Lead-Scoring Assignment

By:

Manne Ashrit Kumar.

Sujata swain.

About Case Study

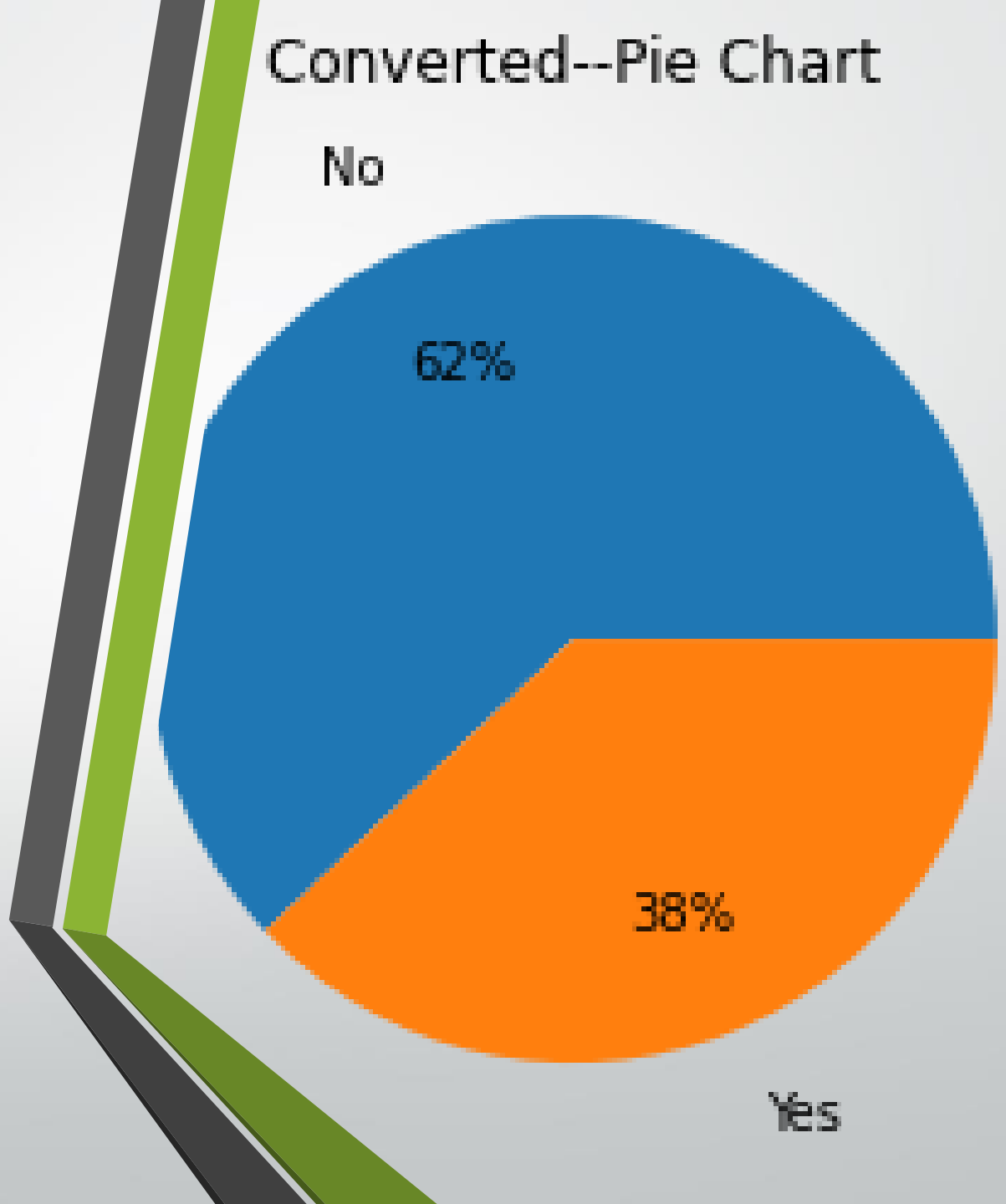
- This case study is about an online education company named X Education which sells online courses to industry professionals.
- The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for 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.
- 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%
- X Education's lead conversion rate is very poor. For more efficiency, the company wishes to identify the most potential leads, also known as 'Hot Leads'.
- The company requires us to build a model where in we need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

Objective

- 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 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.

About Dataset

- A leads dataset from the past with around 9000 data points has been provided.
- This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not.
- The target variable, in this case, is the column 'Converted' which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted.



