# Lead Scoring Case Study

**Kishlay Pandey**
**Sayujya Vartak**

**DSC-38**

# Problem Statement

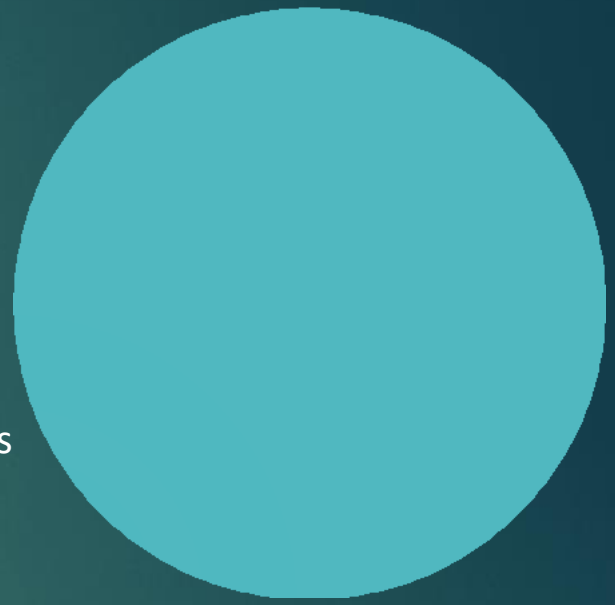
- X Education sells online courses to industry professionals.
- The company markets its courses on several websites and gets a lot of leads, Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30% which is very poor.
- 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.

# Business Objective

- X education wants to know most promising leads.
- The company requires a model to be built wherein 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 Solving Methodology

- Loading Datasets
- Data Understanding
    1. Checking Structure of datasets
    2. Checking other attributes like using info(), describe(), etc
- Data Cleaning
    1. Deleting Redundant columns and rows
    2. Data Quality check - Analyzing missing values, improper data types, duplicated rows
    3. Imputing Missing values
    4. Changing improper data types
- Handling Outliers
- Data Analysis (EDA)
    1. Imbalance Analysis
    2. Defining functions for plotting
    3. Univariate analysis of each variable



- Data Preparation
    1. Dummy Variable Creation
    2. Train Test Split
    3. Feature Scaling
    4. Looking at correlations
- Data Modelling
- Model Evaluation
    1. Calculating important metrics

       (sensitivity, specificity, recall, precision, F1-score, etc.)
    2. Plotting ROC curve
    3. Plotting precision and recall trade off
    4. Finding optimal cut off probability
    5. Making predictions using test dataset
- Final Observations and Summary

# Data Cleaning Approach

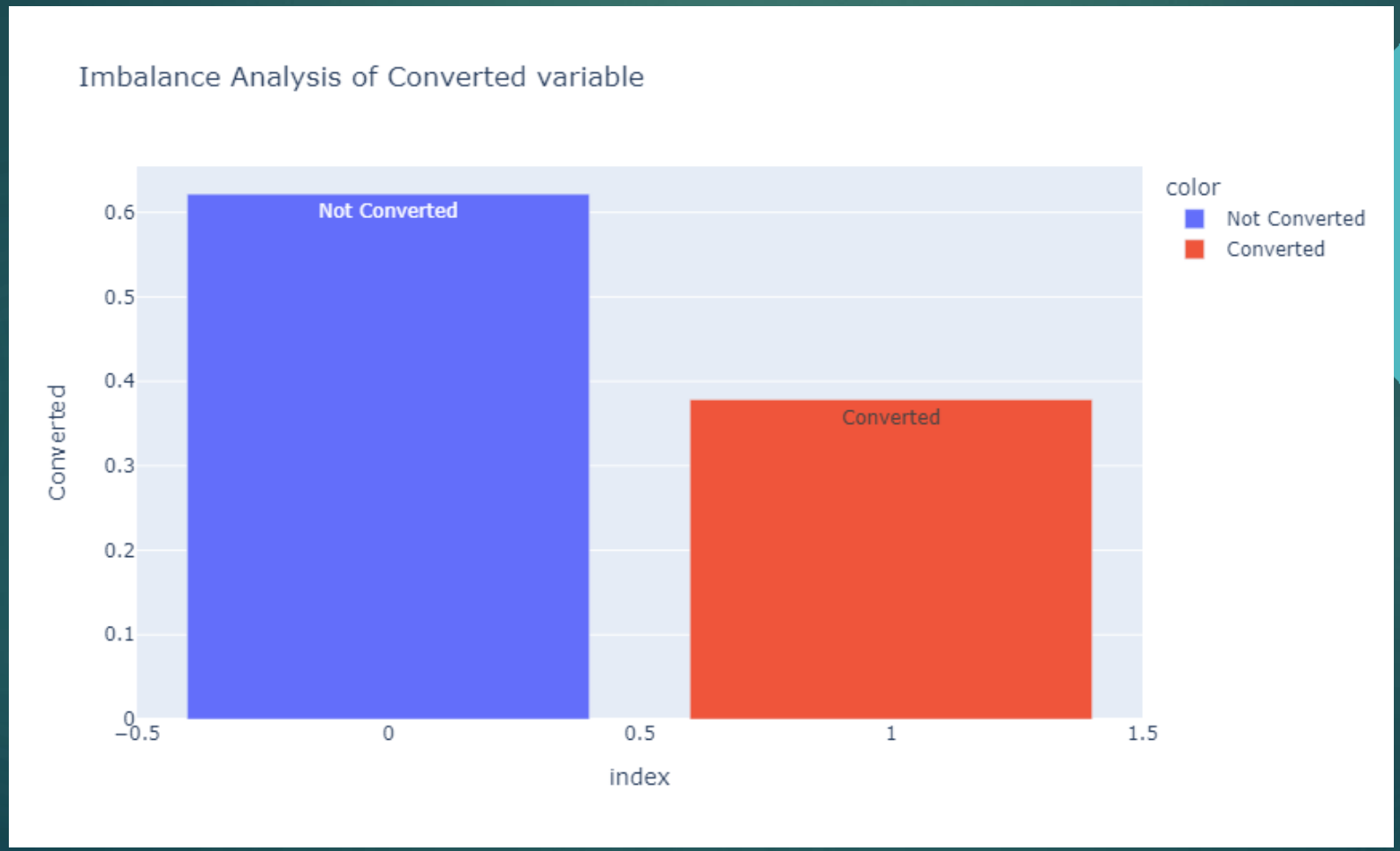
- Columns having more than 35% data missing have been dropped
- Columns having highly skewed data have been dropped
- Columns with missing values between 20% to 35% but have highly skewed data have been dropped
- For columns having less than 1% data missing the rows containing missing values have been dropped (About 98% data was retained after dropping such rows)
- For Categorical columns missing values we imputed using mode of that column
- All Numerical columns had missing values less than 1% so the rows were dropped containing missing values
- Also there were no duplicated rows
- If a row had more than 70% fields missing then it was as not of any use and should be dropped, but there we no such rows

