Lead Score Conversion- Logistic Regression Model Summary

To build a lead generation model for X Education, the historical data provided was first read into python. Variables with a large skew were dropped as they would not contribute towards variation in the dependent variable 'Converted'.

The next step was to observe the null values in the data and then impute the missing data through appropriate values: mode for categorical data, mean value for numerical data with no outliers and median for numerical data with outliers.

After this, dummy variables were created for categorical variables and outliers were checked for, following which the data was divided into 'train' and 'test' sets with a proportion of 70:30. The next step was to scale the continuous variables and then check for variables with high correlation values.

After these variables with very high correlation values were dropped, we performed modelling. Modelling begins with RFE selection of the 15 most important variables which are then inputted to the logistic regression model.

After dropping variables with a p-value greater than 5 percent (these variables are not significant predictors of the dependent variable), we check if the VIF values (a test for multicollinearity) are within the threshold of 5.

The next step was to create a data-frame with the actual Converted values and the predicted probabilities based on probability thresholds. We obtained accuracy, specificity, sensitivity metrics for both a high probability threshold of 0.8 and a lower probability threshold of 0.5.

The ROC curve was also plotted with an obtained area of 0.89, suggesting decent model performance. In order to find the optimum value of cut off for the probability, we plotted the accuracy, sensitivity, specificity thresholds on the same graph and obtained the value of 0.38 as the optimum threshold. From the train sets we obtained precision of 73.8 percent and a recall/sensitivity of 77.6 percent for a probability threshold of 0.38. The specificity figure for the train set was 83 percent.

Predictions were made on the test set, with a sensitivity of 77 percent (only 3 percent less than the ballpark figure of 80 percent provided by the CEO) and a specificity of 80 percent.

Using the model built we recommended that :

- For an aggressive marketing strategy, the team use a lower probability cut off of 0.38 to pursue leads.
- For a conservative marketing strategy for definite conversion of leads, the team use a higher proability threshold of 0.6, keeping in mind that sensitivity decreases as the probability value increases and hence the threshold chosen needs to be optimum.