**Project Title**

**BANK MARKETING: Predicting Whether the Customer Will Subscribe to Term Deposit (FIXED DEPOSIT) or not.**

**1. PROBLEM DEFINITION**

The financial landscape is ever-evolving, and banks must continuously innovate to sustain and grow their revenue streams. For a Portuguese bank facing a revenue decline, a detailed investigation revealed that clients were not depositing as frequently as before. Term deposits, in particular, were identified as a crucial product for the bank. A term deposit is a cash investment held at a financial institution for a fixed amount of time, providing the bank with a steady capital influx. The bank can then invest these funds in higher-gain financial products to generate profit. Additionally, term deposit clients are more likely to purchase other financial products, such as insurance or investment funds, thus increasing the bank's revenue further.

Given the effectiveness of telephonic marketing campaigns but their high cost due to the need for extensive call centers, it is essential to identify clients with a higher likelihood of subscribing to term deposits. By focusing marketing efforts on these clients, the bank can optimize its resources and enhance its campaign's success rate.

Our objective in this project is to develop a machine learning model that can predict whether a client will subscribe to a term deposit based on various attributes, including personal details and call information.

**2. DATA ANALYSIS**

The dataset used for this project is related to direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if a client will subscribe to a term deposit product. We are provided with two files: train.csv for training the model and test.csv for making predictions.

**Dataset Attributes:**

* **ID**: Unique client ID
* **age**: Age of the client
* **job**: Type of job
* **marital**: Marital status of the client
* **education**: Education level
* **default**: Credit in default
* **housing**: Housing loan
* **loan**: Personal loan
* **contact**: Type of communication
* **month**: Contact month
* **day\_of\_week**: Day of week of contact
* **duration**: Contact duration
* **campaign**: Number of contacts performed during this campaign to the client
* **pdays**: Number of days that passed by after the client was last contacted
* **previous**: Number of contacts performed before this campaign
* **poutcome**: Outcome of the previous marketing campaign
* **Subscribed (target)**: Whether the client subscribed to a term deposit (YES/NO)

**3. EDA CONCLUDING REMARKS**

Exploratory Data Analysis (EDA) is a crucial step in understanding the dataset and identifying patterns, anomalies, and relationships between variables. Through EDA, we can make informed decisions about data preprocessing and feature engineering, which directly impact the performance of our machine learning models.

**Key Findings from EDA:**

1. **Age Distribution**:
   * The age of clients ranges widely, with a notable concentration in the 30-40 age bracket. Younger and older clients exhibit different subscription rates, potentially indicating different financial priorities or awareness levels.
2. **Job and Marital Status**:
   * Clients' job types and marital statuses show varying subscription rates. For instance, clients in managerial positions or married clients might have different financial stability and goals compared to students or single clients.
3. **Education Level**:
   * Education levels also play a role, with higher-educated clients potentially being more aware of financial products like term deposits.
4. **Previous Outcomes and Contact Duration**:
   * The outcome of previous campaigns and the duration of the current contact are strong indicators of subscription likelihood. Clients who responded positively in the past or had longer conversations are more likely to subscribe.
5. **Month and Day of Contact**:
   * The time of the year and day of the week when the client is contacted can influence the outcome. There are peaks and troughs that could align with clients' availability and financial planning cycles.

**4. PREPROCESSING PIPELINE**

Preprocessing is a critical step to ensure the dataset is clean and suitable for machine learning algorithms. Our preprocessing pipeline involves the following steps:

**Step 1: Encoding Categorical Variables**

Categorical variables, such as job type, marital status, and education, need to be converted into numerical representations. We use Label Encoding for this purpose.

**label\_encoders = {}**

**categorical\_features = ['job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', 'poutcome']**

**for feature in categorical\_features:**

**if feature in train\_data.columns:**

**label\_encoders[feature] = LabelEncoder()**

**train\_data[feature] = label\_encoders[feature].fit\_transform(train\_data[feature])**

**Step 2: Encoding Target Variable**

The target variable, subscribed, is also encoded to convert the binary 'YES/NO' responses into numerical format (1 for YES, 0 for NO).

