Lead Scoring Case Study

Problem Statement

An education company named 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. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. 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.

Goals of the Case Study

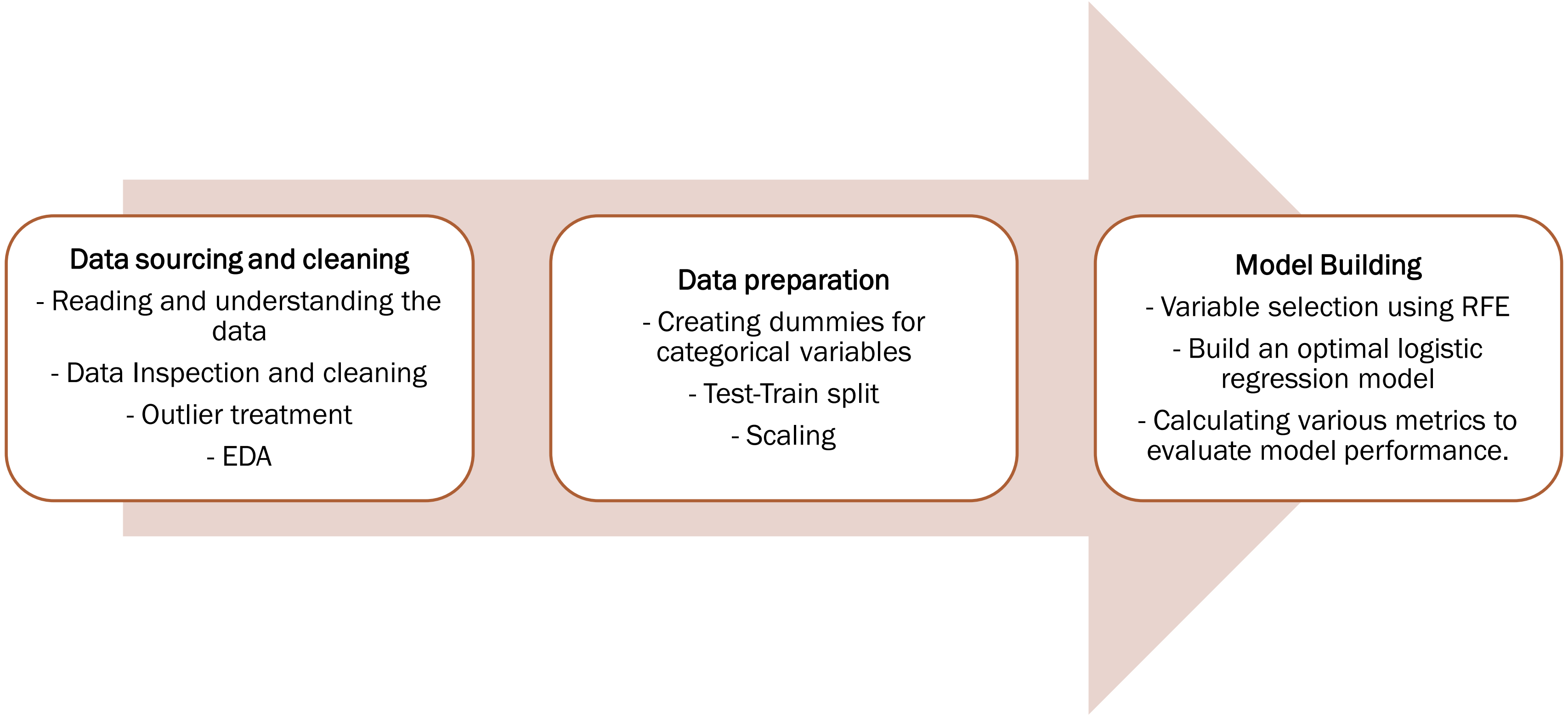
To help X Education select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance 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%.

We need to 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.

Steps involved

1. Reading and Understanding the Data
2. Data cleaning
3. Exploratory Data Analysis
4. Data Preparation (Dummy Variable Creation)
5. Test-Train Split
6. Rescaling the features with Min-Max Scaling
7. Model Building using Stats Model & RFE
8. Plotting the ROC Curve
9. Finding Optimal Cutoff Point
10. Making predictions on the test set
11. Evaluating Overall Metrics
12. Drawing Inferences from the model

Problem solving methodology



Data sourcing and cleaning

- Reading and understanding the data
- Data Inspection and cleaning
- Outlier treatment
- EDA

Data preparation

- Creating dummies for categorical variables
- Test-Train split
- Scaling

Model Building

- Variable selection using RFE
- Build an optimal logistic regression model
- Calculating various metrics to evaluate model performance.

