SUMMARY

The process involves building and refining a model for Company X Education with the goal of converting potential users. The primary focus is on data analysis and prediction to identify effective strategies for increasing the conversion rate. The following steps were undertaken:

1. Data understanding ,Data Cleaning and Preparation:

Handling Missing Values:

- **Conducted an assessment of null values in the dataset.**
- **Drop all the columns in which greater than 3000 missing values are present**
- **Dropped columns which seems to be of less relevance such as Prospect ID, Lead Number etc**
- **Handled columns having more select values**

Numerical Variables and Outliers:

- **Explored and processed numerical variables to enhance model performance.**
- **Identified and addressed outliers to prevent them from unduly influencing the model.**
- **Auto EDA was done using sweetwiz**

2. Dummy Variables Creation:

- **Employed dummy variables to convert categorical variables into a format suitable for modeling using get_dummies**
- **Drop the variables for which the dummy variables were created**

These EDA steps lay the foundation for a more robust analysis and model building process. The overarching objective is to refine the dataset, ensuring it is well-prepared for subsequent stages of the project.

3. Train-Test Split & Scaling:

Data Division:

- **Segregated the dataset into training and testing subsets, allocating 70% of the data for training and reserving 30% for testing.**
- **This division ensures that the model is trained on a substantial portion of the data while retaining a separate set for evaluation, gauging the model's generalization performance.**

Variable Scaling:

- **Focused on specific variables—['TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website']—for scaling**
- **Looked at the correlations as the number of variables are pretty high**

By performing this train-test split and scaling, we establish a robust methodology for model training and testing, promoting the model's ability to generalize well to new, unseen data. The scaling of specific variables further enhances the model's stability and performance across the entire dataset.

4. Model Building:

Feature Selection using RFE:

- **Employed Recursive Feature Elimination (RFE) as a method for feature selection.**
- **Executed RFE to identify and retain the top 15 most relevant variables in the dataset.**
- **Taken a look at which variables are selected by RFE**
- **Now using these variables we build a logistic regression model using statsmodels**

Manual Variable Removal based on VIF and p-value:

- **Conducted a secondary refinement by manually eliminating the remaining variables.**
- **Decisions on variable removal were guided by considerations such as Variance Inflation Factor (VIF) values and p-values.**
- **This step ensured a streamlined set of variables with minimal multicollinearity and statistical significance.**

In summary, the model building process involved a combination of automated feature selection through RFE and manual refinement based on VIF and p-values.

5. Model Evaluation:

Sensitivity – Specificity Evaluation:

On Training Data Performance metrics:

- **Accuracy: 78.86%**
- **Sensitivity:73.94%**
- **Specificity:83.43%**

Used ROC function and plots to determine best suitable cutoff.
Performance Metrics at Cutoff 0.42:

- **Accuracy: 79.08%**
- **Sensitivity: 79.33%**
- **Specificity: 78.84%**

Predictions On Test set:
On Test Data Performance metrics:

- **Accuracy: 78.45%**
- **Sensitivity: 77.94%**
- **Specificity: 78.91%**

Precision Recall View:

- **Precision: 80.57%**
- **Recall: 73.94%**

Precision and Recall Tradeoff:
Training set:

- **Accuracy: 78.95%**
- **Precision: 78.40%**
- **Recall: 77.71%**

Test set:

- **Accuracy: 78.66%**
- **Precision: 78.28%**
- **Recall: 76.74%**

These evaluations provide a nuanced understanding of the model's performance, offering insights into different aspects of its predictive capabilities. The choice between the two optimal cut-off values depends on the specific goals and considerations of the application.

The model demonstrates effective prediction of the conversion rate, with key variables identified as significant contributors. These include metrics related to lead sources, lead origin, and last activity. Notable sources of influence on conversion include total visits, time spent on the website, and engagement through the lead add form.

This analysis provides actionable insights for decision-making, allowing the company to make informed calls based on the identified influential factors. The model's robust performance instills confidence in its predictive capabilities, enhancing the company's ability to strategize effectively for conversion optimization.