Lead Scoring Case Study

By

Gopalakrishnan Narayanan

Pranav Sainath

Table of Contents

Problem Statement

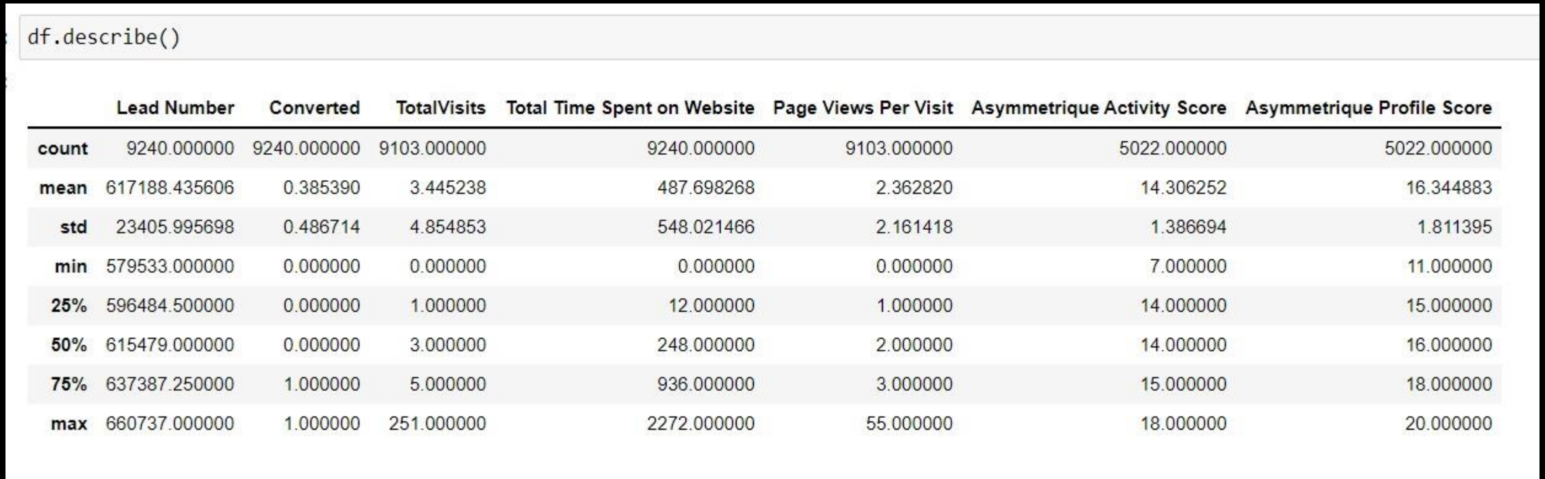
- An education company named X Education sells online courses to industry professionals.
- Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- The company wants to increase the conversion rate to 80% as a target given by the CEO

Objectives

- To build a logistic regression model to assign a lead score to each lead
- Higher score indicates lead is hot i.e., more likely to convert and a lower lead indicates less likely to convert

Approach

- Analyzing Patterns
  - Using Exploratory Data Analysis, we have analyzed the patterns present in the Dataset which will provide us intuition that the which features will help in driving the lead conversion.
- Driving Factors
  - Looking at the below data we get an intuition that how the variables are distributed.



