LEAD SCORE CASE STUDY

Logistic regression model

By

Anilkumar M

17 May 2021

CONTENT

1. Problem statement
2. Approach
3. Attributes analysis
    1. Numerical
    2. Categorical
4. Model building
    1. Final model
    2. Model evaluation
    3. Optimal cutoff point
    4. Precision and recall
5. Prediction-Test set-inference
6. Recommendation

1. Problem statement

The company requires you 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%.

2. Approach

- **Data cleaning and data manipulation**

1. Check and handle duplicate data.
2. Check and handle NA values and missing values.
3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
4. Check and handle outliers in data.
5. Imputation of the values, if necessary.

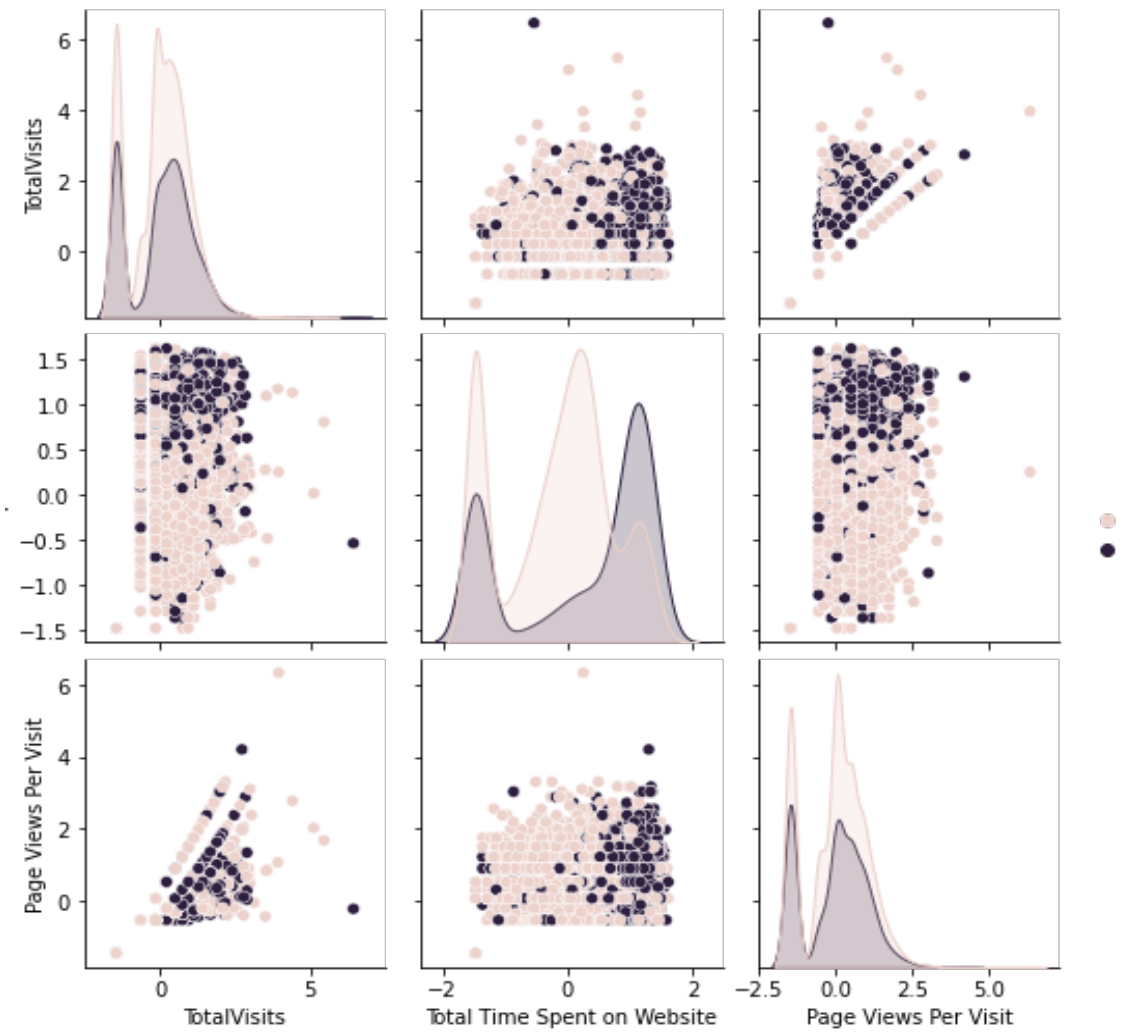
- **Exploratory Data Analysis**

1. Univariate data analysis: value count, distribution of variable etc.
2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.

