Lead Scoring Case Study

BY-

SWATI KUMAR

SAMBIT SEKHAR SAHU

“Start by doing what’s necessary, then what’s possible and suddenly you are doing the impossible”

- **SAINT FRANCIS**

PROBLEM STATEMENT

- An education company named X Education sells online courses to industry professionals & the typical lead conversion rate is around 30%. Although the company 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’. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

OBJECTIVE

- The objective is to help X Education select the most promising leads by building a model and assigning 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%.

Model Building → Predicting Hot Leads → Higher Conversion Rate

APPROACH OF ANALYSIS

- Data Reading & Understanding
- Data Cleaning
- Exploratory Data Analysis
- Data Preparation
- Model Building
- Model Evaluation

Data Reading & Understanding

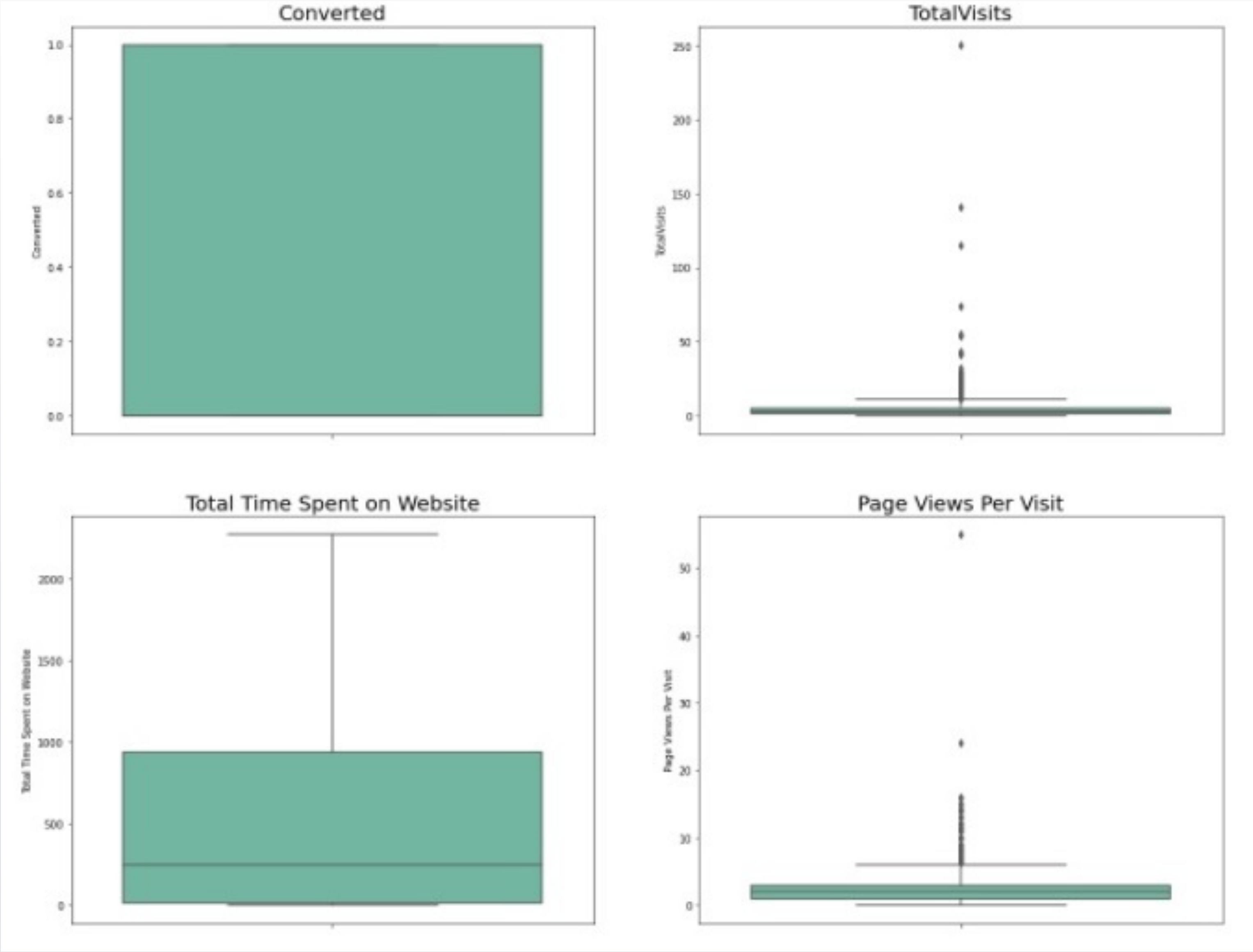
- Firstly, we imported the data & stored it into a data frame.
- Then we looked at its shape, size, info, & description to begin with.

Data Cleaning

- Those columns with null value percentages greater than 40% were removed.
- Categorical columns having only one category were removed, as they won’t be helpful for analysis.
- Columns having highly skewed data were also removed, to counter any possible data-imbalance.
- Columns having too many sub-category levels were treated so that when dummy variables are created, we don’t have too many dummy variables to deal with. Sub-category levels having very low frequency were clubbed into one group.
- Apart from column wise null value analysis, row wise null value analysis was also done. Rows having greater than 50% null value presence were removed (if any).
- Finally proper null value imputation technique was used.

Outlier Treatment

- 'TotalVisits' and 'Page Views Per Visit' had outliers and were treated.
- Here is the attached box plot of the columns after outlier treatment is done.

