LEAD SCORE CASE STUDY

LOGISTIC REGRESSION

Varun Malhotra | Anubhav Agarwal | Lalita Rathod

Problem Statement

X education is an organization which sells online courses to industry professionals. Many professionals who are interested in the courses land on their website and browse for courses. The company marks its courses on several popular websites like Google.

X education wants to select most promising leads that can be converted to paying customers.

The company generates a lot of leads through various ways, but the lead conversion rate is very poor. To make this more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.

The whole process has led to 30% conversion rate for the company by the process of turning leads into customers by approaching those leads which are to be found having interest in taking the course. However, the implementation process of lead generating attributes are not efficient in helping conversions

Business Goals

A logistic model is to be built for the company by assigning a lead score between 0 to 100 to each of the leads which can be used by the company to target the potential leads.

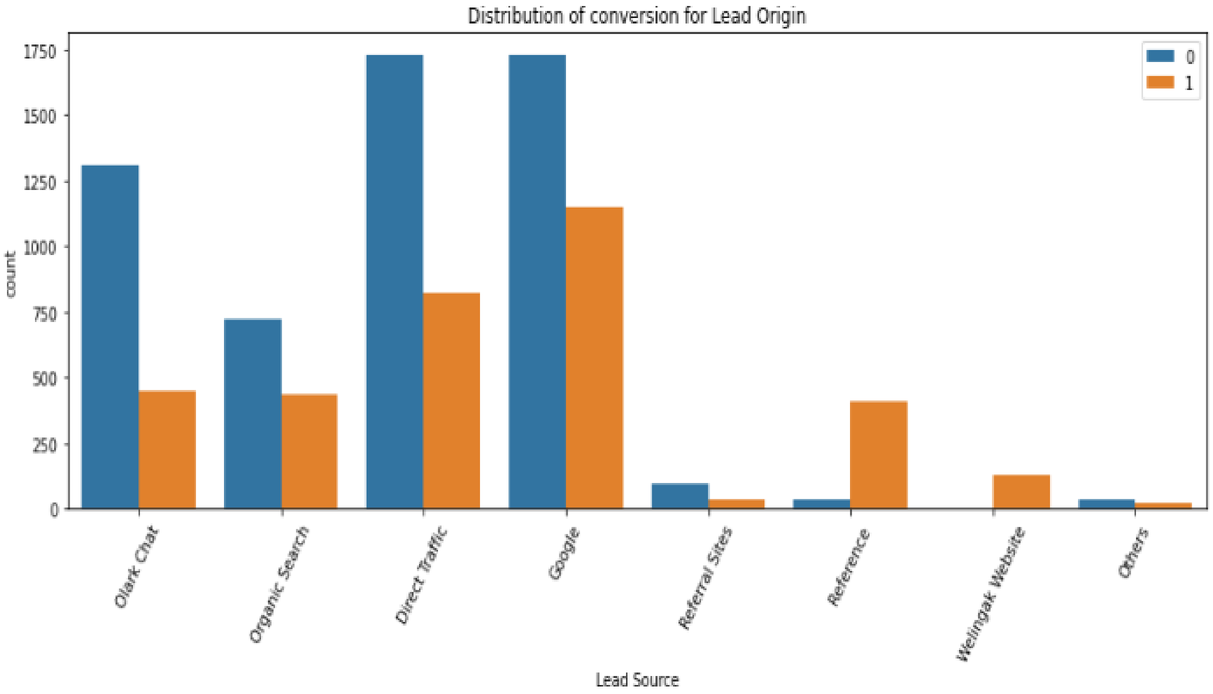
A higher score would mean that the lead is hot i.e., is most likely to convert.

A lower score would mean that the lead is cold i.e., will mostly not get converted.

These lead scores given indicates how promising the lead could be.

