**Assignment Summary**

This is project on analysing and building model on the data of “Leads” and predicts which industry professional is going to join the course in a X Education. The data contains all kind of features which describes the complete activity of the leads like total time spent in website, total visits and conversion rate etc.

The following are the steps involved in this project:

**Data Understanding and Cleaning:**

As we all know that the data which we get initially may not be in our required structure for modelling so we need to clean the data according to out requirement. In our case the data which we got is structured in all aspects except for “Null values” and a category of feature “Select”. So we removed the features which have high rate of null values and made certain imputation techniques on the features which have low nulls. Also there are few columsn with skewed data which we need to take care of like combining them to single category etc.

**Exploratory Data Analysis (EDA):**

In order to go further we need to understand the data that what are the columns or features of the leads are mostly effecting the lead to convert or not. So in order to find out those insights we need perform EDA on the dataset to draw the insights. The insights which we found are:

* The numeric variable seems to be good with no outliers.
* Some of the categorical variables are irrelevant with respect to conversion.
* Some of the features of both numeric and categorical variables have significant impact on conversion rate like (TotalVists, Time spent, and Specialization etc.)

**Dummy Variables:**

After we are done with EDA for all the categorical variables we have created dummies. And the original features and variables with missing values are removed.

**Train – Test Split :**

Now we divided the data into two parts such as Train\_set and Test\_set. The data is divided in 70% and 30% ratio. And the

**Feature Selection:**

As we have so many features after creating dummies we have made the required feature selection for modelling using RFE (Recursive Feature Elimination) method. We have fixed the number of variables to 15. We also scaled the numeric variables of the data using “MinMaxScaler”.

**Model Building:**

We build the model with logistic regression using “Statsmodels”. From the variables we got in RFE, we dropped the remaining features manually based on p|z| values > 0.05 and VIF (Variance inflation factor) > 5.

**Model Evaluation:**

We found the values of evaluation metrics which are Accuracy, Sensitivity and Specificity to describe the performance of the model. To find these we have created the confusion matrix. The values of the metrics came around 80% with Accuracy and Specificity, 65% with sensitivity. We have found the cut-off for the conversion probability using ROC curve.

**Model Predictions:**

With the cut-off of 0.32 we found the predicted values of metrics in train set are around 80% and with test set are also around 80% which seems to be a good model.

**Model Inferences:**

The features which contribute high in defining the potential leads are:

* Total Visits
* Total time spent on website
* In Lead source
* Leads Source\_Welingak Website
* Leads Source\_Reference
* Leads Source\_Olark Chat
* In Occupation
* Working Professional
* Other

The X Education sales managers should focus on the leads who have the above features high as they have high rate to be potentials leads.