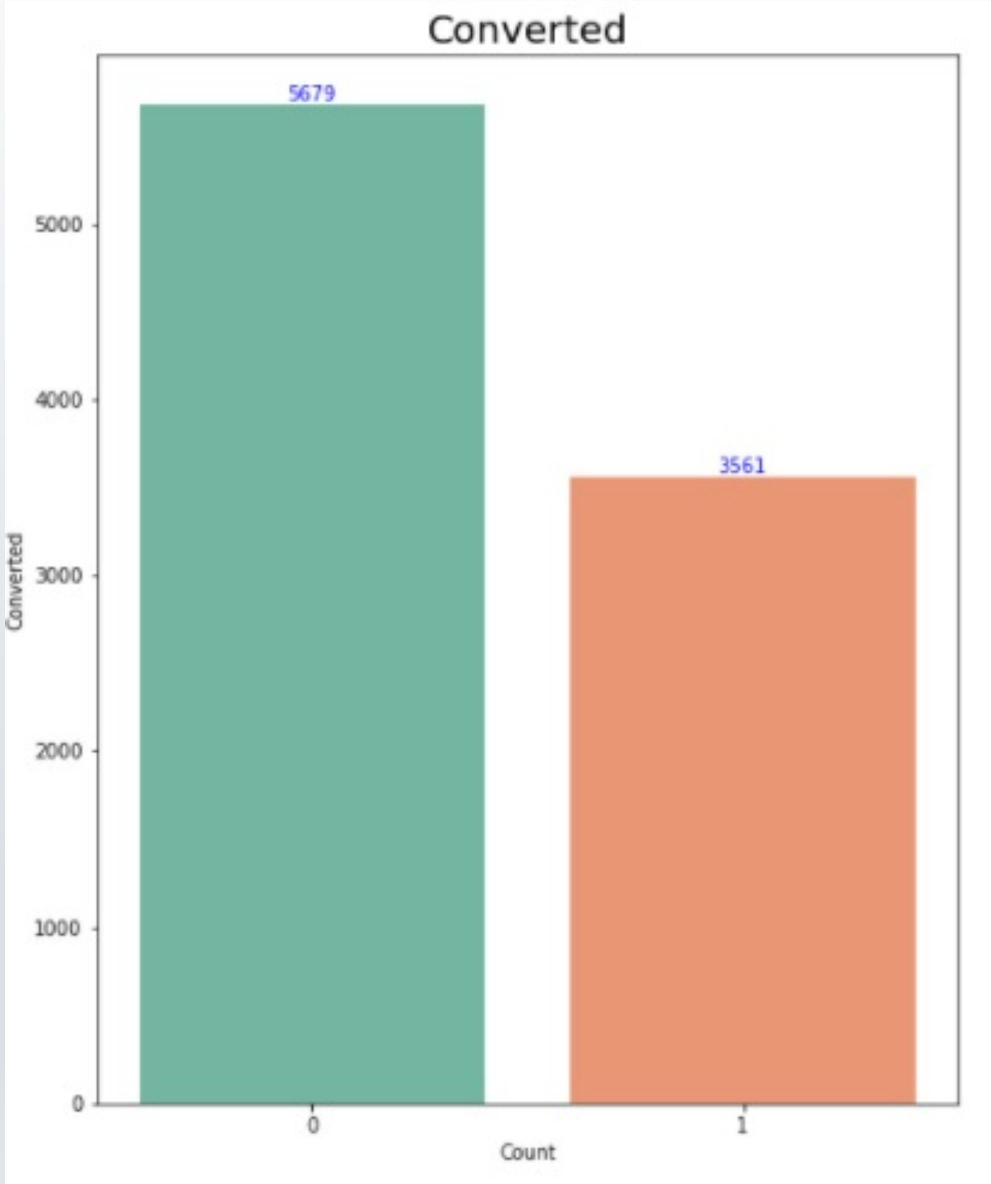


Exploratory Data Analysis

- After data cleaning we proceeded with EDA where we took a closer look at different variables to observe features in detail.
- We found out sales team specific generated data won't be helpful in analysis as the model being built will be used by sales team before hand to identify Hot leads. So, we removed those columns (if any present till EDA) before moving onto Data Preparation.