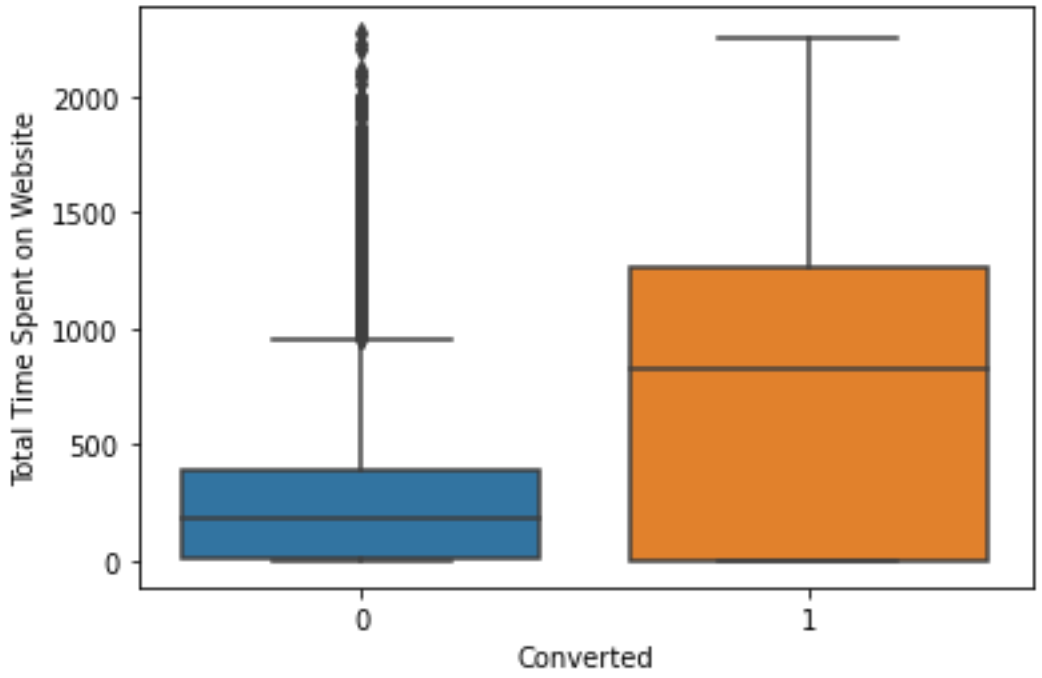
- **Feature Scaling & Dummy Variables and encoding of the data.**
- **Classification technique: logistic regression used for the model making and prediction.**
- **Validation of the model. Model presentation. Conclusions and recommendations.**

3.1 Attribute analysis-Numerical

- Total Visits and Page Views Per Visit seems to be a hot lead convertor variable

