Lead Scoring Case Study

Submitted by:-

Jayati Chowdhury

Abbas Bilgrami

Tilak Hanchate

Problem Statement

An X Education need help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires us 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%.

Goal of 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.

Followed Below Steps

- Data Cleaning & Processing
- Exploratory Data Analysis
- Data Preparations
- Model Building
- Model Evaluation

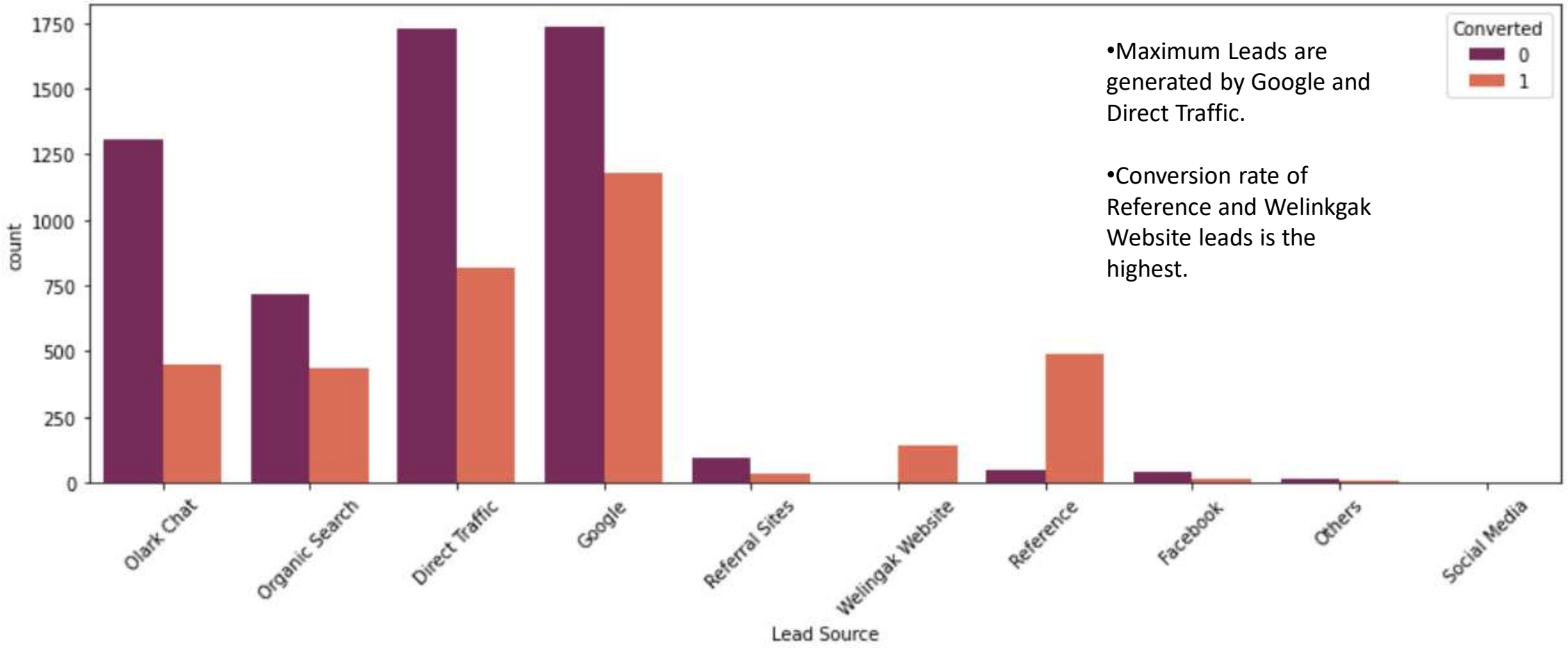
Data Visualization

Data Analysis

Categorical Attribute Analysis

Numerical Attribute Analysis

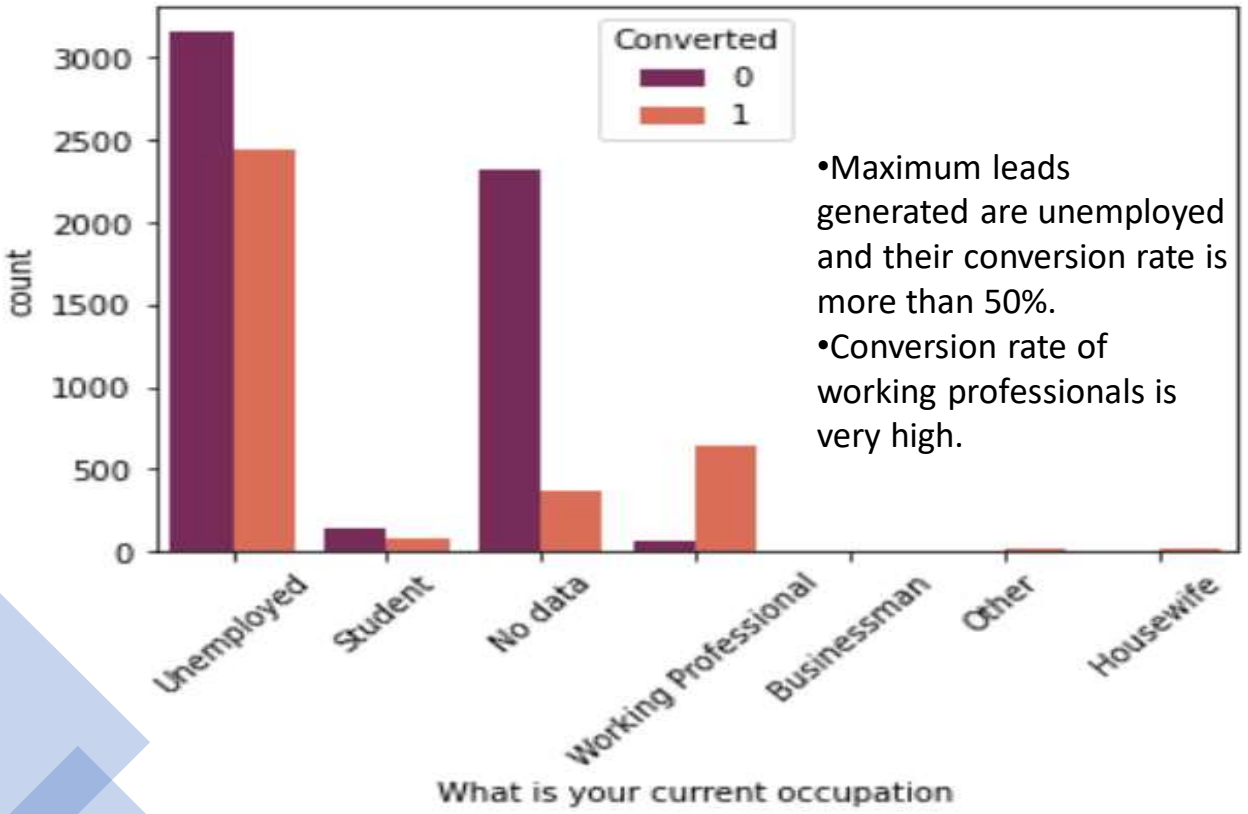
Lead Source Distribution Against Conversion



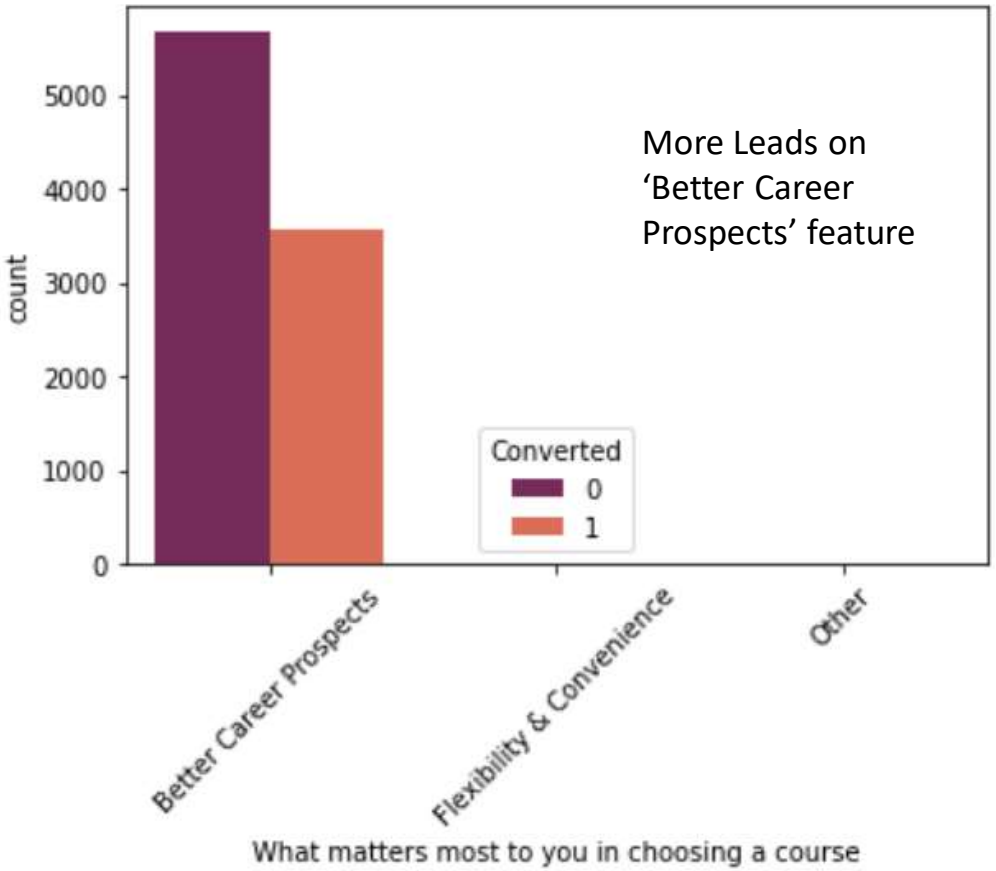
- Maximum Leads are generated by Google and Direct Traffic.

