LeadScoringCaseStudy
(LogisticRegression)

Submitted by : Vijaya
Hiep Dang
Saravanan R

Business Objective

- This Logistic Regression model is designed to select the most promising leads.

- This logistic regression model aims 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.

Datasets Used

- *1. ‘Leads.csv'* contains all the information of the user who has visited the website.

- *2. 'Leads Data Dictionary'* is data dictionary which describes the meaning of the variables / or columns .
- This is provided to understand ‘Leads.csv’ dataset.

Data Understanding and Preparation

- Basic Data Understanding:
- --- checking for columns and their data types
- --- checking shape of data
- --- identifying required columns based on Business objective

- Data Cleaning:
- ---A few columns seem to have Missing Values disguised as 'Select' , replacing 'select' with null values.
- --- Checking for Missing values [if Percentage of missing values is greater then 40% , check the columns , if they do not seem much related , drop the columns]

Data Cleaning (continue…..)

- Data Cleaning (continue…..)
- Data Cleaning for Binary Columns:
- --- Dropping highly skewed columns like 'Magazine','Newspaper Article','Do Not Call','Search',where only one value is holding majority of data.
- --- Replacing null values of some columns like 'Lead Source' with mode value and then handling the skewness by combining all the multiple categorical values constituting very less to the data in a single variable like 'Social Media'.

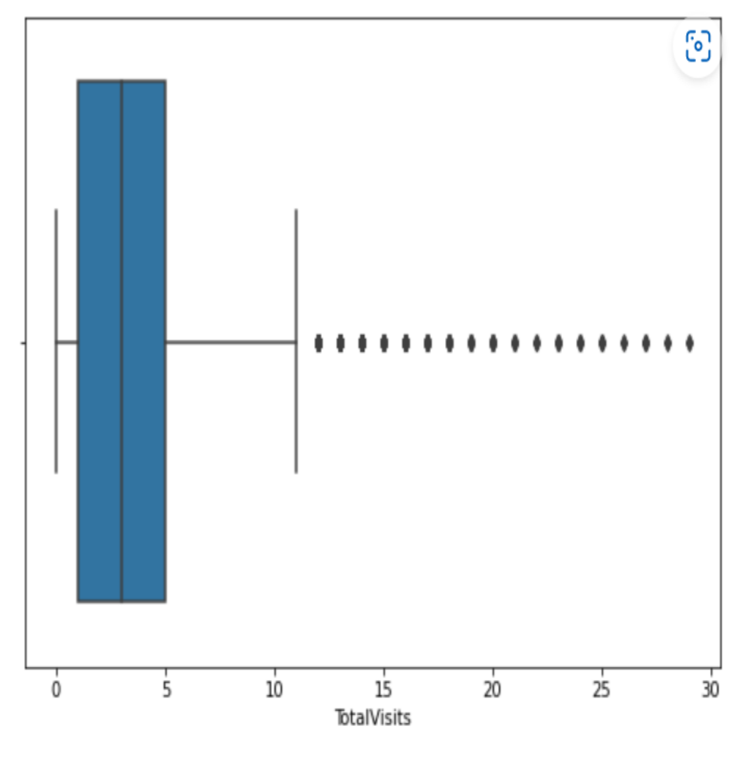
- Data cleaning for numerical columns:
- Replacing missing values of some columns like 'Page Views Per Visit' with median values.

