Lead Score Report

Insights & Analysis

Sagarika, Jude and Rama Krishna Kohli

Problem Statement

- X Education have been provided with a leads dataset from the past with around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or 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 a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

Model Design

Linear Regression Model

- Using generalised linear regression model, we have selected some variables which is going to help in predicting lead conversion ‘0’ or 1.
- The variables are selection through RFE feature elimination given by the model itself.
- The variables are further narrowed down on the basis of p value and VIF values.

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6451
Model Family:	Binomial	**Df Model:**	16
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2609.6
Date:	Sat, 14 Oct 2023	**Deviance:**	5219.2
Time:	15:26:13	**Pearson chi2:**	8.12e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.7911	0.149	-5.302	0.000	-1.084	-0.499
Do Not Email	-1.1811	0.182	-6.492	0.000	-1.538	-0.824
Total Time Spent on Website	1.0651	0.040	26.711	0.000	0.987	1.143
Lead Origin_Landing Page Submission	-1.0227	0.128	-7.972	0.000	-1.274	-0.771
Lead Origin_Lead Add Form	2.8029	0.203	13.794	0.000	2.405	3.201
Lead Source_Olark Chat	1.0993	0.123	8.940	0.000	0.858	1.340
Lead Source_Welingak Website	2.4629	0.750	3.285	0.001	0.993	3.932
Occupation_Unknown	-1.0818	0.088	-12.357	0.000	-1.253	-0.910
Occupation_Working Professional	2.3966	0.190	12.627	0.000	2.025	2.769
Last Activity_Email Opened	0.7288	0.110	6.636	0.000	0.514	0.944
Last Activity_Olark Chat Conversation	-0.6068	0.191	-3.169	0.002	-0.982	-0.231
Last Activity_Other Activity	2.2419	0.488	4.592	0.000	1.285	3.199
Last Activity_SMS Sent	1.8672	0.111	16.782	0.000	1.649	2.085
Last Activity_Unreachable	0.8487	0.368	2.303	0.021	0.126	1.571
Last Activity_Unsubscribed	1.3906	0.485	2.865	0.004	0.439	2.342
Specialization_Hospitality Management	-0.9951	0.327	-3.040	0.002	-1.637	-0.353
Specialization_Others	-0.9785	0.123	-7.927	0.000	-1.220	-0.737

Model Design

Logistic Regression Model

- Based on variables selected by Linear Regression Model, we have run the train data set through Logistic Regression Algorithm.
- Y values (Lead converted or not) is predicted using the model generated by Logistic Regression Algorithm already inbuilt in python.
- Please refer the python coding file fc more details

	Converted_IND	Converted_Prob	Prospect_IND
0	0	0.523486	1871
1	0	0.113305	6795
2	0	0.336733	3516
3	0	0.818686	8105
4	0	0.292254	3934

Model Evaluation

Accuracy, Sensitivity and Specificity

- Accuracy: Percentage of correctly predicted labels.

