

Lead Scoring Case Study

To Identify Hot-Leads from initial pool of leads for X
Education to improve conversion ratio

- Target Audience: Chief Data Scientist

Agenda

- Objective
- Background
- Action Plan
- Key Findings
- Model Evaluation and Results

Objective

- Assign lead score which will serve as a measure of lead quality.
- Identify Hot-Leads in order to improve conversion ratio to around 80%.
- Identify the important factors contributing to the successful conversion of a lead.

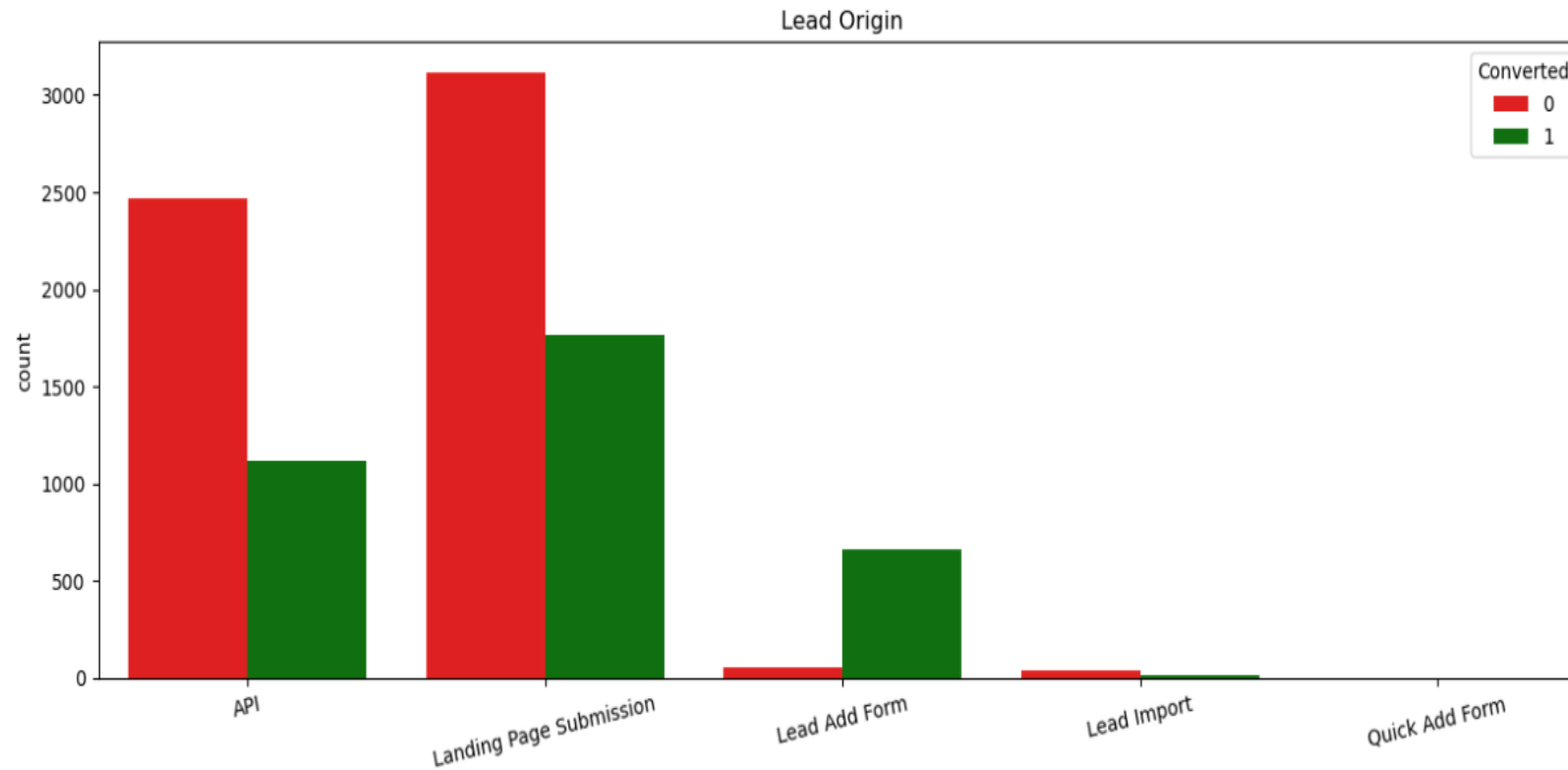
Background

- An education company, X Education, sells online courses to industry professionals by marketing its courses on several websites and search engines like Google.
- Once these people land on the website, they might browse the courses or watch some videos and then fill up a form for the course to be classified as a lead.
- 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%.

