Lead Scoring Case Study – Logistic Regression

Group Assignment

Group Members

- **Sajid Ahmed** | sajid.amd@gmail.com
- **Sahin Sultana Khan**| sahin.mailme@gmail.con

Upgrad & IIITB | Data Science Program – August 2023 | DSC59

Lead Scoring Case Study – Lead Conversion Prediction

Problem Statement:

- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. 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.
- There are some more problems presented by the company which the model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well.

Data Provided:

- ***lead.csv*** contains all the information of the leads. The data is about whether a particular lead got converted or not.

- ***Lead Data Dictionary.csv*** contains definition of columns present in the lead.csv file

Lead Scoring Case Study – Lead Conversion Prediction

Data Cleanup

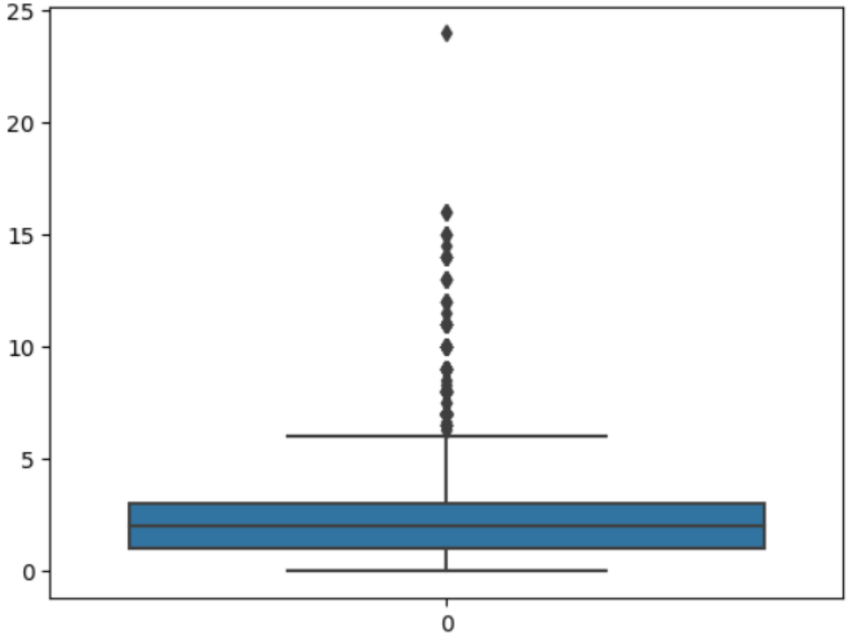
Remove columns:

- having missing values >= 40%
- Others as they are non-relevant to our analysis

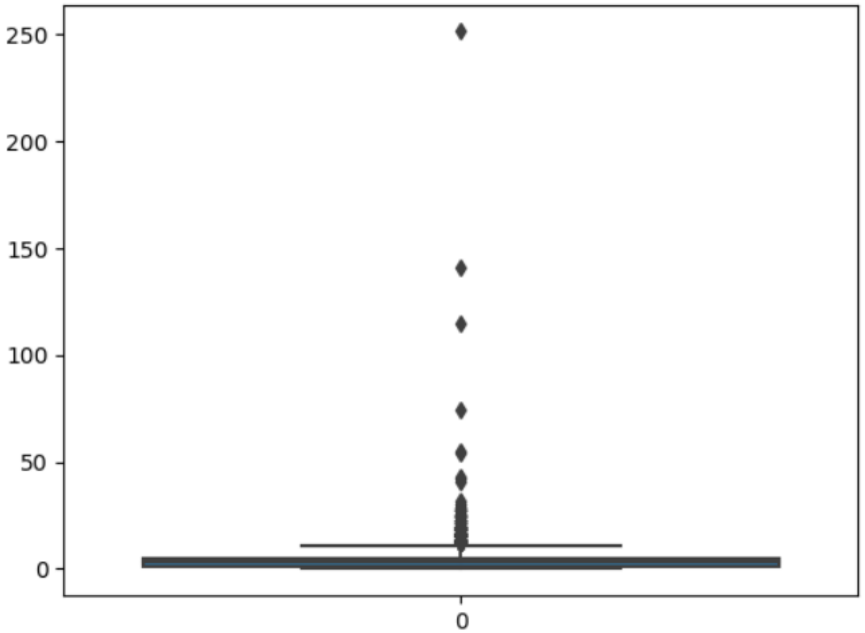
1. Lead Quality
2. Asymmetrique Activity Index
3. Asymmetrique Profile Index
4. Asymmetrique Activity Score
5. Asymmetrique Profile Score

Outlier Treatment

Page Views Per Visit > 15



Remove records where TotalVisits > 30



Lead Scoring Case Study – Lead Conversion Prediction

Data Cleanup

Impute columns with missing values due to:

- ➢ High in volume
- ➢ To include them for improved calculation/ analysis

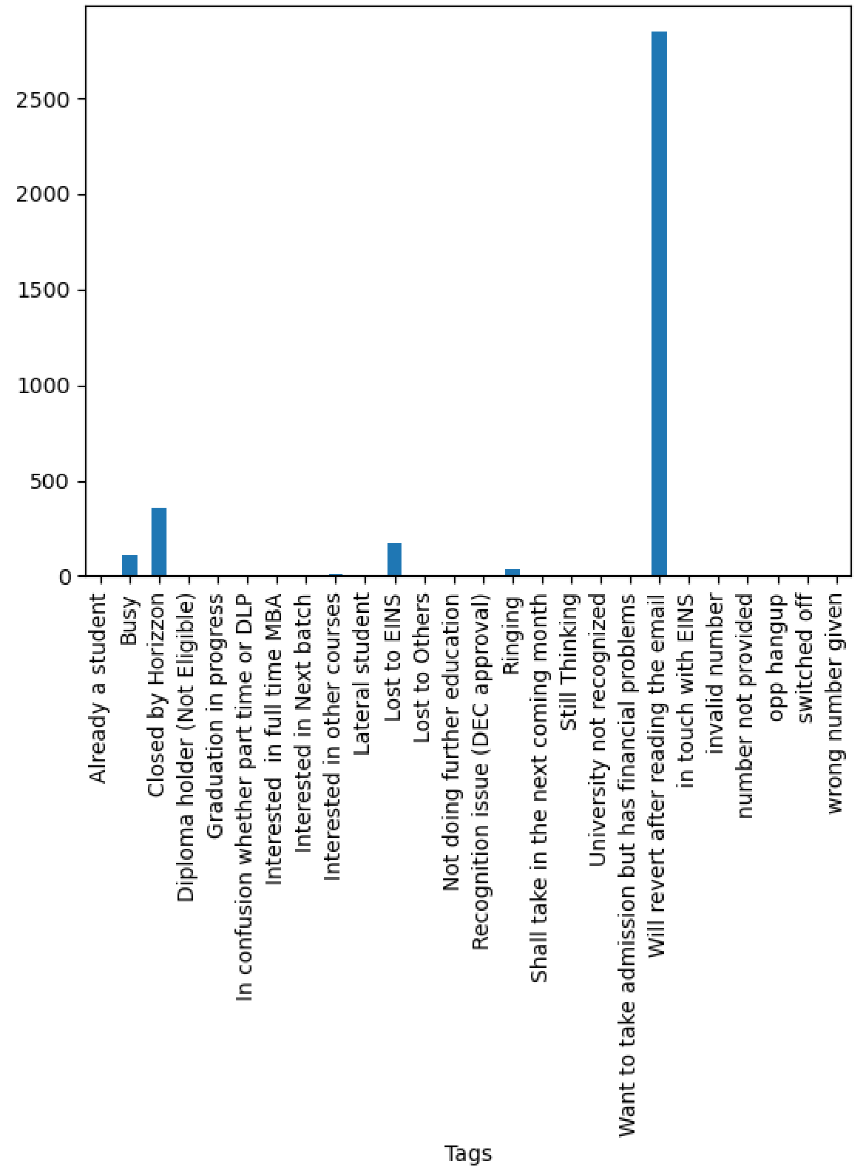
- **Specialization**: Consider 'Select' as equivalent to NULL & impute with more relevant/ frequent Specialization 'Finance Management'
- Column '**What matters most to you in choosing a course**' having NULLs can be imputed with more relevant 'Better Career Prospects'
- Missing **Tags** can be marked as 'Will revert after reading the email' in order to include them in further calculation
- Mark all the missing **Lead Profile** to 'Potential Lead'
- Mark all the missing **What is your current occupation** to 'Blank_Occupation'
- Mark all the missing **How did you hear about X Education** to Blank_Lead_Source'. Similarly impute **Last Activity**
- Impute **Page Views Per Visit** with median
- For the missing **Lead Source,** mark them as Google, which is most likely search platform

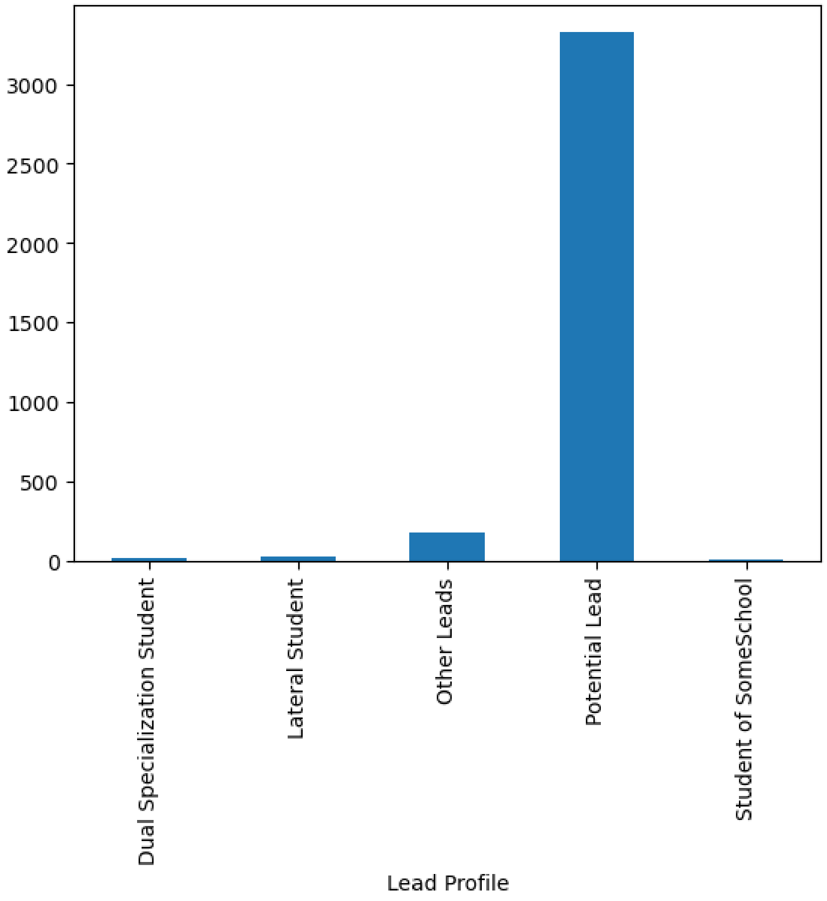
Transform columns with binary values (Yes->1/No->0)

