SUMMARY REPORT

This report covers the key steps involved, from data exploration and feature selection using Information Value (IV) and correlation, to model optimization, with an emphasis on practical applications and enhancing business insights.

Data Exploration:

The initial phase involves a thorough examination of the dataset. This includes identifying duplicates and missing values, which can distort results, and analyzing outliers using visual tools like boxplots to ensure a clean dataset, which is essential for accurate modeling.

Feature Selection with Correlation and IV:

Selecting the right features is vital for improving model performance. We use Pearson correlation to detect multicollinearity and Information Value (IV) to evaluate the predictive power of features. The ***select_correlated_iv_features*** function filters out features **with high correlation (above 0.6)** but **low IV**, focusing on the most relevant variables. This reduces overfitting and enhances model interpretability and accuracy.

Automatic Binning of Variables:

Transforming continuous variables into discrete bins through optimal binning can boost model performance, particularly in logistic regression. The ***auto_binning_process*** function automates this process for both numeric and categorical variables, also calculating IV to identify the most significant predictors for the final model.

Model Performance & Summary:

After feature selection, various models are evaluated. A **baseline logistic regression model** is trained with an initial accuracy of 0.81818 and an AUC score of 0.88224. **Post-optimization**, the logistic regression model's accuracy improves to 0.86851, with the AUC score rising to 0.93027. **The XGBoost model**, however, achieves a little higher accuracy of 0.8869 and an AUC score of 0.95726, showing superior predictive capabilities.

Threshold Optimization for XGBoost:

To further enhance the XGBoost model, threshold optimization is conducted. By testing different thresholds, the optimal value is determined to be 0.64999, which improves accuracy to 0.88853. This step ensures the model aligns with business goals, such as minimizing false positives or negatives, depending on the context.

Practical Applications and Business Insights:

In summary, a comprehensive approach to data preprocessing, feature selection, and model optimization is essential for developing robust predictive models. This process begins with thorough data exploration to ensure a clean and reliable dataset by identifying duplicates, missing values, and outliers. Effective feature selection, using techniques **like Pearson correlation** and **Information Value (IV)**, helps in retaining the most relevant variables and reducing overfitting. Automatic binning further refines continuous variables into discrete bins, enhancing model performance.

Testing and optimizing various models reveal their effectiveness, with XGBoost demonstrating superior predictive performance compared to logistic regression. Additionally, threshold optimization for XGBoost improves model accuracy and aligns with business objectives.

In practical applications, incorporating datetime columns allows for the analysis of trends and behavior changes, providing deeper insights. Adding variables such as lead interaction frequency and referral sources can significantly enhance decision-making. Prioritizing referrals and adapting strategies for specific groups, like students, can further optimize lead targeting and conversion strategies. Not only that, modern AI and machine learning tools can automate and improve these processes, making the entire pipeline more efficient and scalable.

Overall, combining a structured data processing strategy with advanced feature selection, model optimization, and modern AI/ML tools significantly **boosts model accuracy** and **provides valuable business insights**, **improving lead targeting and decision-making processes.**