



# Lead Scoring Case Study

Submitted By- Sanchit Kumar Sharma  
Arko Mallick  
Ankur Vishnoi

# 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

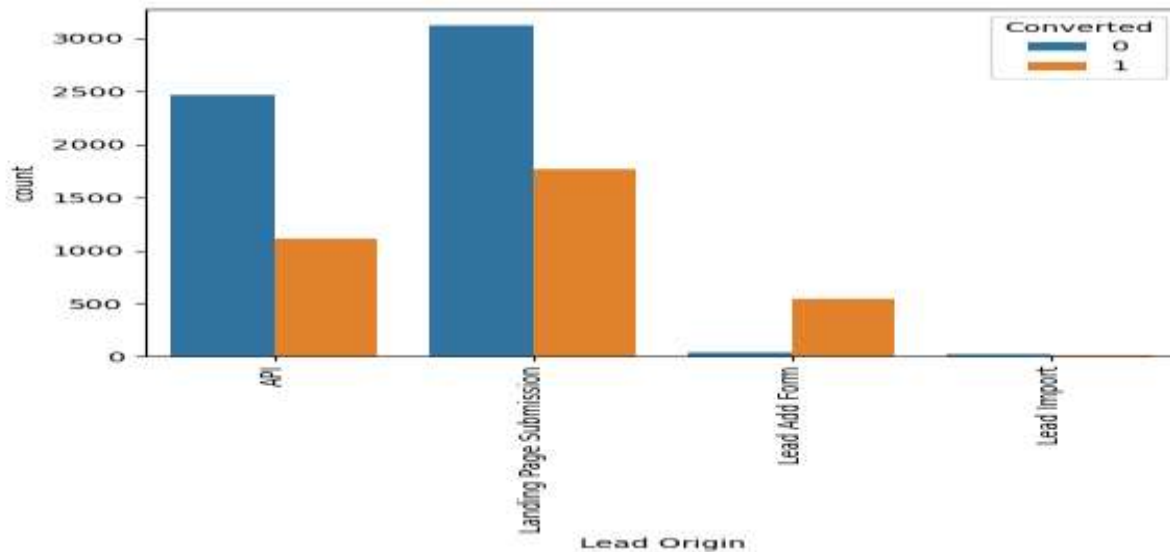
- X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company needs a model wherein you a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and 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%.

# Analysis Done and Steps

- Understanding the data
- Data preparation and cleaning
- Explanatory data analysis
- Univariate and Bivariate Analysis and finding Outliers in the data
- Prepare the data for model building
- Feature Scaling
- Model Building
- Testing the model on train and test set
- Evaluate model by different measures and metrics
- Measure the accuracy of the model and other metrics of evaluation.

