Lead score case study

Problem Statement

- An education company named X Education sells online courses to industry professionals.
- 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.
- X Education gets a lot of leads, its lead conversion rate is very poor.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.

Problem Statement - cont

As you can see, there are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

Solution:

- build a model wherein you need to assign a lead score 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.
- We will use Logistic regression model

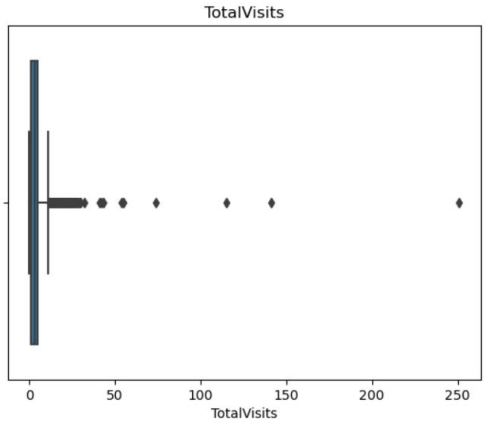
Goals of the Case Study

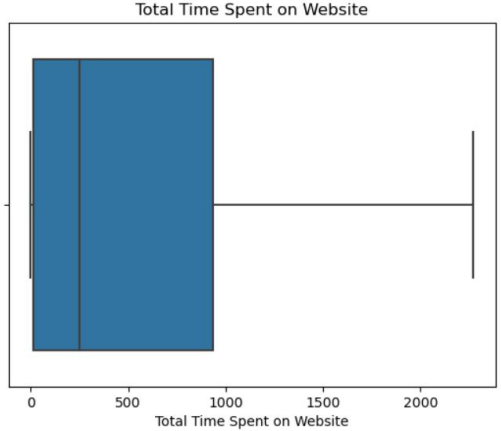
- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
- There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

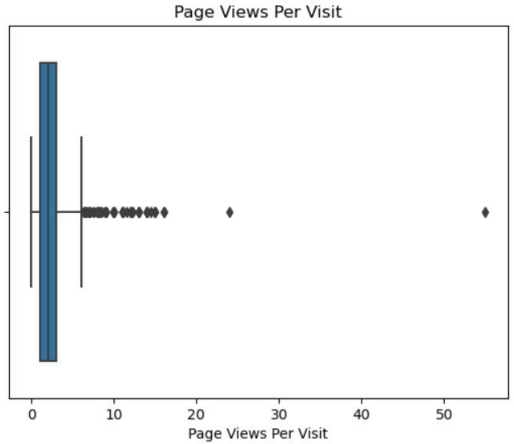
Data understanding, cleaning

- In the first step of data cleaning we will replace value select by np.NaN
- There are 10 columns which has more than 30 percent missing data, we are removing those columns
- There are some columns with 20 to 30 % missing information, for those we are going to replace null values with not provided.
- There are some columns with less than 4 percent missing values and we imputed those with mode or mean.
- Also minor fix was done to some column values.