Solution Methodology

Data Cleaning And Preparation

- Handling Missing values by dropping columns(high % of NULL VALUES), Imputation.
- Standardizing variables.
- Fixing Invalid Data-types and Filter Data(Correcting Data-types, Quality Checks.)
- Handling Outliers.
- Feature Scaling of numerical columns.
- Creation of dummy variables for categorical columns.

Building and Evaluation of ML Model

- Logistic Regression for building and prediction.
- Using manual approach and RFE automated search for feature/variable reduction to get desired model which used for prediction.
- Validation of Model
- Conclusion.

Data Cleaning And Preparation

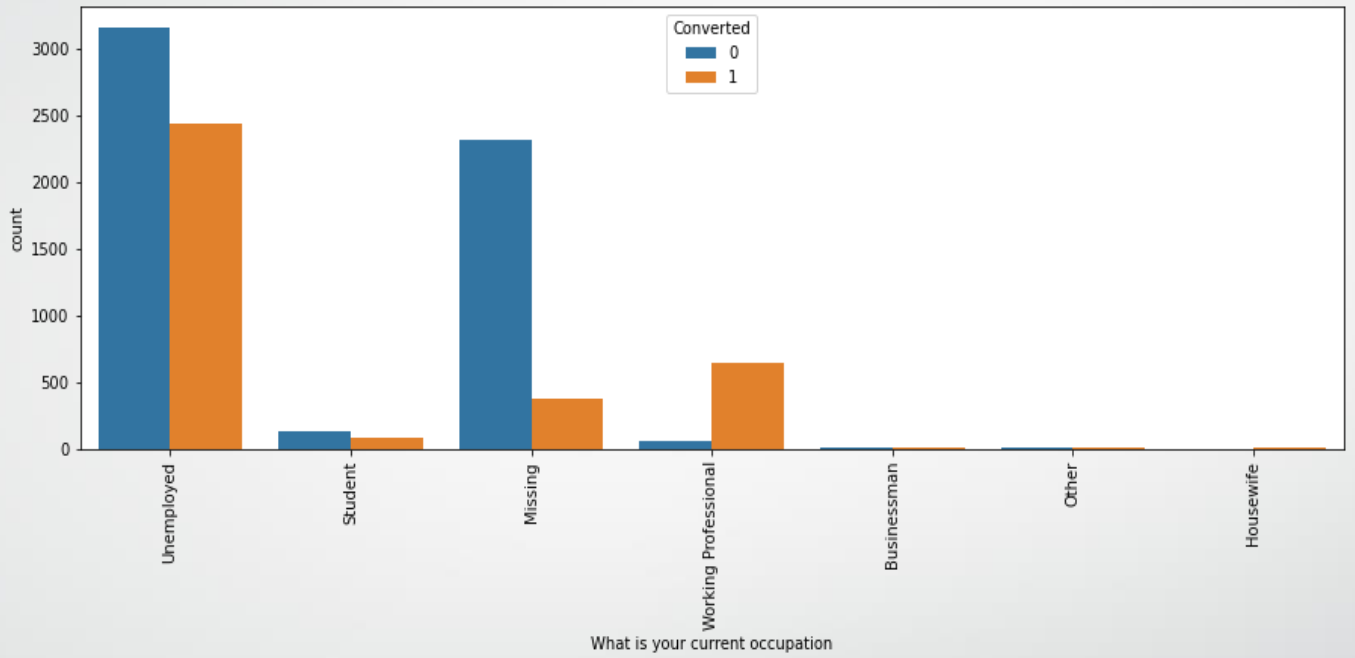
Final Data-Set

```
 30  Lead Source_Social Media                          8953 non-null   uint8
 31  Lead Source_Welingak Website                      8953 non-null   uint8
 32  Last Activity_Converted to Lead                   8953 non-null   uint8
 33  Last Activity_Email Bounced                       8953 non-null   uint8
 34  Last Activity_Email Link Clicked                  8953 non-null   uint8
 35  Last Activity_Email Opened                        8953 non-null   uint8
 36  Last Activity_Form Submitted on Website           8953 non-null   uint8
 37  Last Activity_Olark Chat Conversation             8953 non-null   uint8
 38  Last Activity_Page Visited on Website             8953 non-null   uint8
 39  Last Activity_SMS Sent                            8953 non-null   uint8
 40  Last Notable Activity_Email Link Clicked          8953 non-null   uint8
 41  Last Notable Activity_Email Opened                8953 non-null   uint8
 42  Last Notable Activity_Modified                    8953 non-null   uint8
 43  Last Notable Activity_Olark Chat Conversation     8953 non-null   uint8
 44  Last Notable Activity_Page Visited on Website     8953 non-null   uint8
 45  Last Notable Activity_SMS Sent                    8953 non-null   uint8
dtypes: float64(2), int64(3), uint8(41)
memory usage: 778.1 KB
```

```
: lead_df3.shape
```

