Lead Scoring Case Study

By Vivek Chauhan, Vishesh Divya and Aryan Jain

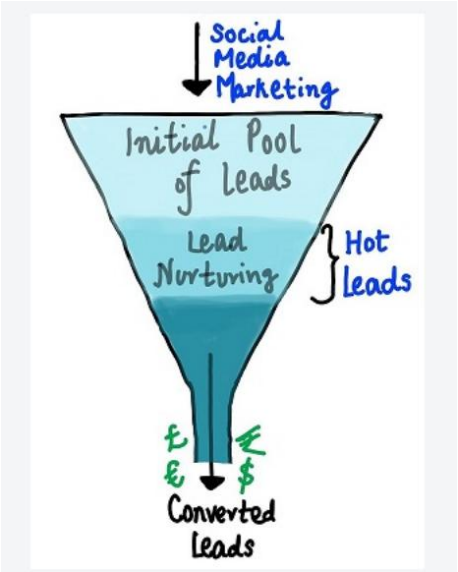
Business Objective

An education company named X Education sells online courses to industry professionals. The Company markets itself on search engines like Google, Bing and referral sites. Basis the marketing the company receives details of potential buyers (Leads) from these sites or past referrals.

Currently the company invests a lot of effort into marketing with low conversion rate (30%). To make this process more efficient, the company wishes to identify the most potential leads, also known as Hot Leads for a more **focused marketing and conversion strategy**. Thus, aiming for a **target lead conversion rate of 80% or higher**

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.

Current Lead Conversation Process



Proposed Lead Conversion Process

Identification of **Hot Leads**

Proactive Outreach to Hot Leads

Focused Marketing

Higher Lead Conversion

Goal of Case Study

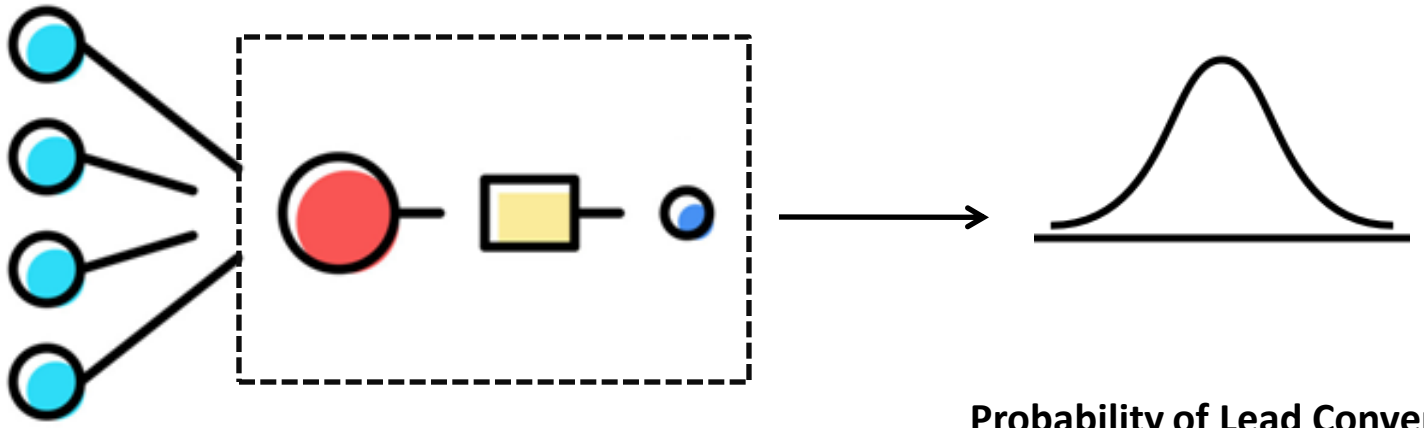
Build a logistic regression model to assign a lead score to all leads. Customers with higher lead scores will have a higher chance of conversion.

Target Lead Conversion Rate = 80%

Input:

Data for previous Leads with important attributes

- Lead Source
- Time Spend on
- Website
- Current Occupation
- Converted etc.



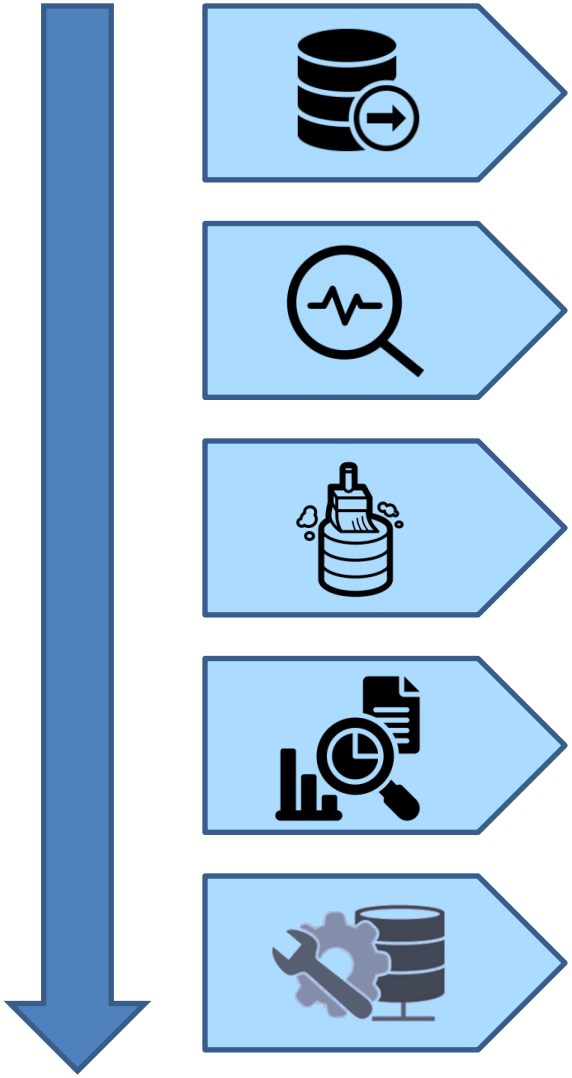
Logistic Regression Model:

- Understand pattern in past data
- Identify important attributes
- Predict Probability (Lead Score) of a Lead getting converted

Probability of Lead Conversion (Lead Score) :

Apply a calculated Probability cutoff to identify hot Leads

Solving for Data Nuances



1. Importing Data

2. Understanding Data

Understanding Data Types, understanding data values

3. Data Cleaning

Handling Null Values, Treatment of **Outliers**

4. Exploratory Data Analysis

Univariate and Bivariate Analysis, Creating Visualization for **Deeper understanding** of data, Identifying important variables for next step

5. Data Preparation & Feature Engineering

Removing unimportant variables, Creating **Dummy Variables** for Categorical data, Creating Separate sets for **Training and Test Data**, **Standardizing** Numerical Variables

Exploring Data

Firstly, we've taken the 'leads' dataset for analysis. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not. The target variable, in this case, is the column 'Converted' which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted.

Reading and understanding the input data

In [2]:
```
# reading the input data file
leads= pd.read_csv('Leads.csv')
```

In [3]:
```
# Checking Shape of Data
leads.shape
```

Out[3]: (9240, 37)

In [4]:
```
# setting parameters to see all the columns
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)

# viewing data head
leads.head()
```

Out[4]:

	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	What is you curren occupatio
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	Unemployee
1	2a272436-5132-4136-86fa-dcc88c88f482	660728	API	Organic Search	No	No	0	5.0	674	2.5	Email Opened	India	Select	Select	Unemployee
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	Studen
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	Unemployee
4	3256f628-e534-4826-9d63-4a8b88782852	660681	Landing Page Submission	Google	No	No	1	2.0	1428	1.0	Converted to Lead	India	Select	Other	Unemployee

Data Cleaning

DATA CLEANING -

Before proceeding with the analysis data needs to be cleaned to make it fit for further processing.

