LEAD SCORING CASE STUDY

PROBLEM SOLVING USING LOGISTIC REGRESSION

By – ANUSHREE CHOWDHURY

BUSINESS OBJECTIVE

• X Education generates a high number of leads but has a low conversion rate. For example, if they generate 100 leads in a day, only around 30 of them will convert.
•To improve efficiency, the organization aims to discover high-potential leads, or 'Hot Leads'. If they are successful in identifying this set of prospects, the lead conversion rate should increase because the sales team will be focusing more on connecting with the potential leads rather than making calls to everyone.

• We use logistic regression to find potential leads.

SOLUTION METHODOLOGY

- ✓ Data Cleaning and manipulation
- ✓ Exploratory Data Analysis
- ✓ Model Building
- ✓ Model Evaluation
- ✓ Model Prediction on Testset
- ✓ Inferences
- ✓ Recommendation

DATA CLEANING

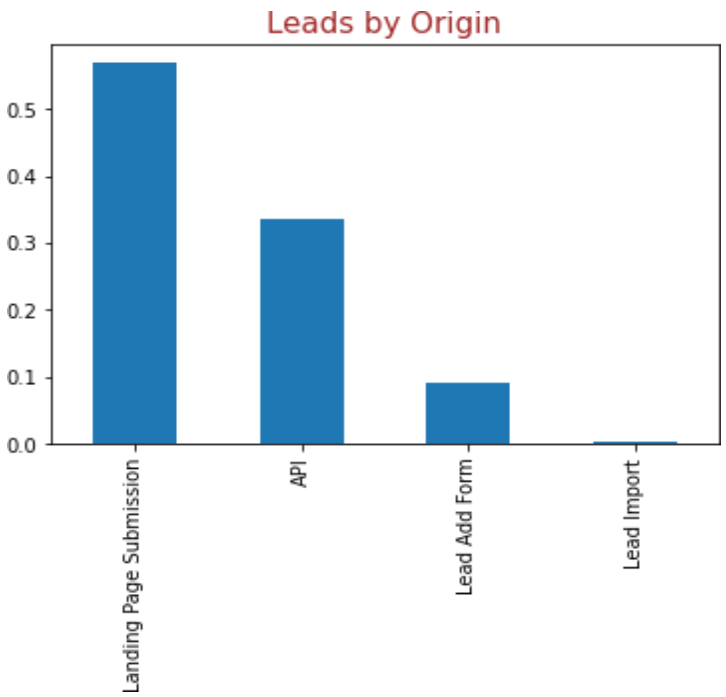
There were many additional rows and columns so we have Eliminated columns with 30% missing values from a dataset of over 9000 data points.

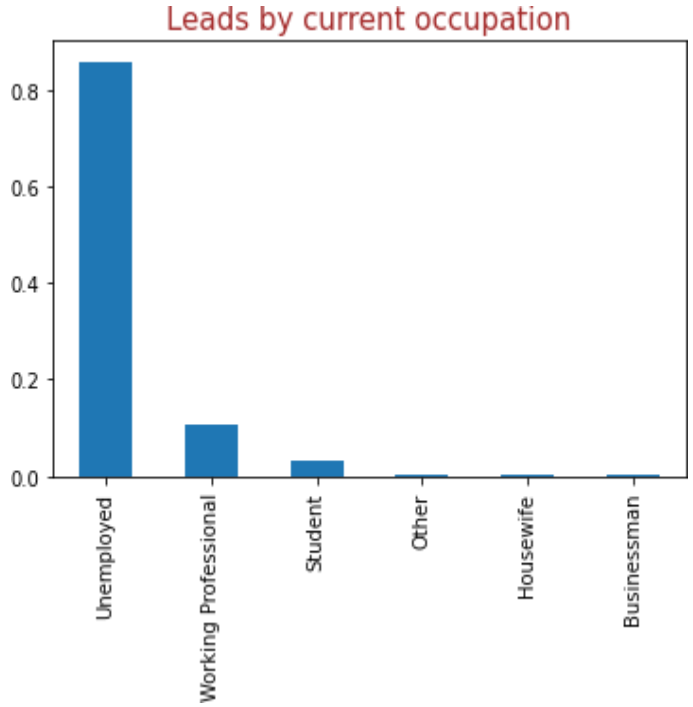
Removed City and Country variables as they are irrelevant to the company's offerings.

