LEAD SCORING CASE STUDY

SUBMITTED BY

TANMAY DAS

HARSHITH HAREESH

HARSH NEVATIA

PROBLEM STATEMENT

- X Education, an online course provider, faces a low lead conversion rate of around 30%. Despite acquiring numerous leads daily through various marketing channels, only a small percentage convert into paying customers. The company markets its courses on several websites and search engines like Google.
- Once potential customers land on the website, they might browse the courses, fill out a form, or watch videos. When these individuals provide their email address or phone number, they are classified as leads. Additionally, the company receives leads through past referrals. The sales team then engages with these leads through calls and emails, but the conversion rate remains low.
- To improve the lead conversion rate to approximately 80%, X Education aims to identify and prioritize 'Hot Leads'—leads with the highest potential for conversion. By focusing on these promising leads, the sales team can enhance their efforts and increase the overall conversion rate.

BUSINESS OBJECTIVE

- The goal is to build a logistic regression model with the given dataset which contains about 9000 data points to assign a lead score between 0 and 100 to each lead, helping the company focus on the most promising leads and improve the overall conversion rate.
- 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.

SOLUTION METHODOLOGY

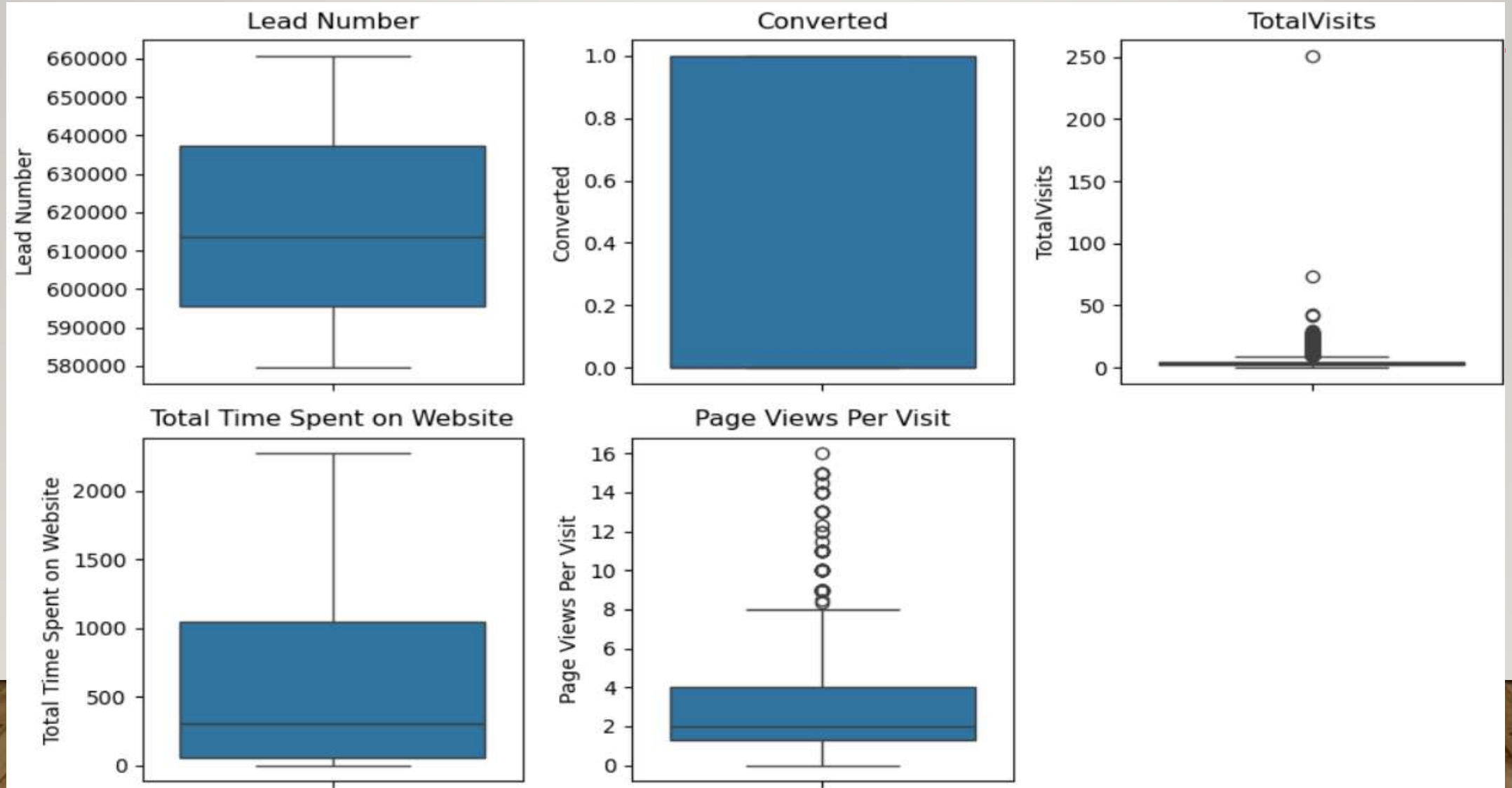
- DATA PREPARATION (handling null values, removing unwanted columns, handling outliers, imputing data)
- EXPLORATORY DATA ANALYSIS
- DUMMY VARIABLE CREATION
- FEATURE SCALING
- MODEL BUILDING (LOGISTIC REGRESSION,STATS MODEL,RFE,P-VALUES0
- MODEL EVALUATION
- MAKING PREDICTIONS