EDA continues varaible

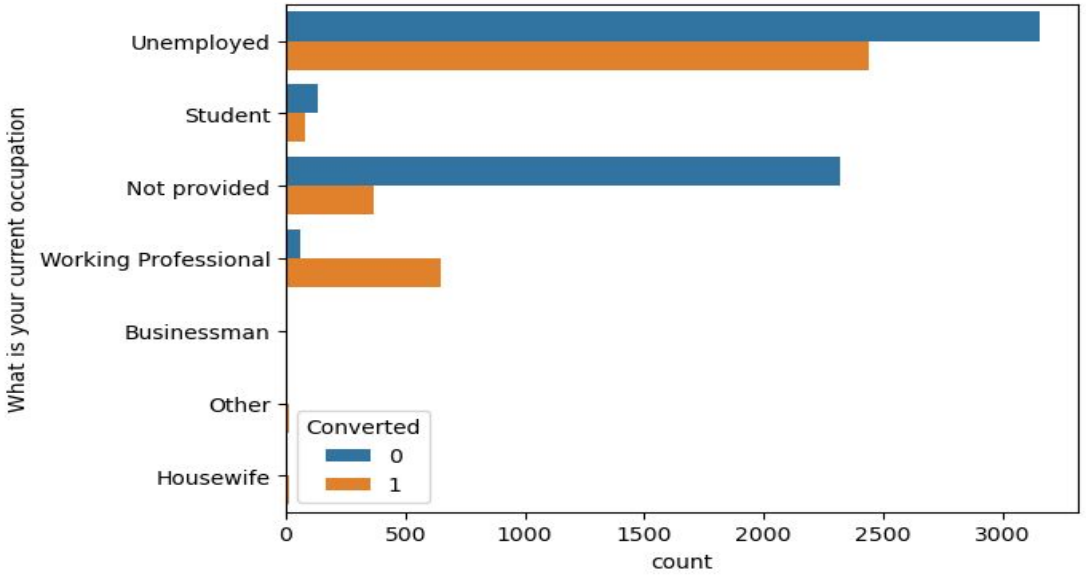


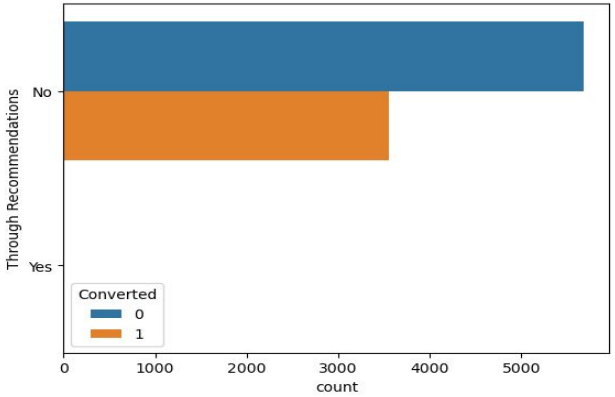


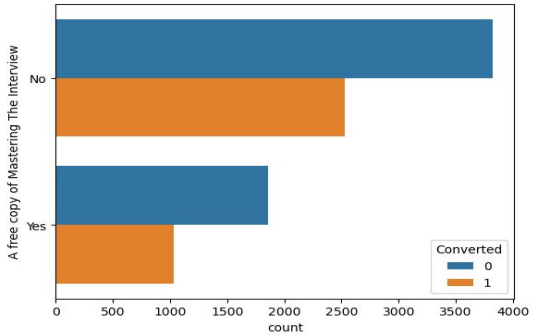


There are some outliers but we will deal those while scalling

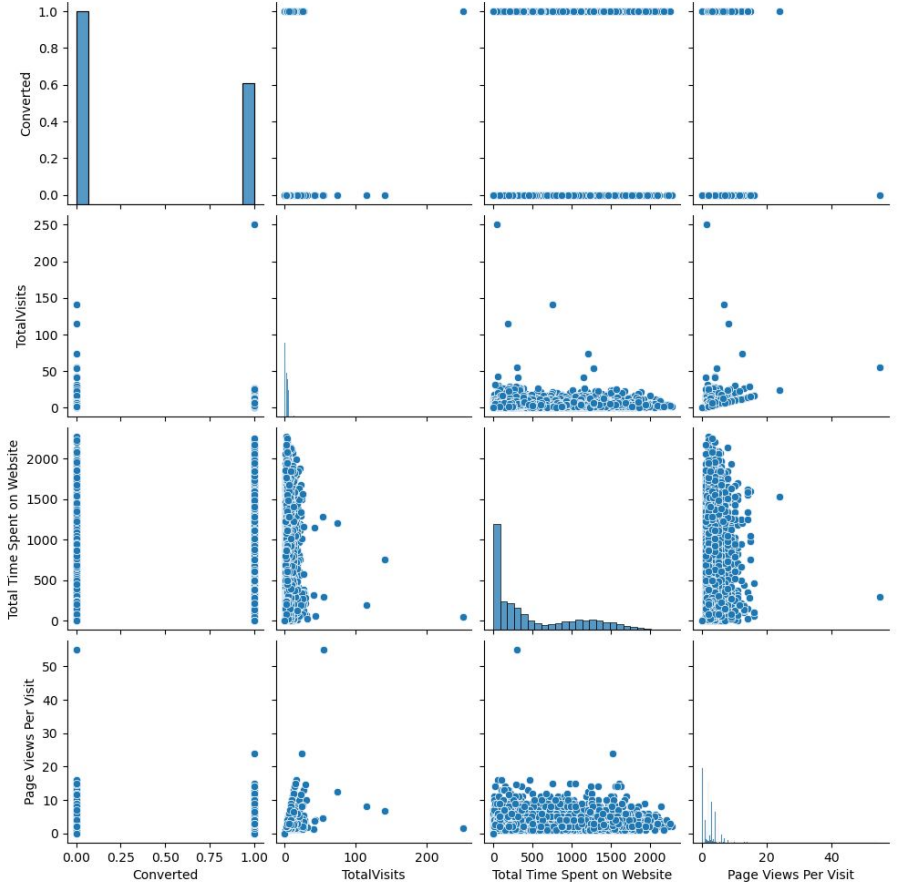
EDA categotrical variable



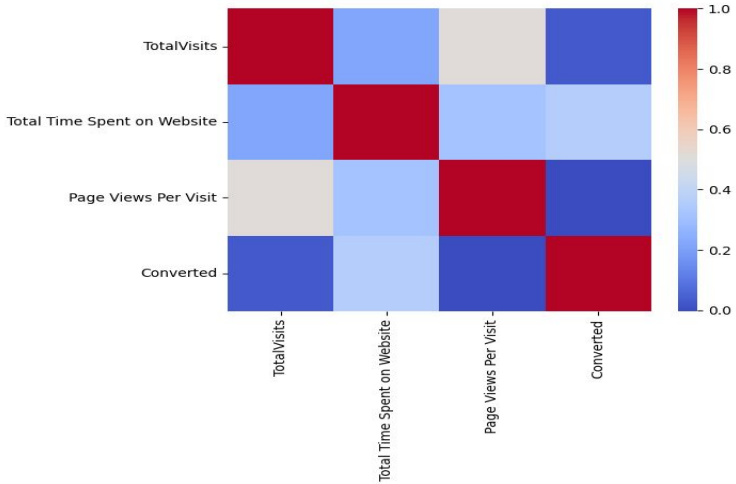


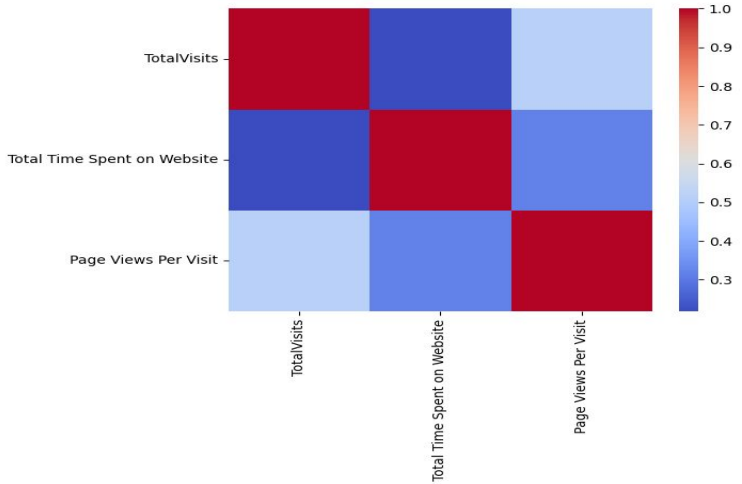


EDA pair plot



EDA heatmap





Since we didn't covert more categorical variables to 0,1 we cant able to see correlation, but we will use VIF to eliminate variables with high correlation.

Logistic Regression Model

After performing RFE, we got 20 variables and after checking VIF, P value, we got a final model.

Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6454
Model Family:	Binomial	**Df Model:**	13
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2665.4
Date:	Sun, 18 Feb 2024	**Deviance:**	5330.8
Time:	22:17:53	**Pearson chi2:**	6.68e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.3989
