International
Institute of Information
Technology Bangalore

Lead Score – Case Study

Submitted By:

- **Vikrant Singh**
- **Manuj Tyagi**

Problem Statement

An X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, 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 at X education is around 30%.**

Now, although X Education 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'.
The company requires you to build a model wherein 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%.**

Goals of the Case Study

There are quite a few goals for this case study.

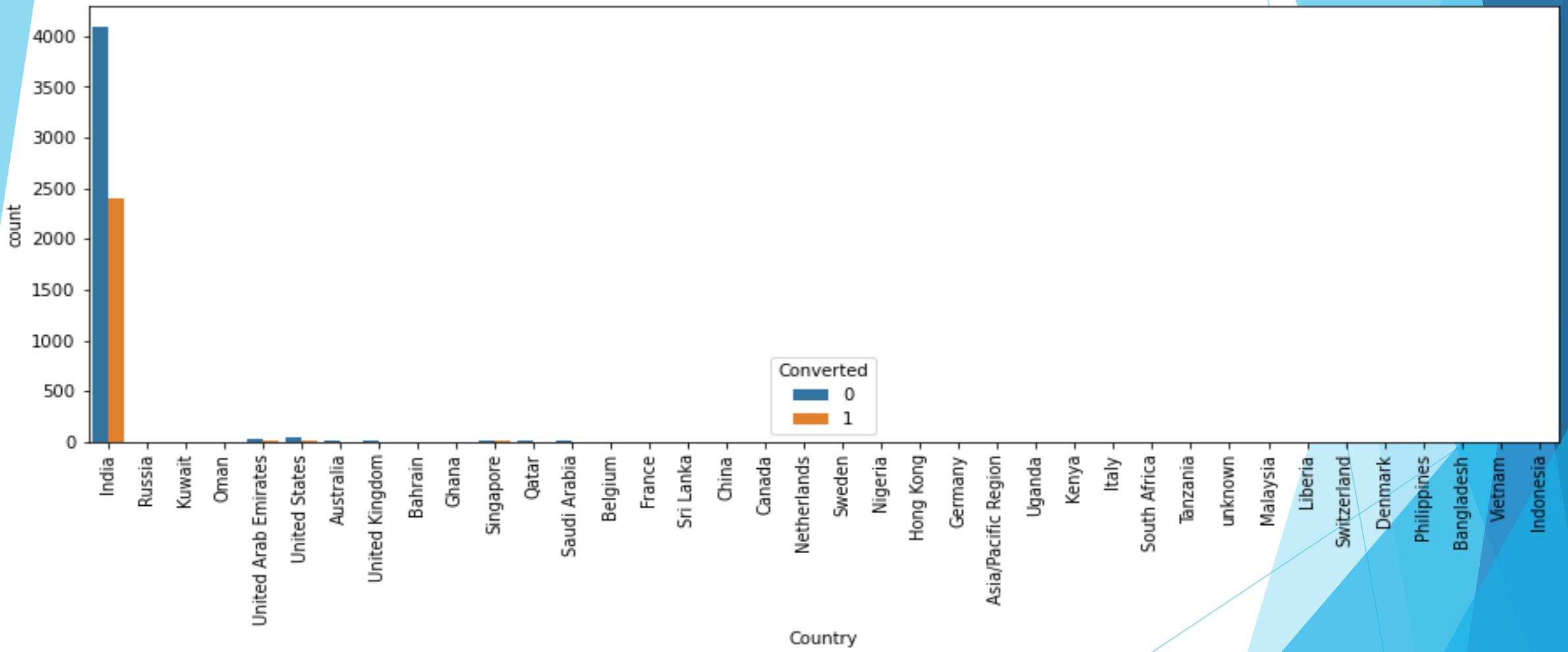
1.Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

