Lead Scoring Case Study

Rajat Soni and Vibha Kashyap

Problem Statement

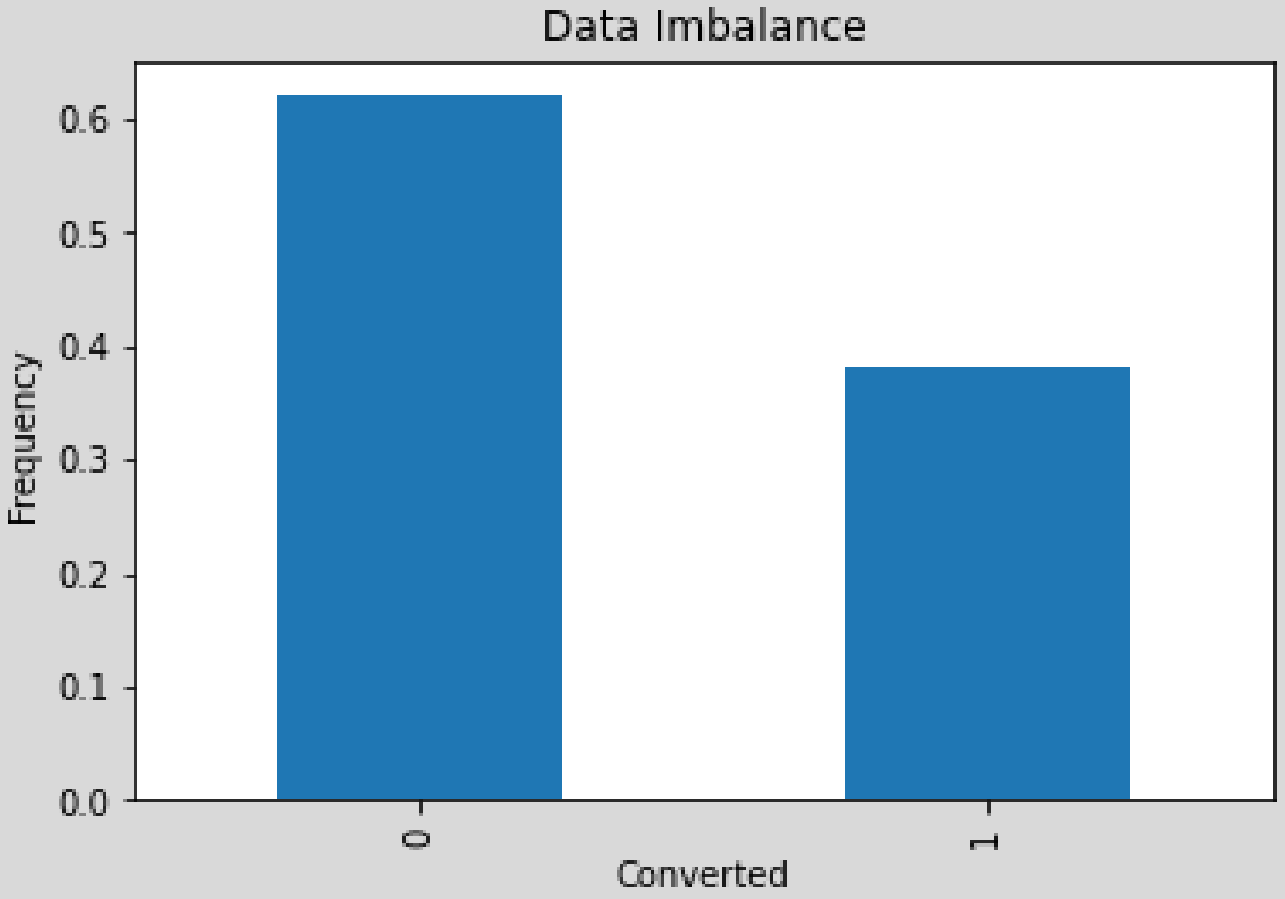
- An education company named X Education sells online courses to industry professionals.
- The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. 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%.
- 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. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Business Goals

The company requires you to build a model wherein you 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. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

