LEAD SCORING

CASE STUDY

Presented by:
Surya Nag S
Hausna Maliakkal
Sushil patil

LEAD SCORING PROCESS

On-page Search

Downloads

Page Views

Email Open Rate

Webinars

Problem Statement:

- X Education sessls online courses to industry professional.

- X Education gets a lot of leads, it's 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

BUSINESS OBJECTIVE

- X Education wants to know most promising leads.
- For that they want to build a model which identifies the hot leads.
- Deployement of the model for the future use.

Times New Roman

SOULUTION METHEDOLOGY

Data cleaning & Data manipulation

- Check and handle duplicate data.
- Check and handle NA values and missing values.
- Drop columns, if it contains a large number of missing values and are not useful for the analysis
- Imputation of the values, if necessary
- Check and handle outliers in data.

Exploratory Data Analysis

- Univariate data analysis: value count, distribution of variables, etc.
- Bivariate data analysis: Correlation coefficients and pattern between the variables ect.
- Feature scaling & Dummy variables and encoding of the data.
- Classification technique: Logistic regression is used for model making & prediction.
- Validation of the model.
- Model presentation.
- Conclusion and recommendations

DATA MANIPULATION

- Total Number of rows = 9240, Total no. of columns = 37.

- Single value features like "Magazine", "Receive More Updates About Our Courses", "Update mu supply"

- Chain content, Get updates on DM content, "i agree to pay the amount through cheque" etc.

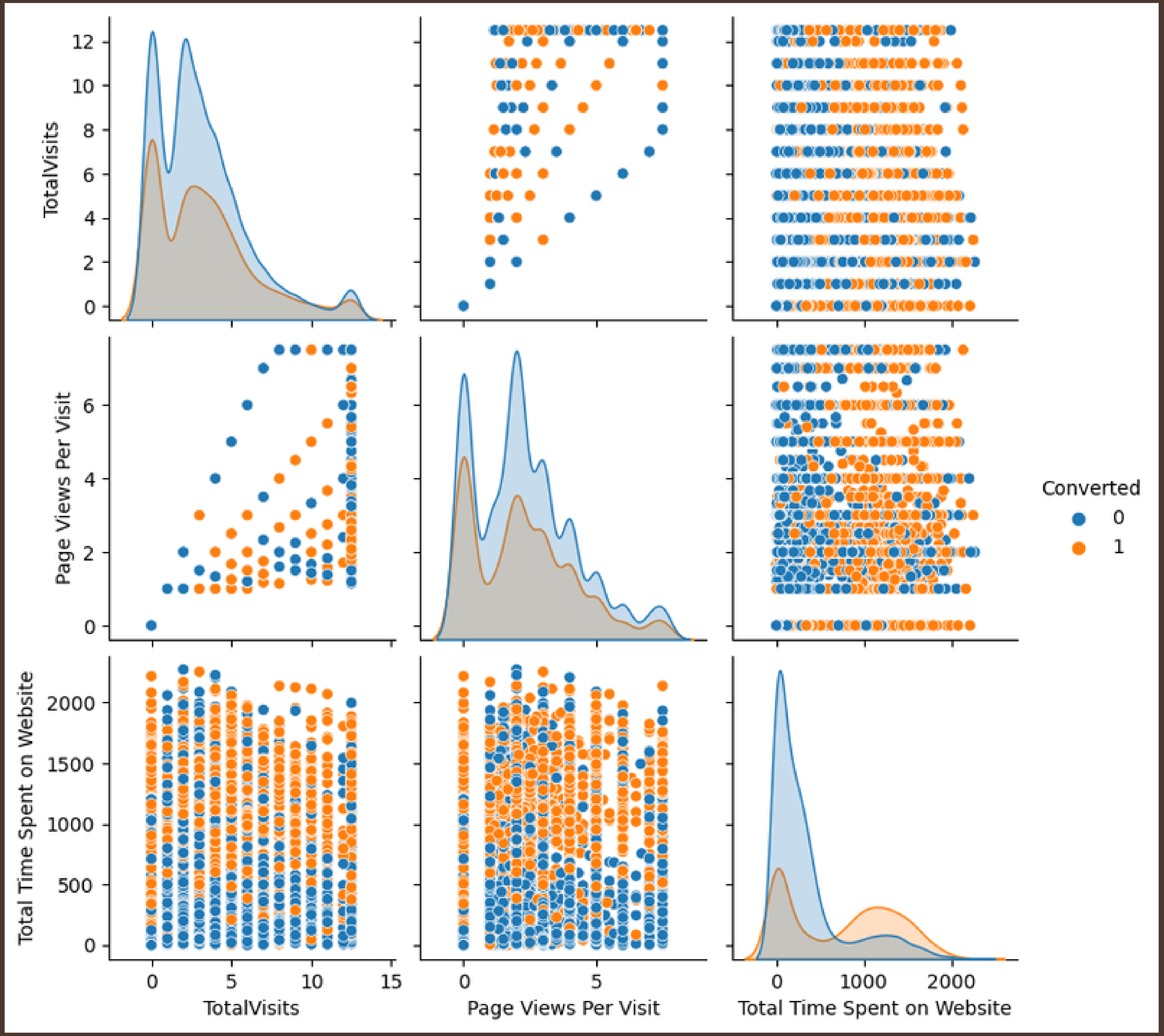
- Removing the "Prospect ID and "Lead Number" which are not necessary for the analysis.