Result

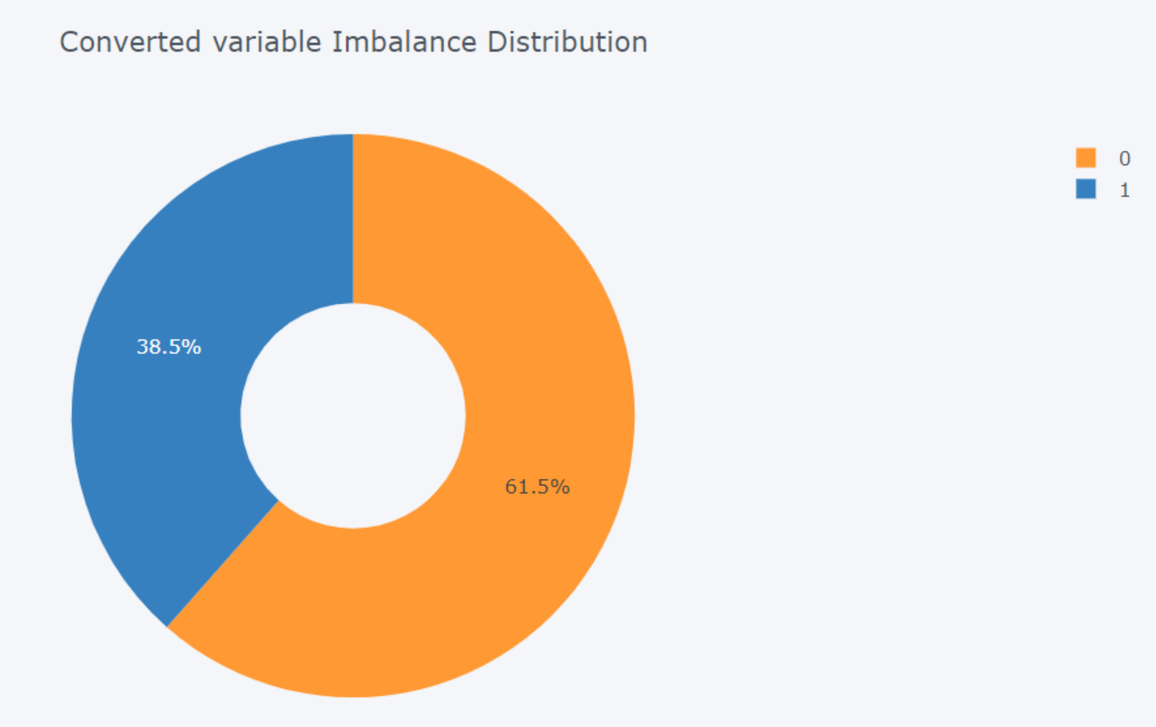
Determine the lead score and check if target final predictions amounts to 80% conversion rate.

- Evaluate the final prediction on the test set using cut off threshold from sensitivity and specificity metrics