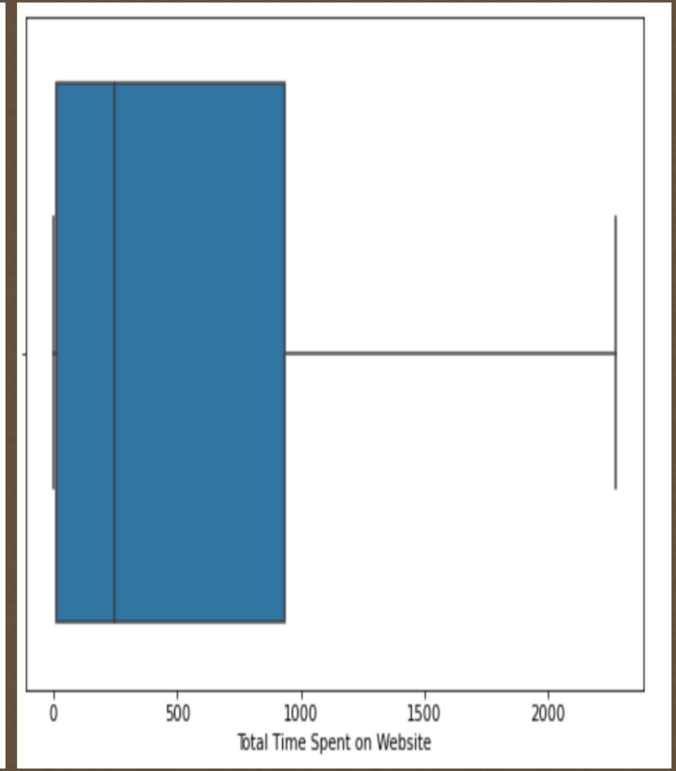
Exploratory Data Analysis

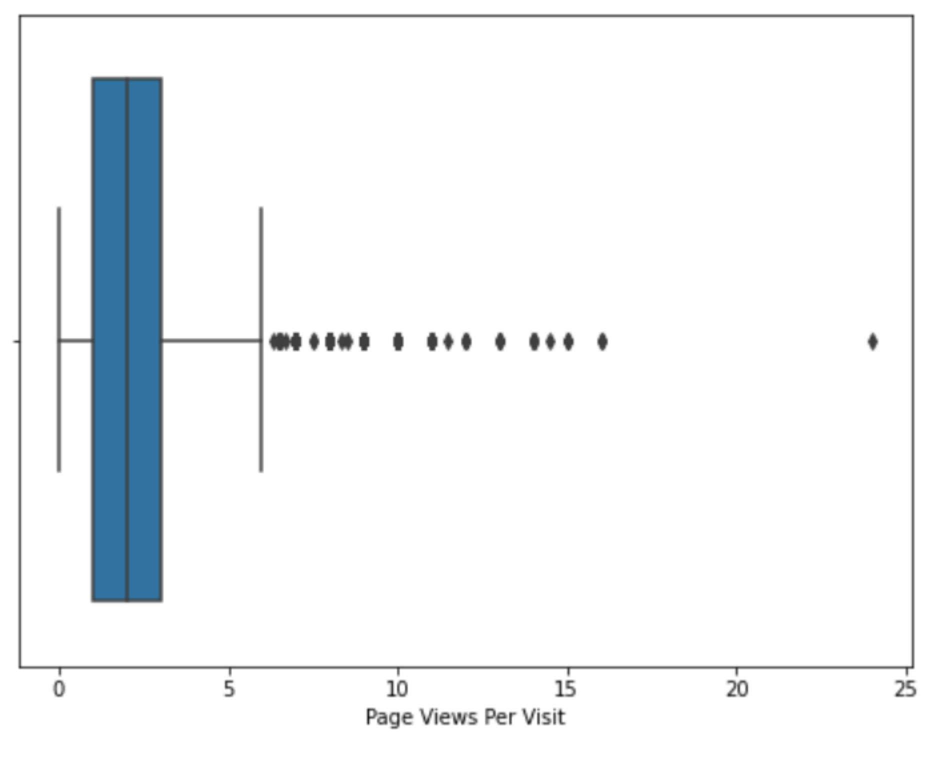
- Univariant Analysis:
- --- Analysing variables like TotalVisits, Total Time Spent on Website, Page Views Per Visit etc . To find hidden patterns in them.

- Bivariant Analysis:
- --- Checking for two variables like ‘Lead Origin’ and ‘Total Time Spent on Website’ , as how this is affecting conversion rate.
- --- Checking for two variables like ‘Lead Source’ and ‘Total Time Spent on Website’ , as how this is affecting conversion rate. Etc.

