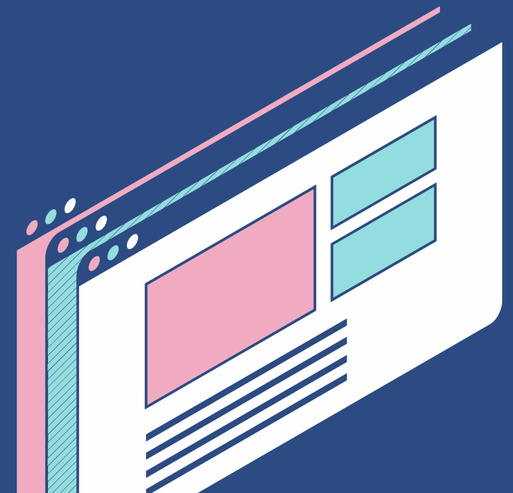
LEAD SCORING CASE STUDY



Presented by- Samiksha Yadav

Problem Statement

X EDUCATION WANTS TO BUILD A MACHINE LEARNING MODEL WHERE THEY ASSIGN A LEAD SCORE TO EACH LEAD SUCH THAT THE CUSTOMERS WITH A HIGHER LEAD SCORE HAVE A HIGHER CONVERSION PROBABILITY. THE BUSINESS REQUIREMENT IS TO INCREASE THE LEAD CONVERSION RATE TO AROUND 80%.

ASSUMPTIONS :

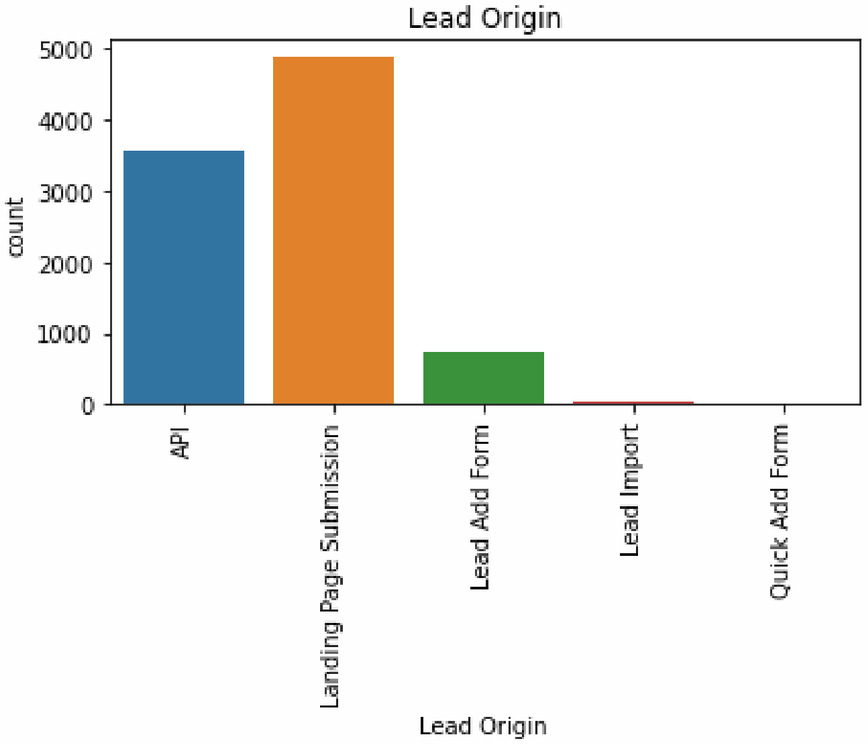
WE REMOVE VARIABLES THAT HAVE UNIQUE VALUES (LIKE 'PROSPECT ID') OR SINGLE VALUES BECAUSE THEY DON'T GIVE US USEFUL INFORMATION. WE ALSO IGNORE VARIABLES WITH A LOT OF MISSING DATA OR IMBALANCED DATA BECAUSE THEY CAN INTRODUCE BIASES AND MAKE OUR ANALYSIS LESS RELIABLE.

