LEAD SCORING CASE STUDY

PROBLEM STATEMENT

❖X Education faces a low lead conversion rate of around 30% despite a high number of leads.

❖Objective: Increase efficiency by identifying 'Hot Leads' for a higher conversion rate.

❖Tasked with building a model to assign lead scores, prioritizing leads with a higher likelihood of conversion.

❖CEO's target: Achieve an 80% lead conversion rate for improved sales efficiency.

BUSINESS OBJECTIVE

- Outline the approach briefly:

- Data Overview: 9000 data points, key attributes, 'Converted' as the target variable.

- Logistic Regression Model: Assign lead scores between 0 and 100.

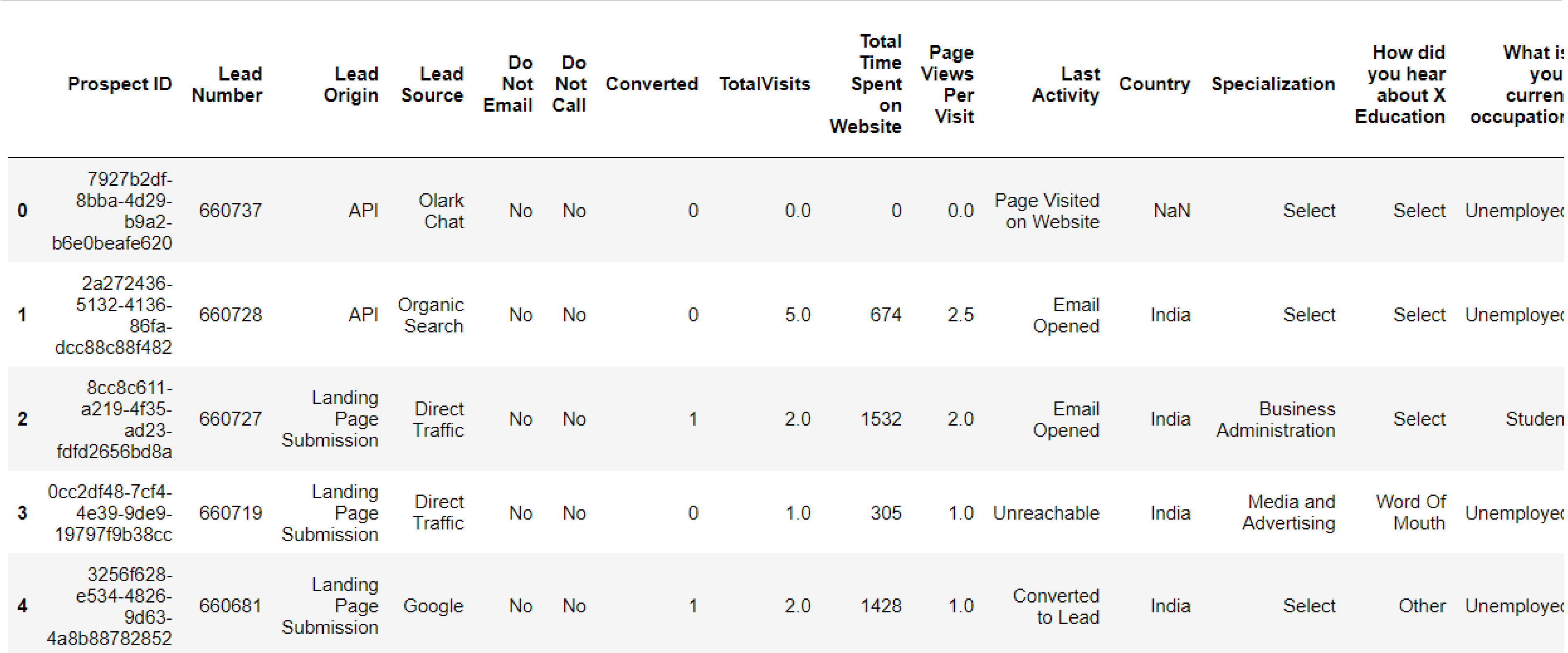
- Results: Conversion predictions, evaluation metrics (accuracy, precision, recall, F1-score).

- Conclude with key recommendations for X Education based on the model's insights.

- Optionally, include a visual representation of the lead conversion process funnel.

DATA SET

- 9000 data points with various attributes: Lead Source, Total Time Spent, Total Visits, Last Activity, etc.
- Target variable: 'Converted' (1 for converted, 0 for not converted).
- Check categorical variables for levels, especially 'Select' (considered as null value).
- Refer to the data dictionary in the provided zip folder for detailed dataset insights.



	Prospect ID	Lead Number	Lead Origin	Lead Source	Do Not Email	Do Not Call	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Last Activity	Country	Specialization	How did you hear about X Education	What is you curren occupatio
0	7927b2df-8bba-4d29-b9a2-b6e0beafe620	660737	API	Olark Chat	No	No	0	0.0	0	0.0	Page Visited on Website	NaN	Select	Select	Unemploye
1	2a272436-5132-4136-86fa-dcc88c88f482	660728	API	Organic Search	No	No	0	5.0	674	2.5	Email Opened	India	Select	Select	Unemploye
2	8cc8c611-a219-4f35-ad23-fdfd2656bd8a	660727	Landing Page Submission	Direct Traffic	No	No	1	2.0	1532	2.0	Email Opened	India	Business Administration	Select	Studen
3	0cc2df48-7cf4-4e39-9de9-19797f9b38cc	660719	Landing Page Submission	Direct Traffic	No	No	0	1.0	305	1.0	Unreachable	India	Media and Advertising	Word Of Mouth	Unemploye
4	3256f628-e534-4826-9d63-4a8b88782852	660681	Landing Page Submission	Google	No	No	1	2.0	1428	1.0	Converted to Lead	India	Select	Other	Unemploye

