Lead Scoring – A Case Study

Submitted By:

Md Liyakat
Bhanu Pratap
Nidhi Tripathi
Shadab Hussain

Abstract

- An education company named “X Education” sells online courses to industry professionals
- Interested Professional visit their website and browsing the courses and/ or fill up a form for the course or watch some videos.
- Few of them fill up a form providing their email address or phone number, they are classified to be a lead.
- The company also gets leads through past referrals.
- Once these leads are acquired, employees from the sales team start making calls, writing emails, etc.
- Through this process, some of the leads get converted while most do not. The typical lead conversion rate of X education is around 30%.
- Overall there are a lot of leads generated in the initial stage but only a few of them come out as paying customers from the bottom.

Business Line Objective

- To select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- To build a model wherein score can be assigned to each lead 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%.

Data Source

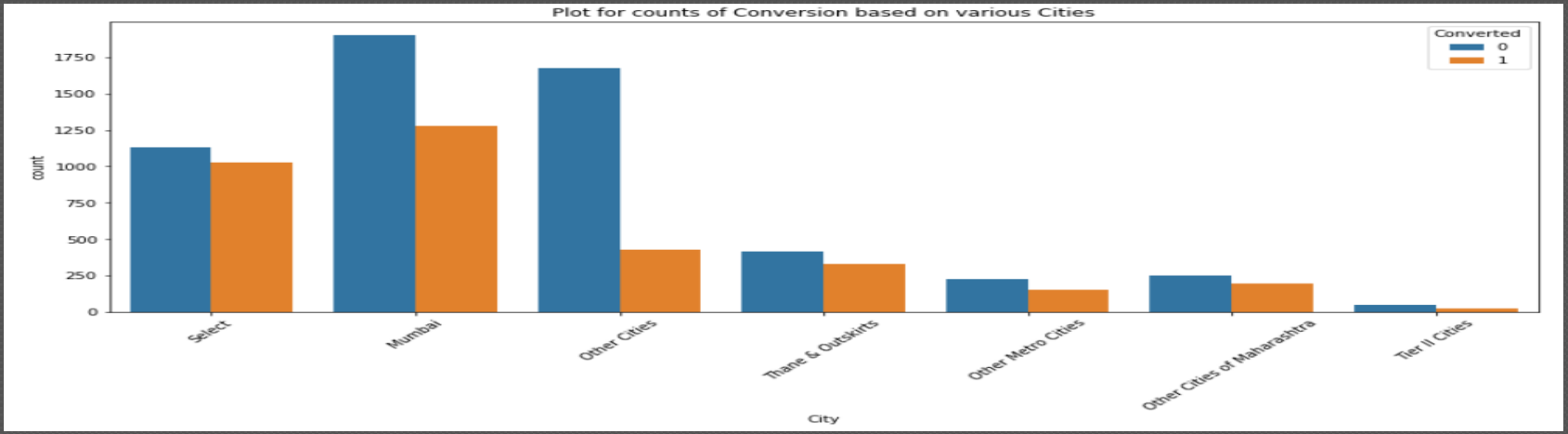
- For this case study we analyzed dataset: Leads.csv
- A leads dataset from the past with around 9000 data points.
- This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc.
- The target variable, is the column ‘Converted’ which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn’t converted.

Analysis Approach

Problem Solving Methodology:

