BANK TERM DEPOSITS ANALYSIS & PREDICTION

Introduction

Term deposits are a type of fixed investment where customers deposit money with a financial institution for a predetermined period at an agreed interest rate. These deposits are a reliable and significant source of income for banks, as they provide a stable inflow of funds while offering customers a secure way to grow their savings.

Business Problem

Telephonic marketing is an effective but costly method for promoting term deposits. To optimize resources and improve conversion rates, the bank needs to identify customers most likely to subscribe to a term deposit before reaching out, ensuring more targeted and cost-efficient marketing efforts

Objectives

1. Identify Key Predictors: Analyze customer demographics, previous interactions, and campaign details to determine the most important factors influencing a customer's decision to subscribe to a term deposit.

2. Develop the Best Prediction Model: Build and evaluate multiple classification models to identify the most accurate model for predicting whether a customer will subscribe to a term deposit, ensuring optimal performance.

3. Optimize Marketing Strategies: Use the insights from the model to recommend targeted telephonic marketing strategies, focusing on high-potential customers to maximize conversion rates and reduce unnecessary outreach.

Metrics of success

- 1. Model Accuracy: Achieve over 80% accuracy in predicting customer subscription to a term deposit.
- 2. Precision & Recall: High precision and recall for identifying customers likely to subscribe (Class 1).
- 3. Feature Importance: Identify key factors influencing subscription decisions to optimize marketing strategies.
- 4. Conversion Rate: Increase conversion rates by targeting high-potential customers.
- 5. Outreach Efficiency: Reduce unnecessary outreach to non-potential customers.
- 6. Cross-validation Performance: Ensure consistent performance across different data subsets to avoid overfitting.

Dataset Overview

This dataset is related to the direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were conducted through phone calls, often requiring multiple contacts with the same client to determine whether they would subscribe to the product (bank term deposit).

Data source: https://www.kaggle.com/datasets/thedevastator/bank-term-deposit-predictions?resource=download

The target variable indicates whether the client subscribed (`"yes"`) or not (`"no"`).

The data folder contains two datasets:

- train.csv: Contains 45,211 rows and 18 columns, ordered by date (from May 2008 to November 2010).

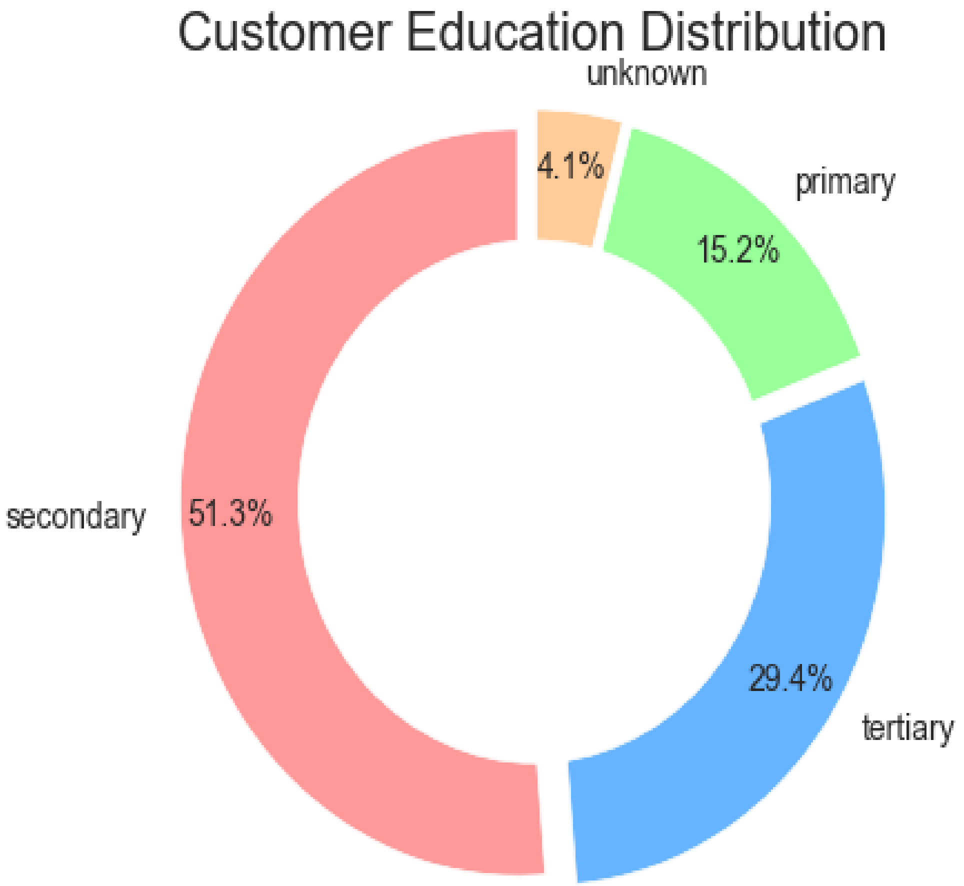
- test.csv: Contains 4,521 rows and 18 columns, representing 10% of the examples, randomly selected from `train.csv`.