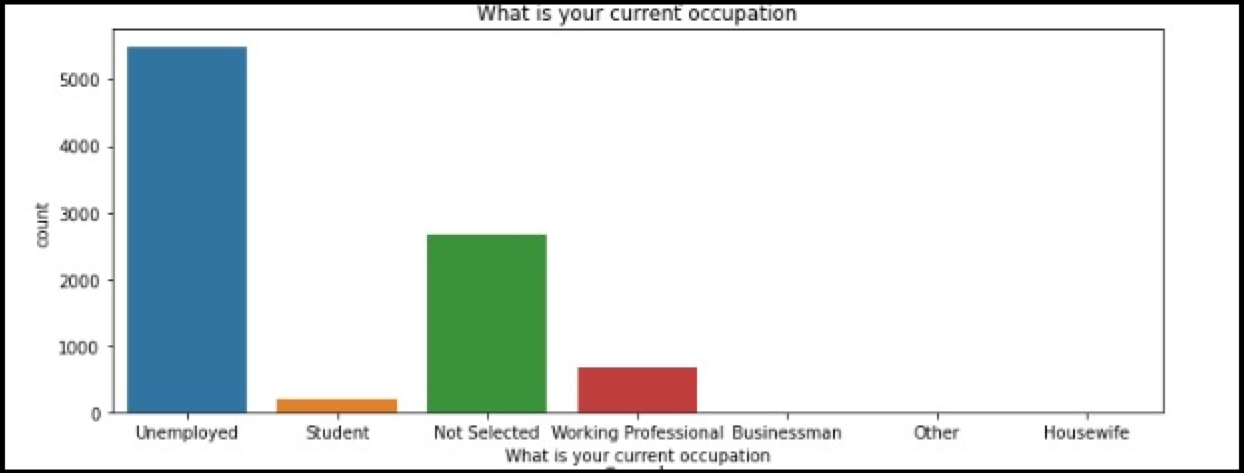
```
df.describe()
```

	Lead Number	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	Asymmetrique Activity Score	Asymmetrique Profile Score
count	9240.000000	9240.000000	9103.000000	9240.000000	9103.000000	5022.000000	5022.000000
mean	617188.435606	0.385390	3.445238	487.698268	2.362820	14.306252	16.344883
std	23405.995698	0.486714	4.854853	548.021466	2.161418	1.386694	1.811395
min	579533.000000	0.000000	0.000000	0.000000	0.000000	7.000000	11.000000
25%	596484.500000	0.000000	1.000000	12.000000	1.000000	14.000000	15.000000
50%	615479.000000	0.000000	3.000000	248.000000	2.000000	14.000000	16.000000
75%	637387.250000	1.000000	5.000000	936.000000	3.000000	15.000000	18.000000
max	660737.000000	1.000000	251.000000	2272.000000	55.000000	18.000000	20.000000

Approach

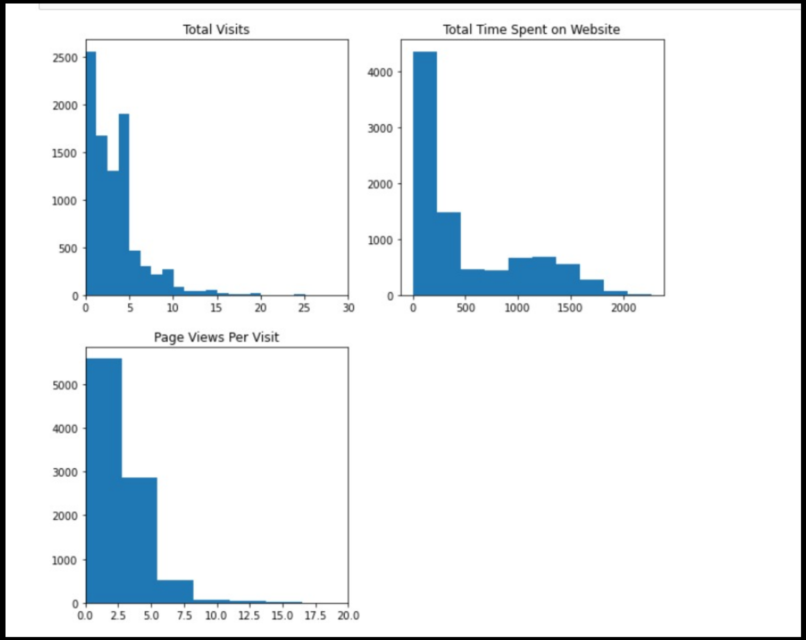
- Correlation
  - Identifying correlations amongst variables to identify the variability in data and identify most important features that can help in driving the conversion of leads.
- Recommendations
  - Focus on features that can expedite the conversion of leads.

