Lead Score Assignment

Business Problem

An education company named X Education sells online courses to industry professionals. 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.

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 a higher lead score have a higher conversion chance and the customers with a 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%.

Solution Approach

As mention in Business problem statement, We need to assign a lead score to each and every lead on the basis of probability of conversion. On the basis of Lead Score, we need to classify whether a lead is a hot lead or not.

As it’s a classification problem, we need to use logistic regression model to classify the leads likely to be converted or not.

So, we have created a logistic Regression Model with below outcomes –

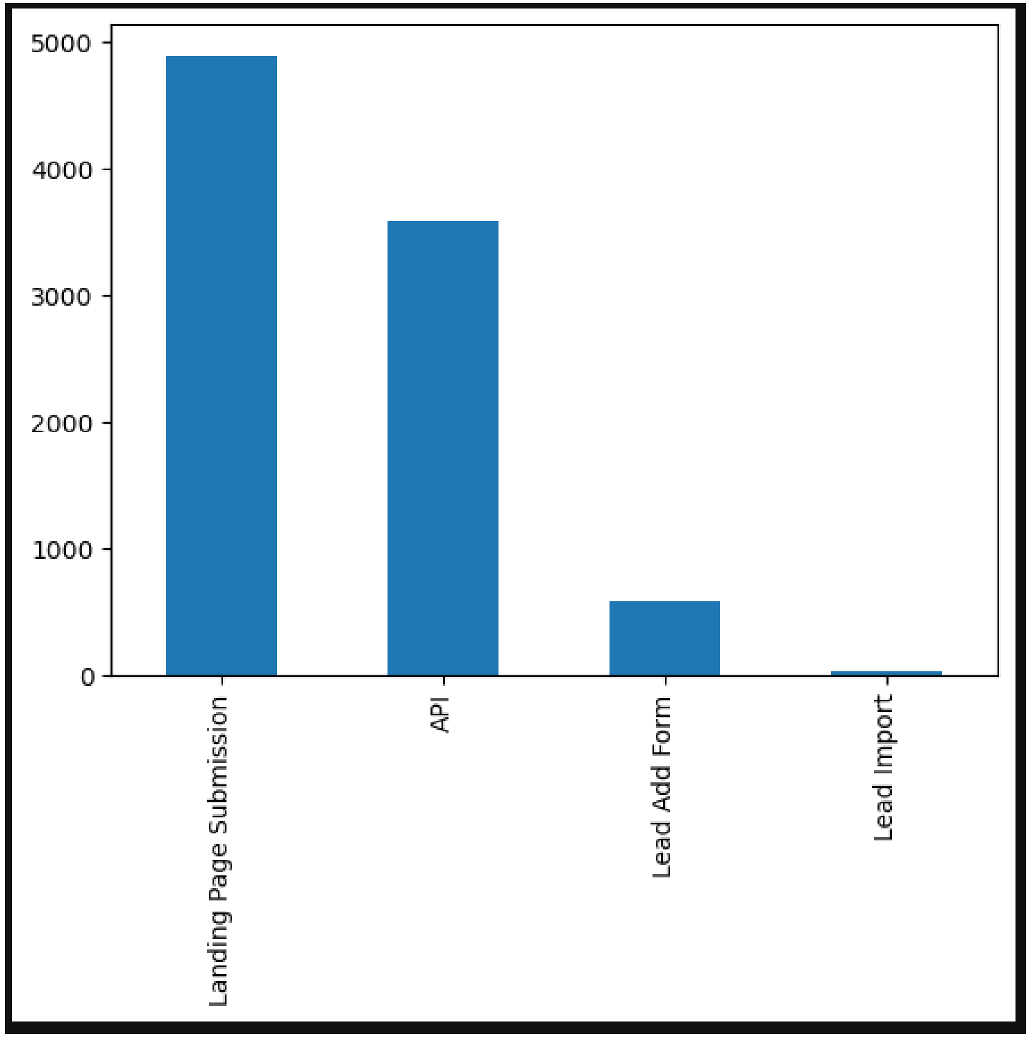
1- Cut-Off is 0.37
2- Accuracy is 0.9 on train data set and 0.89 on test dataset
3- Sensitivity is 0.88 on train dataset and 0.87 on test dataset
4- Specificity is 0.9 on train dataset and 0.89 on test dataset

We need to focus on two metrics i.e. Accuracy which defines the stability of the model and Sensitivity as in this case, we need to focus to minimize the False Negative. Hence, as good as our sensitivity metrics is, our model can identify more hot leads.

We follow the below steps to build our Machine Learning model.

1- Load the dataset and check Null values.
2- Treat Null values by deleting the columns which having Null values more than 40% or deleting rows having Null values less than 2%.
3- Replace "Select" values to NaN to those columns which having Select values
4- Treat Null values by deleting the columns or impute mean/mode values
5- Check and treat outliers
6- Univariate analysis- delete columns which having almost same values
7- Bivariate analysis
8- train and Test split
9- Scaling the features
10- Feature selection using RFE
11- Train the model (Iterative approach from step 11 to 13)
12- Check p-values and VIF
13- Delete columns having high p-values and VIF
14- Check Accuracy and other metrices
15- Evaluate the model on test dataset
16- Summary

Univariate Analysis:



The bar chart gives us an insight on how a customer is originated and turned to a lead.

