LEAD SCORING CASE STUDY

Business Objective

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.

Social Media Marketing
Selection of Hot Leads
Initial Pool of leads Nurturing
Converted Leads

Goals of the Case Study

1. Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

2. There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well.

- .

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. A typical lead conversion process can be represented using the following funnel:

As you can see, there are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. 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%.

METHODOLOGY

- To build a Logistic Regression model that assigns lead scores to all leads such that the customers with higher lead score have a higher conversion chance and vice versa. Target Lead Conversion Rate ≈ 80%

importing libraries
importing libraries

Reading and Understanding the Data

Importing Libraries

Exploratory Data Analysis

Data Preparation

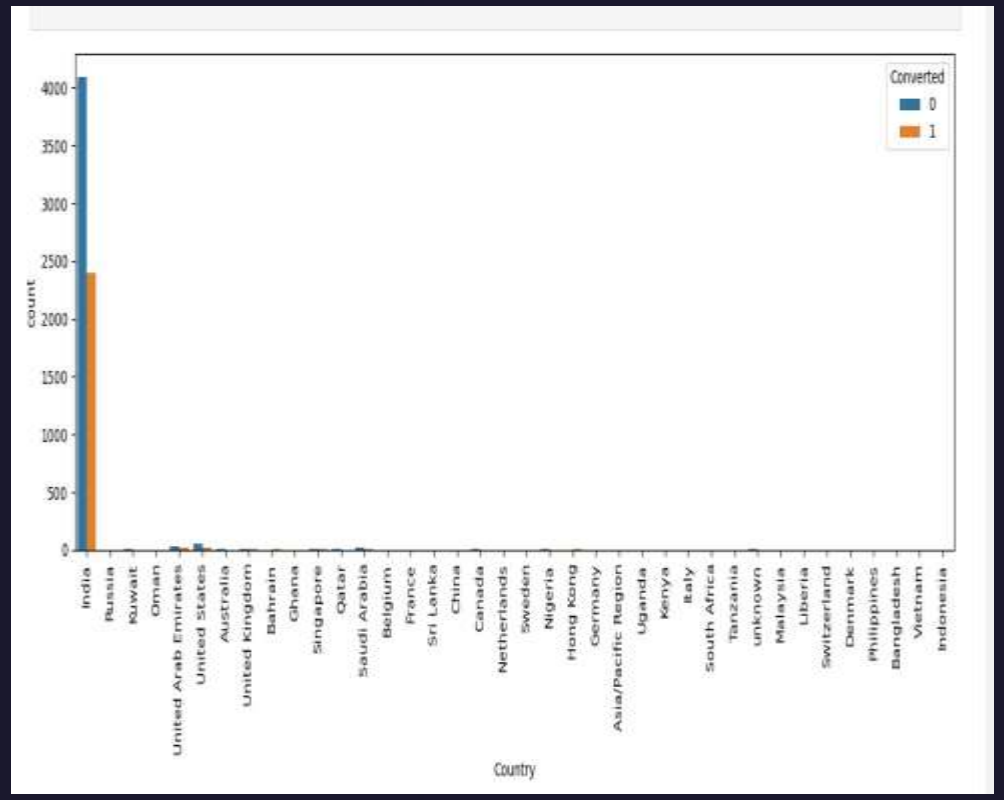
Scaling of Data

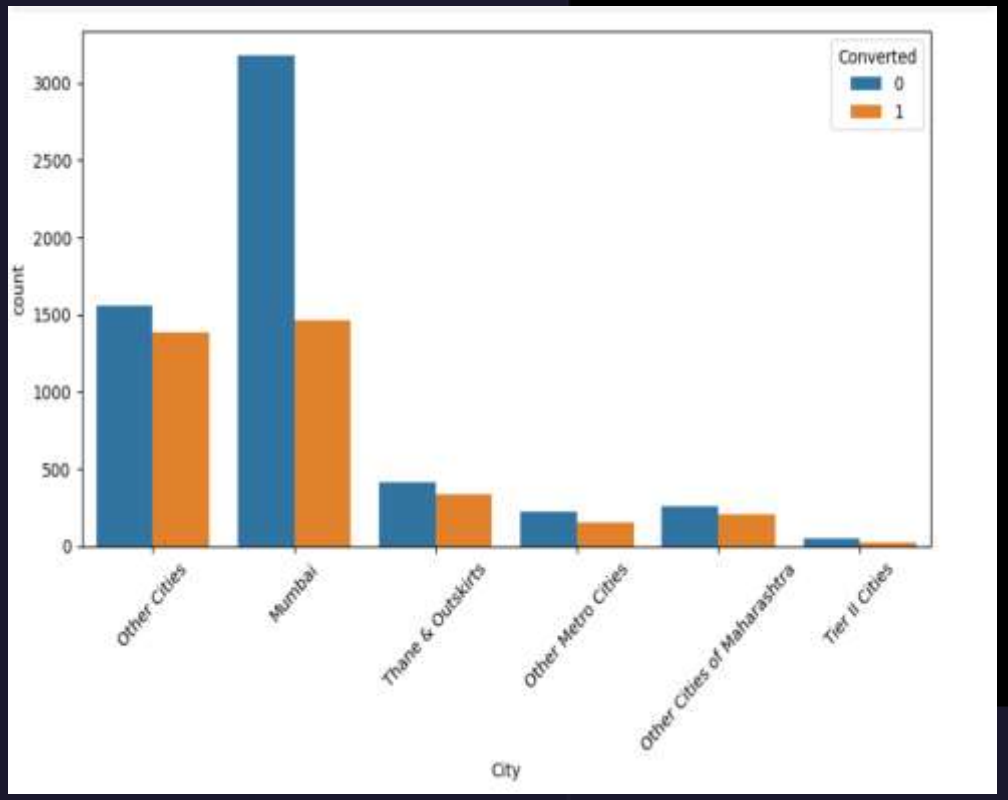
Model Building

Model Evaluation

Categorical Attributes Analysis And Null value treatments

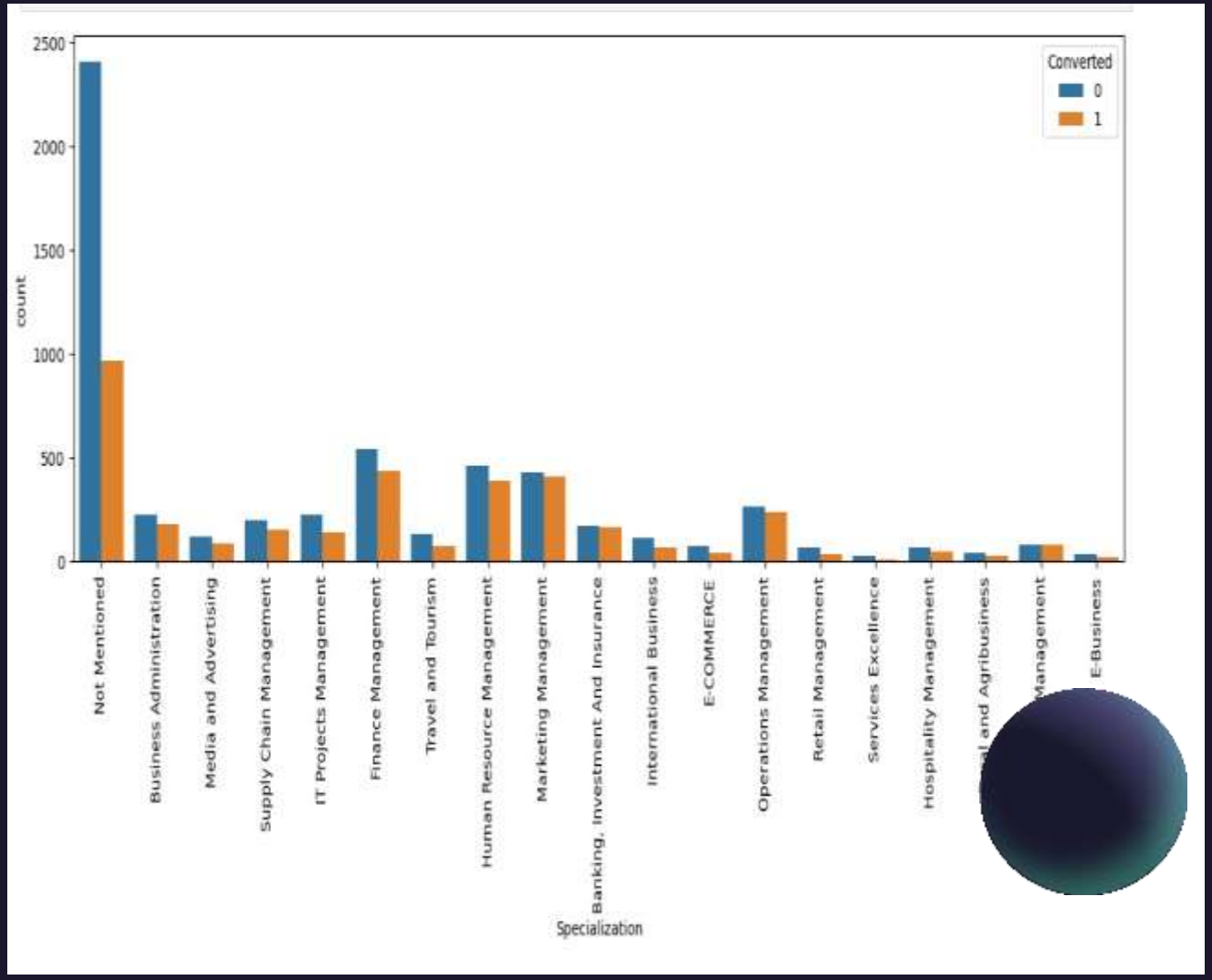
As we can see the Number of Values for India are quite high (approx. 95%+ of the Data), this column can be dropped





