Case Study - Leads Scoring

By

1. Alistair Dsilva

2. Chaman Jha

3. Pranjal Satish Tikande

PROBLEM STATEMENT

Introduction

An education company, X Education sells online courses to industry professionals. The company markets its courses on various websites and search engines such as Google

When visitors arrive on the website, they might browse through the courses, complete a form for a course, or watch some videos. If they fill out a form with their email address or phone number, they become classified as leads. Additionally, the company obtains leads from past referrals. Once these leads are acquired, sales team members begin contacting them through calls and emails. At X Education, the typical lead conversion rate is about 30%.

BUSINESS GOALS:

Company wishes to identify the most potential leads, also known as “Hot Leads”

The company needs a model wherein a lead score is assigned to each of the leads such that the customer with higher lead score have a higher conversion chance and customer with lower lead score have a lower conversion chance

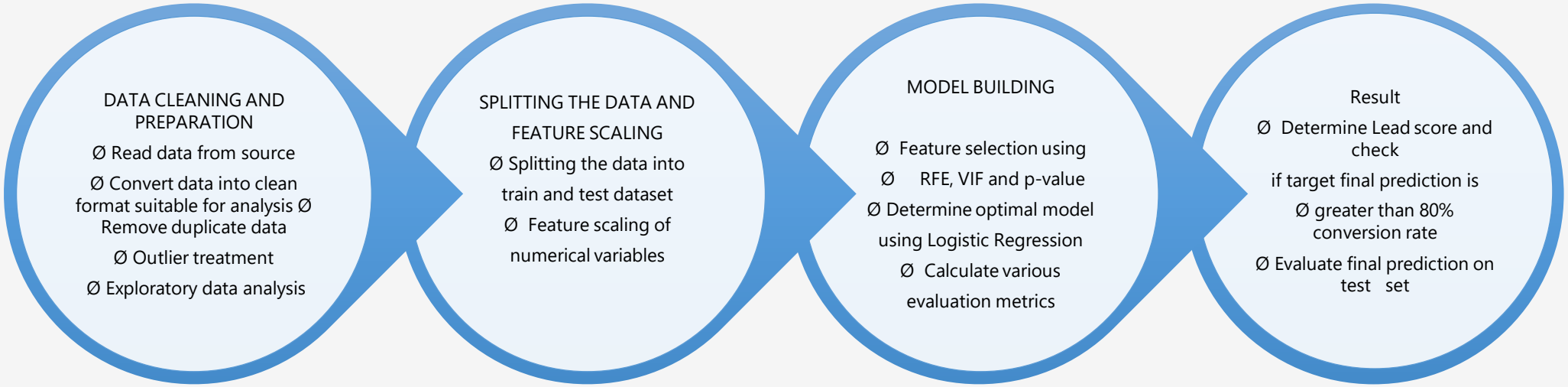
The CEO, in particular, has given a ballpark number for the lead conversion rate i.e. 80%

APPROACH USED IN Analysis

1. DATA CLEANING AND IMPUTING MISSING VALUES
2. EXPLORATORY DATA ANALYSIS : UNIVARIATE , BIVARIATE and MULTIVARIATE ANALYSIS
3. FEATURE SCALING AND DUMMY VARIABLE CREATION
4. LOGISTIC REGRESSION MODEL BUILDING
5. MODEL EVALUATION : SPECIFICITY , SENSITIVITY, PRECISION and RECALL
6. CONCLUSION AND RECOMMENDATION

PROBLEM SOLVING METHODOLOGY

How it works:



DATA CONVERSION

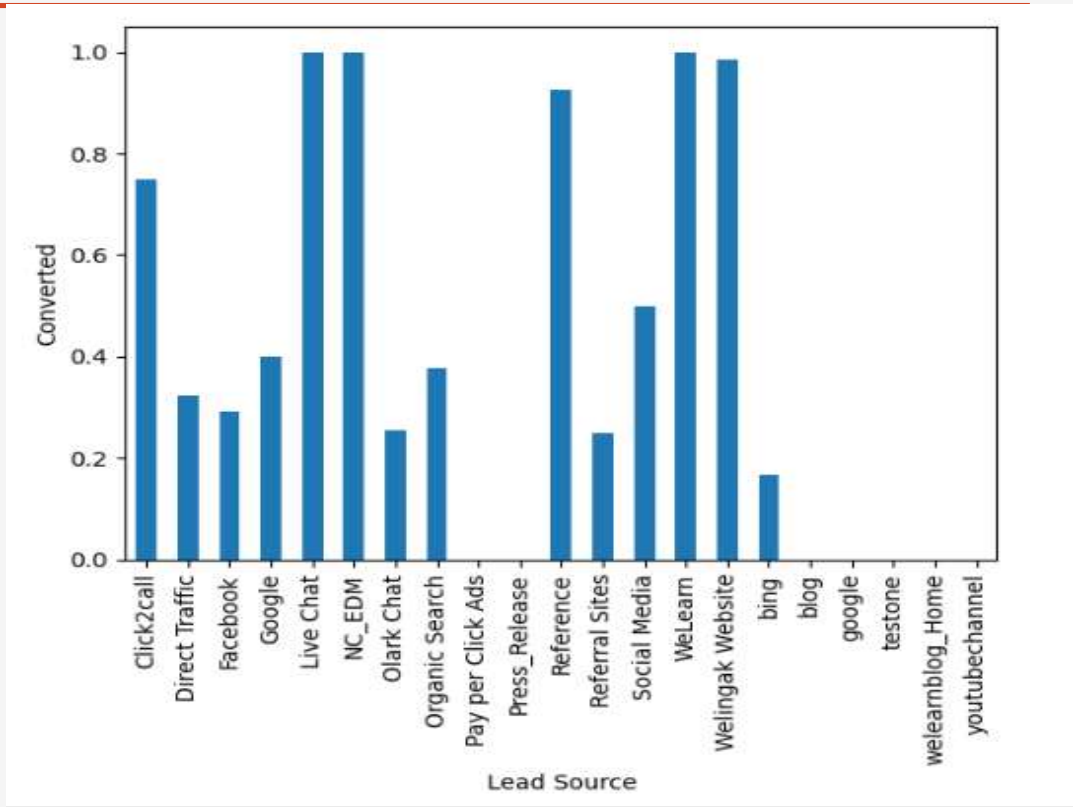
- CONVERTING THE VARIABLE WITH VALUES YES/NO to 1/0s
- CONVERTING THE MISSING COLUMNS INTO unknown AND Other
- CONERTING SELECT LEADS INTO Other Leads
- ASSIGNING THE VALUE 1 TO HOT LEADS REST 2 AND 3.
- DROPIING THE COLUMNS HAVING >70% OF NULL VALUES
- DROPPING UNNECESSARY COLUMNS
- DROPPING THE ROWS AS THE NULL VALUES WERE <2%

EXPLORATORY DATA ANALYSIS

Univariate Analysis and Multivariate analysis

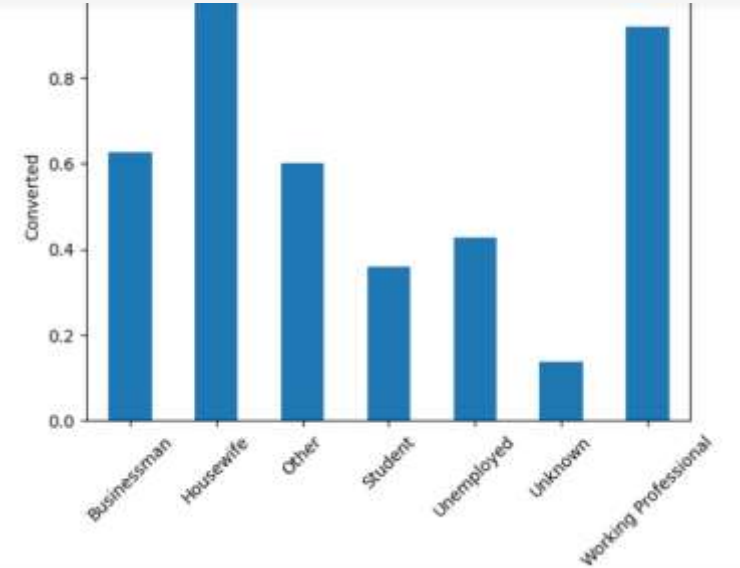
The highest number of leads come from Live Chat ,Welingak Webside , weLearn, NC_EDM Google and Direct Traffic.