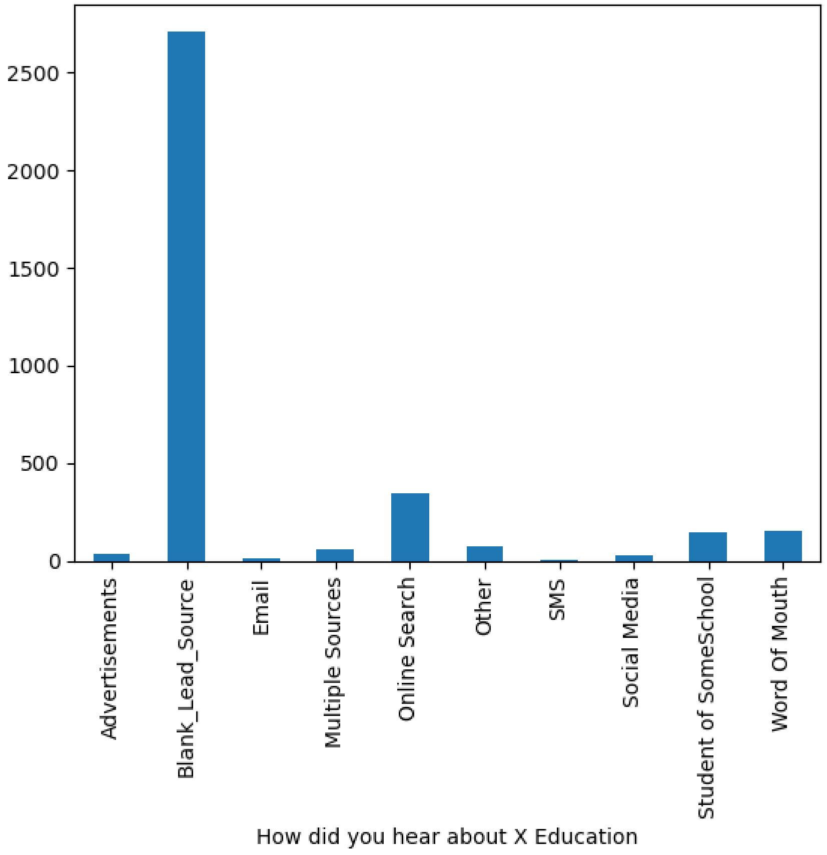
Do Not Email
Do Not Call
Search
Magazine
Newspaper Article
X Education Forums
Newspaper
Digital Advertisement
Through Recommendations
Receive More Updates About Our Courses
Update me on Supply Chain Content
Get updates on DM Content
I agree to pay the amount through cheque
A free copy of Mastering The Interview

