Lead Scoring Case Study

Submitted by – ANKIT DATTA & RHITHIK PR.

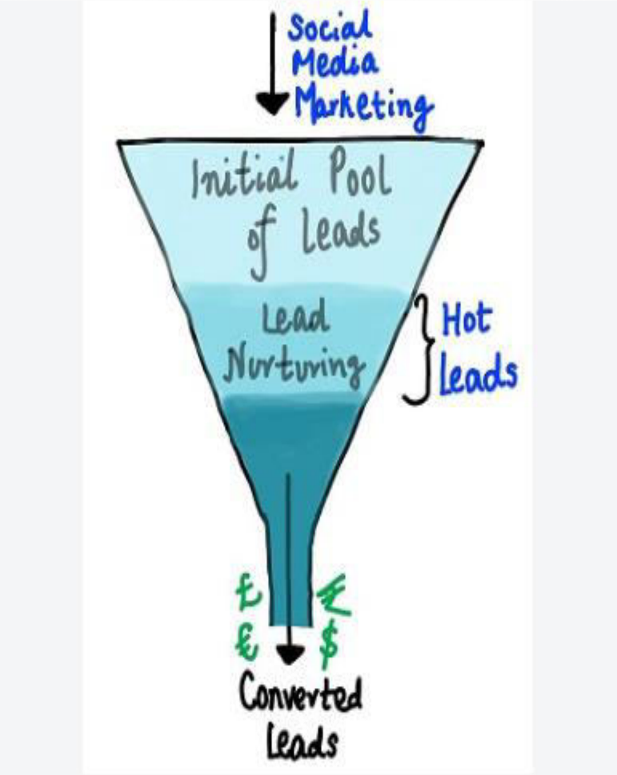
Problem Statement

- X Education sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e., the leads that are most likely to convert into paying customers.
- The company needs a model wherein you a lead score is assigned 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 lead Conversion rate is 38% only.

Business Objective

- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Data Understanding

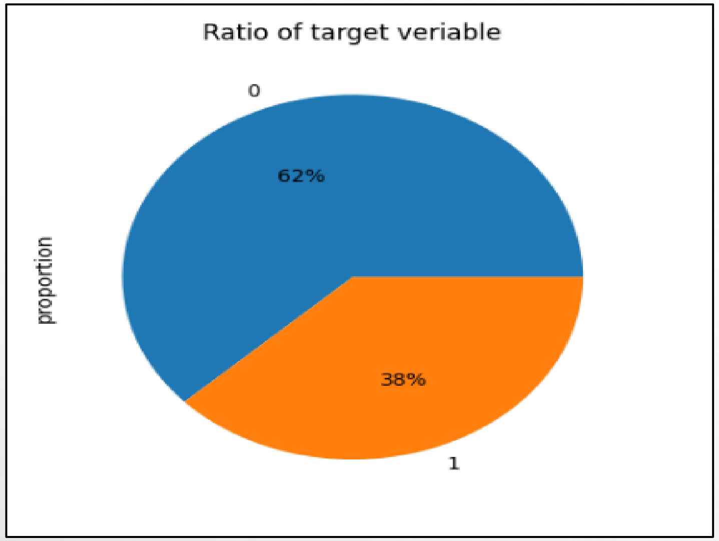


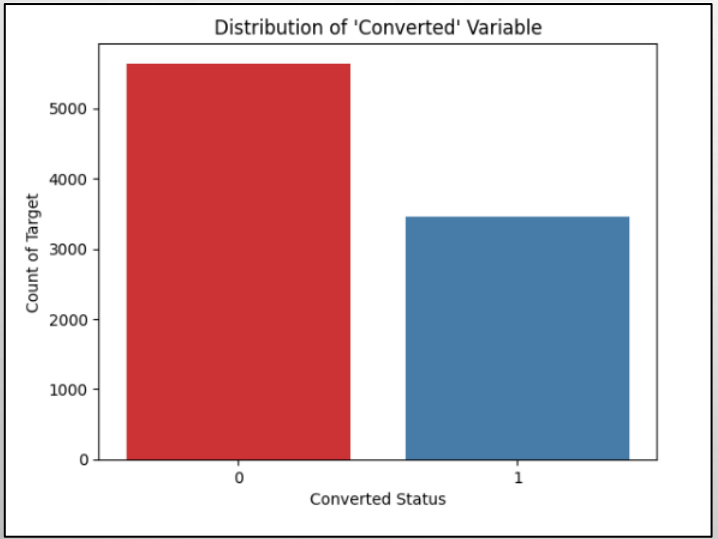
Lead Conversion Process - Demonstrated as a funnel

- We are 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 target variable, in this case, is the column ‘Converted’ which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn’t converted.
- You can learn more about the dataset from the data dictionary provided in the zip folder at the end of the page.
- Another thing that you also need to check out are the levels present in the categorical variables. Many of the categorical variables have a level called 'Select' which needs to be handled because it is as good as a null value.

