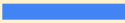


Lead Scoring Case Study



Contributors

Rishav Bhattacharyya

Rajat Roy

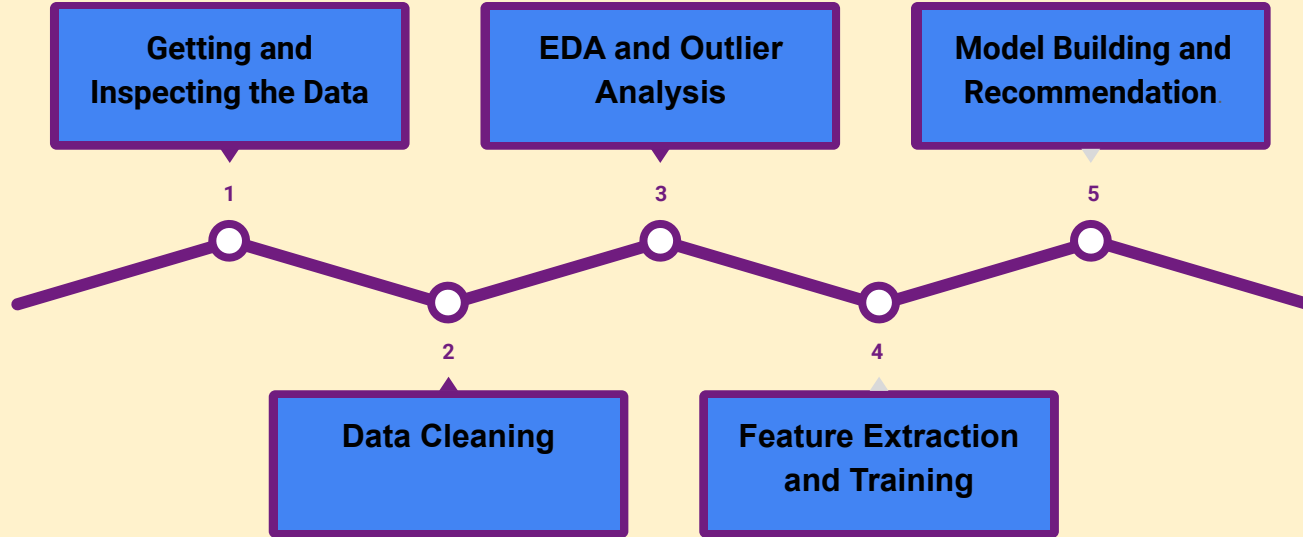
Problem Statement

- **X Education** is a platform which sells online courses to industry professionals.
- People on this platform fill up a form providing their email address or phone number, they are classified to be a **lead**.
- **Lead Conversion Rate(LCR)** = $\text{Total Students acquired} / \text{Total Leads Produced}$
- LCR for X education is very poor(~**30%**)
- In order to increase the LCR. X education need to identify the most potential leads(**Hot Leads**) i.e. which have very high chance of getting converted.

Scope of this deck

- Analysing the factors which are responsible for a lead to be highly convertible.
- Identifying the most promising or potential leads for X education using Machine learning Techniques.

Flow of Analysis



How the Data looks like..

| | Prospect ID | Lead Number | Lead Origin | Lead Source | Do Not Email | Do Not Call | Converted | TotalVisits | Total Time Spent on Website | Page Views Per Visit | Last Activity | Country | Specialization | How did you hear about X Education |
|---|--------------------------------------|-------------|-------------------------|----------------|--------------|-------------|-----------|-------------|-----------------------------|----------------------|-------------------------|---------|-------------------------|------------------------------------|
| 0 | 7927b2df-8bba-4d29-b9a2-b6e0beafe620 | 660737 | API | Olark Chat | No | No | 0 | 0.0 | 0 | 0.0 | Page Visited on Website | NaN | Select | Select |
| 1 | 2a272436-5132-4136-86fa-dcc88c88f482 | 660728 | API | Organic Search | No | No | 0 | 5.0 | 674 | 2.5 | Email Opened | India | Select | Select |
| 2 | 8cc8c611-a219-4f35-ad23-fdfd2656bd8a | 660727 | Landing Page Submission | Direct Traffic | No | No | 1 | 2.0 | 1532 | 2.0 | Email Opened | India | Business Administration | Select |
| 3 | 0cc2df48-7cf4-4e39-9de9-19797f9b38cc | 660719 | Landing Page Submission | Direct Traffic | No | No | 0 | 1.0 | 305 | 1.0 | Unreachable | India | Media and Advertising | Word Of Mouth |

Each Row corresponding to a unique lead

Total Leads: **9240**
Independent Variables: **36**
Dependent/Target - **1**

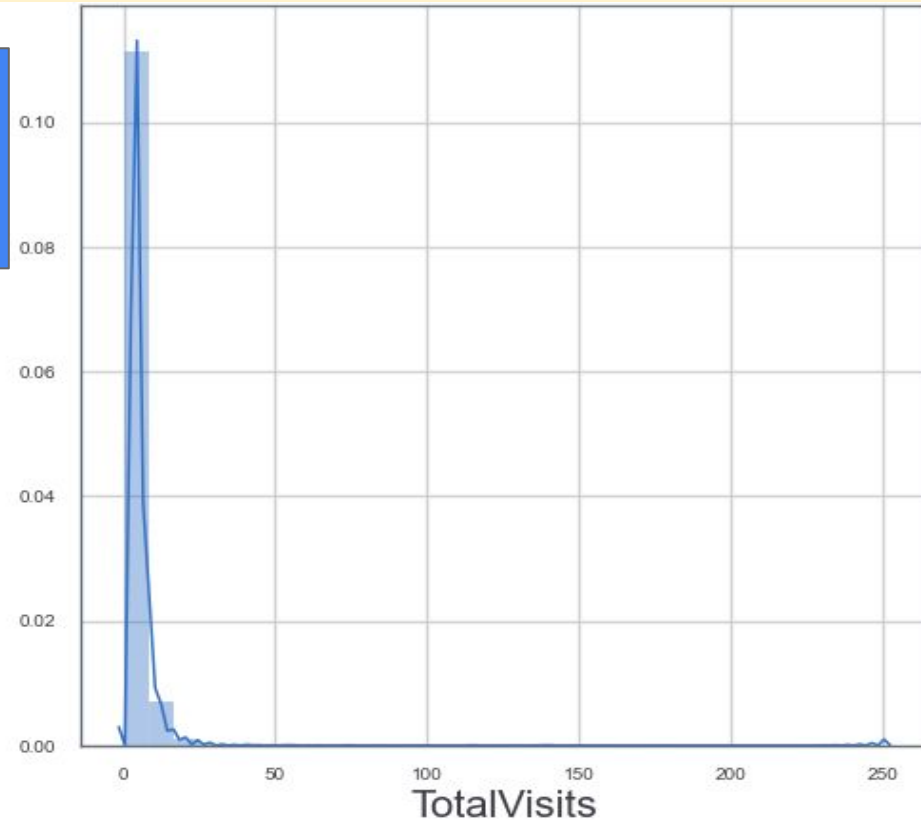
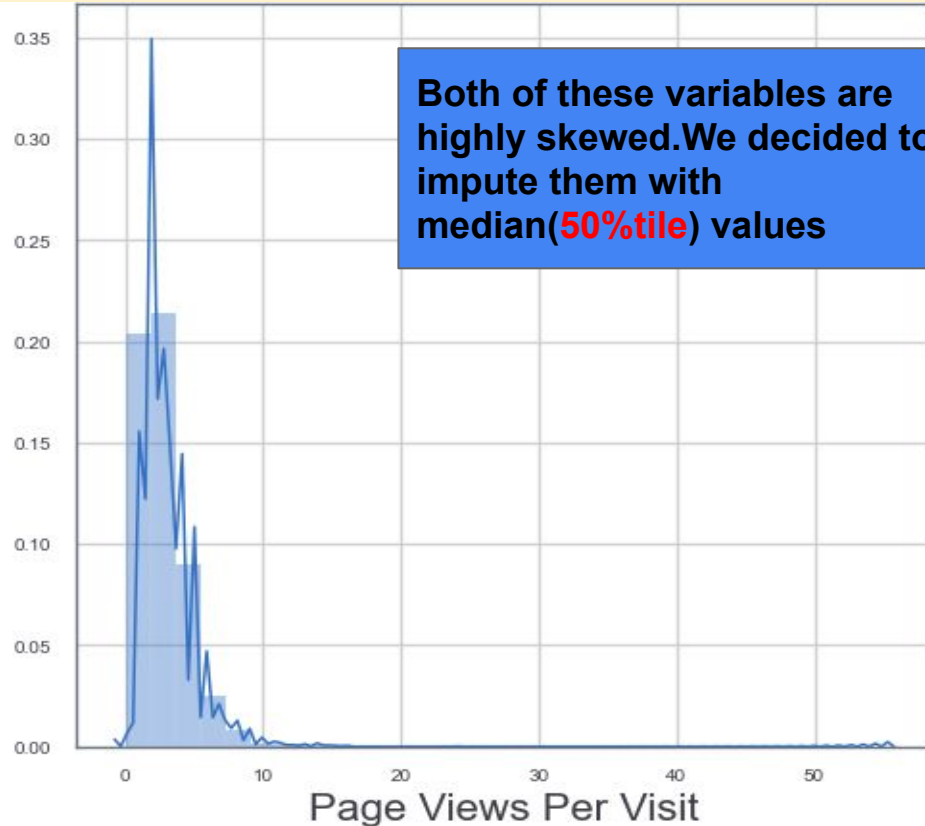
Data Cleaning

