Summary - Lead Scoring Model Development

This project aimed to build a logistic regression model to assign lead scores for X Education and identify hot leads likely to convert into paying customers. With a conversion rate of 30%, the goal was to optimize lead management by focusing on leads more likely to convert, with a target rate of around 80%.

Data Cleaning and Preprocessing

The dataset contained ~ 9,000 leads with attributes such as Lead Source, Total Time Spent on Website, and Last Activity. Several columns had missing or placeholder values like 'Select'.
To handle this:
i) Columns such as City and Country with Select or NaN values were imputed with Unknown.
ii) Categorical variables like Specialization and What matters most to you in choosing a course were imputed with Other.
iii) Columns with more than 45% missing values, like Lead Profile and How did you hear about X Education, were dropped.

Exploratory Data Analysis (EDA)

Univariate Analysis, Categorical Features, Bivariate Analysis, Multivariate Analysis provided insights into lead behavior. For example,
i) Google and Direct Traffic sources contributed the most conversions, while Facebook and Referral Sites had lower conversion rates.
ii) Leads spending more time on the website were more likely to convert, making this a strong predictor.
iii) Features like Last Activity (e.g., email opened or SMS sent) also indicated higher conversion likelihood, emphasizing the importance of lead engagement.
Outliers in numerical variables, such as in Total Visits and Total Time Spent on Website, were capped at the 95th percentile to minimize their influence on the model.

After cleaning and EDA, we applied one-hot encoding for categorical variables and scaled numerical features like Total Time Spent on Website and Total Visits using MinMaxScaler.

Logistic Regression Model

We split the data into training and test sets and used Recursive Feature Elimination (RFE) to select the most relevant features, such as Total Time Spent on Website, Lead Source, and Last Activity. This helped reduce complexity and avoid overfitting.

The logistic regression model was built using scikit-learn, and performance was evaluated using key metrics: accuracy, precision, recall, and ROC-AUC score. The trade-off between precision and recall was explored to suit the business objective. Higher recall identifies more potential conversions but may result in more false positives, while higher precision focuses on quality over quantity.

Learnings and Recommendations

A critical takeaway was the importance of data cleaning, particularly handling categorical variables with missing or placeholder values. This significantly impacted model performance.

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

Using the logistic regression model, X Education can 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.

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