

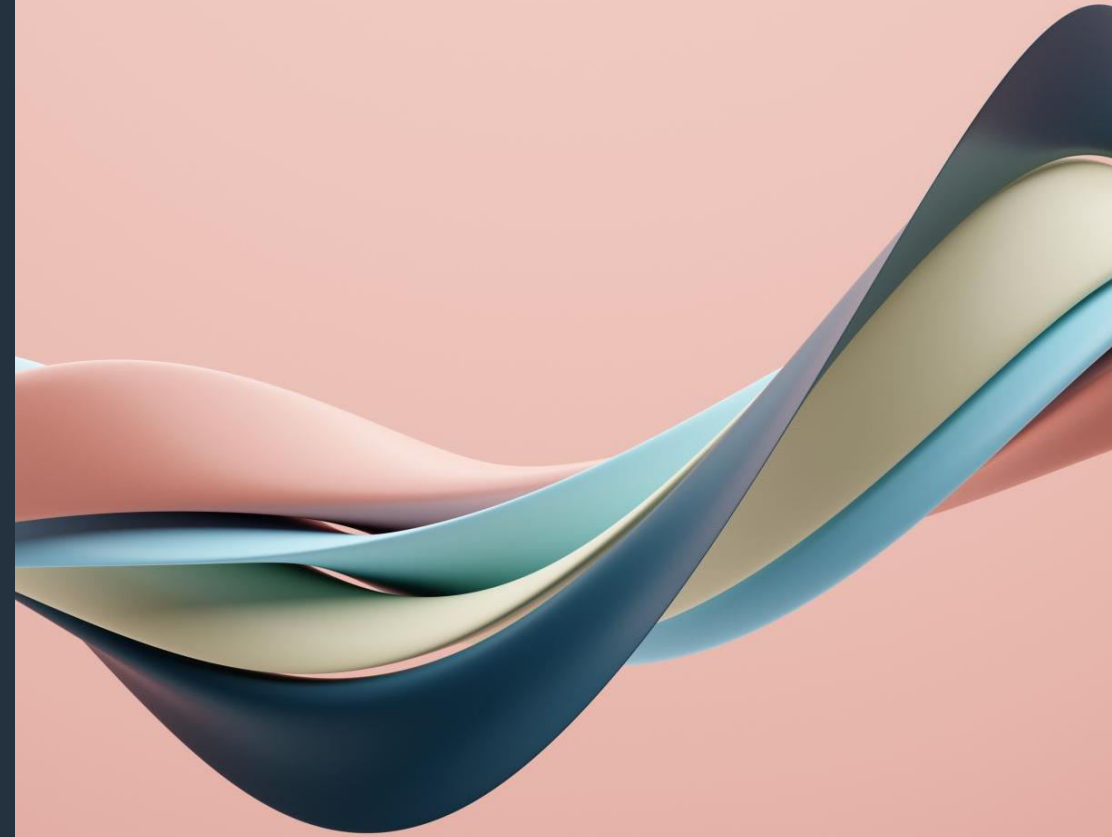
Lead Scoring Case Study

Submitted By:

Archana K

Anshika Sharma

Madhavalatha Arkatla



Problem Statement :

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%.

Business Goal:

X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.

The company needs a model wherein you a lead score is assigned 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%.