Analysis Approach

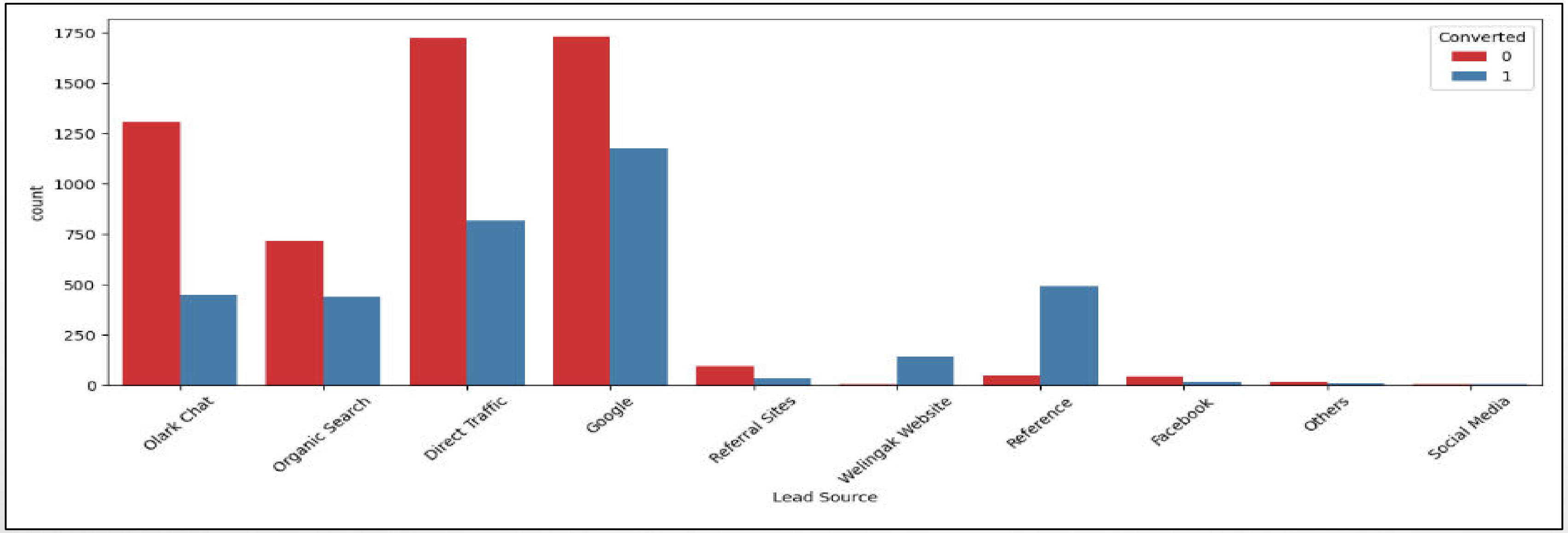
- Data Reading and Understanding
- Data Cleaning
- Data Transformation
- Dummy Variables Creation
- Test-Train Split
- Feature Rescaling
- Model Building
- Conclusion





