Problem Statement

- An education company named X Education sells online courses to industry professionals.
- The company markets its courses on several websites and search engines like Google.
- Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos.
- When these people fill up a form providing their email address or phone number, they are classified to be a lead. 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

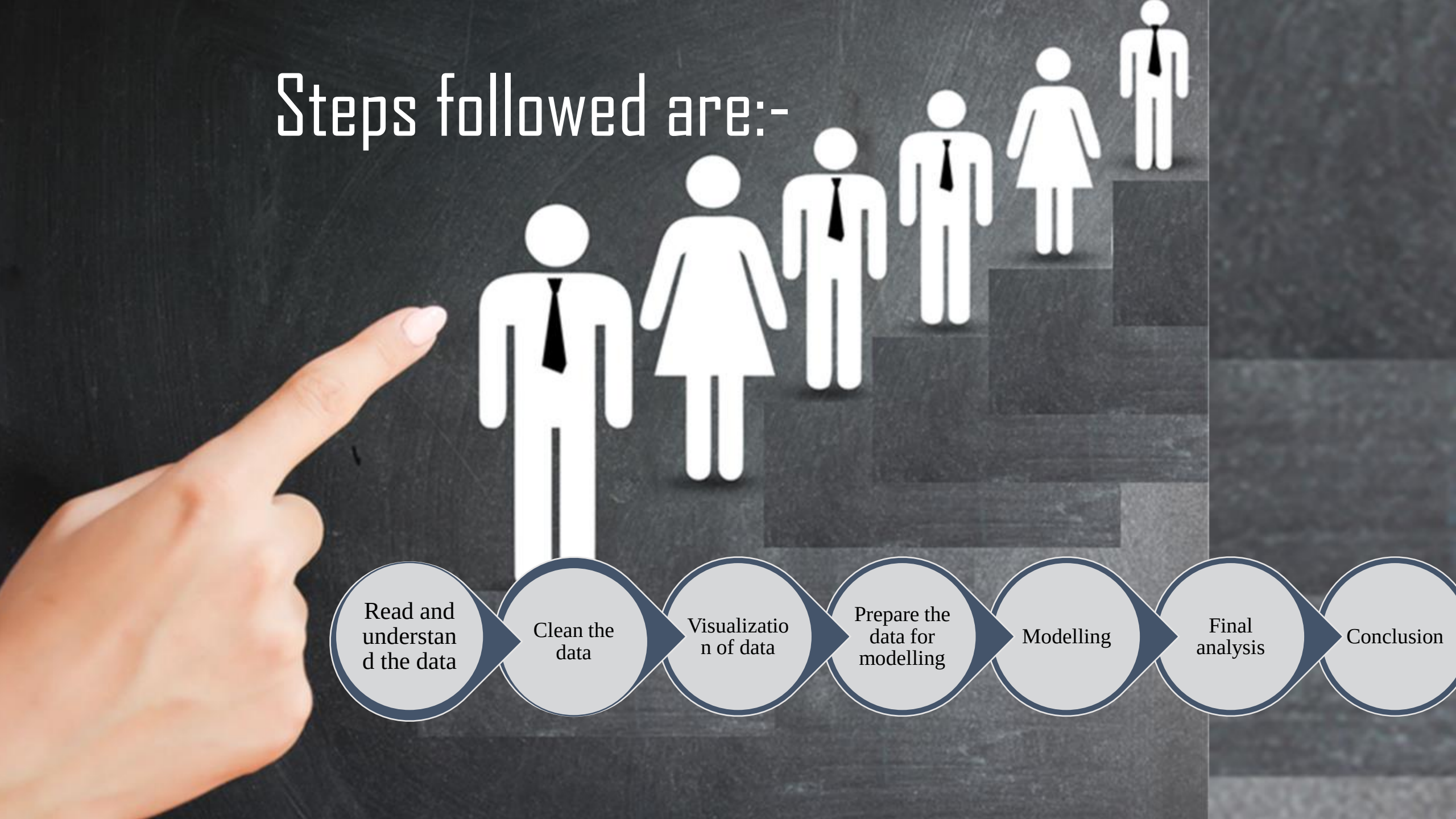
OVERVIEW

Help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.

Build a model where customers with higher lead score have a higher conversion

and Make target lead conversion rate to be around 80%.

Steps followed are:-



1. Read and understand the data
2. Clean the data
3. Visualization of data
4. Prepare the data for modelling
5. Modelling
6. Final analysis
7. Conclusion

READ AND UNDERSTANDING DATA

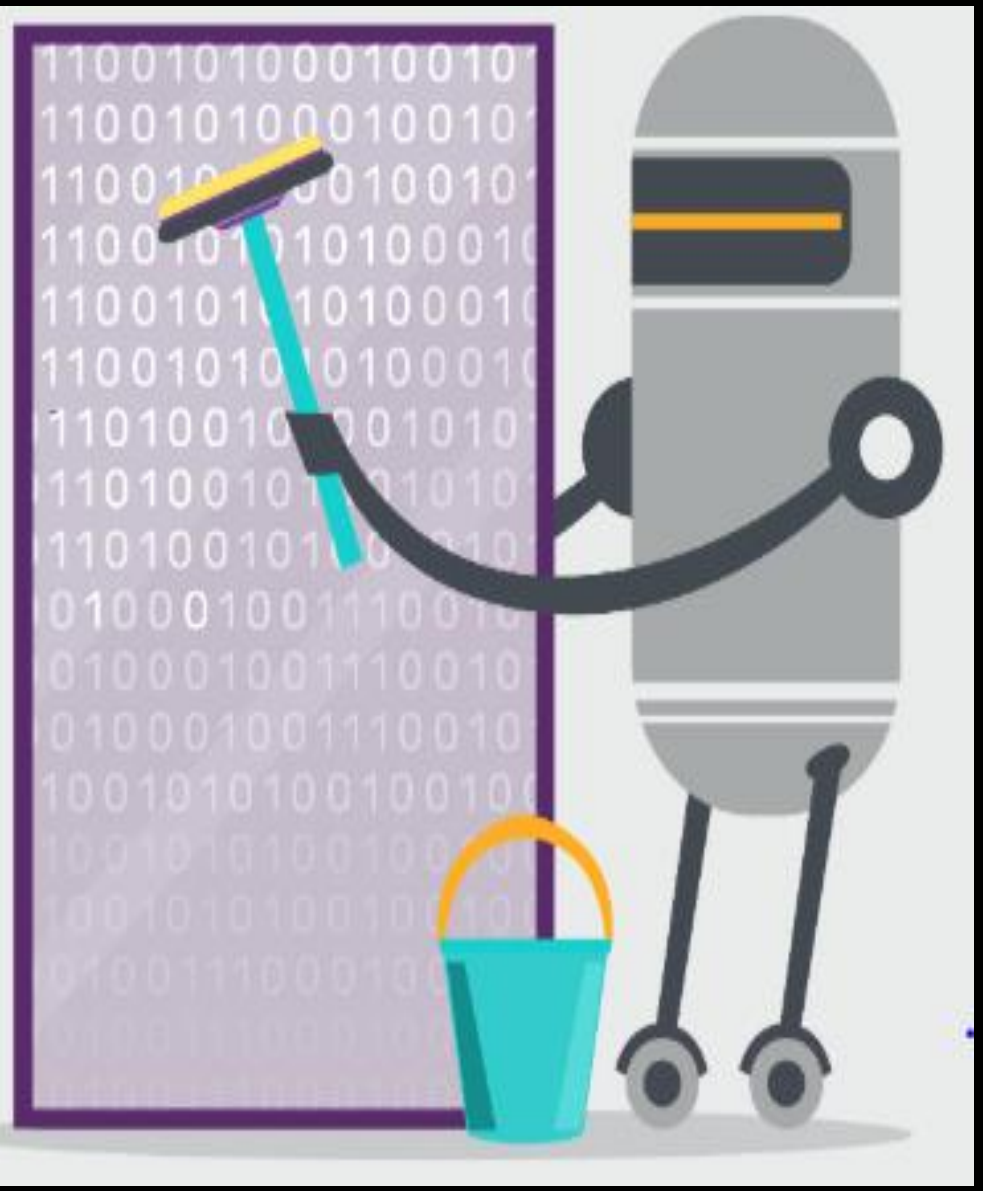
This fig shows the data has 37 columns and 9240 rows

```
# Importing all datasets
leads_data = pd.read_csv("Leads.csv")
leads_data
```

