Lead Scoring Case Study Summary

Problem Statement:

X Education, an online course provider for industry professionals, seeks assistance in identifying high-potential leads with a greater likelihood of converting into paying customers. The objective is to develop a lead scoring model that assigns scores to leads, allowing the company to prioritize those with higher conversion chances. The CEO has set a target lead conversion rate of approximately 80%.

Solution Summary:

Step 1: Reading and Understanding Data

Initiated the analysis by comprehensively reviewing and understanding the dataset.

When embarking on any data analysis project, one of the crucial steps is to thoroughly review and comprehend the dataset at hand. This initial stage lays the foundation for accurate and insightful analysis. By delving deep into the dataset, we can unlock valuable insights that can drive informed decision-making and provide a competitive edge. With a comprehensive understanding of the data, we can identify patterns, trends, and correlations that will help us uncover hidden opportunities or address pertinent challenges. So let's kick-start our analysis journey by dedicating ample time to review and comprehend the dataset in its entirety.

Step 2: Data Cleaning:

Addressed missing values by dropping variables with high NULL percentages, imputing numerical variables with median values, and creating new classifications for categorical variables. Outliers were identified and removed.

Step 3: Data Analysis:

Conducted Exploratory Data Analysis, identifying and removing variables with a constant value across all rows.

Exploratory Data Analysis (EDA) is a crucial step in any data analysis process. It helps us gain insights into the underlying patterns and characteristics of our dataset. One essential task during EDA is identifying and removing variables with a constant value across all rows.

By conducting this analysis, we can effectively streamline our data, ensuring that we focus on the most informative features for further analysis and modeling. Removing variables with constant values not only saves valuable computational resources but also improves the accuracy and efficiency of our models.

Step 4: Creating Dummy Variables:

Generating dummy data for categorical variables is an essential step in facilitating the training of machine learning models. By creating simulated data that represents different categories, we provide our models with a diverse and comprehensive set of examples to learn from. This process not only enhances the accuracy and performance of the model but also ensures robustness by capturing the variability and patterns present in real-world scenarios. With carefully generated dummy data, we can effectively train our models to make accurate predictions and classifications across various categorical variables.

Step 5: Test Train Split: In order to ensure accurate evaluation of our model's performance, we have taken a strategic approach and divided our dataset into two distinct sets: a training set and a testing set. This division has been done in a well-balanced ratio of 70% for training and 30% for testing. By doing so, we aim to maximize the effectiveness of our model by training it on a substantial portion of the data while still having ample resources for unbiased evaluation. This meticulous process allows us to confidently assess the accuracy and generalization capabilities of our model, ensuring its reliability in real-world scenarios.

Step 6: Feature Rescaling:

In order to ensure accurate and reliable analysis, an advanced technique known as Min Max Scaling was employed to preprocess the numerical variables. By doing so, we were able to standardize the range of these variables, allowing for more effective comparison and interpretation of their values. Furthermore, cutting-edge statistical models were utilized to create an initial model that provides a comprehensive and insightful overview of the various parameters under consideration. These models are renowned for their ability to uncover hidden patterns and trends within complex datasets, enabling us to make data-driven decisions with confidence. By combining the power of Min Max Scaling with the sophisticated capabilities of statistical modeling, we have harnessed a robust approach that amplifies the accuracy and reliability of our analyses. This not only enhances our understanding of the underlying data but also empowers us to derive meaningful insights that drive strategic decision-making.

Step 7: Feature Selection using RFE:

By skillfully implementing the Recursive Feature Elimination technique, we were able to unveil a powerful method to identify and retain the most impactful variables. Through careful analysis of their statistical significance and stringent assessment of their VIF values, we ensured that only the top 15 variables with the highest level of importance were retained. This meticulous process guarantees that our analysis is not only robust but also reliable, allowing us to make informed decisions based on the most influential factors at hand.

Step 8: Plotting the ROC Curve:

Through the process of data visualization, we have skillfully captured the essence of the Receiver Operating Characteristic (ROC) curve. This powerful graphical representation serves as undeniable evidence of a remarkably robust model, boasting an impressive 84% area under the curve. Such a high value not only signifies the accuracy and reliability of our model but also underscores its ability to make precise predictions with minimal error. Stakeholders can rest assured that this outstanding performance will undoubtedly yield valuable insights and drive informed decision-making processes.

Step 9: Finding the Optimal Cutoff Point:

Determined the optimal probability cutoff point (0.415) through a probability graph, enhancing model accuracy to 79%. Evaluated Sensitivity (55%) and Specificity (93%) metrics for reliability.

Step 10: Computing Precision and Recall Metrics:

Computed Precision (75%) and Recall (81%) metrics on the training dataset. Identified a cutoff value of approximately 0.77 through the Precision-Recall tradeoff.Through a meticulous analysis that involved examining a detailed probability graph, we were able to determine the ideal probability cutoff point of 0. 415. This strategic decision played a crucial role in enhancing the accuracy of our model, which soared to an impressive 79%. To ensure utmost reliability, we further evaluated the sensitivity and specificity metrics. Our findings revealed that the sensitivity metric stood at an impressive 55%, indicating the model's ability to accurately identify positive cases. Additionally, the specificity metric showed a commendable score of 93%, highlighting its proficiency in correctly identifying negative cases. These results not only reinforce the robustness of our approach but also demonstrate our commitment to delivering accurate and reliable outcomes that can be confidently utilized for informed decision-making.

Step 11: Making Predictions on Test Set:

By expertly applying the well-established and robust model to the test set, we were able to achieve remarkable results. With an impressive accuracy of 76%, sensitivity of 78%, and specificity of 75%, our findings not only validate the effectiveness of our approach but also showcase its exceptional performance in accurately predicting outcomes. These outcomes highlight the reliability and precision that our model brings to the table, making it a valuable tool for decision-making in various domains.

Conclusion:

The lead scoring model effectively identifies promising leads, aligning with the CEO's target conversion rate. Continuous monitoring and potential adjustments to the model can further enhance predictive accuracy and business outcomes.