LEAD SCORING-X EDUCATION

GROUP MEMBERS:

- ABHAY TOMER
- ANUJ GROVER

PROBLEM STATEMENT

- X Education sells online courses to industry professionals.
- 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.

BUSINESS OBJECTIVE:

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

SOLUTION METHODOLOGY:

- Data Cleaning and Data Manipulation

1. Check and handle duplicate data
2. Check and handle missing values
3. Drop columns, if contains missing values
4. Imputation of values
5. Check and handle outliers

- EDA

a. Univariate Data Analysis: Value count, distribution of variable etc

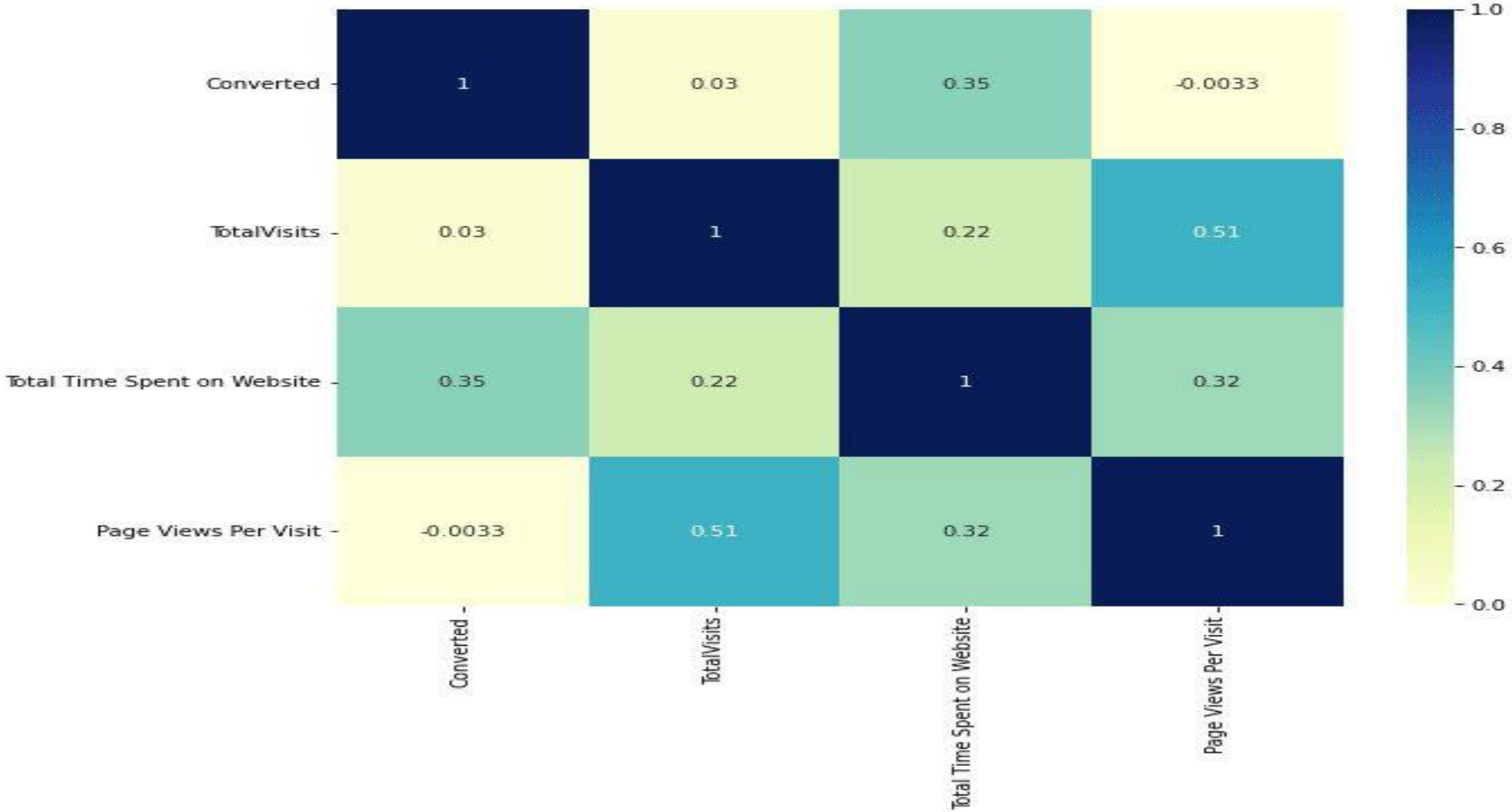
b. Bivariate Data Analysis: correlation coefficients and pattern between the variables etc.