- Firstly, we checked null values present across all the columns.
- As mentioned in the problem statement, many of the categorical variables have a level called 'Select' which are as good as a null value. So, we replaced them with ‘NaN’.
- Then, we had dropped columns with 30% or more missing values.
- For columns with a low frequency of null values, we dropped the rows containing those.
- Then, we segregated the categorical and numerical columns for better analysis.
- We analysed each column individually to figure out a way to deal with their respective missing values.
- We had dropped all the highly skewed columns as they wouldn’t contribute much to our model.

```
In [7]: leads.isnull().sum()

Out[7]: Prospect ID                                    0
        Lead Number                                    0
        Lead Origin                                    0
        Lead Source                                   36
        Do Not Email                                   0
        Do Not Call                                    0
        Converted                                      0
        TotalVisits                                  137
        Total Time Spent on Website                    0
        Page Views Per Visit                         137
        Last Activity                                103
        Country                                     2461
        Specialization                              1438
        How did you hear about X Education          2207
        What is your current occupation             2690
        What matters most to you in choosing a course 2709
        Search                                         0
        Magazine                                       0
        Newspaper Article                              0
        X Education Forums                             0
        Newspaper                                      0
        Digital Advertisement                          0
        Through Recommendations                        0
        Receive More Updates About Our Courses         0
        Tags                                        3353
        Lead Quality                                4767
        Update me on Supply Chain Content              0
        Get updates on DM Content                      0
        Lead Profile                                2709
        City                                        1420
        Asymmetrique Activity Index                 4218
        Asymmetrique Profile Index                  4218
        Asymmetrique Activity Score                 4218
        Asymmetrique Profile Score                  4218
        I agree to pay the amount through cheque       0
        A free copy of Mastering The Interview         0
        Last Notable Activity                          0
        dtype: int64
```

Data Understanding Summary

The Data has 9240 records and 37 Columns

- Prospect ID and Lead ID for identification purposes

- Target Variables - Converted

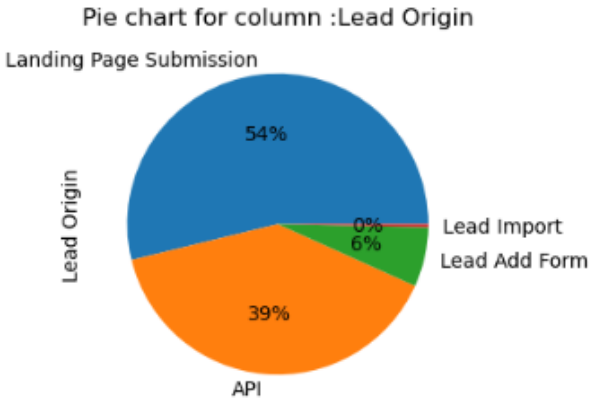
- Numerical Variables
  TotalVisits, Total Time Spent on Website, Page Views Per Visit

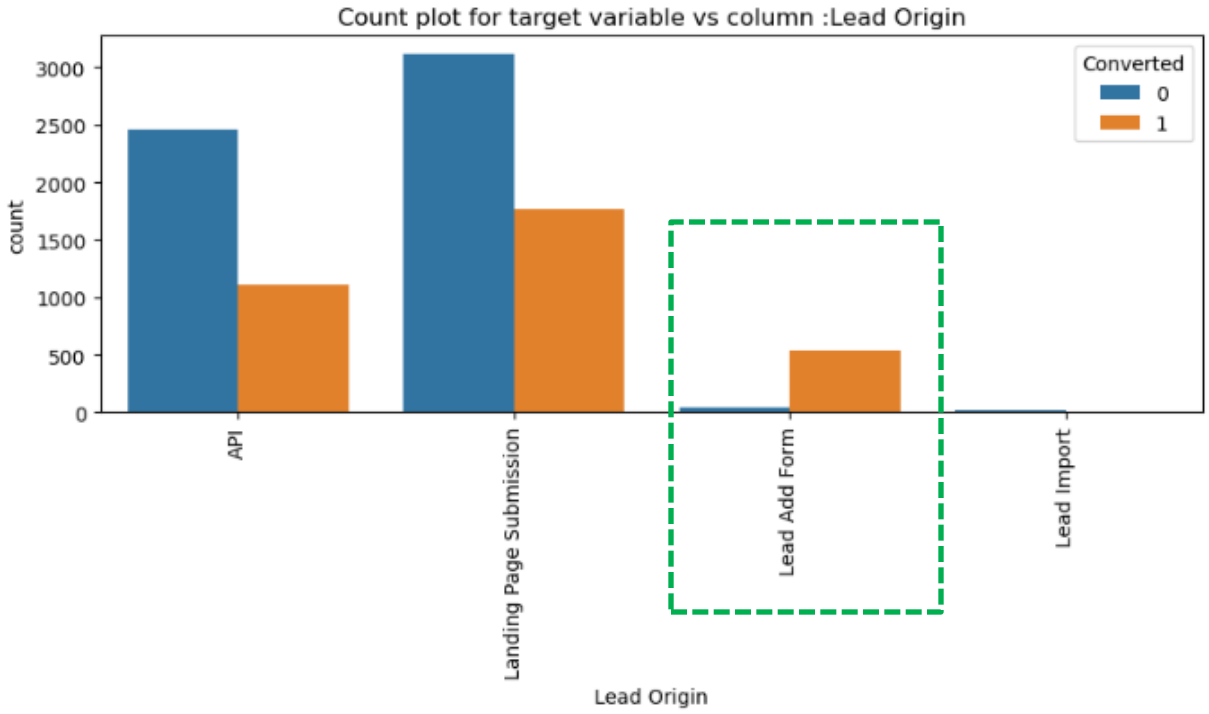
- Important Categorical Variables
  Lead Origin, Lead Source, Last Activity, What is your current occupation, **Last Notable Activity**. There are more but limiting the list to important ones identified in the first view

- Not Important Variables :
  Many Variables had **high percentage of null / select (to be treated as null) .** Finally we dropped all those columns which had more than 30% null values**.** Examples - Lead Quality, Lead Profile, City, Asymmetrique Activity Index, Asymmetrique Profile Index, Asymmetrique Activity Score, Asymmetrique Profile Score

  Many variables has only one value corresponding to 100% frequency. These do not add much value and can be dropped. Examples - Do Not Call, What matters most to you in choosing a course, Search, Magazine etc