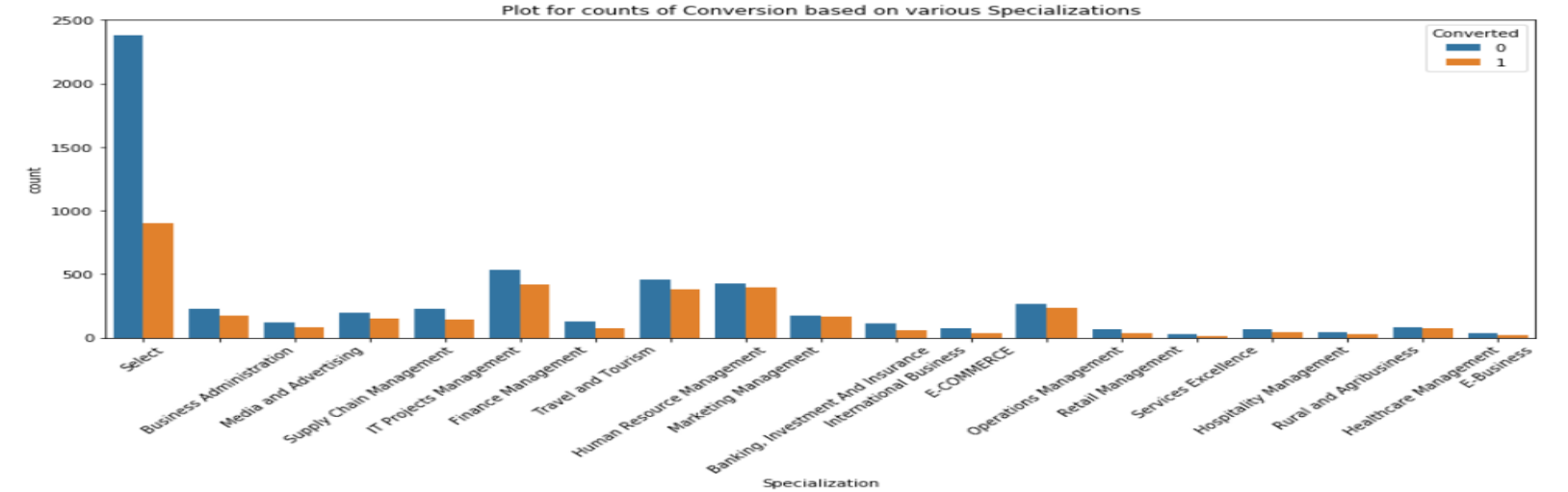
- To analyze the data, we have followed the below steps:
  - Importing the Dataset.
  - Perform Data cleaning.
  - Check the Null Values & fill with ‘other’.
  - Do the Exploratory Data Analysis.
  - Check the Outliers and do the Outlier treatment.
  - Visualize the data using the matplotlib , seaborn libraries and correlation matrix.
  - Create the dummy Variables.
  - Do the Test Train Split.
  - Do the Feature scaling.
  - Perform the RFE Analysis on Data.
  - Perform the VIF analysis.
  - Plotting the ROC Curve.
  - Finding Optimal Cutoff Point for a desired Precision
  - Making predictions on the test set.

Exploratory Data Analysis



- The above bar graph is visualizing the lead conversion.
- This is a plot for counts of Conversion based on various Cities.
- 1 means it was converted and 0 means it wasn't converted.
- From the above graph we can see that the Mumbai has the highest count of '0' and '1'.

Exploratory data Analysis:

