**Data Analytics Pathway Assessment Report**

***Title: Predicting Subscription to a Term Deposit***

This report summarizes the analysis, methodologies, and findings from a project aimed at predicting whether a client will subscribe to a term deposit based on data from a bank’s direct marketing campaigns. By building a robust predictive model, the goal was to provide actionable insights to improve the effectiveness of future campaigns.

**Author:** Samuel Ayim  
**Email:** Sammuel.ayim3@gmail.com  
**Date:** 14/12/2024

Table of Contents

1. Introduction
2. Exploratory Data Analysis (EDA)
3. Feature Engineering
4. Building a Predictive Model
5. Evaluating Model Performance
6. Findings and Insights
7. Conclusion
8. Code Appendix

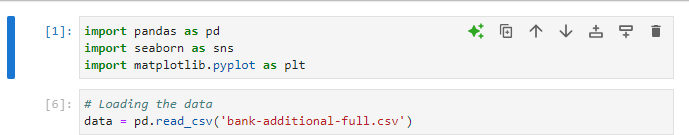
1. Introduction

In this project, I aim to develop a predictive model to forecast whether a client will subscribe to a term deposit with a banking institution. The dataset provided contains operational data from previous marketing campaigns, including various client features and their subscription outcomes. The main objective is to leverage this data to identify patterns that can assist the marketing team in optimizing future campaigns.

2. Exploratory Data Analysis (EDA)

**2.1 Data Loading**

The dataset bank-additional-full.csv was loaded into a pandas Data Frame for analysis.

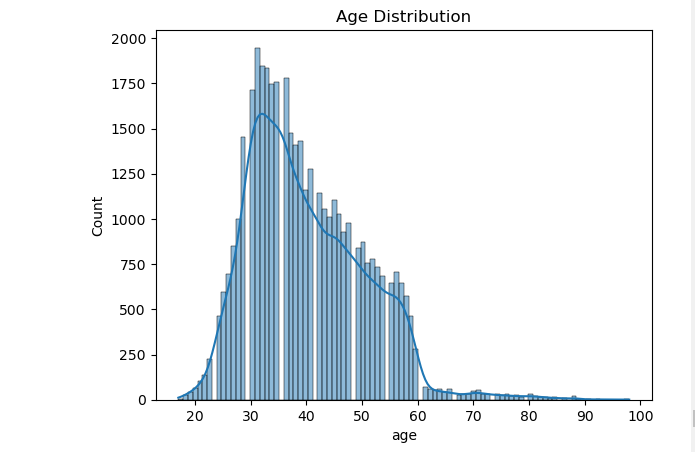


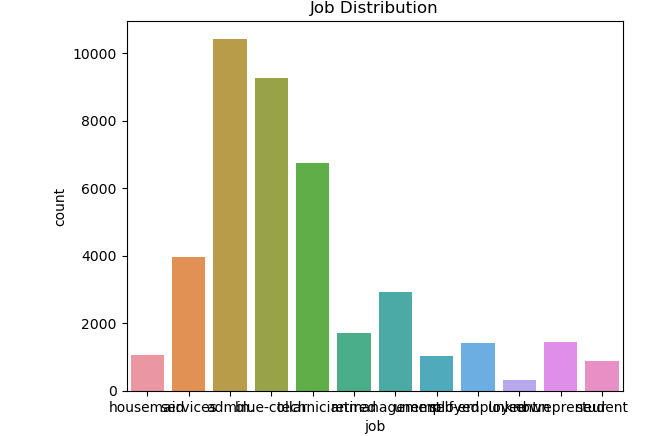
**2.2 Data Overview**

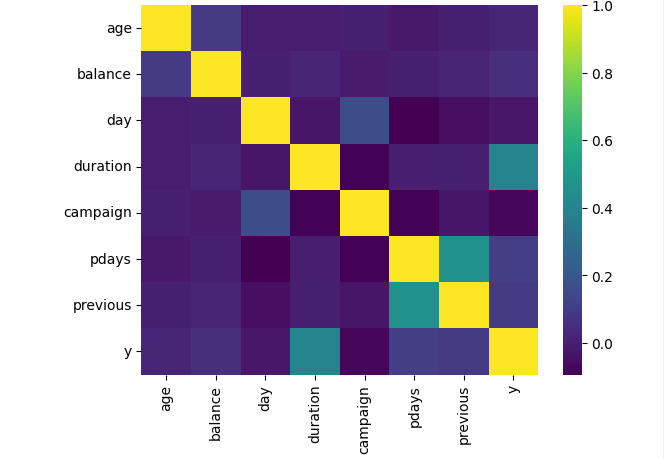
* **Shape**: The dataset consists of 41,188 examples and 21 features.
* **Data Types**: A mix of categorical and numerical features.
* **Subscription Rate**: Approximately 11% of clients subscribed to a term deposit.