- After checking for the value counts for some of the object type type variables, we find some of the features which have enough variance, which have dropped, the features are:

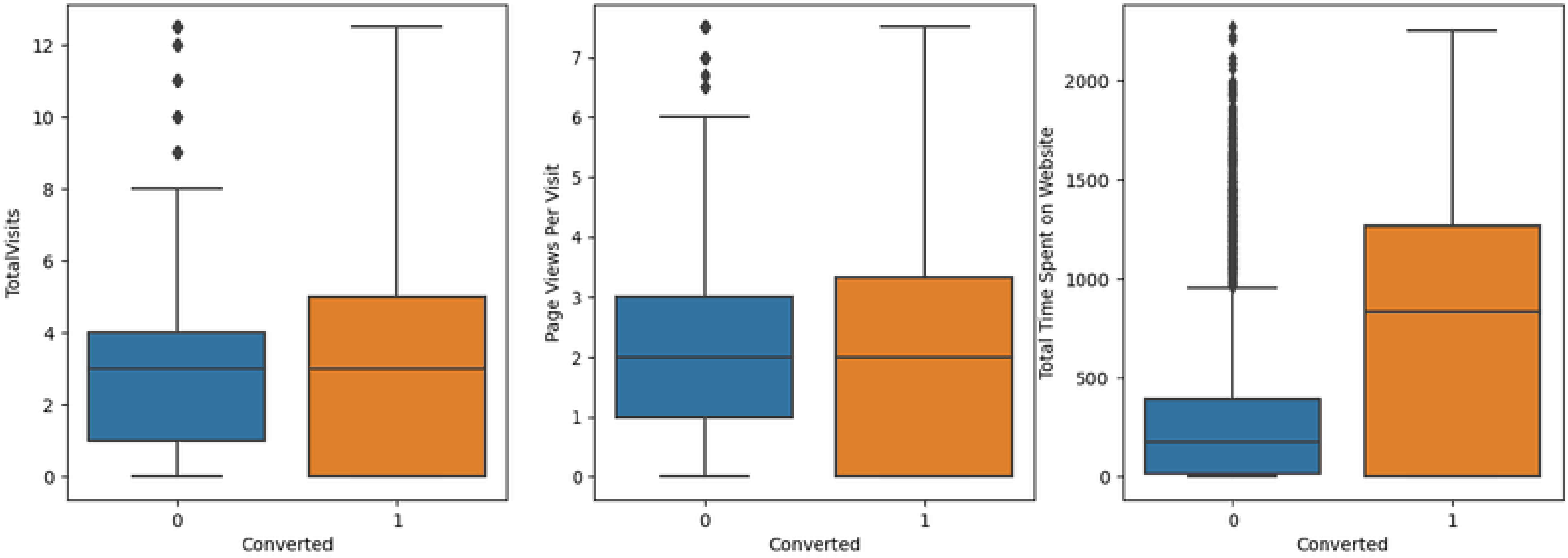
"Do Not Call", "What matters most to you in choosing course", "Search", "Newspaper Article", " X Education Forums", "Newspaper", "Digital Advertisement" etc.