# Explanatory Data Analysis

- Lead Origin Vs Converted

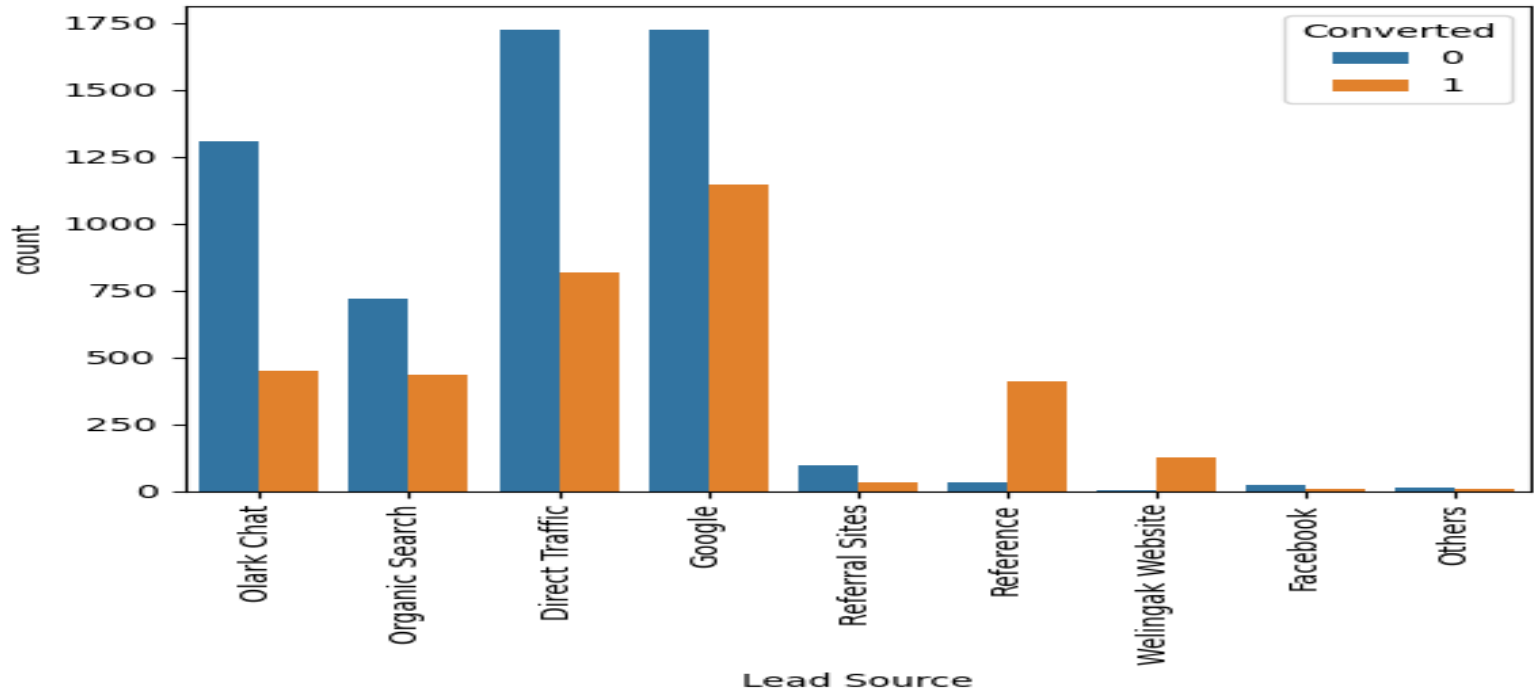


API and Landing Page Submission have 30-35% conversion rate but count of lead originated from them are considerable.

Lead Add Form has more than 90% conversion rate but count of lead are not very high.

