LEAD CONVERSION ANALYSIS

UTSAV KUMAR

OBJECTIVE

1. EDA
2. DUMMY MODEL AND DATA SPLITITNG
3. MODEL TRAINING
4. MODEL EVALUATION
5. PREDICTION AND RECALL

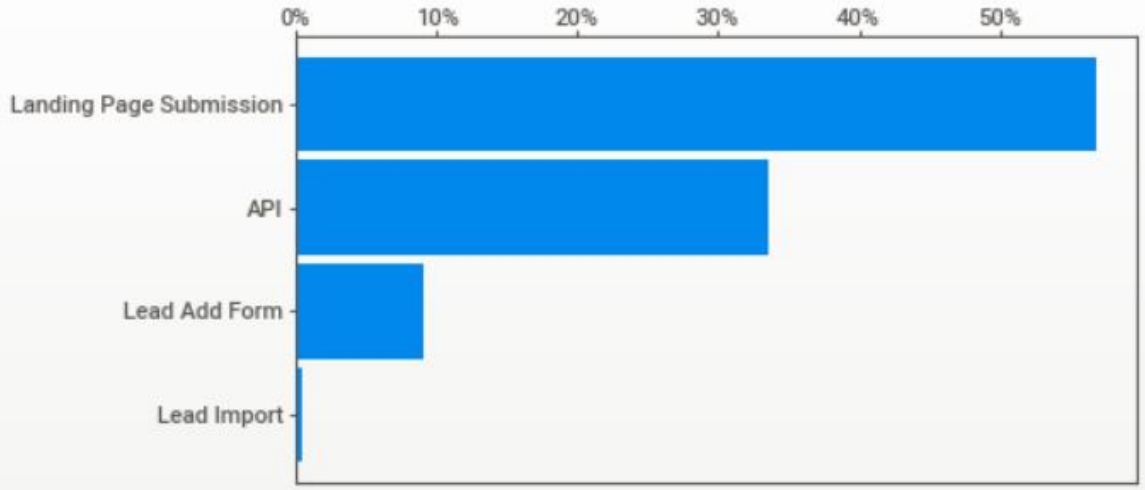
EDA

After performing all missing values and irregularities removed, we used an auto EDA method (SWEETVIZ) which helps in autocreation and proper alignment of data with respective graphs (mostly count boxplots) to give a better understanding.
Note this is just for visualization and all the cleaning and EDA processes are done manually.
The forward graphs which are created can be used for process determination.

EDA

Lead Origin

MISSING: ---



TOP CATEGORIES

Landing Page Submission	3,625	57%
API	2,140	34%
Lead Add Form	581	9%
Lead Import	27	<1%
ALL	6,373	100%

CATEGORICAL ASSOCIATIONS
(UNCERTAINTY COEFFICIENT, 0 to 1)

Lead Origin PROVIDES INFORMATION ON...

Lead Source	0.35
A free copy of Mastering The ...	0.28
Specialization	0.13
Converted	0.07
Last Activity	0.04
What is your current occupati...	0.03
Do Not Email	0.01
Last Notable Activity	0.01

THESE FEATURES GIVE INFORMATION ON Lead Origin:

Lead Source	0.61
Specialization	0.32
A free copy of Mastering The ...	0.19
Last Activity	0.06
Converted	0.05
Last Notable Activity	0.02
What is your current occupati...	0.01
Do Not Email	0.00

NUMERICAL ASSOCIATIONS
(CORRELATION RATIO, 0 to 1)

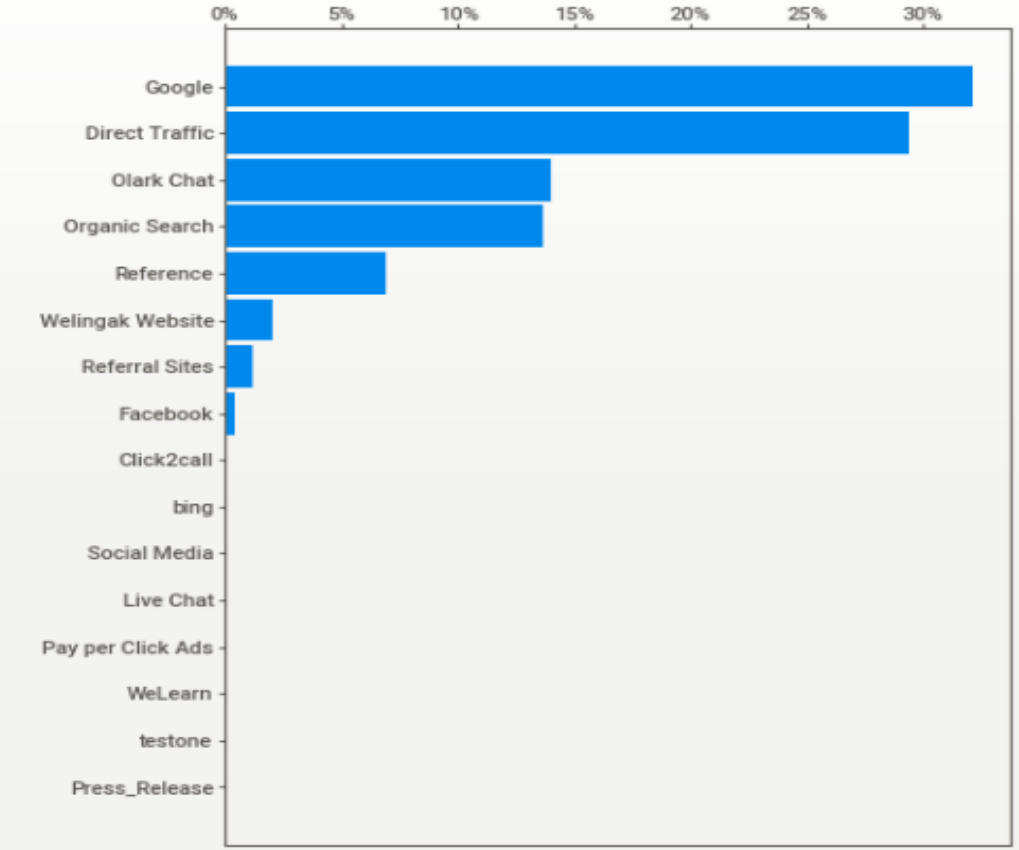
Lead Origin CORRELATION RATIO WITH...