Action Plan

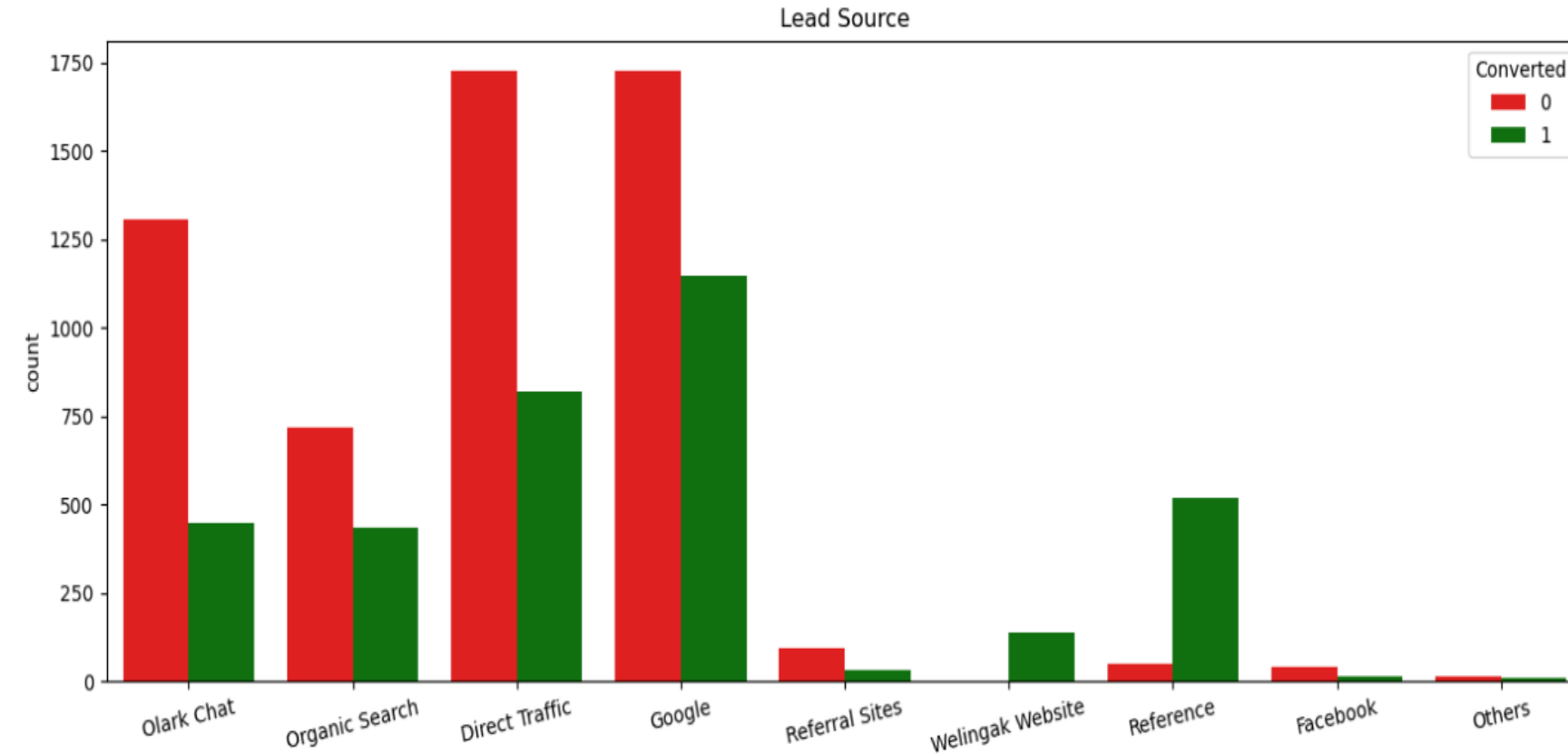
1. Reading and understanding data.
2. Cleaning the data.
3. Data Analysis and Visualisation.
4. Handling Outliers.
5. Data Preparation for model building
 - Test-Train split
 - Rescaling the features with Min-Max Scaler
6. Model Building and Feature selection
7. Finding Optimal Cut-off Probability
8. Making prediction on Test set
9. Assigning Scores based on probability of conversion