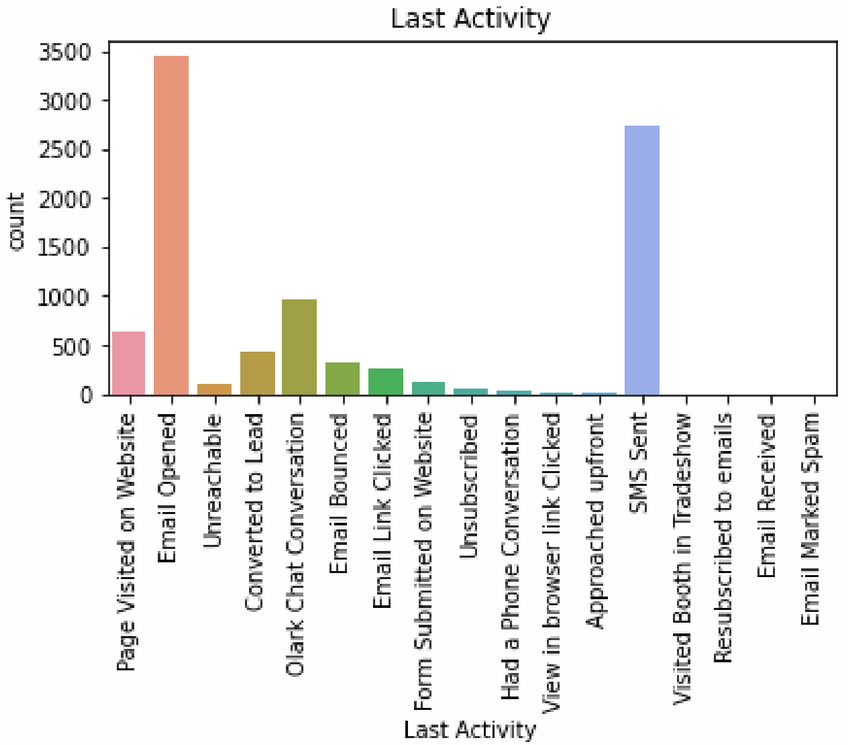
APPROACH

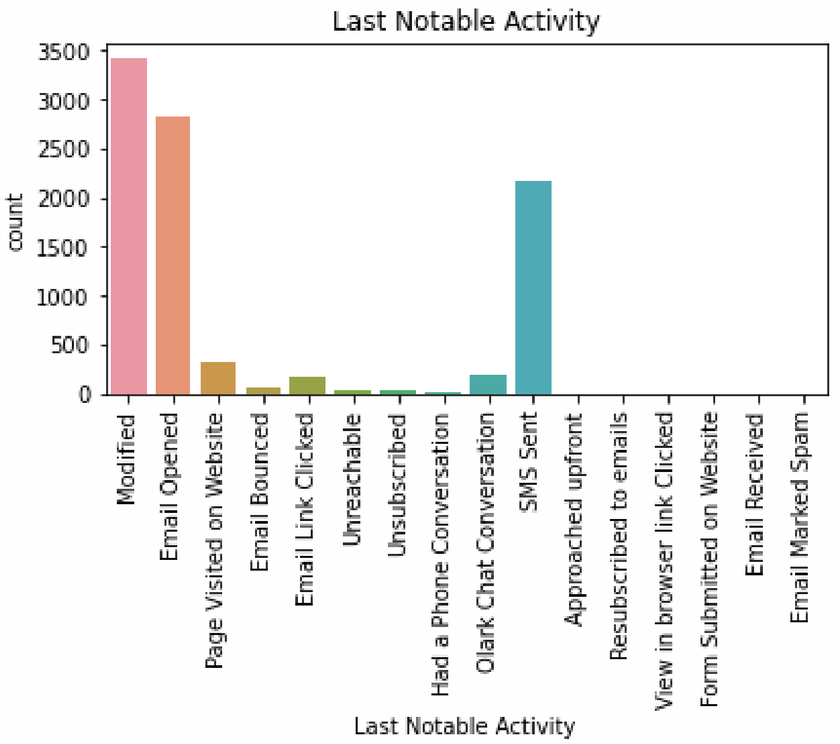
1. Data Cleaning and Data Manipulation
2. Exploratory Data Analysis
3. Data Pre-processing
4. Model Building- Logistic Regression Model
5. Model Evaluation
6. Predictions
7. Conclusions and Recommendations

