Bank Marketing Dataset Analysis Report

1. Overview

The primary goal is to build a predictive model to classify whether a customer will subscribe to a term deposit (`y`). This involves identifying important features, optimizing model performance, and ensuring interpretability.

2. Data and Model

Data

- The dataset contains information collected from a marketing campaign by a Portuguese bank. It includes customer details (e.g., age, job, marital status) and campaign-specific attributes (e.g., contact method, duration). The target variable is binary: yes (subscribed) or no (not subscribed).

1. **age** (numeric)
2. **job** : type of job (categorical: "admin.","unknown","unemployed","management","housemaid","entrepreneur","student","blue-collar","self-employed","retired","technician","services")
3. **marital** : marital status (categorical: "married","divorced","single"; note: "divorced" means divorced or widowed)
4. **education** (categorical: "unknown","secondary","primary","tertiary"
5. **default**: has credit in default? (binary: "yes","no")
6. **balance**: average yearly balance, in euros (numeric)
7. **housing**: has housing loan? (binary: "yes","no")
8. **loan**: has personal loan? (binary: "yes","no")

related with the last contact of the current campaign:

9. **contact**: contact communication type (categorical: "unknown","telephone","cellular")
10. **day**: last contact day of the month (numeric)
11. **month**: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
12. **duration**: last contact duration, in seconds (numeric)
13. \

other attributes:

14. **campaign**: number of contacts performed during this campaign and for this client (numeric, includes last contact)
15. **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)
16. **previous**: number of contacts performed before this campaign and for this client (numeric)
17. **poutcome**: outcome of the previous marketing campaign (categorical: "unknown","other","failure","success")

Output variable (desired target):

18. **y** : has the client subscribed a term deposit? (binary: "yes","no").

- After cleaning, the data was split into training and testing sets (80-20 split).

Approach

1. **Data Cleaning:**
   - Checked for missing values and there were no missing values.
   - Verified data types, there are no columns where type conversions is necessary.

check for missing values

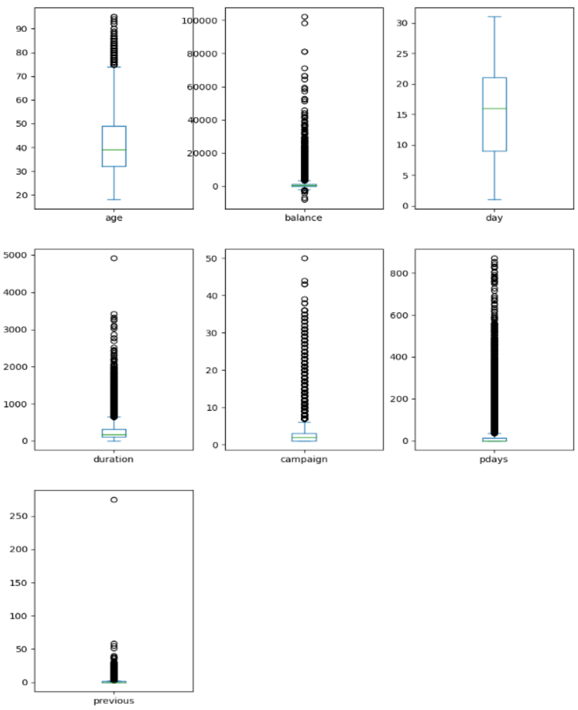
```
data.isnull().sum()

#there are no missing values
```