2.There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

Strategy

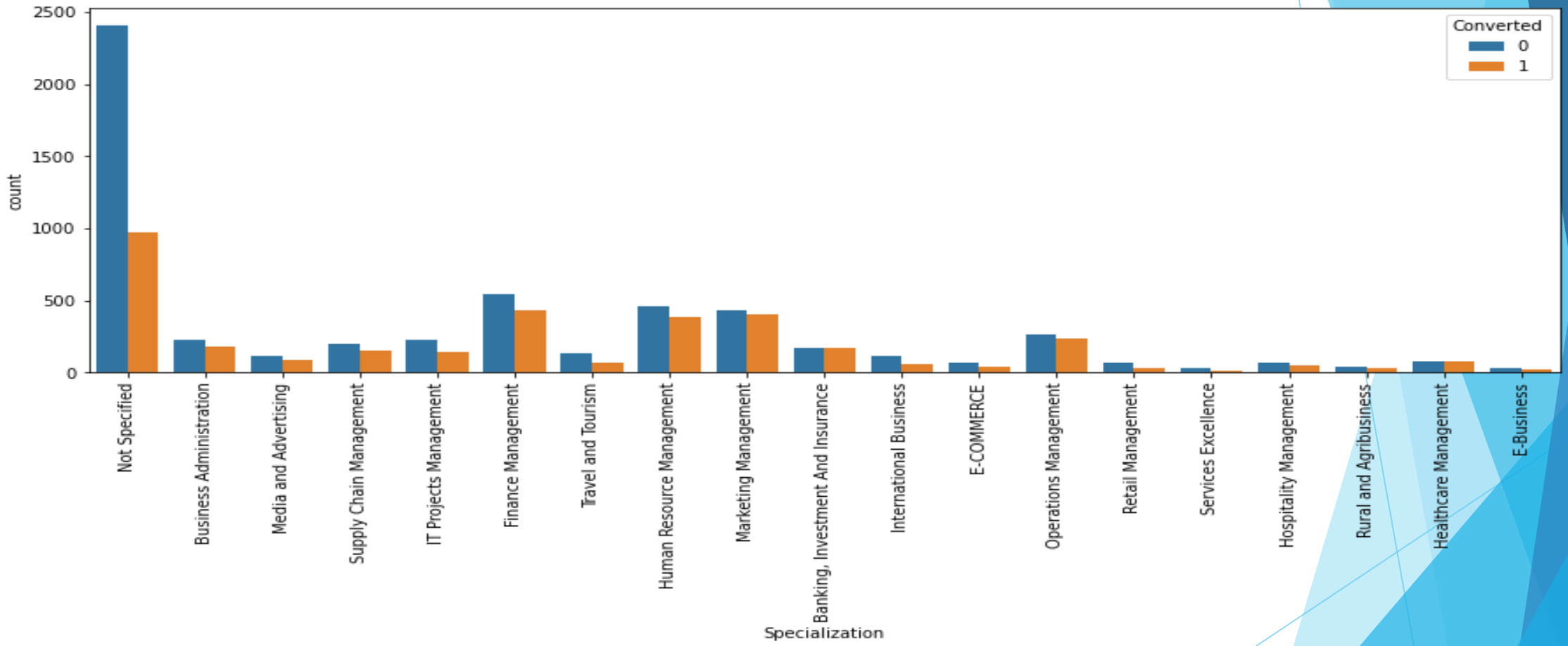
- Importing and understanding the data
- Cleaning the data and preparing it for Analysis
- Exploratory Data Analysis
- Splitting data into training set and test set
- Scaling
- Building logistic regression model with the best features
- Evaluating the model on training set
- Finding the optimal cut-off to get the best accuracy, sensitivity and specificity.
- Evaluating the model on test set
- Assigning lead score to leads based on model created

Exploratory Data Analysis



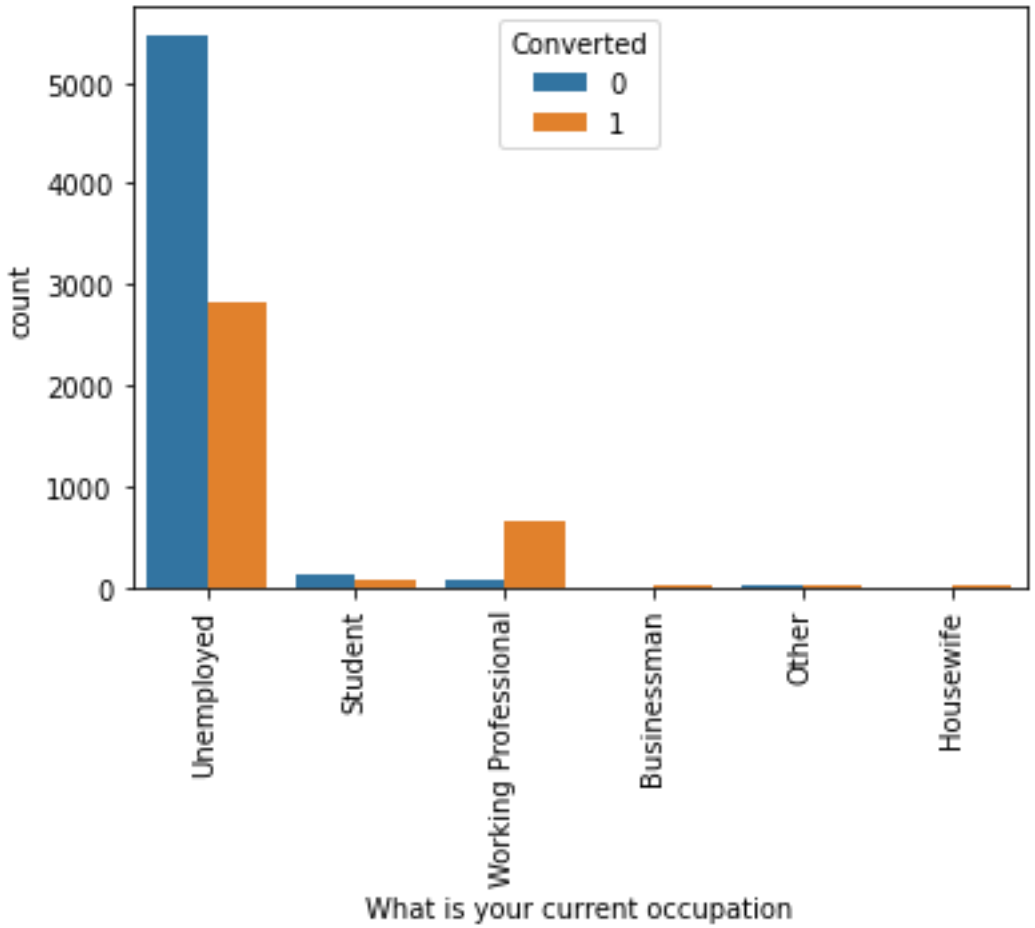
As we can see the Number of Values for India are quite high (about 97% of the Data), so we drop this column

