Lead Score Case Study

Aishwarya khatri
Kumar Pratibadh
Praveen Thomas

Contents

Problem Statements

Objectives

Approach

Data Insights

Model Metrics

Conclusion

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.

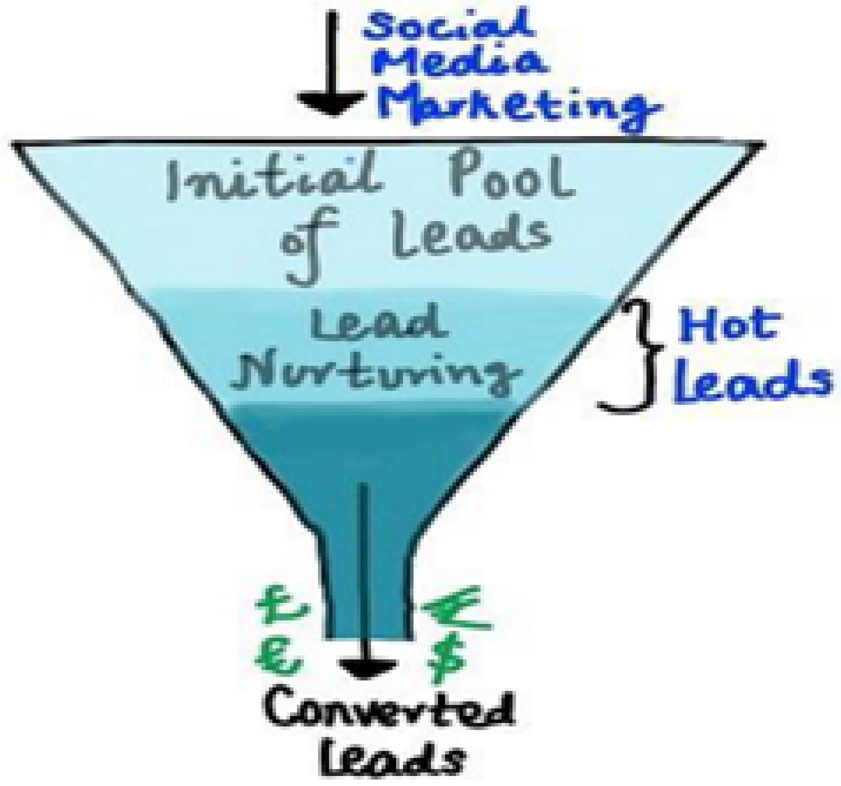


Fig. Lead Conversion Process

Objectives

- To assist the business in choosing the most qualified leads, also referred to as "Hot Leads," with a lead conversion rate of about 80%.
- To create a model in which a lead score is given to each lead, with higher lead scores representing customers who are more likely to convert and lower lead scores representing customers who are less likely to convert.
- Assist the sales team in refocusing on prospective leads and keeping them from making pointless phone calls.

