OBJECTIVE

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score 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%.

STEPS FOLLOWED TO ARRIVE AT THE SOLUTION:

Data Cleaning:

The overall data was checked wherein a high number of null values were found. Hence, the variables with 50% and more were filtered and then checked the value count. The qualitative null value was replaced by ‘not provided’ based on the columns. Then the columns which have null value more than 35% were removed based on effective data. Some null values were replaced with mode, mean and median value as per qualitative and quantitative data.

Data understanding using Exploratory Data Analysis (EDA):

EDA was done to check the condition of data. First the Univariate Analysis was performed for both Continuous and Categorical variables and then performed Bivariate Analysis with respect to Target variable (lead converted). After studying the EDA, it was observed that some columns are irrelevant for the study and these variables were dropped.

Dummy Variable Creation:

This step entailed the creation of dummy variables for categorical variables.

Test-Train Split of the Data:

The dataset at hand was split into the Test set and the Train set wherein 70% of the data is the train set and 30% of the data is the test set.

Scaling of the variables:

A few variables had values that were in the range of 0 to 1 while the others were not. This could cause collinearity issues hence the variables were scaled using Min Max scaler.

Feature Selection and Model Building:

The process of feature selection was done using Recursive Feature Selection. 25 best features were first selected and then the p-value and the Variance Inflation factor was looked at to further eliminate features and arrive at a good model.

The 25 variables selected are: 'Lead Number','Total Time Spent on Website', 'Page Views Per Visit','Lead Origin_Landing Page Submission', 'Lead Origin_Lead Add Form','Lead Source_ Direct Traffic', 'Lead Source_Google','Lead Source_Olark Chat', 'Lead Source_Organic Searc h','Lead Source_Reference', 'Lead Source_Referral Sites','Lead Source_Welingak Website', ' Do Not Email_Yes','Specialization_Finance Management',‘Specialization_Human Resource Management','Specialization_Marketing Management','Specialization_Operations Manageme nt', 'Specialization_Select','Specialization_Travel and Tourism', 'Specialization_not provided', 'What is your current occupation_Unemployed','What is your current occupation_Working Pr ofessional','What is your current occupation_not provided','What matters most to you in choo sing a course_not provided' and 'A free copy of Mastering The Interview_Yes'.
The logistic regression model was run (GLM) and variables that had a p-value greater than 0. 05 (insignificant variables) and variables with Variance Inflation Factor (VIF) of greater than 5 were dropped. The final model chosen has a total of 13 variables and these variables are: To tal Time Spent on Website', 'Lead Origin_Landing Page Submission','Lead Origin_Lead Add Form', 'Lead Source_Direct Traffic','Lead Source_Olark Chat', 'Lead Source_Referral Sites',' Lead Source_Welingak Website', 'Do Not Email_Yes','Specialization_Select', 'Specialization _not provided','What is your current occupation_Working Professional','What is your current occupation_not provided' and 'TotalVisits'

<u>Model Evaluation</u>:

The y values were first predicted and a separate data frame with the actual converted flag and the predicted probabilities was created. The confusion matrix was the formed which is as follows:

Predicted	Not converted	Converted
Actual		
Not converted	3531	424
Converted	864	1649

The accuracy was then found, and the value obtained was 80% which is a good value but looking just at the accuracy is not enough to assess the goodness of the model. Hence, other metrics were also looked at. The sensitivity value is 0.66 and specificity value is 0.89. The ROC curve was derived. A good classifier is that the curve must be as close to the boundary as possible. Here, the curve is far from the diagonal line towards the upper-left corner. The area under the curve of the ROC is 0.87 which is a pretty good value and hence the model at hand looks like a good model. The cut off was then determined at a point wherein the Accuracy, Sensitivity and specificity intersect which in this case was at around 0.39 and was used as the cut-off. The accuracy, sensitivity and specificity were then found using this cut-off and the values were .81,.74,.85 respectively. The lead score was calculated, and the conversion rate was looked at. This conversion rate was found to be 74% which is a significant increase in the conversion rate when compared to the conversion rate of 39%.

<u>Precision- Recall</u>:

The precision and recall score were looked at and the precision score was found to be 0.79 and the recall score was found to be 0.65. The precision-recall trade-off was then looked at. The precision and the recall were plotted, and they were found to intersect at 0.4 which was then used as the cut off and the accuracy, sensitivity and specificity was looked at which was found to be 0.81,0.734 and 0.85 respectively.

Predictions on the Test data:

The final step was to make predictions on the test data set. 0.4 was used as the cut-off, lead score was found, and the conversion rate was calculated. The conversion rate on the test set is 72%. Accuracy, sensitivity, and specificity values were calculated on the test set. The accuracy value on the test set is 0.81 which is a good value, the sensitivity value is 0.72, and the specificity value is 0.86. The choice of cut-off has worked out well as the sensitivity and specificity has worked out well on the test set. Precision and Recall on the test data was calculated and the precision score is 0.76 and the recall score is 0.72. The precision and recall intersect at 0.39.

CONCLUSION

The Accuracy, Sensitivity and Specificity values of the test set are around 81%, 72% and 86% which are approximately close to the values obtained from the train set implying that the model is good. The variable TotalVisits has the highest coefficient value implying that this variable has the highest significant positive influence on whether the lead will be converted or not. The variable total time spent on the website has a high coefficient value implying that it has the highest positive influence on whether the lead will be converted or not. Leads originating from add form also has a high positive impact on whether the lead will be converted or not. The lead source being Welingak Website also has a positive impact on whether a particular lead converts. Working professionals seem to have a high positive impact on whether a particular lead converts. The lead source being Olark Chat also has a positive impact on whether a particular lead converts. Variables that have a negative impact on whether a lead will be converted or not are the leads originating from landing page submission, direct traffic, Specialization being Select, don’t email being yes, Lead Source being Referral Sites, specialisations and occupations that are not provided and the last notable activity that is modified. Since the conversion rate is 74% on the train test and 72% on the test set which is a significant increase in the rate of conversion as compared to the conversion rate of 39%, the company X Education can look into the impact that these variables have on the conversion of leads and work towards improving the conversion rate by focusing on leads with a higher lead score.