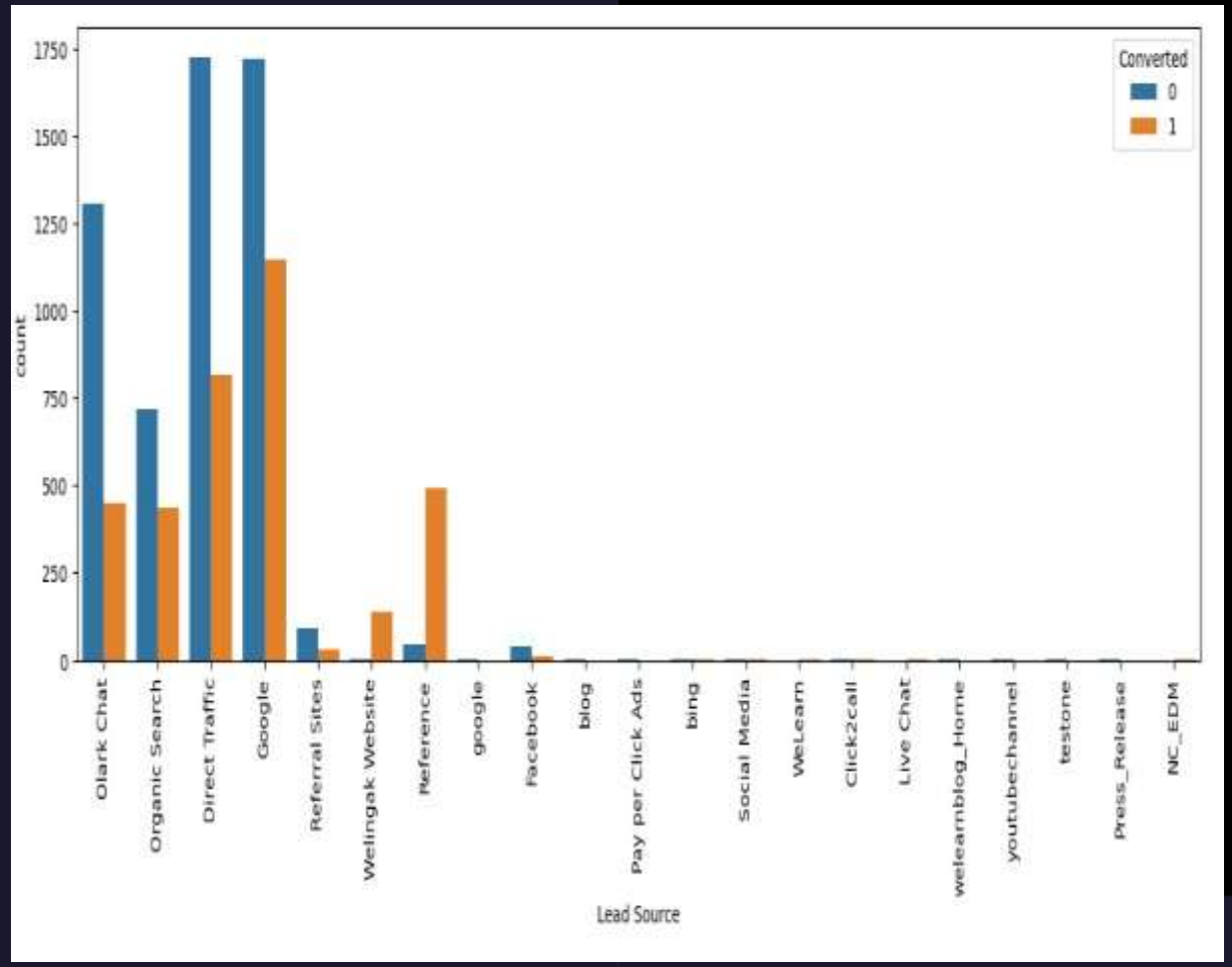
Categorical Attributes Analysis And Null value treatments

We observe that specializations with a focus on management exhibit a notably higher number of generated leads, and a substantial proportion of these leads are successfully converted. Consequently, it becomes evident that this 'Specialization' variable holds significant importance and should not be considered for elimination.