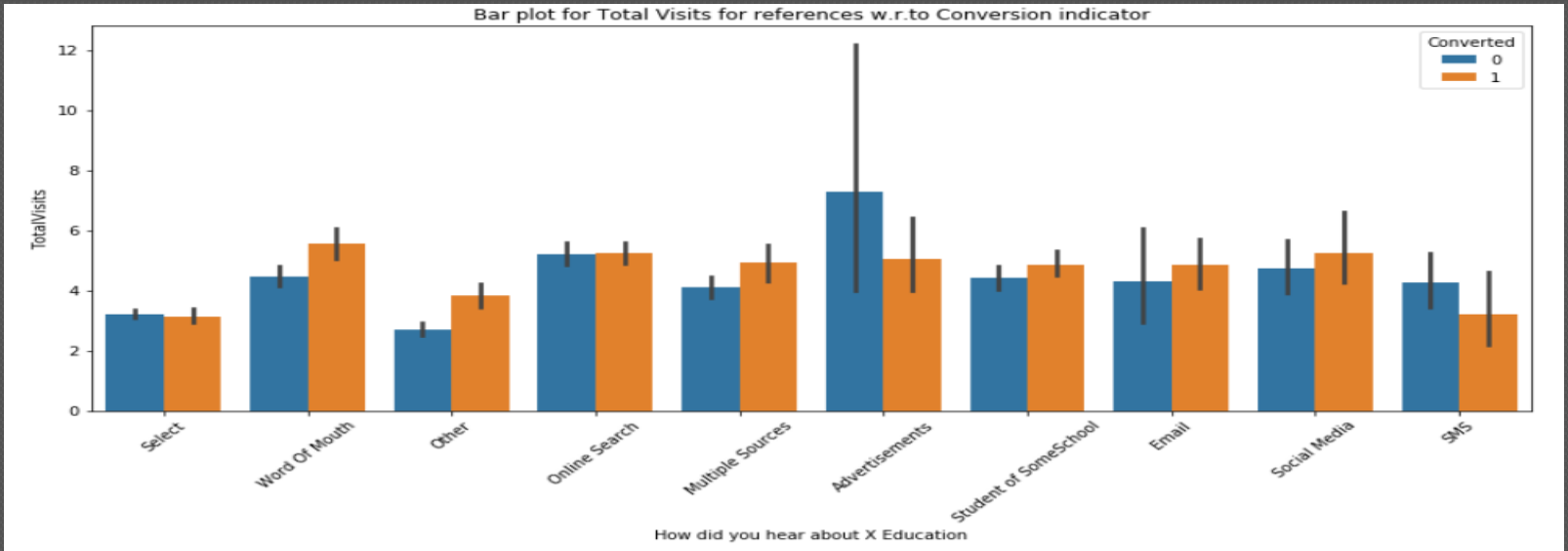


❖ This plot for counts of Conversion is based on different specialization professionals

❖From the above graph we can see that the Select has the highest count of ‘0’ and ‘1’.

❖Here, we can see that through the Finance Management having more lead.

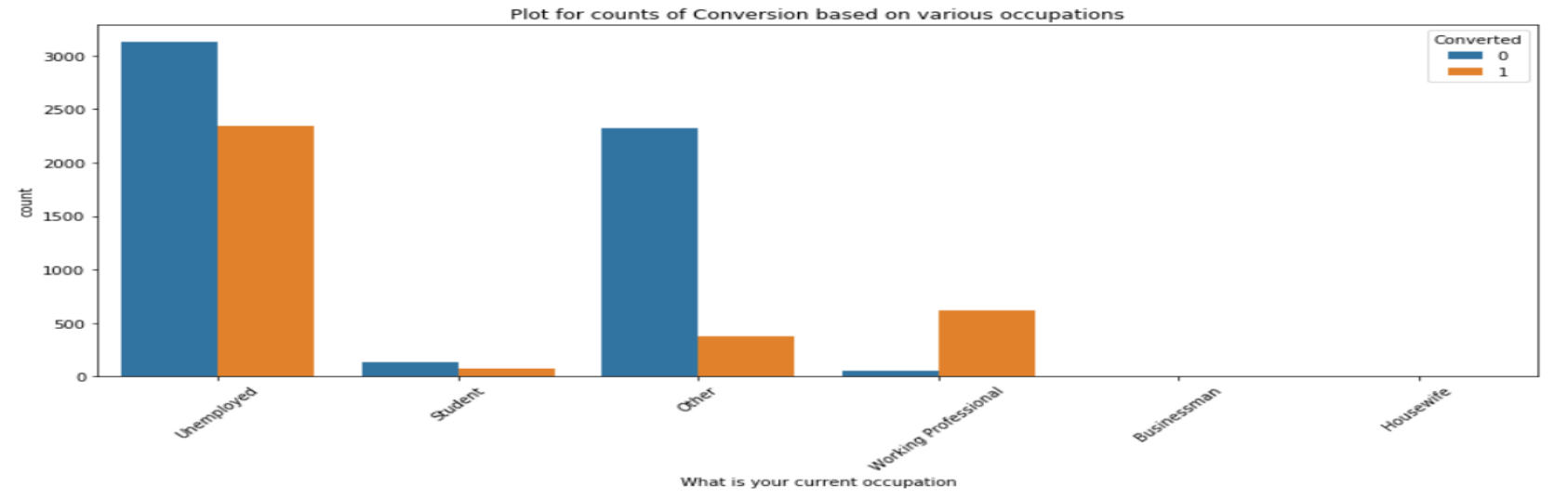
Exploratory data Analysis:



❖ This is a Bar plot for Total Visits for references w.r.to Conversion indicator.