DATA PREPARATION

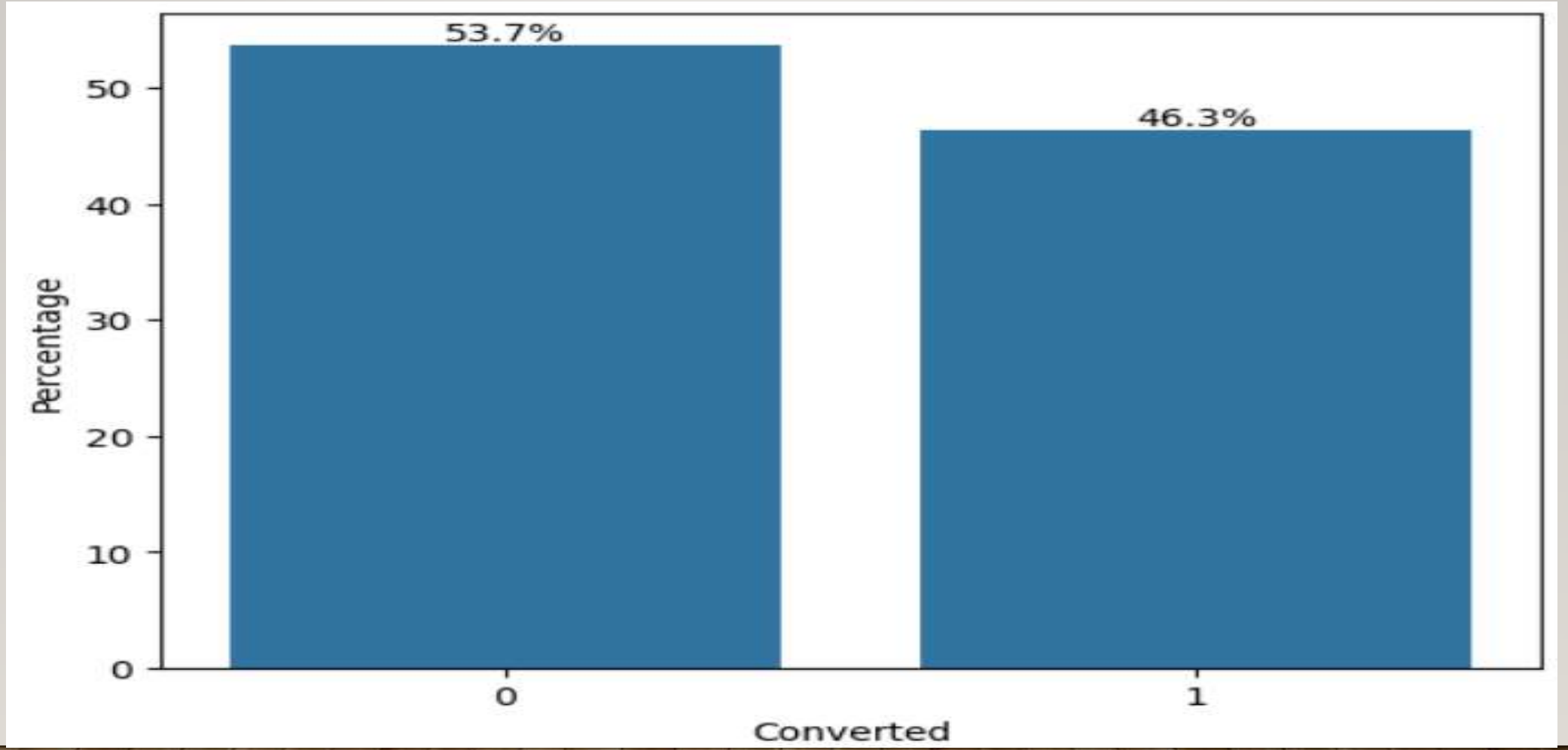
- 9270 rows and 37 columns
- "**Select**" appears to have been mistakenly recorded during data collection, despite it not representing a valid data point. This can be substituted with "**Unknown**" to maintain data integrity and ensure consistency in analysis.
- DROPPING columns having more than 30% null values
- Dropped Category Columns with Skewed Distribution of sub-categories or single values only
- Imputing NULL VALUES OF categorical and numerical columns with mode and median respectively

