Machine learning-1 > Group Case Study-1

Lead Scoring Case Study - Assignment

Riddhi Agarwal
Tushar Sharma
PGDDS-C12
IIIT-B & UpGrad

INDEX

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 company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not.

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%.

What Needs To Be Done

The problem statement comprises of an online education agency who is gathering the Data of leads by collecting the online information they fill while visiting the site or clicking the ads on google/any other site like no. of clicks, time spent on the web-page, age, gender etc.

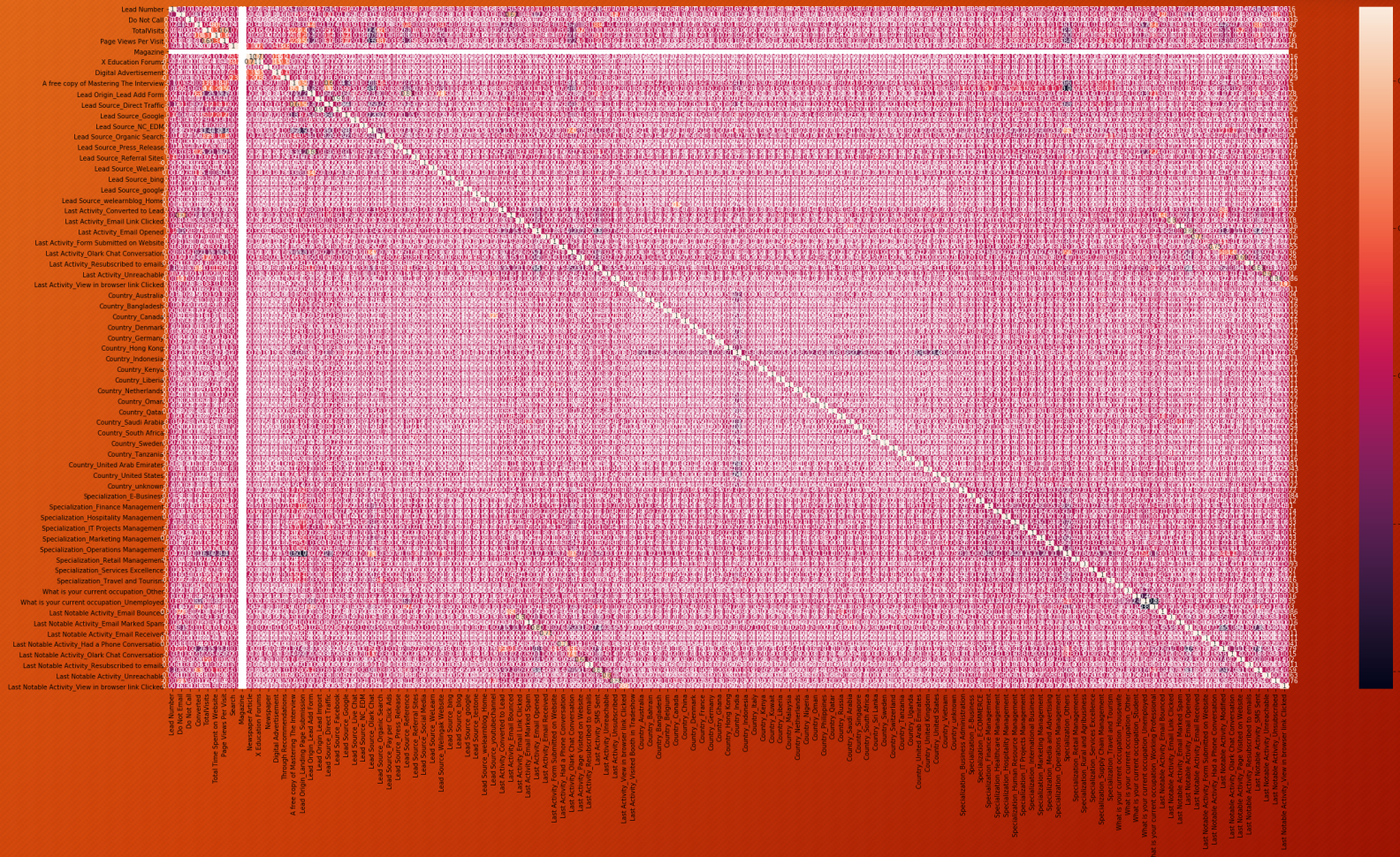
Analyzing the same to classify the leads based on their seriousness and interest. The genuine leads shall be contacted and will be tried to convert them into a potential prospect

