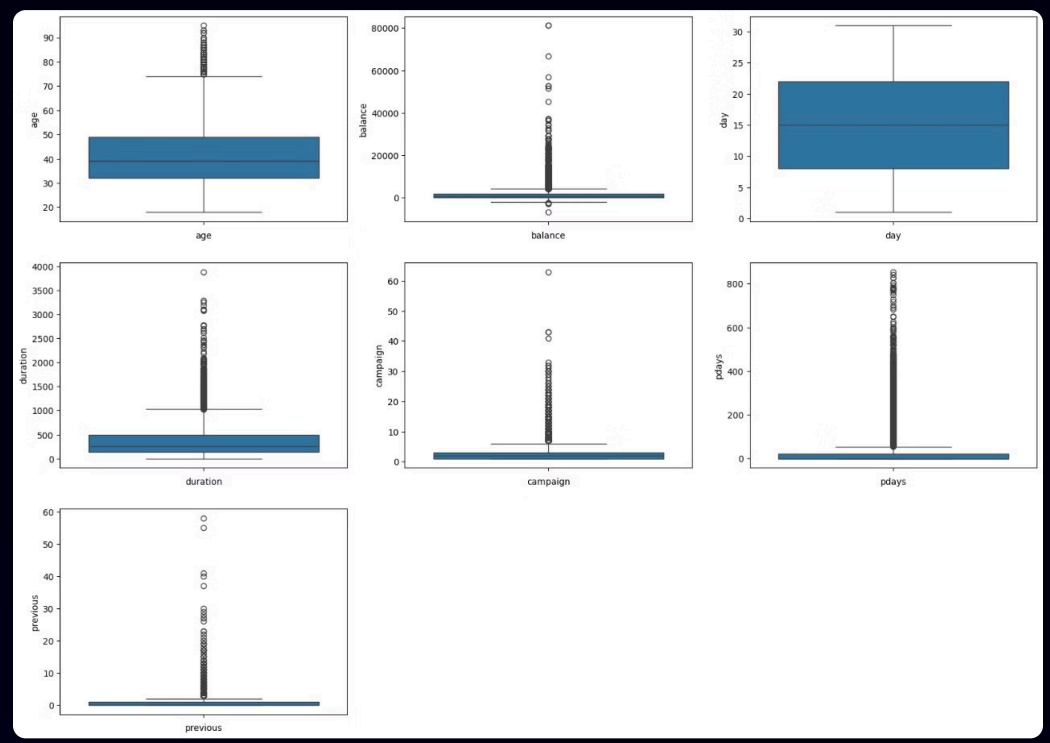
Data Preprocessing

- **Describing the Data frame**:
  - A thorough statistical summary was generated, revealing the dataset's composition, including numerical data (e.g., age, balance) and categorical variables (e.g., job, education, deposit status). This helped in understanding the range, distribution, and central tendencies of the data.
- **Finding Unique Elements**:
  - Each column's unique values were examined to identify the variability and potential patterns in the dataset. This exploration was critical in distinguishing between continuous variables and categorical features.
- **Unwanted Columns**:
  - We evaluated the relevance of each column. Certain columns, like those containing only a single unique value or too many missing values, were flagged as potentially redundant. Further analysis may suggest whether these columns should be removed to streamline the dataset.
- **Handling "Unknown" Values**:
  - Instead of traditional null or empty fields, the dataset contains the value "unknown" in multiple columns, such as job, education, and contact. These entries were treated as missing values and addressed during the preprocessing phase to avoid skewing the model.
- **Empty Fields**:
  - We scanned the dataset for completely empty fields, finding no blank cells. This ensured the data was complete in terms of structural integrity, minimizing the need for extensive data cleaning.
- **Fields with Only One Value**:
  - Several columns were checked for uniqueness, and any that held a single constant value across all rows were flagged for removal, as they don't contribute meaningful variability to the model.
- **Exploring Categorical Features**:
  - Categorical columns like `job`, `marital status`, `education`, and `contact method` were analyzed to understand their distribution and relationship to the target variable (`deposit`). Encoding strategies will be applied to convert these categorical features into numerical representations for machine learning algorithms.