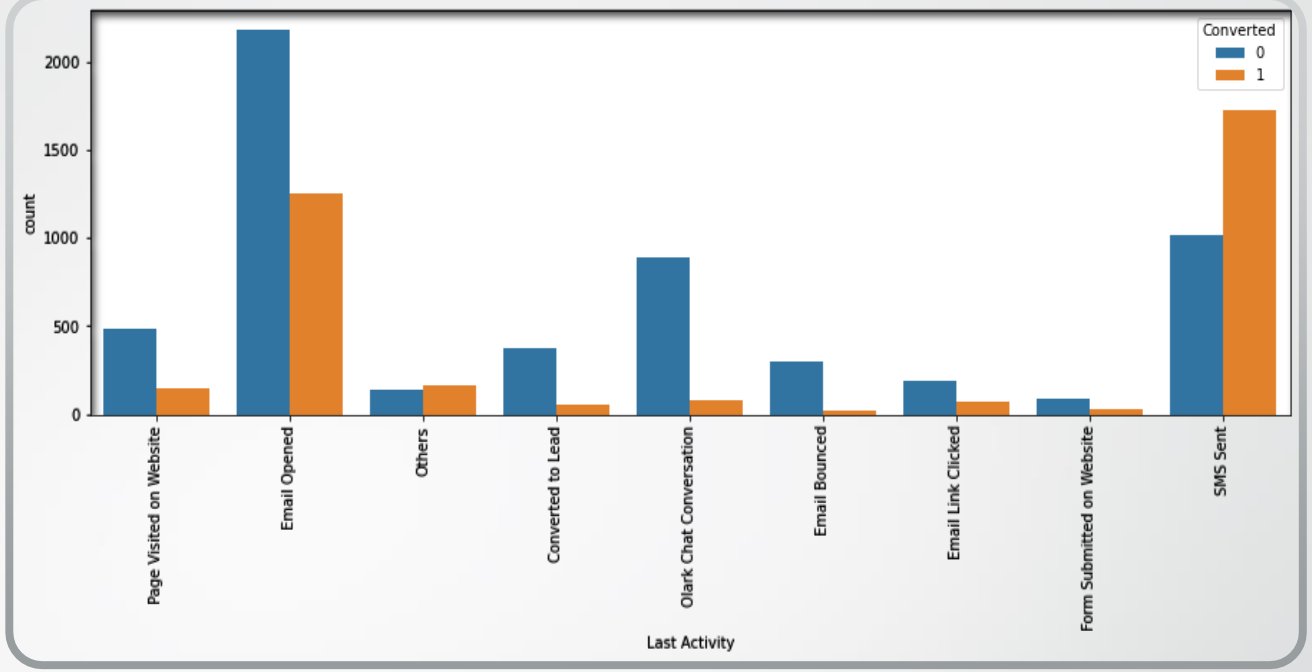
```
: (8953, 46)
```

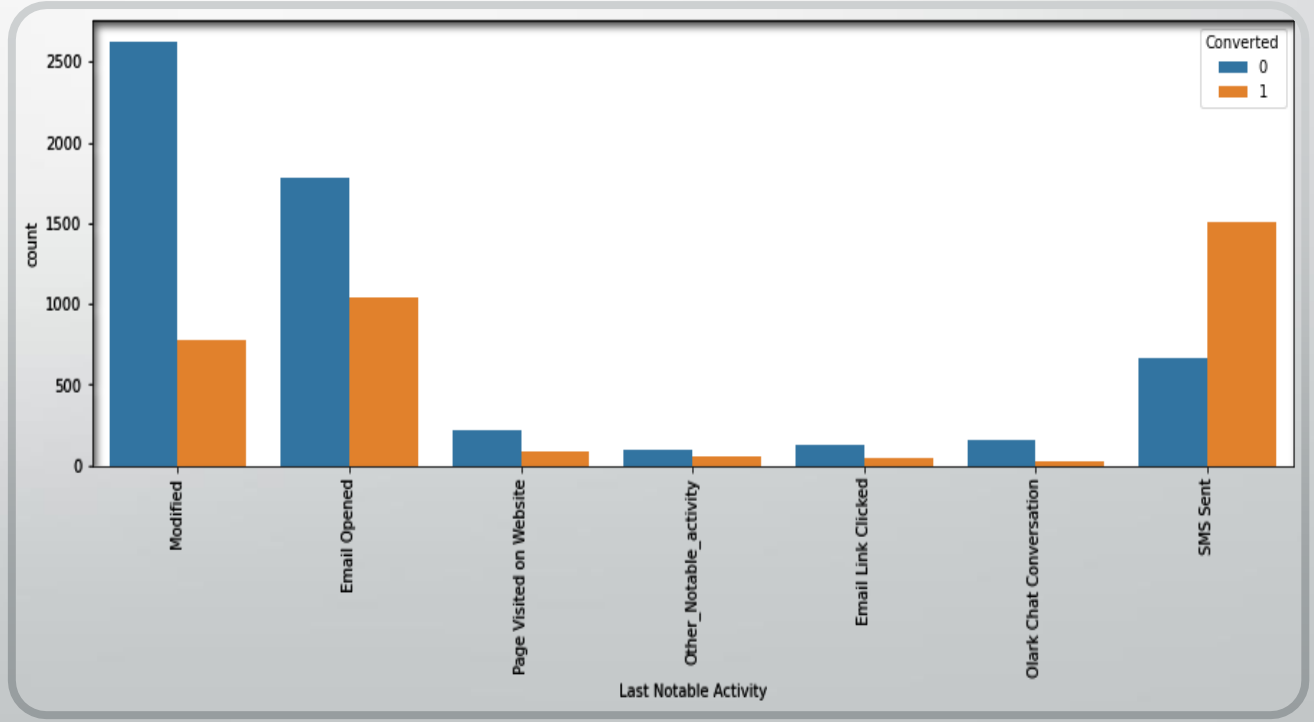