$$Accuracy = \frac{Correctly\ Predicted\ Labels}{Total\ Number\ of\ Labels}$$

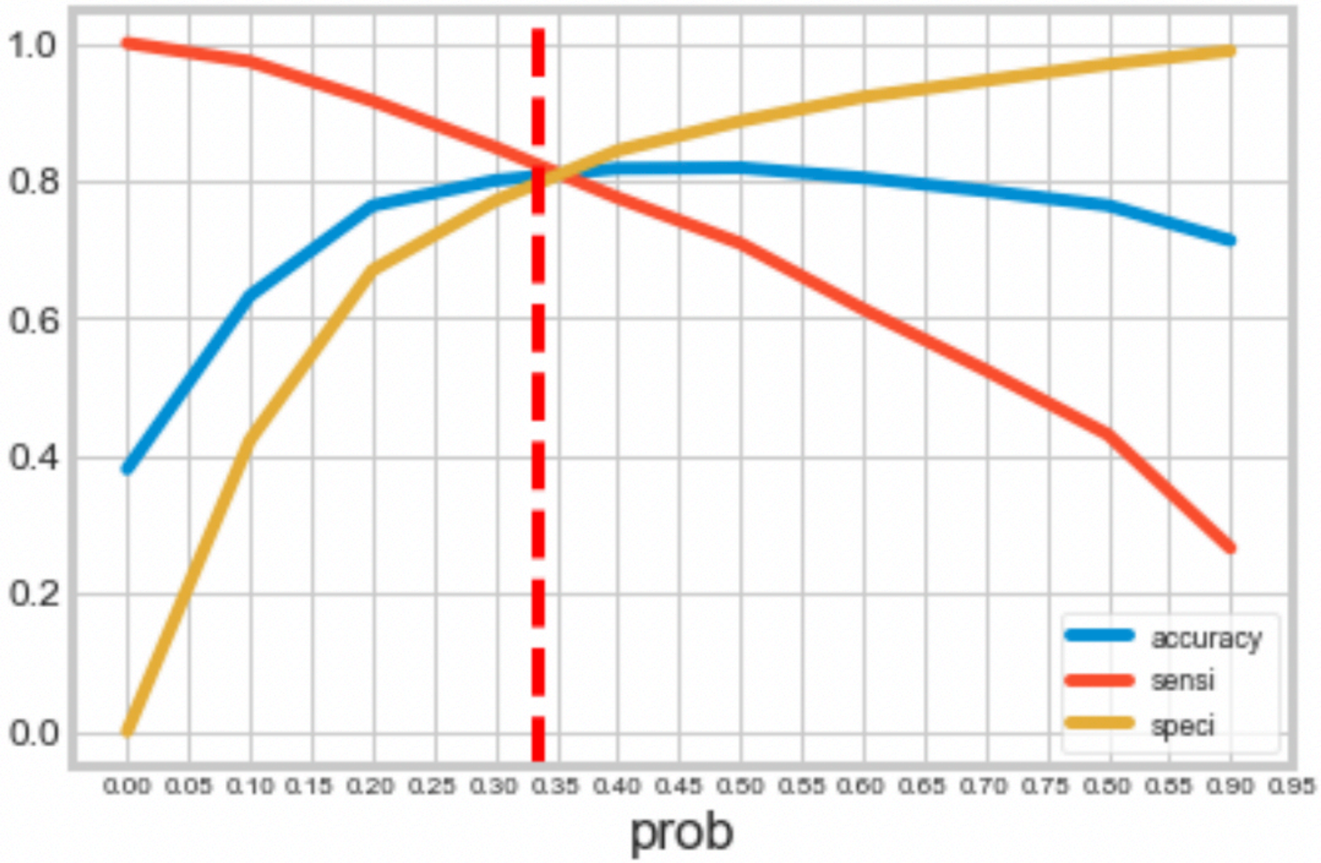
- Sensitivity: Percentage of true '1' to total number of '1' predicted.

$$Sensitivity = \frac{Number\ of\ actual\ Yeses\ correctly\ predicted}{Total\ number\ of\ actual\ Yeses}$$

- Specificity: Percentage of true '0' to total number of '0' predicted.

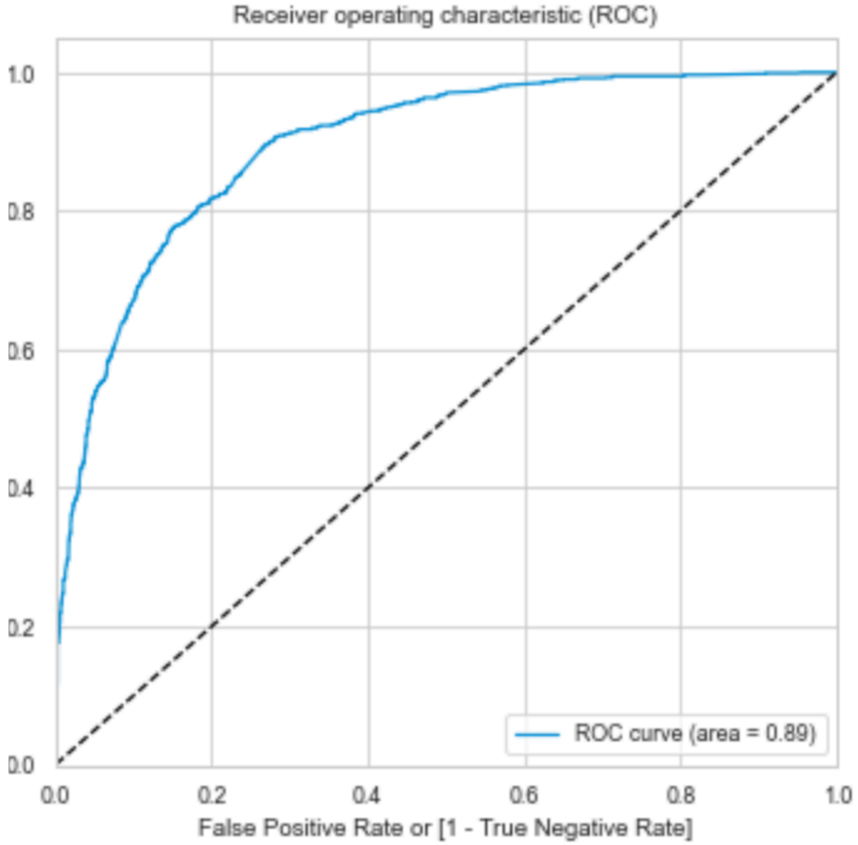
$$Specificity = \frac{Number\ of\ actual\ Nos\ correctly\ predicted}{Total\ number\ of\ actual\ Nos}$$

