Summary of Logistic Regression Model for Lead Scoring

Objective: Build a logistic regression model to assign a lead score (0-100) to each lead, indicating their likelihood to convert. Higher scores suggest "hot" leads, while lower scores indicate "cold" leads.

Steps Taken to Reach Solution

1. **Exploratory Data Analysis**:

- **Library Imports**: Utilized libraries such as NumPy, pandas, matplotlib, seaborn, Scikitlearn, and stats models.
- **Data Cleaning**:
- Dropped columns with over 30% missing values and city variable.
- Removed "select" options in columns like "Lead Profile" and "How did you hear about X Education".
- Filled missing values in categorical columns with mode and updated "specialization" column.
- Dropped rows with null values in "Page views per visit".
- Removed "Prospect ID" and "Lead Number" columns.
- **Outlier Detection and Removal**: Used boxplots to detect and drop outliers in "TotalVisits" and "Page Per visits".
- **Visualization and Correlation**: Created various plots and heatmaps to understand data distribution and collinearity.
- **Dummy Variables**: Created and concatenated dummy variables for categorical columns.

2. **Model Building**:

- **Train-Test Split**: Split data into train (70%) and test (30%) sets.
- **Feature Scaling**: Applied MinMaxScaler to rescale features.
- **Feature Selection**: Used Recursive Feature Elimination (RFE) to select 15 significant features.
- **Logistic Regression Model**: Added a constant, created a Generalized Linear Model (GLM), and refined by removing insignificant variables using VIF and p-values.

3. **Model Prediction**:

- Predicted probabilities using the logistic model.
- Created a data frame to store conversion probabilities and applied a 0.5 threshold to classify leads.

4. **Model Evaluation**:

- **Confusion Matrix**: Evaluated model performance by calculating accuracy, sensitivity, and specificity from the confusion matrix.
- **Finding Optimal Cutoff**:
- Created ROC curve and identified optimal cutoff based on the trade-off between accuracy, sensitivity, and specificity.
- Updated threshold and recalculated performance metrics.
- **Test Dataset Evaluation**:
- Transformed test dataset, added constant, and predicted using the logistic model.
- Evaluated test predictions using confusion matrix, sensitivity, specificity, precision, and recall.
- Made final predictions on test dataset with 0.42 as the cutoff.

5. Conclusion

1. Key Variables Influencing Lead Conversion:

- TotalVisits: 1.454840
- Total Time Spent on Website: 4.713648
- Lead Origin_Lead Add Form: 4.972153
- Lead Origin_Lead Import: 1.834902
- Lead Source_Olark Chat: 1.603159
- Last Activity_Had a Phone Conversation: 2.070172
- Last Activity_Olark Chat Conversation: -1.515929
- Last Activity_SMS Sent: 1.384397
- Current Occupation Working Professional: 2.879075
- Last Notable Activity Unreachable: 2.158080
- Do Not Email: -1.474431

2. the final model, with a cutoff threshold of 0.42, successfully segments leads into "hot" and "cold", aiding the sales team in targeting efforts more efficiently.

3. **Total Time Spent on Website**: This is a strong indicator with a high coefficient (4.713648), suggesting that more time spent on the website correlates with a high likelihood of conversion.

4. lead origin lead add form has the highest coefficient, which implies that the variable is best indicator for lead conversion