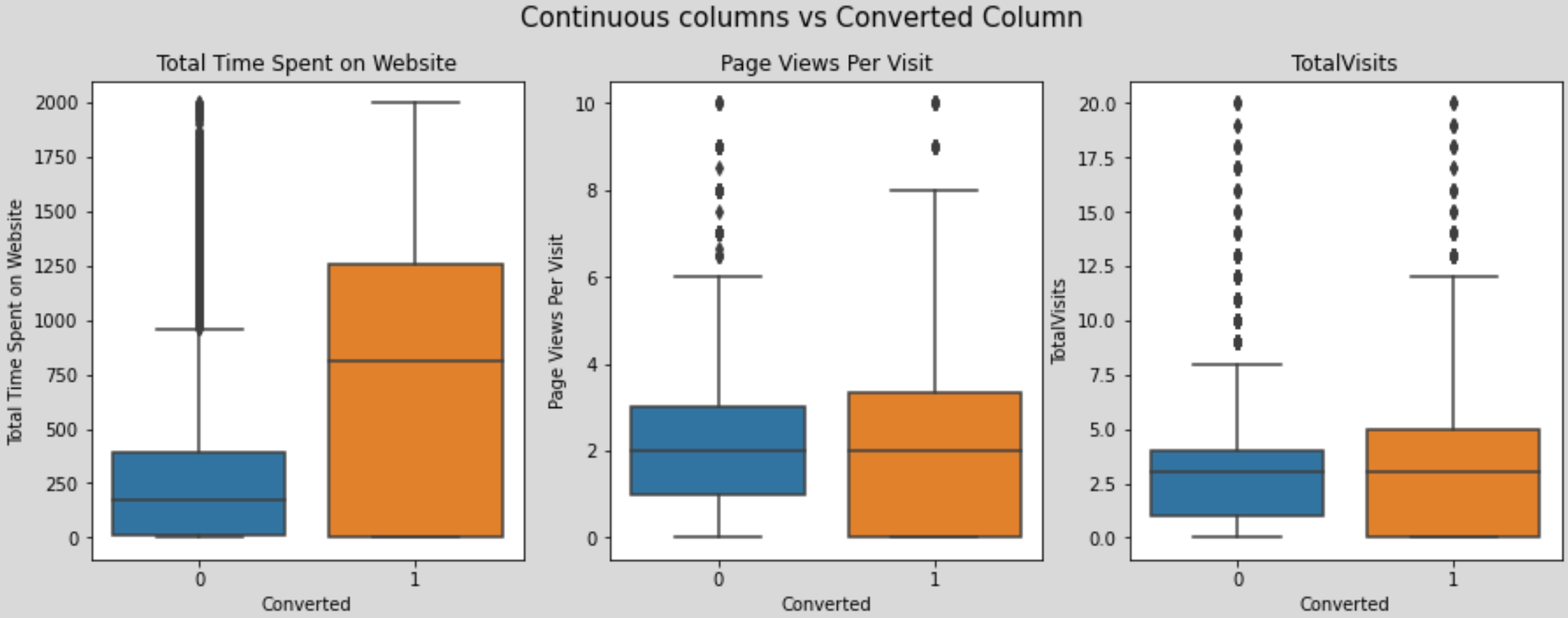
Steps Involved

- Imported required libraries and Understand the data
- Data cleaning.
- EDA
- Data Pre-Processing (Creating Dummy Variables. Train-Test Split, Scaling, Looking at Correlation)
- Feature Selection using RFE
- Model Building and Checking P-Value and VIE
- Plotting ROC Curve and finding optimal probability cutoff
- Model Evaluation on Test Set
- Finding the lead scores (100*probability of conversion)
- Performance Comparison for train and test scores
- Summary