	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	...	Get updates on DM Content	Lead Profile	City	Asymmetrique Activity Index	Asymm Profil
0	7927b2df-8bba-4d29-b9a2-b6e0beafe620	660737	API	Olark Chat	No	No	0	0.0	0	0.00	...	No	Select	Select	02.Medium	02.I
1	2a272436-5132-4136-86fa-dcc88c88f482	660728	API	Organic Search	No	No	0	5.0	674	2.50	...	No	Select	Select	02.Medium	02.I
2	8cc8c611-a219-4f35-ad23-fdfd2656bd8a	660727	Landing Page Submission	Direct Traffic	No	No	1	2.0	1532	2.00	...	No	Potential Lead	Mumbai	02.Medium	
3	0cc2df48-7cf4-4e39-9de9-19797f9b38cc	660719	Landing Page Submission	Direct Traffic	No	No	0	1.0	305	1.00	...	No	Select	Mumbai	02.Medium	
4	3256f628-e534-4826-9d63-4a8b88782852	660681	Landing Page Submission	Google	No	No	1	2.0	1428	1.00	...	No	Select	Mumbai	02.Medium	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
9235	19d6451e-fcd6-407c-b83b-48e1af805ea9	579564	Landing Page Submission	Direct Traffic	Yes	No	1	8.0	1845	2.67	...	No	Potential Lead	Mumbai	02.Medium	
9236	82a7005b-7196-4d56-95ce-a79f937a158d	579546	Landing Page Submission	Direct Traffic	No	No	0	2.0	238	2.00	...	No	Potential Lead	Mumbai	02.Medium	
9237	aac550fe-a586-452d-8d3c-f1b62c94e02c	579545	Landing Page Submission	Direct Traffic	Yes	No	0	2.0	199	2.00	...	No	Potential Lead	Mumbai	02.Medium	
9238	5330a7d1-2f2b-4df4-85d6-64ca2f6b95b9	579538	Landing Page Submission	Google	No	No	1	3.0	499	3.00	...	No	NaN	Other Metro Cities	02.Medium	02.I
9239	571b5c8e-a5b2-4d57-8574-f2ffb06fdeff	579533	Landing Page Submission	Direct Traffic	No	No	1	6.0	1279	3.00	...	No	Potential Lead	Other Cities	02.Medium	

9240 rows × 37 columns

DATA CLEANING



- ‘select ‘ as input converted into null values
- Checking null values in columns and remove columns in which null values are greater than 50%, (3 columns deleted)
- Dropping unnecessary columns i.e. in which only 1 unique input value is given (5 columns deleted)
- Drop those columns which are created by sales department people and no use in analysing data (4 columns deleted)
- Drop highly skewed categorical column i.e. in which only 1 category is present at higher rate than others (9 columns deleted)
- Have 2 unique columns i.e. prospect id and lead number , so we dropped prospect id and worked with lead number .
- Dropped the rows in which null values are higher than 5%.
- Working with those categorical columns in which categories are more than 4 , and also having null values
    1. In these columns null values replaced with median/mode.
    2. And rest of the categories whose count are too less than others combined to make a single category (binning)
- In continuous columns null values are replaced with median.(as it has outliers)

After cleaning data , left with rows: 9233 ,columns:12 → 99.99% data left for analysis

Check correlation (multicollinearity) using heatmap

Exploratory Data Analysis

Univariate Analysis-I (categorical)

Univariate Analysis-II

Bivariate Analysis (Categorical - Categorical Variables) –I

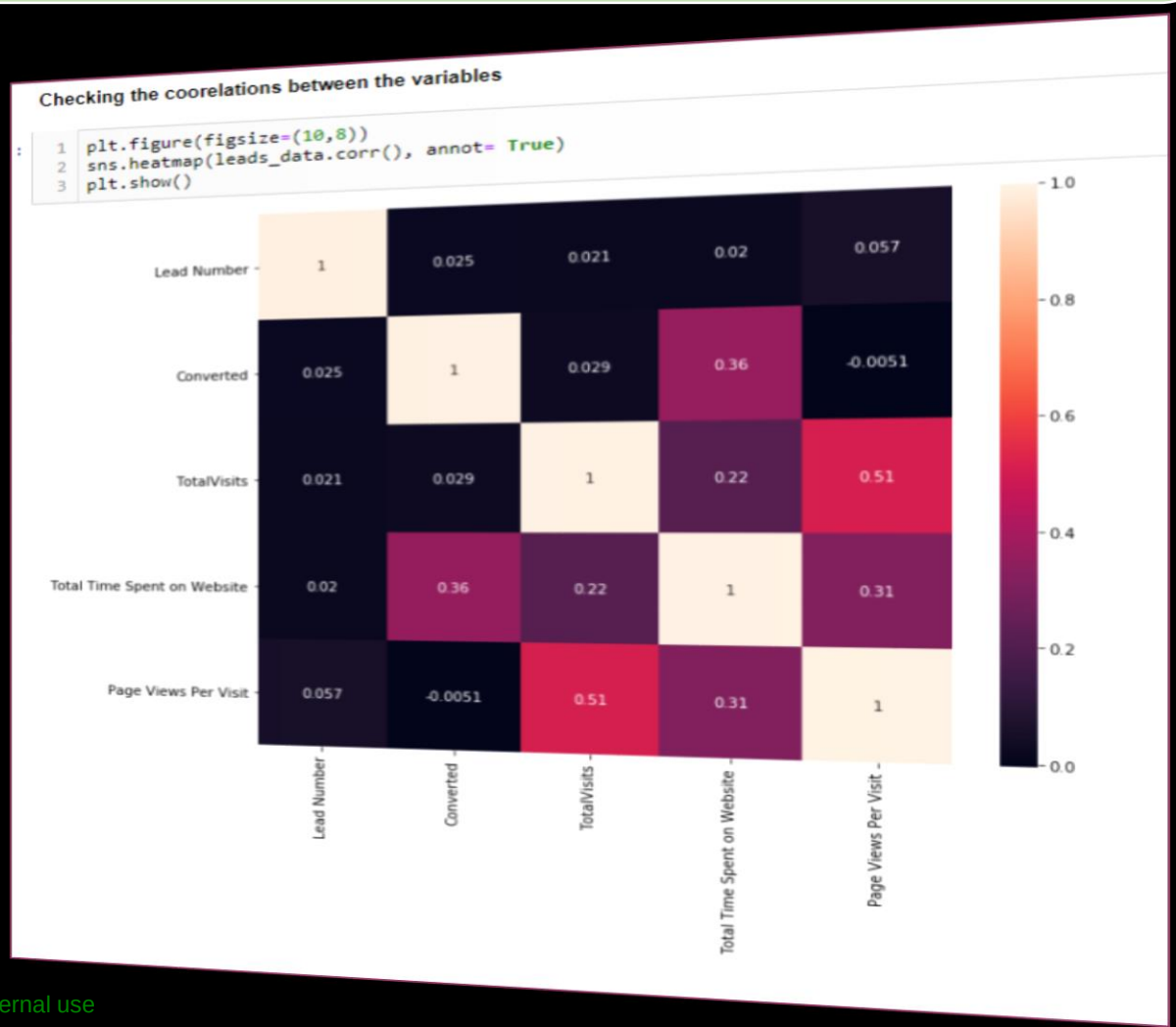
Bivariate Analysis (Categorical - Categorical Variables) – II

Bivariate Analysis (Numerical - Categorical Variables)

Check correlation (multicollinearity) using heatmap

If there is some obvious multicollinearity going on, this is the first place to catch it

