Problem Statement -

- X Education is an online education company that sells its online courses to various industry professionals. The company uses search engines like Google and other websites to market its courses.
- The business agenda for the marketing team is to target those leads who are most likely to convert into paying customers. To achieve this, we are assigned the task of building a model to predict leads and assign lead scores to all prospective leads, so that leads with higher scores will be the ones most likely to convert and those with least scores result in non-conversion.
- The target lead conversion rate should be around 80%, as given by the CEO.

Summary of Steps Taken -

Step 1: Inspecting the Dataframe

- Reading and Understanding the Dataset.
- Examining the number of rows and columns.
- Inspecting the column types i.e., numerical, categorical or object type.
- Getting numerical summary of numerical columns.

Step 2: Data Preparation

- Checking for duplicate data in the entire dataset and remediation.
- Replacing ‘Select’ values in the dataset as missing values (NaN).
- Check for missing (NaN) values in the dataset and its remediation.
  - Drop out the variables having more than 40% of missing values.
  - Imputing missing values with the mode (maximum frequency value occurring in the column) in case of categorical column.
  - Imputing missing values with the median in case of numerical column.
  - Imputing missing values in certain columns as ‘No Information’ or ‘Not Mentioned’, wherever applicable.
  - Grouping together low frequency values in certain categorical columns as ‘Others’.
- Dropping columns having Imbalanced Data i.e., one value in majority and other one in minority.
- Outlier Treatment for numerical columns by eliminating outliers above the 99% quantile.
- Percentage of data retained after data cleaning was 98.38%.

Step 3: Data Analysis

- Data Analysis of categorical columns using count plots to determine most significant feature variables for lead conversion.

Step 4: Dummy Variable Creation

- Mapping ‘Yes/No’ variables as ‘1/0’ respectively.
- Creating new Dummy Variables for categorical columns with multiple levels and dropping the original/parent columns.

Step 5: Test-Train Split

- The entire dataset was split into the train and test datasets, with the split proportion of 70 : 30 % respectively.

Step 6: Feature Scaling

- Scaling the original numerical columns 'TotalVisits','Total Time Spent on Website' and 'Page Views Per Visit' using the Standard Scaler.

Step 7: Model Building

- Building the initial model using logistic regression algorithm of ‘statsmodels’ library.

Step 8: Feature Selection Using RFE

- Using the automated Recursive Feature Elimination, the 20 most important features were selected. Using the manual elimination technique from the descriptive statistics generated, we eliminated features with P-values greater than 0.05 to narrow down the most significant features for the model.
- Finally, the 14 most significant variables were selected for the final model based on their VIF values also to prevent multicollinearity among independent variables.
- A data frame was created having converted probability values using an initial assumption that probability values over 0.5 means 1 else 0.
- Using the above assumption, the Confusion Metrics was generated and other metrics such as ‘Accuracy’, Sensitivity’ and the ‘Specificity’ of the model were calculated.

Step 9: Plotting the ROC Curve

- Further, the ROC curve was plotted for the features and the we obtained a good curve with an area under the curve of 95.86% leading to a good model.

Step 10: Finding the Optimal Cutoff Point

- Then a probability graph was plotted at different probability values for the ‘Accuracy’, ‘Sensitivity’, and ‘Specificity’ respectively. The intersection point of these graphs is taken as the optimal cutoff threshold point. This cutoff point gave a value equal to 0.3.
- Based on this new value close to 92% values were observed to be rightly predicted by the model.
- The new values of obtained were, ‘accuracy=89.6%, ‘sensitivity=89.2%’, ’specificity=89.8%’.
- The lead score was calculated and resulted in the final predicted variables giving a target lead prediction of approximately 89% on the test data.
- Computing the Precision and Recall metrics
  - Also, the Precision and Recall metrics values calculated were 90.8% and 83.1% respectively on the train data set.
  - The Precision and Recall tradeoff, gave a cut off value of approximately 0.37.

Step 11: Making Predictions on Test Set

- Then the learnings were implemented on the test model and the conversion probability based on the Sensitivity and Specificity metrics were used to calculate the metrics on the test dataset. Observed values were ‘accuracy=90.0%’, ‘sensitivity=89.3%’, ‘specificity= 90.4%’.