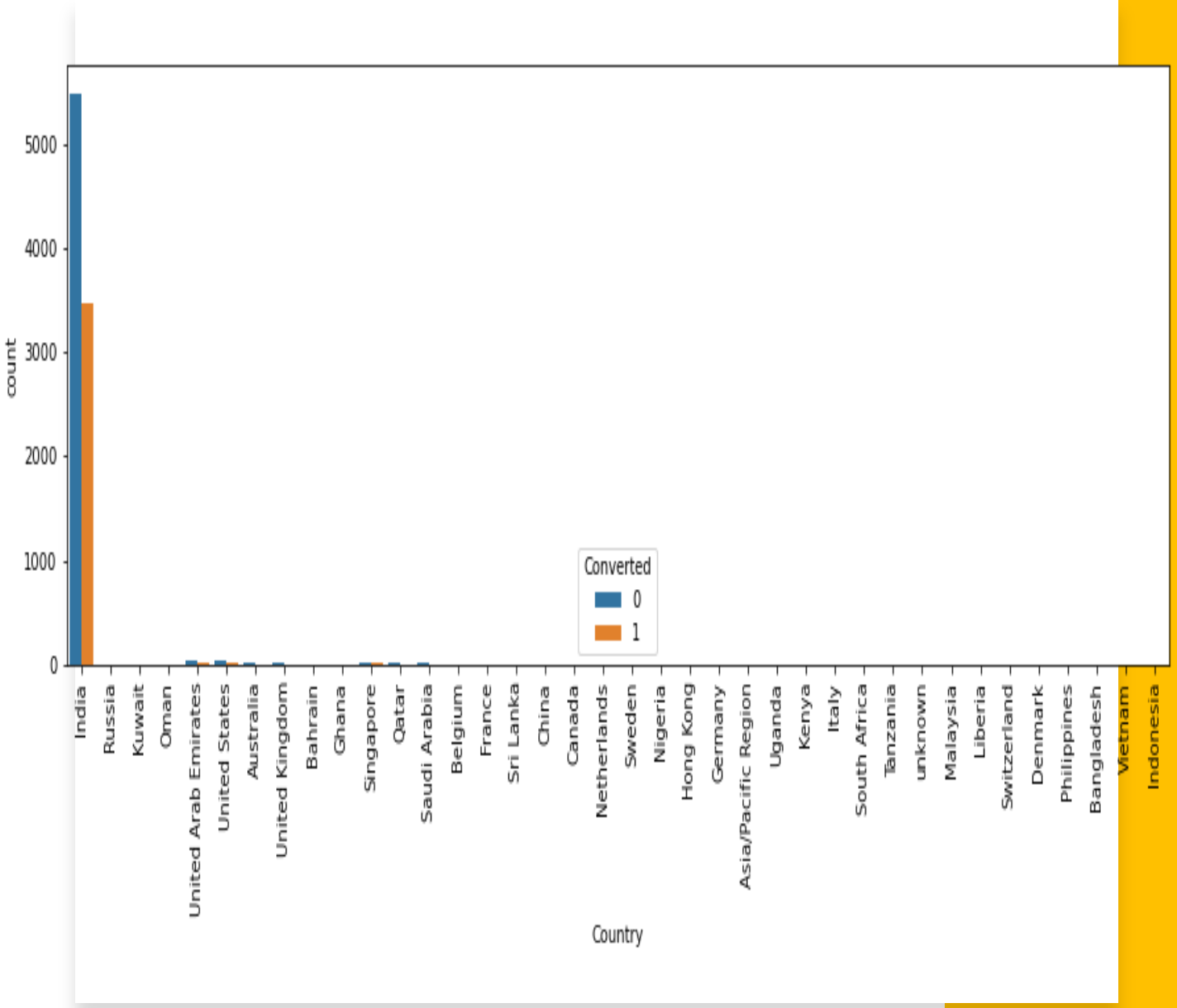


- Leads spending more time on the website are more likely to be converted.
- Website should be made more engaging to make leads spend more time.