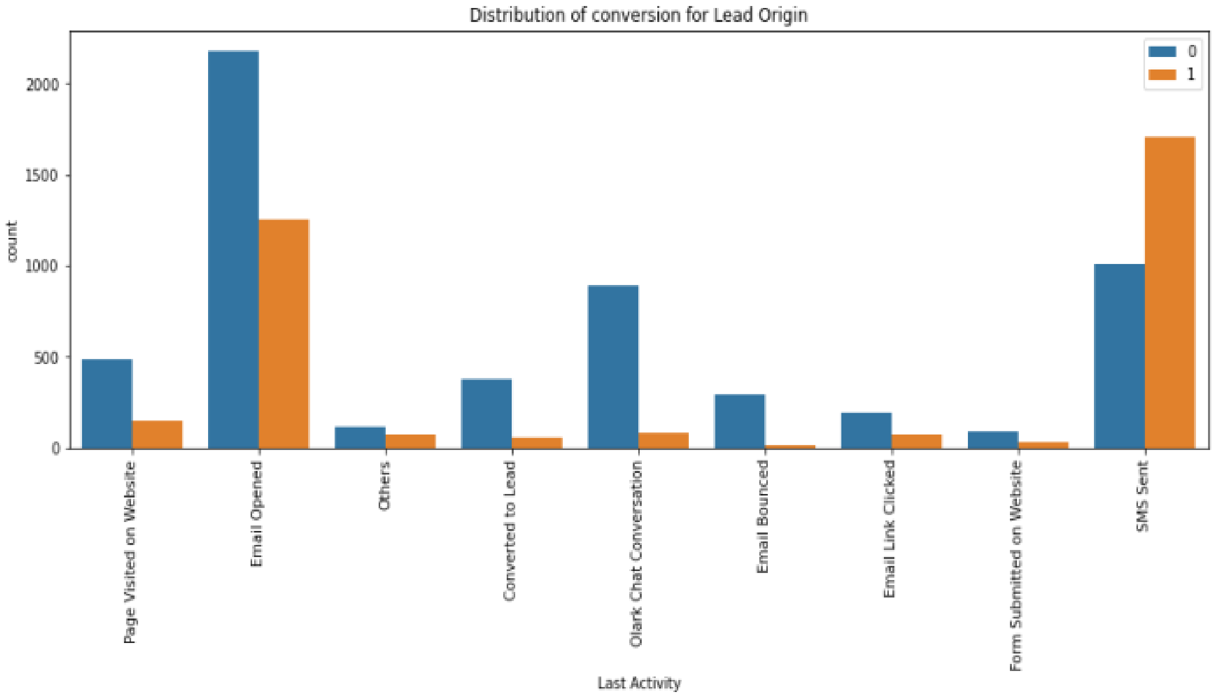
The CEO, has given a ballpark of the target lead conversion rate to be around 80%