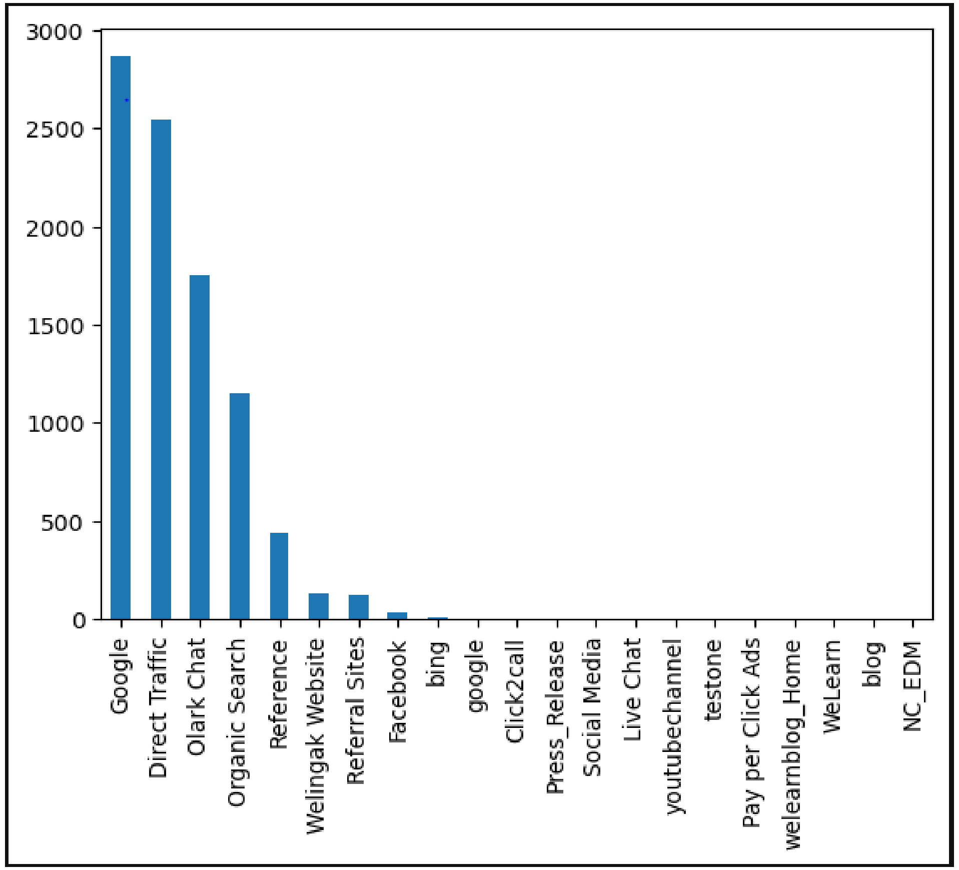
Overall it can be seen that the ‘Landing Page Submission’ has the highest contribution to identify a customer to be a lead, followed by ‘API’.

On the other hand ‘Lead Import’ has the lowest contribution to turn a customer to a lead.

When it comes to source of the leads we can clearly see that ‘Google’ generates the highest number of lead source.

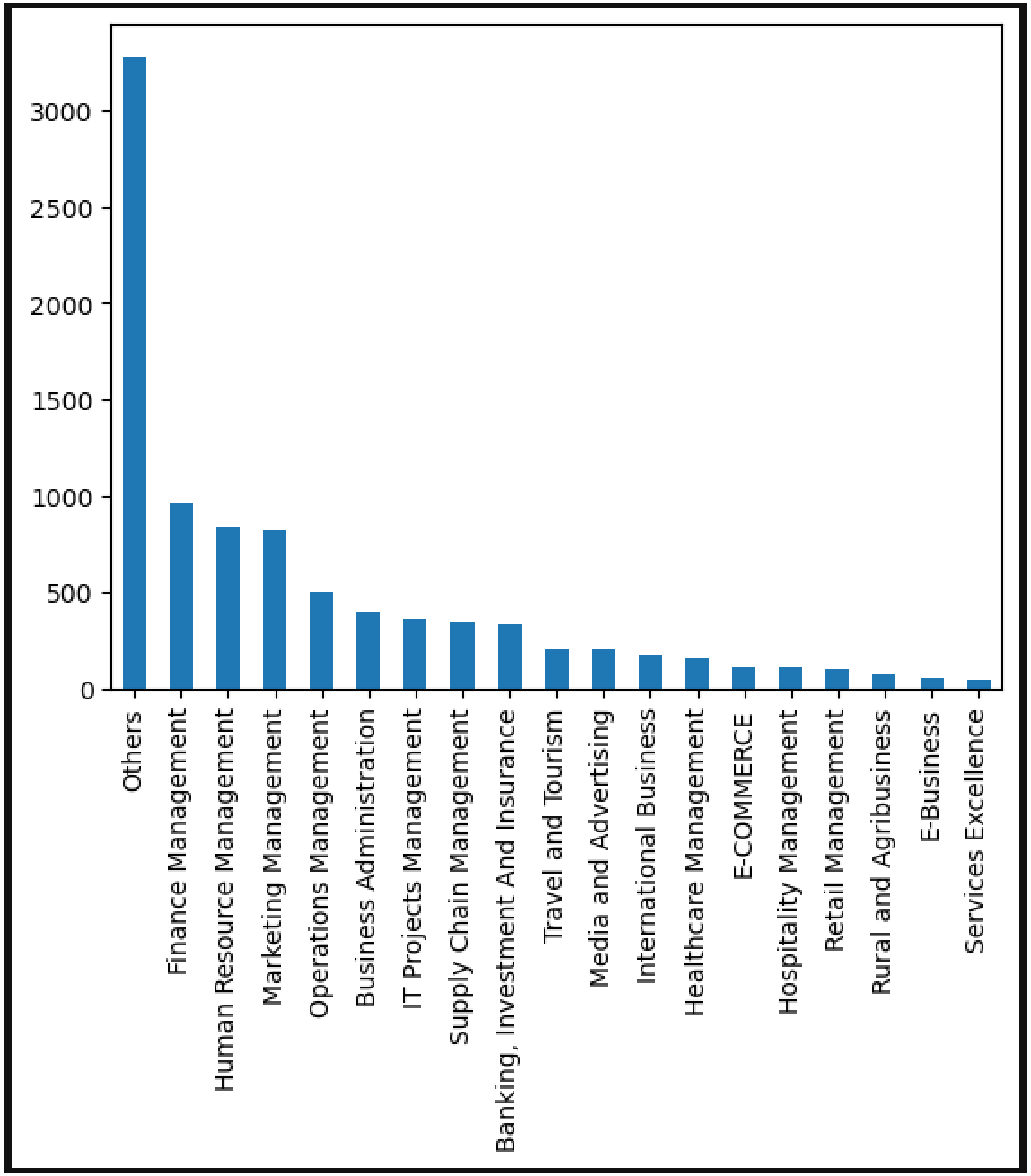
Whereas ‘Direct Traffic’, ‘Olark Chat’ and ‘Organic search’ has also contributed to generate leads.

