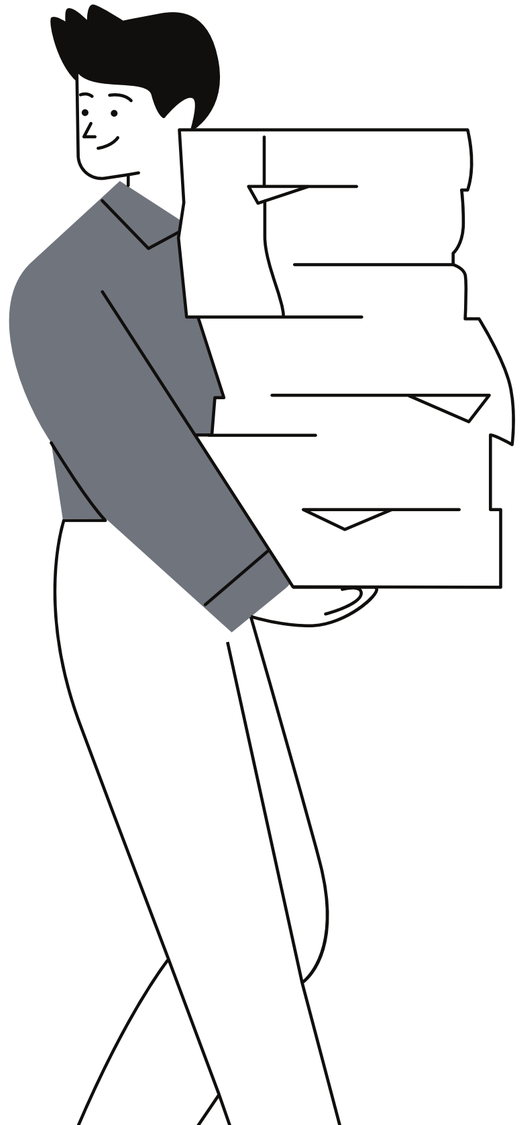
Lead Score Case Study



Vybhava | Sanjana | Shubham

Problem Statement

X Education wants help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires to build a machine learning model, wherein a lead score needs to be assigned to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a 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%.

Business Objective

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Methodology

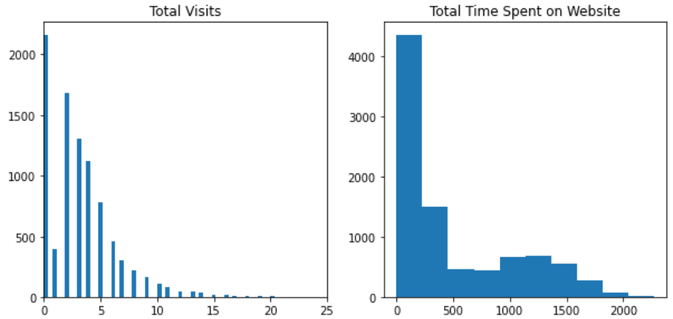
- Data Cleaning
- Handling outliner and missing values
- EDA:- Univaraiate analysis, Bivariate Analysis, Corelation Matrix and Data Visualisation
- Data Preparation
- Feature Scaling
- Model Building :- Feature selection using RFE
- Determine the model's features and train it