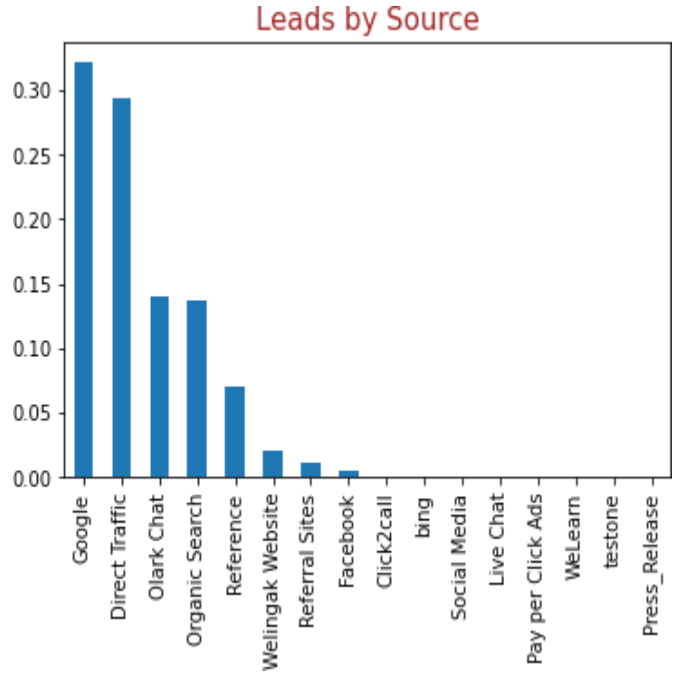
After cleaning the data, we discovered a 48% conversion rate for online courses. Prospect ID and Lead Number were removed as they are only record identifiers.

We also removed any columns with skewed data points as they lack predictability.