However the least amount of leads are generated from ‘Facebook’ , ‘Bing’ and ‘Referral sites’

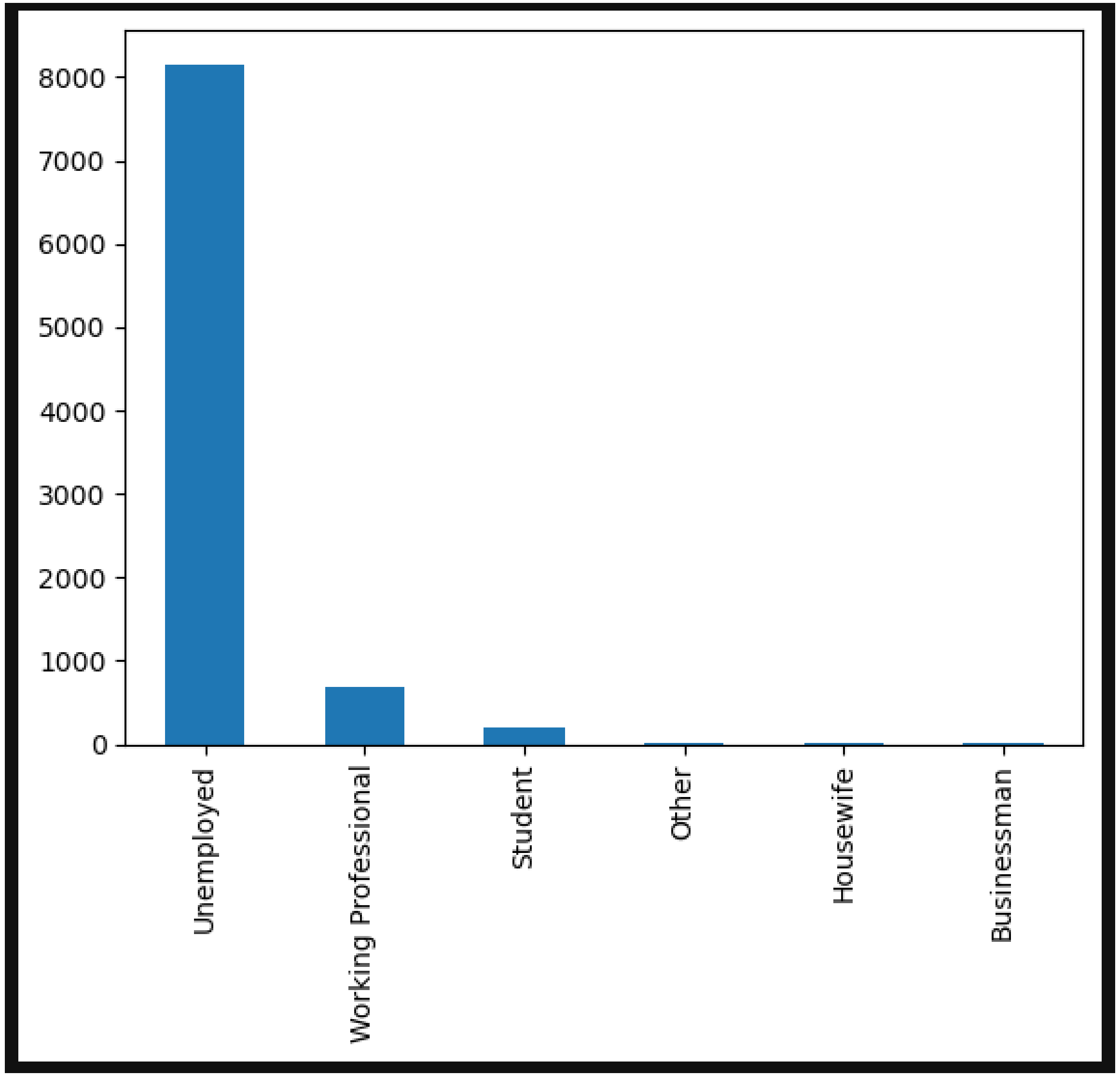


A good number of leads which we are receiving are from ‘Finance management’ and ‘Human Resource management’, followed by Marketing and Operations background.