EDA

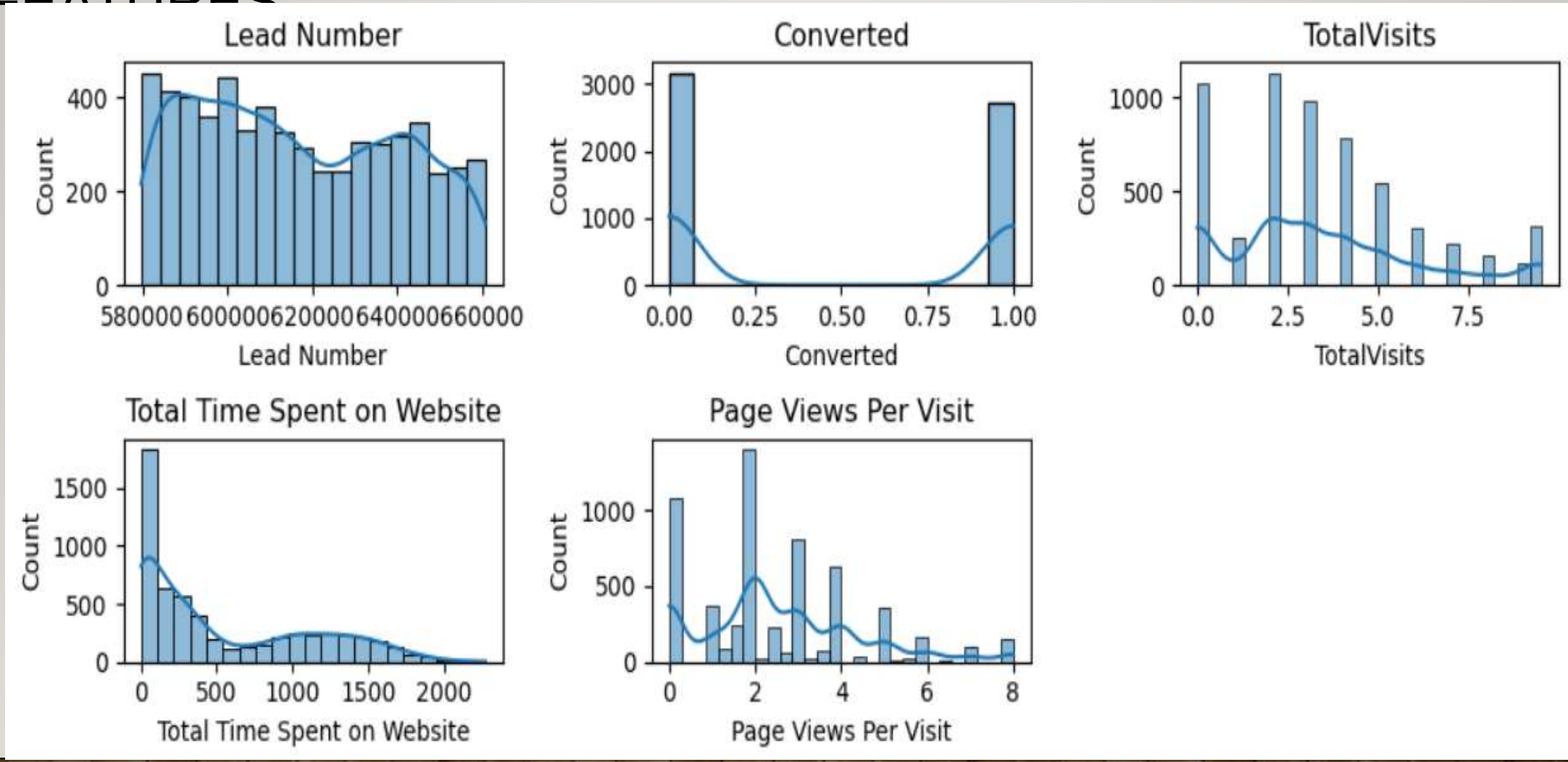
- BOX PLOT WITH NUMERICAL COLUMNS



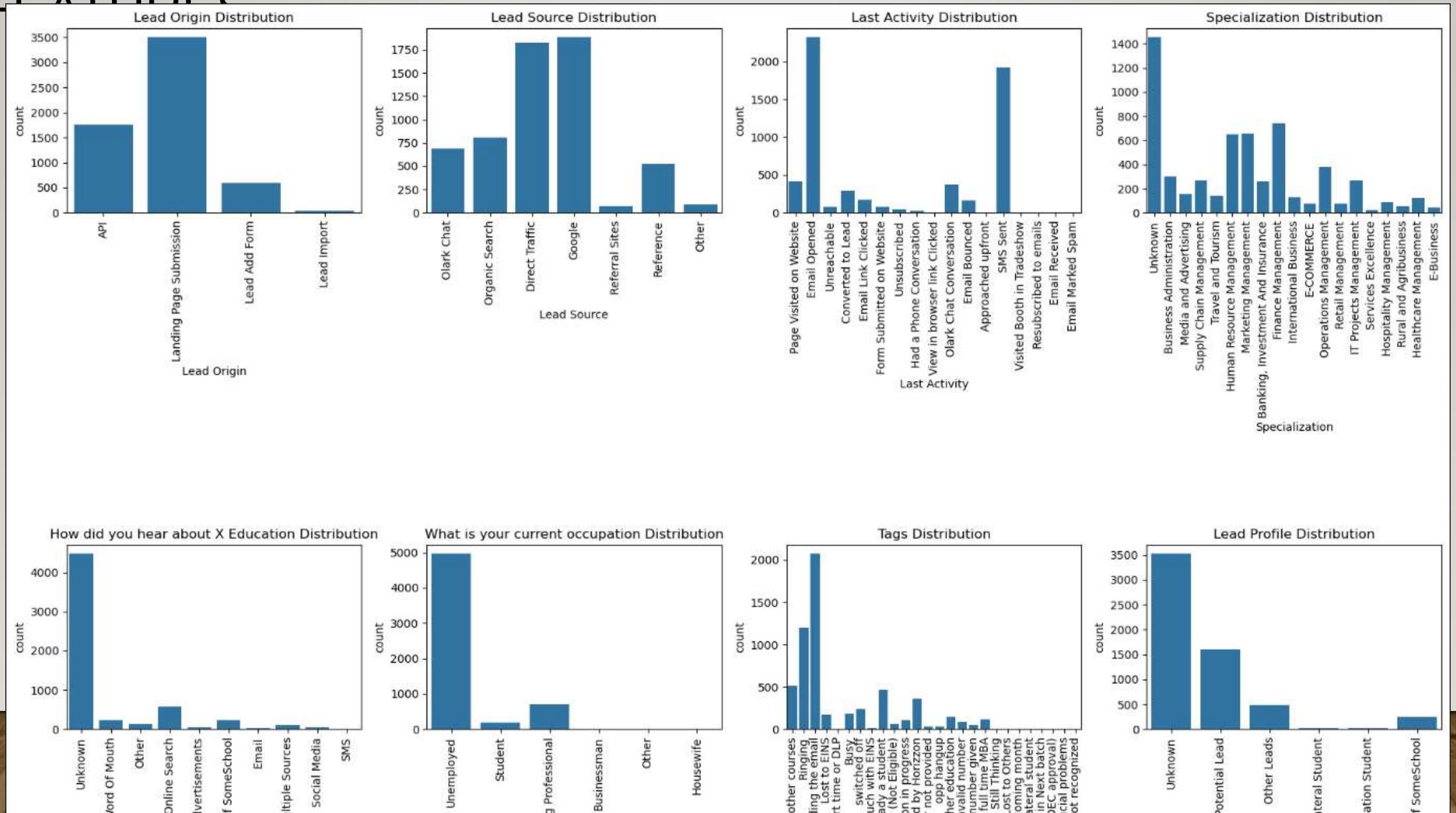
- BAR PLOT WITH TARGET VARIABLE- CONVERTED



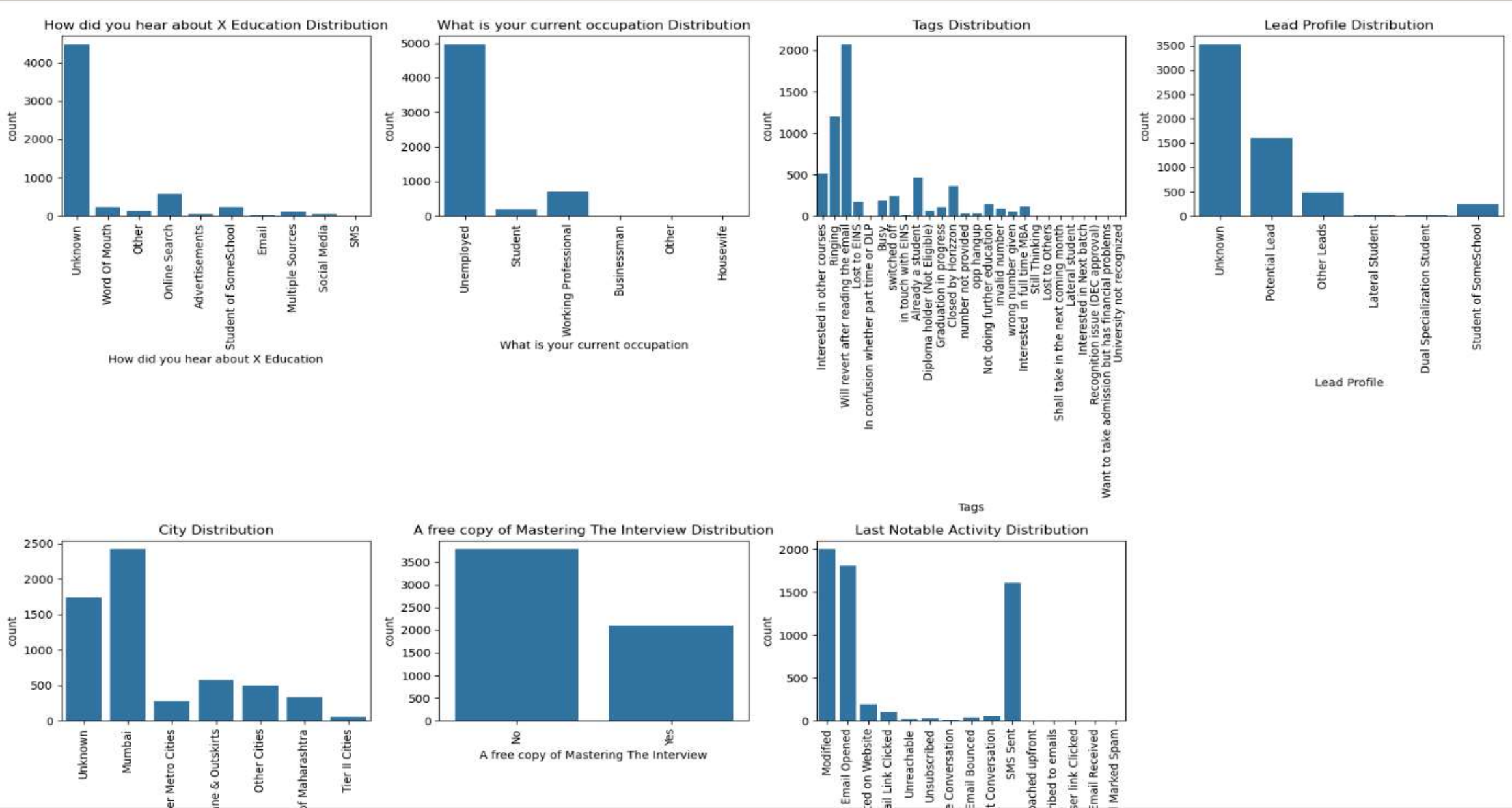
UNIVARIATE ANALYSIS - WITH NUMERICAL FEATURES