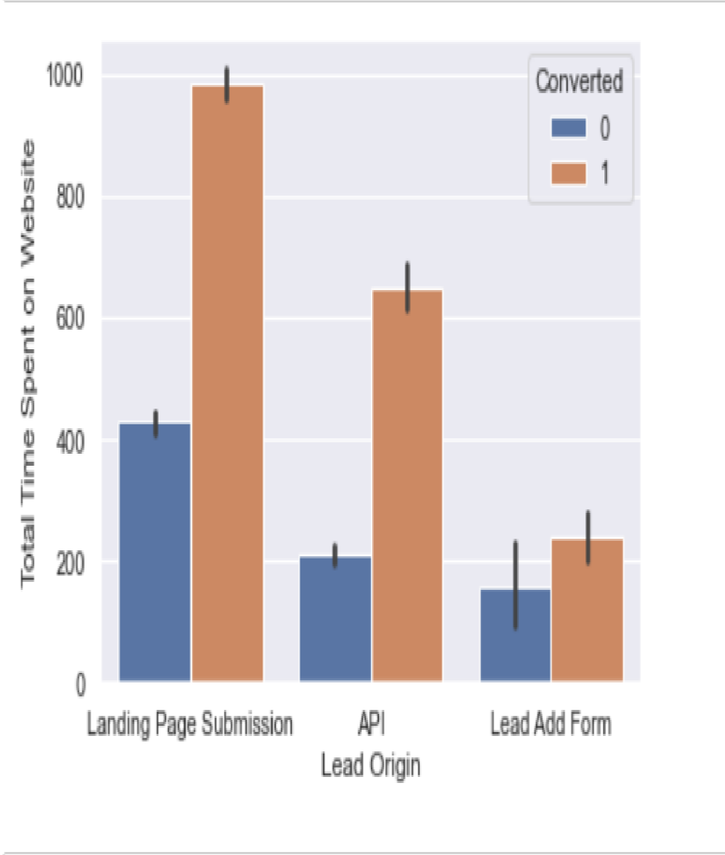
Exploratory Data Analysis (continue …)
Univariant Analysis