- Dropping the column shaving more than 35% as missing values such as "How did you hear about X Education" and "Lead Profile"

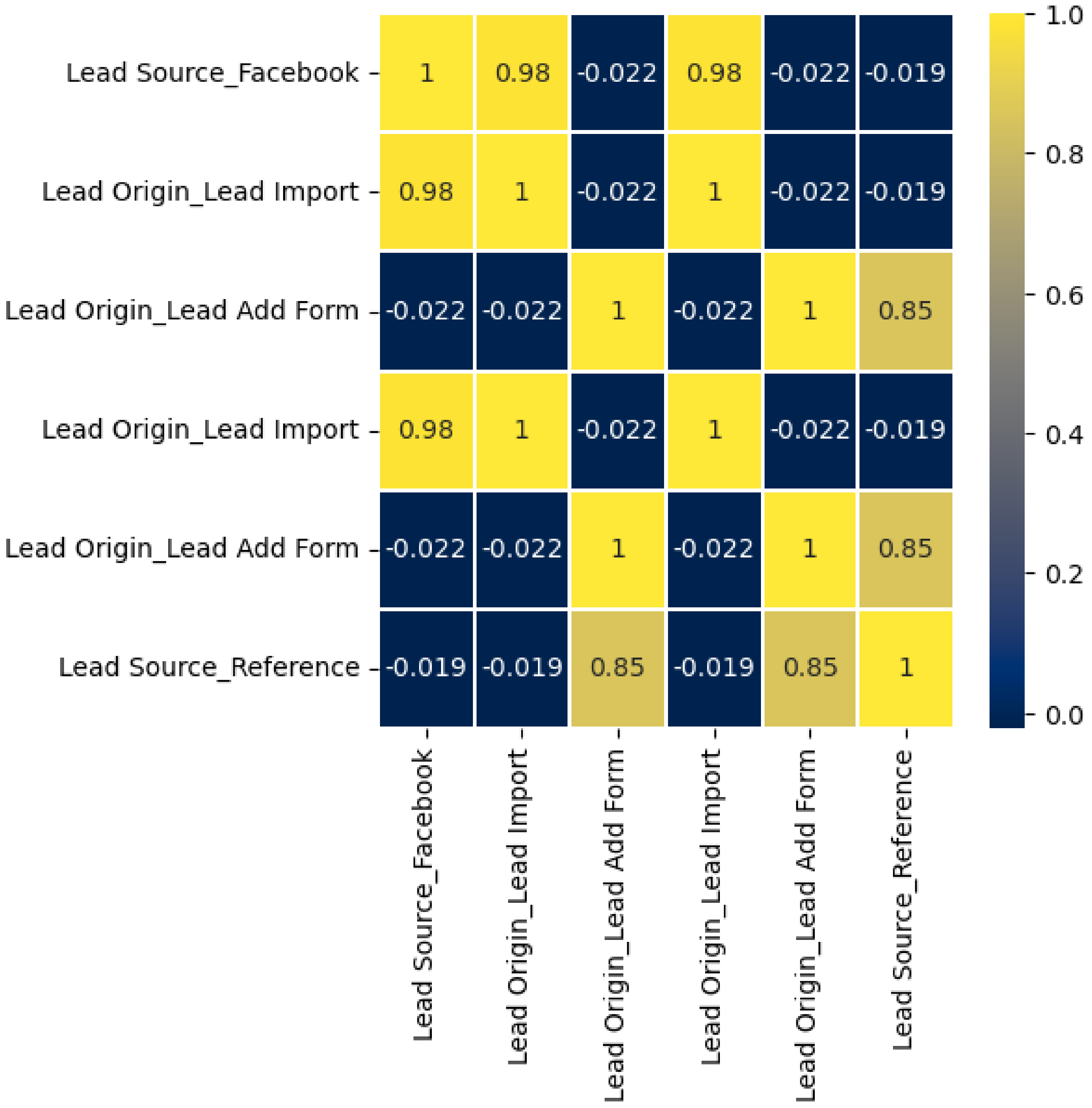
EXPLORATORY DATA ANALYSIS (EDA)