Missing Value Percentages

| columns | missing_perc |
|---|--------------|
| Lead Quality | 51.590909 |
| Asymmetrique Activity Index | 45.649351 |
| Asymmetrique Profile Score | 45.649351 |
| Asymmetrique Activity Score | 45.649351 |
| Asymmetrique Profile Index | 45.649351 |
| Tags | 36.287879 |
| Lead Profile | 29.318182 |
| What matters most to you in choosing a course | 29.318182 |
| What is your current occupation | 29.112554 |
| Country | 26.634199 |
| How did you hear about X Education | 23.885281 |
| Specialization | 15.562771 |
| City | 15.367965 |
| Page Views Per Visit | 1.482684 |
| TotalVisits | 1.482684 |
| Last Activity | 1.114719 |
| Lead Source | 0.389610 |

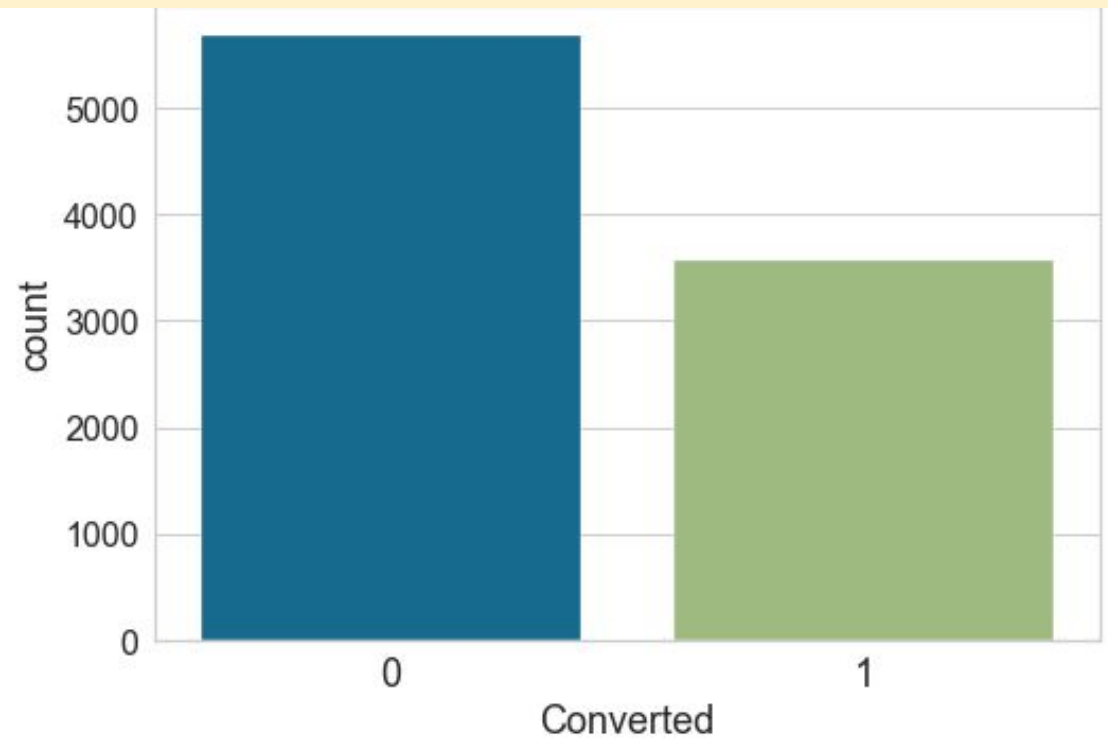
Columns with more than **45%** values missing have been dropped

Numerical Column Distribution(With Missing Values)



Exploratory Data Analysis

Overall Converted Data



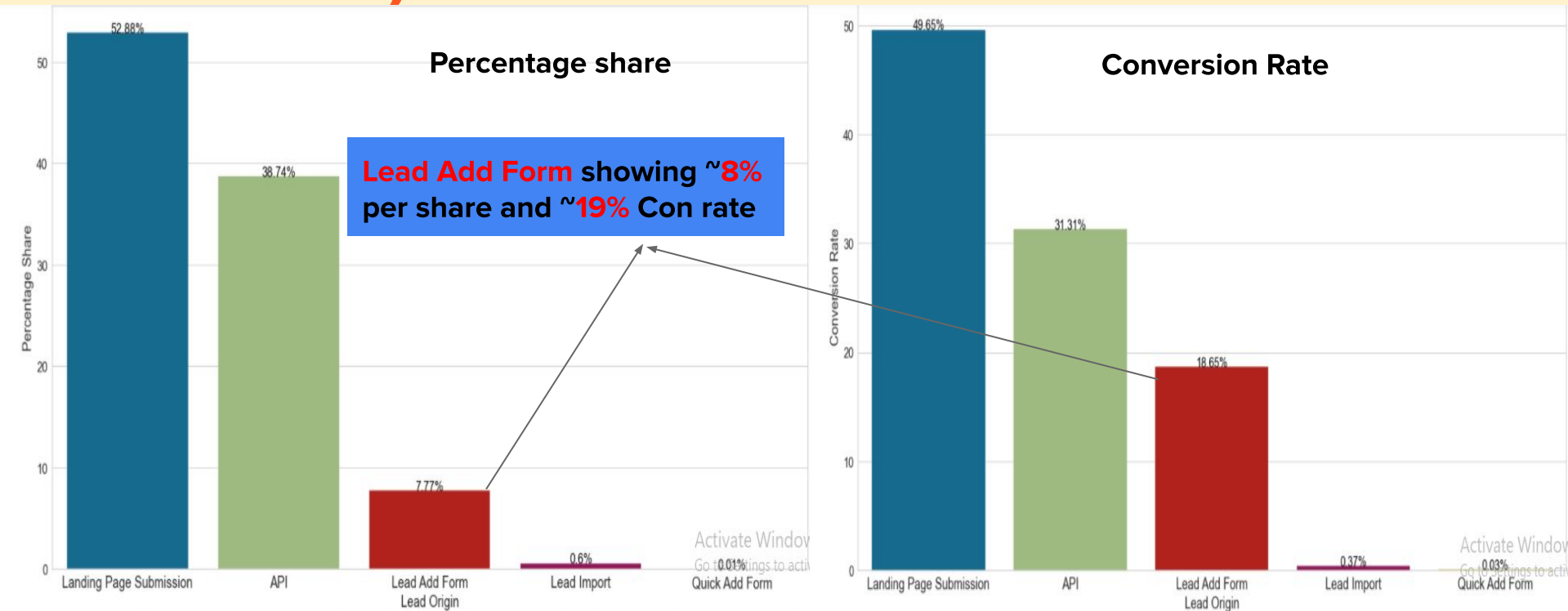
Data is slightly imbalanced with **~40%** of leads converted

Conversion rate(For each category)
= (Total sum of conversions from the category)/(Overall sum of conversions)

Metric: Lead Origin

Definition: The origin identifier with which the customer was identified to be a lead.

Lead Origin(Max. Conversion for Landing Page Submission)