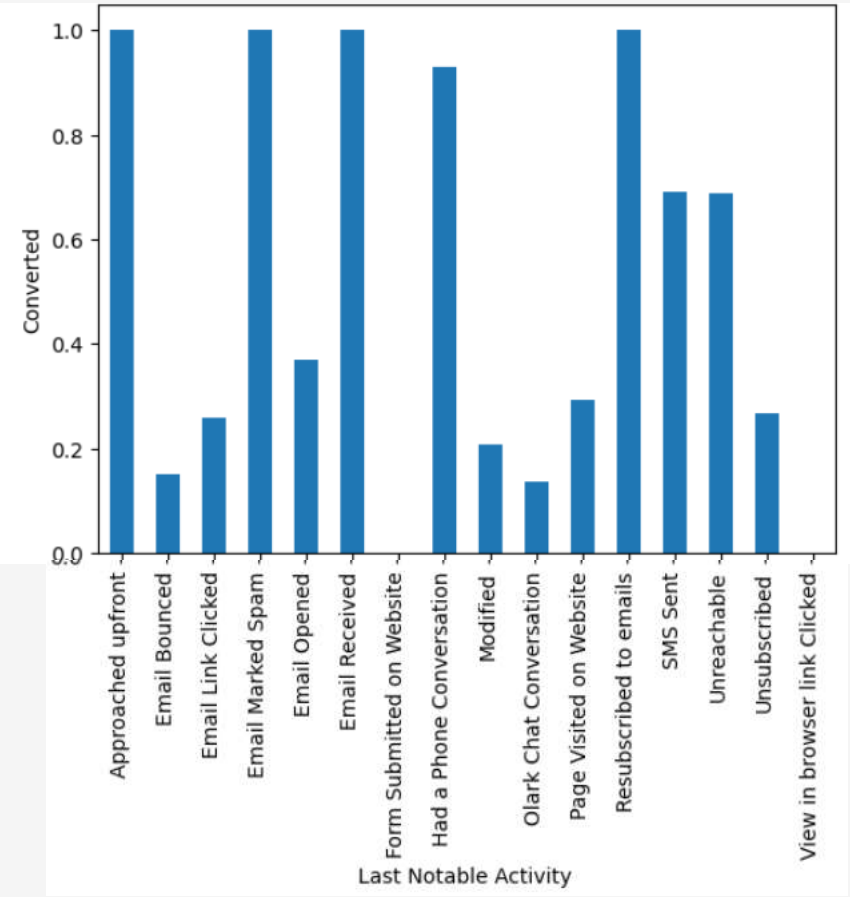
Leads from References and the Welingak Website have the highest conversion rates. To boost the overall lead conversion rate, efforts should be made to improve the conversion rates for leads from Google, Olark Chat, Organic Search, and Direct Traffic. Additionally, increasing the number of leads from References and the Welingak Website would be beneficial.



EXPLORATORY DATA ANALYSIS

Occupation wise conversion rate is high in working professional after Housewife , need to more focus on them as hot leads.