- Conversion rate of Reference and Welinkgak Website leads is the highest.

Occupation & Prospects Against Conversion

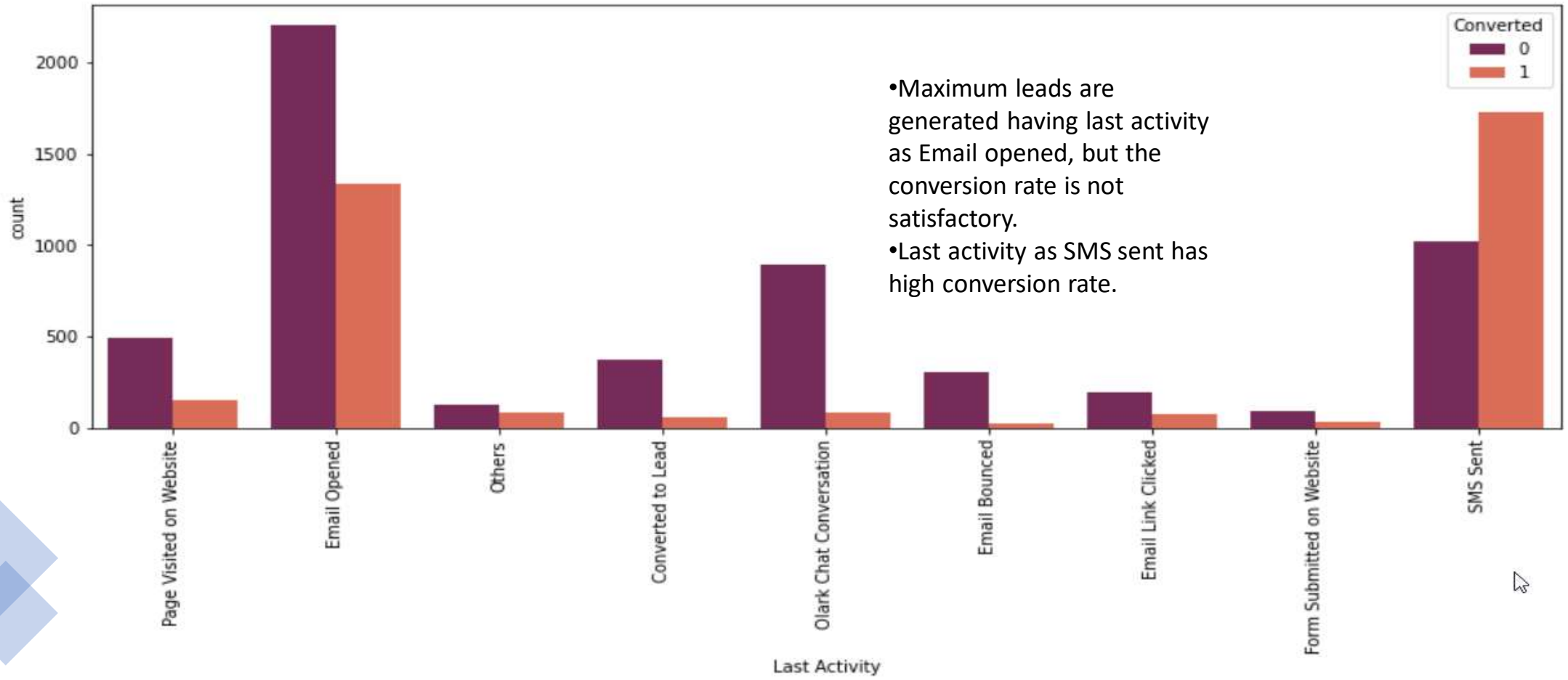


- Maximum leads generated are unemployed and their conversion rate is more than 50%.
- Conversion rate of working professionals is very high.