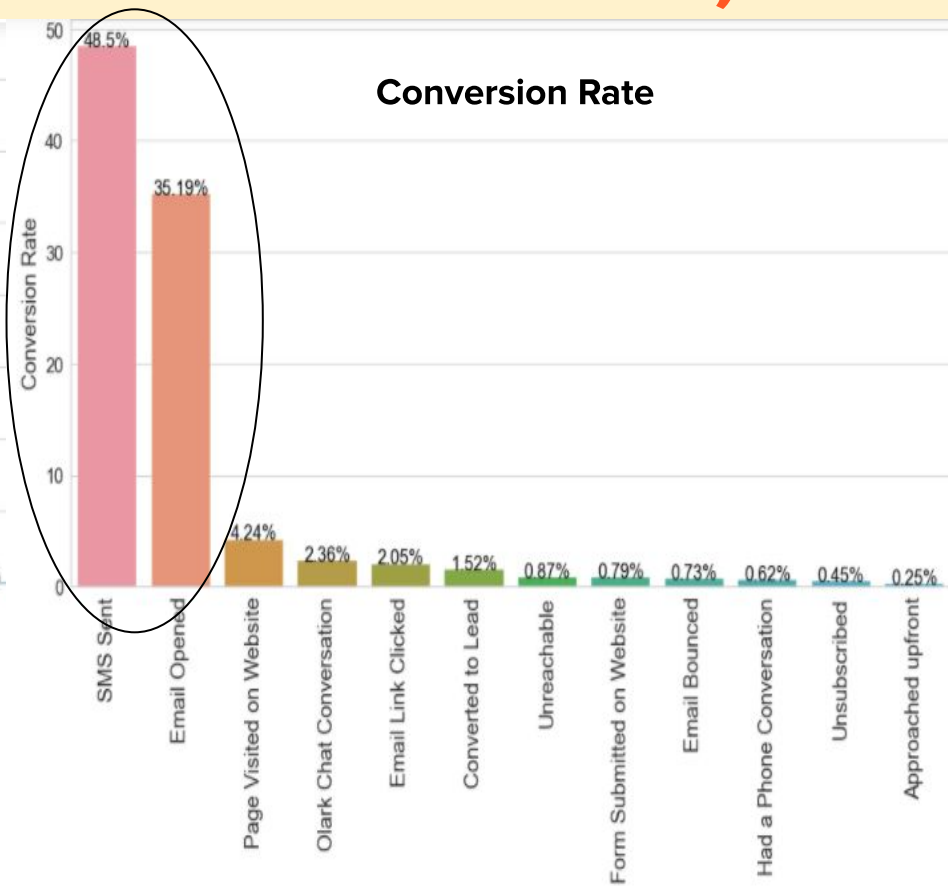
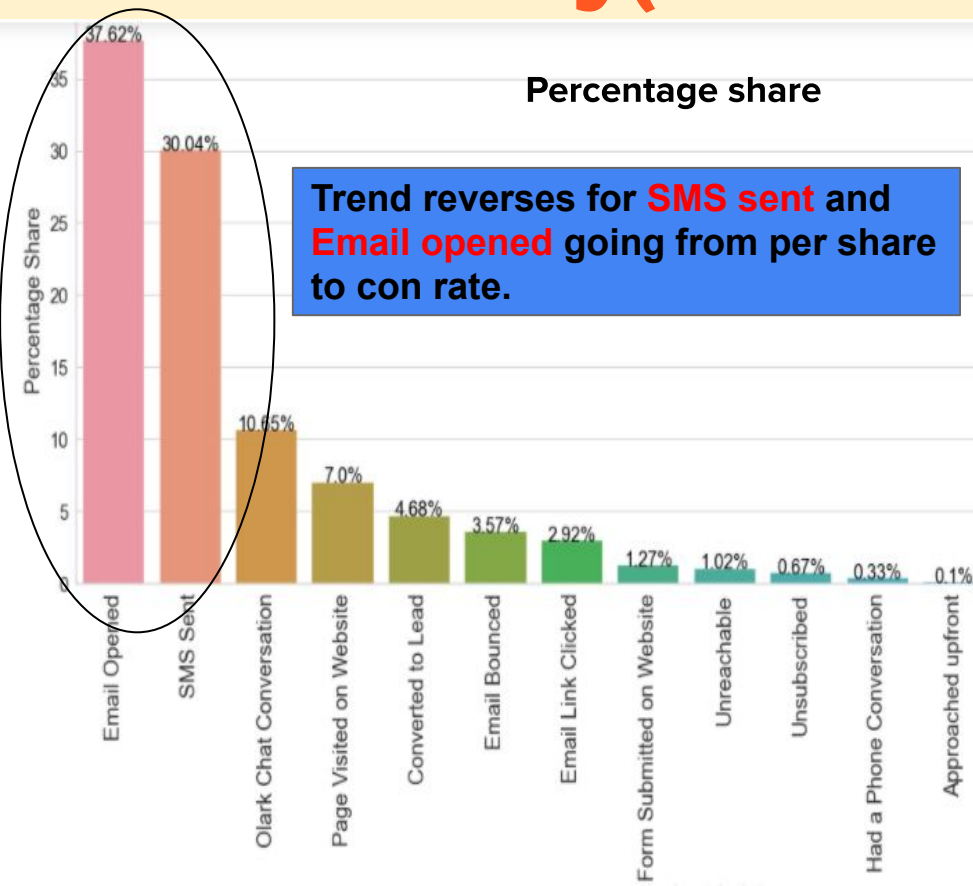


Recommendation: Increase Lead Add Form percentage share in order to get more conversion.

Metric: Last Activity

**Definition: Last activity
performed by the
customer.**

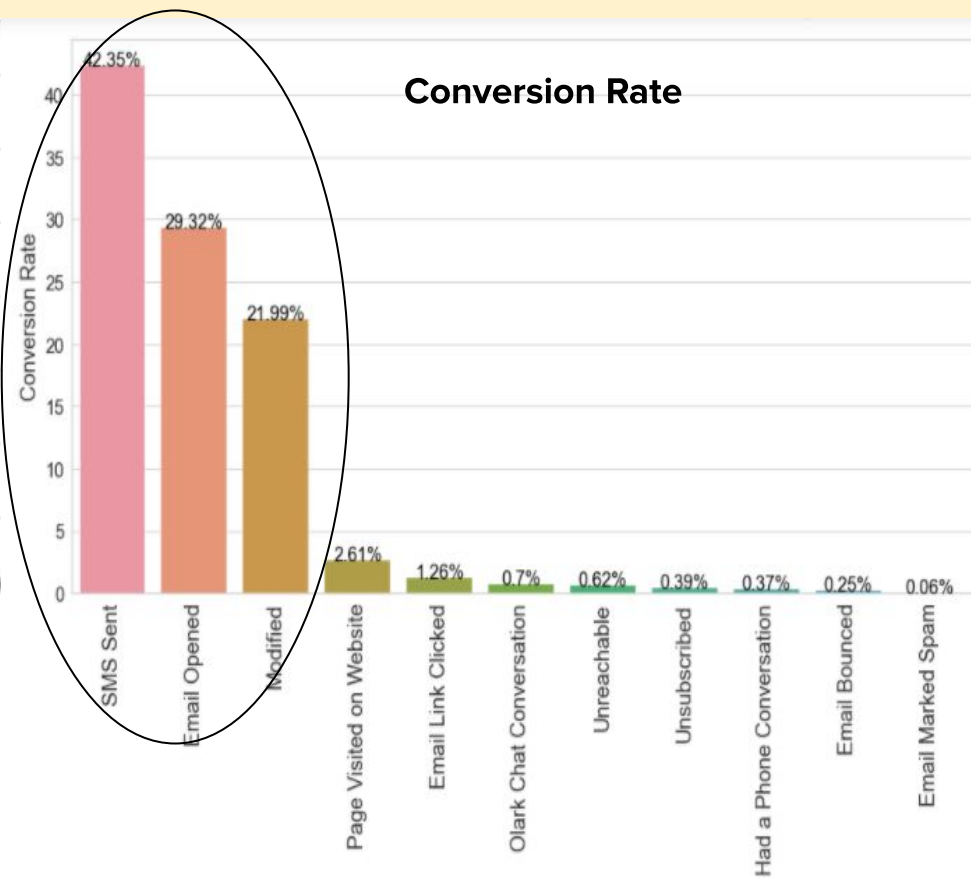
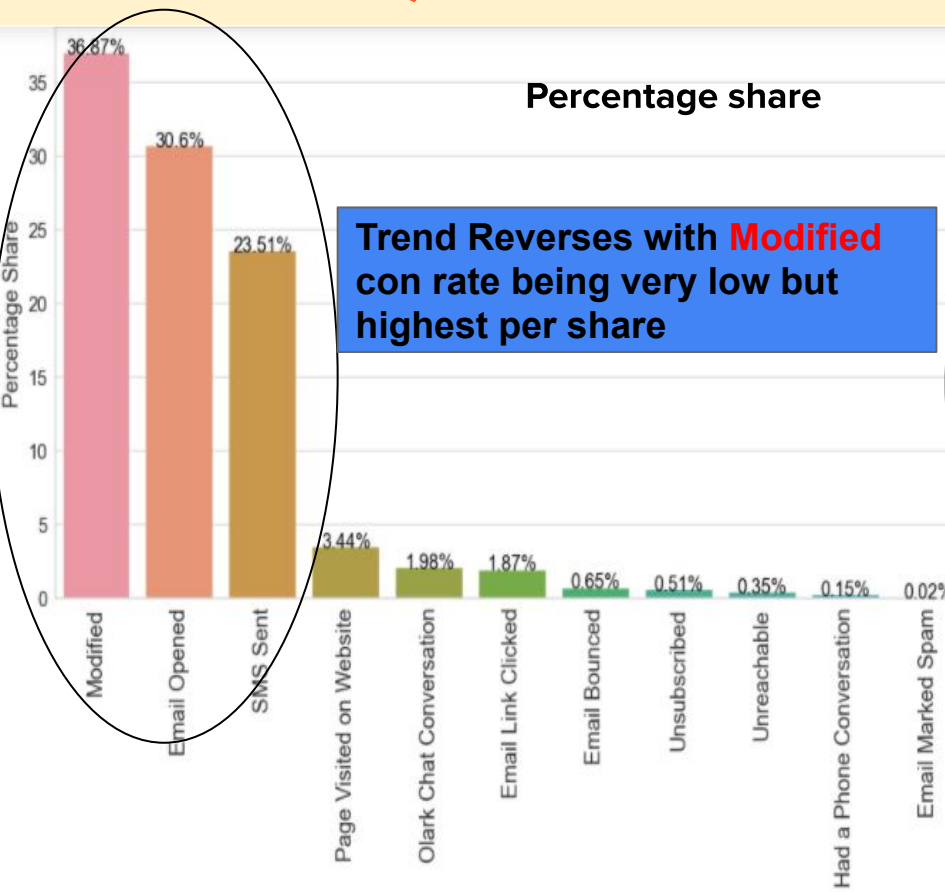
Last Activity(Max. Conversion for SMS sent)



