Lead Scoring Case Study

Submitted by:
Ethish M
Vamsi K

Problem statement

- X Education is an online course provider for industry professionals. The company markets its courses on several websites and search engines like Google.
- Once a potential customer lands on the website, they can browse the courses or fill out a form to receive more information. If they provide their email address or phone number, they are classified as a lead. X Education also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls and writing emails to them.
- Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X Education is around 30%.
- When the sales team is able to identify a specific group of leads, it is expected that the lead conversion rate will increase. This is because the team will be able to concentrate their efforts on communicating with potential leads rather than spending time making calls to a broad range of individuals.

Objective

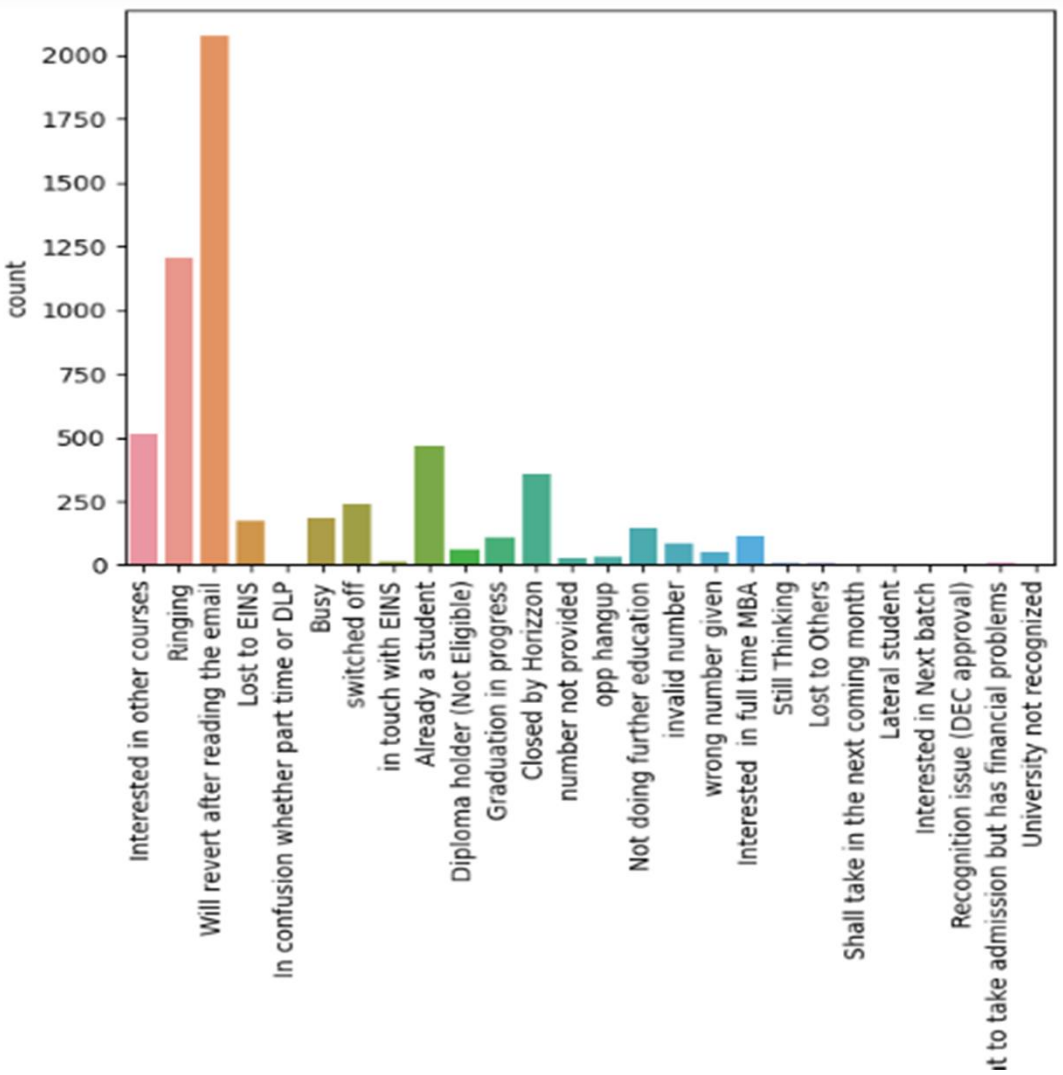
- Identifying a specific group of leads can lead to an increase in the conversion rate as the sales team can focus their efforts on communicating with potential leads rather than wasting time on a broad range of individuals.
- X Education requires assistance in selecting the most promising leads that are likely to convert into paying customers. A lead scoring model needs to be implemented where each lead is assigned a score, with higher scores indicating a higher chance of conversion.
- The CEO has set a target lead conversion rate of approximately 80%.

