Summary - Lead Scoring Model Development

Lead Scoring Model Development:

This project aimed to develop a logistic regression model to assign lead scores for X Education, identifying hot leads likely to convert into paying customers. With an initial conversion rate of 30%, our goal was to optimize lead management and target a conversion rate of around 80%.

Data Cleaning and Preprocessing: The dataset contained 9,240 leads with attributes like Lead Source, Total Time Spent on Website, and Last Activity. We addressed missing or placeholder values by:

- Imputing columns such as City and Country with 'Unknown.'
- Imputing categorical variables like Specialization and What Matters Most to You in Choosing a Course with 'Other.'
- Dropping columns with more than 45% missing values, like Lead Profile and How Did You Hear About X Education.

Exploratory Data Analysis (EDA): Univariate and bivariate analysis provided insights into lead behavior. For example:
Google and Direct Traffic sources had the highest conversions, while Facebook and Referral Sites had lower rates.
Leads spending more time on the website were more likely to convert.
Features like Last Activity (e.g., email opened or SMS sent) indicated higher conversion likelihood, emphasizing the importance of lead engagement.

Handling Outliers: Outliers in numerical variables, such as Total Visits and Total Time Spent on Website, were capped at the 95th percentile to minimize their influence on the model.

Feature Engineering: We applied encoding for categorical variables and scaled numerical features using MinMaxScaler.

Model Building: The data was split into training and test sets. Recursive Feature Elimination (RFE) was used for feature selection, highlighting Total Time Spent on Website, Lead Source, and Last Activity as crucial features. This reduced complexity and avoided overfitting.

Model Evaluation: The logistic regression model was built using scikit-learn and evaluated using key metrics: accuracy, precision, recall, and ROC-AUC score. The precision-recall trade-off was explored to suit business objectives: higher recall identified more potential conversions but increased false positives, while higher precision focused on quality over quantity.

Learnings and Recommendations: Key takeaways included:

The importance of data cleaning, particularly handling categorical variables with missing or placeholder values, significantly impacted model performance.

Understanding the precision-recall trade-off is crucial, as it influences how the sales team should prioritize leads.

Using the logistic regression model allows X Education to focus on leads with a higher likelihood of conversion, potentially improving the overall conversion rate.

Future enhancements could involve testing advanced models like decision trees or random forests to further improve predictive accuracy.

This project emphasized the importance of thorough data preprocessing, feature selection, and aligning model performance with business objectives to create actionable insights for the sales team.