**train\_data['subscribed'] = train\_data['subscribed'].map({'yes': 1, 'no': 0})**

**Step 3: Separating Features and Target**

We separate the features (independent variables) from the target variable (dependent variable).

**X = train\_data.drop(columns=['ID', 'subscribed'])**

**y = train\_data['subscribed']**

**Step 4: Splitting Data into Training and Validation Sets**

To evaluate our model's performance, we split the data into training and validation sets.

**X\_train, X\_val, y\_train, y\_val = train\_test\_split(X, y, test\_size=0.2, random\_state=42)**

**Step 5: Scaling Numerical Features**

Scaling numerical features ensures that features with larger ranges do not dominate those with smaller ranges. We use StandardScaler for this purpose.

**scaler = StandardScaler()**

**numerical\_features = ['age', 'duration', 'campaign', 'pdays', 'previous', 'balance']**

**X\_train[numerical\_features] = scaler.fit\_transform(X\_train[numerical\_features])**

**X\_val[numerical\_features] = scaler.transform(X\_val[numerical\_features])**

**5. BUILDING MACHINE LEARNING MODELS**

With the preprocessed data, we proceed to build and train our machine learning model. For this project, we use a RandomForestClassifier due to its robustness and ability to handle both numerical and categorical data effectively.

**Training the Model**

**clf = RandomForestClassifier(random\_state=42)**

**clf.fit(X\_train, y\_train)**

**Evaluating the Model**

We evaluate our model using accuracy score, confusion matrix, and classification report on the validation set.

**y\_pred = clf.predict(X\_val)**

**print("Accuracy:", accuracy\_score(y\_val, y\_pred))**

**print("Confusion Matrix:\n", confusion\_matrix(y\_val, y\_pred))**

**print("Classification Report:\n", classification\_report(y\_val, y\_pred))**

**Saving the Model**

To deploy our model, we save it along with the scalers and label encoders.

**joblib.dump(clf, 'term\_deposit\_model.joblib')**

**joblib.dump(scaler, 'scaler.joblib')**

**joblib.dump(label\_encoders, 'label\_encoders.joblib')**

**Preprocessing and Predicting on Test Data**

We preprocess the test dataset similarly and make predictions using our trained model.

**for feature in categorical\_features:**

**if feature in test\_data.columns:**

**test\_data[feature] = label\_encoders[feature].transform(test\_data[feature])**

**X\_test = test\_data.drop(columns=['ID'])**

**X\_test[numerical\_features] = scaler.transform(X\_test[numerical\_features])**

**test\_predictions = clf.predict(X\_test)**

**test\_data['subscribed'] = test\_predictions**

**test\_data['subscribed'] = test\_data['subscribed'].map({1: 'yes', 0: 'no'})**

**test\_data[['ID', 'subscribed']].to\_csv('term\_deposit\_predictions.csv', index=False)**

**6. CONCLUDING REMARKS**

The successful prediction of whether a client will subscribe to a term deposit can significantly enhance the marketing strategies of the Portuguese bank. By focusing on clients with a higher likelihood of subscription, the bank can optimize its resources, reduce costs associated with telephonic marketing, and improve overall campaign effectiveness.

**Summary of Findings:**

* **EDA Insights**: Age, job type, marital status, education level, previous campaign outcomes, and contact duration are critical factors influencing subscription likelihood.
* **Preprocessing**: Proper encoding and scaling of features are essential for preparing the dataset for machine learning algorithms.
* **Model Building**: RandomForestClassifier proves to be an effective model, offering robust performance and handling mixed data types.
* **Model Evaluation**: The model's accuracy and detailed classification report highlight its potential in accurately predicting client behavior.

By leveraging machine learning, the bank can drive better decision-making and tailor its marketing efforts to maximize client engagement and revenue growth. This project underscores the transformative potential of data-driven approaches in the banking sector, paving the way for more intelligent and efficient marketing strategies.