UNIVARIATE ANALYSIS

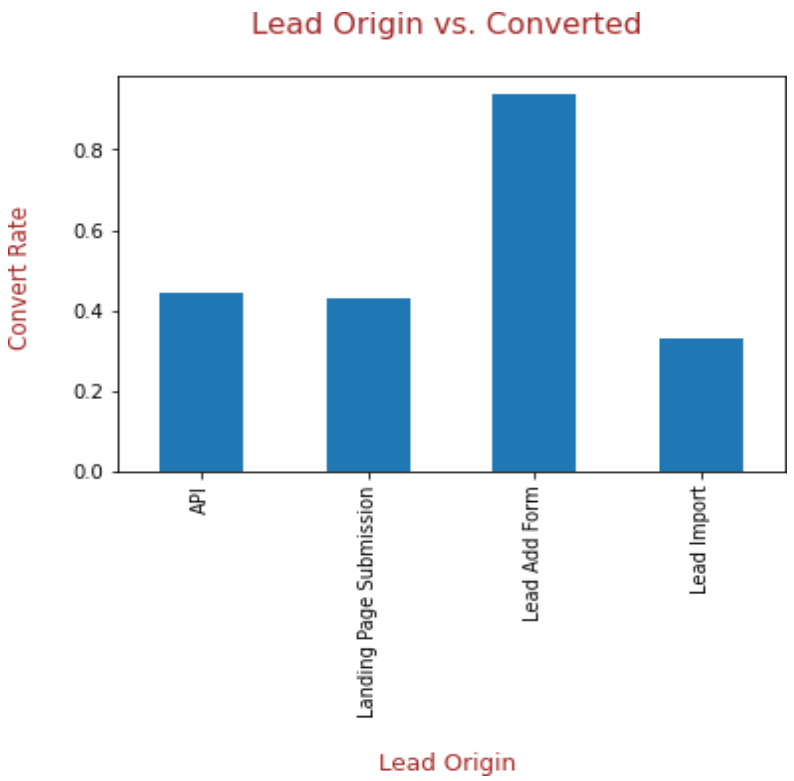


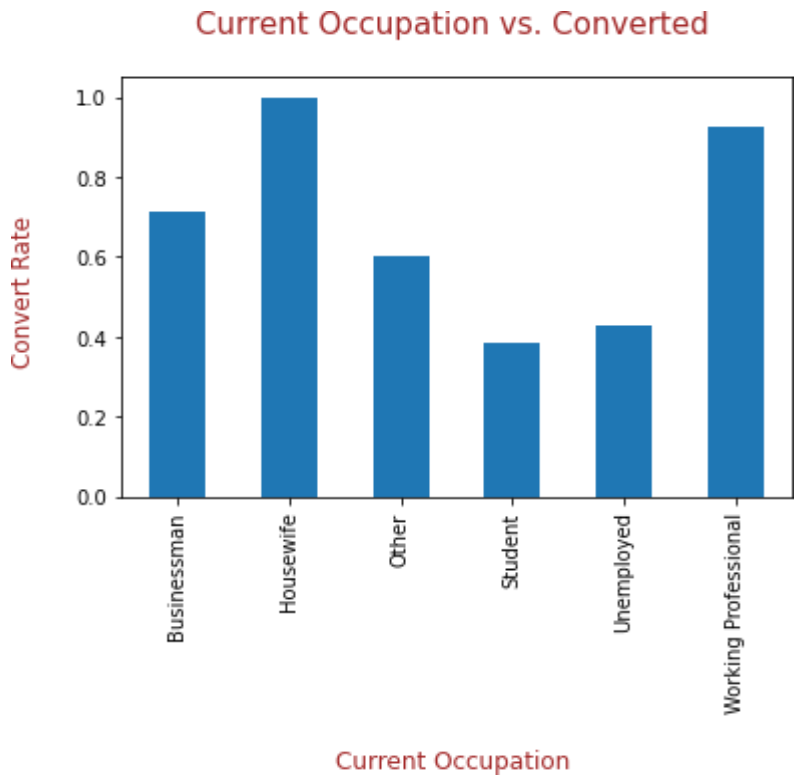


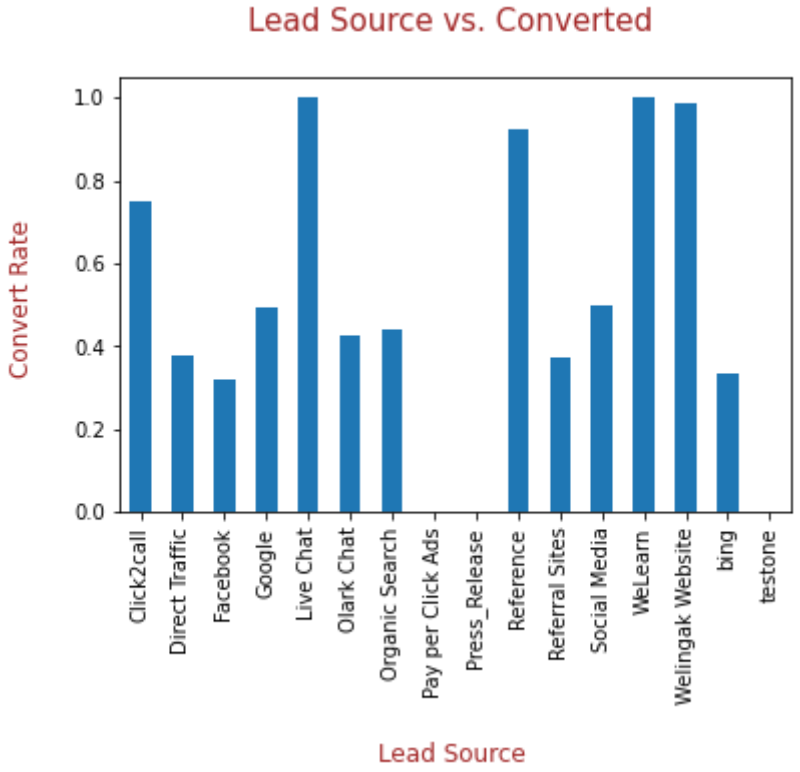


- ✓ Majority of leads are originated from Landing Page Submission followed by API
- ✓ More leads are received from 'Google' and 'Direct Traffic' followed by Olark Chat and Organic Search
- ✓ More leads are received from Unemployed customers

BI-VARIATE ANALYSIS







- ✓ Lead originated from Add Form are more likely to be converted
- ✓ Working Professional and Housewife are more likely to be converted
- ✓ Lead sources from Live Chat, Reference, WeLearn and Welingak Website are more likely to be Converted

MODEL BUILDING

- ✓ Divide data into train and test sets using a 70:30 ratio.
- ✓ Scale numerical features using the MinMax scaler.
- ✓ Use Recursive Feature Elimination (RFE) to find the top 15 features.
- ✓ Use p-value and variance inflation factor to remove insignificant characteristics.
- ✓ We ended up with 11 characteristics for the model.
- ✓ We developed a lead score (conversion probability multiplied by 100) that ranges from 0 to 100 points. A higher score suggests a heated lead with a higher chance of lead conversion.