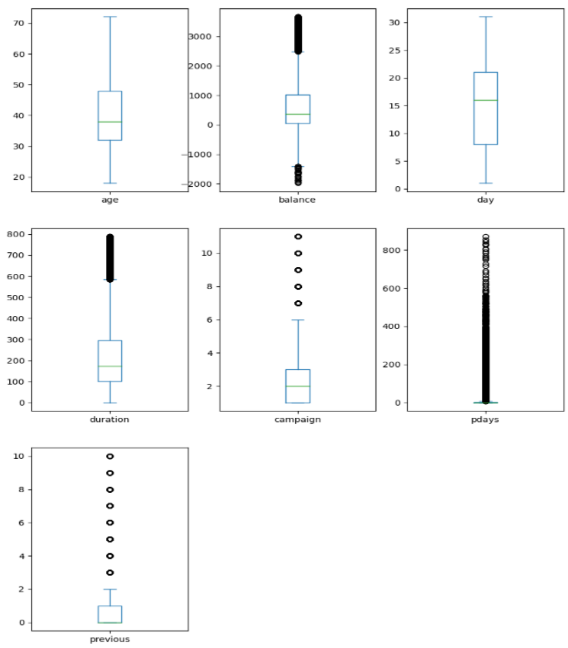
```
age          0
job          0
marital      0
education    0
default      0
balance      0
housing      0
loan         0
contact      0
day          0
month        0
duration     0
campaign     0
pdays        0
previous     0
poutcome     0
y            0
dtype: int64
```

```
data.info()
```