Lead Scoring Case Study – Lead Conversion Prediction

Key Data Metrics – post clean up

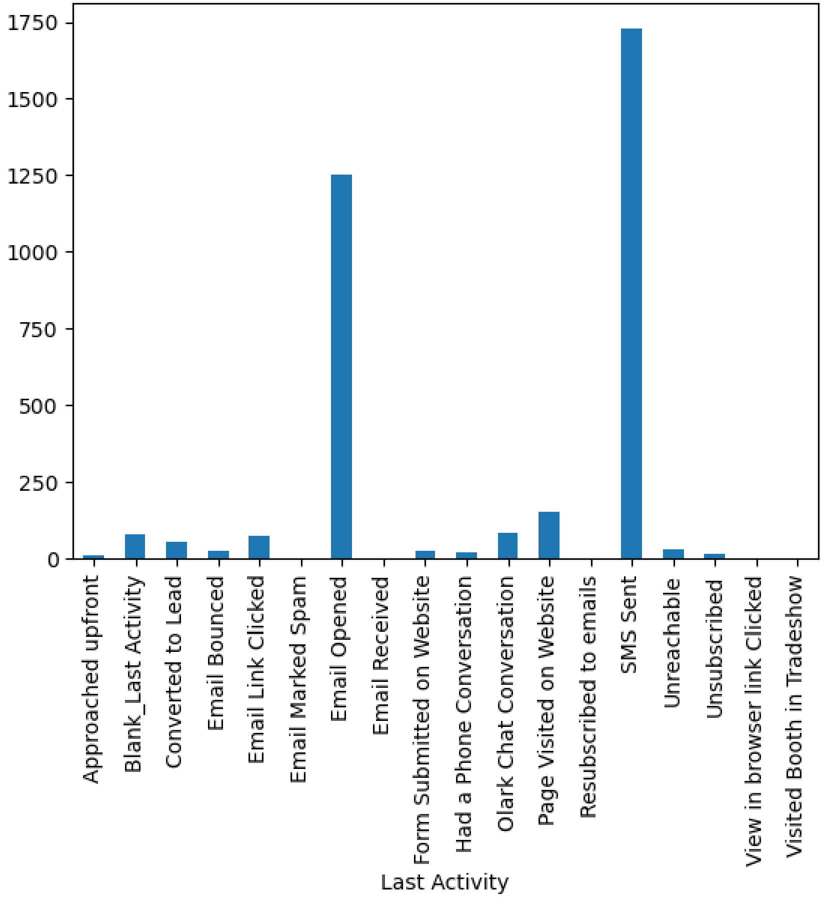


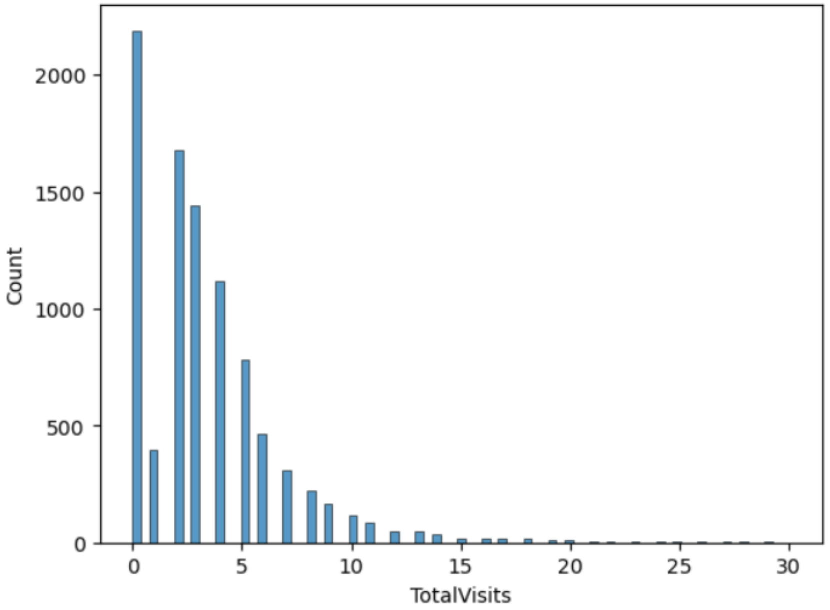


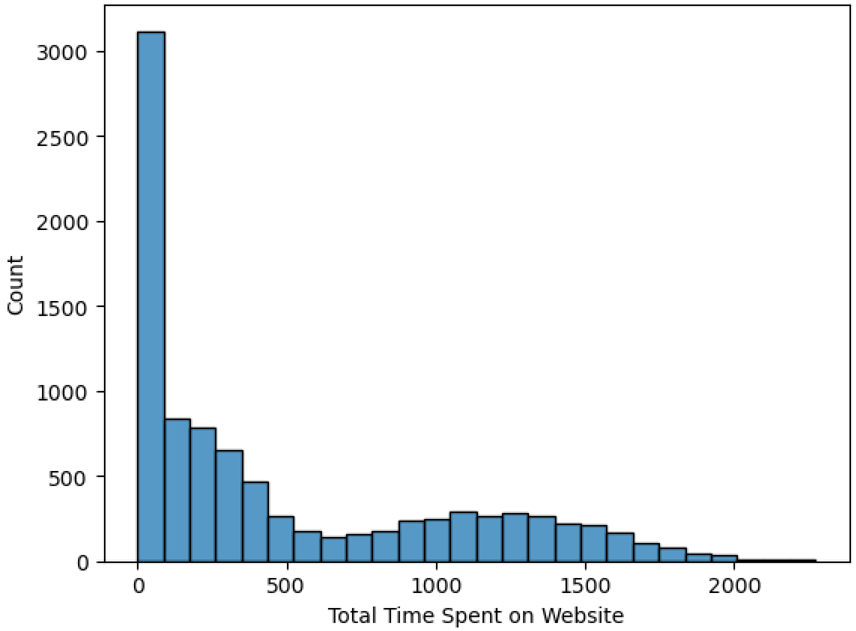


Lead Scoring Case Study – Lead Conversion Prediction

Key Data Metrics – post clean up

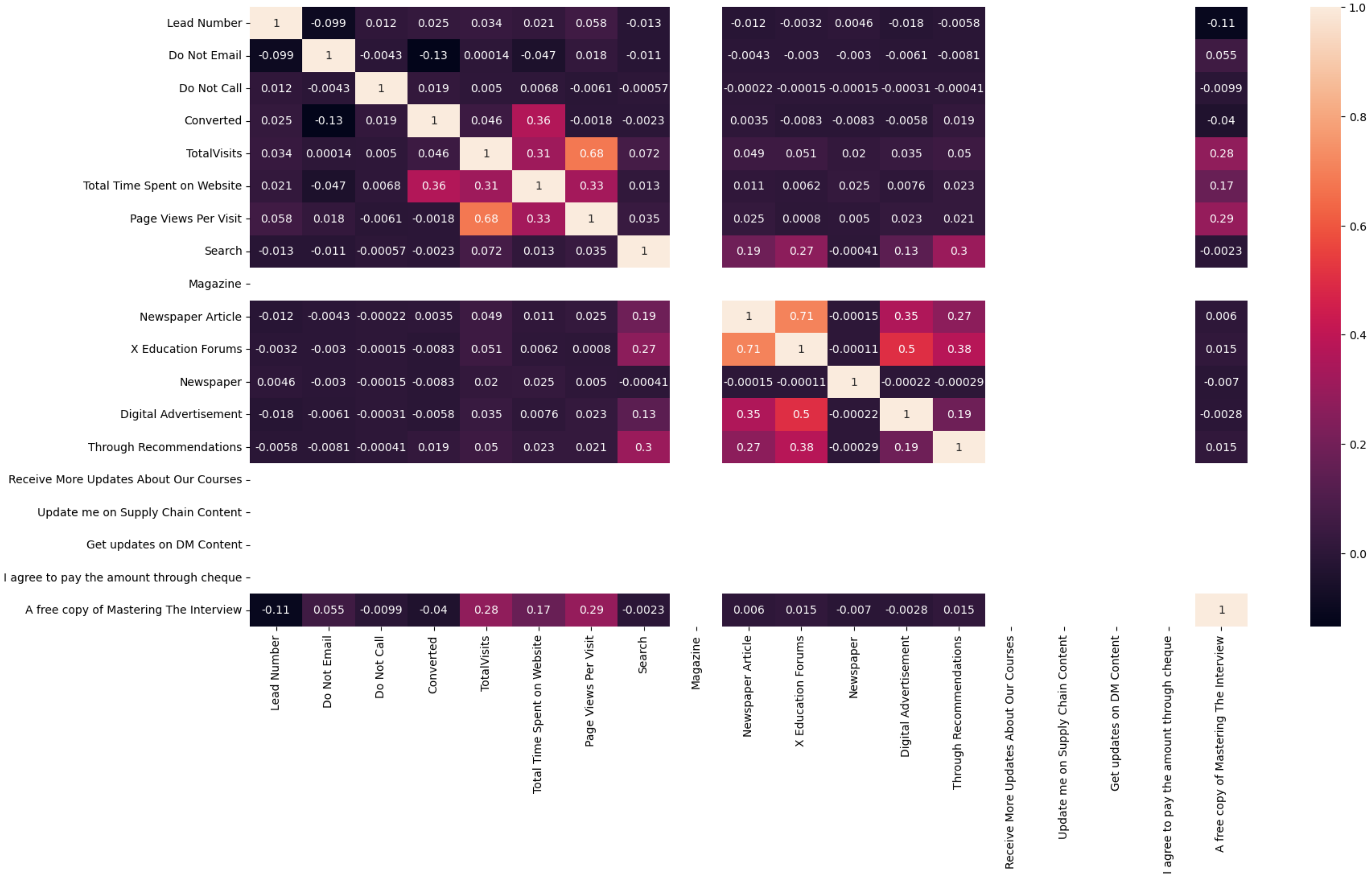