BOX PLOT


TotalVisits
0
2
4
6
8
10
12
Converted
0
1
Page Views Per Visit
0
1
2
3
4
5
6
7
Converted
0
1
Total Time Spent on Website
0
500
1000
1500
2000
Converted
0
1

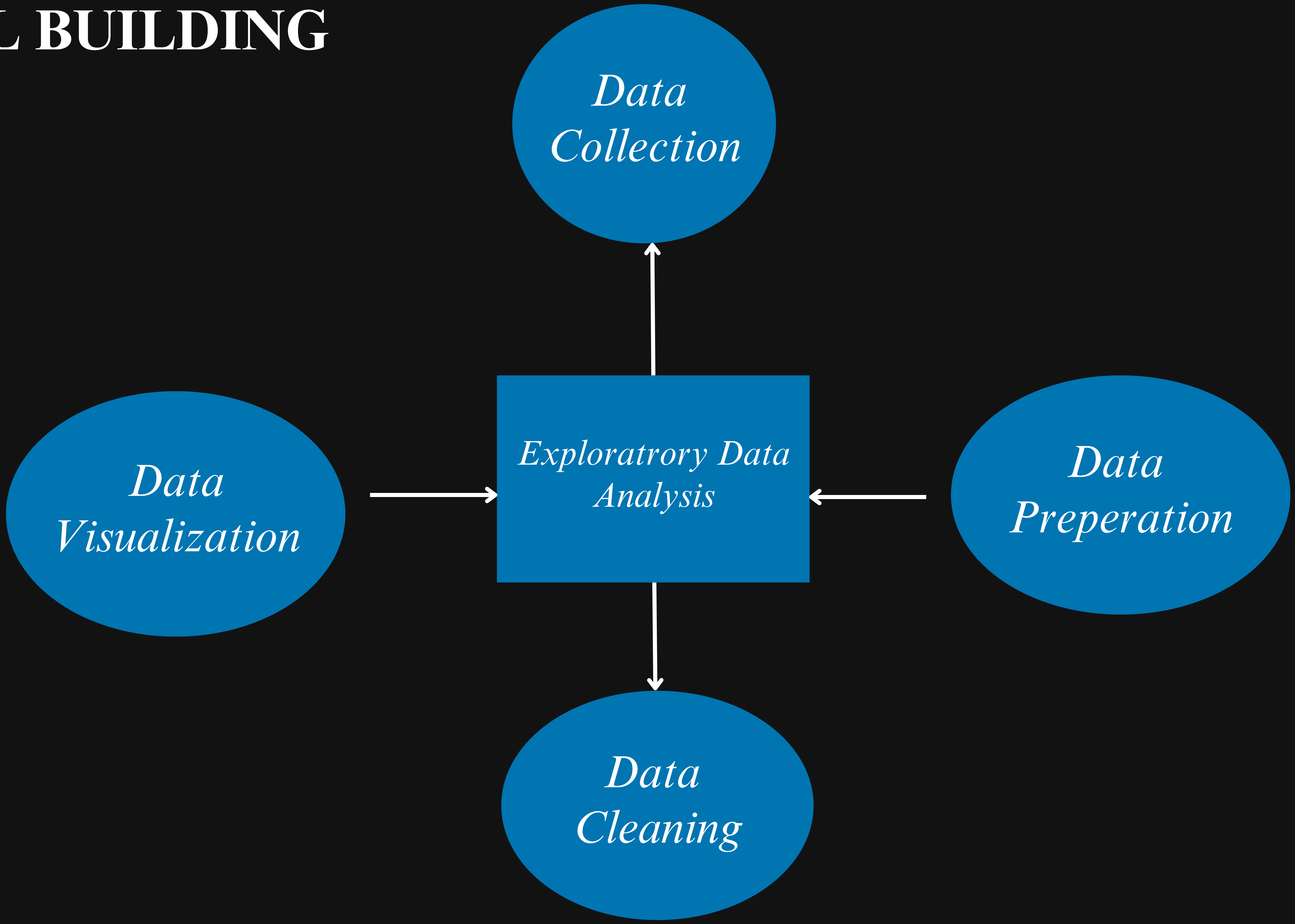
HEAT MAP


Lead Source_Facebook
Lead Origin_Lead Import
Lead Origin_Lead Add Form
Lead Origin_Lead Import
Lead Origin_Lead Add Form
Lead Source_Reference
1 0.98 -0.022 0.98 -0.022 -0.019
0.98 1 -0.022 1 -0.022 -0.019
-0.022 -0.022 1 -0.022 1 0.85
0.98 1 -0.022 1 -0.022 -0.019
-0.022 -0.022 1 -0.022 1 0.85
-0.019 -0.019 0.85 -0.019 0.85 1
Lead Source_Facebook
Lead Origin_Lead Import
Lead Origin_Lead Add Form
Lead Origin_Lead Import
Lead Origin_Lead Add Form
Lead Source_Reference
1.0
0.8
0.6
0.4
0.2
0.0

DATA CONVERSION

- Numerical Variables are normalized
- Dummy Variables are created for object type variables
- Total Rows for Analysis- 37
- Total Columns for Analysis- 9240

MODEL BUILDING



MODEL BULDING

- Splitting the Data in to Training & Testing Sets.

- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.