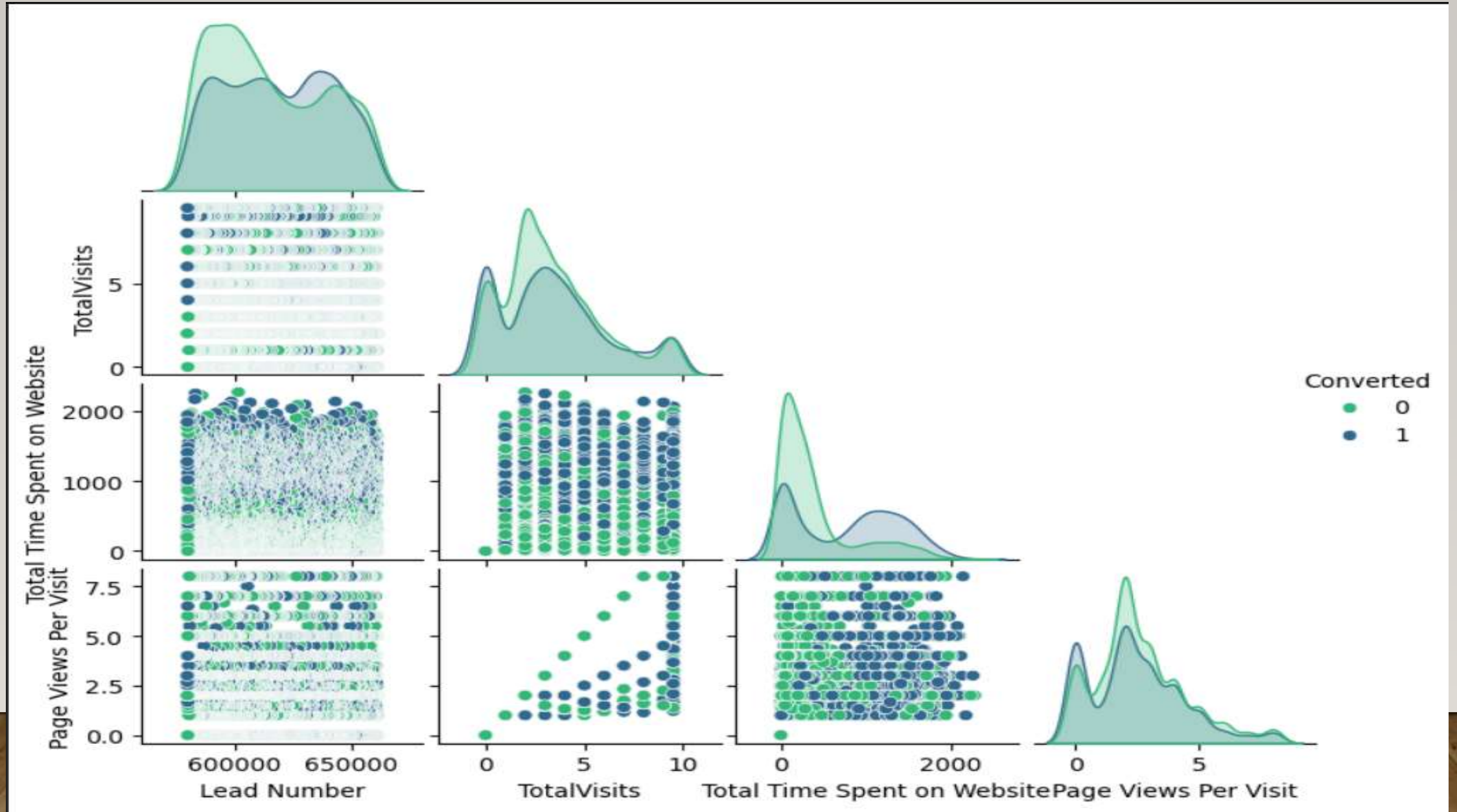
UNIVARIATE ANALYSIS -WITH CATEGORICAL FEATURES