Column Description

- 1. age`: Age of the client (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"` includes divorced or widowed).
- 4. `education`: Level of 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 a housing loan? (binary: `"yes"`, `"no"`).
- 8. `loan`: Has a personal loan? (binary: `"yes"`, `"no"`)

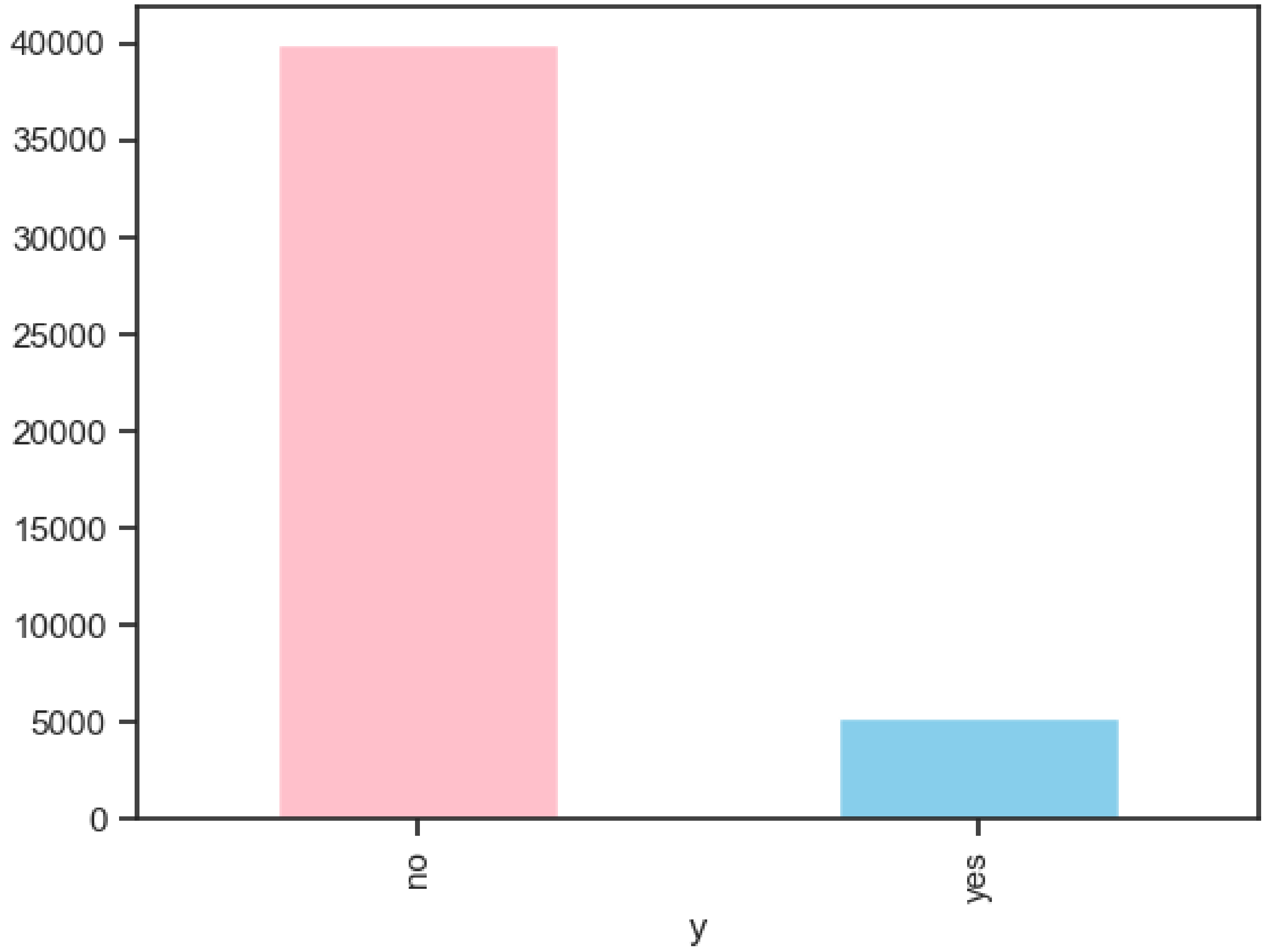
Exploratory Data Analysis

- Most customers had attained Secondary school level of education i.e., 51.3% followed by those with Tertiary level of education at 29.4%