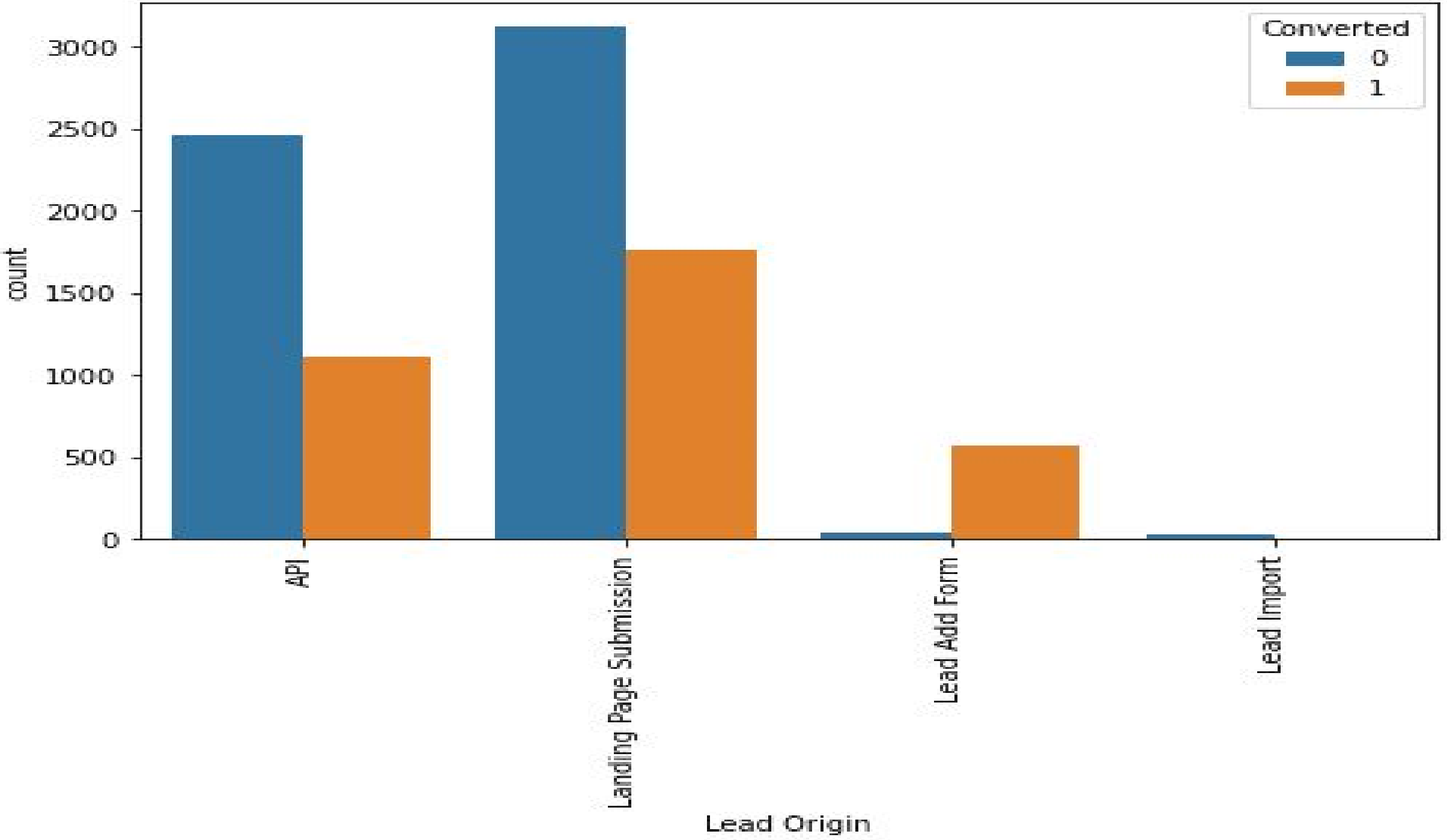
- Feature Scaling and Dummy variables and encoding of the data
- Classification technique: logistic regression used for making predictions
- Validation of the model
- Conclusion and recommendations