Exploratory Data Analysis

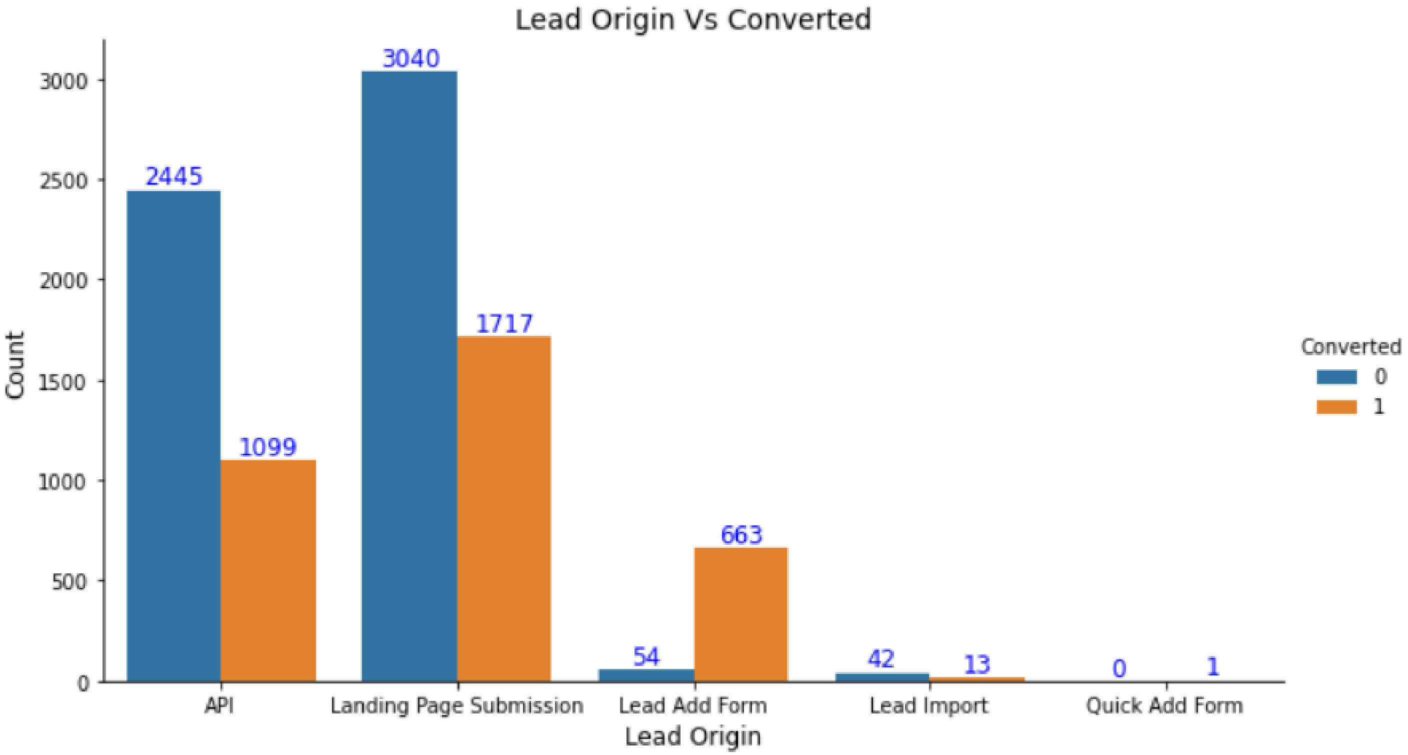
Variable: Converted

We can see that the conversion rate is 38.5% .