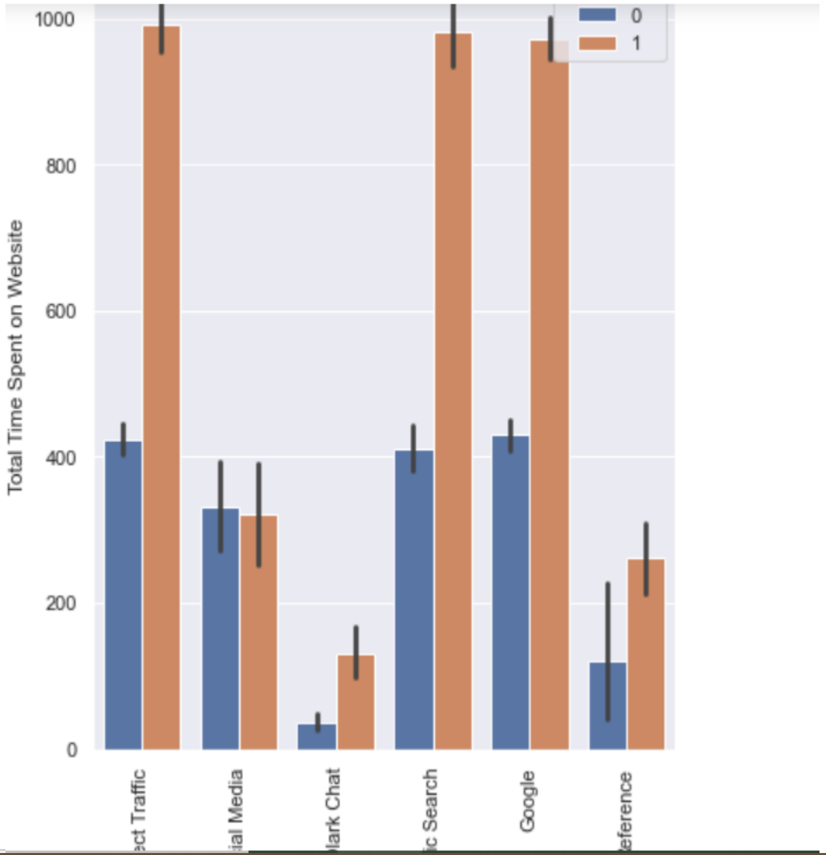


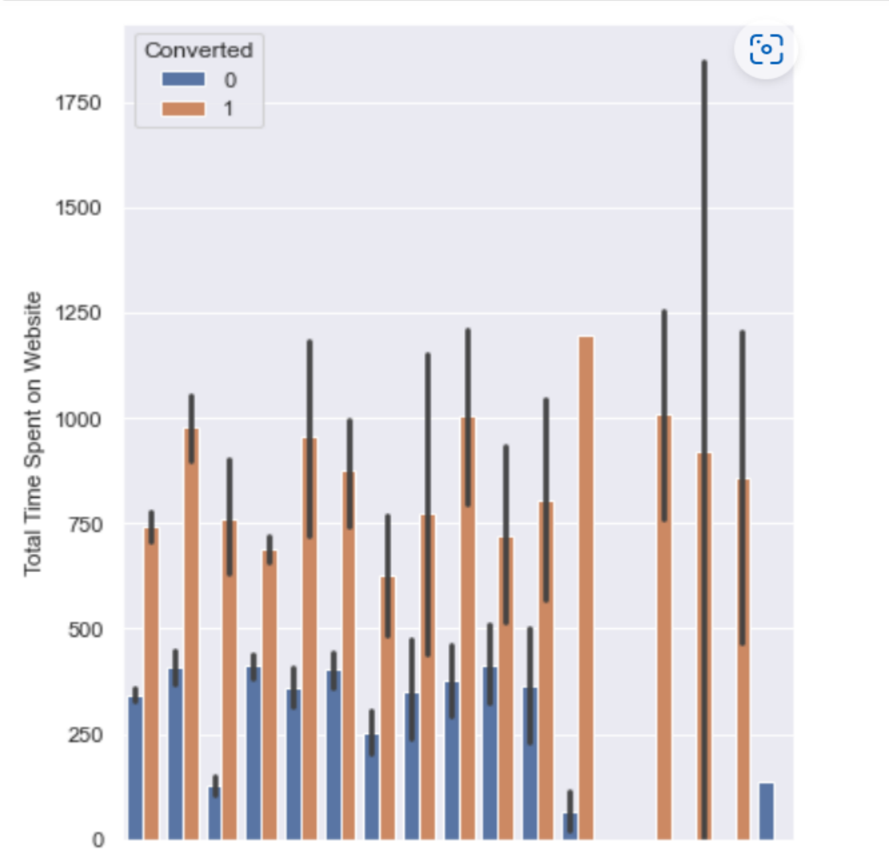




Exploratory Data Analysis (continue …)
Multivariant Analysis







Data Preparation for Logistic Regession Model

--- Creating dummies for categorical Variable.

Identifying all the categorical columns and creating dummies for these columns

--- Performing Train-Test split(70-30 ratio).

Splitting the entire data set into train_data (70%) and test_data(30%).

--- Scaling the data.

