Bank Marketing Prediction:
Predict if a Customer will Subscribe the Term Deposit after Direct Marketing from a Bank

Kuan-Hsuan(Jennifer) Chiu
Master of Information Systems
New York University

I. Project Background

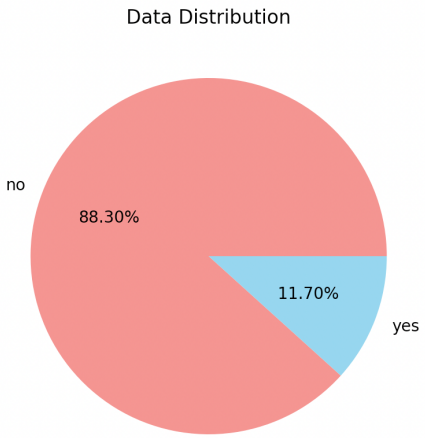
In the financial industry, bank marketing is crucial. It helps banks build and maintain relationships with their existing customers by promoting and selling banking products and services. However, banks often face budget constraints when it comes to disseminating information to their customers.

In this project, our aim is to apply data analytics to predict customers who are more likely to accept new promotions and, ultimately, make the most of the budget. We utilize the UCI Bank Marketing Prediction Dataset to forecast the likelihood of customer subscription to banking products. By leveraging a combination of machine learning techniques, including logistic regression, decision trees, and random forests, we develop predictive models to identify the key factors that influence subscription decisions.

II. Data Overview

The dataset is associated with direct marketing campaigns (phone calls) conducted by a Portuguese banking institution. The classification goal is to predict whether the client will subscribe to a term deposit (variable y). The dataset consists of 45,211 instances, including 39,922 instances labeled as "no" and 5,289 instances labeled as "yes," resulting in an imbalanced dataset.

The dataset can be categorized into three main types:
1. Bank client data: age, job, marital, education, default, balance, housing, loan
2. Attributes related to the last contact of the current campaign: contact, day, month, duration
3. Other attributes: campaign, pdays, previous, poutcome



III. Data Prediction

The primary objective of this project is to pinpoint the most responsive customers, enabling the company to efficiently allocate its resources. To achieve this goal, we commence by cleansing the data and subsequently implementing a classification model. The outcomes will be categorized into two main groups: those who respond "yes" to term deposits and those who respond "no" to term deposits.

Data Preparation

1. Check for any missing values.
2. Transform categorical data types into ordinal data types by applying dummy coding.
3. Address the imbalanced data by employing the oversampling method, which leads to a total of 63,884 instances, comprising 31,942 instances labeled as "no" and 31,942 instances labeled as "yes."
4. Split the dataset into 80% training data and 20% testing data.

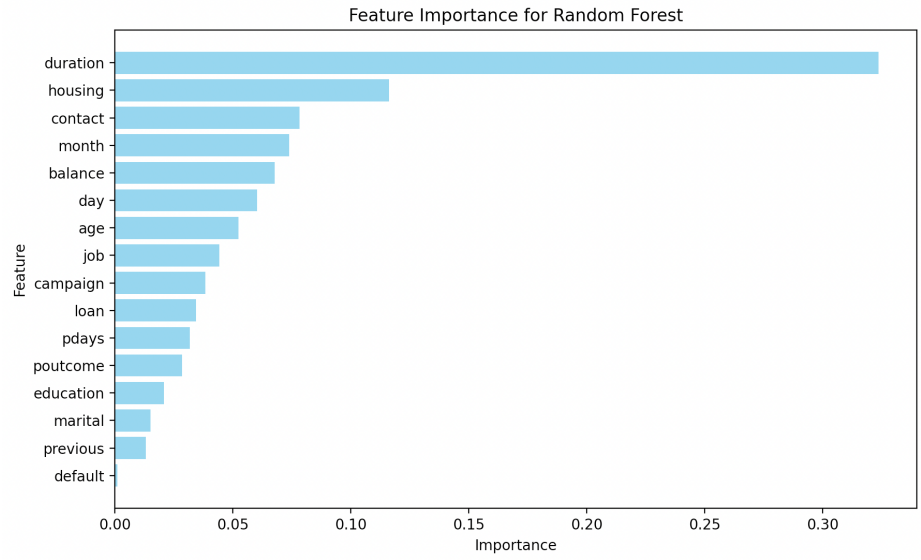
Modeling

We employ three classification algorithms: logistic regression, decision tree, and random forest. We identify the best-performing model and utilize it to discern the features that influence customers' decision behavior.
(Python Code: https://github.com/jenniferchiutw/predictive_analytics/tree/main/bank_marketing)