Exploratory Data Analysis



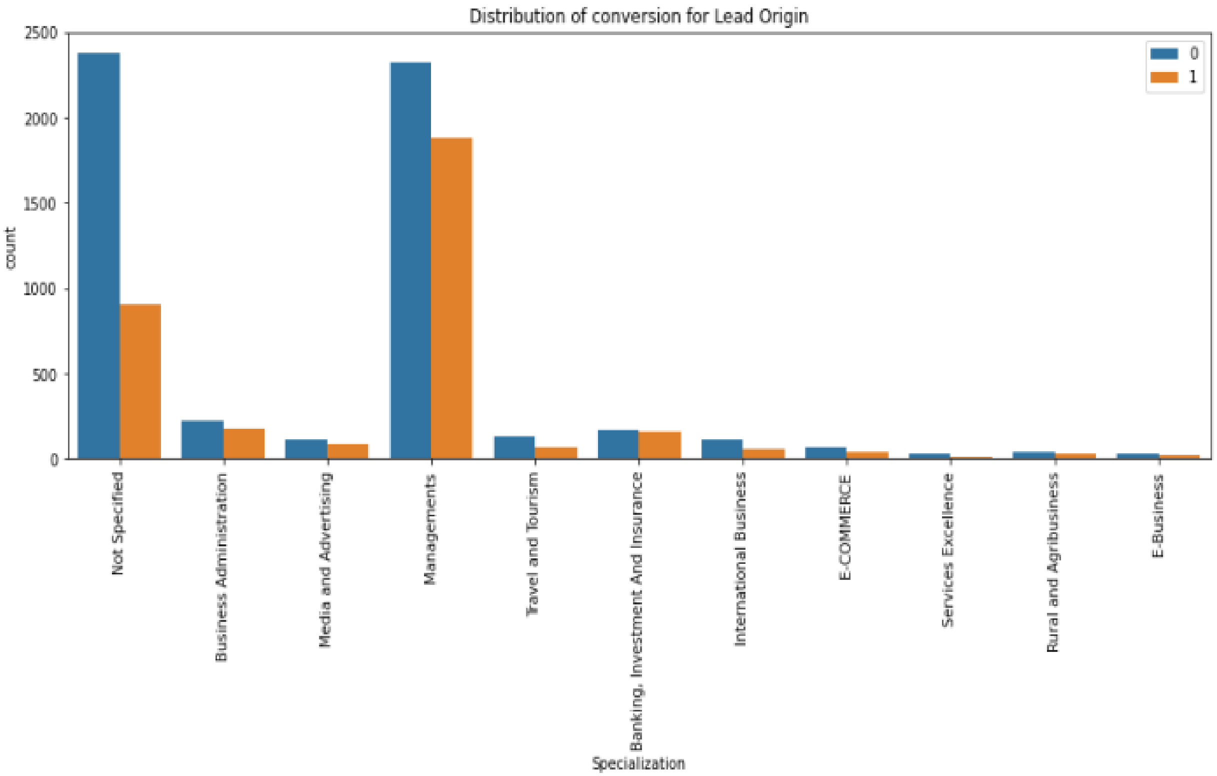
Most of the leads are sourced from google showing that this has had high conversion rate when compared to other modes.