# Data Analysis (EDA)

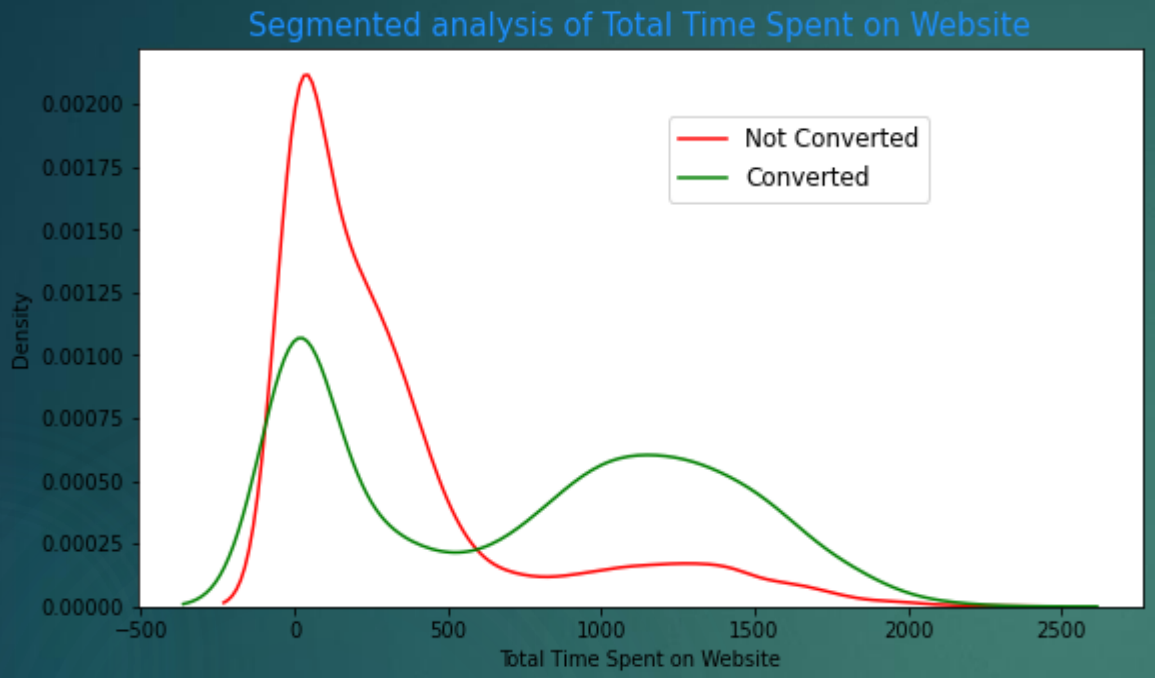
1. Imbalance Analysis Of Converted Variable
2. Segmented Analysis Of Numerical Variables With Respect To Converted Variable
3. Segmented Analysis Of Categorical Variables With Respect To Converted Variable

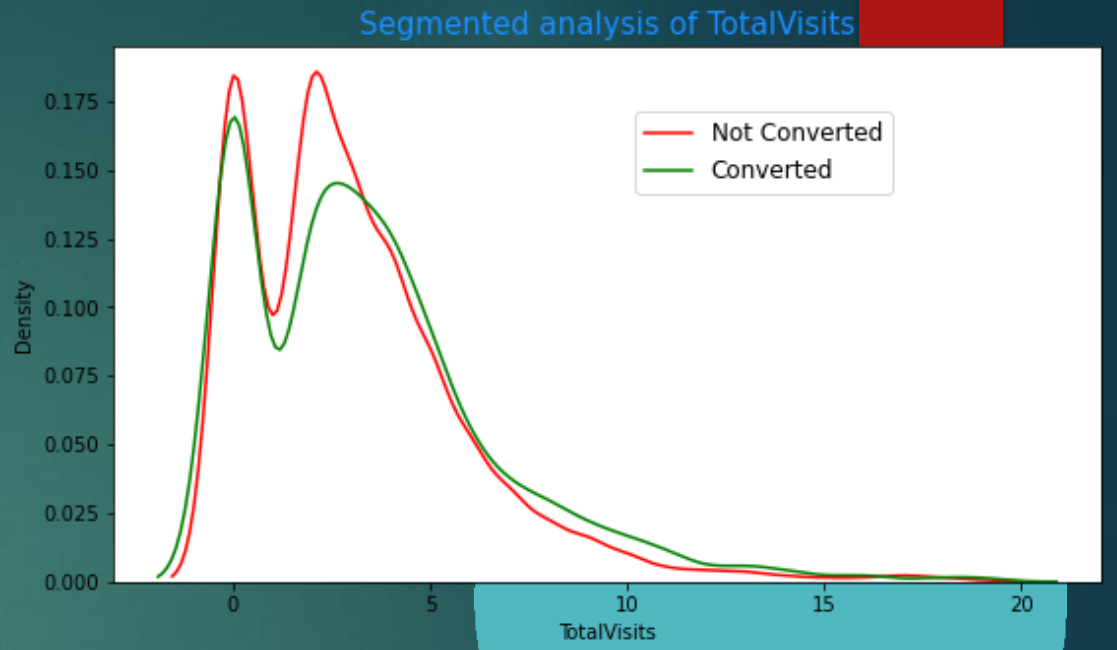
# Imbalance Analysis

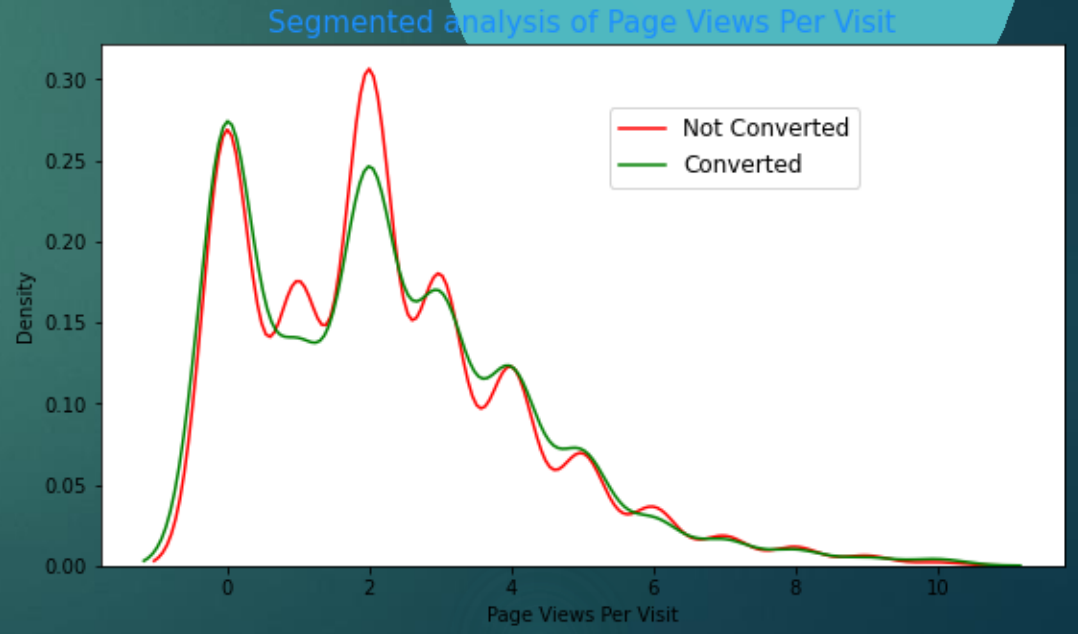
Only 38% Of Total Leads Have Been Converted Whereas 62% Are Not Converted



