Lead Score Case Study Summary

We conducted this case study for X Education with the objective of providing them insights to successfully convert leads they received into students of the platform. Our goal was to generate a machine learning based model that identifies the highest probability of converting customers to students.

We started this by assessing the data and identifying various columns, understanding what they mean by assessing the data dictionary. We initially saw that there were a large number of 'Select' values spread across the dataset. These values meant that the user had made no selection in those fields. We decided to replace these values with NaN values and treated them appropriately later. Then we conducted an analysis on the percentage of null columns in the dataset, tackling the ones with the highest percentage of null values first. We removed columns that had over 70% null values and for the remaining columns assessed them individually.

While assessing our columns individually we found that various columns were actually summarized into one column already. Therefore, it did not make sense for us to keep these columns and we decided to drop them entirely. Examples of such columns are Search, Newspaper Article etc., they are already represented in the 'Lead Source' column. The distribution represented in these individual columns was very well represented by the data in the Lead Source column. There were a few columns that had highly skewed data, i.e. data pointing in one direction only. The Country, What matters most to you in choosing a course, are a few examples. Most of the leads, 95% and above, mentioned that they were from India and were looking for better career prospects. We dropped these columns as well.
The tendency of skewed data to sway the model heavily towards its direction makes the model incapable of predicting the results correctly.

While analyzing the various columns we performed univariate and bivariate analysis on them. Bivariate analysis was carried out with the Converted column as a benchmark. This analysis yielded some very important insights that we have mentioned below. Key being that the longer the user stayed on the website, the higher the chances of them converting.

Of particular interest to us was the 'Total Time Spent on The Website' column. This column had highly varied data that we had to properly convert to correct metrics to make better sense of it.

Certain columns had a large mix of values, some outliers and a small number of null values. We had to perform appropriate outlier & null value treatment for each of these. In certain cases such as that of the Specialization column, we could not have taken the Mode value to impute the null columns. This is because we had to consider the fact that the mentioned options in the form might not have represented the applicants specialization correctly. We decided to club these values into one field and later assess them.
In the case of numerical columns like TotalVisits we imputed the null values with the median value. This is because the difference between the median and mean was very less. We also capped any outliers present instead of dropping them so that all the rows of data are retained. Capping was done with by replacing the lowest values with the 1%ile & highest with the 95%ile value in the column.

Having conducted our EDA and prepared our data free of any anomalies we proceeded to preparing our data for the Logistic Regression Model.

We created dummy variables from our final 12 variables and correctly dropped all the original columns, other category variables that we had created. Then we performed the train-test split using the 70-30 method for splitting.
Using the StandardScaler we scaled our numerical columns so that all the variables follow similar units. Scaling helps us in standardizing our dataset and preventing any features that have higher units to skew the model in its favor. We assessed the split datasets and plotted a correlation heatmap to identify any variables with high collinearity. We found 2 such variables and dropped them from both the training and test sets.

Finally, we proceeded to creating our logistic regression model.
Firstly, it is important that we define our model acceptance criteria:
- The model does not over-fit
- The model is simple enough to be understood
- The model is built using significant features.
- The VIF value is under 3 & the p value is under 0.05 for each feature
- The accuracy, sensitivity and specificity of our model after test are at least 80%(+- 1% between all 3 parameters)

We created a basic logistic regression model with all our features from the scaled training dataset. The GLM summary report from this model provided us the base benchmarks for our model. Based on the above criteria we performed RFE with 20 variables and began creating and the models. We eliminated any variables that had high p values and VIF values. Eventually we generated a model with 15 variables that we performed training and testing on.

On both our training and testing models we predicted the probability score for converting the leads and correctly added them to a new table along with the customer id. Once we had our scores and probability of converting a lead, we checked for the accuracy, specificity, sensitivity, precision and recall. Our model performed very well on these statistics. WE generated the ROC curve and the probability cutoff curve to find our optimal cutoff, which is 0.33.

Assessing our model at the optimal cutoff we saw that our model satisfied our condition of 80% sensitivity.
We then compared then ran our model against our test set and achieved a similar result of 80% sensitivity.
Thus confirming our model is correct and completing the modeling.

Finally, we generated the table which contains lead scores for the leads in the original dataset. These scores can now be used to convert leads to students of the platform. The bench mark being that the leads above a score of 33 have a higher chance of converting. The higher the score, the better the chances of conversion.

Through the course of this case study for X Education we have identified the following key aspects:

- Most applicants would like to join a course to have better career prospects
- X Education has the highest conversion rate of individuals who are referred to them
- Overall it is safe to say that the more time the user spends on the website, the better their chances of becoming a student.
- Hot Leads are identified as 'Customers having lead score of 33 or above'
- Sales Team of the company should first focus on the 'Hot Leads'
- Higher the Lead Score, higher the chances of conversion of 'Hot Leads' into 'Paying Customers'

- The 'Cold Leads'(Customer having lead score < 33) should be focused after the Sales Team is done with the 'Hot Leads'