Lead Scoring Case Study – Lead Conversion Prediction

Key Data Metrics – Looking for Correlations

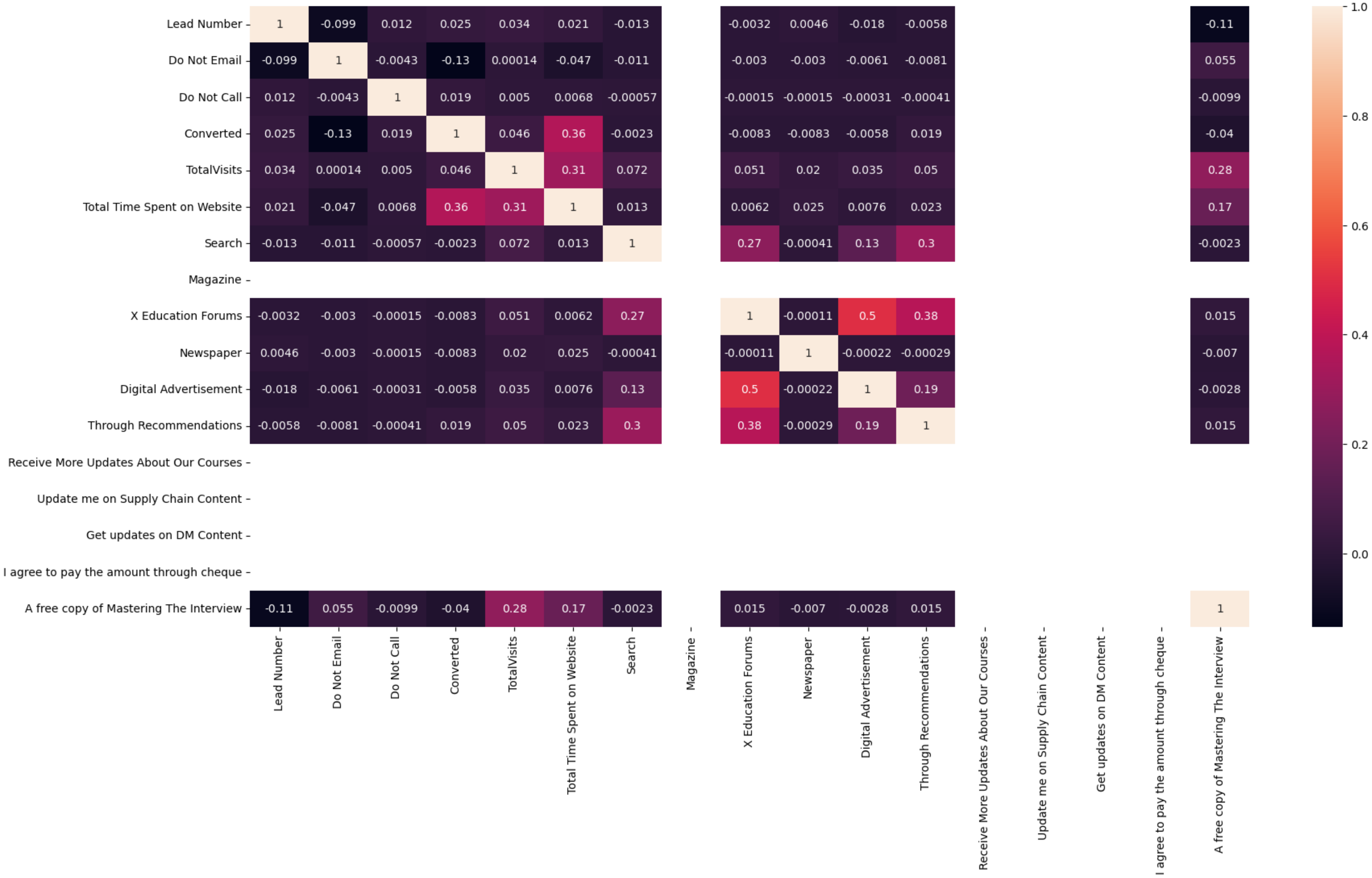


Remove following highly correlated variables

- **Page Views Per Visit**
- **Newspaper Article**

Lead Scoring Case Study – Lead Conversion Prediction

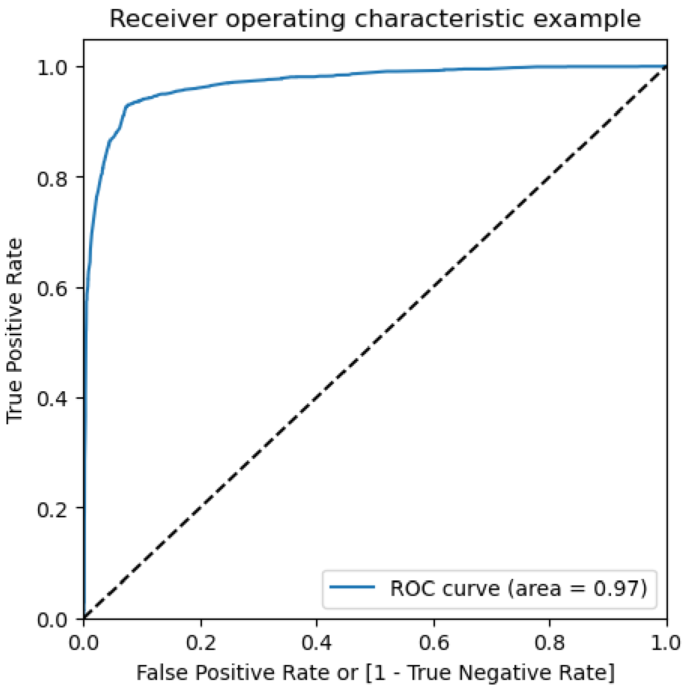
Key Data Metrics – After dropping highly Correlated variables