Exploratory Data Analysis



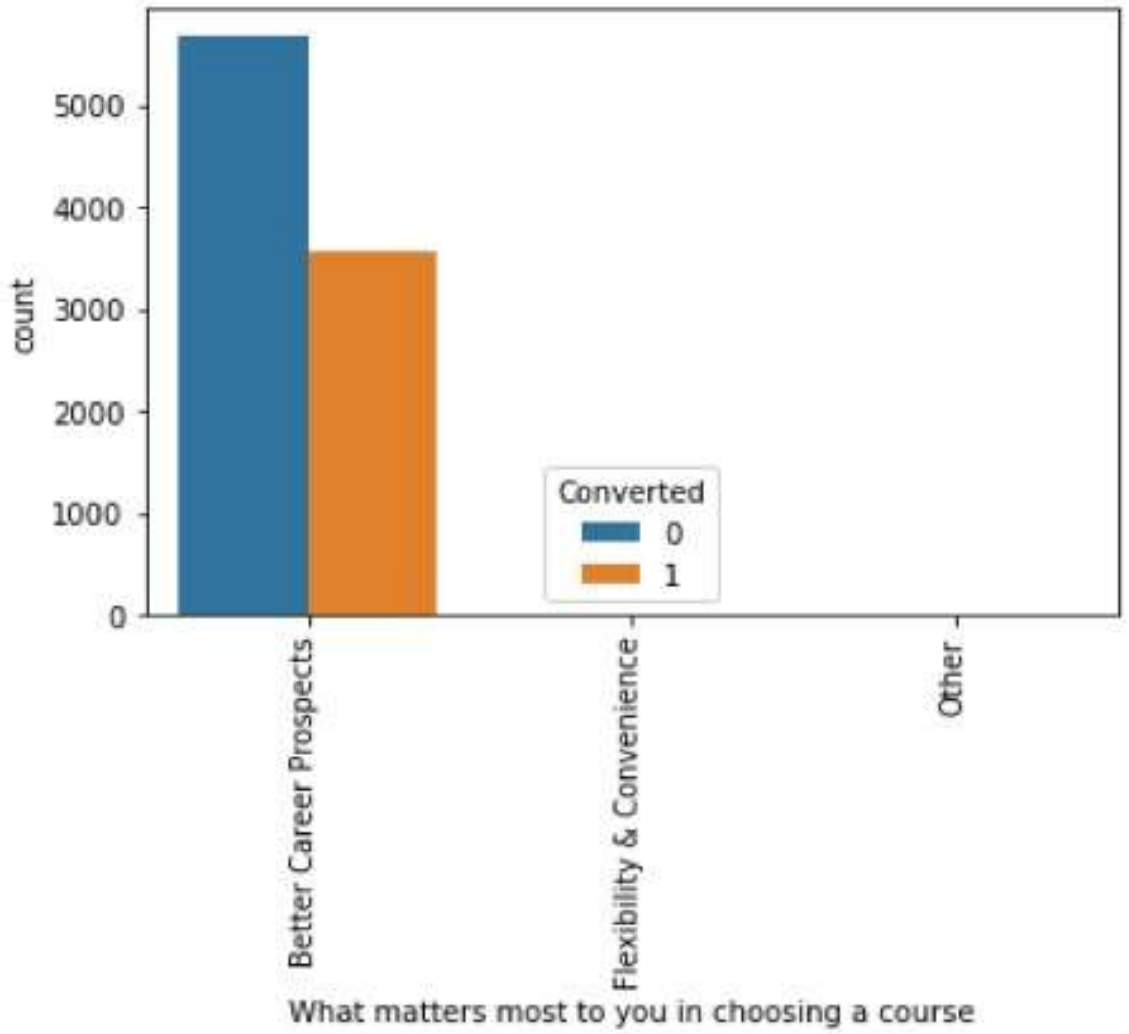
as We see that specialization with Management in them have higher number of leads as well as leads converted. So this is definitely a significant variable.