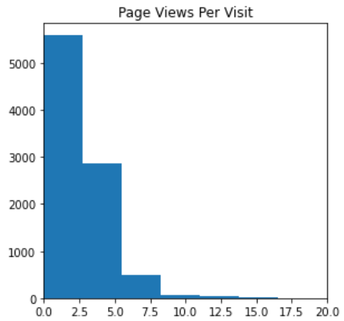
Explaing the Data

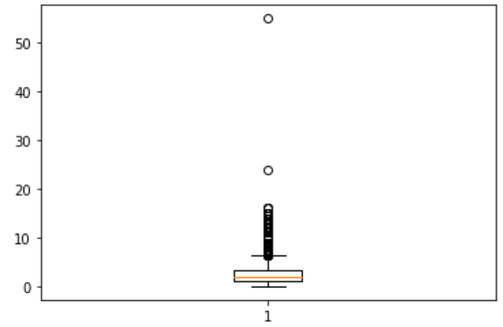
- Leads.csv has 9240 rows and 37 columns.
- The data types of the columns are float64(4), int64(3), object(30)
- The target variable is '**converted**'
- Dropped the variables which have more than 35% of missing data:
- Dropping the '**country**' column as the data is biased towards india values, '**TotalVisits**' column has outliers hence dropping the outlier values as the number of rows with those values are very less.
- The listed colums were imputed with mode values to handle missing values:
    - 'Specialization', 'What is your current occupation', 'What matters most to you in choosing a course', 'Tags'

Outlier analysis

	TotalVisits	Total Time Spent on Website	Page Views Per Visit
count	9071.000000	9071.000000	9071.000000
mean	3.401279	482.937934	2.369127
std	3.660471	545.311359	2.159857
min	0.000000	0.000000	0.000000
25%	1.000000	11.000000	1.000000
50%	3.000000	246.000000	2.000000
75%	5.000000	923.500000	3.185000
90%	7.000000	1373.000000	5.000000
95%	10.000000	1557.000000	6.000000
99%	17.000000	1839.000000	9.000000
max	74.000000	2272.000000	55.000000

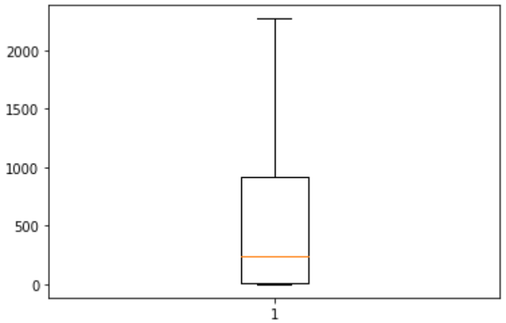






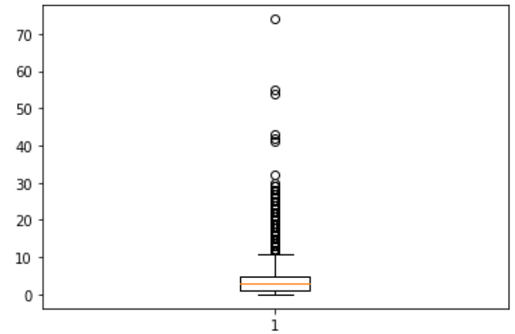
'Page per visit'

Most of the points lie within the 0 to 10 range with few outliers in the range between 10 to 20. We see one outlier point in 55 values



'Total time spent on website'

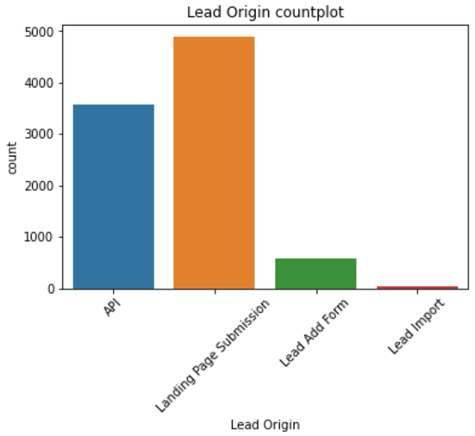
The mean of the boxplot is around 250 and the quantile range is within 100 to 1000 Very few outliers might be present arount 2000 value