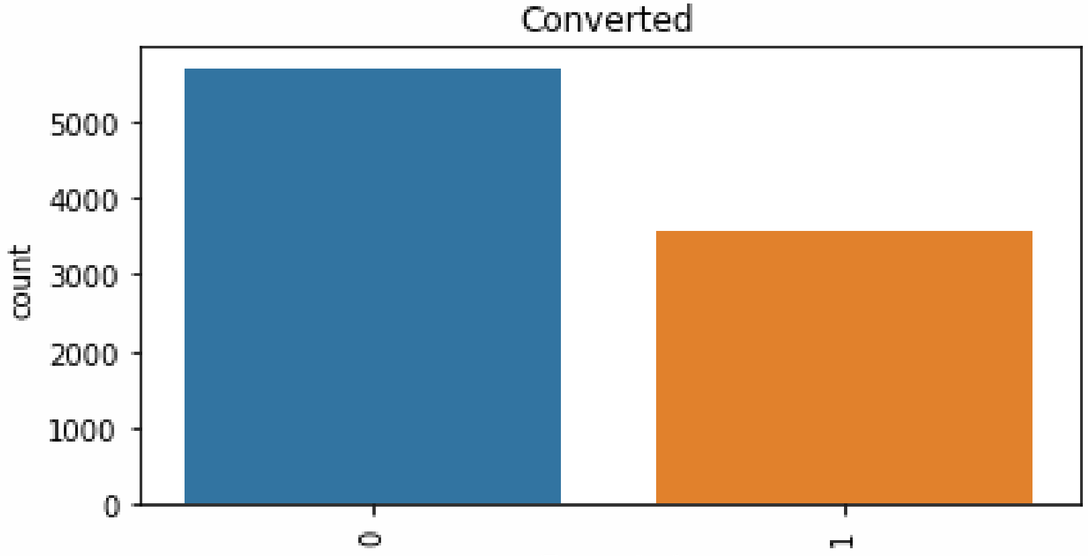
EDA Conclusion

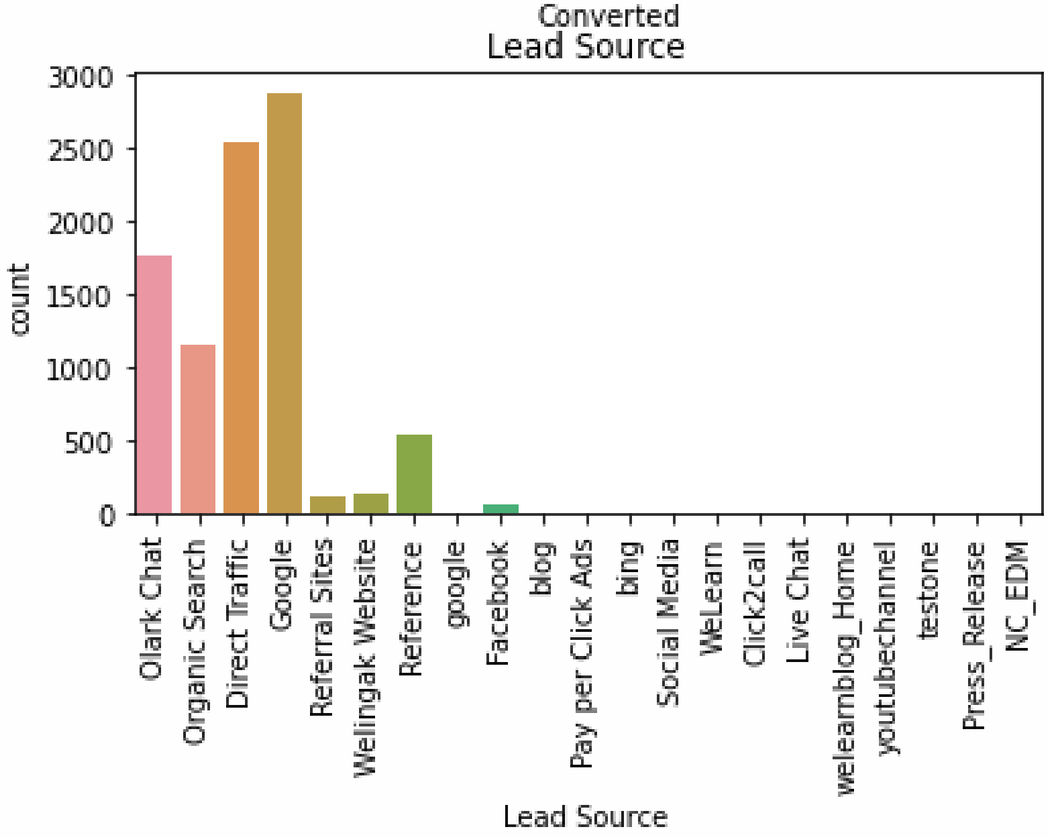
Key Variables with Significant Information about the Dataset

