Understanding the Problem

X Education, an education company, is committed to enhancing its lead conversion rate by focusing on "Hot Leads," which are those prospects with a higher likelihood of converting into customers. The primary objective is to develop a robust model that assigns a lead score to each prospect. This scoring system will enable the sales team to prioritize their efforts efficiently, with the aim of achieving a conversion rate of approximately 80%.

Approach

Loading Necessary Libraries and Reading the Dataset

The first step in our approach involved loading the necessary libraries and reading the dataset to perform an initial inspection. During this inspection, we identified columns containing the value "Select," which needed to be replaced with null values. Key columns requiring attention included Specialization, How did you hear about X Education, Lead Profile, and City.

Data Cleaning and Preprocessing

To ensure the dataset was ready for analysis, we addressed missing values and redundant information. Columns containing more than 30% null values were removed. These included 'How did you hear about X Education', 'Tags', 'Lead Quality', 'Lead Profile', 'City', 'Asymmetrique Activity Index', 'Asymmetrique Profile Index', 'Asymmetrique Activity Score', and 'Asymmetrique Profile Score'. Additionally, columns that were optional or irrelevant for the analysis were discarded.

Univariate and Bivariate Analysis

We conducted comprehensive univariate and bivariate analyses to understand the distribution of individual variables and the relationships between different variables. This step was crucial in identifying significant features that could influence lead conversion rates.

Splitting the Dataset

To evaluate the model's performance and ensure its generalization capability, we split the dataset into training and testing subsets. This approach allowed us to build the model on the training data and test its accuracy on unseen data.

Feature Selection with RFE

We used Recursive Feature Elimination (RFE) for feature selection. RFE helped us identify the most important features for model building by recursively removing less significant features and building the model iteratively.

Model Building and Evaluation

Using the selected features, we built a classification model to predict the lead score. We evaluated the model using various metrics, including accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). Additionally, we assessed the variance inflation factor (VIF) to ensure there was no multicollinearity among the features.

Performance Evaluation with ROC Curve

We generated the Receiver Operating Characteristic (ROC) curve to evaluate the model's performance. The ROC curve helped us identify the optimal cutoff point to balance correct predictions and errors in the classification model.

Generating Lead Scores

Based on the model's predictions, we generated lead scores for all the prospects. These scores enabled the sales team to identify and prioritize the hottest leads, thereby increasing their chances of conversion.

Recommendations

Based on our findings, we provided X Education with several recommendations to improve their lead conversion rate. These included:

- **Targeted Outreach**: Focus on leads with high scores to maximize conversion rates.
- **Personalized Communication**: Customize communication strategies based on lead scores to engage prospects more effectively.

Conclusion

By implementing a data-driven approach to lead scoring, X Education can significantly enhance its lead conversion rate. This model not only prioritizes high-potential leads but also provides actionable insights for the sales team to optimize their efforts. Achieving a conversion rate of approximately 80% is within reach, enabling the company to grow its customer base and increase revenue.

Learnings

By solving the case study, we grasped various aspects of model building and successfully met the business requirements. This provided us with valuable, real-world experience.