Lead Scoring Case Study

PRESENTED BY:

- TUSHAR MESTRY
- ANKIT SHRIVASTAV
- RUMANA FATHIMA

BATCH – DS-56 MAY -2023

Lead Scoring Case Study

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

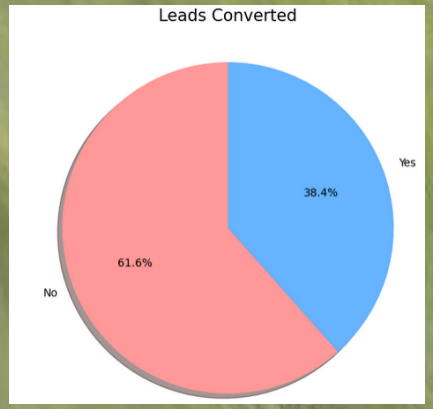
Problem Statement

1.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.

2.There are some more problems presented by the company which your 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. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

Loading the Data Set & Performing Exploratory Data Analysis

Rows & Columns of Data Set (9240, 37)

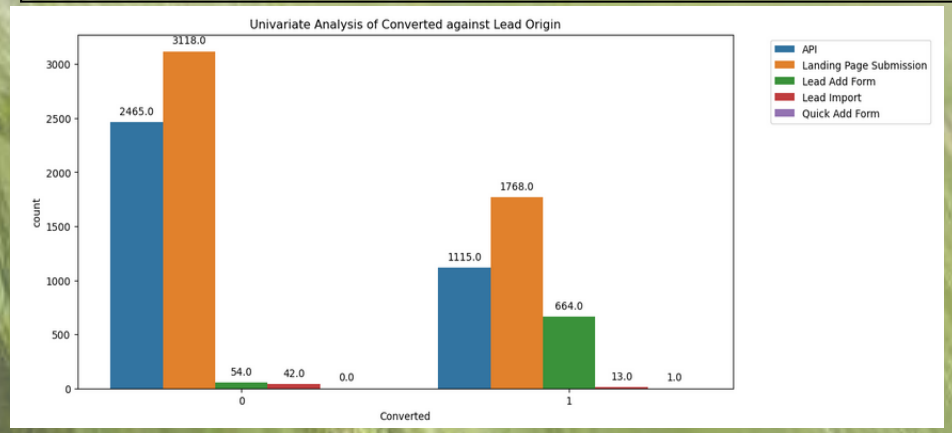


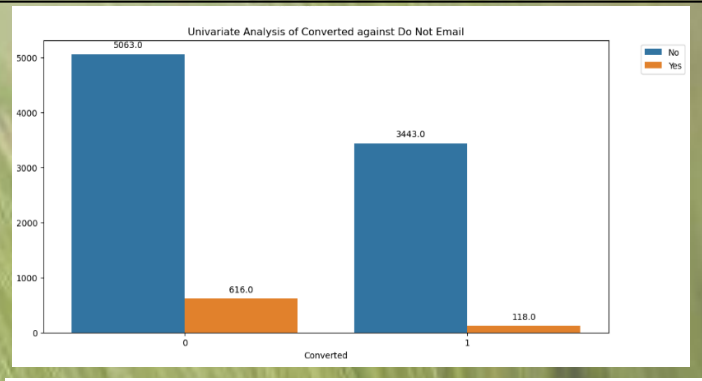
Treatment of Missing Values & Columns

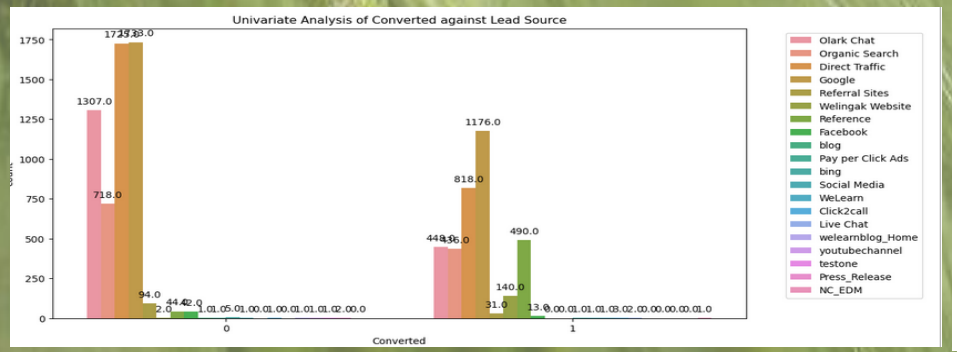
Columns having missing values > 40%

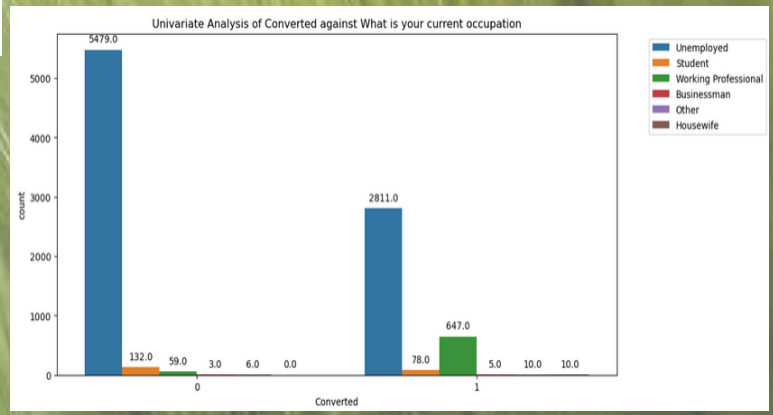
['How did you hear about X Education', 'Lead Quality', 'Lead Profile', 'Asymmetrique Activity Index', 'Asymmetrique Profile Index', 'Asymmetrique Activity Score', 'Asymmetrique Profile Score']