We need to apply Logistic Regression Model on our Dataset to achieve the desired result by finding out the conversion probability using the lead score and create a pool of potential candidates who will be contacted for conversion with 80% chance.

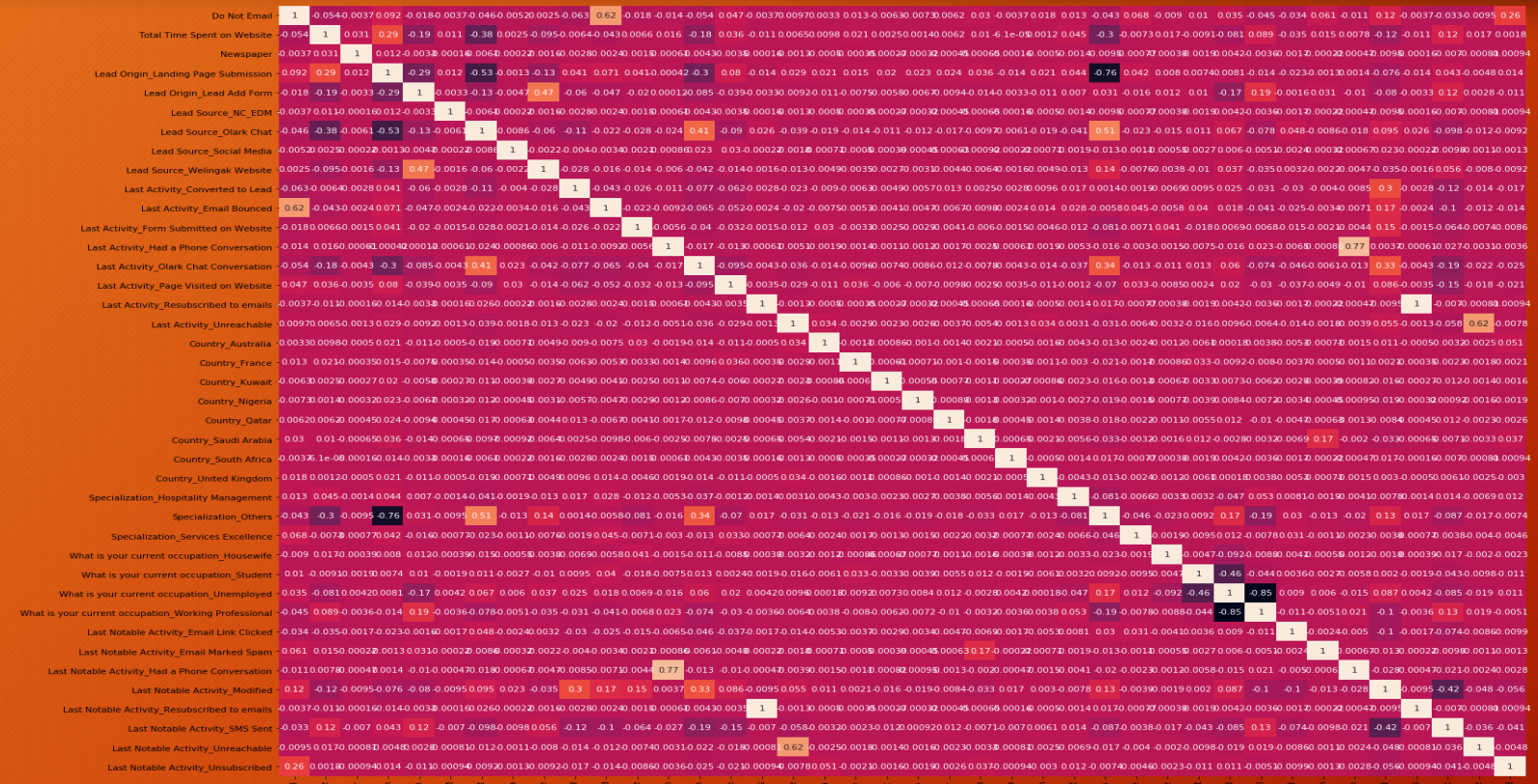
Analysis Approach

Data Cleaning: The first step was to check (using .head(), .describe(), .info(), .shape)and clean the Data by dropping the irrelevant variables, checking the null values and replacing them if required.
Data Preparation: Mapping the YES/NO to 1's & 0's , Creating the Dummy variables for the applicable columns, checking for outliers using Quantiles function.
Train-test Split: The data split in Train and Test Set (70:30 ratio) to perform the regression.
Feature Scaling: Feature Scaling is performed on the Dataset.
Correlation Matrix & RFE: Correlation matrix is plotted to see the correlation among the variables but due to high number of variables RFE is applied then correlation was checked again.
Linear Logistic Regression Model: Regression model is applied on the Train Set and after 7 iterations we got the desired values and accuracy.
Confusion Matrix: To check the overall accuracy confusion matrix technique is applied and precision score was checked which comes out to be .78 and accuracy score was .81. Therefore, lead conversion rate is 78% and model's overall accuracy is 80%
ROC Curve: To find out the optimum cut-off ROC curve was plotted and it showed 60 is the optimum lead score at which lead conversion rate is more than 80%.

Correlation Matrix Before RFE



Correlation Matrix After RFE



RFE

In the previous 2 Slides the correlation matrix is shown before RFE and after RFE.