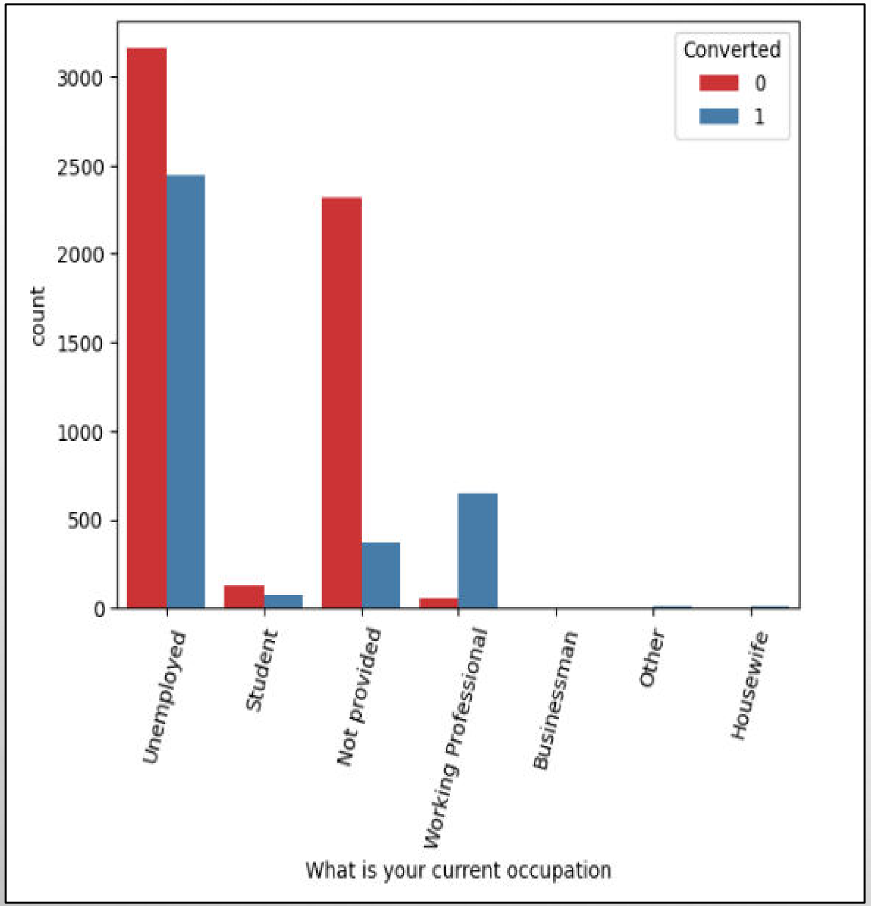
We can see in the graphs to the left,

- The Lead Score, which ranges from 0 to 100, indicates the probability of the customer being converted or not.
- The lead Conversion rate is 38%.
- Only 38% of the total leads gets converted.
- 62% of the total leads doesn't gets converted.
- There is data imbalance in the dataset.