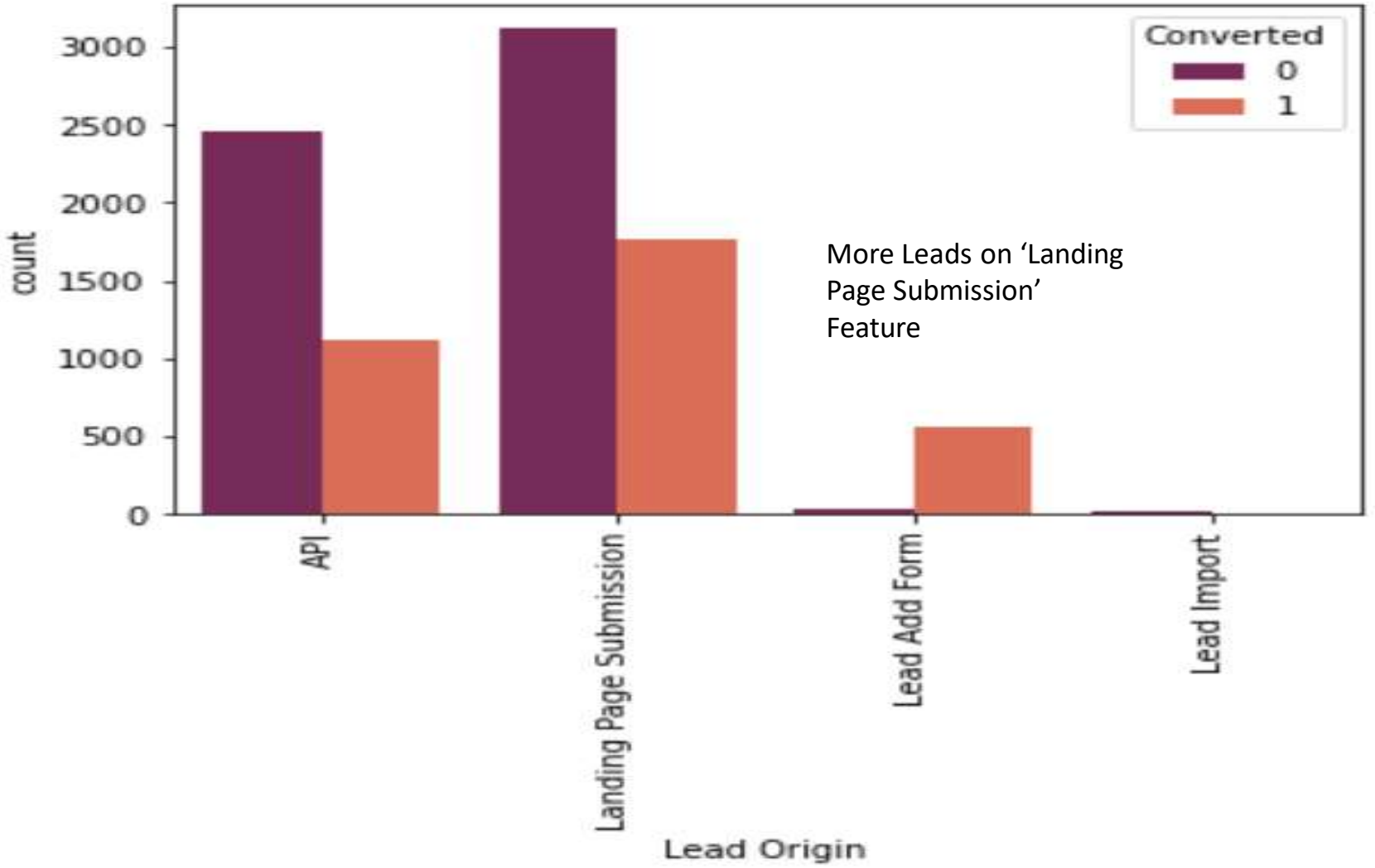
More Leads on ‘Better Career Prospects’ feature

