Lead Score Case Study

Problem Statement:

An X Education need help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires us to build a model wherein you need to assign a lead score 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%.

Goals and Objectives:

There are quite a few goals for this case study.

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. 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. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

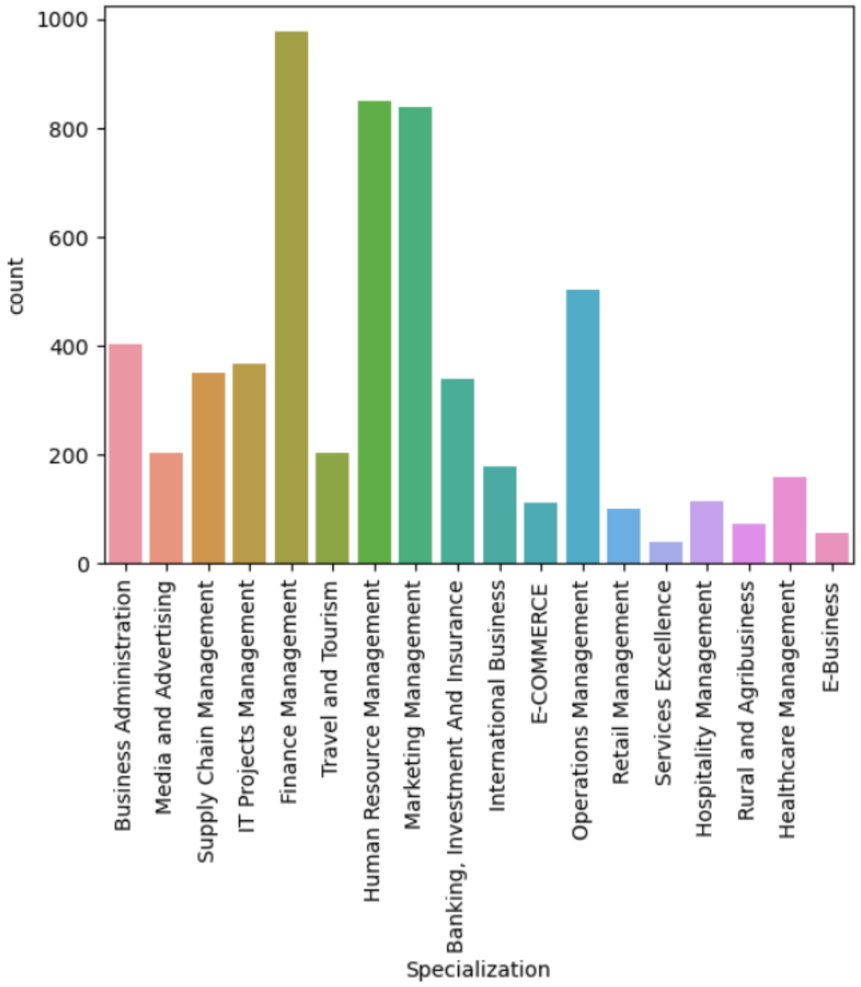
Data preparation or Data Cleaning

- Before making any changes to the original data frame let's take a copy of the same.
- Some of the columns contains values like select. Which means customer didn't select any of the options.
- So, we can consider this select value as null value. Replace select value with null.
- Let's see the null value percentage in each column or feature in the dataset BEFORE replacing select with null value
- Replace select string with null value
- Let's see the null value percentage in each column or feature in the dataset AFTER replacing select with null value