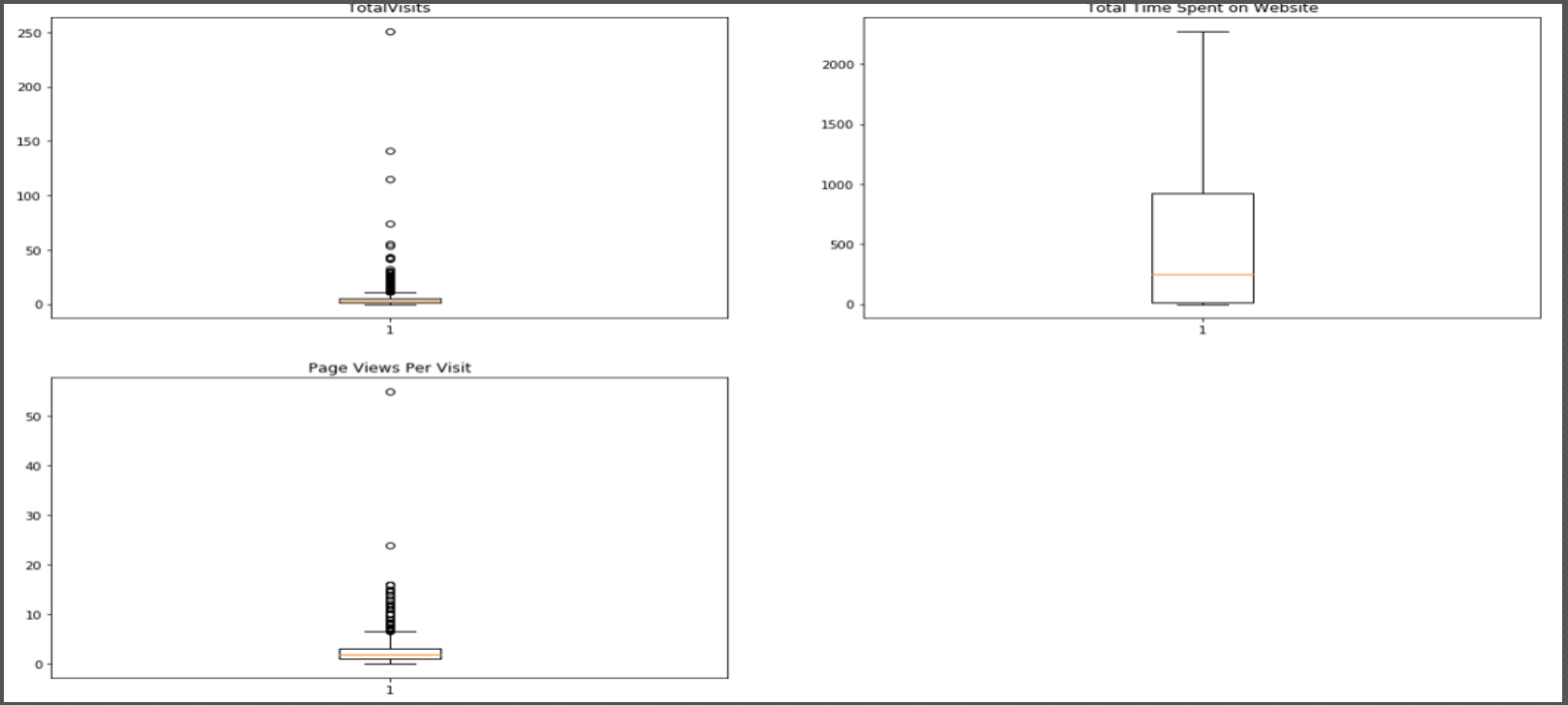
❖ Here, we can see that through the adervertisements having more lead.