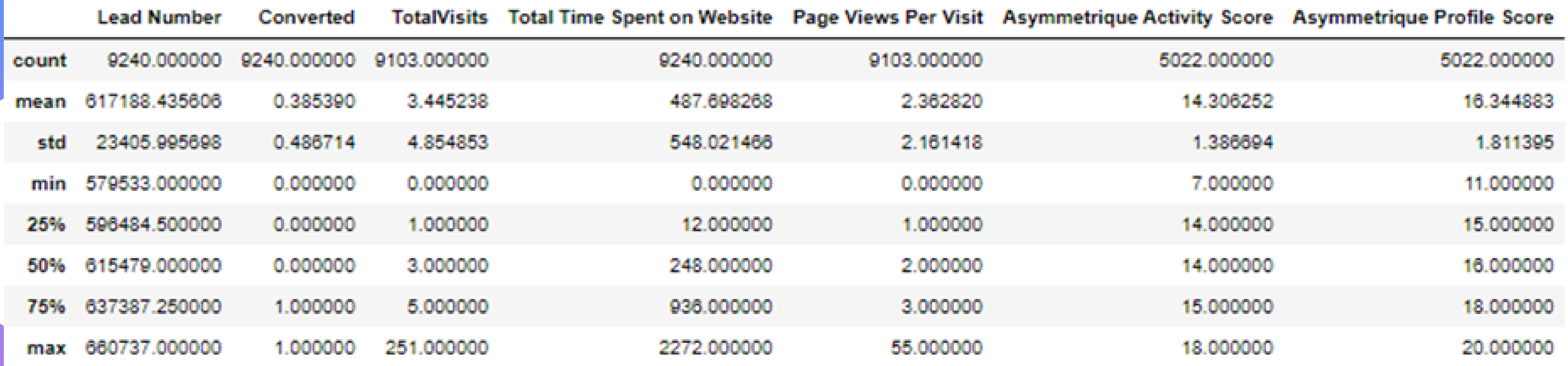
Approach

Analysing Patterns:

We have examined the patterns found in the dataset using exploratory data analysis, which has given us an intuitive sense of the features that will help in promoting lead conversion.

Driving Factors:

By examining the data below, we can get a sense of how the variables are distributed.



	Lead Number	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Asymmetrique Activity Score	Asymmetrique Profile Score
count	9240.000000	9240.000000	9103.000000	9240.000000	9103.000000	5022.000000	5022.000000
mean	617188.435606	0.385390	3.445238	487.698268	2.362820	14.306252	16.344883
std	23405.995698	0.486714	4.854853	548.021466	2.161418	1.386694	1.811395
min	579533.000000	0.000000	0.000000	0.000000	0.000000	7.000000	11.000000
25%	596484.500000	0.000000	1.000000	12.000000	1.000000	14.000000	15.000000
50%	615479.000000	0.000000	3.000000	248.000000	2.000000	14.000000	16.000000
75%	637387.250000	1.000000	5.000000	936.000000	3.000000	15.000000	18.000000
max	660737.000000	1.000000	251.000000	2272.000000	55.000000	18.000000	20.000000

Correlations:

Finding correlations between variables will help you determine data variability and the key characteristics that can help convert leads.

Recommendations:

Pay attention to characteristics that might increase lead conversion rates.