Data Insights



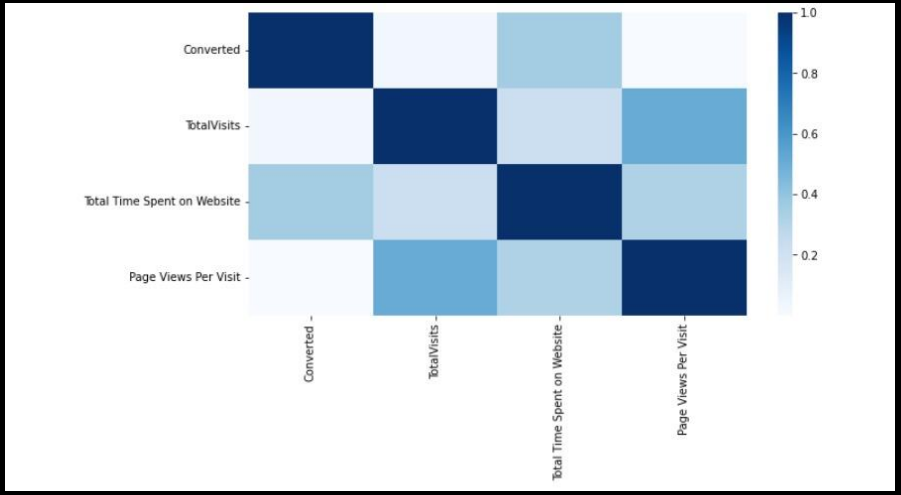
We see that the most responses are from the people who are currently unoccupies (ie looking for a job), followed by students and working professionals.

Data Insights



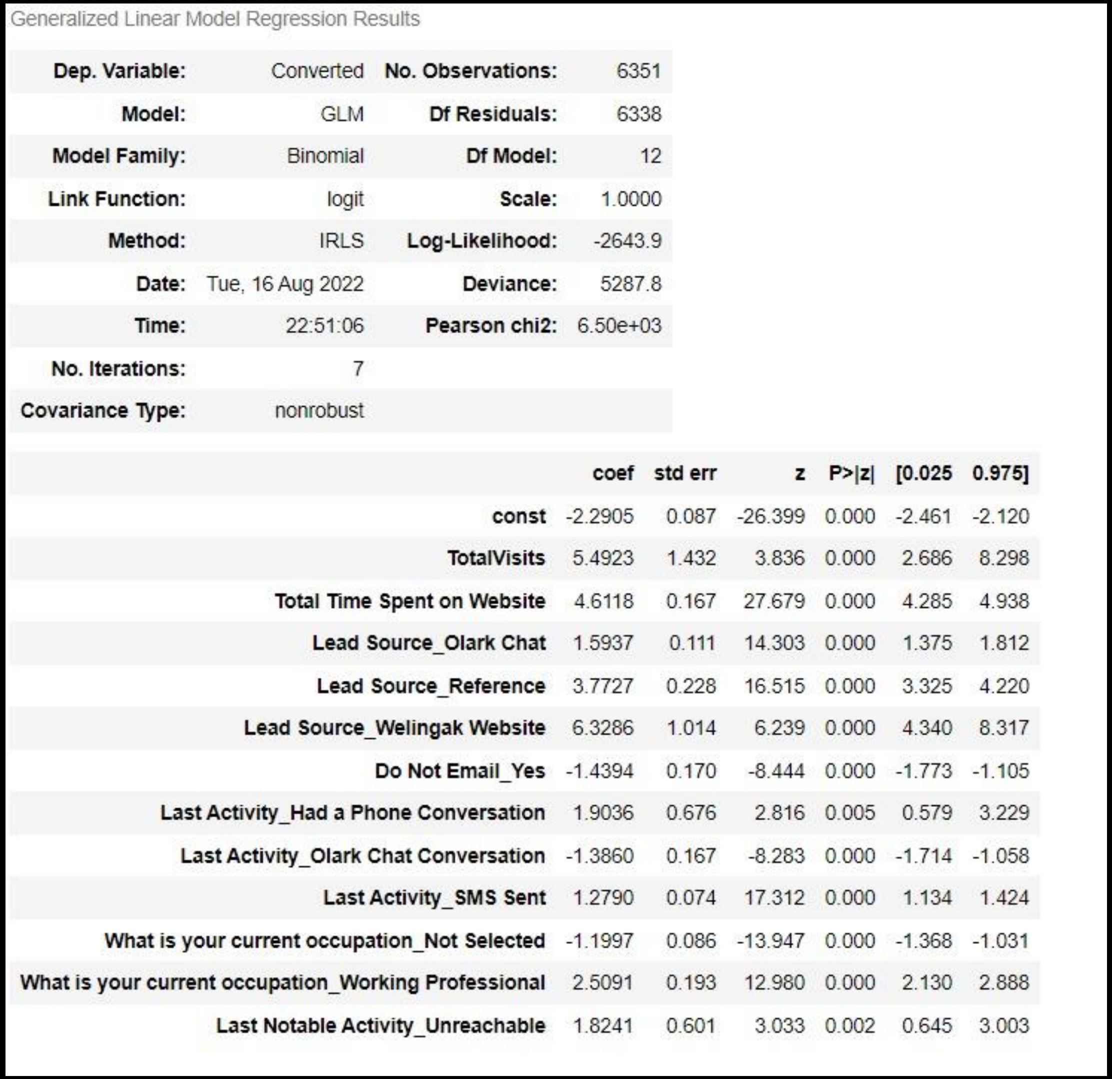
This graphs show us the habit of people who visit the sites based on the number of times they visit, the no of pages
They visit and the amount of time spent on the site