Whereas the highest number of leads are from ‘Other’ specialization

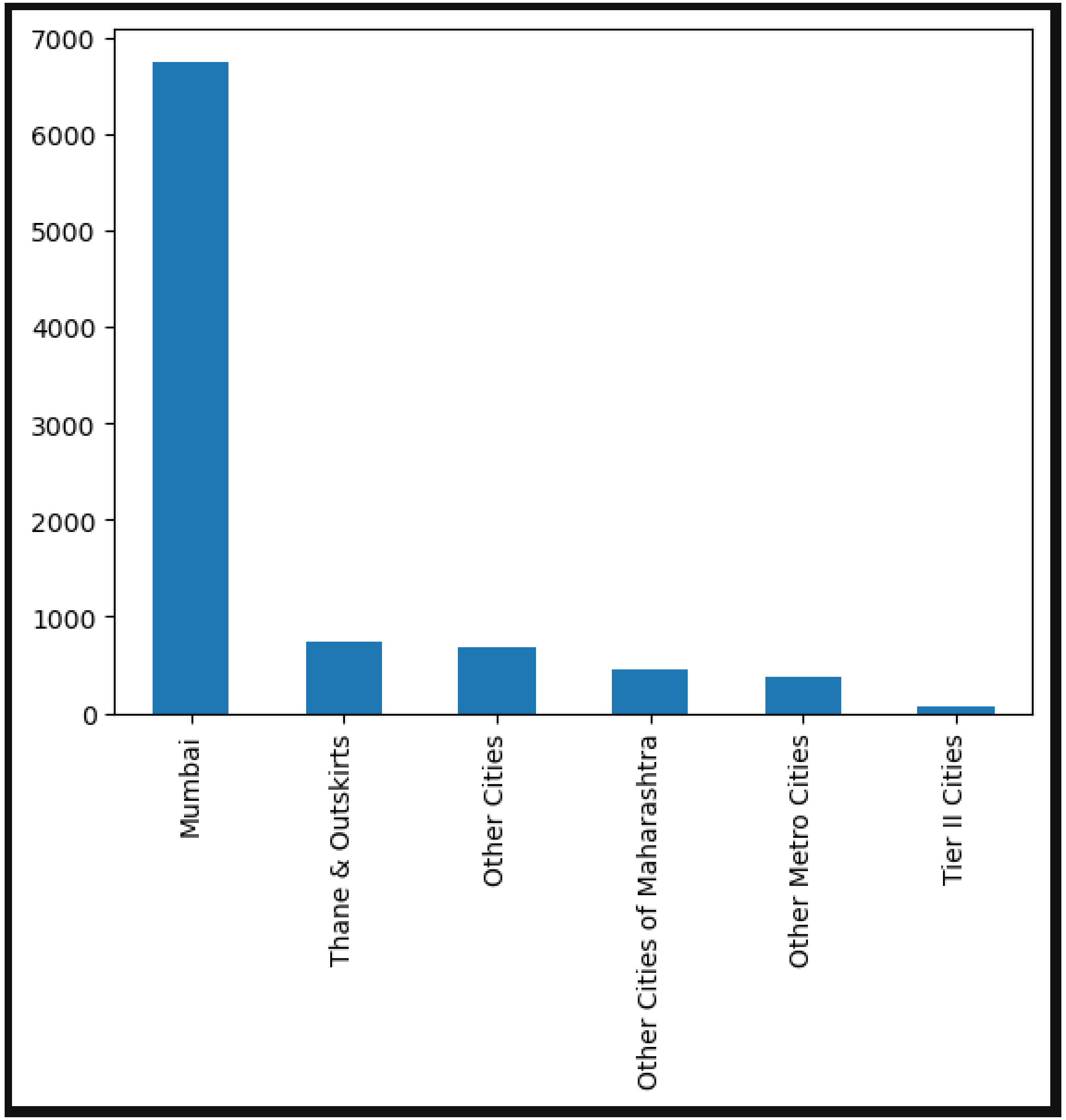


As we can see that the leads who are Unemployed are the highest in number and more interested to join our course subsequent to Working professional and Students