**Metric: Last Notable
Activity**

**Definition: The last notable
activity performed by the
student.**

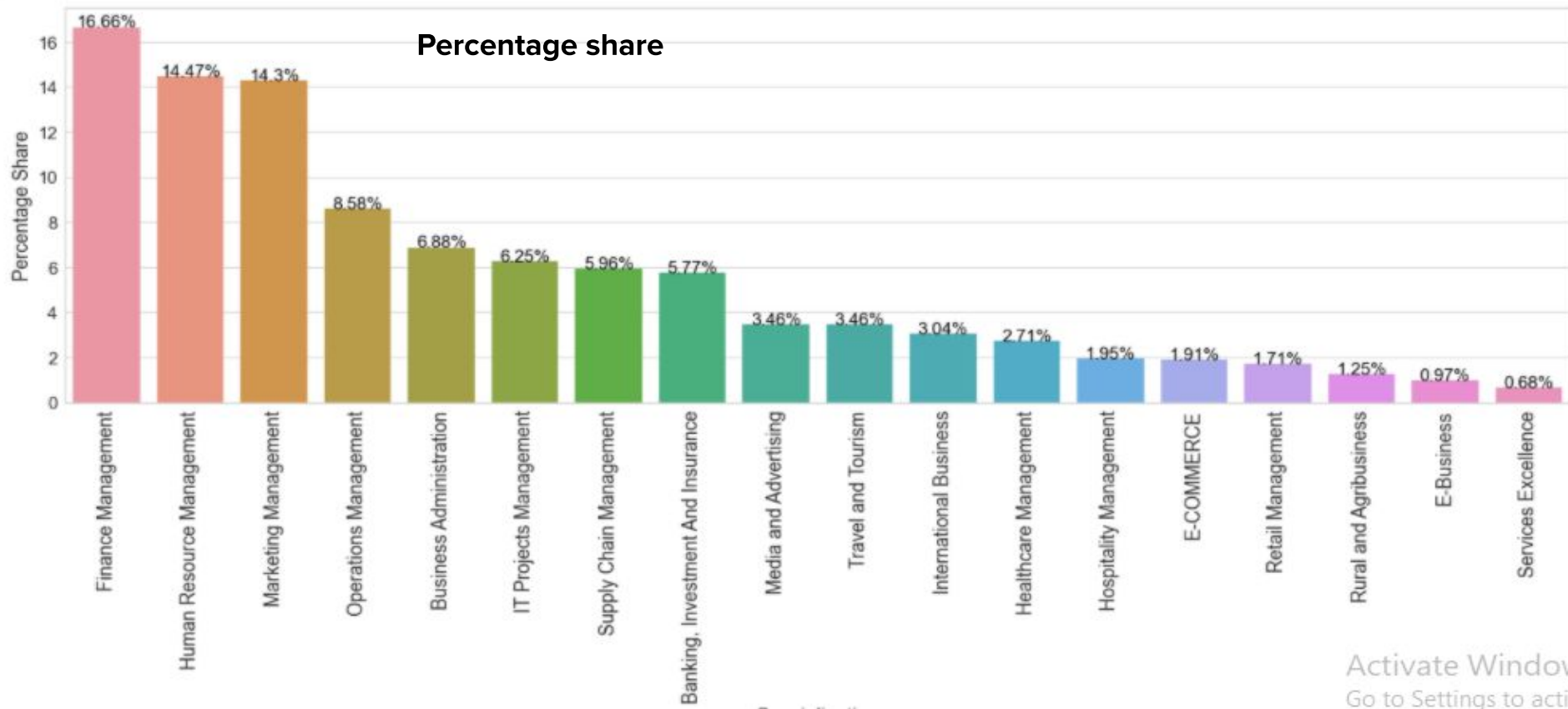
Last Notable Activity(Max. Conversion for SMS sent)



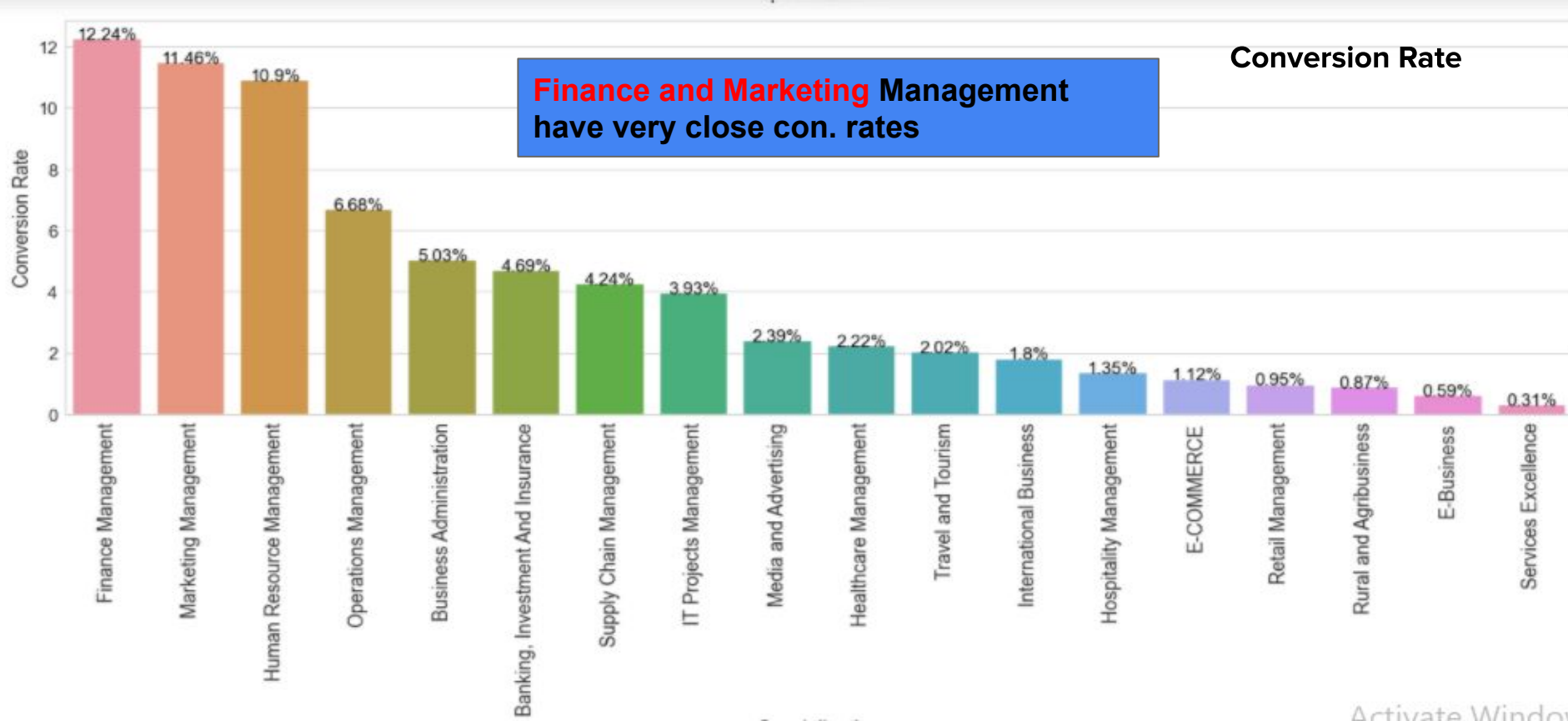
Metric: Specialisation

Definition: The industry domain in which the customer worked before.