MODEL EVALUATION

Generalized Linear Model Regression Results

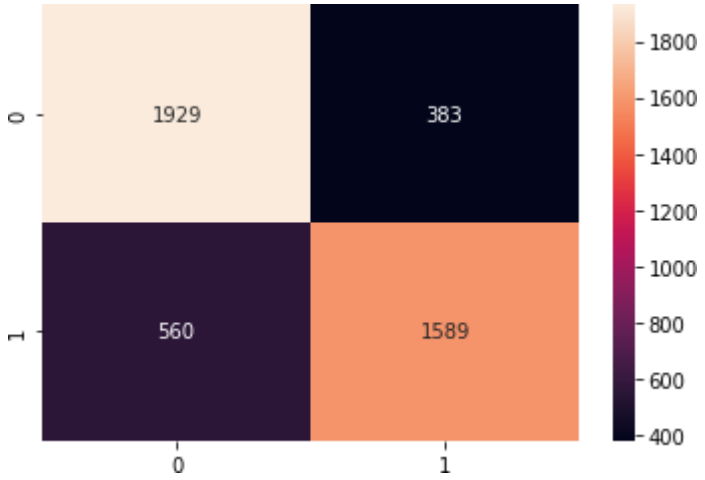
Dep. Variable:	Converted	**No. Observations:**	4461
Model:	GLM	**Df Residuals:**	4449
Model Family:	Binomial	**Df Model:**	11
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2079.1
Date:	Mon, 14 Nov 2022	**Deviance:**	4158.1
Time:	15:08:31	**Pearson chi2:**	4.80e+03
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2040	0.196	1.043	0.297	-0.179	0.587
TotalVisits	11.1489	2.665	4.184	0.000	5.926	16.371
Total Time Spent on Website	4.4223	0.185	23.899	0.000	4.060	4.785
Lead Origin_Lead Add Form	4.2051	0.258	16.275	0.000	3.699	4.712
Lead Source_Olark Chat	1.4526	0.122	11.934	0.000	1.214	1.691
Lead Source_Welingak Website	2.1526	1.037	2.076	0.038	0.121	4.185
Do Not Email_Yes	-1.5037	0.193	-7.774	0.000	-1.883	-1.125
Last Activity_Had a Phone Conversation	2.7552	0.802	3.438	0.001	1.184	4.326
Last Activity_SMS Sent	1.1856	0.082	14.421	0.000	1.024	1.347
What is your current occupation_Student	-2.3578	0.281	-8.392	0.000	-2.908	-1.807
What is your current occupation_Unemployed	-2.5445	0.186	-13.699	0.000	-2.908	-2.180
Last Notable Activity_Unreachable	2.7846	0.807	3.449	0.001	1.202	4.367

	Features	VIF
9	What is your current occupation_Unemployed	2.82
1	Total Time Spent on Website	2.00
0	TotalVisits	1.54
7	Last Activity_SMS Sent	1.51
2	Lead Origin_Lead Add Form	1.45
3	Lead Source_Olark Chat	1.33
4	Lead Source_Welingak Website	1.30
5	Do Not Email_Yes	1.08
8	What is your current occupation_Student	1.06
6	Last Activity_Had a Phone Conversation	1.01
10	Last Notable Activity_Unreachable	1.01

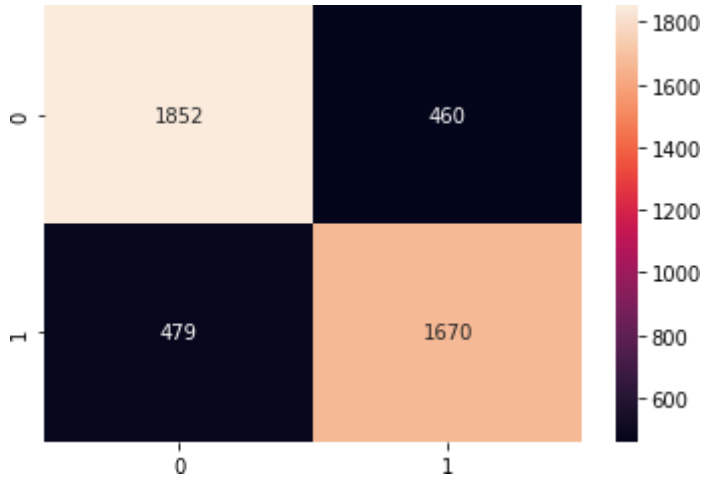
MODEL EVALUATION

TRAINING SET



Accuracy	78.86%
Sensitivity	73.94%
Specificity	83.43%
Precision	80.58%
Recall	73.94%

