# Lead_Scoring_Summary

**Problem Statement:** X Education specializes in selling online courses to professionals within various industries. The company is seeking assistance in identifying the most promising leads, those that have a higher likelihood of converting into paying customers. The objective is to develop a lead scoring model that assigns a score to each lead. This score is designed to distinguish between leads with higher and lower conversion potential, aligning with a conversion rate of around 80%.

**Solution Steps:**

1. **Data Understanding:** The initial step involved comprehending the dataset and its characteristics.

2. **Data Cleaning:**
- Variables with a substantial proportion of missing values were dropped.
- Missing values were imputed using median values for numerical variables.
- New categorical variables were generated when required.
- Outliers were detected and addressed.

3. **Data Analysis:**
- Exploratory Data Analysis (EDA) provided insights into data distribution and patterns.
- Variables with only a single value across all rows were dropped.

4. **Creating Dummy Variables:** Categorical variables were transformed into dummy variables to enable their use in modeling.

5. **Test-Train Split:** The dataset was divided into training and testing subsets, with a 70-30% split.

6. **Feature Rescaling**: Min-Max Scaling was applied to numeric variables to standardize their range. An initial model was developed using statistical techniques to understand feature significance.

7. **Feature Selection using RFE:**
- Recursive Feature Elimination (RFE) was employed to select the top 20 important features.
- Insignificant features were eliminated based on recursive P-value analysis.

- The final set of 15 significant variables was chosen, considering VIF (Variance Inflation Factor) values.

8. **Probability Calculation and Confusion Metrics:**
- A Data Frame with converted probability values was constructed.
- Conversion probabilities above 0.5 were categorized as 1; otherwise, 0.
- Confusion Metrics were computed to determine the model's overall accuracy.
- Sensitivity and Specificity matrices were calculated to assess model reliability.

9. **ROC Curve Plotting:** An ROC curve was plotted, revealing an area under the curve (AUC) of 86%, indicating the model's effectiveness.

10. **Optimal Cutoff Point Identification:**
- Probability graphs for Accuracy, Sensitivity, and Specificity were plotted.
- The intersection of these graphs determined the optimal cutoff point (0.42).
- With this new cutoff, around 79% of predictions were accurate.
- Accuracy, Sensitivity, and Specificity metrics were updated to reflect these changes.

11. **Precision and Recall Metrics:**
- Precision and Recall values were calculated, yielding 78% and 77% on the training dataset.
- A cutoff value of approximately 0.41 was determined, considering the trade-off between Precision and Recall.

12. **Predictions on Test Set:**
- Insights gained from the training model were applied to the test dataset.
- Conversion probabilities were computed, aligning with Sensitivity and Specificity metrics.
- The test model demonstrated an accuracy of 79%, Sensitivity of 79%, and Specificity of 78%.

**Conclusion:** This comprehensive approach resulted in a lead scoring model that effectively predicted lead conversion potential, leading to more targeted and efficient sales efforts.