Data Visualization

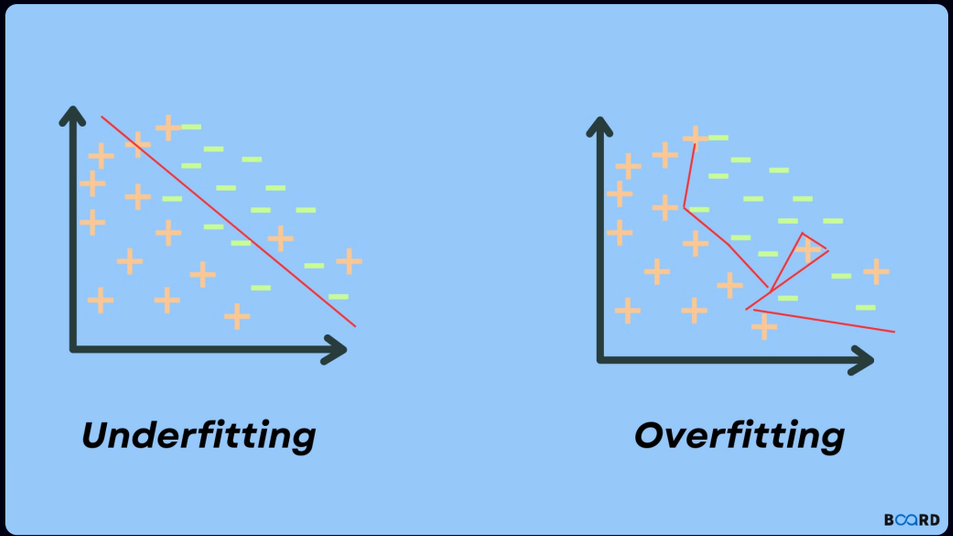


- The image shows several bar plots visualizing the distribution of categorical features from a dataset, including variables like **job**, **marital status**, **education**, **contact method**, and **loan status**.
- Some key insights include a higher proportion of clients being **married** and **employed in admin jobs**, with most communications happening via **cellular** and a significant number of **unknown** outcomes in the marketing campaign.



- The image shows a series of **box plots** for numerical features in the dataset, including **age**, **balance**, **duration**, **campaign**, **pdays**, and **previous**. The plots highlight the distribution of values and the presence of outliers.
- Notably, there are significant outliers in features like **balance**, **duration**, **campaign**, **pdays**, and **previous**, indicating a wide range of values or potentially abnormal cases in the dataset.

Overfitting vs Underfitting



deposit	default	
no	no	5753
	yes	115
yes	no	5236
	yes	52

1. **Avoiding overfitting:** The 'default' variable may strongly correlate with deposit outcomes, leading the model to overly rely on it, which could reduce its ability to generalize to unseen data.
2. **Simplifying the model:** Removing 'default' reduces the complexity of the model, allowing it to focus on other, potentially more informative features, which may improve prediction accuracy for future campaigns.

Predictive Modeling Techniques

1 Logistic Regression

A widely used statistical method for binary classification, predicting the probability of a customer responding to a campaign.

2 Decision Trees

Tree-based models provide interpretable decision rules for predicting customer behavior based on specific features.

3 Random Forests

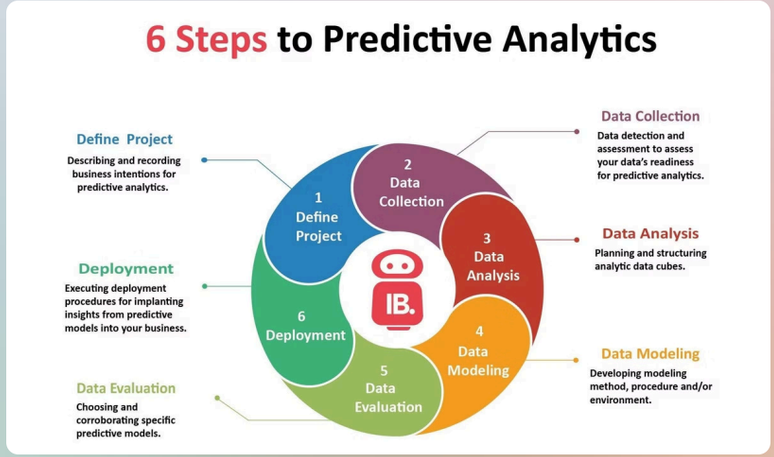
An ensemble learning technique that combines multiple decision trees to improve prediction accuracy and handle complex relationships.

