**Data Analytics Pathway Assessment Report**

As a member of the Azubi Africa data analytics team, my task was to build a predictive model that determines the likelihood of a client subscribing to a term deposit based on the features provided in the dataset attached to the assessment. This is a brief report detailing the methods used in building the predictive model, key findings of the Exploratory Data Analysis (EDA), model insights, common characteristics of clients likely to subscribe and actionable recommendations for the marketing team.

**Methods Used in Building the Predictive Model**

1. **Data preparation and exploratory data analysis (EDA)**

The dataset had no missing values. 12 duplicate rows were removed, leaving 41,176 unique rows, and column names were standardized for consistency (e.g., "emp.var.rate" became "emp\_var\_rate"). Summary statistics and missing value checks were performed. Histograms were used to visualise the distribution of the numerical variables and bar plots visualized distributions of the categorical variables. Correlation heatmaps and boxplots were also used to identify relationships and outliers.

1. **Feature Engineering**

The categorical variables were encoded using label encoding. The data was split into training and test sets, with features scaled using StandardScaler.

1. **Modeling**

A Random Forest Classifier was trained on the dataset.

1. **Model Evaluation**

Performance was assessed using confusion matrices, classification reports, and ROC-AUC score. Visualised the ROC curve and confusion matrix.

**Key Findings of the Exploratory Data Analysis (EDA)**

1. **Numerical Insights:**

**Duration**: Clients with longer call durations were more likely to subscribe.

**Age**: A higher proportion of clients in the 30-60 age range subscribed.

**Previous**: The number of previous contacts positively influenced subscription likelihood.

High correlation was observed between **euribor3m**, **emp\_var\_rate**, and **nr\_employed**.

1. **Categorical Insights:**

**Job**: Higher subscription rates among clients with jobs in management, blue-collar, and technician roles.

**Marital Status**: Married clients were less likely to subscribe than single or divorced ones.

**Education**: Higher education levels correlated with increased subscriptions.

**Month**: May had the highest number of calls, while March had a higher subscription rate.

**Model Insights**

1. **Feature Importance:**

Using a Random Forest Classifier, the most impactful features for predicting subscriptions are:

‘Duration’ of the call.

‘Contact method’ (e.g., cellular).

‘Month’ of the campaign.

‘Previous’, ‘pdays’ (days since last contact), and ‘euribor3m’ (interest rate) also influenced the outcome significantly.

1. **Model Performance**:

**Confusion Matrix** shows reasonable classification results.

**Classification Report** indicates that Precision, recall, and F1 scores vary, but the model performs best in identifying non-subscribers.

**ROC AUC Score**: 0.94 indicates good predictive power.

**Common Characteristics of Clients Likely to Subscribe**

* **Age:** 30-60 years old.
* **Job:** Management, blue-collar, and technician roles.
* **Education:** Higher education levels.
* **Previous Contacts:** More previous interactions with the bank.
* **Call Duration:** Longer call durations.

**Actionable Recommendations for the Marketing Team**

1. **Targeting**:

* Focus on clients aged 30-60, especially in management or technical roles.
* Prioritize those with a history of multiple previous contacts.

1. **Timing**:

* Focus on clients aged 30-60, especially in management or technical roles.
* Prioritize those with a history of multiple previous contacts.
* Target campaigns in March and avoid low-performing months like December.

1. **Engagement**:

* Aim for longer, more personalized calls to improve subscription rates.
* Develop educational content for clients with lower education levels to boost understanding and trust.

## **Summary**

* The model achieved an overall accuracy of 91%, with a strong ability to correctly identify clients who would not subscribe.
* The model’s ROC-AUC score of 0.944 indicates a high level of discrimination between those who will and will not subscribe.
* However, the model had a lower recall for predicting positive cases (clients who subscribe), with a recall of 51% and a precision of 65%, indicating room for improvement in identifying potential subscribers.

**Conclusion**

This prediction has attained a solid accuracy of 91%, showing that it is effective in identifying clients who are less likely to subscribe.

However, the model’s precision and recall for clients who subscribe suggest that there is room for improvement, particularly in reducing the number of false negatives (as seen in the confusion matrix), by refining the model and exploring alternative algorithms.