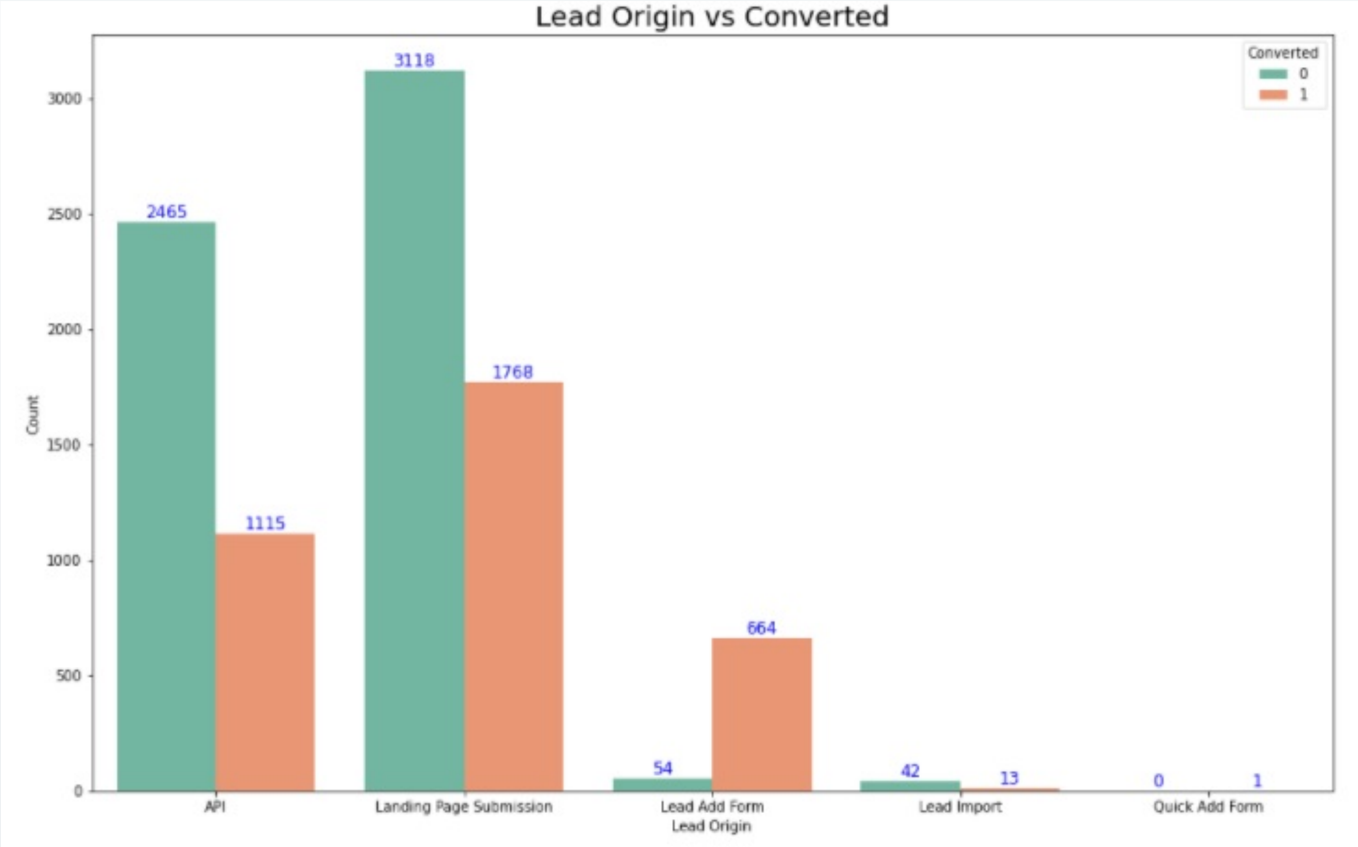
Conversion rate from previous data

- From the graph we can observe the conversion rate to be 38.5%



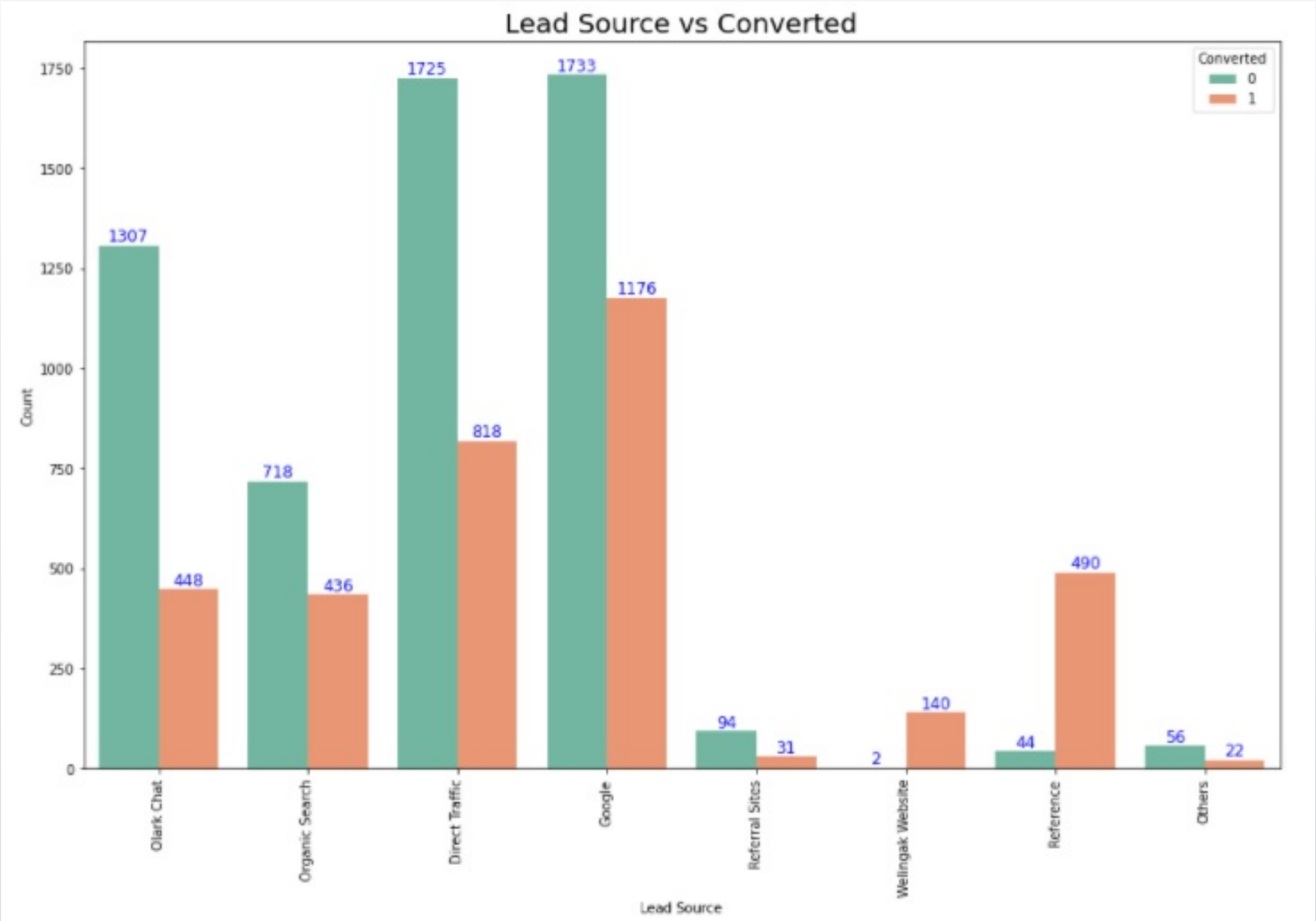
Lead Origin vs Converted

- From the graph we can see that the maximum conversion happened by landing page submission.



