LEAD SCORING CASE STUDY

USING – LOGISTIC REGRESSION

Presented By:

Satya Iyer.
Saurav Dilip Jagtap
Saurav Negi

Problem Statement

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.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. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

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. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.

If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone. A typical lead conversion process can be represented using the following funnel:

Our Business Goal

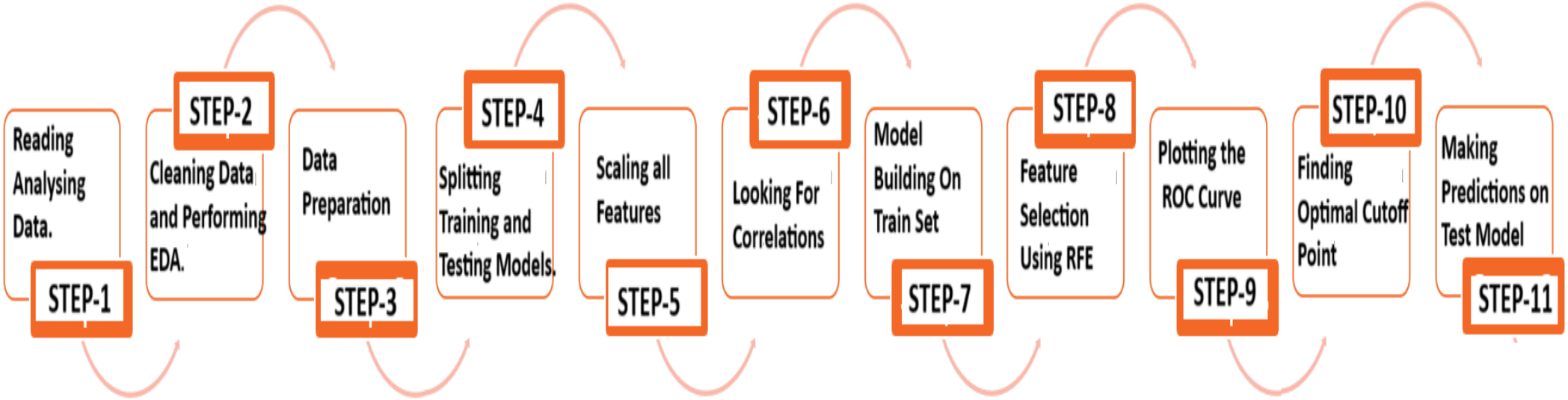
Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads

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.

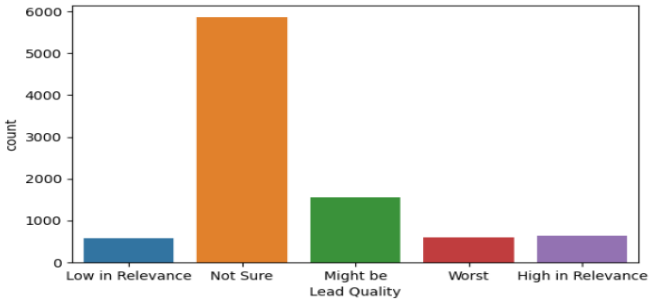
We are building a model where model conversion rate is above 80%.

Strategy


Reading Analysing Data.
STEP-1
STEP-2
Cleaning Data and Performing EDA.
Data Preparation
STEP-3
STEP-4
Splitting Training and Testing Models.
Scaling all Features
STEP-5
STEP-6
Looking For Correlations
Model Building On Train Set
STEP-7
STEP-8
Feature Selection Using RFE
Plotting the ROC Curve
STEP-9
STEP-10
Finding Optimal Cutoff Point
Making Predictions on Test Model
STEP-11

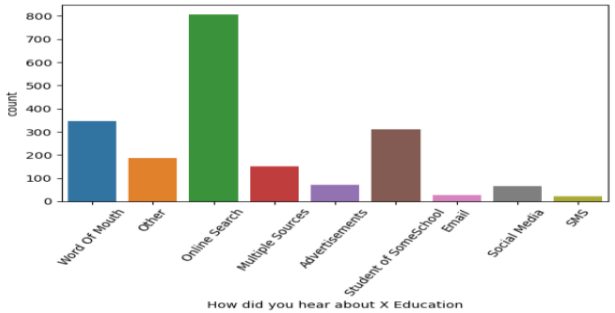
Exploratory Data Analysis(Univariate Analysis)

1.Lead_Quality:



Maximum Lead is **Not sure wheather to be converted or not.**

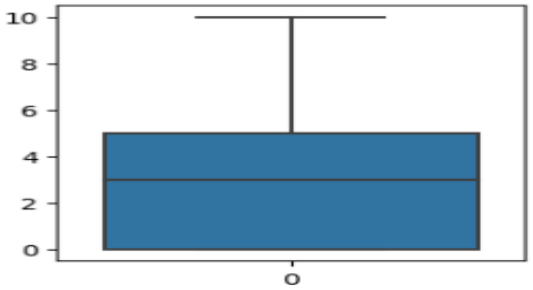
2 Hear about X Education :



Online Search Platform ,**Word of Mouth and Advertisement** provide a good Leads.

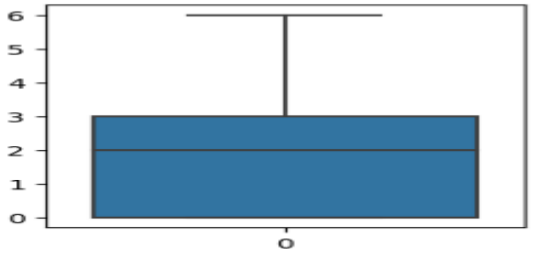
.

3.Total Visits :



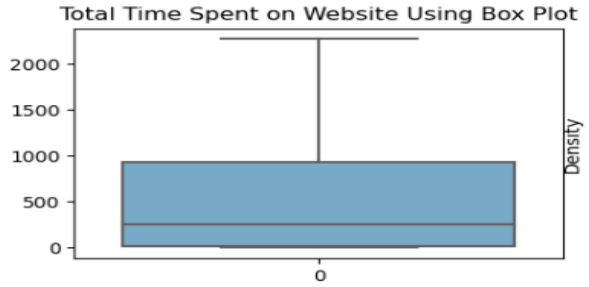
Number of Lead's **TotalVisits** is same that is lies in median in both the cases if a lead may converted or not.