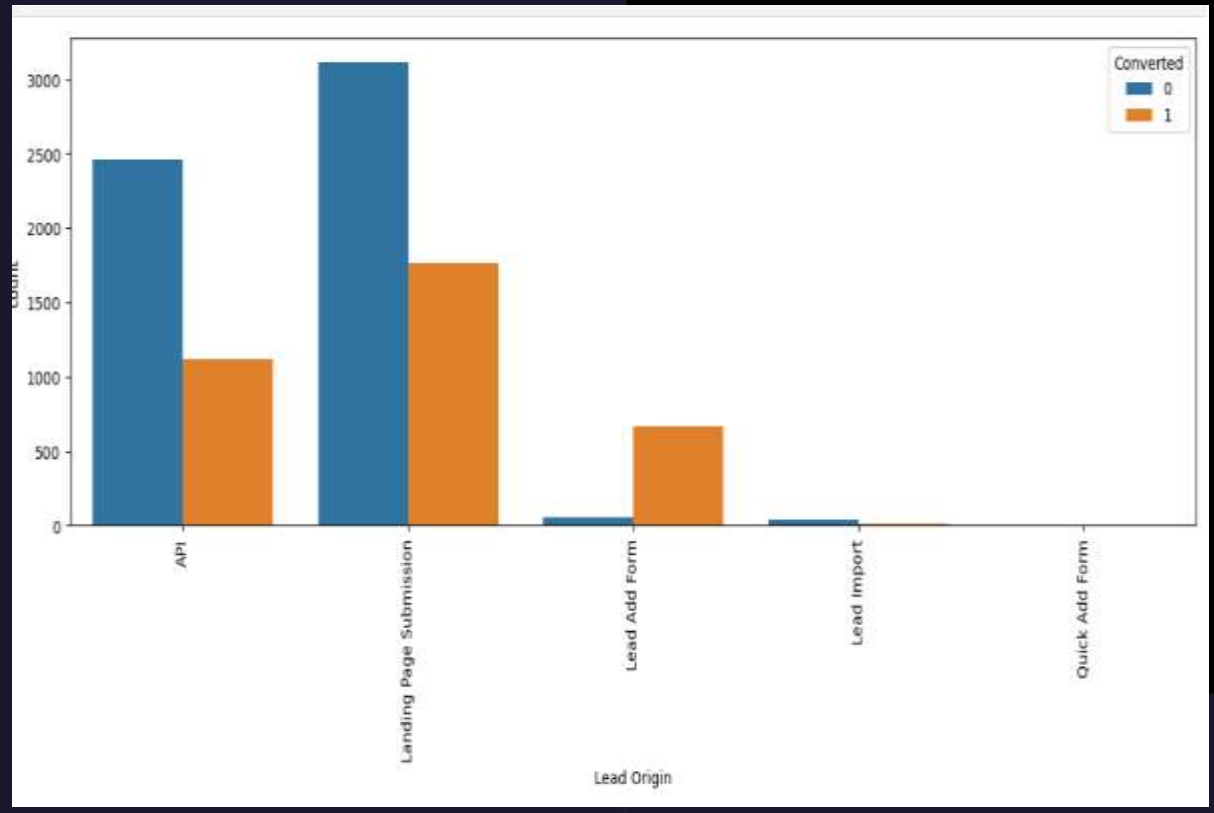
Categorical Attributes Analysis And Null value treatments

- Google and Direct traffic sources generate the highest number of leads.
- The conversion rate for reference leads and leads through the Welingak website is notably high.
- To enhance the overall lead conversion rate, prioritize improving the conversion rates for Olark chat, organic search, direct traffic, and Google leads.
- Additionally, aim to increase lead generation from reference sources and the Welingak website.