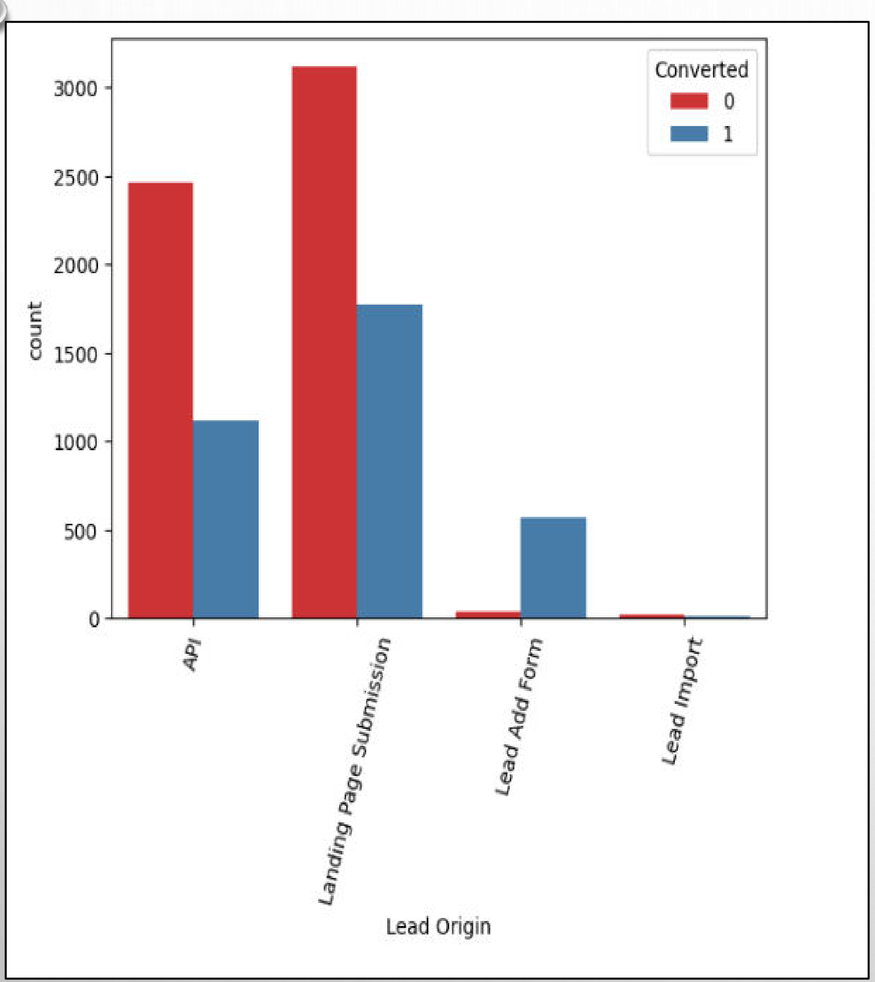
We can see in the above graph,

- Maximum Leads are generated by Google and Direct Traffic.
- Conversion rate of Reference leads and Welinkgak Website leads is very high.



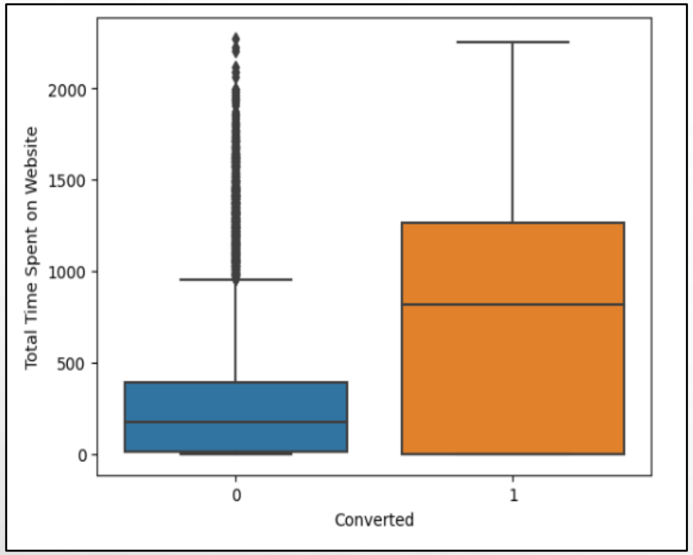
We can see in the graph to the left,

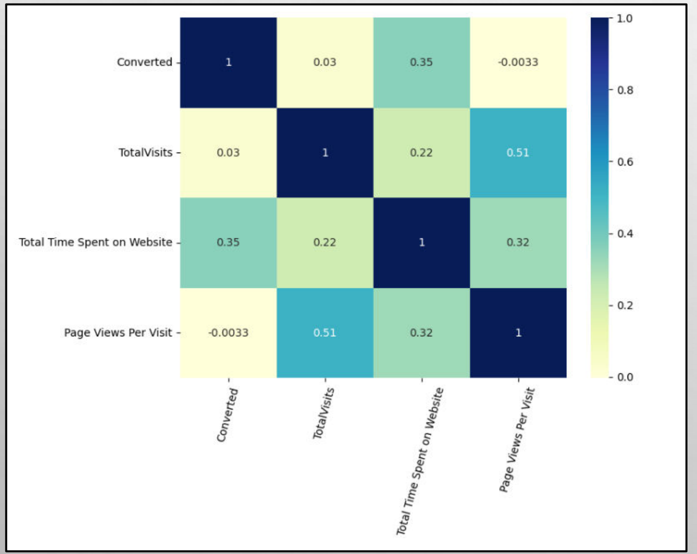
- Maximum leads generated are unemployed and their conversion rate is more than 50%.
- Conversion rate of working professionals is very high.



We can see in the graph to the left,

- Maximum leads are generated from 'Landing Page Submission' but conversion rate is not too good.
- Lead Add Form has high conversion rate.