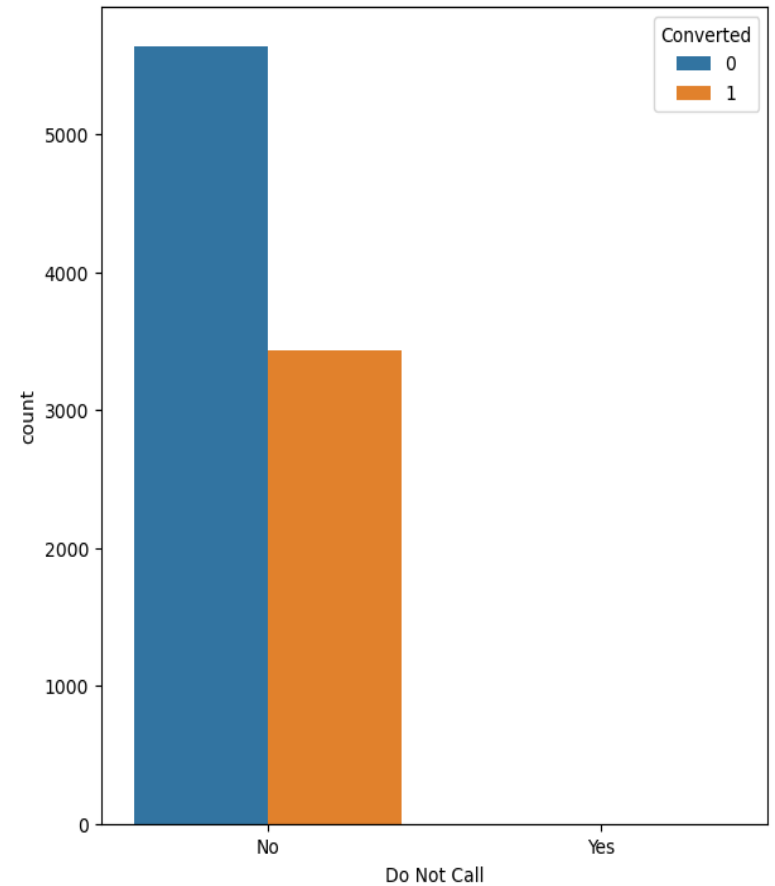
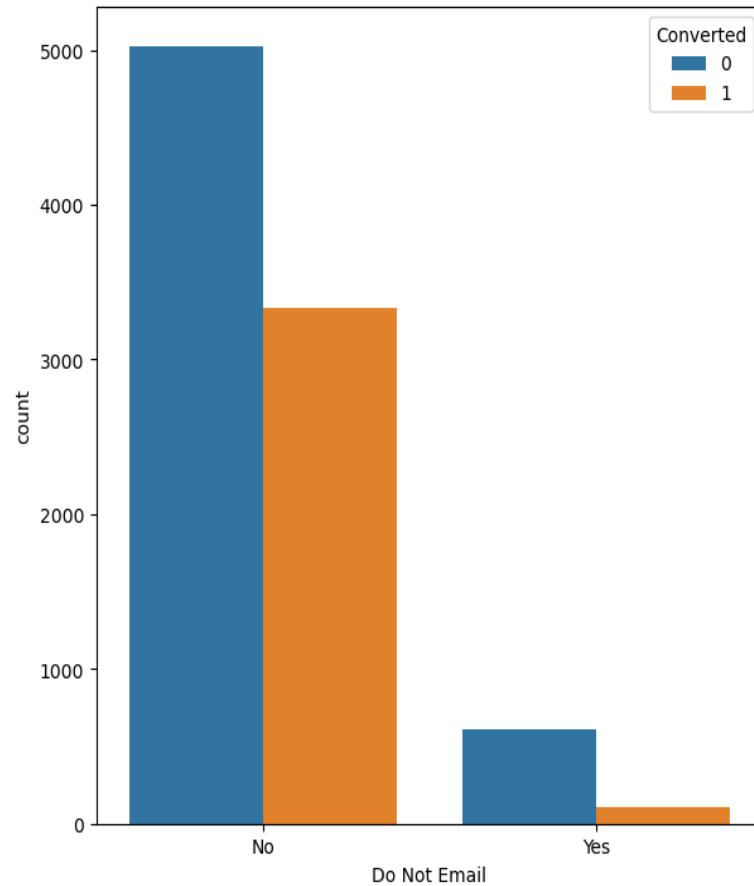
Lead Import are very less in count.

## Lead Source Vs Converted

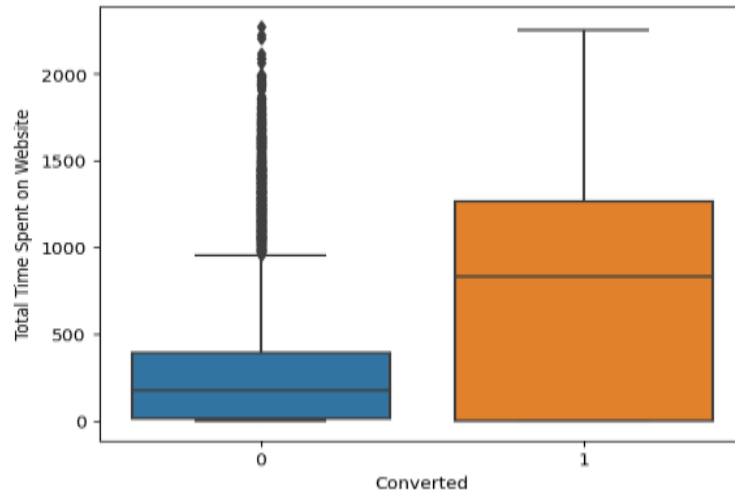


Google and Direct traffic generates maximum number of leads.  
Conversion Rate of reference leads and leads through welingak website is high.