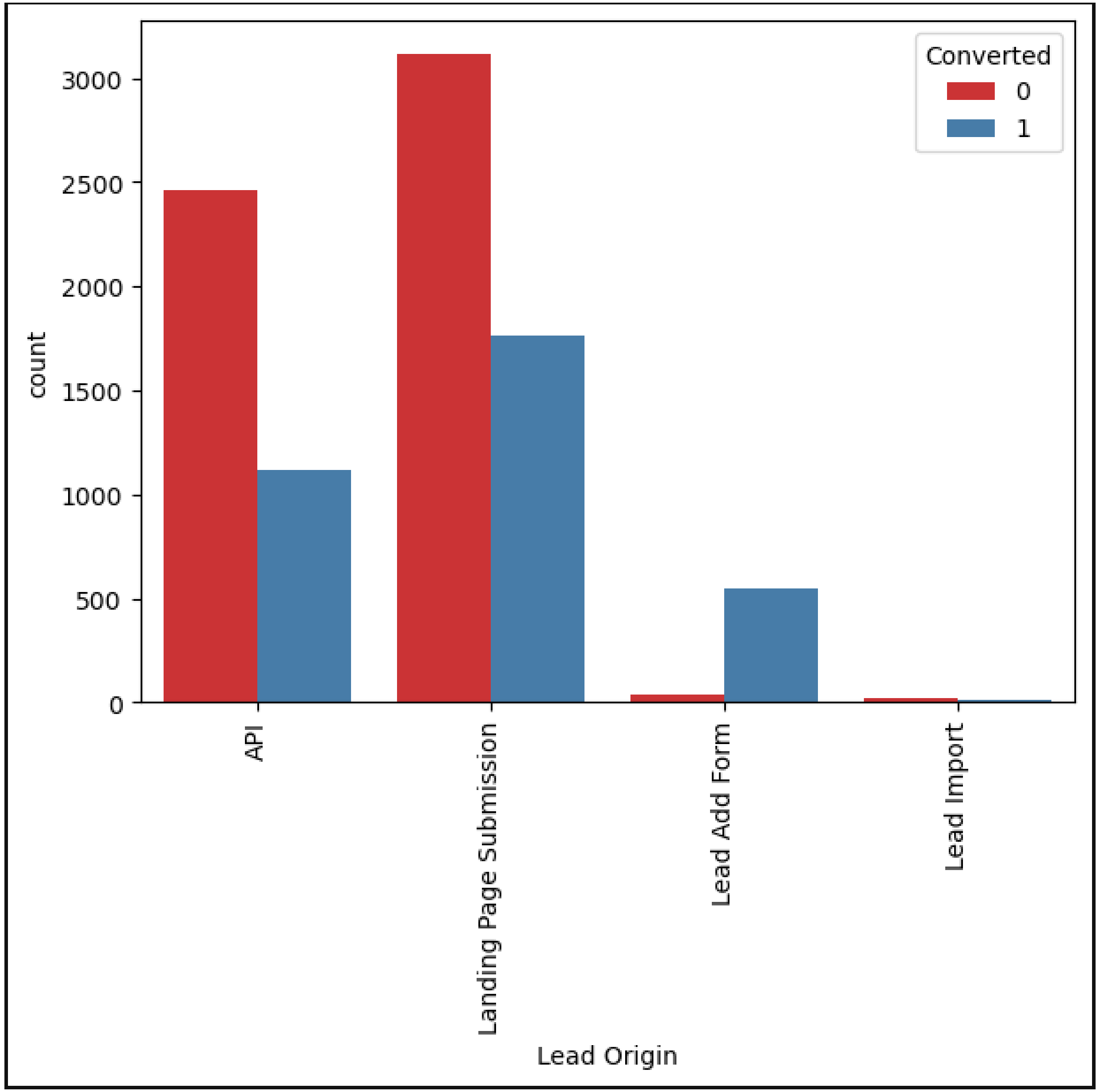
Most of the leads generated are from Mumbai city, and next in line are Thane and the outskirts of Thane.

Whereas other cities have approximately equal contribution of generating leads and Tier II cities has the least contribution of leads