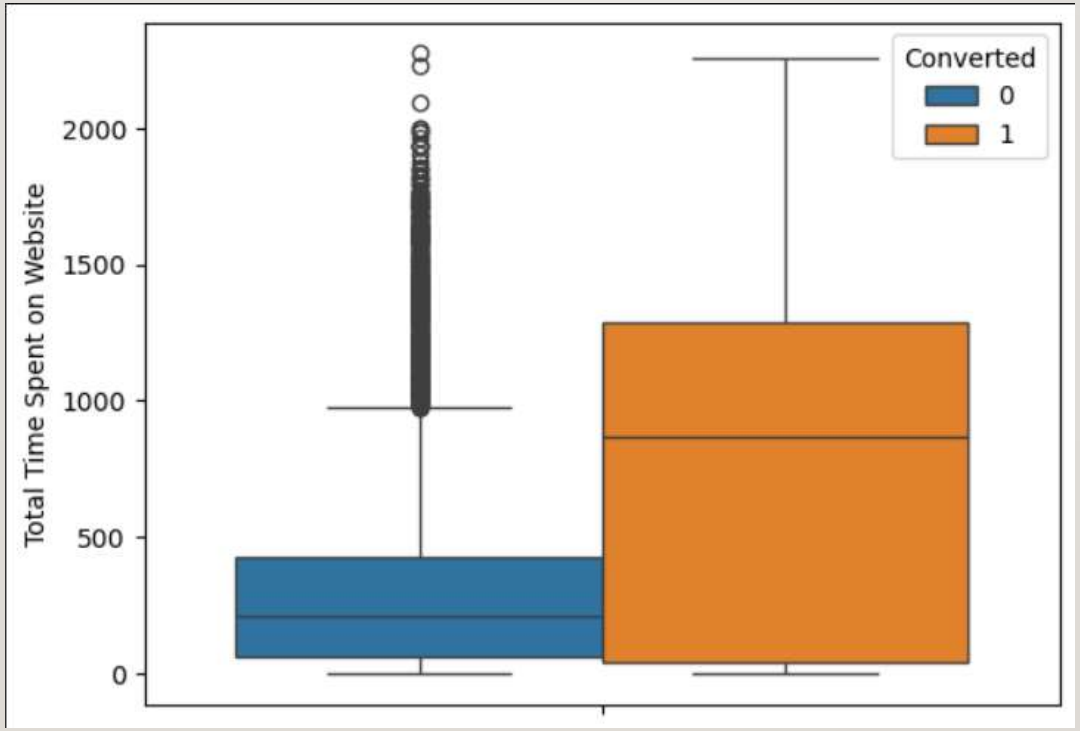
UNIVARIATE ANALYSIS -WITH CATEGORICAL



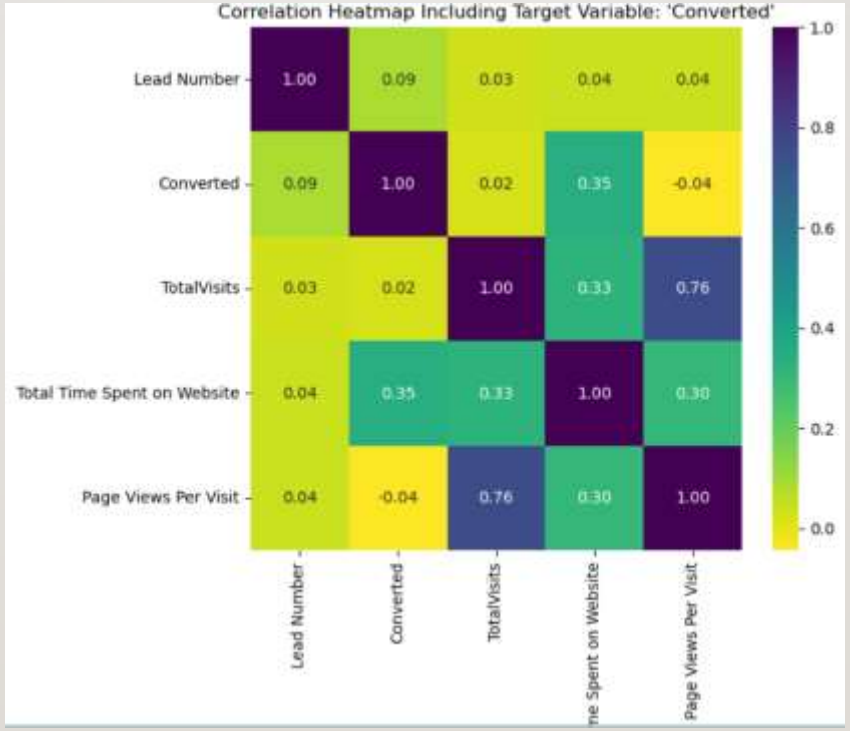
BIVARIATE ANALYSIS - WITH NUMERICAL FEATURES



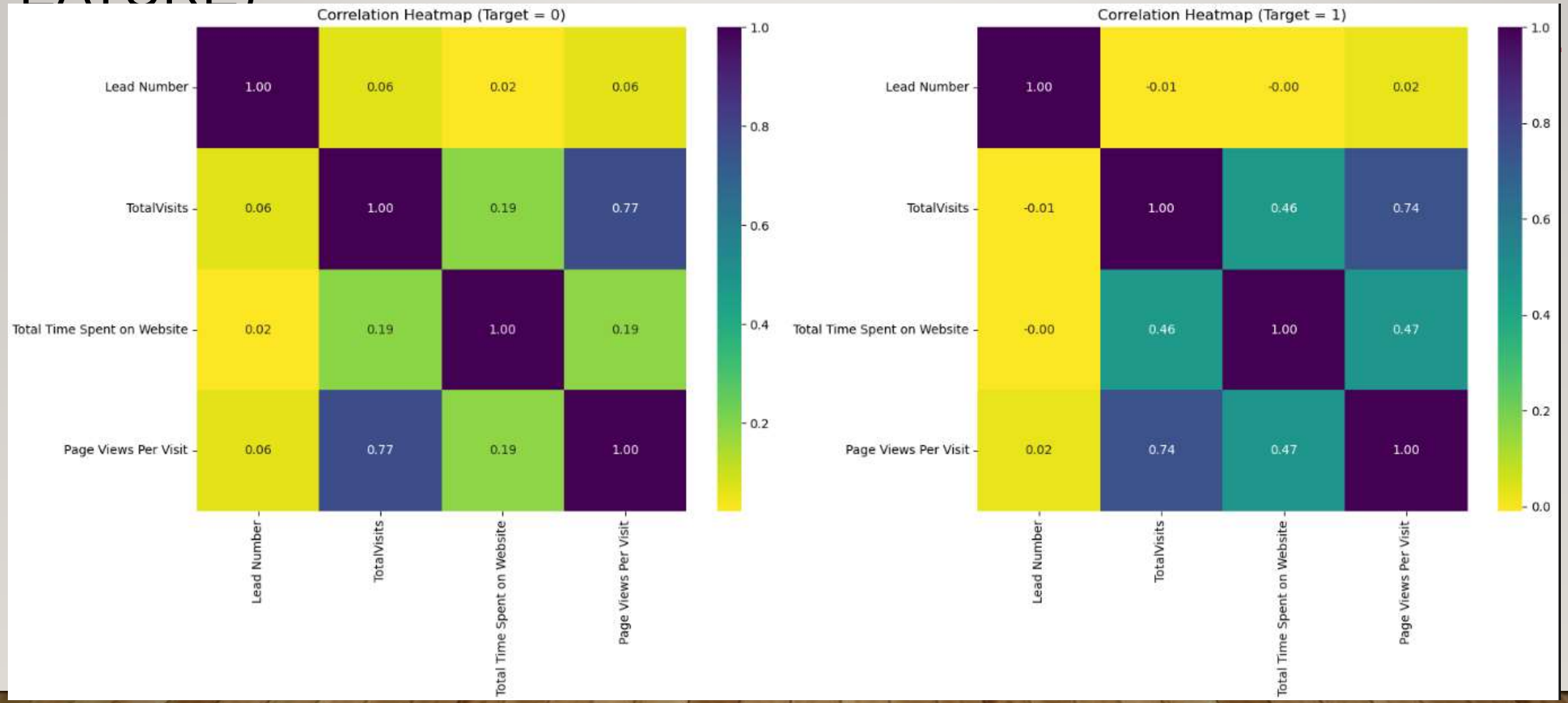
TOTAL TIME SPENT ON WEBSITE VS CONVERTED



MULTIVARIATE ANALYSIS



CORRELATION HEATMAP - (SEGREGATED BY TARGET FEATURE)