4 Support Vector Machines (SVMs)

A powerful technique for identifying optimal hyperplanes to separate different classes of customers based on their features.

5 XGBoost

A gradient boosting algorithm known for its high accuracy and efficiency in handling large datasets. It iteratively combines weak learners to create a strong predictive model.



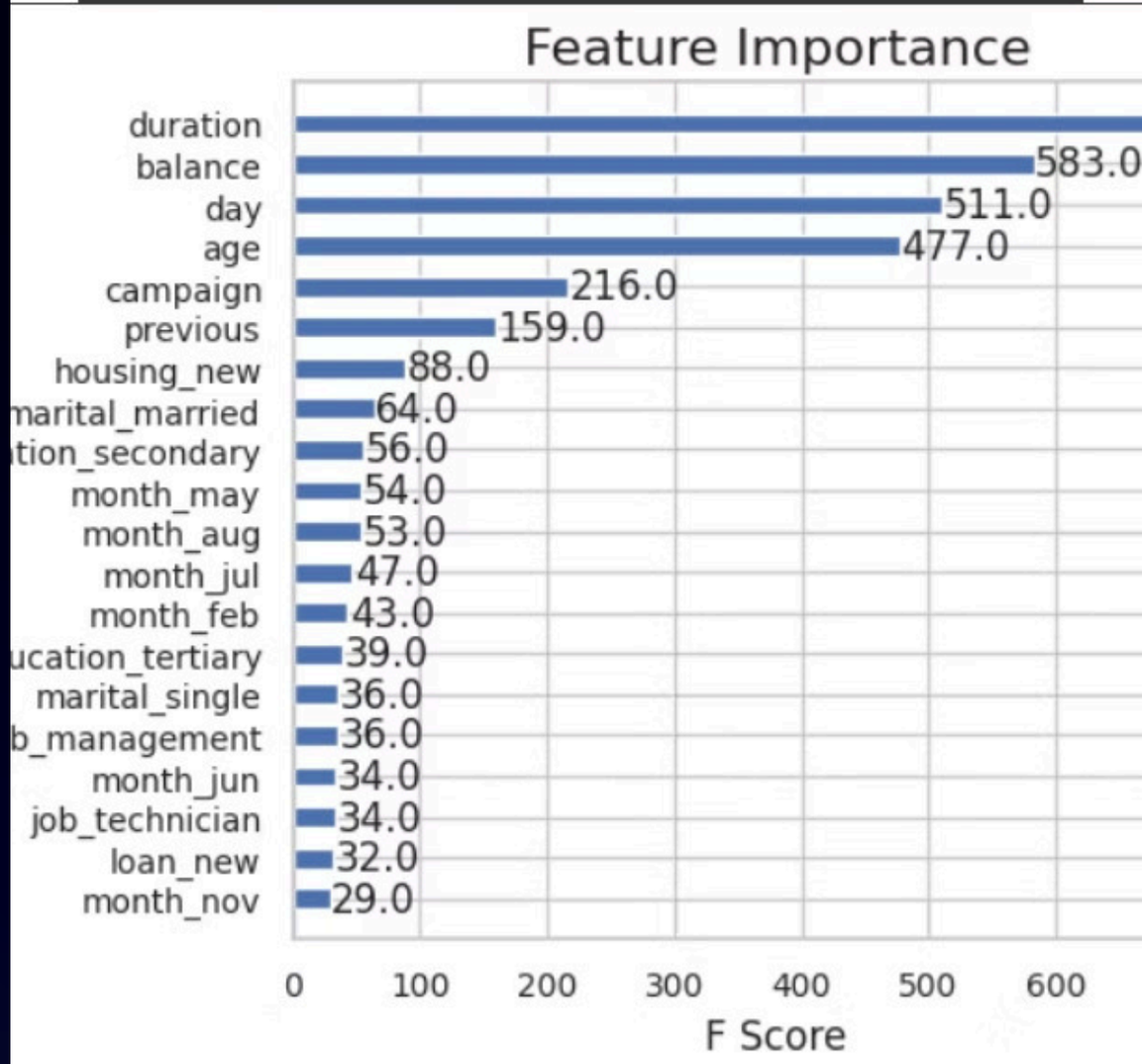
Feature Engineering and Selection

Feature	Description	Importance
Age	Customer's age	High
Balance	Customer's Bank Balance	High
Education Level	Customer's highest education level	Moderate
Previous Campaign Contacts	Number of previous contacts with the customer	High
Contact Duration	Duration of the last contact with the customer	High

```
Training Accuracy: 95.68%
Testing Accuracy: 84.90%
              precision    recall  f1-score   support

           0       0.86      0.85      0.85      1728
           1       0.84      0.85      0.84      1616

    accuracy                           0.85      3344
   macro avg       0.85      0.85      0.85      3344
weighted avg       0.85      0.85      0.85      3344
```



Forecasting Customer Responses to Future Campaigns

1

Data Cleaning and Preprocessing

