Summary

Lead Scoring Case Study

Accuracy : 80%

Sensitivity : 79%

Specificity : 81%

Problem Statement

X Education, an online course provider catering to industry professionals, seeks assistance in identifying the most prospective leads, those with the highest likelihood of becoming paying customers. They require a lead scoring model that distinguishes leads by assigning scores, with the intent of favoring higher scores for customers demonstrating greater conversion potential and lower scores for those with diminished conversion prospects. The CEO has set a challenging goal, aiming for a target lead conversion rate of approximately 80%.

Solution Summary

1. **Data Familiarization** : Initiated the process by comprehensively reading and inspecting the dataset.
2. **Data Cleansing :** Addressed data imbalance, redundancy, and missing values.
3. **Data Transformation :** Constructed dummy variables for categorical attributes and streamlined the dataset by eliminating duplicate and redundant variables.
4. **Train-Test Data Split :** Segregated the dataset into training and testing subsets, maintaining a 70-30% ratio.
5. **Feature Scaling via** Standardscaler and **Correlation with** Heatmap**.**

Model Building and Evaluation

1. Applied Logistics Regression for this problem.

2. Employing Recursive Feature Elimination, we meticulously curated a selection of the 15 most pivotal features.

3. We rigorously examined the generated statistics, iteratively analyzing P-values to identify and retain essential variables while discarding the insignificant ones.

4. Ultimately, we distilled the dataset to encompass a formidable roster of 12 paramount variables, noting their commendable Variance Inflation Factors (VIFs).

5. In the culmination of our efforts, we meticulously gauged the optimal probability cutoff through iterative points analysis, methodically assessing accuracy, sensitivity, and specificity.

6. Subsequently, we visualized the Receiver Operating Characteristic (ROC) curve for the selected features, attaining a notably commendable area under the curve (AUC) of 88%, further underpinning the robustness of our model.

7. In the crucible of validation, we rigorously ascertained if the model accurately forecasted approximately 80% of cases, drawing insights from the converted column.

8. The final model faced a rigorous assessment, subjecting precision, recall, accuracy, sensitivity, and specificity to meticulous scrutiny on the training dataset.

9. Navigating the nuanced terrain of the precision and recall trade-off, we pinpointed an optimal cutoff value, approximating 0.41.

10. Translating our insights into the test model, we calculated conversion probability based on Sensitivity and Specificity metrics, culminating in an accuracy of 80.68%, with Sensitivity at 79.55% and Specificity at 81.39%.

Conclusion

- The lead scoring on the test dataset yielded an impressive conversion rate and there are 479 leads which are having "lead conversion rate" of greater than or equal to 80%.

- A high sensitivity value in our model enhances our ability to identify the most promising leads, aligning with our conversion optimization goals.

- Notably, the features with the most substantial contributions to the probability of lead conversion are identified as Welingak Website, Reference source, Working Professional , Phone Conversation, SMS, underlining their significance in the conversion prediction process.