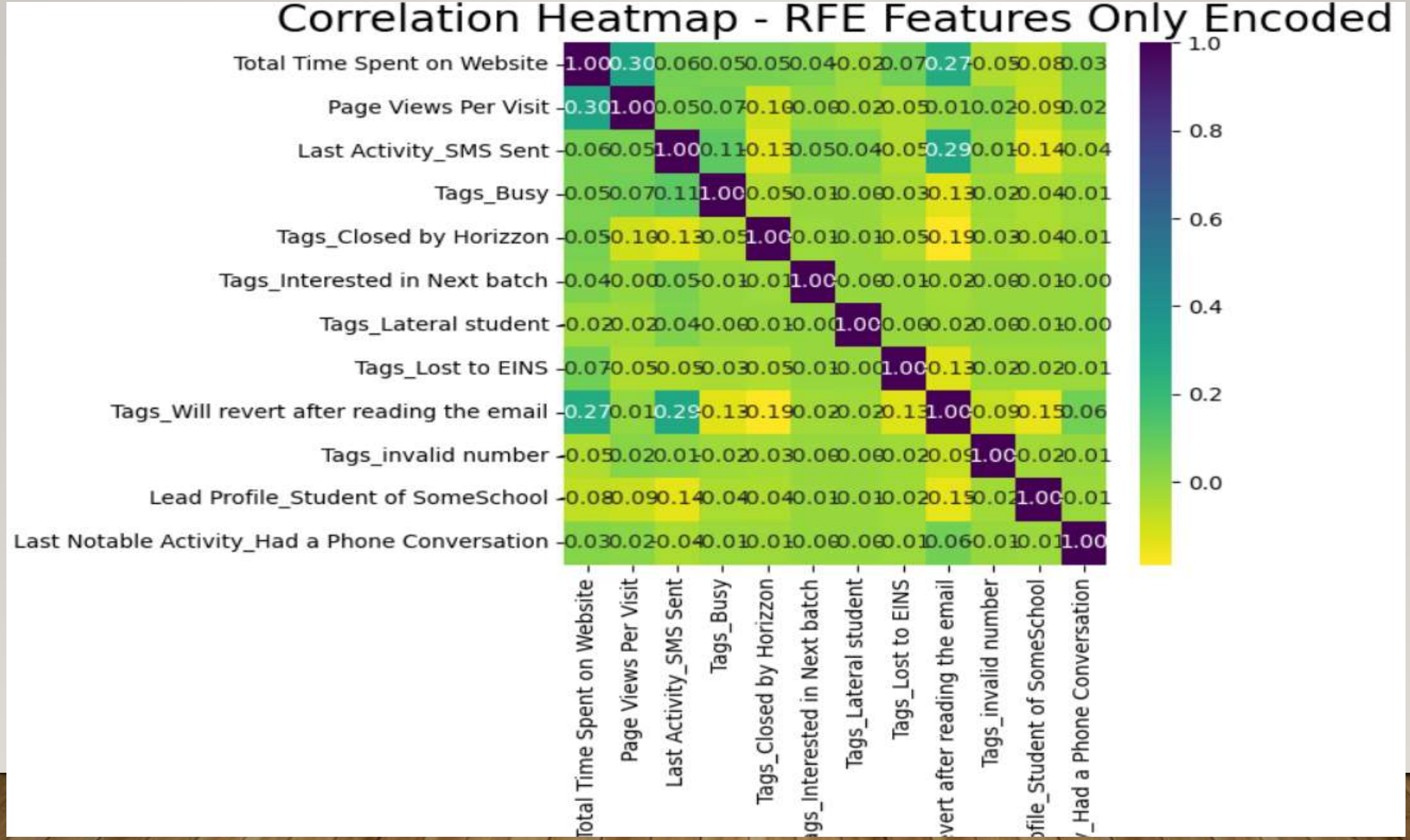


FINDINGS FROM CORRELATION

- There was high correlation noticed between Page Views Per Visit & Total Time Spent on Website
- There was good Correlation also be noticed between Total Time Spent on Website & Converted
- Those who show a strong interest in buying an education program are likely to visit the website more often and spend more time exploring the programs

MODEL BUILDING USING LOGISTIC REGRESSION

- CREATED DUMMY VARIABLES FOR CATEGORICAL COLUMNS
- APPLYING TRAIN TEST SPLIT ON THE DATASET IN THE RATION 7:3
- APPLYING MINMAXSCALER ON NUMERICAL COLUMNS (EXCLUDING LEAD NUMBER)
- USING RFE TECHNIQUE TO ELIMINATE FEATURES
- CHOOSING A MODEL WHOSE P-VALUES AND VIF ARE ACCEPTABLE