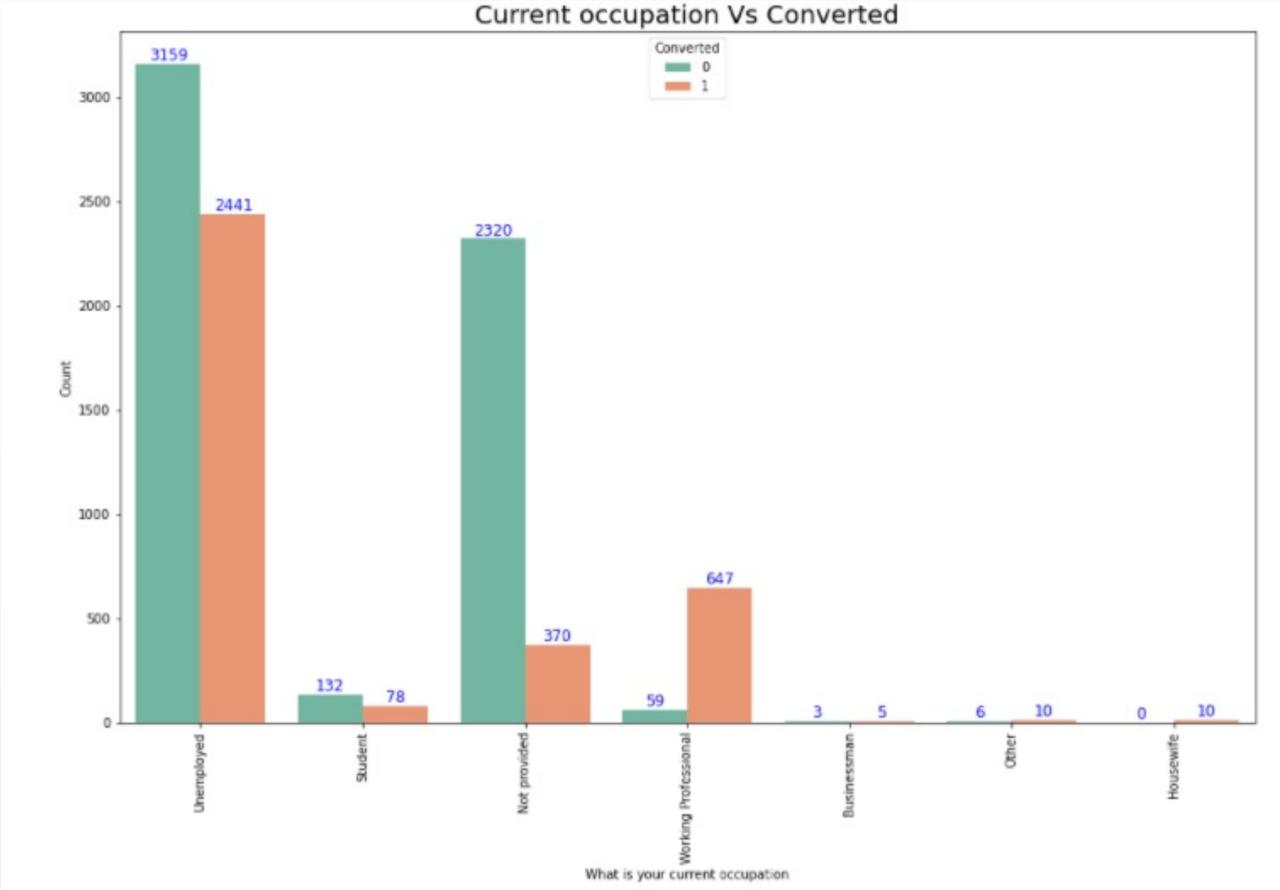
Lead Source vs Converted

- From the graph we can see Google contributed the most in lead conversion



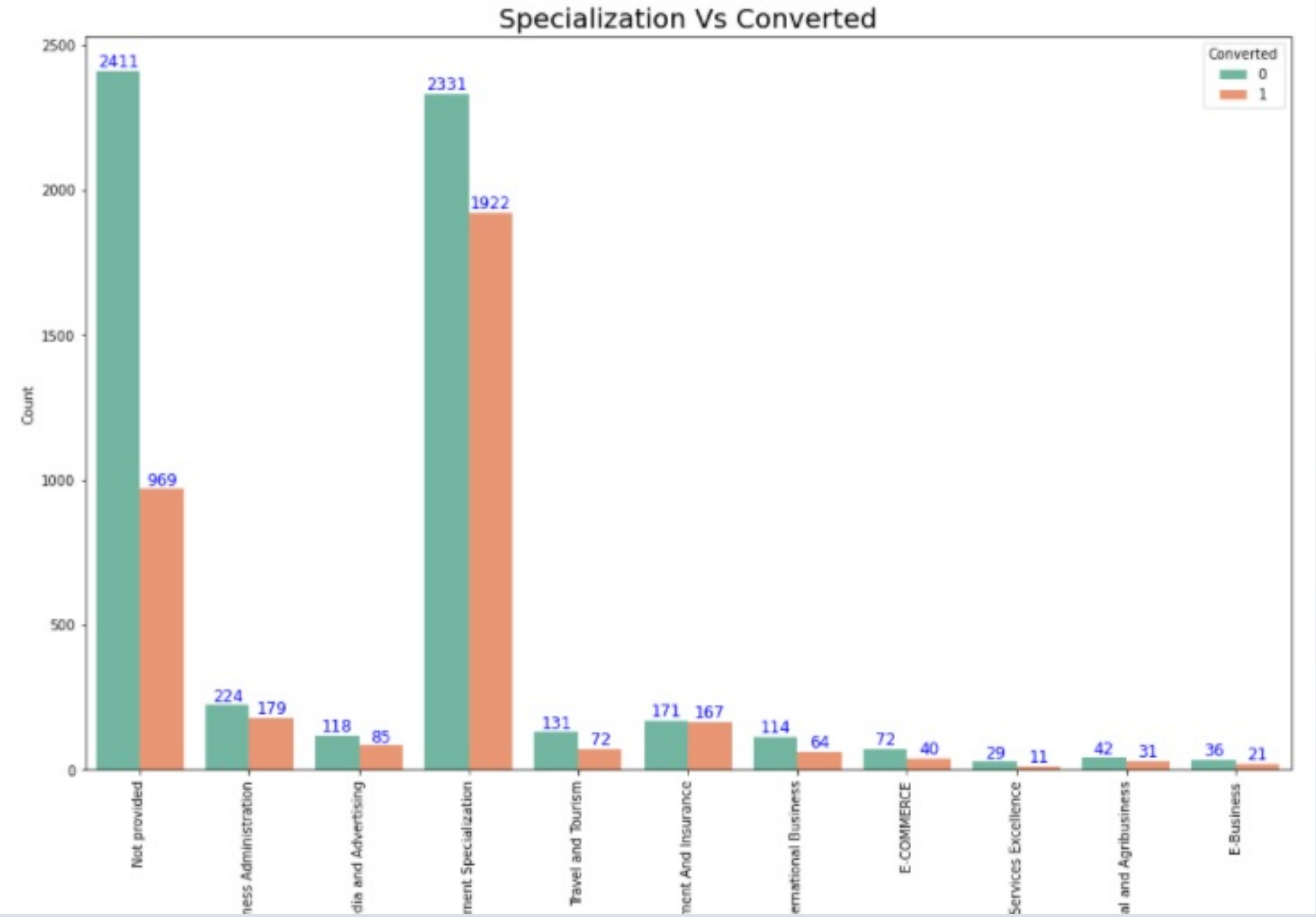
Current Occupation vs Converted

- As we can see, more conversions happened for unemployed.