Lead Origin



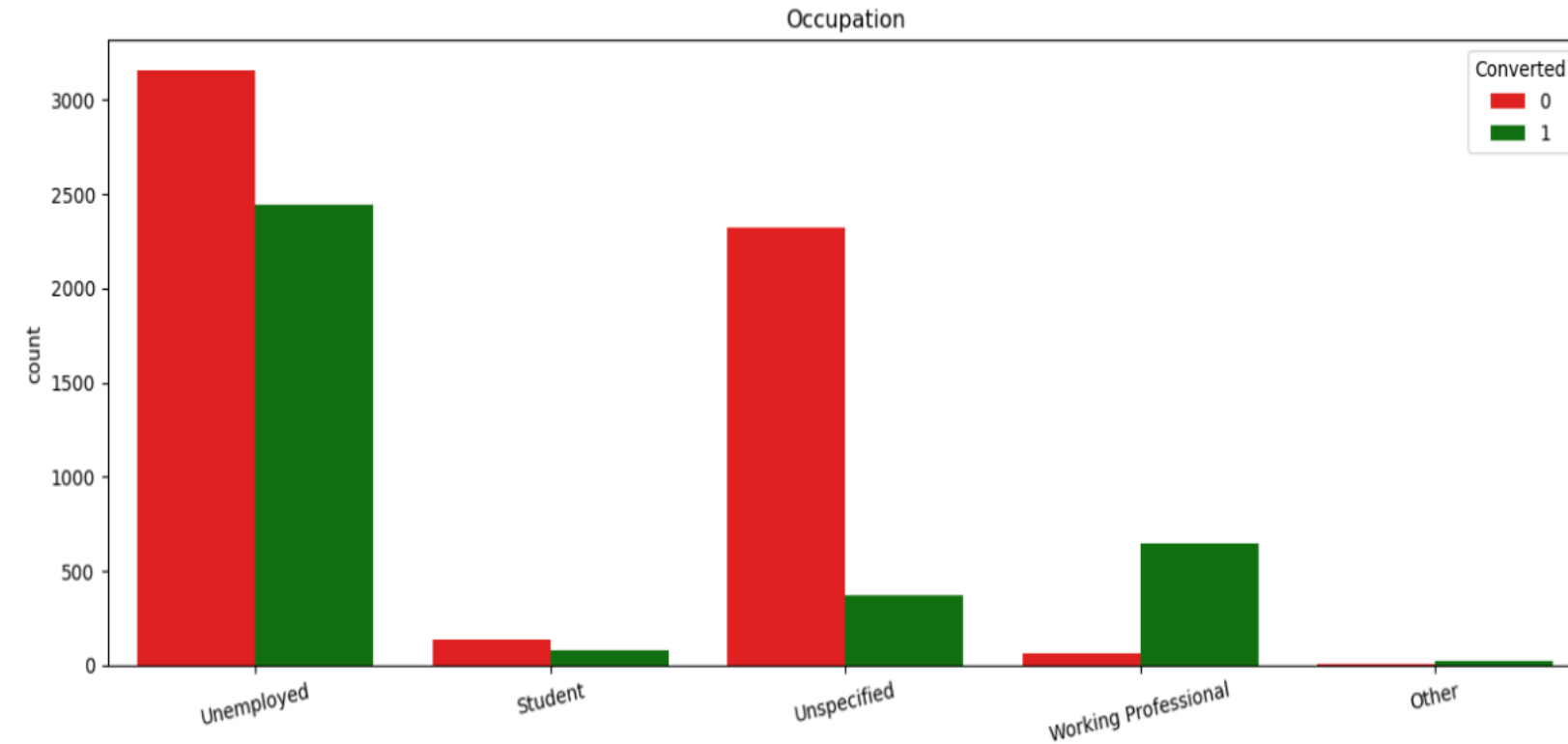
- "Landing page submission" and "API" have higher contribution to the leads.
- "Lead Add Form" have higher percentage of conversion.

