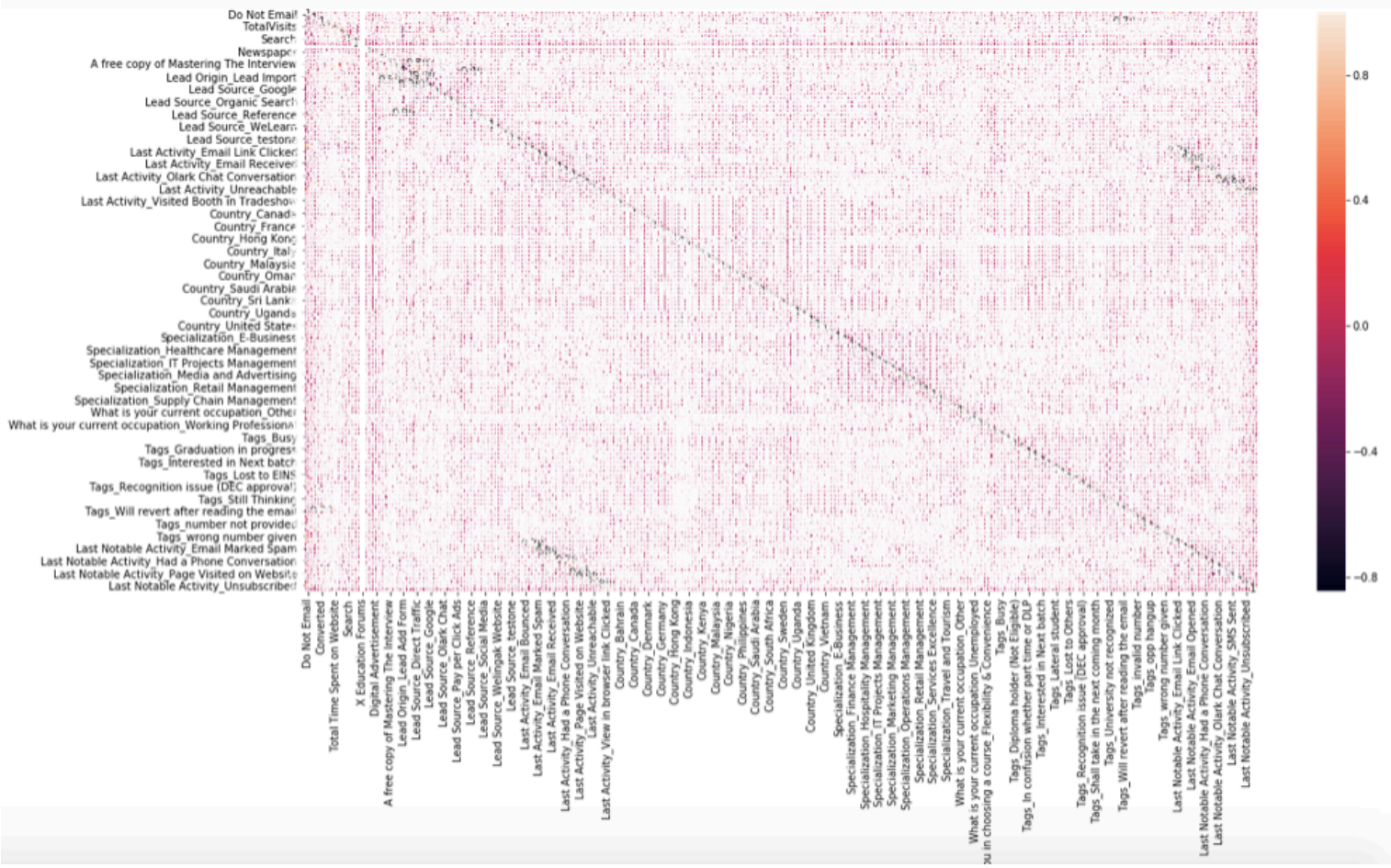
Lead Scoring

X Education

26 August 2019



Problem Statement

An education company with the name of X Education sells online courses to industry professionals. The issue is that their lead conversion rate is very low due to the huge amount of leads they are able to gather through various marketing channels. Hence the education platform wants to build a model which can assign a lead score to each of the leads such that a customer with a higher lead score have a higher conversion chance and vice versa. Through this lead score the CEI wants to focus on leads with a higher score for a better conversion rate - something which is higher than 80%.

