Lead Scoring – Log regression assignment

Akash Tandon

Vinay pant

Pritish Kumbhare

Problem Statement

- An education company named X Education sells online courses to industry professionals.
- Though company gets plenty of leads, Its lead conversion is poor
- company wishes to identify the most potential leads, also known as ‘Hot Leads’ and drive sales teams focus on interacting with these “potential” students

Objective is to 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.

As per CEO, target lead conversion rate should be around 80%.

Approach

- We had a dataset provided with ~9k records

Data Cleansing

a. Handling Nulls

- We started off by cleansing by deleting any columns with high (40%) nulls and proceeded with data imputation for remaining columns with >1% to 40% null values
- Post that any rows with null values were dropped

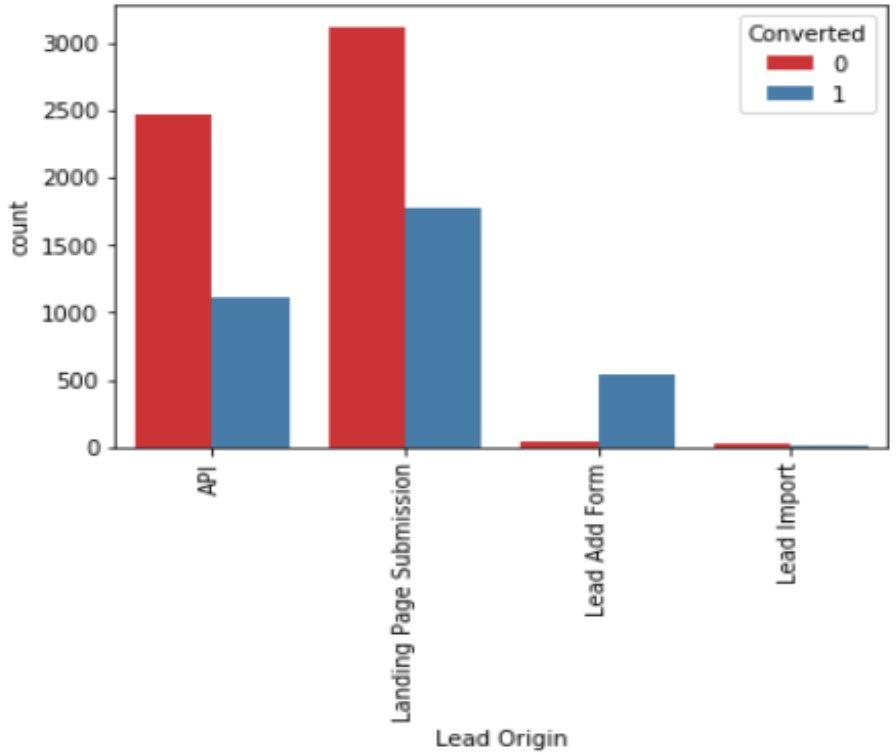
b. Cat Variables

- All cat variables were scanned to see their distribution as per lead conversion status
- Dropped any columns which were heavily skewed as Conversion as 0 or 1
- Rest were cleansed to group the columns with low frequencies
- Any comment columns which are irrelevant for model building were also dropped
- Created the dummy variables for all the relevant cat variables
- Dropped the original columns for which dummy var are created