Exploratory Data Analysis : Important Categorical Variables

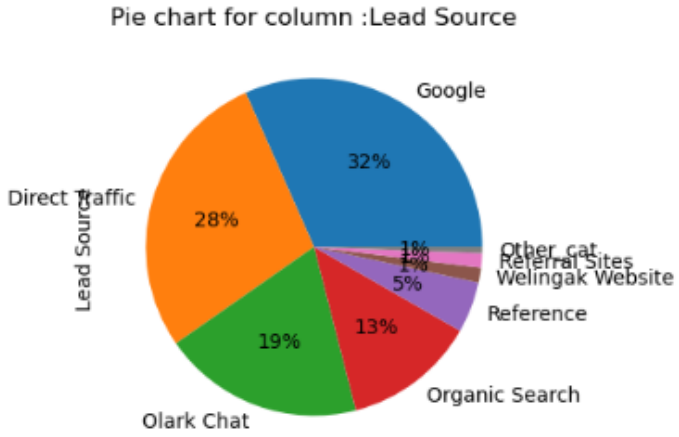


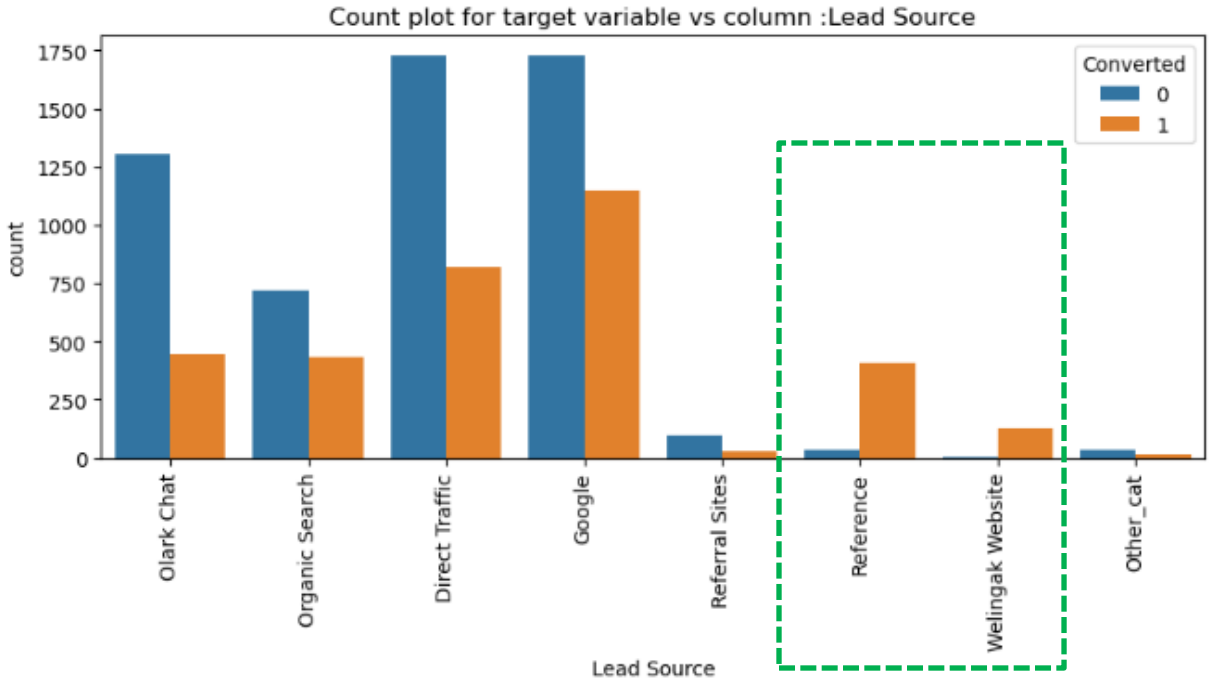


Understanding Variable **Lead Origin**

1. It tells about The origin identifier with which the customer was identified to be a lead. Includes API, Landing Page Submission, etc.
2. **Lead Add Form** happens to have the highest conversion rate. This can be important input to identify hot leads
3. These are the hottest leads which have much higher probability of conversion, this is followed by API and Landing Page Submission
4. Lead Import has negligible count

Exploratory Data Analysis : Important Categorical Variables

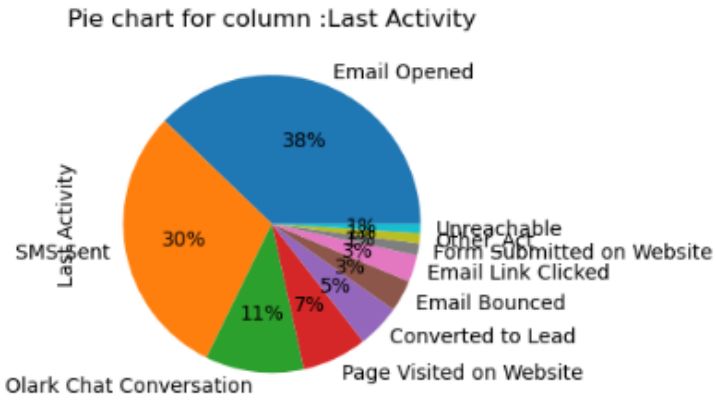


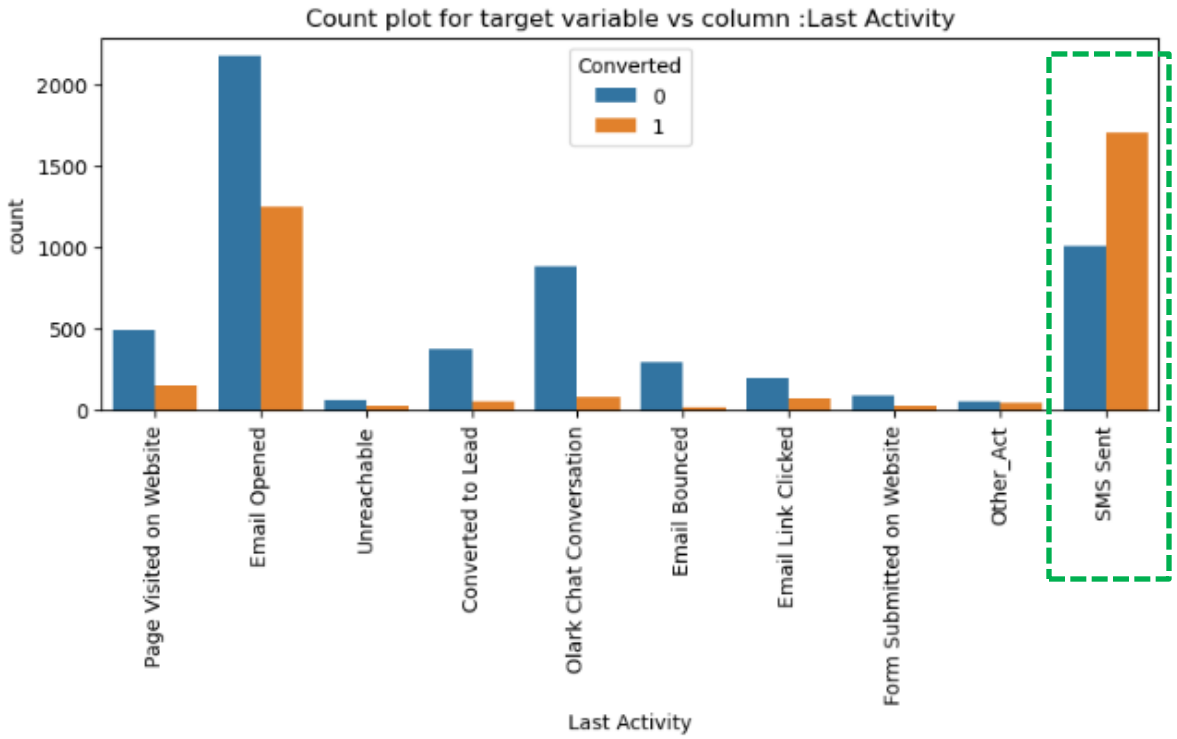


Understanding Variable **Lead Source**

1. It tells about the source of the lead - Includes Google, Organic Search, Olark Chat, etc.
2. Conversion Rate is highest for Lead Source - **Welingak Websit and Reference**. This could be an important input as we plan to maximize Conversion Rates
3. Top 3 lead generating sources are - Google, Direct Traffic and Olark Chat
4. Also, there were many sources with negligible frequency which we have **grouped together as Other_cat**