- Also 10 housewives applied and all got converted.



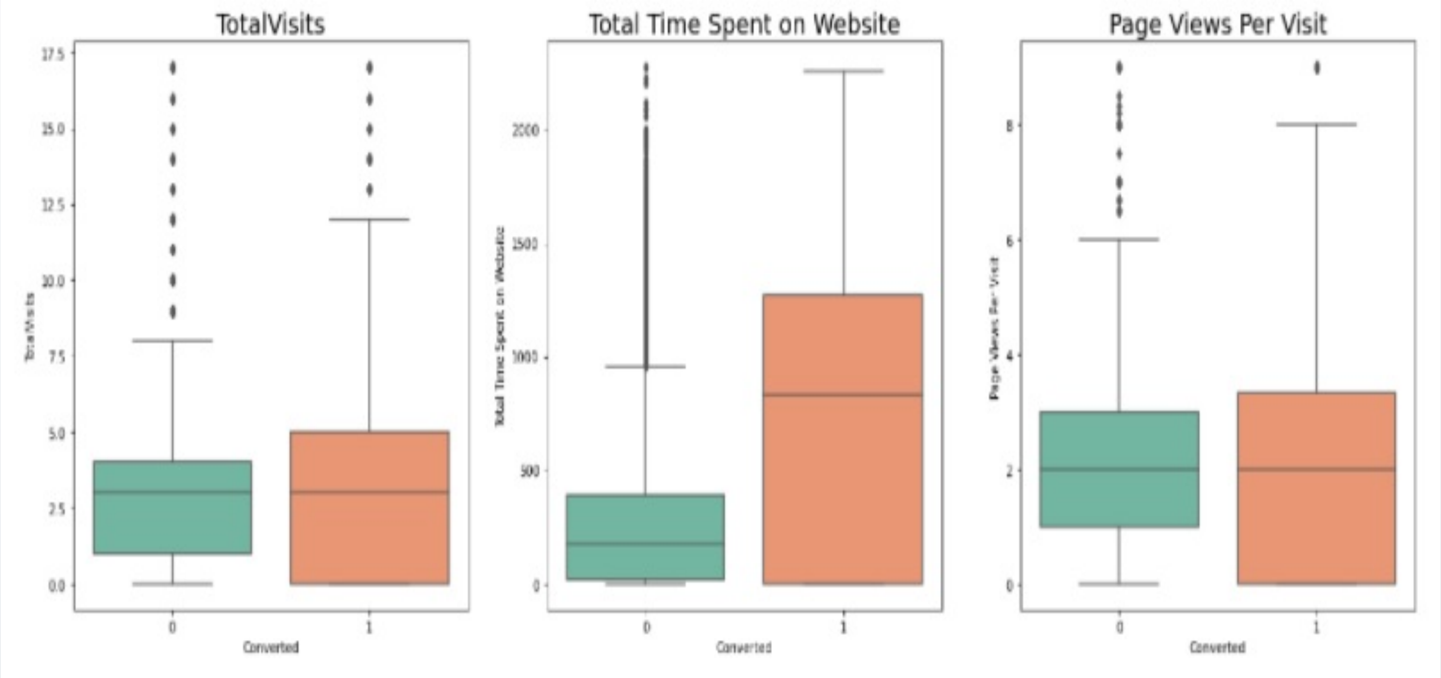
Specialization vs Converted

- As we can see, the most conversions happened when specialization is 'Management Specialization'.