We cleaned and preprocessed the data to address missing values, inconsistencies, and outliers. This ensures data quality and reliability for analysis.

2

Model Training

We trained the an XGBoost model, on the bank data to identify patterns and relationships between features and customer responses.

3

Model Evaluation

The performance of the trained models was evaluated using metrics such as accuracy, precision, recall, and F1-score.

4

Forecasting

Once trained and validated, the model can be used to predict customer responses to future marketing campaigns based on new data.

Recommending Key Features for Effective Targeting

Duration

Targeting customers based on the duration of their last contact can reveal patterns in their engagement with marketing campaigns.

Balance

Customers with higher bank balances may be more receptive to specific products or services, allowing for tailored marketing campaigns.

Age

Targeting specific age groups can be effective, considering their financial needs, product preferences, and life stages.

Campaign Contacts

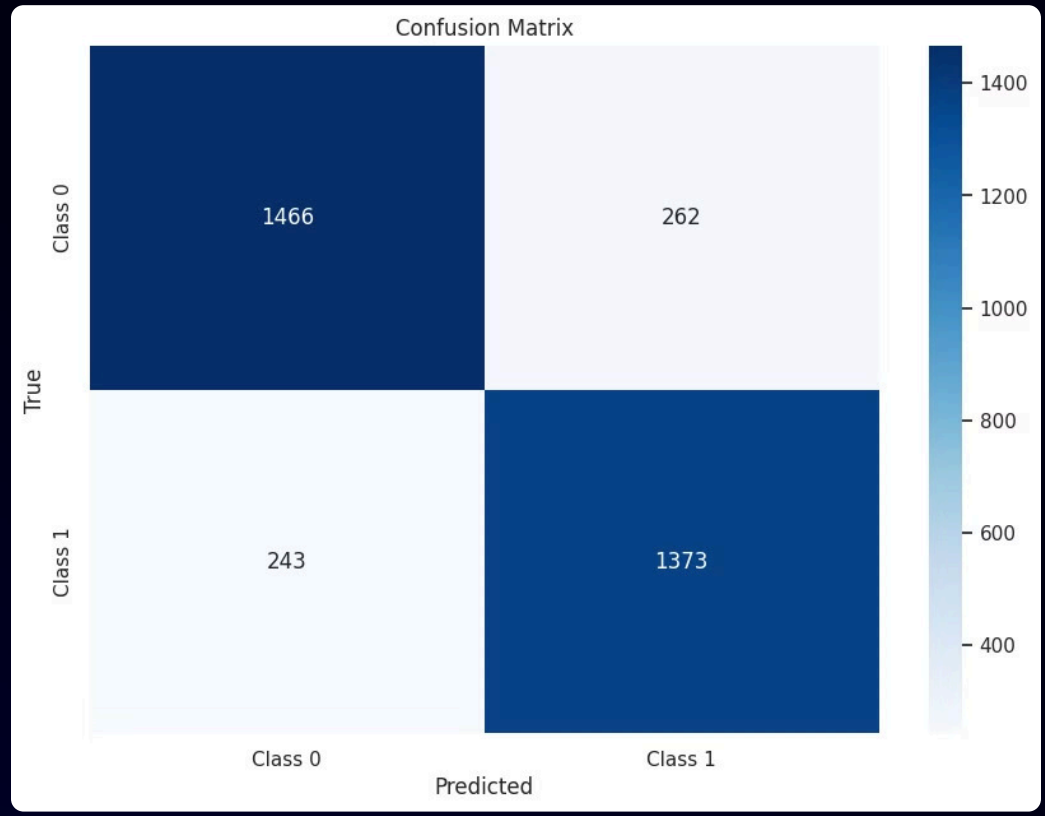
Understanding the number of previous contacts with a customer can help personalize future campaigns and tailor messages.