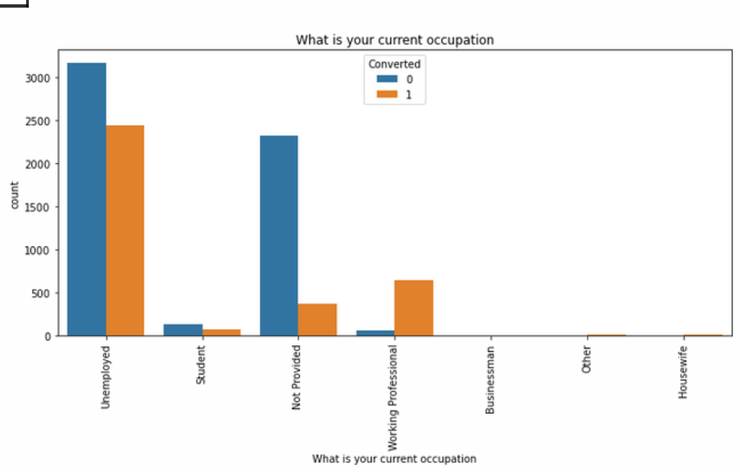


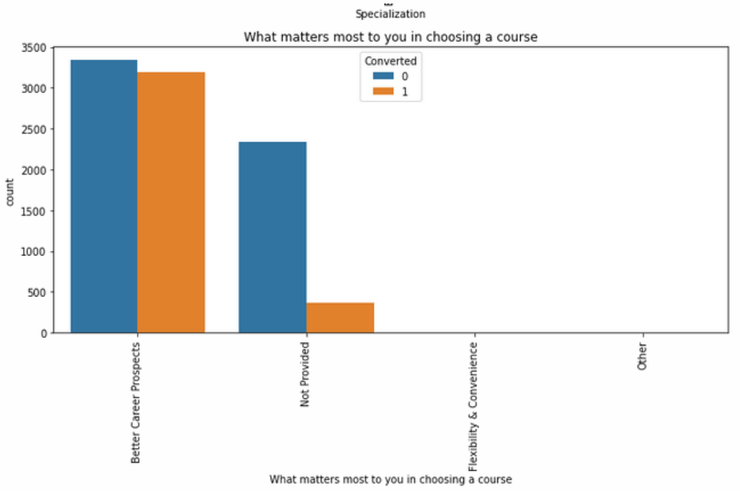


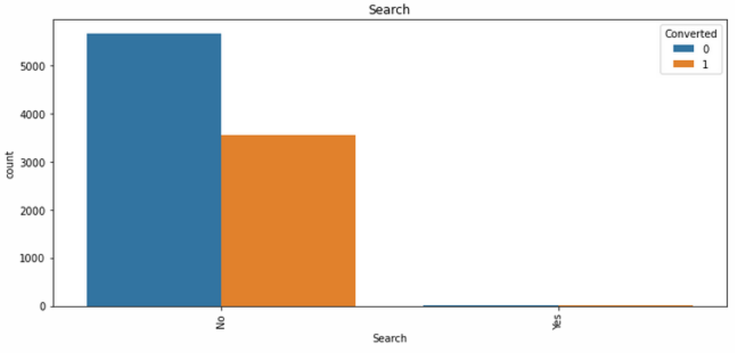


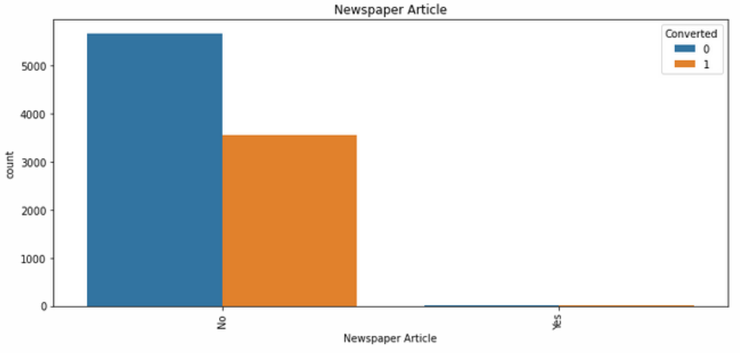

Converted
count
0
1000
2000
3000
4000
5000
0
1


Converted
Lead Source
count
0
500
1000
1500
2000
2500
3000
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Welingak Website
Reference
google
Facebook
blog
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
weleamblog_Home
youtubechannel
testone
Press_Release
NC_EDM
Lead Source


What is your current occupation
Converted
0
1
count
0
500
1000
1500
2000
2500
3000
Unemployed
Student
Not Provided
Working Professional
Businessman
Other
Housewife
What is your current occupation