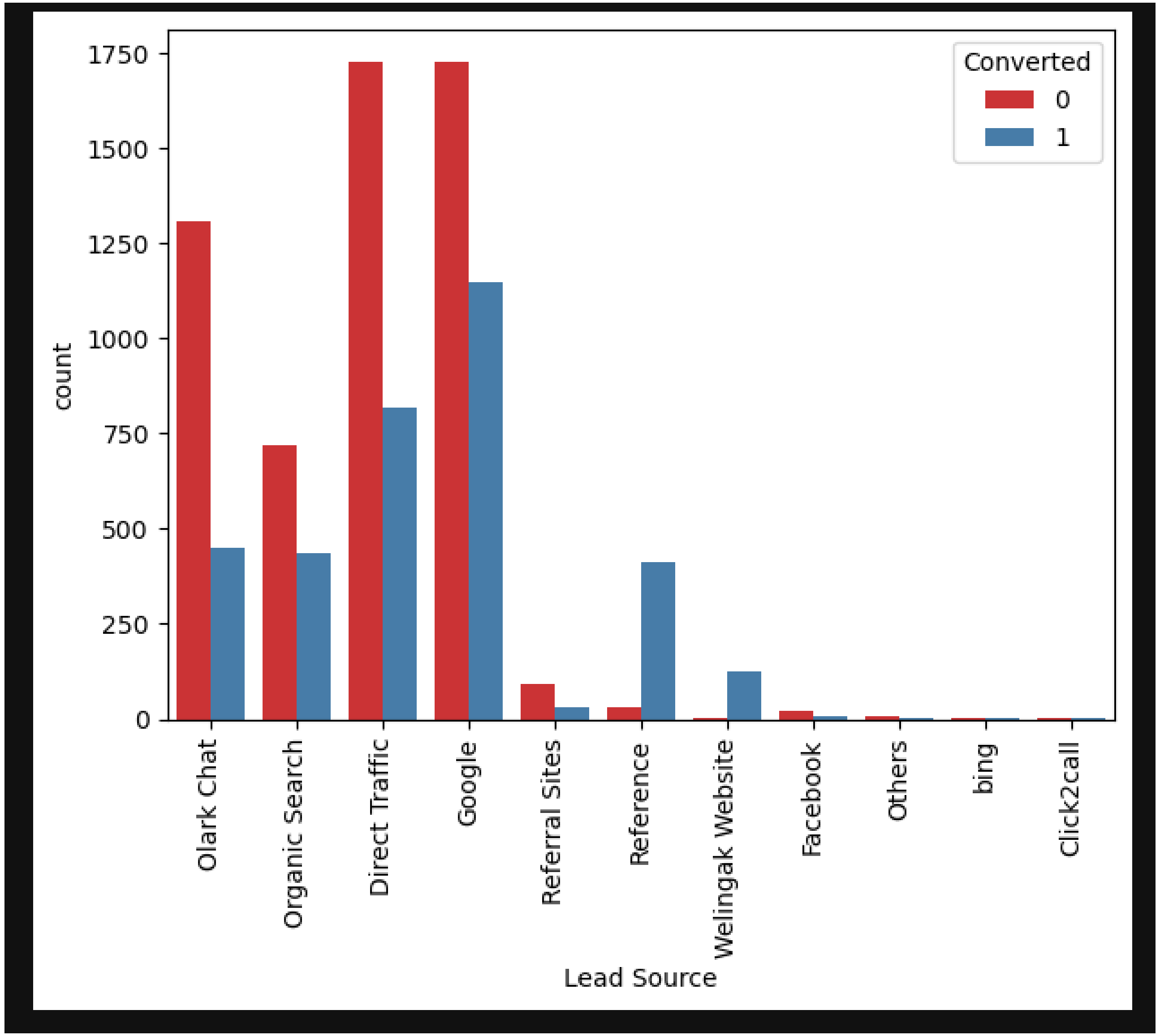


Bivariate Analysis :

The bar chart depicts that the leads generated from Lead Add form have very high conversion rate and are more likely to enroll as compared to leads generated from API and Land page Submission