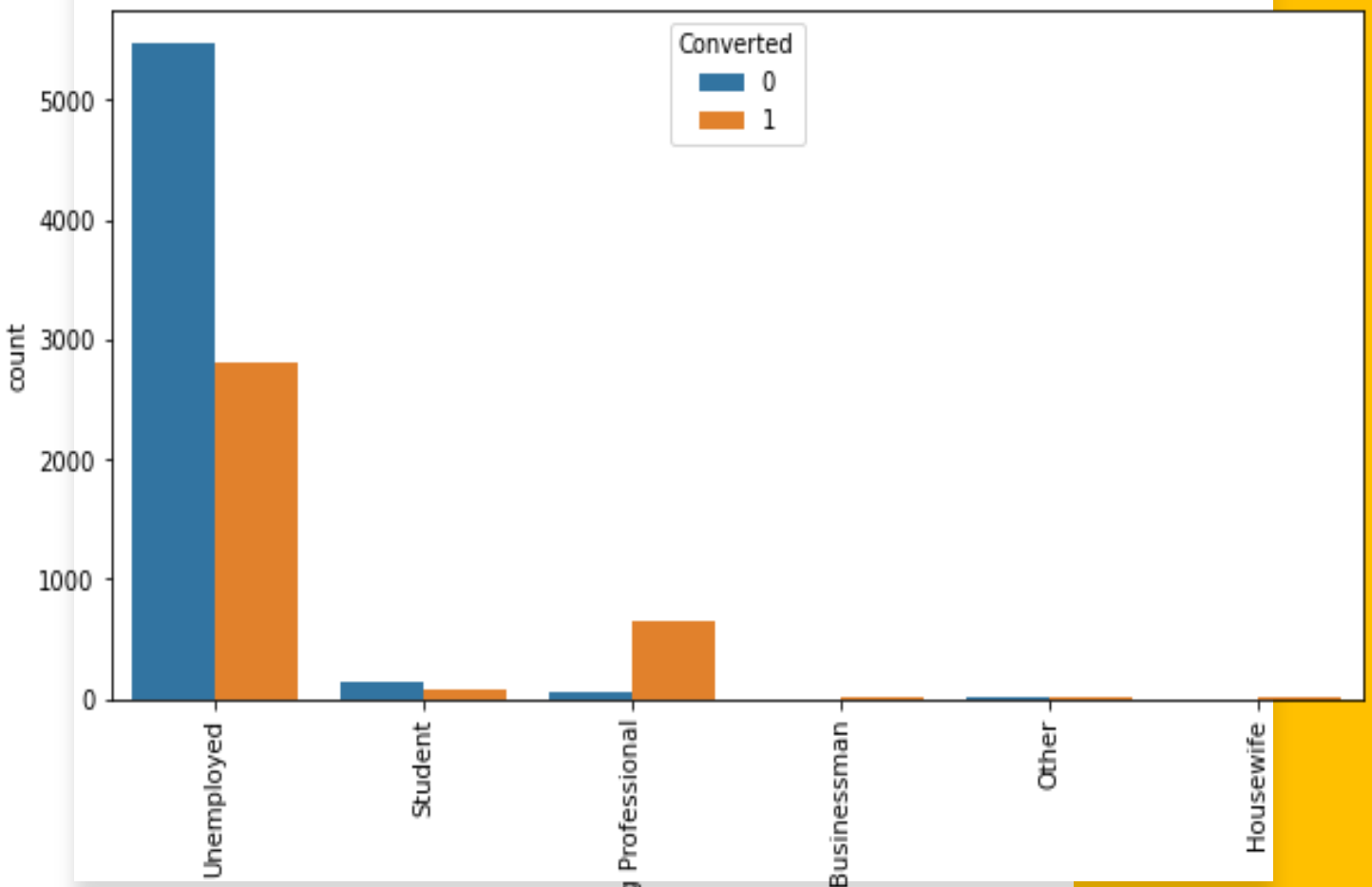


3.2 Attribute analysis-Categorical

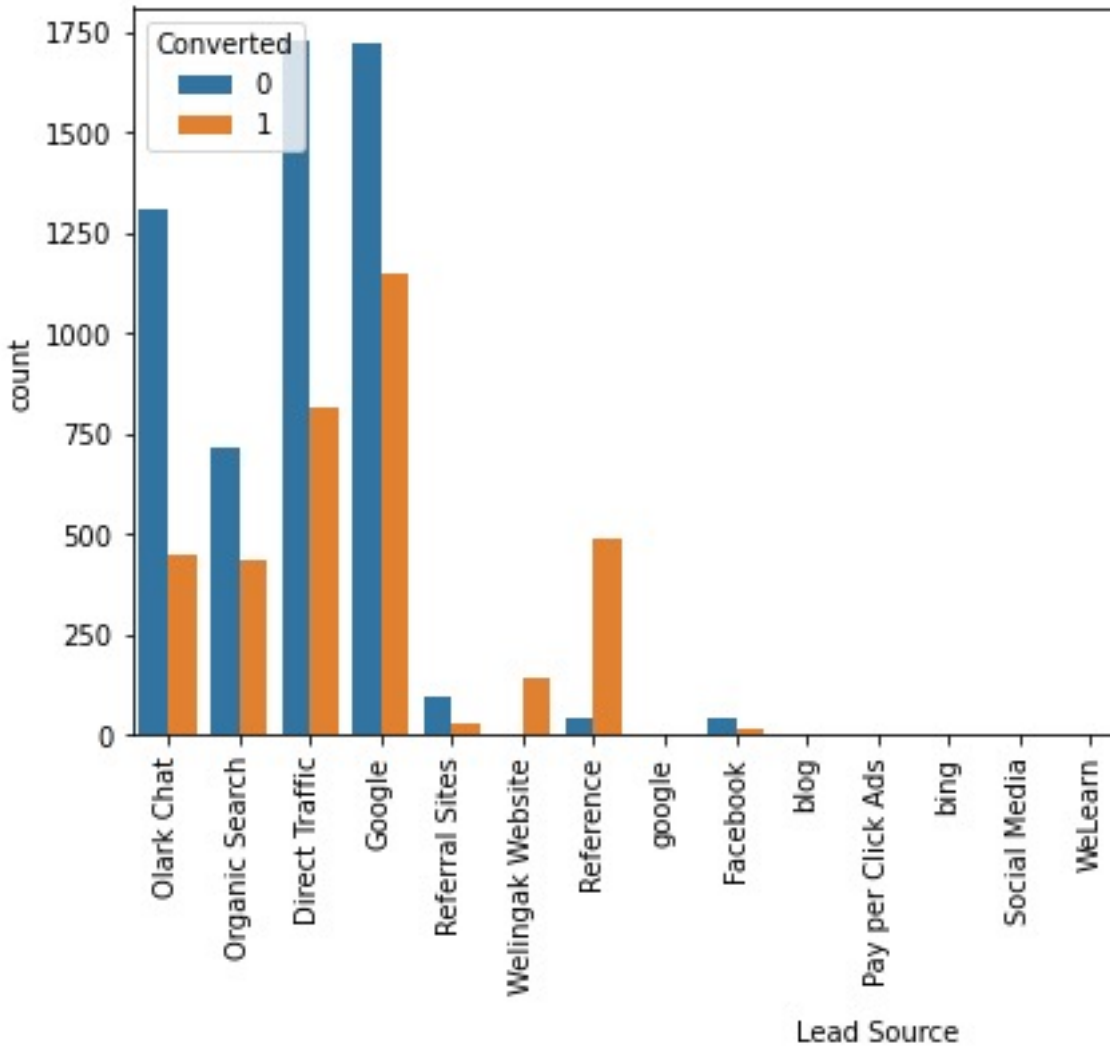
- We can see the Number of Values for India are quite high(nearly 97%), we can drop this column easily