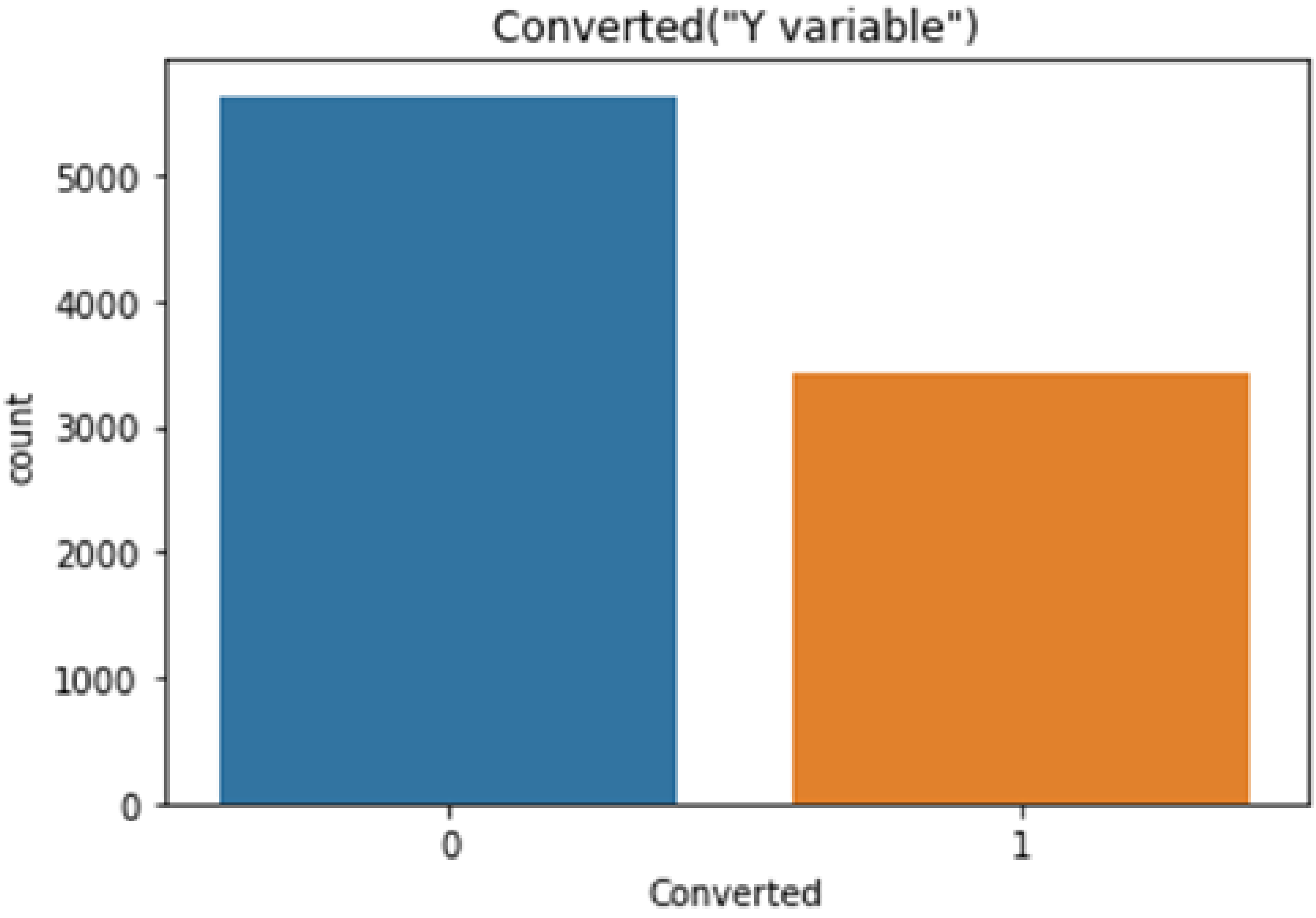
Data Insights

We need to determine which entries among the 9240 total entries of unique customers have the best likelihood of being converted.