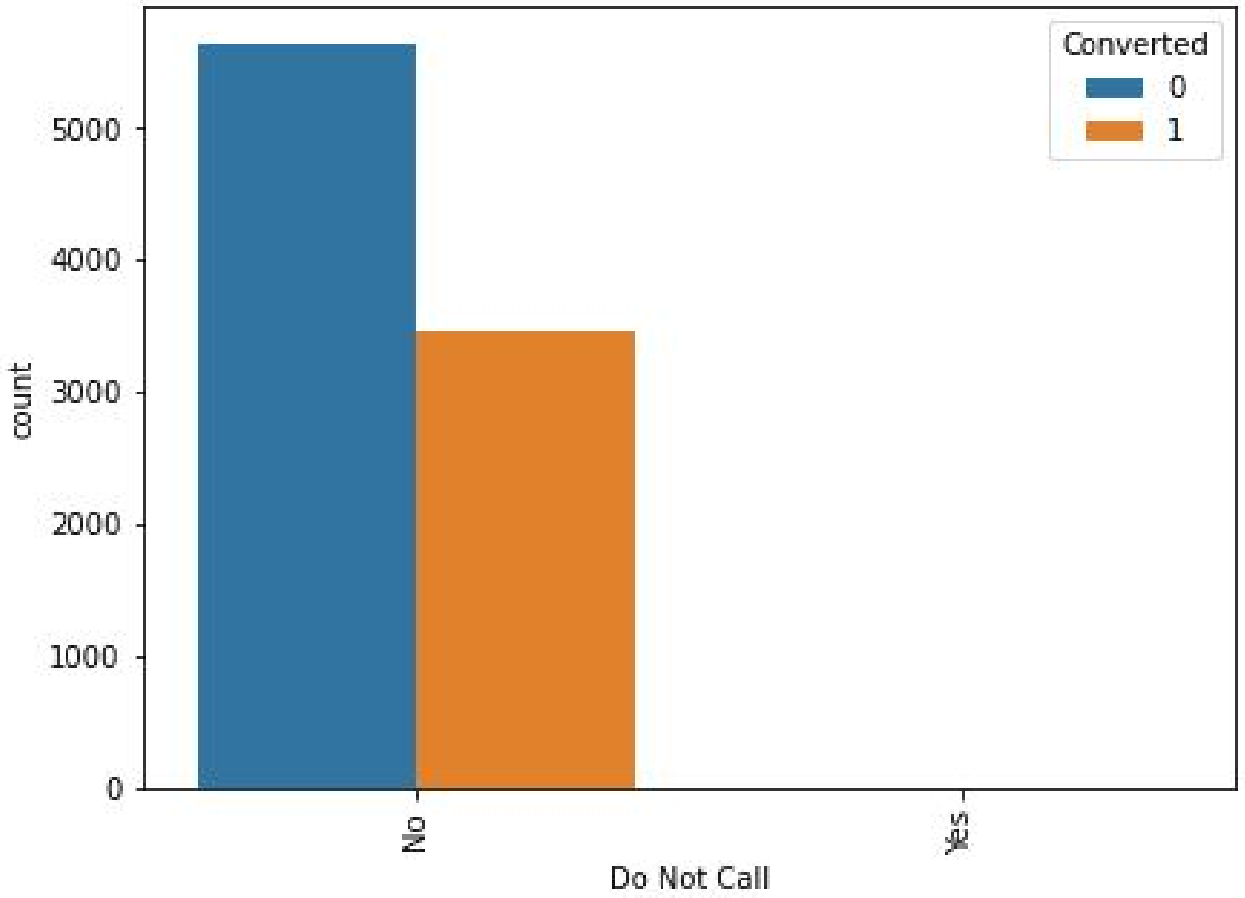
DATA MANIPULATION

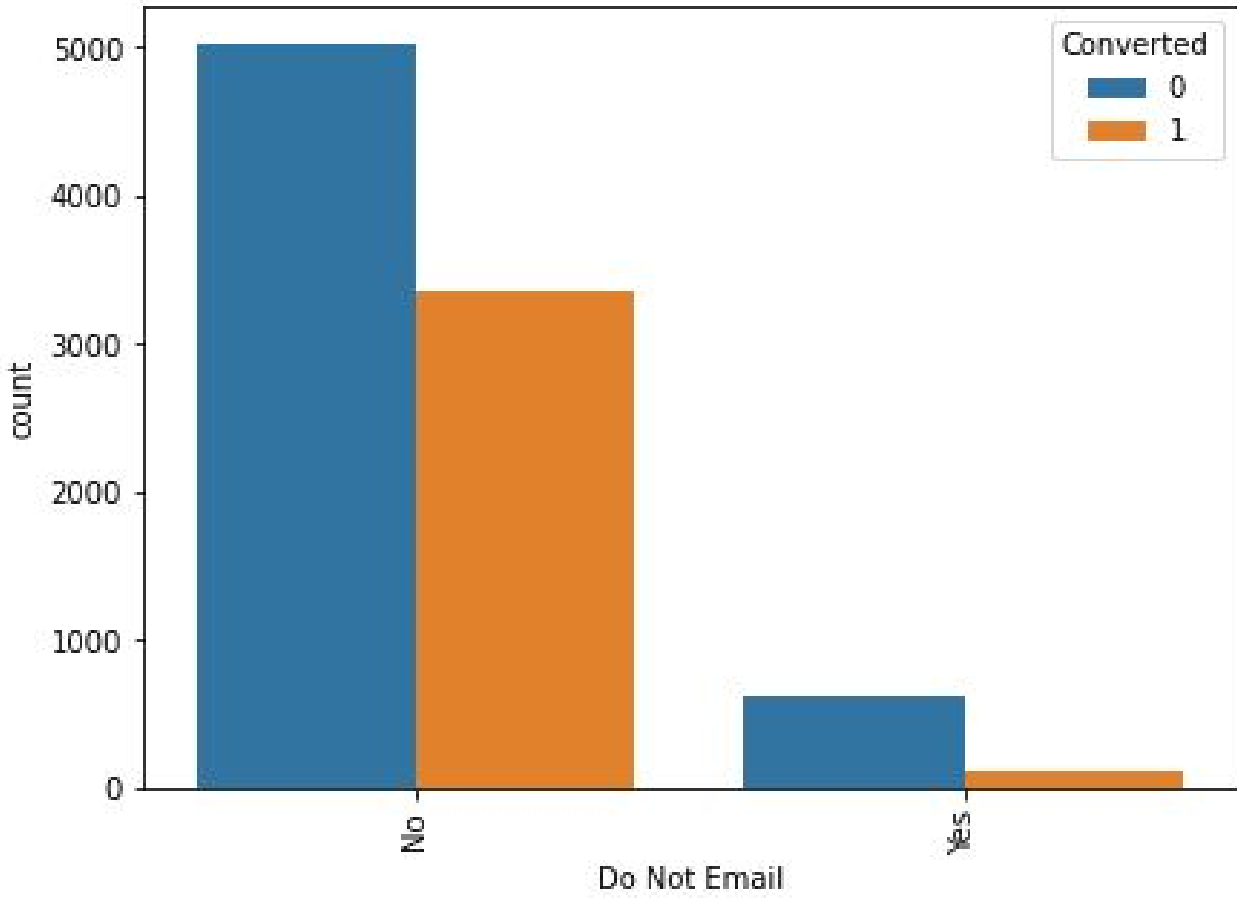
- Total number of Rows = 37, Total number of columns = 9240
- Single values feature like ’Magazine’, ‘Receive more updates about our courses’, ‘update me on supply’.
- ‘Chain content’, ‘ Get updates on DM Content’, ‘I agree to pay the amount through cheque’ have been dropped
- Removing the ‘Prospect ID’ and ‘Lead Number’ which is not neceassory for the analysis.
- After checking for the value counts for some of the object type variables, we find some of the features which has no variance, which we have dropped the features are: ‘Do not call’
- Dropping the columns having more than 35% as missing values such as ‘How did you hear about X Education and ‘Lead Profile’