Last Activity Against Conversion



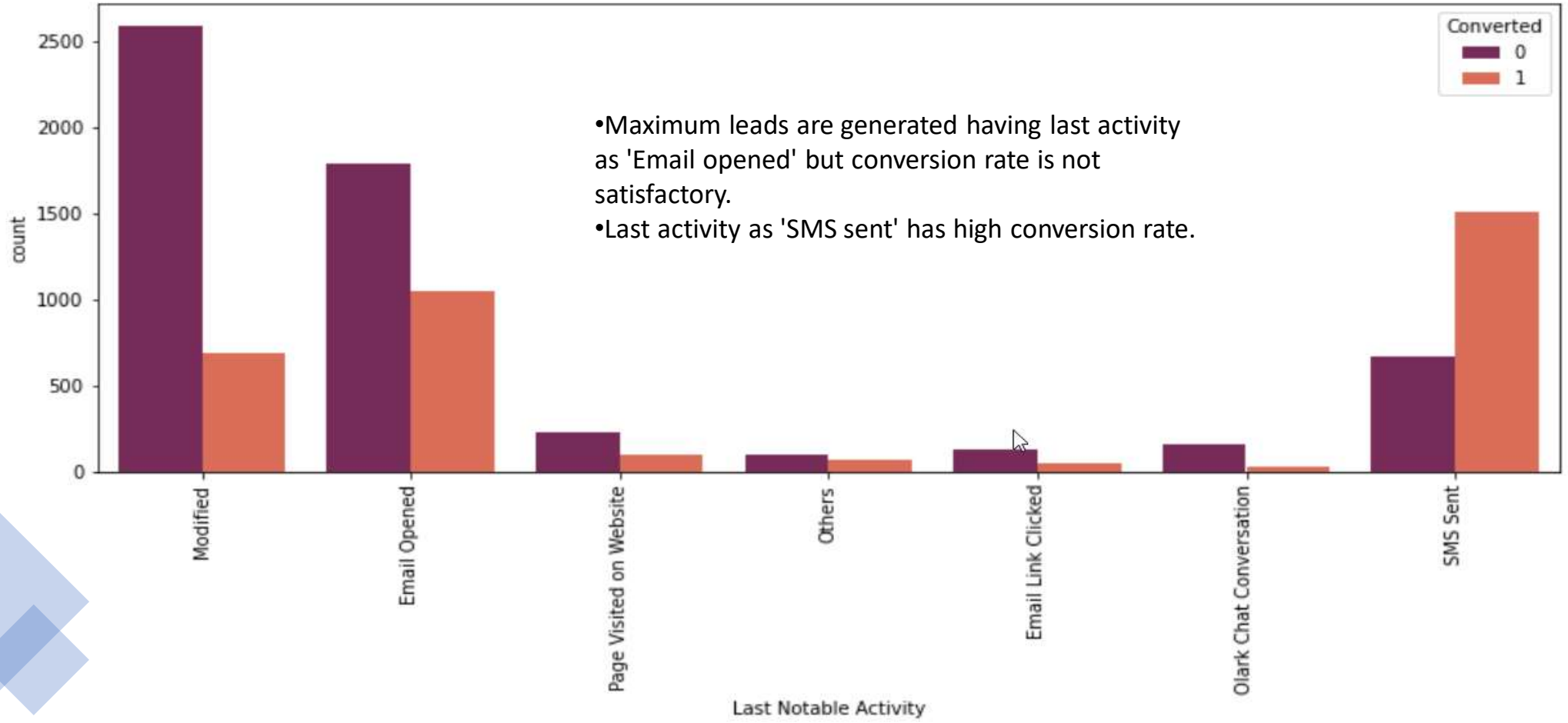
- Maximum leads are generated having last activity as Email opened, but the conversion rate is not satisfactory.
- Last activity as SMS sent has high conversion rate.