Problem Approach

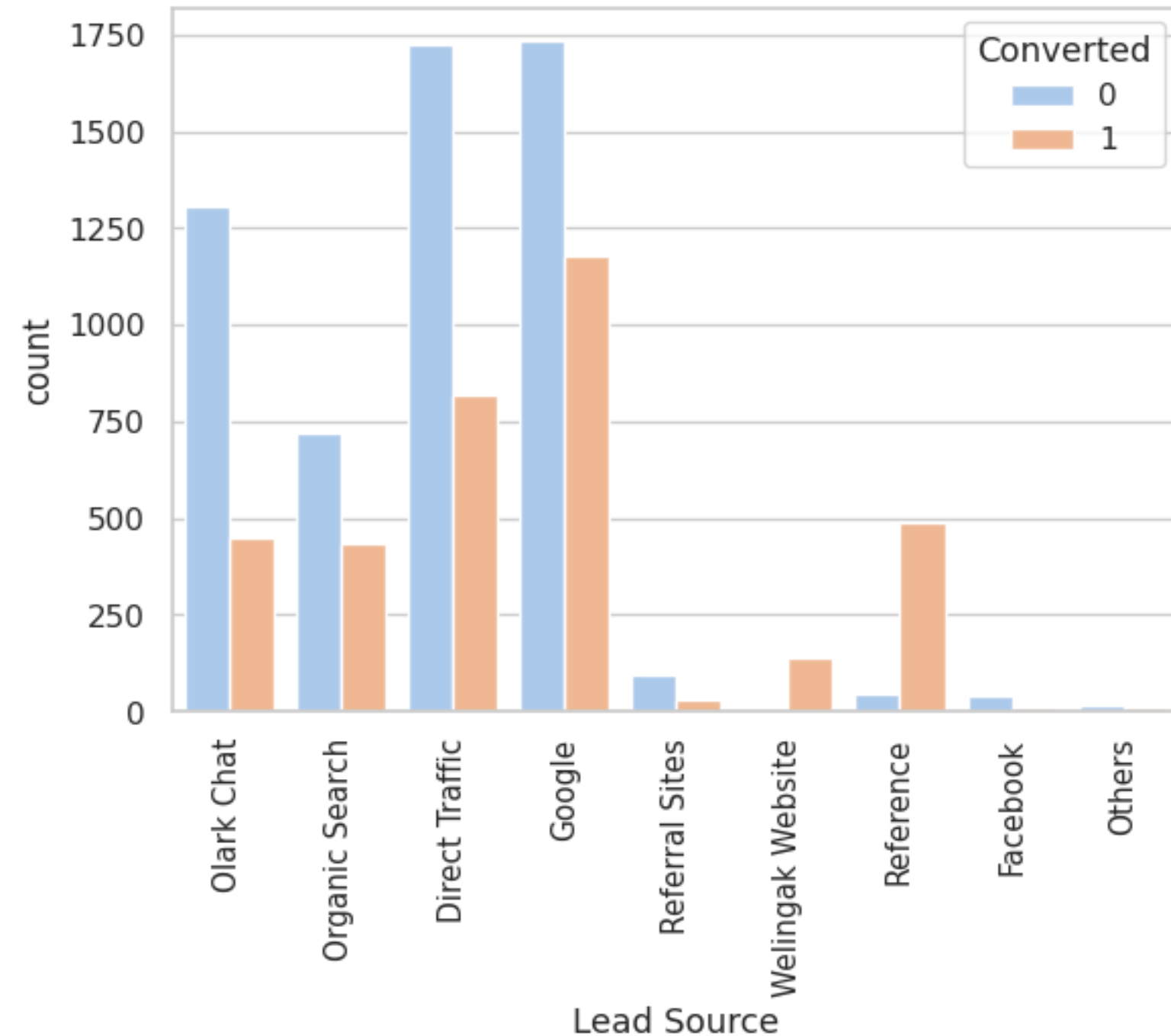
1. Read and understand the data
2. Clean and prepare the data
3. Exploratory Data Analysis
4. Prepare the data for Model Building
5. Building a logistic Regression model and calculate Lead Score.
6. Evaluating the model by using different metrics -Specificity and Sensitivity or Precision and Recall.
7. Applying the best model in Test data based on the Sensitivity and Specificity Metrics.

In our dataset, the value 'Select' appears as a default placeholder in several columns. This indicates unselected options in form dropdowns. We need to replace these 'Select' values with NaN to ensure data quality and consistency. Dropped the columns having nulls greater than 45%. Imputing the mean, median, or mode values as appropriate for each respective feature.