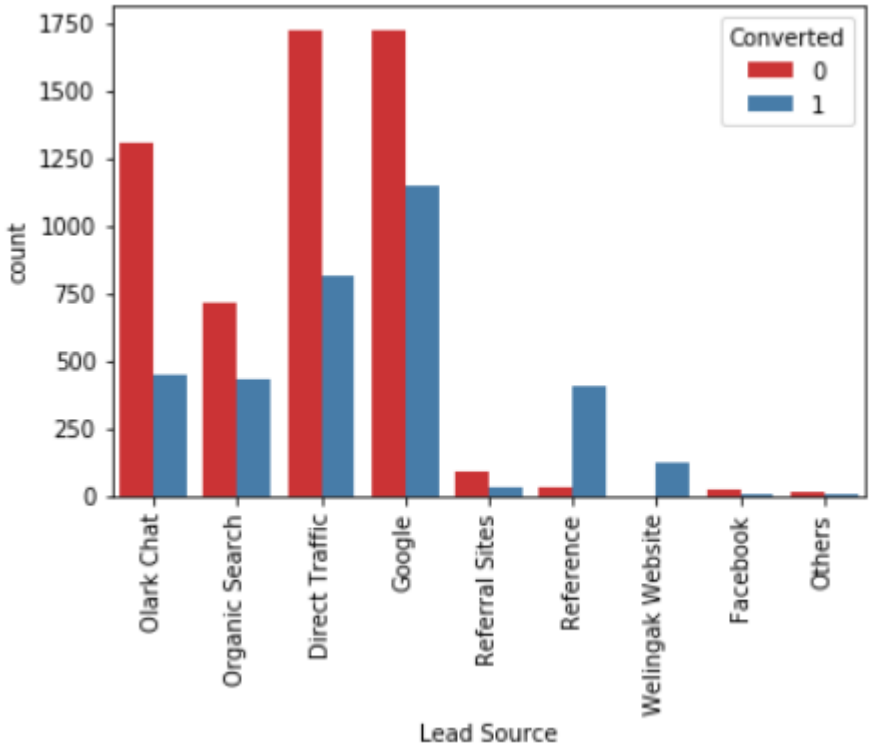
b. Num Variables

- Scanned columns for outliers and limited the outliers to 95%

Learnings

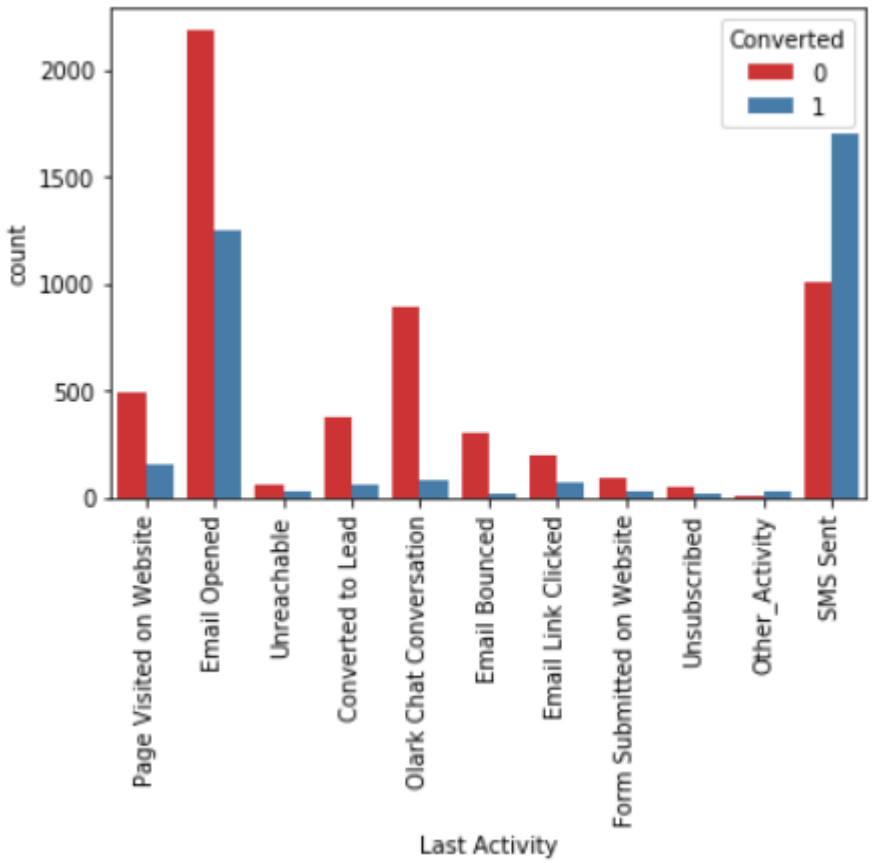


API and landing page submission seems to negatively affect conversion

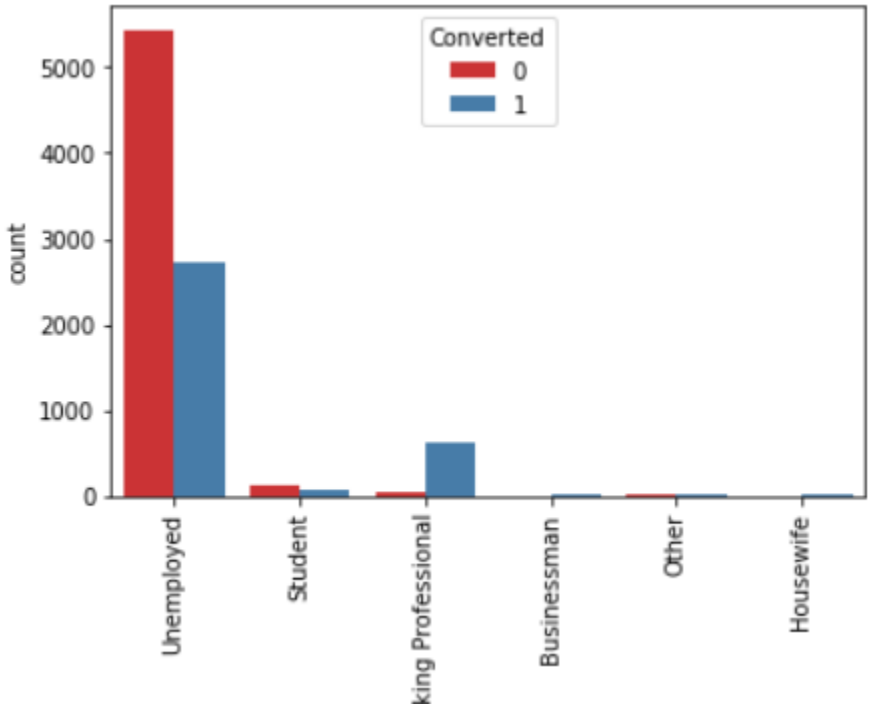


Direct traffic, google and olark chat are the major lead source

Learnings

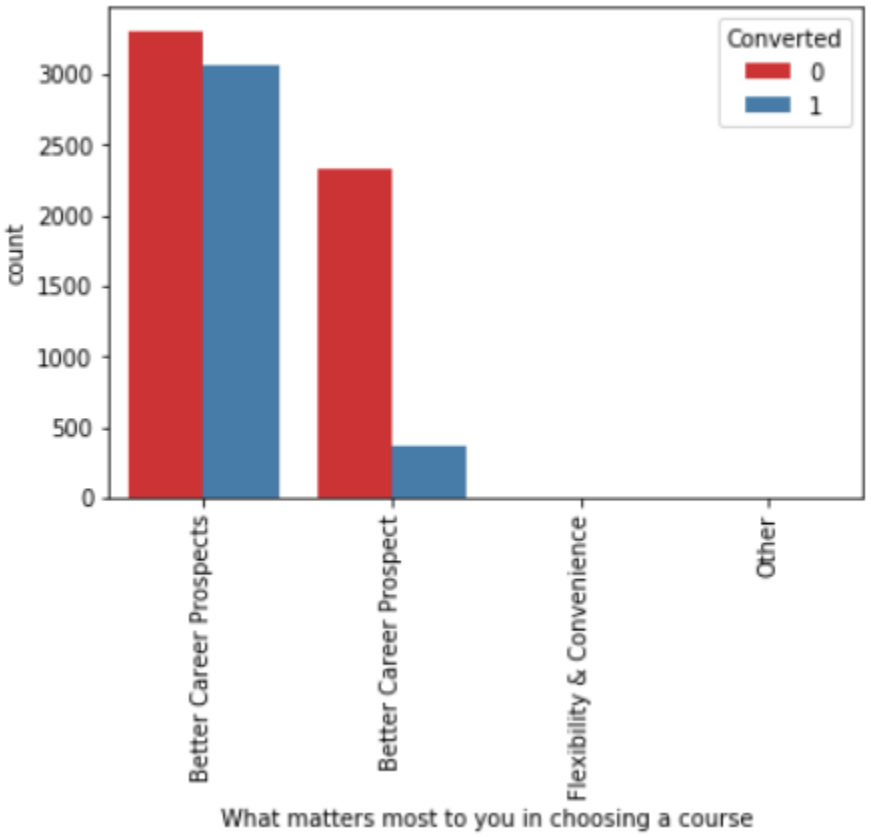


Major last activity is sms sent, email opened



Most prospects are unemployed however working professionals' conversion is high

Learnings



Most are looking for better career prospects

Model Development

- Split the dataset into 70%:30% split for Train:Test data
- Did the feature selection with RFE an created the model. Below model was selected based on the regression results for further model evaluation

Model Evaluation

- Model was evaluated by looking at confusion matrix of training data
- Model was checked for Sensitivity, Specificity, False positive rate, Positive Predictive Value, Negative Predicted values
- Optimal cut off was taken as 0.35 as per balanced sensitivity and specificity

Final Model