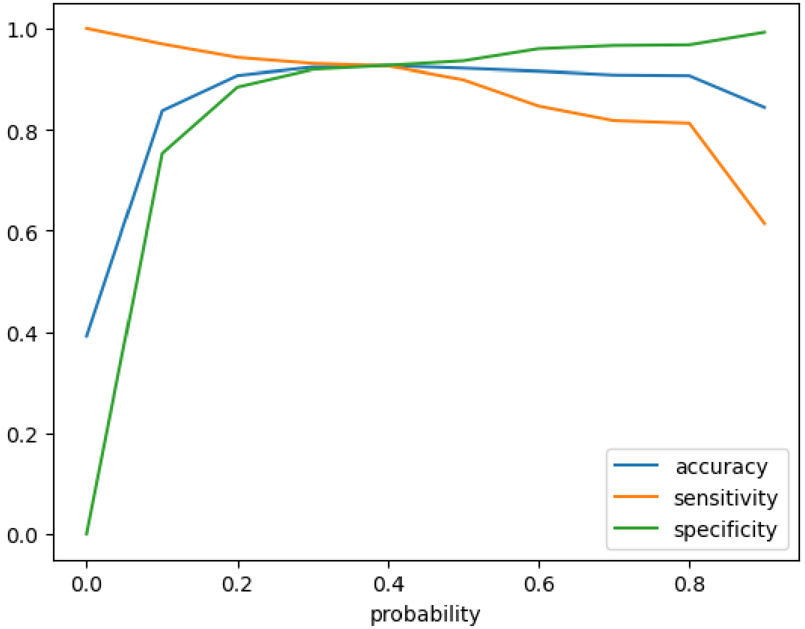


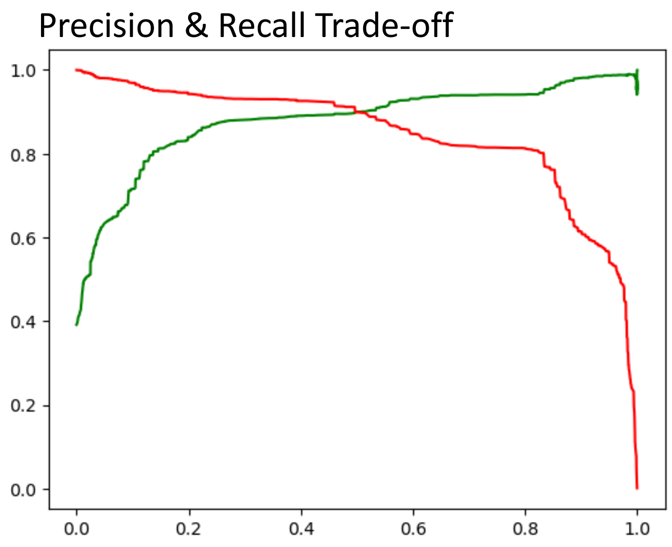
	Lead Number	Do Not Email	Do Not Call	Converted	TotalVisits	Total Time Spent on Website	Search	Magazine	X Education Forums	Newspaper	Digital Advertisement	Through Recommendations	Receive More Updates About Our Courses	Update me on Supply Chain Content	Get updates on DM Content	I agree to pay the amount through cheque	A free copy of Mastering The Interview
Lead Number	1	-0.099	0.012	0.025	0.034	0.021	-0.013		-0.0032	0.0046	-0.018	-0.0058					-0.11
Do Not Email	-0.099	1	-0.0043	-0.13	0.00014	-0.047	-0.011		-0.003	-0.003	-0.0061	-0.0081					0.055
Do Not Call	0.012	-0.0043	1	0.019	0.005	0.0068	-0.00057		-0.00015	-0.00015	-0.00031	-0.00041					-0.0099
Converted	0.025	-0.13	0.019	1	0.046	0.36	-0.0023		-0.0083	-0.0083	-0.0058	0.019					-0.04
TotalVisits	0.034	0.00014	0.005	0.046	1	0.31	0.072		0.051	0.02	0.035	0.05					0.28
Total Time Spent on Website	0.021	-0.047	0.0068	0.36	0.31	1	0.013		0.0062	0.025	0.0076	0.023					0.17
Search	-0.013	-0.011	-0.00057	-0.0023	0.072	0.013	1		0.27	-0.00041	0.13	0.3					-0.0023
Magazine																	
X Education Forums	-0.0032	-0.003	-0.00015	-0.0083	0.051	0.0062	0.27		1	-0.00011	0.5	0.38					0.015
Newspaper	0.0046	-0.003	-0.00015	-0.0083	0.02	0.025	-0.00041		-0.00011	1	-0.00022	-0.00029					-0.007
Digital Advertisement	-0.018	-0.0061	-0.00031	-0.0058	0.035	0.0076	0.13		0.5	-0.00022	1	0.19					-0.0028
Through Recommendations	-0.0058	-0.0081	-0.00041	0.019	0.05	0.023	0.3		0.38	-0.00029	0.19	1					0.015
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																	
A free copy of Mastering The Interview	-0.11	0.055	-0.0099	-0.04	0.28	0.17	-0.0023		0.015	-0.007	-0.0028	0.015					1

Lead Scoring Case Study – Lead Conversion Prediction

Key Data Metrics – training data set







An **ROC curve** demonstrates several things:

- It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
- The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

