SUMMARY REPORT

In this assignment we have to basically find the probabilities of customer visiting the X- Education website turning into lead that is the leads that are most likely to convert into paying customers.

We then went through the data set in detail. There are total 37 column and 9240 rows. There are Categories named 'Select' in all the categorical columns; This is nothing but the empty field left by the customers and it is as good as a null entry. Therefore we proceeded replacing the 'Select' field by np.nan.

We checked the data frame for missing values and skeweness. The highly skewed categorical columns and columns with high missing values (greater than 50%) are removed. There were few levels in categorical columns with very less amount of percentage so those categories are clubbed in a separate category called other and such categories are removed from all the categorical columns. Also the outliers from continuous columns are capped within the soft range.

In the next step we carried out the exploratory data analysis and univariate and multivariate analysis is performed. After the data cleaning and EDA the percentage of retained rows is checked. The percentage of retained rows turned out to be 84%; The 15% of the unnecessary data was removed.

We then proceeded with the data preparation; In this step we converted the binary column into zero's and one's. All the categorical columns are replaced by their respective dummies. After the dummification we divided the data in the ratio of 0.7 : 0.3.

The next step is Feature Scaling in this step all the numeric columns are scaled using standard scaler function. The correlation metrics is also checked.

In model building step we used statsmodel library. We checked the variables with their VIF value and P value. The columns with high P (greater than 0.05) and VIF value (greater than 5) were removed in various iterations until the satisfactory results were obtained.

The ROC curve is also plotted to check whether the model is performing good or not. Optimal cutoff point is found for the metrics sensitivity, specifity and

accuracy. Optimal cutoff is also found using precision and recall metrics. Finally we used our model on the test data to check its performance and we found the satisfactory results.

Major Learning's

- All the columns present in the data frame may not be useful. Columns with high amount of missing values lead to wrong results.
- Only few variables in the data frame have an influence over the target variable.
- Columns with high VIF value have high correlation and might not produce the satisfactory results.
- Columns with high P value Indicates that their presence is insignificant and target variable has no correlation with the respective variable and their coefficient value is just by chance.
- Optimal cutoff point is very important in order to correctly classify or predict the actual probabilities rather than manually deciding the cutoff values and compromising the model accuracy.
- Metrics such as sensitivity-specificity and precision - recall give the similar results and are very important to evaluate the performance of the logistic regression model.