Lead Origin Against Conversion



More Leads on ‘Landing Page Submission’ Feature

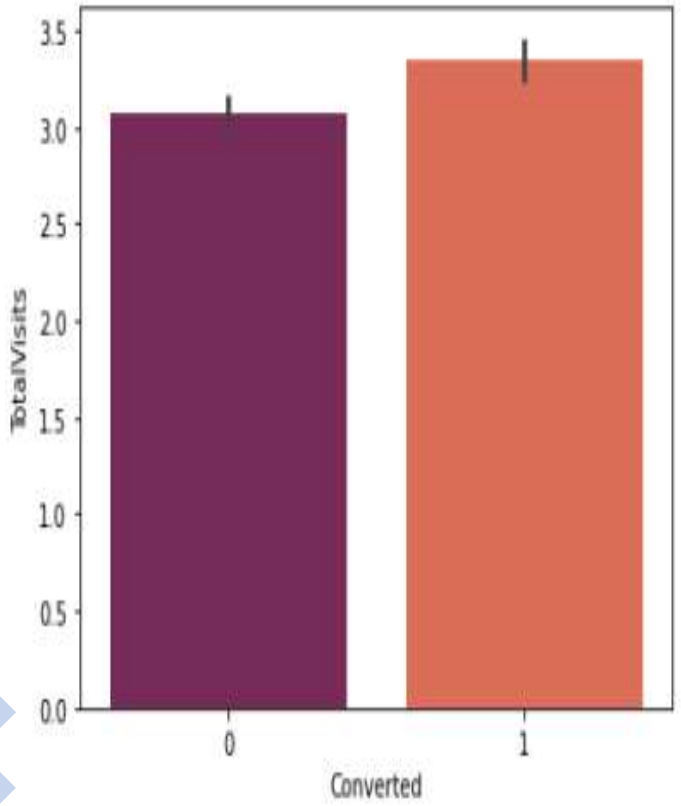
Last Notable Activity Against Conversion

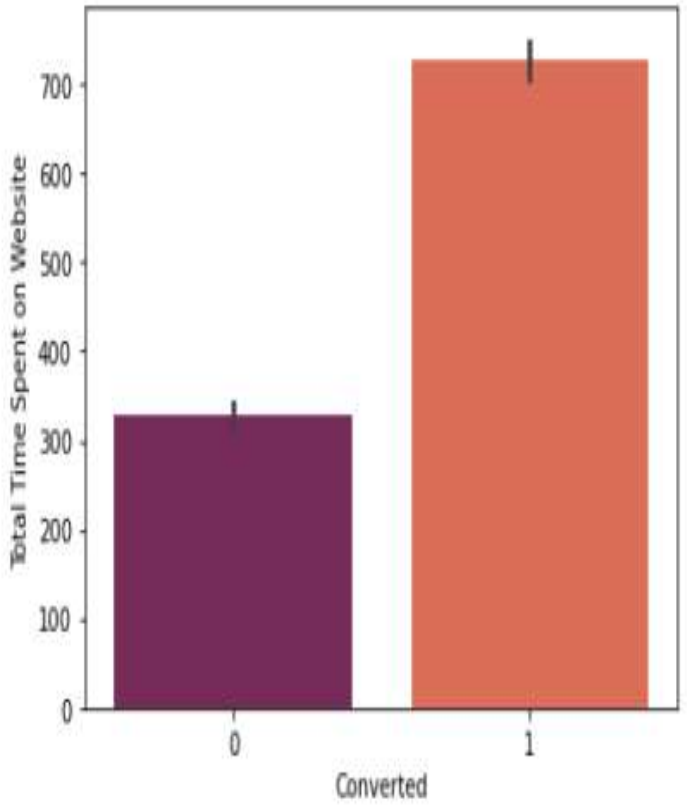


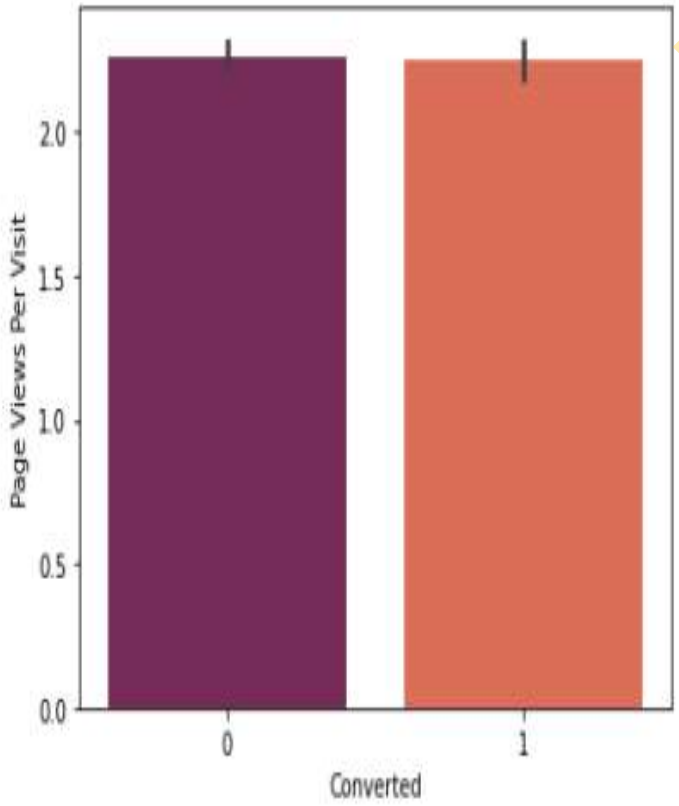
- Maximum leads are generated having last activity as 'Email opened' but conversion rate is not satisfactory.
- Last activity as 'SMS sent' has high conversion rate.

