Lead Scoring Case Study

Agenda

1. Problem Statement
2. Data Cleaning & Sanity Check
3. Exploratory Data Analysis(EDA)
4. Data Pre-Processing, Train-Test split & Feature scaling
5. Model Building & Model Evaluation
6. Conclusions & Recommendations

Problem Statement:

1. X Education sells online courses to industry professionals.
2. The company markets its courses on several websites and search engines like Google. 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.
3. 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.

Business Objective:

1. X Education wants to know what factors will help them to know most promising leads.
2. The CEO, in particular, has given a ballpark of the target lead conversion rate to. be around 80%.

Data Cleaning & Sanity Check:

1. Around 16 features were having missing values, we have removed features having missing value % more than 40%. For the remaining features, we have deep dive-in and categorized similar and missing values into ‘Others’ category.
2. We have dropped highly skewed and imbalanced features like 'Search','Do Not Call','Magazine','Newspaper Article','X Education Forums’ etc.
3. Few numerical columns like “TotalVisits” & “Page Per Views” had outliers, we have used the method of Soft-Capping to handle outliers where we remove the values that are >Q3 & <Q1.

Exploratory Data Analysis:

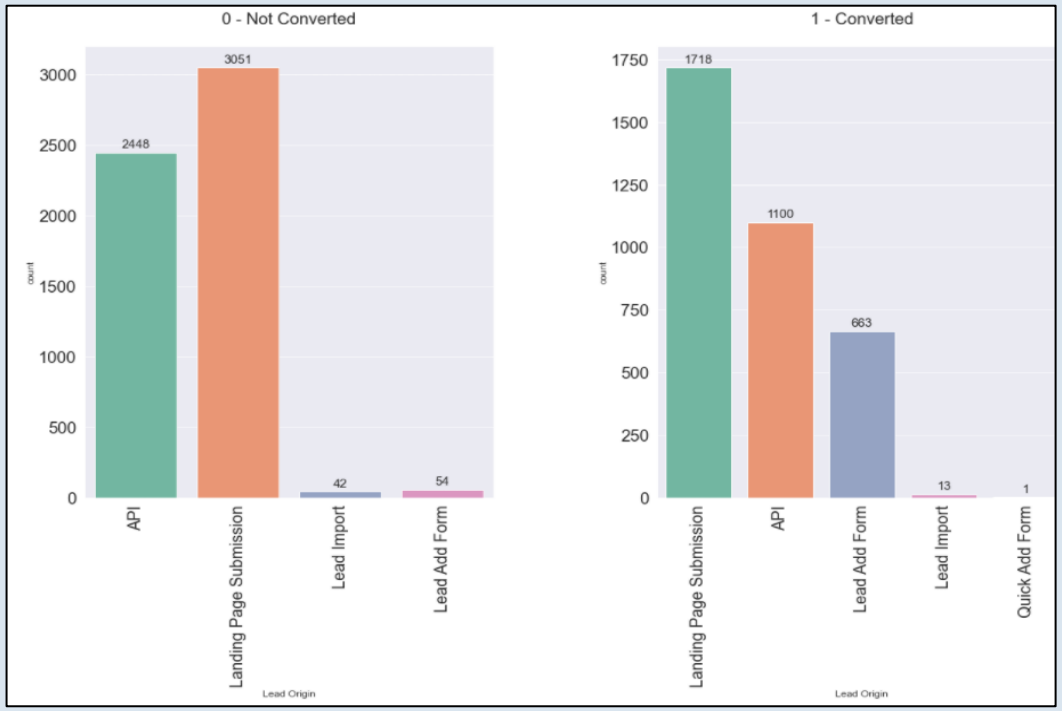
1. Conducted Univariate Analysis of Categorical columns :Lead Origin', 'Lead Source', 'Last Activity', 'Specialization', 'What is your current occupation’ etc. Using count plots.
2. Conducted Univariate Analysis of Numerical columns : 'Converted','TotalVisits','Total Time Spent on Website’, 'Page Views Per Visit’ etc. Using Distribution plots.