Exploratory data Analysis:



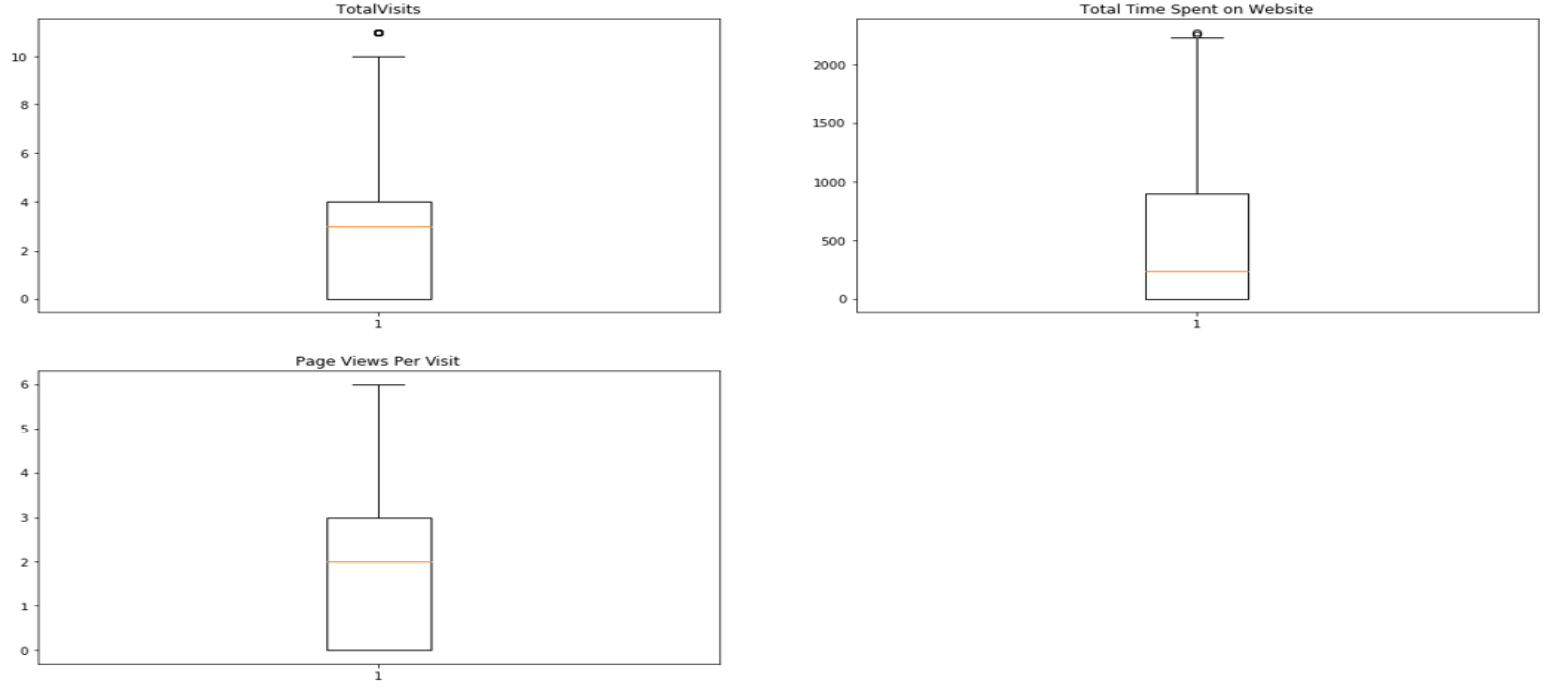
- This is a Bar plot for Occupation wise conversion counts.
- Here, we can see that through unemployed section having more conversions.

Outliers:



❖ There are clearly some outliers in the columns ‘TotalVisits’ and ‘Page Views Per Visit’.

After Outliers Treatment:



❖Here , we can see that our data is free of outliers.