Evaluation

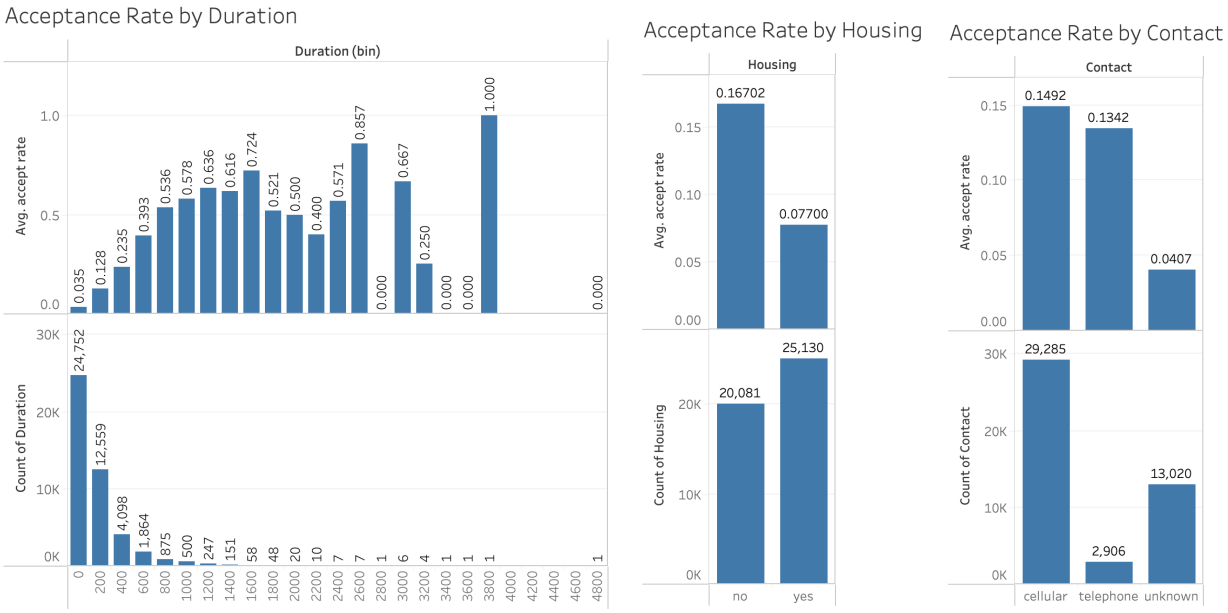
Model	**Accuracy**	**Recall**	**F1 Score**	**ROC AUC**
Logistic Regression	0.79	0.69	0.44	0.82
Decision Tree	0.85	0.58	0.48	0.74
Random Forest	0.88	0.65	0.57	0.91

Since the random forest model is the best, we employ it to ascertain the importance of features, identifying those that have the most significant impact on customers' decisions.



IV. Business Findings

Based on the analysis, the most responsive customers exhibit the following characteristics:
Feature 1 – Duration: Individuals with longer last contact durations.
Feature 2 – Housing: Individuals without housing loans.
Feature 3 – Contact: Individuals who use cellular communication.



V. Conclusion

The primary objective of this project is to identify the most responsive customers. We applied logistic regression, decision tree, and random forest models, and determined that the random forest model outperformed the others. Subsequently, we utilized this model to identify the features that have the greatest impact on customers' behavior. Our findings indicate that duration, housing, and contact are the most influential factors. As a result, when conducting bank marketing, companies can prioritize targeting these types of customers.

References

1. Bank Marketing Dataset: https://archive.ics.uci.edu/dataset/222/bank+marketing
2. A Predictive Analytics Framework to Anomaly Detection: https://ieeexplore.ieee.org/document/9179589