Exploratory Data Analysis

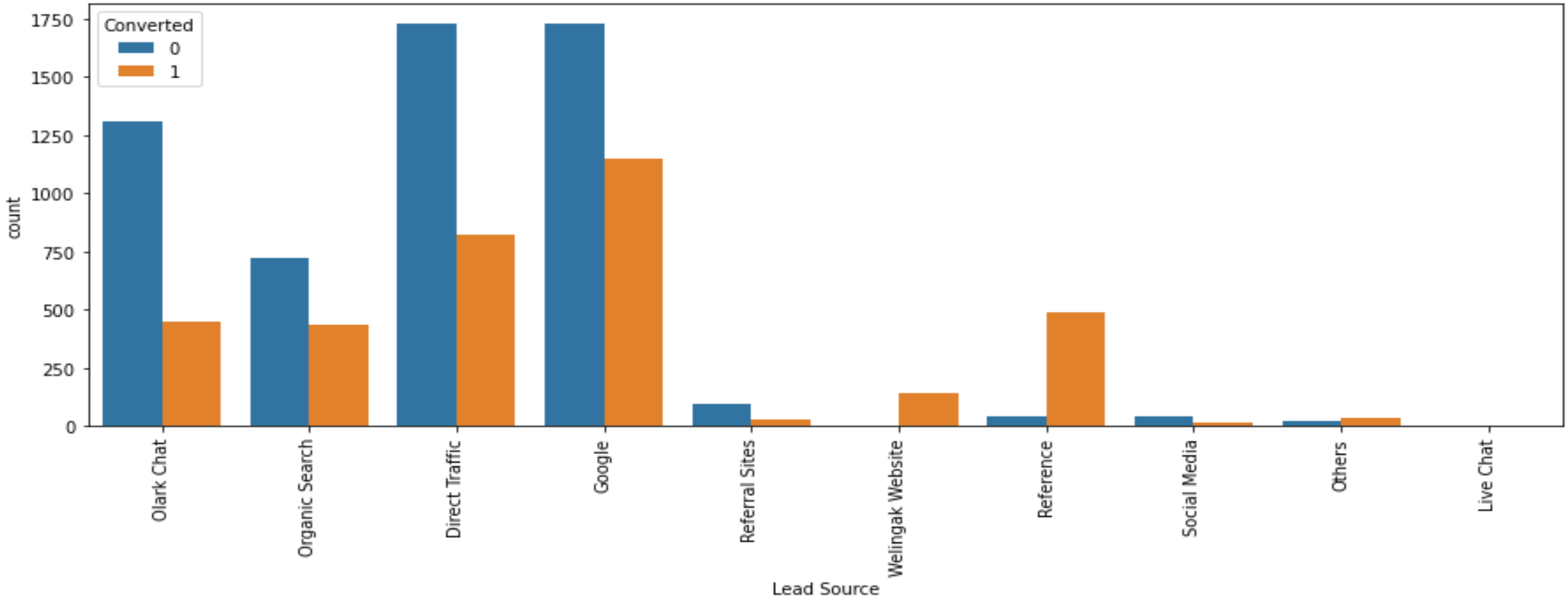


Working Professionals going for the course have high chance of joining it and Unemployed leads are the most in terms of Absolute numbers.