- Use RFE for Feature Selection

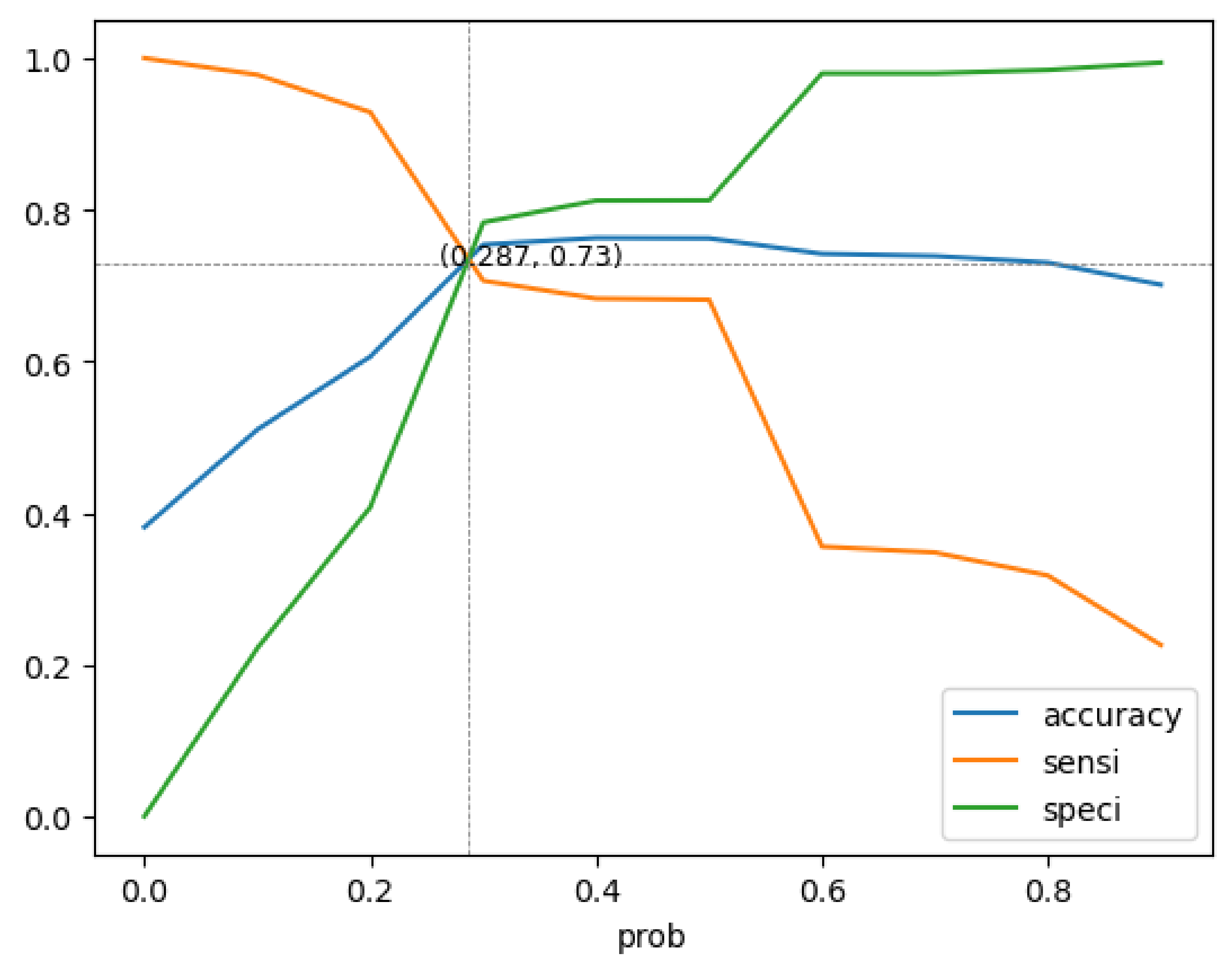
- Running RFE with 15 variables as output

- Building Model by removing the variable whose P-value is grater than 0.05 and vi value is greater than 5

- Predictions on test data set.

- Over all accuracy 81%.