Model Building:

- For model Building , first do the test train split and feature scaling.
- For feature selection we have used the RFE Model.
- In model 1, the column Last Activity Approached upfront is insignificant as p value is greater than 0.05.
- In model 2, the column Lead Profile Dual Specialization Student is insignificant as p value is greater than 0.05.
- In model 3, the column Lead Profile Lateral Student is insignificant as p value is greater than 0.05.
- In model 4, the column What is your current occupation Housewife is insignificant as p value is greater than 0.05.
- In model 5, the column Last Activity Had a Phone Conversation is insignificant as p value is greater than 0.05.
- In model 6, the column Last Notable Activity Unreachable is insignificant as p value is greater than 0.05.
- In model 7, all the features in the model are significant as p value is greater than 0.05.
- Drop the column from the model.

Final Model:

- All the P-values are greater than 0.05.
- All the features in the model are significant.

Dep. Variable:	Converted	**No. Observations:**	5959
Model:	GLM	**Df Residuals:**	5944
Model Family:	Binomial	**Df Model:**	14
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2254.0
Date:	Sat, 02 Mar 2019	**Deviance:**	4507.9
Time:	16:41:33	**Pearson chi2:**	5.97e+03
No. Iterations:	7	**Covariance Type:**	nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.1663	0.096	-12.209	0.000	-1.353	-0.979
Total Time Spent on Website	1.0912	0.044	25.062	0.000	1.006	1.177
Lead Origin_Landing Page Submission	-0.3885	0.097	-3.990	0.000	-0.579	-0.198
Lead Origin_Lead Add Form	2.9352	0.245	11.999	0.000	2.456	3.415
Lead Source_Olark Chat	1.1376	0.129	8.795	0.000	0.884	1.391
Lead Source_Welingak Website	3.1363	1.035	3.031	0.002	1.108	5.165
Last Activity_Converted to Lead	-1.3035	0.235	-5.544	0.000	-1.764	-0.843
Last Activity_Email Bounced	-2.1838	0.394	-5.540	0.000	-2.956	-1.411
Last Activity_Olark Chat Conversation	-1.3774	0.175	-7.891	0.000	-1.719	-1.035
What is your current occupation_Other	-0.8991	0.095	-9.454	0.000	-1.086	-0.713
What is your current occupation_Working Professional	2.3394	0.203	11.513	0.000	1.941	2.738
Lead Profile_Potential Lead	1.5967	0.107	14.892	0.000	1.387	1.807
Lead Profile_Student of SomeSchool	-2.0221	0.424	-4.774	0.000	-2.852	-1.192
Last Notable Activity_Had a Phone Conversation	3.1678	1.207	2.625	0.009	0.802	5.533
Last Notable Activity_SMS Sent	1.4420	0.086	16.814	0.000	1.274	1.610

VIFs:

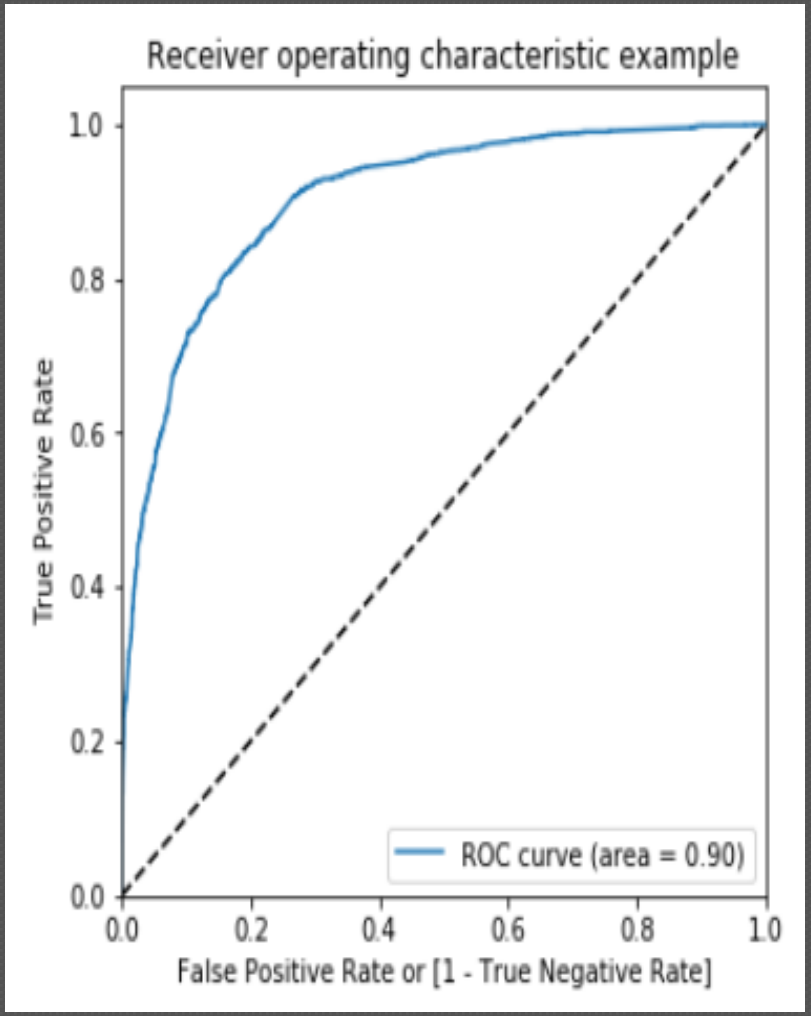
- All variables have a good value of VIF. So we need not drop any more variables and we can proceed with making predictions using this model.
- All variables have a good value of VIF.