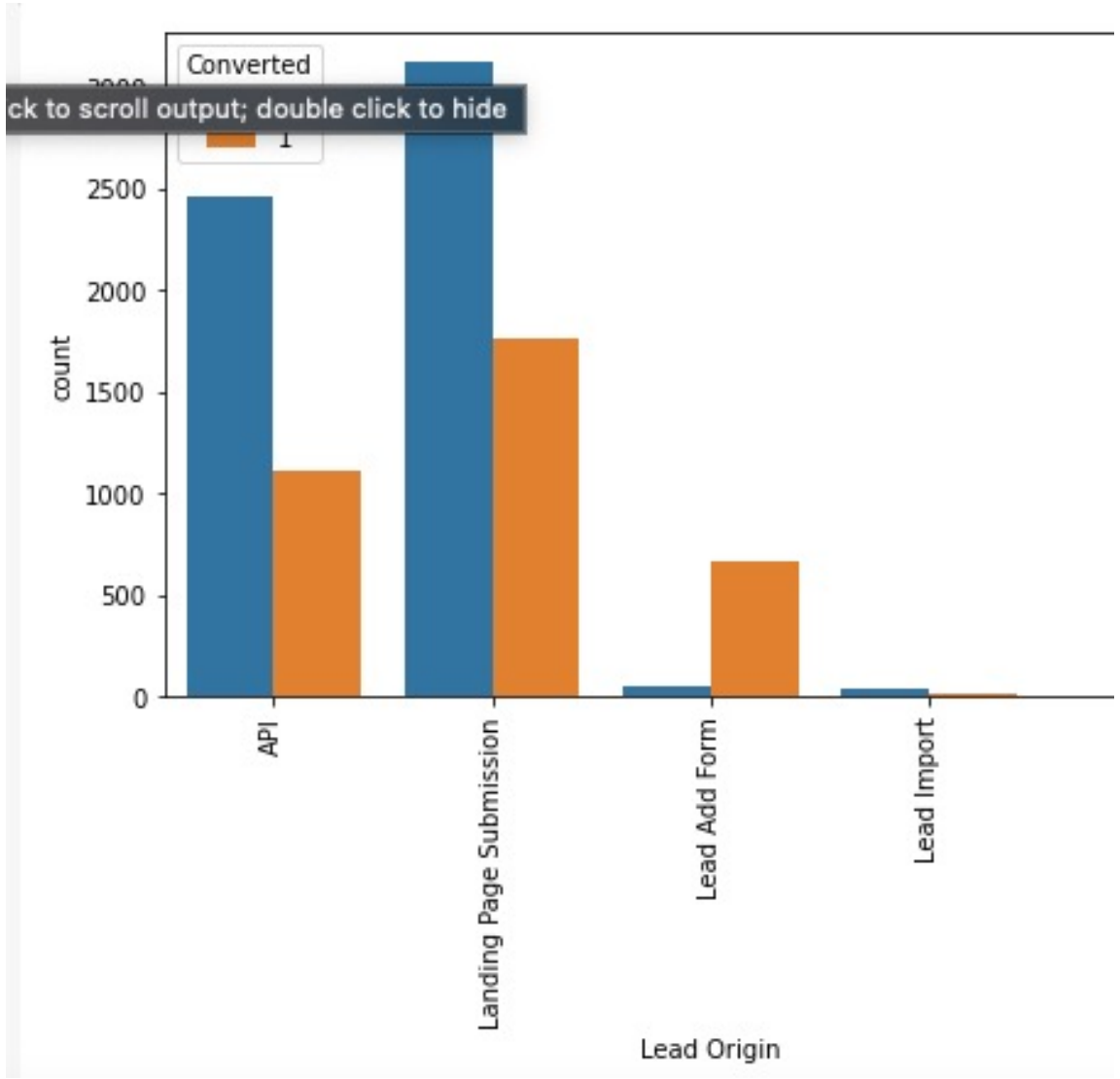
- Working Professionals are higher in term of hot leads
- Unemployed leads are the most