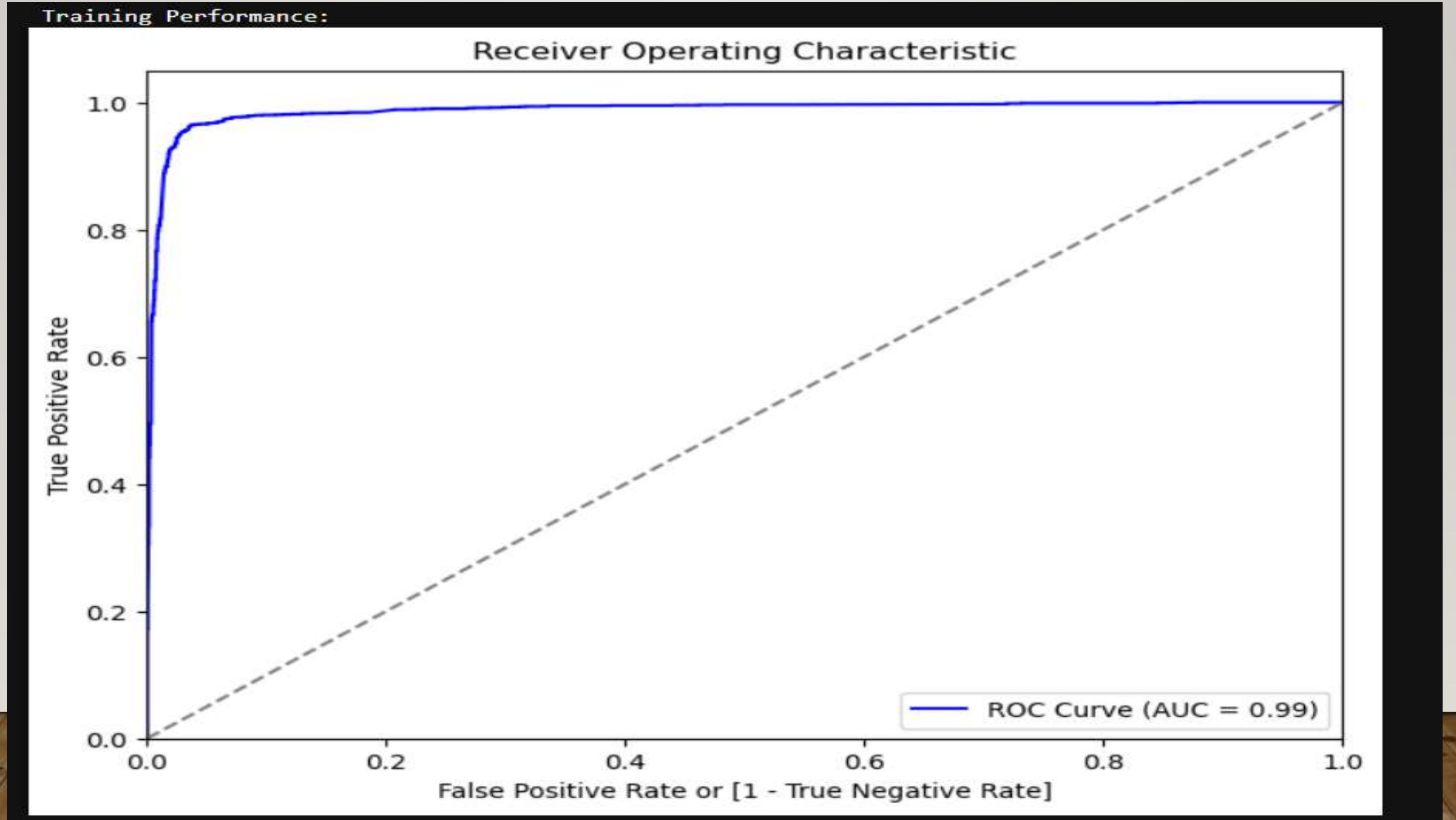

Correlation Heatmap - RFE Features Only Encoded
Total Time Spent on Website
Page Views Per Visit
Last Activity_SMS Sent
Tags_Busy
Tags_Closed by Horizzon
Tags_Interested in Next batch
Tags_Lateral student
Tags_Lost to EINS
Tags_Will revert after reading the email
Tags_invalid number
Lead Profile_Student of SomeSchool
Last Notable Activity_Had a Phone Conversation
1.0
0.8
0.6
0.4
0.2
0.0

MODEL EVALUATION

```
Training Performance:
              precision    recall  f1-score   support

           0       0.96      0.97      0.96      2502
           1       0.96      0.96      0.96      2207

    accuracy                           0.96      4709
   macro avg       0.96      0.96      0.96      4709
weighted avg       0.96      0.96      0.96      4709

Confusion Matrix (Training):
[[2416   86]
 [  90 2117]]
```

PLOT ROC AUC

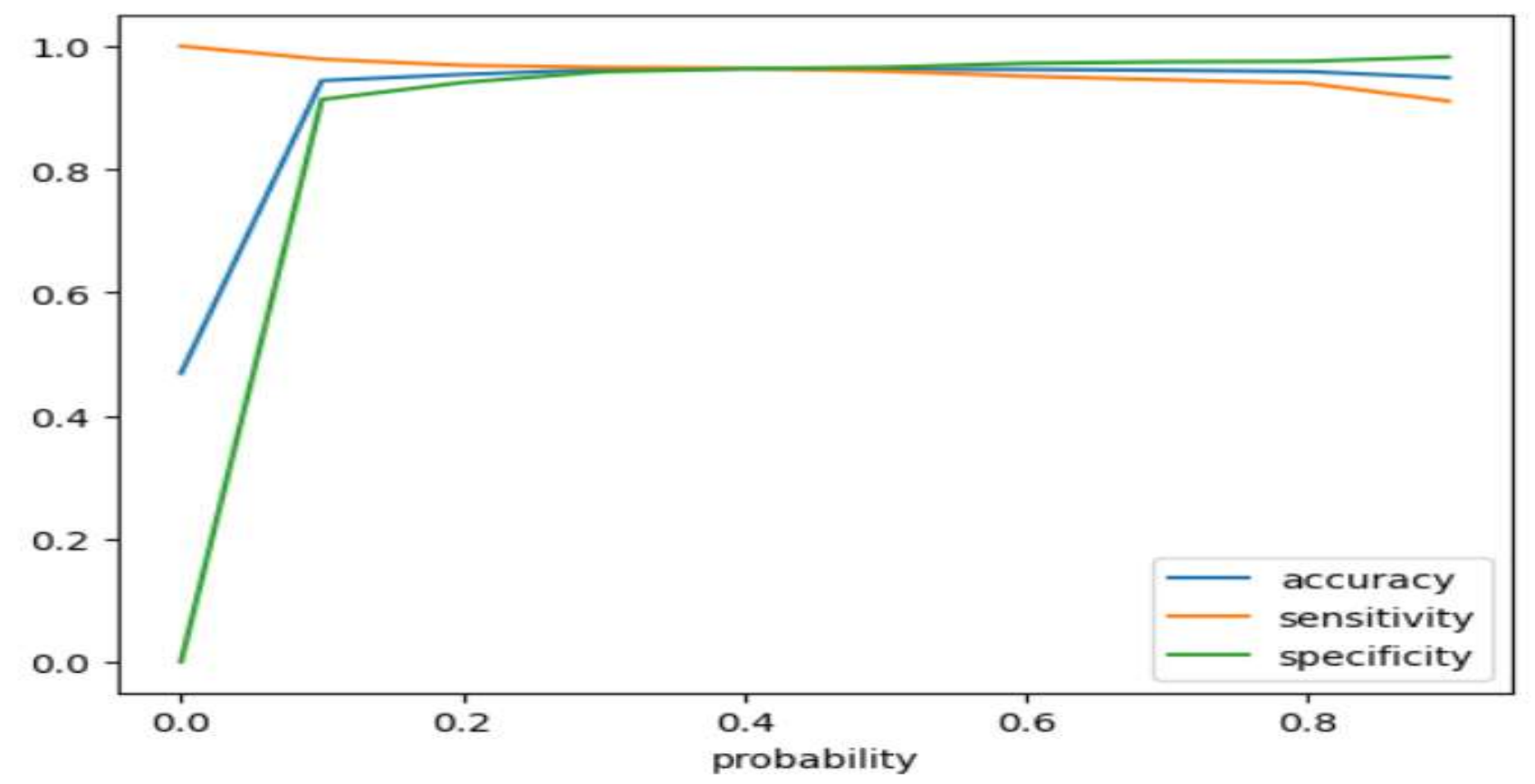
FOR THE ABOVE EVALUATION, WE CHOSE A THRESHOLD VALUE OF 0.5. NEXT, WE WILL ATTEMPT TO DETERMINE THE OPTIMAL THRESHOLD VALUE