Specialisation(Percentage Share)



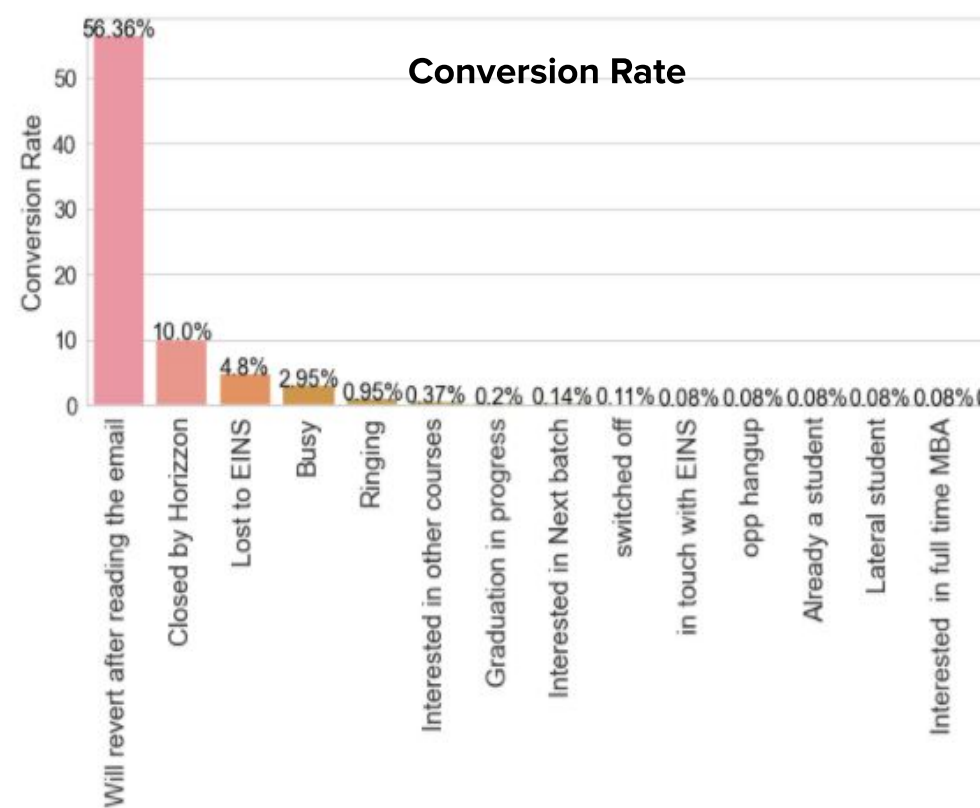
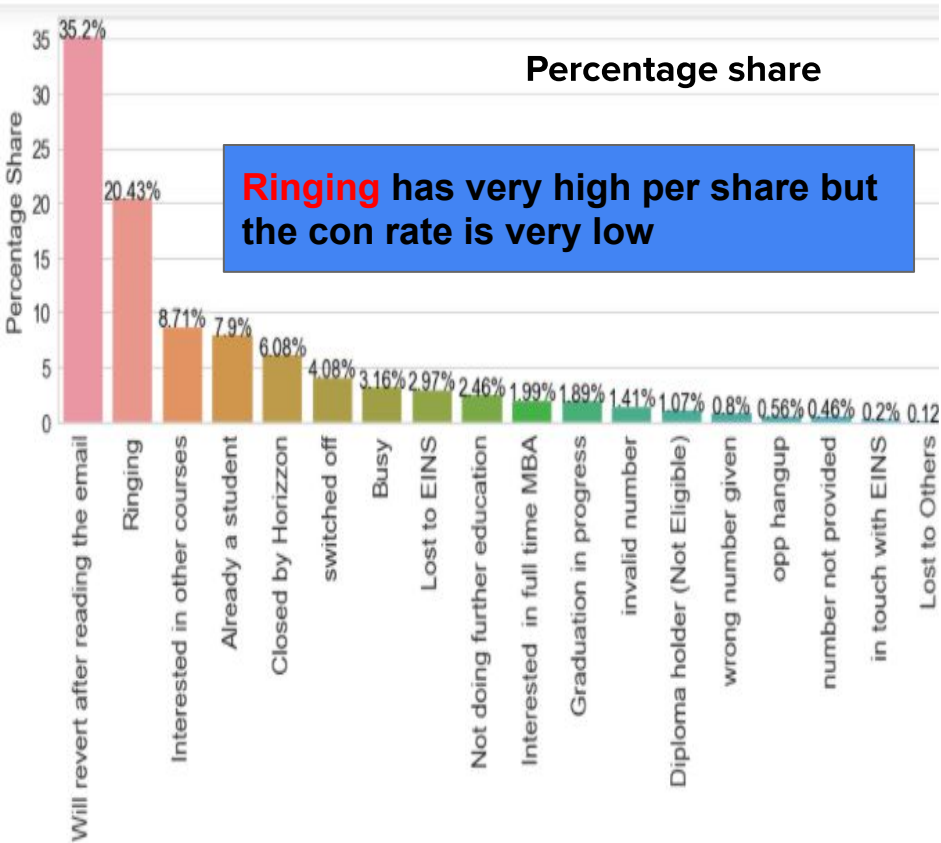
Specialisation (Conversion Rate Highest for Finance Management)



Metric: Tags

Definition: Tags assigned to customers indicating the current status of the lead.

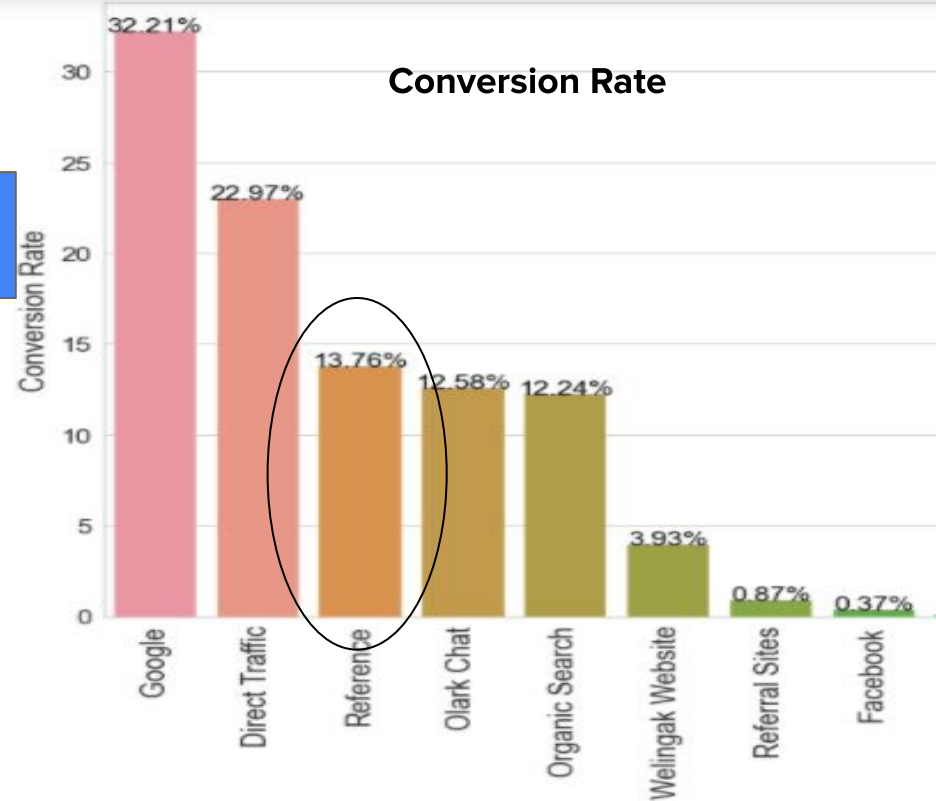
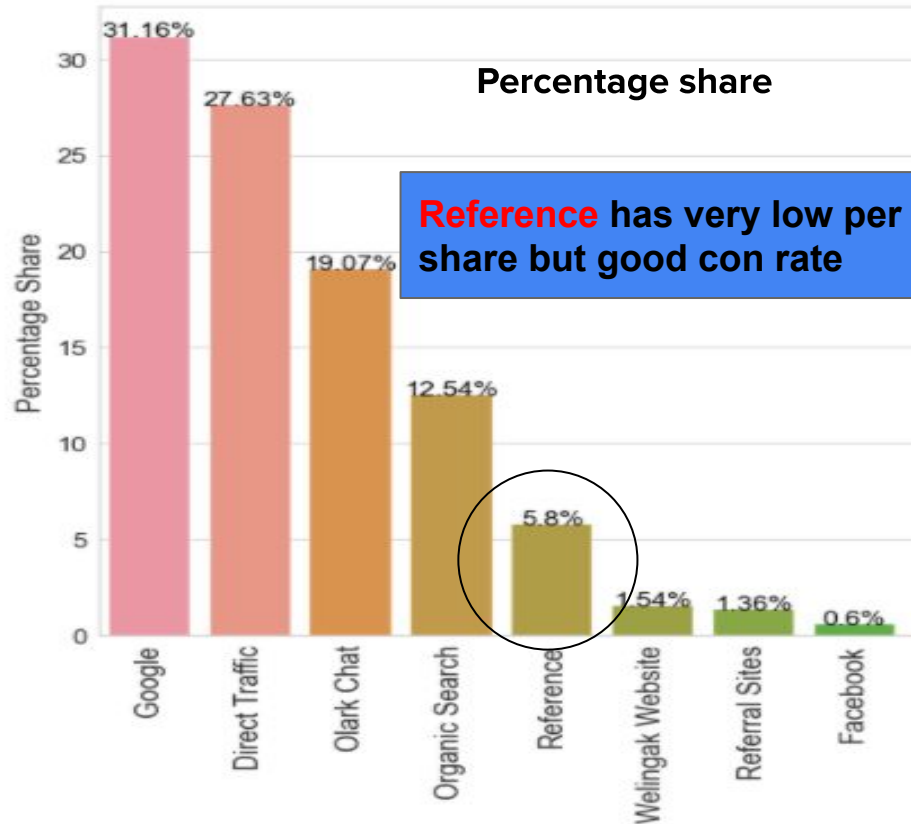
Tags(Highest Con. Rate for Will revert after reading the mail)