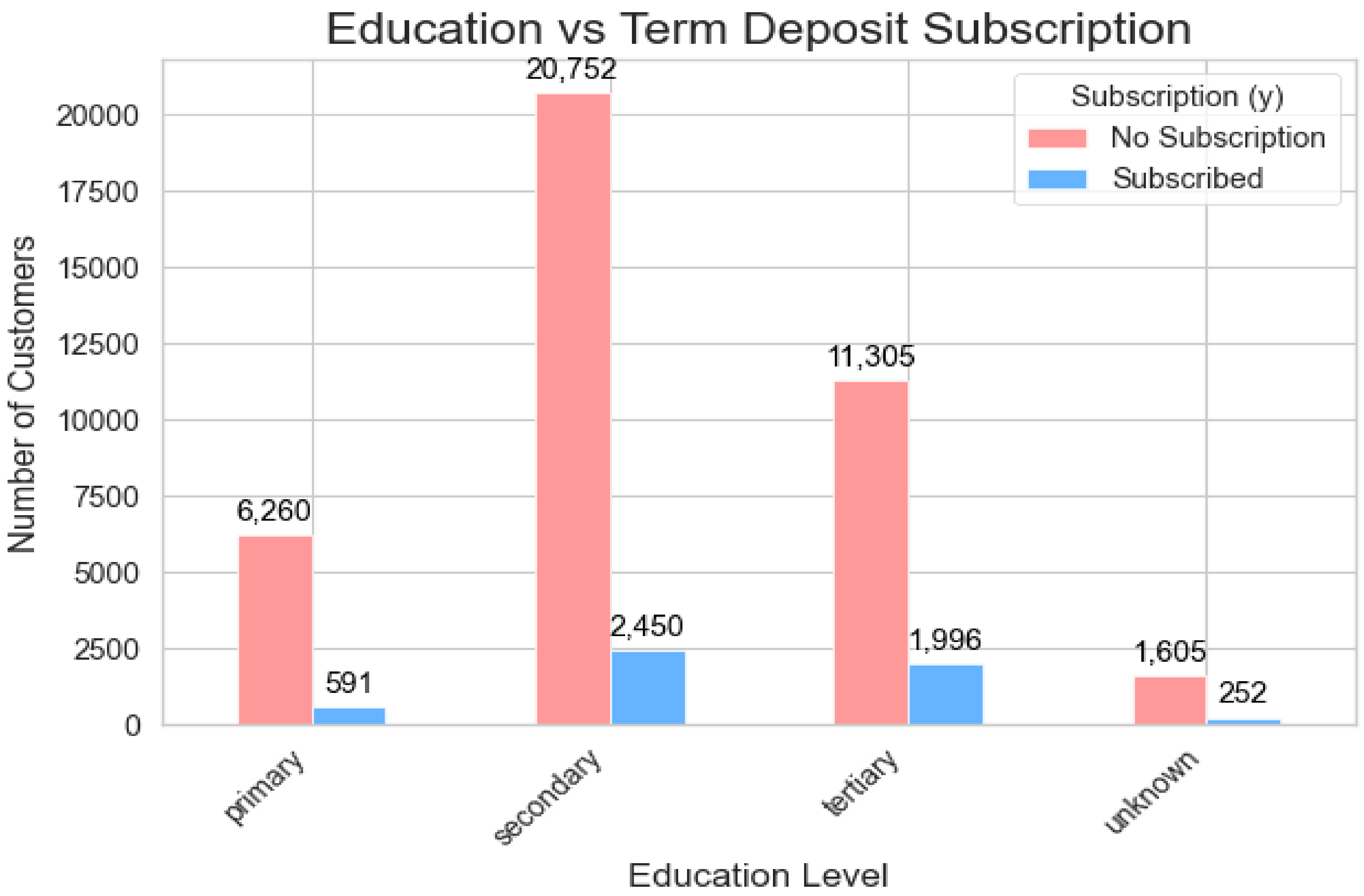
Analysis of 'y' Target Variable

- Majority of customers were non-subscribers to term-deposits



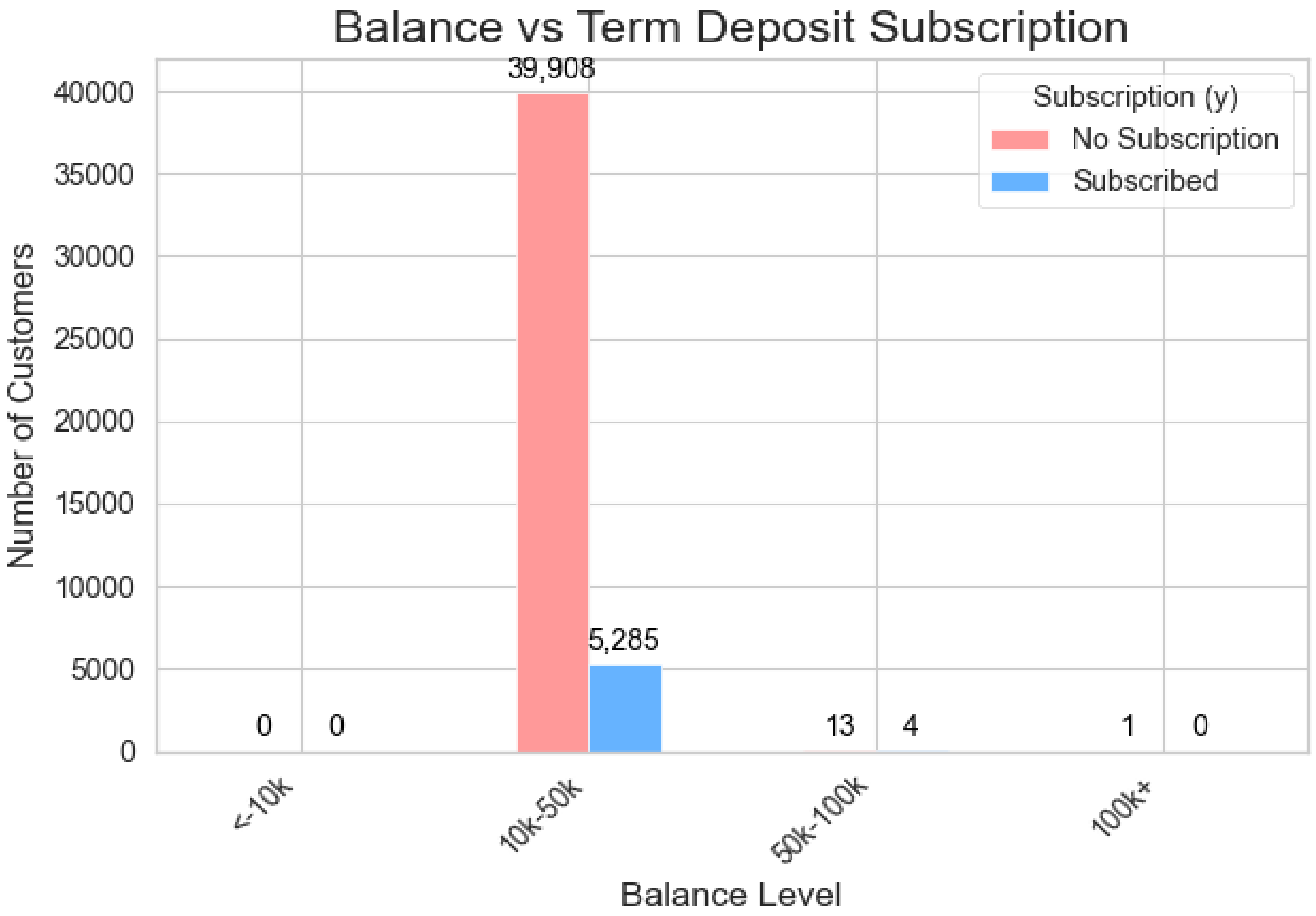
Education level vs Deposit Subscription

- Customers with Secondary school education level comprised the majority of those who had subscribed to term deposits followed by those with tertiary level of education.