Previous Contact

Analyzing previous contact methods can help identify the most effective channels for reaching customers.

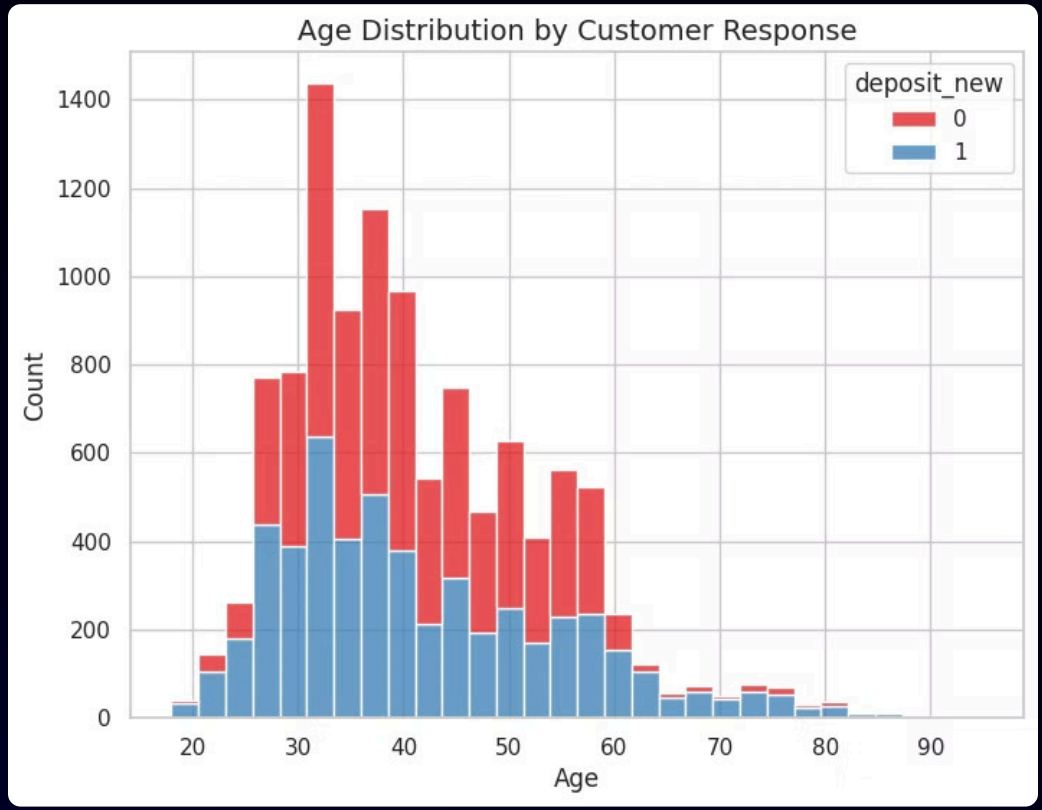
Insights Made

- **True Positives**: 1,373 customers were correctly predicted to make a deposit.
- **True Negatives**: 1,466 customers were correctly predicted not to make a deposit.
- **False Positives:** 262 customers were incorrectly predicted to make a deposit but didn't.
- **False Negatives:** 243 customers were incorrectly predicted not to make a deposit but did.