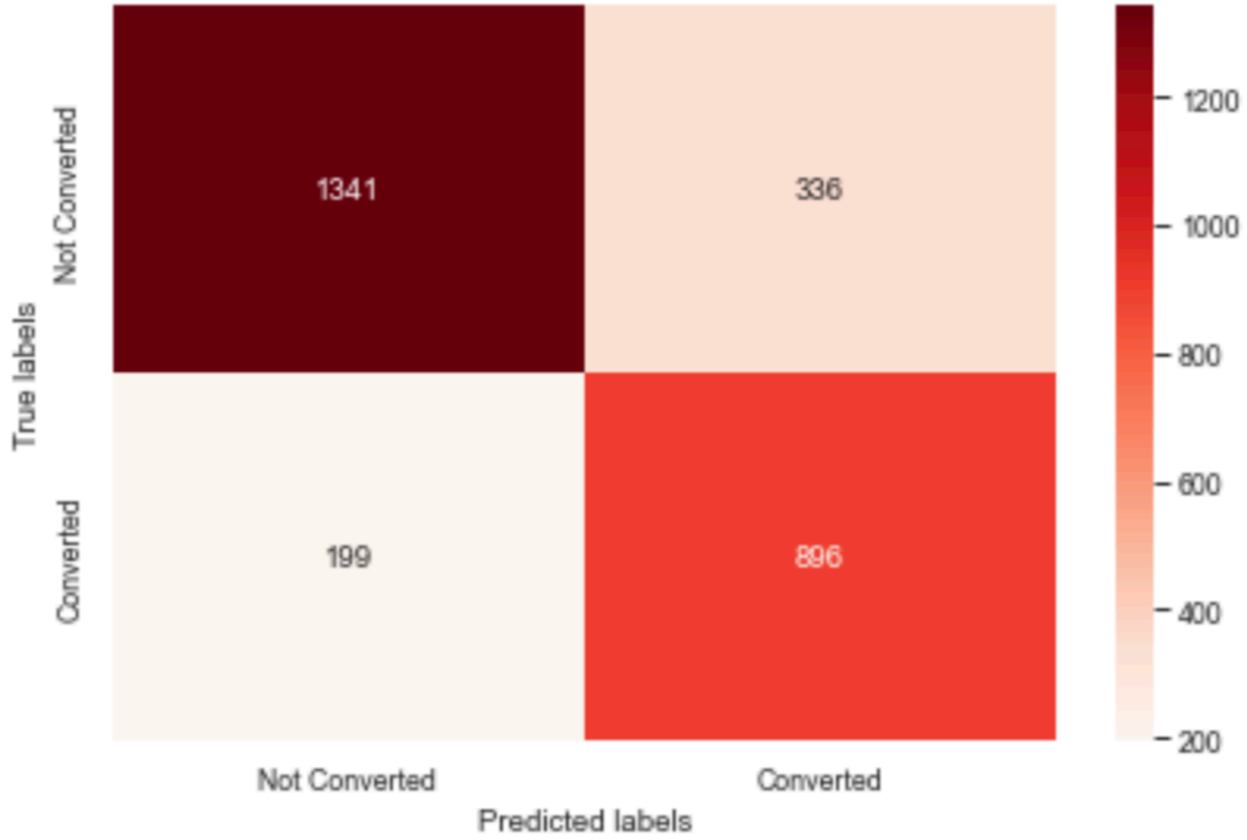
- Method: We are predicting accuracy, sensitivity and specificity for different values of conversion probability such as 0.0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9.
- Optimal Cut off Probability: Point on the graph where accuracy, sensitivity and specificity lines intersects. Please refer the graph to the right to check that optimal cut off point for our model is 0.335.
- This means out of 10 leads, about 3 leads are converting.