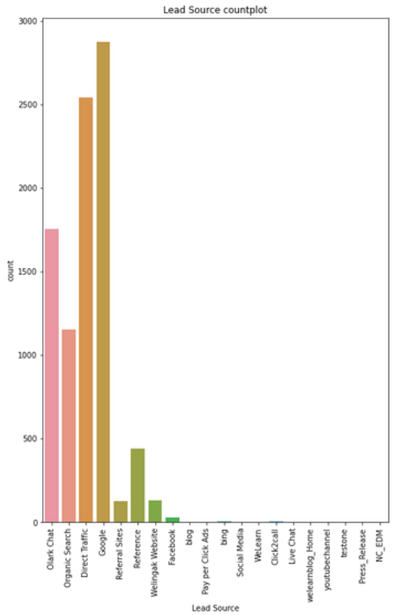


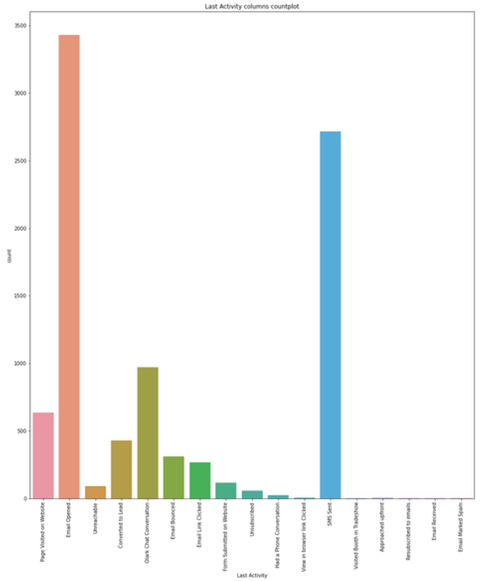
'TotalVisits'

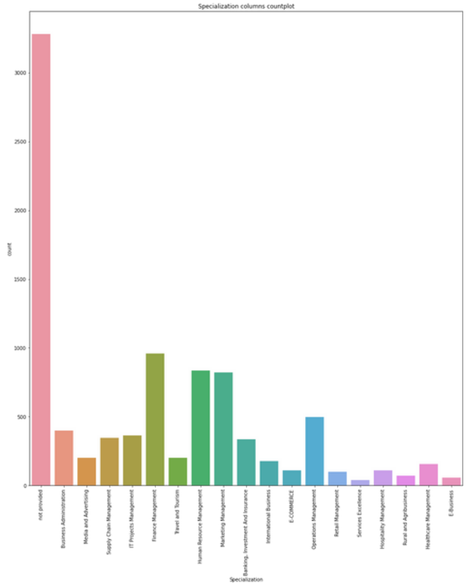
The column had few points in more than 100 values , dropped them as there was many outlier points in the column, the mean of box plot is aournd 5 , the box plot extends from 0 to 10 and outliers extending above 70

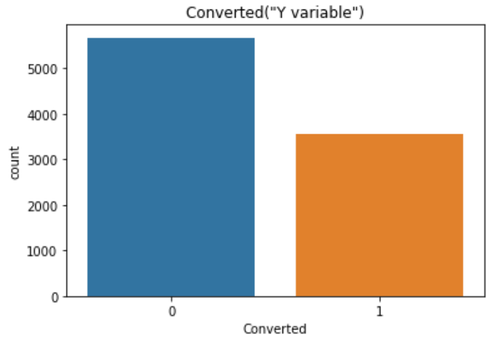
Categorical Varibales - Univariate analysis











'Lead Origin countplot'

We can observe that 'Landing Page Submission' has the highest count in this category

'Lead Source countplot'

In this case, Google is the highest followed by Direct Traffic and Olark Chat

'Last Activity'

Email opened and SMS sent has the highest count

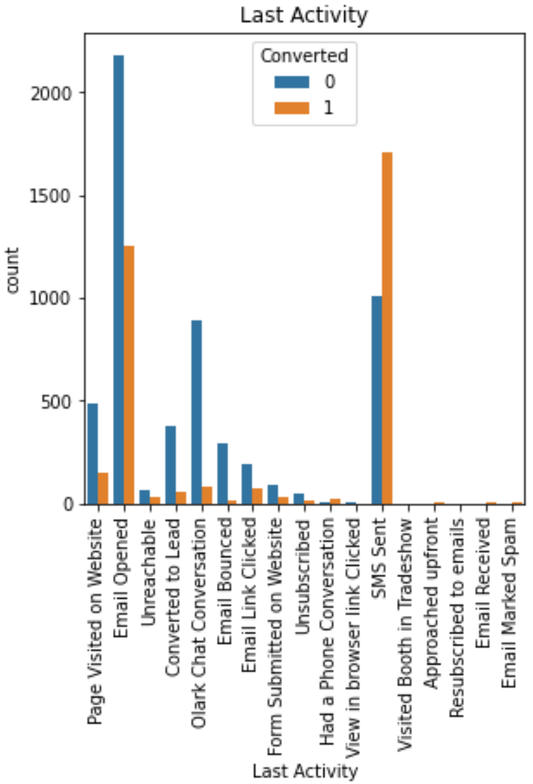
'Specialization'