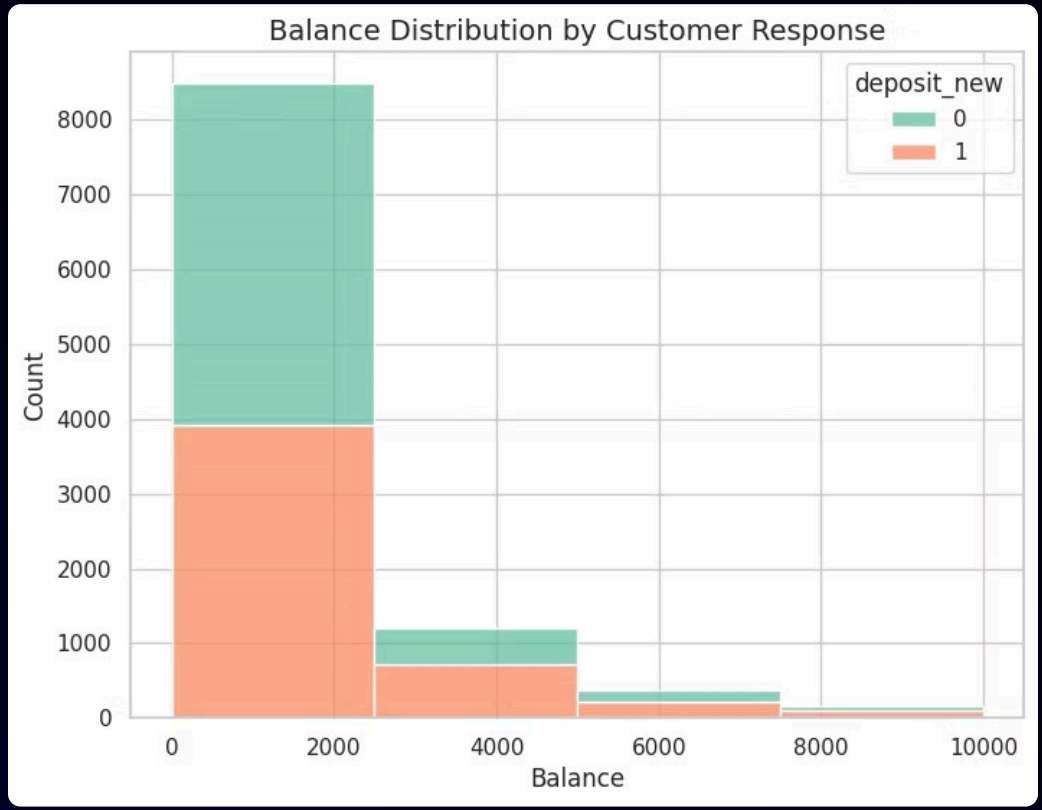
Insights Made

1. **High non-deposit rate in younger age groups:** Customers between 20 and 40 years show a higher proportion of red bars, indicating more customers did not subscribe to deposits in this age group.
2. **Older age groups (40-60) are more likely to subscribe:** As age increases beyond 40, especially between 40 and 60, the proportion of customers who subscribed to deposits (blue bars) increases, showing a more balanced or even higher subscription rate.
3. **Fewer responses from older customers:** There is a sharp decline in both deposit and non-deposit customers after age 60, indicating fewer customers in the older age groups are involved in the marketing campaigns.