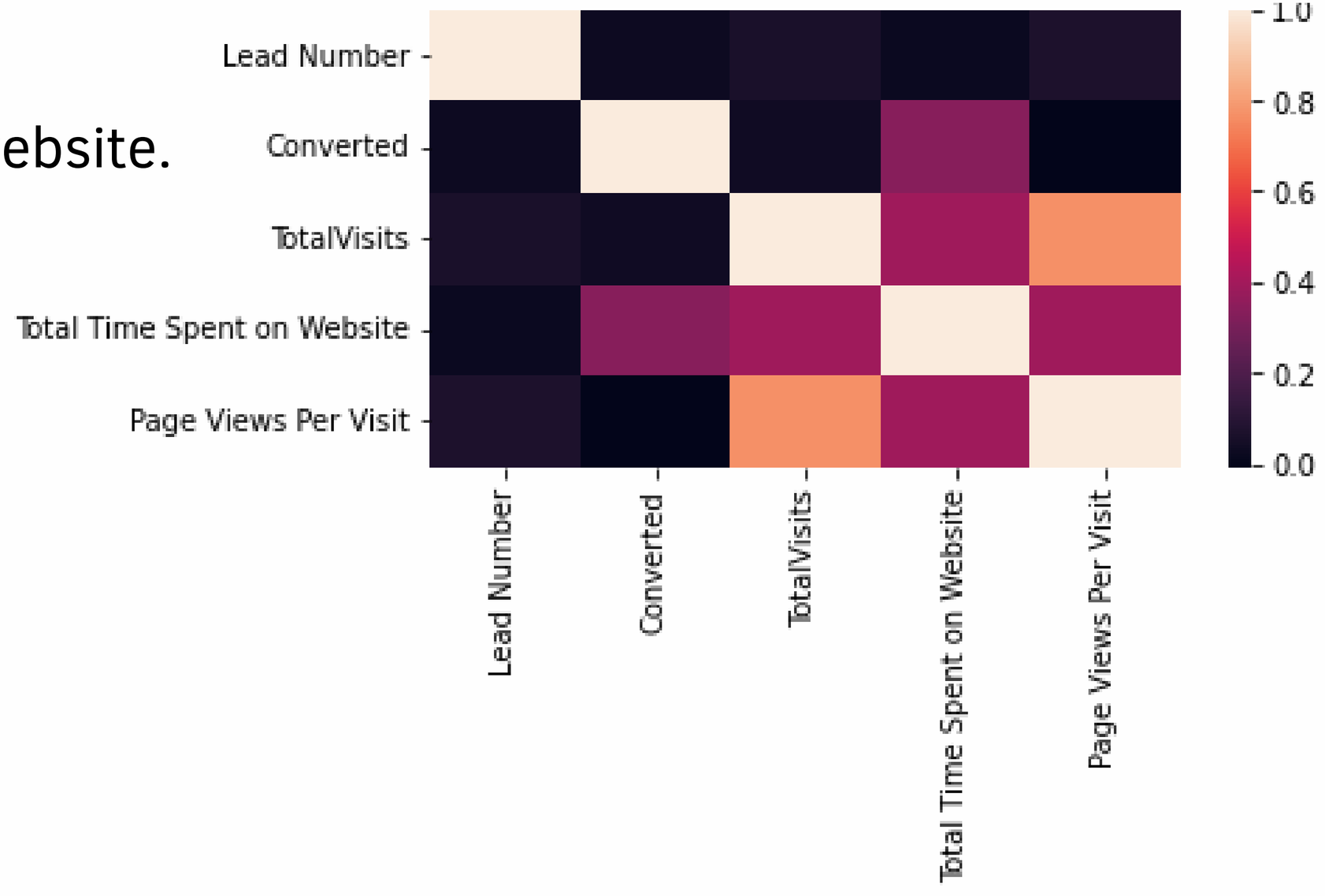

Specialization
What matters most to you in choosing a course
Converted
0
1
count
0
500
1000
1500
2000
2500
3000
3500
Better Career Prospects
Not Provided
Flexibility & Convenience
Other
What matters most to you in choosing a course


