Lead Scoring Case Study Using Logistic Regression

Submitted By:

1. Dwarika Dash
2. Diksha Katagi
3. Devishree CM

Problem Statement

- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- 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%.

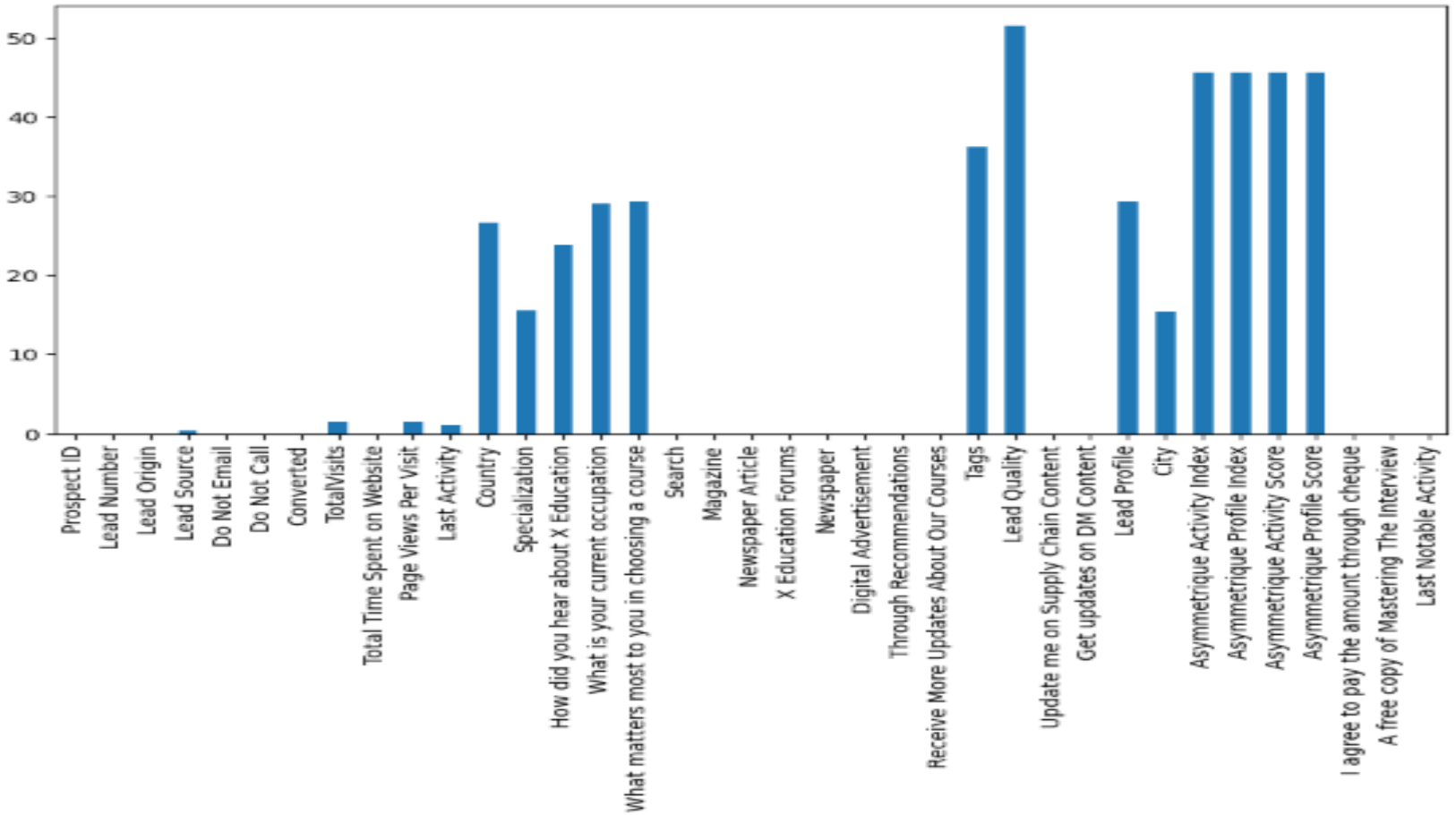
Goals of the Case Study

- 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.