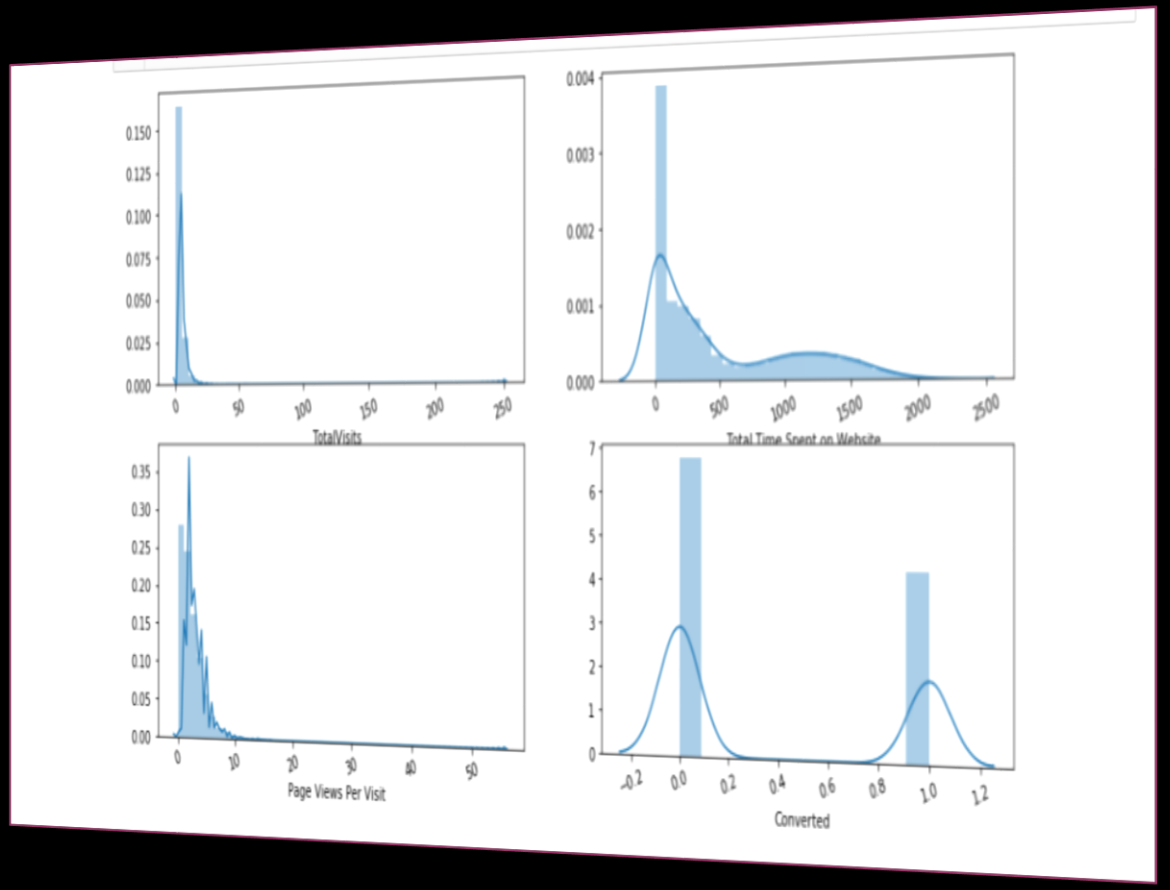
Here's where we will also identify if some predictors directly have a strong association with the outcome variable



Exploratory Data Analysis

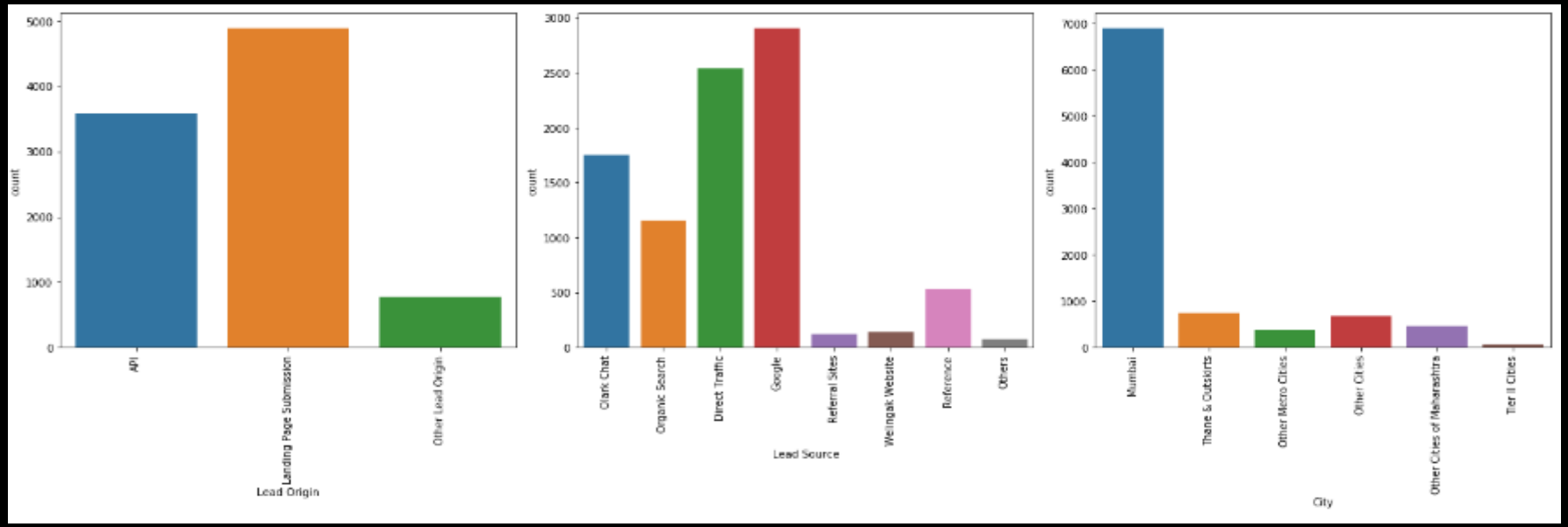
Univariate Analysis

Performing univariate analysis on numerical variables.



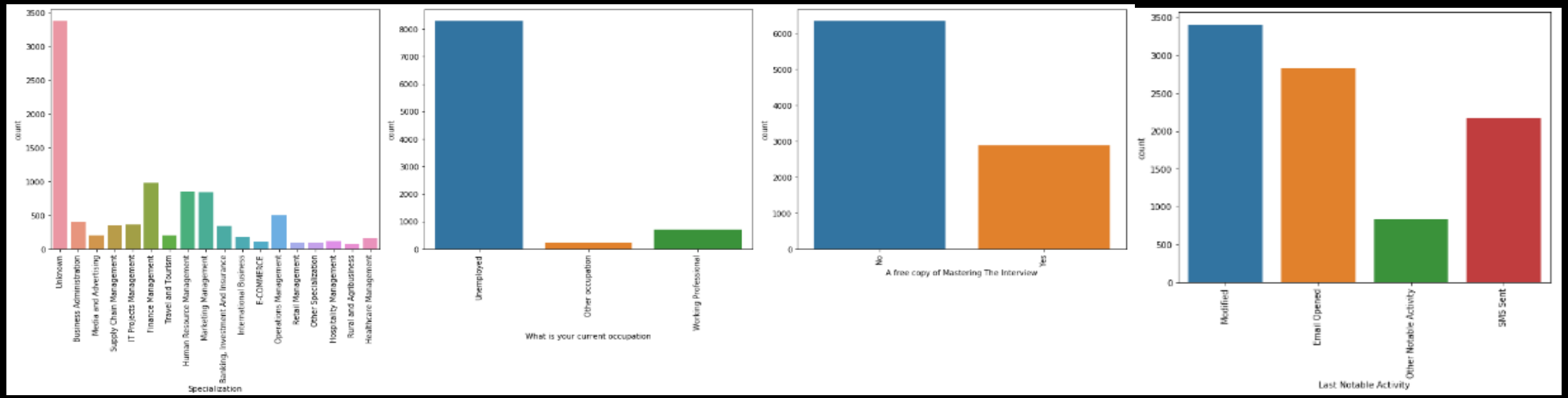
Univariate Analysis-I (categorical)

Lead Origin	• Most of the Leads are originated from API and Landing Page Submission.
Lead Source	• Google, Direct traffic, olark chat and organic search generates maximum number of leads.
City	• Mumbai city generates maximum number of leads



Univariate Analysis-II

Specialization	• Unknown, Finance Management, Human Resource Management, Marketing Management, Operations Management etc. generates maximum number of leads
What is your current occupation	• Unemployed generates maximum number of leads.
A free copy of Mastering The Interview	• Most of the people have selected No.
Last Notable Activity	• Most of the lead have their Email opened, SMS Sent and modified as their last activity