# Numerical Variables



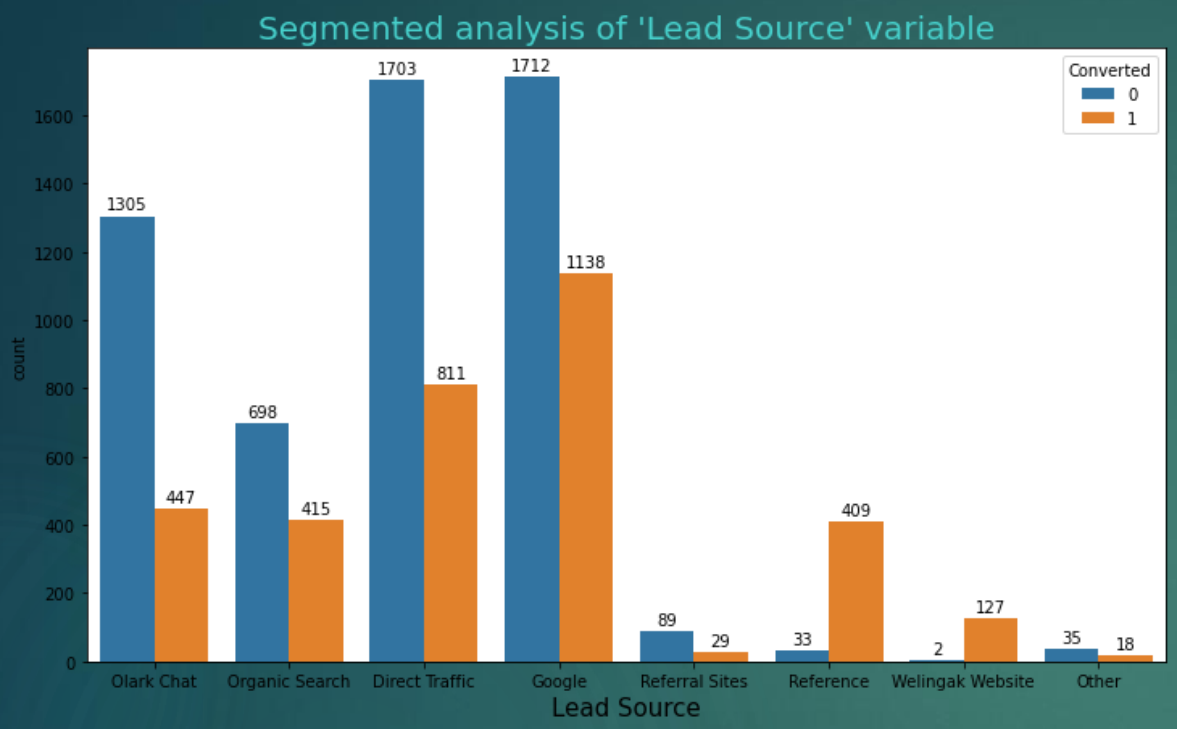



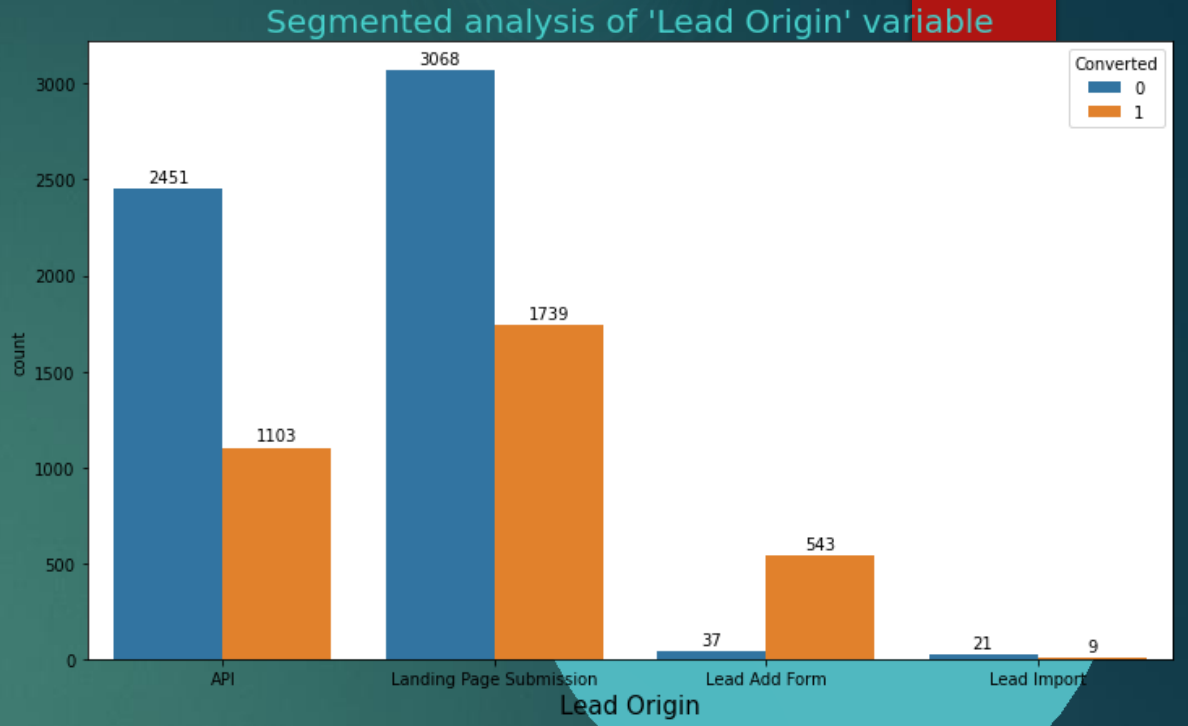



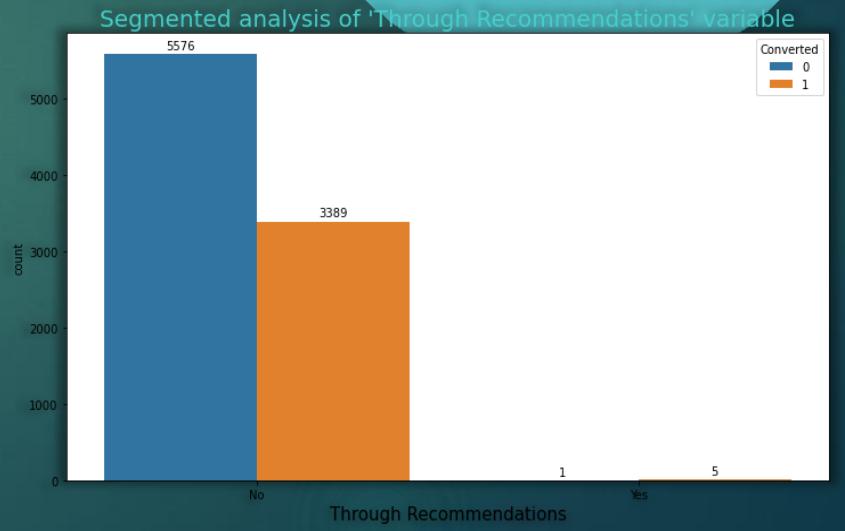

## Inference

1. People which spend more time(about 10 Minutes or more) on the website are more likely to convert than those which spend less than 10 minutes on website
2. For other numerical features(`TotalVisits` and `Page Views Per Visit`), there is not a significant divide between converted/ not-converted customers, but we can infer that, more the pages viewed by customer per visit or more the total visits of a customer, more is the chance of him being converted