Problem Solving Approach

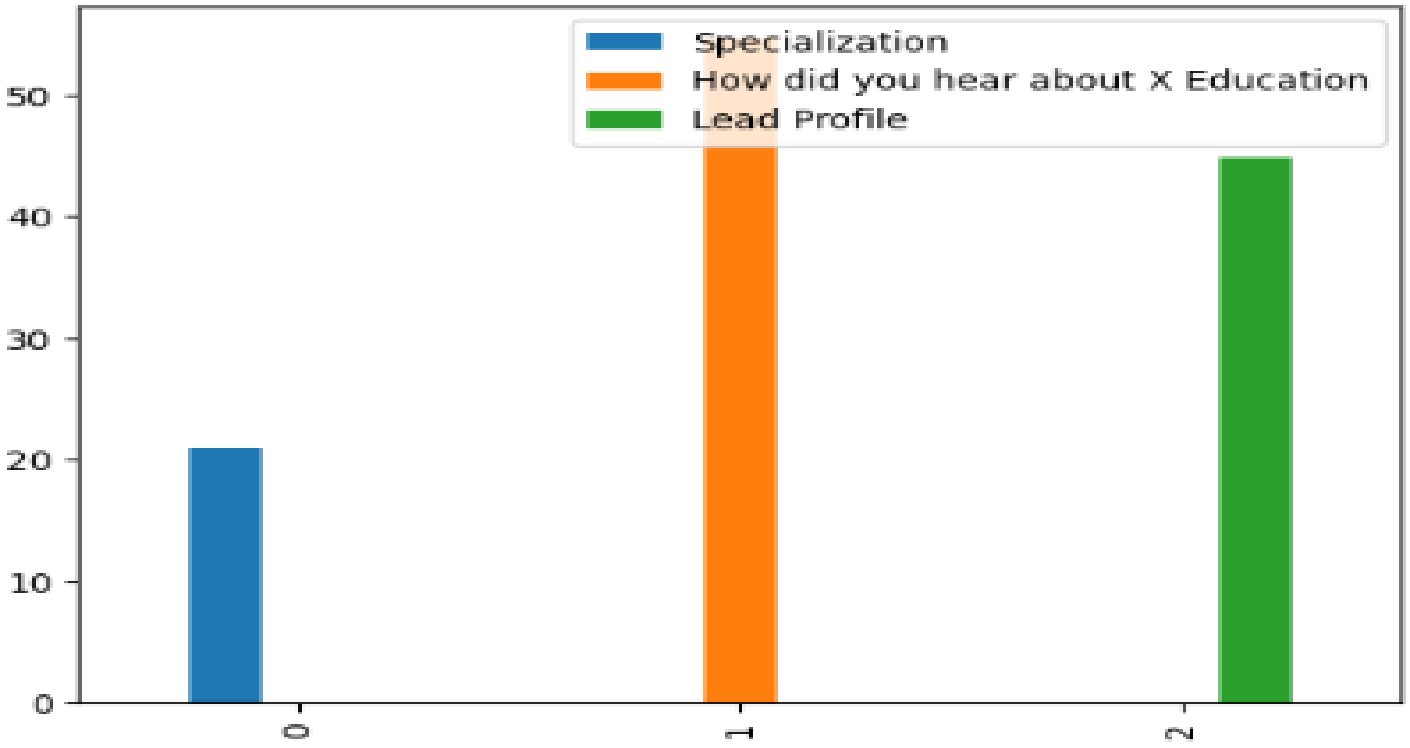
- Data Sourcing
- Data Preparation
- Data Cleaning
- EDA
- Dummy Variable Creation
- Test-Train split
- Feature scaling
- Correlations
- Model Building
- Model Evaluation
- Prediction