4. Page Views Per Visit:



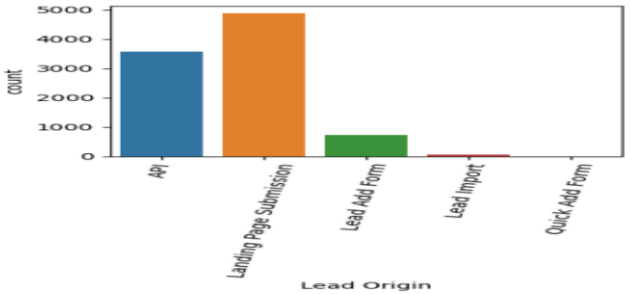
Leads revisiting the **website** are more likely to be converted.

5.Total Time spent on Website:



Leads spending more time on the **Website** are more likely to be converted.

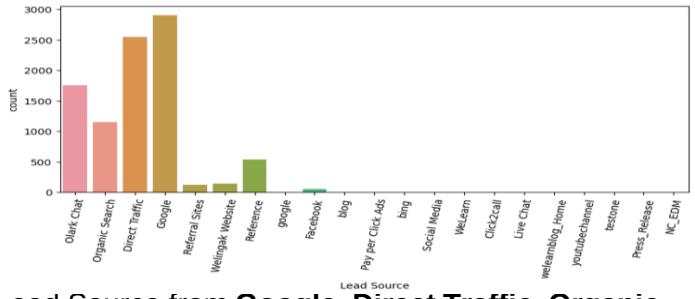
6. Lead Origin:



API and Landing Page Submission have 40-50% conversion rate.

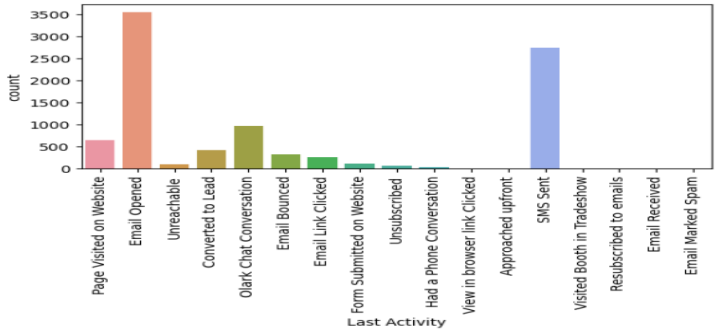
Exploratory Data Analysis(Univariate Analysis)Contd..

7. Lead_Source:



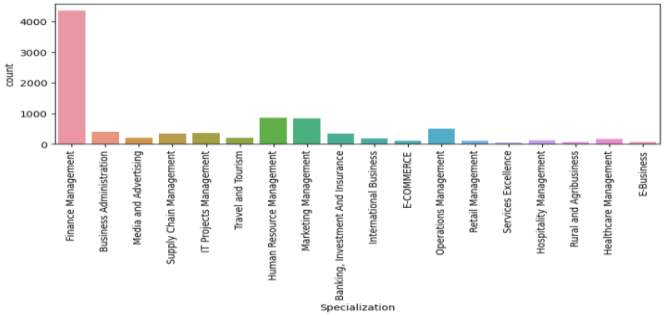
Lead Source from **Google ,Direct Traffic, Organic Search, Olark Chat, Reference** has a very high chance of getting converted.

8. Last Activity :



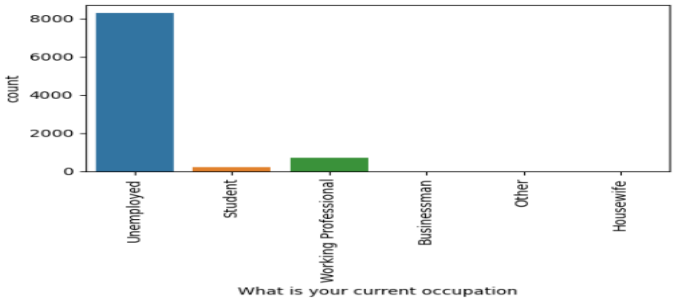
Most of the lead have their **Email opened and SMS Sent** is most frequent Last activity by many leads.

9.Specialization:



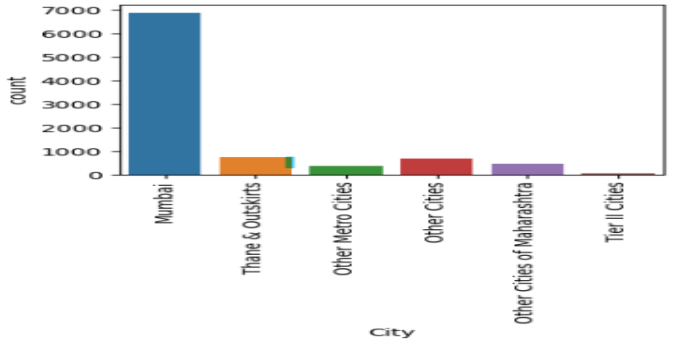
Finance Management specialization are most likely to be converted than others..

10.What is your current occupation:



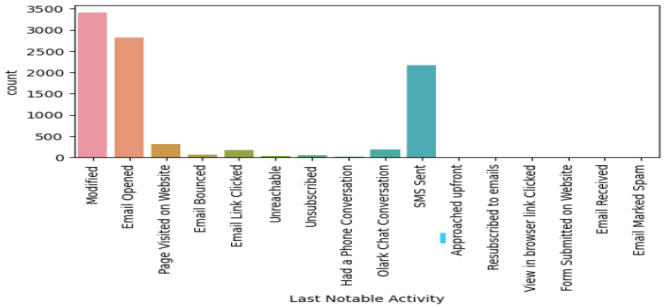
Unemployed and Working Professionals have high chances of becoming a lead.

11.City:



From **Mumbai City** we can get a majority of Leads..

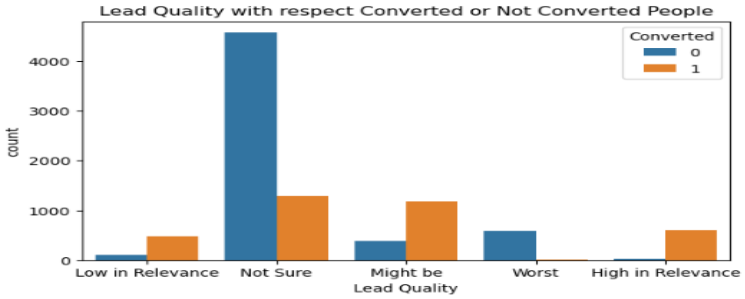
12. Last notable Activity:



Last Notable Activity is Email Openend ,SMS Sent,or Modified and they also get converted.