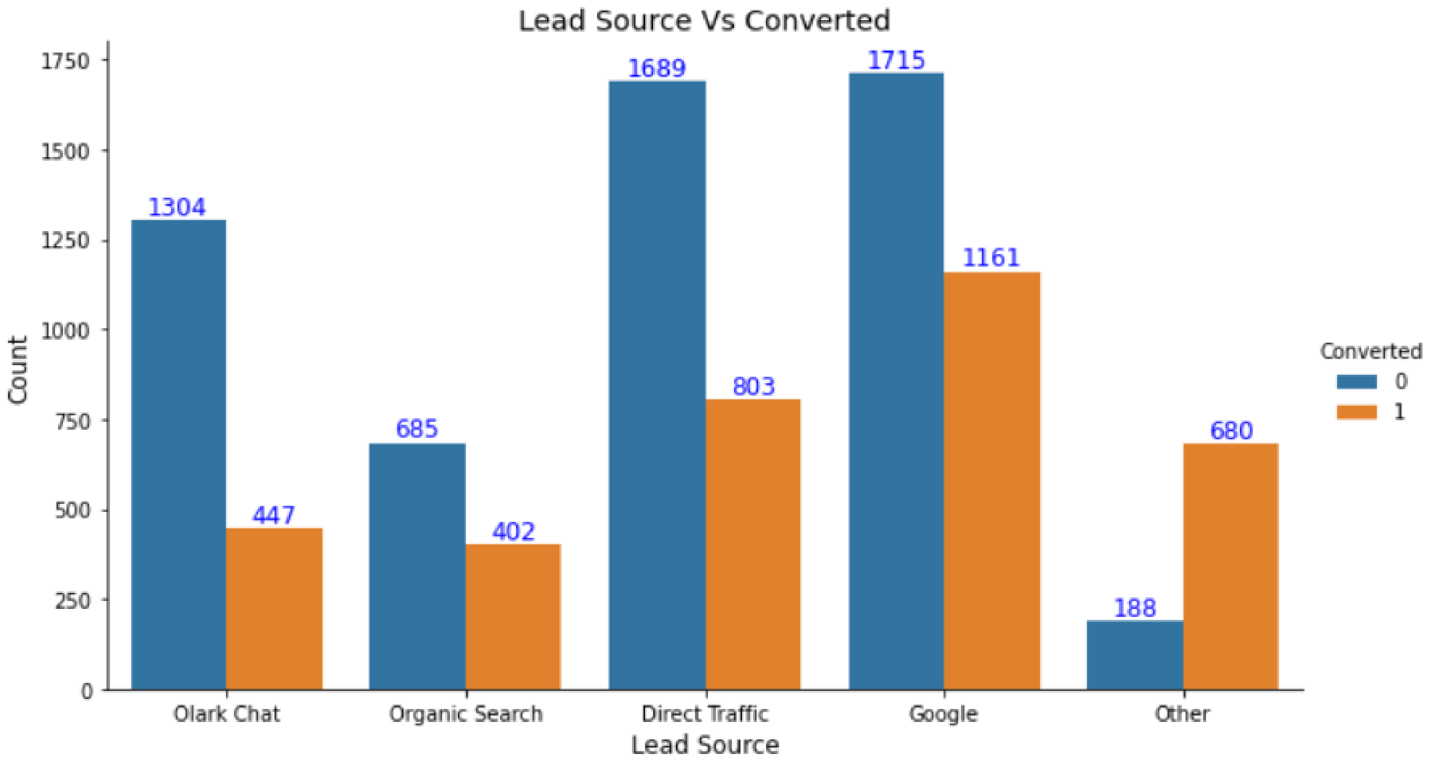
Exploratory Data Analysis

From the graph, it can be seen that the maximum conversion happened from Landing Page Submission.



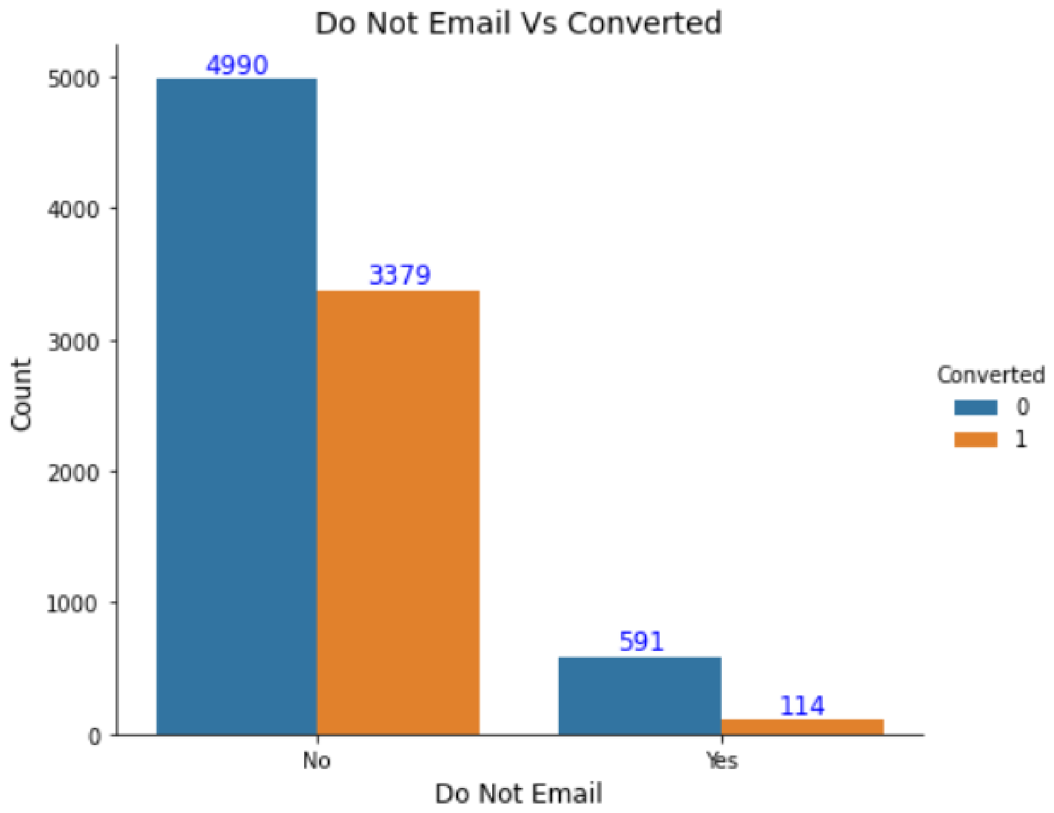
Exploratory Data Analysis

From the plot we can see that maximum conversions are there when lead is obtained from Google.