Page Views Per Visit	0.50
Total Time Spent on Website	0.32
TotalVisits	0.29

EDA

Lead Source

MISSING: ---



TOP CATEGORIES

Google	2,048	32%
Direct Traffic	1,873	29%
Olark Chat	892	14%
Organic Search	870	14%
Reference	443	7%
Welingak Website	129	2%
Referral Sites	75	1%
Facebook	28	<1%
Click2call	4	<1%
bing	3	<1%
Social Media	2	<1%
Live Chat	2	<1%
Pay per Click Ads	1	<1%
WeLearn	1	<1%
testone	1	<1%
Press_Release	1	<1%
ALL	6,373	100%

CATEGORICAL ASSOCIATIONS
(UNCERTAINTY COEFFICIENT, 0 to 1)

Lead Source PROVIDES INFORMATION ON...

Lead Origin	0.61
A free copy of Mastering The ...	0.37
Converted	0.08
Specialization	0.08
What is your current occupati...	0.05
Last Activity	0.05
Do Not Email	0.03
Last Notable Activity	0.02

THESE FEATURES GIVE INFORMATION ON Lead Source:

Lead Origin	0.35
A free copy of Mastering The ...	0.15
Specialization	0.11
Last Activity	0.05
Converted	0.03
What is your current occupati...	0.02
Last Notable Activity	0.01
Do Not Email	0.00

NUMERICAL ASSOCIATIONS
(CORRELATION RATIO, 0 to 1)

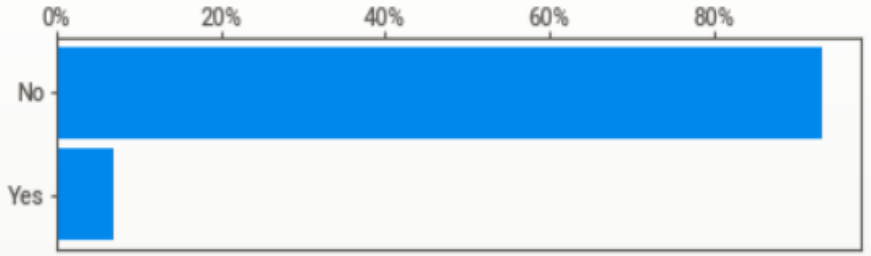
Lead Source CORRELATION RATIO WITH...

Page Views Per Visit	0.62
Total Time Spent on Website	0.45
TotalVisits	0.38

EDA

Do Not Email

MISSING: ---



TOP CATEGORIES

No	5,938	93%
Yes	435	7%
ALL	6,373	100%

CATEGORICAL ASSOCIATIONS
(UNCERTAINTY COEFFICIENT, 0 to 1)

Do Not Email PROVIDES INFORMATION ON...

Last Activity	0.07
Last Notable Activity	0.04
Converted	0.02
Lead Origin	0.00
Lead Source	0.00
What is your current occupati...	0.00
Specialization	0.00
A free copy of Mastering The ...	0.00

THESE FEATURES GIVE INFORMATION ON Do Not Email:

Last Activity	**0.43**
Last Notable Activity	**0.23**
Converted	0.05
Lead Source	0.03
Specialization	0.02
Lead Origin	0.01
What is your current occupati...	0.01
A free copy of Mastering The ...	0.01

NUMERICAL ASSOCIATIONS
(CORRELATION RATIO, 0 to 1)

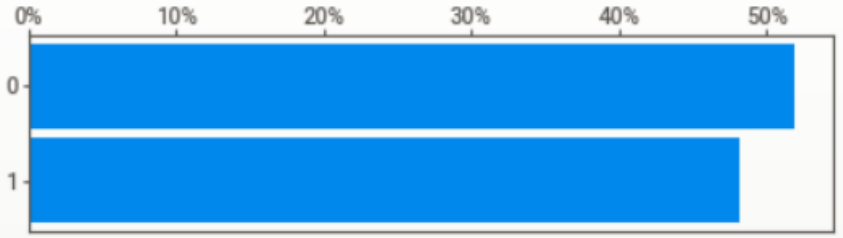
Do Not Email CORRELATION RATIO WITH...

Total Time Spent on Website	0.05
Page Views Per Visit	0.04
TotalVisits	0.03

EDA

Converted

MISSING: ---



TOP CATEGORIES

0	3,308	52%
1	3,065	48%
ALL	6,373	100%

CATEGORICAL ASSOCIATIONS
(UNCERTAINTY COEFFICIENT, 0 to 1)

Converted PROVIDES INFORMATION ON...

What is your current occupati...	0.11
Lead Origin	0.05
Do Not Email	0.05
Last Activity	0.04
Last Notable Activity	0.04
Lead Source	0.03
A free copy of Mastering The ...	0.01
Specialization	0.00

THESE FEATURES GIVE INFORMATION ON Converted:

Last Activity	0.10
Last Notable Activity	0.08
What is your current occupati...	0.08
Lead Source	0.08
Lead Origin	0.07
Do Not Email	0.02
Specialization	0.01
A free copy of Mastering The ...	0.01

NUMERICAL ASSOCIATIONS
(CORRELATION RATIO, 0 to 1)

Converted CORRELATION RATIO WITH...