In this case Finance Management is the highest followed right after operations management. *note that 'not provided' has null values.*

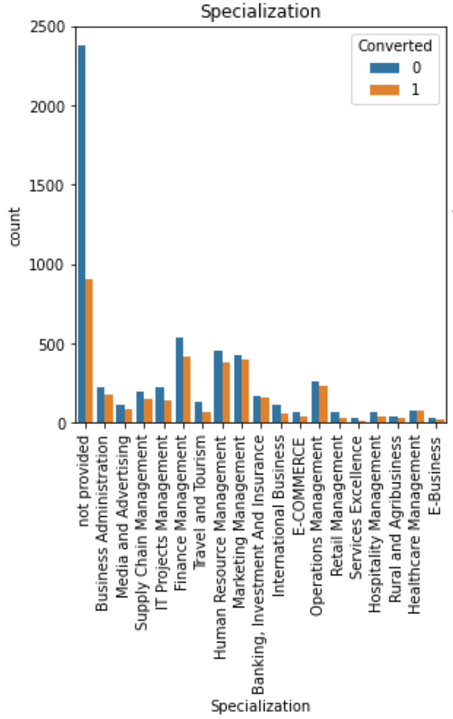
'Converted'

'Converted' values is our target variable, and as the result shows that the non-coverted values overpower the converted values.

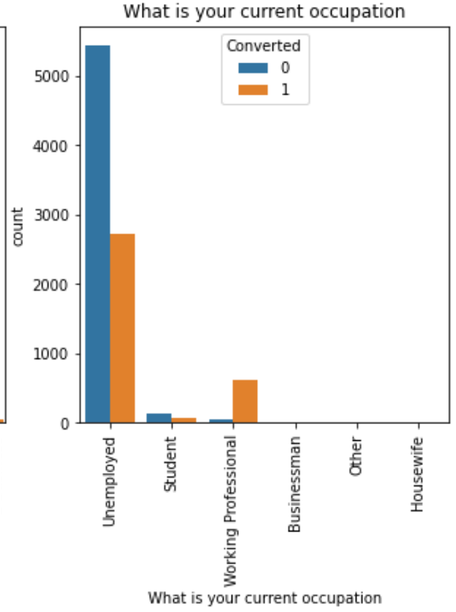
Relating categorical variables to Target Variable



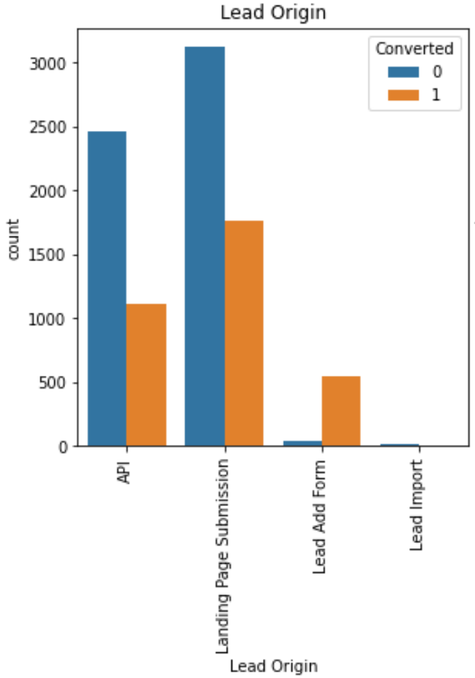
Last Activity : In this graph we see people who were targeted via SMS has converted more, and rought 50% or more customers has been converted via email opened category