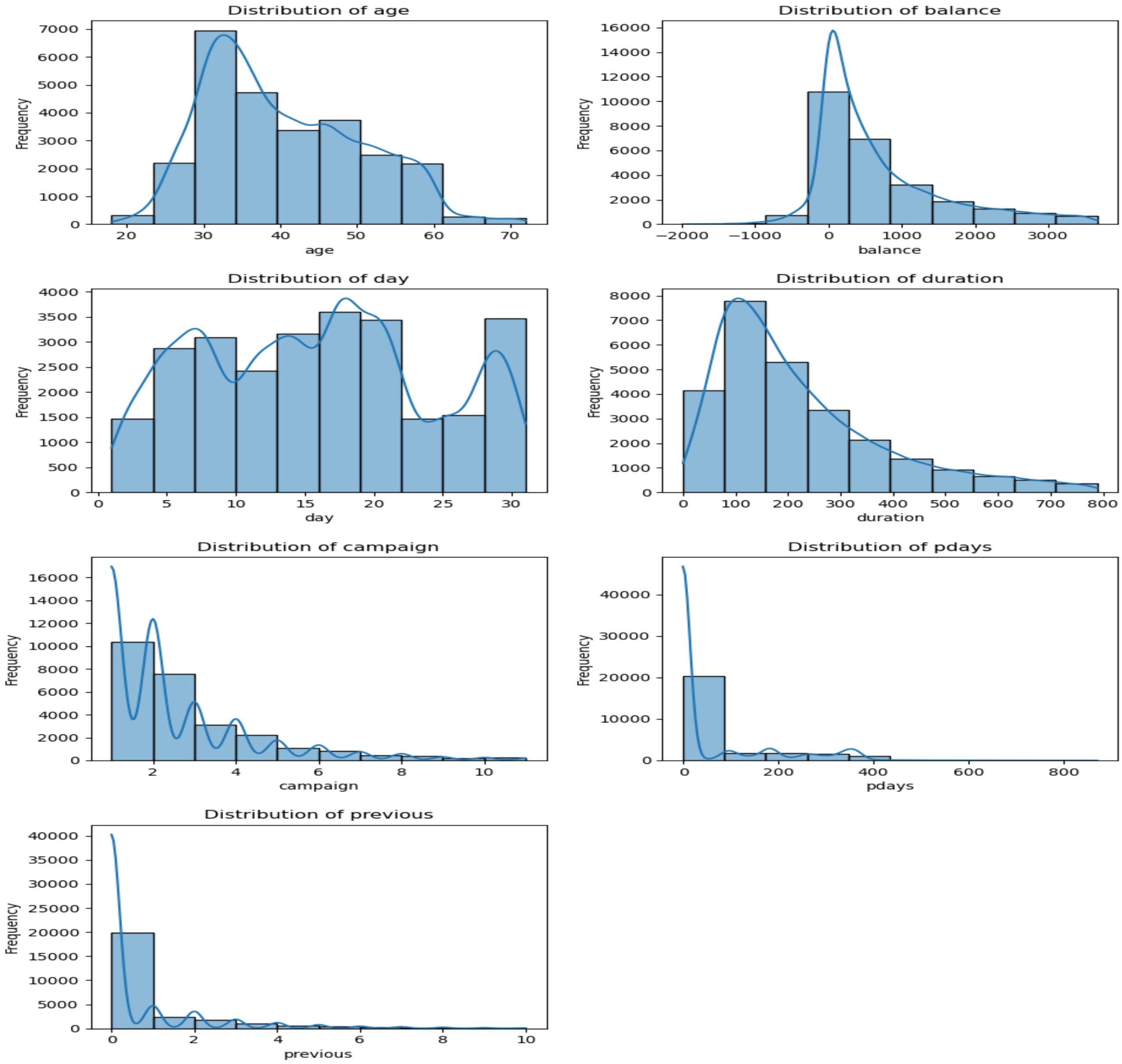
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45211 entries, 0 to 45210
Data columns (total 17 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   age        45211 non-null  int64
 1   job        45211 non-null  object
 2   marital    45211 non-null  object
 3   education  45211 non-null  object
 4   default    45211 non-null  object
 5   balance    45211 non-null  int64
 6   housing    45211 non-null  object
 7   loan       45211 non-null  object
 8   contact    45211 non-null  object
 9   day        45211 non-null  int64
 10  month      45211 non-null  object
 11  duration   45211 non-null  int64
 12  campaign   45211 non-null  int64
 13  pdays      45211 non-null  int64
 14  previous   45211 non-null  int64
 15  poutcome   45211 non-null  object
 16  y          45211 non-null  object
dtypes: int64(7), object(10)
memory usage: 5.9+ MB
```

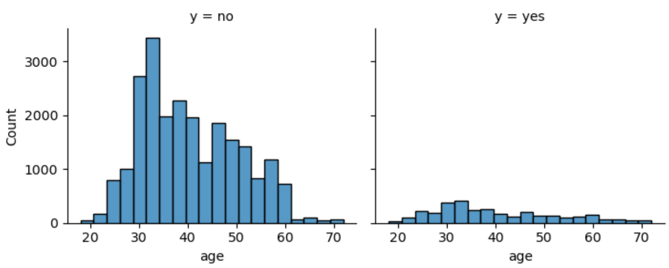
   - Identified and Handled Outliers using boxplot and IQR method respectively.

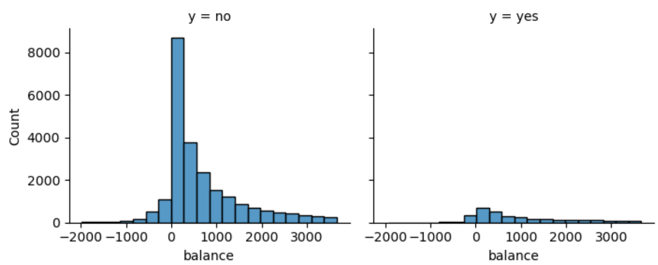