Numerical Attribute Analysis Against Conversion

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







Model Evaluation

Final Model Summary –All P Values Are Zero

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.7857	0.168	10.635	0.000	1.457	2.115
Do Not Email	-0.3635	0.041	-8.794	0.000	-0.445	-0.282
TotalVisits	0.2184	0.048	4.554	0.000	0.124	0.312
Total Time Spent on Website	0.9817	0.035	27.669	0.000	0.912	1.051
Page Views Per Visit	-0.3261	0.051	-6.378	0.000	-0.426	-0.226
Lead Origin_Lead Add Form	3.7164	0.212	17.567	0.000	3.302	4.131
No data	-3.6812	0.184	-20.047	0.000	-4.041	-3.321
Other	-2.5191	0.660	-3.819	0.000	-3.812	-1.226
Student	-2.5483	0.272	-9.379	0.000	-3.081	-2.016
Unemployed	-2.4361	0.172	-14.141	0.000	-2.774	-2.098

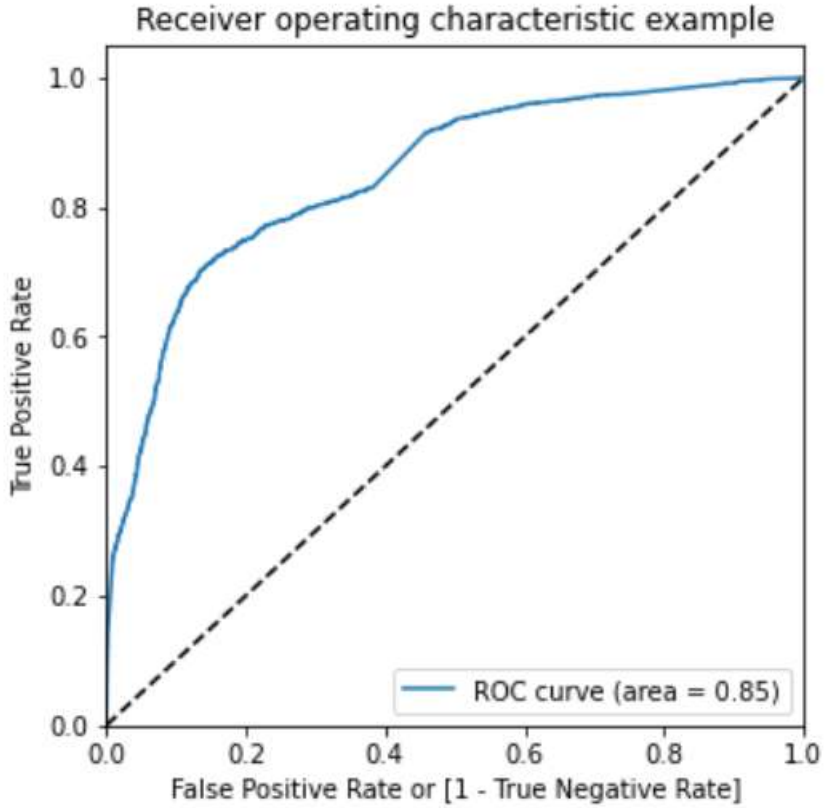
Important features depending on the co-efficient of Final Model are a)Lead Origin_Lead Add Form ,b)Total Time Spent on website ,c)Total Visits