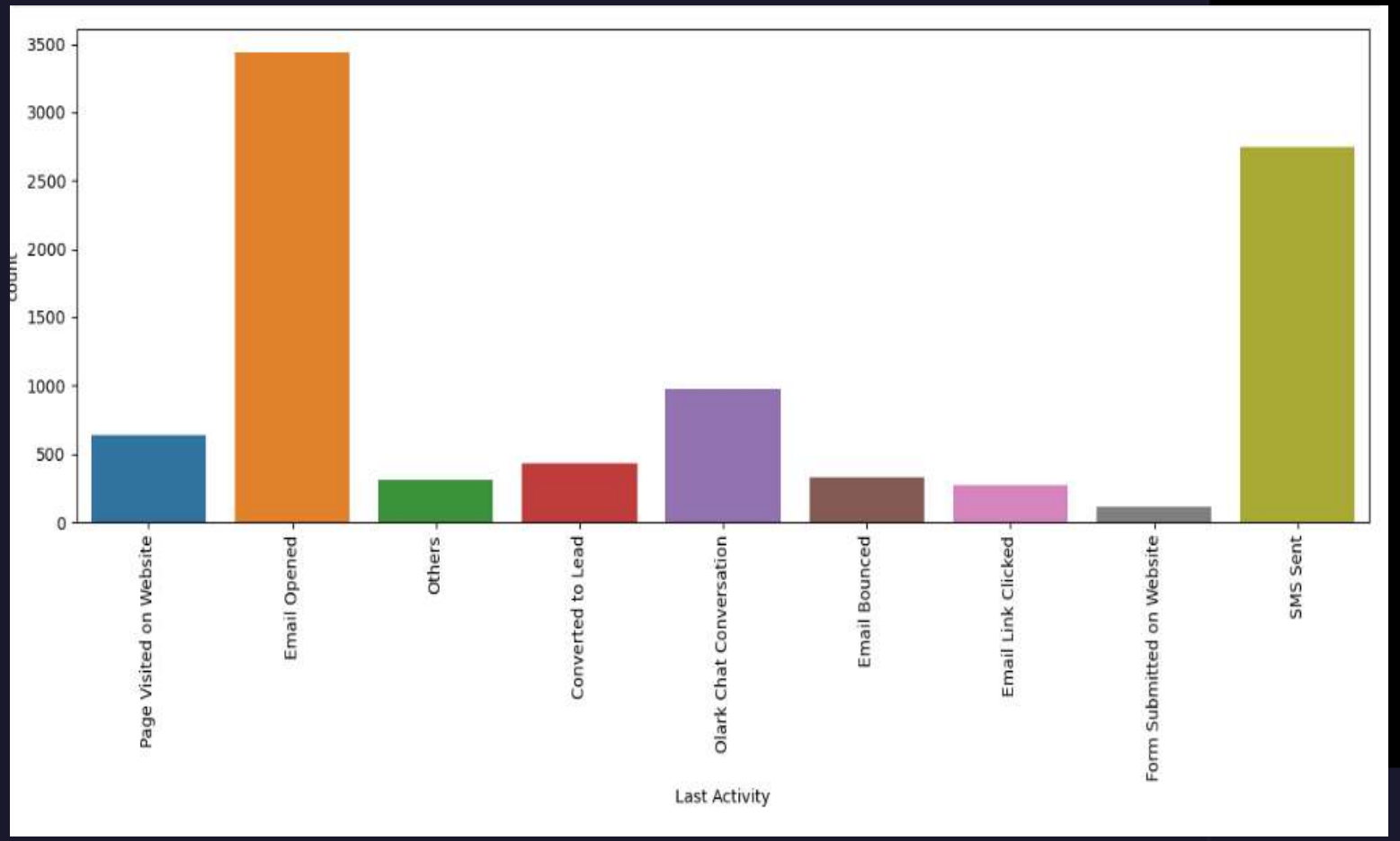


Categorical Attributes Analysis And Null value treatments

- API and Landing Page Submission yield a high number of leads and conversions.
- Lead Add Form has a remarkable conversion rate, even though the total lead count is not very high.
- Lead Import and Quick Add Form generate comparatively fewer leads.
- To enhance the overall lead conversion rate, focus on improving the conversion of leads from API and Landing Page Submission sources.
- Prioritize efforts to generate more leads through the Lead Add Form.