References has had high conversion rate when compared percentage wise.



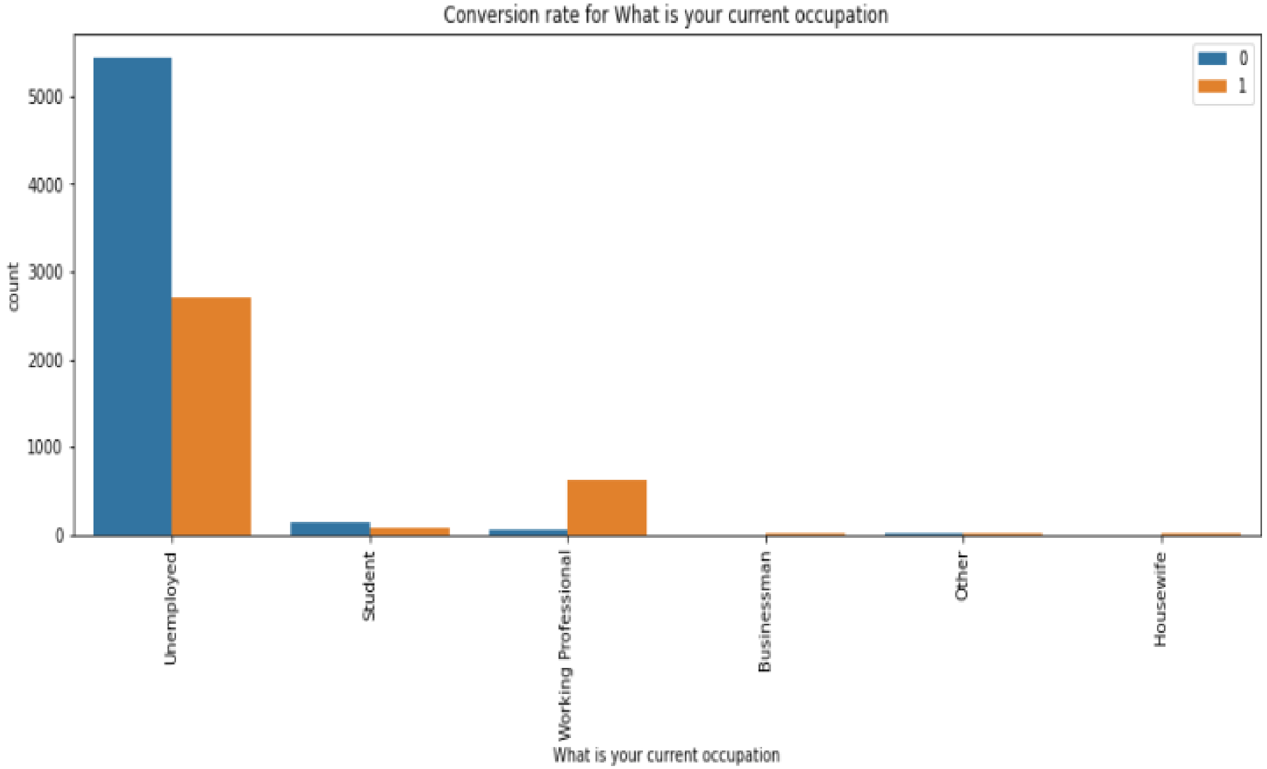
“SMS” has shown to be a promising mode for getting higher confirmed leads, followed by “Email Opened” mode.