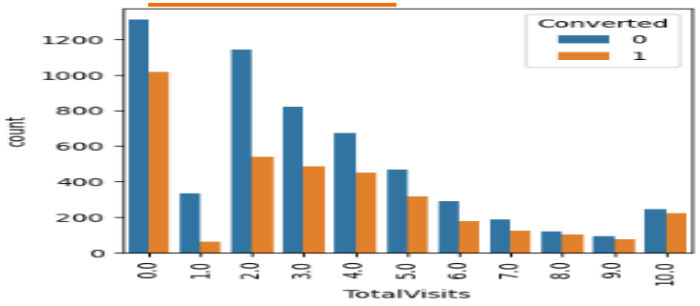
Exploratory Data Analysis(Bivariate Analysis)

1.Lead Quality:



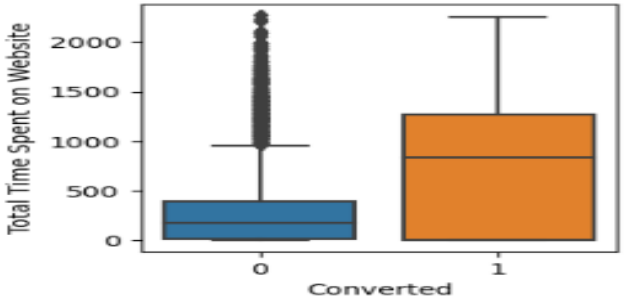
Lead is **Not sure** whether they will convert or not. But Lead with **High in Relevance** is majorly get converted

3.Total Visits :



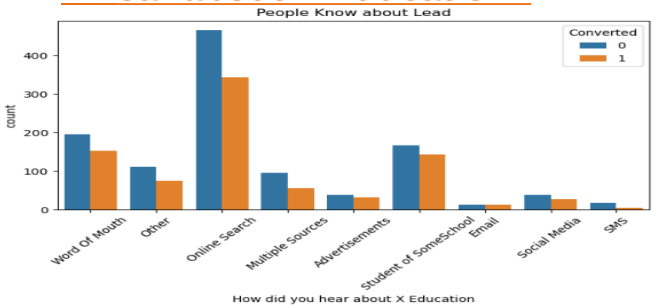
Number of Lead's **TotalVisits** is same that is lies in median in both the cases if a lead is converted or not.

5.Total Time spent on Website:



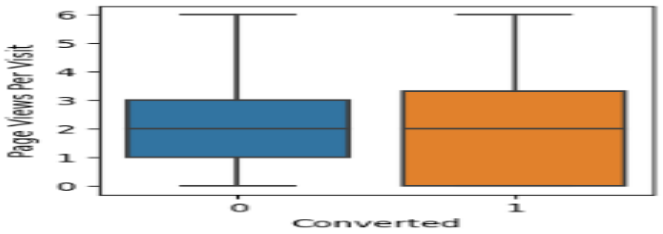
Leads spending more time on the **Website** are more likely to be converted.

2 Hear about X Education :



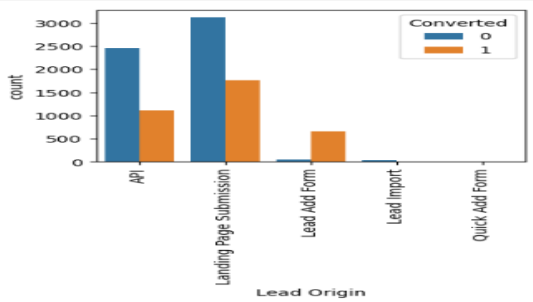
Online Search Platform and there chances of getting converted is also very high. **Word of Mouth and Advertisement** also provide a good Leads.

4. Page Views Per Visit:



Leads revisiting the **website** are more likely to be converted.

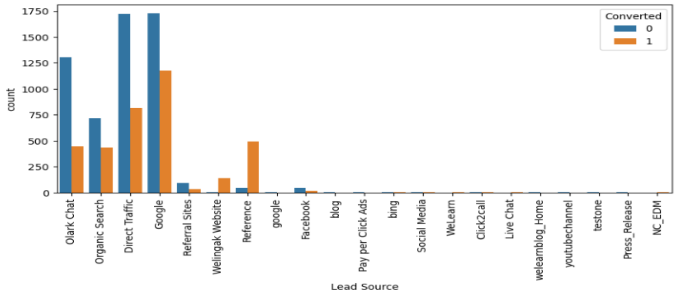
6. Lead Origin:



API and Landing Page Submission have 40-50% conversion rate.

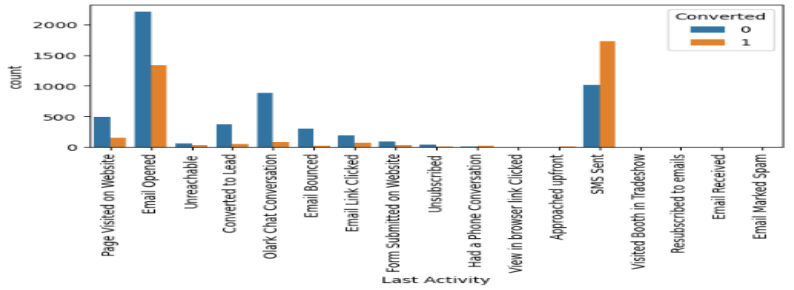
Exploratory Data Analysis(Bivariate Analysis)Contd..

7. Lead_Source:



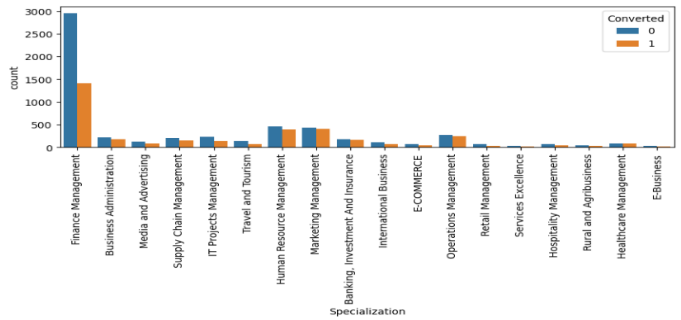
Lead Source from **Google** has a very high chance of getting converted.