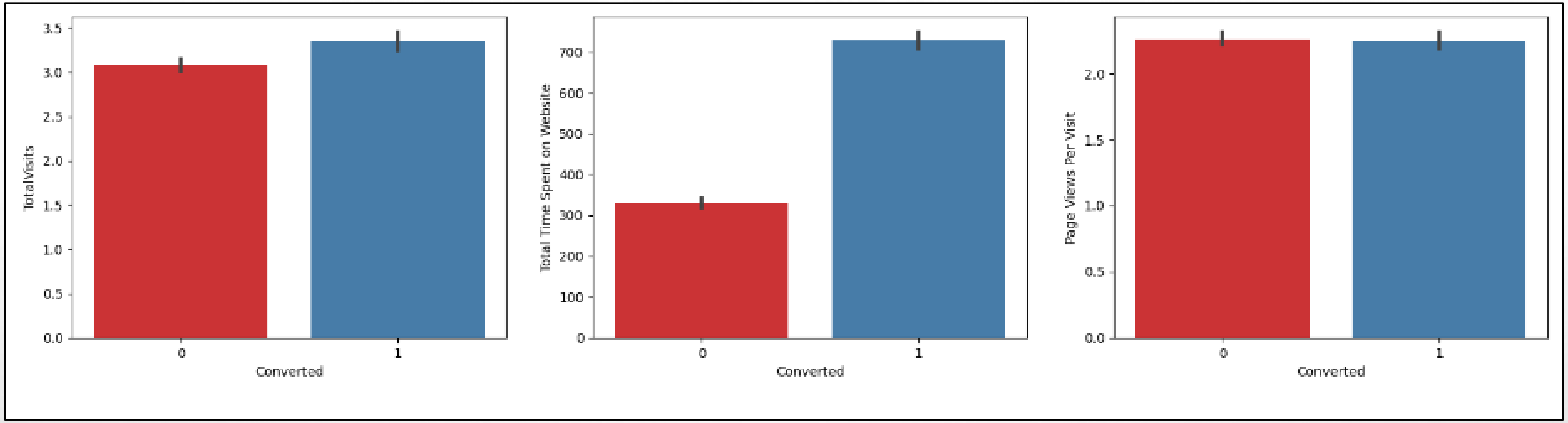


We can see in the graph to the left,

- Leads spending more time on website are more likely to convert , thus website should be made more engaging to increase conversion rate

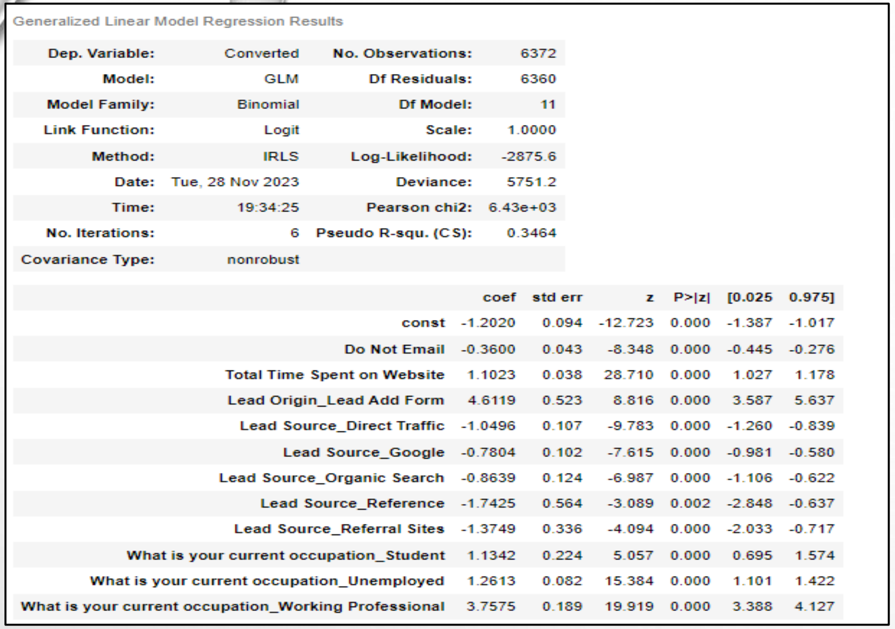
We can see in the graph to the left,

- We can notice that while the variables don't exhibit strong correlations with each other, there is still multicollinearity present among certain features.



We can see in the above graph,

- The conversion rate is high for Total Visits, Total Time Spent on Website and Page Views Per Visit.