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.2794	0.083	-27.364	0.000	-2.443	-2.116
TotalVisits	7.6024	2.016	3.771	0.000	3.651	11.554
Total Time Spent on Website	4.5442	0.166	27.411	0.000	4.219	4.869
Lead Origin_Lead Add Form	2.1464	0.202	10.650	0.000	1.751	2.541
Lead Origin_Lead Import	-1.5943	0.461	-3.458	0.001	-2.498	-0.691
Lead Source_Welingak Website	2.7066	1.027	2.635	0.008	0.694	4.720
Do Not Email_Yes	-1.4951	0.165	-9.061	0.000	-1.819	-1.172
Last Activity_Olark Chat Conversation	-1.2487	0.163	-7.660	0.000	-1.568	-0.929
Last Activity_SMS Sent	1.3259	0.074	18.014	0.000	1.182	1.470
Country_Not provided	1.4897	0.111	13.394	0.000	1.272	1.708
What is your current occupation_Working Professional	2.4809	0.185	13.399	0.000	2.118	2.844
What matters most to you in choosing a course_Not provided	-1.2059	0.086	-13.966	0.000	-1.375	-1.037
Last Notable Activity_Had a Phone Conversation	3.5666	1.124	3.173	0.002	1.364	5.770
Last Notable Activity_Unreachable	2.3589	0.699	3.377	0.001	0.990	3.728