Exploratory Data Analysis

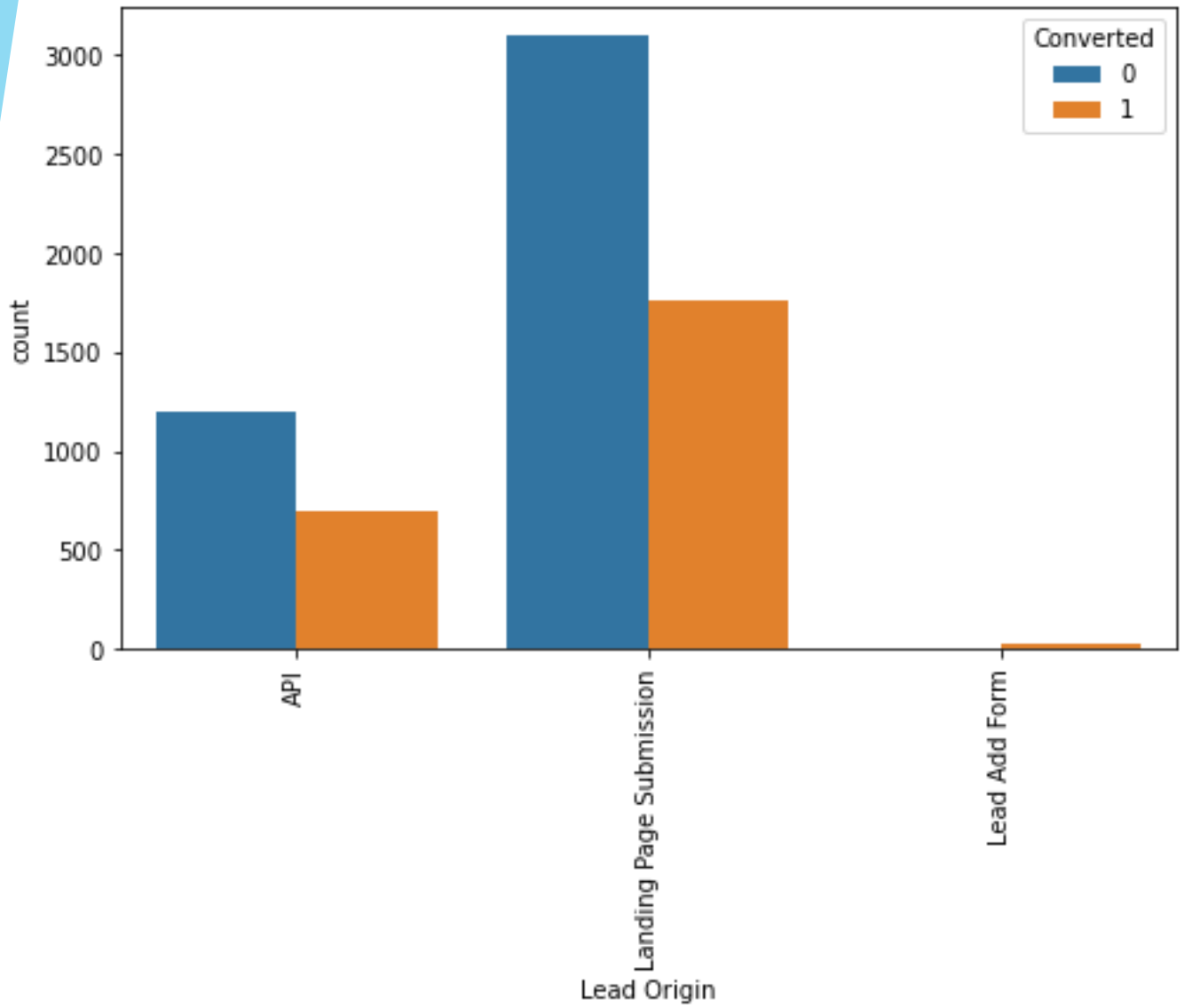


Exploratory Data Analysis



As we can see the Maximum number of leads are generated by Google and Direct traffic and the Conversion Rate of reference leads and leads through welingak website is high. To improve overall lead conversion rate our focus should be on improving lead conversion of olark chat, organic search, direct traffic, and Google leads and generate more leads from reference and welingak website.

Exploratory Data Analysis



1. API and Landing Page Submission bring a higher number of leads as well as conversion.

2. Lead Add Form has a very high conversion rate but the count of leads is not very high.

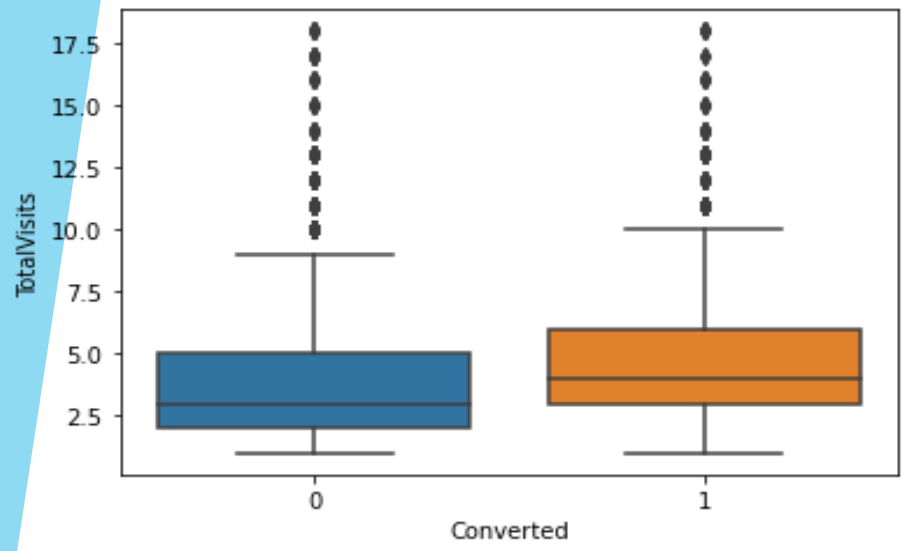
3. Lead Import and Quick Add Form get very few leads.

4. In order to improve the overall lead conversion rate, we have to improve the lead conversion of API and Landing Page Submission origin and generate more leads from the Lead Add Form.