Categorical Univariate Analysis



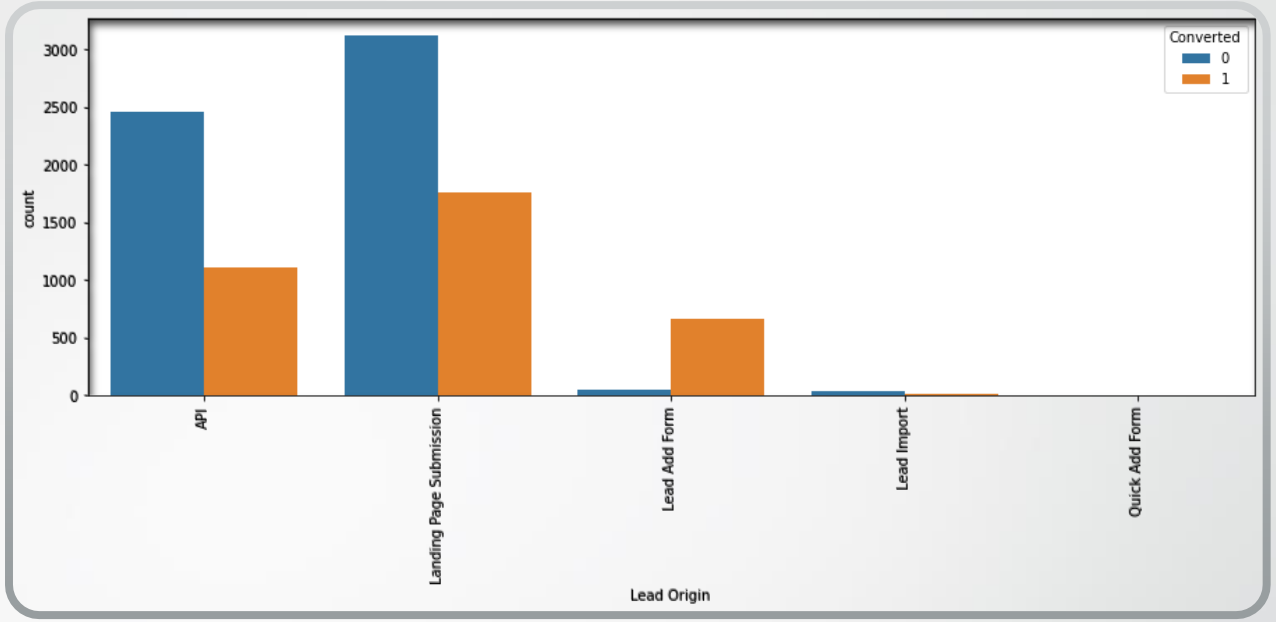
Occupation