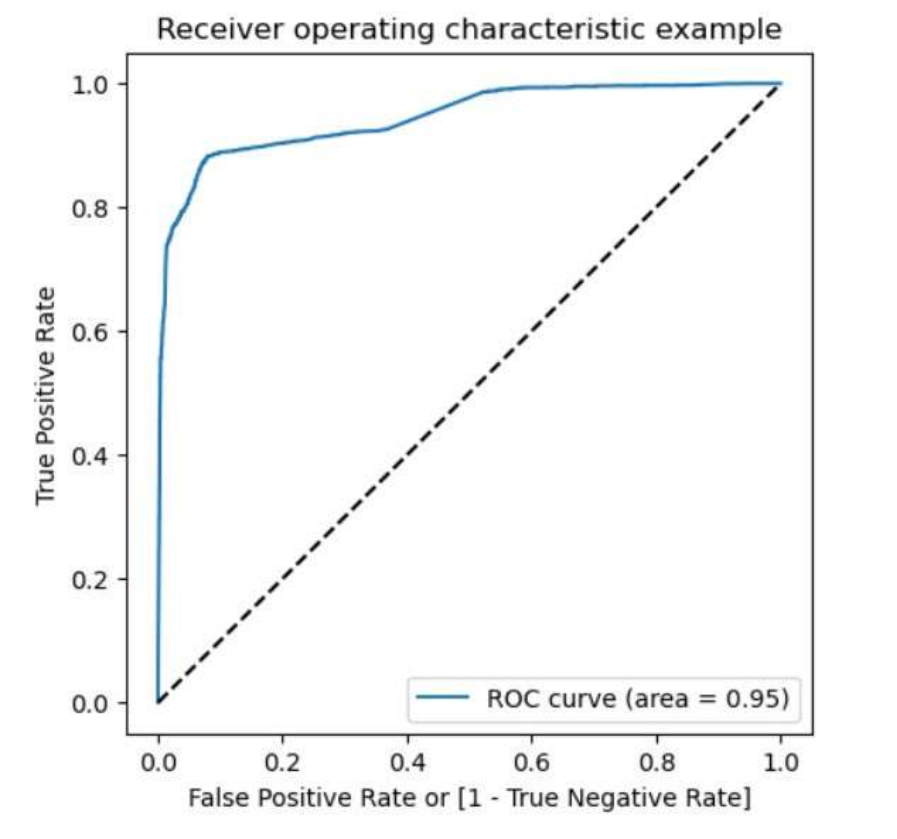


Email Opened and SMS sent are 2 most activity seen by the leads.