i)Decision Criteria:

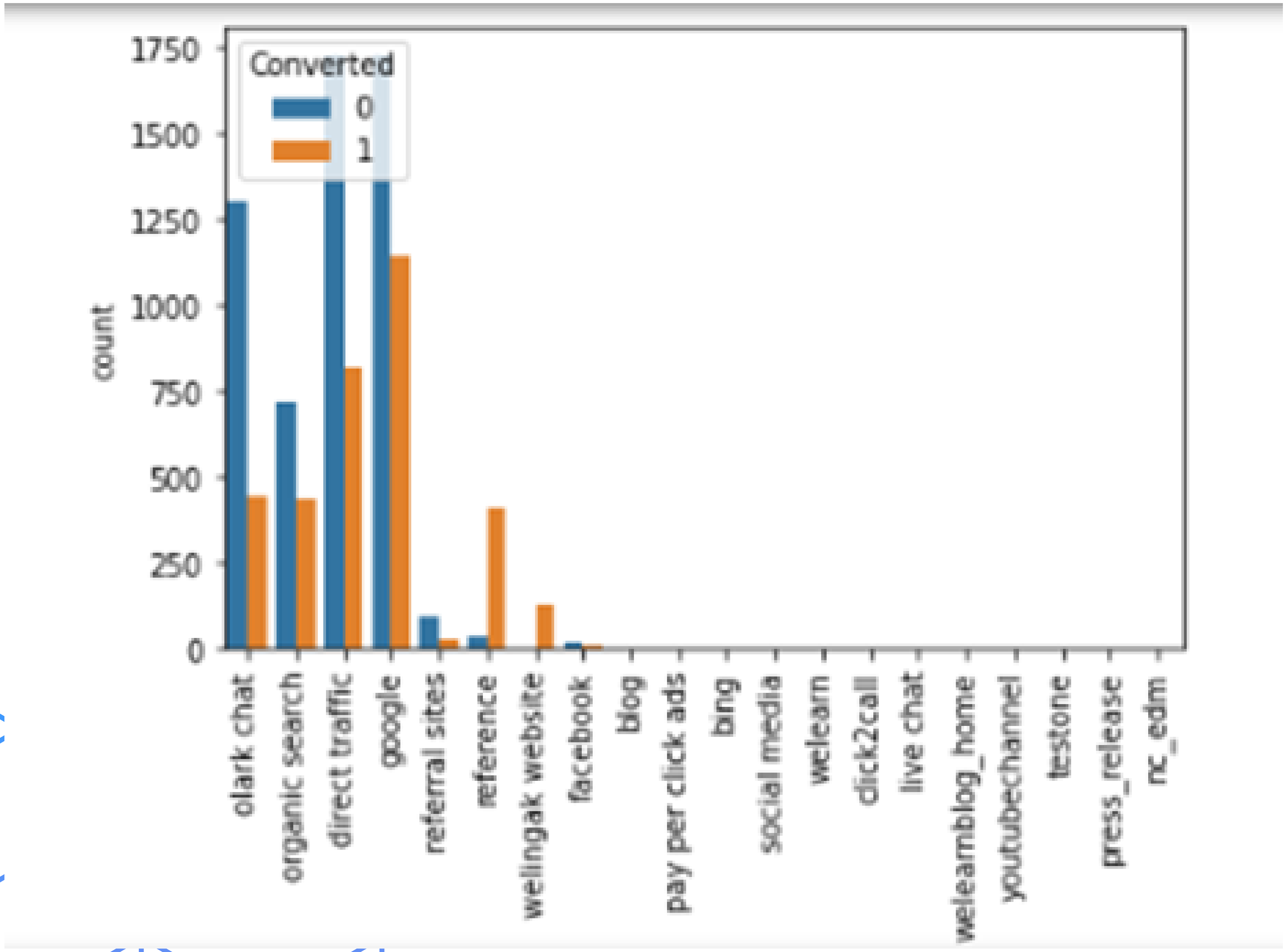
We must choose the submissions with the highest chances of conversion from the 9240 unique customer entries in total.