Data Insights



There is correlation between the variables “Total Visits” and “Page view per Visits

Factor Responsible in Driving Leads

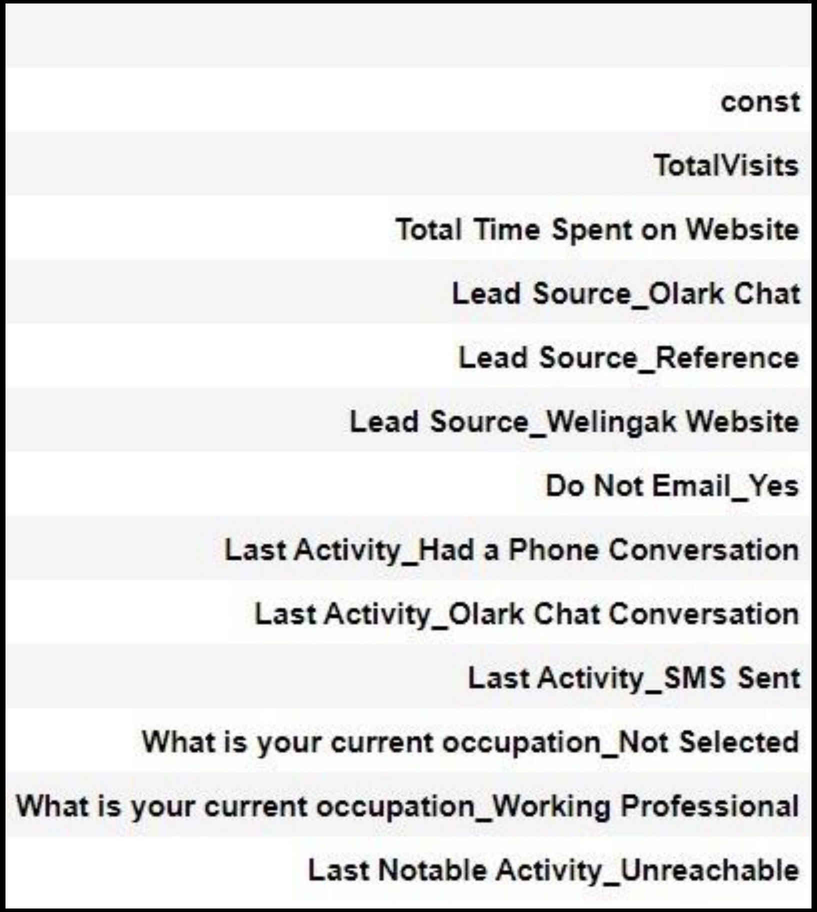


Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6351
Model:	GLM	**Df Residuals:**	6338
Model Family:	Binomial	**Df Model:**	12
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2643.9
Date:	Tue, 16 Aug 2022	**Deviance:**	5287.8
Time:	22:51:06	**Pearson chi2:**	6.50e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	-2.2905	0.087	-26.399	0.000	-2.461	-2.120
TotalVisits	5.4923	1.432	3.836	0.000	2.686	8.298
Total Time Spent on Website	4.6118	0.167	27.679	0.000	4.285	4.938
Lead Source_Olark Chat	1.5937	0.111	14.303	0.000	1.375	1.812
Lead Source_Reference	3.7727	0.228	16.515	0.000	3.325	4.220
Lead Source_Welingak Website	6.3286	1.014	6.239	0.000	4.340	8.317
Do Not Email_Yes	-1.4394	0.170	-8.444	0.000	-1.773	-1.105
Last Activity_Had a Phone Conversation	1.9036	0.676	2.816	0.005	0.579	3.229
Last Activity_Olark Chat Conversation	-1.3860	0.167	-8.283	0.000	-1.714	-1.058
Last Activity_SMS Sent	1.2790	0.074	17.312	0.000	1.134	1.424
What is your current occupation_Not Selected	-1.1997	0.086	-13.947	0.000	-1.368	-1.031
What is your current occupation_Working Professional	2.5091	0.193	12.980	0.000	2.130	2.888
Last Notable Activity_Unreachable	1.8241	0.601	3.033	0.002	0.645	3.003

Important factors which influence the conversion of leads



const
TotalVisits
Total Time Spent on Website
Lead Source_Olark Chat
Lead Source_Reference
Lead Source_Welingak Website
Do Not Email_Yes
Last Activity_Had a Phone Conversation
Last Activity_Olark Chat Conversation
Last Activity_SMS Sent
What is your current occupation_Not Selected
What is your current occupation_Working Professional
Last Notable Activity_Unreachable

Terminologies Required

- Converted Categorical columns to numerical
  - This step is done as our algorithm runs only on numerical data.
- Feature Scaling
  - This is done to bring our data into same scale.
- Data Splitting
  - We have split the data into 70:30 and named it as train data and test data. We run model on train data and validate our model on test data.
- Confusion Matrix
  - True positive (TP): correct positive prediction False positive (FP): incorrect positive prediction True negative (TN): correct negative prediction False negative (FN): incorrect negative predictio

Terminologies Required

- Accuracy = (True Negative + True Positive)/Total
  - This metrics provides the accuracy of the model, where total is TP + FN + FP +FN.