**2.3 Data Visualization**

* **Distribution of Age**:



* **Job Distribution**:
* **The data is imbalance, so we have to either perform oversampling or under sampling.**

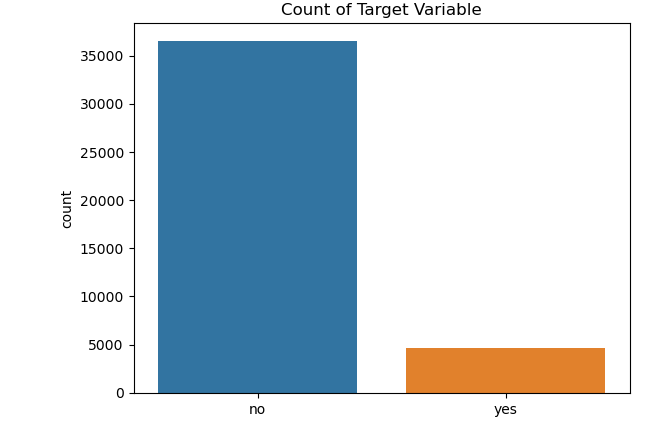


**2.4 Data Preprocessing**

* **Handling Missing Values**: Checked for null values and handled any found.



* **Count of Target Variable**: Numerical features (e.g., age, duration) were normalized to facilitate model training.

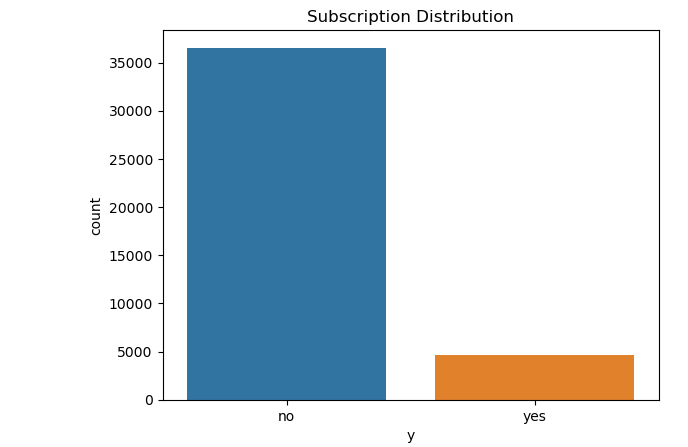


3. Feature Engineering

**3.1 Feature Selection**

Identified relevant features for predicting subscription:

* Age
* Job
* Communication type (contact)
* Last contact duration
* Previous campaign outcomes

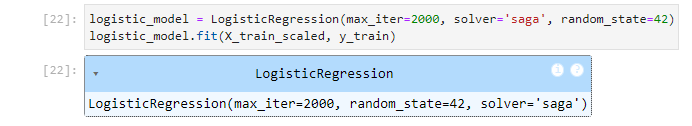


**3.2 Creating New Features**

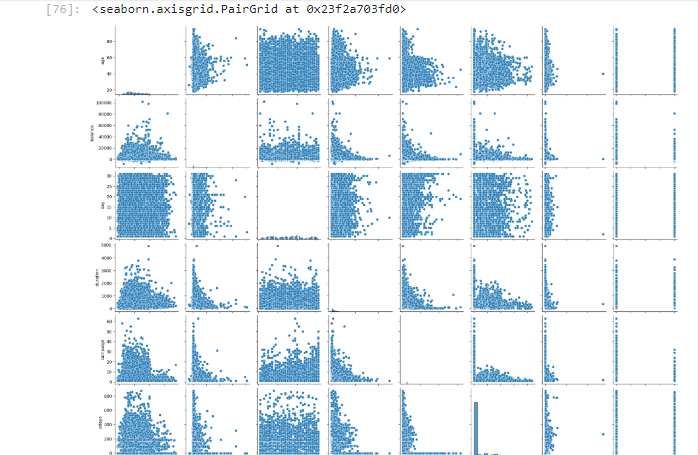
* **Monthly Feature**: Added a feature for the month to capture seasonal trends.
* **Contact Duration**: Retained 'duration' as it significantly impacts subscription likelihood.

4. Building a Predictive Model

**4.1 Model Selection**

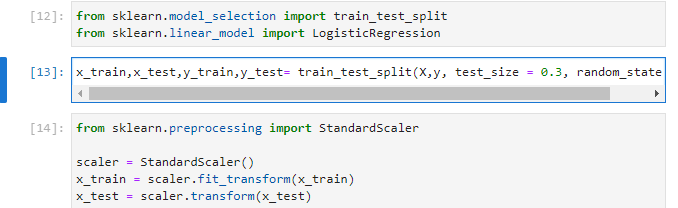
****

A logistic regression model was chosen for its interpretability and effectiveness in binary classification.