Bi-variate Segmented Analysis:

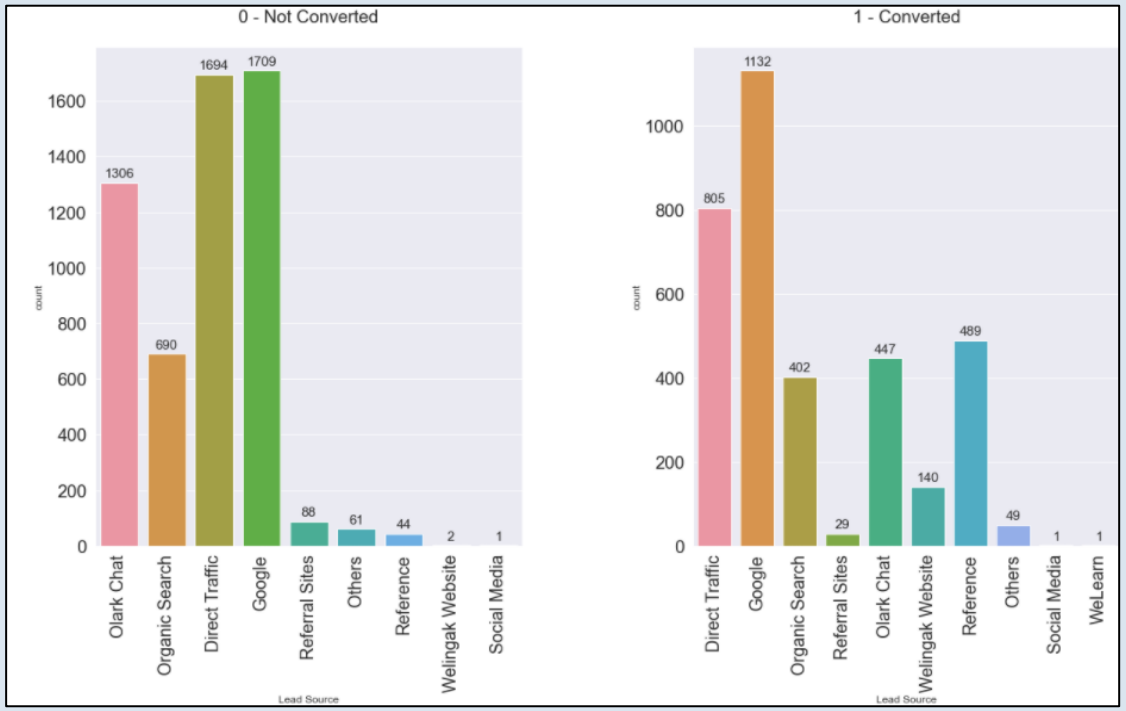
1. The entire data was divided into 2 parts, Set-0 had data where converted=0, Set-1 had data where converted=1.
2. Segmented analysis was conducted on above 2 datasets using count plots with data labels.

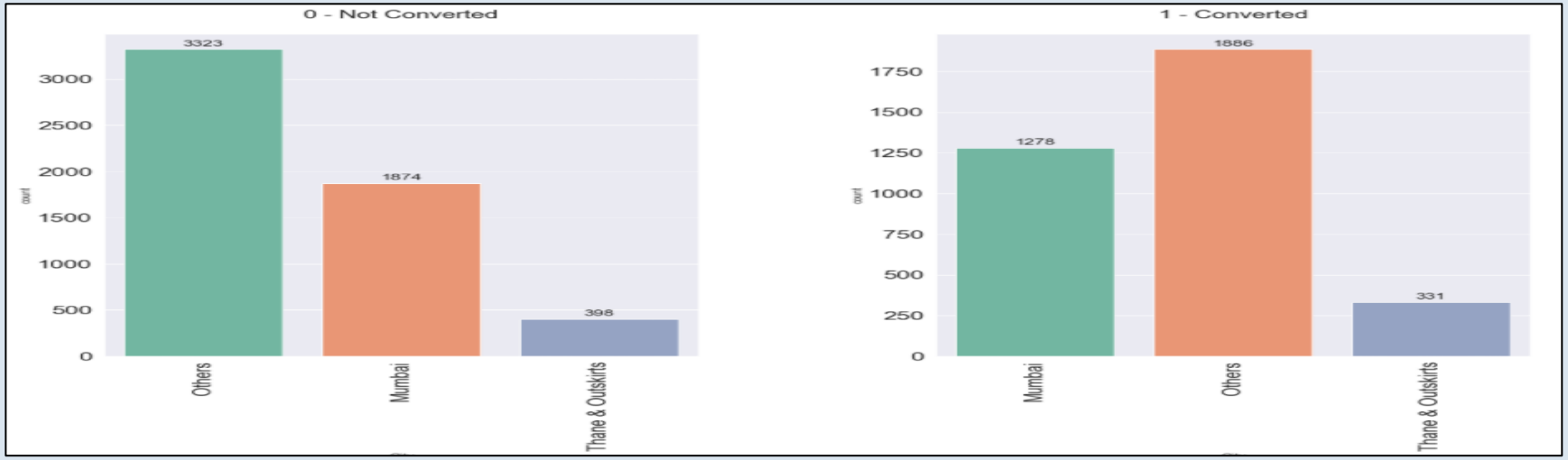
Bi-variate Segmented Analysis:

Lead Origin



Lead Source

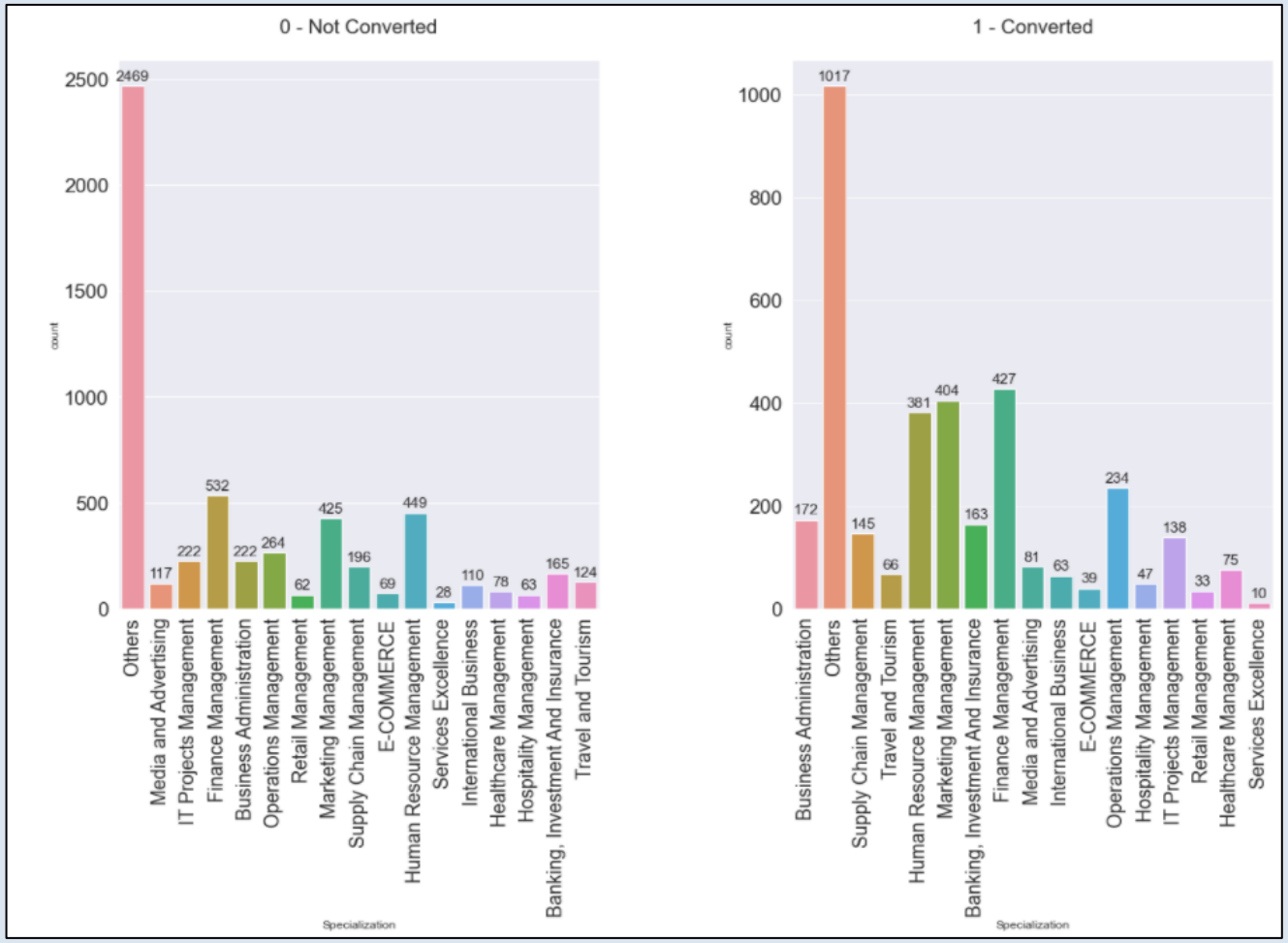