Lead Source



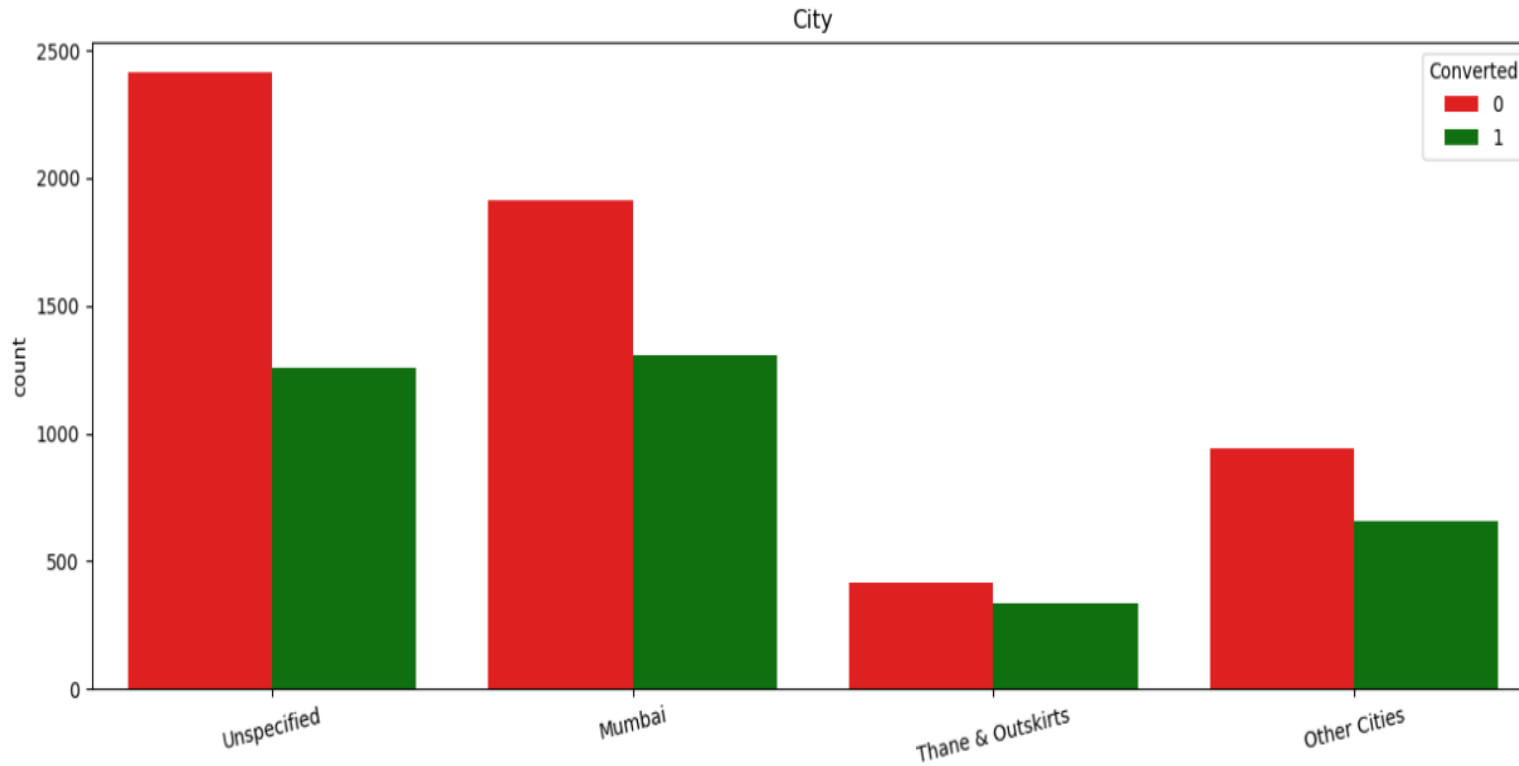
- "direct traffic" & "Google" have highest contribution
- "reference" and 'welingak website' ensures the conversion in most of the cases.