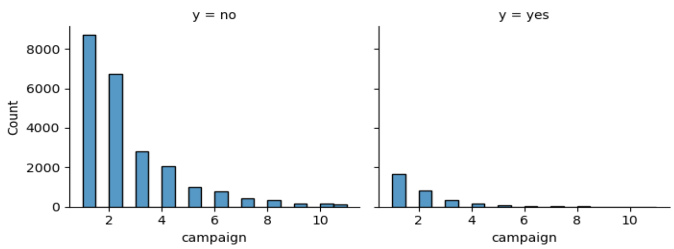


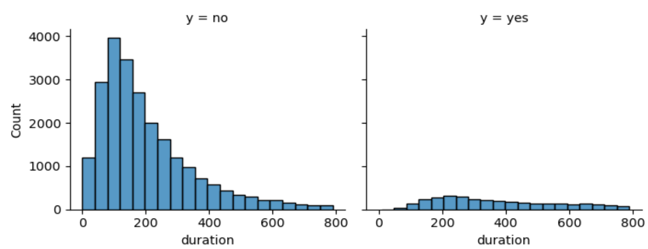
2. **Exploratory Data Analysis (EDA):**
    - Analysed distributions of numerical variables (e.g., age, balance).
    - Where age, balance, duration, campaign, pdays, previous all are right skewed

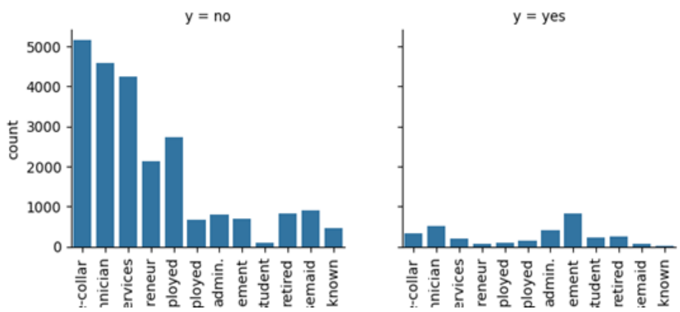


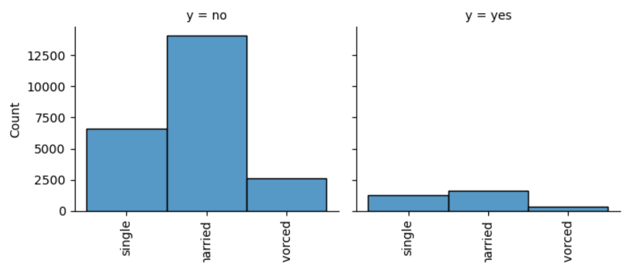


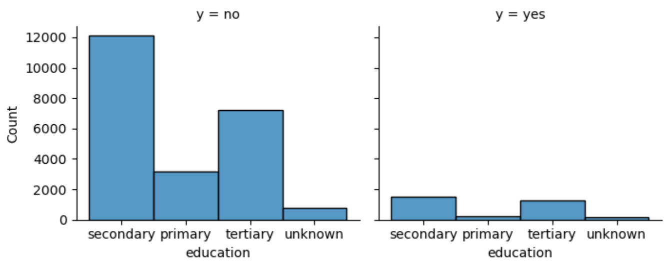


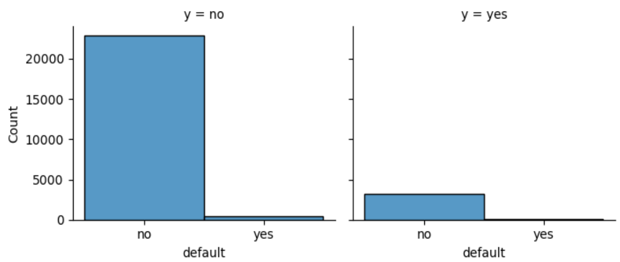


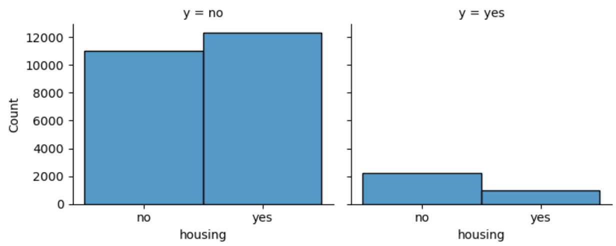


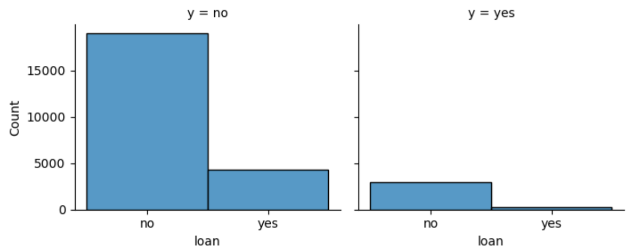


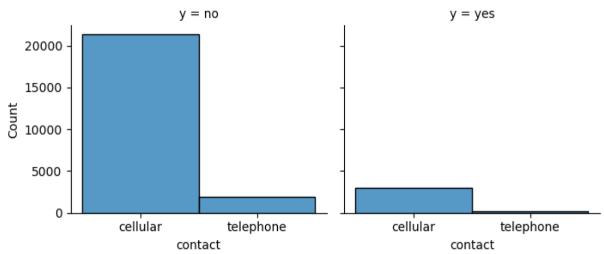






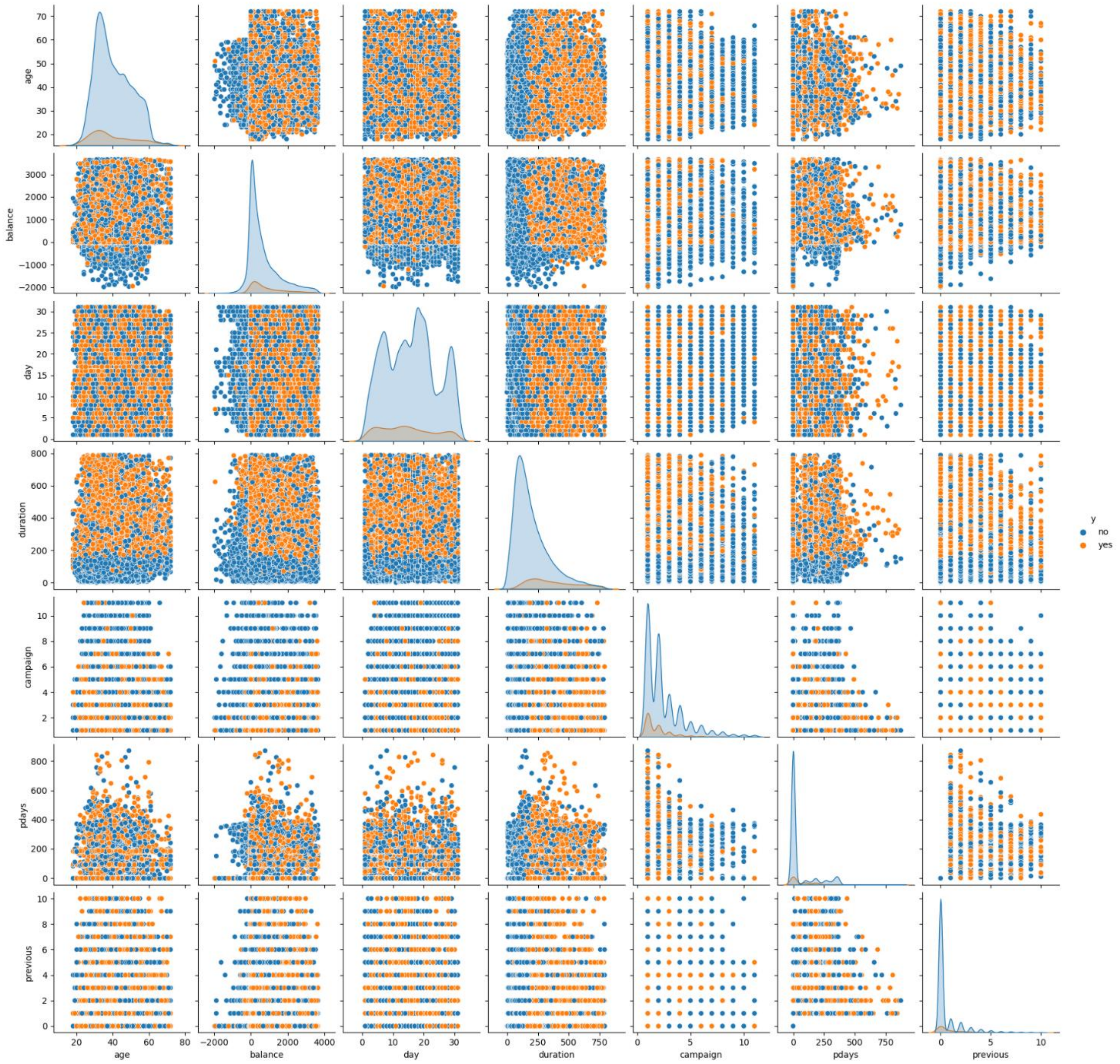






- **Age :** Younger demographic (age less than 40-50) have a higher chance to subscribe to a term deposit.
- **Balance :** Customers with a balance between 0 and 1000 are more likely to use a term deposit.
- **Campaign :** During the campaign to the customers with contact of 2 to 4 times have took subscription.
- **Duration :** The duration from 300 to 500 (sec) are more likely to subscribe to a term deposit by the customer.
