Lead Scoring Case Study

Ankur Sugandhi

Problem Statement

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. The typical lead conversion rate at X education is around 30%.
The main Agenda is to help sales team and make the process more efficient by identifying “Hot Leads” based on the data provided with ball park target of lead conversion around 80%.

Steps involved

Below are the steps involved in analyzing and understanding the outcome:

a) **Step 1: Reading and Understanding the Data :**Data Summary including type, shape and info.

b) **Step 2: Data Cleaning & EDA :** Handling missing values, replacing less important attributes with others, Graphical representation

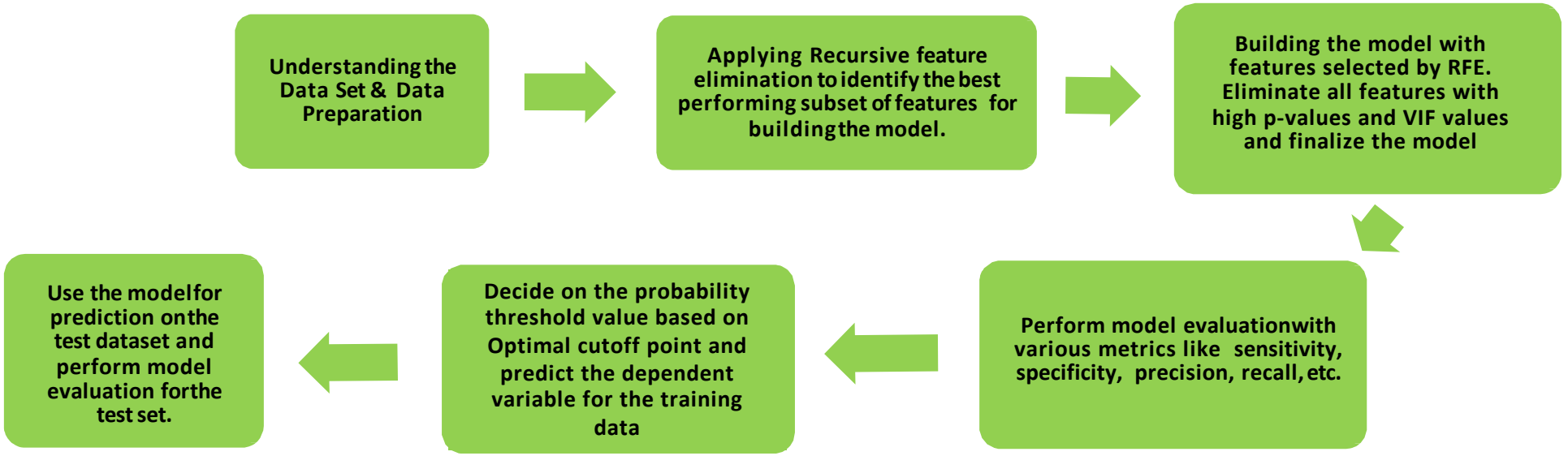
c) **Step 3: Data Preparation :** Dummy variable, Probability check, Standard scaling, correlation

d) **Step 4: Model Building :** Logistic regression, RFE, VIF & p-values, Optimal probability, confusion matrix, Model performance

e) **Step 5: Final Analysis :** Lead Score

Solving Methodology

The entire case study into various checkpoints to meet each of the sub-goals. The checkpoints are represented in a sequential flow as below:



List of Column Dropped

The following data preparation processes were applied to make the data dependable so that it can provide significant business value by improving Decision Making Process:

Columns Dropped -> 'How did you hear about X Education', 'Lead Profile', 'Lead Quality'
Reason: More than 50% Null values

Columns dropped -> 'Last Activity', 'Country', 'What is your current occupation', 'What matters most to you in choosing a course', 'Tags', 'Do Not Call', 'Search', 'Magazine', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations', 'Receive More Updates About Our Courses', 'Update me on Supply Chain Content', 'Get updates on DM Content', 'I agree to pay the amount through cheque', 'City'
Reason: As more than 80% values are not unique

Columns dropped -> 'Prospect ID', 'Asymmetrique Activity Index', 'Asymmetrique Profile Index', 'Asymmetrique Activity Score', 'Asymmetrique Profile Score'
Reason: Sales Team Generated information (Not system Generated)