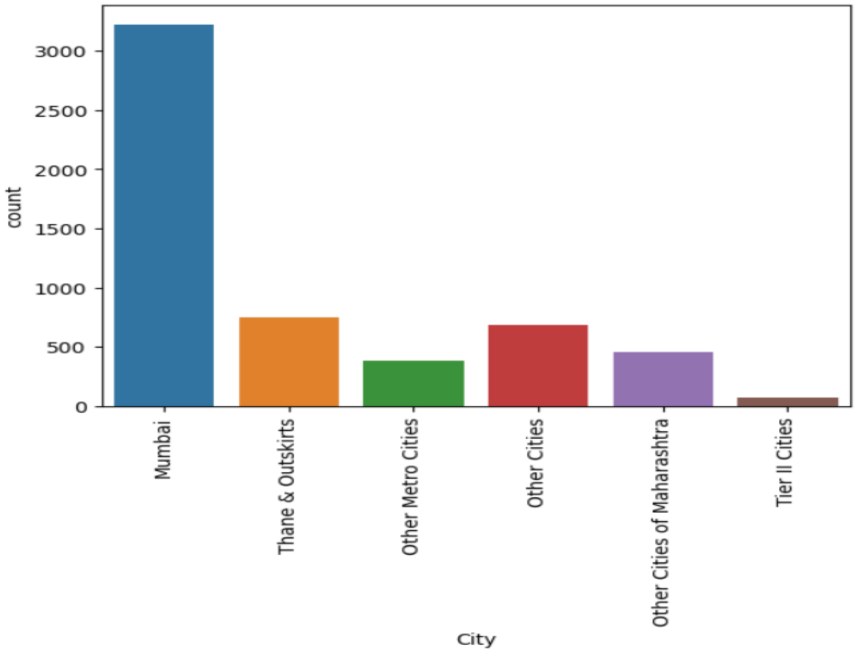
Dropping Columns

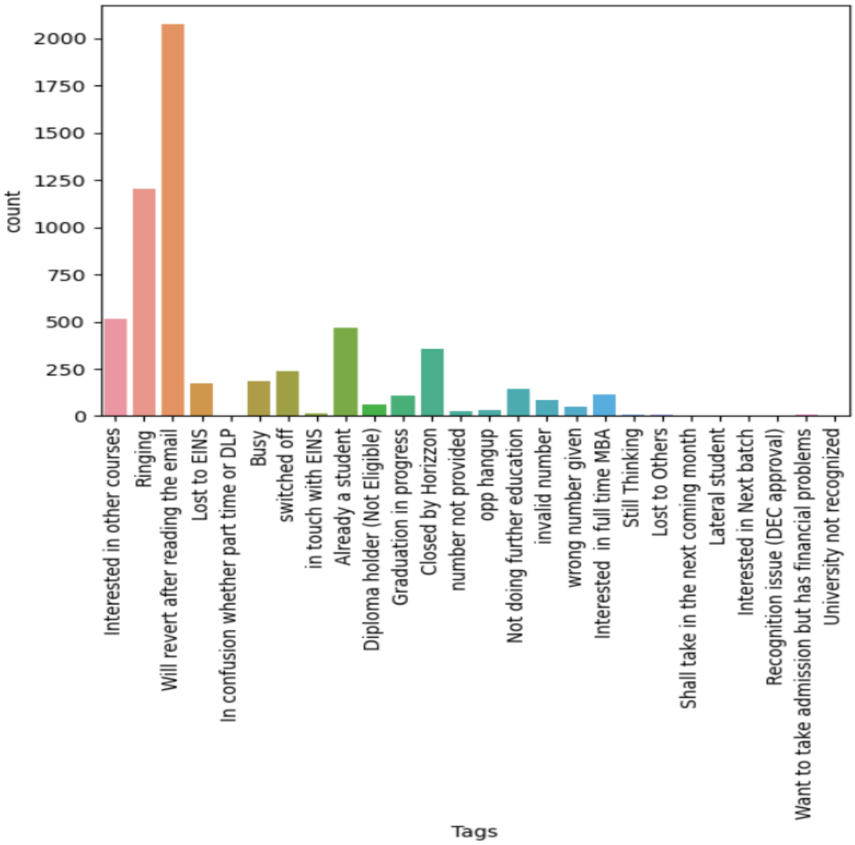
- We are dropping few columns which contains high null values. Here we are considering above 40% null value columns