- **Job :** Most of the customer who took term deposit are working in management, technician, administration and blue-collar.
- **Marital :** Married customers are interested in term deposite, followed by singles.
- **Education :** Majority customers contacted were secondary and tertiary educated.
- **Default :** Most of the customers who took subscription are not defaulter so this column can be dropped.
- **House Loan :** Majority of customers who took term deposit don't have house loan.
- **Loan(Personal loan) :** Majority customers who took subscription do not have personal loan.
- **Contact :** Majority customers were contacted via cellular.

o Examined relationships between numerical variables (e.g., job, marital status) and the target variable.



- **Age vs Balance** : There is non-linear relationship between age and balance. Subscribers are more likely to have positive balance regardless of age. among subscribers, middle aged customers (between 40 to 50 years) tend to have higher balances.
- **Age Vs Duration** : Relationship is non linear. Subscriptions are strongly concentrated in cases with longer durations of calls (>300 to <500) seconds across all age groups. middle aged customers (>40 to <50) tend to have longer calls leading to subscriptions.
- **Balance Vs Duration** : Relationship is non-linear. Subscriptions occur when balances are positive (>0) and durations are long (>300 seconds). Customers with high balances (>1000) are more likely to subscribe if the call duration is long.
- **Age Vs Campaign** : Non-linear negative relationship. Subscriptions are common in the first few campaigns(1-4), regardless of age. Middle aged customers (40 - 50 years) have higher subscription rates at lower campaign counts.
- **Balance Vs Campaign** : Relationship is non-linear. Positive balances and lower campaign counts(1-4) leads to subscriptions. Few subscriptions occur at higher campaign counts(>4) even for customers with high balances.

3. **Feature Engineering:**
   - Encoded categorical variables using label encoding and one-hot encoding.