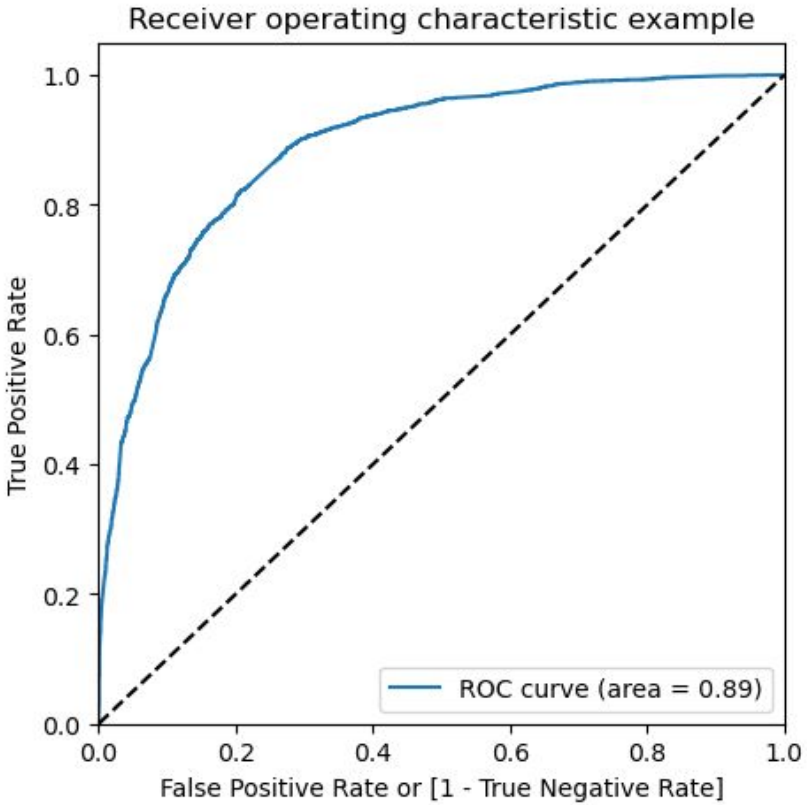
Model evaluation

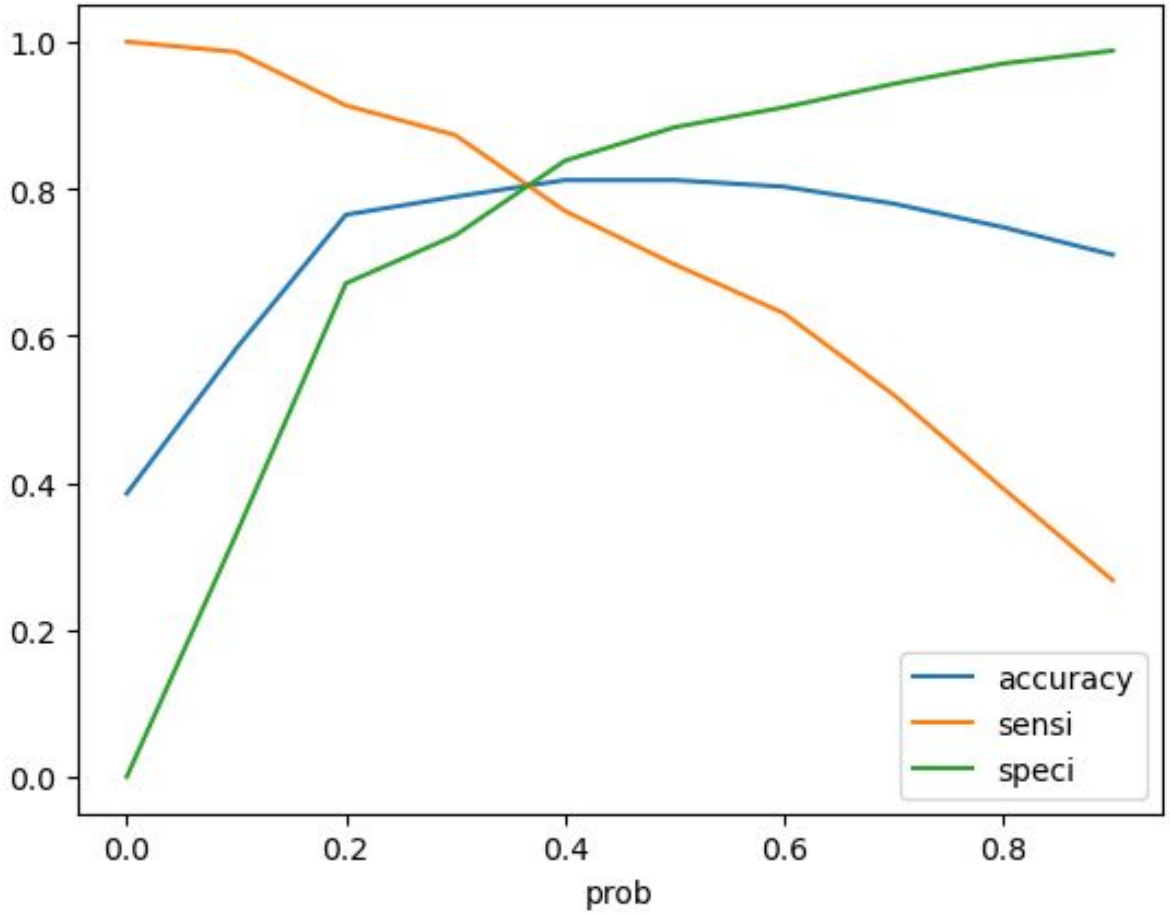
We kept 0.5 as cut off and this is the result i got.