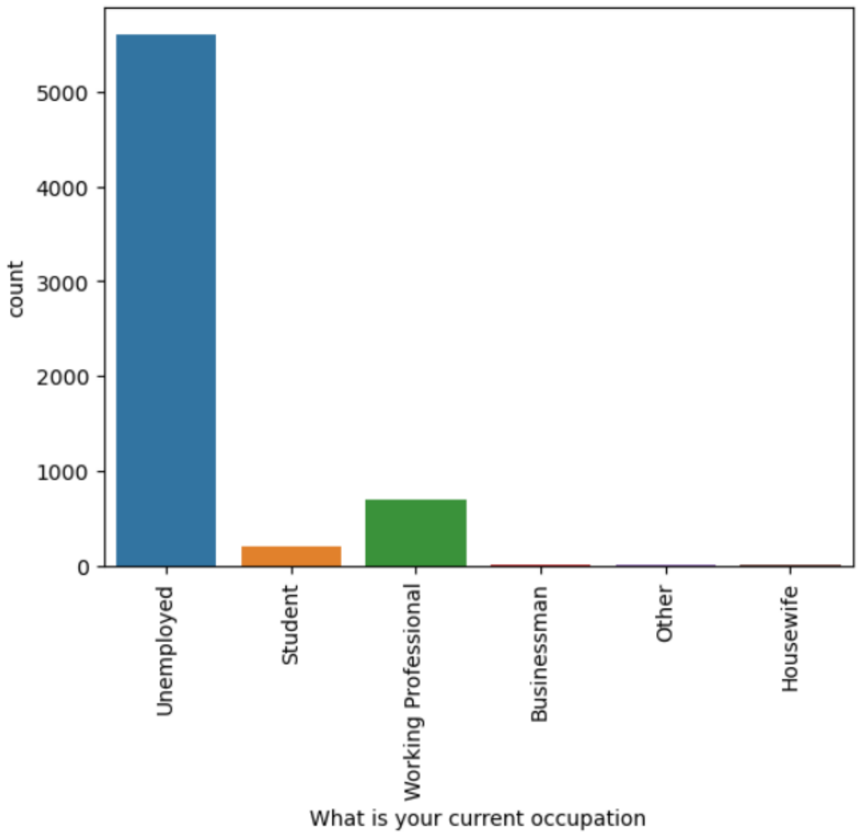
- We can also drop columns which contains more than 35% null value columns. But don't drop those columns right away
- We have few more columns with high percentage of nulls. Let's look at them individually
- We can use either histogram or count plot from seaborn for the same purpose



- We have around 36% of missing data for specialization. We can replace with mode. However, the customer might not have interest
- In existing course. for time being I'm replacing nulls with other.
- Replacing null values in the city column with mode value







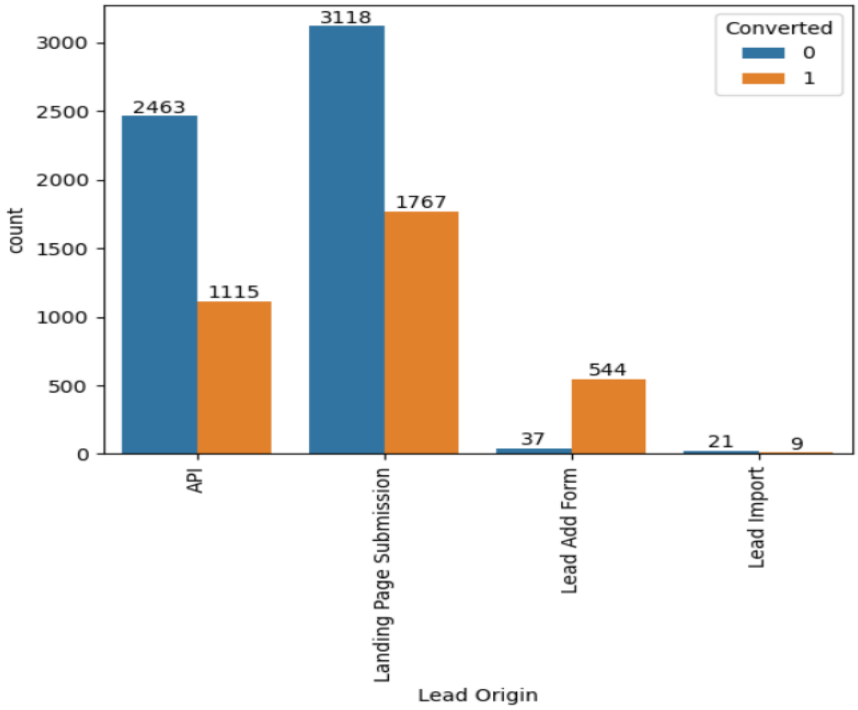
- Replacing null values in the 'What is your current occupation' column with mode value
- This feature is dominated by 1 value. We call it as skewed column. Let's drop this column
- we have to check if we have such columns in our dataset
- Country column is also skewed towards India. Hence, we are dropping this column
- Dropping records which contains null values. We are losing only less information hence dropping them
- Some columns contain only 1 value across the records. These kind of columns or features will not give any information to us. We can drop such columns

Exploratory Data Analysis

- Let's do univariate and bivariate analysis to understand the impact of that feature on our target variable.
- If the feature doesn't have any impact on the target, we can drop such features.
- Let's do the data analysis before doing data conversions.
- Columns which contain unique values across the dataset will not give much information to our problem.
- We can drop such columns. Here we have 2 features with unique values 'Prospect ID' and 'Lead Number'
- We can't drop 'Prospect ID' as it's useful to identify the customer.