Occupation



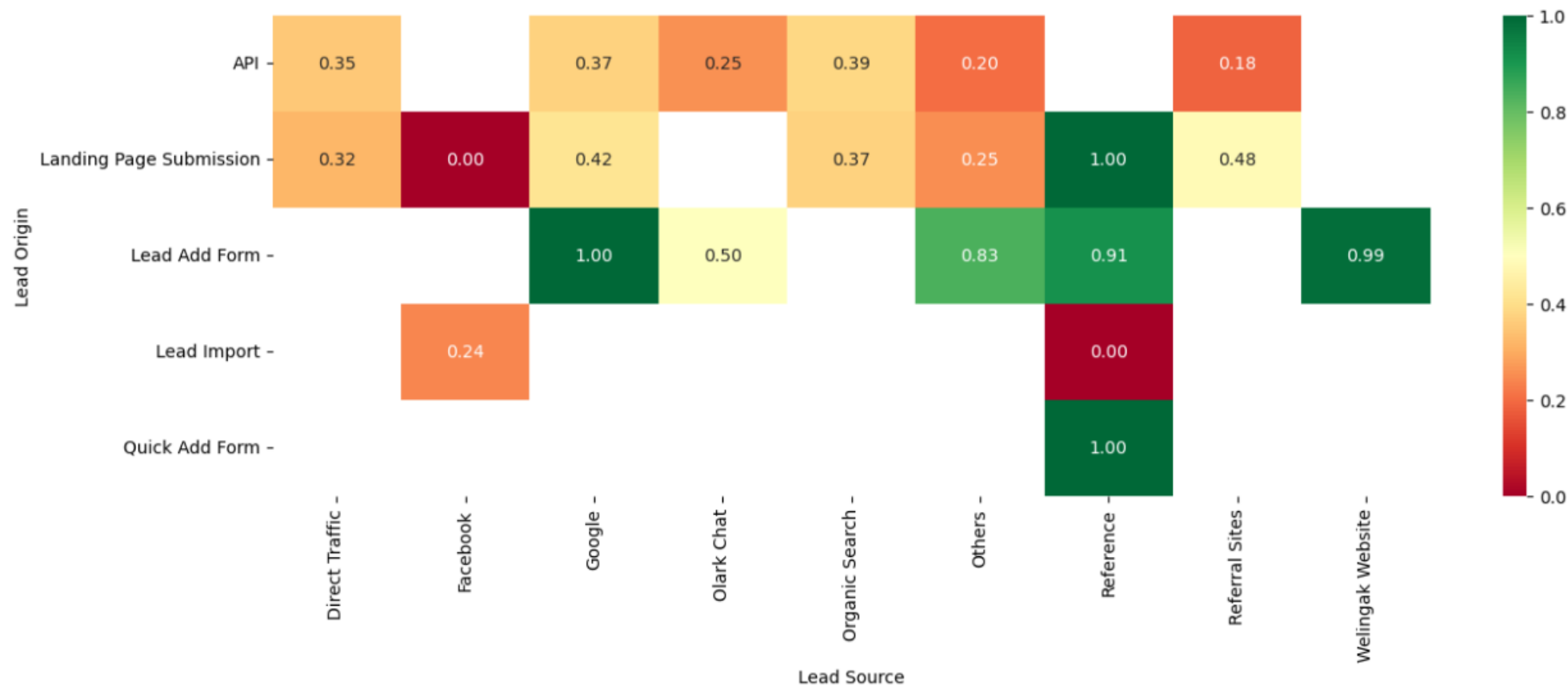
- A lot of clients are unemployed but the conversion ratio is around 40%
- There is quite high chance of conversion if client is employed (working professional).
- If the lead does not have 'Occupation' mentioned, then it's a major red flag.

