Lead Score Case Study Summary Report

Submitted by

- Sanyam Gupta
- Saran G
-Arun Santhosh Vanam

Problem

Company X Education sales online courses, the courses are advertised over various online platforms like different search engines and social networking sites to facilitate the person for easy registration for availing the course online.

A person who want to enroll for the course have to register their details for initial discussion regarding the course. That person is treated as a lead for the company, when the person shows some more interest while discussion treated as a hot lead and the sales team target those people and follow up to make them join in the course.

However, here the problem occurs that sales person is not able to identify the hot leads. So, they do take follow ups of random leads who may or may not get convert in the feature, which continues to achieve the target conversion rate as 30 percent throughout the year as they call all of them without understanding the lead behavior.

In such case company wants a way to identify the potential candidates as hot leads so that the sales team could only focus on these candidates, whose probability of getting converted is high. Accordingly, the company can achieve their annual targets as early as possible and engage the team for other tasks as well.

As per their company CEO, the current conversion rate is 30 % and want to increase it till 80%.

Lead Score Case Study Learning Summary

When we first received the problem statement and data, our initial step was to explore the dataset and attempt to build a logistic regression model. However, we encountered several data-related challenges. These included missing values, which appeared in different forms such as 'NaN' and as the default dropdown option 'Select' (indicating that some fields were not filled out). Additionally, the data was biased and lacked sufficient variance.

The missing data was likely due to leads leaving certain fields blank, as they were optional, or because salespeople did not fill them out when leads showed little interest in joining.

Steps:

- Reading and understanding data
- Data Quality Check
- Data Preparation and Cleanup
- Outlier Treatment
- Dummy Variables Creation
- Test-Train split of data
- Scaling
- Check correlations
- Model building (GLM, RFE, VIF)
- Model Evaluation
- Make Predictions on the Test Set
- Precision-Recall view
- Add lead score

First model Learnings: -

Initially, when building the first model, instead of imputing or using random values for nulls, we addressed the issue by removing features with more than 30% missing values and excluding some records with fewer missing values. Our goal was to retain as many records as possible while building the model. This approach resulted in a model with approximately 80% accuracy, 79% sensitivity, and 79% specificity.

All variables had a p-value of less than 0.05, indicating statistical significance. However, the variance inflation factor (VIF) for one variable, *Page Views Per Visit*, was high, suggesting multi-collinearity. To address this, we dropped the column and re-ran the model.

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	4392
Model:	GLM	**Df Residuals:**	4376
Model Family:	Binomial	**Df Model:**	15
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1992.4
Date:	Mon, 11 Jan 2021	**Deviance:**	3984.8
Time:	01:19:48	**Pearson chi2:**	4.81e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.5730	0.118	-13.278	0.000	-1.805	-1.341
TotalVisits	2.2692	0.340	6.680	0.000	1.603	2.935
Total Time Spent on Website	4.4914	0.191	23.455	0.000	4.116	4.867
Page Views Per Visit	-1.2627	0.294	-4.296	0.000	-1.839	-0.687
Lead Origin_Lead Add Form	2.1505	0.223	9.638	0.000	1.713	2.588
Lead Source_Direct Traffic	-1.6529	0.150	-10.993	0.000	-1.948	-1.358
Lead Source_Google	-1.2775	0.151	-8.447	0.000	-1.574	-0.981
Lead Source_Organic Search	-1.4947	0.185	-8.071	0.000	-1.858	-1.132
Lead Source_Referral Sites	-1.5400	0.419	-3.674	0.000	-2.362	-0.718
Lead Source_Welingak Website	2.6249	1.034	2.538	0.011	0.598	4.652
Last Activity_Email Bounced	-1.0959	0.410	-2.675	0.007	-1.899	-0.293
Last Activity_Email Opened	0.9902	0.108	9.182	0.000	0.779	1.202
Last Activity_Had a Phone Conversation	2.8198	0.876	3.220	0.001	1.103	4.536
Last Activity_SMS Sent	1.9187	0.113	16.967	0.000	1.697	2.140
What is your current occupation_Working Professional	2.5692	0.194	13.218	0.000	2.188	2.950
Last Notable Activity_Unreachable	3.4554	0.816	4.232	0.000	1.855	5.056

	Features	VIF
2	Page Views Per Visit	6.56
0	TotalVisits	4.59
5	Lead Source_Google	3.53
4	Lead Source_Direct Traffic	3.05
1	Total Time Spent on Website	2.47
6	Lead Source_Organic Search	2.41
12	Last Activity_SMS Sent	2.09
10	Last Activity_Email Opened	2.06
3	Lead Origin_Lead Add Form	1.71
8	Lead Source_Welingak Website	1.32
13	What is your current occupation_Working Profes...	1.21
7	Lead Source_Referral Sites	1.11
9	Last Activity_Email Bounced	1.09
11	Last Activity_Had a Phone Conversation	1.01
14	Last Notable Activity_Unreachable	1.01

Second Model Learning: -

In the second model, after dropping the *Page Views Per Visit* column, we re-ran the analysis. The p-values for all remaining variables were below 0.05, and the VIF scores for all variables were under 5, indicating no multi-collinearity issues. Therefore, this model is robust and suitable for proceeding further.