Observation 1:

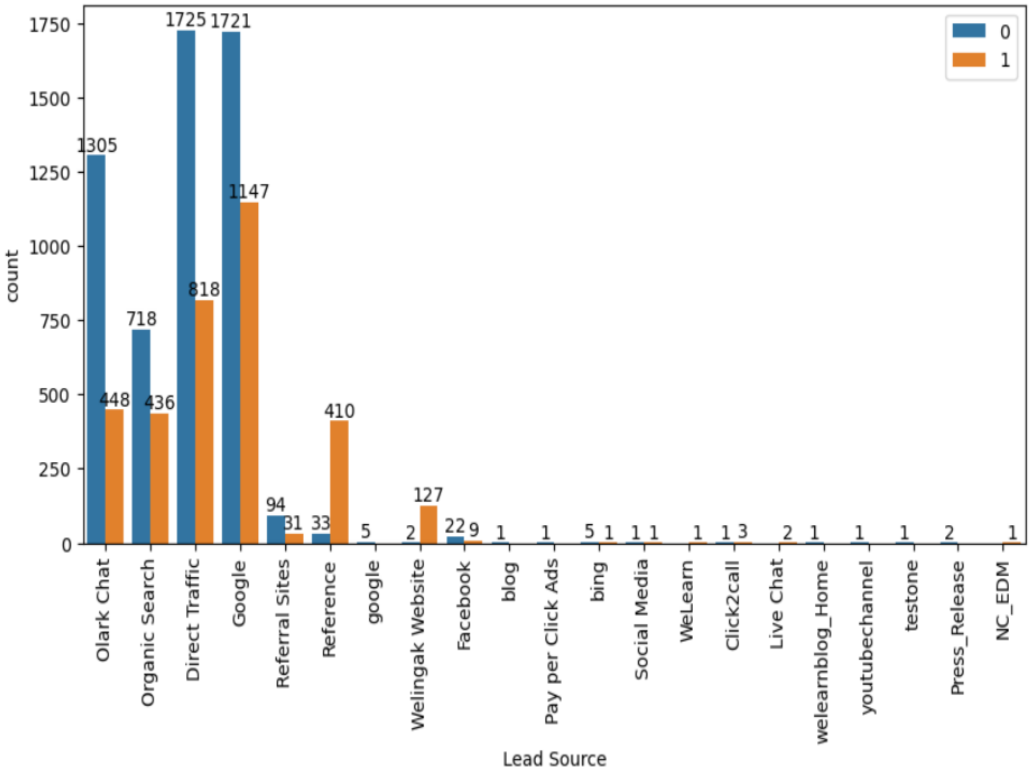
We have around 38% lead conversion rate in our dataset



Observation 2

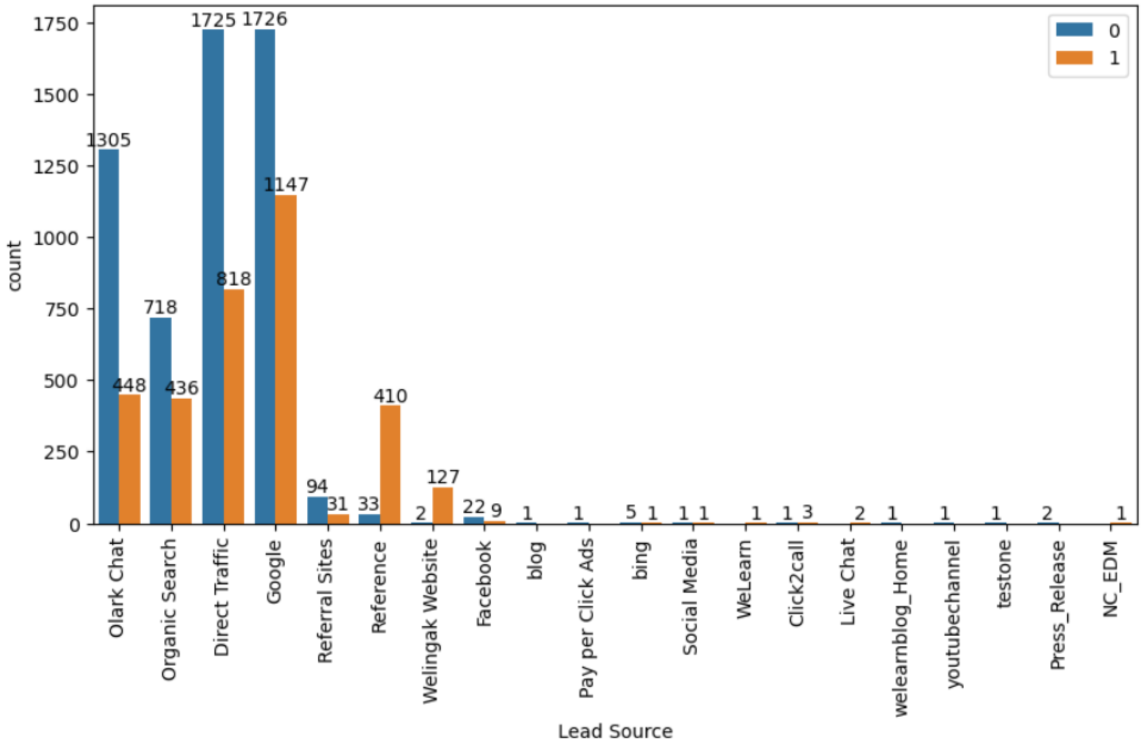
In Lead Origin feature, for 'Lead Add Form' the conversion rate is higher. It stands at 93%. If we can increase the incoming flow through this channel, there is higher chance of conversion rate