using MinMaxScaler to scale the dataset.

--- Defining X_variable: Independent Variables which will affect target variable.

--- Defining Y-Variable : Dependent/Target variable.

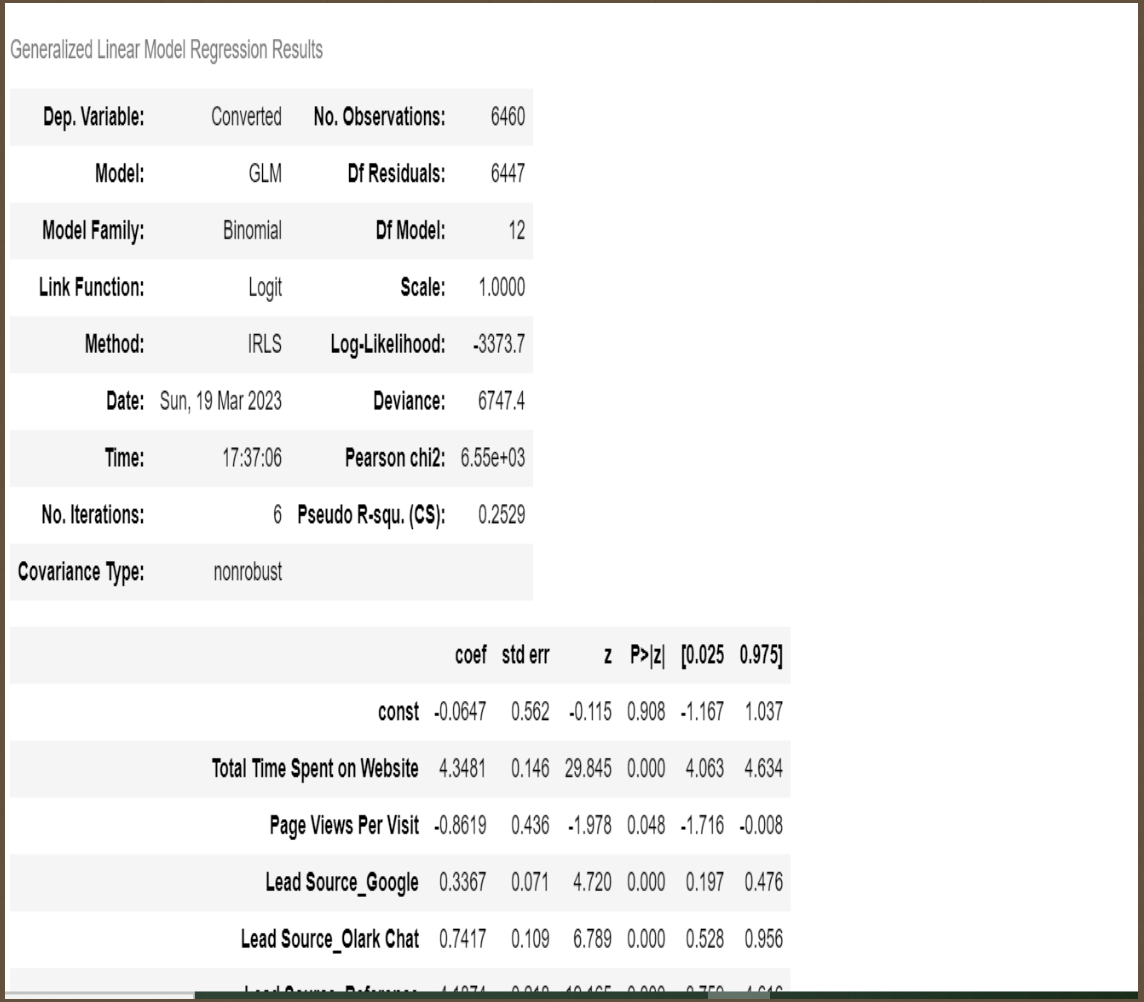
--- Checking correlation between the numerical columns to understand the data better and dropping the columns such as 'Lead Origin_Landing Page Submission' , 'TotalVisits' ,' A free copy of Mastering The Interview' etc which are causing high collinearity.