# Categorical Variables



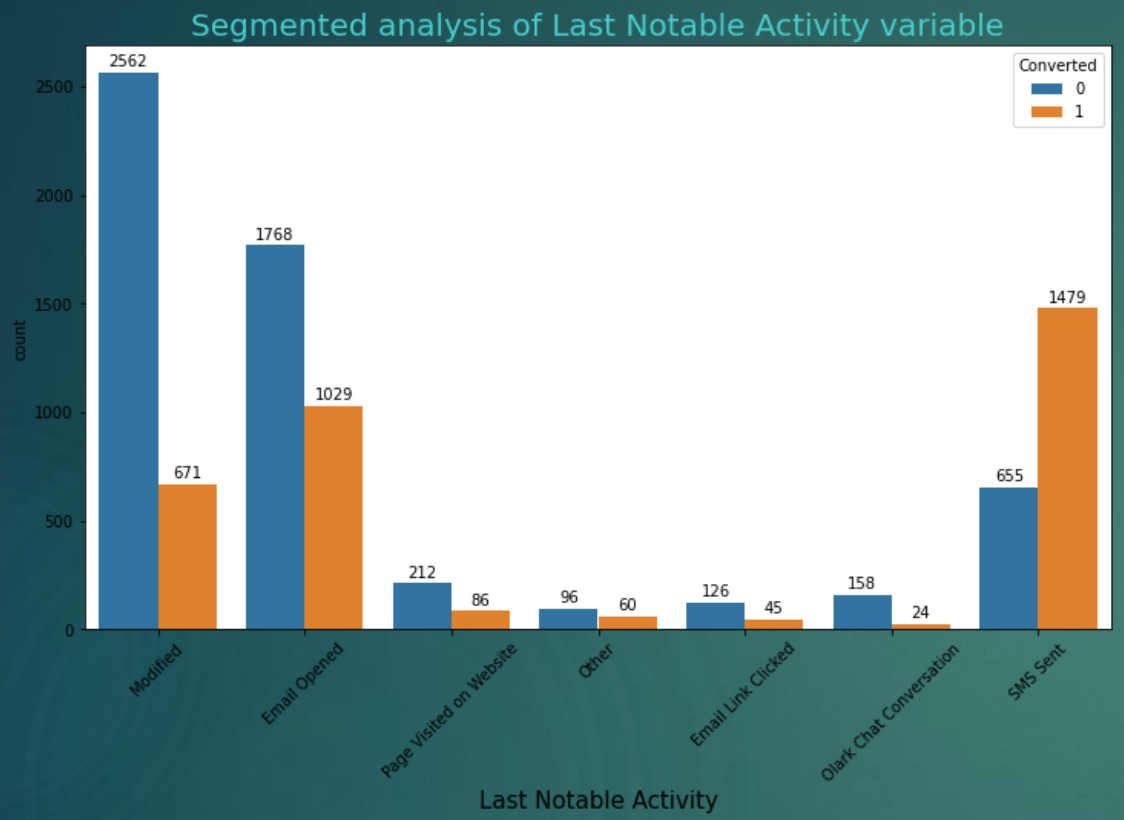



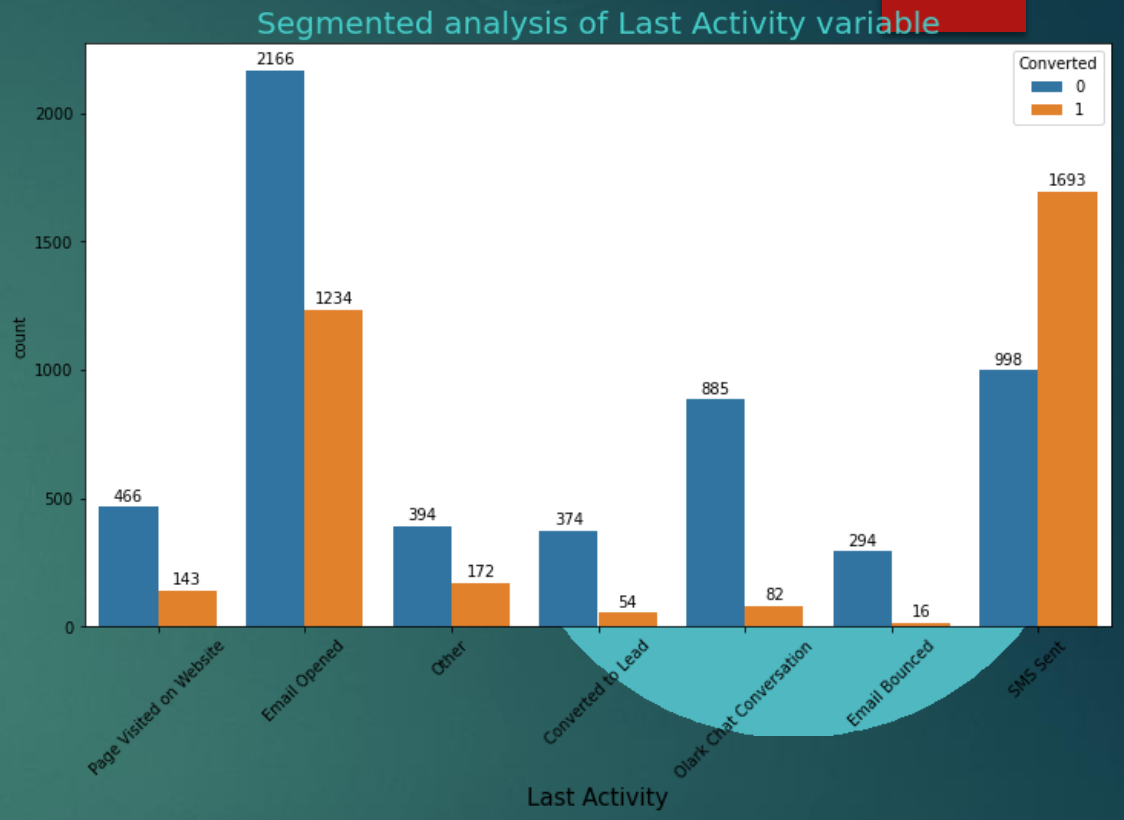




## Inference

- Leads generated from `reference`, `Welingak Website` and `Lead Add Form Lead Origin` can be considered as hot leads as the conversion rate is more than 90%
- Leads which came through recommendation of other customers can also be considered as hot leads due to high conversion rate
- In `Through Recommendations` column `Yes` category though having very few entries, has the highest conversion rate, greater than 83%