# Do Not Email & Do Not Call

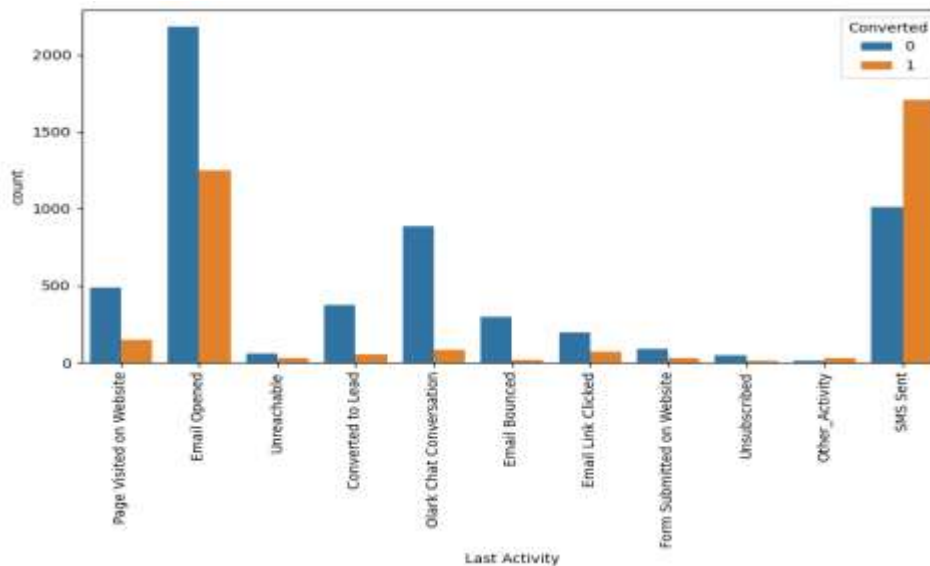


## Total time spent on website



Leads spending more time on the website are more likely to be converted.

## Last Activity

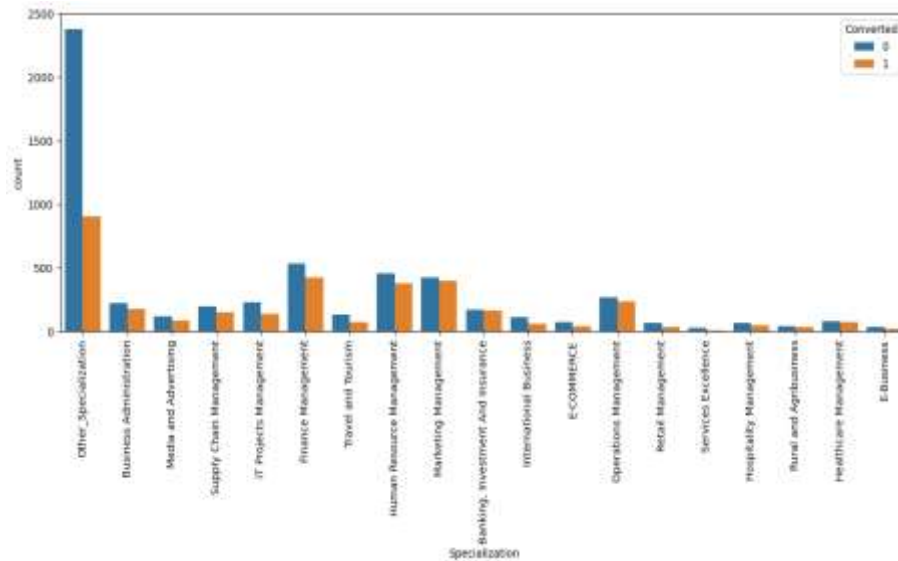


Most of the lead have their Email opened as their last activity.