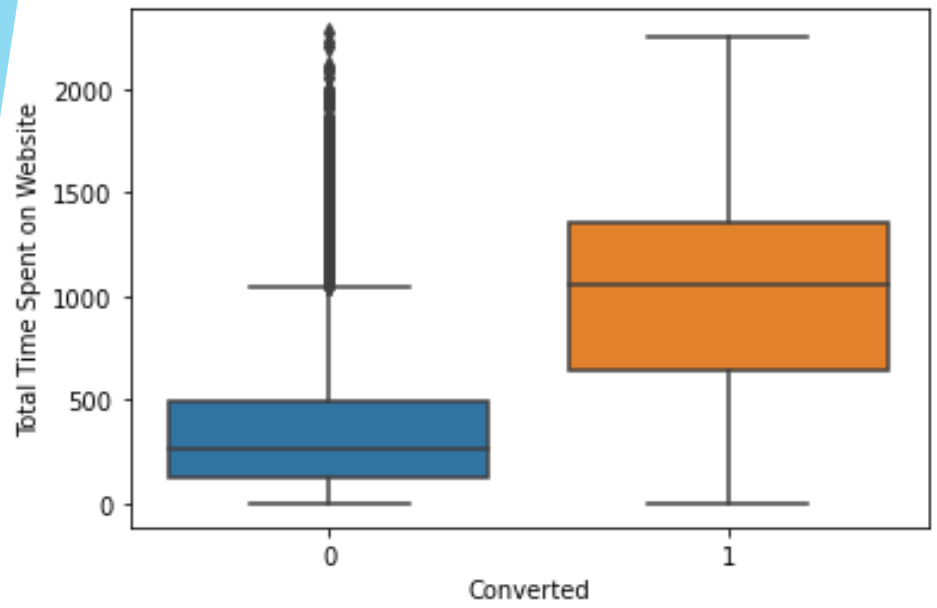
Exploratory Data Analysis

correlations of numeric values

Exploratory Data Analysis

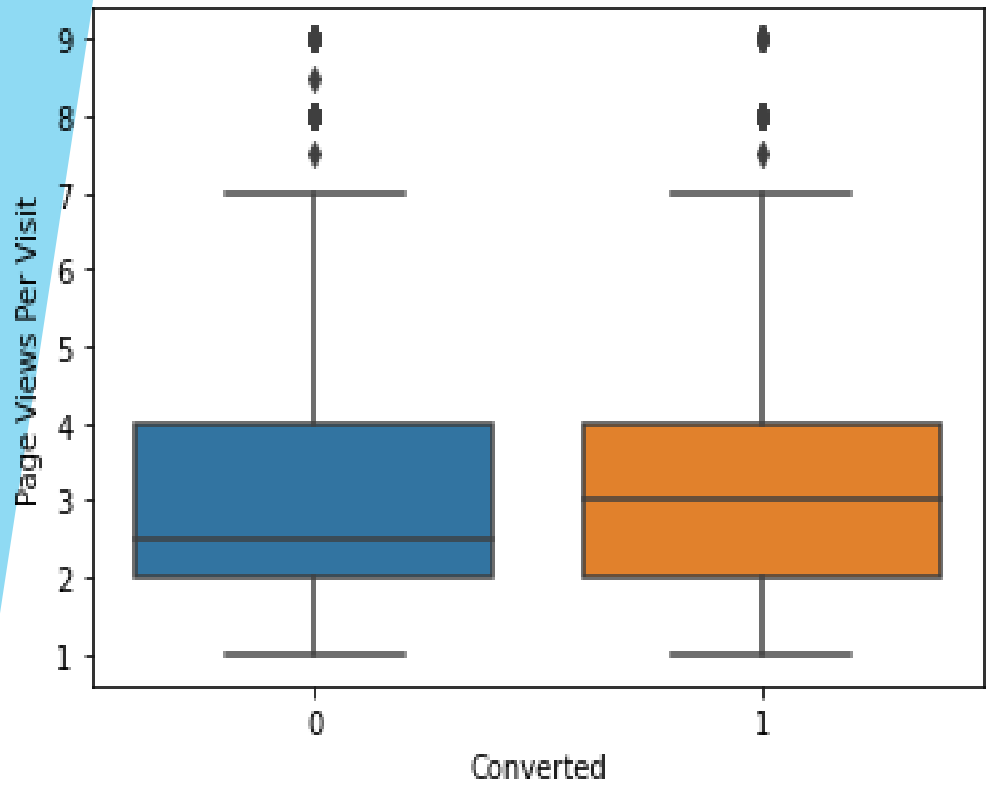


As we can see Median for converted and not converted leads are the close and Nothing conclusive can be said on the basis of Total Visits



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

Exploratory Data Analysis



The median for converted and unconverted leads is the same but Nothing can be said specifically for lead conversion from Page Views Per Visit

Model Building Approach

- Splitting data into train and test set
- Initial Model built with all the selected feature
- Feature selection using RFE(Top 15)
- Building Subsequent models based on these features
- Eliminating features based on p-value and VIF and Rebuilding the model with remaining features.
- Evaluating accuracy and other metrics on the training set
- Finding the optimal cut-off using the intersection of accuracy, sensitivity, and specificity on the graph.
- Evaluating the model on test data
- Comparing the Test and Train metrics.
- Assigning a lead score to the leads.

Most impacting variables

According to the summary obtained, we can find the most impacting variables:

- Tags_Lost to EINS
- Tags_Will revert after reading the email
- What is your current occupation_Working Professional

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	4648
Model:	GLM	**Df Residuals:**	4636
Model Family:	Binomial	**Df Model:**	11
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-938.81
Date:	Mon, 17 Oct 2022	**Deviance:**	1877.6
Time:	18:51:33	**Pearson chi2:**	6.12e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	1.0332	0.196	5.274	0.000	0.649	1.417
Do Not Email	-1.1948	0.289	-4.139	0.000	-1.761	-0.629
Total Time Spent on Website	1.1381	0.063	18.077	0.000	1.015	1.262
Last Activity_SMS Sent	1.4964	0.129	11.583	0.000	1.243	1.750
What is your current occupation_Working Professional	1.8338	0.438	4.189	0.000	0.976	2.692
Tags_Interested in other courses	-4.5763	0.456	-10.045	0.000	-5.469	-3.683
Tags_Lost to EINS	2.7168	0.637	4.264	0.000	1.468	3.966
Tags_Not Specified	-2.4183	0.201	-12.005	0.000	-2.813	-2.024
Tags_Other_Tags	-4.7697	0.294	-16.215	0.000	-5.346	-4.193
Tags_Ringing	-5.6556	0.333	-16.959	0.000	-6.309	-5.002
Tags_Will revert after reading the email	2.1120	0.277	7.621	0.000	1.569	2.655
Last Notable Activity_Modified	-1.0873	0.136	-8.006	0.000	-1.354	-0.821