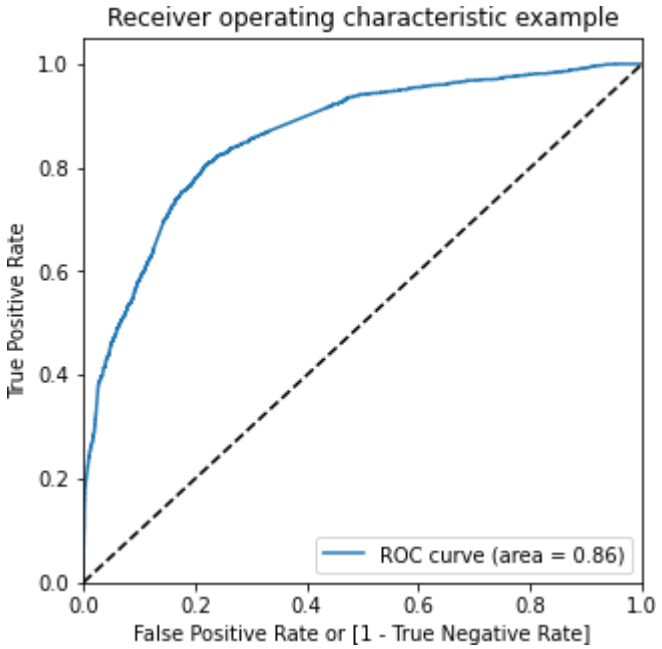
TEST SET

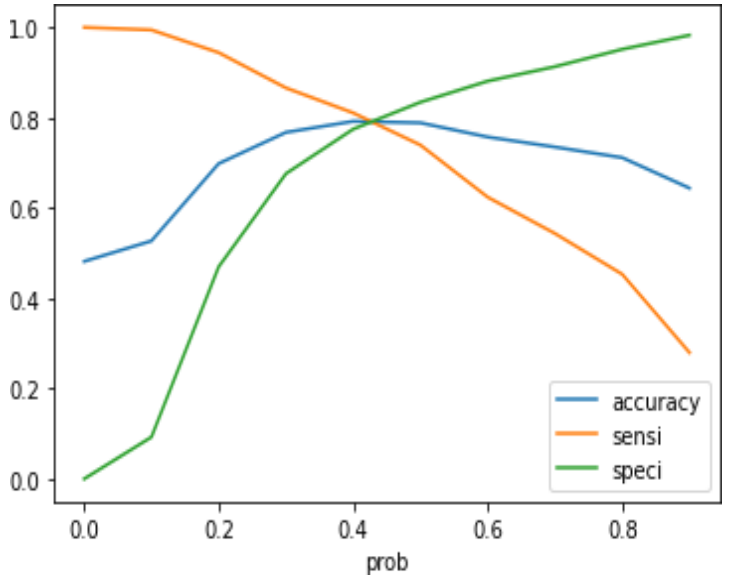


Accuracy	78.95%
Sensitivity	77.71%
Specificity	80.10%
Precision	78.40%
Recall	77.71%

MODEL EVALUATION - ROC/CUTOFF

	prob	accuracy	sensi	speci
0	0	48.17%	100.00%	0.00%
0.1	0.1	52.70%	99.44%	9.26%
0.2	0.2	69.83%	94.42%	46.97%
0.3	0.3	76.75%	86.60%	67.60%
0.4	0.4	79.20%	81.06%	77.47%
0.5	0.5	78.86%	73.94%	83.43%
0.6	0.6	75.72%	62.40%	88.11%
0.7	0.7	73.50%	54.35%	91.31%
0.8	0.8	71.15%	45.32%	95.16%
0.9	0.9	64.40%	27.97%	98.27%





INFERENCES

Top three variables in your model that contribute most to the probability of a lead being converted.

i. Total number of visits,

ii. Time spent on website.

iii. Lead Origin_Lead Add Form.

The top three categorical/dummy variables in the model that should be prioritized to maximize the likelihood of lead conversion

i. Lead Origin_Add Form

ii. Last Activity: Phone Conversation

iii. Lead Source: Welingak Website.

RECOMMENDATION

Depending on the requirements the model needs to be tweaked such that

Scenario 1:
So, as the company has more interns, we need to lower the cutoff criterion so that our model can anticipate practically all leads. The downside of this lower threshold is that we will misclassify some non-conversions as conversions, but this is a reasonable trade-off assuming that we have sufficient people to handle it.

Scenario 2:
Typically, when the organization has fewer personnel calling potential consumers, it is preferable to have more precise predictions, in which case the model specificity should be significantly higher. This means that, based on the graph above, we must select a considerably higher cutoff value. The trade-off is that we will miss some leads, but because the organization has fewer employees, they can focus on accurately forecasted leads.

Scenario 3:
To save money on lead conversion, the organization could automate SMS and email campaigns to potential leads during low-manpower periods.

THANK YOU