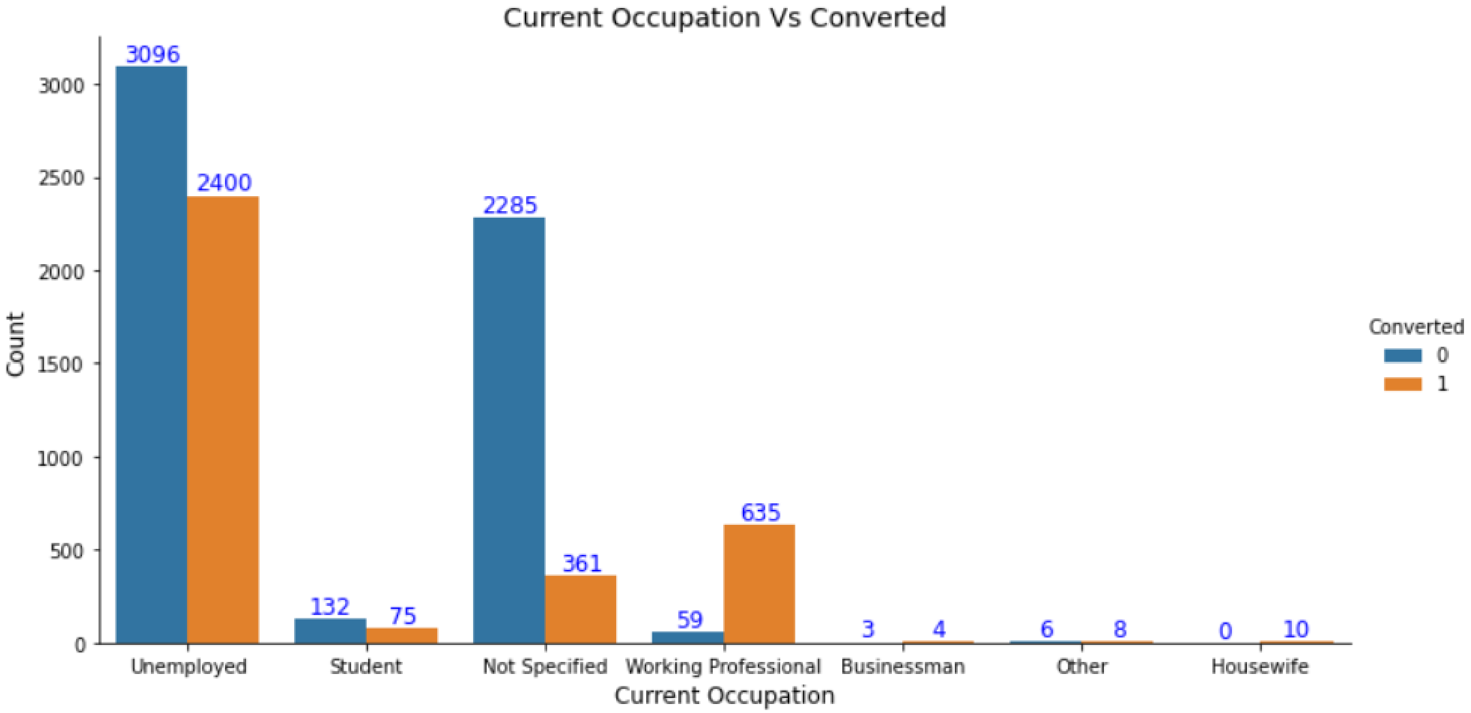
Exploratory Data Analysis

Majority of customers opted to be contacted through emails and the conversion rate is also higher for these customers.
Also, conversion rate is very for customers opting to not receive email from the company.