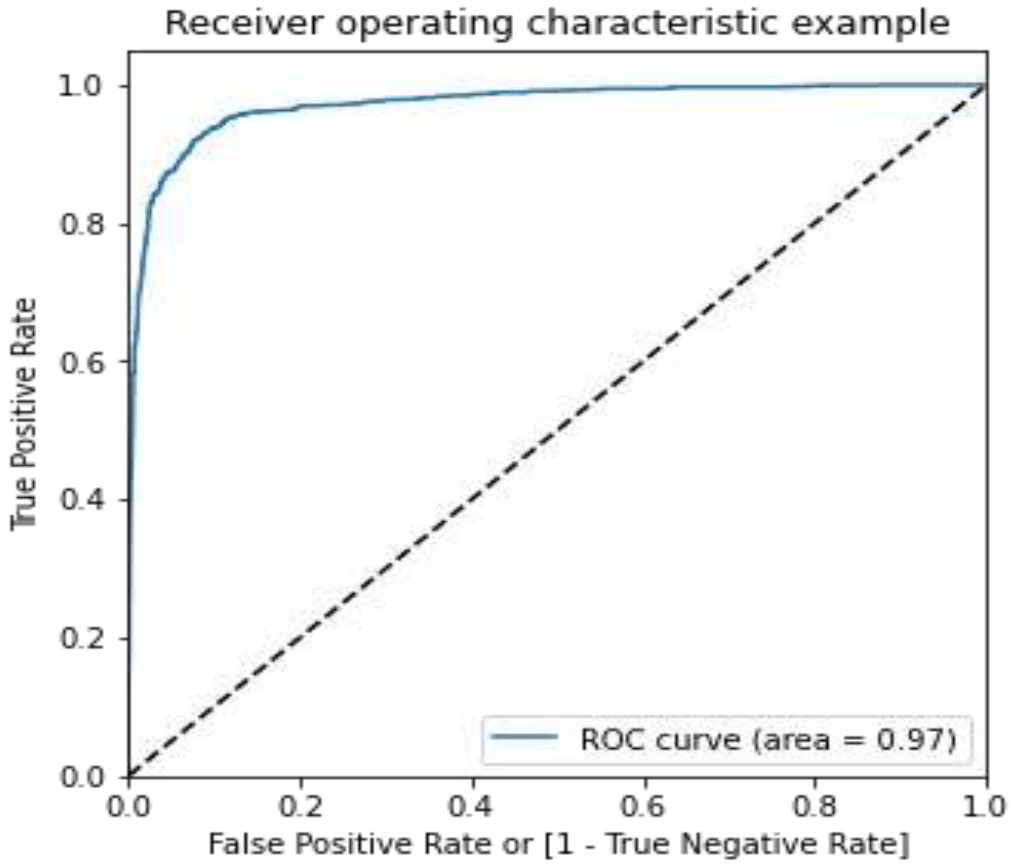
Finding optimal cutoff

	prob	accuracy	sensitivity	specificity
0.0	0.0	0.362952	1.000000	0.000000
0.1	0.1	0.840577	0.970954	0.766295
0.2	0.2	0.908348	0.950800	0.884161
0.3	0.3	0.919105	0.921755	0.917595
0.4	0.4	0.921687	0.896266	0.936170
0.5	0.5	0.924484	0.873740	0.953394
0.6	0.6	0.922978	0.841138	0.969605
0.7	0.7	0.914157	0.805572	0.976022
0.8	0.8	0.898236	0.751630	0.981763
0.9	0.9	0.870482	0.662715	0.988855

- Calculated Accuracy, Sensitivity, and specificity for all the cut-off values.
- Looking at the data we can say a **cut-off of 0.3** would be optimal as per the business needs.
- Findings: 0.2 to 0.7 is the cut-off range anything out of this will decline in accuracy.

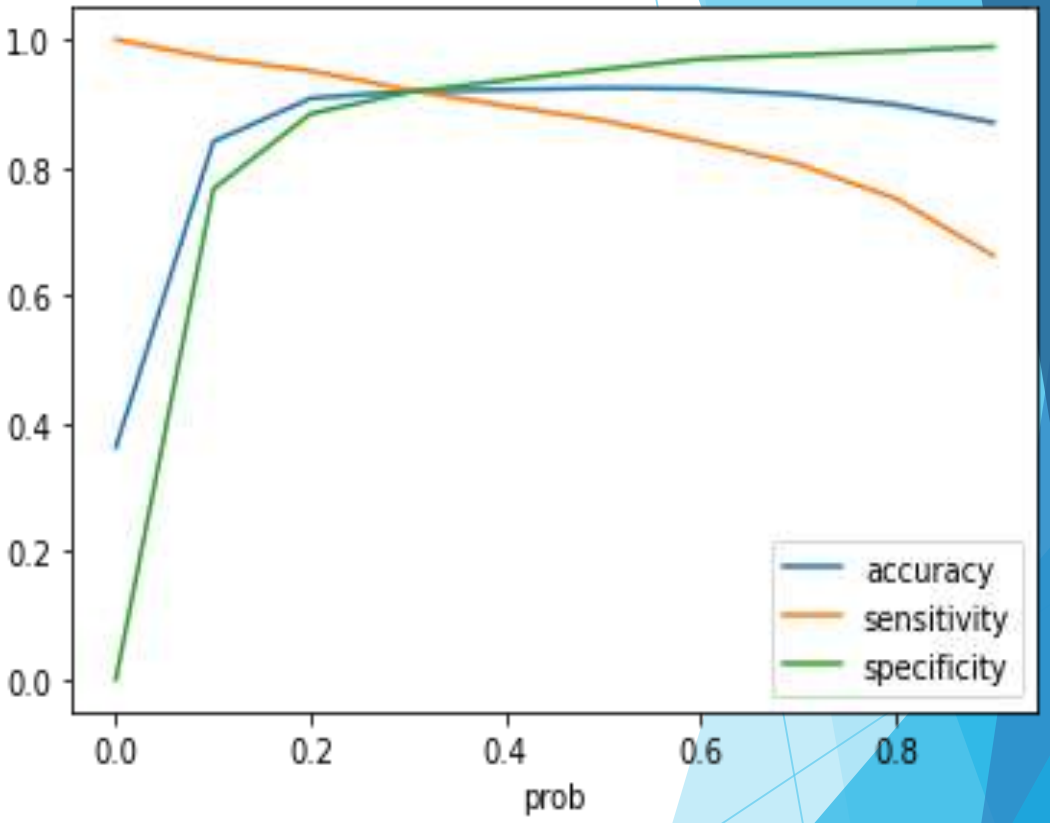
Model Evaluation -Sensitivity and Specificity on Train Data Set