Bivariate Analysis (Categorical- Categorical Variables) –I

Lead Origin

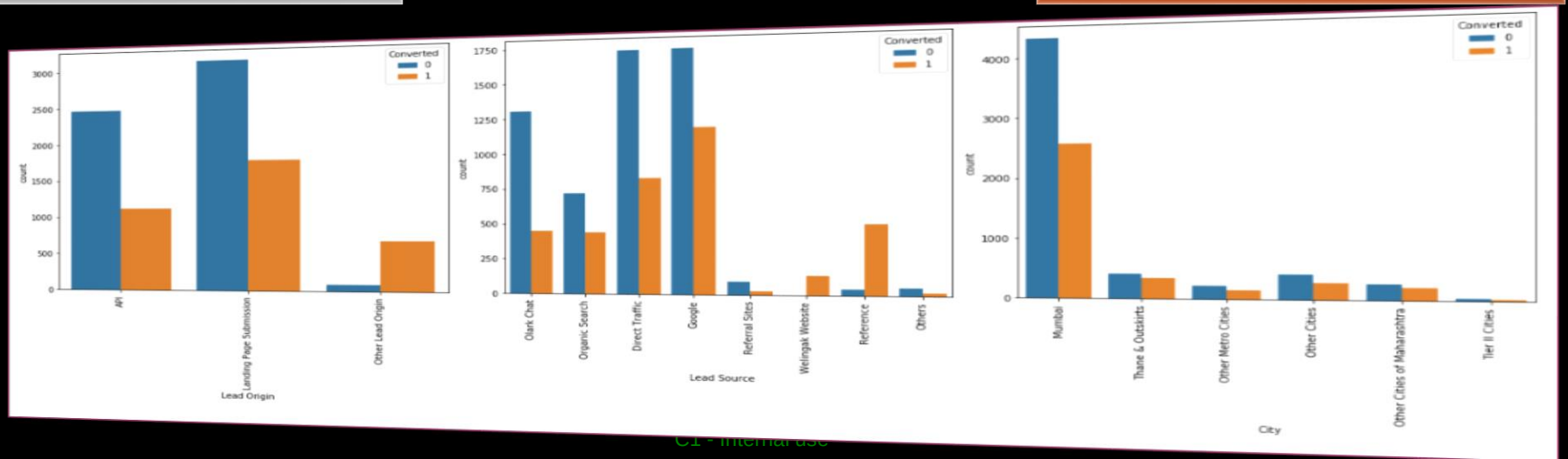
- API and Landing Page Submission have more conversion rate but count of lead originated from them are considerable
- Other Lead Origin has more conversion rate but count of lead are not very high.
- Lead Import are very less in count.
- To improve overall lead conversion rate, we need to focus more on improving lead conversion of API and Landing Page Submission origin and generate more leads from Other Lead Origin.

Lead Source

- Google and Direct traffic generates maximum number of leads.
- Conversion Rate of reference leads and leads through welingak website is high.
- To improve overall lead conversion rate, focus should be on improving lead conversion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.

City

- Mumbai city generates maximum number of leads.
- To improve overall lead conversion rate, focus should be on improving lead conversion of Thane & Outskirts, Other Cities, Other Cities of Maharashtra, Other Metro Cities.



Bivariate Analysis (Categorical- Categorical Variables) –II

Specialization

- Unknown, Finance Management and Marketing Management generates maximum number of leads.
- Healthcare Management conversion rate is high.
- To improve overall lead conversion rate, focus should be on improving lead converion of Human Resource Management, Operations Management, Business Administration etc

What is your current occupation

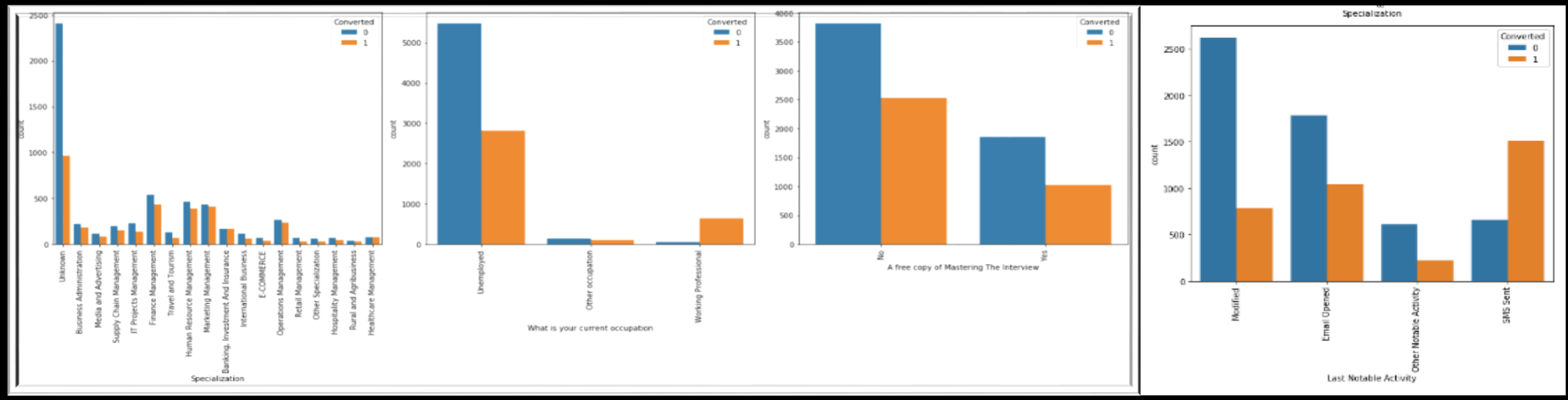
- Unemployed generates maximum number of leads.
- Working professional conversion rate is high.
- To improve overall lead conversion rate, focus should be on improving lead converion of Unemployed, other occupations

A free copy of Mastering The Interview

- Yes generates maximum number of leads
- To improve overall lead conversion rate, focus should be on improving lead converion of Yes and No

