Problem Statement:

X Education sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company needs a model wherein you a lead score is assigned 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, in particular, has given a ballpark of the target lead conversion rate to be around 80%

Solution

Exploratory Data Analysis and Data Cleaning:

1.The dataset was loaded into the python notebook and explored statistically and visually to get an idea of outliers, distribution of data, feature redundancies, and so on.

2.Correlations between the variables were also identified using heatmap .

3.The redundant variables were removed.

4.Some of the columns had string values“Select”which meant that the user did not select a particular value in the form. This means it is as good as null values.

5.Columns with more than 40% null values were removed from model building.

6.Outliers were removed statistically from the data to avoid noise in the model building.

Data Preparation:

1.Categorical variables were converted into numerical data of 0 or 1 using dummy variables.

2.The dataset was split into training and testing datasets in a ratio of 70:30

3.Data points were standardised to similar scale using Standard scaler for numerical features to avoid bias of some variables in higher scales in the model.

Model Building:

1.Initial model was created by selecting 15 variables using the RFE technique.

2.Insignificant variables that were identified were removed to optimize the model for better modelling. VIFs were also checked to see the multicollinearity between the model features.

3.An important step in logistic regression is to find the optimal cutoff for the probability to fit the business needs and to have good accuracy, sensitivity and specificity. We obtained 0.398 as the cutoff from the plot between accuracy, sensitivity and specificity and probability range.

4.Recall and Precision view was then considered to finalise the cutoffs for the prediction in test dataset. The cutoff in the Recall-Precision graph was obtained as 0.445.Model evaluation was completed by checking the values of all the performance measures namely ROC curve, between accuracy, sensitivity and specificity, Recall and precision. All of them were in acceptable range. Optimal point 0.398 was selected as it had higher sensitivity and recall of 81% .

Model Evaluation:

Predicted values were calculated using the model. Lead score is 100 multiplied by the log odd which is predicted, to have a value between 0 and 100.

Conclusions and Recommendations for the Company Strategy:

1.The model evaluation steps revealed that the accuracy, precision and recall parameters are acceptable values. The Recall score is a bit greater than precision score as well. This fits the business needs for the future.

2.Focus on the leads from Add form as they contribute most to conversion

3.People who are working professionals should be ideal target

4.Leads from Welingak website are convertible, so advertising revenue should be focused on this website.

5.People who don’t want to be mailed should be avoided.

6. People with no specific specialization should be avoided.