Search
Converted
0
1
count
0
1000
2000
3000
4000
5000
No
Yes
Search

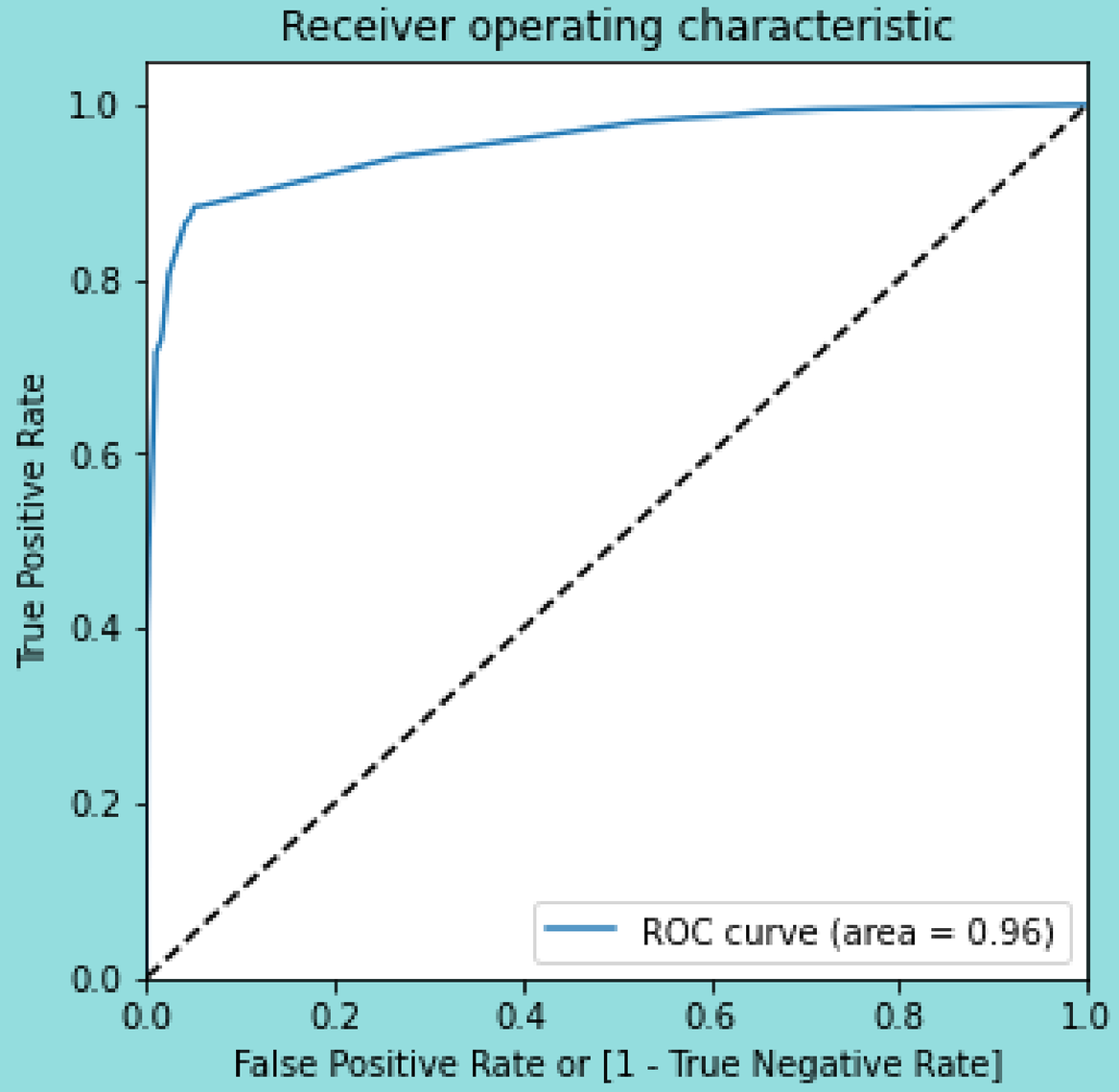

Newspaper Article
Converted
0
1
count
0
1000
2000
3000
4000
5000
No
Yes
Newspaper Article

The correlation matrix reveals strong correlations between 'Converted' and the following variables:

- Total visits
- Page views per visit
- Time spent on the website.



RESULT



ROC CURVE

- THE ROC CURVE SHOWS THAT 96% OF THE AREA IS UNDER THE CURVE.
- THE MODEL'S CLASSIFICATION PROBABILITY FOR LEAD CONVERSION IS EXCEPTIONALLY HIGH.