```
data_encoded.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 26502 entries, 12657 to 45209
Data columns (total 32 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   age                26502 non-null  int64
 1   education          26502 non-null  int64
 2   balance            26502 non-null  int64
 3   day                26502 non-null  int64
 4   month              26502 non-null  int64
 5   duration           26502 non-null  int64
 6   campaign           26502 non-null  int64
 7   pdays              26502 non-null  int64
 8   previous           26502 non-null  int64
 9   y                  26502 non-null  int64
 10  job_admin.         26502 non-null  bool
 11  job_blue-collar    26502 non-null  bool
 12  job_entrepreneur   26502 non-null  bool
 13  job_housemaid      26502 non-null  bool
 14  job_management     26502 non-null  bool
 15  job_retired        26502 non-null  bool
 16  job_self-employed  26502 non-null  bool
 17  job_services       26502 non-null  bool
 18  job_student        26502 non-null  bool
 19  job_technician     26502 non-null  bool
 20  job_unemployed     26502 non-null  bool
 21  marital_divorced   26502 non-null  bool
 22  marital_married    26502 non-null  bool
 23  marital_single     26502 non-null  bool
 24  default_no         26502 non-null  bool
 25  default_yes        26502 non-null  bool
 26  housing_no         26502 non-null  bool
 27  housing_yes        26502 non-null  bool
 28  loan_no            26502 non-null  bool
 29  loan_yes           26502 non-null  bool
 30  contact_cellular   26502 non-null  bool
 31  contact_telephone  26502 non-null  bool
dtypes: bool(22), int64(10)
memory usage: 2.8 MB
```

4. **Data Visualization:**
   - Used bar charts, histograms, and boxplots to identify trends.
   - Created correlation pair plot to detect multicollinearity among numerical features.
5. **Model Selection:**
   - Logistic Regression (baseline model).
   - Tree-based methods (Random Forest, Decision Tree, Extreme Gradient Boosting).