Before RFE – The correlation matrix is very dense and its not possible to find out the correlation between the variables die to large number of columns. To get rid of this problem RFE is applied.

After RFE – The correlation matrix after RFE is much comprehensible and can be further processed for Logistic Regression

Final Regression Model Result

The final independent variables for the prediction of dependent variable i.e. Lead Score is shown on the right side .
All the P-Values are under the limit and further elimination is not required.

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.2094	0.060	-20.042	0.000	-1.328	-1.091
Do Not Email	-1.3399	0.214	-6.249	0.000	-1.760	-0.920
Total Time Spent on Website	1.1176	0.040	27.979	0.000	1.039	1.196
Lead Origin_Lead Add Form	3.8600	0.220	17.556	0.000	3.429	4.291
Lead Source_Olark Chat	1.1374	0.102	11.167	0.000	0.938	1.337
Lead Source_Welingak Website	2.7830	1.034	2.692	0.007	0.757	4.809
Last Activity_Converted to Lead	-0.9405	0.212	-4.426	0.000	-1.357	-0.524
Last Activity_Email Bounced	-0.7486	0.392	-1.912	0.056	-1.516	0.019
Last Activity_Olark Chat Conversation	-1.2871	0.163	-7.875	0.000	-1.607	-0.967
Last Activity_Page Visited on Website	-0.4690	0.146	-3.211	0.001	-0.755	-0.183
Last Activity_Unreachable	-1.3473	0.650	-2.072	0.038	-2.622	-0.073
What is your current occupation_Working Professional	2.7229	0.183	14.879	0.000	2.364	3.082
Last Notable Activity_Had a Phone Conversation	3.3463	1.107	3.022	0.003	1.176	5.517
Last Notable Activity_Modified	-0.3092	0.091	-3.415	0.001	-0.487	-0.132
Last Notable Activity_SMS Sent	1.3435	0.086	15.594	0.000	1.175	1.512
Last Notable Activity_Unreachable	3.0769	0.830	3.709	0.000	1.451	4.703
Last Notable Activity_Unsubscribed	1.1946	0.488	2.448	0.014	0.238	2.151