Null Check



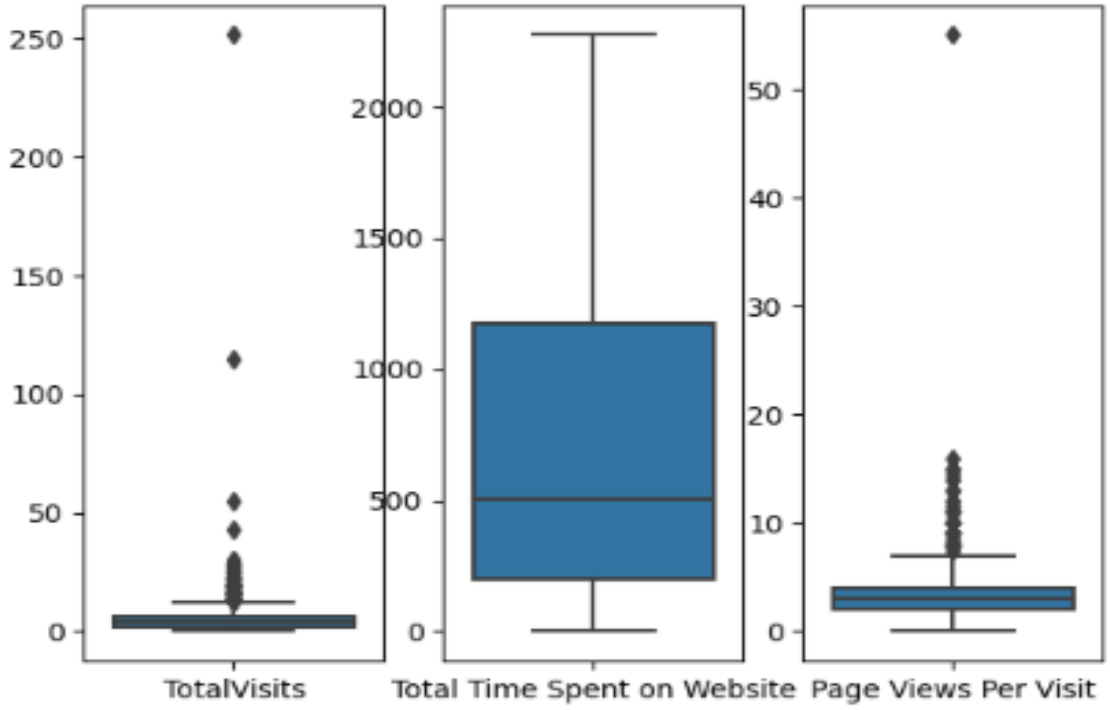
Dropped columns having more than 30% null values

Level ‘Select’ Check



Dropped columns having more than 30% ‘Select’ level
Dropped records with ‘Select’ level in other columns

Outliers Analysis



Dropped records having more than 20 Total Visits

- Dropped columns which contains maximum unique values
- Dropped records having null values
- Mapped ‘Yes’, ‘No’ to binary numbers
- Created dummy variables for all categorical variables
- Dropped highly correlated columns
- Train –Test split
- Feature selection using RFE
- Dropped columns with high P value and high VIF