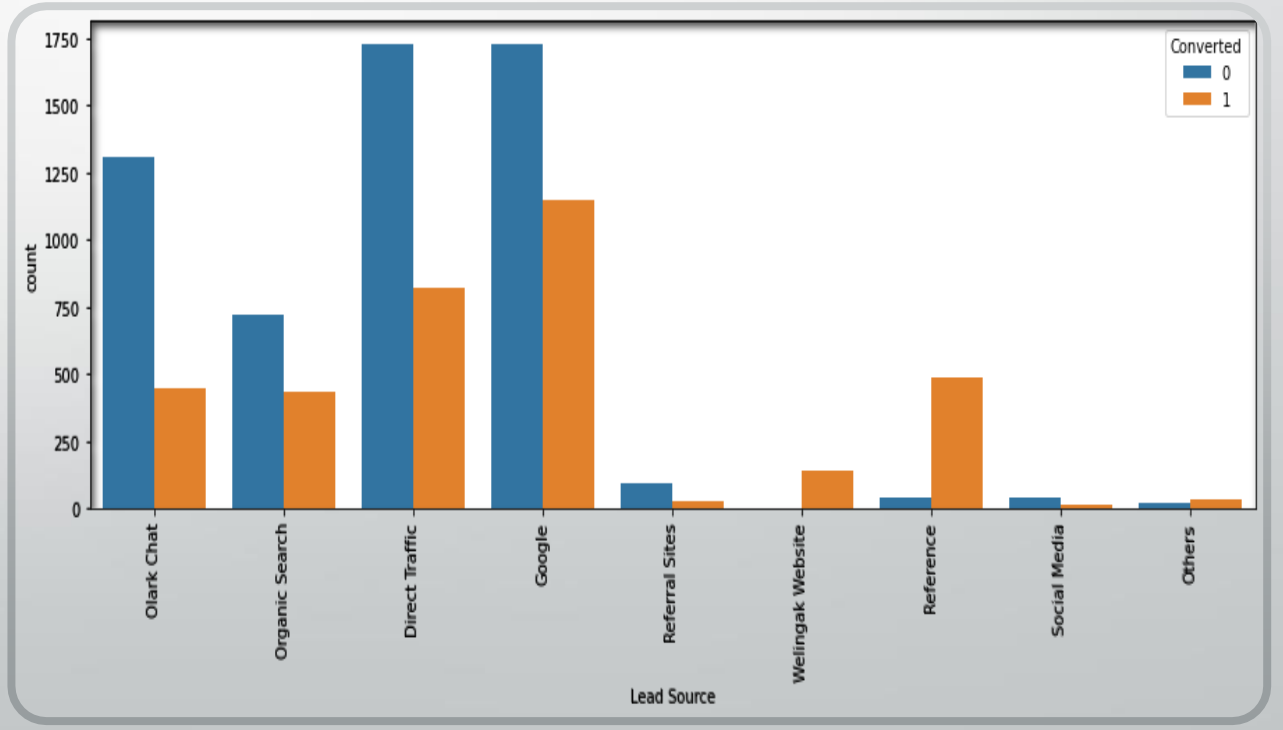
Last Activity and Last Notable Activity (hue-Converted)