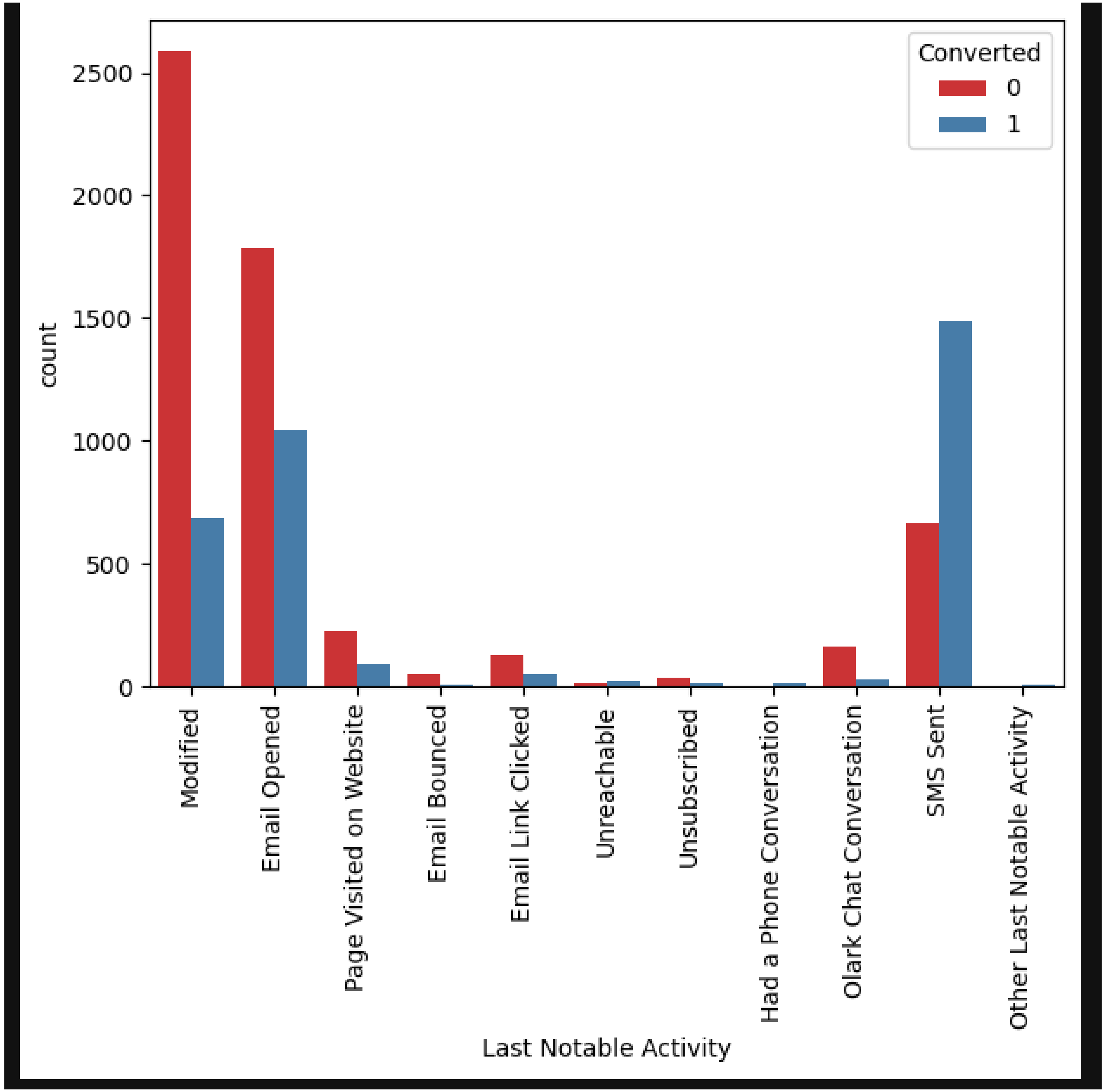


The conversion rate for Reference is the highest as per the chart i.e. leads who are referred by others have a good chance to enroll to our course as compared to other lead source



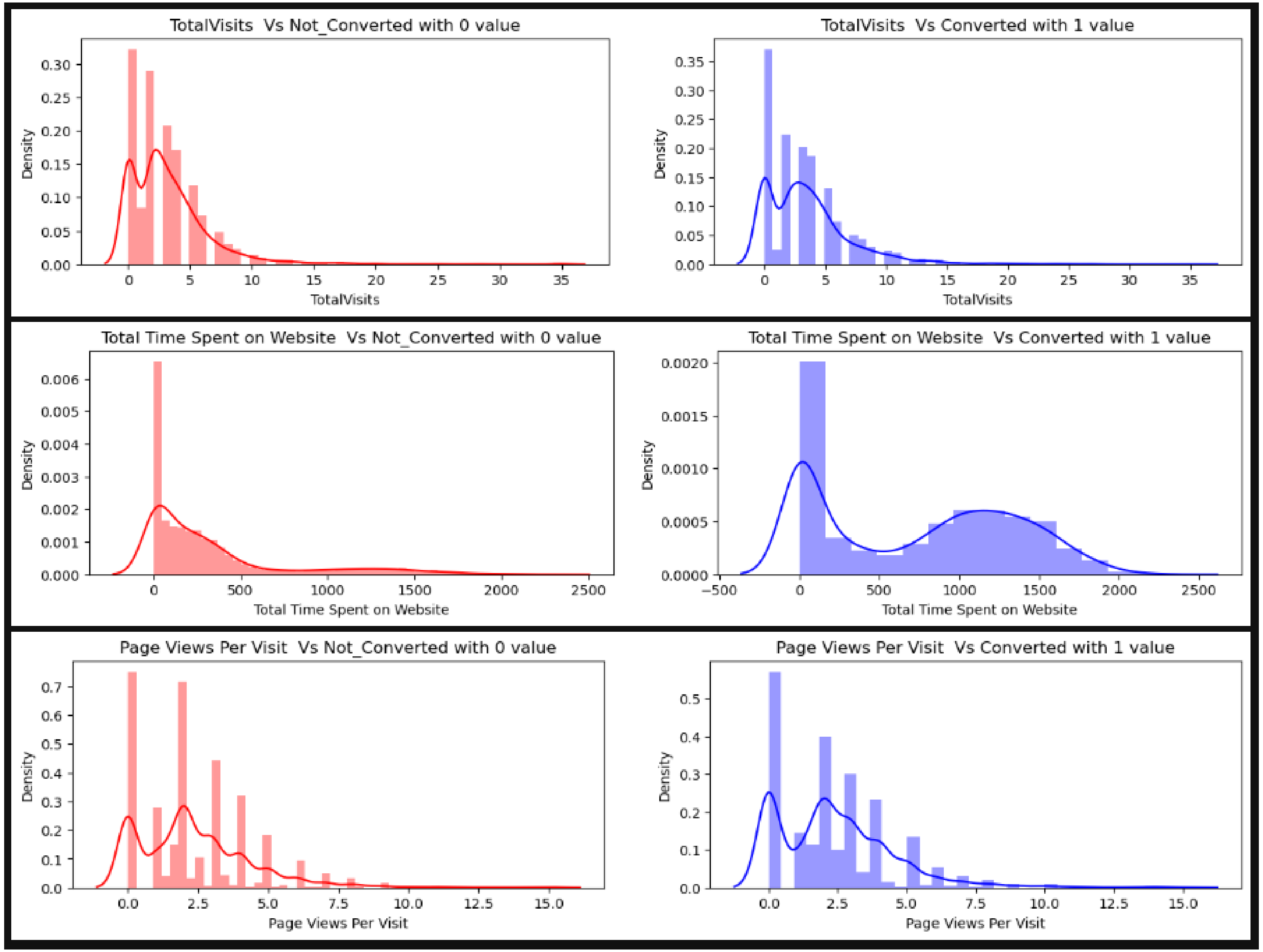
As we can see that when the customers are sent a SMS there is high chance that they will enroll to our course.