ROC Curve



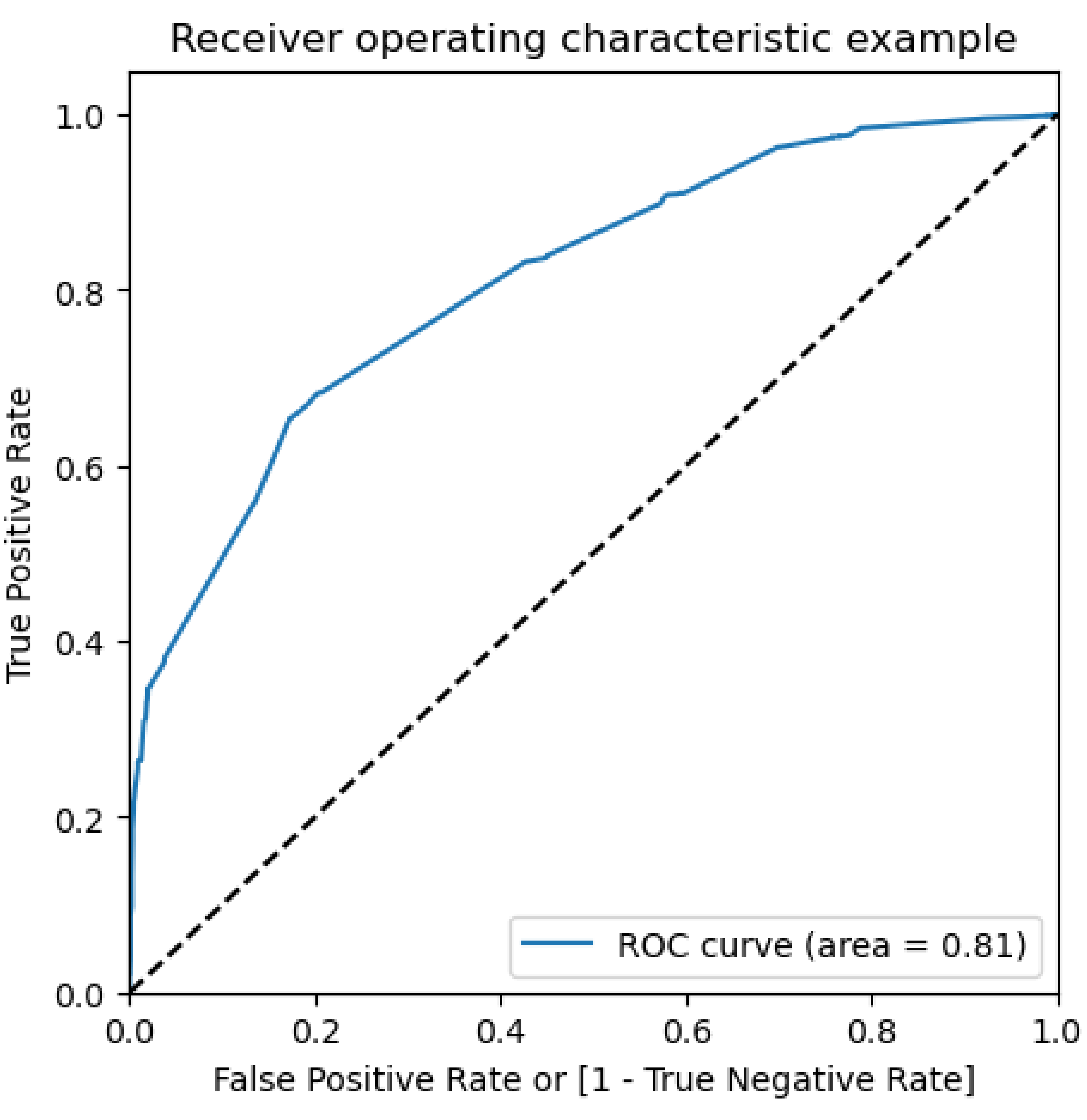
- Finding Optimal cut off Point

- Optimal cut-off probability is that

- Probability where we get balanced sensitivity and specificity

- From the second graph it is visible that the optimal cut off is at 0.35

ROC Curve


Receiver operating characteristic example
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate or [1 - True Negative Rate]
ROC curve (area = 0.81)

PREDICTION ON TEST SET

The evaluation matrix are pretty close to each other so it indicates that the model is performing consistently across different evaluation metrics in both test and train dataset.

For the Test Set:

- **Accuracy: 76.24%**
- **Sensitivity: 68.13%**
- **Specificity: 81.23%**
- These metrics are very close to train set, so out final model logm4 is performing with good consistency on both Train & Test Set.

- This shows that our test prediction is having accuracy, precision and recall scores in an acceptable range.

CONCLUSION

- It was found that the variables that mattered the most in the potential buyers are (In descending order):
- The total time spent on the website.
- Total number of visits.
- When the lead source was: Google, Direct traffic Organic search Welingak Website
- When the last activity was: SMS, Olark chat conversion
- When the lead origin is lead add format.
- When their current occupation is as a working professional.
- Keeping these in mind X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.

THANK YOU