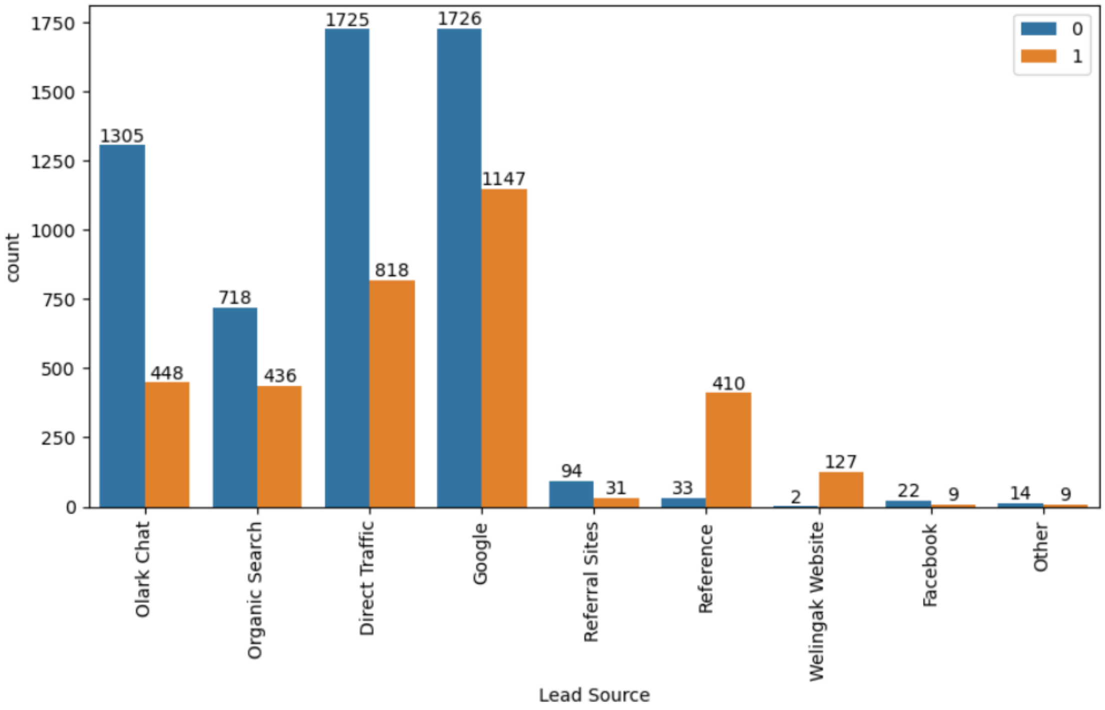
Both for 'API' & 'Landing Page Submission' segments conversion rate stands at 31% and 36%. We have a scope to increase the lead conversion in these 2 areas.



