Summary

X Education sells online courses to the industry professionals. Even after getting lot of leads the lead conversion rate is very poor. To make this process more efficient, we will identify the most potential leads, also known as “Hot Leads”. After successful identification of these leads, the conversion rate will go up as the sales team will now be focussing more on communicating with the potential leads rather than making calls to everyone which results in waste of resource and time.

We will classify the hot leads from all the leads by building a Logistic regression model. In this process, the first step is to analyse the data and understand the shape and its values. Upon the inspection of data we found many values retained as “select”. We converted them as missing values and afterwards total missing values were found out. We removed the columns with more than 40% missing values.

For the other features where null values were present, we inspected each features one by one and in some of the features they were grouped and designated as a different class that is “not given”, and for the columns where there were less than 1% missing values, they were replaced by the mode or median of those features. We again dropped some columns due to data imbalance as there was no pattern found.

We visualised the Categorical variables and Continuous variables with the target variable to find patterns in the data and draw insights from it. We also checked for outliers in the continuous variables and then outlier treatment was done.

Next comes the stage where we need to do data preprocessing so it it ready for model building. We converted features with binary class to 0 and 1 and One hot encoding was done for multi-class categorical variables. We split the data into training and test sets with a ratio of 7:3. We then scaled the continuous variables by using MinMaxScaler. We further checked the correlations between the features by drawing a heat map.

Now we are ready to build our ML model. We applied Recursive Feature Elimination(RFE) to find the top 20 features. Next, we used GLM(Generalised Linear Model) and chi square test to find the most statistically significant features on which inferences can be drawn. This was done by selecting the variables with p-value less than 0.05 and a Variance Inflation Factor less than 5. We predicted the conversion probabilities for the train set and by taking a threshold value of 0.5, train set metrics were determined.

- **Accuracy score - 81.3%**
- **Sensitivity - 71%**
- **False Positive Rate - 12%**
- **Positive Predictive value - 80%**
- **Negative Predictive value - 82%**

We also need to figure out the best threshold that can be taken so that the sensitivity of the model is as high as possible and the False Predictive rate is as low as possible. For this purpose, we drew an ROC(Receiver Operating Characteristic) curve of area 0.88 and found an optimum cut-off point as 0.38. We used this cut-off point to calculate the performance metrics one more time, and they were,

- **Accuracy score - 81%**
- **Sensitivity - 79.5%**
- **False Positive Rate - 18%**
- **Positive Predictive Value - 74%**
- **Negative Predictive Value - 86%**

It is also important to check the model with other performance metrics such as Precision and Recall. And they were obtained as

- **Precision score - 79.3%**
- **Recall score - 71%**

The optimum cut off point from precision recall curve is approximately 0.42.

Now our model is ready for predictions on unseen data. We used our model to predict conversion probabilities on the test set, and we found the performance metrics as,

- **Accuracy score - 80.6%**
- **Sensitivity - 76%**
- **False Positive Rate - 16.7%**
- **Precision Score - 72.7%**
- **Recall Score (Sensitivity) - 76%**