Approach & Methodology:

- Checking the missing values
- Handling outliers.
- Differentiates numerical columns and categorical columns.
- Univariate and Bivariate analysis.
- Correlations.
- Data Preparations
- Train Test Split
- Feature Scaling
- Model Building
- Checking Variance Inflation Factor (V.I.F)
- Confusion Matrix
- Plotting ROC Curve
- Finding optimal cut-off point
- Accuracy, Sensitivity, Specificity
- Precision And Recall

DATA CLEANING:

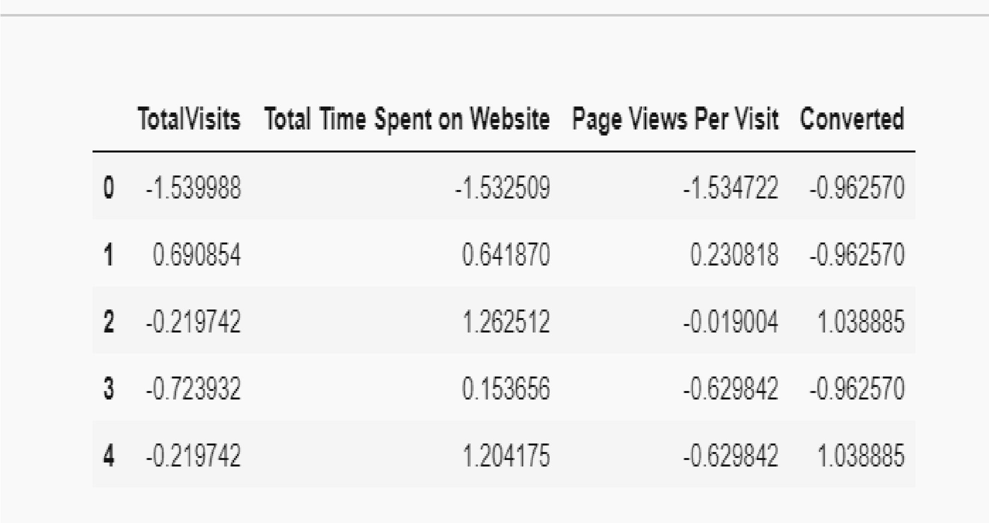
There is a lot of columns with high number of missing values and since we have around 9000+ data points we can eliminate the columns with 30% missing values;

We dropped City and Country variables since it's of no use to us as the company provides online courses;

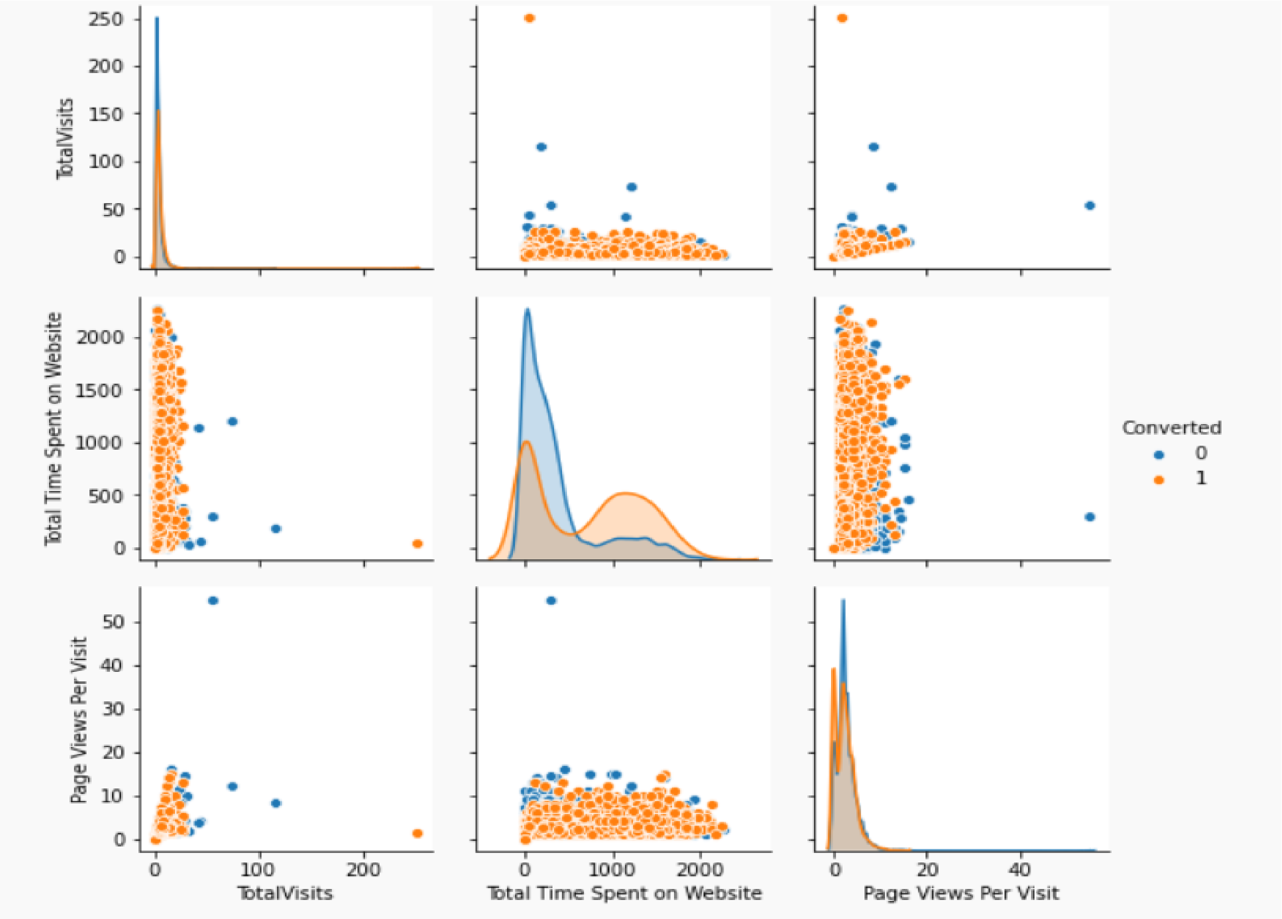
Prospect ID and Lead Number are just records identifier and as hence dropped.