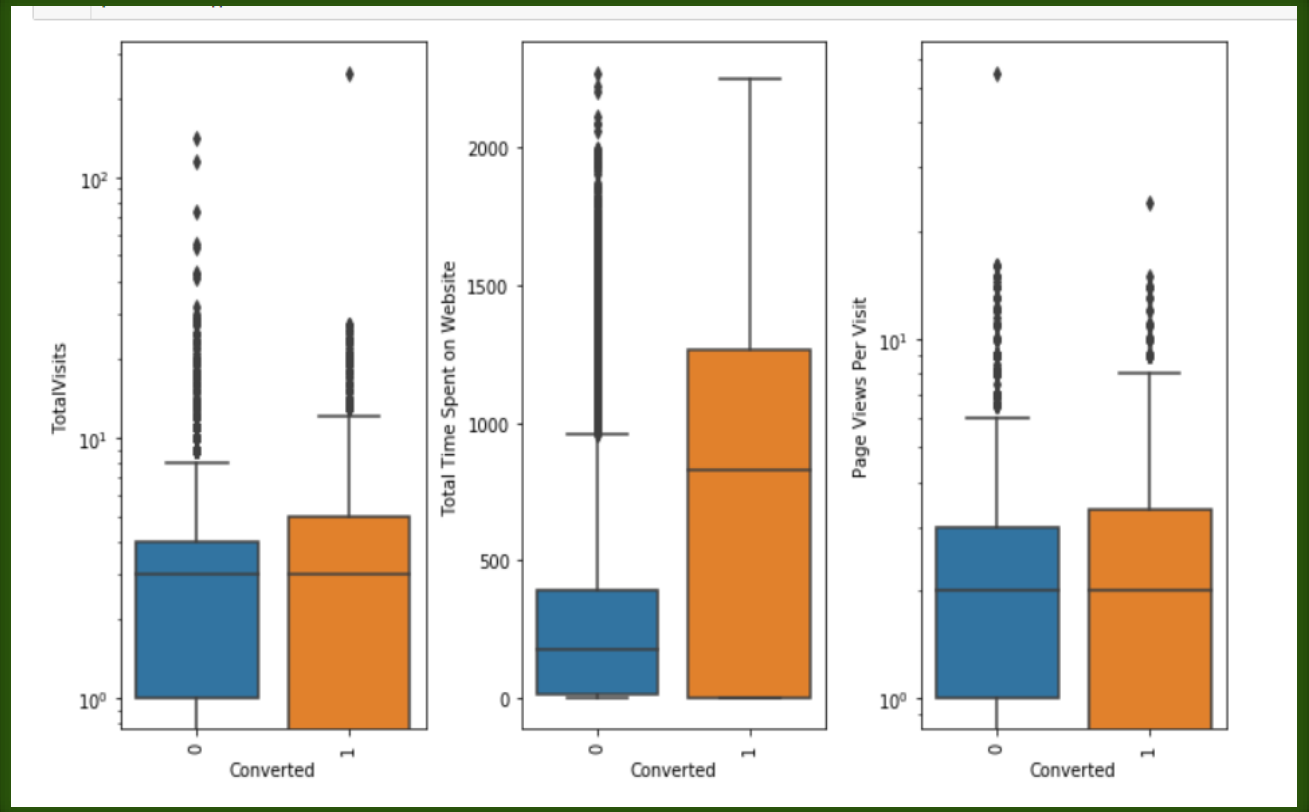
Last Notable Activity

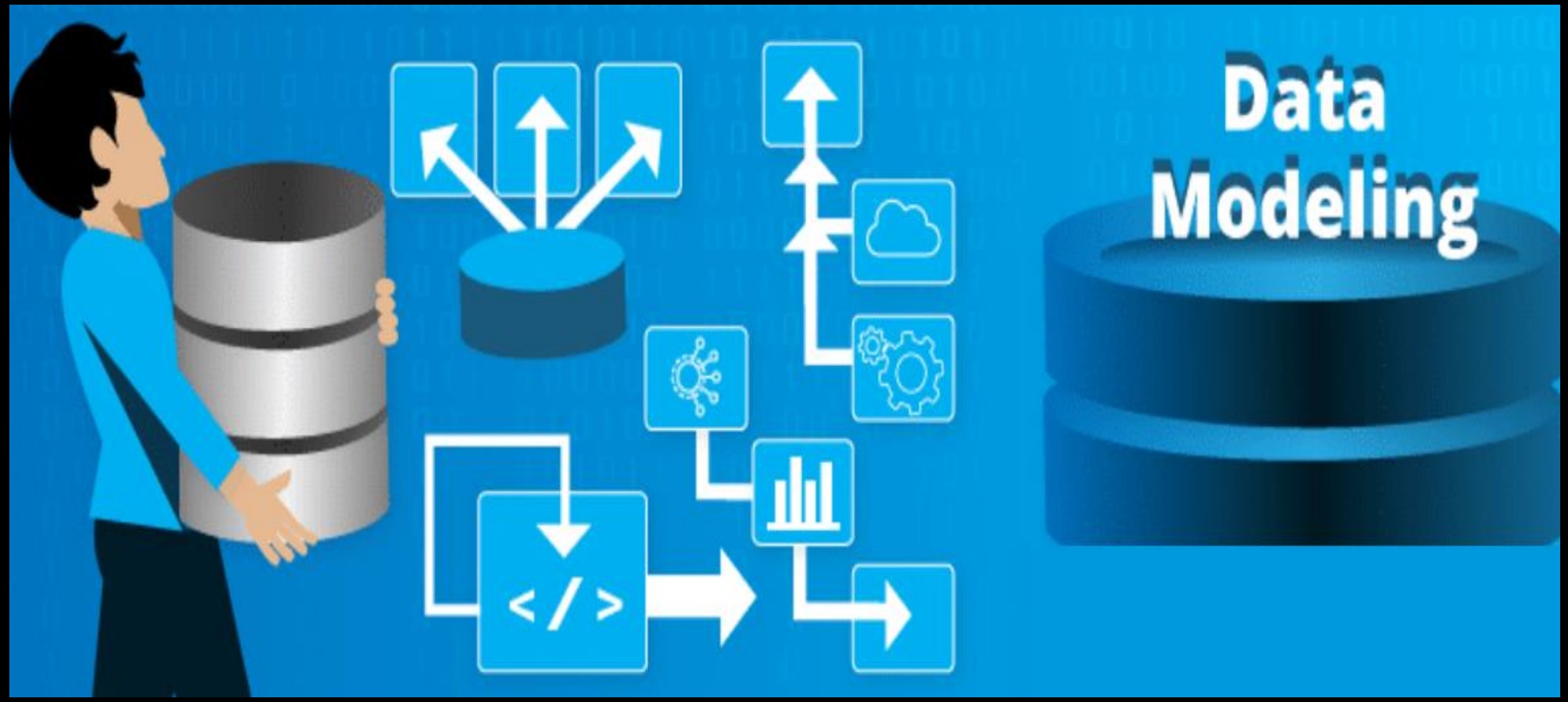
- SMS Sent generates maximum number of leads and conversion rate is also high.
- Most of the lead have their Email opened and modified as their last activity.
- To improve overall lead conversion rate, focus should be on improving lead converion of Email Opened, Modified

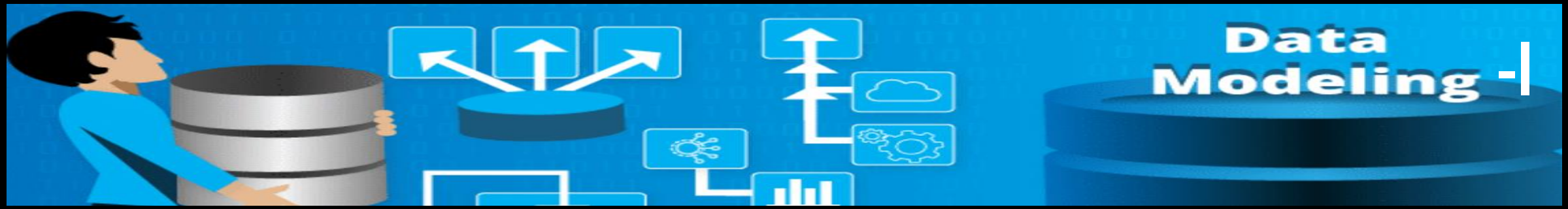


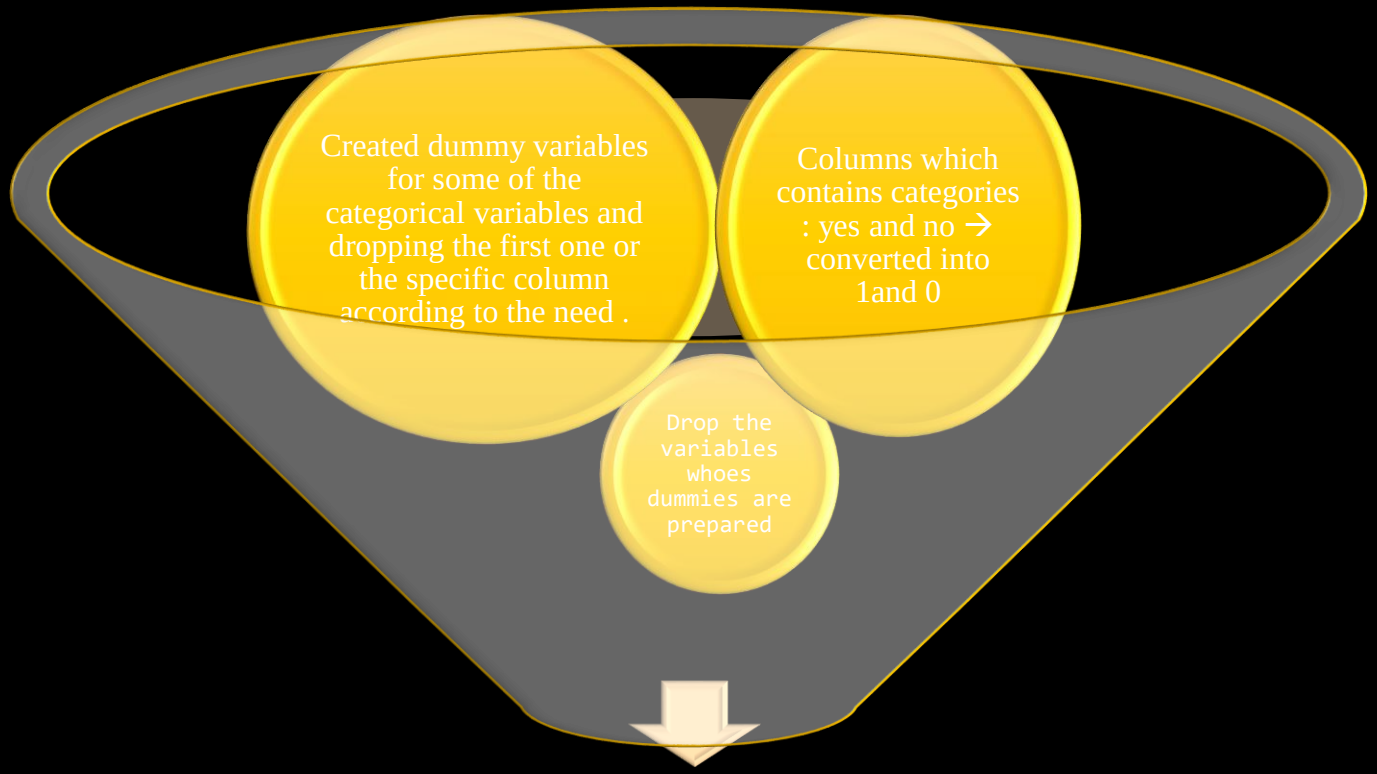
Bivariate Analysis (Numerical - Categorical Variables)

Total Visits	Median for converted and not converted leads are the same	Nothing conclusive can be said on the basis of Total Visits.
Total Time Spent on Website	Leads spending more time on the website are more likely to be converted	Website should be made more engaging to make leads spend more time.
Page Views Per Visit	Median for converted and not converted leads are the same.	Nothing can be said specifically for lead conversion from Page Views Per Visit