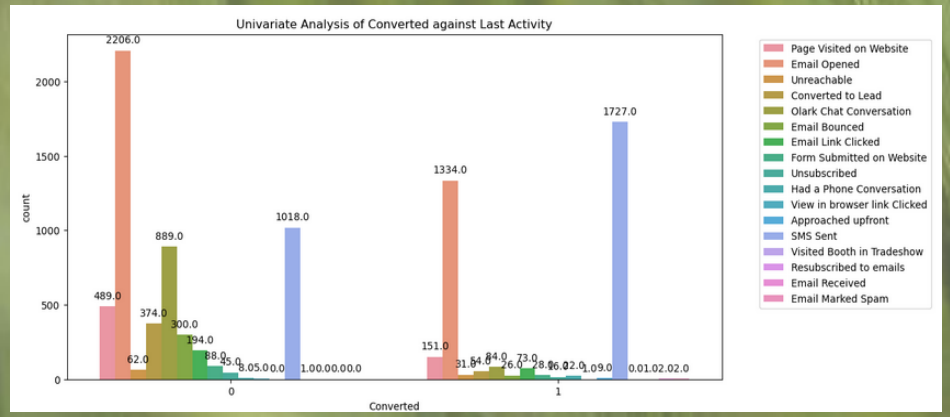
Uni-Variate Analysis of Category Columns