Exploratory Data Analysis : Important Categorical Variables

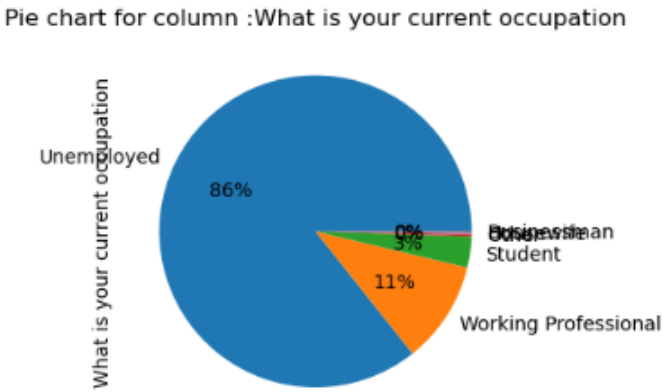


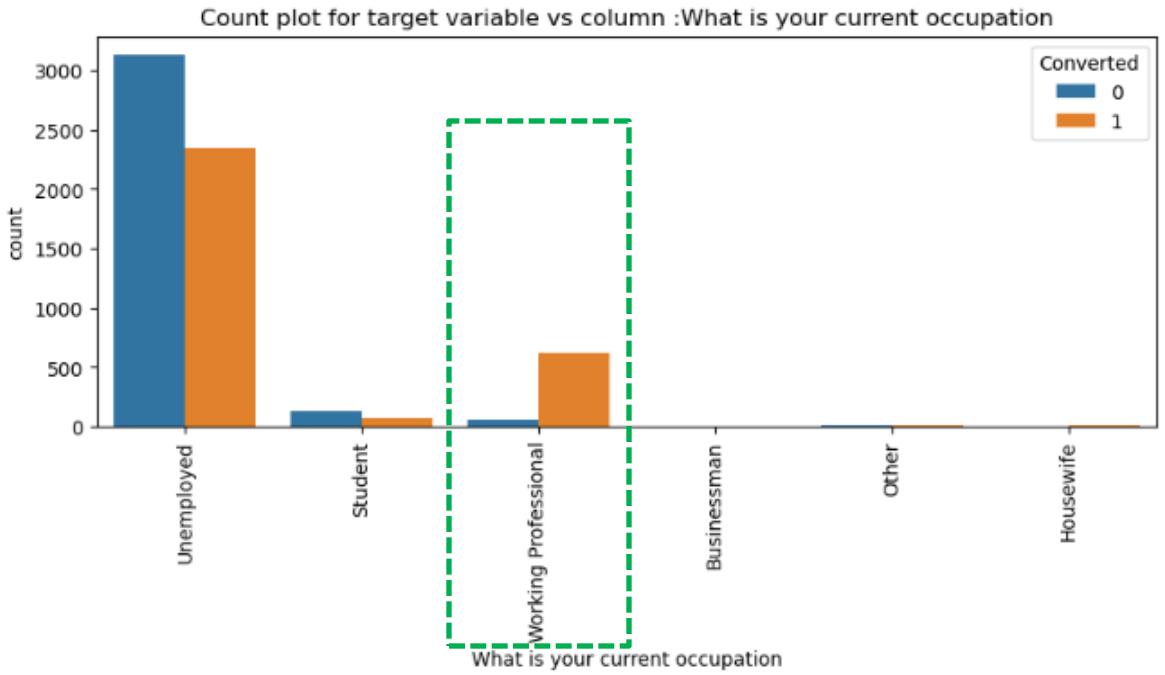


Understanding Variable **Last Activity**

1. **SMS sent** has highest conversion rate. This is an important input to identify hot leads
2. Olark Chat Conversation, Converted to Lead, Email Bounced have lowest conversion rates. These can be important to identify non conversions
3. Top 3 activities - Email Opened, SMS Sent, Olark Chat Conversation
4. Also, there were many activities with negligible frequency which we have **grouped together as Other_Act**

Exploratory Data Analysis : Important Categorical Variables

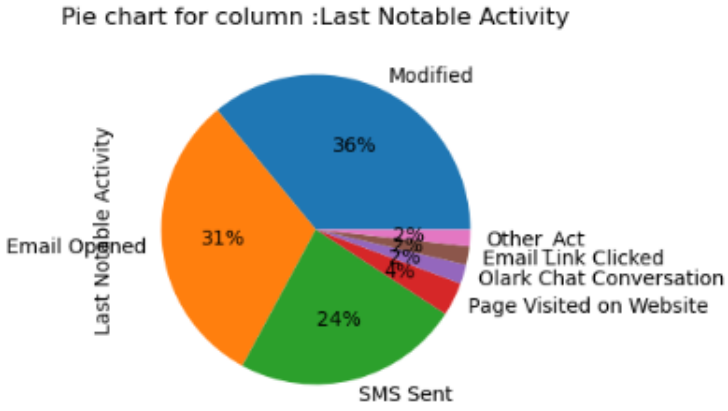


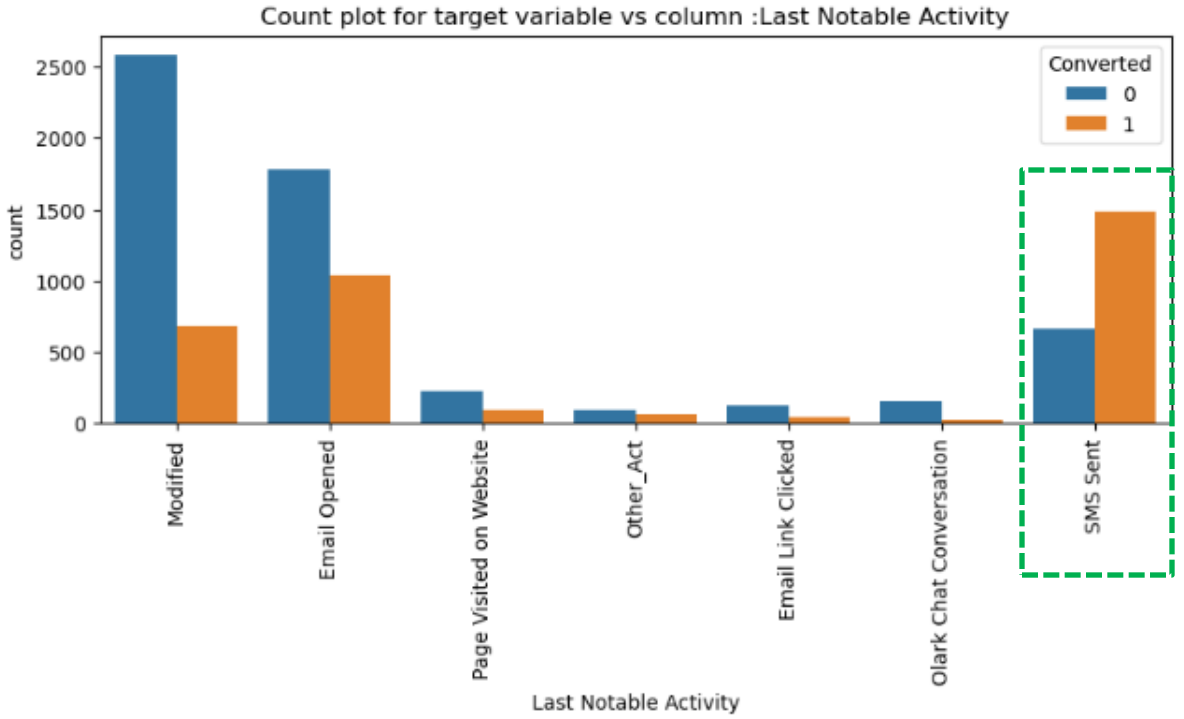


Understanding Variable **What is your current occupation**

1. Maximum values are **Unemployed - 86%**
2. **Working Professional** have very high conversion rate. This can be an important input to identify hot leads
3. There are many NULLs, now 86% of the Occupation is Unemployed. So Nulls can be replaced with Unemployed

