Summary Report

Problem Statement:

- X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- This analysis is done to find ways to get more students and working professionals to join their courses. The basic data provided to us tells us lot about how the potential customers visit their website, the time they spend here, how they reached the site and the conversion rate.

Solution Summary:

Step1: Inspecting the data frame Read and analyze the data

Step2: Data Cleaning All the variables that had high percentage of null values in them are dropped. This step also included the missing values. Also, the outliers were identified and removed.

Step3: EDA Then we started Exploratory Data Analysis to get a feel of how the data is oriented. It was found that a lot of elements in the categorical variable were irrelevant. Those variables were dropped.

Step4: Dummy Variable Creation Then we went on creating dummy variable for the categorical variables. For numeric values, we used MinMax scaler.

Step5: Test Train Split We split the dataset into 70% and 30% for train and test respectively.

Step6: Feature Rescaling We used MinMax scaler to scale the original numerical variables. Then using stats model we created our first model, which would give us a complete statistical view of all the parameters of our model.

Step7: Model building using statsmodel RFE was done to attain the top 15 relevant variables. Later, the rest of the variables were removed on the basis of their VIF and P-value.

Step8: Model Evaluation We created a confusion matrix. Optimum cutoff value(using ROC curve), was used to determine accuracy, sensitivity and specificity.

Step9: Making Prediction on Test Set Prediction was done test data frame and with and optimum cutoff of 0.35 with accuracy, sensitivity and specificity of 77%, 76% and 60% respectively.

Step10: Computing the Precision and Recall metrics We implemented the learnings to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be 77%.