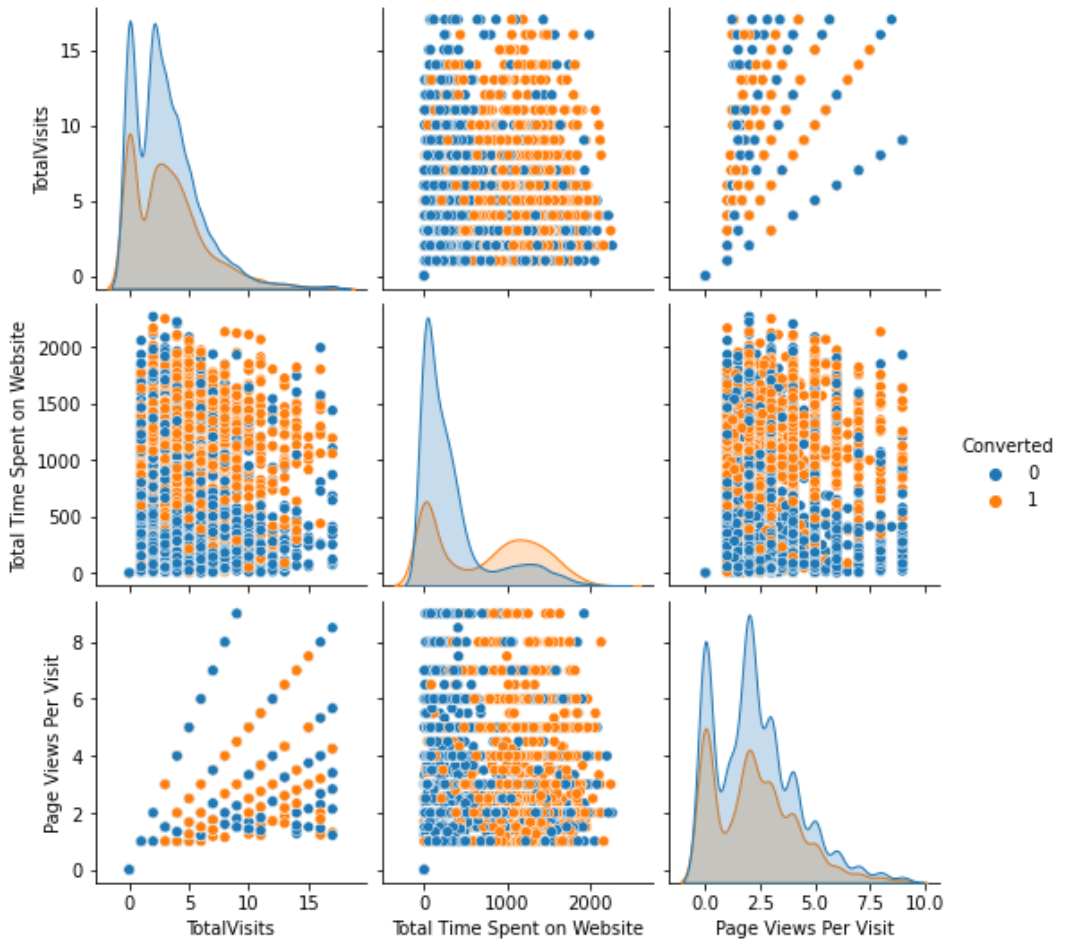




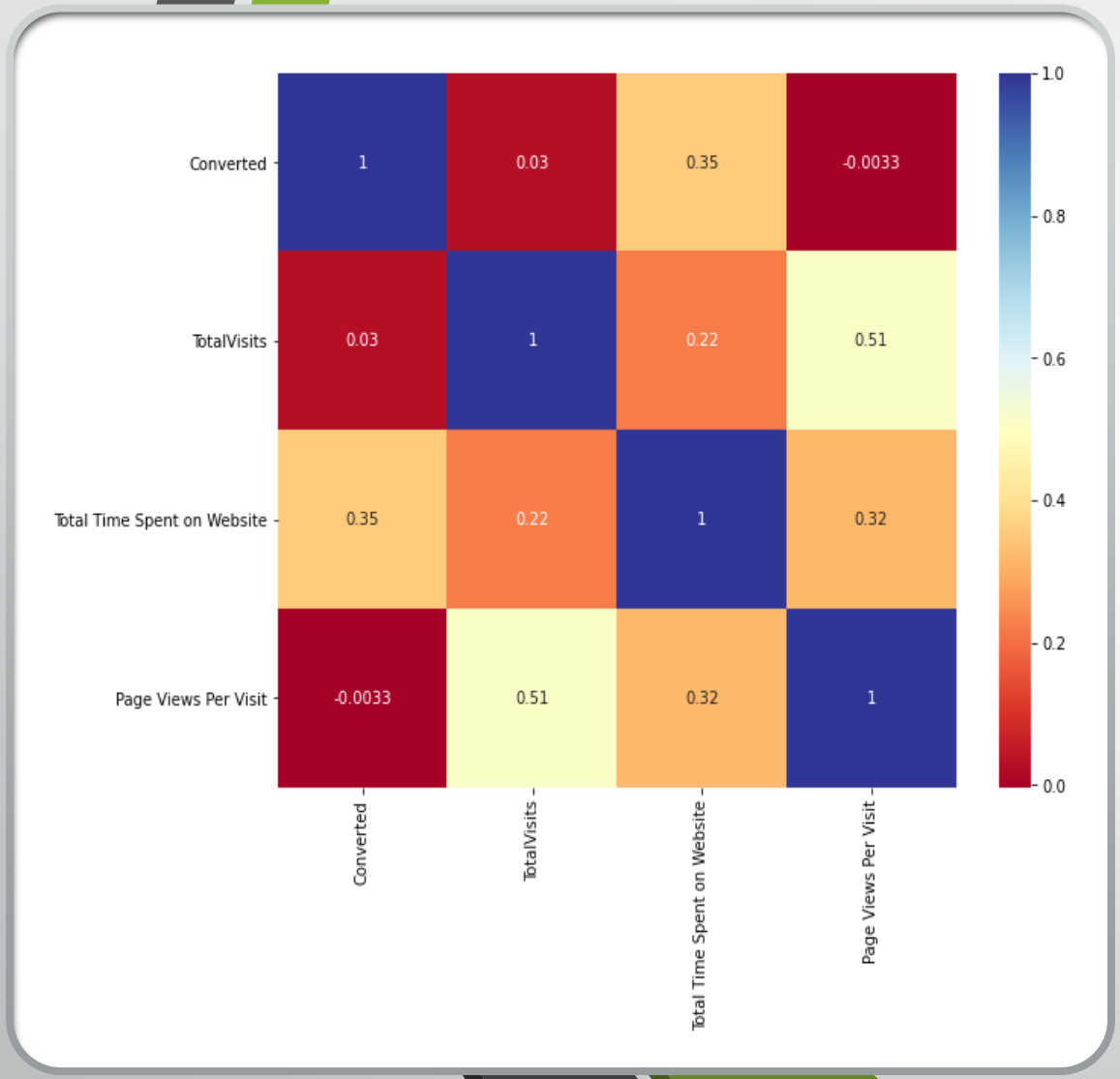
Lead Source and Lead Source Activity(hue-Converted)







