Project Summary

This project involves a Logistic Regression model we prepared in order to help the X Education company to increase the sales of their product by assigning a lead score to the customers who visits their website. Lead score is assigned between (0 to 100) which can be used by the company to target their potential lead.

A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

The followed approach and the steps used to prepare this model are shared below:

Data Reading and understanding:

Here we tried to get the look and feel of the data, we observed following things:

- Data types of each column
- First few rows and how the data looks & dimensions
- Statistical aspects of the numerical features given in the data.
- Understood the features given in the dataset using data dictionary excel file.

Data Cleaning and Preparation:

Here we checked for discrepancies in the dataset:

- We have replaced some ”Select” values with NULL values as those are Nulls as per problem statement and business understanding.
- We have dropped the features with majority same value throughout all the records.
- Deleted the columns with missing data considering the threshold as 60%.
- Handled country feature, categorizing every country into "Others" which share very less value counts in the feature.

Dealt with multi-label and binary label features:

We created dummy variables for 14 multi-label features and mapped the dummy variables to 0 and 1 and the same we did for 2 binary features as well which didn’t have labels in the form of 0 and 1.

Missing Value Imputation and Outlier Treatment:

- Checked for null values and imputing them with appropriate methods
  - We used mode imputation for categorical columns.
  - We used mean imputation for numerical columns, if there is no skewness in data.
- We used median imputation for numerical columns, if there is skewness in the data.

EDA:

A quick EDA was done to check the condition of our data to proceed further for Train-Test split. We specially used box plots to make sure there are no such outliers affecting our data. The numeric values seems good and no outliers were found.

Train-Test Split:

The split was done at 70% and 30% for train and test data respectively.

Feature Scaling:

We used Standard Scalar technique using our sklearn library to standardize the data into -1 to 1 range.

Looked at Correlation between each variables:

Looked at correlation matrix using heat map before proceeding with the model building part.

Model Building:

Firstly, RFE was done to attain the top 20 relevant variables. Later the rest of the variables were removed manually depending on the VIF values and p-value (Thevariables with VIF < 5 and p-value < 0.05 were kept).

Model Evaluation:

A confusion matrix was made. Later on the optimum cut off value of 0.35 (using ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around (~ 92%) each. We focus in Recall as the business objective is to not miss out on any possible Hot Lead. So out of all prospects which can be converted (TP + FN), want to maximize the hot leads conversion following score of our model (TP)

Prediction:

Prediction was done on the test data frame and with same cut off with specificity of around (~93%).

Recommendations (few examples):

- The company **should make calls** to the leads coming from the `lead sources` `"Welingak Websites"` and `"Reference"` as these are more likely to get converted.
- The company **should make calls** to the leads who are the `"working professionals"` as they are more likely to get converted.
- The company **should make calls** to the leads who spent `"more time on the websites"` as these are more likely to get converted.
- The company **should not make calls** to the leads whose `last activity` was `"Olark Chat Conversation"` as they are not likely to get converted.
- The company **should not make calls** to the leads whose `lead origin` is `"Landing Page Submission"` as they are not likely to get converted.