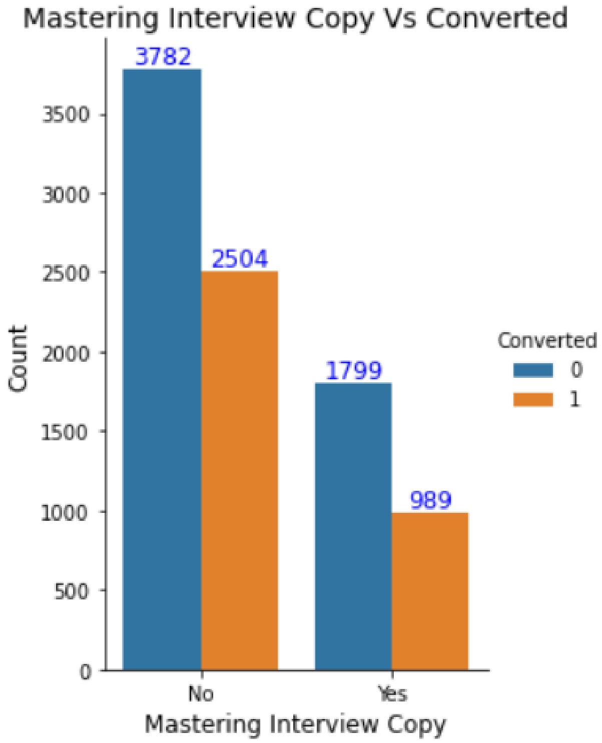
Exploratory Data Analysis

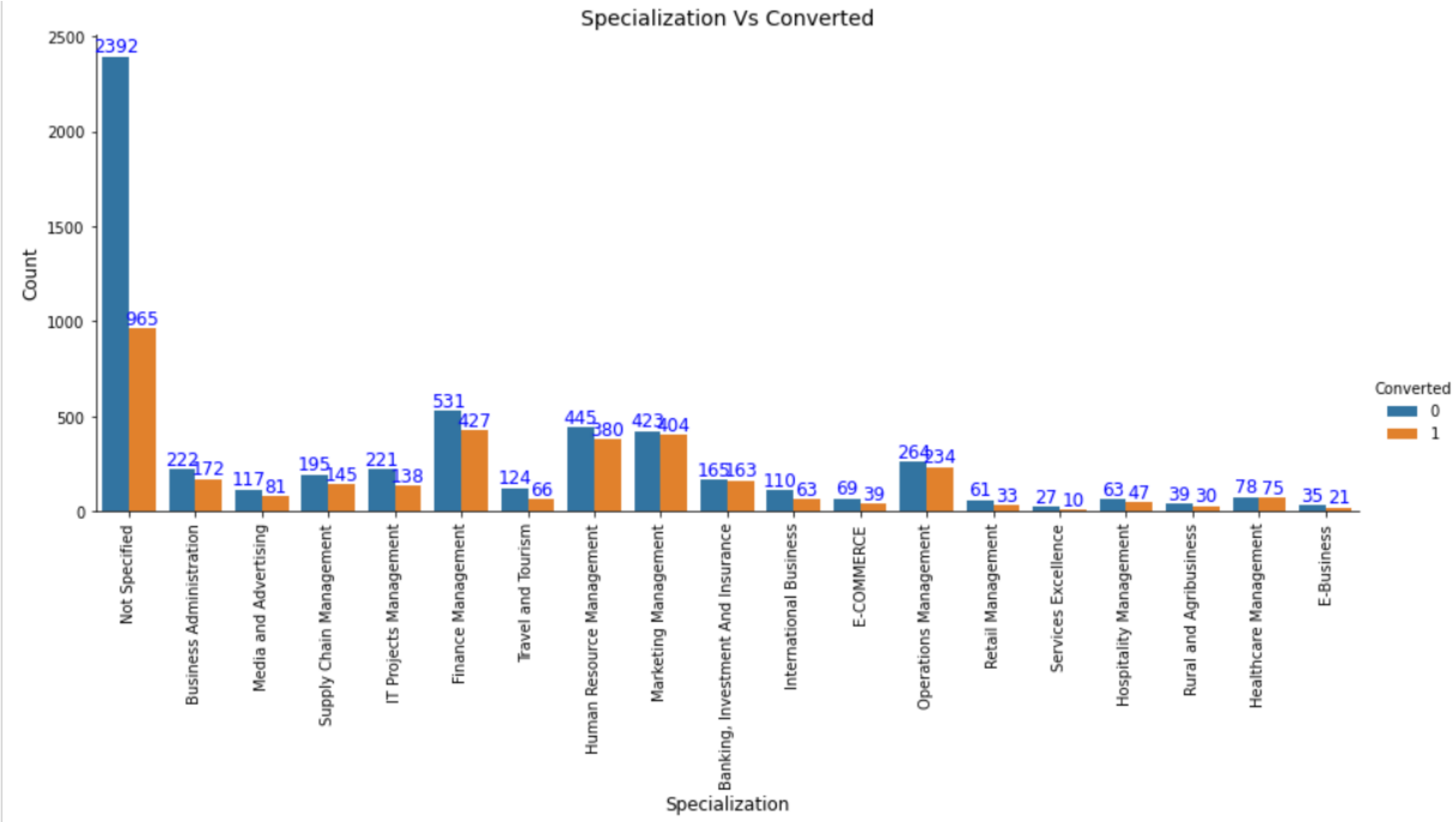
Maximum customers that converted are unemployed. We can also observe that working professionals are more likely to convert where customers who chose not to share their occupation are less likely to be converted.