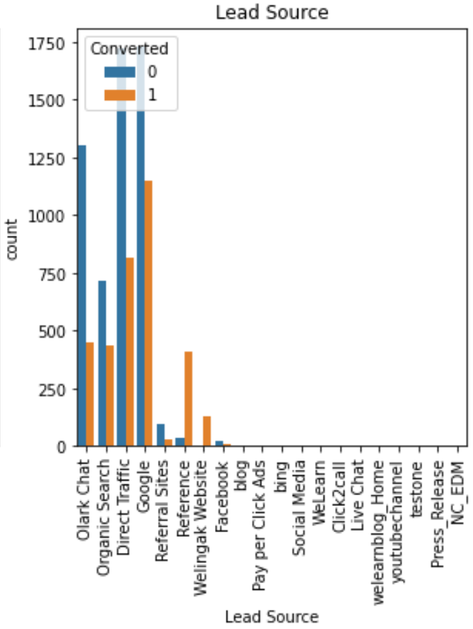
Specialization: Here we obersver people who opted for finance managment are the most converted leads. *The null values are indicated using 'not provided'.*



What is your current occupation: Here we observe that people who are unemployed are the most interested followed by working professional



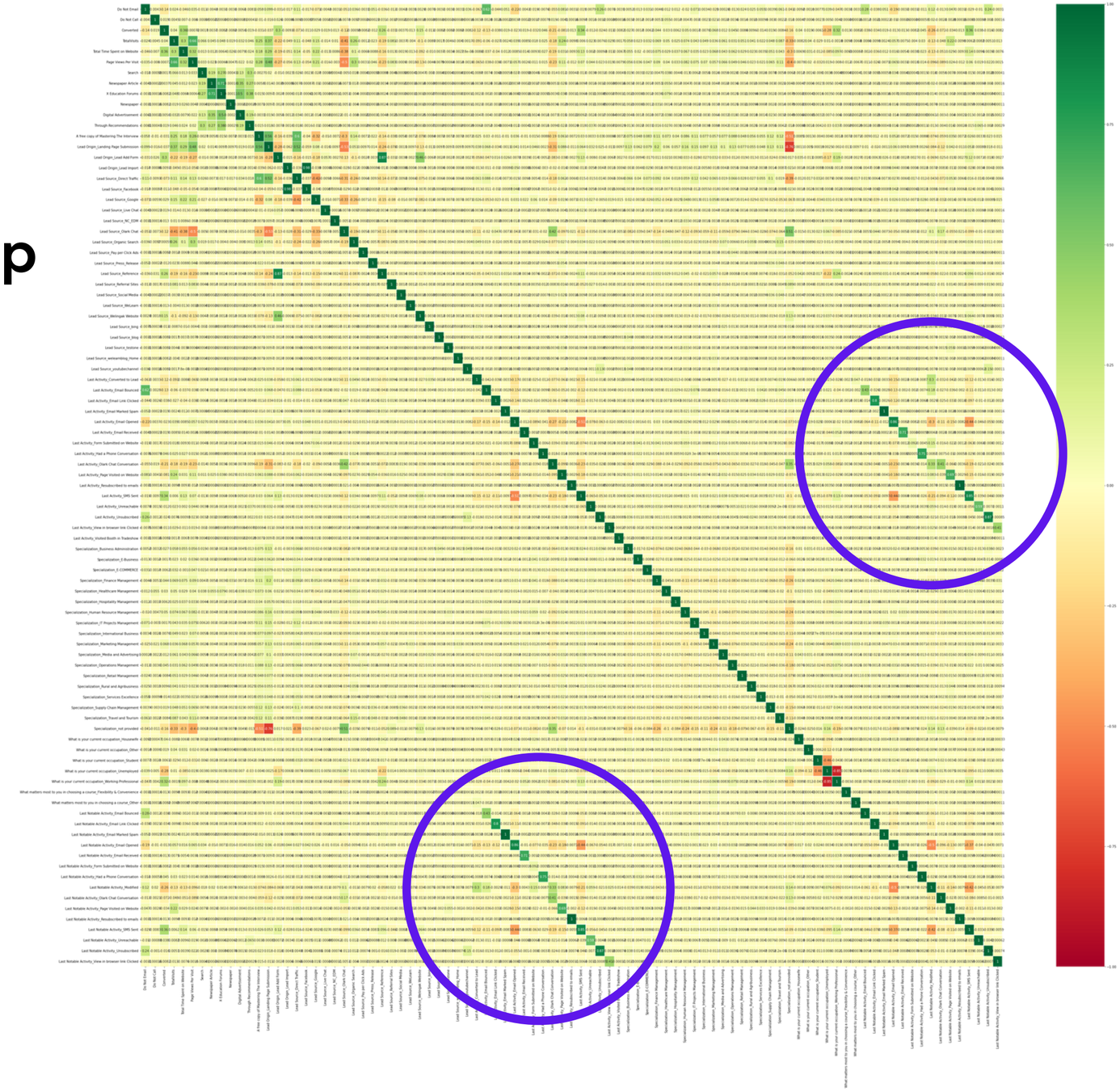
Lead Origin: In this case, *'API'* and *'Landing page submission'* category the conversion rate is higher.