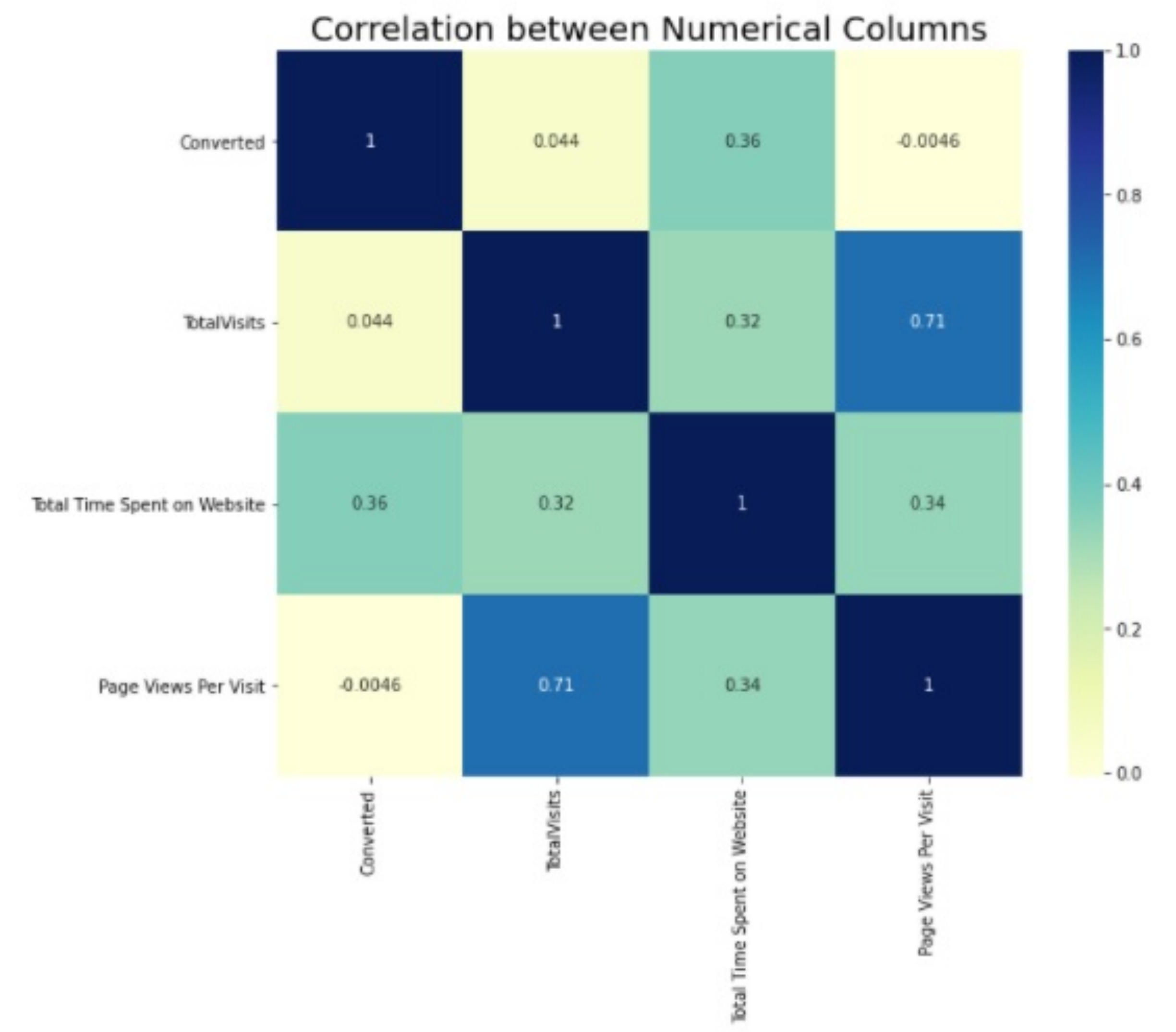
Numerical Features vs Converted

- Median for converted and not converted leads is almost same in the case of 'TotalVisits' and 'Page Views Per Visit'.
- It seems leads spending more time on the website have more chances of getting converted.



Correlation between Numerical Features

- Correlation of 'TotalVisits' and 'Page Views Per Visit' is the highest.
- 'Converted' has a decent correlation with 'Total Time Spent on Website'.



Data Preparation

- Dummy variables were created for categorical columns having different sub-category levels.
- For numerical columns scaling was done using Standard Scaler to bring data into one scale.
- Data was split into Train & Test sets with 70-30 ratio.

Model Building

- Balanced approach was used while building the model i.e., Automated (RFE) + Manual Feature Elimination (p-value & VIF).
- Firstly, RFE was used to attain top 20 features.
- Then p-value & VIF value were used with backward feature elimination method i.e., one feature was removed at a time until we reached optimum model with all features having p-value < 0.05 and VIF value < 3.