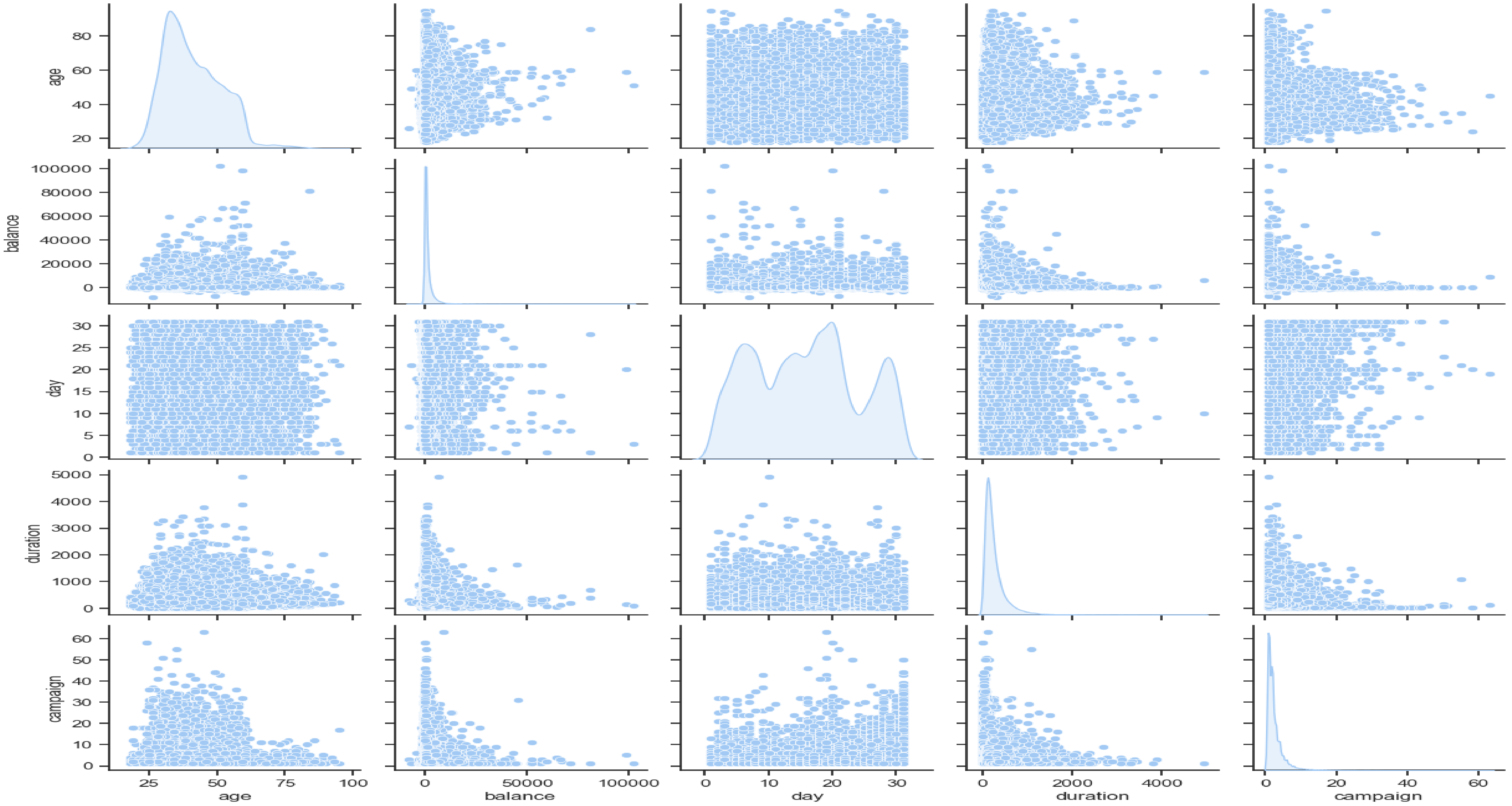
Bank account Balance vs Term deposit subscription

- Customers with bank account balance between 10K-10K formed the majority of those who subscribed to term deposits and this category of customers also comprised super majority of the non-subscribers



