Summary Report on Lead Scoring Model Development

Objective & Approach

The objective of this assignment was to improve the lead conversion rate for X Education, an online course provider targeting industry professionals. The company observed that their typical conversion rate of 30% was subpar, and the CEO aimed for an improved conversion rate of 80%. The dataset provided contained 9,240 records with 37 columns, some of which had missing or irrelevant data. The goal was to analyze and clean this data, build a predictive model, and optimize lead scoring for better conversions.

Data Cleaning & Preprocessing

The first step in the process was data cleaning. We began by examining the percentage of missing values in each column. Columns with more than 25% null values were dropped, leading to the removal of 8 columns. Additionally, rows containing null values were discarded to ensure the quality of the data. Irrelevant columns, such as those related to city, country, and lead number, were also removed since they did not contribute to meaningful insights.

It was noted that certain columns contained default values like 'Select', which added no value to the analysis. These were similarly removed. After this preprocessing, the dataset was cleaned and ready for further exploration.

Feature Selection & Modeling

After cleaning the data, we explored the relationships between variables using a correlation matrix. Variables with strong relationships were considered for the modelling phase. The data was then split into training and testing sets, with 70% allocated for training and 30% for testing.

Next, the numeric features were scaled using StandardScaler to normalize the data. Feature selection was carried out using Recursive Feature Elimination (RFE) to identify the 25, 20, and finally 15 most significant variables. This process helped in narrowing down the features that contributed the most to the model's predictive power.

The modelling process involved using a logistic regression model, which is suitable for classification problems. Before finalizing the model, columns with high p-values and high Variance Inflation Factor (VIF) were dropped, ensuring the removal of multicollinearity and enhancing model stability. The chosen threshold for p-value was 0.05, and for VIF, it was below 5.

Model Evaluation & Threshold Optimization

The model was evaluated using the ROC curve, with an area under the curve (AUC) of 0.83, indicating a good overall performance. However, the initial threshold of 0.5 for classification did not provide balanced specificity and sensitivity values. Upon examining the Sensitivity-Specificity curve, it was determined that an optimal threshold value of 0.4 would improve the balance, yielding an accuracy of 77.72% with sensitivity and specificity both close to 77.7%.

Final Model Performance

The final logistic regression model achieved an accuracy of 75.85% on the test dataset. The balance between sensitivity and specificity at the optimized threshold indicated that the model was well-calibrated for predicting lead conversion, which is crucial for improving business outcomes.

Learnings & Missing Aspects

This exercise provided valuable insights into the importance of data preprocessing, feature selection, and model evaluation in predictive modelling. Key learnings include the necessity of:

- Handling missing and irrelevant data effectively to improve model quality.

- Carefully selecting features and tuning hyperparameters for better performance.

- Evaluating multiple thresholds to identify the optimal point for classification tasks.

However, a few elements were missing from this solution that could improve future models:

1. **Cross-Validation**: While the data was split into training and testing sets, using cross-validation would have provided a better estimate of the model's generalizability.

2. **Model Comparison**: Exploring other models, such as decision trees or random forests, could offer more robust performance, especially for more complex patterns in the data.

3. **Feature Engineering**: Additional feature engineering, such as creating interaction terms or exploring non-linear transformations, could potentially improve model performance.

In conclusion, the lead scoring model developed achieved a respectable accuracy of 75.85%, with a well-calibrated threshold, providing actionable insights for X Education to target and convert more leads effectively. Further refinement through cross-validation, model comparison, and advanced feature engineering could push the performance closer to the desired target conversion rate of 80%.