Conversion rate for leads with last activity as SMS Sent is almost 60%.

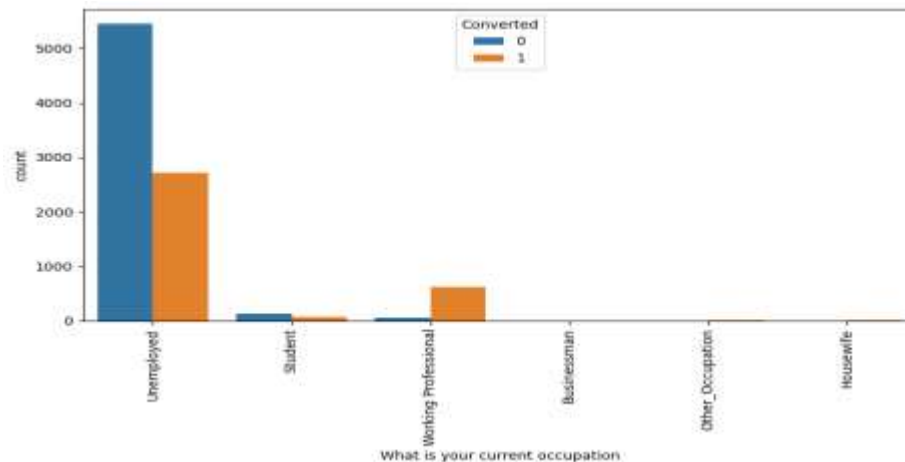


# Specialization



Focus should be more on the Specialization with high conversion rate.