Approach

- Data sourcing, Cleaning and Preparation (Read, convert, remove duplicates, outliner)
- Exploratory Data Analysis
- Splitting the data into Test and Train dataset.
- Building a logistic Regression model and compute the Lead Score.
- Evaluating the model by using different metrics -Specificity and Sensitivity or Precision and Recall.
- Applying the best model in Test data based on the Sensitivity and Specificity Metrics.

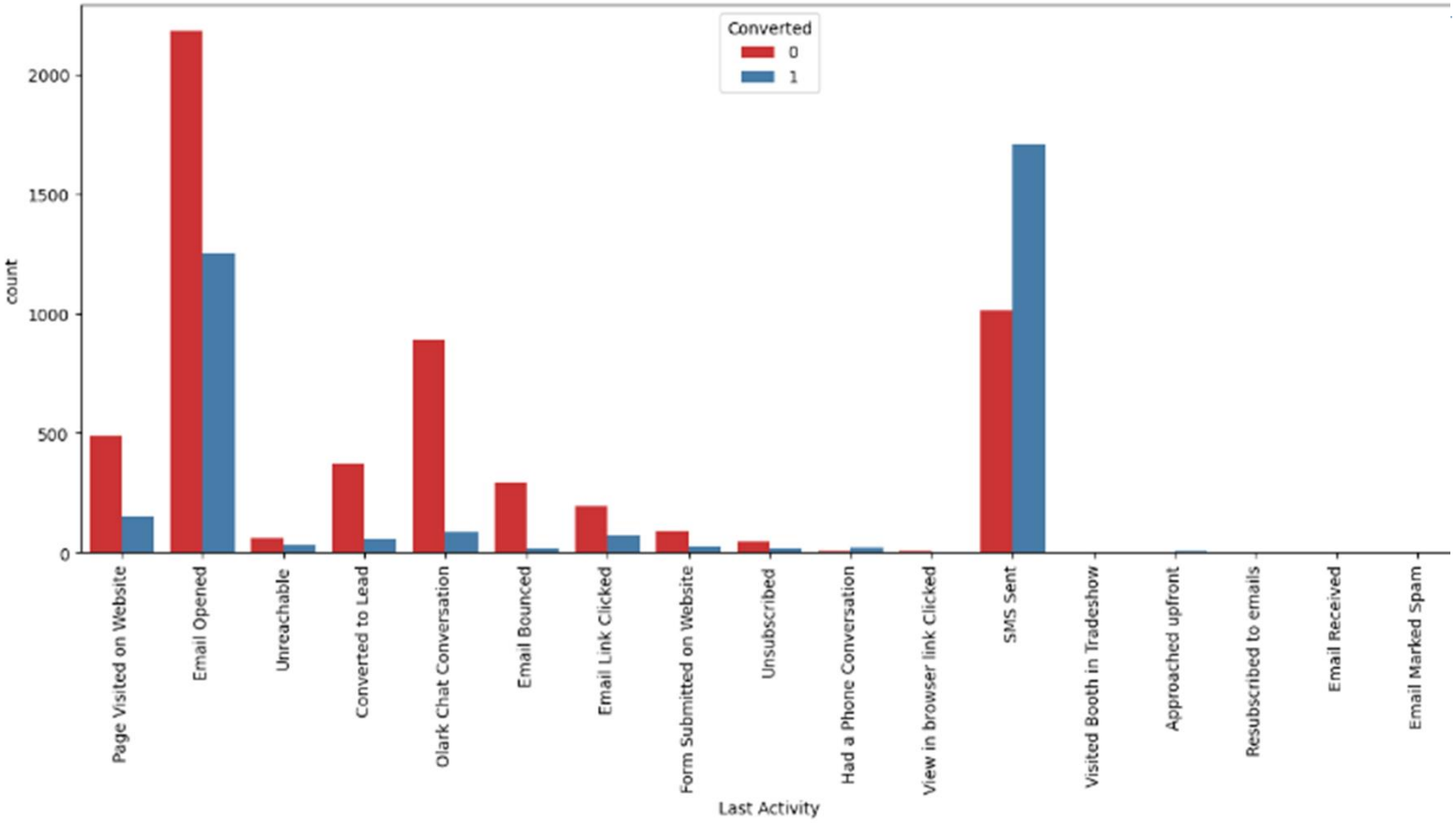
Exploratory Data Analysis

Major conversion has happened from Emails sent and Calls made