8. Last Activity :



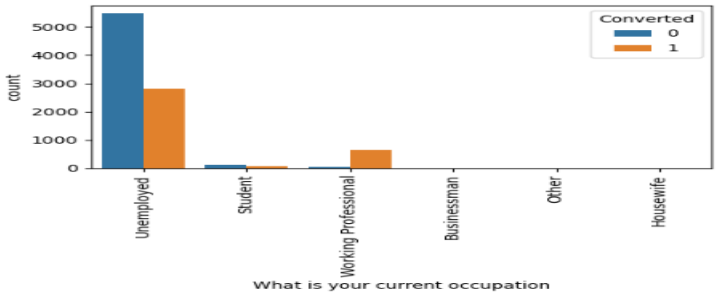
Most of the lead have their **Email opened** as their last activity. Conversion rate for last activity as **SMS Sent** is almost 65%..

9.Specialization:



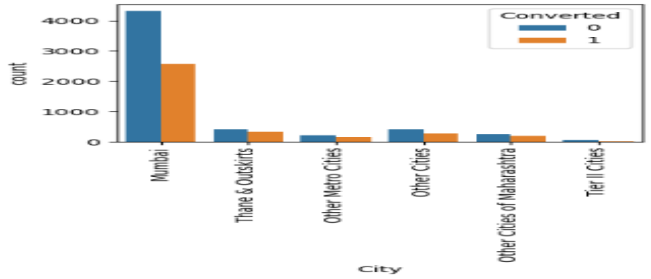
Finance Management specialization are most likely to be converted than others..

10.What is your current occupation:



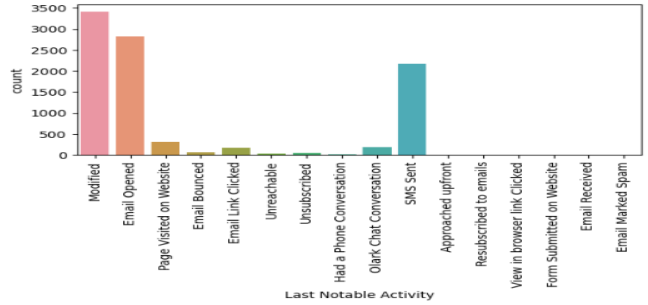
Working Professionals have high chances of joining it. Unemployed leads are the most in numbers but has around 30-35% conversion rate..

11.City:



From **Mumbai City** we can get a majority of Leads.Overall through Maharashtra we can have a good chance of getting Leads .

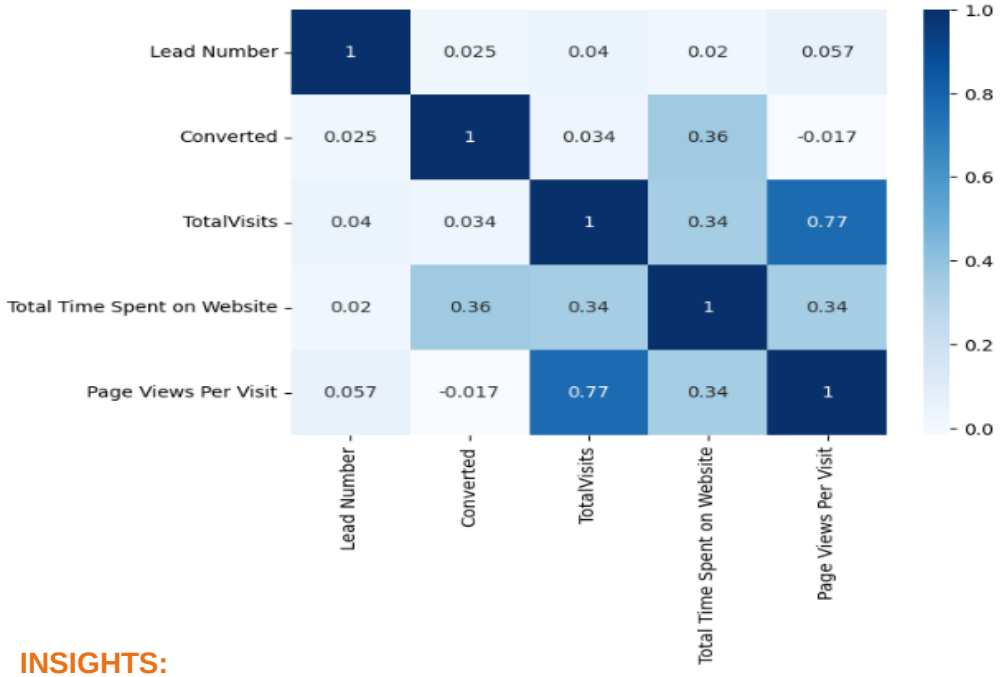
12. Last notable Activity:



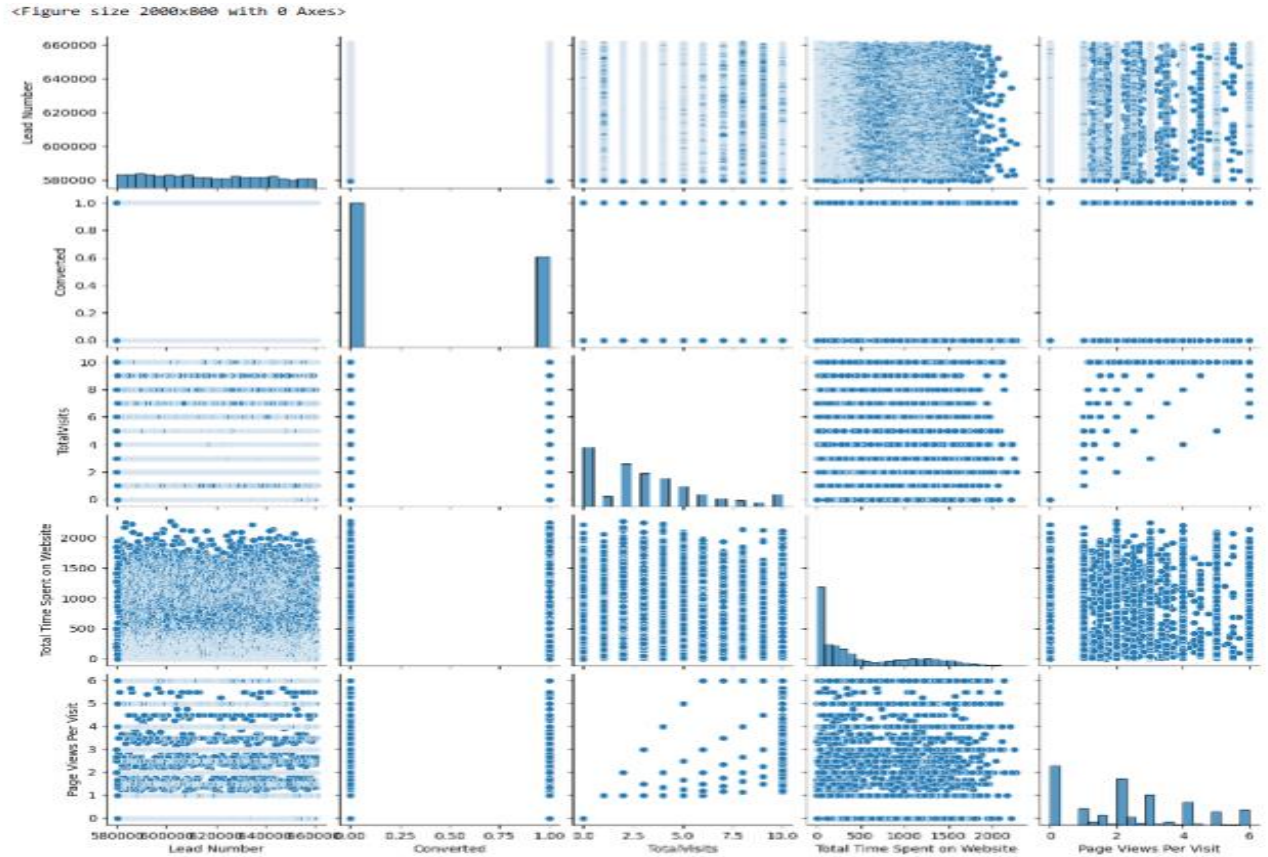
Last Notable Activity is Email Openend ,SMS Sent,or Modified and they also get converted.

EDA(Multivariate Analysis)

1. HEAT MAP



2.PAIR PLOT:

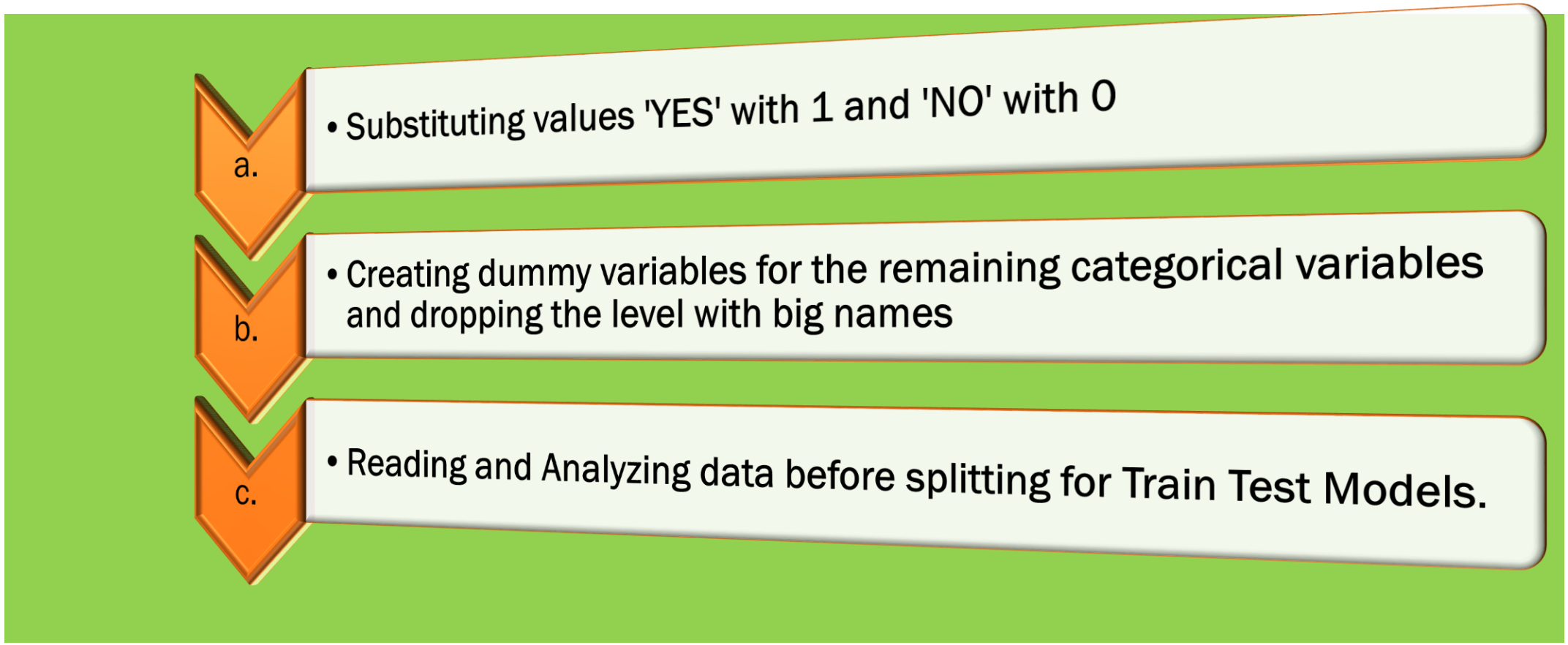


INSIGHTS:

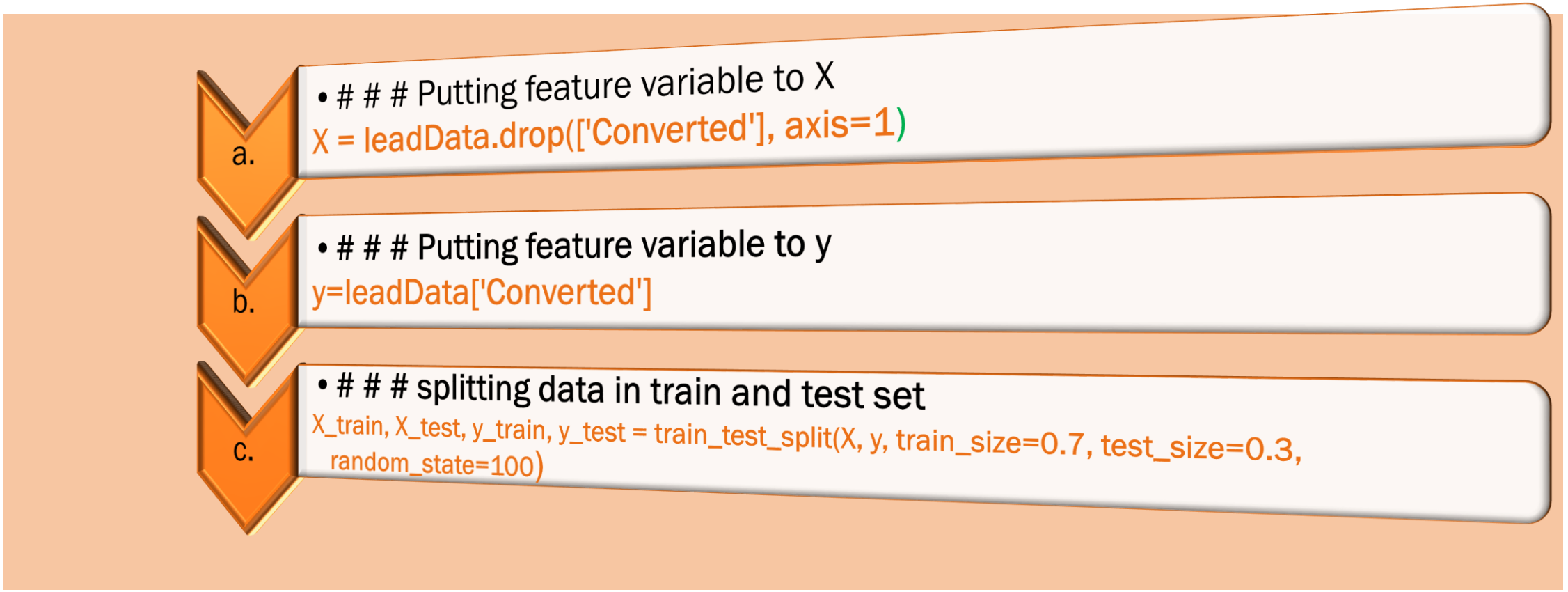
From above it is clearly understandable that all parameters are independent of each other and they don't have any correlation. Hence we have to consider all parameters we cannot drop anyone. Where as TotalVisits and Page Views Per Visit have a fair correlation of 0.77 ,But on this we cannot delete anyone of them ,as they don't have strong correlation.

.