EDA

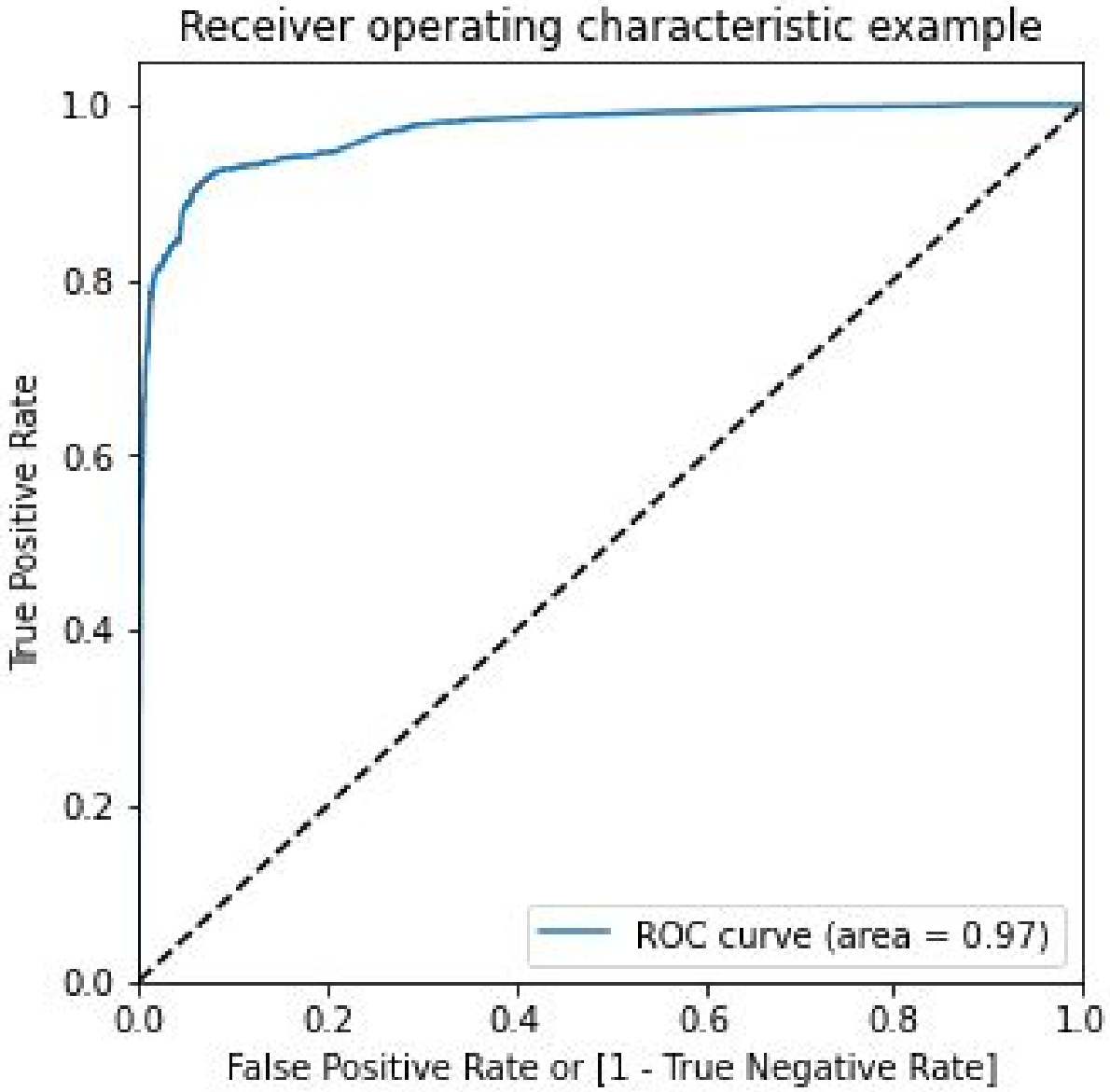

Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
1
0.03
0.35
-0.0033
0.03
1
0.22
0.51
0.35
0.22
1
0.32
-0.0033
0.51
0.32
1
Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
1.0
0.8
0.6
0.4
0.2
0.0


Converted
0
1
3000
2500
2000
1500
1000
500
0
count
API
Landing Page Submission
Lead Add Form
Lead Import
Lead Origin