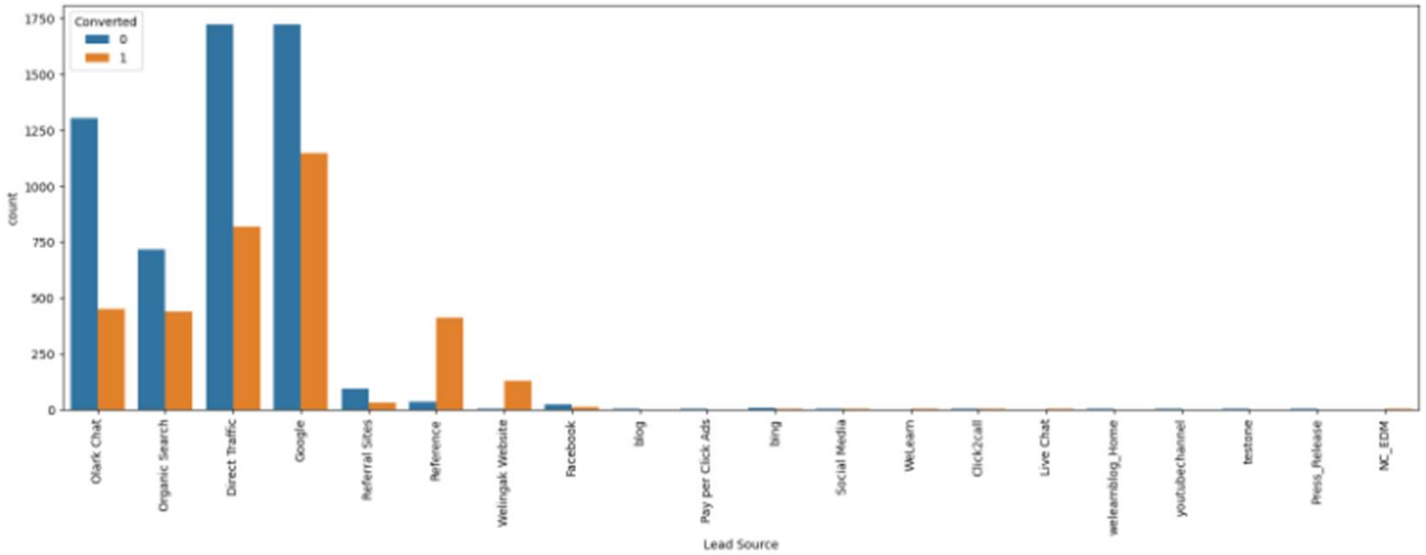
Exploratory Data Analysis

Conversion rate for leads with last activity as SMS Sent is High

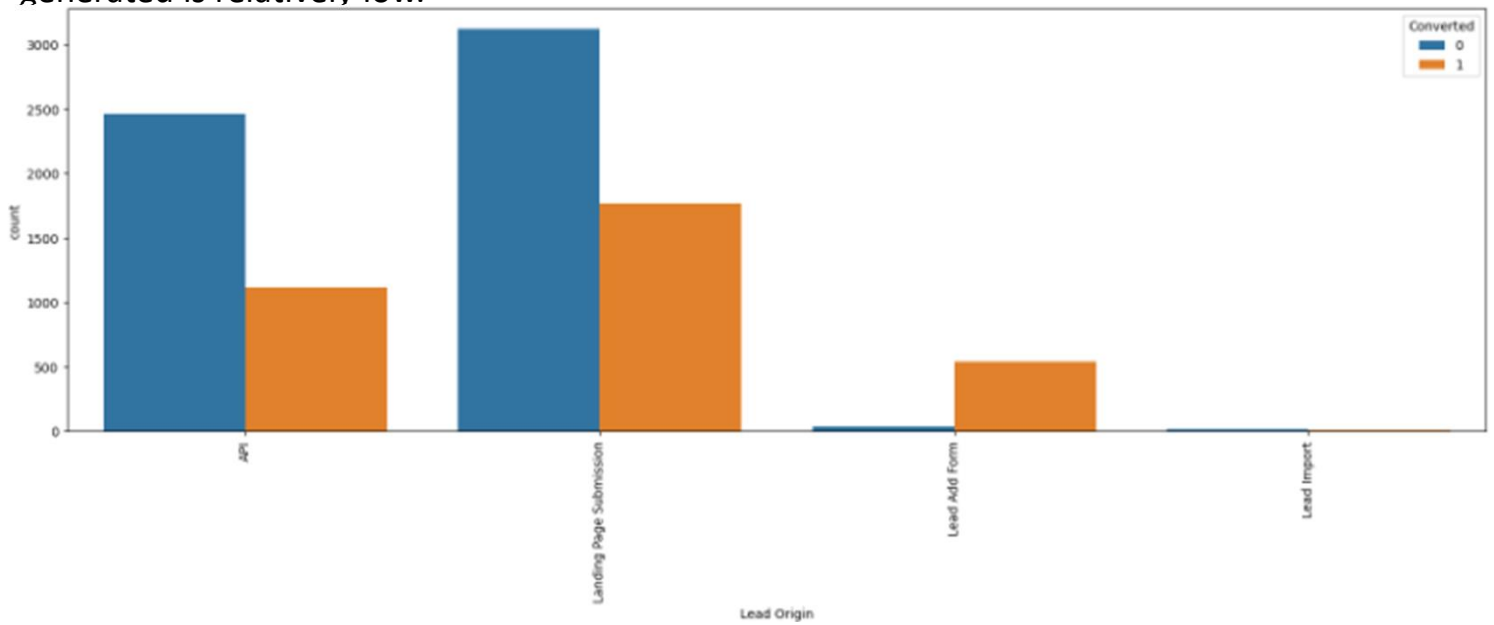


Major Leads

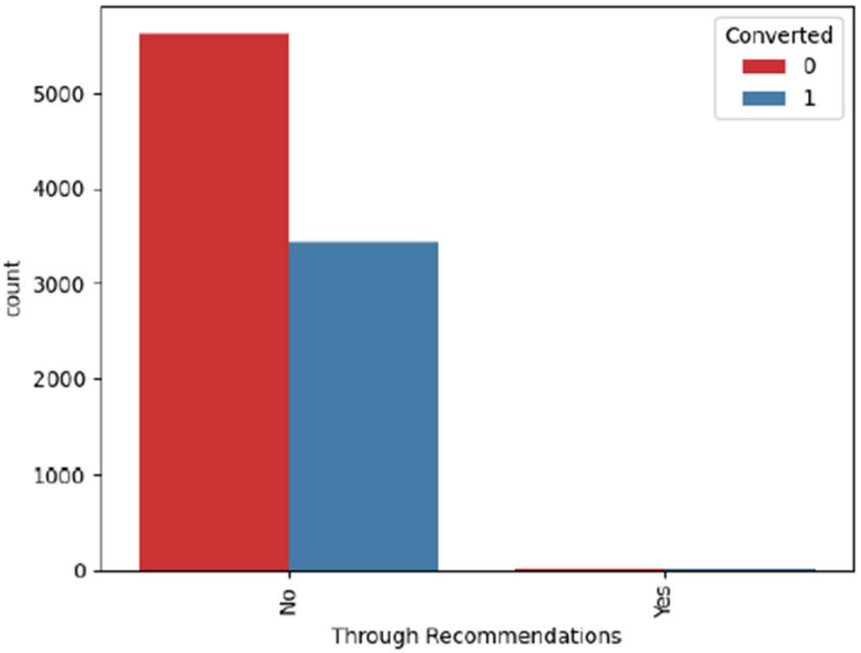
Conversion in the lead source is from Google is high