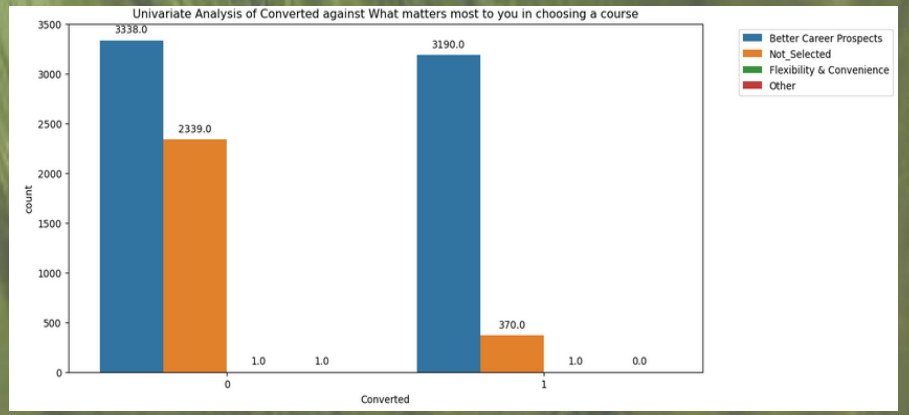






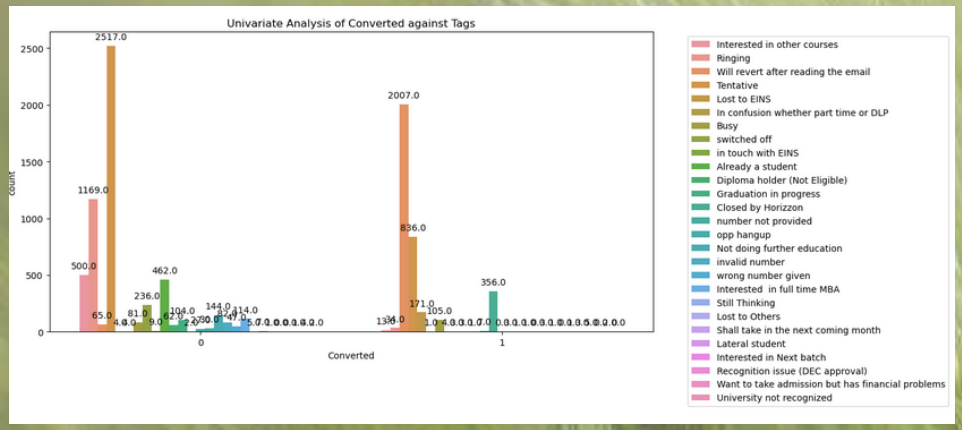


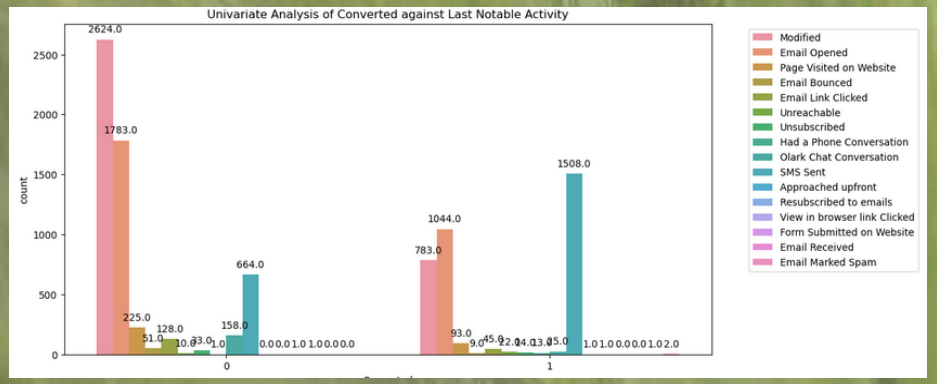


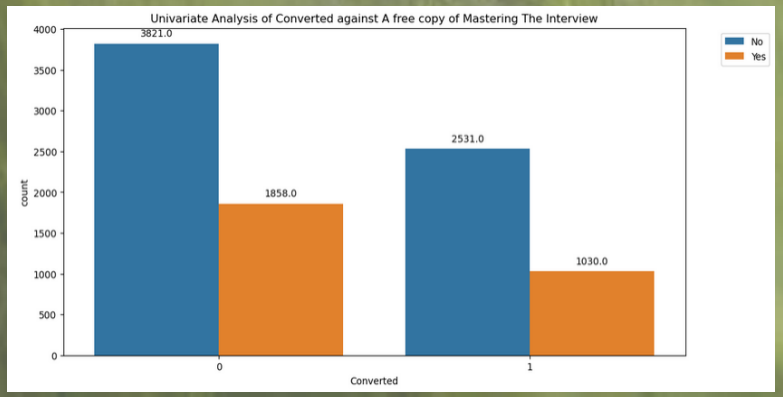


Loading the Data Set & Performing Exploratory Data Analysis

Uni-Variate Analysis of Category Columns

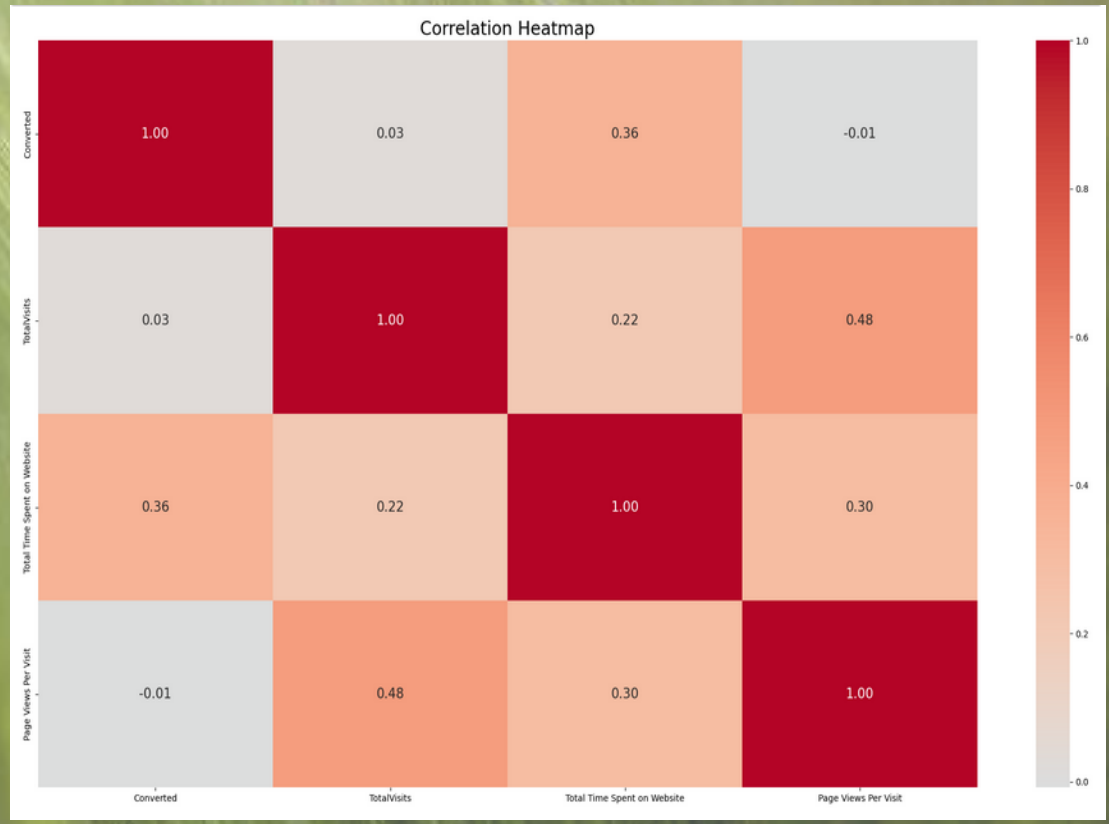




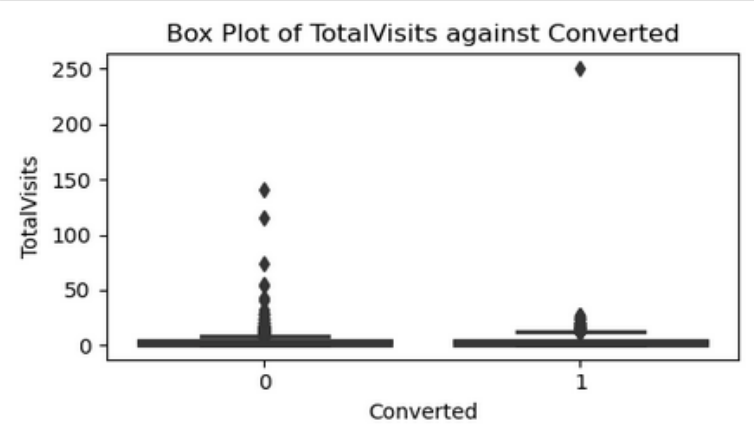


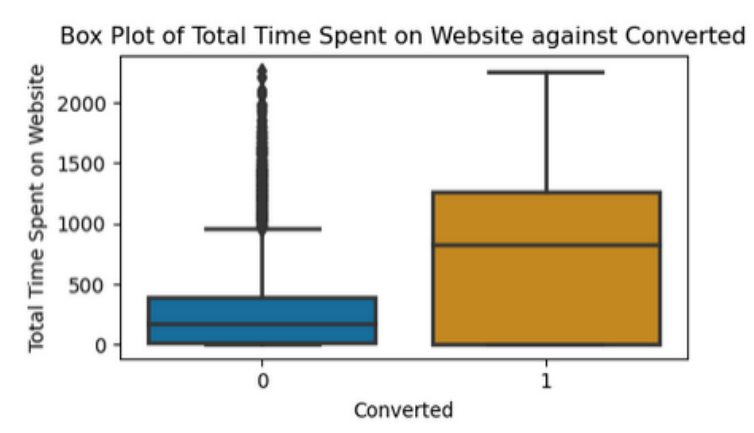
Dropping Highly Skewed Columns
[Magazine', 'Do Not Call', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations', 'Receive More Updates About Our Courses', 'Update me on Supply Chain Content', 'Get updates on DM Content', 'I agree to pay the amount through cheque']

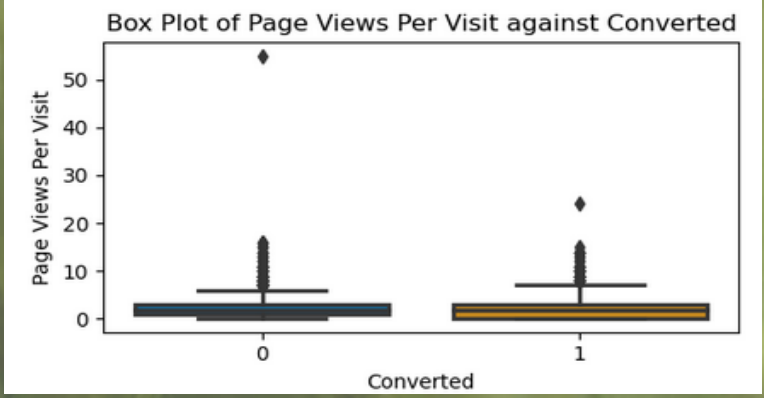
Correlation Analysis for all Numerical features



Bi-Variate Analysis for all Numerical features for analyzing Outliers







Post
1. **Creating Dummy variables**
2. **Scaling Numerical features**
3. **Split Data in "Training" & "Test" Data Sets**

Preparing Data Frame for Model Building

	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Lead Origin_Landing Page Submission	Lead Origin_Lead Add Form	Lead Origin_Lead Import	Lead Source_Direct Traffic	Lead Source_Facebook	Lead Source_Google	Lead Source_Live Chat	S
4958	0.665513	-0.034860	1.575017	1	0	0	0	0	1	0	
8957	-0.384904	-0.548908	-0.085221	0	0	0	0	0	0	0	
8274	0.315374	-0.862844	-0.085221	1	0	0	1	0	0	0	
8909	-1.085182	-0.884875	-1.192046	0	0	0	0	0	0	0	
9173	-0.735043	-0.787573	-0.638634	1	0	0	1	0	0	0	

5 rows × 193 columns

Removing Outliers from DatacSet

Original Data Frame shape: (9240, 16)
Data Frame shape after removing outliers: (9037, 16)

