**Lead Scoring Case Study Summary Report**

Company X Education creates online courses for industry professionals and markets them on several websites and search engines. People filling forms with their information are called leads. Sales teams contact these people and help them in selecting and pay for courses. Out of all those who register and pay are said to be converted.

The purpose of this study was to increase lead conversion rate by identifying highly promising leads(hot leads) to focus on by the sales team. We were required to build a model to assign a lead score between 0 to 100 to all the leads such that the higher the score more will be their chances of buying the course, and the target conversion rate was 80%.

We read and understood the data, performed EDA, built a logistic regression model, divided data into test and train sets, trained our model, and predicted on test data to see patterns to finally recommend solutions.

The data set had 37 columns with ‘Converted’ as the target variable representing whether the lead has converted. After replacing select values with null, columns containing more than 45% of null values were dropped. Also ‘Prospect ID’ and ‘Lead Number’ were dropped as they won’t affect conversion chances. In columns containing null values less than 45%, these were replaced by the most occurred value.

From EDA we got to know that most of the leads originated from landing page submission and API out of which the conversion rate of landing page submission is greater than API. Olark chat, organic search, direct traffic, and Google were lead sources that can create and were able to convert the maximum number of leads. The maximum number of leads converted did not want a free copy of mastering the interview, so It was not a useful strategy.

We split the data into Target and Independent variables and with the help of train\_test\_split from sklearn we split it further into training and testing data sets. Numeric variables from the training data set were scaled to bring in an equal range. The RFE algorithm is used with initial 20 variables of the training data set. The output of the RFE algorithm is combined with the columns of training data along with ranking. Columns that are ranked 1 by RFE were selected and all other columns were dropped. Constant was added and Generalized Linear Model was used to fit the training data.

We found that all the variables have P values less than 0.05, which indicates that these variables are good enough for prediction. There is no redundant variable present in the data as the VIF of all variables is less. The ROC curve has an area of 0.96 under it indicating a good predictive model.

The performance of the model on training data is:

•Accuracy 89.33%

•Sensitivity 93.53%

•Specificity 86.73%

The performance of the model on the test data set is:

•Accuracy 89.14%

•Sensitivity 92.7%

•Specificity 86.98%

From this, we can say that the model seems to predict the conversion rate very well and we are able to get a conversion accuracy of more than 80%, which was required. Focusing more on the sources of leads like Direct Traffic, the last notable activity like Olark chat conversation, or paying attention to the current status of the customer will provide better results.