Here we have two versions of Google & google. let's convert them to single string

We can club few categorizes into one category as they are small in number



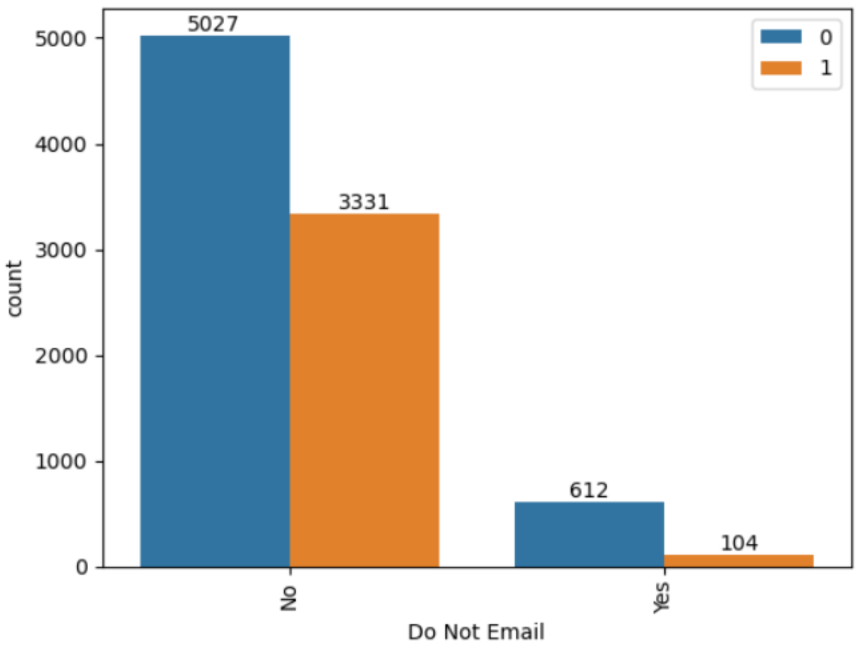


Observation 3:

We have high conversion rate for References. It's stands at 92%. However, we have only 5% of references are present in entire dataset. If we can increase the references we have higher chances of conversion

Direct Traffic & Google have good conversion rates of 32% & 39%. We have opportunity here to improve the conversion rates.

We also have higher conversion rate for Welingak Website. It's stands at 98%. However, we have only 1.5% of this data in the entire dataset. If we are able to increase the Welingak Website inflow we can have higher conversion rate thru this.

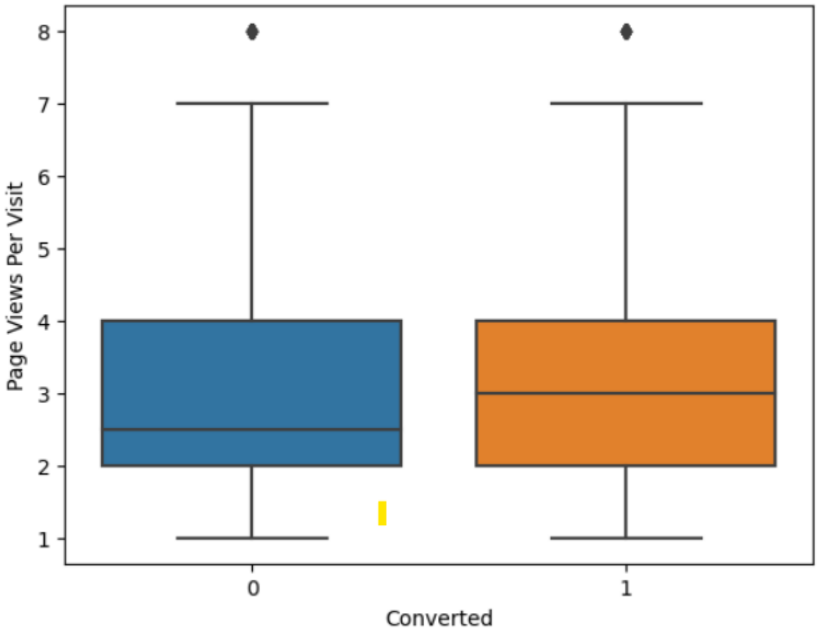


- Dropping 'Do Not Email' feature. It gives no information to us and it's also skewed
- 99% of data is related to No's in Do Not Call category. We can drop this column
- Dropping 'Do Not Call' feature. It gives no information to us and it's also skewed