We dropped all columns which have skewed data points as it wont have any predictability

DATA MODELING:

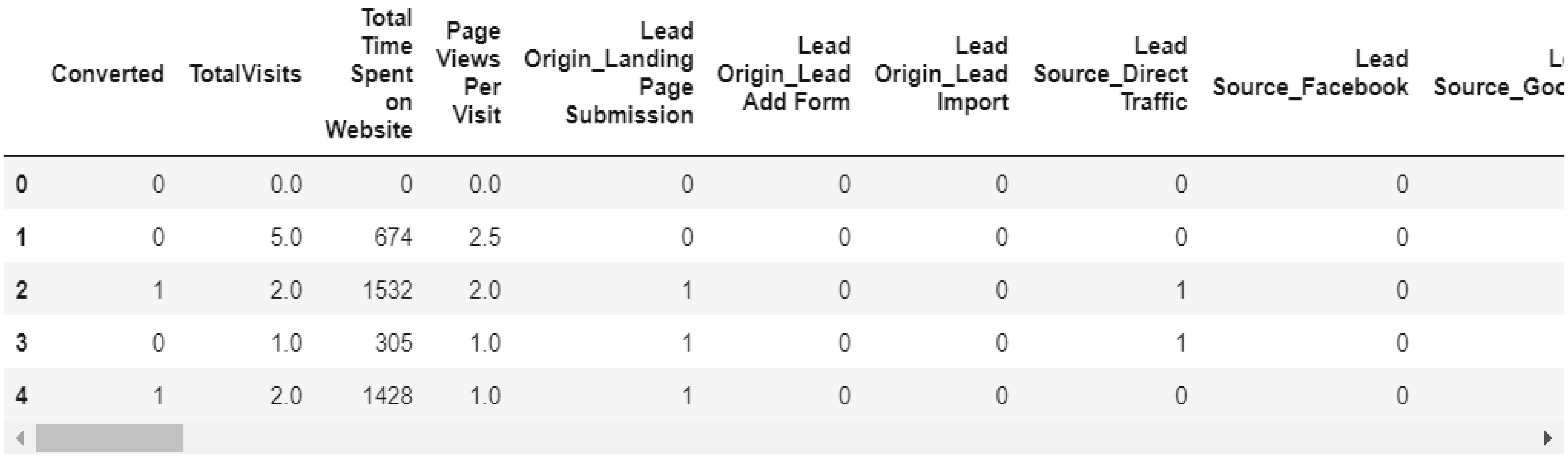


	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Converted
0	-1.539988	-1.532509	-1.534722	-0.962570
1	0.690854	0.641870	0.230818	-0.962570
2	-0.219742	1.262512	-0.019004	1.038885
3	-0.723932	0.153656	-0.629842	-0.962570
4	-0.219742	1.204175	-0.629842	1.038885



Data Preparation

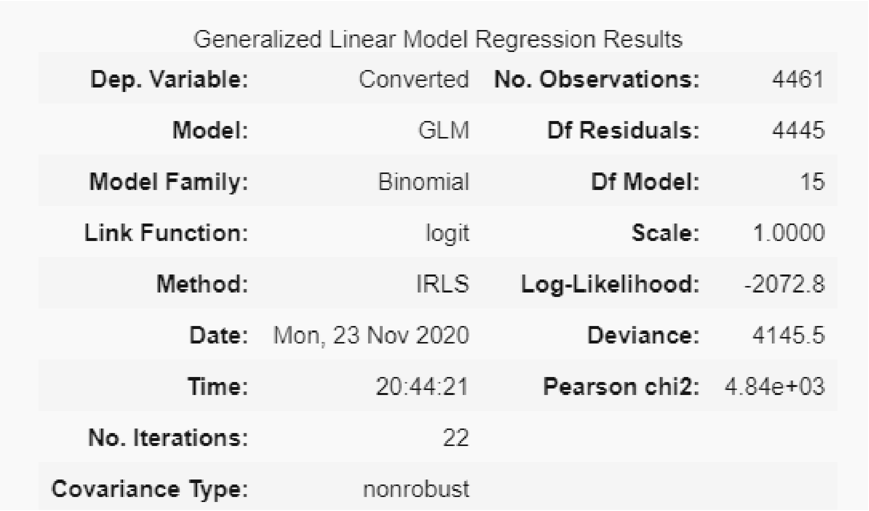
- After dummification : -



	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Lead Origin_Landing Page Submission	Lead Origin_Lead Add Form	Lead Origin_Lead Import	Lead Source_Direct Traffic	Lead Source_Facebook	L Source_Goc
0	0	0.0	0	0.0	0	0	0	0	0	
1	0	5.0	674	2.5	0	0	0	0	0	
2	1	2.0	1532	2.0	1	0	0	1	0	
3	0	1.0	305	1.0	1	0	0	1	0	
4	1	2.0	1428	1.0	1	0	0	0	0	