Lead Source: And in this case, the conersion rate from *'google'*, *'direct traffic'* and *'olark chat'* is the highest.

Corelation Graph - Heat Map

- Majority of the number of variables which have higher correlation values. Hence, we are dropping the variables after RFE selection.
- We observe two clusters of higher corrected variables in the graph.
- The variables are positively as well as negatively correlated.



Model Building

Firstly, RFE was done to attain 20 variables. Later the insignificant variables were removed based on p-values and VIF values. The variables with p-values <0.05 and VIF<5 were retained in the model.

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6349
Model:	GLM	**Df Residuals:**	6330
Model Family:	Binomial	**Df Model:**	18
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2603.3
Date:	Tue, 24 Jan 2023	**Deviance:**	5206.6
Time:	10:22:32	**Pearson chi2:**	6.52e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.7581	0.134	5.652	0.000	0.495	1.021
Do Not Email	-1.2656	0.199	-6.363	0.000	-1.655	-0.876
Total Time Spent on Website	1.0868	0.040	26.914	0.000	1.008	1.166
Lead Origin_Landing Page Submission	-1.0147	0.132	-7.661	0.000	-1.274	-0.755
Lead Origin_Lead Add Form	2.1724	0.231	9.414	0.000	1.720	2.625
Lead Source_Direct Traffic	-1.2381	0.148	-8.365	0.000	-1.528	-0.948
Lead Source_Google	-1.0129	0.126	-8.032	0.000	-1.260	-0.766
Lead Source_Organic Search	-1.1982	0.148	-8.095	0.000	-1.488	-0.908
Lead Source_Referral Sites	-1.3867	0.345	-4.025	0.000	-2.062	-0.711
Lead Source_Welingak Website	2.4744	0.758	3.264	0.001	0.989	3.960
Last Activity_Converted to Lead	-1.1412	0.213	-5.369	0.000	-1.558	-0.725
Last Activity_Email Bounced	-1.7371	0.635	-2.735	0.006	-2.982	-0.492
Last Activity_Olark Chat Conversation	-1.5428	0.171	-9.038	0.000	-1.877	-1.208
Specialization_not provided	-1.1463	0.123	-9.311	0.000	-1.388	-0.905
What is your current occupation_Working Professional	2.7095	0.197	13.774	0.000	2.324	3.095
Last Notable Activity_Email Bounced	2.3095	0.787	2.934	0.003	0.767	3.852
Last Notable Activity_Had a Phone Conversation	2.9265	1.126	2.600	0.009	0.720	5.133
Last Notable Activity_SMS Sent	1.5810	0.081	19.464	0.000	1.422	1.740
Last Notable Activity_Unreachable	1.4619	0.495	2.954	0.003	0.492	2.432