Analysis

Importing and Inspecting the Data

We have started by importing and inspecting our lead data. We have checked few attributes like total number of rows and columns, data types of all the columns and basic overview of the dataset.

Data Cleaning

Post gathering few essentials, we moved on to data cleaning as it is the crucial part of modelling. We see that a few columns have 'Select' which is as good as null value. This is because since the data is gathered from a web form which in general has the option of 'Select' if nothing is selected in the form. Hence we convert this type of entries as null values and then find out the percentage of null values. From this we see that four columns have very low percentage of null values. Hence we drop these null values by rows as we have a lot of data to work with.

```
# Removing rows wherein below mentioned columns have null values. As count/percentage of null values is very less.
df = df[~pd.isnull(df['Lead Source'])]
df = df[~pd.isnull(df['TotalVisits'])]
df = df[~pd.isnull(df['Page Views Per Visit'])]
df = df[~pd.isnull(df['Last Activity'])]
```

Next we drop columns with very high null values (null value percentage more than 40%) as dropping rows will result in huge data loss and imputing them will result in huge bias. Then we checked the counts of values in variable - Country and Specialization. Since

India was an outlier in this data we imputed the remaining nulls with India whereas in Specialization we have imputed the remaining nulls with 'Others'. This was done as we thought that people might not have found their educational specialisation as most of the leads are from India and as we know India has a lot of graduation streams which are not mentioned in the Specialization column like Bachelor of Arts/Science/Engineering etc.

Outliers, Categorical Data Conversion and Splitting the Data

Further we checked for outliers using the describe function with the percentiles. We did see that there were outliers in three of the numerical variables - TotalVisits, Total Time Spent on Website and Pages Views Per Visit. However we didn't remove these outliers as these outliers can increase the probability of a lead getting converted.

	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit
count	5666.000000	5666.000000	5666.000000	5666.000000
mean	0.452877	3.724497	543.632369	2.582031
std	0.497818	4.921537	563.463335	2.062967
min	0.000000	0.000000	0.000000	0.000000
25%	0.000000	2.000000	56.000000	1.250000
50%	0.000000	3.000000	298.000000	2.000000
75%	1.000000	5.000000	1026.750000	4.000000
90%	1.000000	8.000000	1438.000000	5.000000
95%	1.000000	10.000000	1597.500000	6.000000
99%	1.000000	17.350000	1850.350000	9.000000
max	1.000000	251.000000	2272.000000	16.000000

Post outliers correction, we have converted binary variables to columns with 0/1 entries as working with 0/1 as an indicator is easier. Then we have converted the remaining categorical variables with multiple levels to 0/1 using the get_dummies function and dropping the first level and the actual variable.

Once our data became ready for modelling, we used the train_test_split function from sklearn library to split the dataset into two groups - training data and testing data. After splitting we have scaled our numerical variable only excluding the 0/1 columns and

checked the data imbalance or conversion rate which came out to 45.28% which is pretty good to use for modelling.

Correlation

Before we began with modelling, we tried checking correlation. However since we had a lot of columns our approach here didn't work.