# Categorical Variables

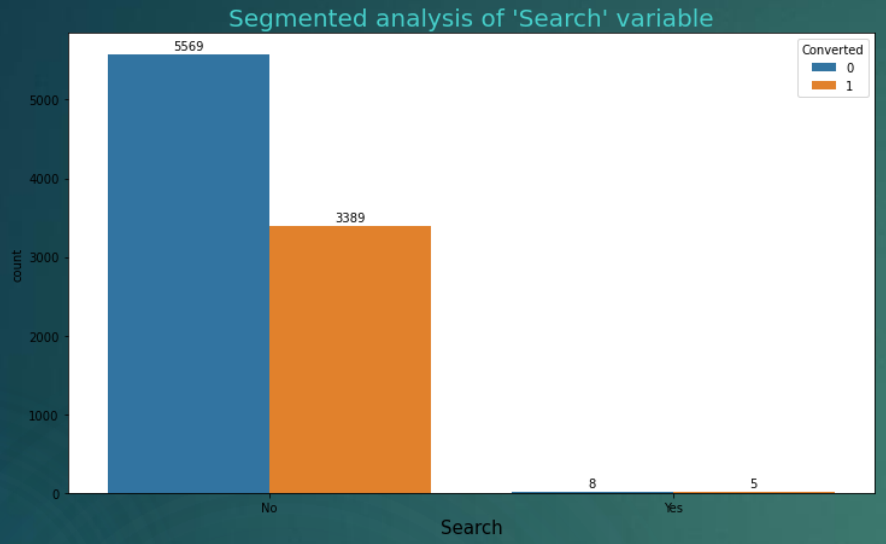


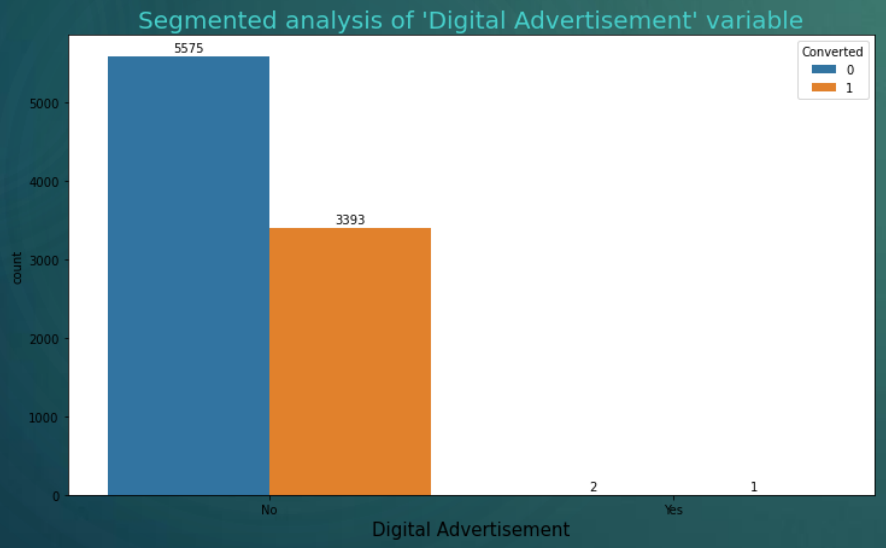



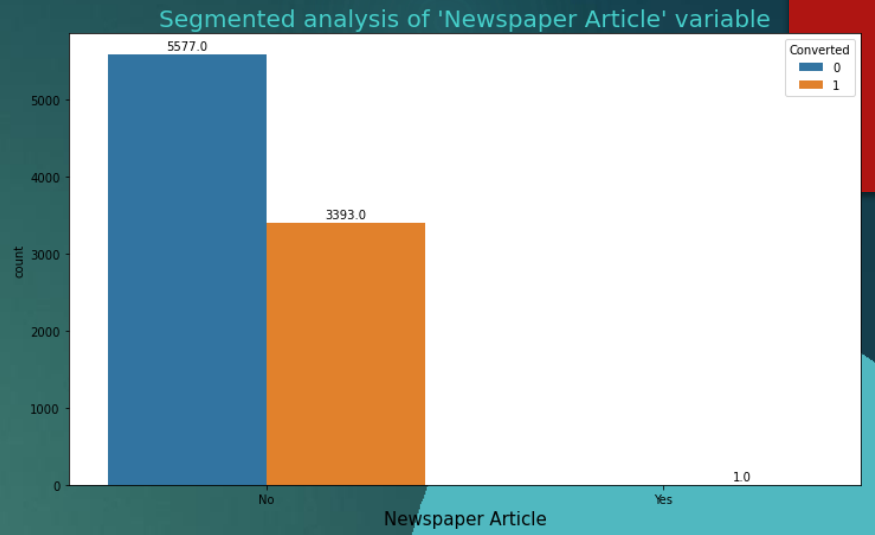

## Insights

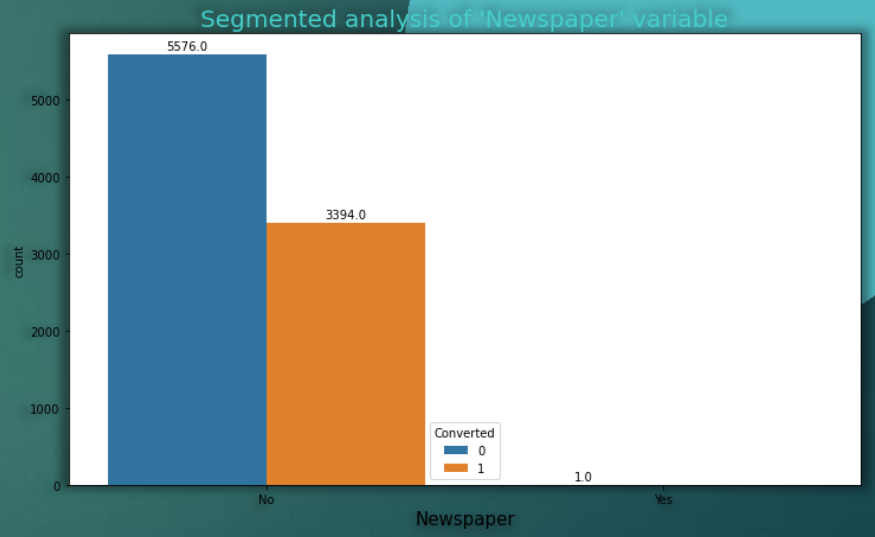
- In `Last Activity` column `SMS Sent` category has the highest conversion rate of 63% while `Olark Chat Conversation` and `Email Bounced` categories have the lowest conversion rates of 9% and 8%