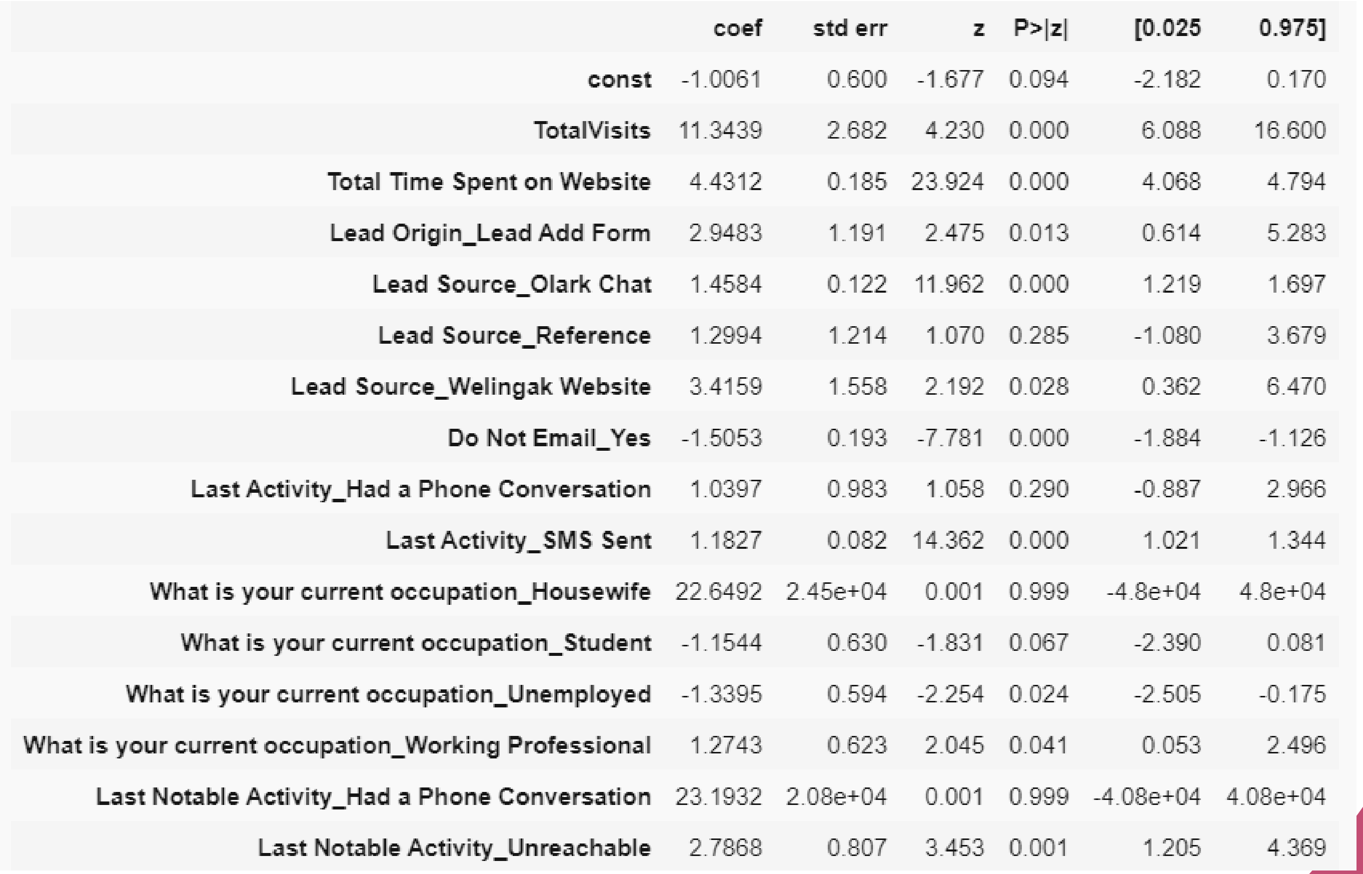
Model Building

After creating a RFE we got are model as shown below :



Generalized Linear Model Regression Results

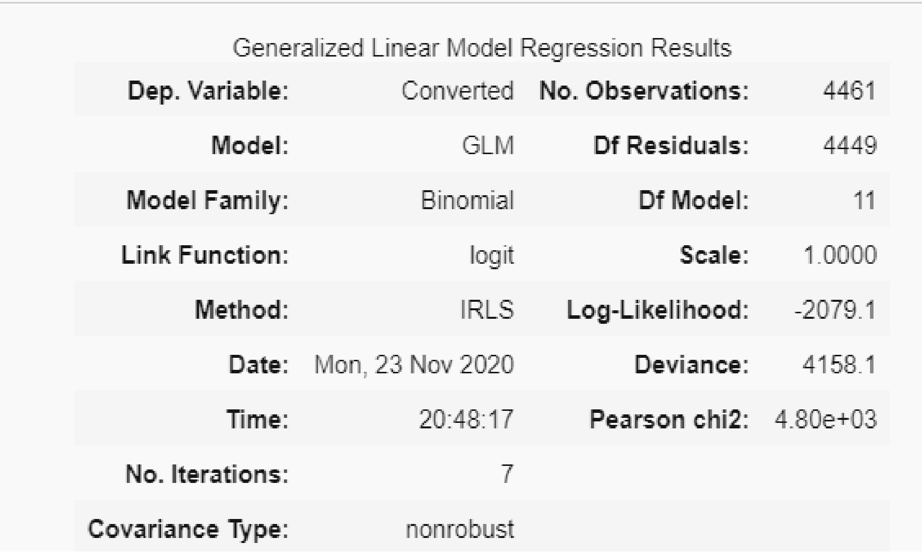
Dep. Variable:	Converted	**No. Observations:**	4461
Model:	GLM	**Df Residuals:**	4445
Model Family:	Binomial	**Df Model:**	15
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2072.8
Date:	Mon, 23 Nov 2020	**Deviance:**	4145.5
Time:	20:44:21	**Pearson chi2:**	4.84e+03
No. Iterations:	22		
Covariance Type:	nonrobust		