The count of lead's last activity as "Email Opened" is maximum
The conversion rate of SMS sent as last activity is maximum

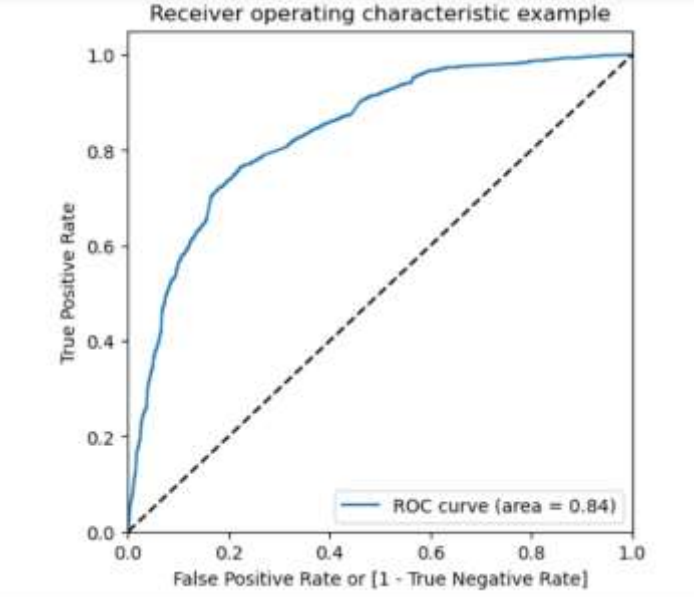
The majority of leads have approach upfront , email received then the valuable hot leads are "Email Opened" as their last activity. Leads with "SMS Sent" as their last activity have the highest conversion rate. To improve results, we should aim to increase the conversion rate for leads with "Email Opened" as their last activity by following up with a call. Additionally, we should try to increase the number of leads whose last activity is "SMS Sent."

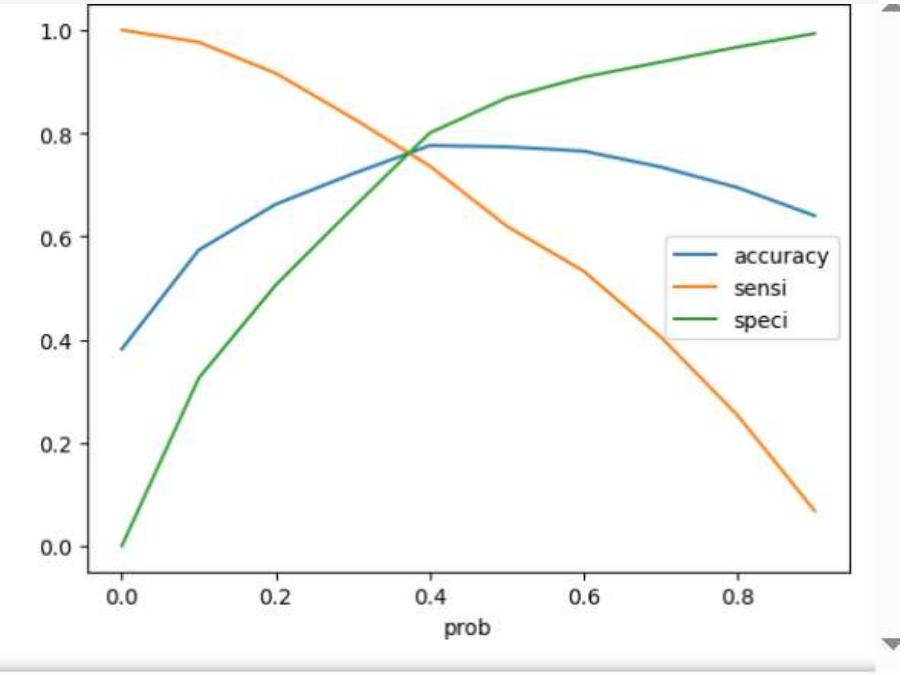
MODEL BUILDING

SPLITTING THE DATA INTO TEST AND TRAINING SETS

WE HAVE CHOSEN THE TRAIN_TEST SPLIT RATIO AS 70:30

USING RFE TO CHOOSE TOP 15 VARIABLES
UILD MODEL BY REMOVING THE VARIABLES WHOSE p-VALUE
0.05 AND VIF > 5
PREDICTIONS ON TEST DATASET
OVERALL ACCURACY IS 76.0 %





The graph indicates that the best threshold for making predictions is 0.35.