- In `Last Notable Activity` column `SMS Sent` category has the highest conversion rate of about 70%

# Categorical Variables



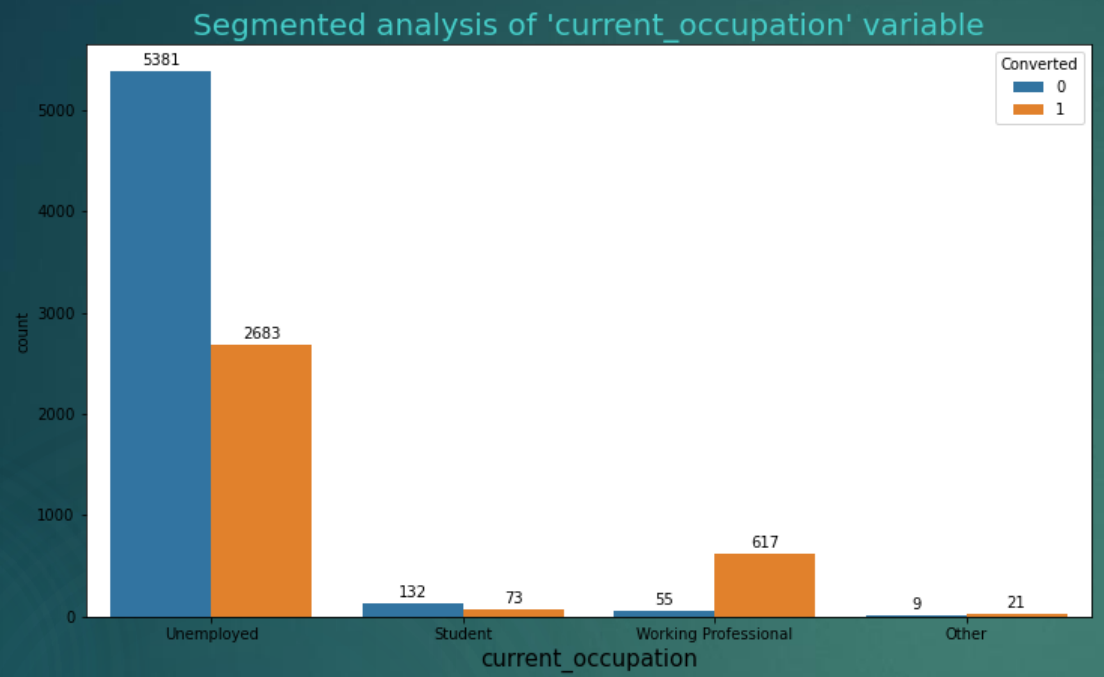



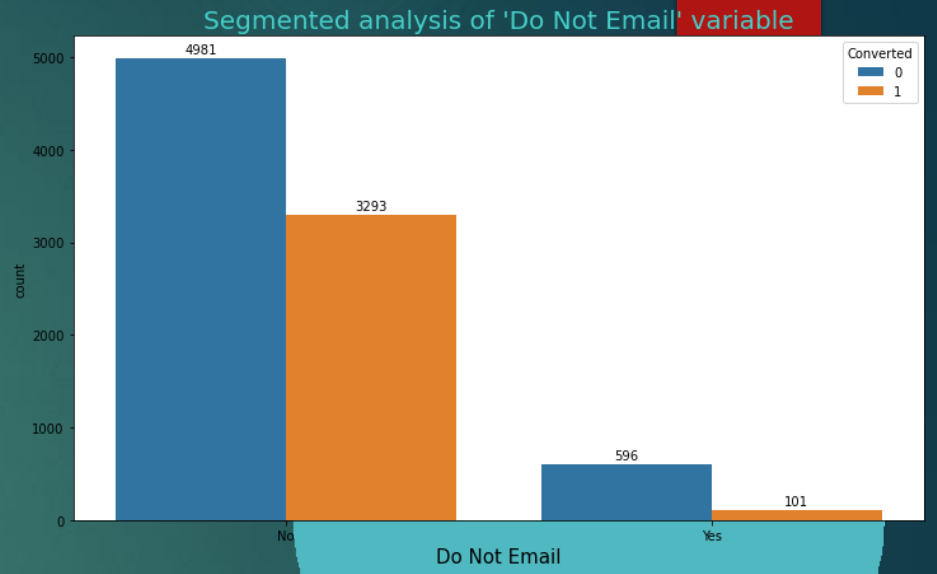



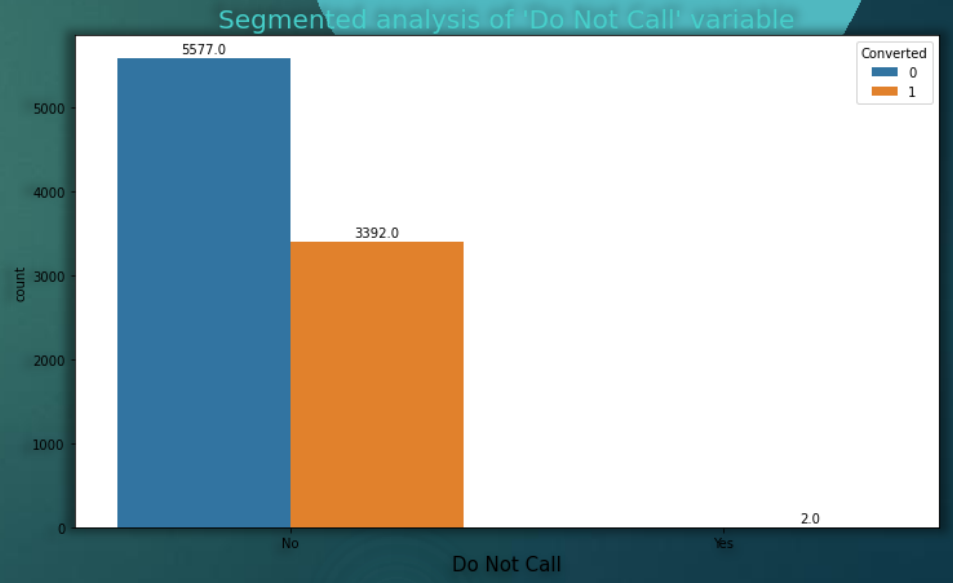

## Insights

- About 99% of customers hadn't seen any ad through above given channels
- All the columns above have about 38% conversion rate for the category 'No'