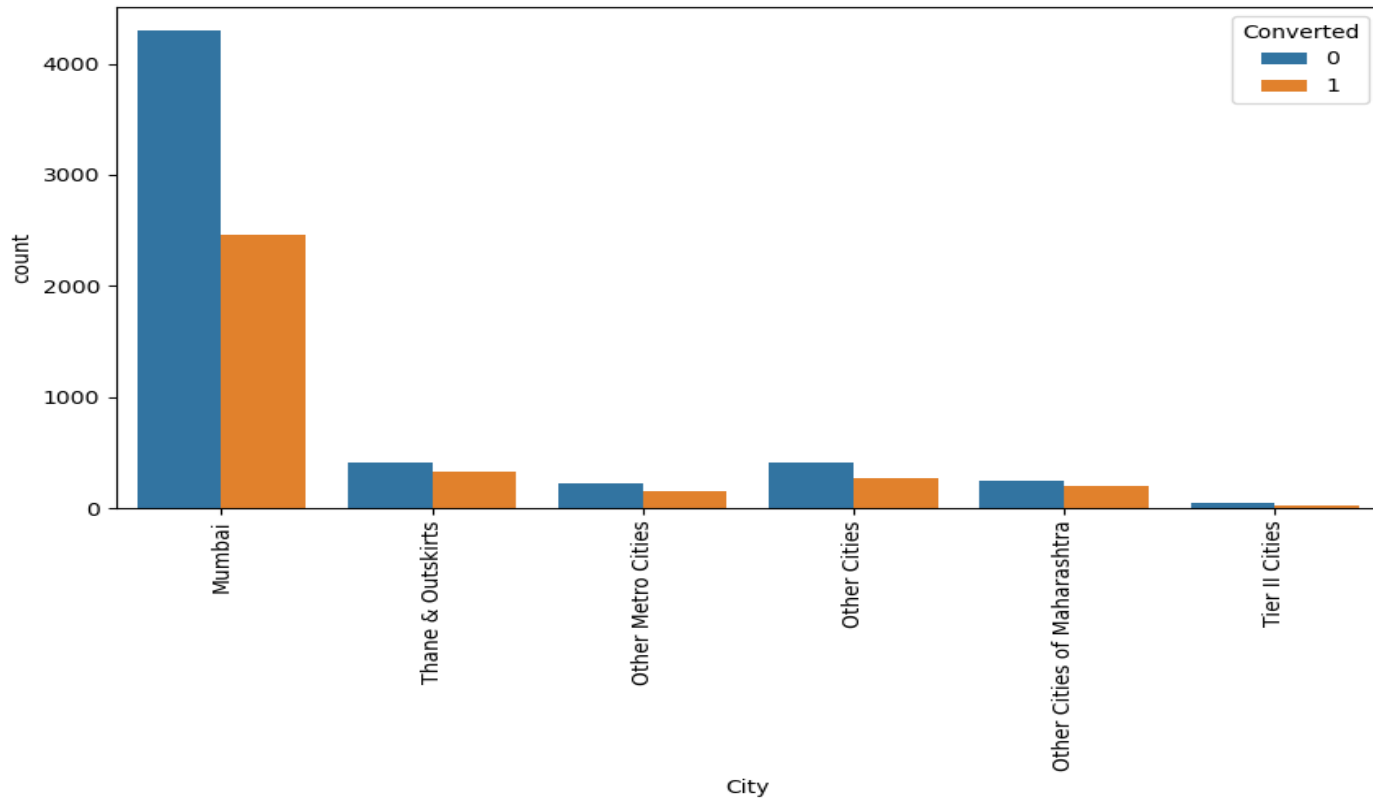
# Occupation



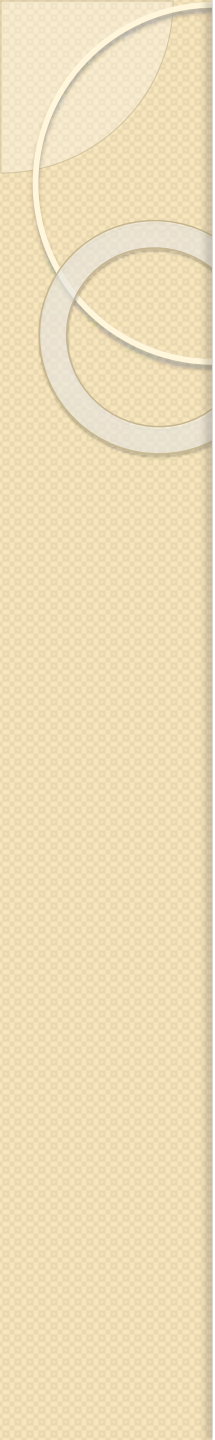
Working Professionals going for the course have high chances of joining it.

Unemployed leads are the most in numbers but has around 30-35% conversion rate.

# City



Most leads are from Mumbai with around 30% conversion rate.

- 
- Based on the Explanatory Data analysis we have seen that many columns are not adding any information to the model, hence we can drop them for further analysis

# Model Building

- Splitting into train and test set
- Scale variables in train set
- Build the first model
- Use RFE to eliminate less relevant variables
- Build the next model
- Eliminate variables based on high p-values
- Check VIF value for all the existing columns
- Predict using train set
- Evaluate accuracy and other metric
- Predict using test set
- Precision and recall analysis on test predictions
- ROC curve analysis