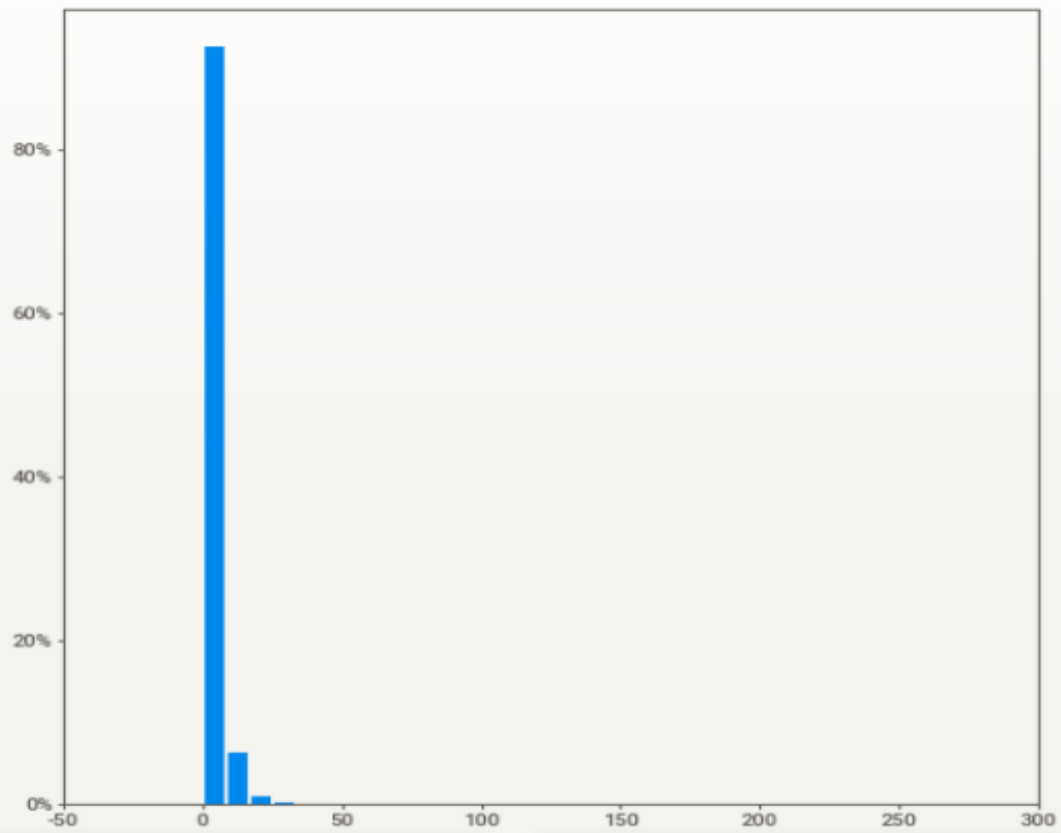
Total Time Spent on Website	0.31
Page Views Per Visit	0.06
TotalVisits	0.01

EDA

TotalVisits

MISSING: ---

Auto | 5 | 15 | 30



NUMERICAL ASSOCIATIONS
(PEARSON, -1 to 1)

Page Views Per Visit	0.49
Total Time Spent on Website	0.20

CATEGORICAL ASSOCIATIONS
(CORRELATION RATIO, 0 to 1)

Lead Source	0.38
Lead Origin	0.29
Last Notable Activity	0.24
Last Activity	0.24
Specialization	0.23
A free copy of Mastering The ...	0.19
What is your current occupati...	0.05
Do Not Email	0.03
Converted	0.01

MOST FREQUENT VALUES

0.0	1,347	21.1%
2.0	1,225	19.2%
3.0	937	14.7%
4.0	823	12.9%
5.0	578	9.1%
6.0	337	5.3%
1.0	260	4.1%
7.0	225	3.5%
8.0	162	2.5%
9.0	128	2.0%
10.0	76	1.2%
11.0	70	1.1%
13.0	39	0.6%
12.0	31	0.5%
14.0	28	0.4%

SMALLEST VALUES

0.0	1,347	21.1%
1.0	260	4.1%
2.0	1,225	19.2%
3.0	937	14.7%
4.0	823	12.9%
5.0	578	9.1%
6.0	337	5.3%
7.0	225	3.5%
8.0	162	2.5%
9.0	128	2.0%
10.0	76	1.2%
11.0	70	1.1%
12.0	31	0.5%
13.0	39	0.6%
14.0	28	0.4%

LARGEST VALUES

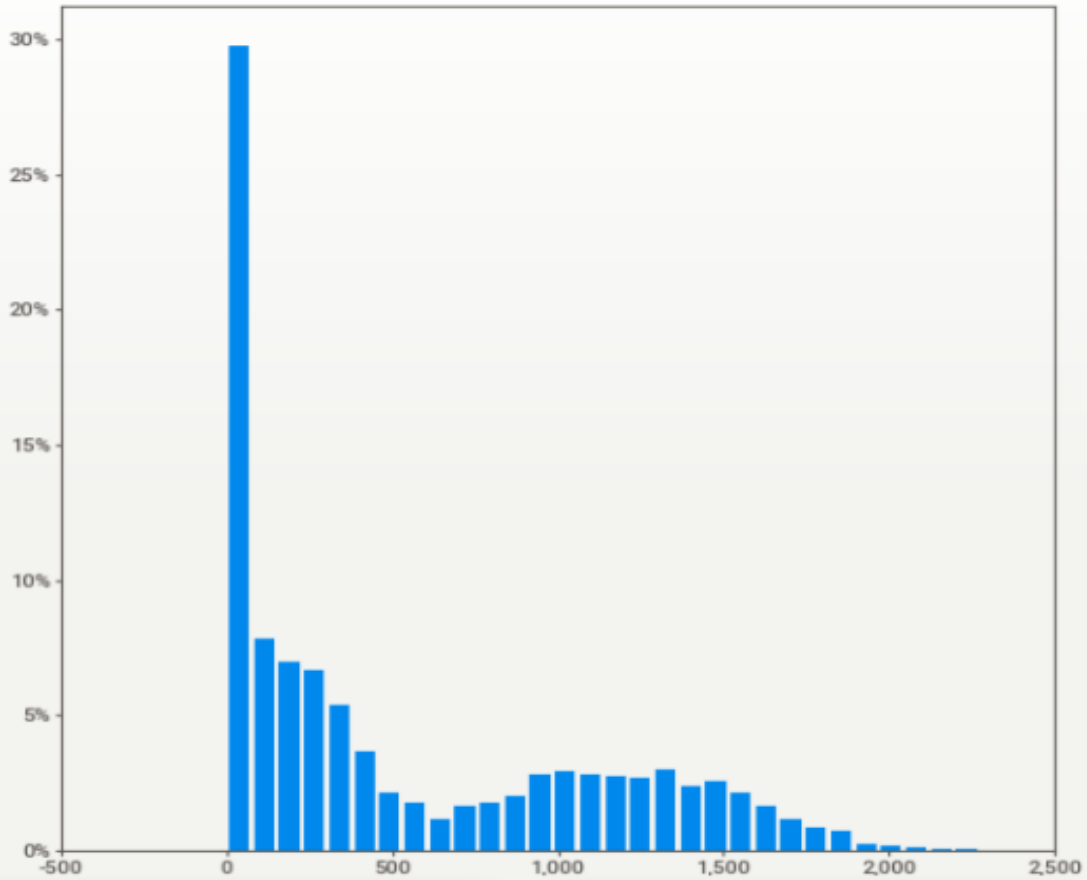
251.0	1	<0.1%
115.0	1	<0.1%
74.0	1	<0.1%
55.0	1	<0.1%
43.0	1	<0.1%
42.0	1	<0.1%
32.0	1	<0.1%
30.0	1	<0.1%
29.0	2	<0.1%
28.0	1	<0.1%
27.0	5	<0.1%
26.0	2	<0.1%
25.0	3	<0.1%
24.0	3	<0.1%
23.0	5	<0.1%