# Categorical Variables




## Insights

- In `Do Not Email` column major conversion(about 40%) has happened when customer opted for it
- In `Do Not call` column major conversion(about 38%) has happened when customer opted for it. Also though only 2 customers opted to not get a call but they both got converted
- In `current_occupation` column `Working Professional` category has the highest conversion rate, greater than 67% followed by `Other` category having conversion rate of 60%
- Working Professionals are most likely to convert

# Variables Impacting Conversion Rate

## Customer Filled Variables Which Impact Conversion Rate

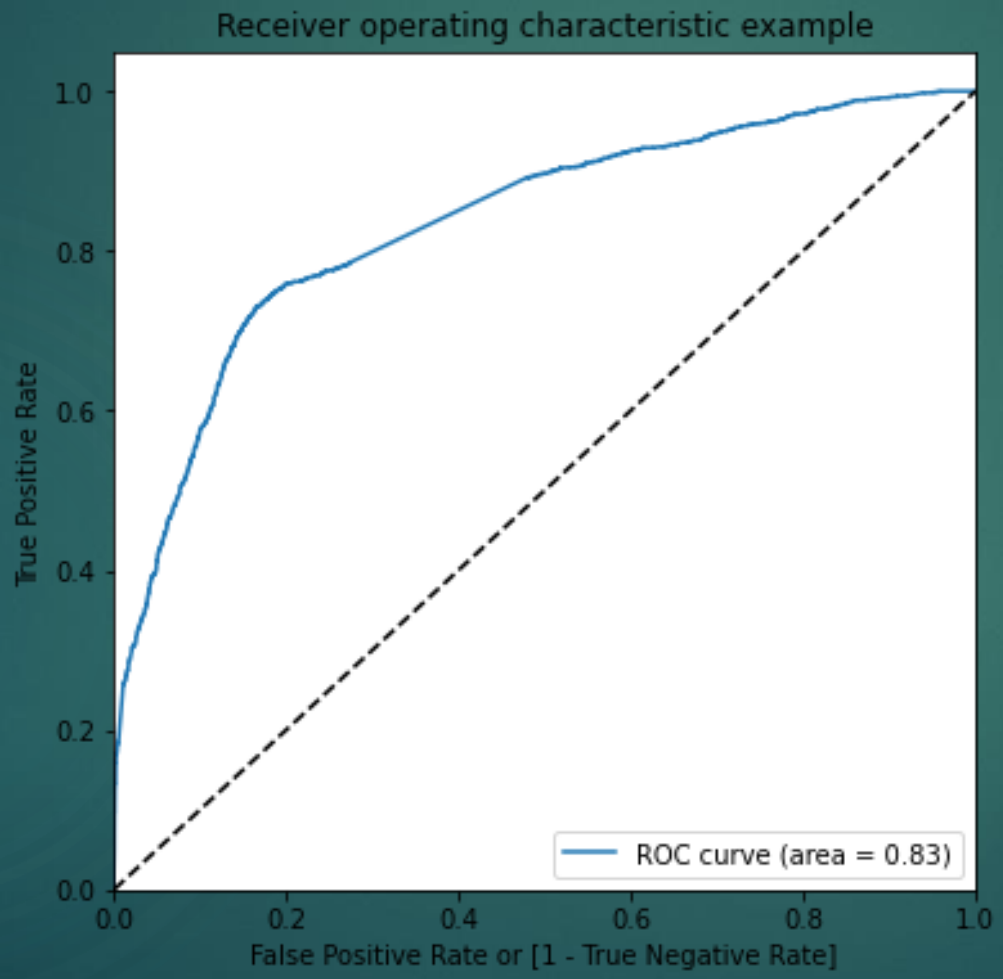
- Do Not Email
- Totalvisits
- Total Time Spent on Website
- Lead Origin_lead Add Form
- Lead Source_Direct Traffic
- Lead Source_Google
- Lead Source_Organic Search
- Lead Source_Referral Sites
- Lead Source_Welingak Website
- current_occupation_Working Professional

## Sales Team Generated Variables which Impact Conversion Rate

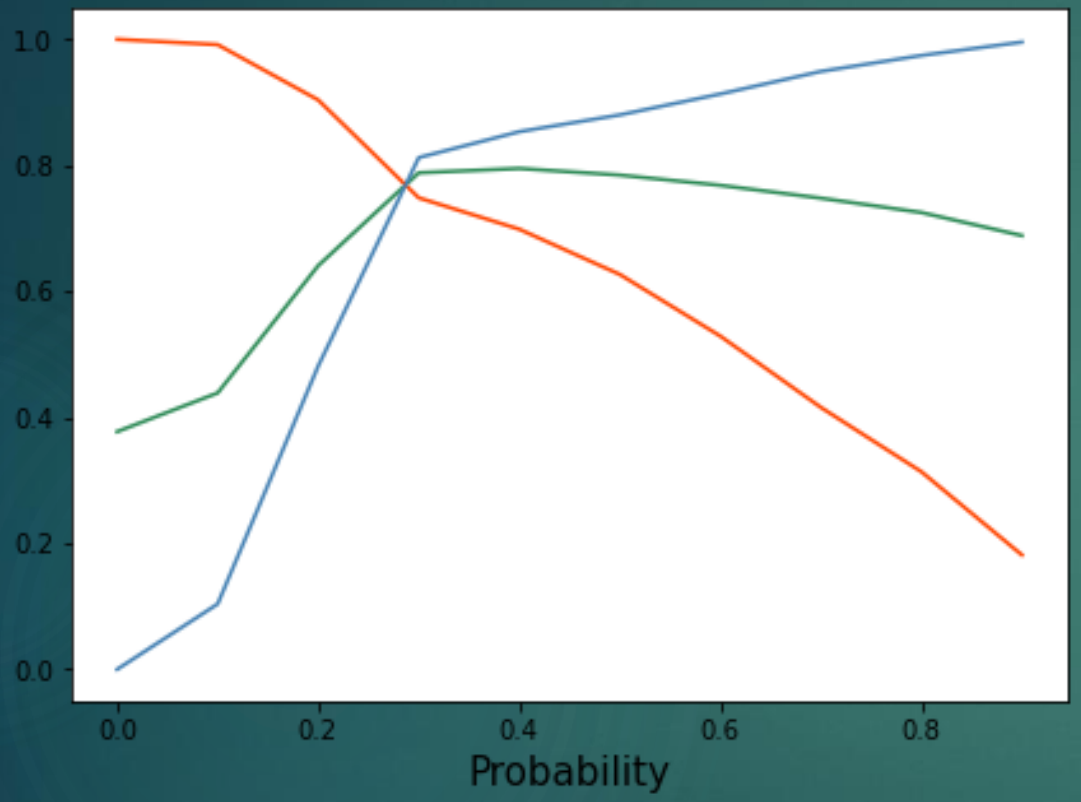
- Last Activity_SMS Sent
- Last Notable Activity_SMS Sent