	Features	VIF
0	const	15.83
3	Lead Origin_Landing Page Submission	3.69
5	Lead Source_Direct Traffic	3.65
13	Specialization_not provided	2.95
6	Lead Source_Google	2.89
7	Lead Source_Organic Search	1.99
11	Last Activity_Email Bounced	1.89
4	Lead Origin_Lead Add Form	1.87
1	Do Not Email	1.65
2	Total Time Spent on Website	1.33
9	Lead Source_Welingak Website	1.32
12	Last Activity_Olark Chat Conversation	1.29
15	Last Notable Activity_Email Bounced	1.24
14	What is your current occupation_Working Profes...	1.15
17	Last Notable Activity_SMS Sent	1.13
8	Lead Source_Referral Sites	1.08
10	Last Activity_Converted to Lead	1.06
18	Last Notable Activity_Unreachable	1.01
16	Last Notable Activity_Had a Phone Conversation	1.00

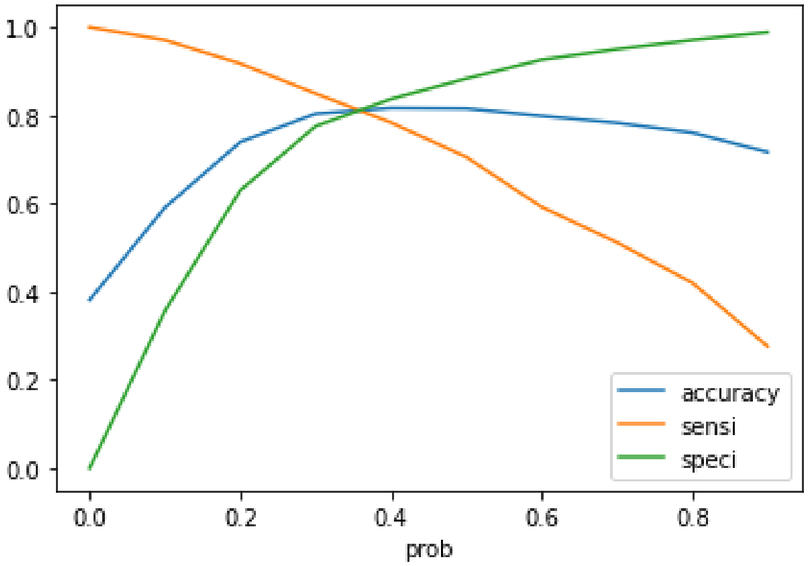
Model Evaluation

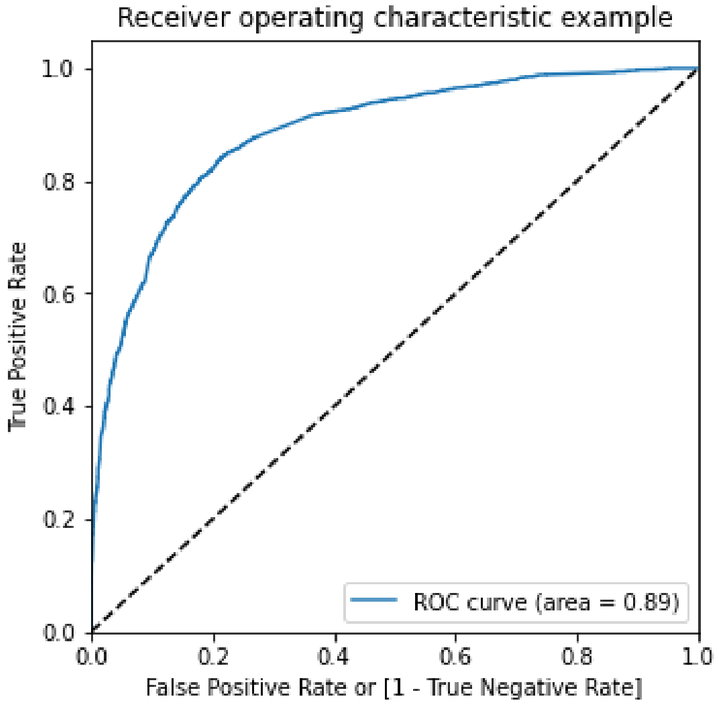
A confusion matrix was made. Later the optimum cut-off value (based on the ROC curve) was used to find the Accuracy, sensitivity, and specificity which came to around 80% for each of the metrics.