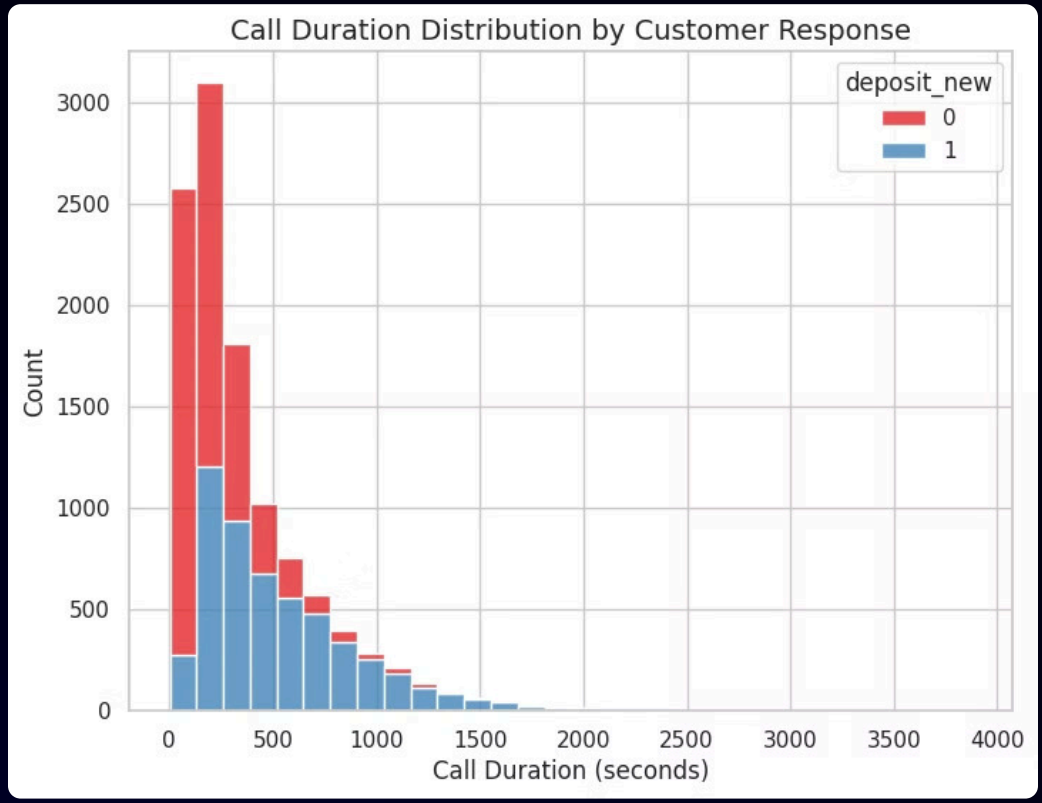
Insights Made

1. **High concentration of low balances:** Most customers, regardless of whether they subscribed to the deposit (green for no, red for yes), have very low balances, concentrated around 0 to 5,000.
2. **Higher balance increases likelihood of deposit subscription:** As balance increases beyond 0, there is a visible rise in the proportion of customers subscribing to deposits (red), especially in the range of 5,000 to 10,000.
3. **Few high-balance customers:** Very few customers have balances beyond 20,000, but those who do are more likely to subscribe to deposits, as indicated by the larger share of red bars in the higher balance ranges.



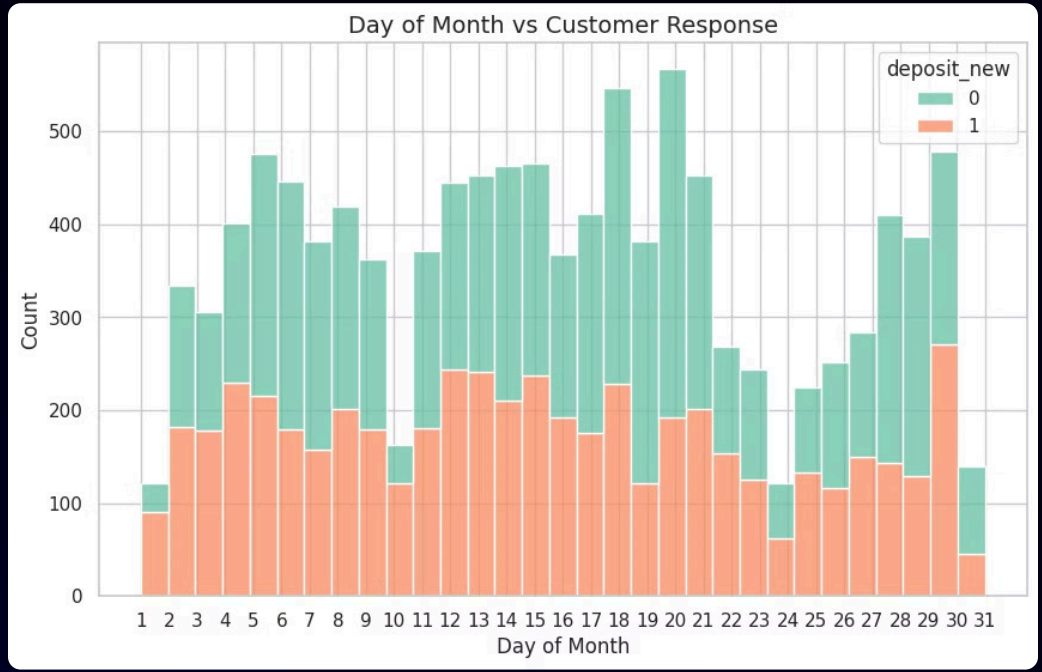
Insights Made

1. **Shorter Calls Are More Frequent:** The majority of calls have a duration of less than 500 seconds. This indicates that many customer interactions are relatively quick and efficient.
2. **Longer Calls Are Associated with Higher Conversions:** The distribution of call durations for customers who make a deposit (response=1) extends further to the right compared to those who do not (response=0). This suggests that longer calls may be more likely to result in a positive outcome (e.g., a deposit).