Data Cleaning

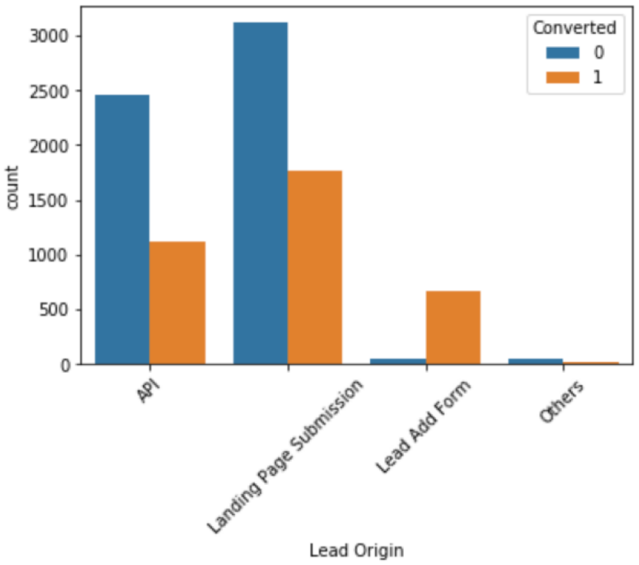
Converting Yes, No values in column with 1 & 0
For better analysis (Columns -> 'Do Not Email', 'A free copy of Mastering The Interview')

dropping rows having more than 3 null values

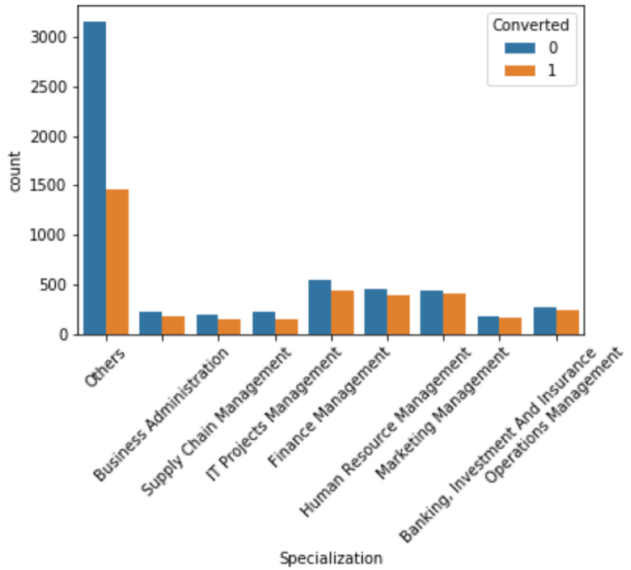
Imputing missing values with imputation techniques:
Lead Source – Mode
TotalVisits – Mode
Page Views Per Visit – Mean

	Lead Number	Lead Origin	Lead Source	Do Not Email	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Specialization	A free copy of Mastering The Interview	Last Notable Activity
0	660737	API	Olark Chat	0	0	0	0	0.0	Others	0	Modified
1	660728	API	Organic Search	0	0	5	674	2.5	Others	0	Email Opened
2	660727	Landing Page Submission	Direct Traffic	0	1	2	1532	2.0	Business Administration	1	Email Opened
3	660719	Landing Page Submission	Direct Traffic	0	0	1	305	1.0	Others	0	Modified
4	660681	Landing Page Submission	Google	0	1	2	1428	1.0	Others	0	Modified

