Lead Scoring Case Study – Summary

Steps Followed:

- Importing necessary libraries
- Importing the provided dataset
- Data Wrangling
- Conducting Exploratory Data Analysis (Variable Inspection)
- Preparing the Data
- Building the Model (Logistic Regression)
- Evaluating the Model (Logistic Regression Metrics)
- Testing the Model
- Inferring from the Model
- Drawing Conclusions Based on Results

Data Wrangling:

- Import the dataset.
- Examine the entire dataset and make key observations.
- Check the overall dimensions of the dataset.
- Inspect column formats and correct any irregularities found in the dataset.
- Identify and handle any NULL values present in the dataset.
- Address NULL values by imputing those rows or replacing them with mean or median values.

Exploratory Data Analysis (EDA):

- Checked for data imbalance, revealing a ratio of 1:1.6 (converted to not converted).
- Conducted univariate and multivariate categorical analysis on all features, displaying count plots.
- Dropped columns with significant data imbalance.
- Performed univariate and multivariate numerical analysis on all numerical columns, plotting a pair plot and heatmap for visualization.
- Utilized boxplot analysis to identify and treat outliers.

Data Preparation:

- Binary-level categorical columns were already mapped to 1/0 in previous steps.
- Created dummy features (one-hot encoded) for categorical variables: Lead Origin, Lead Source, Do Not Email, Last Activity, Specialization, Current Occupation, Tags, City, A Free Copy of Mastering the Interview, Last Notable Activity.
- Split Train & Test Sets with a 70:30 ratio.
- Utilized feature scaling with the Standardization method to scale the features.
- Checked correlations, dropping predictor variables highly correlated with each other.

Model Building:

- Acknowledged the dataset's extensive dimensions and features, potentially impacting model performance and computation time.
- Recognized the importance of Recursive Feature Elimination (RFE) to select only significant columns.
- Utilized Manual Feature Reduction by dropping variables with p-values greater than 0.05.
- Identified Model 3 as stable after three iterations, with significant p-values (< 0.05) and no signs of multicollinearity (VIFs < 5).
- Designated Model 3 as the final model for Model Evaluation and predictions.

Model Evaluation:

- The final trained model achieved an accuracy score of 91%, Precision score of 89%, F1 score of 93%, and ROC curve area of 97% after selecting the optimal cut-off at 0.35 from the graph of accuracy, sensitivity, and specificity.
- Assigned lead scores for the trained data.

Model Testing:

- Tested the built model on the test data, yielding an accuracy score of 82%, sensitivity of 80%, and an F1 Score of 77%, indicating model stability.
- Assigned lead scores for the tested data.

Conclusion:

- Landing Page Submissions and Lead Add Form lead to more conversions.
- Conversions are higher for leads from Google, Organic Search, Direct Traffic, and Referrals.
- SMS and Email marketing leads have higher conversions.
- Finance, HR, Marketing, Operations, and Banking sector leads tend to convert more.
- "Better Career Prospects" option for career outcome leads to higher conversions.
- Leads spending more time on the website tend to convert more.
- Reducing website bounce rate can increase customer engagement time and conversions.
- Lead Add Form generates qualifying leads and should be used across key areas.
- Sales team should focus on working professionals for higher conversions.
- Leads with a Lead Score >0.35 tend to convert more, with a model accuracy score of 91%.