the ROC Curve



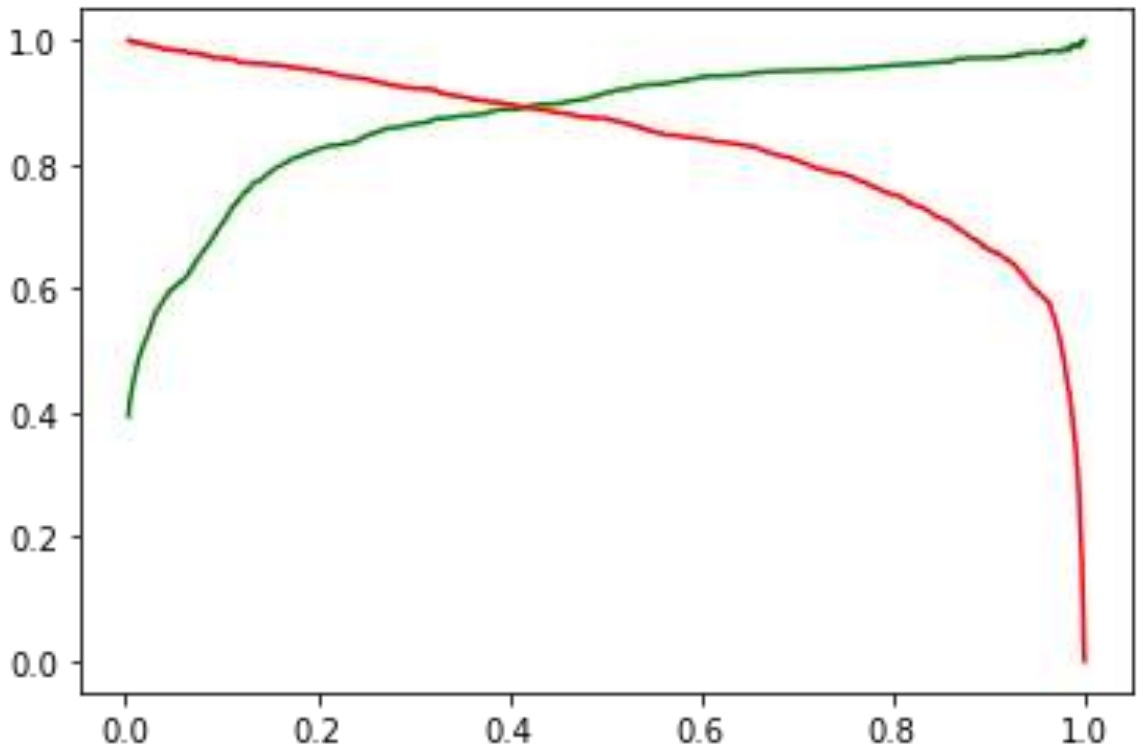
In the ROC plot it can be seen that the curve is going close to Y axis and near the value of 1. Also the area under the curve is very high. so the model is reliable.

Optimal Cutoff Point



form the above we can conclude that the cutoff of 0.3 is optimal as it shows the balance between all the metrics. so we can take 0.3 as the cutoff probability.

Model Evaluation -Precision and Recall on Train Data Set



The above graph shows an optimal cut off 0.4 based on Precision and Recall

- Precision- 95.03%
- Recall- 80.55 %

Comparing the train and test.

Train **Data:**

Accuracy : 91.9 %

Sensitivity : 92.1 %

Specificity : 91.7 %

Test Data:

Accuracy : 91.6 %

Sensitivity : 91.6 %

Specificity : 91.4 %

Suggestions and Recommendations:

- Sales team must put more focus on the leads who are working professions, as there is a high probability of conversion.
- Continuous reminder Emails must be sent to the potential leads.
- Should not waste resources on those leads who are interested in other courses.
- Try to make the platform more interactive so that the lead spends more time on the website.

Conclusion:

- The logistic Regression model is used to predict the probability of the conversion of a lead.
- The model has an accuracy of 91%, a sensitivity of 92%, and a specificity of 91%.
- This model will significantly improve the conversion rate from the existing 30%.
- In Business terms the model has the capability to adjust to the company's requirements in the future.