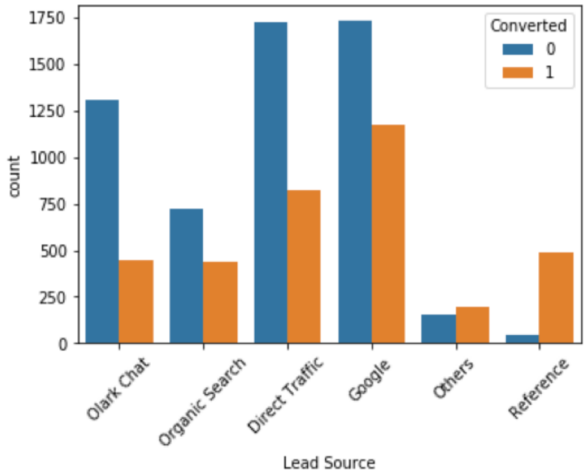
Graphical Representation



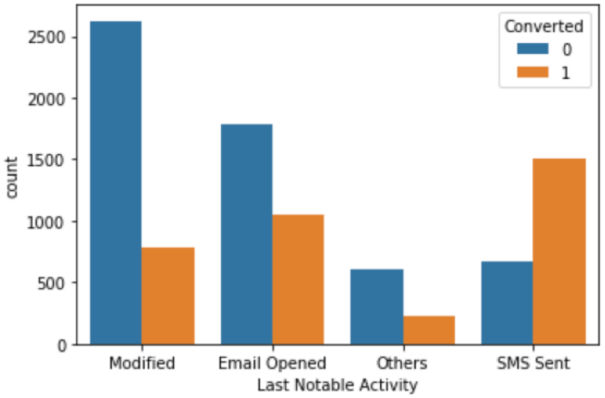
Lead Add Form has better conversion rate



Not much Significance



Reference has better conversion rate



SMS Sent has better conversion rate

Pair Plot

Total time spent on the website increases the probability on conversion

Recursive feature elimination

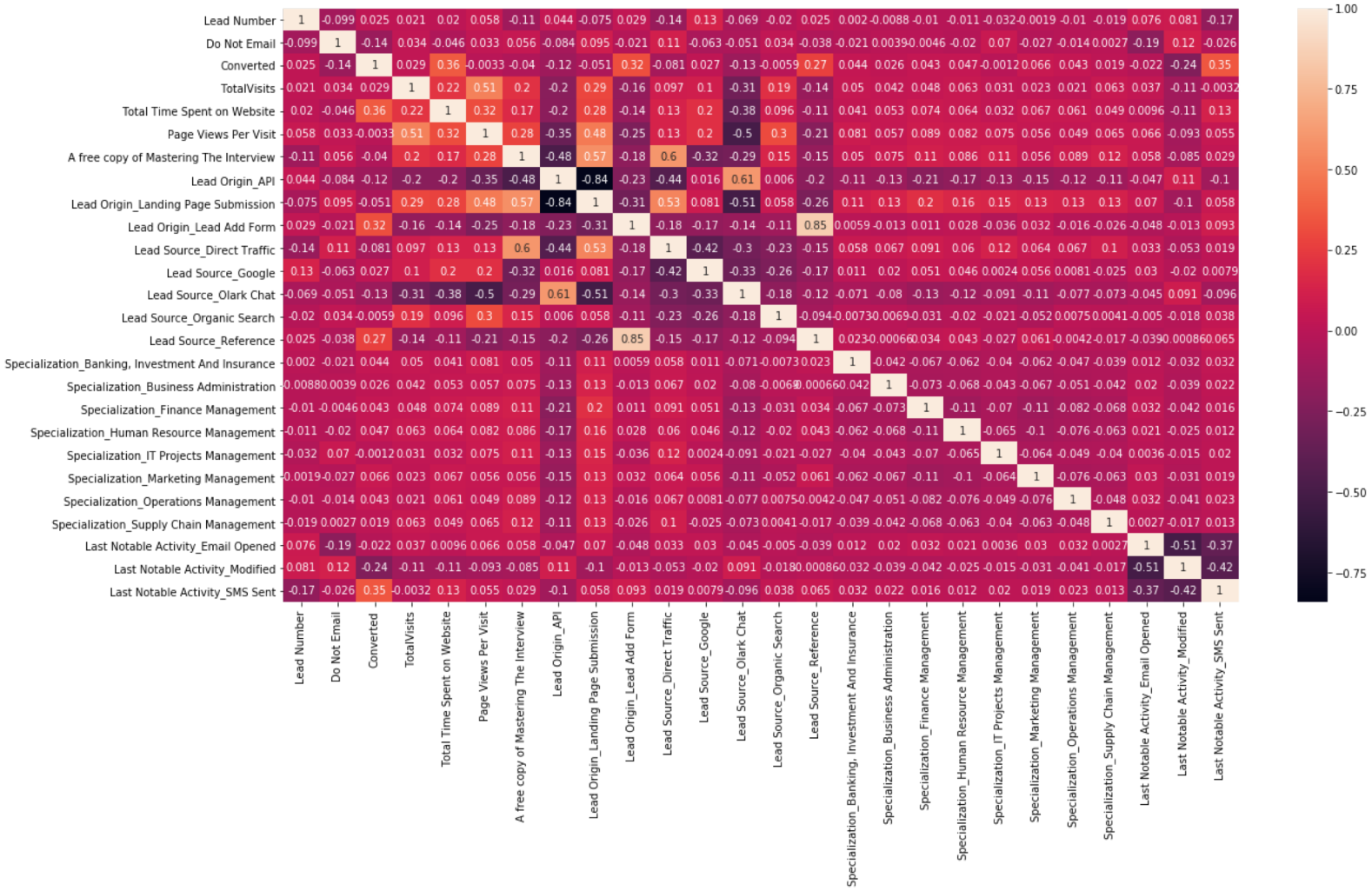
•**Recursive feature elimination** is an optimization technique for finding the best performing subset of features. It is based on the idea of repeatedly constructing a model and choosing either the best (based on coefficients), setting the feature aside and then repeating the process with the rest of the features. This process is applied until all the features in the dataset are exhausted. Features are then ranked according to when they were eliminated**.**

Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6452
Model Family:	Binomial	**Df Model:**	15
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2889.8
Date:	Sat, 24 Oct 2020	**Deviance:**	5779.7
Time:	19:00:01	**Pearson chi2:**	6.72e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.1131	0.084	-13.188	0.000	-1.279	-0.948
Do Not Email	-1.2063	0.160	-7.549	0.000	-1.519	-0.893
Total Time Spent on Website	1.1129	0.038	28.947	0.000	1.038	1.188
Lead Origin_Landing Page Submission	-0.5807	0.098	-5.906	0.000	-0.773	-0.388
Lead Origin_Lead Add Form	3.9230	0.190	20.639	0.000	3.550	4.296
Lead Source_Olark Chat	0.7869	0.112	7.002	0.000	0.567	1.007
Specialization_Banking, Investment And Insurance	0.8074	0.186	4.351	0.000	0.444	1.171
Specialization_Business Administration	0.5946	0.168	3.532	0.000	0.265	0.925
Specialization_Finance Management	0.8317	0.123	6.788	0.000	0.592	1.072
Specialization_Human Resource Management	0.6255	0.126	4.971	0.000	0.379	0.872
Specialization_IT Projects Management	0.8022	0.183	4.383	0.000	0.444	1.161
Specialization_Marketing Management	0.6825	0.124	5.510	0.000	0.440	0.925
Specialization_Operations Management	0.7581	0.150	5.051	0.000	0.464	1.052
Specialization_Supply Chain Management	0.6460	0.178	3.637	0.000	0.298	0.994
Last Notable Activity_Modified	-0.6787	0.079	-8.561	0.000	-0.834	-0.523
Last Notable Activity_SMS Sent	1.4115	0.081	17.343	0.000	1.252	1.571

Running RFE with the output number of the variable equal to 15

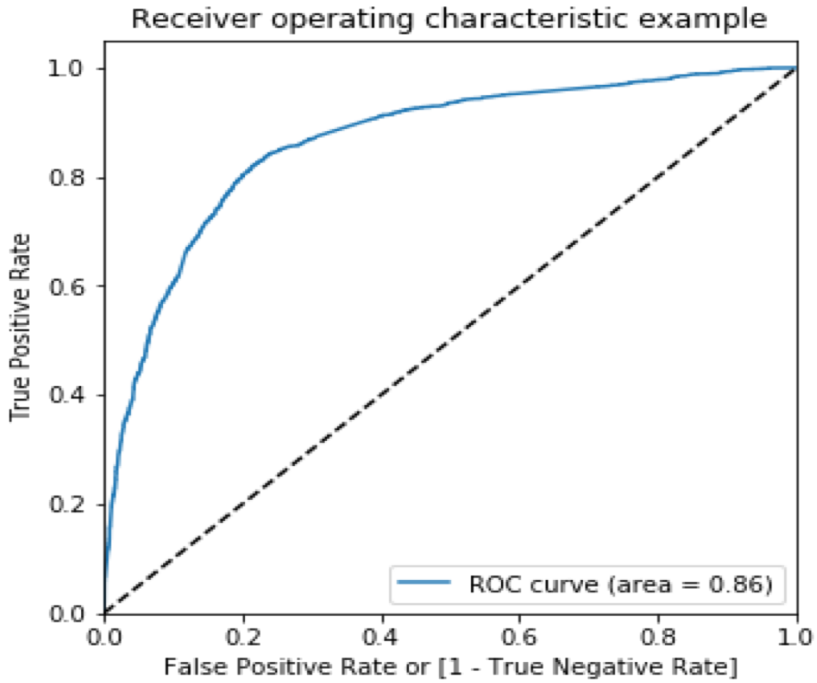
Correlation of the variables



After doing all the data cleaning we did our correlation analysis of the attributes and further understanding the correlation between all the variables using heat map.

