Summary Report

Introduction-

The following is a summary of the process followed and learnings from deploying a logistic regression model for the case study where X Education has appointed us to help them select the most promising leads. The company required us to build a model wherein we needed to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO has given a goal of the target lead conversion rate to be around 80%, whereas the typical lead conversion rate at X education is around 30%.

Process followed in Case Study -

- **Importing Libraries** *- As a first step, we imported the relevant libraries and modules such as numpy, pandas, matplotlib,seaborn sklearn, statsmodels for all the data processing, cleaning, visualization and model building steps. Moreover the warnings were also ignored.*

- **Reading the Data** *- Secondly, the "Leads" csv file was imported using pandas and the columns, data types, null values, statistical summary and shape of data frame was inspected.*

- **Data Preparation** *- Data was processed by cleaning the data to take care of null values through removing columns with null values in excess of 40%. Also, the 'select' level was handled that is present in a lot of categorical variables by replacing it with np.nan. Also, other categorical columns were removed based on skewness, unique levels and no value addition. Levels within categorical columns like Specialization were also merged under umbrella levels based on logical deductions.*

- **EDA** *- Exploratory data analysis was performed wherein numerical and categorical variables were visualized w.r.t. The target variable conversion and outlier treatment and other data processing was executed.*

- **Dummy Variable Creation** *- For the categorical columns, dummy variables were created. For these variables dummy creation was performed - 'Lead Origin','Lead Source','Last Activity','Specialization','City','Last Notable Activity'*

- **Test-Train Split** *- The dataset was divided into train and test data set in the proportion of 70% and 30% respectively.*

- **Feature Scaling** - *For the numerical variables i.e. 'TotalVisits','Total Time Spent on Website','Page Views Per Visit' feature scaling was performed using Standard Scaler.*

- **Checking Correlation** - *Top 15 Correlations were found out and highly correlated dummy variables were dropped.*

- **Model Building** - *We first automated feature elimination using RFE for top 15 variables. Post that we built 5 Logistic Regression Models by understanding statistical significance through p values and multicollinearity by VIF to manually drop redundant features. In the final model the p values for all variables was 0 and also VIF was less than 5. The final model had 11 dependent variables. Also scores were assigned based on the predicted probabilities of conversion for the y-train.*

- **Model Evaluation** - *We built a confusion matrix to understand the performance metrics of model like accuracy, sensitivity and specificity. Also the ROC curve was plotted and there was a good amount of area under the curve as 0.86 indicating a decent model built. Post that also optimal probability cutoff was found out using the intersection of curves between accuracy, sensitivity and specificity. We also inspected the precision-recall curve for the model in this step.*

Learnings from the Model -

- *We learnt that the respective variables that were significant in determining the conversion of a lead were - 'Total Time Spent on Website', 'Lead Origin_Lead Add Form', 'Lead Origin_Lead Import', 'Lead Source_Olark Chat', 'Last Activity_Converted to Lead', 'Last Activity_Email Bounced', 'Last Activity_Had a Phone Conversation', 'Last Activity_Olark Chat Conversation', 'Last Activity_SMS Sent', 'Last Notable Activity_Modified', Last Notable Activity_Olark Chat Conversation'.*
- *It can be concluded that some factors like total time spent on website and some levels of lead origin and lead source positively impact the probability of a lead to convert. However, some levels of last activity and last notable activity are detrimental to the probability of lead conversion.*