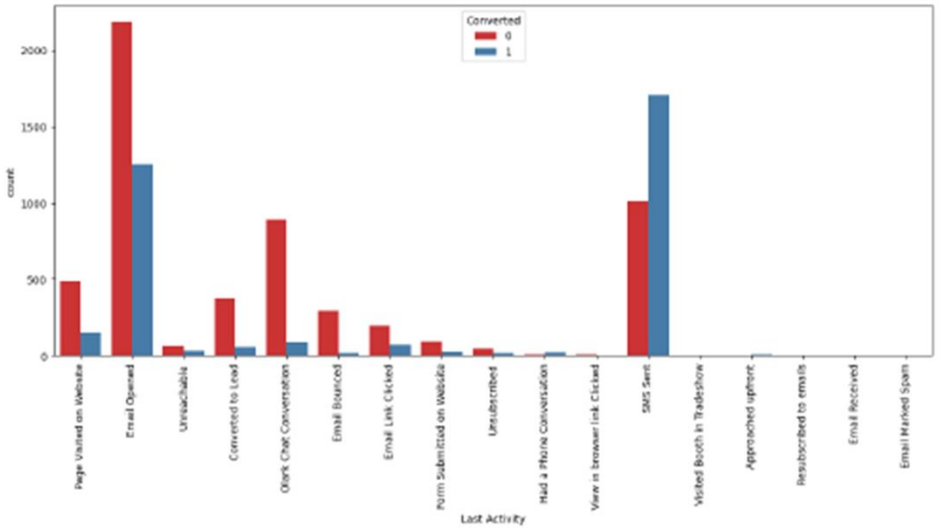


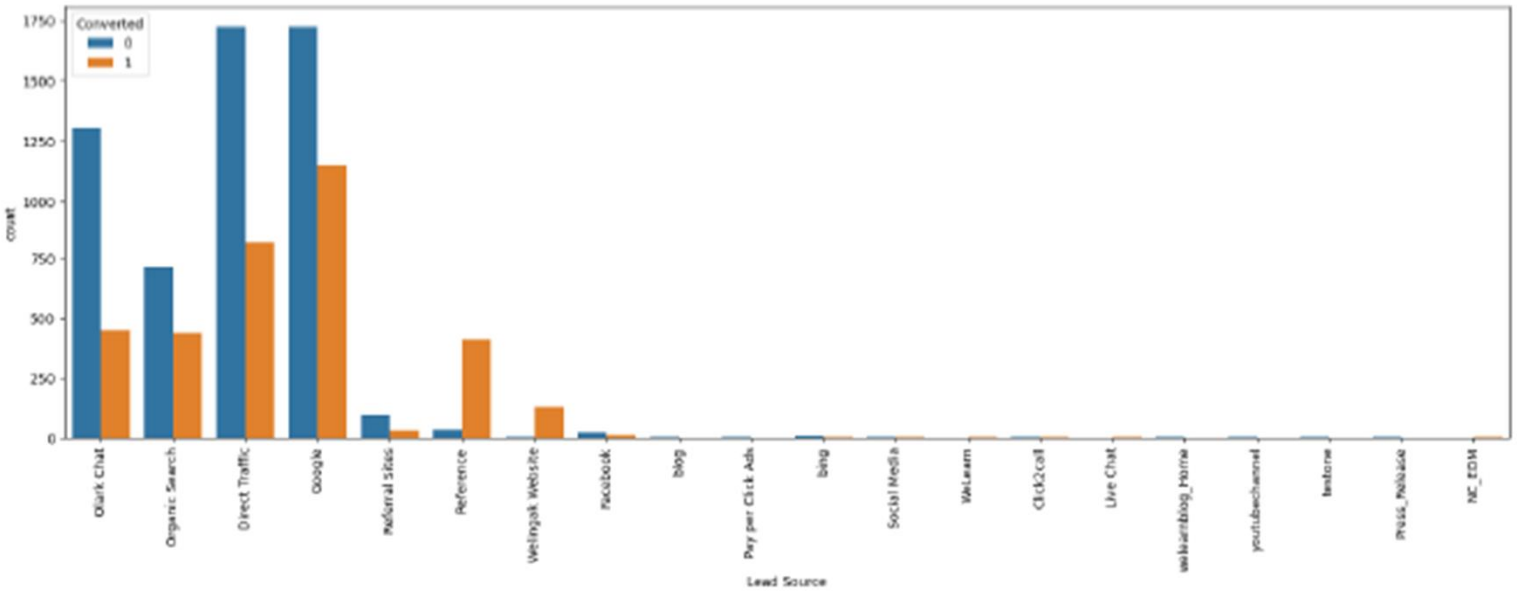
It's important to prioritize the Lead Add form as it has a conversion rate of over 90%, but the number of leads generated is relatively low.