Numerical Bi-Variate Analysis



Correlation Heatmap(Numerical Variables)

ML– Model Building

- Splitting the final data into train and test data.
- Train-Test split ratio 70:30.
- Building a first model using logistic regression.
- Using RFE for Feature Selection with 20 variables as output variable.
- Eliminating variables with p-value > 0.05 and VIF(Variance Inflation Factor)>5.
- Training Train Dataset
- Predictions on Test Dataset.

Final Model--Summary

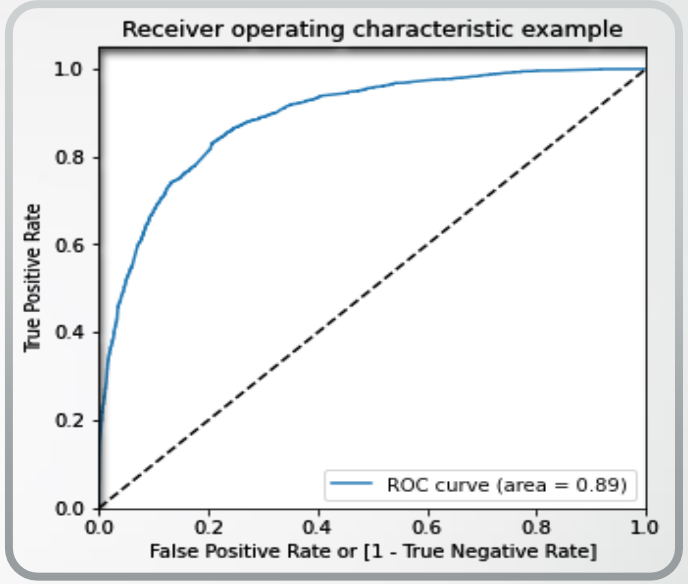
Generalized Linear Model Regression Results