City



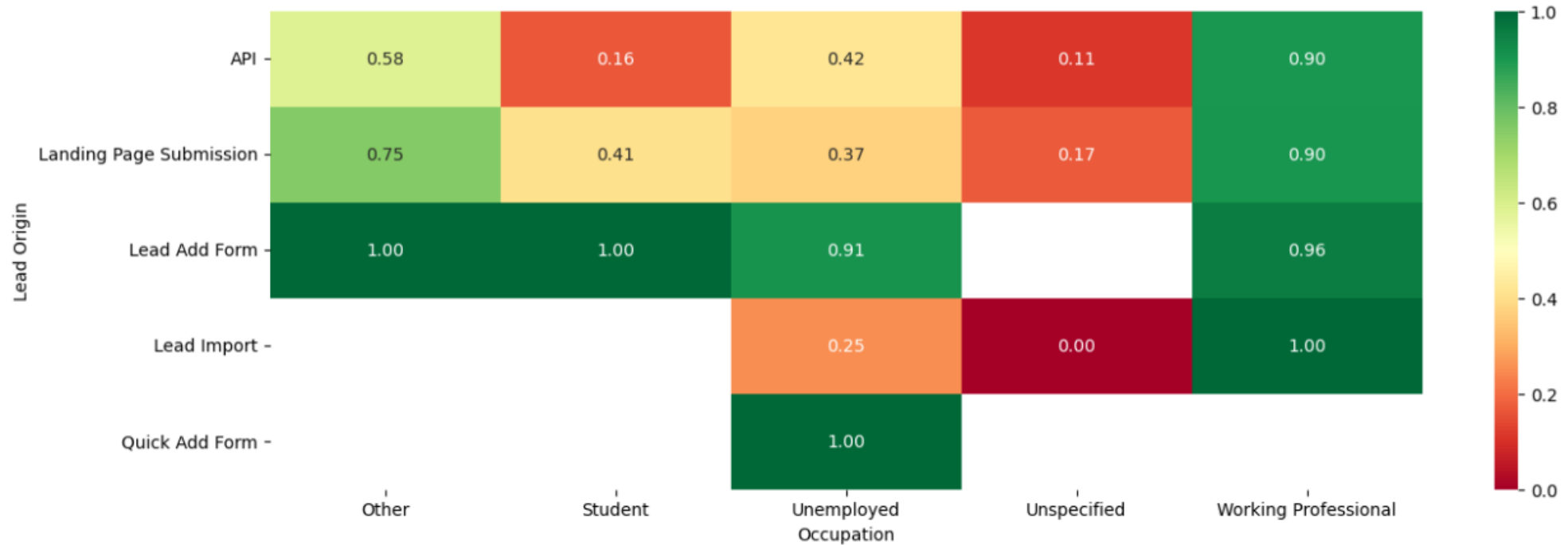
- Most of the people are from Mumbai.
- The conversion rate is quite similar for non-mumbai or other cities.
- If the city is not mentioned, then there is high chance that the lead won't get converted.

Lead Origin Vs Lead Source



- A Lead with Origin as 'Landing Page Submission' or 'Lead Add Form' with source as 'Reference' has good chance of conversion
- A Lead with Origin as 'Lead Add Form' with source as 'Google' or 'Welingak Website' has high chance of conversion.