Model Building

- Steps involved in Model Building:

- --- Variable selection using RFE
- --- Building a Logistic Model with good sensitivity
- --- Check p-value and VIF.
- --- Find optimal probability cut-off
- --- Check model performance over test data
- ---Generate Score variable

Model Building (continue…)
First Model (Left)

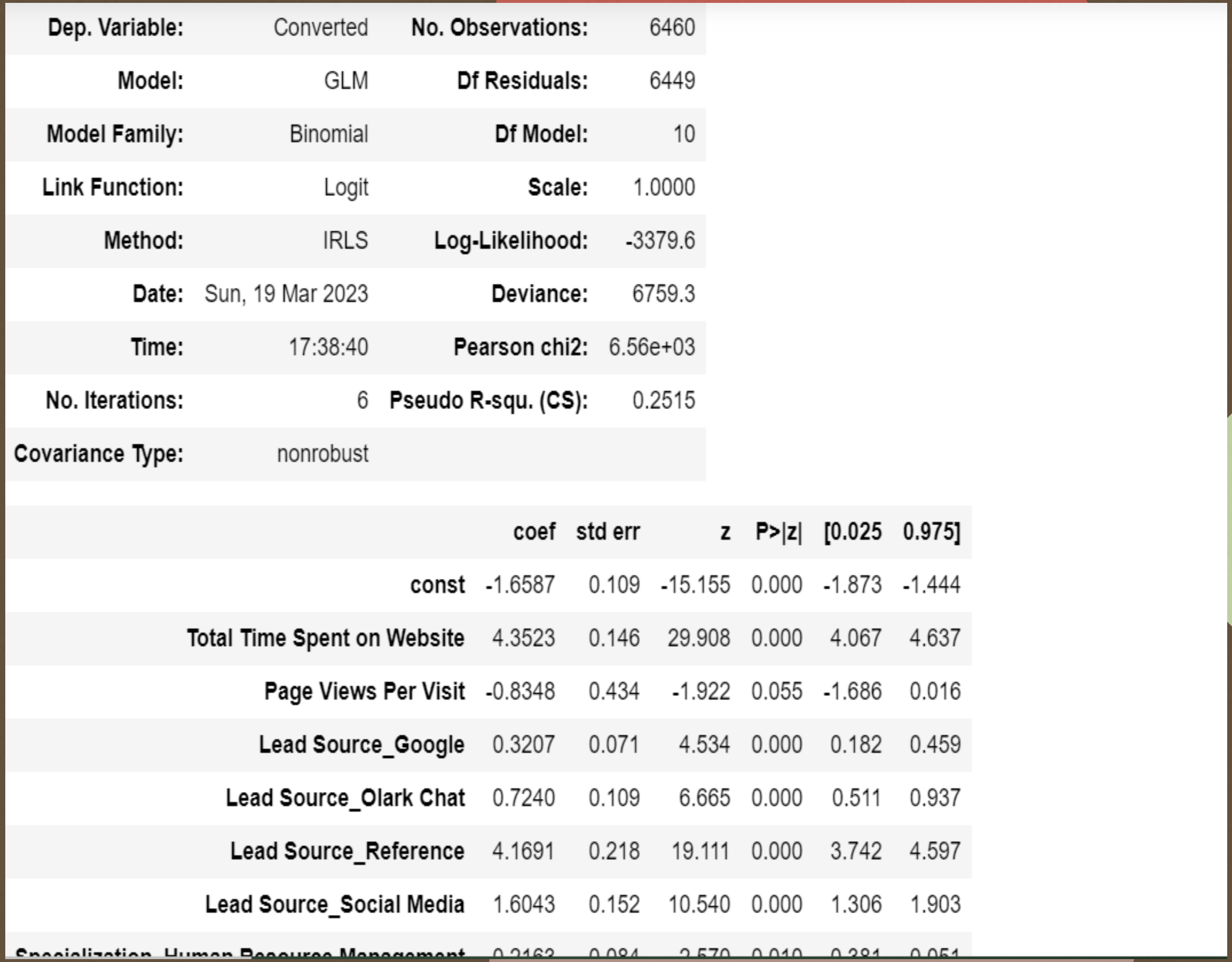


Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6460
Model:	GLM	**Df Residuals:**	6447
Model Family:	Binomial	**Df Model:**	12
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-3373.7
Date:	Sun, 19 Mar 2023	**Deviance:**	6747.4
Time:	17:37:06	**Pearson chi2:**	6.55e+03
No. Iterations:	6	**Pseudo R-squ. (CS):**	0.2529
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0647	0.562	-0.115	0.908	-1.167	1.037
Total Time Spent on Website	4.3481	0.146	29.845	0.000	4.063	4.634
Page Views Per Visit	-0.8619	0.436	-1.978	0.048	-1.716	-0.008
Lead Source_Google	0.3367	0.071	4.720	0.000	0.197	0.476
Lead Source_Olark Chat	0.7417	0.109	6.789	0.000	0.528	0.956
Lead Source_Reference	4.1874	0.218	19.165	0.000	3.750	4.616

Final Model (after dropping high VIF and P-value columns



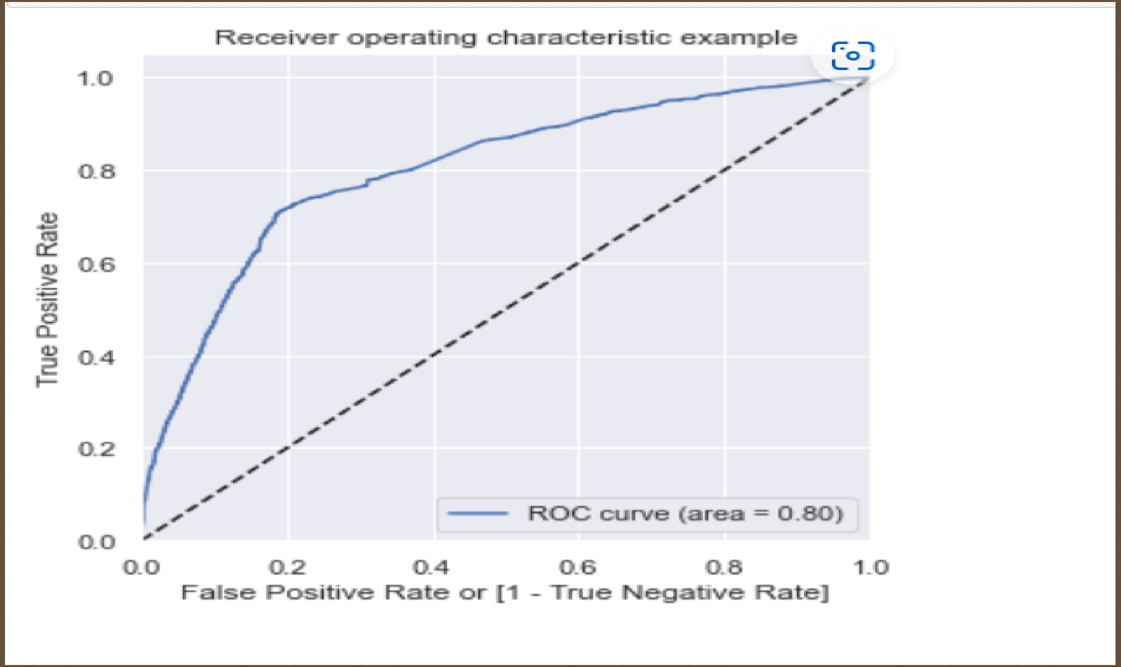
Dep. Variable:	Converted	**No. Observations:**	6460
Model:	GLM	**Df Residuals:**	6449
Model Family:	Binomial	**Df Model:**	10
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-3379.6
Date:	Sun, 19 Mar 2023	**Deviance:**	6759.3
Time:	17:38:40	**Pearson chi2:**	6.56e+03
No. Iterations:	6	**Pseudo R-squ. (CS):**	0.2515
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.6587	0.109	-15.155	0.000	-1.873	-1.444
Total Time Spent on Website	4.3523	0.146	29.908	0.000	4.067	4.637
Page Views Per Visit	-0.8348	0.434	-1.922	0.055	-1.686	0.016
Lead Source_Google	0.3207	0.071	4.534	0.000	0.182	0.459
Lead Source_Olark Chat	0.7240	0.109	6.665	0.000	0.511	0.937
Lead Source_Reference	4.1691	0.218	19.111	0.000	3.742	4.597
Lead Source_Social Media	1.6043	0.152	10.540	0.000	1.306	1.903
Specialization_Human Resource Management	0.2163	0.084	2.570	0.010	0.381	0.051