Confusion matrix of Train Data

Converted/not converted	converted	**not converted**
converted	3470	457
not converted	713	1709

Accuracy	80.94%
Sensitivity	81.79%
Specificity	80.41%





Perdiction

The prediction was done on the test data frame with an optimum cut-off of 3.5 with accuracy, sensitivity, and specificity of about 80%.

Confusion matrix of Test Data

Converted/not converted	converted	not converted
converted	1373	337
not converted	188	824

Performance results

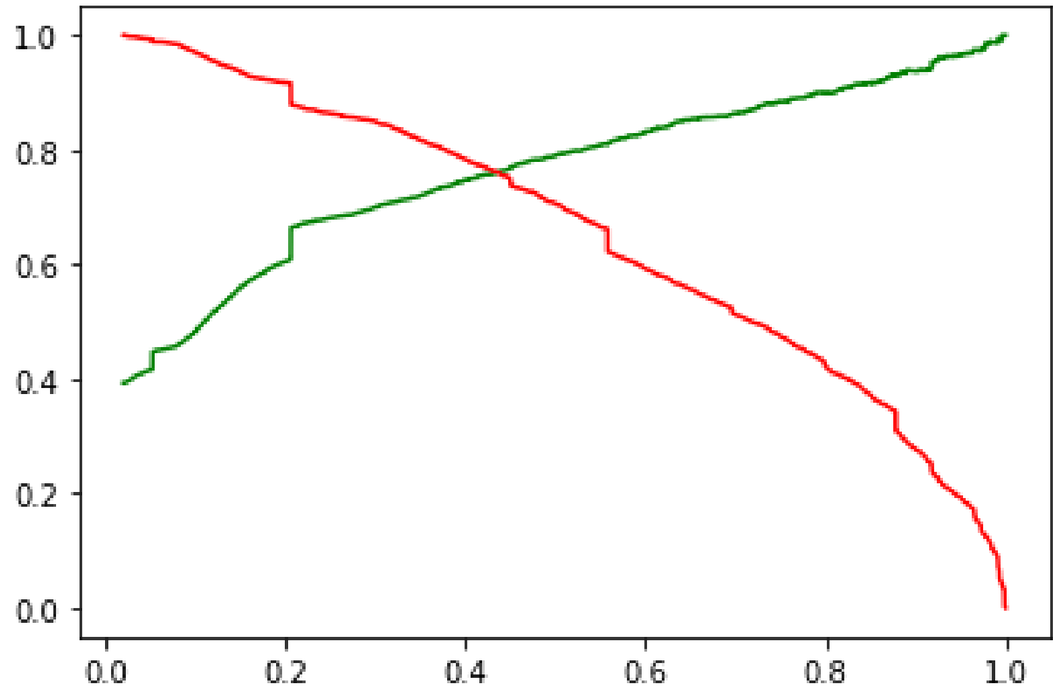
Accuracy	80.71%
Sensitivity	81.42%
Specificity	80.29%

Predicted Dataframe

	Lead ID	Converted	Converted_probability	final_predicted	Lead Score	Visitor_type
0	4703	1	0.383500	1	38.0	Normal Visitor
1	5544	0	0.603716	1	60.0	Hot Lead
2	5520	0	0.088284	0	9.0	Normal Visitor
3	1342	1	0.028975	0	3.0	Normal Visitor
4	4101	0	0.001984	0	0.0	Normal Visitor

Precision-Recall

This method was used to recheck the model performance. A cut-off of 4.1 with a Precision of around 78% and a Recall of around 70% was found.



precision - recall cut-off curve

Results

It was found that the following are the important features that help in predicting the most promising leads which X Education can focus on to bring in more customers to buy their courses.

- Total time spent on a website.
- The lead source was: Google, Direct traffic, Organic search, Welingak website
- The last activity was: SMS, Olarck chat conversation
- Their current occupation being a working professional and even unemployed people

Learning

- Understood the practical usage of the logistic regression model
- Practiced the implementation of the logistic regression model using python
- Understood the variable performance metrics calculation using confusion matrix
- Practiced usage of the GIT and GitHub

Thank you