Exploratory Data Analysis : Important Categorical Variables

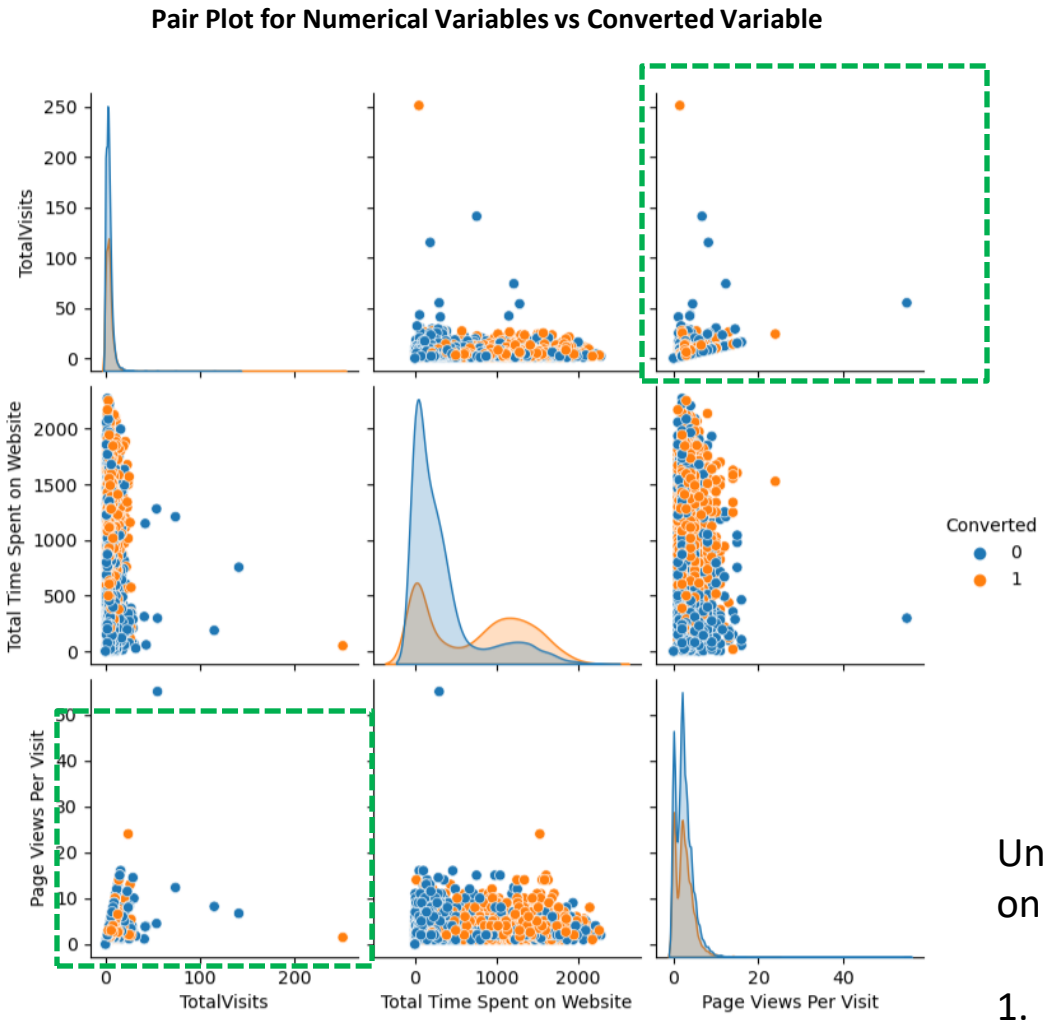




Understanding Variable **Last Notable Activity**

1. Top 3 Last Notable Activity are - Modified, Email Opened and SMS Sent
2. **SMS Sent** has the highest Conversion Rate and this can be an important input to identify hot leads
3. Also, there were many activities with negligible frequency which we have **grouped together as Other_Act**

Exploratory Data Analysis : Numerical Variables

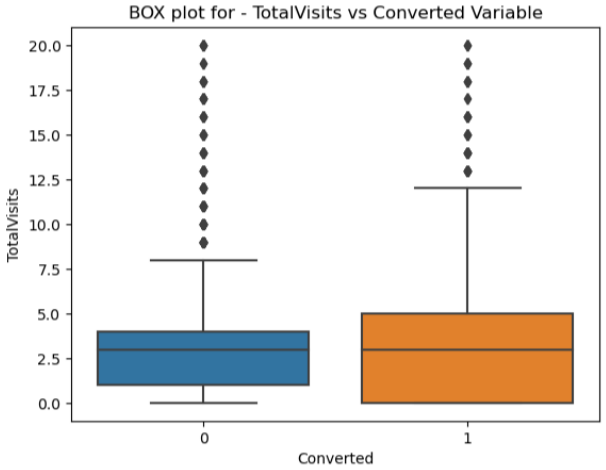


Understanding Variables - 'TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit'

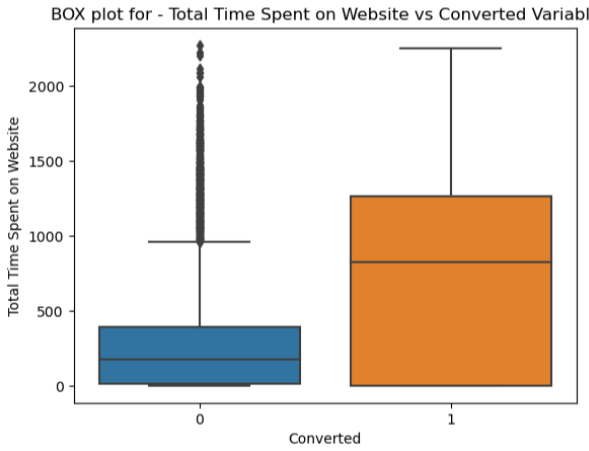
1. We see very high correlation between Total Visits and Page Views Per Visit
2. We will handle this in the Model Building phase via VIF

Exploratory Data Analysis : Numerical Variables

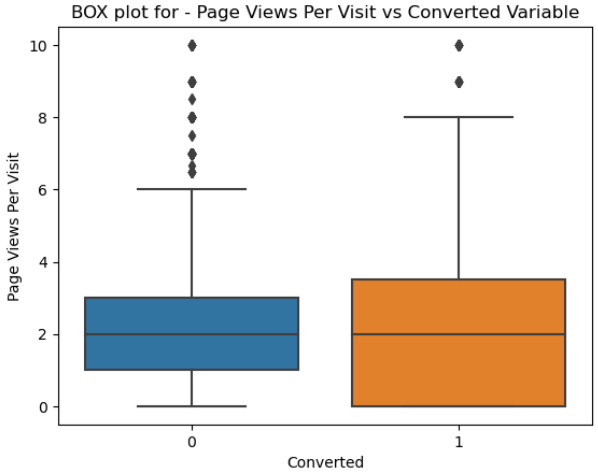
We identified outliers in the numerical values and remove them. Post outlier removal we are analyzing the numerical columns vis a vis the target variable - converted