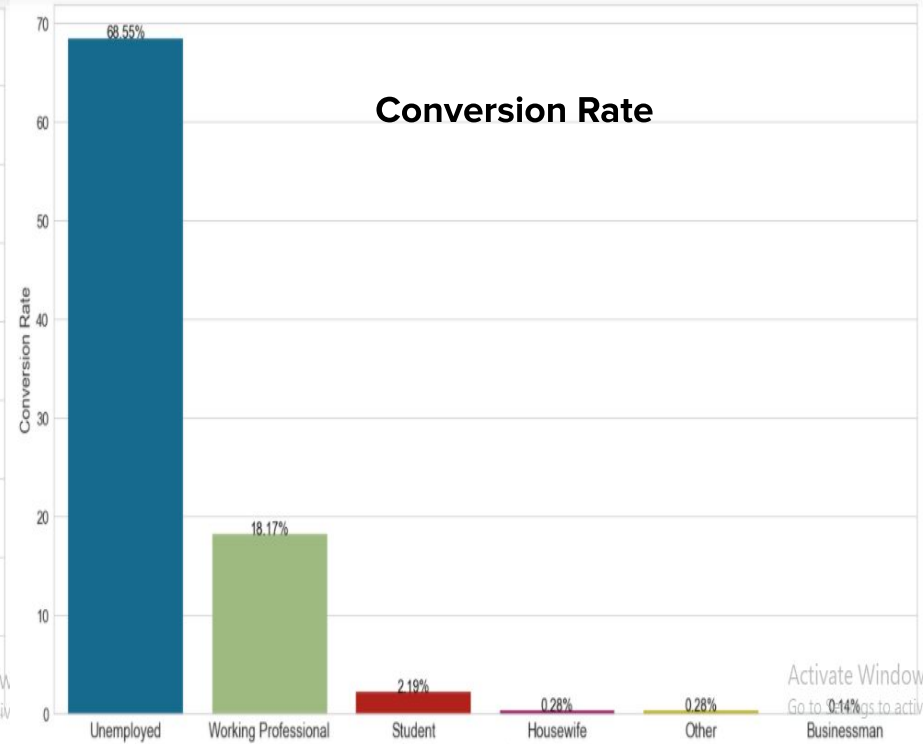
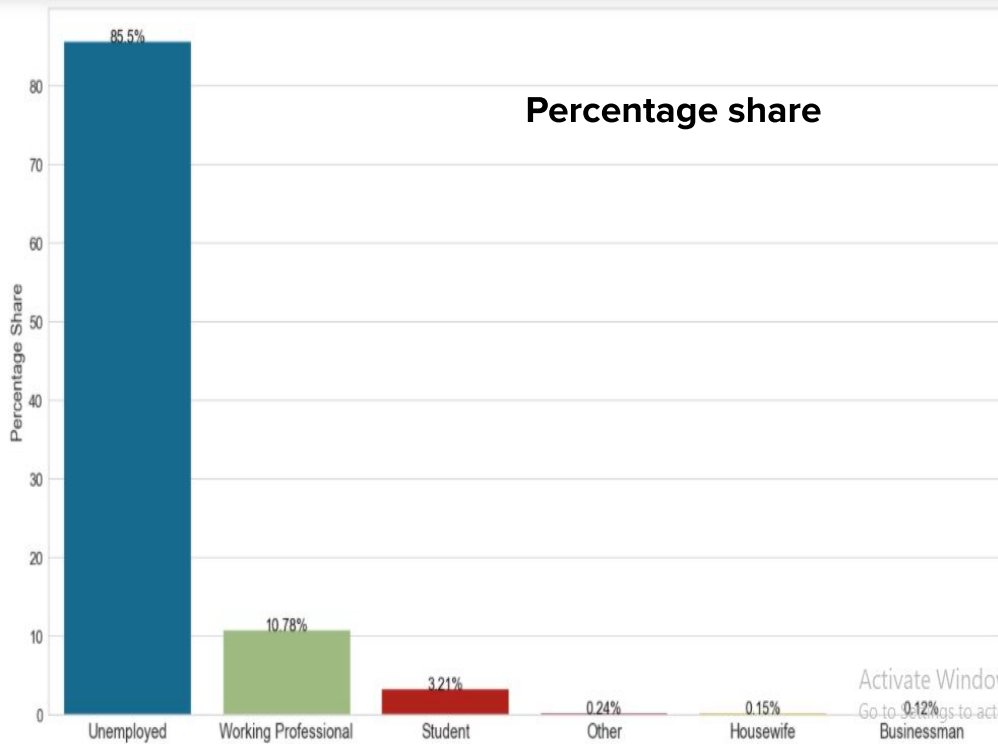
Metric: Lead Source

Definition: The source of the lead.

Lead Source (Google has highest Conversion rate)



What is your current occupation?(Max. Conversion observed for Unemployed)

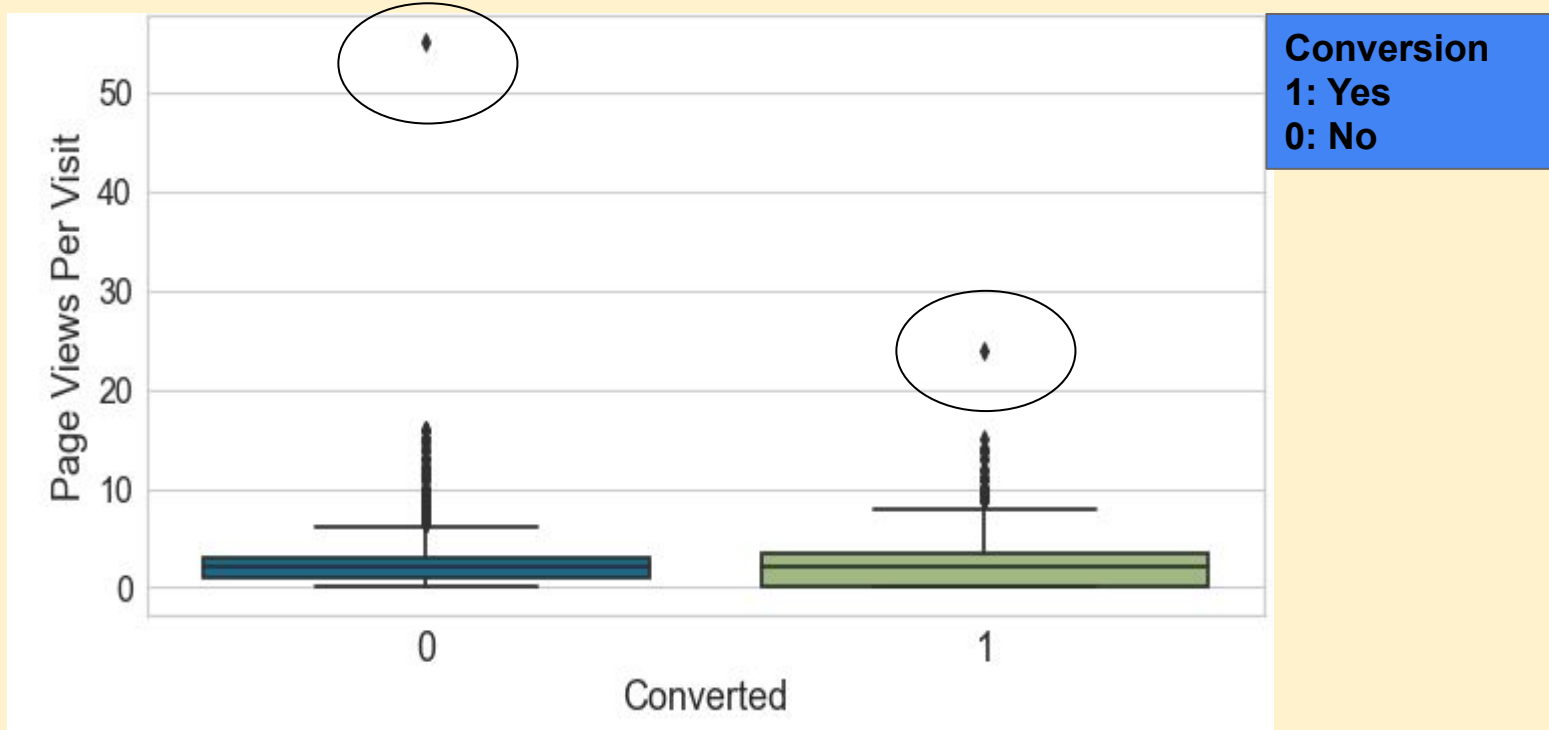


Even though X education is for **working professionals**. Max. users who are converting are **Unemployed**

