Lead Score Analysis – Summary Report

The objective of this project was to build a predictive model to determine the likelihood of a lead converting into a customer for X Education, an online course provider. The goal was to optimize the company's sales funnel by identifying high-potential leads, thus enabling better targeting and resource allocation by the sales team.

Approach

The assignment was approached in a systematic manner following the standard data science workflow:

1. Understanding the Problem Statement

The first step was to understand the business context—X Education receives leads from multiple sources, and not all leads convert to customers. Identifying high-quality leads through predictive modeling can significantly improve conversion rates and reduce sales efforts.

2. Data Exploration and Cleaning

The raw dataset was thoroughly examined for null values, duplicate entries, and irrelevant columns. Columns with a high percentage of missing values (like ‘Lead Profile’, ‘Tags’) or with only one unique value (like ‘What matters most to you in choosing a course’) were dropped. Data types were standardized, and categorical variables were formatted appropriately.

3. Feature Engineering

Dummy variables were created for categorical columns using one-hot encoding. Redundant columns were removed based on correlation, domain knowledge, or high multicollinearity (checked via VIF). The target column, ‘Converted’, was retained as the binary outcome.

4. Model Building and Evaluation

Logistic regression was chosen due to its interpretability and suitability for binary classification problems. The data was split into training and test sets (70:30 ratio), and the model was trained on the training data. Backward elimination was used to refine the feature set, guided by p-values and VIF.

Performance was evaluated using:

- **Accuracy**
- **Precision, Recall, F1-Score**
- **ROC-AUC curve**
- **Confusion matrix**

Threshold tuning was also conducted to optimize recall without sacrificing too much precision, as the business preferred capturing more potential leads even at the risk of some false positives.

Key Learnings

1. **Business-Driven Data Cleaning:**
   Real-world datasets are often messy. Understanding the business impact of each feature helped decide whether to keep or drop it, beyond just missing value percentages.
2. **Importance of Feature Selection:**
   Removing multicollinear variables improved model performance and interpretability. VIF and p-value analyses were critical in refining the model.
3. **Model Interpretability Matters:**
   Logistic regression's clear coefficients and odds ratios helped explain the influence of different features, aligning with the needs of non-technical stakeholders.

4. **Threshold Optimization:**
   The classification threshold significantly impacts business outcomes. By adjusting it, we achieved a balance between recall and precision that supports sales strategies.
5. **End-to-End Pipeline Development:**
   The project strengthened my ability to handle the full lifecycle—from data preprocessing to model deployment preparation and business interpretation.

Conclusion

This project provided a comprehensive exposure to solving a real-world business problem using logistic regression. It sharpened my technical skills in data preprocessing, model building, and evaluation while reinforcing the importance of aligning machine learning work with business goals. The final model is capable of scoring leads and assisting the sales team in prioritizing follow-ups, potentially boosting conversion rates and improving revenue efficiency.