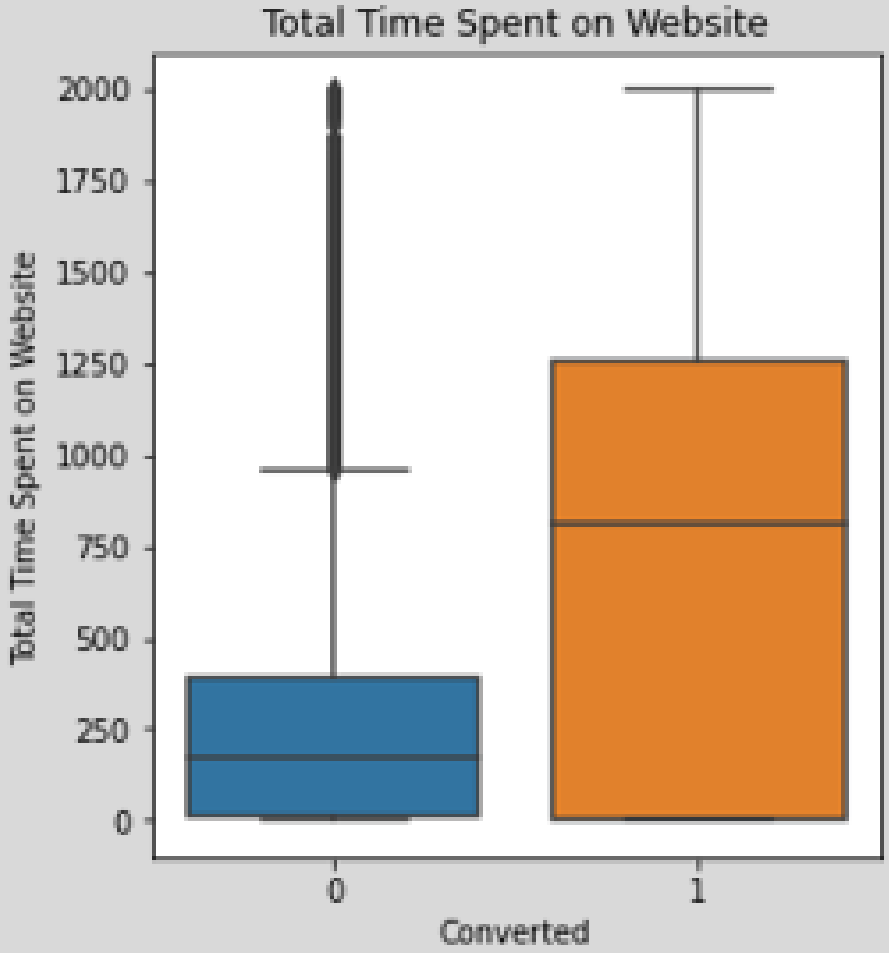
There is a data imbalance of 62-63 percent.

This implies that the conversion rate is 37-38%



The above plots show the conversion rates related to how much time a lead spends on the website, how many pages are viewed per visit and the number of visits.

- This shows that the leads that spend more time on the website are more likely to convert to a student.



FINAL MODEL SUMMARY

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6294
Model:	GLM	Df Residuals:	6279
Model Family:	Binomial	Df Model:	14
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2656.2
Date:	Mon, 13 Jun 2022	Deviance:	5312.4
Time:	19:21:10	Pearson chi2:	6.68e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.1490	0.523	-2.195	0.028	-2.175	-0.123
Total Time Spent on Website	4.1072	0.147	27.882	0.000	3.818	4.396
Lead Origin_Lead Add Form	4.3566	0.227	19.225	0.000	3.912	4.801
Lead Source_Facebook	1.3645	0.480	2.843	0.004	0.424	2.305
Lead Source_Olark Chat	1.2487	0.103	12.124	0.000	1.047	1.451
Last Activity_Converted to Lead	-1.1995	0.208	-5.766	0.000	-1.607	-0.792
Last Activity_Email Bounced	-1.9288	0.313	-6.161	0.000	-2.542	-1.315
Last Activity_Had a Phone Conversation	1.3305	0.839	1.586	0.113	-0.314	2.975
Last Activity_Olark Chat Conversation	-1.5930	0.168	-9.482	0.000	-1.922	-1.264
What is your current occupation_Student	-1.0775	0.573	-1.881	0.060	-2.200	0.045
What is your current occupation_Unemployed	-1.2514	0.522	-2.398	0.016	-2.274	-0.229
What is your current occupation_Working Professional	1.4920	0.553	2.698	0.007	0.408	2.576
Last Notable Activity_Had a Phone Conversation	2.2386	1.390	1.610	0.107	-0.486	4.963
Last Notable Activity_SMS Sent	1.5304	0.080	19.090	0.000	1.373	1.687
Last Notable Activity_Unreachable	2.1787	0.566	3.847	0.000	1.069	3.289

TRAINING AND TEST DATA METRICS

	Train Data	Test Data
Accuracy	**81.5%**	**81.2%**
Specificity	89.0%	88.4%
Sensitivity	68.9%	69.8%
Precision	79.3%	79.0%
Recall	68.9%	69.8%

CONCLUSION

- As visible from Model summary, the two variables which influence conversion rates the most are:
    - Total Time Spent on Website
    - Lead Origin_Lead Add Form – The leads that fill in the form available on the website.
    - Last Notable Activity_Had a Phone Conversation – Whether the leads were contacted by the sales team.