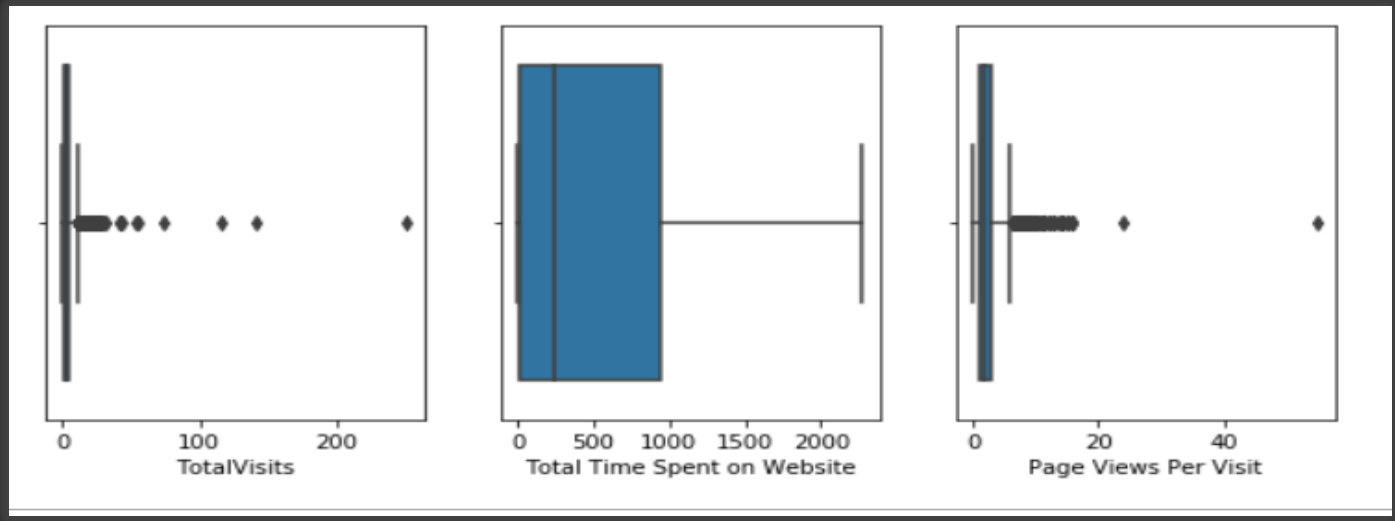




Left with rows : 9233 and columns: 42(all numeric data)

Checking for outliers

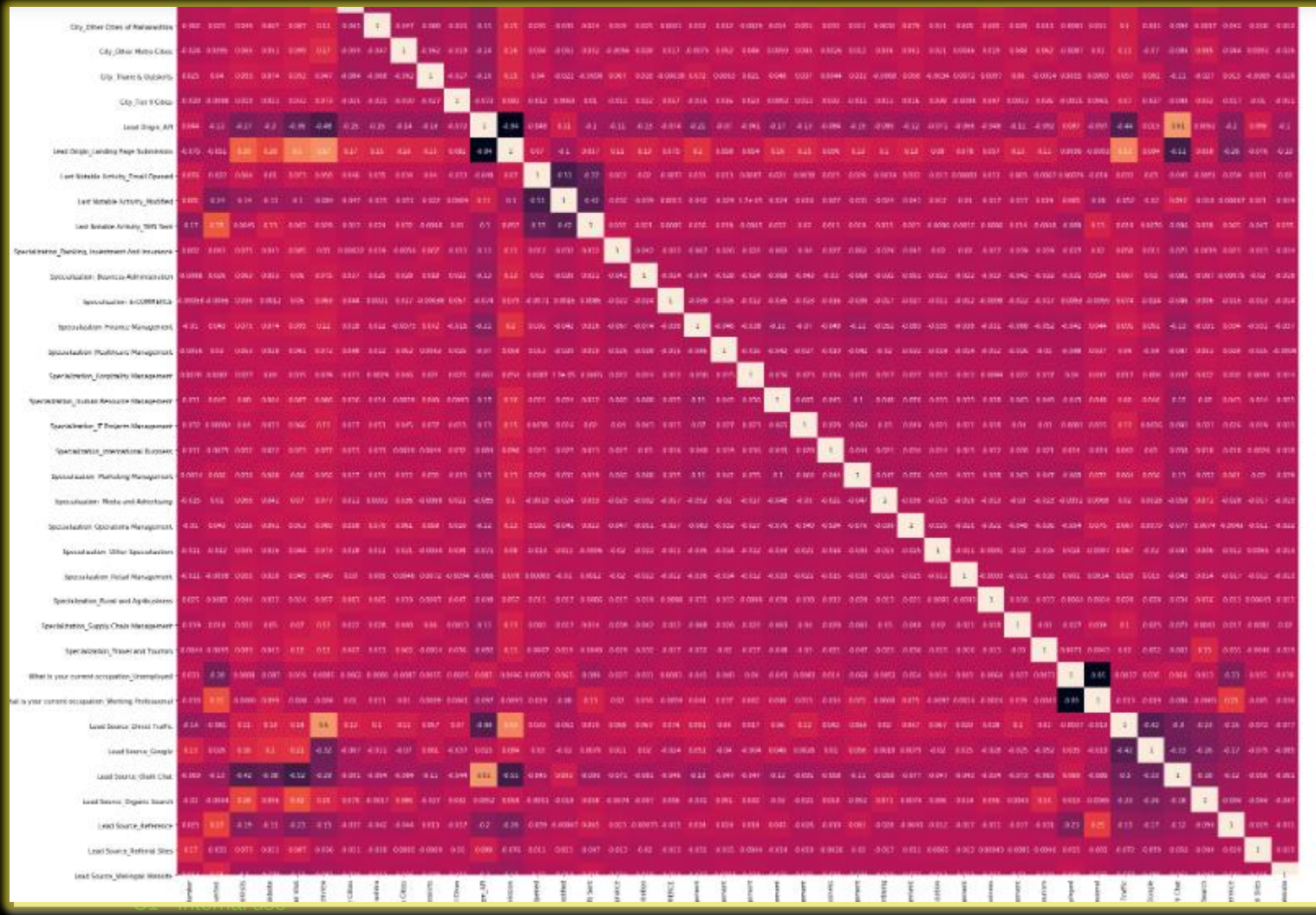
Do outlier treatment using IQR method on “total visits” and “page views per visit” using quantiles(low:0.05 and high: 0.95)

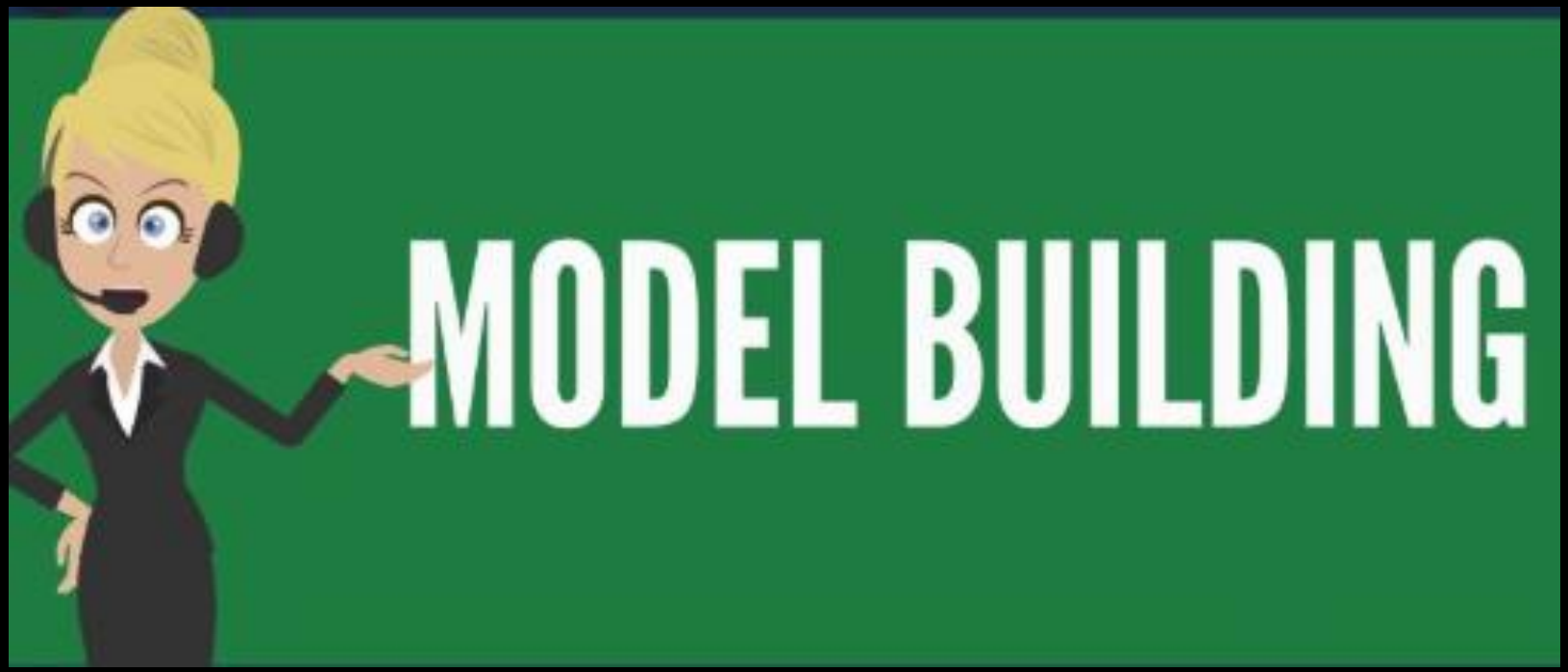


Data distributed in Train dataset (70% of main dataset) and test dataset(30% of main dataset), using train data set model will be generated and then applied on test dataset

Scaling of train dataset using standard -scaler because we need need data scaled in between 1-0 only

Drop highly correlated columns after seeing the Correlation of remaining column.





