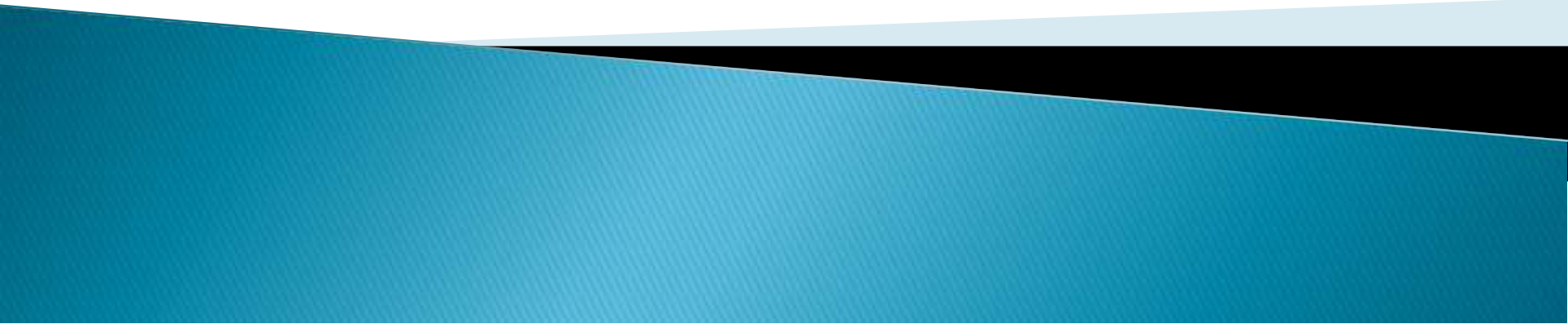
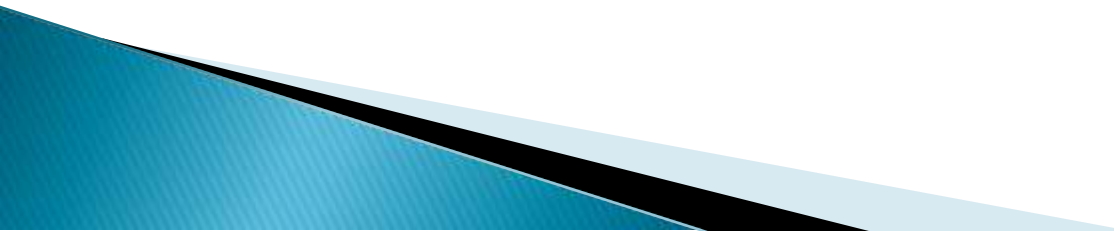


X Education Lead Scoring Case Study

By Abhishek, Sapna, Sanjeev



Introduction

- ▶ X Education which is a online Industry Professional courses seller markets courses on several websites like Google. When a person lands on the website to browse courses they fill up a form providing email and phone numbers, These are classified as leads. The Company also gets leads from referrals as well. Once the leads are gathered the sales team steps in to contact the lead and tries to convert them.
- 

Problem Statement

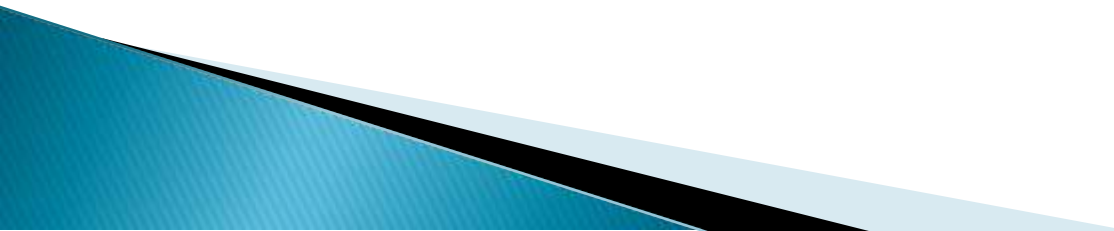
- ▶ The Sales team has a typical conversion rate of 30%. X Education would like to know the possible leads that are sure to convert and most of the efforts are provided to those leads with the high lead score.

Solution and Approach

- ▶ Solution is to build a logistic regression model that can be used for analysing the leads and provide a lead score for the team to concentrate on the High quality leads and increase the conversion rate.