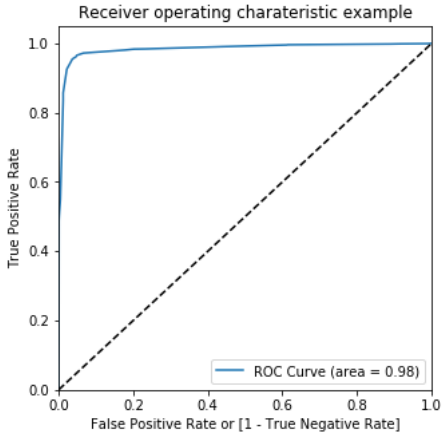
Modelling with RFE and Statsmodels

We started our modelling process by using RFE and LogisticRegression modules from sklearn to reduce the effort. We built our first model using only top 15 variables. Next we imported statsmodels for statistical assessment of the model. Through this we saw our model had few variables with very high p-values but when we checked their VIFs it came out too low. However we have also created confusion matrix using 0.5 as the cutoff and assessing accuracy of the model on this cutoff.

Since our model had variables with very high p-values, we started dropping them one by one after rebuilding our model with one less variable each time. We did eight iterations to come to an optimal model which had p-values less than 0.05 and VIFs less than 5.

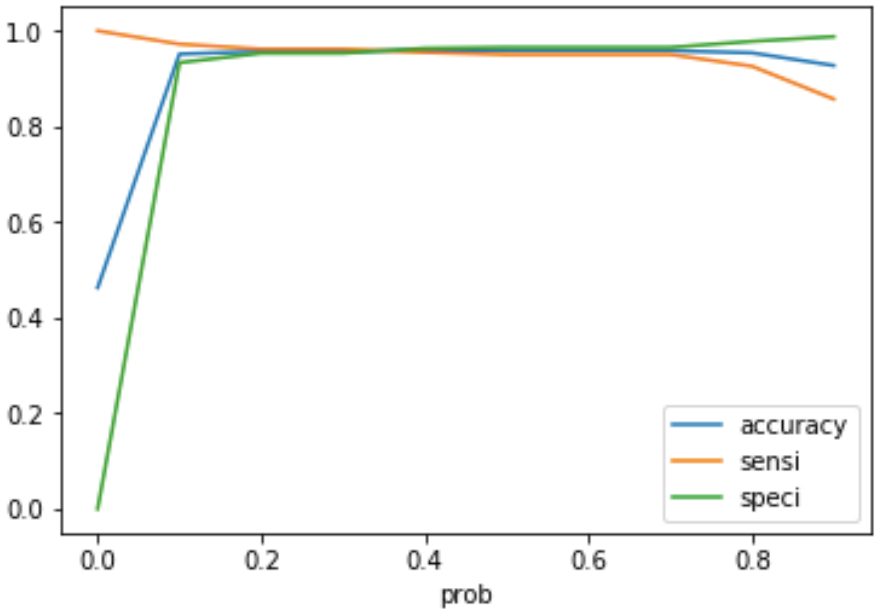
Metrics - Sensitivity and Specificity

Once we were finished with modelling our data, we moved to assessing our model in terms of sensitivity as the goal of our model is to increase the lead conversion rate which means high sensitivity. Hence we checked the sensitivity of the model using the same 0.5 as the cutoff and sensitivity came out to 95% which was better than the requirement. Next we plotted the ROC curve on our assumed cutoff. Below is the representation.



Optimal Cutoff Point

Even though we got a good sensitivity percentage, still we would like to increase our models efficiency in terms of sensitivity. Hence we create the matrix of accuracy, sensitivity and specificity with various probability cutoffs and then plot it for visual inference.



As we see, 0.4 is the optimal cutoff point. Hence we update our training model with the same and re-calculate the confusion matrix and sensitivity which came out to be 95.47%.

Prediction

Now we use the eighth model, predict function and optimal cutoff on our test dataset. Next we build another confusion matrix and check the sensitivity which came out to be 94.93% which is pretty good since we had 80% as ur target. Hence we could say that our model can easily predict which leads will get converted or not.

Business Model

The most notable variables are Tags_Closed by Horizzon, Tags_Lost to EINS which essentially means that students/ industry professionals were actually looking a course. However due to some reason our lead was lost to other educational platforms.

The next important variables are Tags_Will revert after reading the email and Last Activity_SMS Sent. Sales team should also focus on those leads who ask for an email/sms or follow-up.