Lead Origin Vs Occupation

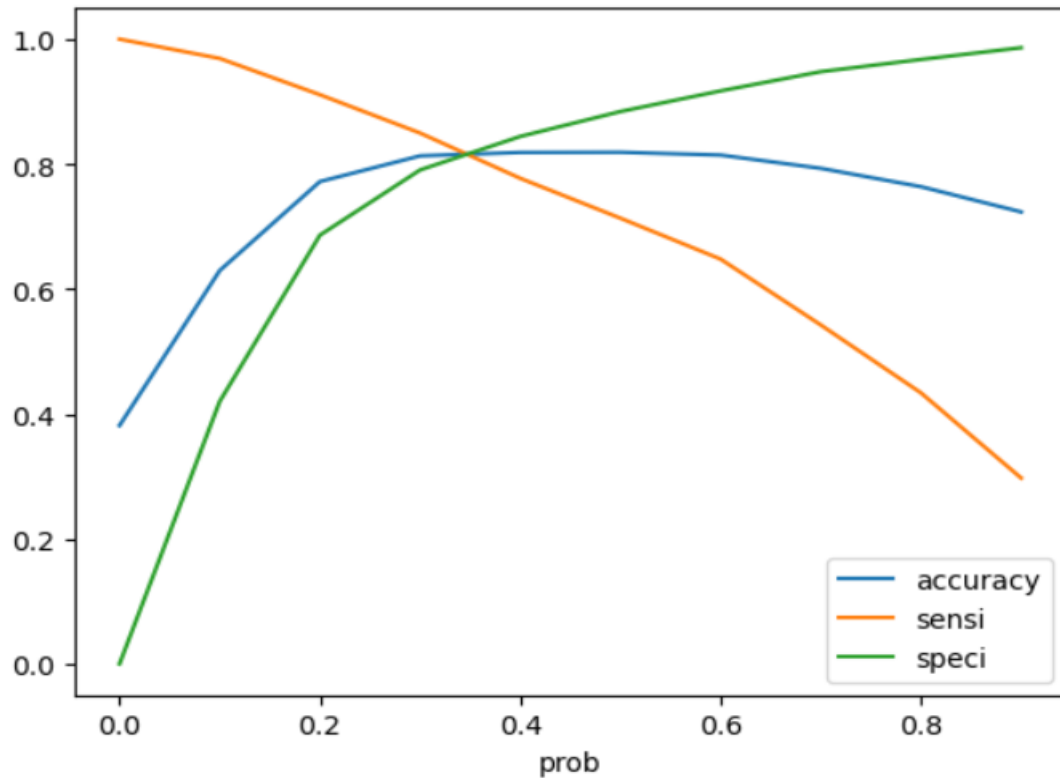


- Majority of conversions happen when Lead origin is 'Lead Add Form', 'Landing Page Submission' or 'API' across all 'Occupation' except 'Unspecified'.
- Leads in which Occupation is 'Unspecified', should be our last priority.
- Working professionals have almost perfect conversion ratio across all the Lead Origin

Important Features with Coefficients

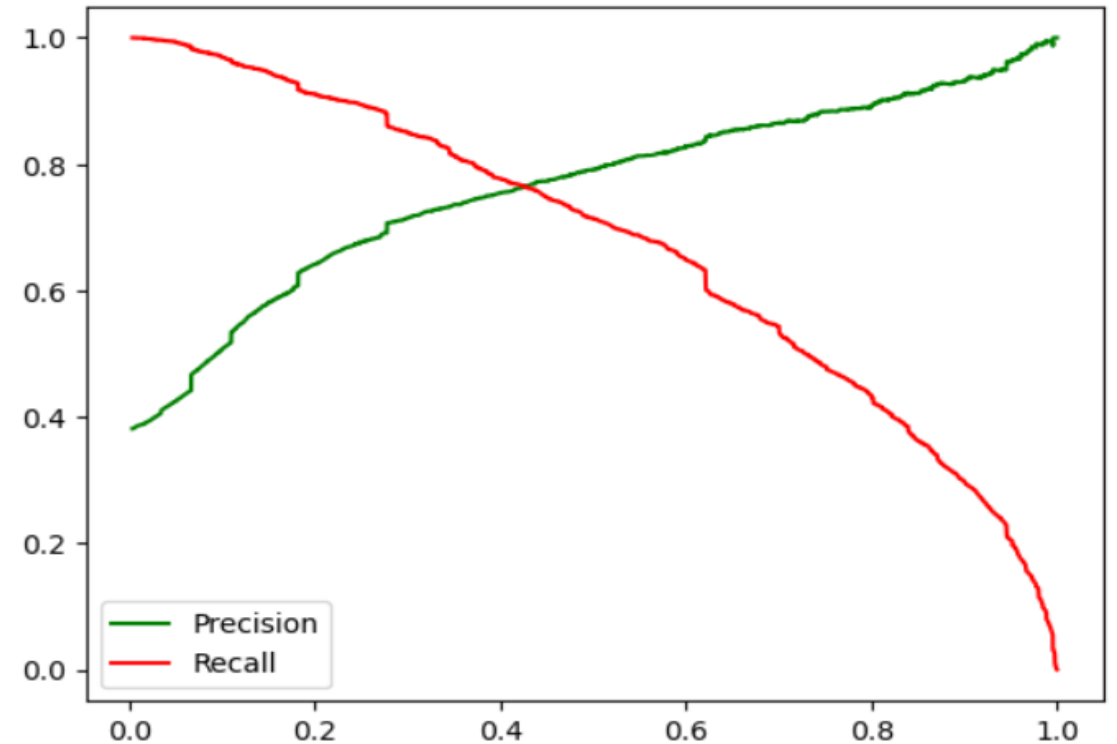
TotalVisits	0.962941
Total Time Spent on Website	4.484820
Do Not Email	-1.458675
Occupation_Other	2.764631
Occupation_Student	1.105638
Occupation_Unemployed	1.137331
Occupation_Working Professional	3.621710
City_Mumbai	0.256931
City_Other Cities	0.431175
City_Thane & Outskirts	0.513857
Lead_Origin_Lead Add Form	2.352516
Lead_Source_Direct Traffic	-1.797766
Lead_Source_Facebook	-1.522164
Lead_Source_Google	-1.257957
Lead_Source_Organic Search	-1.675021
Lead_Source_Referral Sites	-1.229667
Lead_Source_Welingak Website	3.039217
Last_Notable_Activity_Modified	-0.547565
Last_Notable_Activity_Olark Chat Conversation	-1.254379
Last_Notable_Activity_SMS Sent	1.450356
Last_Notable_Activity_Unreachable	2.037565
Last_Notable_Activity_Unsubscribed	1.355158