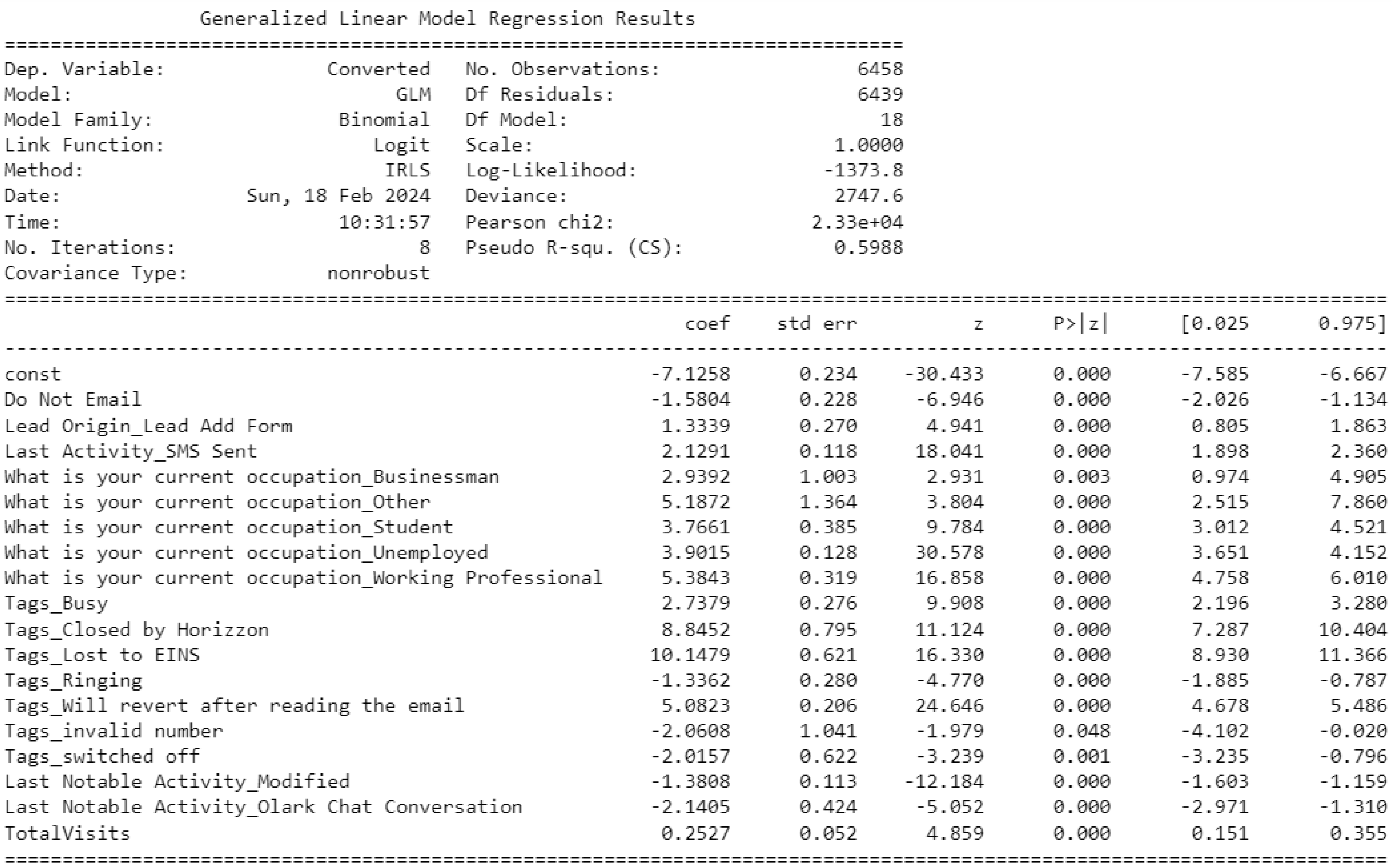
- **Initial approach** was with the **cut-off of 0.5**, which resulted into **model accuracy** of **92.13**
- Upon plotting accuracy, sensitivity & specificity Vs the probabilities in between 0 – 0.9, got the **new cut-off point of 0.4**
- The **model accuracy** overall has seen **slight improvement to 92.66** with the new cut-off of 0.4

Lead Scoring Case Study – Lead Conversion Prediction

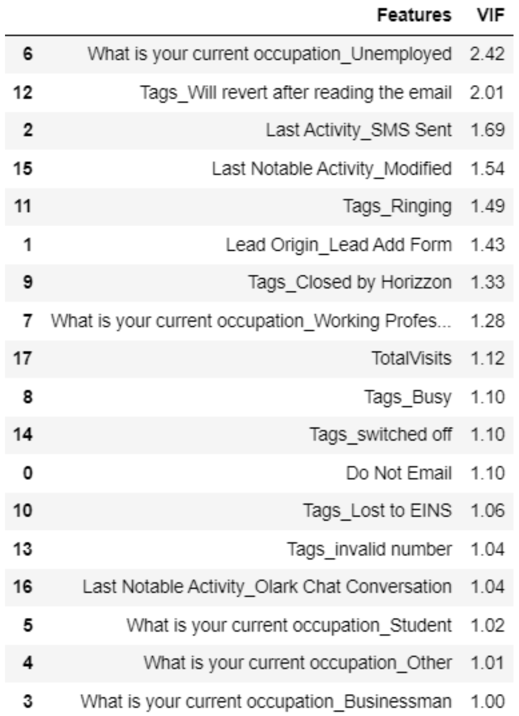
Key Data Metrics – training data set

Total Number of Features in scope via RFE & manual feature selection = **21**

Final model – Post feature elimination (manual & VIF)