	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0061	0.600	-1.677	0.094	-2.182	0.170
TotalVisits	11.3439	2.682	4.230	0.000	6.088	16.600
Total Time Spent on Website	4.4312	0.185	23.924	0.000	4.068	4.794
Lead Origin_Lead Add Form	2.9483	1.191	2.475	0.013	0.614	5.283
Lead Source_Olark Chat	1.4584	0.122	11.962	0.000	1.219	1.697
Lead Source_Reference	1.2994	1.214	1.070	0.285	-1.080	3.679
Lead Source_Welingak Website	3.4159	1.558	2.192	0.028	0.362	6.470
Do Not Email_Yes	-1.5053	0.193	-7.781	0.000	-1.884	-1.126
Last Activity_Had a Phone Conversation	1.0397	0.983	1.058	0.290	-0.887	2.966
Last Activity_SMS Sent	1.1827	0.082	14.362	0.000	1.021	1.344
What is your current occupation_Housewife	22.6492	2.45e+04	0.001	0.999	-4.8e+04	4.8e+04
What is your current occupation_Student	-1.1544	0.630	-1.831	0.067	-2.390	0.081
What is your current occupation_Unemployed	-1.3395	0.594	-2.254	0.024	-2.505	-0.175
What is your current occupation_Working Professional	1.2743	0.623	2.045	0.041	0.053	2.496
Last Notable Activity_Had a Phone Conversation	23.1932	2.08e+04	0.001	0.999	-4.08e+04	4.08e+04
Last Notable Activity_Unreachable	2.7868	0.807	3.453	0.001	1.205	4.369

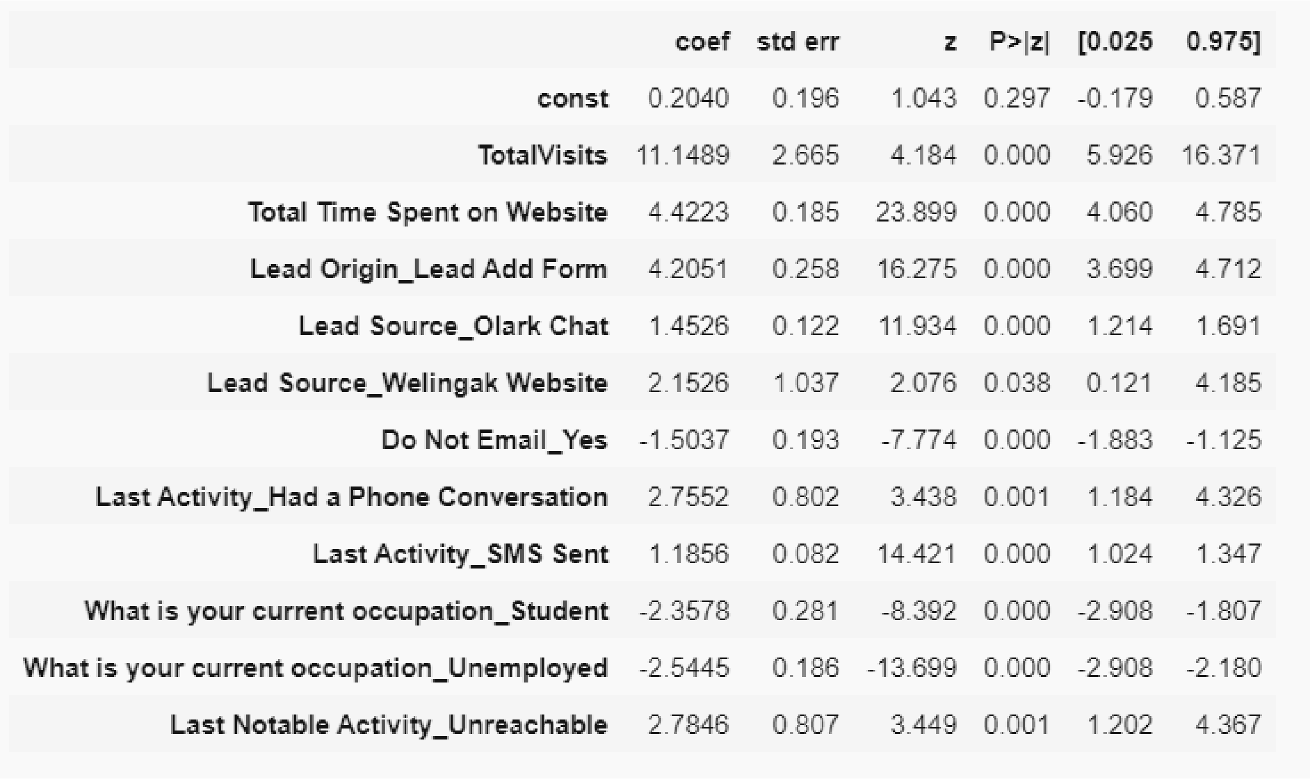
Model Building

After Removing the variables with high p-value finally we got are final model as shown below:



Generalized Linear Model Regression Results

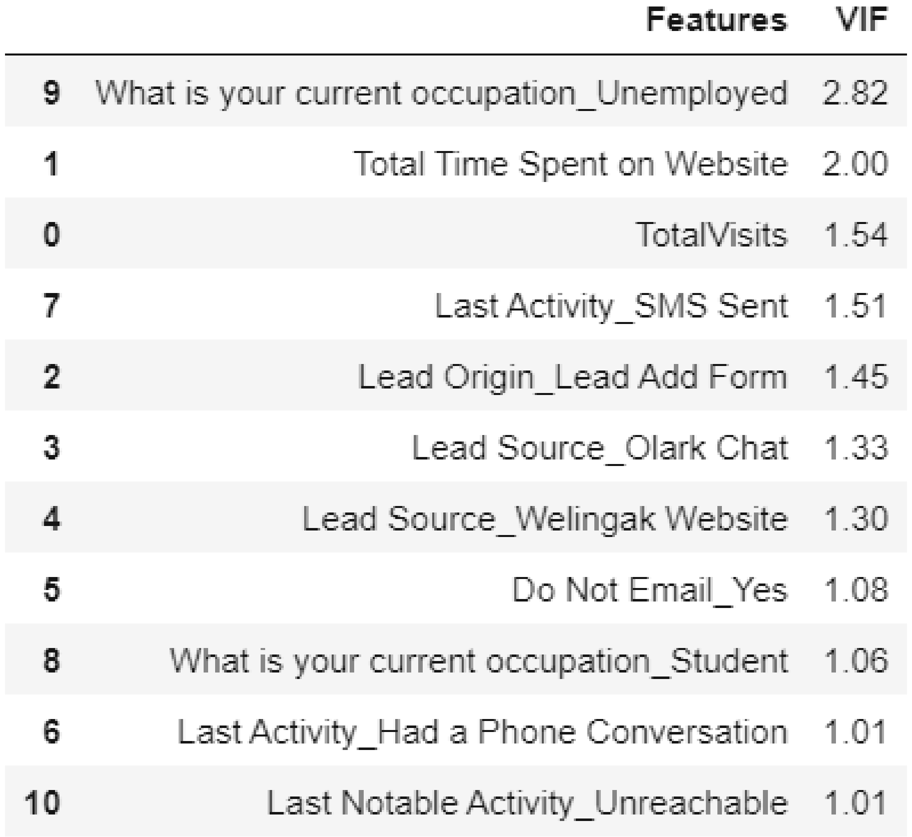
Dep. Variable:	Converted	**No. Observations:**	4461
Model:	GLM	**Df Residuals:**	4449
Model Family:	Binomial	**Df Model:**	11
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2079.1
Date:	Mon, 23 Nov 2020	**Deviance:**	4158.1
Time:	20:48:17	**Pearson chi2:**	4.80e+03
No. Iterations:	7		
Covariance Type:	nonrobust		



	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2040	0.196	1.043	0.297	-0.179	0.587
TotalVisits	11.1489	2.665	4.184	0.000	5.926	16.371
Total Time Spent on Website	4.4223	0.185	23.899	0.000	4.060	4.785
Lead Origin_Lead Add Form	4.2051	0.258	16.275	0.000	3.699	4.712
Lead Source_Olark Chat	1.4526	0.122	11.934	0.000	1.214	1.691
Lead Source_Welingak Website	2.1526	1.037	2.076	0.038	0.121	4.185
Do Not Email_Yes	-1.5037	0.193	-7.774	0.000	-1.883	-1.125
Last Activity_Had a Phone Conversation	2.7552	0.802	3.438	0.001	1.184	4.326
Last Activity_SMS Sent	1.1856	0.082	14.421	0.000	1.024	1.347
What is your current occupation_Student	-2.3578	0.281	-8.392	0.000	-2.908	-1.807
What is your current occupation_Unemployed	-2.5445	0.186	-13.699	0.000	-2.908	-2.180
Last Notable Activity_Unreachable	2.7846	0.807	3.449	0.001	1.202	4.367

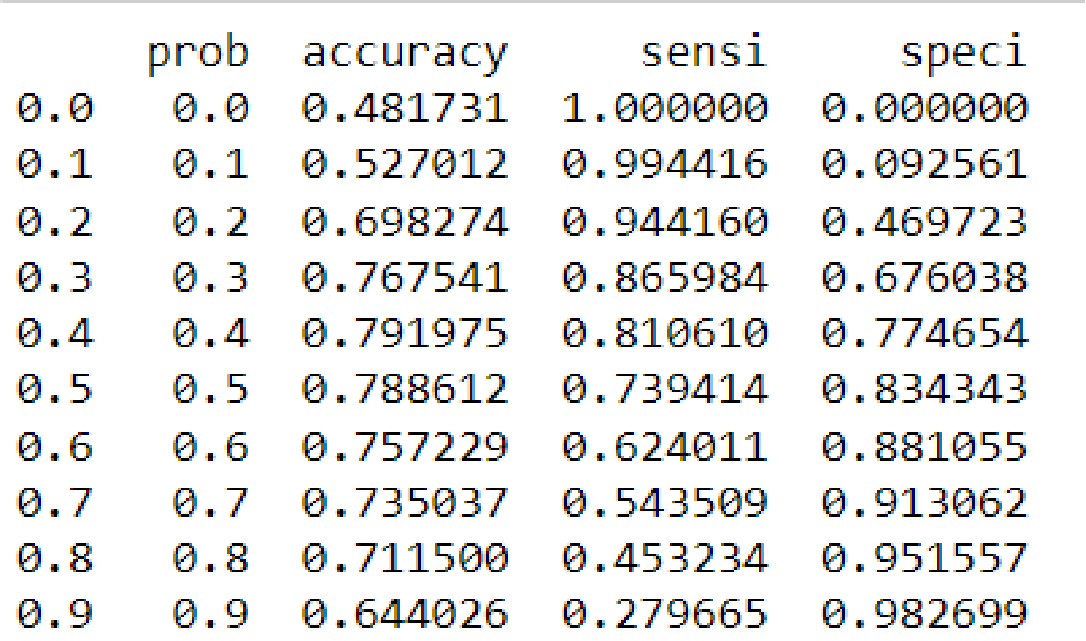
Variance Influence Factor (V.I.F)

V.I.F Values for the final model :