Logistic Regression Model Building

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6325
Model:	GLM	Df Residuals:	6151
Model Family:	Binomial	Df Model:	173
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	nan
Date:	Sun, 26 Nov 2023	Deviance:	71288.
Time:	19:45:38	Pearson chi2:	3.49e+18
No. Iterations:	100	Pseudo R-squ. (CS):	nan
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	5.06e+15	6.31e+07	8.02e+07	0.000	5.06e+15	5.06e+15
TotalVisits	9.321e+13	1.29e+06	7.24e+07	0.000	9.32e+13	9.32e+13
Total Time Spent on Website	3.938e+14	1.04e+06	3.79e+08	0.000	3.94e+14	3.94e+14
Page Views Per Visit	-1.379e+14	1.38e+06	-1e+08	0.000	-1.38e+14	-1.38e+14
Lead Origin_Landing Page Submission	-6.637e+13	4.07e+06	-1.63e+07	0.000	-6.64e+13	-6.64e+13
Lead Origin_Lead Add Form	6.673e+13	1.3e+07	5.15e+06	0.000	6.67e+13	6.67e+13
Lead Origin_Lead Import	-3.012e+13	2.99e+07	-1.01e+06	0.000	-3.01e+13	-3.01e+13
Lead Source_Direct Traffic	-3.268e+15	4.27e+07	-7.65e+07	0.000	-3.27e+15	-3.27e+15
Lead Source_Facebook	-3.494e+15	3.83e+07	-9.13e+07	0.000	-3.49e+15	-3.49e+15
Lead Source_Google	-3.272e+15	4.26e+07	-7.68e+07	0.000	-3.27e+15	-3.27e+15
Lead Source_Live Chat	9.447e+14	6.26e+07	1.51e+07	0.000	9.45e+14	9.45e+14
Lead Source_NC_EDM	2.081e+14	7.95e+07	2.62e+06	0.000	2.08e+14	2.08e+14
Lead Source_Olark Chat	-2.958e+15	4.28e+07	-6.91e+07	0.000	-2.96e+15	-2.96e+15
Lead Source_Organic Search	-3.316e+15	4.28e+07	-7.76e+07	0.000	-3.32e+15	-3.32e+15
Lead Source_Pay per Click Ads	12.5022	6.89e-07	1.82e+07	0.000	12.502	12.502
Lead Source_Press_Release	-19.0820	4.82e-07	-3.96e+07	0.000	-19.082	-19.082
Lead Source_Reference	-3.372e+15	4.08e+07	-8.26e+07	0.000	-3.37e+15	-3.37e+15
Lead Source_Referral Sites	-3.327e+15	4.34e+07	-7.67e+07	0.000	-3.33e+15	-3.33e+15
Lead Source_Social Media	-2.373e+15	6.4e+07	-3.7e+07	0.000	-2.37e+15	-2.37e+15
Lead Source_WeLearn	1.819e+14	7.97e+07	2.28e+06	0.000	1.82e+14	1.82e+14
Lead Source_Welingak Website	-1.66e+15	4.13e+07	-4.02e+07	0.000	-1.66e+15	-1.66e+15
Lead Source_bing	-2.714e+15	5.44e+07	-4.99e+07	0.000	-2.71e+15	-2.71e+15
Lead Source_blog	-8.08e+15	7.97e+07	-1.01e+08	0.000	-8.08e+15	-8.08e+15
Lead Source_testone	148.4038	1.53e-06	9.73e+07	0.000	148.404	148.404

As we see there many features which are populated in our first Model, hence using "Recursive Feature Elimination"

List of "15"columns identified thru "RFE"
['Lead Source_Welingak Website', 'Last Activity_SMS Sent', 'What matters most to you in choosing a course_Not_Selected', 'Tags_Busy', 'Tags_Closed by Horizzon', 'Tags_Lost to EINS', 'Tags_Not doing further education', 'Tags_Ringing', 'Tags_Tentative', 'Tags_Will revert after reading the email', 'Tags_invalid number', 'Tags_switched off', 'Tags_wrong number given', 'Last Notable Activity_Modified', 'Last Notable Activity_Olark Chat Conversation']

Tags_Not doing further education
Tags_invalid number
Tags_wrong number given
These feature are removed as those are insignificant due to 'p'value>0.05

Building Model-1

Model-1

```
1 X_train_sm = sm.add_constant(X_train[col])
2 logm1 = sm.GLM(y_train,X_train_sm, family = sm.families.Binomial())
3 res = logm1.fit()
4 res.summary()
```

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6325
Model:	GLM	Df Residuals:	6309
Model Family:	Binomial	Df Model:	15
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1227.9
Date:	Sun, 26 Nov 2023	Deviance:	2455.8
Time:	19:45:47	Pearson chi2:	1.20e+04
No. Iterations:	24	Pseudo R-squ. (CS):	0.6112
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.9981	0.196	-15.287	0.000	-3.383	-2.614
Lead Source_Welingak Website	2.7675	0.753	3.676	0.000	1.292	4.243
Last Activity_SMS Sent	2.0721	0.121	17.090	0.000	1.834	2.310
What matters most to you in choosing a course_Not_Selected	-2.6183	0.146	-17.889	0.000	-2.905	-2.331
Tags_Busy	2.4824	0.283	8.772	0.000	1.928	3.037
Tags_Closed by Horizzon	8.9140	0.741	12.026	0.000	7.461	10.367
Tags_Lost to EINS	9.5897	0.770	12.446	0.000	8.080	11.100
Tags_Not doing further education	-22.0213	2.04e+04	-0.001	0.999	-4e+04	4e+04
Tags_Ringing	-1.6250	0.295	-5.501	0.000	-2.204	-1.046
Tags_Tentative	3.5748	0.225	15.913	0.000	3.135	4.015
Tags_Will revert after reading the email	6.6164	0.260	25.424	0.000	6.106	7.126
Tags_invalid number	-23.4287	2.68e+04	-0.001	0.999	-5.26e+04	5.25e+04
Tags_switched off	-2.2076	0.623	-3.544	0.000	-3.429	-0.987
Tags_wrong number given	-23.6730	3.77e+04	-0.001	0.999	-7.4e+04	7.4e+04
Last Notable Activity_Modified	-1.6250	0.126	-12.848	0.000	-1.873	-1.377
Last Notable Activity_Olark Chat Conversation	-1.6406	0.431	-3.807	0.000	-2.485	-0.796