EDA

Total Time Spent on Website

MISSING: ---

Auto | 5 | 15 | 30



NUMERICAL ASSOCIATIONS
(PEARSON, -1 to 1)

Page Views Per Visit	0.30
TotalVisits	0.20

CATEGORICAL ASSOCIATIONS
(CORRELATION RATIO, 0 to 1)

Lead Source	0.45
Lead Origin	0.32
Converted	0.31
Specialization	0.26
A free copy of Mastering The ...	0.15
Last Activity	0.15
Last Notable Activity	0.13
What is your current occupati...	0.09
Do Not Email	0.05

MOST FREQUENT VALUES

0	1,351	21.2%
60	15	0.2%
127	14	0.2%
96	13	0.2%
87	13	0.2%
74	13	0.2%
62	12	0.2%
94	12	0.2%
69	12	0.2%
234	11	0.2%
36	11	0.2%
176	11	0.2%
129	11	0.2%
213	11	0.2%
158	11	0.2%

SMALLEST VALUES

0	1,351	21.2%
1	6	<0.1%
2	10	0.2%
3	8	0.1%
4	7	0.1%
5	10	0.2%
6	3	<0.1%
7	7	0.1%
8	5	<0.1%
9	7	0.1%
10	8	0.1%
11	5	<0.1%
12	10	0.2%
13	5	<0.1%
14	11	0.2%

LARGEST VALUES

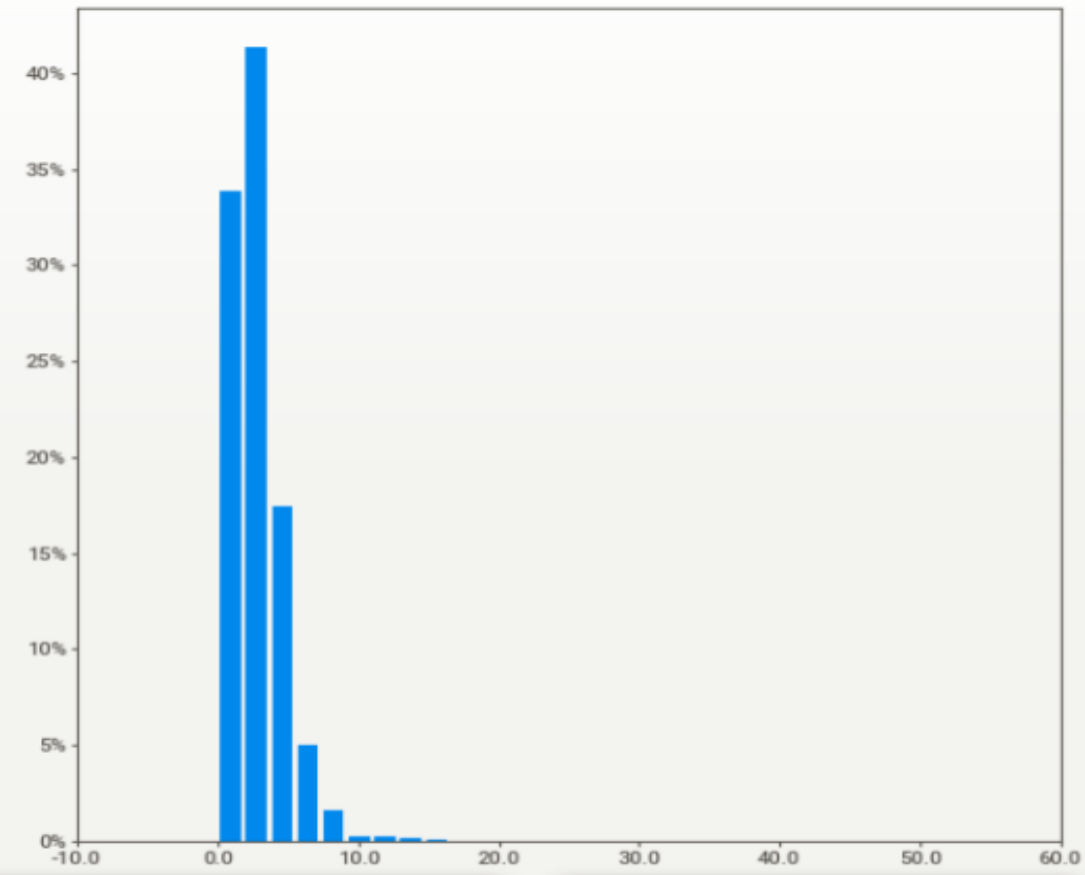
2272	1	<0.1%
2253	1	<0.1%
2226	1	<0.1%
2170	1	<0.1%
2140	1	<0.1%
2137	1	<0.1%
2125	1	<0.1%
2111	1	<0.1%
2094	1	<0.1%
2090	1	<0.1%
2069	1	<0.1%
2059	1	<0.1%
2058	1	<0.1%
2037	1	<0.1%
2020	1	<0.1%

EDA

Page Views Per Visit

MISSING: ---

Auto | 5 | 15 | 30



NUMERICAL ASSOCIATIONS
(PEARSON, -1 to 1)

TotalVisits	0.49
Total Time Spent on Website	0.30

CATEGORICAL ASSOCIATIONS
(CORRELATION RATIO, 0 to 1)

Lead Source	0.62
Lead Origin	0.50
Specialization	0.35
A free copy of Mastering The ...	0.25
Last Activity	0.19
Last Notable Activity	0.10
Converted	0.06
What is your current occupati...	0.06
Do Not Email	0.04

MOST FREQUENT VALUES

0.0	1,347	21.1%
2.0	1,324	20.8%
3.0	874	13.7%
4.0	659	10.3%
1.0	423	6.6%
5.0	379	5.9%
1.5	203	3.2%
6.0	183	2.9%
2.5	181	2.8%
7.0	104	1.6%
3.5	63	1.0%
8.0	60	0.9%
1.33	52	0.8%
2.33	48	0.8%
1.67	41	0.6%