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:              Converted   No. Observations:                 6458
Model:                            GLM   Df Residuals:                     6439
Model Family:                Binomial   Df Model:                           18
Link Function:                  Logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -1373.8
Date:                Sun, 18 Feb 2024   Deviance:                       2747.6
Time:                        10:31:57   Pearson chi2:                 2.33e+04
No. Iterations:                     8   Pseudo R-squ. (CS):             0.5988
Covariance Type:            nonrobust
========================================================================================================================
                                                           coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------------------------------
const                                               -7.1258      0.234    -30.433      0.000      -7.585      -6.667
Do Not Email                                        -1.5804      0.228     -6.946      0.000      -2.026      -1.134
Lead Origin_Lead Add Form                            1.3339      0.270      4.941      0.000       0.805       1.863
Last Activity_SMS Sent                               2.1291      0.118     18.041      0.000       1.898       2.360
What is your current occupation_Businessman          2.9392      1.003      2.931      0.003       0.974       4.905
What is your current occupation_Other                5.1872      1.364      3.804      0.000       2.515       7.860
What is your current occupation_Student              3.7661      0.385      9.784      0.000       3.012       4.521
What is your current occupation_Unemployed           3.9015      0.128     30.578      0.000       3.651       4.152
What is your current occupation_Working Professional     5.3843      0.319     16.858      0.000       4.758       6.010
Tags_Busy                                            2.7379      0.276      9.908      0.000       2.196       3.280
Tags_Closed by Horizzon                              8.8452      0.795     11.124      0.000       7.287      10.404
Tags_Lost to EINS                                   10.1479      0.621     16.330      0.000       8.930      11.366
Tags_Ringing                                        -1.3362      0.280     -4.770      0.000      -1.885      -0.787
Tags_Will revert after reading the email             5.0823      0.206     24.646      0.000       4.678       5.486
Tags_invalid number                                 -2.0608      1.041     -1.979      0.048      -4.102      -0.020
Tags_switched off                                   -2.0157      0.622     -3.239      0.001      -3.235      -0.796
Last Notable Activity_Modified                      -1.3808      0.113    -12.184      0.000      -1.603      -1.159
Last Notable Activity_Olark Chat Conversation       -2.1405      0.424     -5.052      0.000      -2.971      -1.310
TotalVisits                                          0.2527      0.052      4.859      0.000       0.151       0.355
========================================================================================================================
```



	Features	VIF
6	What is your current occupation_Unemployed	2.42
12	Tags_Will revert after reading the email	2.01
2	Last Activity_SMS Sent	1.69
15	Last Notable Activity_Modified	1.54
11	Tags_Ringing	1.49
1	Lead Origin_Lead Add Form	1.43
9	Tags_Closed by Horizzon	1.33
7	What is your current occupation_Working Profes...	1.28
17	TotalVisits	1.12
8	Tags_Busy	1.10
14	Tags_switched off	1.10
0	Do Not Email	1.10
10	Tags_Lost to EINS	1.06
13	Tags_invalid number	1.04
16	Last Notable Activity_Olark Chat Conversation	1.04
5	What is your current occupation_Student	1.02
4	What is your current occupation_Other	1.01
3	What is your current occupation_Businessman	1.00

Lead Scoring Case Study – Lead Conversion Prediction

Key Data Metrics – training data set

Final model – Accuracy & Other measures

Existing Lead Conversion Rate = **38.58**

Overall model Accuracy on train data = **92.66**

Sensitivity = 0.926
Specificity = 0.927
False Positive Rate = 0.073
Positive Predictive Value = 0.890
Negative Predictive Value = 0.951

Confusion Matrix with cut-off = **0.5**

	Predicted	
Actual	3723	207
	332	2196

Confusion Matrix with cut-off = **0.4**

	Predicted	
Actual	3643	287
	187	2341

- There is a clear **gain in model performance** when we shifted to 0.4 from 0.5 as the probability cut-off
- As we observe, there is a slight increase in the False Positive Rate, which in the case of Leads are ok to have more number of leads classified for conversion.
- We also observe the False Negatives have significantly gone down which has subsequently improved the model to detect higher number of conversions in the data
- The model predicted 0.926 (Sensitivity) which is a great number for the model in it’s ability to detect total conversions over actual number of conversions

Lead Scoring Case Study – Lead Conversion Prediction

Key Data Metrics – Applying Model on test data

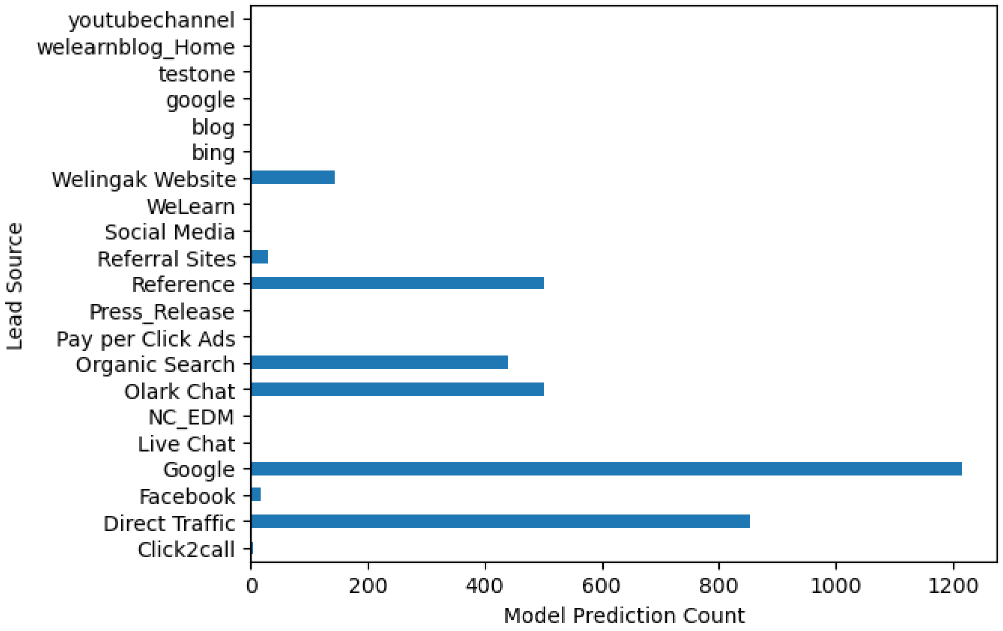
Overall Accuracy on test data = **92.85**

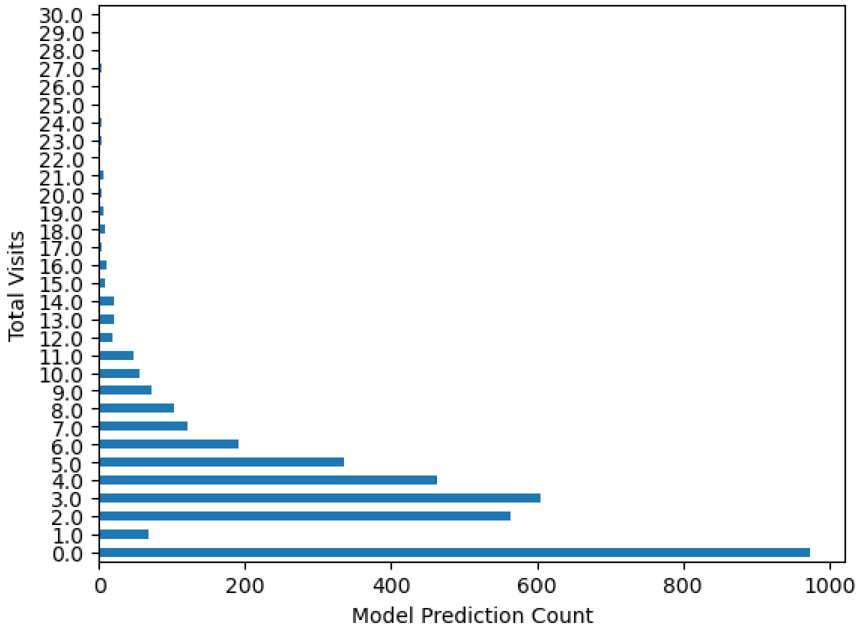
Sensitivity = 0.928
Specificity = 0.929
False Positive Rate = 0.165
Positive Predictive Value = 0.885
Negative Predictive Value = 0.956

Confusion Matrix with cut-off = **0.4**

	Predicted	
Actual	1613	124
	74	957

Conclusion: The model has proven to be generalized and performed very well on the test (unseen) data set





Thank You !