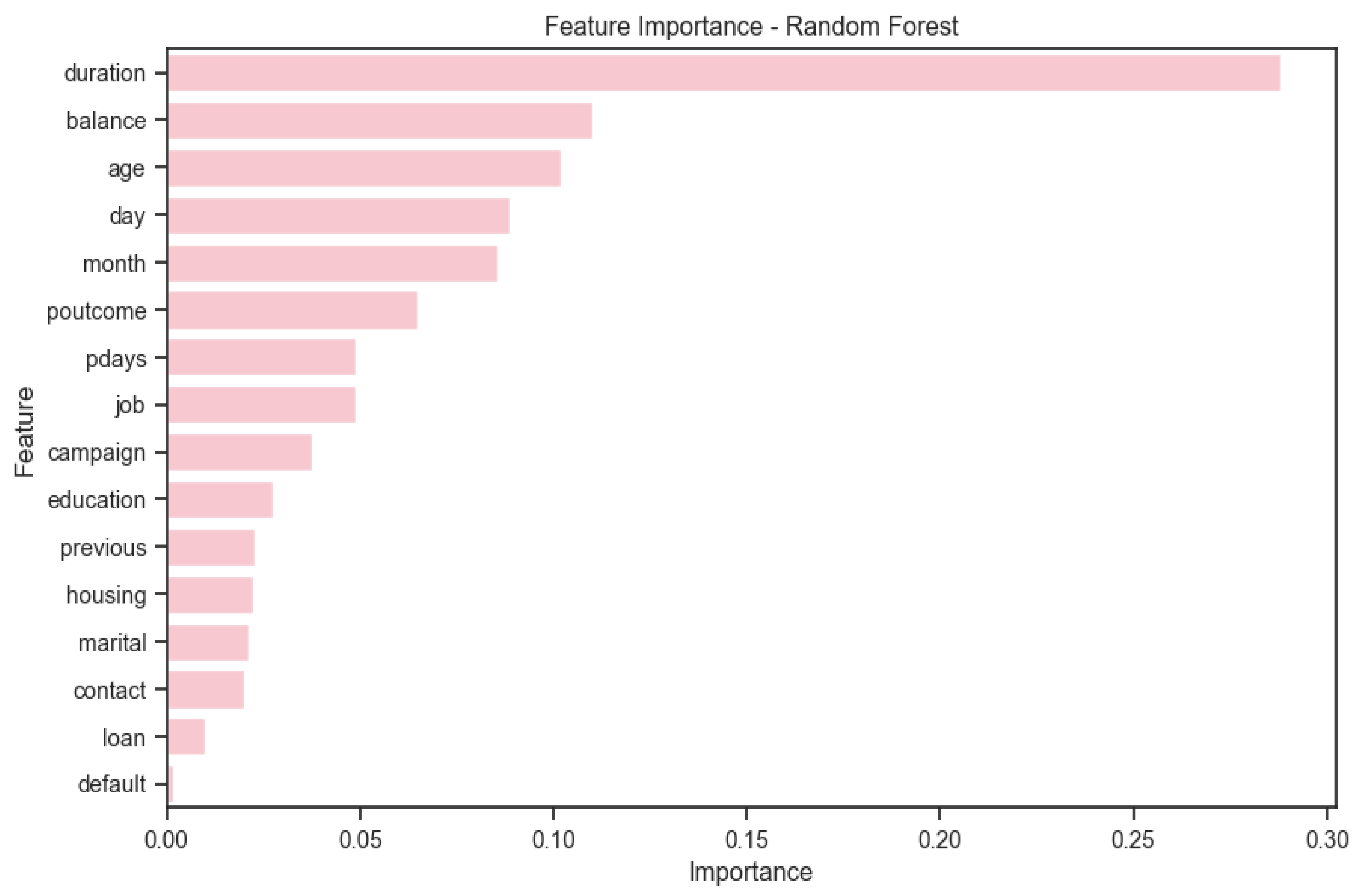
Distribution of Age

Pair plots



age
balance
day
duration
campaign
age
balance
day
duration
campaign
20
40
60
80
0
20000
40000
60000
80000
100000
0
5
10
15
20
25
30
0
1000
2000
3000
4000
5000
0
10
20
30
40
50
60
25
50
75
100
0
50000
100000
0
10
20
30
0
2000
4000
0
20
40
60