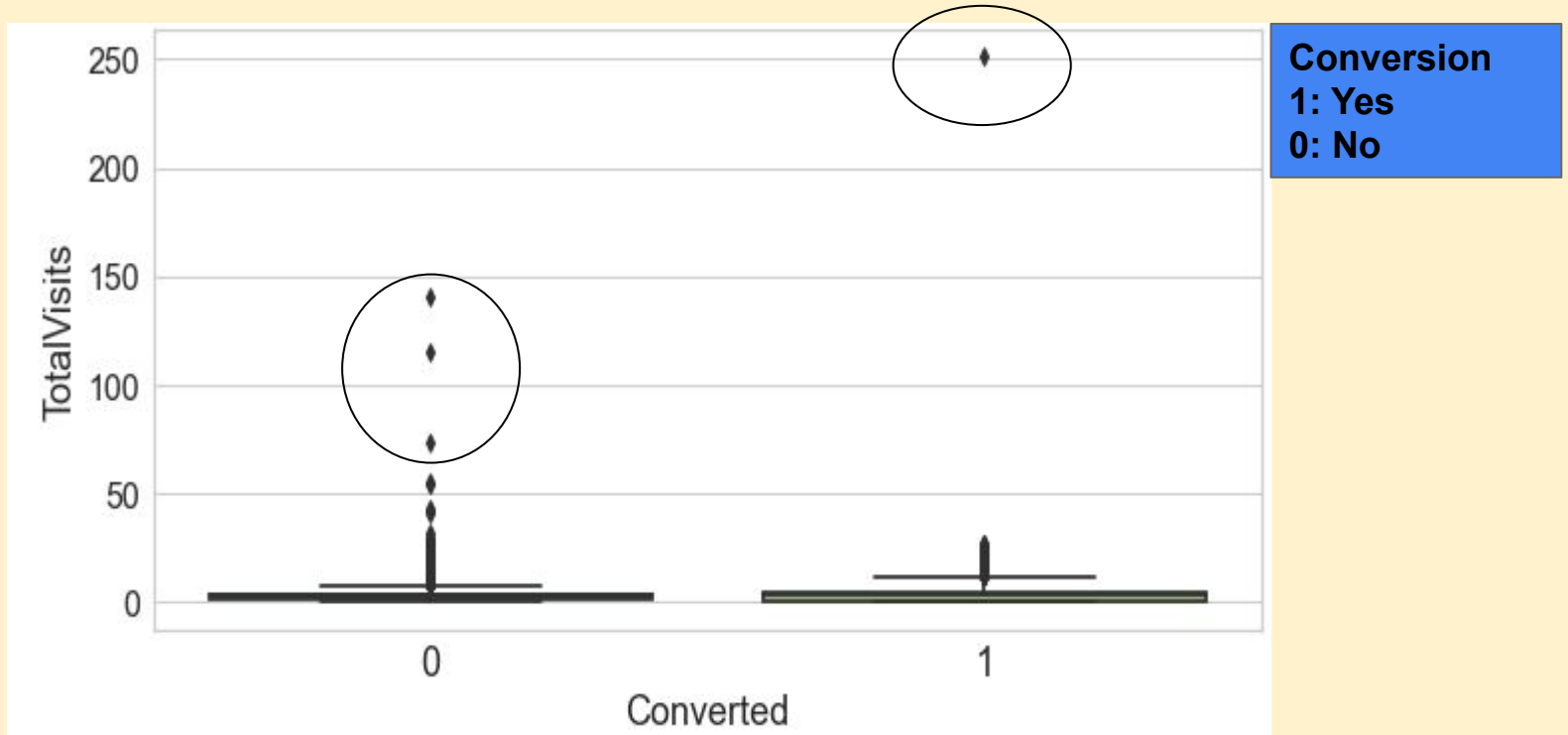
A free copy of Mastering The Interview



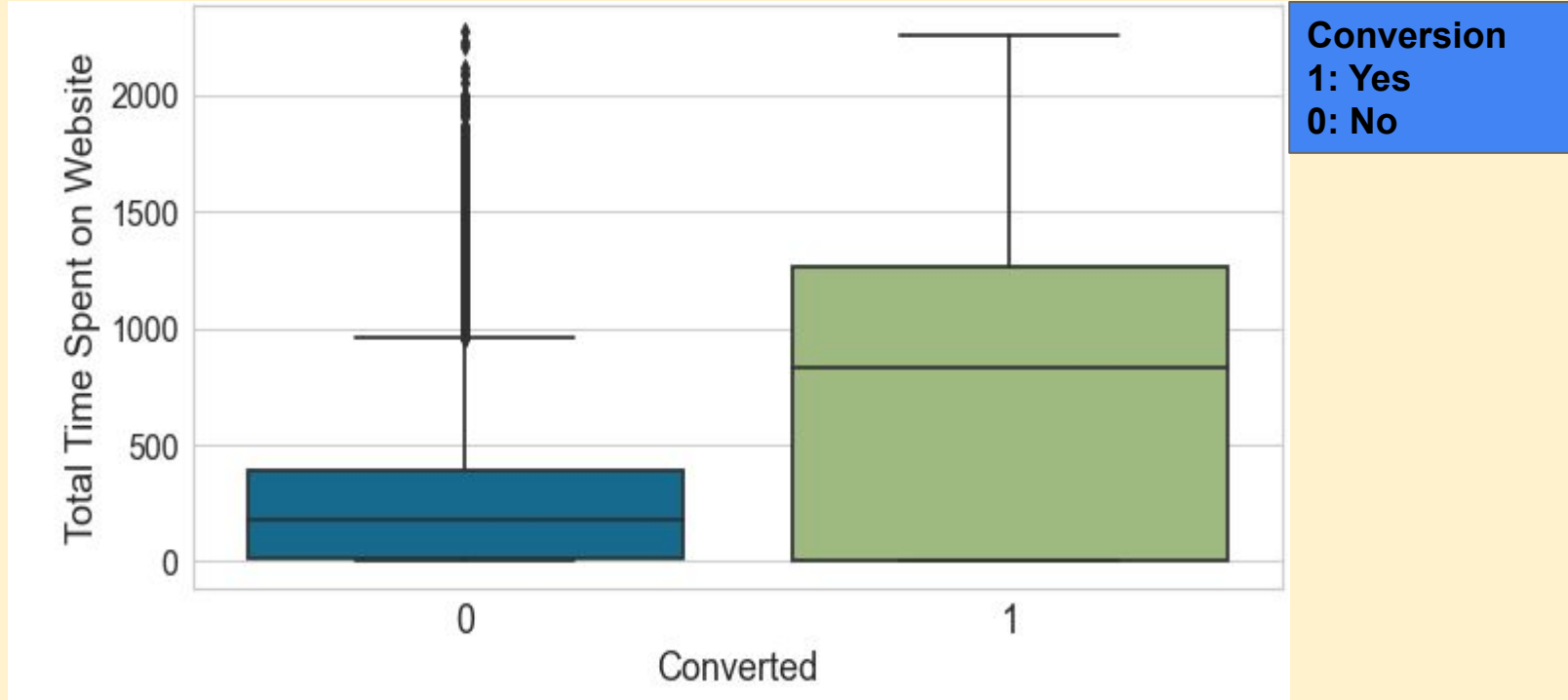
Page Views Per Visit (Distribution with Outliers)



Total Visits(Distribution with Outliers)



Total Time Spent on Website (Distribution with Outliers)