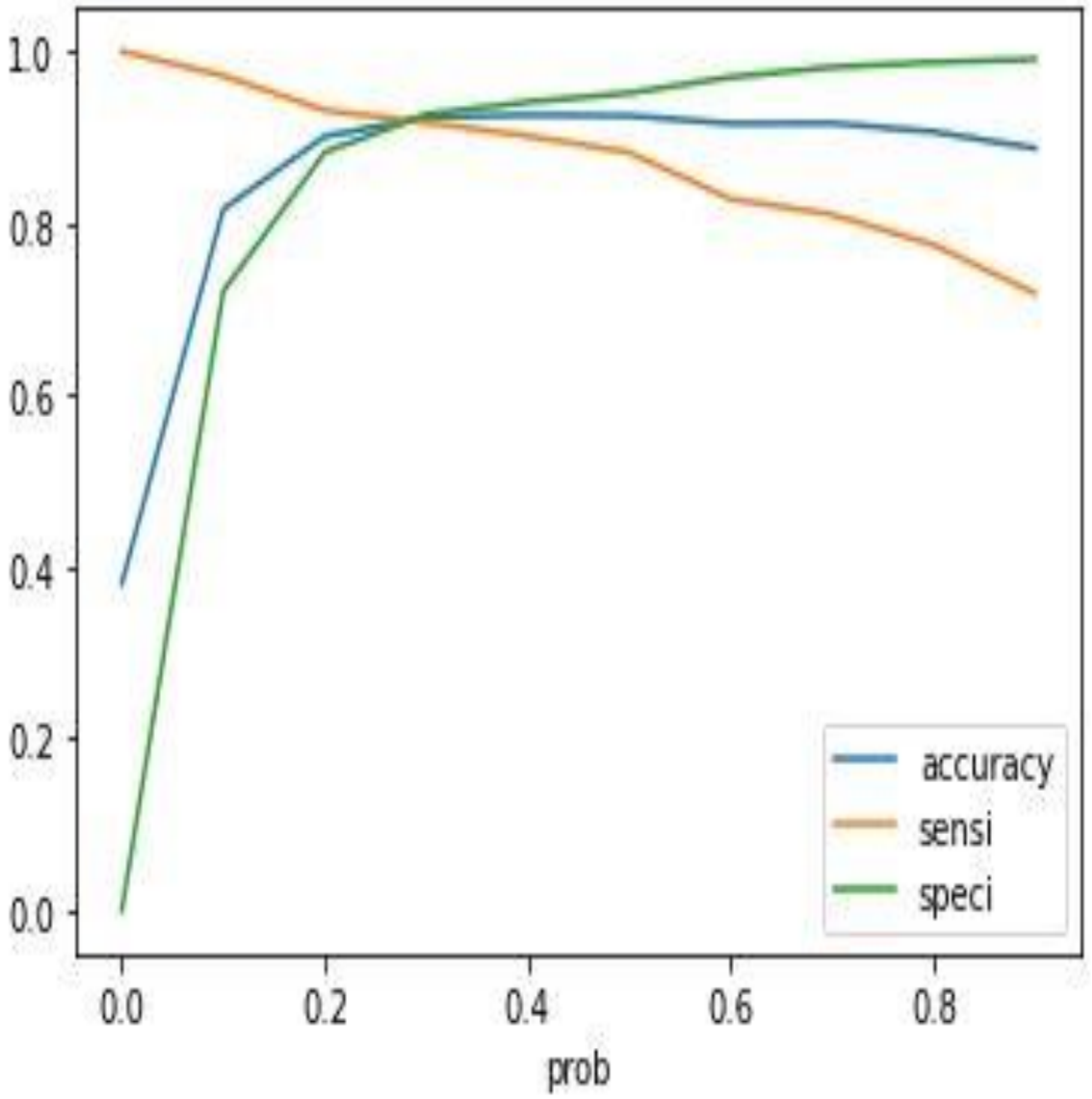

Converted
0
1
5000
4000
3000
2000
1000
0
count
No
Yes
Do Not Call


Converted
0
1
5000
4000
3000
2000
1000
0
count
No
Yes
Do Not Email

ROC CURVE



The ROC Curve should be a value close to 1. We are getting a good value of 0.97 indicating a good predictive model.



So as we can see that the model seems to be performing well. The ROC curve has a value of 0.97, which is very good. We have the following values for the Train Data: Accuracy : 92.29% Sensitivity : 91.70% Specificity : 92.66% Some of the other Stats are derived below, indicating the False Positive Rate, Positive Predictive Value, Negative Predictive Values, Precision & Recall.

CONVERSION

- Numerical Variables are Normalised
- Dummy Variables are created for object type variable

CONCLUSION

Our Logistic Regression Model is decent and accurate enough, when compared to the model derived using PCA. X Education Company needs to focus on following key aspects to improve the overall conversion rate: a. Increase user engagement on their website since this helps in higher conversion b. Increase on sending SMS notifications since this helps in higher conversion c. Get Total Visits increased by advertising etc. since this helps in higher conversion d. Improve the Olark Chat service since this is affecting the conversion negatively.

MODEL BUILDING

- Splitting The data into Training and Testing Data
- The first basic step is to perform Regression in train-test split, we have chosen 70:30 ratio.
- Using RFE with 15 variable as output.
- Building Model whose p-value is greater than 0.5 and vif value is greater than 5
- Predictions on test data set
- Overall accuracy 93%


THANK YOU