- Accuracy score 80%
- Sensitivity as 69.7%
- Specificity as 88%

Now, we will use ROC curve to know the best cutoff value.

ROC curve


Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.89)

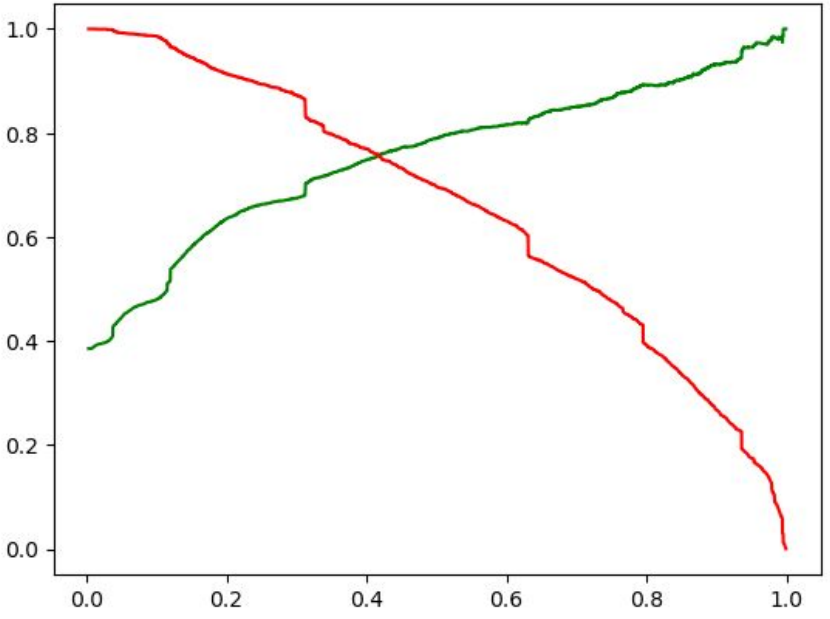


From the above graph we can conclude 0.37 is the ideal cutoff

Model evaluation with cutoff 0.37

Now with 0.37 as cut off and this is the result i got.

- Accuracy score 80%
- Sensitivity as 78%
- Specificity as 81%
- Precision as 73%
- Recall as 78.6%



Model evaluation for test data

The result we got from test data

- Accuracy score 80%
- Sensitivity as 78%
- Specificity as 81%
- Precision as 73%
- Recall as 78%

Summary

below are the list of variables which are important for our model:

- TotalVisits
- Total Time Spent on Website
- Lead Origin_Lead Add Form
- Lead Origin_Lead Import
- Lead Source_Welingak Website
- Do Not Email_Yes
- Last Activity_Olark Chat Conversation
- Last Activity_SMS Sent
- Country_Not provided
- What is your current occupation_Working Professional
- What matters most to you in choosing a course_Not provided
- Last Notable Activity_Had a Phone Conversation
- Last Notable Activity_Unreachable

Ideal cutoff : 0.37