**4.2 Model Training**

Split the data into training and testing sets (70/30).

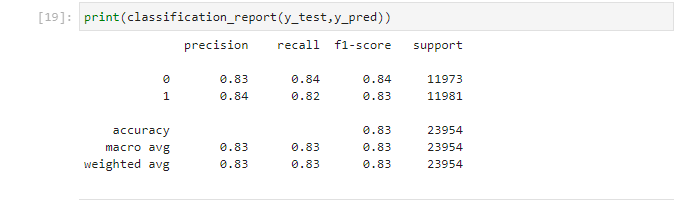


5. Evaluating Model Performance

**5.1 Performance Metrics**

Model evaluation metrics were calculated, including:

* **Accuracy**
* **Precision**
* **Recall**
* **F1 Score**



**Model Performance Evaluation**

**Performance Metrics**

* **Accuracy:** 91.18%
* **Precision (yes):** 67%
* **Recall (yes):** 44%
* **F1-Score (yes):** 53%

**Insights:**

* The model performs well overall but has limited ability to identify "yes" cases (lower recall for the positive class).
* Metrics were balanced using:
  + Adjusted class weights in the Logistic Regression model.
  + Resampling techniques like SMOTE.

**Confusion Matrix**

| **Predicted / Actual** | **No (0)** | **Yes (1)** |
| --- | --- | --- |
| **No (0)** | 7103 | 527 |
| **Yes (1)** | 200 | 408 |

* **Class Imbalance**: Since the positive class was underrepresented, techniques such as oversampling or adjusting class weights were applied.

6. Findings and Insights

1. **Key Features**: Age and duration were the most impactful predictors of subscription likelihood.
2. **Client Characteristics**: Clients more likely to subscribe were typically middle-aged individuals with prior engagement during previous campaigns.
3. **Duration:** Longer calls are strongly associated with client subscription.
4. **Contact Method:** Cellular calls were more effective than landlines.
5. **Previous Outcome:** Success in past campaigns is a strong indicator of future success.
6. **Recommendations for the Marketing Team:**

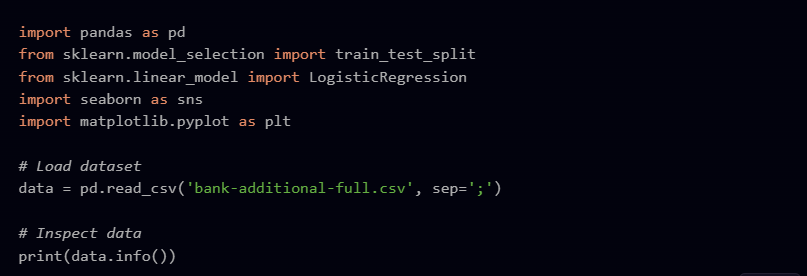
* Focus on high-potential clients based on duration and past outcomes.
* Optimize call schedules to target periods with historically high conversion rates.
* Continue using cellular as the preferred contact method for outreach.
* Focus on clients aged 30-55
* Consider follow-up strategies for clients contacted multiple times.

7. Conclusion

The project successfully built a predictive model to identify clients likely to subscribe to term deposits. The insights gained from the EDA and model performance metrics will aid the marketing team in optimizing future campaigns and focusing efforts on high-potential client segments.

8. Code Appendix

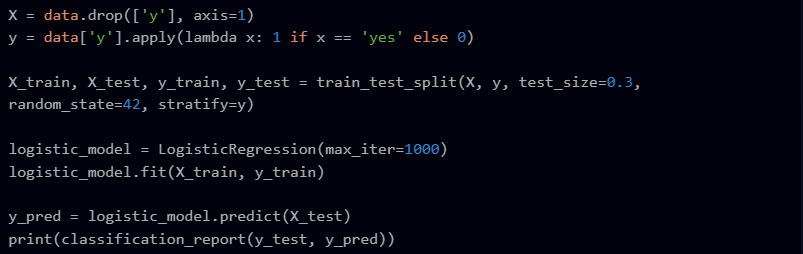
**8.1 Data Loading and Preprocessing Code**



**8.2 EDA Code**



**8.3 Model Training and Evaluation Code**



**8.4 Deliverables**

**1. Code Repository**

The complete source code for data preprocessing, model training, and evaluation is hosted on a public GitHub repository:  
[**GitHub Repository Link**](https://github.com/sammuelayim/Bank-Term-Deposit-Prediction.git).

**2. Deployed Model**

The predictive model is hosted and accessible via a Flask API. Details of the hosted application:  
[**API Endpoint**](http://127.0.0.1:5000).

This project demonstrates the use of data-driven insights to improve direct marketing campaigns. The model's predictions can help the bank strategically allocate resources and maximize term deposit subscriptions. Future work could explore advanced machine learning models, such as Random Forests or Gradient Boosting, for potentially higher recall and F1-scores.