Observation 4¶

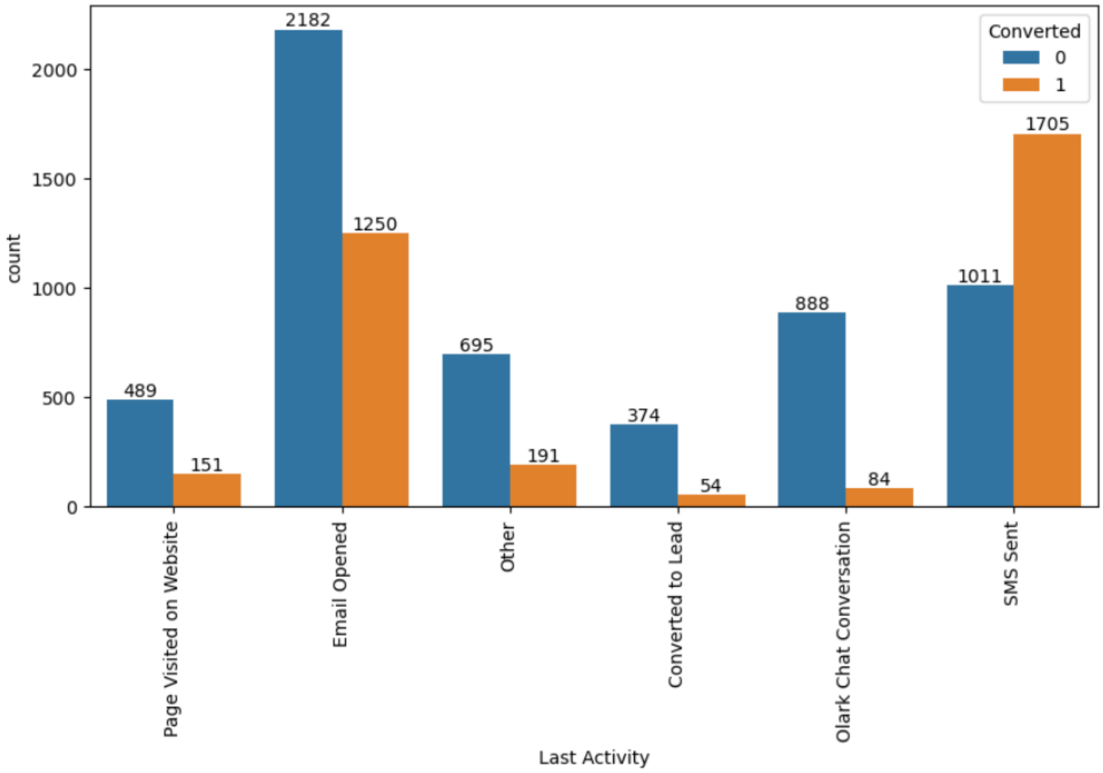
We have lot of outliers in the dataset. Total Visits 75% is 5 and maximum value is 251. which says we have lot of outliers. let's consider only till 90 percentile or 95 percentiles

Total Visits feature doesn't give much information. Median for both converted (3) and non-converted (4) looks similar. We can drop this column as well