Using logistic regression with binomial families applied on training data set.

Our latest model have the following features:

- All variables have p-value < 0.05.
- All the features have very low VIF values, meaning, there is hardly any multicollinearity among the features. This is also evident from the heat map.
- The overall accuracy of 80.43 at a probability threshold of 0.05 is also very acceptable.

Use logistic regression model with RFE technique , so that top 15 columns which are best suited for model will be selected by RFE itself

Later on using P-value and VIF values , drop those columns which are not required

Repeat these steps till we get the best top 10 columns for the model and also accuracy of model is also good

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6457
Model:	GLM	Df Residuals:	6446
Model Family:	Binomial	Df Model:	10
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2844.9
Date:	Sun, 31 May 2020	Deviance:	5689.8
Time:	13:09:43	Pearson chi2:	6.78e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.9250	0.084	-11.013	0.000	-1.090	-0.760
Total Time Spent on Website	1.0908	0.038	28.506	0.000	1.016	1.166
Last Notable Activity_Email Opened	0.7783	0.078	9.983	0.000	0.626	0.931
Last Notable Activity_SMS Sent	2.0382	0.086	23.764	0.000	1.870	2.206
What is your current occupation_Working Professional	2.7388	0.185	14.815	0.000	2.376	3.101
Lead Source_Direct Traffic	-1.1945	0.106	-11.250	0.000	-1.403	-0.986
Lead Source_Google	-0.7780	0.101	-7.676	0.000	-0.977	-0.579
Lead Source_Organic Search	-0.9508	0.125	-7.625	0.000	-1.195	-0.706
Lead Source_Reference	2.8397	0.207	13.722	0.000	2.434	3.245
Lead Source_Referral Sites	-1.1019	0.332	-3.320	0.001	-1.752	-0.451
Lead Source_Welingak Website	4.8397	0.724	6.688	0.000	3.421	6.258

	Features	VIF
1	Last Notable Activity_Email Opened	1.53
2	Last Notable Activity_SMS Sent	1.48
5	Lead Source_Google	1.41
4	Lead Source_Direct Traffic	1.38
3	What is your current occupation_Working Profes...	1.18
6	Lead Source_Organic Search	1.17
7	Lead Source_Reference	1.17
0	Total Time Spent on Website	1.09
9	Lead Source_Welingak Website	1.03
8	Lead Source_Referral Sites	1.01

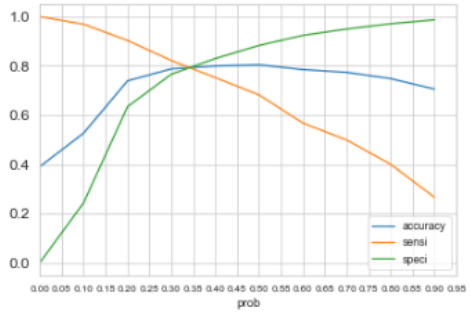
Finding Optimal Cutoff Point

Plotted the ROC Curve : The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test. So by seeing our ROC curve we can say that our test is accurate.

Calculating the area under the curve(GINI): Since we got a value of 0.86, our model seems to be doing well on the test dataset.

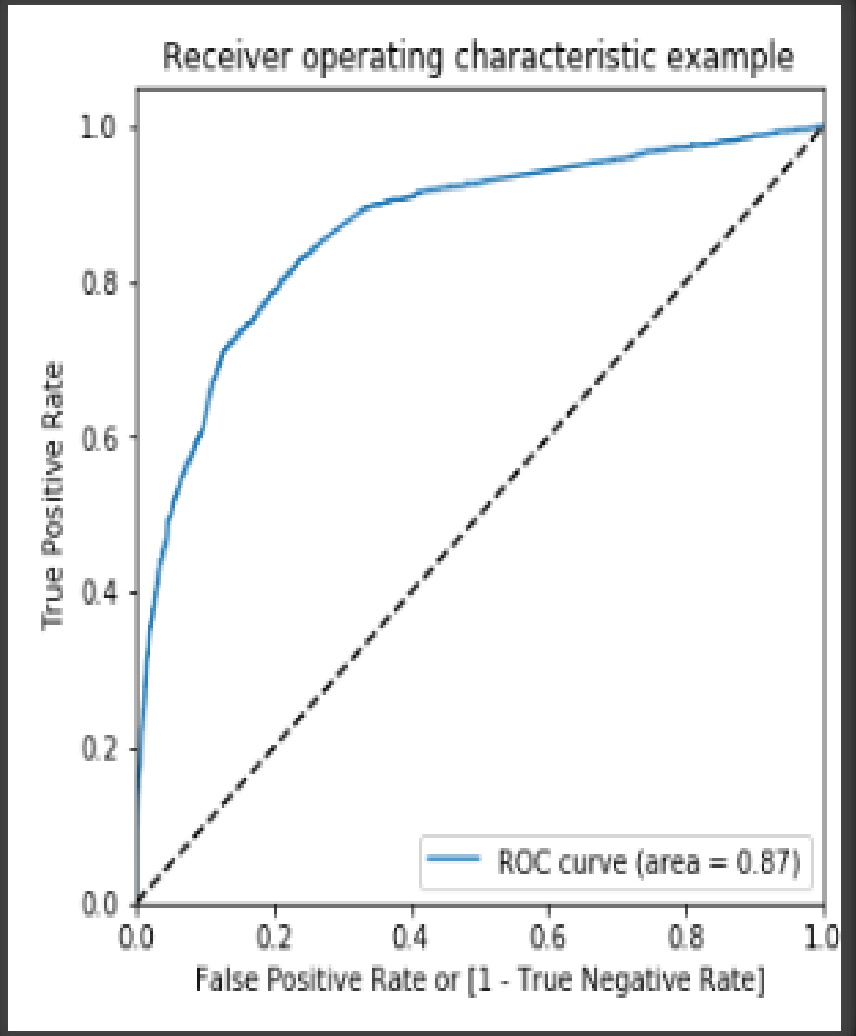
Finding Optimal Cutoff Point: (So that we get our sensitivity more than 80%) From the below curve , we selected 0.33 as the optimum point to take it as a cutoff probability

```
# Let's plot accuracy sensitivity and specificity for various probabilities.
sns.set_style("whitegrid")
sns.set_context("paper")
cutoff_df.plot.line(x='prob', y=['accuracy','sensi','speci'])
plt.xticks(np.arange(0, 1, step=0.05), size = 8)
plt.yticks(size = 12)
plt.show()
```



From the curve above, 0.33 is the optimum point to take it as a cutoff probability.

	prob	accuracy	sensi	speci
0.0	0.0	0.390274	1.000000	0.000000
0.1	0.1	0.526715	0.969048	0.243586
0.2	0.2	0.739043	0.903175	0.633985
0.3	0.3	0.787672	0.821429	0.766066
0.4	0.4	0.799442	0.752381	0.829566
0.5	0.5	0.804398	0.681746	0.882906
0.6	0.6	0.784730	0.567460	0.923800
0.7	0.7	0.773579	0.498413	0.949708
0.8	0.8	0.748180	0.400000	0.971044
0.9	0.9	0.705591	0.265873	0.987046