Exploratory Data Analysis

Customers who do not want a free copy of Mastering the Interview have higher conversion rate.

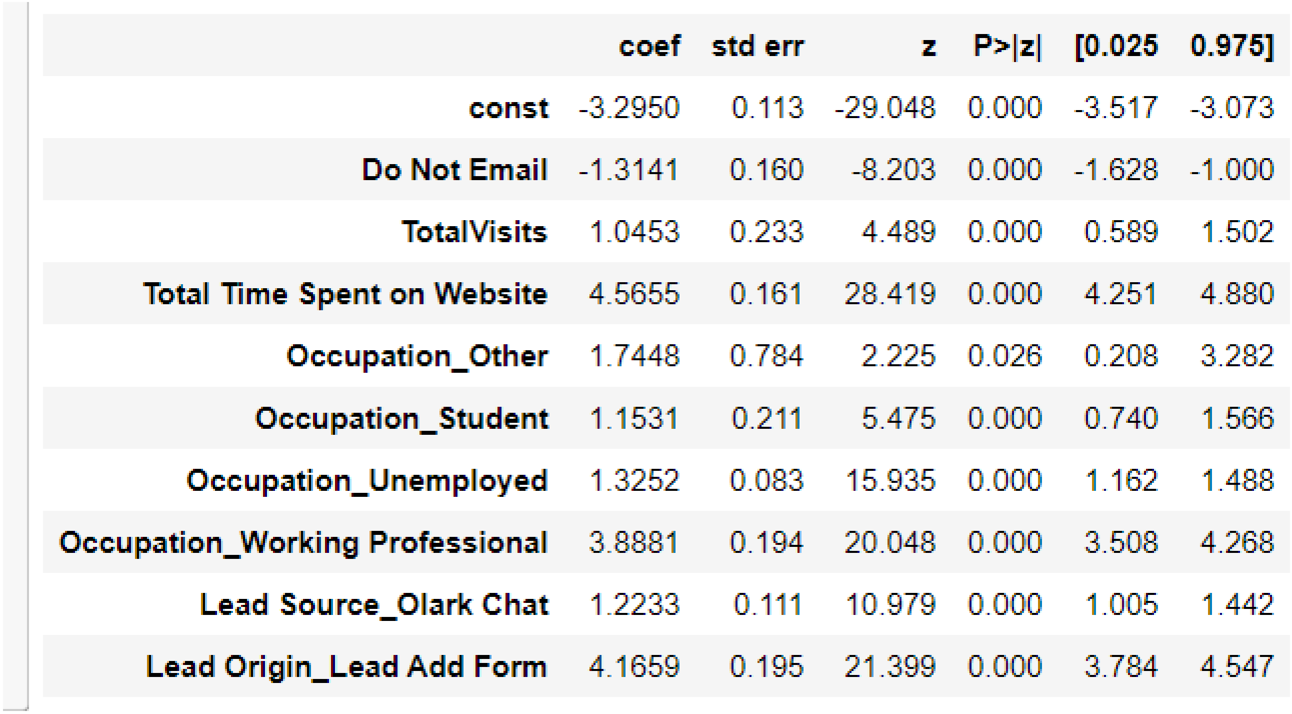




Though majority leads are the customers who have not specified their Specialization, the conversion rate is very low.