# Model Evaluation on Train Dataset

## ROC Curve



# Sensitivity-Specificity



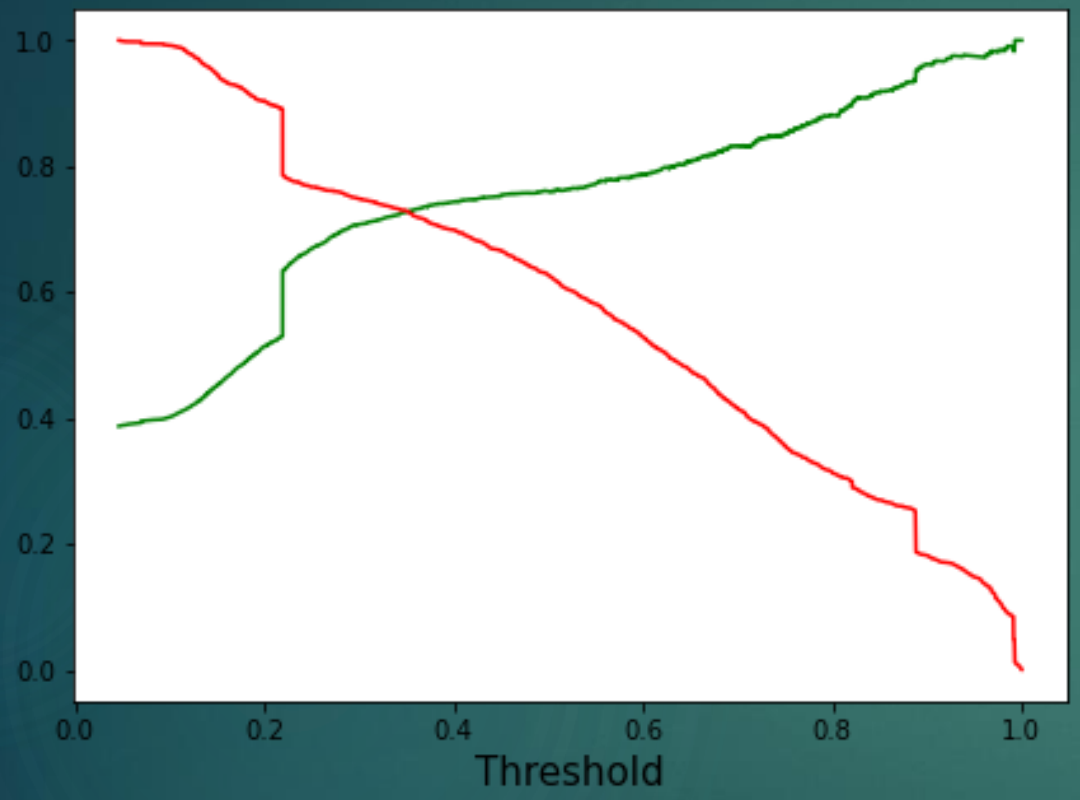

When using the Sensitivity-Specificity-Accuracy plot, we found out that the optimal cutoff point is 0.3.

Confusion Matrix

| | |
|---|---|
| 3176 | 734 |
| 596 | 1773 |

- Accuracy Score of 78.8%
- Specificity = 81.2%
- Sensitivity = 74.8%
- False Positive Rate = 18%
- Positive Predictive Rate = 70.7%
- Negative Predictive Rate = 84%

# Precision-Recall Tradeoff




When using the Precision-Recall Tradeoff,
we found out that the optimal cutoff point is 0.35.

Confusion Matrix

| | |
|---|---|
| 3265 | 645 |
| 643 | 1727 |

- Accuracy Score of 79.5%
- Precision = 72.8%
- Recall = 72%
- Specificity = 83.5%

# Model Evaluation On Test Dataset

Using the sensitivity-specificity we found out optimal cutoff point was 0.3.
When we plotted the precision-recall tradeoff, we got the optimal cutoff about 0.35
As we have a bit more emphasis on sensitivity as we want to predict hot leads we will choose 0.3 as our final optimal cutoff

- Accuracy Score of 80.2%
- Specificity = 82.9%
- Sensitivity = 75.8%

Confusion Matrix


The Cutoff 0.3 represents that if a lead has a leads score of 30 or above then that lead has high probability of converting and should be considered as a hot lead

# Conclusion

- Before building the machine learning model we had a conversion rate of 38% but after we built the model we were able to predict 80% of leads which were converted giving us a conversion rate of 80%, i.e. an increase of 42%.
- A lead having Lead Score **greater than 30** should be considered as a **Hot Lead**
- Customers which spend about 10 minutes or more on the website or have Lead Origin as 'Lead add form' should be considered as hot leads as they have a high probability of getting converted.
- Customers having Lead Source as 'Reference' or 'Wellingak Website' OR customers which are working professional have a high probability of getting converted and should be considered as Hot Leads
- Leads generated from customers which visit the website more often (more than 5 times) have a moderate to high chance of getting converted

# Recommendations

- There are a lot of leads generated in the initial stage but only a few of them come out as paying customers. 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.

- When the leads conversion needs to be aggressive sales team can adopt the strategy of connecting with all the leads having a lead_score above 30 and educate them about the product maintain a constant communication.

  If all customers having a lead_score above 30 have been targeted then the sales team can also target customers with lead_score of 20 and above as these customers can also convert

- When the sales team needs to minimize the rate of useless phone calls they can choose only the Leads having a lead_score above 60 or 70

  If a customer has a lead score lower than 60 or 70 but he/she has a combination of attributes which are, visited the website multiple times or spent more than 10 minutes on the website or is a working professional or has Lead Source as ‘Welingak Website’ or has Lead Origin as ‘Lead Add Form’ or has come on the website through someone’s recommendation, then they are most likely to convert and such customers can also be targeted