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:**	-2598.8
Date:	Mon, 15 Jan 2024	**Deviance:**	5197.6
Time:	20:24:57	**Pearson chi2:**	6.30e+03
No. Iterations:	7	**Covariance Type:**	nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.4580	0.147	-9.888	0.000	-1.747	-1.169
Total Time Spent on Website	1.1104	0.040	27.508	0.000	1.031	1.190
Lead Origin_Landing Page Submission	-1.0274	0.127	-8.062	0.000	-1.277	-0.778
Lead Origin_Lead Add Form	3.0183	0.232	13.034	0.000	2.564	3.472
Lead Source_Olark Chat	1.2865	0.124	10.404	0.000	1.044	1.529
Lead Source_Welingak Website	2.4603	0.759	3.240	0.001	0.972	3.948
Last Activity_Email Bounced	-2.0558	0.381	-5.401	0.000	-2.802	-1.310
Last Activity_Olark Chat Conversation	-1.3306	0.168	-7.927	0.000	-1.660	-1.002
Last Activity_Other_Activity	1.7982	0.463	3.884	0.000	0.891	2.706
Last Activity_SMS Sent	1.2469	0.074	16.741	0.000	1.101	1.393
Specialization_Others	-0.9302	0.125	-7.423	0.000	-1.176	-0.685
What is your current occupation_Working Professional	2.3792	0.190	12.514	0.000	2.007	2.752
What matters most to you in choosing a course_Better Career Prospects	1.2266	0.088	13.908	0.000	1.054	1.400