1. The median of TotalVisits is similar for Non Converted and Converted Prospects
2. Inter Quartile Range is much higher for Converted Prospects
3. Also, we do not see any high outliers and the data is more of less in continuous pattern



1. Non Converted Prospects have much lower Median and IQR
2. On other hand, Converted Prospects have much high Median and IQR.
3. This indicates that CMs who spend higher time on website have higher changes of conversion. Which can be used to identify hot leads



1. There is no difference in the median for Not Converted vs Converted Prospects
2. Converted prospects have higher IQR

Feature Engineering

For Numerical Variables – we have done **Standardization using StandardScaler()** so that they are on similar scale and not impact model coefficients

scaler = StandardScaler()
X_train[['TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit']]= scaler.fit_transform(X_train[['TotalVisits',
'Total Time Spent on Website', 'Page Views Per Visit']])

Out[70]:

	Do Not Email	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Lead Origin_Landing Page Submission	Lead Origin_Lead Add Form	Lead Origin_Lead Import	Lead Source_Google	Lead Source_Olark Chat	Lead Source_Organic Search	Lead Source_Other_cat	Sou
5491	0	0.572234	-0.370985	0.099392	1	0	0	0	0	0	0	
3850	0	-1.058274	-0.890176	-1.203491	0	0	0	0	1	0	0	

For Categorical Variables – we have created dummy variables **using pd.get_dummies()**

var_dum= pd.get_dummies(leads[['Lead Origin','Lead Source', 'Last Activity', 'What is your current occupation', 'Last Notable Activity']], drop_first=True)

This will help us get the features be desired form for optimal modeling exercise

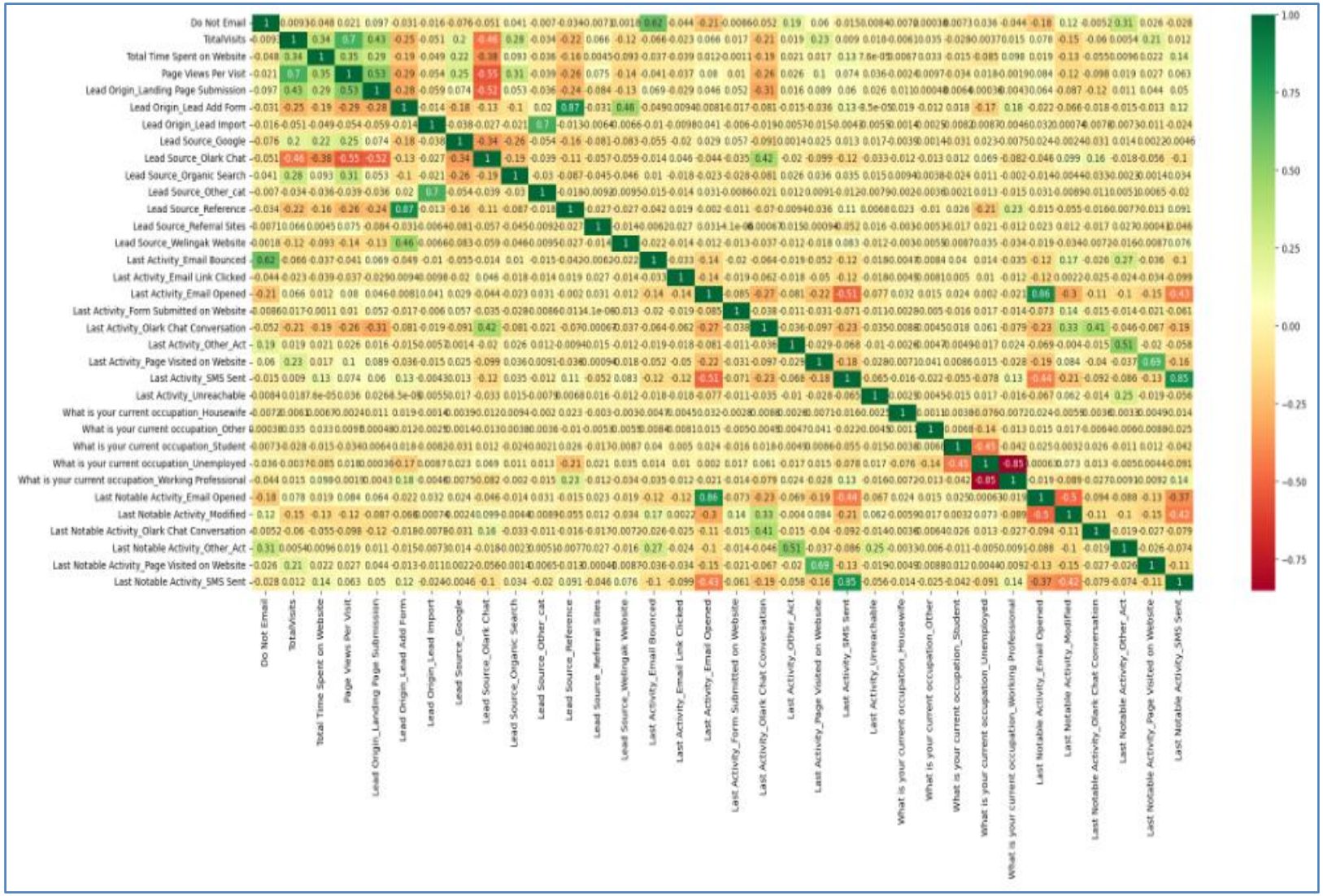
In [156]:
```
leads1.head()
```

Out[156]:

	Prospect ID	Lead Number	Do Not Email	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Lead Origin_Landing Page Submission	Lead Origin_Lead Add Form	Lead Origin_Lead Import	Lead Source_Google	Lead Source_Olark Chat	L Source_Org Se
0	7927b2df-8bba-4d29-b9a2-b6e0beafe620	660737	0	0	0.00	0	0.00	0	0	0	0	1	
1	2a272436-5132-4136-86fa-dcc88c88f482	660728	0	0	5.00	674	2.50	0	0	0	0	0	

Understanding Correlation between Features

- We see some darker shades which signify high correlation between variables.
- We will address these in our modelling exercise using VIF scores.



Building the Model

We will create two sets of Data – Training Data (70%) and Test Data (30%).
This will help us evalaute the model performance on unseen data

 # We have used the option stratify = y to ensure that training and test data has proportionate values of Output variable y
X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.7, test_size=0.3, random_state=100, stratify=y)

Since there are several variables we have used RFE to pick the top 20 significant variables.

from sklearn.feature_selection import RFE
rfe = RFE(estimator=logreg, n_features_to_select= 20)
rfe = rfe.fit(X_train, y_train)

We will then refine the model by doing multiple iterations and understanding model statistics – Pvalue and VIF to remove less significant / correlated variables.