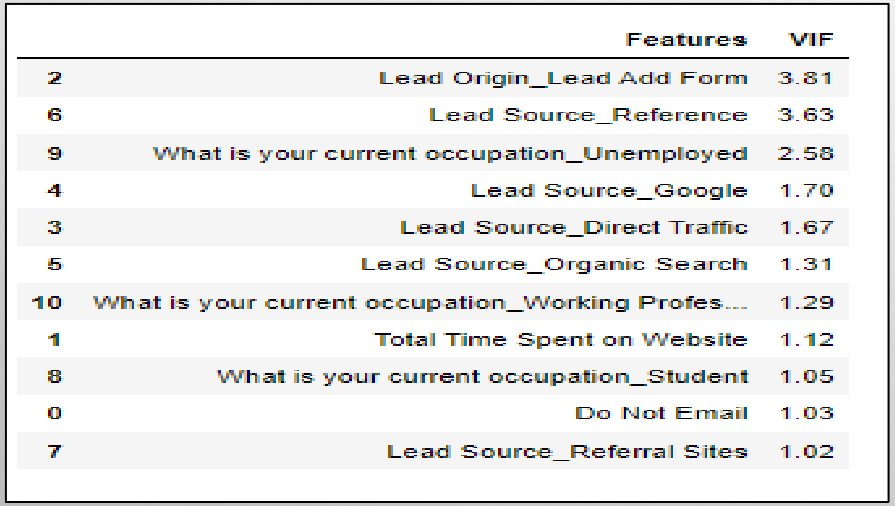
Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6372
Model:	GLM	Df Residuals:	6360
Model Family:	Binomial	Df Model:	11
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2875.6
Date:	Tue, 28 Nov 2023	Deviance:	5751.2
Time:	19:34:25	Pearson chi2:	6.43e+03
No. Iterations:	6	Pseudo R-squ. (CS):	0.3464
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.2020	0.094	-12.723	0.000	-1.387	-1.017
Do Not Email	-0.3600	0.043	-8.348	0.000	-0.445	-0.276
Total Time Spent on Website	1.1023	0.038	28.710	0.000	1.027	1.178
Lead Origin_Lead Add Form	4.6119	0.523	8.816	0.000	3.587	5.637
Lead Source_Direct Traffic	-1.0496	0.107	-9.783	0.000	-1.260	-0.839
Lead Source_Google	-0.7804	0.102	-7.615	0.000	-0.981	-0.580
Lead Source_Organic Search	-0.8639	0.124	-6.987	0.000	-1.106	-0.622
Lead Source_Reference	-1.7425	0.564	-3.089	0.002	-2.848	-0.637
Lead Source_Referral Sites	-1.3749	0.336	-4.094	0.000	-2.033	-0.717
What is your current occupation_Student	1.1342	0.224	5.057	0.000	0.695	1.574
What is your current occupation_Unemployed	1.2613	0.082	15.384	0.000	1.101	1.422
What is your current occupation_Working Professional	3.7575	0.189	19.919	0.000	3.388	4.127

We can see in the final model to the left,

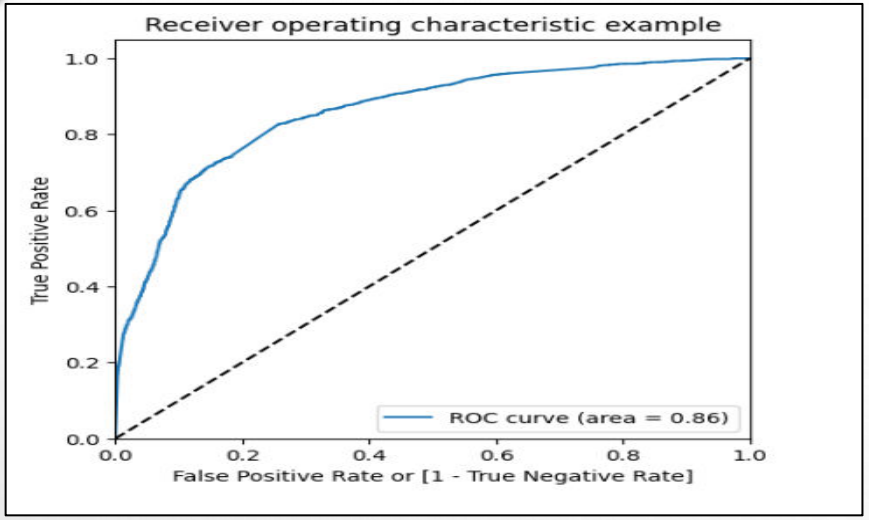
- Model 5 seems to be stable with significant p-values.
- Variables with high p-values are dropped.
- We are using Model 5 for further analysis.