Category “SMS Sent” has the highest conversion rate.



Maximum of the leads have no information on specialization.

On the other hand, managements has high conversion rates, implying people from these specializations can be promising leads.



We see that “Unemployed” is the most occurring occupation.

Conversion rate for “Working Professional” is highest compared to other occupations.

Model Building

First, we will split the data randomly into train and test set where train_size = 0.7 and test _size = 0.3.

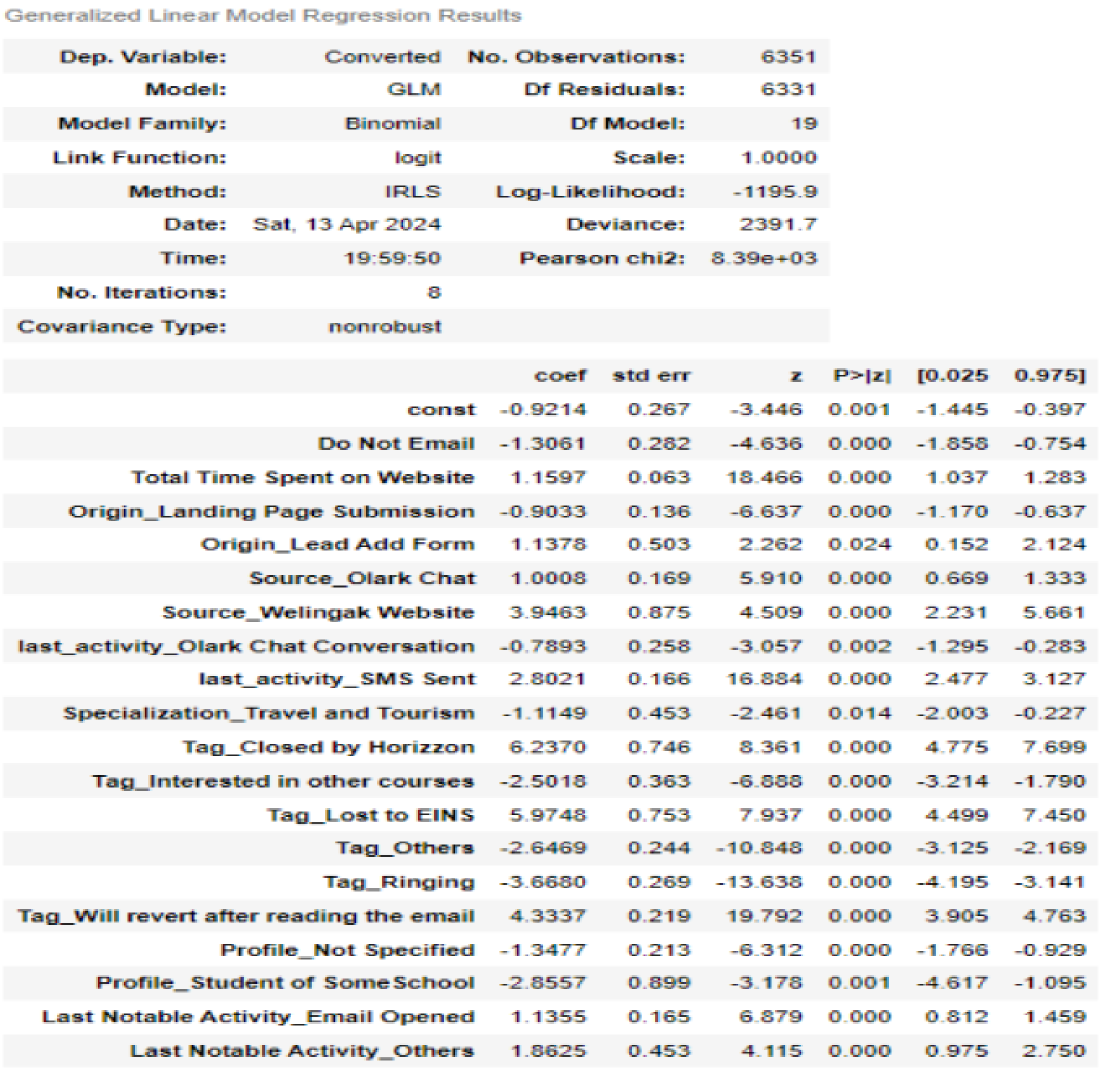
Then we scale the variables in the train set and build the first model.