	Features	VIF
3	Lead Source_Olark Chat	1.82
1	Lead Origin_Landing Page Submission	1.70
2	Lead Origin_Lead Add Form	1.70
8	What is your current occupation_Other	1.52
10	Lead Profile_Potential Lead	1.45
7	Last Activity_Olark Chat Conversation	1.43
13	Last Notable Activity_SMS Sent	1.39
0	Total Time Spent on Website	1.34
4	Lead Source_Welingak Website	1.34
9	What is your current occupation_Working Profes...	1.24
5	Last Activity_Converted to Lead	1.07
6	Last Activity_Email Bounced	1.06
11	Lead Profile_Student of SomeSchool	1.06
12	Last Notable Activity_Had a Phone Conversation	1.00

ROC Analysis:

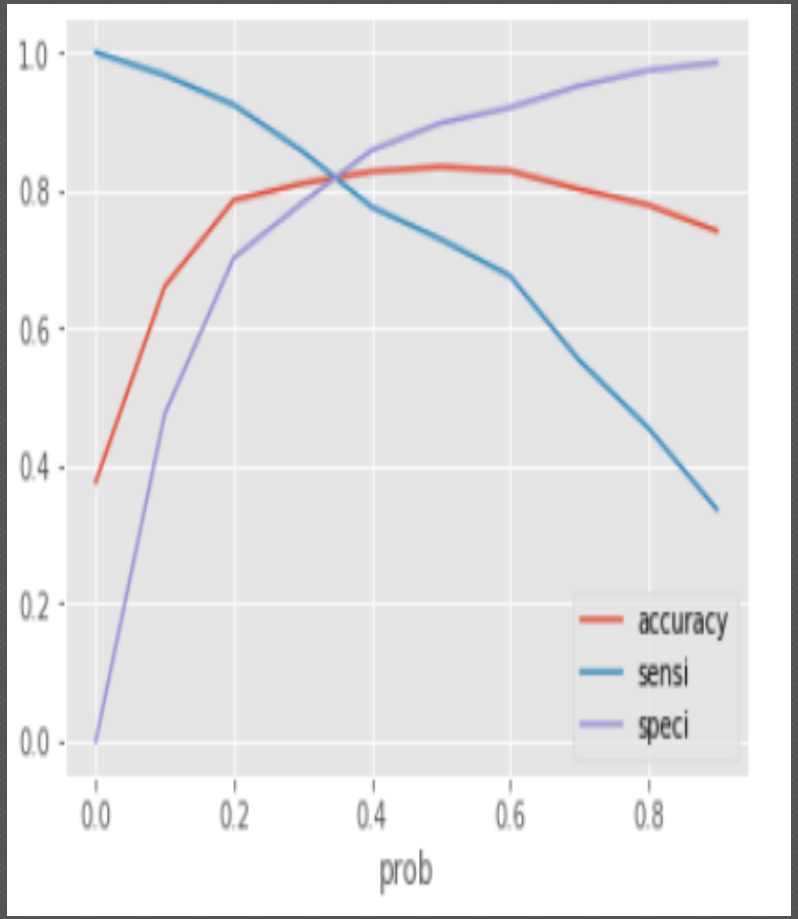
An ROC curve demonstrates several things:

- It shows the tradeoff between sensitivity and specificity.
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.