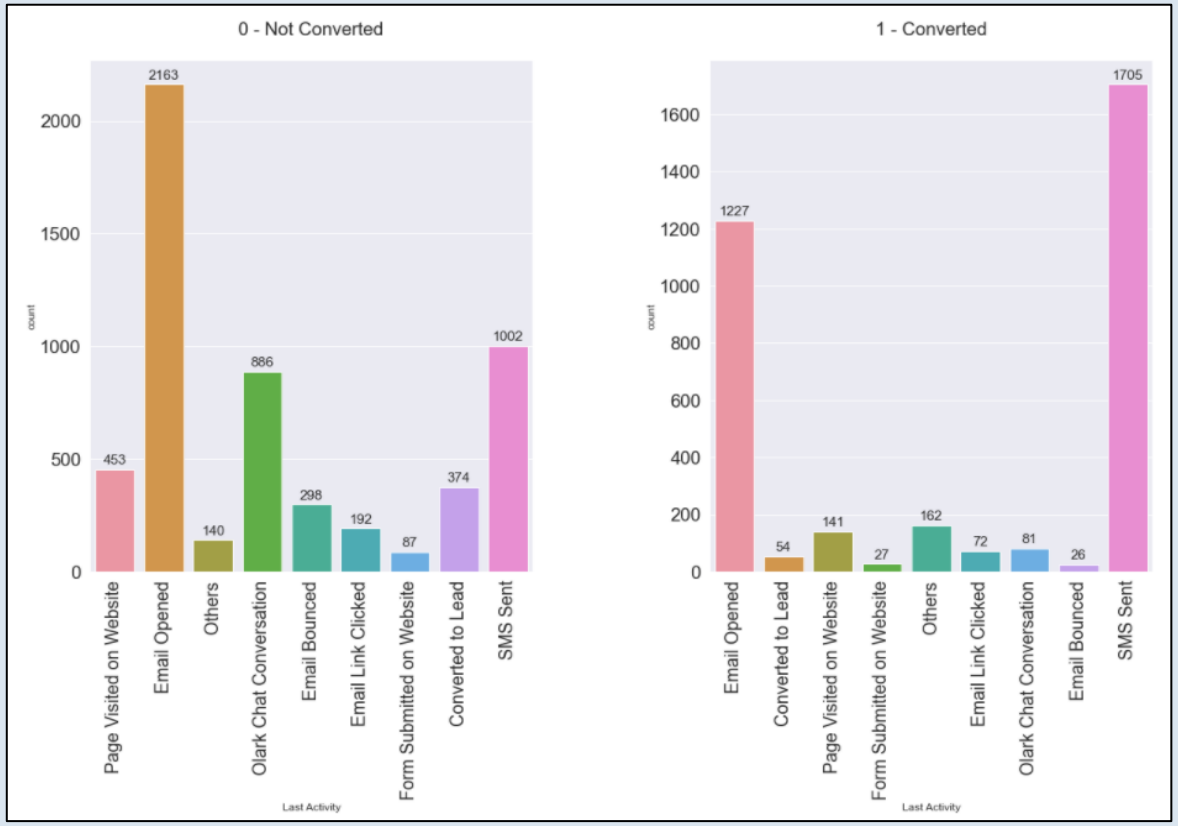
City

Bi-variate Segmented Analysis:

Specialization

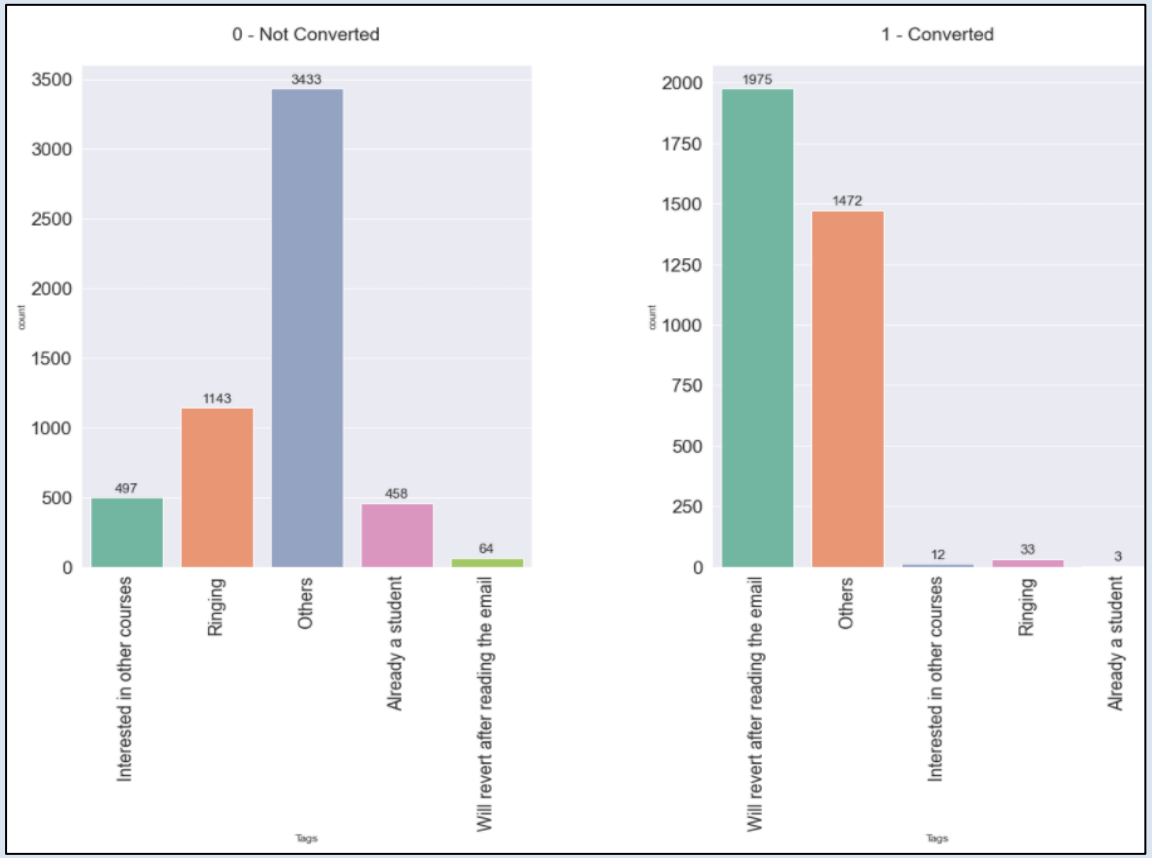


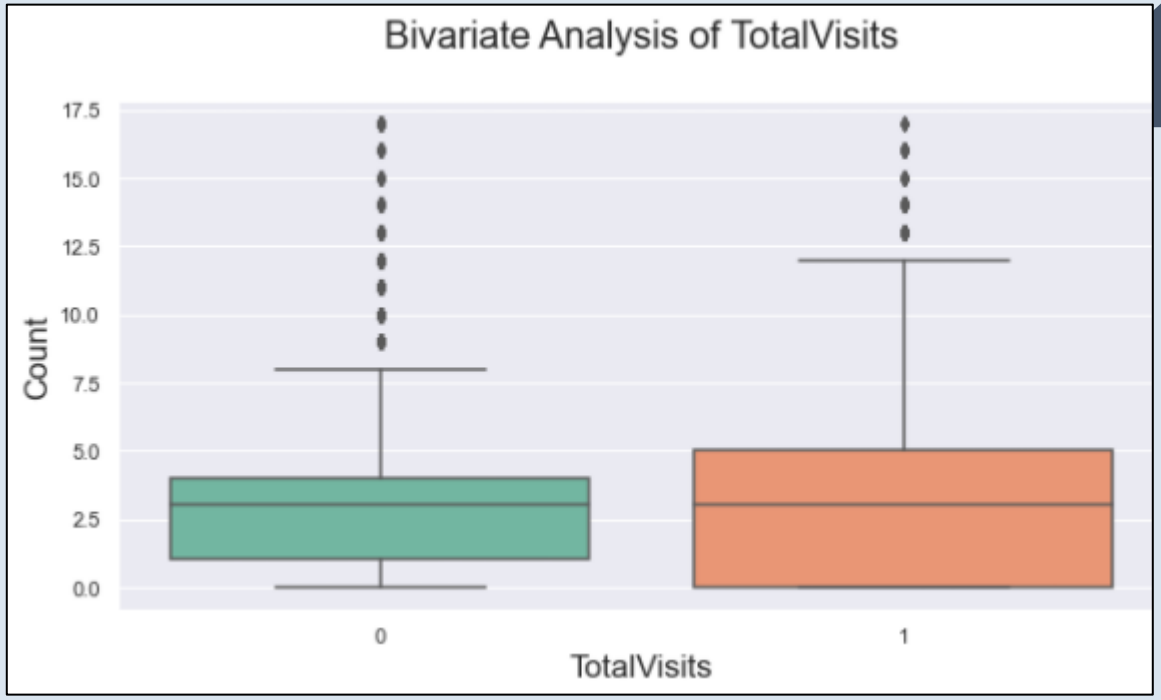
Last Activity

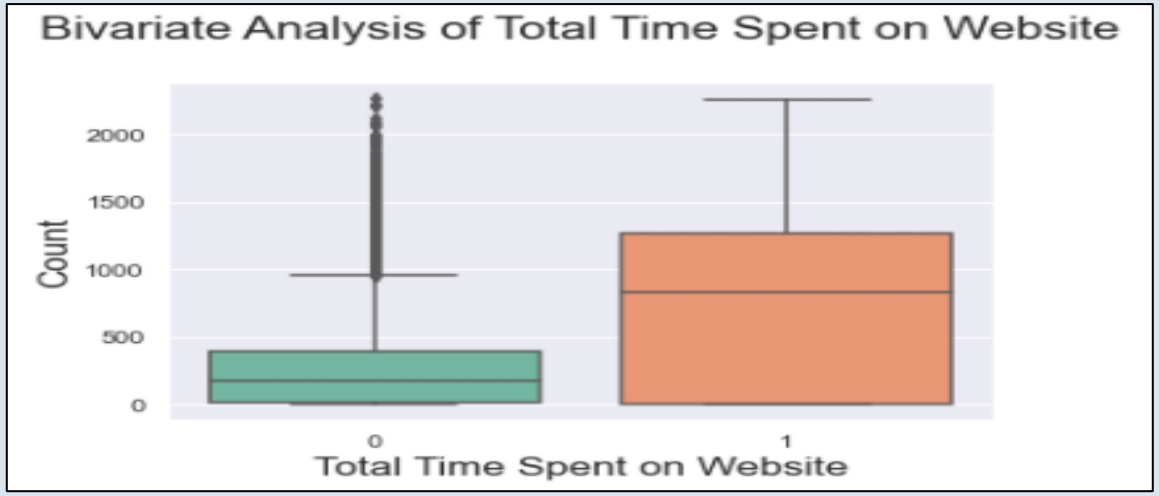


Bi-variate Segmented Analysis:

Tags







Insights From EDA:

1. "Landing Page Submission" has highest chances of getting converted compared to APIs & Lead Add Form.

2. "Google" & "Direct Traffic" leads must be given more emphasis.

3. Chances are that if the lead has opened the email & not responded then he/she may not be converted. But the company should focus on sending "SMS" as it has high chances of lead conversion.

4. Leads who are from "Financial Management, HR Management & Marketing Management" specializations will be more likely to react positively towards the sale.

5. The leads with status "Will revert after reading the email" have more conversion rate comparatively.

6. "TotalVisits" has almost same median for both converted and non-converted leads, although leads who made more than 5 visits at least are potential conversions.

7. It's clear that the greater the "Time spent on website" the more leads are interested and more the conversion.

Data Pre-Processing, Train-Test split & Feature scaling

1. We have converted some categorical columns having Yes/No values to 1/0, like 'Do Not Email', 'A free copy of Mastering The Interview' features.
2. We have created dummy variables & dropped original variables, Variables with suffix 'Others'. Finally we have 44 columns & 9090 rows for model building.
3. We used 'train_test_split()' method of sklearn to split data into train(70%) & test(30%) sets.
4. We have applied Standard-Scaling method for scaling the numeric features.