We dropped the column 'Lead Origin_API' as it is highly correlated.

ROC Curve



It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

In this case we understood that Scaling was not close to 45 deg hence it was quite OK

Confusion Matrix:

Confusion Matrix:

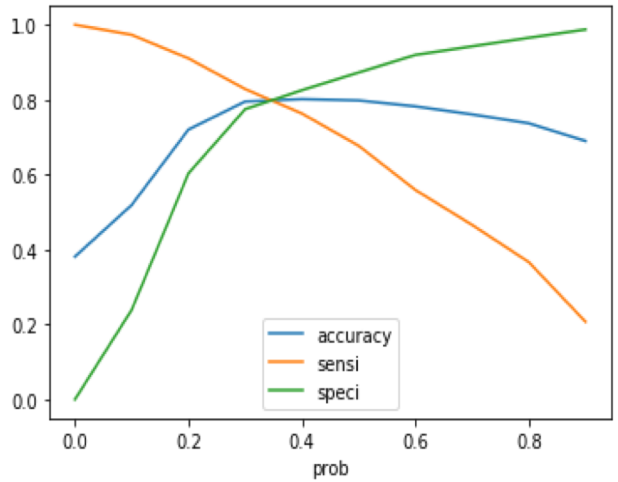
Sensitivity -> 82.8
Specificity -> 77.5
Precision -> 76.6
Recall -> 67.7

Analysing model performance over test data:

Accuracy -> 80.4
Sensitivity -> 74.6
Specificity -> 84.2

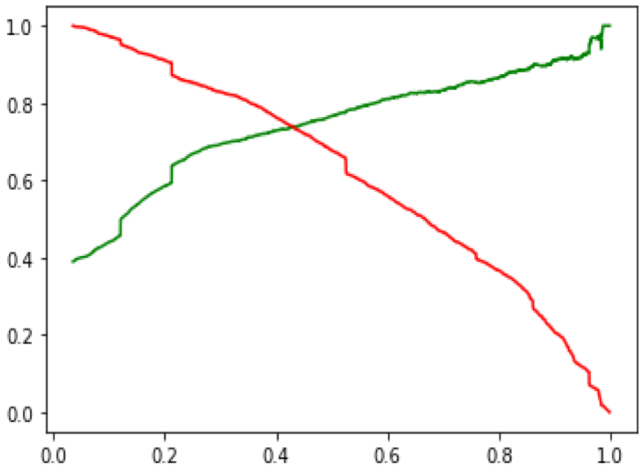
	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Optimal Cut-off Point



Plotting Accuracy, Sensitivity, and specificity Graph
Plots show values of sensitivity and specificity at all of the possible thresholds that could be used to define a positive test result. Typically, sensitivity (true positive rate) is plotted against 1-specificity (false positive rate): each point represents a different threshold in the same group of patients.

The cutoff probability is more than **0.3** which is used for making the final predicted number



Trade-off Graph
The threshold Graph shows that the last prediction probability was OK

Lead Score

Lead Score = 100 * Conversion Probability

```
In [160]: y_pred_final['Lead Score'] = list(map(lambda x: x*100 , y_pred_final['Converted_Prob']))

In [162]: y_pred_final.head(100)
```

Out[162]:

	Converted	Lead Number	Converted_Prob	final_predicted	Lead Score
0	1	4269	0.797669	1	79.766928
1	1	2376	0.962167	1	96.216683
2	1	7766	0.208786	0	20.878609
3	0	9199	0.120240	0	12.024036
4	1	4359	0.924633	1	92.463312
5	1	9186	0.686966	1	68.696576
6	1	1631	0.553645	1	55.364489
7	1	8963	0.174698	0	17.469786
8	0	8007	0.138766	0	13.876586
9	1	5324	0.421566	1	42.156638
10	0	2558	0.414429	0	41.442872

The figure showing Lead Score for top 10 records from the data set.

- The train and test dataset is concatenated to get the entire list of leads available.
- The Conversion Probability is multiplied by 100 to obtain the Lead Score for each lead.
- Higher the lead score, higher is the probability of a lead getting converted and vice versa,
- Since, we had used 0.33 as our final Probability threshold for deciding if a lead will convert or not, any lead with a lead score of 34 or above will have a value of '1' in the final_predicted column.

Thank You