   - Evaluated models using metrics such as Accuracy, Classification Report, Confusion Matrix, Roc Curve, AUC Score .
6. **Model Evaluation:**
   - Plotted ROC curves and calculated AUC to compare models.
   - The AUC (Area Under the ROC Curve) score summarizes the model's performance across all thresholds. A higher AUC indicates a better-performing model.
   - Analysed feature importance for interpretability.

3. Results

Model Performance:

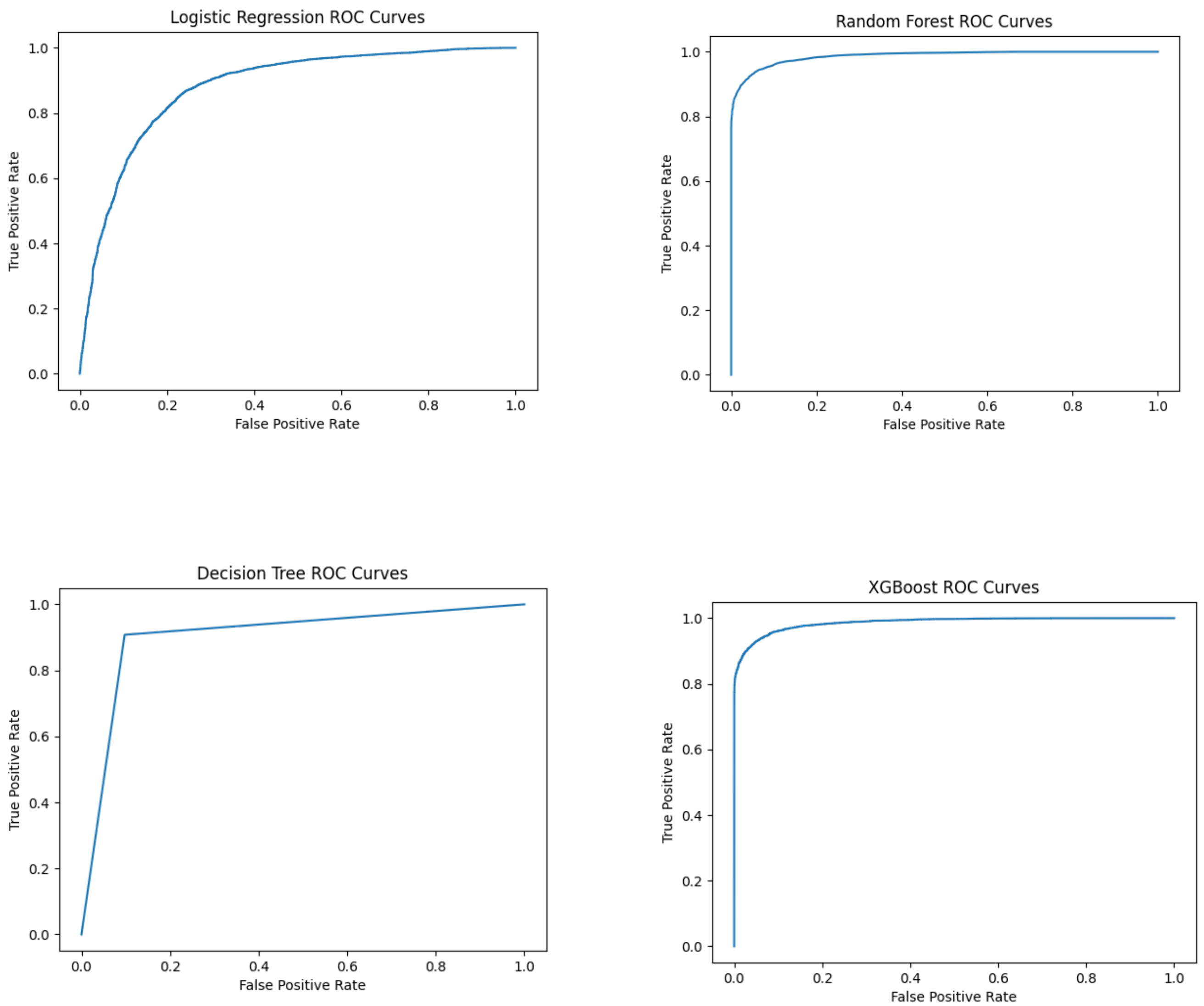
- **Logistic Regression:** Baseline accuracy of 80.30%
- **Random Forest:** Accuracy of 93.93%, with the highest AUC (0.9867).
- **Decision Tree:** Accuracy of 90.50%, with lower recall and precision compared to Random Forest.
- **Extreme Gradient Boosting:** Accuracy of 93.89%, but with slightly lower recall compared to Random Forest.

Feature Importance:

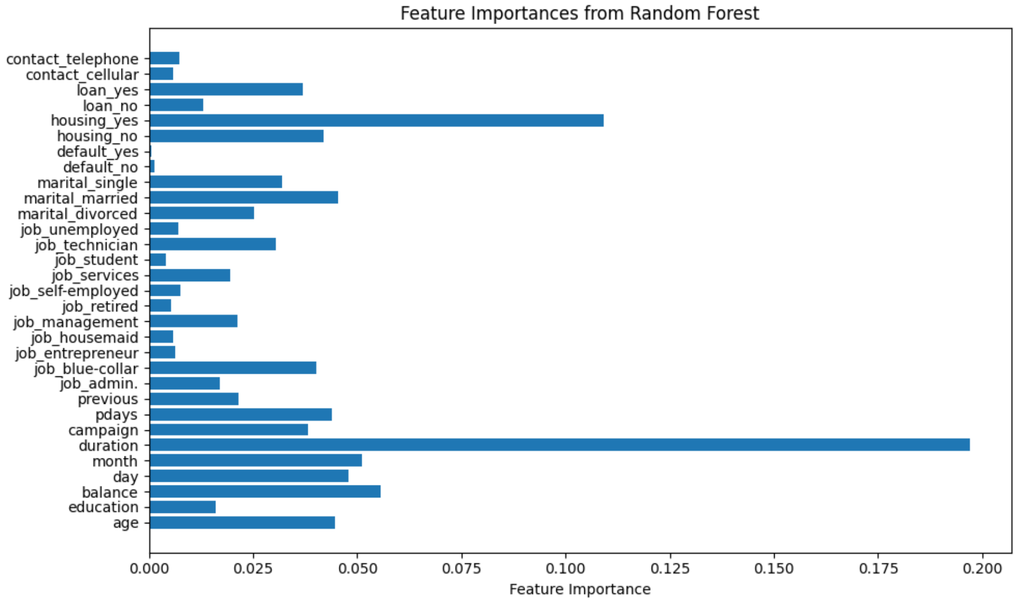
- The most important features influencing the prediction include **duration**, **age**, **balance**, and **campaign**.
- Partial dependence plots indicate that longer call durations and higher balances are positively associated with subscription.
- Categorical features like job and marital status also contributed significantly.

Visual Insights:

- **Figure 1:** ROC curves for all models.



- **Figure 2:** Feature importance plot from the Random Forest model.
    - Feature importance plot is a great way to understand which features have the most influence on our model's predictions.



(Important Features : duration, housing loan, balance, marital married, age, job blue collar)

4. Conclusion

- Random Forest outperformed other models in predictive accuracy and interpretability.
- Focus marketing efforts on clients with longer call durations and higher balances.
- Tailor strategies based on important features to improve campaign effectiveness.
- Continuously monitor and update the model to ensure its accuracy and relevance with new data.