SMALLEST VALUES

0.0	1,347	21.1%
1.0	423	6.6%
1.14	1	<0.1%
1.17	1	<0.1%
1.2	3	<0.1%
1.22	2	<0.1%
1.23	2	<0.1%
1.25	17	0.3%
1.27	1	<0.1%
1.31	1	<0.1%
1.33	52	0.8%
1.38	3	<0.1%
1.4	8	0.1%
1.43	1	<0.1%
1.45	1	<0.1%

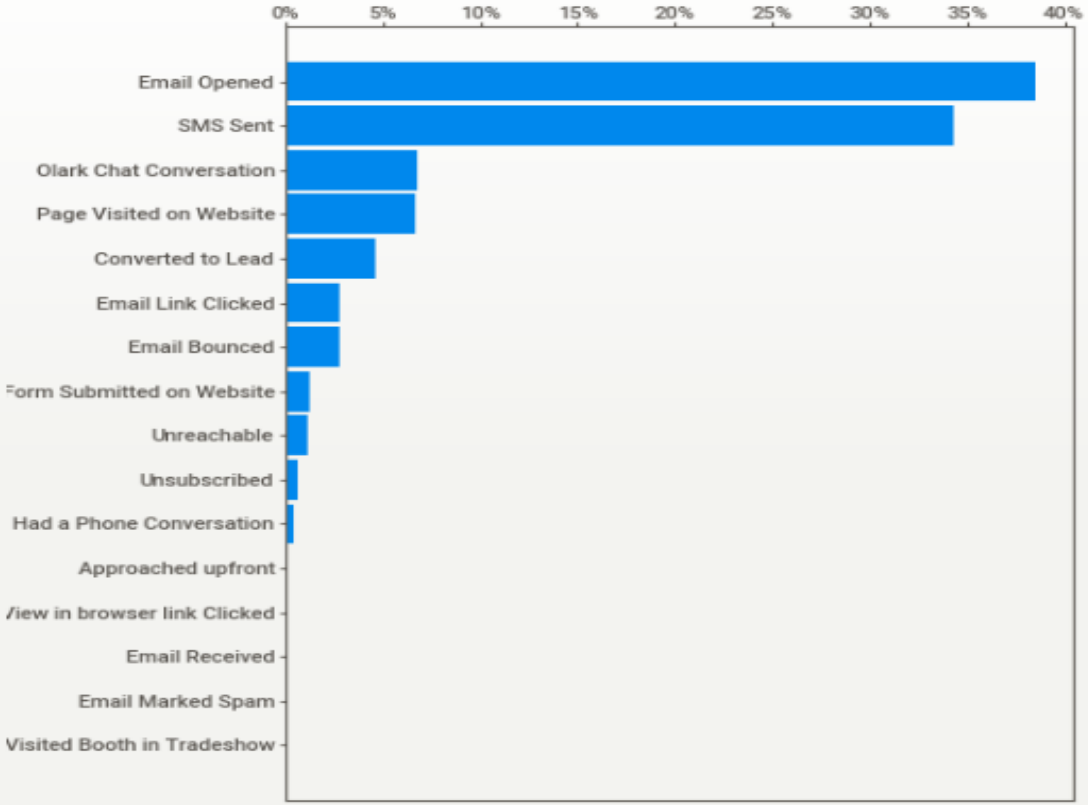
LARGEST VALUES

55.0	1	<0.1%
16.0	1	<0.1%
15.0	4	<0.1%
14.5	1	<0.1%
14.0	4	<0.1%
13.0	5	<0.1%
12.33	1	<0.1%
12.0	2	<0.1%
11.5	1	<0.1%
11.0	14	0.2%
10.0	18	0.3%
9.0	37	0.6%
8.5	1	<0.1%
8.33	1	<0.1%
8.21	1	<0.1%

EDA

Last Activity

MISSING: ---



TOP CATEGORIES

Email Opened	2,455	39%
SMS Sent	2,189	34%
Olark Chat Conversation	428	7%
Page Visited on Website	427	7%
Converted to Lead	292	5%
Email Link Clicked	178	3%
Email Bounced	175	3%
Form Submitted on Website	81	1%
Unreachable	71	1%
Unsubscribed	40	<1%
Had a Phone Conversation	23	<1%
Approached upfront	5	<1%
View in browser link Clicked	4	<1%
Email Received	2	<1%
Email Marked Spam	2	<1%
Visited Booth in Tradeshow	1	<1%
ALL	6,373	100%

CATEGORICAL ASSOCIATIONS
(UNCERTAINTY COEFFICIENT, 0 to 1)

Last Activity PROVIDES INFORMATION ON...

Last Notable Activity	0.69
Do Not Email	0.43
Converted	0.10
Lead Origin	0.06
Lead Source	0.05
What is your current occupati...	0.04
A free copy of Mastering The ...	0.03
Specialization	0.02

THESE FEATURES GIVE INFORMATION ON Last Activity:

Last Notable Activity	0.61
Do Not Email	0.07
Lead Source	0.05
Converted	0.04
Lead Origin	0.04
Specialization	0.04
What is your current occupati...	0.01
A free copy of Mastering The ...	0.01

NUMERICAL ASSOCIATIONS
(CORRELATION RATIO, 0 to 1)