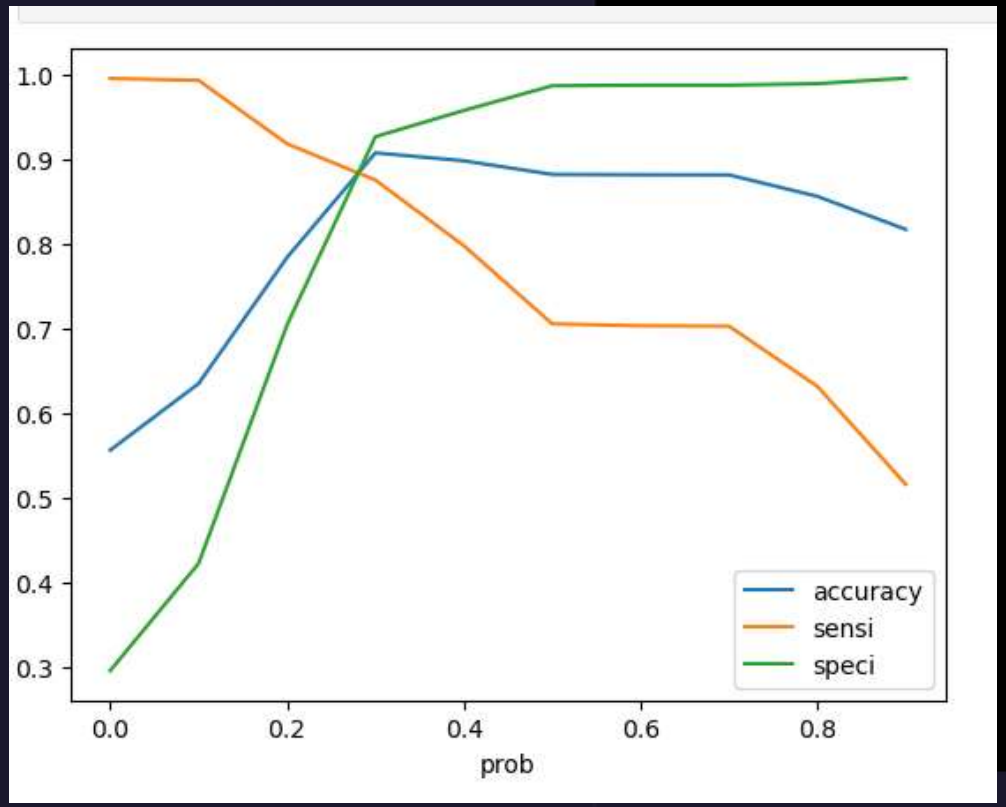


Numerical Attributes Analysis and Null value treatment

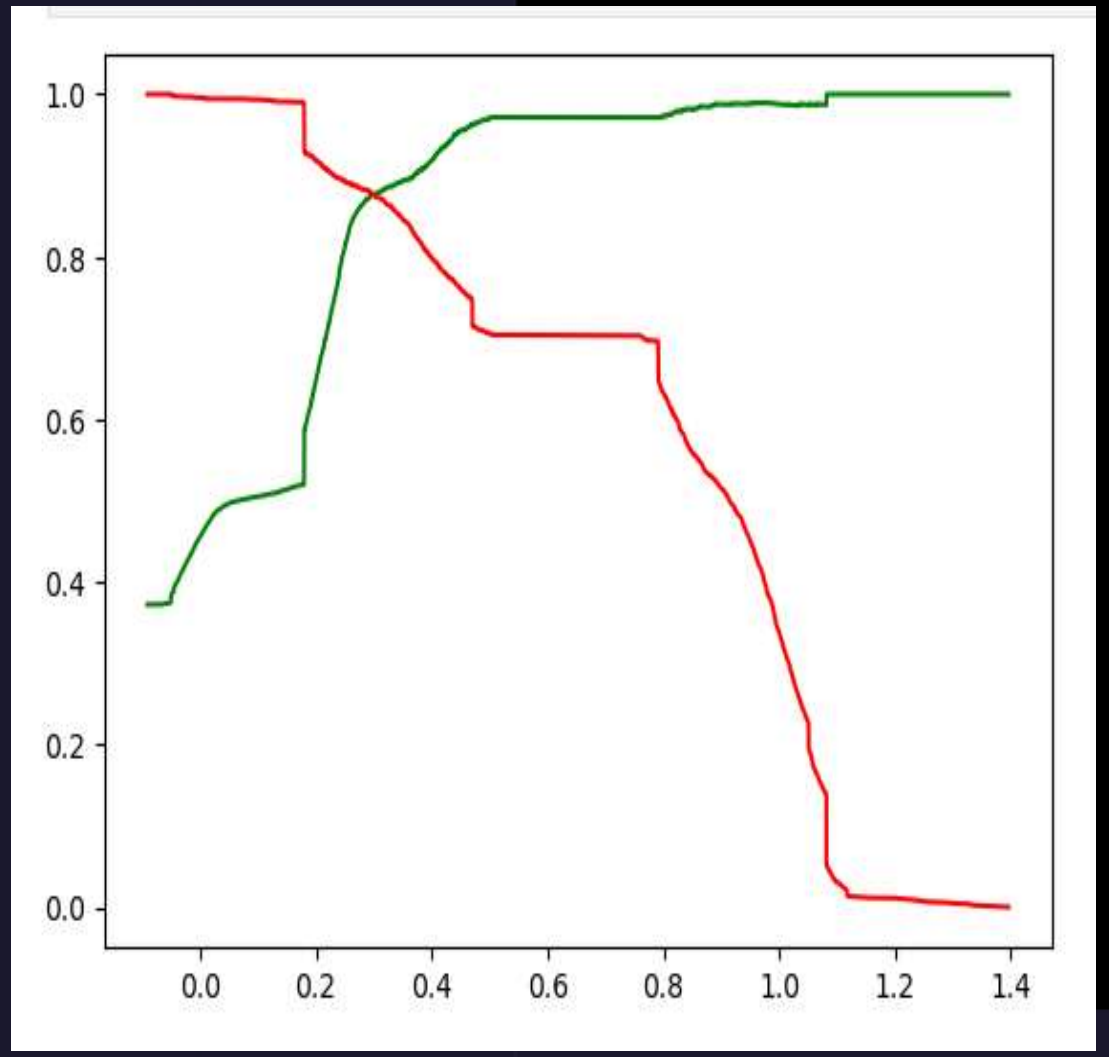
As we know ROC curve should be close to1. We are getting a good score of 0.95.



The below graph display an approximate cutoff at 0.3



This plot is the precision and recall values at different of train dataset, that shows the probability thresholds to visualize the trade-off between precision and recall in a binary classification model.