Data Loading and Data Cleaning

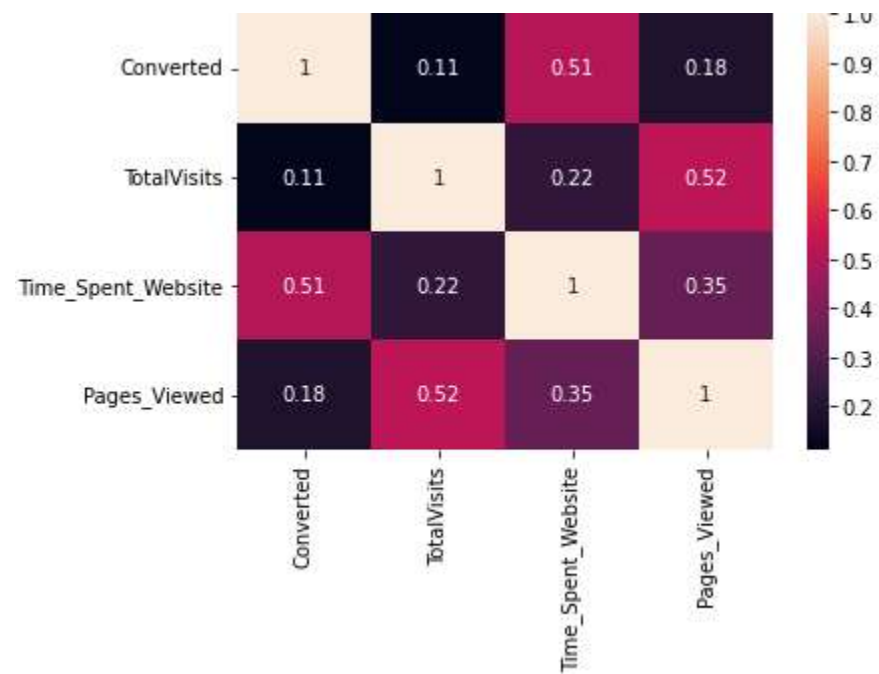
- ▶ The given data is loaded into the jupyter notebook and all required libraries are loaded for the processing
 - ▶ Before the actual model building, first the data was analyzed and found that some of the columns had more than 30% of missing values. Therefore the columns with high missing values are removed from the data frame. This was done using the null check and missing value technique.
- 

Contd.

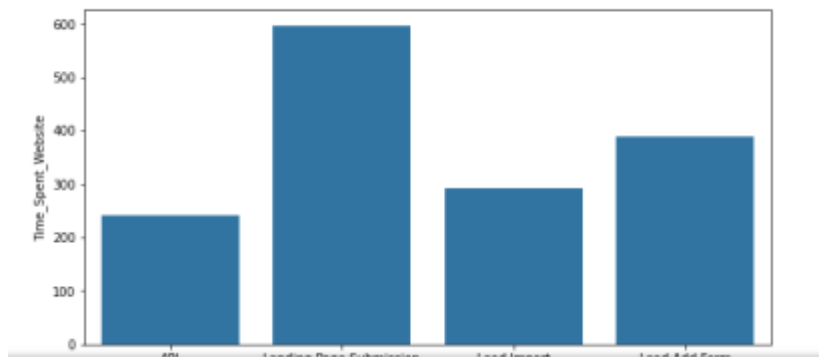
- ▶ Few data were missing in rows wise as well therefore Data processing techniques like replacing NaN values with Median and Mode were used to fill the gaps.
- ▶ Columns name were not proper to be used for processing, so columns were also rename

EDA – Exploratory Data Analysis

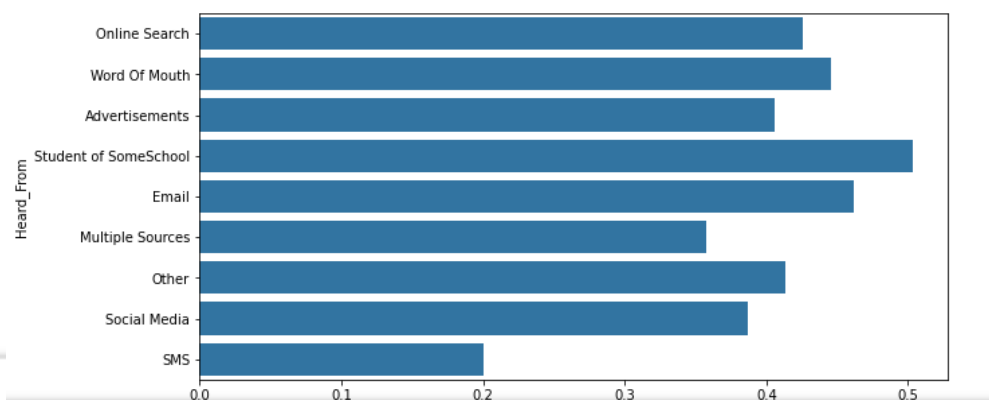
- Once the data was cleaning and processed. EDA was performed to find the correlations between the columns and the strength of the correlation.