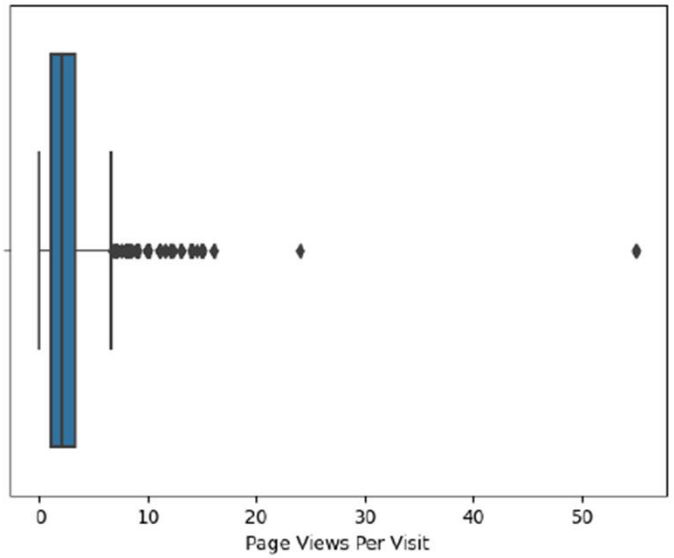
Other Major Leads

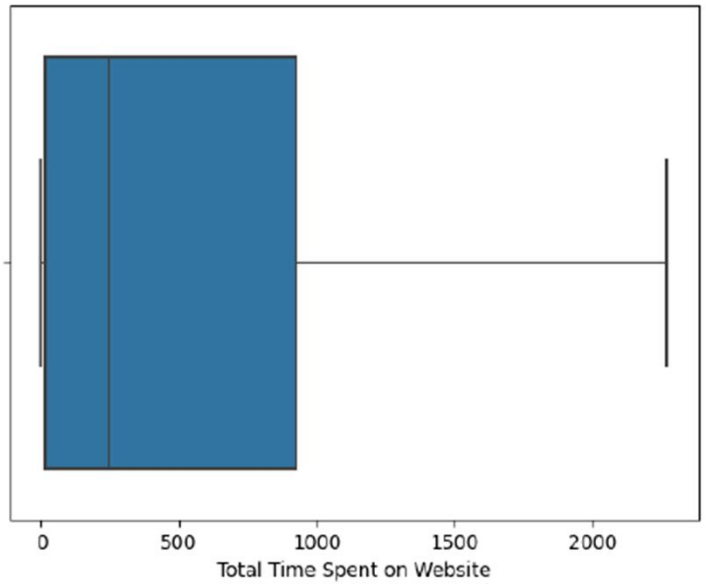


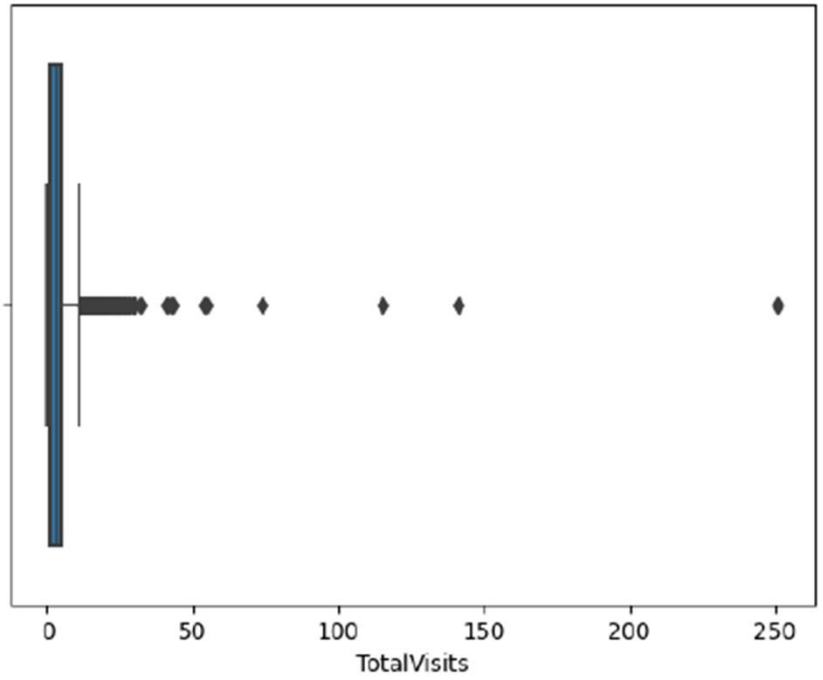


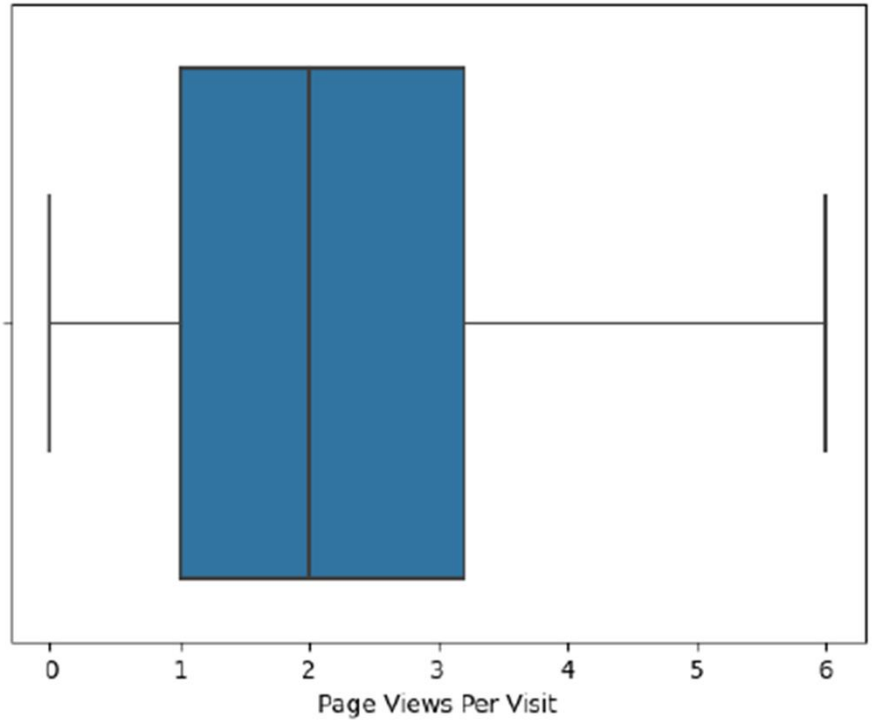


Box Plots









from above we can see there are outliers

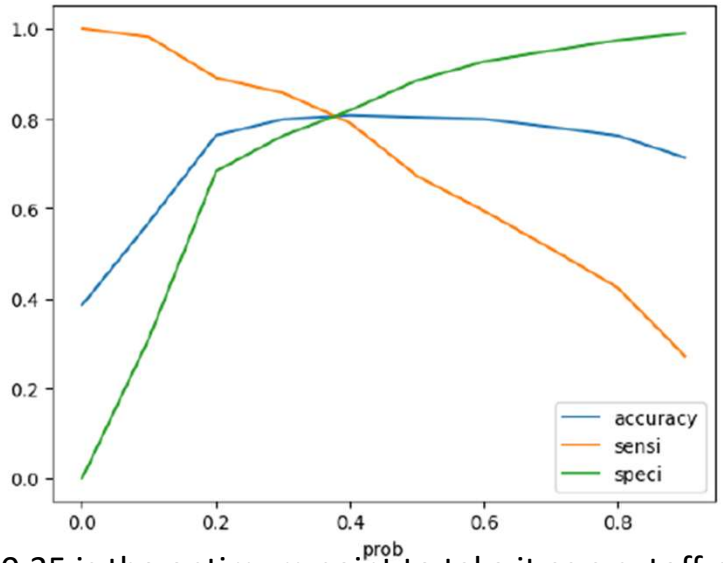
Model Building

- Pvalue below variables are high consequently we dropped following variables
  - What is your current occupation_Housewife
  - Lead Source_Referral Sites
  - Page Views Per Visit
  - Total visits
  - Last Activity_Had a Phone Conversation
- Pvalues for following variables are 0 and VIF values are low for all variables, thus we have below variables in our model.
  - Lead Origin_Landing Page Submission
  - Last Notable Activity_Modified
  - Specialization_UnKnown
  - Last Activity_Olark Chat Conversation
  - Total Time Spent on Website
  - Do Not Email_Yes
  - Last Notable Activity_Email Opened
  - Last Activity_Email Bounced
  - Lead Origin_Lead Add Form
  - Lead Source_Welingak Website
  - Last Notable Activity_Olark Chat Conversation
  - Last Activity_Converted to Lead
  - What is your current occupation_Working Profes
  - Last Notable Activity_Page Visited on Website
  - Last Notable Activity_Email Link Clicked