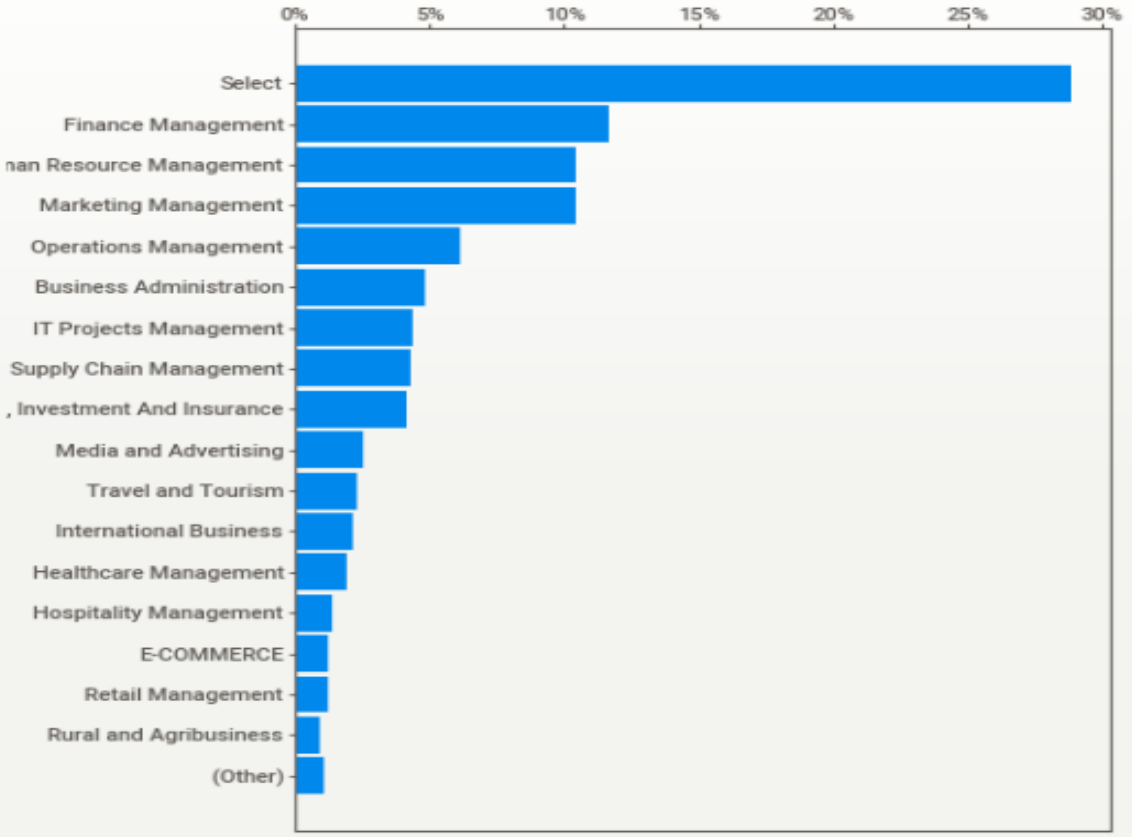
Last Activity CORRELATION RATIO WITH...

TotalVisits	0.24
Page Views Per Visit	0.19
Total Time Spent on Website	0.15

EDA

Specialization

MISSING: ---



TOP CATEGORIES

Select	1,838	29%
Finance Management	745	12%
Human Resource Management	665	10%
Marketing Management	663	10%
Operations Management	391	6%
Business Administration	310	5%
IT Projects Management	278	4%
Supply Chain Management	275	4%
Banking, Investment And Insurance	266	4%
Media and Advertising	161	3%
Travel and Tourism	149	2%
International Business	136	2%
Healthcare Management	122	2%
Hospitality Management	90	1%
E-COMMERCE	80	1%
Retail Management	78	1%
Rural and Agribusiness	58	<1%
(Other)	68	1%
ALL	6,373	100%

CATEGORICAL ASSOCIATIONS
(UNCERTAINTY COEFFICIENT, 0 to 1)

Specialization PROVIDES INFORMATION ON...

Lead Origin	0.32
A free copy of Mastering The ...	0.24
Lead Source	0.11
What is your current occupati...	0.07
Last Activity	0.04
Do Not Email	0.02
Last Notable Activity	0.02
Converted	0.01

THESE FEATURES GIVE INFORMATION ON Specialization:

Lead Origin	0.13
Lead Source	0.08
A free copy of Mastering The ...	0.06
Last Activity	0.02
What is your current occupati...	0.01
Last Notable Activity	0.01
Converted	0.00
Do Not Email	0.00

NUMERICAL ASSOCIATIONS
(CORRELATION RATIO, 0 to 1)

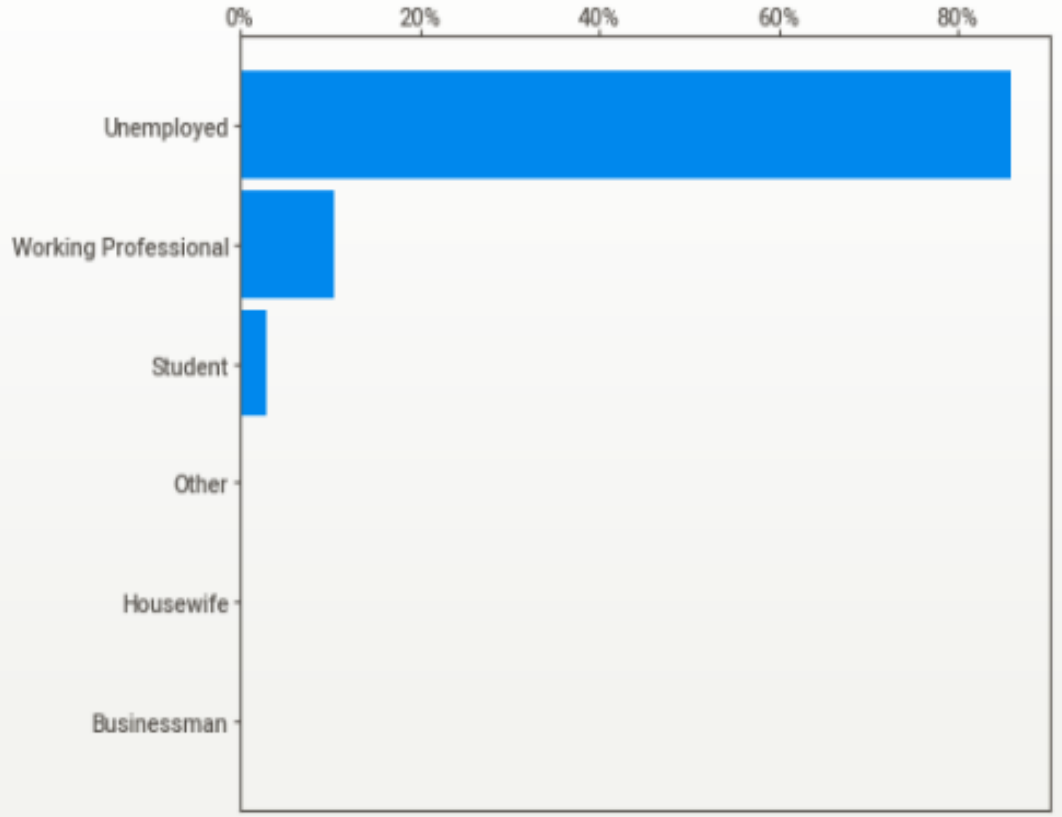
Specialization CORRELATION RATIO WITH...

Page Views Per Visit	0.35
Total Time Spent on Website	0.26
TotalVisits	0.23

EDA

What Is your current occupation

MISSING: ---



TOP CATEGORIES

Unemployed	5,476	86%
Working Professional	673	11%
Student	193	3%
Other	15	<1%
Housewife	9	<1%
Businessman	7	<1%
ALL	6,373	100%

CATEGORICAL ASSOCIATIONS
(UNCERTAINTY COEFFICIENT, 0 to 1)

What Is your current occupation PROVIDES INFORMATION ON...