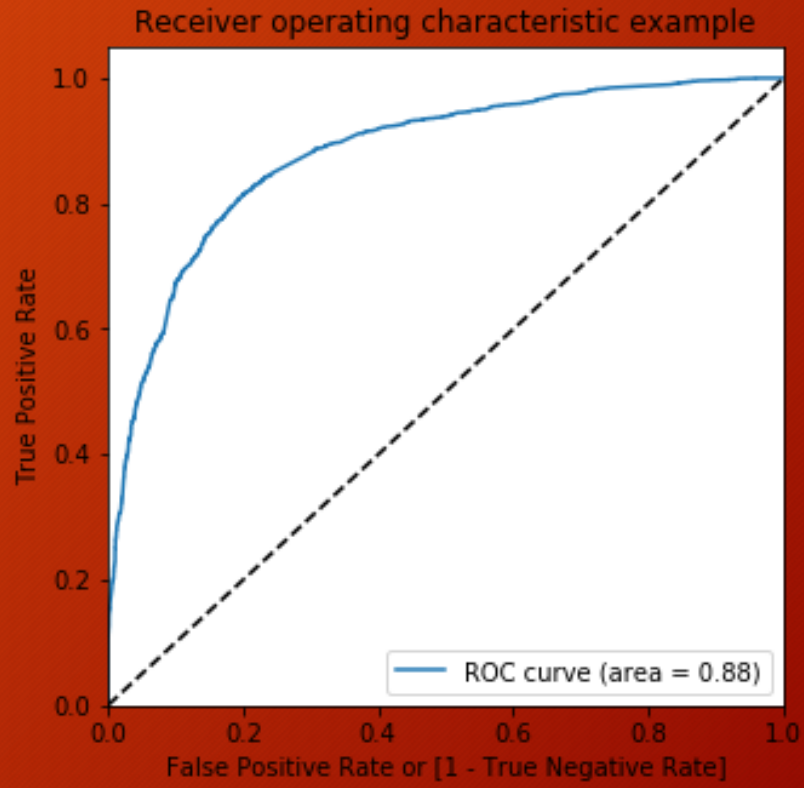
VIF Score Of Final Regression Model Result

Similarly , The final independent variables' VIF Score is less than 3 in this case , which is highly favorable for the analysis.

	Features	VIF
0	Do Not Email	2.06
12	Last Notable Activity_Modified	1.95
6	Last Activity_Email Bounced	1.89
9	Last Activity_Unreachable	1.68
14	Last Notable Activity_Unreachable	1.65
7	Last Activity_Olark Chat Conversation	1.61
3	Lead Source_Olark Chat	1.59
2	Lead Origin_Lead Add Form	1.55
4	Lead Source_Welingak Website	1.32
1	Total Time Spent on Website	1.27
5	Last Activity_Converted to Lead	1.26
13	Last Notable Activity_SMS Sent	1.21
10	What is your current occupation_Working Profes...	1.16
15	Last Notable Activity_Unsubscribed	1.15
8	Last Activity_Page Visited on Website	1.12
11	Last Notable Activity_Had a Phone Conversation	1.00

ROC Curve

Plotted ROC curve to find the optimum value of lead score at which the accuracy of the model will be best.



Result Analysis

The aim of the case study is to make lead conversion rate 80%, means out of total converted leads predicted by the model, actual converted leads should be more than 80%.
So the precision of the model should be 80% and above.
For this, calculated the precision score at various lead score and for lead score 60 and above precision is coming out to be 83%.
So, leads having lead score more than 60 are the hot leads and they should be focused to convert them into customers.