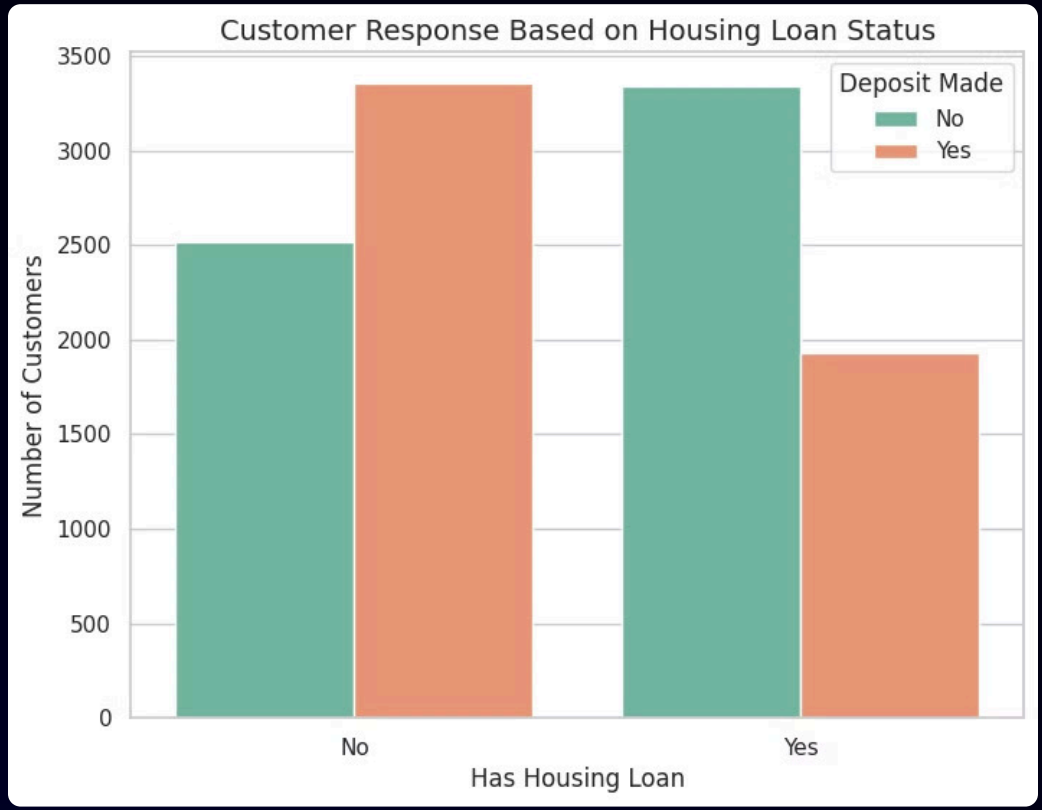
Insights Made

1. **Consistent Pattern Throughout the Month:** There doesn't appear to be a strong, consistent pattern between the day of the month and customer response. The number of deposits (response=1) fluctuates throughout the month, with no clear peaks or troughs.
2. **No Clear Correlation:** Overall, the data does not suggest a clear correlation between the day of the month and customer response.



Insights Made

- Customers with housing loans are more likely to make deposits.
- More customers without housing loans overall.
- Housing loan status is a significant factor in deposit decisions.



Conclusion and Next Steps

Key Findings:

- **Duration** and **Balance** are highly influential.
- **Age**, **Day**, and **Housing Loan** have moderate impact.
- Other features vary in importance.

Recommendations:

- Prioritize interactions with longer durations.
- Consider customer balance and housing loan status in decisions.
- Explore the impact of age and day on the target variable.

Next Steps:

- Perform deeper analysis on key features.
- Develop targeted strategies based on insights.
- Continuously monitor and refine models.