Converted	0.08
Lead Source	0.02
Specialization	0.01
Lead Origin	0.01
Last Activity	0.01
Last Notable Activity	0.01
Do Not Email	0.01
A free copy of Mastering The ...	0.00

THESE FEATURES GIVE INFORMATION ON What Is your current occupation:

Converted	0.11
Specialization	0.07
Lead Source	0.05
Last Activity	0.04
Last Notable Activity	0.03
Lead Origin	0.03
Do Not Email	0.00
A free copy of Mastering The ...	0.00

NUMERICAL ASSOCIATIONS
(CORRELATION RATIO, 0 to 1)

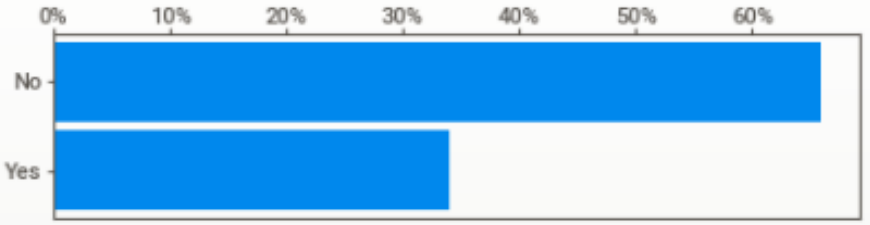
What Is your current occupation CORRELATION RATIO WITH...

Total Time Spent on Website	0.09
Page Views Per Visit	0.06
TotalVisits	0.05

EDA

A free copy of Mastering The Interview

MISSING: ---



TOP CATEGORIES

No	4,202	66%
Yes	2,171	34%
ALL	6,373	100%

CATEGORICAL ASSOCIATIONS
(UNCERTAINTY COEFFICIENT, 0 to 1)

A free copy of Mastering The Interview PROVIDES INFORMATION ON...

Lead Origin	0.19
Lead Source	0.15
Specialization	0.06
Last Activity	0.01
Converted	0.01
Do Not Email	0.01
Last Notable Activity	0.00
What is your current occupati...	0.00

THESE FEATURES GIVE INFORMATION ON A free copy of Mastering The Interview:

Lead Source	0.37
Lead Origin	0.28
Specialization	0.24
Last Activity	0.03
Last Notable Activity	0.01
Converted	0.01
Do Not Email	0.00
What is your current occupati...	0.00

NUMERICAL ASSOCIATIONS
(CORRELATION RATIO, 0 to 1)

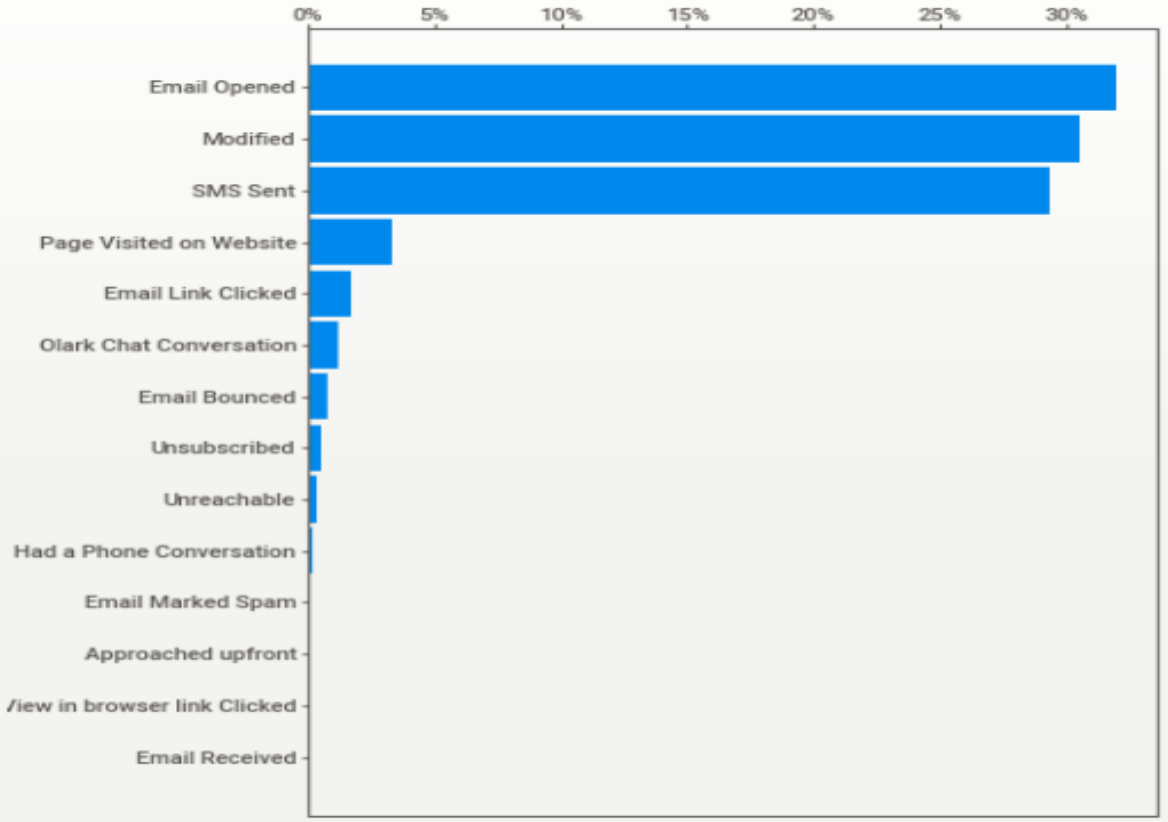
A free copy of Mastering The Interview CORRELATION RATIO WITH...

Page Views Per Visit	0.25
TotalVisits	0.19
Total Time Spent on Website	0.15

EDA

Last Notable Activity

MISSING: ---



TOP CATEGORIES

Email Opened	2,038	32%
Modified	1,947	31%
SMS Sent	1,869	29%
Page Visited on Website	213	3%
Email Link Clicked	110	2%
Olark Chat Conversation	76	1%
Email Bounced	49	<1%
Unsubscribed	31	<1%
Unreachable	22	<1%
Had a Phone Conversation	13	<1%
Email Marked Spam	2	<1%
Approached upfront	1	<1%
View in browser link Clicked	1	<1%
Email Received	1	<1%
ALL	6,373	100%

CATEGORICAL ASSOCIATIONS
(UNCERTAINTY COEFFICIENT, 0 to 1)

Last Notable Activity PROVIDES INFORMATION ON...