P-value & VIF value for Final features

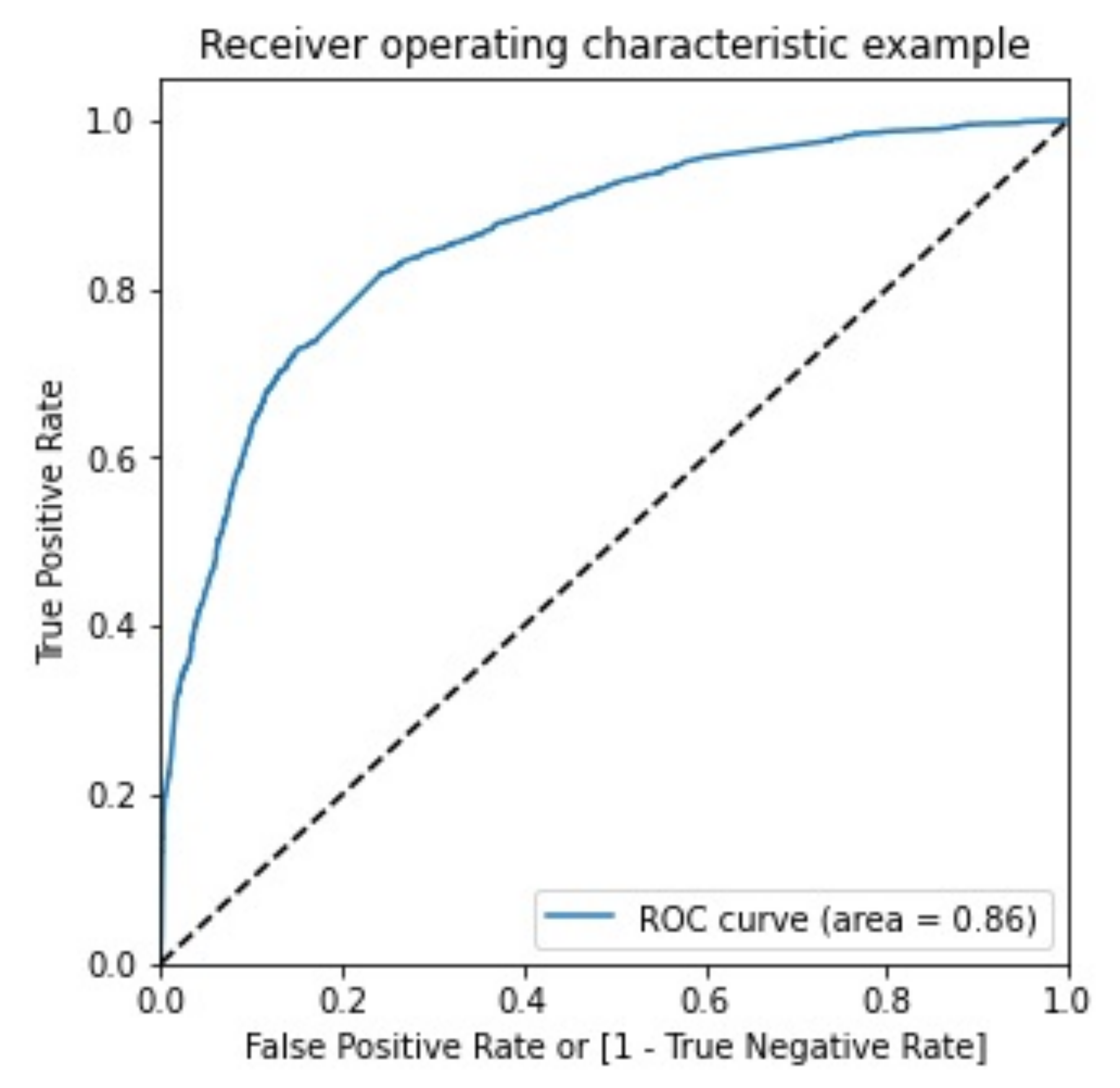
Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6459
Model Family:	Binomial	Df Model:	8
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2906.7
Date:	Sun, 13 Jun 2021	Deviance:	5813.4
Time:	18:16:20	Pearson chi2:	7.40e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.9833	0.078	-25.430	0.000	-2.136	-1.830
Do Not Email	-1.3416	0.159	-8.421	0.000	-1.654	-1.029
Total Time Spent on Website	1.0973	0.038	28.913	0.000	1.023	1.172
Lead Origin_Lead Add Form	3.5239	0.185	19.067	0.000	3.162	3.886
Lead Source_Olark Chat	0.9598	0.095	10.154	0.000	0.775	1.145
Lead Source_Welingak Website	2.0741	0.741	2.800	0.005	0.622	3.526
CurrentOccupation_Student	1.0607	0.221	4.795	0.000	0.627	1.494
CurrentOccupation_Unemployed	1.2166	0.081	14.955	0.000	1.057	1.376
CurrentOccupation_Working Professional	3.7467	0.191	19.668	0.000	3.373	4.120

	Features	VIF
2	Lead Origin_Lead Add Form	1.47
3	Lead Source_Olark Chat	1.28
6	CurrentOccupation_Unemployed	1.26
1	Total Time Spent on Website	1.24
4	Lead Source_Welingak Website	1.24
7	CurrentOccupation_Working Professional	1.14
0	Do Not Email	1.05
5	CurrentOccupation_Student	1.02