Calculating VIF –on Final Model

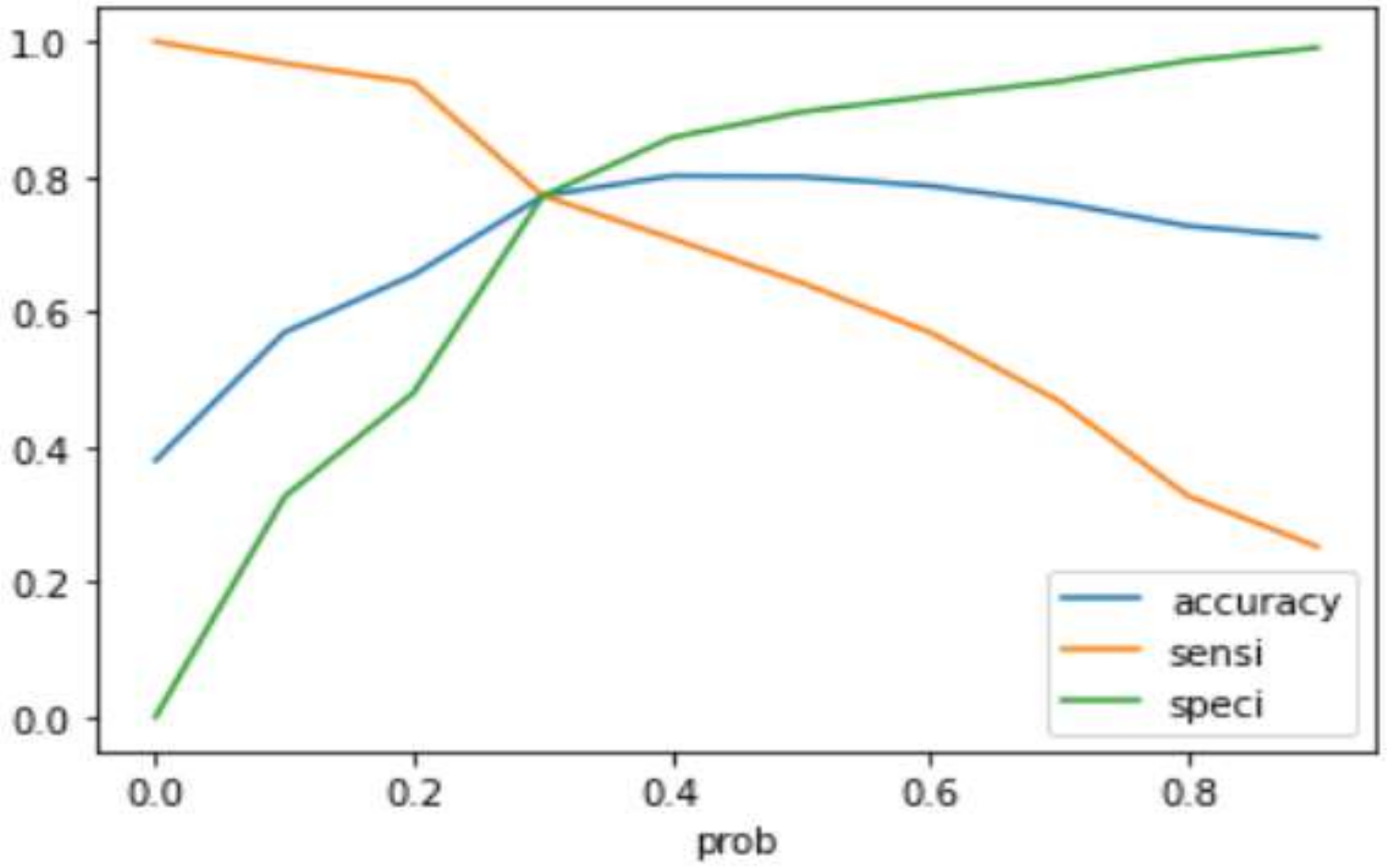
	Features	VIF
3	Page Views Per Visit	2.54
1	TotalVisits	2.48
4	Lead Origin_Lead Add Form	1.21
2	Total Time Spent on Website	1.20
8	Unemployed	1.09
5	No data	1.02
0	Do Not Email	1.01
6	Other	1.00
7	Student	1.00

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

ROC Curve-Area Under Curve 0.85



Finding Optimal Threshold



Showing changes in Sensitivity, Specificity, Accuracy with changes in the probability threshold values.

Optimal Cut-off = 0.30

Assigning Lead Score

	Prospect ID	Converted	Converted_prob	Lead_Score	final_Predicted
0	3504	0	0.310289	31	1
1	4050	1	0.921415	92	1
2	7201	0	0.179481	18	0
3	1196	0	0.288291	29	0
4	8219	1	0.360747	36	1

Conclusions

Inferences & Recommendations

After running the model on the Test Data these are the figures we obtain:-

- Accuracy : 77.00%
- Sensitivity :76.2%
- Specificity : 77.50%
- 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%, 76% and 77.5% 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%
- And three important feature depending on the co-efficient of Final Model are

a)Lead Origin_Lead Add Form ,b)Total Time Spent on website ,c)Total Visits

These three features are contributing more towards the probability of the Final Model.

So from the analysis this is clear that mainly using these features like Lead Origin_Lead Add Form ,Total Time Spent on website ,Total Visits

Student , Unemployed, Page Views Per Visit we can calculate the final probable lead conversions.

Thank You