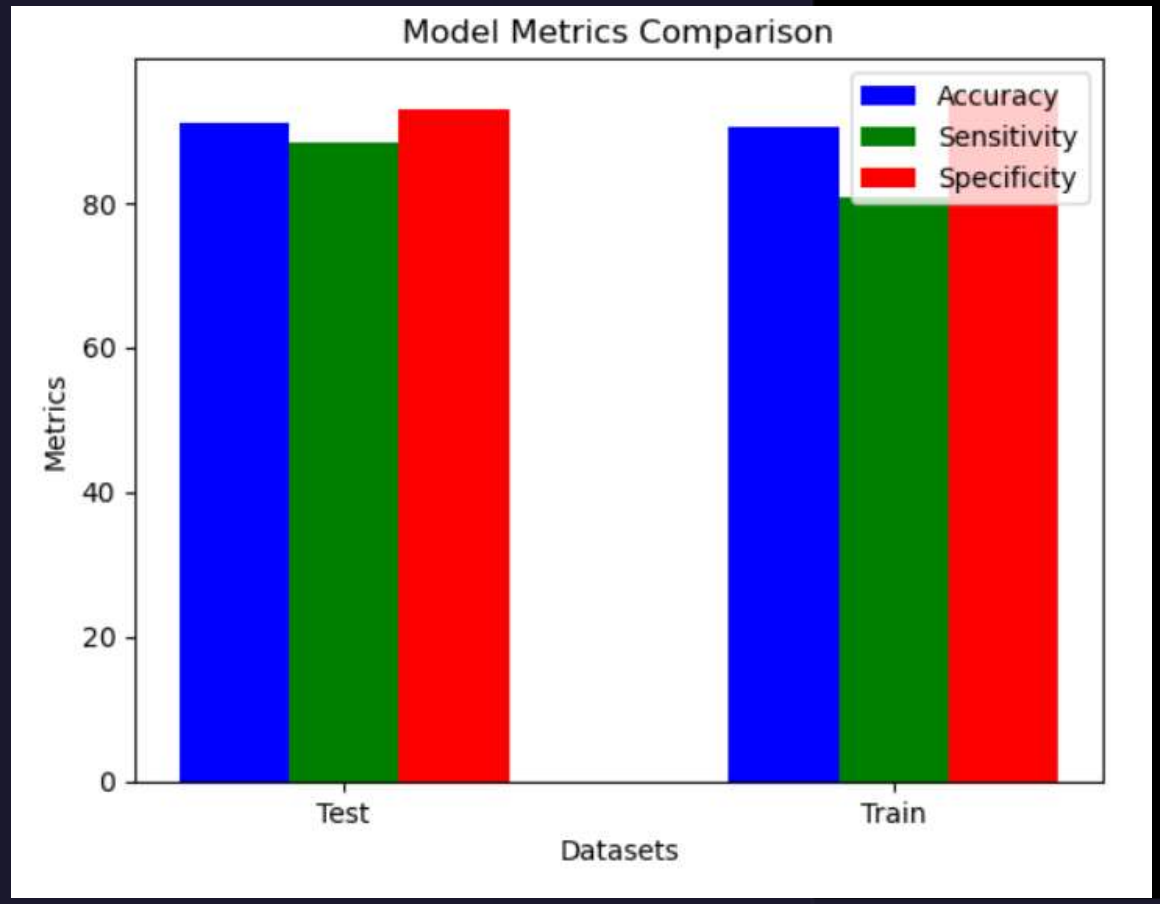
Model evaluation

Train Data

- Accuracy : 90.62%
- Sensitivity: 80.75%
- Specificity : 95.16%

Test Data

- Accuracy: 91.08%
- Sensitivity:88.34%
- Specificity:92.85%



Conclusion

The above metrics reflect the model's ability to make accurate predictions in both datasets. The accuracy values for the training and test sets are close, indicating that our model generalizes well to new, unseen data. Additionally, the model demonstrates strong specificity and sensitivity, indicating that it effectively identifies both true positives and true negatives.

- Our model demonstrates robust performance on both training and test datasets, with accuracy values around 91%.
- Sensitivity and specificity are well-balanced, indicating effective lead classification.
- The lead scores closely match actual conversion rates, making it a valuable tool for optimizing marketing efforts.
- In summary, our model is reliable and suitable for lead scoring, optimizing customer targeting and boosting conversion rates.

Thank You

Team Members:
Debanik Kundu
Shwetha Shankar
Leema Soaphy