MODEL EVALUATION

Based on the Sensitivity-Specificity-Accuracy plot, a probability of 0.27 appears to be the optimal cutoff. Meanwhile, the Precision-Recall Curve suggests that 0.3 might be the best threshold.

We have decided to use 0.27 as the optimal cutoff probability for assigning Lead Scores in the training data

	probability_score	accuracy_score	sensitivity_score	specificity_score	precision_score
0.0	0.0	0.381515	1.000000	0.000000	0.381515
0.1	0.1	0.573768	0.976475	0.325356	0.471691
0.2	0.2	0.662415	0.916220	0.505855	0.533526
0.3	0.3	0.721304	0.829137	0.654786	0.597028
0.4	0.4	0.776098	0.736690	0.800407	0.694823
0.5	0.5	0.773579	0.619893	0.868381	0.743933
0.6	0.6	0.765234	0.532811	0.908605	0.782424
0.7	0.7	0.734687	0.405283	0.937882	0.800979
0.8	0.8	0.694536	0.252992	0.966904	0.825034
0.9	0.9	0.640214	0.068097	0.993126	0.859375

TRAIN DATA - CONFUSION MATRIX

Predicted Actual	Not Converted	Converted
Not Converted	2441	1487
Converted	370	2053

CALCULATED ACCURACY, SENSITIVITY AND SPECIFICITY FOR VARIOUS PROBABILITY CUTOFFS FROM 0.1 TO 0.9

Ø AS PER THE GRAPH AND LOOKING AT THE OTHER SCORES, IT CAN BE SEEN THAT THE OPTIMAL POINT IS 0.27

Accuracy	**70 %**
PRECISION	57 %
SENSITIVITY	84 %
SPECIFICITY	62 %

MODEL PREDICTION

TRAIN DATA - CONFUSION MATRIX		
Predicted Actual	**Not Converted**	**Converted**
Not Converted	1247	464
Converted	196	816

Accuracy	**61 %**
PRECISION	74 %
SENSITIVITY	80 %
SPECIFICITY	72 %

	feature	importance	abs_importance
5	Total Time Spent on Website	4.374200	4.374200
11	What is your current occupation	3.201829	3.201829
22	Lead Quality	-3.191956	3.191956
21	Tags	2.573936	2.573936
1	Lead Source	2.396875	2.396875
6	Page Views Per Visit	-2.109061	2.109061
12	What matters most to you in choosing a course	-1.660608	1.660608
0	Lead Origin	1.523987	1.523987
2	Do Not Email	-1.383406	1.383406
8	Country	1.348959	1.348959
25	Lead Profile	1.263985	1.263985
4	TotalVisits	0.739957	0.739957
30	Last Notable Activity	0.552490	0.552490
19	Through Recommendations	0.395692	0.395692
29	A free copy of Mastering The Interview	-0.313191	0.313191
13	Search	-0.309671	0.309671
10	How did you hear about X Education	0.262690	0.262690
3	Do Not Call	0.250413	0.250413
18	Digital Advertisement	-0.173389	0.173389
9	Specialization	-0.153621	0.153621
27	Asymmetrique Activity Index	-0.107226	0.107226
7	Last Activity	0.093095	0.093095
26	City	-0.069020	0.069020
14	Magazine	0.000000	0.000000
23	Update me on Supply Chain Content	0.000000	0.000000

CONCLUSION

The logistic regression model effectively predicts the likelihood of lead conversion. Sensitivity-specificity and Precision-Recall metrics were considered, with the final model cut-off based on sensitivity-specificity. The lead score calculated shows a conversion rate of around 82% on test data compared to 78 % on training data. The model is capable of meeting the company's future needs.

Key factors contributing to lead conversion include:

- Total Time Spent on Website
- Tags
- What is your current occupation

Overall, the model is robust and aligns well with the company's requirements