Where as the when it comes to other last activities performed by customer Email and Modified show a negative correlation.



As our analysis gives an insight that customer spending more time on website shows a high correlation with customer getting converted

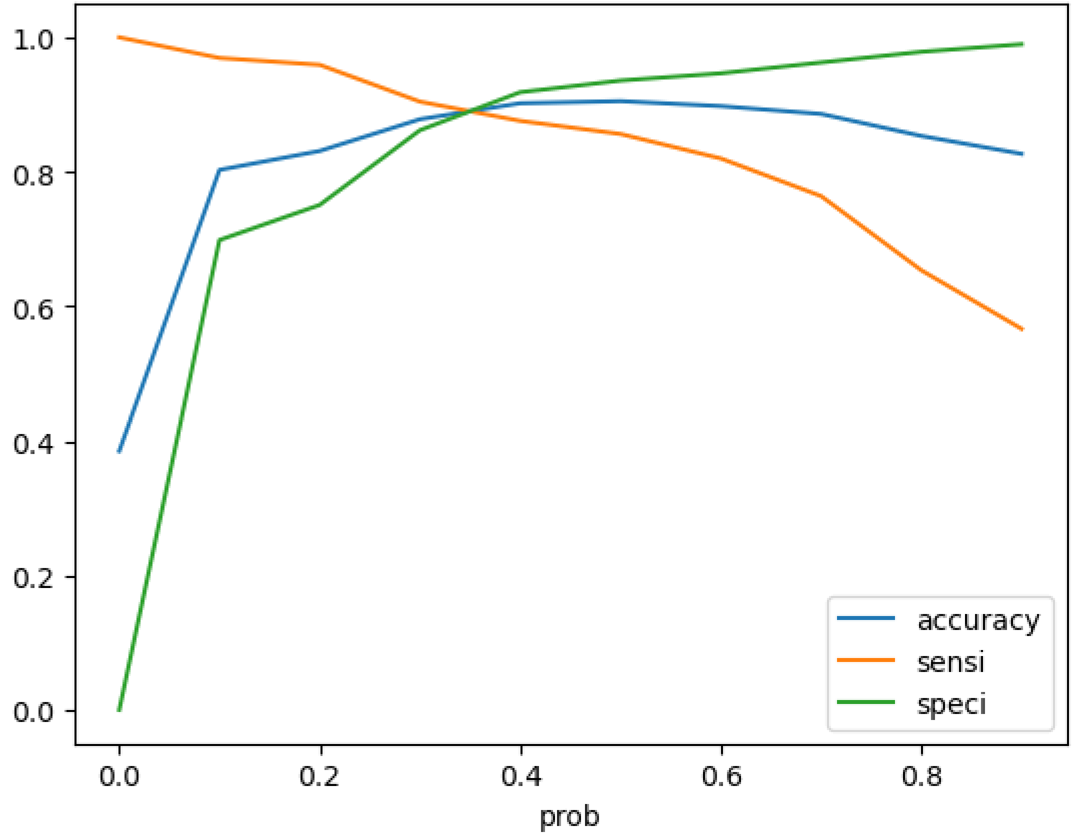
To summarize when a customer spends more time on our website there is a high chance that they will enroll to our course



Model Evaluation:

Our cut off is around 0.37 - Thus we can safely choose to consider any Prospect Lead with Conversion Probability higher than 37 % to be a hot Lead

Leads with predicted conversion probabilities above 37% are deemed more likely to convert, and therefore, they can be prioritized for further action, such as targeted marketing efforts or personalized follow-ups.



OBSERVATION:

Train Data:

- Accuracy: 90%
- Sensitivity: 88%
- Specificity: 90%

Test Data:

- Accuracy: 89%
- Sensitivity: 87%
- Specificity: 89%

Final features for our model

- 'Do Not Email'
- 'Total Time Spent on Website'
- 'Page Views Per Visit'
- 'Last_Notable_Activity_dummies__Modified'
- 'Last_Notable_Activity_dummies__Olark Chat Conversation'
- 'Lead_Quality__Not Sure'
- 'Lead_Quality__Worst'
- 'Tags__Busy'
- 'Tags__Closed by Horizzon'
- 'Tags__Lost to EINS'
- 'Tags__Ringing'
- 'Tags__Will revert after reading the email'
- 'Tags__switched off'
- 'Occupation__Working Professional'
- 'Specialization__Others'
- 'Last_Activity__Had a Phone Conversation'
- 'Last_Activity__SMS Sent'
- 'Lead Source__Welingak Website'
- 'Lead Origin__Landing Page Submission'
- 'Lead Origin__Lead Add Form'

CONCLUSION

- We see that the conversion rate is high for Lead Add form
- But very low for API and Landing submission comparatively. Therefore we can intervene that we need to focus more on the leads originated from Lead Add form
- We see max number of leads are generated by Google and Direct traffic. Max conversion ratio is by Reference
- Leads who spent more time on website, more likely to convert.
- Highest rate of conversion is when SMS is sent to the customer.
- Customers who are unemployed have a high chance of enrolling