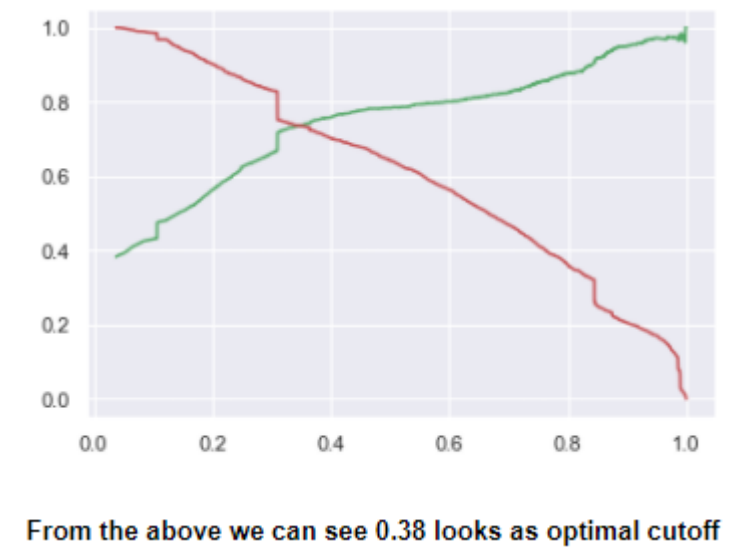
Model Building & Evaluation

1. We used RFE(Recursive Feature Elimination) technic to get the top 15 variables from 44 variables.
2. We created 6 models in an iterative process feature elimination based on P-value & VIF Scores evaluation to determine best fit model, Model6 was the best well-balanced model.
3. We have plotted ROC Curve:
   a) The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
   b) The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

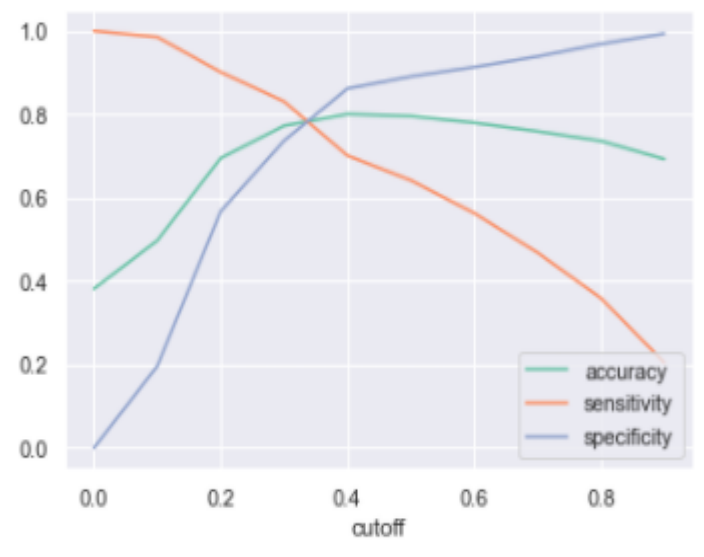
The ROC Curve looks good as area under the curve is 0.86

Model Building & Evaluation

4. We conducted Precision/Recall trade-off & Sensitivity/Specificity/Accuracy trade-off to determine the optimal cut-off and we got **0.38** as optimal cut-off in both.