	Features	VIF
2	Lead Origin_Lead Add Form	3.81
6	Lead Source_Reference	3.63
9	What is your current occupation_Unemployed	2.58
4	Lead Source_Google	1.70
3	Lead Source_Direct Traffic	1.67
5	Lead Source_Organic Search	1.31
10	What is your current occupation_Working Profes...	1.29
1	Total Time Spent on Website	1.12
8	What is your current occupation_Student	1.05
0	Do Not Email	1.03
7	Lead Source_Referral Sites	1.02

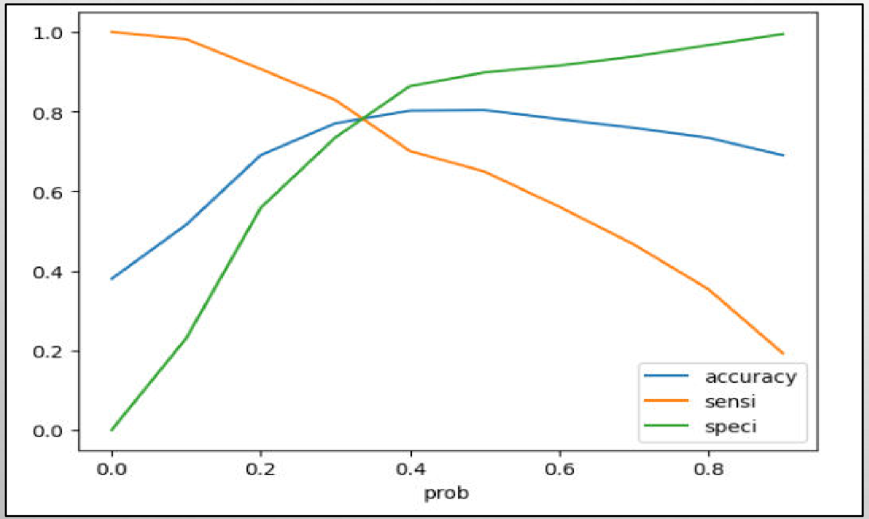
We can see in the VIF values of the feature variables to the left,

- All variables have a good value of VIF. So we need not drop any more variables and we can proceed with making predictions using this model.



We can see in the graph to the left,

- The ROC Curve should be a value close to 1. We are getting a good value of 0.86 indicating a good predictive model.



We can see in the graph to the left,

- 0.3 is the optimum point to take it as a cutoff probability.

Inference:

- ❑ The ROC curve has a value of 0.86, which is very good. We have the following values for the Train Data:
- ▪ Accuracy : 77.05%
- ▪ Sensitivity :82.89%
- ▪ Specificity : 73.49%
- ❑ The Train data conversion rate is approximately 83%.
- ❑ After running the model on the Test Data these are the figures we obtain:
- ▪ Accuracy : 77.52%
- ▪ Sensitivity :83.01%
- ▪ Specificity : 74.13%
- ❑ The Test data conversion rate is approximately 83%.
- ❑ We can see that the final prediction of conversions have a target rate of 83% (almost same as predictions made on training data set)

Conclusion:

- While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
- Accuracy, Sensitivity and Specificity values of test set are around 77%, 83% and 74% which are approximately closer to the respective values calculated using trained set.
- Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 80%
- Hence overall this model seems to be good.

Recommendations:

- The company should make calls to the leads coming from the lead sources "Welingak Websites" and "Reference" as these are more likely to get converted.
- The company should make calls to the leads who are the "working professionals" as they are more likely to get converted.
- The company should make calls to the leads who spent "more time on the websites" as these are more likely to get converted.
- The company should make calls to the leads coming from the lead sources "Olark Chat" as these are more likely to get converted.
- The company should make calls to the leads whose last activity was SMS Sent as they are more likely to get converted.
- The company should not make calls to the leads whose last activity was "Olark Chat Conversation" as they are not likely to get converted.
- The company should not make calls to the leads whose lead origin is "Landing Page Submission" as they are not likely to get converted.
- The company should not make calls to the leads whose Specialization was "Others" as they are not likely to get converted.
- The company should not make calls to the leads who chose the option of "Do not Email" as "yes" as they are not likely to get converted.

Thank you !