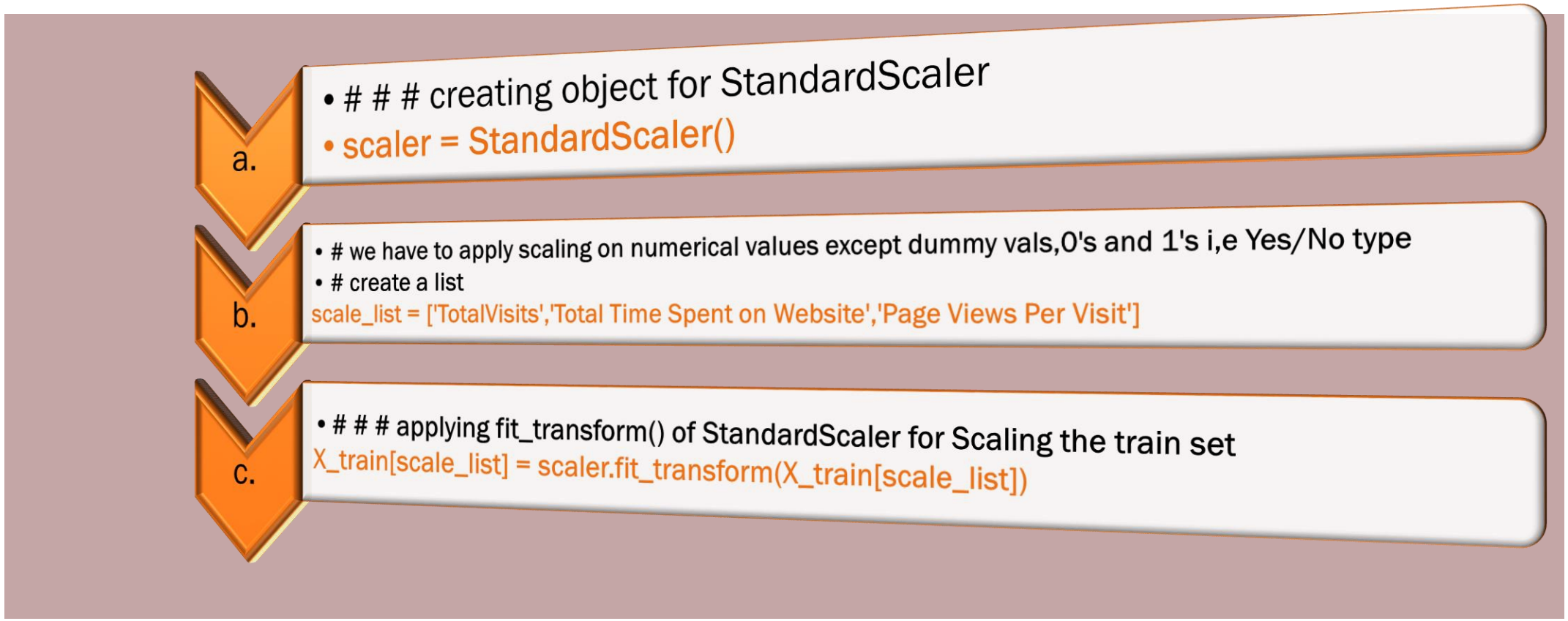
Step-3 Data Preparation



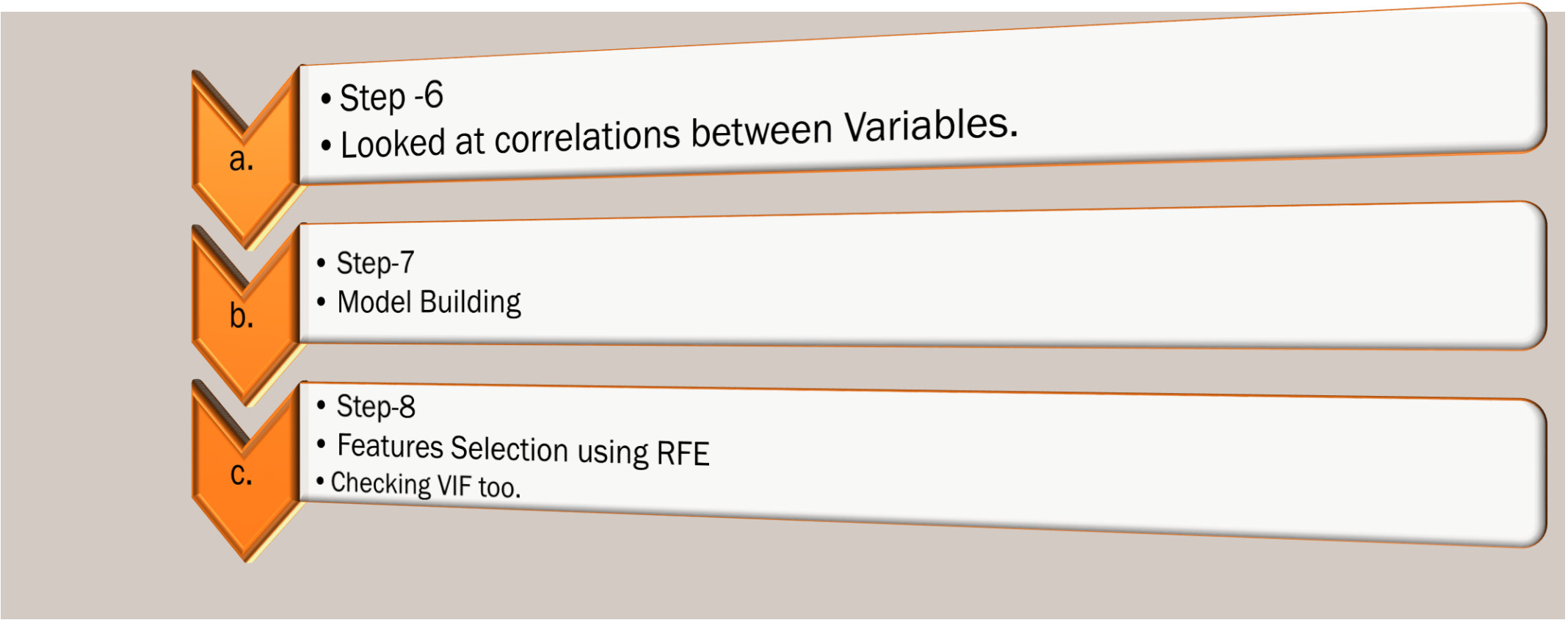
Step-4 Splitting and Creating Training and Testing sets



Step-5 Scaling All Features.



Steps -6,7,8


a.
• Step -6
• Looked at correlations between Variables.
b.
• Step-7
• Model Building
c.
• Step-8
• Features Selection using RFE
• Checking VIF too.

Final Training Model After doing Step 6,7,8

1. Accuracy :
   - 0.8923 i,e 89%
2. Confusion Matrix:

# Actual/Predicted	not_converted	converted
# not_converted	3726	276
# converted	420	2046

3. Calculate Metrics beyond simply accuracy.:
   - i).Sensitivity – 0.8296 i,e 82%
     ii).Specificity – 0.9310 i,e 93%
     iii).False Positive rate – 0.0689
     iv).Positive Predictive Value – 0.8811
     v).Negative Predictive value – 0.8986

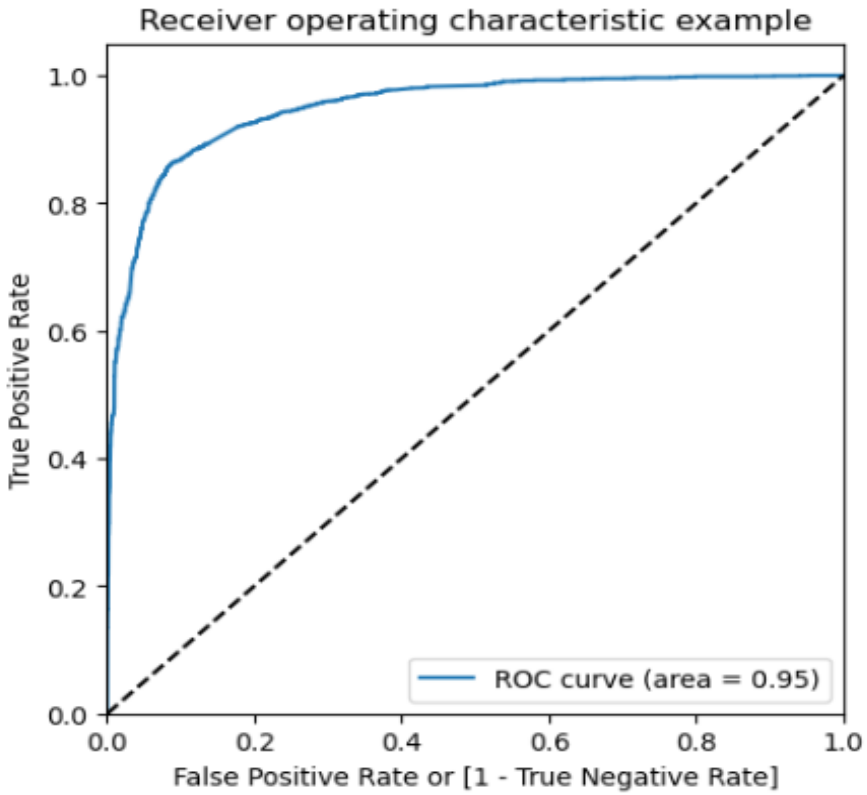
Step-9 Plotting ROC Curve

An ROC curve demonstrates several things:

•It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).

•The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.

•The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.