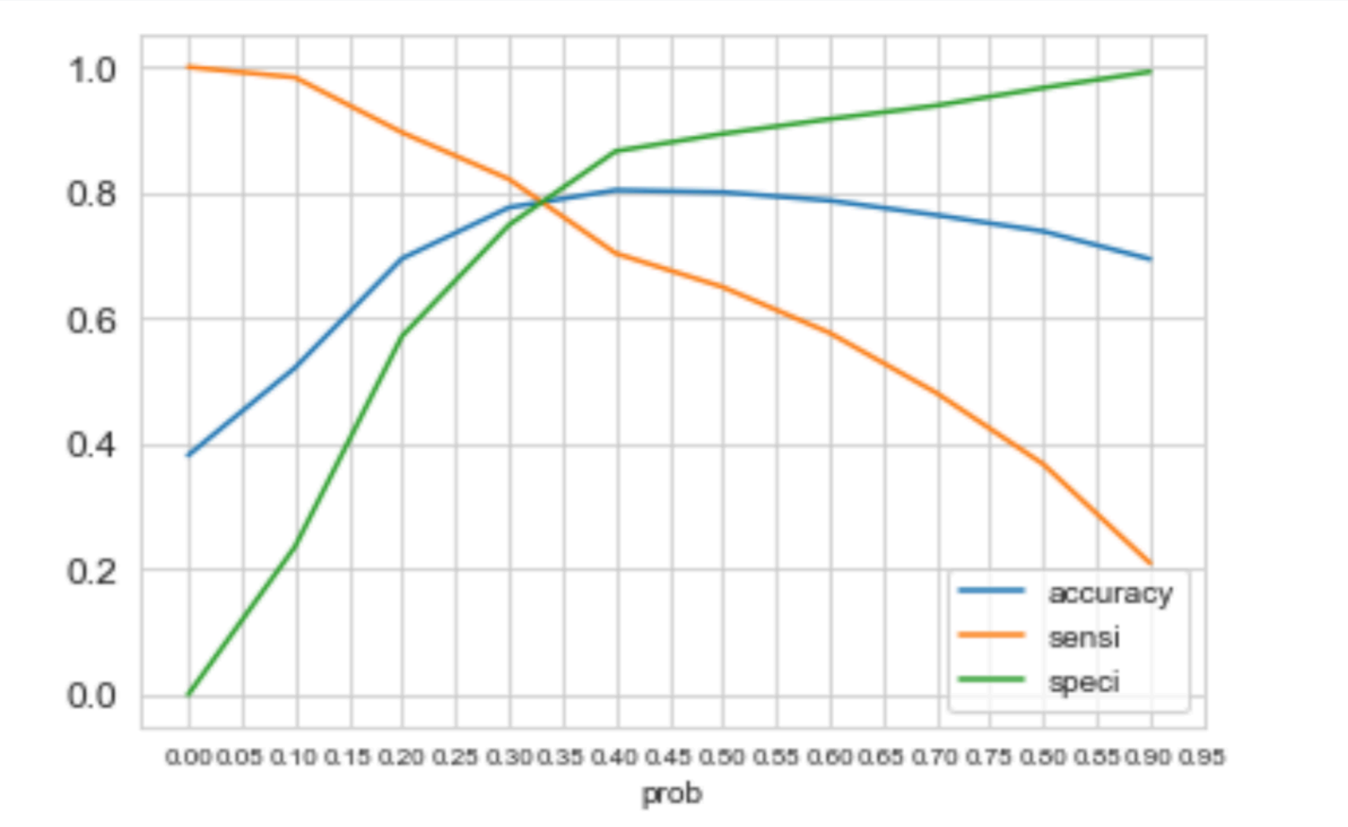
ROC Curve



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

Finding Optimal Cut-off

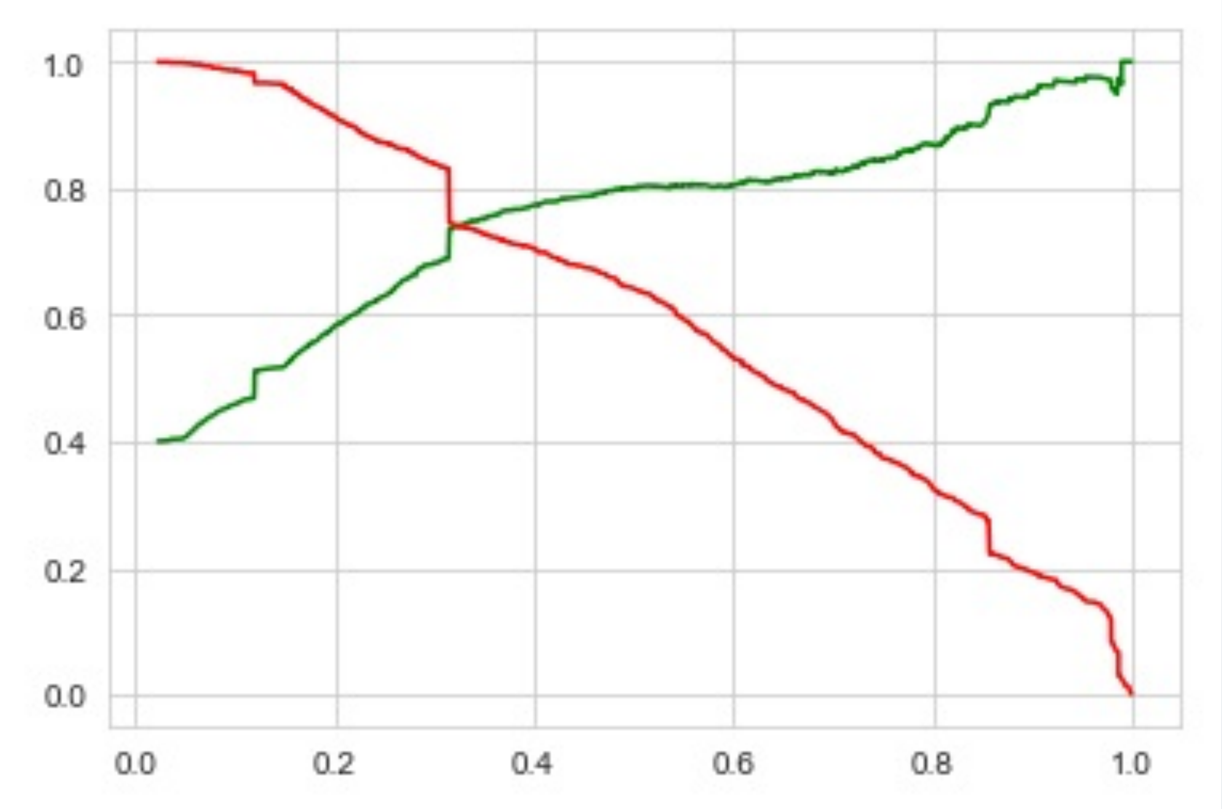
- From the graph, the point where accuracy, sensitivity & specificity intersect gives the optimal cut-off.
- 0.314 is the optimal cut-off.



Model Evaluation

- To check the model’s performance, the model was trained on train set & was then tested on test set.
- Parameters like Sensitivity, Specificity, Accuracy, Precision & Recall were used for evaluating model performance.

Plotting to Observe trade-off between Precision & Recall



Result

Train Set

- Sensitivity: 81.71%
- Specificity: 75.78%
- Accuracy: 78.04%
- Precision: 78.99%
- Recall: 64.69%

Test Set

- Sensitivity: 83.1%
- Specificity: 75.43%
- Accuracy: 78.46%
- Precision: 68.83%
- Recall: 83.1%

Conclusion

Top 3 Variables that contributed the most towards the probability of a lead getting converted-

1. CurrentOccupation_Working Professional
2. Lead Origin_Lead Add Form
3. Lead Source_Welingak Website

Lead conversion rate: **83.10%**