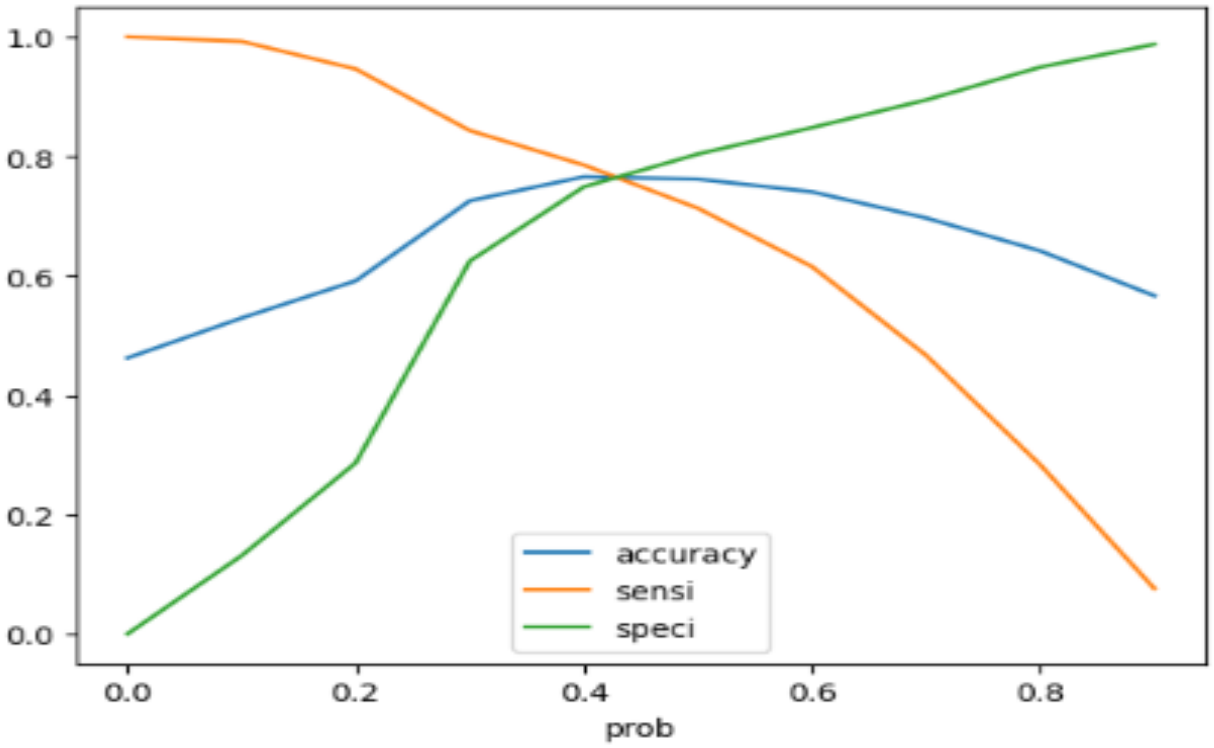
Final Model

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	2811
Model:	GLM	**Df Residuals:**	2803
Model Family:	Binomial	**Df Model:**	7
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1486.5
Date:	Sat, 13 Apr 2024	**Deviance:**	2973.1
Time:	17:24:14	**Pearson chi2:**	2.90e+03
No. Iterations:	5	**Pseudo R-squ. (CS):**	0.2760
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.6414	0.178	-3.601	0.000	-0.991	-0.292
Do Not Email	-1.5628	0.212	-7.355	0.000	-1.979	-1.146
TotalVisits	1.5093	0.383	3.939	0.000	0.758	2.260
Total Time Spent on Website	4.5977	0.202	22.705	0.000	4.201	4.995
Lead Origin_Landing Page Submission	-1.1249	0.161	-6.993	0.000	-1.440	-0.810
Lead Source_Olark Chat	-1.0164	0.374	-2.719	0.007	-1.749	-0.284
Lead Source_Referral Sites	1.3419	0.635	2.112	0.035	0.097	2.587
Last Activity_Converted to Lead	-1.1034	0.234	-4.724	0.000	-1.561	-0.646

Determining Final Cut-off



Final cut-off determined at 0.43

Evaluation Metrics

Train Data

- Accuracy – 76%
- Sensitivity – 76%
- Specificity – 76%

Test Data

- Accuracy – 76%
- Sensitivity – 75%
- Specificity – 76%

Conclusion

Focus on following attributes:

- Do Not Email
- TotalVisits
- Total Time Spent on Website
- Lead Origin_Landing Page Submission
- Lead Source_Olark Chat
- Lead Source_Referral Sites
- Last Activity_Converted to Lead