Approach and Results of Building the Predictive Model

In this assignment, we followed a structured process to build a logistic regression model to predict lead conversion. The process included data preparation, feature selection, model training, and evaluation. Below is a detailed explanation of each step.

---

1. Importing Libraries and Data Exploration

The first step was to import essential Python libraries like **pandas** and **numpy** for data manipulation, **matplotlib** and **seaborn** for visualizations, and **sklearn** for machine learning. These tools helped us clean, analyze, and build the model. The dataset was loaded into a **pandas DataFrame**, and we examined the first rows to understand the structure, data types, and any immediate issues like missing values or irrelevant columns.

---

2. Data Preprocessing

To prepare the data for modeling, we performed the following steps:

- **Converting Binary Values:** Columns with "Yes/No" values were transformed into 0s and 1s.
- **Handling Missing Data:** Missing values in numerical columns were filled with the mean, while categorical columns used the most frequent value.
- **Removing Unnecessary Columns:** Irrelevant columns like IDs that did not contribute to predictions were dropped.
- **Outlier Treatment:** Outliers were removed using **Interquartile Range (IQR)** to prevent skewing the model.
- **Standardizing Values:** Numerical features were scaled using **StandardScaler** for consistency.
- **Creating Dummy Variables:** Categorical variables were converted into numerical dummy variables using one-hot encoding.
- **Visualizations:** Heatmaps and boxplots were used to identify relationships, skewed data, and multicollinearity issues.

---

3. Feature Selection and Model Building

- **Feature Selection:** We used **Recursive Feature Elimination (RFE)** to identify the 15 most important features and avoid overfitting.

- **Model Training:** A **logistic regression** model was trained using these 15 features. The model achieved **78.99% accuracy** on the training data, showing that it learned useful patterns.
- **P-Value Analysis:** Features with high p-values (> 0.05), such as *Last_Activity_Approached upfront* and *Country_Qatar*, were removed as they did not contribute meaningfully to the predictions.
- **Variance Inflation Factor (VIF):** VIF values were checked for multicollinearity, and all values were below 3, confirming no problematic correlations.
- **ROC Curve Analysis:** The model's **ROC curve** showed excellent performance with a high **AUC score**, proving it could effectively differentiate between positive and negative leads.

---

4. Model Evaluation on Test Data

The model was tested on unseen data, and different **decision thresholds** were tried to optimize its accuracy:

- A threshold of **0.42** initially resulted in an accuracy of **79.65%**.
- After testing thresholds like 0.3, 0.4, and 0.5, we found that **0.4** gave the best accuracy of **80.45%**.
  This result showed that the model generalized well to new data and avoided overfitting, making its predictions reliable.

Conclusion

This project followed a systematic process to clean the data, select the best features, and build a logistic regression model. By combining feature selection techniques like **RFE**, p-value analysis, VIF checks, and performance evaluation using the **ROC curve**, we ensured the model was accurate and reliable. The final model achieved **80.45% accuracy** on test data, demonstrating its effectiveness in predicting lead conversion. This model can be a valuable tool for prioritizing leads and improving business efficiency.