PROBABILTY, ACCURACY, SENSITIVITY, AND SPECIFICITY AT DIFFERENT CUT-OFFS

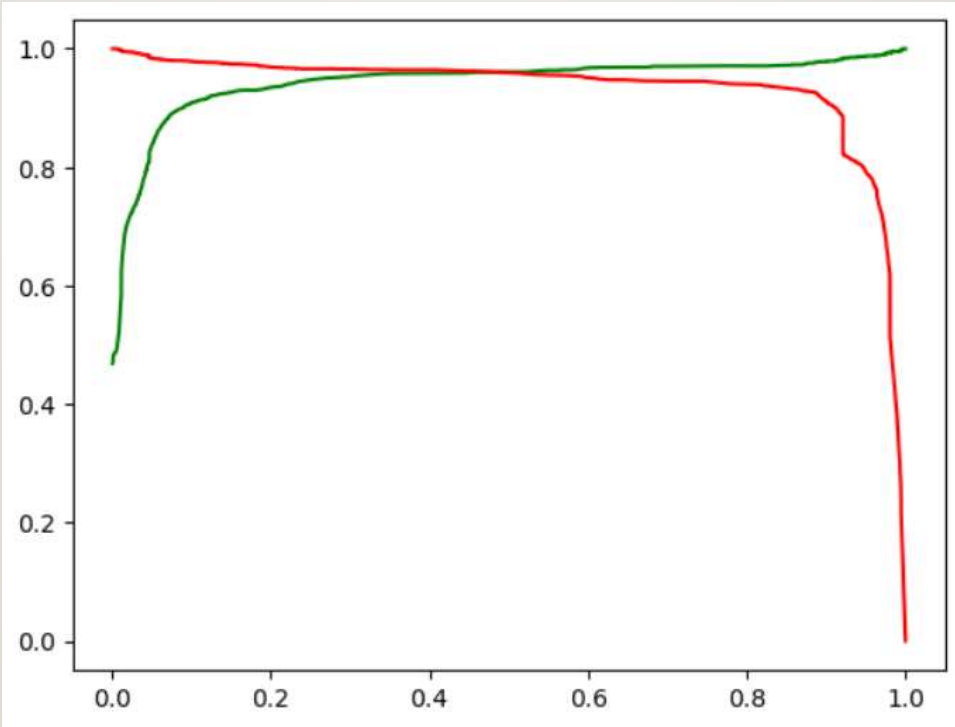
	probability	accuracy	sensitivity	specificity
0.000	0.000	0.469	1.000	0.000
0.100	0.100	0.944	0.979	0.913
0.200	0.200	0.954	0.969	0.941
0.300	0.300	0.962	0.966	0.958
0.400	0.400	0.964	0.964	0.963
0.500	0.500	0.963	0.959	0.966
0.600	0.600	0.962	0.951	0.972
0.700	0.700	0.961	0.945	0.974
0.800	0.800	0.959	0.940	0.975
0.900	0.900	0.949	0.911	0.982

VISUALISATION OF THE METRICS



PRECISION & RECALL TRADEOFF

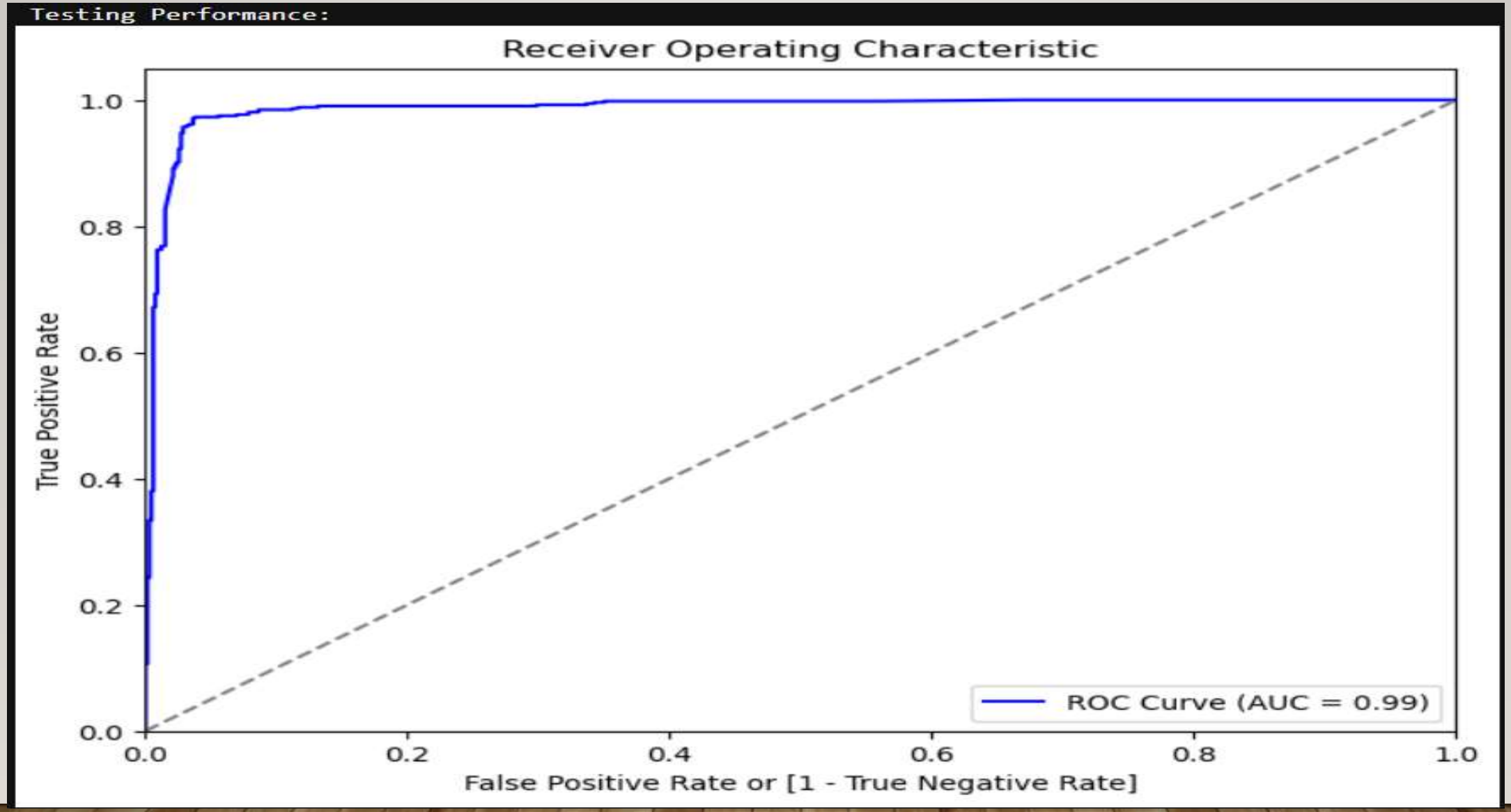
AS PER THE ABOVE CURVE, WE CAN SEE THAT THE OPTIMAL CUT-OFF IS AROUND 0.45



PREDICTIONS ON TEST SET

```
Testing Performance:
              precision    recall  f1-score   support

           0       0.98      0.96      0.97       660
           1       0.95      0.97      0.96       518

    accuracy                           0.97      1178
   macro avg       0.97      0.97      0.97      1178
weighted avg       0.97      0.97      0.97      1178

Confusion Matrix (Testing):
[[635  25]
 [ 14 504]]
```

Testing Performance:



CONCLUSION

- Leads who are more likely to convert are –
  - The ones who spent more on website
  - Who visits the the page often
  - When the lead source was google, direct traffic, organic search
  - When the last activity was sms
  - Focus more on working professionals as they will have higher budgets to spend on fees