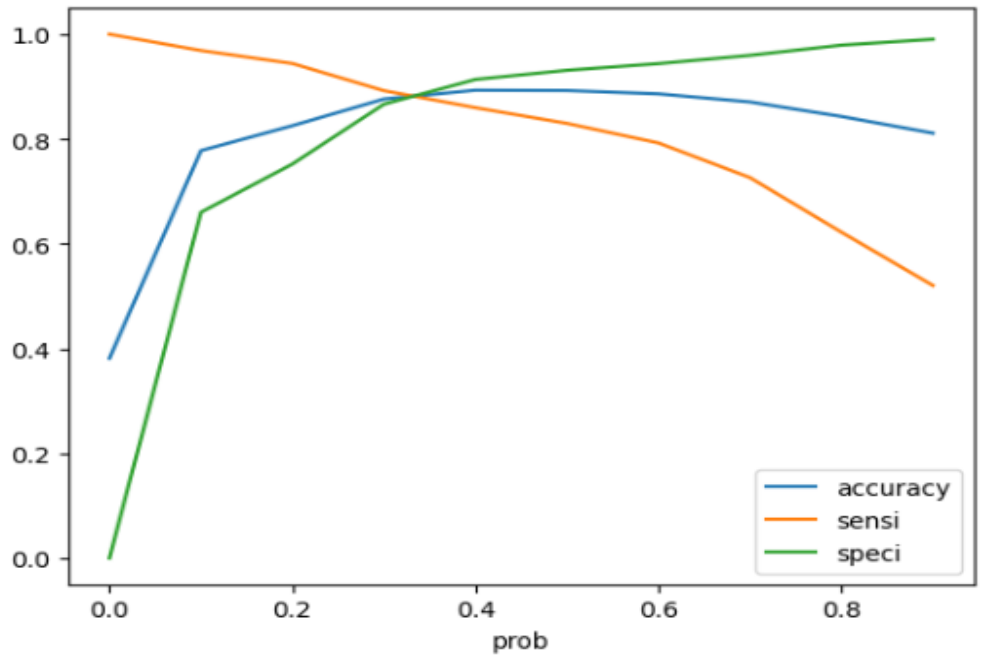
Step-10 Calculating cut-off point

Optimal cutoff probability is that prob where we get balanced sensitivity and specificity

	prob	accuracy	sensi	speci
0.0	0.0	0.381262	1.000000	0.000000
0.1	0.1	0.777675	0.968775	0.659920
0.2	0.2	0.825294	0.944039	0.752124
0.3	0.3	0.876160	0.892133	0.866317
0.4	0.4	0.893012	0.859692	0.913543
0.5	0.5	0.892393	0.829684	0.931034
0.6	0.6	0.886054	0.792376	0.943778
0.7	0.7	0.870594	0.726277	0.959520
0.8	0.8	0.842919	0.622060	0.979010
0.9	0.9	0.811070	0.520276	0.990255

From the curve above, 0.375 is the optimum point to take it as a cutoff probability.

```
In [183]: # Let's plot accuracy sensitivity and specificity for various probabilities.
          cutoff_df.plot.line(x='prob', y=['accuracy','sensi','speci'])
          plt.show()
```



Calculating Again keeping cut-off at 0.375

1 **Accuracy** :

- 0.8900 i,e 89%

2. **Confusion Matrix:**

# Actual/Predicted	not_converted	converted
# not_converted	3621	381
# converted	330	2136

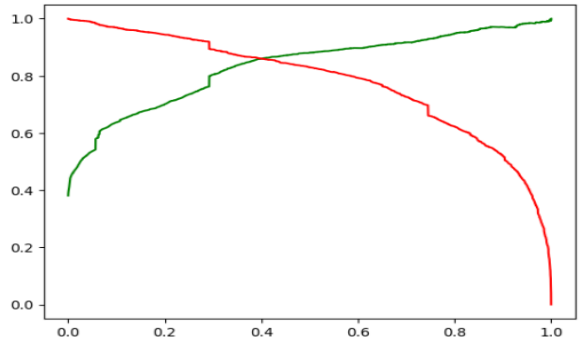
3. **Calculate Metrics beyond simply accuracy.:**

- i).Sensitivity – 0.8666 i,e 86%
  ii).Specificity – 0.9047 i,e 90%
  iii).False Positive rate – 0.095
  iv).Positive Predictive Value – 0.8486
  v).Negative Predictive value – 0.9164

4.. **Percision And Recall Values:**

- i).Precision (TP/TP+FP) – 0.8811
- ii).Recall (TP/TP + FN) – 0.8296
- iii).F1-Score – 0.8546 i,e 85%.

Precision and recall tradeoff :



Lead Score:

	Converted	Lead_Score
0	0	1
1	0	24
2	0	1
3	0	16
4	0	6
...	...	...
6463	1	33
6464	1	98
6465	1	29
6466	0	1
6467	0	6

6468 rows × 2 columns

Step-11 Making Prediction on Test Set

1. Accuracy :
   - 0.8870 i,e 88%
2. Confusion Matrix:

# Actual/Predicted	not_converted	converted
# not_converted	1514	163
# converted	150	945

3. Calculate Metrics beyond simply accuracy.:
   - i).Sensitivity -- 0.8630 i,e 86%
     ii).Specificity -- 0.9028 i,e 90%
     iii).False Positive rate -- 0.097
     iv).Positive Predictive Value -- 0.8528
     v).Negative Predictive value -- 0.9098

Conclusions and Recommendations for the Company Strategy

1.The model evaluation steps revealed that the accuracy, precision and recall parameters are acceptable values.

2. The top three variables in your model which contribute most towards the probability of a lead getting converted are:

a. What is your current occupation. [positively co-relation]
b. Tags. [Both negatively and positively co-related]
c. Lead Source. [Both negatively and positively co-related]

3.In order to increase the time that a user spends on webpage, the company to employ more web developers and UI/UX designers to improve the experience for the user and thereby luring the users to spend more time on the webpage, exploring the contents.

4.The other valuables like Lead Source especially Google ,Direct Traffic, Organic Search, Olark Chat, Reference has a very high chance of getting converted. Welingak Website and Reference has an importance influence on the Lead score. It would be a good strategy to increase the marketing budget on these lead sources.

5.It is also important not to waste a lot of time on some factors that do not contribute much or negatively affect the Lead scores.