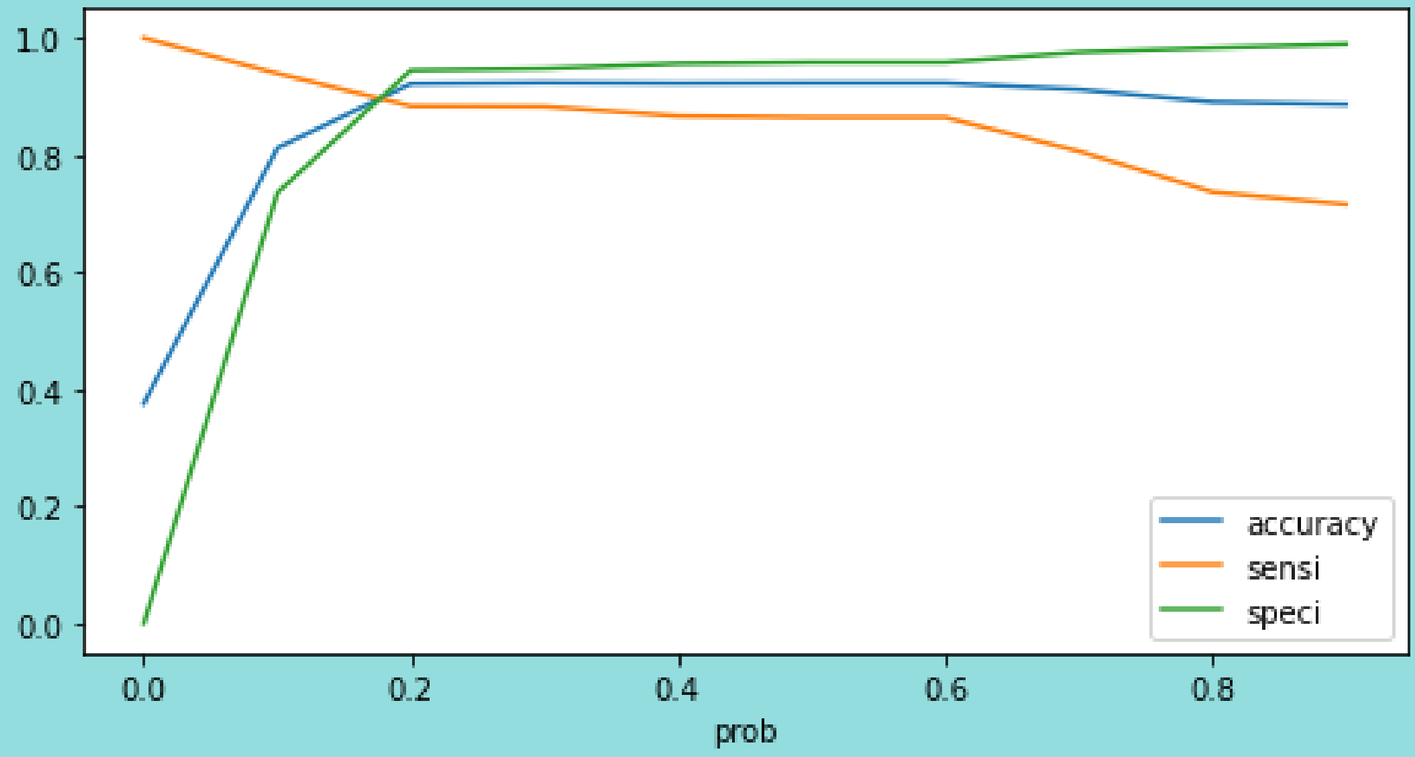
Optimal Probability Cut-off

With 0.2 cut-off, the model has:

Accuracy – 92%

Sensitivity – 88%

Specificity – 94%



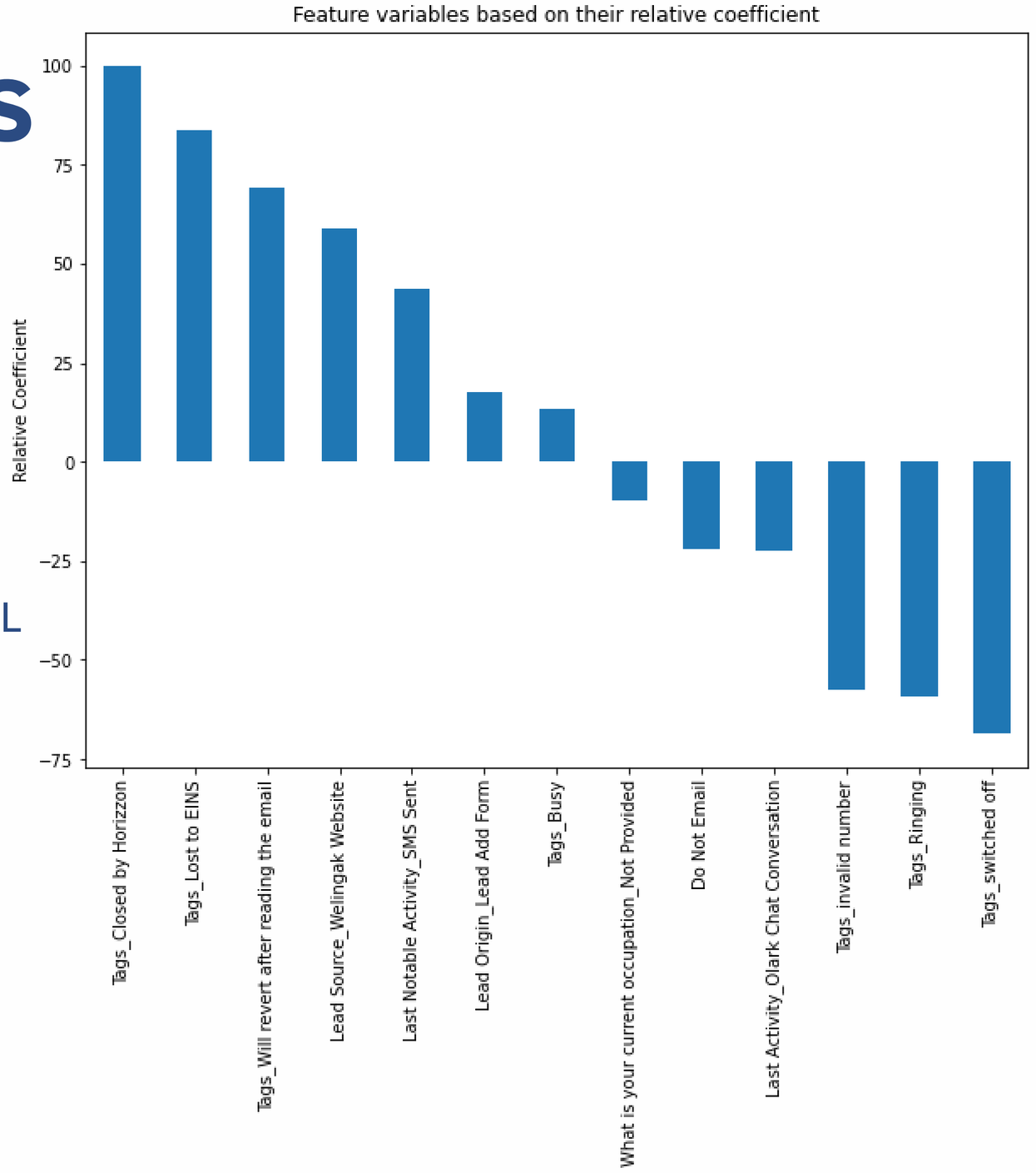
Important Features

TOP 3 VARIABLES WITH HIGH LEAD CONVERSION PROBABILITY:

- TAGS: CLOSED BY HORIZZON, LOST TO EINS, WILL REVERT AFTER READING THE EMAIL
- LEAD SOURCE: WELINGAK WEBSITE
- LAST NOTABLE ACTIVITY: SMS SENT

TOP 3 VARIABLES NEEDING IMPROVEMENT IN CONVERTING QUALITY LEADS:

- TAGS: INVALID NUMBER
- TAGS: RINGING
- TAGS: SWITCHED OFF



RECOMMENDATIONS

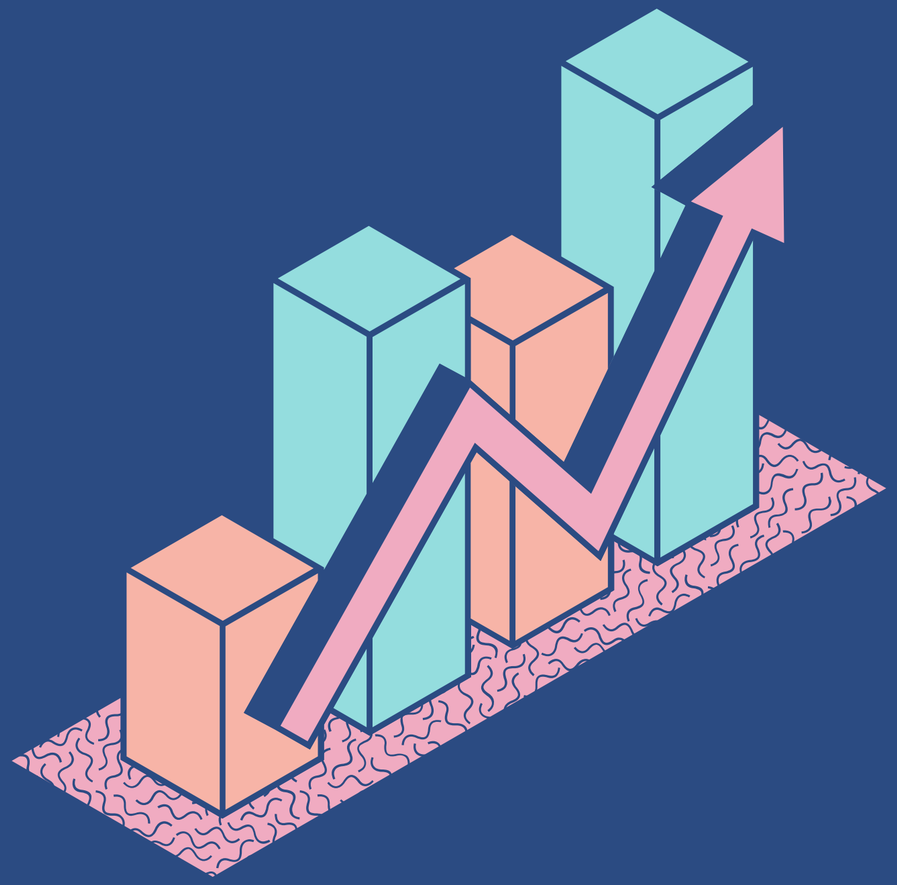
Key learnings from this assignment include:

·Understanding the process of data exploration and handling missing values.

·Recognizing the importance of performing EDA and data pre-processing.

·Implementing a systematic approach for model building and feature selection, considering the impact on both training and test datasets.

Successfully solving problems through teamwork and leveraging individual strengths.



THANK YOU