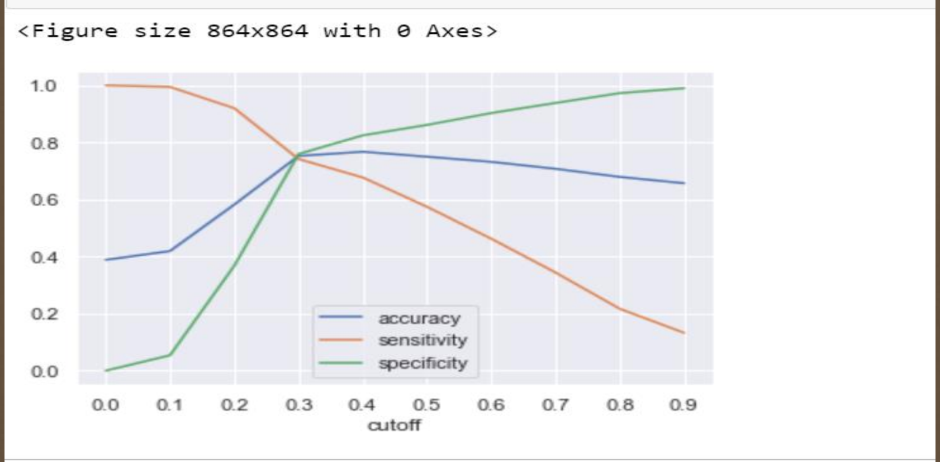
Model Evaluation

- Assessing the Model:
- --- Prediction of Hot_leads on Train_set first.

- ROC Curve:
- ---



- **Optimal Threshold**
- **---** Probability Threshold is 0.29 (closer to 0.3) But considering 0.25 to tradeoff sensitivity against accuracy



```
<Figure size 864x864 with 0 Axes>
```

- **Other Steps Involved in assessing the model:**
- --- Calculating confusion matrix
- --- Checking Sensitivity
- Sensitivity for Training set – 79.9%
- --- Checking Accuracy
- Accuracy on training set : 69.96%
- --- Checking specificity
- Specificity on train set : 63.73
- --- Checking for precison Score
- --- Checking for Recall Score