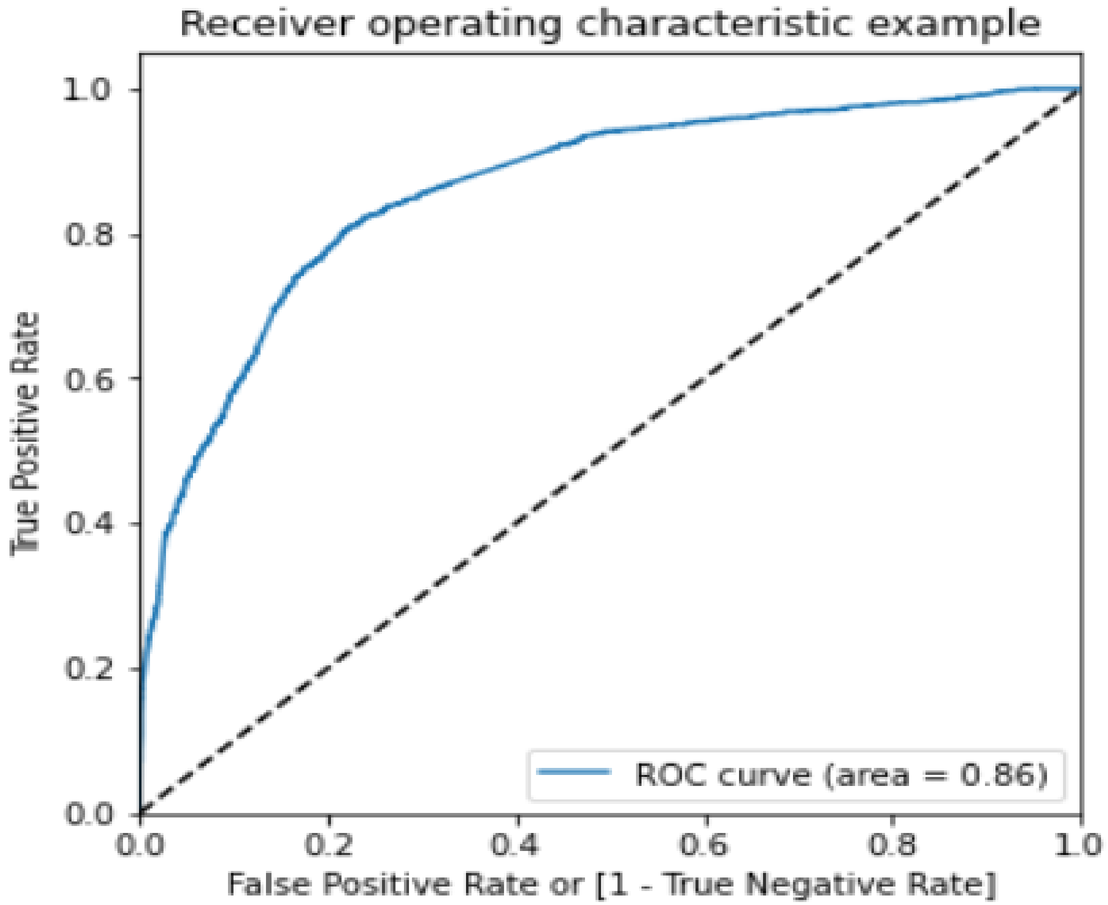
	Features	VIF
9	What is your current occupation_Unemployed	2.82
1	Total Time Spent on Website	2.00
0	TotalVisits	1.54
7	Last Activity_SMS Sent	1.51
2	Lead Origin_Lead Add Form	1.45
3	Lead Source_Olark Chat	1.33
4	Lead Source_Welingak Website	1.30
5	Do Not Email_Yes	1.08
8	What is your current occupation_Student	1.06
6	Last Activity_Had a Phone Conversation	1.01
10	Last Notable Activity_Unreachable	1.01

MODEL EVALUATION

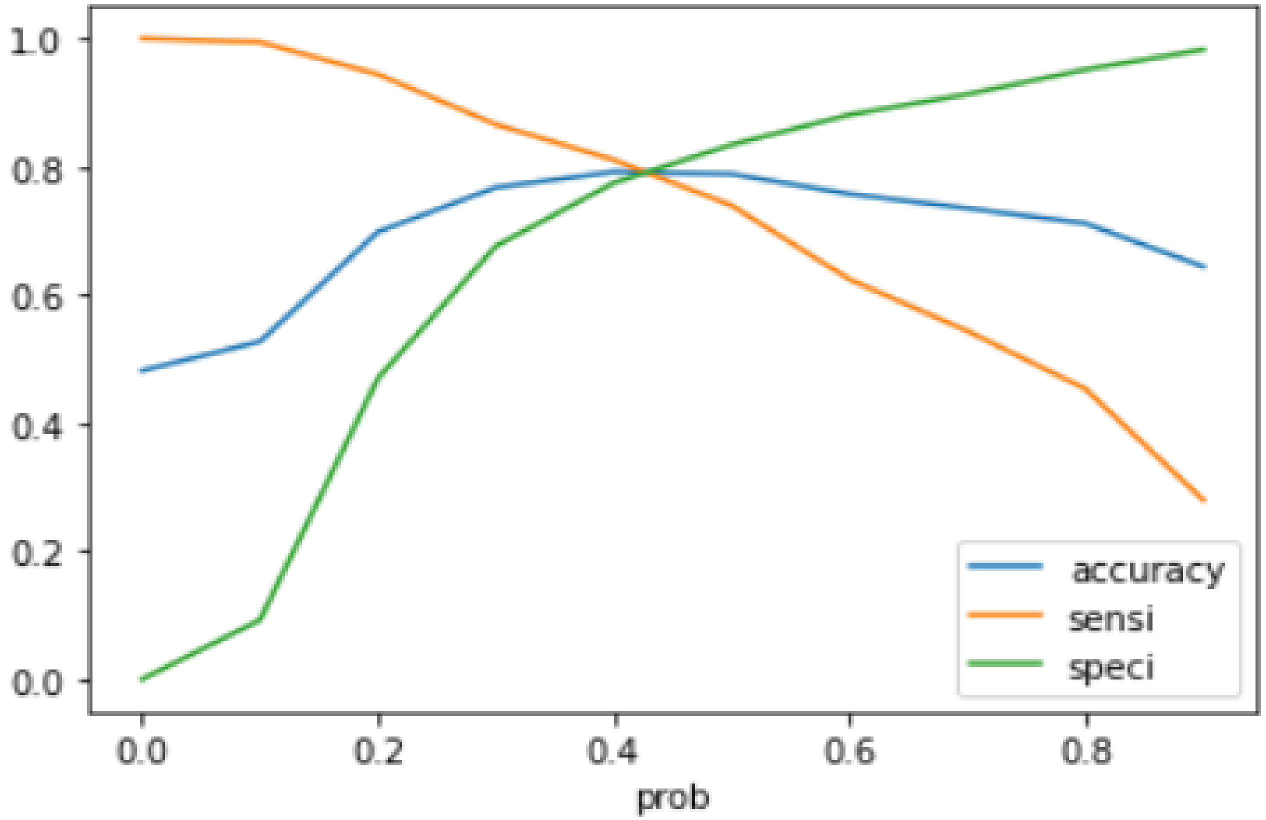


```
     prob  accuracy     sensi     speci
0.0   0.0  0.481731  1.000000  0.000000
0.1   0.1  0.527012  0.994416  0.092561
0.2   0.2  0.698274  0.944160  0.469723
0.3   0.3  0.767541  0.865984  0.676038
0.4   0.4  0.791975  0.810610  0.774654
0.5   0.5  0.788612  0.739414  0.834343
0.6   0.6  0.757229  0.624011  0.881055
0.7   0.7  0.735037  0.543509  0.913062
0.8   0.8  0.711500  0.453234  0.951557
0.9   0.9  0.644026  0.279665  0.982699
```