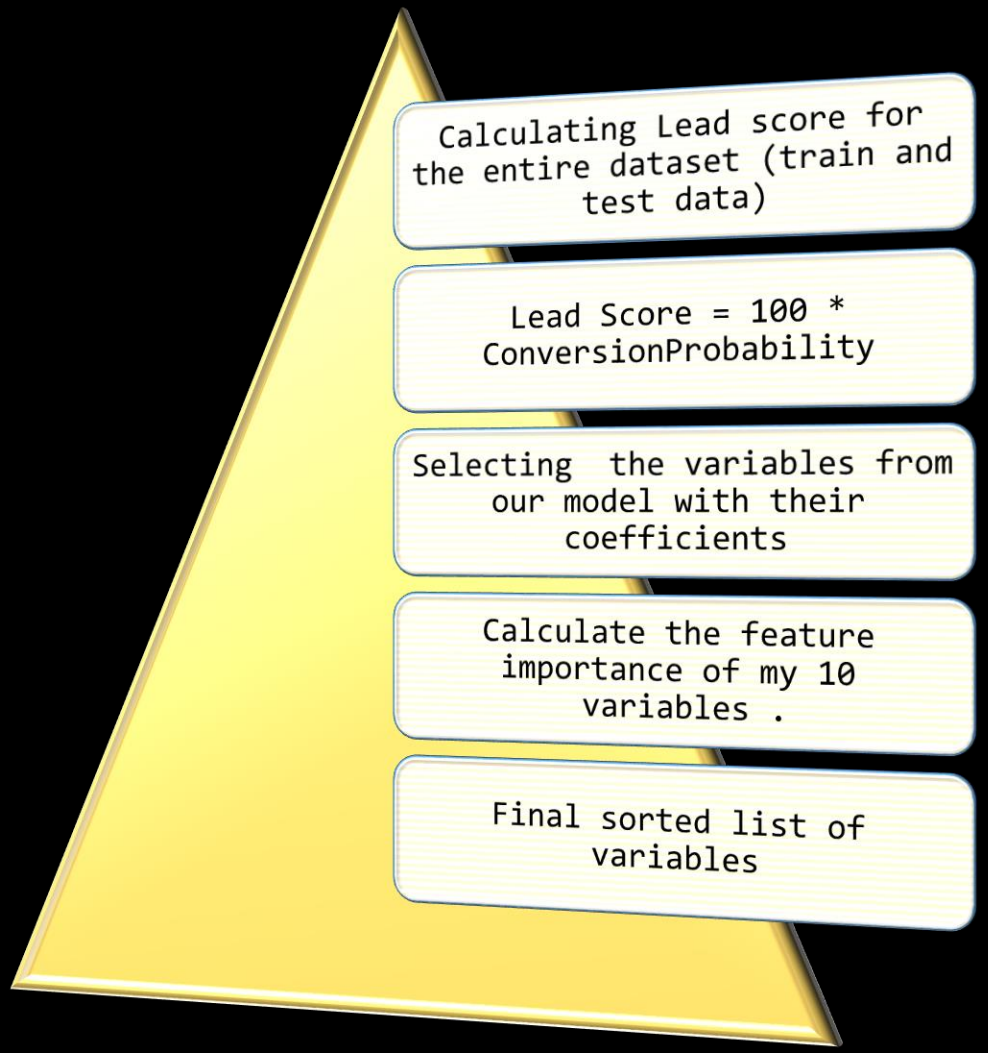
Calculated Metrics beyond simply accuracy of TRAIN data

METRICES	DEFINATION	Percentage (%)
1.Sensitivity	The proportion of observed converted leads that were predicted to be converted	**80.59**
2. Specificity	The proportion of observed non-converted leads that were predicted to be non-converted	**78.21**
3. False positive rate:	The proportion of observed non-converted leads are predicted as converted leads	**21.78**
4.Positive predictive value	The converted lead predictive value tells you how often a converted lead test represents a true converted lead	**70.30**
5. Negative predictive value	The non- converted predictive value tells you how often a non-converted test represents a true non-converted lead.	**86.3**
6. Precision	Positive predictive value	**70.30**
7. Recall	Sensitivity	**80.59**
8. F1 Score	is a measure of a test's accuracy using precision and recall	**75.1**

Calculated Metrics beyond simply accuracy of TEST data

METRICES	DEFINATION	Percentage (%)
1.Sensitivity	The proportion of observed converted leads that were predicted to be converted	**83.41**
2. Specificity	The proportion of observed non-converted leads that were predicted to be non-converted	**79.26**
3. False positive rate:	The proportion of observed non-converted leads are predicted as converted leads	**20.73**
4.Positive predictive value	The converted lead predictive value tells you how often a converted lead test represents a true converted lead	**70.66**
5. Negative predictive value	The non- converted predictive value tells you how often a non-converted test represents a true non-converted lead.	**88.86**
6. Precision	Positive predictive value	**70.66**
7. Recall	Sensitivity	**83.41**
8. F1 Score	is a measure of a test's accuracy using precision and recall	**76.51**

LEAD SCORE




```
# Calculating the Lead Score value
# Lead Score = 100 * Conversion_Prob
lead_full_pred['Lead_Score'] = lead_full_pred['Conversion_Prob'].apply(lambda x : round(x*100))
lead_full_pred.head()
```

	LeadID	Converted	Conversion_Prob	final_predicted	Lead_Score
0	1052	1	0.730041	1	73
1	8706	0	0.130804	0	13
2	4876	0	0.051798	0	5
3	6157	1	0.951349	1	95
4	5217	1	0.237318	0	24

Selecting Top 10 features which contribute most towards the probability of a lead getting converted

```
feature_importance = new_params
pd.DataFrame(feature_importance).reset_index().sort_values(by=0,ascending=False).head(10)
```

	index	0
9	Lead Source_Welingak Website	4.84
7	Lead Source_Reference	2.84
3	What is your current occupation_Working Profes...	2.74
2	Last Notable Activity_SMS Sent	2.04
0	Total Time Spent on Website	1.09
1	Last Notable Activity_Email Opened	0.78
5	Lead Source_Google	-0.78
6	Lead Source_Organic Search	-0.95
8	Lead Source_Referral Sites	-1.10
4	Lead Source_Direct Traffic	-1.19

Selecting Top 10

Features which contribute most towards the probability of a lead getting converted

Selecting Top 10 features which contribute most towards the probability of a lead getting converted

```
feature_importance = new_params
pd.DataFrame(feature_importance).reset_index().sort_values(by=0,ascending=False).head(10)
```

	index	0
9	Lead Source_Welingak Website	4.84
7	Lead Source_Reference	2.84
3	What is your current occupation_Working Profes...	2.74
2	Last Notable Activity_SMS Sent	2.04
0	Total Time Spent on Website	1.09
1	Last Notable Activity_Email Opened	0.78
5	Lead Source_Google	-0.78
6	Lead Source_Organic Search	-0.95
8	Lead Source_Referral Sites	-1.10
4	Lead Source_Direct Traffic	-1.19

Based on our model, some features are identified which contribute most to a Lead getting converted successfully.

The conversion probability of a lead increases with increase in values of the following features in descending order:

The conversion probability of a lead increases with decrease in values of the following features in descending order:

Features with Positive Coefficient:

1) Lead Source_Welingak Website: 4.84
2) Lead Source_Reference: 2.84
3) What is your current occupation_Working Professional: 2.74
4) Last Notable Activity_SMS Sent: 2.04
5) Total Time Spent on Website: 1.09
6) Last Notable Activity_Email Opened: 0.78

Features with Negative Coefficient:

1) Lead Source_Google: -0.78
2) Lead Source_Organic Search: -0.95
3) Lead Source_Referral Sites: -1.10
4) Lead Source_Direct Traffic: -1.19
1) Lead Source_Google -16.08

Metrices	Train data set	Test data set
Sensitivity	80.59	83.41
Specificity	78.21	79.26
Accuracy	80.44	80.81
F1- Score	75.1	76.51

The first two features are from Lead Source category. So we need to pay more attention on leads coming from Welingak Website and Reference.

We need to start promoting more on these websites of Wenlingak Website by popping adds with offers to tempt the customers

We need to call people who have enrolled to our courses and offer them certain discounts in their fee if they provide references to others who turn out Hot Leads.

We need to focus more on person those are working as they have money and esteem to grow in their current profile so they can be our Hot leads.

We can move into the fourth variable SMS sent back by student. We need to monitor who all sent back SMS response and let our sales team call them and provide offers and convert them into Hot Leads.

The next variable being time spent by the clients on website. More the time they spend more the chances of them trying to understand the course and more the chance of becoming a Hot Lead. So we find out who spends most time and let the sales team call them and give them some discounts or offers.

The next variable is Last Notable Activity was Email Opened, sales team can forward some nurturing emails regarding our courses and attractive offers to join the courses to people those who spend more time in reading the emails so that they can be converted to Hot leads.

By following all the above mentioned points we have a higher chances of achieving conversion rate of Leads close to 80%.

X- Education company:- For increasing lead conversion rate (more than 80%) target the potential leads whose leads score are in the range of 60-100(higher score would mean that the lead is hot, i.e. is most likely to convert) and also try to get leads from the top 10 features which are selected by the model as they contribute most to a Lead getting converted successfully.

Recommendations

1. <u>If the company already reached its target for a quarter before the deadline.</u>
Since the company already reached its target for a quarter before the deadline. The company can tune the model in such a way that they get moderate sensitivity and high specificity. So that they don’t need to make unnecessary call.

<u>2.X Education has a period of 2 months every year during which they hire some interns. If company want to make the lead conversion more aggressive.</u>
The company can tune the model in such a way that they get high sensitivity and moderate specificity. Since there are more interns the company can make more phone calls.

Thank you