- Sensitivity = True Positive / (True Positive + False Positive)
  - Sensitivity (SN) is calculated as the number of correct positive predictions divided by the total number of positives. It is also called recall (REC) or true positive rate (TPR).
  - The best sensitivity is 1.0, whereas the worst is 0.0.
- Specificity = True Negative/ (True Negative + False Negative)
  - Specificity (SP) is calculated as the number of correct negative predictions divided by the total number of negatives. It is also called true negative rate (TNR). The best specificity is 1.0, whereas the worst is 0.0

Terminologies Required

- Precision = True Positive/ (True Positives +False Positives)
  - Precision is defined as the number of true positives divided by the number of true positives plus the number of false positives.
- Recall = True Positives/(True Positives +False Negatives)
  - The precise definition of recall is the number of true positives divided by the number of true positives plus the number of false negatives.
  - True positives are data point classified as positive by the model that are positive (meaning they are correct), and false negatives are data points the model identifies as negative that are positive (incorrect).

Model Metrics

- Running Model on features selected, we get following metrics :
  - Train Data :
    - Accuracy – 81 percent
    - Sensitivity – 78 percent
    - Specificity – 82 percent
    - Precision - 75.7 percent
    - Recall - 75.2 percent
    - Confusion Matrix :
      - No of Not Converted Leads
        - Not Converted : 2993
        - Converted : 942
      - No of Converted Leads
        - Not Converted : 473
        - Converted : 1955

Model Metrics

- Running Model on features selected, we get following metrics :
    - Test Data :
        - Accuracy – 80.7 percent
        - Sensitivity – 81 percent
        - Specificity – 80.5 percent
        - Precision - 74 percent
        - Recall - 76 percent
        - Confusion Matrix :
            - No of Not Converted Leads
                - Not Converted : 1405
                - Converted : 339
            - No of Converted Leads
                - Not Converted : 186
                - Converted : 793

Conclusion

- Company should focus on following features to increase the leads
  - Lead Source_Welingak Website
  - TotalVisits
  - Total Time Spent on Website
- Company should also focus on Lead Score (which are the probabilities obtained via algorithm) which are greater than 80% to expedite the conversion rate