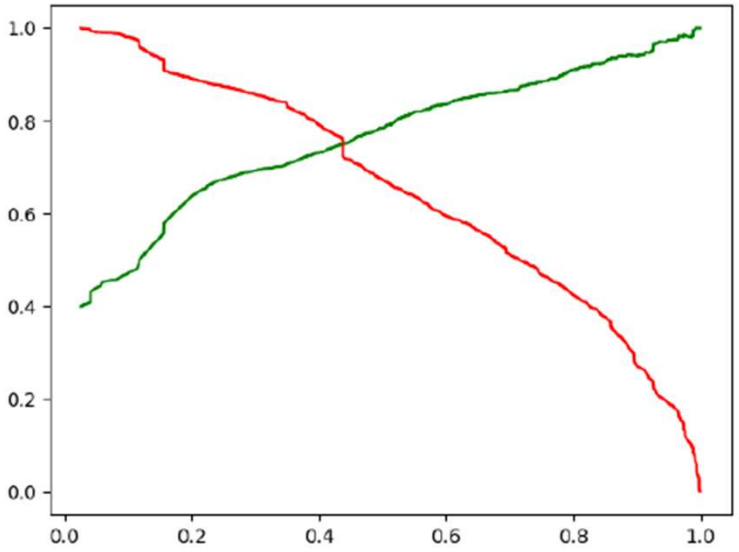
Features vs VIF

	Features	VIF
1	Lead Origin_Landing Page Submission	2.88
12	Last Notable Activity_Modified	2.69
8	Specialization_UnKnown	2.15
7	Last Activity_Olark Chat Conversation	1.93
0	Total Time Spent on Website	1.88
4	Do Not Email_Yes	1.85
11	Last Notable Activity_Email Opened	1.83
6	Last Activity_Email Bounced	1.76
2	Lead Origin_Lead Add Form	1.49
3	Lead Source_Welingak Website	1.37
13	Last Notable Activity_Olark Chat Conversation	1.37
5	Last Activity_Converted to Lead	1.23
9	What is your current occupation_Working Profes...	1.18
14	Last Notable Activity_Page Visited on Website	1.10
10	Last Notable Activity_Email Link Clicked	1.05

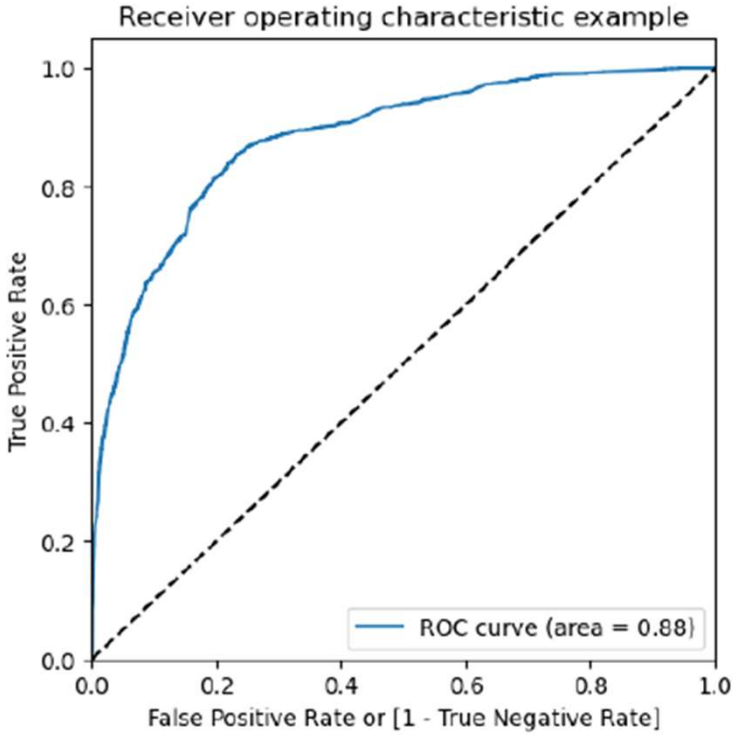
Model Assessment



0.35 is the optimum point to take it as a cutoff probability.



Precision and recall Curve



Our model is considered good as evidenced by the higher (0.88) area under the ROC curve

Recommendations

- The company should make calls to the leads who spent "more time on the websites" as these are more likely to get converted.
- The company should make calls to the leads coming from the lead sources "Welingak Websites" and "Reference" as these are more likely to get converted.
- The company should make calls to the leads who are the "working professionals" as they are more likely to get converted.
- The company should not make calls to the leads whose last activity was "Olark Chat Conversation" as they are not likely to get converted.
- The company should not make calls to the leads whose lead origin is "Landing Page Submission" as they are not likely to get converted.
- The company should not make calls to the leads whose Specialization was "Unknown" as they are not likely to get converted.
- The company should not make calls to the leads who chose the option of "Do not Email" as "yes" as they are not likely to get converted.

Conclusion

- The most significant variables contributing to lead conversion in the model are
  - Total time spent on website
  - Lead Add Form from Lead Origin
- The test set's Accuracy, Sensitivity, Specificity, Precision score values are approximately similar to those calculated using the trained set, with values of around 80%,67%, 88% and 78%.
- Suggests that the model is generally effective