Lead Scoring Case Study - Summary

Importing Libraries and Dataset:

- Required libraries were imported, primarily statsmodels and scikit-learn.
- The dataset was imported.

Data Understanding:

- Null/duplicates inspection was performed, and no duplicates were found, but various columns with nulls were observed.

Data Cleaning:

- Values with "Select" were replaced with null values in all columns.
- Columns with a null percentage value greater than 40% were dropped, and rows with very low percentages of null values (approximately 1-2%) were deleted.
- For columns ranging between 20-40% of null values, the mode value was imputed for categorical columns since the mode value was around 95% in all columns.
- Columns representing the ID of the lead or customer and those with only one unique value were dropped.

Exploratory Data Analysis (EDA):

- Data imbalance was checked, and the ratio was found to be 1:1.6 (converted to not converted).
- Univariate and multivariate categorical analysis was made on all features, and count plots were displayed.
- Columns with high data imbalance were dropped.
- Univariate and multivariate numerical analysis was carried out on all numerical columns, and a pair plot and heatmap were plotted.
- Boxplot analysis was made to handle and treat outliers present.

Data Preparation:

- Dummy variables were created for multi-level categorical columns.
- Data was split into train and test data at a ratio of 70:30.
- Continuous numerical columns were rescaled using MinMaxScaler.
- Variable selection was automated using RFE method, and the top 16 variables were selected for model building.

Model Building:

- Models were built by considering the P-values and Variance Inflation Factor (VIF) for manual feature elimination.

Model Evaluation:

- The final trained model had an accuracy score of 80%, sensitivity of 80%, F1 score of 75%, and ROC curve area of 88% after choosing the optimal cutoff at 0.35 from the graph of accuracy, sensitivity, and specificity.
- Lead score was assigned for the trained data.

Model Testing:

- The built model was then tested on the test data where we got an accuracy score of 82%, sensitivity of 80%, and an F1 Score of 77%. Hence the model was stable.
- Lead score was then assigned to the tested data.

Model Inference and Recommendations:

- The top three predictor variables turned out to be:-
  1. Total Time Spent on Website
  2. Lead Origin_Lead Add Form
  3. What is your current occupation_Working Professional.
- High conversion was seen for leads that came through Lead Add Form and Landing Page Submission where customers spent more time on the webpage before submitting the lead form.
- Sales teams can rely on the model and the assigned lead scores.
- Teams should focus more on working professional leads as they have no financial restrictions and tend to purchase the course.