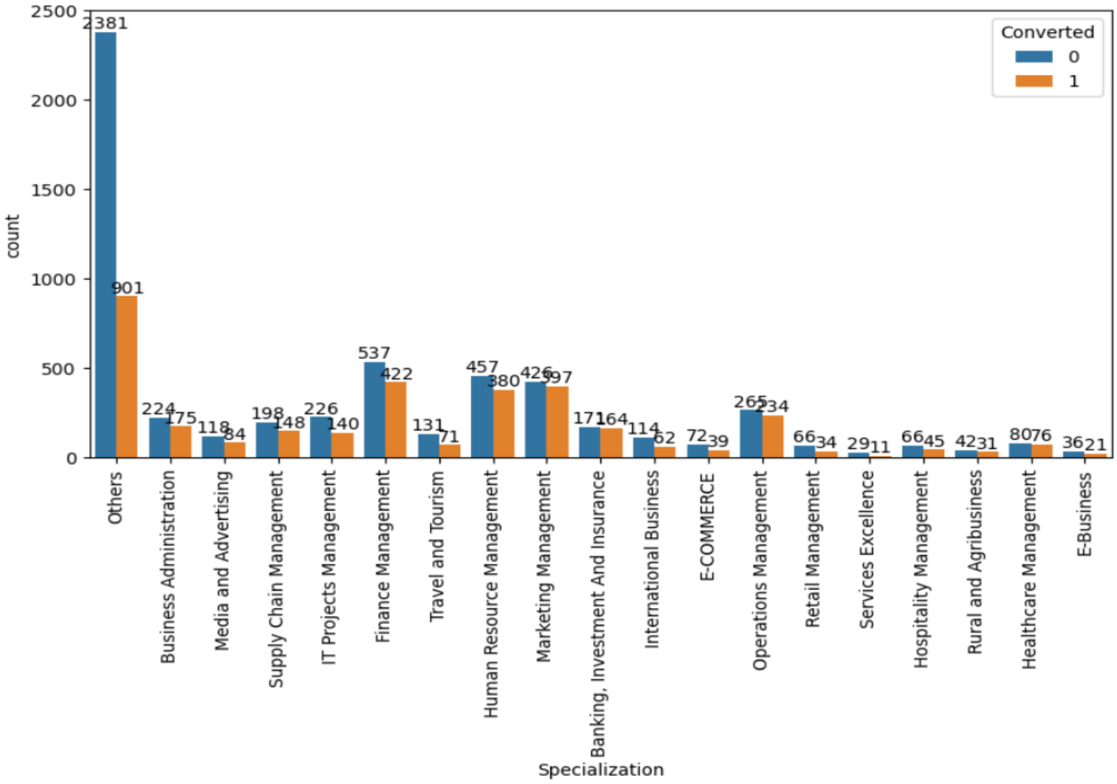
Observation 6

Page Views Per Visit feature median is more or less same for both converted and non-converted leads. It’s giving no info for us



Observation 7

1. SMS Sent category has higher conversion rate. It stands at 63%. If we can increase the influx of this category, we can have better conversion rate
2. Email Opened is another category which has conversion rate at 36%. We have to understand this and try to improve conversion rate for this category



Observation 8

1. Working professionals have higher conversion rate
2. Many unemployed are visiting the website. We can target this category to boost the conversion.

Observation 9

1. SMS Sent has the higher conversion rate
2. Email Opened also has very good conversion rate

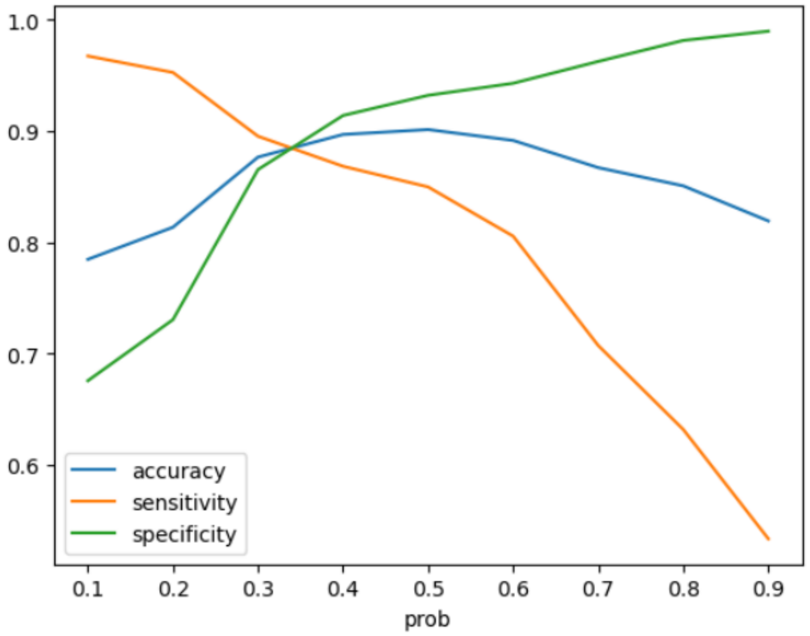
Model Building

- Scaling few columns
- Feature Selection using Recursive Feature Elimination - RFE
- Generic function to calculate VIF of variables
- Dropping 'What is your current occupation_Unemployed' due to high VIF (13). We have to consider only <5 VIF
- Dropping 'Tags_Lateral student' due to high p-value
- Dropping 'Tags_Lost to Others' & 'Tags_number not provided' due to high p-value
- Dropping few columns due to high p-value 'ags_Diploma holder (Not Eligible)' & 'Tags_invalid number' & 'Tags_wrong number given'

Metrics and Evaluation

- Confusion Matrix

- Precision and Recall
- Sensitivity
- Specificity
- False Positive Rate
- True Positive Rate
- ROC Curve



Final Conclusion

Test data

Accuracy : 89%
Precision : 82%
Recall : 89%
Sensitivity : 89%
Specificity : 87%
TruePositive Rate : 89%
False Positive rate : 60%

Train Data

Accuracy 88%
Precision 82 %
Recall 88%
Sensitivity: 88%
Specificity: 88%
False Positive Rate: 60%
True Positive Rate: 88%