- Maximum number of leads are generated by Google and Direct traffic.
- Conversion Rate of reference leads and leads through welingak website is high.
- To improve overall lead conversion rate, focus should be on improving lead converion of olark chat, organics search,direct traffic, and google leads and generate more leads from reference and welingak website.



- API and Landing Page Submission bring higher number of leads as well as conversion.
- Lead Add Form has a very high conversion rate but count of leads are not very high.
- Lead Import and Quick Add Form get very few leads.
- In order to improve overall lead conversion rate, we have to improve lead converion of API and Landing Page Submission origin and generate more leads from Lead Add Form.



4. Model building

- We use RFE method to remove insignificant variables
- Model was built with rfe count 19. We iterated the process until we deleted insignificant variables and VIF was below 5%.

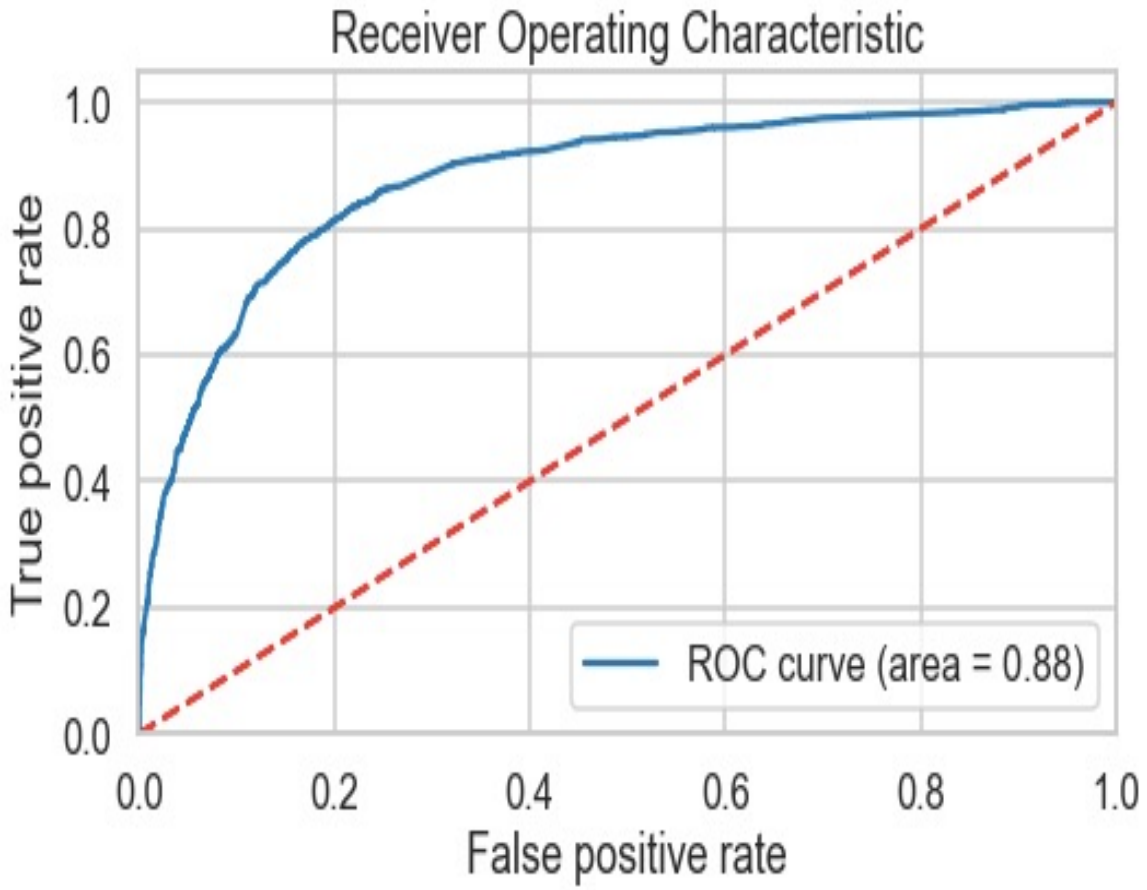
Final model

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.5015	0.010	48.826	0.000	0.481	0.522
Do Not Email	-0.1771	0.018	-9.938	0.000	-0.212	-0.142
Total Time Spent on Website	0.1878	0.005	36.209	0.000	0.178	0.198
Lead Origin_Lead Add Form	0.5584	0.020	28.060	0.000	0.519	0.597
Lead Source_Olark Chat	0.1686	0.014	11.948	0.000	0.141	0.196
Lead Source_Welingak Website	0.1951	0.043	4.486	0.000	0.110	0.280
Last Activity_Olark Chat Conversation	-0.1225	0.020	-6.134	0.000	-0.162	-0.083
What is your current occupation_Working Professional	0.3450	0.018	19.053	0.000	0.310	0.381
Last Notable Activity_Email Link Clicked	-0.3083	0.036	-8.650	0.000	-0.378	-0.238
Last Notable Activity_Email Opened	-0.2242	0.013	-17.588	0.000	-0.249	-0.199
Last Notable Activity_Modified	-0.3012	0.013	-23.495	0.000	-0.326	-0.276
Last Notable Activity_Olark Chat Conversation	-0.2846	0.040	-7.160	0.000	-0.363	-0.207
Last Notable Activity_Page Visited on Website	-0.2673	0.026	-10.180	0.000	-0.319	-0.216

	Features	VIF
5	Last Activity_Olark Chat Conversation	1.89
3	Lead Source_Olark Chat	1.65
9	Last Notable Activity_Modified	1.51
2	Lead Origin_Lead Add Form	1.41
10	Last Notable Activity_Olark Chat Conversation	1.30
4	Lead Source_Welingak Website	1.23
1	Total Time Spent on Website	1.20
6	What is your current occupation_Working Profes...	1.14
0	Do Not Email	1.11
8	Last Notable Activity_Email Opened	1.10
7	Last Notable Activity_Email Link Clicked	1.02
11	Last Notable Activity_Page Visited on Website	1.02