We then used RFE to eliminate the less relevant variables and then came to the final model by eliminating high p-values.

Here in the table, we can see that "Tag_Closed by Horizzon" has the highest coefficient in the model summary implying that it contributes most towards the probability of converting a lead.

When checked the VIF scores, all are below the set limit i.e., VIF < 5.

If we focus on "last_activity_SMS Sent", Source_Welingak Website" and Tag_Lost to EINS" as they have high coefficient of 2.8, 3.94, and 5.9748 respectively we can increase the probability of lead conversion.

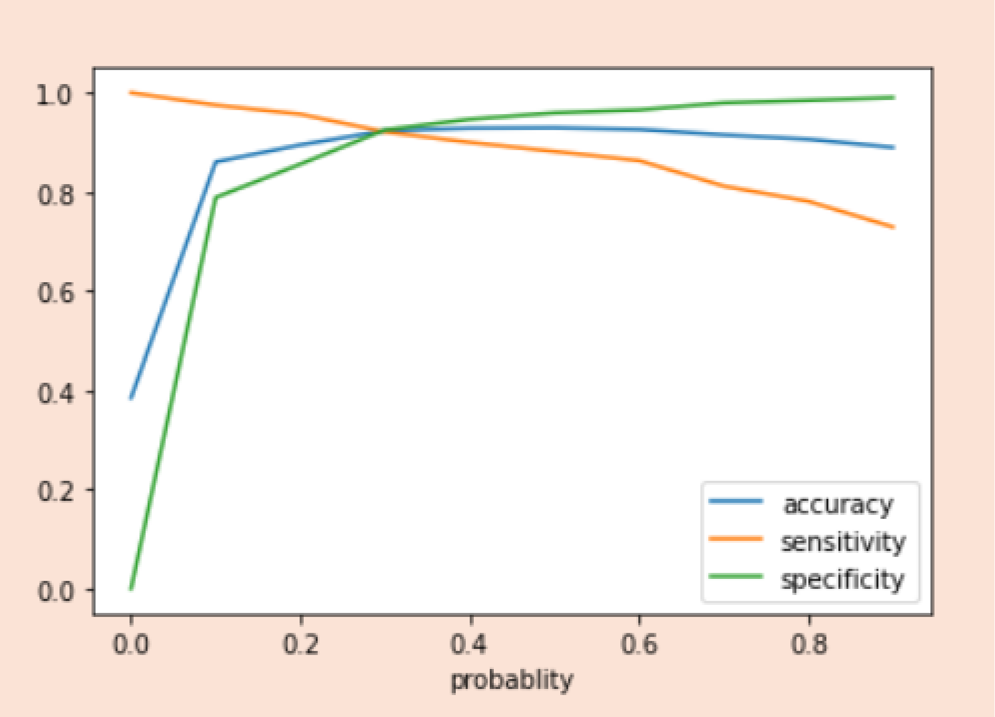


Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6351
Model:	GLM	Df Residuals:	6331
Model Family:	Binomial	Df Model:	19
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1195.9
Date:	Sat, 13 Apr 2024	Deviance:	2391.7
Time:	19:59:50	Pearson chi2:	8.39e+03
No. Iterations:	8		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.9214	0.267	-3.446	0.001	-1.445	-0.397
Do Not Email	-1.3061	0.282	-4.636	0.000	-1.858	-0.754
Total Time Spent on Website	1.1597	0.063	18.466	0.000	1.037	1.283
Origin_Landing Page Submission	-0.9033	0.136	-6.637	0.000	-1.170	-0.637
Origin_Lead Add Form	1.1378	0.503	2.262	0.024	0.152	2.124
Source_Olark Chat	1.0008	0.169	5.910	0.000	0.669	1.333
Source_Welingak Website	3.9463	0.875	4.509	0.000	2.231	5.661
last_activity_Olark Chat Conversation	-0.7893	0.258	-3.057	0.002	-1.295	-0.283
last_activity_SMS Sent	2.8021	0.166	16.884	0.000	2.477	3.127
Specialization_Travel and Tourism	-1.1149	0.453	-2.461	0.014	-2.003	-0.227
Tag_Closed by Horizzon	6.2370	0.746	8.361	0.000	4.775	7.699
Tag_Interested in other courses	-2.5018	0.363	-6.888	0.000	-3.214	-1.790
Tag_Lost to EINS	5.9748	0.753	7.937	0.000	4.499	7.450
Tag_Others	-2.6469	0.244	-10.848	0.000	-3.125	-2.169
Tag_Ringing	-3.6680	0.269	-13.638	0.000	-4.195	-3.141
Tag_Will revert after reading the email	4.3337	0.219	19.792	0.000	3.905	4.763
Profile_Not Specified	-1.3477	0.213	-6.312	0.000	-1.766	-0.929
Profile_Student of SomeSchool	-2.8557	0.899	-3.178	0.001	-4.617	-1.095
Last Notable Activity_Email Opened	1.1355	0.165	6.879	0.000	0.812	1.459
Last Notable Activity_Others	1.8625	0.453	4.115	0.000	0.975	2.750

Model Evaluation



ACCURACY, SENSITIVITY AND SPECIFICITY (TRAIN)

- Accuracy - 92.93%
- Sensitivity - 88.14%
- Specificity - 95.93%

We can see the optimal cutoff point is coming around 0.3, lets see values for other metrics with cutoff point as 0.3.

ACCURACY, SENSITIVITY AND SPECIFICITY (TEST)

- Accuracy – 90.86%
- Sensitivity – 90.09%
- Specificity – 91.29%

For the above observations the metric values are very close for Train and Test data so we can accept this model

Conclusion

People spending higher than average time are promising leads, hence the company should target these customers as this will lead to higher conversion rate.

The mode that has been promising since the start that is observed even in EDA is “SMS” and hence the company should continue with this. Working professional should be targeted since they show better conversion rate and tend to look for better career prospective.

Unemployed leads should be avoided as they have low conversion rate and might not even have enough budget to spend on course. Moreover, they have low conversion rate.

The built logistic regression model shows 91.15% accuracy.

The threshold has been selected from Accuracy, Sensitivity, Specificity measures and Precision and Recall Curves. With this, the model finds correct promising leads and the ones which have very less chance of getting converted.

Hence, we can say that the overall model built is accurate and hence one can achieve desired results with the same