- Of the 9240 entries, about 37.8% of leads are converted, whereas 73% of leads are not.



37.8% of the data marked as "Converted" is 1, meaning that 37.8% of the leads were converted. This indicates that we have adequate lead conversion data for modelling.

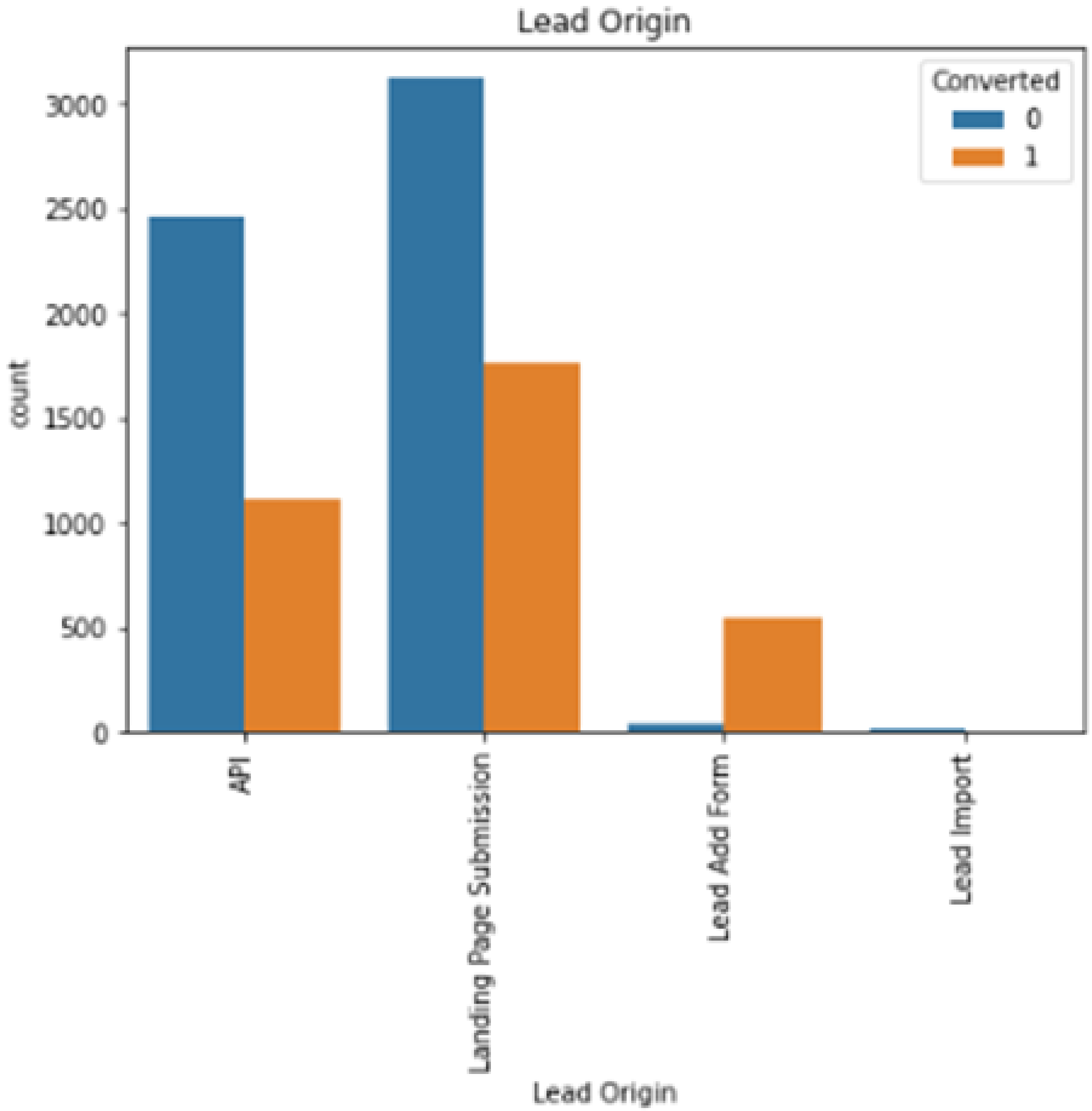
ii)Lead Origin:



"Direct Traffic" and "Google" produce the most leads, while "Reference" and "Welingak Website" have the highest conversion rates.

iii)Examine the distribution of categorical columns relative to converted columns.

- The majority of the leads come from customers who were classified as leads via landing page submission.
- Customers who came from a Lead Add Form are more likely to convert. There are not many of these clients.
- The lowest conversion rates are for lead import and lead origin-API. There are extremely few Lead Import clients.

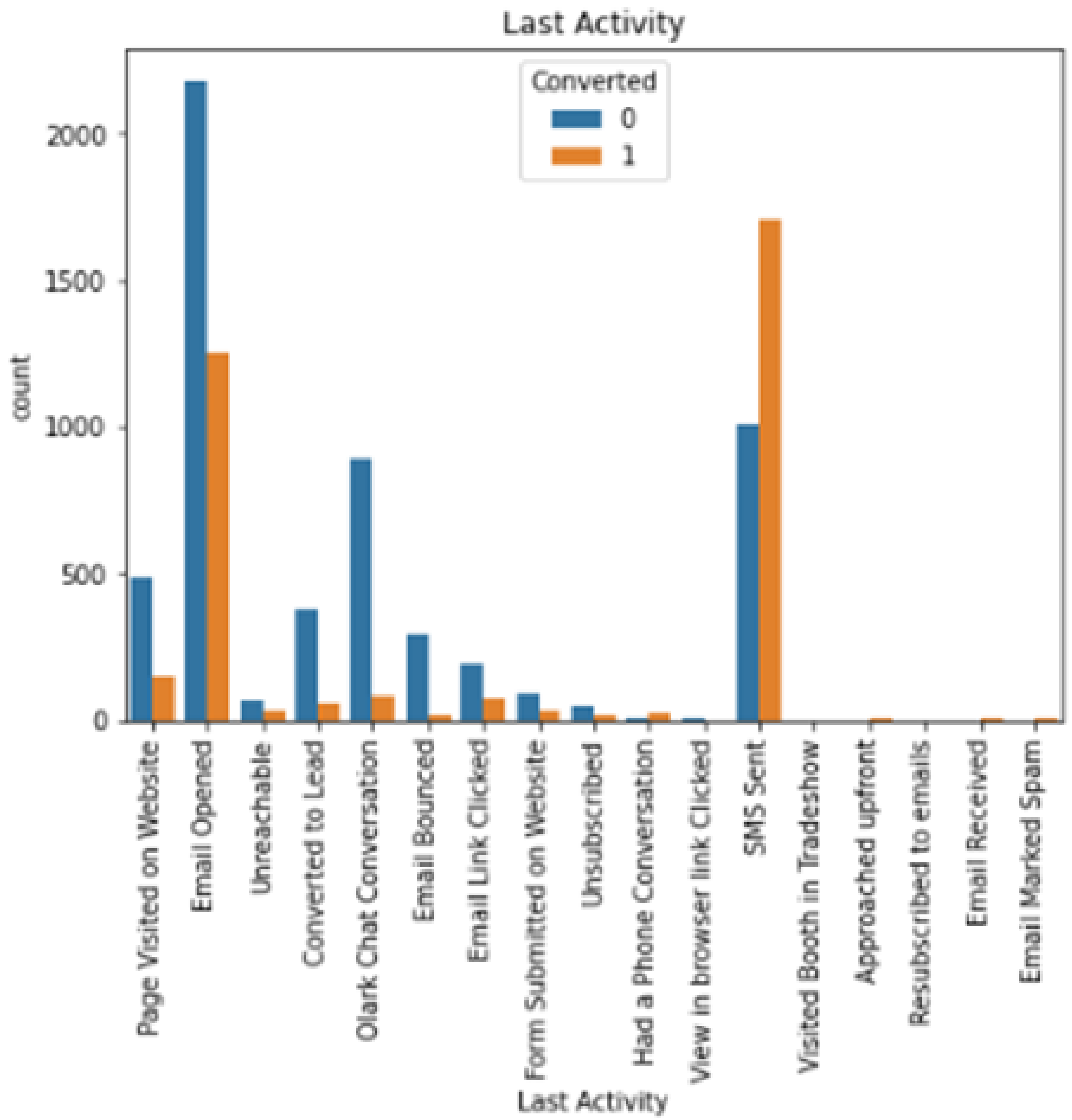


The most leads are generated by "API" and "Landing Page Submission," but these methods only convert about 30% of those leads. While the "Lead Add Form" generates less leads, it has a high conversion rate. In order to generate more leads, we should aim to boost the conversion rates for "API" and "Landing Page Submission" and "Lead Add Form." 'Lead Import' doesn't appear to be all that important.

iv) Last Activity:

- The conversion rate for customers whose most recent activity was sending an SMS is greater, at about 63%.
- The majority of consumers had their email opened as their most recent action. They convert at a rate of about 36%.

Observations for Last Activity:



- The final action that generates the most leads is "Email Opened," but "SMS Sent" has the highest conversion rates. It converts at a remarkably rapid pace.
- The categories that follow "SMS Sent" have a very small impact. They may all be combined into a single category

Observations for EDA

- To analyse the data using visual approaches and how they respond to the target field, an exploratory data analysis was performed (Converted). With the use of statistical summaries and graphical representations, it was used to identify trends, patterns, or to verify assumptions.

- It was found that a lot of attributes in the categorical variables were irrelevant.

- The numeric values had outliers which were handled.

Observations for EDA

- To analyse the data using visual approaches and how they respond to the target field, an exploratory data analysis was performed (Converted). With the use of statistical summaries and graphical representations, it was used to identify trends, patterns, or to verify assumptions.

- It was found that a lot of attributes in the categorical variables were irrelevant.

- The numeric values had outliers which were handled.

Model Metrics:

i)Train-Test split of Data:

The split was done at 80% and 20% for train and test data respectively.

i)Model Building:

- Firstly, RFE was done to attain the top 15 relevant variables.
- Secondly VIF values and p-value were used to remove few more variable which were creating influence (The variables with VIF < 5 and p-value < 0.05 were kept).

ii)Model Evaluation:

- Confusion matrix and ROC was created for train Data, then calculation was made to find out the accuracy, sensitivity and specificity which came to be around

Train Data:

- Accuracy : 79.94%
- Sensitivity : 82.00%
- Specificity : 78.00%

Test Data:

- Accuracy : 80.77%
- Sensitivity : 82.00%
- Specificity : 80.00%

Conclusion

Top four features that indicates hot leads are –

- Total Time Spent on Website
- Lead Origin is lead add form
- Lead Source is welingak website
- Current occupation is Professional