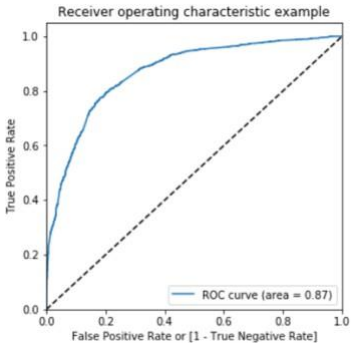
Generalized Linear Model Regression Results

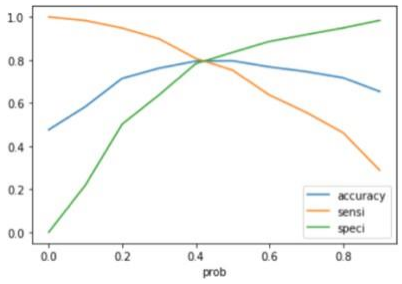
Dep. Variable:	Converted	**No. Observations:**	4392
Model:	GLM	**Df Residuals:**	4377
Model Family:	Binomial	**Df Model:**	14
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2001.8
Date:	Mon, 11 Jan 2021	**Deviance:**	4003.6
Time:	01:19:48	**Pearson chi2:**	4.80e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

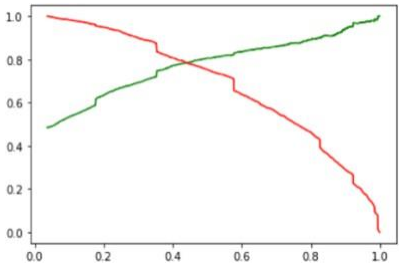
	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.5401	0.118	-13.104	0.000	-1.770	-1.310
TotalVisits	1.5367	0.291	5.288	0.000	0.967	2.106
Total Time Spent on Website	4.4997	0.191	23.520	0.000	4.125	4.875
Lead Origin_Lead Add Form	2.1659	0.223	9.712	0.000	1.729	2.603
Lead Source_Direct Traffic	-1.8843	0.141	-13.330	0.000	-2.161	-1.607
Lead Source_Google	-1.5232	0.141	-10.823	0.000	-1.799	-1.247
Lead Source_Organic Search	-1.7925	0.172	-10.412	0.000	-2.130	-1.455
Lead Source_Referral Sites	-1.8312	0.413	-4.430	0.000	-2.641	-1.021
Lead Source_Welingak Website	2.6191	1.033	2.534	0.011	0.594	4.645
Last Activity_Email Bounced	-1.0851	0.408	-2.662	0.008	-1.884	-0.286
Last Activity_Email Opened	0.9344	0.107	8.767	0.000	0.725	1.143
Last Activity_Had a Phone Conversation	2.7627	0.873	3.166	0.002	1.052	4.473
Last Activity_SMS Sent	1.8508	0.111	16.632	0.000	1.633	2.069
What is your current occupation_Working Professional	2.5755	0.194	13.289	0.000	2.196	2.955
Last Notable Activity_Unreachable	3.4174	0.820	4.166	0.000	1.810	5.025

	Features	VIF
0	TotalVisits	3.39
4	Lead Source_Google	2.93
3	Lead Source_Direct Traffic	2.58
1	Total Time Spent on Website	2.47
11	Last Activity_SMS Sent	2.06
9	Last Activity_Email Opened	2.04
5	Lead Source_Organic Search	2.02
2	Lead Origin_Lead Add Form	1.71
7	Lead Source_Welingak Website	1.32
12	What is your current occupation_Working Profes...	1.21
8	Last Activity_Email Bounced	1.09
6	Lead Source_Referral Sites	1.08
10	Last Activity_Had a Phone Conversation	1.01
13	Last Notable Activity_Unreachable	1.01

We validated all key points, including correlation, p-value statistics, and multi-collinearity, and successfully built a strong model. The model demonstrated an improved ROC curve, an optimal threshold value, and achieved nearly 80% accuracy, sensitivity, and specificity on both the training and test datasets.







Summary of Logistic Regression Model Performance and Insights:

- Model Accuracy: 79.5%
- Sensitivity: 79.6%
- Specificity: 79.4%
- The model assigns a score between 0 and 100 to each record, based on predicted values.
- The accuracy of the model is nearly 80%.
- The analysis offered insights into the practical application of logistic regression in the industry.
- Identified and addressed various data-related challenges during model development.

Conclusion and Key Insights from Logistic Regression Model Development:

In summary, we successfully developed a logistic regression model that demonstrates robust performance, with an accuracy of 79.5%, sensitivity of 79.6%, and specificity of 79.4%. The model is capable of assigning a predictive score between 0 and 100 to each record, which effectively reflects the probability of lead conversion. Throughout the process, we meticulously addressed data challenges such as missing values, multi-collinearity, and biased data, resulting in a refined model with statistically significant variables and low VIF scores.

This project provided valuable insights into the practical implementation of logistic regression in a business context, highlighting the importance of thorough data exploration and validation. The outcome is a reliable model that can be leveraged for strategic decision-making, particularly in optimizing lead conversion efforts.