Precision/Recall Trade-off

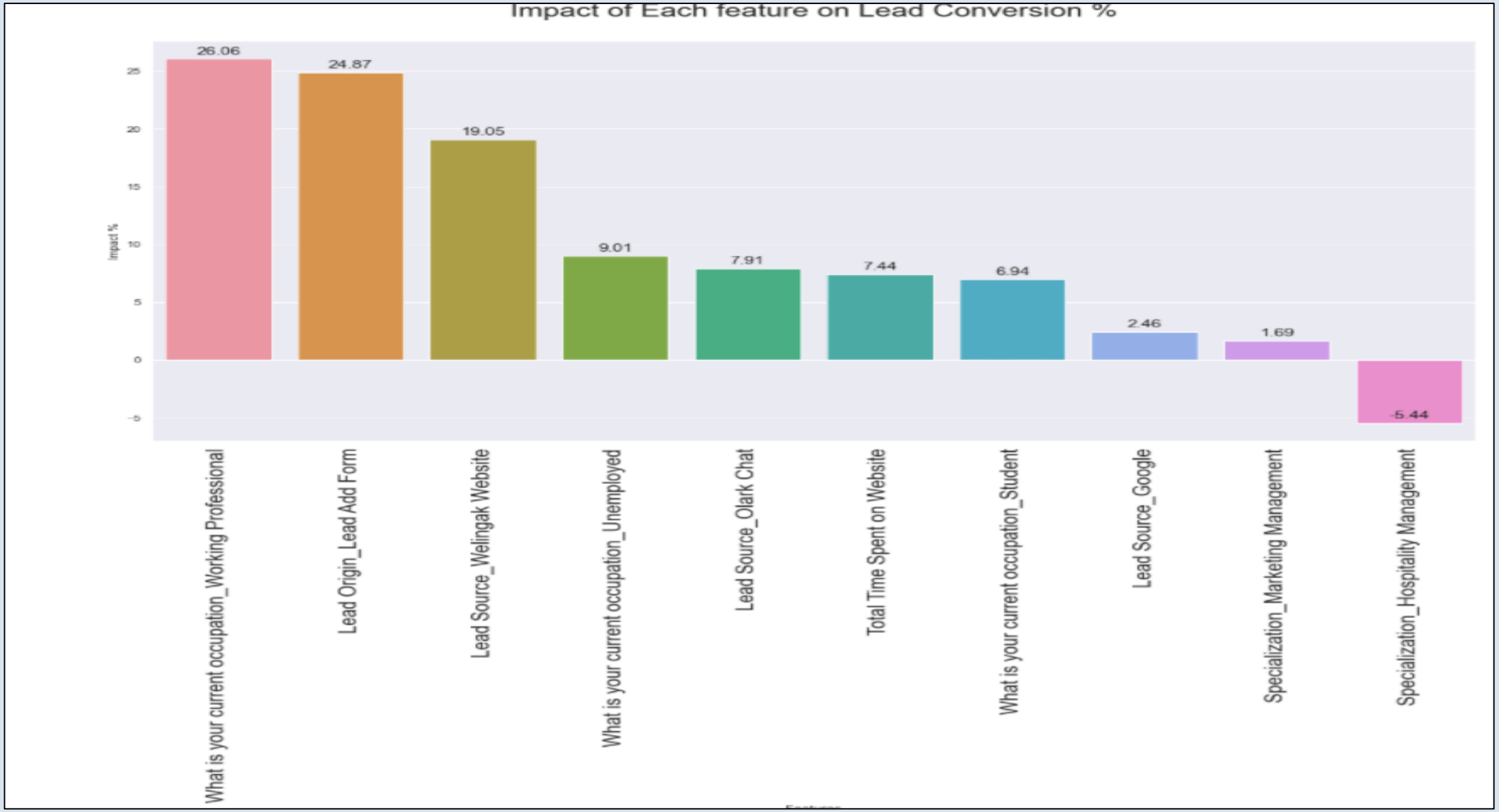


Sensitivity/Specificity/Accuracy Trade-off

5. The Model has Accuracy: 80%, Sensitivity:72%, Specificity:85%. Below is the classification report:

	precision	recall	f1-score	support
0	0.82	0.85	0.83	1660
1	0.75	0.70	0.73	1067
accuracy			0.79	2727
macro avg	0.78	0.78	0.78	2727
weighted avg	0.79	0.79	0.79	2727

Feature's Impact on Conversion Rate as Per Model.



Conclusion:

1. Leads from "**Working Professionals**" & "**Student**" occupation are the ones which will have higher lead conversion probability of **33%** Lead conversion chance compared to other occupations.
2. Leads that have filled the "**Add Form**" are the prospects that needs to be focused more comparatively for higher conversion rate - Approximately **25% Conversion rate.**
3. Leads coming from the background Specialization of "Hospitality Management" will negatively impact the lead conversion rate **(-5.4%)**. Although the leads from "Marketing Management" also has less chances of being converted into a "Hot" lead.
4. Considering the Leadsource we can infer that leads coming from Welingak Website & Olark Chat has higher chances of getting converted **(27% Chance)** compared to Google.
5. "**Time Spent on Website**" is positively contributing towards the lead conversion with **7.4%** Conversion rate.

Recommendations:

1. Leads coming from the source "**Welingak Website**", Filling the "**Add Form**" & belonging to "**Working Professional**" occupation , has "**spent more time on website**" & is coming from "**Marketing Management**" background have **86% Conversion rate** making them a "Hot" lead. That would help CEO to take unanimous decisions.
2. Leads coming from the source "**Google**" and are belonging to "**Hospitality Management"** shall be ignored or less prioritized as they have very low conversion chances (-3%).

Thank You