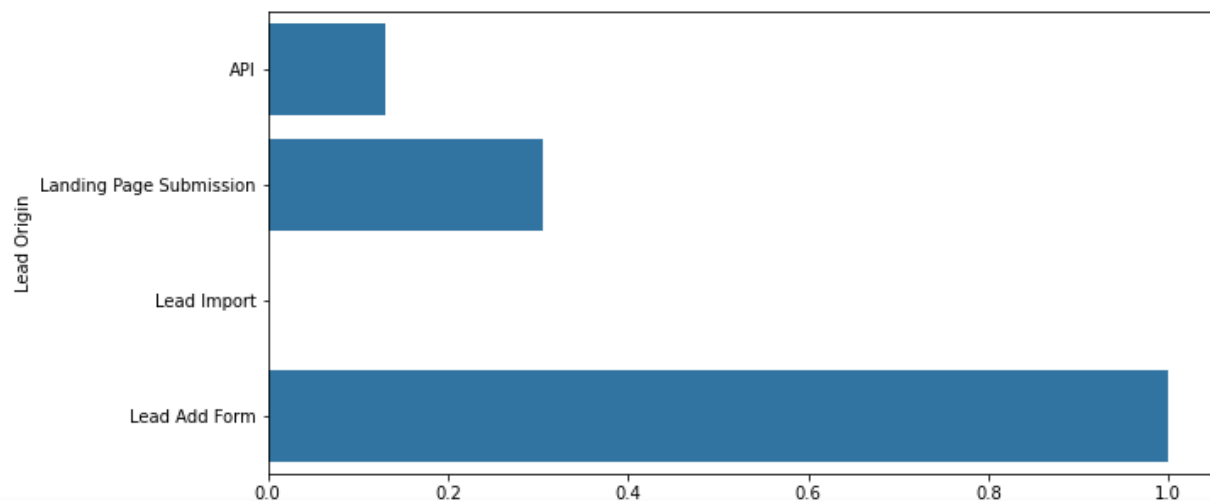
----- Time_Spent_Website Vs Lead Origin -----



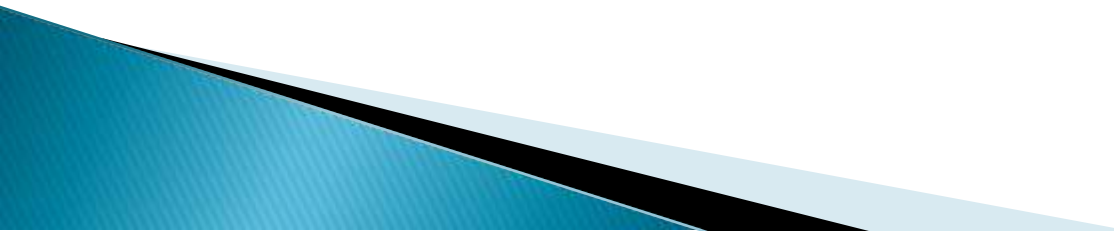
----- Heard_From Vs Converted -----



----- Lead Origin Vs Converted -----



Model Building

- ▶ After EDA was performed, the categorical columns were identified and was used for creating dummies as part of the model building. After that the selected columns were dropped for avoiding redundancy.
 - ▶ The created dummies were appended to the dataframe.
- 

Scaling and Feature Selection

- ▶ The Data was later scaled using the MinMaxScaler
- ▶ The Feature selection was done using the RFE library
- ▶ Later the Train and Test Split was performed on the scaled data



- ▶ The Model was built using statsmodel library and the results were analyzed.
- ▶ After every model building iteration the feature with $P > |z| > 0.05$ value was dropped. After all features with required value was obtained. The model was trained on the test data and predicted.

Generalized Linear Model Regression Results

Dep. Variable:	y	No. Observations:	2595
Model:	GLM	Df Residuals:	2581
Model Family:	Binomial	Df Model:	3
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-993.74
Date:	Tue, 21 May 2024	Deviance:	1987.5
Time:	11:07:58	Pearson chi2:	2.58e+03
No. Iterations:	5	Pseudo R-squ. (CS):	0.2576
Covariance Type:	nonrobust		

	coef	std err	z	P> z	[0.025	0.975]
const	-1.8962	0.134	-14.114	0.000	-2.160	-1.633
Time_Spent_Website	4.8449	0.235	20.639	0.000	4.385	5.305
Heard_From_Online Search	-0.8929	0.130	-6.879	0.000	-1.147	-0.638
Current_Occupation_Working Professional	3.0840	0.428	7.232	0.000	2.255	3.932

	Feature	VIF
0	const	7.415006
1	Time_Spent_Website	1.071733
2	Heard_From_Online Search	1.072824
3	Current_Occupation_Working Professional	1.034980

Thank you