Last Activity	0.61
Do Not Email	0.23
Converted	0.08
What is your current occupati...	0.03
Lead Origin	0.02
Lead Source	0.01
Specialization	0.01
A free copy of Mastering The ...	0.01

THESE FEATURES GIVE INFORMATION ON Last Notable Activity:

Last Activity	0.69
Do Not Email	0.04
Converted	0.04
Specialization	0.02
Lead Source	0.02
Lead Origin	0.01
What is your current occupati...	0.01
A free copy of Mastering The ...	0.00

NUMERICAL ASSOCIATIONS
(CORRELATION RATIO, 0 to 1)

Last Notable Activity CORRELATION RATIO WITH...

TotalVisits	0.24
Total Time Spent on Website	0.13
Page Views Per Visit	0.10

DUMMY MODEL AND DATA SPLITITNG

- The first five rows of the DataFrame show that the categorical variables have been transformed into binary (0 or 1) values. For example, 'Lead Origin' has been transformed into columns like 'Lead Origin_Landing Page Submission', 'Lead Origin_Lead Add Form', etc., with 0 or 1 indicating the absence or presence of the corresponding category.
- The DataFrame now has a total of 75 columns, reflecting the original features along with the newly created dummy variables.
- Next steps could involve splitting the data into training and testing sets, scaling numerical features if necessary, and then proceeding with building and testing your linear regression model.

MODEL EVALUATION:

- Training and Testing Accuracy:

- The training accuracy is around 80%, indicating that the model performs reasonably well on the training data. However, the testing accuracy is significantly lower at around 53%. This suggests that the model might be overfitting the training data.

CONFUSION MATRIX AND CLASSIFICATION REPORT:

- The confusion matrix and classification report provide more detailed insights into the model's performance.
- The model seems to have a high recall for class 1 (Converted), but precision is relatively low, leading to an overall lower F1-score. This indicates potential class imbalance or issues with the model's ability to correctly identify non-converted cases.
- The output of logreg.predict_proba(X_test) shows the predicted probabilities for each class. It seems the model is quite confident in its predictions, often predicting very high probabilities for class 1.

FEATURE SELECTION:

- We have used Recursive Feature Elimination (RFE) to select 15 features. The coefficients from the logistic regression model indicate the impact of each selected feature on the log-odds of conversion.
- Multicollinearity Check (VIF):
- The VIF values are generally below 5, indicating that multicollinearity among the selected features is not a significant issue.

FINDING THE OPTIMAL CUTOFF:

1. **Sensitivity and Specificity:**

   The confusion matrix indicates that the model has good sensitivity (ability to correctly identify conversions) but relatively lower specificity (ability to correctly identify non-conversions).

2. **Accuracy Score:**

   The overall accuracy score is around 78.86%.

3. **Sensitivity (True Positive Rate):**

   The sensitivity is approximately 73.94%.

4. **Specificity (True Negative Rate):**

   The specificity is around 83.43%.

THRESHOLD OPTIMIZATION:

1. **ROC Curve:**
    1. The ROC curve is a visual representation of the trade-off between true positive rate (sensitivity) and false positive rate at different probability cutoffs. The AUC is 0.86, indicating a good model performance.
2. **Optimal Cutoff Selection:**
    1. By analyzing the sensitivity, specificity, and accuracy at different probability cutoffs, you identified 0.42 as the optimal threshold for the final predictions.
3. **Model Evaluation with Optimal Cutoff:**
    1. With the adjusted cutoff at 0.42, the model's accuracy increased to approximately 79.09%.
    2. Sensitivity (True Positive Rate) improved to around 79.34%.
    3. Specificity (True Negative Rate) also increased to approximately 78.85%.
4. **Confusion Matrix:**
    1. The confusion matrix for the final predictions shows:
        1. True Positives: 1705
        2. True Negatives: 1823
        3. False Positives: 489
        4. False Negatives: 444
5. **Sensitivity and Specificity:**
    1. Sensitivity: 79.34%
    2. Specificity: 78.85%

PREDICTIONS ON TEST SET:

- Feature Scaling:

Scaled the numerical columns of the test set using the same scaler used for the training set.

- Feature Selection:

Selected the relevant columns from the test set based on the features used in the training set.

- Prediction:

Used the trained logistic regression model (res) to predict conversion probabilities on the test set.

- Threshold Adjustment:

Chose a threshold of 0.42 for making the final predictions on the test set.

PREDICTIONS ON TEST SET:

- Performance on Test Set:

  Overall accuracy on the test set is approximately 78.45%.

  Sensitivity (True Positive Rate) is around 77.95%.

  Specificity (True Negative Rate) is approximately 78.92%.

- Confusion Matrix:

  True Positives: 714

  True Negatives: 786

  False Positives: 210

  False Negatives: 202

PRECISION-RECALL VIEW:

Cutoff Point Selection:

- Used the precision-recall tradeoff to choose an optimal cutoff point for making final predictions.

Applying Cutoff to Test Set:

- Chose a cutoff point of 0.44 for the training set, considering precision and recall tradeoff.

Performance on Training Set with New Cutoff:

- Overall accuracy on the training set with the new cutoff is approximately 78.95%.
- Precision is around 78.49%.
- Recall is approximately 77.71%.

APPLYING CUTOFF TO TEST SET:

Confusion Matrix for Training Set with New Cutoff:

- True Positives: 1670
- True Negatives: 1852
- False Positives: 460
- False Negatives: 479

Cutoff Point for Training Set:

- Used the trained logistic regression model to predict conversion probabilities on the test set.

Cutoff Point for Test Set:

- Applied the same cutoff point of 0.44 to make final predictions on the test set.

PERFORMANCE ON TEST SET WITH NEW CUTOFF:

- Overall accuracy on the test set with the new cutoff is approximately 78.66%.
- Precision is around 78.29%.
- Recall is approximately 76.75%.

- Confusion Matrix for Test Set with New Cutoff:

  True Positives: 703

  True Negatives: 801

  False Positives: 195

  False Negatives: 213

THANK YOU

Utsav Kumar

utsavkumar3050@gmail.com

UpGrad PGDM