Optimal Cut-off & Results



From the curve above, 0.36 is the optimum point to take it as a intermediate cutoff probability

Accuracy - 81.73 %
Sensitivity - 80.60 %
Specificity - 81.43%



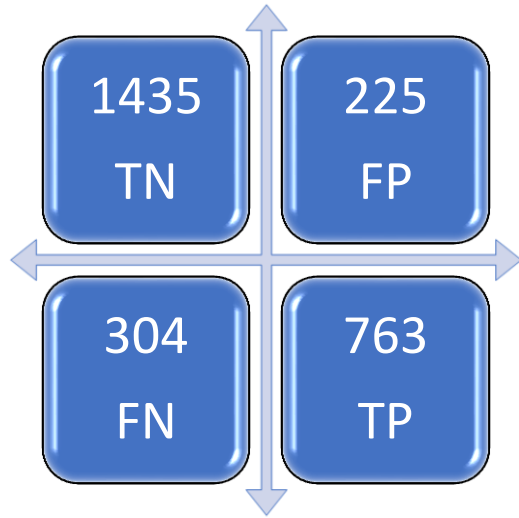
From precision-recall graph, we can take 0.42 as optimal final probability cutoff. but as our target is to predict around 80% converted leads, we can even choose cut-off higher than 0.42 to increase the precision of our model but at the cost of recall.

Lets choose cutoff as 0.47

Precision – 78.04 %
Recall – 73.51 %

Test Scores

Confusion Matrix



1435 TN	225 FP
304 FN	763 TP

Precision : 77.22

Recall : 71.50

- We were able to achieve around 77% precision and around 72% sensitivity on test data set.
- At the cost of 28% less business, we can reduce the work load by almost 64%.

Conclusion

Cut-off Lead Score - 47

- We were able to achieve around 77% precision and around 72% recall on test data set.
- So, initially if we focus only on leads which are predicted as potential leads by our model, then we can successfully convert almost 77% from them and these 77% of the leads will give us 72% of the current revenue/business.
- Also, the model has predicted that out of total leads, 36% will get converted, so initially we do not need to focus on rest of the leads, which reduces the workload by 64%.
- At the cost of 28% business, we can reduce the work load by almost 64%.
- Later on, as per the availability of workforce or time, we can focus on rest of the leads having lead score less than 47 and capitalize rest of the potential leads to increase the business/revenue. But it should be noted that the conversion ratio would gradually keep on decreasing with decrease in cut-off score.