Model Evaluation

Confusion Matrix and ROC Curve





- True Positive Rate: $$True\ Positive\ Rate\ (TPR) = \frac{True\ Positives}{Total\ Number\ of\ Actual\ Positives}$$
- False Positive Rate: $$False\ Positive\ Rate\ (FPR) = \frac{False\ Positives}{Total\ Number\ of\ Actual\ Negatives}$$

Model Evaluation

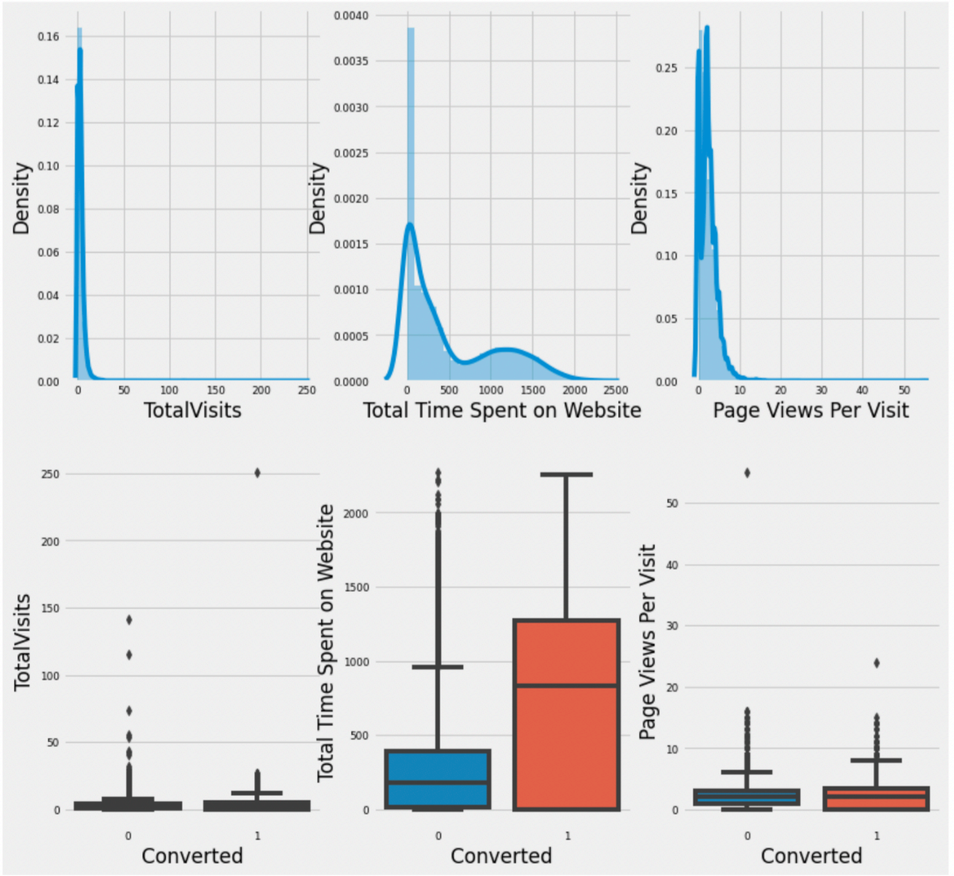
ROC Curve- Analysis

- Interpretation:
  - Area under the ROC Curve is directly proportional to the accuracy model. It shows a trade of between sensitivity and specificity.
  - If sensitivity increases and specificity.
  - Closer the curve comes of left hand border and then to the top border of the ROC space, more accurate is the test. Closer the curve comes to the 45 degree diagonal of the ROC space, the less accurate the test.
- Model Metrics:
  - Model Accuracy value is : 80.7%
  - Model Sensitivity value is : 81.83%
  - For the train data sensitivity was 81.79%, and for the test it is 81.83%. Hence, it can be said that then model is working well on the test data.
  - The area under the ROC Curve is 0.89, which indicates that the model is good.

