Summary report

Data preprocessing, feature selection, and model optimization are essential steps in building robust predictive models. This report outlines the key steps from data exploration, feature selection using Information Value (IV) and correlation, to model optimization, with a focus on real-world applications and improving business insights.

1. Data Exploration

The first step in any data science project is to thoroughly explore the dataset. We begin by checking for duplicates and missing values, as these can lead to biased results or misinterpretations. Outliers are also examined, as they can skew model performance. Visualizing data using boxplots helps in detecting outliers. The process ensures a clean dataset, which is essential for building accurate models.

2. Feature Selection Using Correlation and IV

Feature selection is a crucial process that improves model performance by removing irrelevant or redundant variables. Here, we use Pearson correlation to identify multicollinearity among variables and employ Information Value (IV) to assess the predictive power of each feature. The function **ft_select_corr_iv** selects features with high correlation (above 0.7) but low IV, helping to retain the most relevant variables. This step helps in reducing overfitting, enhancing interpretability, and improving overall model accuracy.

3. Automatic Binning of Variables

Optimal binning is a powerful method for transforming continuous variables into discrete bins, which can improve the predictive performance of models, especially logistic regression. The **ft_auto_binning** function performs this binning automatically for both numeric and categorical variables, ensuring that each variable is transformed optimally. It also calculates the IV for each variable, helping in selecting the most important predictors for the final model.

4. Model Performance Summary

Once the features are selected, we test the performance of various models. Initially, a baseline logistic regression model is trained, yielding an accuracy of 0.81 and an AUC score of 0.88. After optimization, the accuracy of the logistic regression model improves to 0.789, while the AUC score increases significantly to 0.940. Finally, an XGBoost model is trained, achieving a higher accuracy of 0.904 and an AUC score of 0.966, demonstrating superior predictive performance.

5. Threshold Optimization for XGBoost

To further refine the XGBoost model, decision threshold optimization is performed. By iterating through various thresholds, the best one is found to be 0.45, resulting in an accuracy of 0.906. This step ensures that the model not only predicts well but also aligns with the business objectives, such as minimizing false positives or false negatives depending on the context.

6. Real-World Applications and Insights

In real-world applications, it is crucial to examine the dataset over time. Including datetime columns allows for an analysis of trends, seasonality, and behavior changes over time, providing deeper business insights. Moreover, creating additional variables, such as lead interaction frequency, referral sources, and conversion times, can offer a competitive edge in decision-making.

Prioritizing leads referred through references is highly recommended, as these leads tend to have a higher probability of converting. Additionally, while students can be approached, their likelihood of conversion is lower due to the course being industry-focused. However, this challenge can be turned into an opportunity by positioning the course as a tool to enhance industry readiness, which can attract students aiming to prepare for the workforce before graduation.

In conclusion, a well-rounded data processing strategy coupled with advanced feature selection and model optimization techniques can significantly enhance model performance and provide actionable business insights, improving lead targeting and overall decision-making processes.