SUMMARY

Problem Statement:

- X Education provides courses to industry professionals.
- Even after getting a lot of leads the lead conversion rate is around 30% which the company believes is poor.
- The company wants to build a model to assign a lead score for every lead and focus on calling and emailing whose lead score is high.

Data Loading and Understanding:

- Importing warnings and libraries such as numpy pandas matplotlib seaborn sklearn and reading the csv file.

Inspecting the Data Frame:

- Checking columns shape, dimension, size, datatypes, info and statistical summary for the dataframe.

Data Pre-processing and Cleaning:

- Replacing columns with values SELECT as null and checked duplicates.
- Checking missing values and its percentage for all columns and dropping columns which had missing values greater than 40% except Lead Quality. Replacing null values of Lead Quality with Not Sure.
- Imputing missing values of categorical variables with mode and mean or median in case of numerical variables.

E.D.A:

- Performing Uni-variate & Bi-variate analysis.
- Replacing values for some columns to others as there were so many categories for which the count is negligible.
- Performing outlier analysis and removing outliers.
- Dropping certain columns that didn’t seem to be helping in our analysis.

Data Preparation:

- Binary mapping on columns which contains values as Yes or No only.
- Creating dummy variables for categorical columns and dropping the columns for which dummy variables have been created.

Train Test Split:

- Dividing the dataset into X and y and performing train test split with a proportion of 70:30 % values using model selection from sklearn.

Feature Scaling:

- Using Standard Scaling to scale the numerical variables.

Model Building:

- Creating our initial model using statsmodels and analyzing the complete statistical view of our first model.
- Using RFE going ahead and selecting top 10 features.
- Using manual feature elimination, building model by dropping variables which contain high VIF value and p-value.
- Finally coming out with 9 variables for which VIF values as well as p-values seems fine.
- Making predictions on the train data by assigning a probability score.
- Deriving the confusion matrix and overall accuracy of the model come out to be 91.78%. Also calculating Sensitivity (85.6%), Specificity (95.7%) etc and other values for checking and understanding the reliability of the model.

Plotting the ROC curve:

- Plotting the ROC curve of the model features and area of curve came out to be 0.91

Finding optimal cutoff point:

- Plotting probability graph for Accuracy, Sensitivity and Specificity for probability values. The optimal cutoff point came out to be approximately 0.2
- Checking Sensitivity, Specificity etc again and precision recall tradeoff also at 0.2
- Predicting the results on the test set and calculating accuracy (91.31%) Sensitivity (87%) Specificity (93.8%)

Conclusions and Recommendations:

- Assigning a lead score for every lead according to our final model and predicted the values for the test set with 91% accuracy.
- The top 3 features are:
  - Tags_Lost to EINS
  - Tags_Closed by Horizzon
  - Lead Source_Welingak Website
- Focusing on advertisement on welingak website might result into more lead conversions.