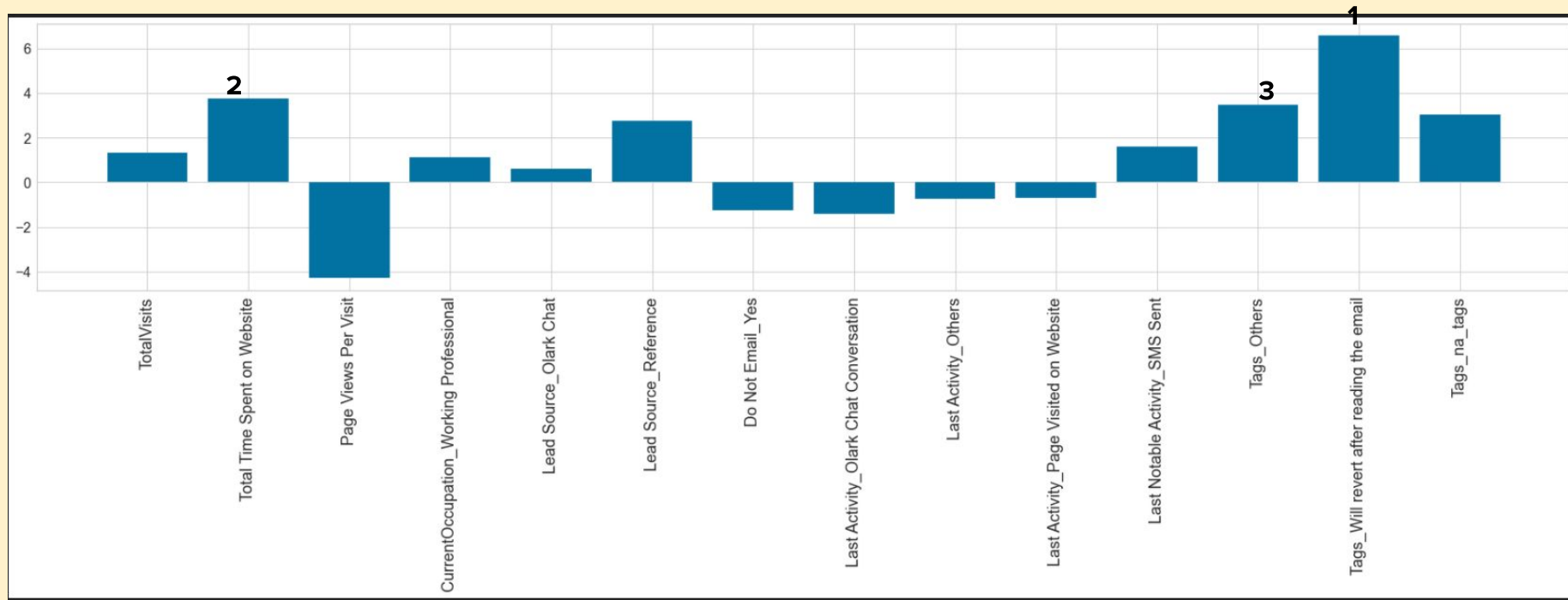
Model Building

Logistic Regression

Final set of features used for Model Building

| | features | importances |
|----|--|-------------|
| 0 | TotalVisits | 1.336255 |
| 1 | Total Time Spent on Website | 3.762874 |
| 2 | Page Views Per Visit | -4.301955 |
| 3 | CurrentOccupation_Working Professional | 1.114571 |
| 4 | Lead Source_Olark Chat | 0.619641 |
| 5 | Lead Source_Reference | 2.761631 |
| 6 | Do Not Email_Yes | -1.258111 |
| 7 | Last Activity_Olark Chat Conversation | -1.415590 |
| 8 | Last Activity_Others | -0.734594 |
| 9 | Last Activity_Page Visited on Website | -0.691319 |
| 10 | Last Notable Activity_SMS Sent | 1.592412 |
| 11 | Tags_Others | 3.475668 |
| 12 | Tags_Will revert after reading the email | 6.569813 |
| 13 | Tags_na_tags | 3.028872 |

Feature Importance (Top 3 Features)



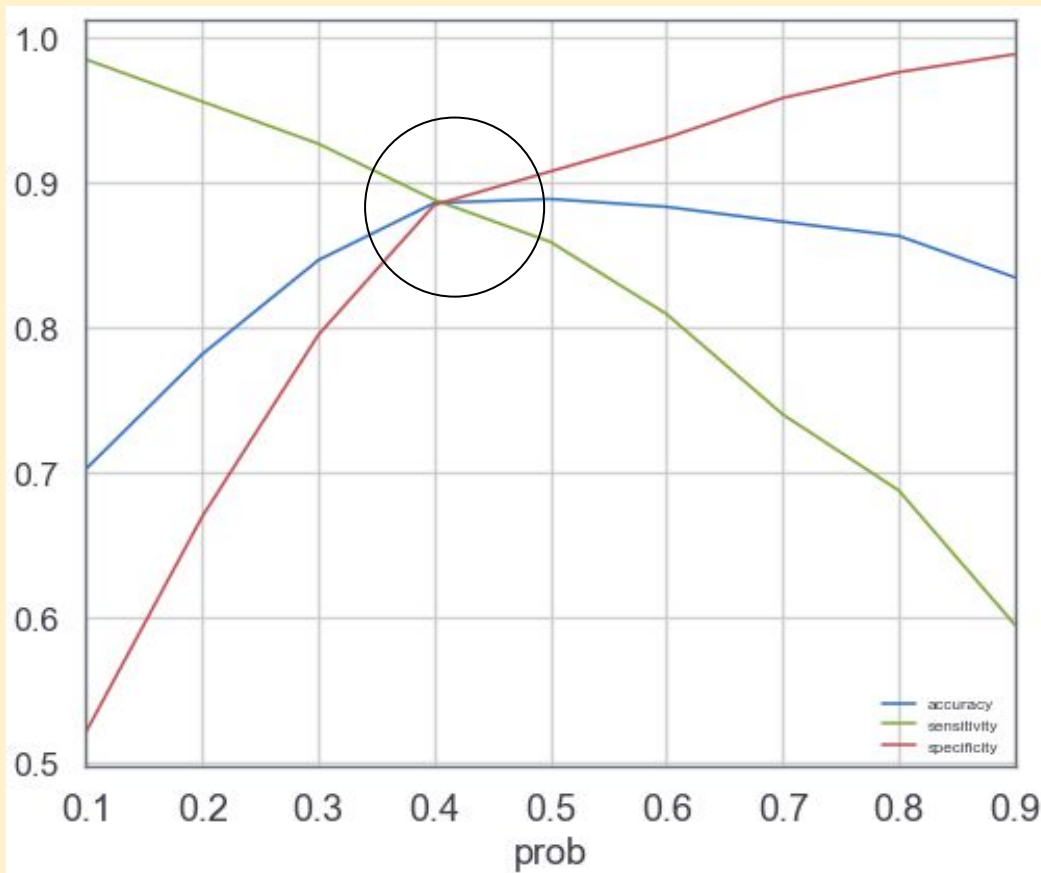
40

Threshold Score for all the leads

>40 - High Chance of Lead Conversion

<40 - Lower Chance of Lead Conversion

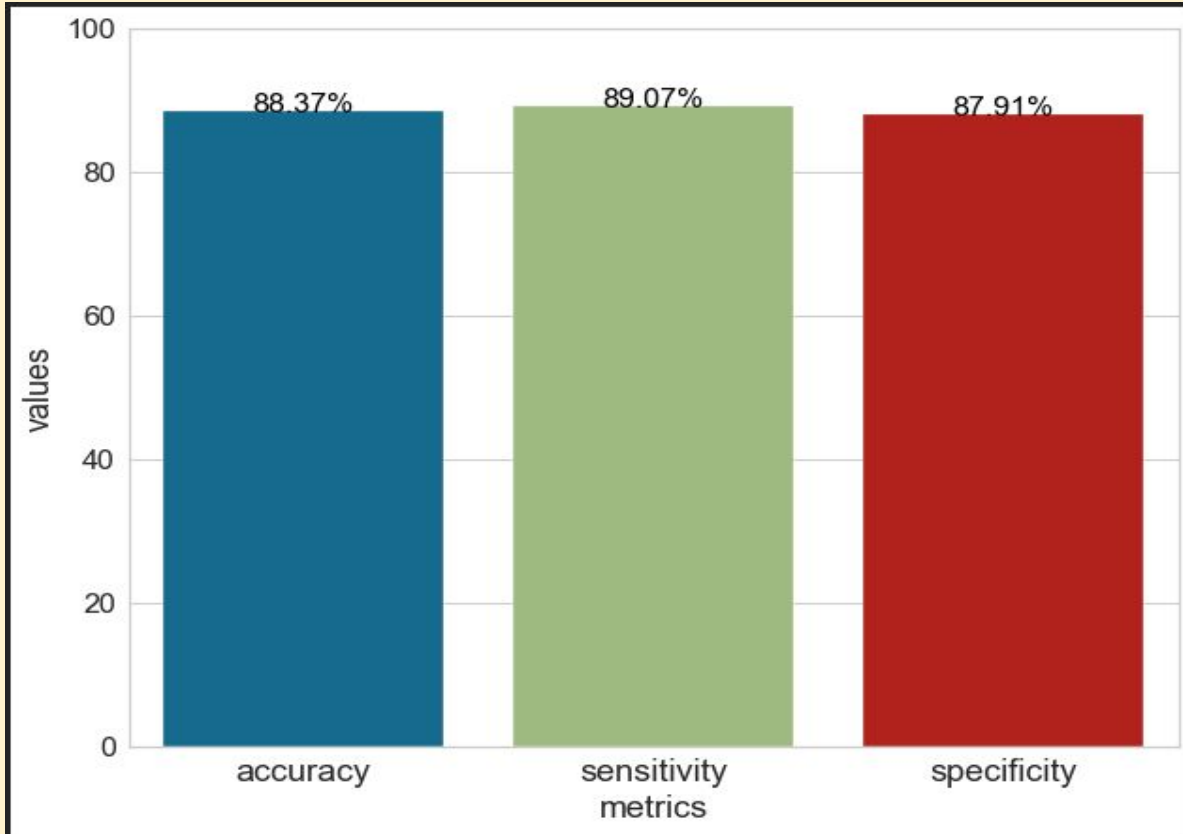
Threshold Probability(0.4)



| | | | |
|--------|---|------------------|------------------|
| Actual | 0 | TN 989 | FP 136 |
| | 1 | FN 79 | TP 644 |
| | | 0 | 1 |
| | | Predicted | |

Confusion Matrix

**We are able to capture 89% of the
Converted leads**



Area Under ROC curve is **0.95**