Logistic Regression Model Building

Model-2

```
# Let's re-run the model using the selected variables
X_train_sm = sm.add_constant(X_train[col])
logm2 = sm.GLM(y_train,X_train_sm, family = sm.families.Binomial())
res = logm2.fit()
res.summary()
```

2]: Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6325
Model:	GLM	Df Residuals:	6310
Model Family:	Binomial	Df Model:	14
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1231.8
Date:	Sun, 26 Nov 2023	Deviance:	2463.7
Time:	19:45:47	Pearson chi2:	1.17e+04
No. Iterations:	24	Pseudo R-squ. (CS):	0.6107
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	
const	-3.1173	0.197	-15.799	0.000	
Lead Source_Welingak Website	2.7632	0.752	3.673	0.000	
Last Activity_SMS Sent	2.0433	0.120	16.987	0.000	1.808 2.279
What matters most to you in choosing a course_Not_Selected	-2.6102	0.146	-17.901	0.000	-2.896 -2.324
Tags_Busy	2.6141	0.282	9.269	0.000	2.061 3.167
Tags_Closed by Horizzon	9.0211	0.742	12.159	0.000	7.567 10.475
Tags_Lost to EINS	9.6913	0.771	12.569	0.000	8.180 11.203
Tags_Not doing further education	-21.8992	2.04e+04	-0.001	0.999	-4.01e+04 4e+04
Tags_Ringing	-1.4829	0.294	-5.050	0.000	-2.058 -0.907
Tags_Tentative	3.6953	0.225	16.399	0.000	3.254 4.137
Tags_Will revert after reading the email	6.7310	0.261	25.747	0.000	6.219 7.243
Tags_invalid number	-23.2895	2.69e+04	-0.001	0.999	-5.27e+04 5.26e+04
Tags_switched off	-2.0637	0.622	-3.318	0.001	-3.283 -0.845
Last Notable Activity_Modified	-1.6106	0.126	-12.751	0.000	-1.858 -1.363
Last Notable Activity_Olark Chat Conversation	-1.6401	0.431	-3.807	0.000	-2.485 -0.796

Model-2

Model-3

```
# Let's re-run the model using the selected variables
X_train_sm = sm.add_constant(X_train[col])
logm3 = sm.GLM(y_train,X_train_sm, family = sm.families.Binomial())
res = logm3.fit()
res.summary()
```

4]: Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6325
Model:	GLM	Df Residuals:	6311
Model Family:	Binomial	Df Model:	13
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1237.2
Date:	Sun, 26 Nov 2023	Deviance:	2474.4
Time:	19:45:47	Pearson chi2:	1.16e+04
No. Iterations:	23	Pseudo R-squ. (CS):	0.6100
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.2840	0.198	-16.572	0.000	-3.672	-2.896
Lead Source_Welingak Website	2.7578	0.751	3.670	0.000	1.285	4.231
Last Activity_SMS Sent	2.0146	0.119	16.867	0.000	1.780	2.249
What matters most to you in choosing a course_Not_Selected	-2.6023	0.145	-17.916	0.000	-2.887	-2.318
Tags_Busy	2.7926	0.281	9.944	0.000	2.242	3.343
Tags_Closed by Horizzon	9.1721	0.743	12.350	0.000	7.716	10.628
Tags_Lost to EINS	9.8373	0.772	12.748	0.000	8.325	11.350
Tags_Not doing further education	-20.7315	1.24e+04	-0.002	0.999	-2.43e+04	2.43e+04
Tags_Ringing	-1.2937	0.292	-4.436	0.000	-1.865	-0.722
Tags_Tentative	3.8623	0.226	17.103	0.000	3.420	4.305
Tags_Will revert after reading the email	6.8911	0.263	26.245	0.000	6.376	7.406
Tags_switched off	-1.8727	0.621	-3.016	0.003	-3.090	-0.656
Last Notable Activity_Modified	-1.5923	0.126	-12.613	0.000	-1.840	-1.345
Last Notable Activity_Olark Chat Conversation	-1.6390	0.431	-3.806	0.000	-2.483	-0.795

Model-3

Model-4

```
# Let's re-run the model using the selected variables
X_train_sm = sm.add_constant(X_train[col])
logm4 = sm.GLM(y_train,X_train_sm, family = sm.families.Binomial())
res = logm4.fit()
res.summary()
```

6]: Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6325
Model:	GLM	Df Residuals:	6312
Model Family:	Binomial	Df Model:	12
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1239.4
Date:	Sun, 26 Nov 2023	Deviance:	2478.9
Time:	19:45:47	Pearson chi2:	1.17e+04
No. Iterations:	8	Pseudo R-squ. (CS):	0.6098
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.3615	0.197	-17.031	0.000	-3.748	-2.975
Lead Source_Welingak Website	2.7589	0.752	3.671	0.000	1.286	4.232
Last Activity_SMS Sent	2.0180	0.119	16.892	0.000	1.784	2.252
What matters most to you in choosing a course_Not_Selected	-2.6032	0.145	-17.913	0.000	-2.888	-2.318
Tags_Busy	2.8690	0.280	10.232	0.000	2.319	3.419
Tags_Closed by Horizzon	9.2526	0.742	12.463	0.000	7.798	10.708
Tags_Lost to EINS	9.9182	0.771	12.857	0.000	8.406	11.430
Tags_Ringing	-1.2187	0.291	-4.185	0.000	-1.790	-0.648
Tags_Tentative	3.9400	0.225	17.501	0.000	3.499	4.381
Tags_Will revert after reading the email	6.9701	0.262	26.617	0.000	6.457	7.483
Tags_switched off	-1.7979	0.621	-2.897	0.004	-3.014	-0.581
Last Notable Activity_Modified	-1.5958	0.126	-12.645	0.000	-1.843	-1.348
Last Notable Activity_Olark Chat Conversation	-1.6373	0.431	-3.800	0.000	-2.482	-0.793