We will be removing variables basis below criteria

1. **Pvalue more than 5%** - this means that the variable is not significant
2. **VIF value more than 5** – This means that the variable shows multi-collinearity with other variables in the model

Final Model – Model Statistics and VIF Scores

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6288
Model:	GLM	Df Residuals:	6274
Model Family:	Binomial	Df Model:	13
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2640.1
Date:	Tue, 18 Jul 2023	Deviance:	5280.2
Time:	15:56:00	Pearson chi2:	6.38e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.3854
Covariance Type:	nonrobust		

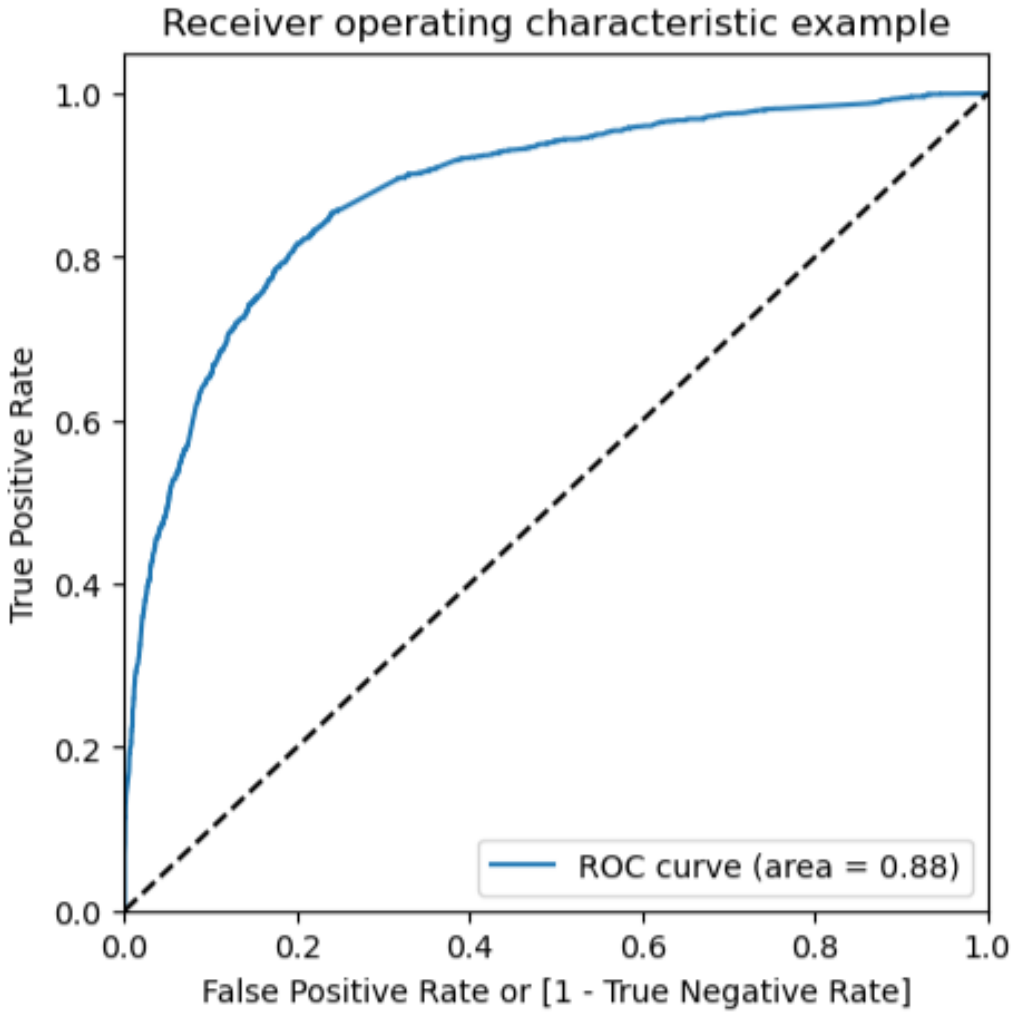
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.2937	0.218	-5.928	0.000	-1.721	-0.866
Do Not Email	-1.3943	0.209	-6.682	0.000	-1.803	-0.985
Total Time Spent on Website	1.1312	0.040	28.152	0.000	1.052	1.210
Lead Source_Olark Chat	1.2305	0.103	11.925	0.000	1.028	1.433
Lead Source_Other_cat	0.9639	0.382	2.523	0.012	0.215	1.713
Lead Source_Reference	4.0761	0.233	17.480	0.000	3.619	4.533
Lead Source_Welingak Website	6.4766	1.012	6.400	0.000	4.493	8.460
Last Activity_Email Bounced	-1.3540	0.423	-3.198	0.001	-2.184	-0.524
Last Activity_Email Opened	0.3601	0.090	3.982	0.000	0.183	0.537
Last Activity_Olark Chat Conversation	-1.0776	0.169	-6.382	0.000	-1.409	-0.747
What is your current occupation_Unemployed	-0.4212	0.210	-2.005	0.045	-0.833	-0.009
What is your current occupation_Working Professional	2.2215	0.273	8.151	0.000	1.687	2.756
Last Notable Activity_Other_Act	1.9295	0.309	6.253	0.000	1.325	2.534
Last Notable Activity_SMS Sent	1.8812	0.100	18.808	0.000	1.685	2.077

Features	VIF
What is your current occupation_Unemployed	3.86
Last Activity_Email Opened	2.48
Last Notable Activity_SMS Sent	2.00
Do Not Email	1.86
Lead Source_Olark Chat	1.80
Last Activity_Email Bounced	1.72
Last Activity_Olark Chat Conversation	1.70
What is your current occupation_Working Profes...	1.38
Total Time Spent on Website	1.31
Lead Source_Reference	1.20
Last Notable Activity_Other_Act	1.16
Lead Source_Welingak Website	1.05
Lead Source_Other_cat	1.01