Data Trends

Exploratory Data Analysis-Numerical Variables

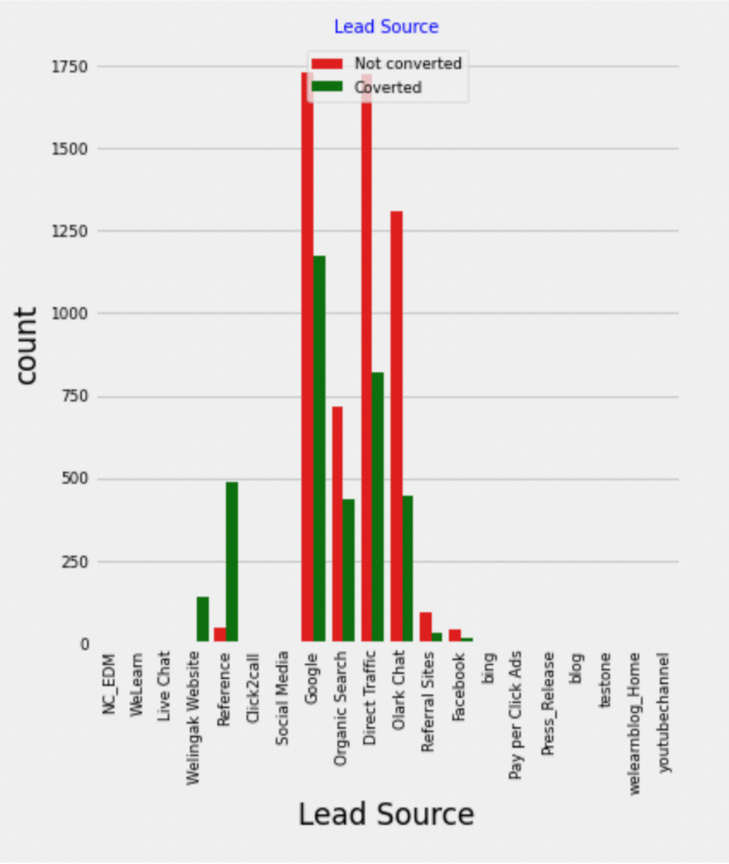
- Total Visits: This variable requires outlier treatment since there is not a lot that can be predicted based on the box plot. But from the graph it is evident that this variable has almost no impact on customer conversion.
- Total Time Spent on Website: As can be seen this variable seems to have an impact on lead conversion rate. As customers who have spent at an average of 1300 minutes on the website of 'X' education institute have a higher chances of purchasing the course.
- Page Views Per Visit: As can be seen this variable has little to no impact on conversion of customers. Page views per visit of a hot and cold lead seems to the similar. Hence, this cannot be used for making any predictions.



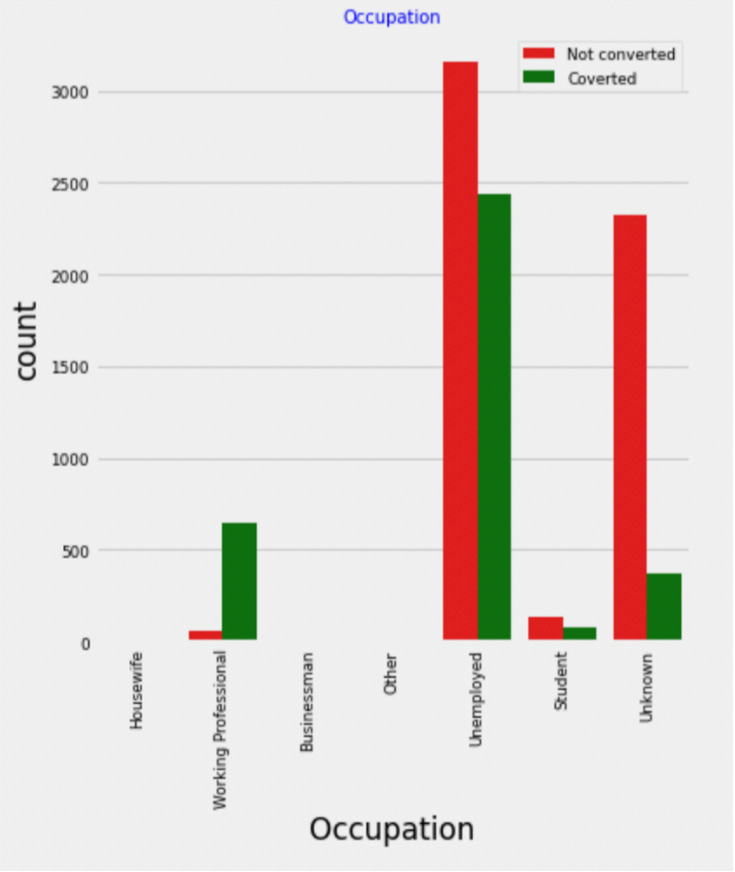
Please Note Outliers of Total Visits and Page Views per visit was taken care of (capped at 95%) since they can have an impact on logistic regression model.