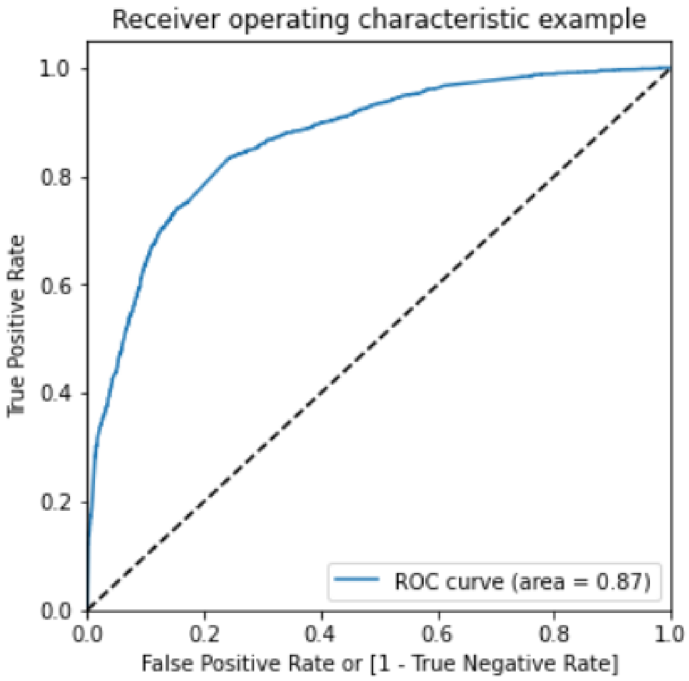
We can also see that lead conversion rate is higher for Management specializations (Finance, marketing, operations) as well as Banking & Insurance.

Variables Impacting the Conversion Rate

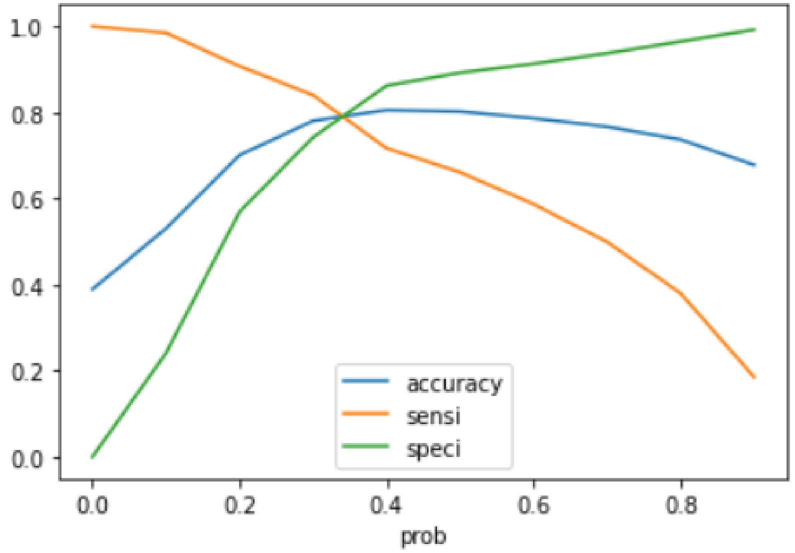


	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.2950	0.113	-29.048	0.000	-3.517	-3.073
Do Not Email	-1.3141	0.160	-8.203	0.000	-1.628	-1.000
TotalVisits	1.0453	0.233	4.489	0.000	0.589	1.502
Total Time Spent on Website	4.5655	0.161	28.419	0.000	4.251	4.880
Occupation_Other	1.7448	0.784	2.225	0.026	0.208	3.282
Occupation_Student	1.1531	0.211	5.475	0.000	0.740	1.566
Occupation_Unemployed	1.3252	0.083	15.935	0.000	1.162	1.488
Occupation_Working Professional	3.8881	0.194	20.048	0.000	3.508	4.268
Lead Source_Olark Chat	1.2233	0.111	10.979	0.000	1.005	1.442
Lead Origin_Lead Add Form	4.1659	0.195	21.399	0.000	3.784	4.547

ROC Curve and Optimal Cutoff Point



Area under ROC curve is 0.87



From the curve above, 0.31 is the optimum point to take it as a cutoff probability.

Metrics (Accuracy, Sensitivity, Specificity) on Train and Test data

Confusion Matrix

2971	962
405	2067

Train Metrics

Accuracy : 78 %
Sensitivity: 84 %
Specificity: 75 %

Confusion Matrix

1247	455
184	837

Test Metrics

Accuracy : 77 %
Sensitivity: 82 %
Specificity: 73 %

Inferences from the model

❑The Sensitivity, Accuracy and Specificity of the model is 82%, 77 and 73% respectively.

❑For this model we have considered higher sensitivity value to achieve better lead conversion rates. Due to higher sensitivity chances of missing the hot leads is lower.

❑The model achieves the target of 80% by predicting 82% of hot leads. Hence this is a good model for Education X company to improve their conversion rate.

❑Based on the coefficient values obtained from our final model, following are the top three variables that contribute most towards the probability of a lead getting converted :

1) Total Time Spent on Website

2) Lead Origin_Lead Add Form (dummy variable of original feature: Lead Origin)

3) Occupation_Working Professional (dummy variable of original feature: What is your current occupation)