## Model Evaluation Test

Precision Score – 0.93

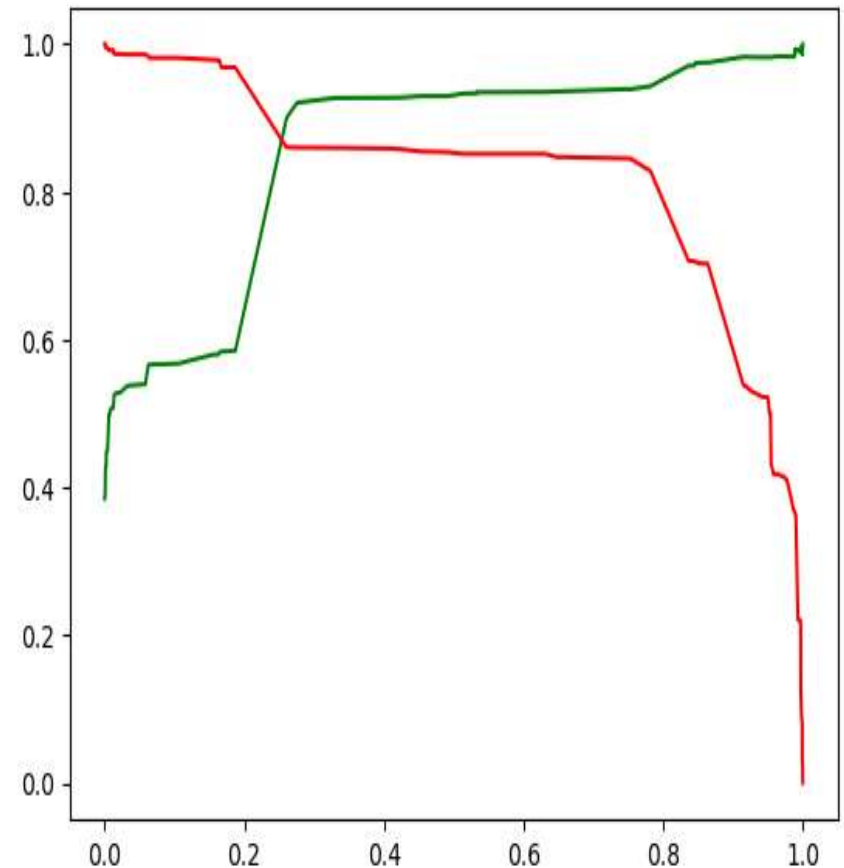
Recall Score – 0.85

Accuracy Sensitivity  
and Specificity

Accuracy- 81.8%

Sensitivity – 53.8%

Specificity – 99.4%



# Model Evaluation Train

**Specificity – 96.2%**

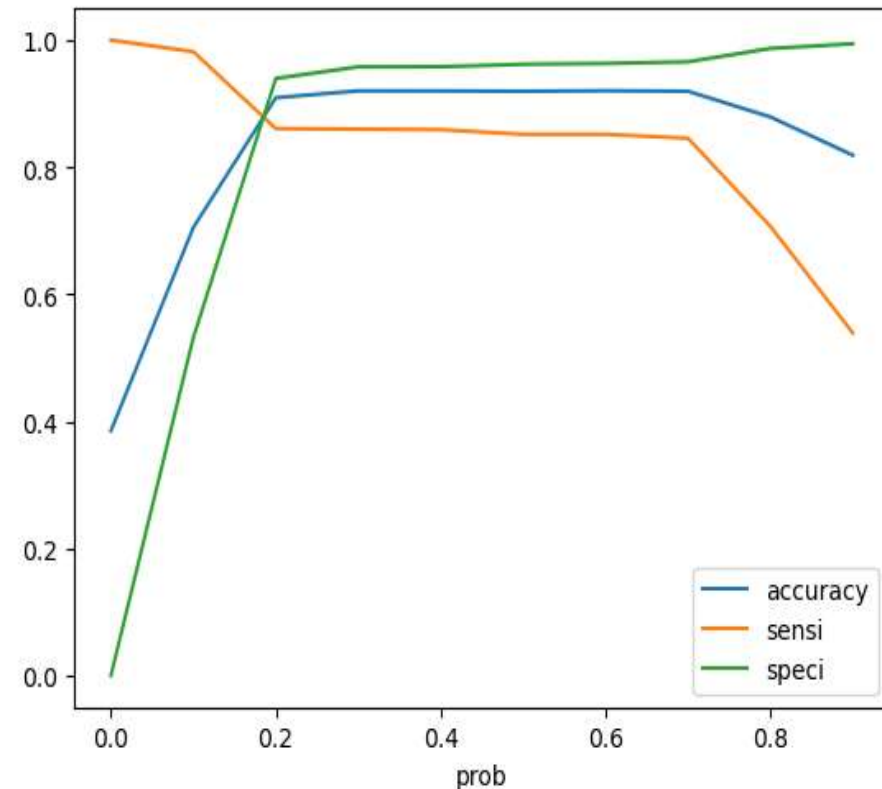
**Sensitivity- 85.1%**

**Negative Predictive Value-  
91.1%**

**Positive Predictive Value-  
93.3%**

**False Positive Rate – 3.8%**

**Accuracy – 91.9%**

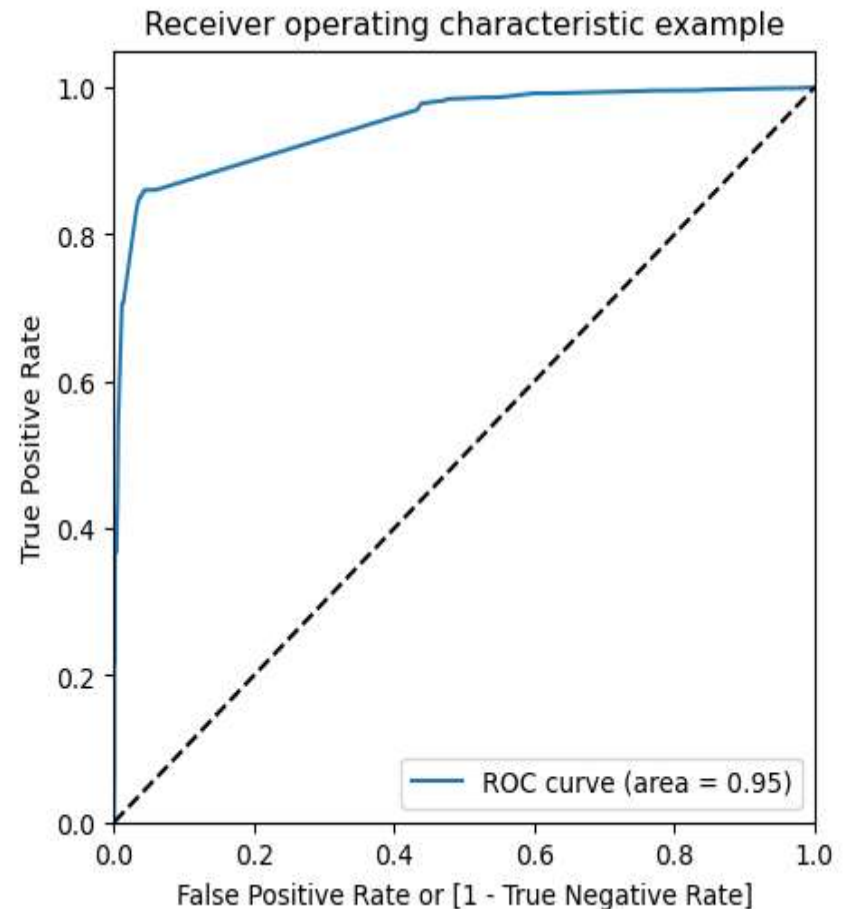


# ROC Curve

It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).

The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.

The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.



## Conclusions of EDA

- To improve overall lead conversion rate, we need to focus more on improving lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.
- To improve overall lead conversion rate, focus should be on improving lead conversion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.
- Median for converted and not converted leads are the same.
- Nothing conclusive can be said on the basis of Total Visits.
- Leads spending more time on the website are more likely to be converted.
- Median for converted and unconverted leads is the same.
- Nothing can be said specifically for lead conversion from Page Views Per Visit



# Logistic Regression Model(Conclusion)

- The Model shows High close to 81.8% accuracy.
- The threshold has been selected from accuracy, sensitivity, specificity measures and precision, recall curves.
- The Model shows sensitivity of 53.8% and specificity of 99.4%.
- The Model helps in finding promising leads and shows the accurate churn rate of the leads.
- According to our analysis this model proves to be accurate.

**Thank You**