Final Model-4

As we check in Model-4 all features are significant

Logistic Regression Model Building

Checking VIF (Variation Inflation Factor) for Model-4

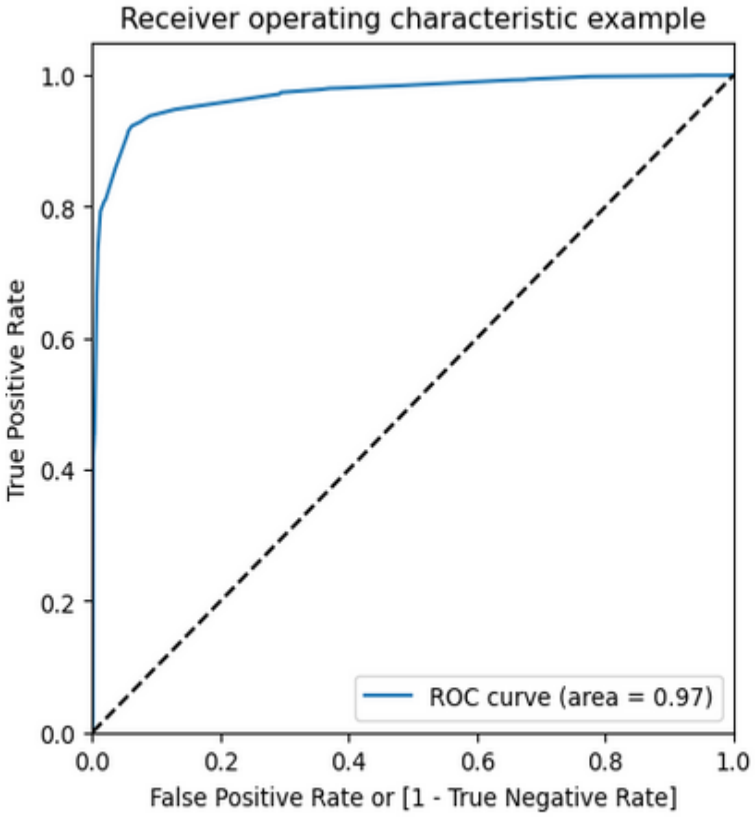
	Features	VIF
0	Lead Source_Welingak Website	1.13
4	Tags_Closed by Horizzon	1.07
5	Tags_Lost to EINS	1.06
11	Last Notable Activity_Olark Chat Conversation	1.05
3	Tags_Busy	1.04
9	Tags_switched off	1.03
2	What matters most to you in choosing a course_...	0.26
8	Tags_Will revert after reading the email	0.12
7	Tags_Tentative	0.11
1	Last Activity_SMS Sent	0.09
6	Tags_Ringing	0.08
10	Last Notable Activity_Modified	0.02

After building the logistic regression model with significant features and low multicollinearity, the next step is to predict outcomes on the training data. This allows us to evaluate the model's performance and its ability to generalize to the data it was trained on

Train Model - Metrics

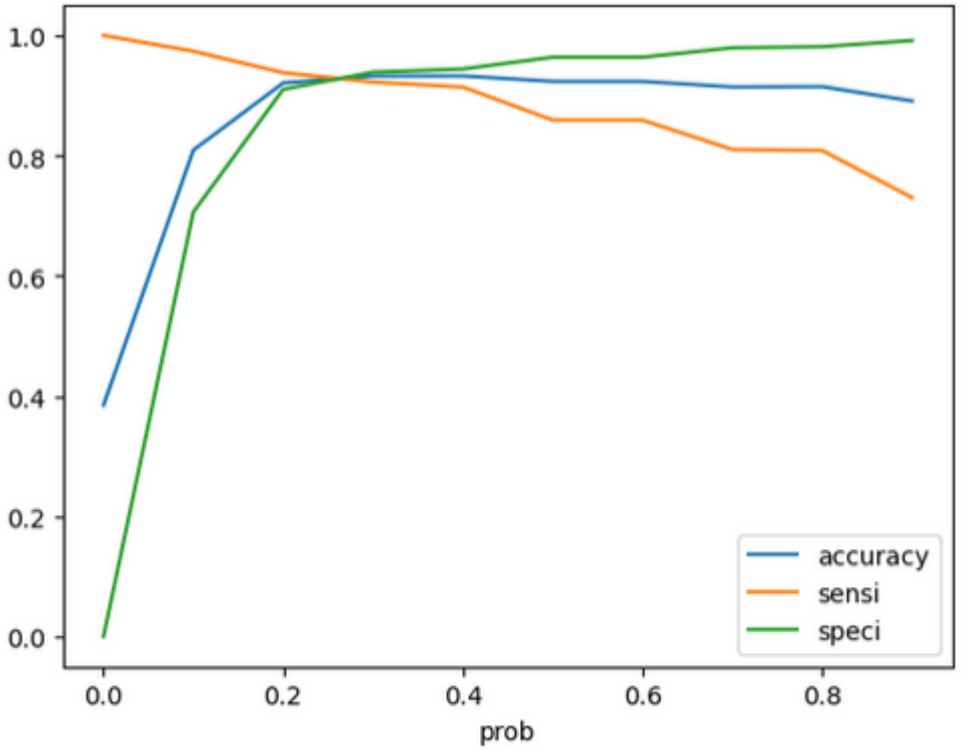
Overall Accuracy of Model = 92.3%
Sensitivity = 85.9%
Specificity = 96.3%
False Positive Rate = 3.6%
True Positive Rate = 93.6%
True Negative Rate = 91.6%