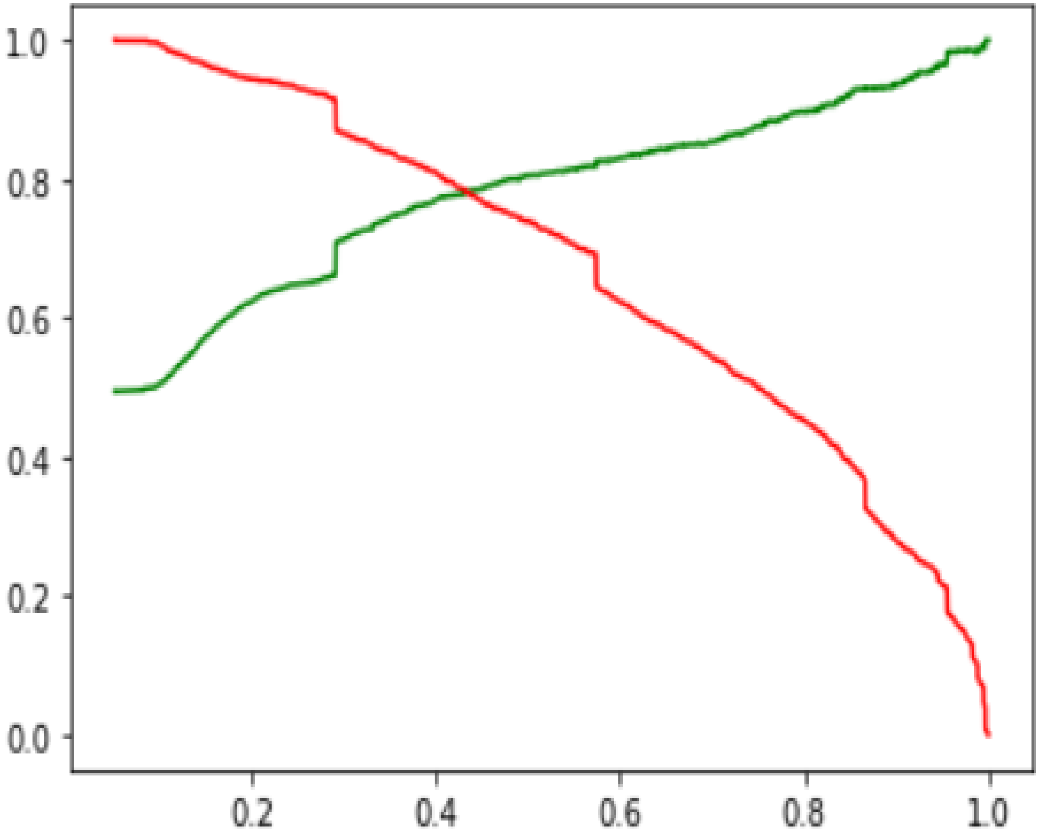
ROC Curve


Receiver operating characteristic example
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate or [1 - True Negative Rate]
ROC curve (area = 0.86)

Optimal Cut-off point – (0.4)



Probability



INFERENCES

Top three variables in your model which contribute most towards the probability of a lead getting converted :

a. Total Visits,
b. Total Time Spent on Website,
c. Lead Origin_Lead Add Form

Top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion :

a. Lead Origin_Lead Add Form
b. Last Activity_Had a Phone Conversation
c. Lead Source_Welingak Website

RECOMMENDATION

Scenario 1:
So when the company has more interns we need have lower cutoff threshold so that our model can predict almost all leads. The flip side to this decrease in threshold will be that we will misclassify some non-conversions as conversions but this is a good tradeoff given we have mode manpower to deal with it.

Scenario 2:
Typically, when the company has less people to call potential customers so its good to have more accurate predictions in which case the model specificity should be much more higher. This would mean form the above graph the we would have to choose a cutoff point which is much higher. The tradeoff of this is that we are going to miss some leads but given that the company has less manpower who can focus more on correctly predicted leads.

Scenario 3:
The company should focus on sending automated SMS and emails to potential leads during the time they have less manpower which allows for cost effective lead conversion without manual intervention.

THANK YOU