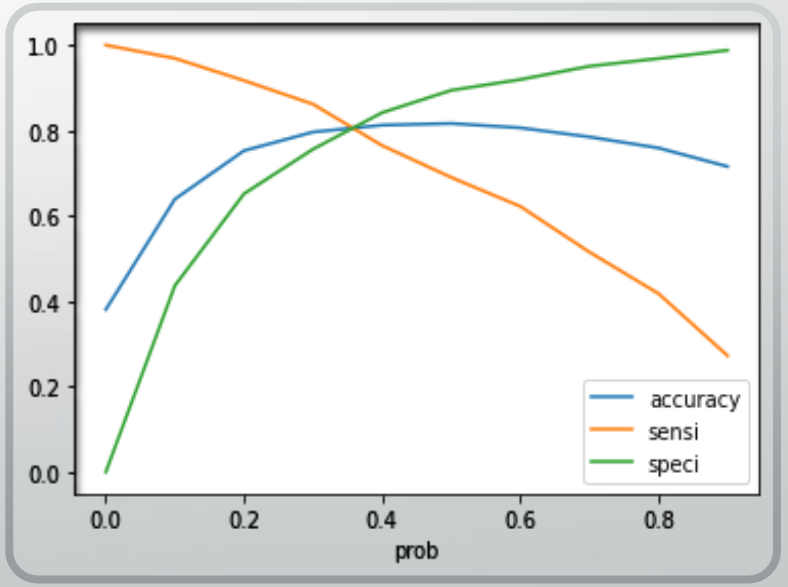
Dep. Variable:	Converted	No. Observations:	6267
Model:	GLM	Df Residuals:	6248
Model Family:	Binomial	Df Model:	18
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2541.7
Date:	Wed, 12 Jan 2022	Deviance:	5083.4
Time:	11:26:25	Pearson chi2:	6.20e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.7389	0.112	-6.596	0.000	-0.958	-0.519
Total Time Spent on Website	1.0859	0.041	26.781	0.000	1.006	1.165
Lead Origin_Lead Add Form	2.6306	0.250	10.536	0.000	2.141	3.120
What is your current occupation_Businessman	2.3791	1.163	2.046	0.041	0.100	4.659
What is your current occupation_Student	1.1429	0.248	4.617	0.000	0.658	1.628
What is your current occupation_Unemployed	1.1605	0.087	13.302	0.000	0.989	1.331
What is your current occupation_Working Professional	3.6440	0.208	17.501	0.000	3.236	4.052
Lead Source_Direct Traffic	-1.6147	0.117	-13.756	0.000	-1.845	-1.385
Lead Source_Google	-1.1279	0.112	-10.054	0.000	-1.348	-0.908
Lead Source_Organic Search	-1.2869	0.135	-9.498	0.000	-1.552	-1.021
Lead Source_Referral Sites	-1.3609	0.344	-3.953	0.000	-2.036	-0.686
Lead Source_Welingak Website	2.2233	1.039	2.140	0.032	0.187	4.260
Last Activity_Email Bounced	-1.4247	0.324	-4.393	0.000	-2.060	-0.789
Last Activity_Olark Chat Conversation	-0.9095	0.198	-4.591	0.000	-1.298	-0.521
Last Activity_SMS Sent	1.1225	0.077	14.522	0.000	0.971	1.274
Last Notable Activity_Email Link Clicked	-0.5812	0.271	-2.146	0.032	-1.112	-0.050
Last Notable Activity_Modified	-0.8008	0.085	-9.387	0.000	-0.968	-0.634
Last Notable Activity_Olark Chat Conversation	-1.1046	0.433	-2.550	0.011	-1.954	-0.255
Last Notable Activity_Page Visited on Website	-0.4481	0.216	-2.074	0.038	-0.872	-0.025

Finding Optimal Cutoff

- For ROC-graph, we are getting a good value of 0.89 indicating a good predictive model
- From second graph, the optimal cutoff turned out to be 0.35.





Final Observation

Train Data:

- Accuracy : 80.54%
- Sensitivity : 83.56%
- Specificity : 78.69%

Test Data:

- Accuracy : 79.97%
- Sensitivity : 86.83%
- Specificity : 75.83%

- After obtaining Lead_Score from table(shown), the sales should actively pursue leads with Lead Score more than 60%.

	Prospect ID	Converted	Converted_prob	Lead_Score	final_Predicted
0	7681	0	0.399226	40	1
1	984	0	0.242113	24	0
2	8135	0	0.413669	41	1
3	6915	0	0.199569	20	0
4	2712	1	0.237945	24	0

Conclusion

Top Features contributing for high Conversion Rate:

- What is your current occupation--Working Professional.
- What is your current occupation--Businessman.
- What is your current occupation--Student.
- Lead Source--Welingak Website.
- Lead Origin--Lead Add Form.
- Total Time Spent on Website.

Conclusion

Top features to focus to increase Conversion Rate:

- Lead Source--Direct Traffic.
- Lead Source--Google.
- Lead Source--Organic Search.
- Lead Source--Referral Sites.
- Last Activity--Email Bounced.