Feature importance



- Less important features include Default, loan, contact & Marital status
- Most important features include Duration, Balance and Age

MODEL PERFOMANCE EVALUATION

Base Model: Logistic Regression Model

Accuracy: 88.56%

Confusion Matrix:

```
[[3909   91]
 [ 426   95]]
```

Classification Report:

	precision	recall	f1-score	support
0	0.90	0.98	0.94	4000
1	0.51	0.18	0.27	521
accuracy			0.89	4521
macro avg	0.71	0.58	0.60	4521
weighted avg	0.86	0.89	0.86	4521

KNN Model

Accuracy: 90.09%

Confusion Matrix:

```
[[3890  110]
 [ 338  183]]
```

Classification Report:

	precision	recall	f1-score	support
0	0.92	0.97	0.95	4000
1	0.62	0.35	0.45	521
accuracy			0.90	4521
macro avg	0.77	0.66	0.70	4521
weighted avg	0.89	0.90	0.89	4521

RANDOM FOREST MODEL

```
Random Forest Accuracy: 96.11%
```

```
Confusion Matrix:
[[3980   20]
 [ 156  365]]
```

```
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.99      0.98      4000
           1       0.95      0.70      0.81       521

    accuracy                           0.96      4521
   macro avg       0.96      0.85      0.89      4521
weighted avg       0.96      0.96      0.96      4521
```

1. Accuracy: The model performs similarly on both training and test sets, with a slight increase in accuracy on the test set (96.11% vs. 95.96%).
2. Class 0 (Negative class):
   - High precision, recall, and F1-score on both sets, showing the model is very effective in correctly identifying the negative class.
3. Class 1 (Positive class):
   - Slight decrease in recall for Class 1 on the test set (70% vs. 68%), indicating the model is slightly better at identifying positive cases in the test set.
   - Precision for Class 1 remains very similar between the training and test sets (around 0.95–0.97).

PREDICTING PROBABILITY OF SUBSCRIPTION FOR EACH CUSTOMER

	subscription_prob
855	0.899611
1469	0.860702
1485	0.848232
3304	0.847622
2159	0.847171

	subscription_prob
30	0.762140
33	0.701522
38	0.754617
70	0.764783
83	0.807538
110	0.713135
156	0.731173
164	0.778260
328	0.726359
450	0.765965

CONCLUSIONS

- Key Factors: Duration of the call, balance, and age are the strongest predictors.
- Campaign Timing: The day, month, and timing of contact matter for successful subscriptions.
- Demographics and Financial Stability: Job type and balance are secondary but important indicators.
- Random Forest The model generalizes well from training to testing with consistent performance(96.11%), particularly in predicting the negative class (Class 0). The slight improvement in recall for the positive class (Class 1) on the test set may indicate better generalization or data characteristics in the test set. The model seems well-calibrated with no significant overfitting or underfitting.
- Customers with high `subscription_prob` values (e.g., 0.892162, 0.871691, 0.870282) have a high likelihood of conversion.
- - Customers with `subscription_prob` values between 0.75 and 0.80 represent the next tier of likely conversions. While not as urgent as the highest-probability customers, this segment still holds significant potential and should be targeted with focused outreach.

RECOMMENDATIONS

- Focus on longer, high-quality interactions

- Tailor campaigns based on age and balance.

- Utilize personalized outreach strategies to improve subscription rates.

- Focus on Random Forest for its overall performance and improve recall for the minority class by addressing class imbalance and tuning model parameters.

- Target High-Probability Customers: Focus marketing efforts on customers with high `subscription_prob` (e.g., 0.892162, 0.871691, 0.870282). These customers are highly likely to subscribe, requiring minimal outreach for better conversion rates.

- Use a Probability Threshold: Consider customers with probabilities around 0.75–0.80 for the next tier of targeted outreach. While not as urgent as the highest-probability customers, they still represent a valuable segment to target.

- Optimize Resource Allocation: Allocate resources efficiently by prioritizing customers with probabilities above 0.8. This ensures higher conversion rates with less effort on low-probability customers.

- Reduce Unnecessary Outreach: Deprioritize customers with lower probabilities (e.g., 0.716083, 0.725284), or approach them with different strategies, such as offering incentives or alternate products.

Thank you

Richard Mokaya
+254707751916
rmokaya1@gmail.com