Pvalue is less than 5% for all variables

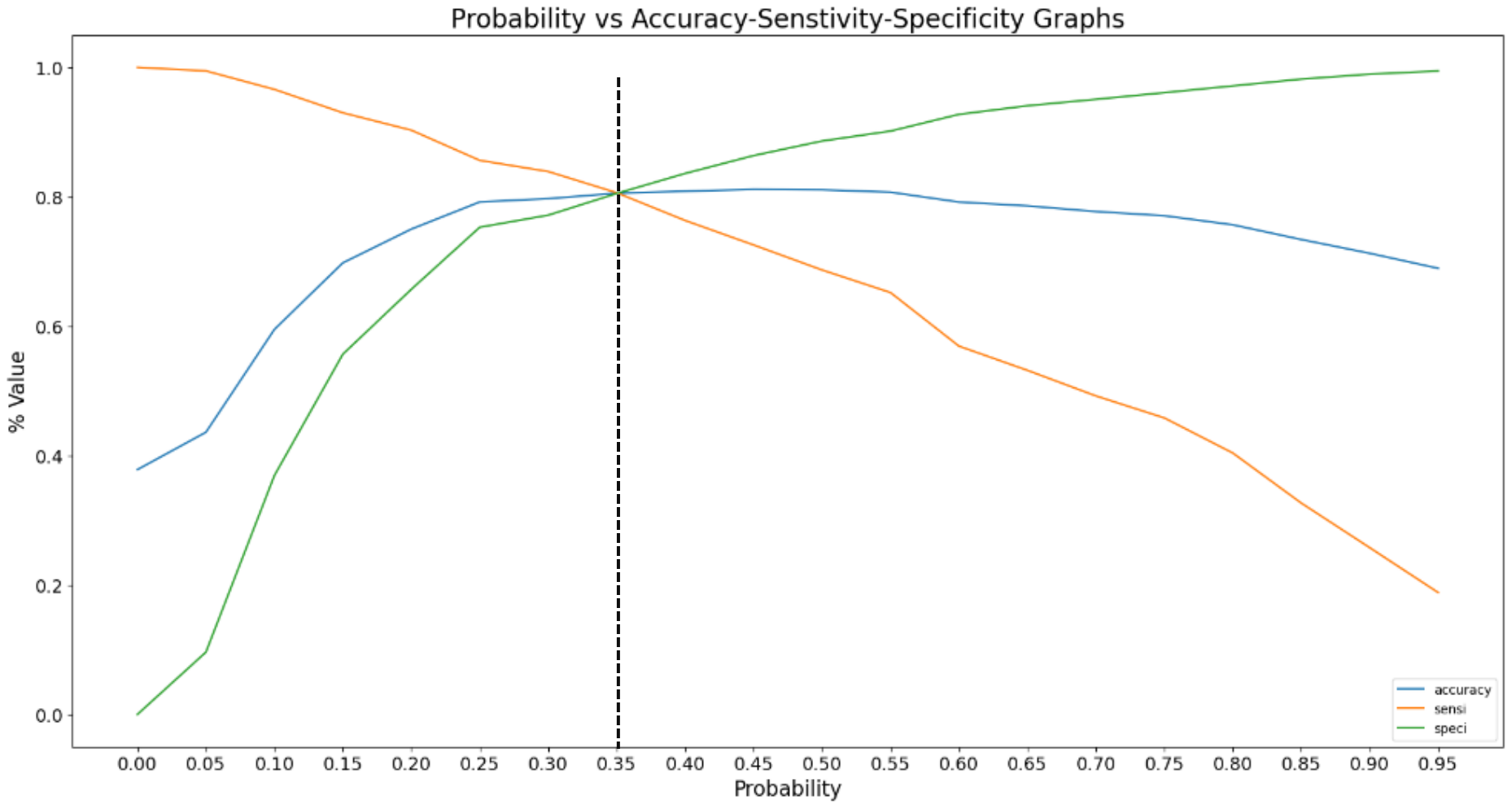
VIF is less than 5 for all variables

ROC Curve



- The area under the curve is 0.88 out of 1 which indicates a good predictive model.
- The curve is close to the top-left corner of the plot, which represents a model with high true positive rate and a low false positive rate at all thresholds

Probability Cutoff vs Accuracy and Other Metrics



From the above graph, approx 0.35 is the optimum point to take it as a cutoff probability

Confusion Matrix and Accuracy Metrics

Using 0.35 as probability threshold, we were able to arrive at below confusion matrix and accuracy scores

Confusion Matrix on Training Data

3147	762
461	1918

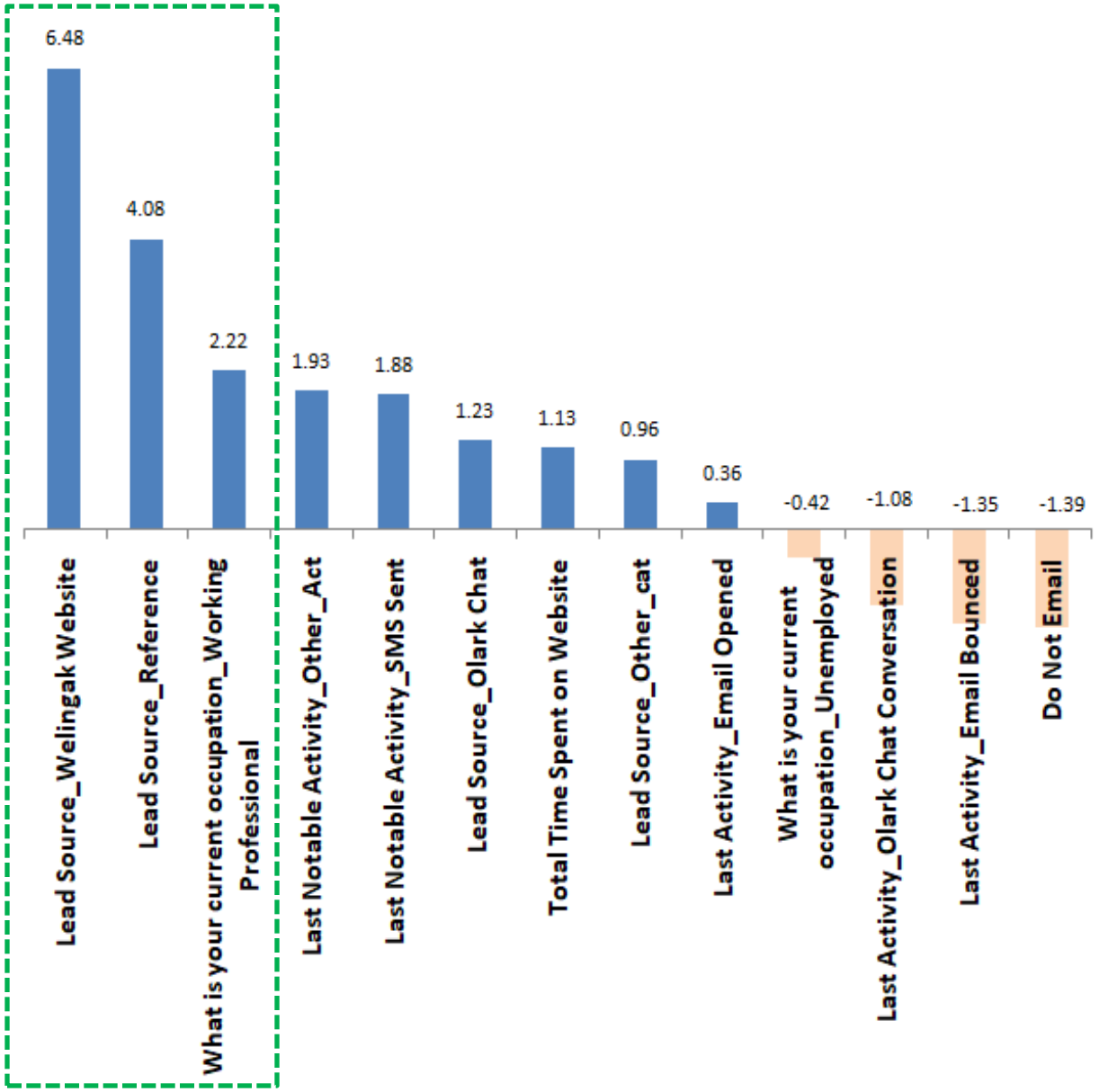
- Accuracy = 80.6%
- Sensitivity = 80.6%
- Specificity = 80.5%
- Precision = 71.6%
- Recall = 80.6%

Confusion Matrix on Test Data

1362	314
208	811

- Accuracy = 80.6%
- Sensitivity = 79.6%
- Specificity = 81.2%
- Precision = 72.0%
- Recall = 79.6%

Final Model - Features



Top Three variables which lead to a higher conversion rate are

- Lead Source_Welingak Website 6.48
- Lead Source_Reference 4.08
- What is your current occupation_Working Professional 2.22

Conclusion and Recommendations

Focus on key features. This can be done as below

- More advertising on **Welingak Website** in terms of advertising, etc.
- Create an **referral program** where past learners can refer new ones
- Create a plan for **Working professionals** to aggressively target them

Depending on workforce available and ROI select the **optimal Probability Cutoff**.

- If focusing on **Sensitivity, we can lower the probability** cutoff to and optimal point where we are able to see the maximum return
- If focusing on **Specificity, we can increase the probability** cutoff to and optimal point where we are able to see the maximum return