Target Columns:

- Specialization
- How did you hear about X Education
- Lead Profile
- City

We removed Lead Number and Prospect ID from the dataset since they are merely indexes and do not enhance the regression model's predictive power.



Top Three Lead Sources:

- Google
- Direct Traffic
- Olark Chat

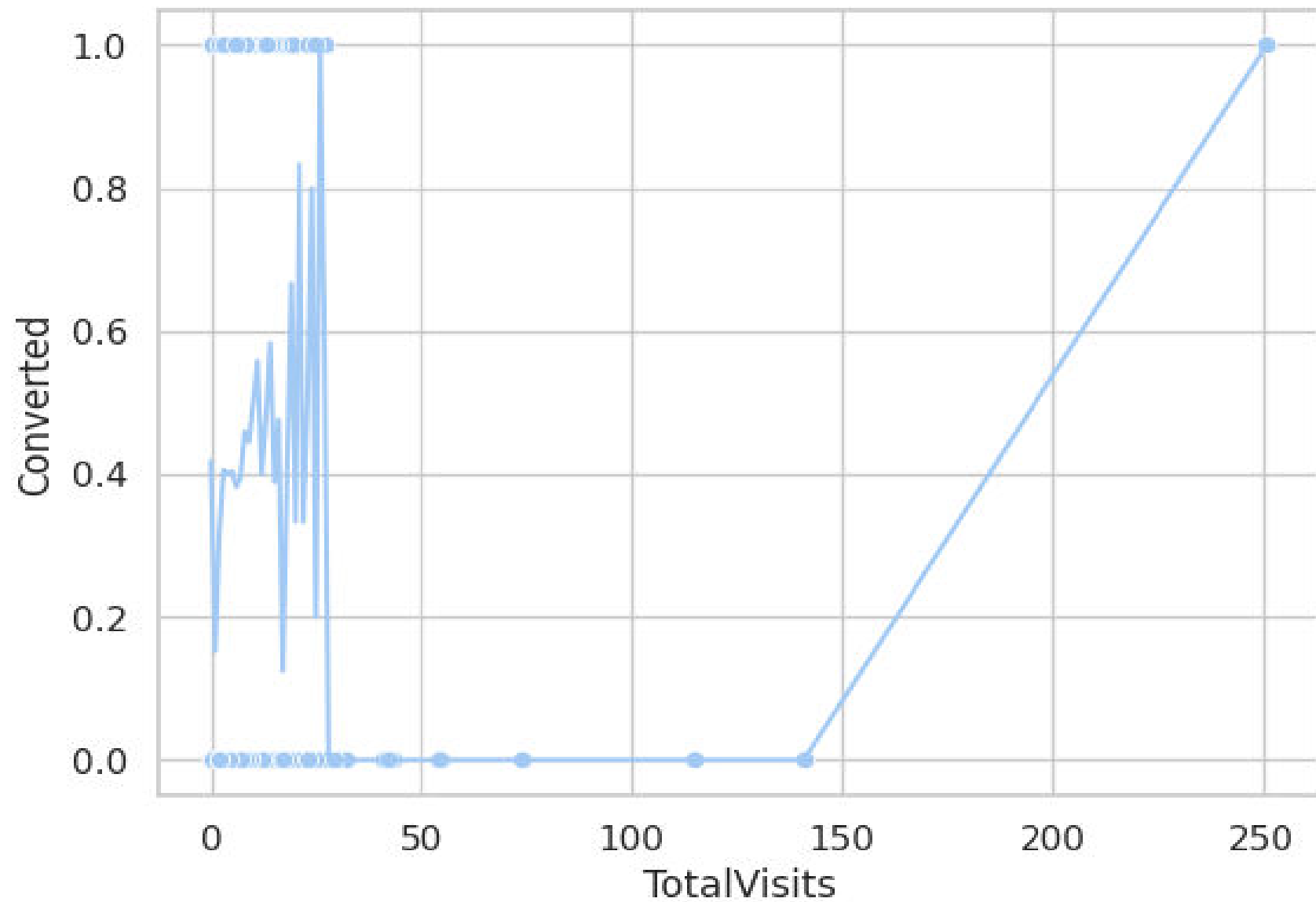
Significant Lead Sources:

- Reference
- Organic Search
- Welingak Website
- Referral Sites

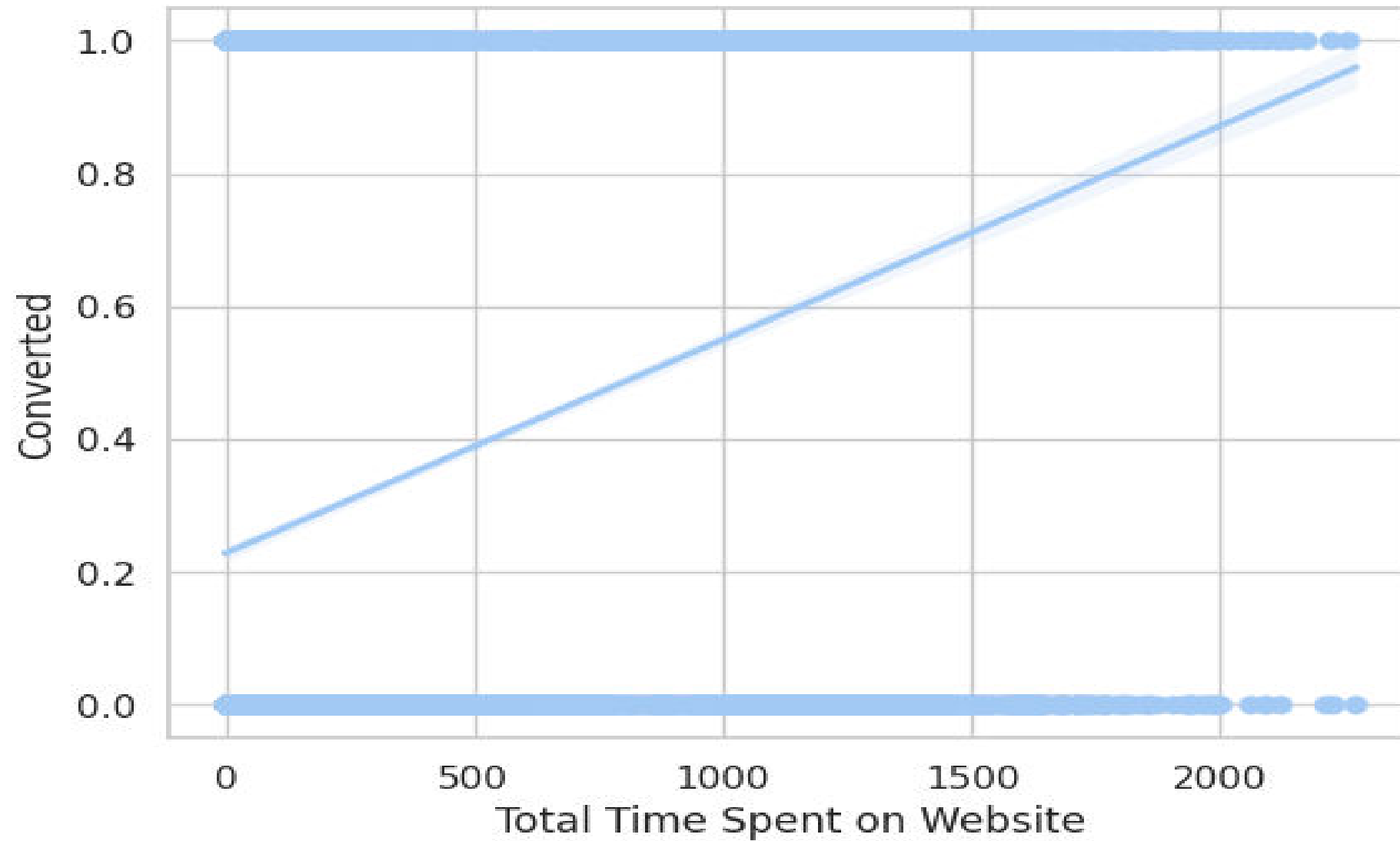
Highest Conversion Rate:

- Reference
- Welingak Website
- Direct Traffic


Total Visits with respect to Converted leads

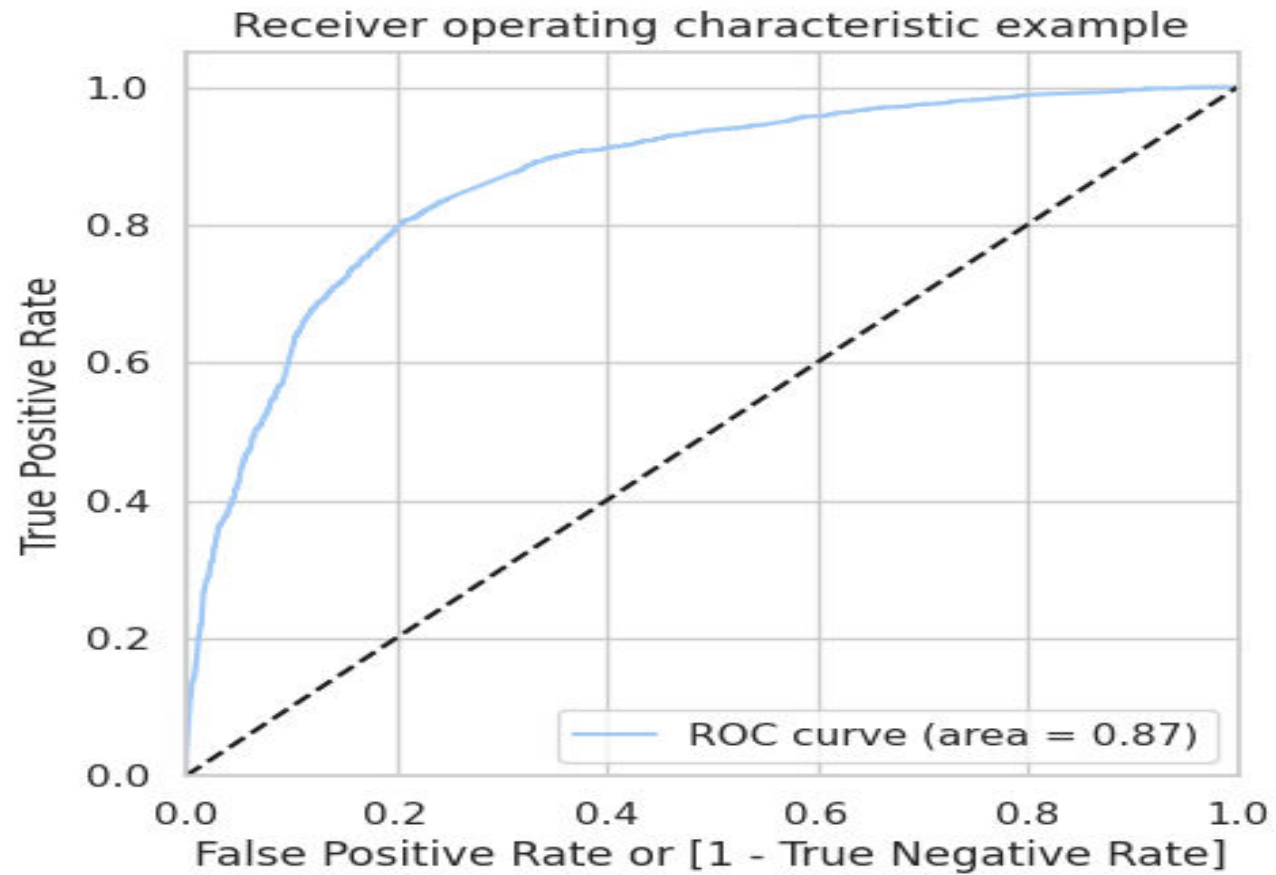


Total Time Spent on Website with respect to converted leads



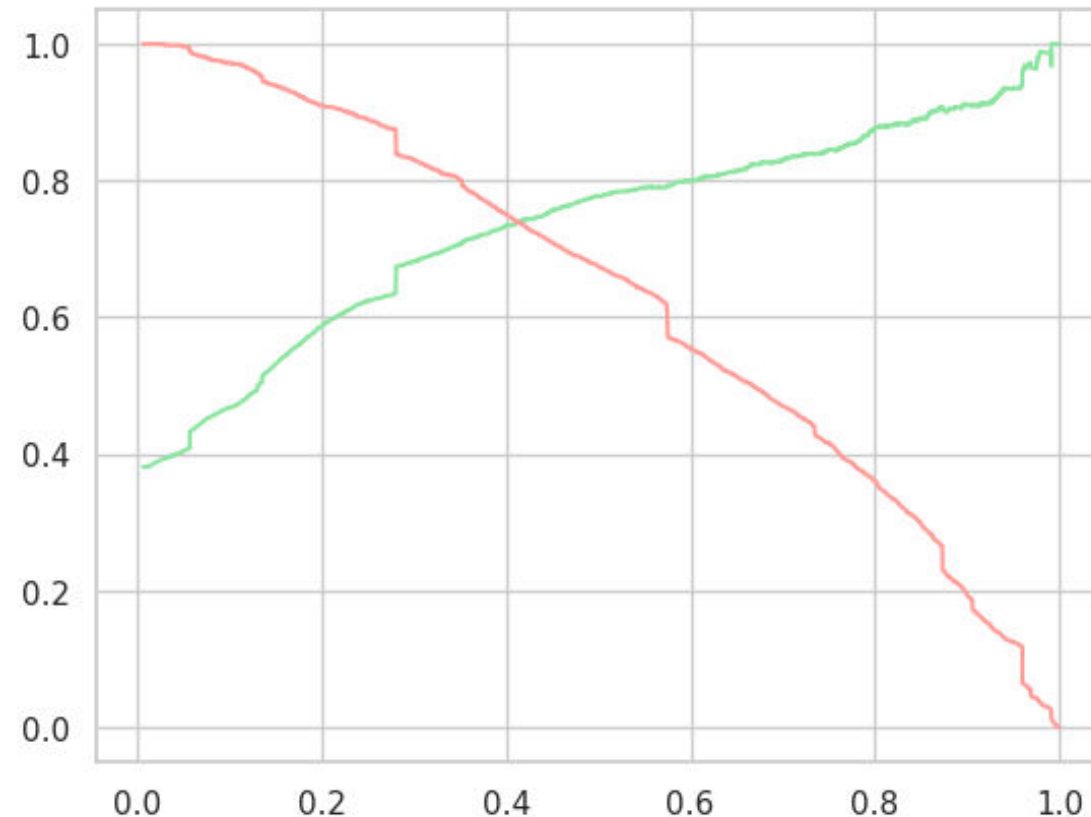
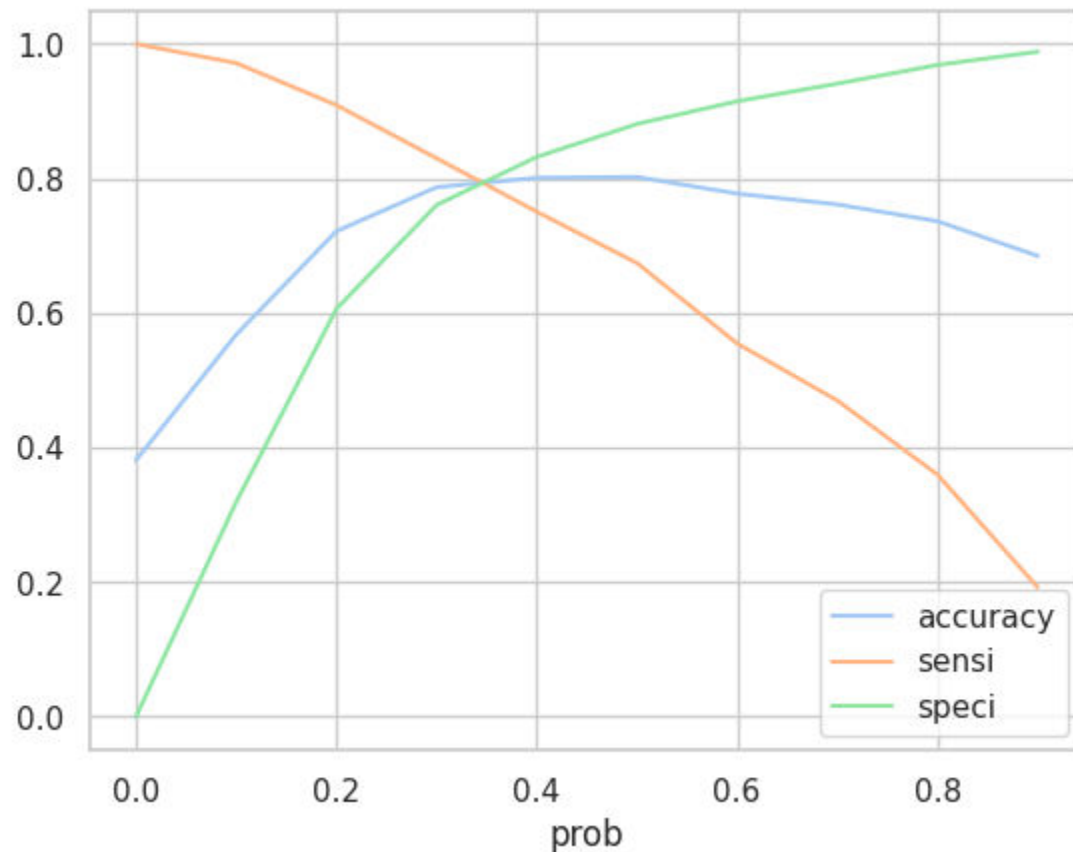
Variables Impacting the Conversion Rate

- Do Not Email
 - Total Visits
 - Total Time Spent On Website
 - Pages Views Per Visit
 - Lead Origin API
 - Lead Origin Landing Page Submission
 - Lead Origin Add Form
 - Lead Origin Import
 - Last Notable Activity Email Opened
 - Last Notable Activity Modified
 - Last Notable Activity Other
 - Last Notable Activity SMS Sent
- 
- A large, solid red curved shape that starts from the bottom right corner and sweeps upwards and to the left, ending near the center of the bottom edge of the slide.



The area under the curve of the ROC is 0.87 which is quite good.

Model Evaluation



As we can see that around 0.38, we get the optimal values of the three metrics. So, let's choose 0.35 as our cutoff. Thus, we can safely choose to consider any prospect lead with conversion probability higher than 42% to be a Hot Lead.

Observations

- While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
- Accuracy, Sensitivity and Specificity values of test set are around 81%, 79% and 82% which are approximately closer to the respective values calculated using trained set.
- The top 3 variables that contribute for lead getting converted in the model are Total time spent on website, Lead Add Form from Lead Origin and Phone Conversation from Last Notable Activity
- Hence overall this model seems to be good.