Plotting the ROC Curve



The Final Model-4 has Area Under Curve (AUC) value of 0.95, which is a very good indicator

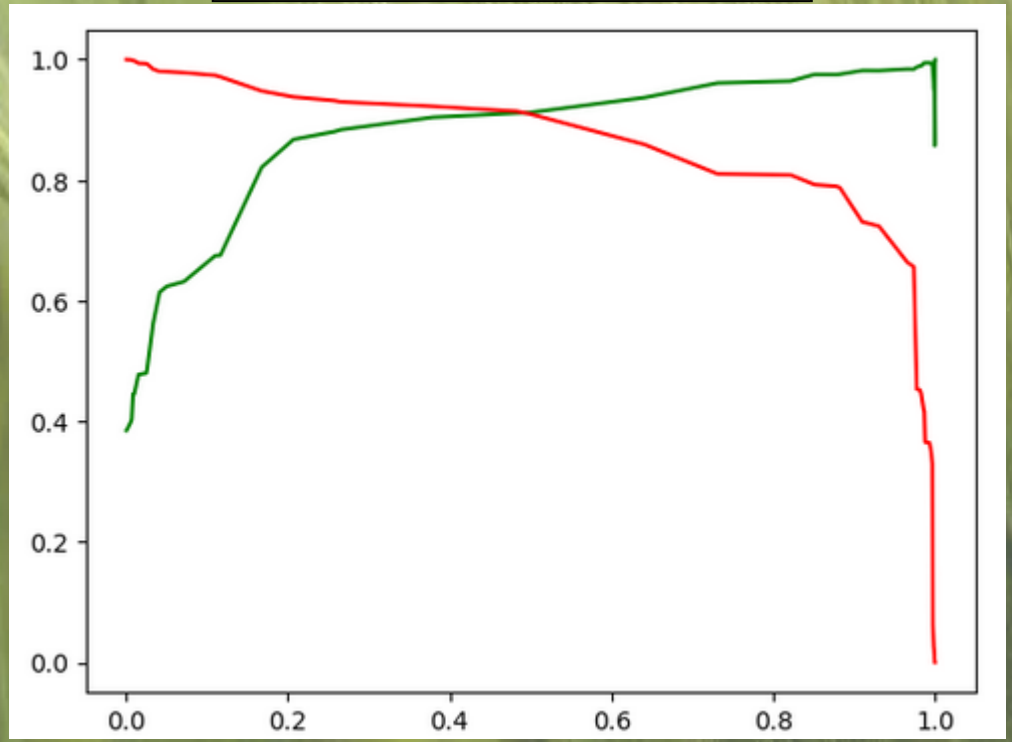
Optimal Cut-Off Point



From the curve above, 0.225 is the optimum point to take it as a cutoff probability.

Precision, Recall & Trade-off within

Precision = 93.6%
Recall = 85.9%
Specificity = 96.3%
False Positive Rate = 3.6%
True Positive Rate = 93.6%
True Negative Rate = 91.6%



NOTE: The intersection point of the curve is the threshold value where the model achieves a balance between precision and recall. It can be used to optimize the performance of the model based on business requirement, Here our probability threshold is 0.45 approx. from above curve.

Making Predictions on the Test Set & Conclusion

Train Model - Metrics

- Overall Accuracy of Model = 92.3%
- Sensitivity = 85.9%
- Specificity = 96.3%
- False Positive Rate = 3.6%
- True Positive Rate = 93.6%
- True Negative Rate = 91.6%
- [TN= 3582, FP= 308],
- [FN= 166, TP= 2269]

Test Model - Metrics

- Overall Accuracy of Model = 93.3%
- Sensitivity = 92.7%
- Specificity = 93.7%
- [TN=1568, FP=105],
- [FN=75, TP=964],

Logistic Regression Model for Lead Conversion

The probability expression of the model can be written as:

$\ln(p1-p)=$

− 3.3615 + 2.7589×Lead Source_Welingak Website + 2.0180×Last Activity_SMS Sent − 2.6032×What matters most to you in choosing a course_Not_Selected + 2.8690×Tags_Busy + 9.2526×Tags_Closed by Horizzon + 9.9182×Tags_Lost to EINS − 1.2187×Tags_Ringing + 3.9400×Tags_Tentative + 6.9701×Tags_Will revert after reading the email − 1.7979×Tags_switched off − 1.5958×Last Notable Activity_Modified − 1.6373×Last Notable Activity_Olark Chat Conversation

This logistic regression model is based on the provided coefficients and predictor variables.