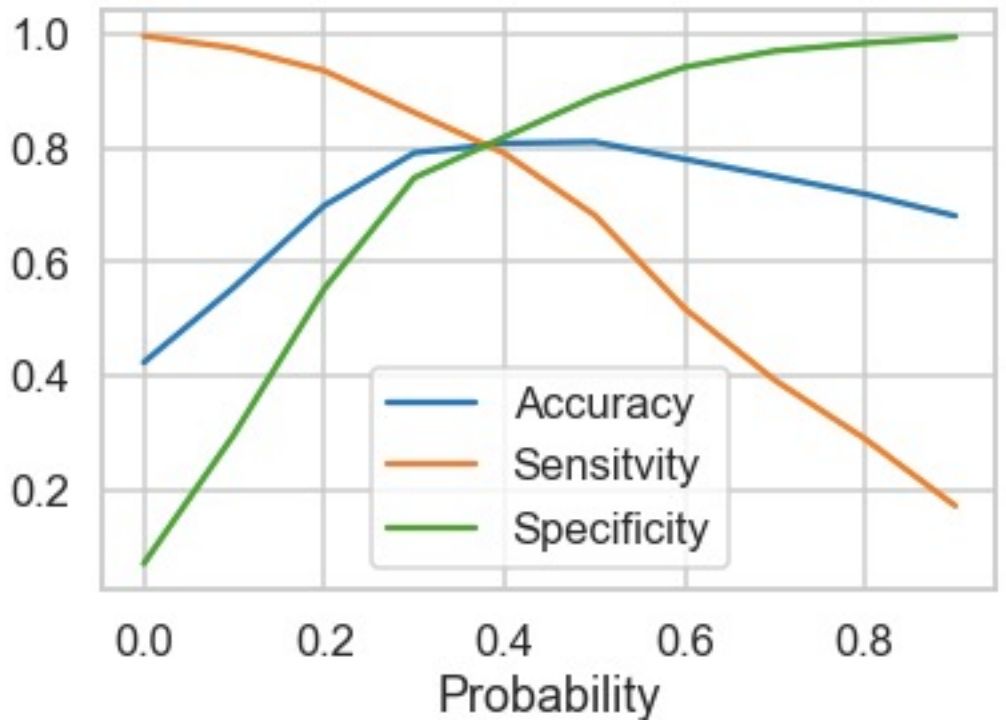
Model evaluation

We can see that curve is closer to the left side of the border than to the right side. Therefore, our model has greater accuracy. Area under the curve is 88% of the total area (ROC curve area).



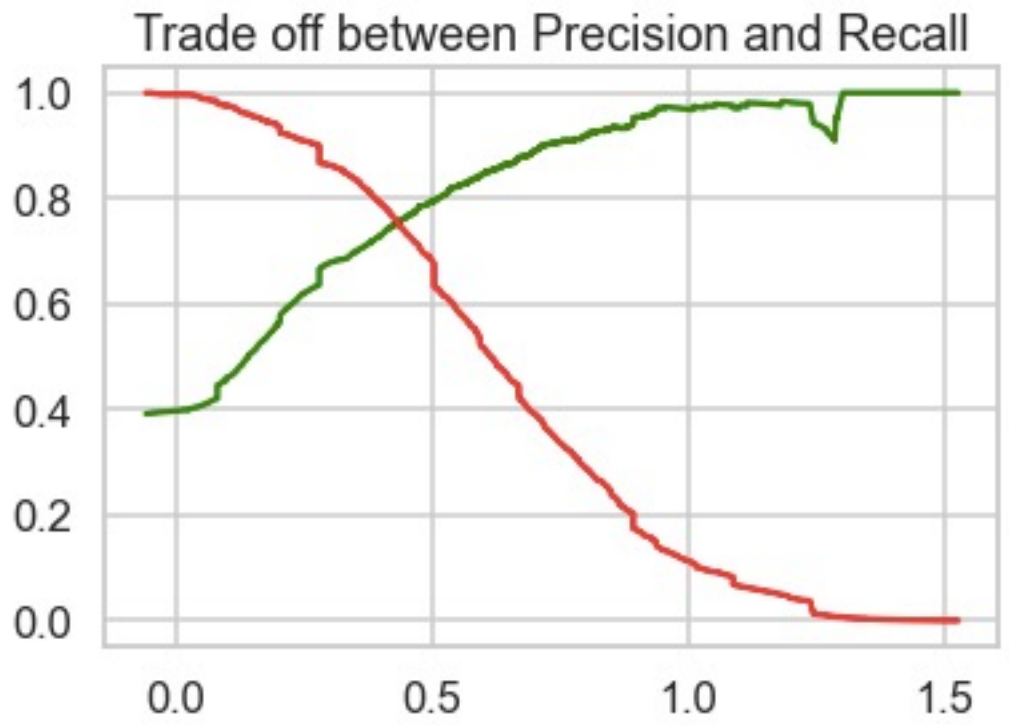
Optimal cutoff point

- From the above curve, 0.4 is the optimum point to take it as a cutoff probability



Precision and Recall

- There is a trade off between precision and recall and they meet near to 0.5
- Precision has 73% score, whereas, Recall has 79% score. So recall is more valuable from the business objective point of view



5. Prediction-Test set-Inference

- Model is stable and suitable for company requirements based on following outcomes:
  - Accuracy, Sensitivity and Specificity values of test set are around 81%, 78% and 83% which are approximately closer to the respective values calculated using trained set. The values are in acceptable range.
  - 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.
  - Also, the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 78%

6. Recommendation

- Company can focus on following categories for Lead conversions:
  - Total Time Spent on websites.
  - Current Occupation of the Customers in which mostly unemployed are the hot Leads.
  - In order to improve Lead Conversion Company should focus in API and Landing Page Submission Origin.
  - Lead Source are one of the feature through which high conversion rates are observed with the help of welingak websites and olark chat.

Thank you