Data Trends

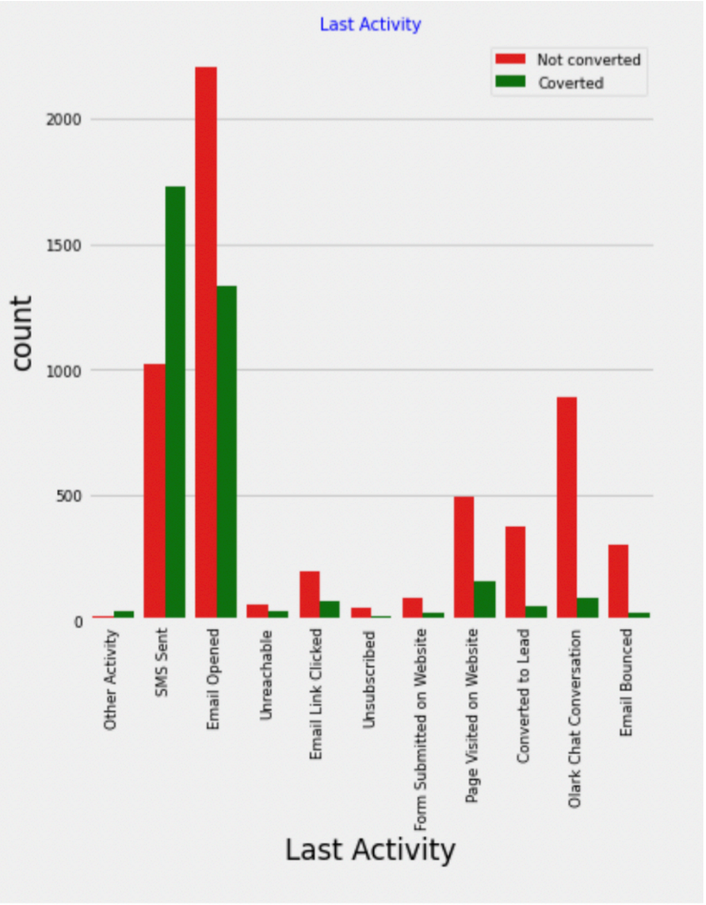
EDA- Categorical Variables



Hot leads-Leads Source Welling Website



Hot leads-Occupation Unemployed



Hot leads-Last Activity SMS sent

Analysis and Insights

Summary

- Business Recommendations
  - Customers who are Unemployed and Working Proffesional as they have very high lead conversion rate.
  - Customers whose last activity is ‘SMS sent’.
  - We do not know as of now what information this activity is conveying to the customer however, this is a very important strategy to keep in mind as this is generating almost 50% lead conversion rate. Customers whose Last activity has been Olark chat conversation should not be contacted, very poor leads.
  - This can be researched further and if required investment to this activity can be removed. There is a strong negative relation to this variable against lead conversion both from EDA and ML insights.
  - In order to address the volume start filtering out the cold leads i.e. people with bounced email, people who do not spend much time on the website etc.
- Assumptions
  - All data is correct and no tampering has been done with it.
  - Recommendations that are more cost effective are preferred.
  - Number online educators are increasing and the competition is fierce. Hence, aside from the recommendations, the product quality or the courses by X education institute is of high quality, competitive and industry relevant.
  - All variables are just predictors, the stories can change with time and other factors.
  - Course Fee is affordable or there is a quick facilitation of education loans by the institute. Fee amount is assumed not be a blocker