Final Step: Making Prediction on Test Set

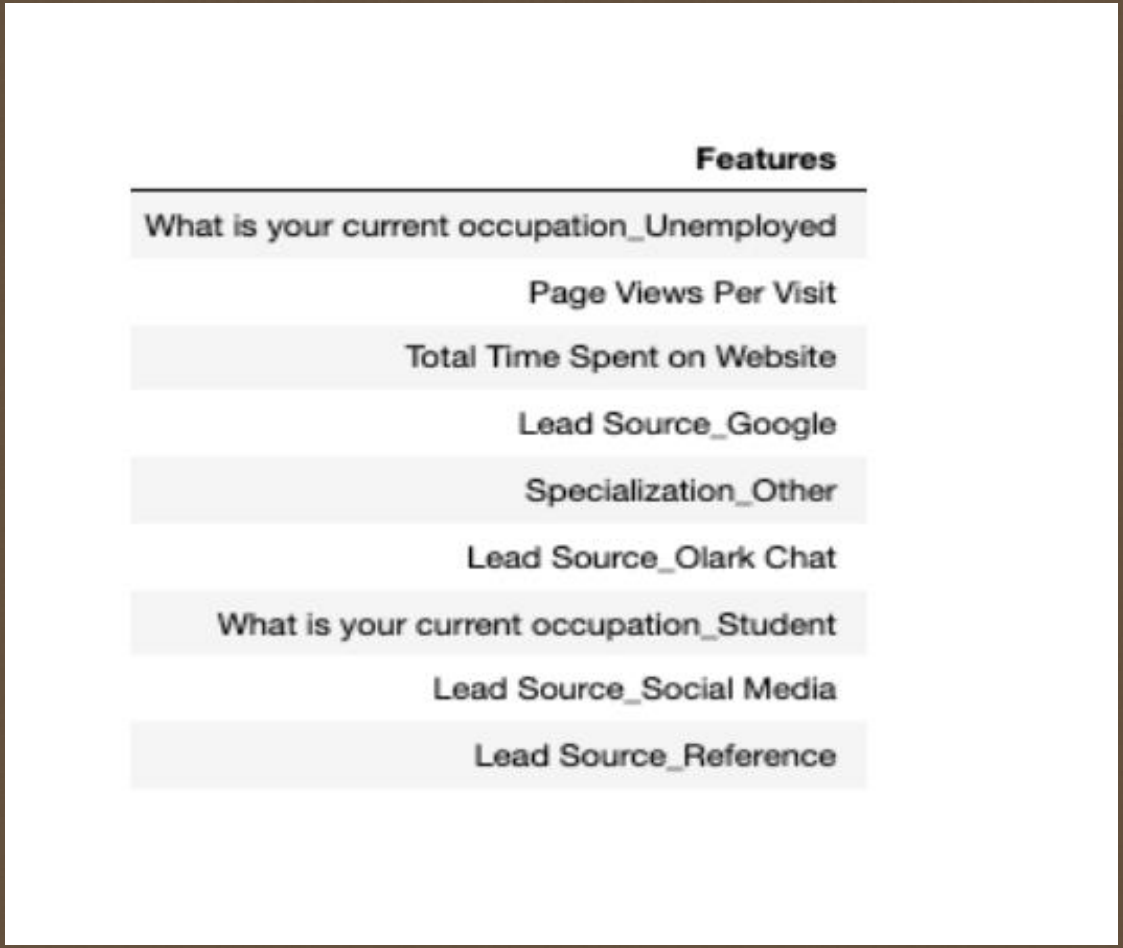
- --- Scaling test data set . (only transform and no fit is done).
- --- Making prediction on test data set using the created model.

- Checking / Assessing the model over test data.
- --- Calculating confusion matrix
- --- Checking Sensitivity
- Sensitivity for Test set – 79.78%
- --- Checking Accuracy
- Accuracy on test set : 69.33%
- --- Checking specificity
- Specificity on test set : 64.78
- --- Checking for precison Score
- --- Checking for Recall Score

- *** **Difference between Sensitivity scores of Train and Test = 2.74%**

-

Variables Impacting the Target variable (Conversion)



Features
What is your current occupation_Unemployed
Page Views Per Visit
Total Time Spent on Website
Lead Source_Google
Specialization_Other
Lead Source_Olark Chat
What is your current occupation_Student
Lead Source_Social Media
Lead Source_Reference

Conclusions:

Some of the Business inferences we can make are :

- --- If we see the model , variable 'Total Time Spent on Website' has high and positive coeffient indicating that maximum the time spent on website by any user , higher is the chance of getting converted.
- --- Variables 'Page Views Per Visit' , 'Specialization_Human Resource Management' , 'What is your current occupation_Unemployed' has negative coefficients.
- --- From Business Perspective we can infer that   Unemployed users , student users, Users with Human Resource specialization are less chances of conversion.
- Lead sources coming from 'Olark Chat' or 'Google' have higher chances of conversion.
- Also Sources coming from 'References' (Lead Source_Reference) has high positive coefficient indicating these leads has higher chance of conversion