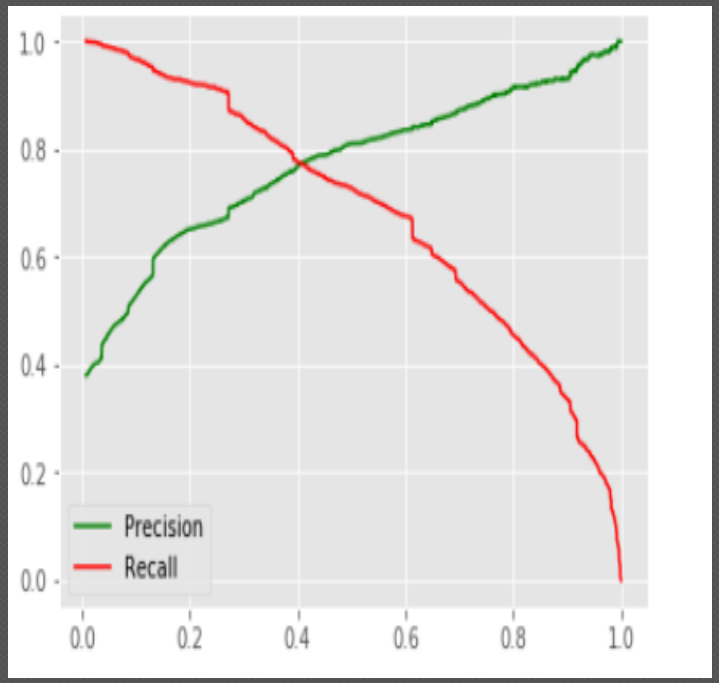
Optimal Cutoff Point:

- Optimal cutoff probability is that prob where we get balanced sensitivity and specificity
- From the curve above, 0.37 is the optimum point to take it as a cutoff probability.
- The optimum cutoff is 0.5 for required precision.



Precision and Recall:

	Train data	Test data
Precision	0.8110	0.89572
Recall	0.72860	0.72860.



Summary

	Train Data	Test Data
sensitivity	0.7286	0.6964
Specificity	0.8974	0.8957
Accuracy	0.8338	0.82067

- The false positive rate of our logistic regression is 0.10255.
- Positive predictive value of our logistic regression 0.8110.
- Negative predictive value of our logistic regression 0.8455.

Lead Scoring

	Prospect ID	Lead Number	Converted	Converted_Prob	Lead Score
0	7927b2df-8bba-4d29-b9a2-b6e0beafe620	660737	0	0.271887	27.188732
1	2a272436-5132-4136-86fa-dcc88c88f482	660728	0	0.314344	31.434409
2	8cc8c611-a219-4f35-ad23-fdfd2656bd8a	660727	1	0.894511	89.451144
3	0cc2df48-7cf4-4e39-9de9-19797f9b38cc	660719	0	0.129704	12.970353
4	3256f628-e534-4826-9d63-4a8b88782852	660681	1	0.274931	27.493085
5	2058ef08-2858-443e-a01f-a9237db2f5ce	660680	0	0.036913	3.691307
6	9fae7df4-169d-489b-afe4-0f3d752542ed	660673	1	0.913160	91.315994
7	20ef72a2-fb3b-45e0-924e-551c5fa59095	660664	0	0.036913	3.691307
8	cfa0128c-a0da-4656-9d47-0aa4e67bf690	660624	0	0.036653	3.665337

Conclusion:

- The top six variables which contribute most towards the probability of a lead getting converted are followings:
  1.) Last Notable Activity
  2.) Lead Source
  3.) Lead Origin
  4.) What is your current occupation
  5.) Lead Profile
  6.) Total Time Spent on the Website