

# LEAD SCORING CASE STUDY

PROBLEM SOLVING USING LOGISTIC REGRESSION

By - Sandeep KS, Dolat Ram and Shrey Toshniwal




## BUSINESS OBJECTIVE

- ✓ X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- ✓ To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.
- ✓ Our objective is to help select promising leads using logistic regression.



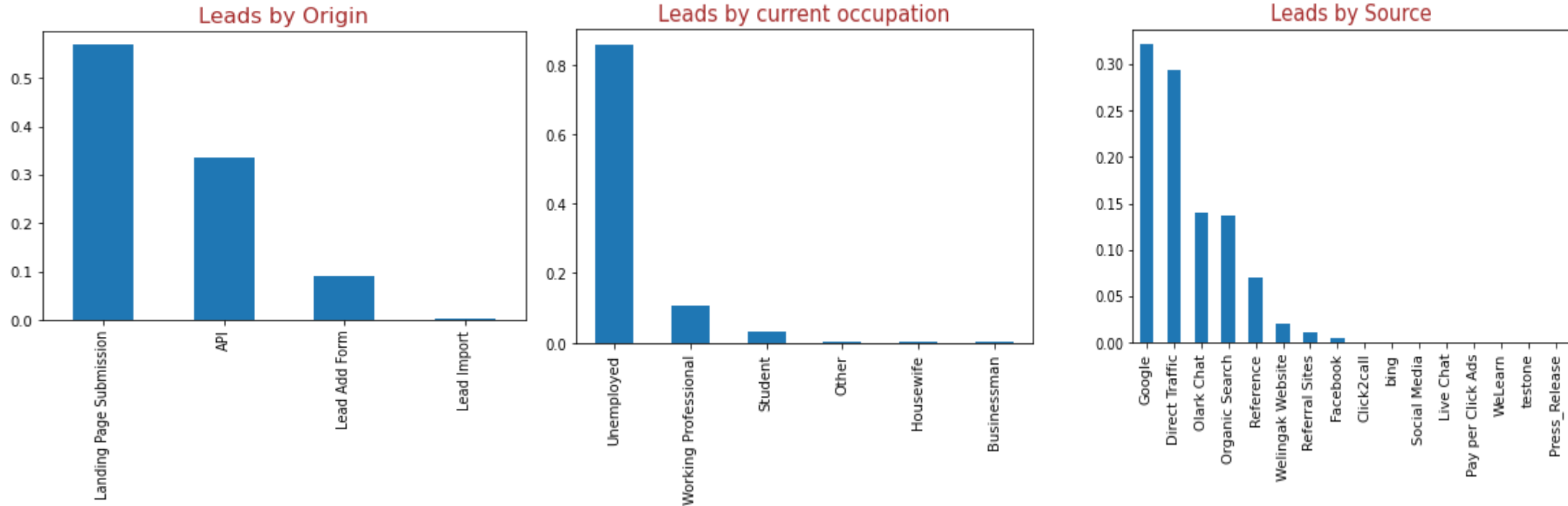
## SOLUTION METHODOLOGY

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

# DATA CLEANING

- ✓ There a lot of columns with high number of missing values and since we have around 9000+ data points we can eliminate the columns with 30% missing values;
- ✓ We dropped City and Country variables since it's of no use to us as the company provides online courses;
- ✓ Prospect ID and Lead Number are just records identifier and as hence dropped;
- ✓ We dropped all columns which have skewed data points as it wont have any predictability value;
- ✓ We have found 48% conversion rate after cleaning the data.

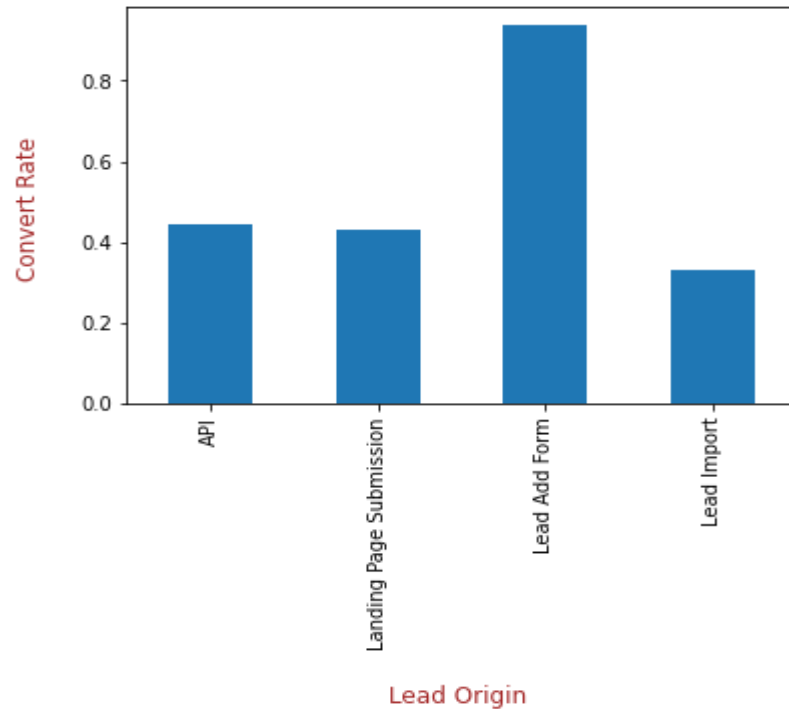
# UNIVARIATE ANALYSIS



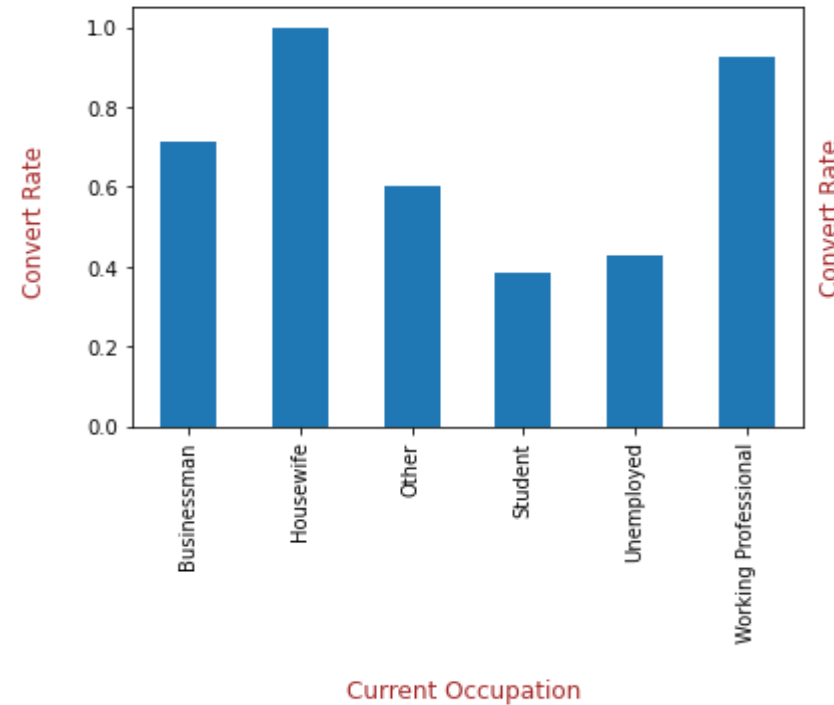
- ✓ Majority of leads are originated from Landing Page Submission followed by API
- ✓ More leads are received from 'Google' and 'Direct Traffic' followed by Olark Chat and Organic Search
- ✓ More leads are received from Unemployed customers

# BI-VARIATE ANALYSIS

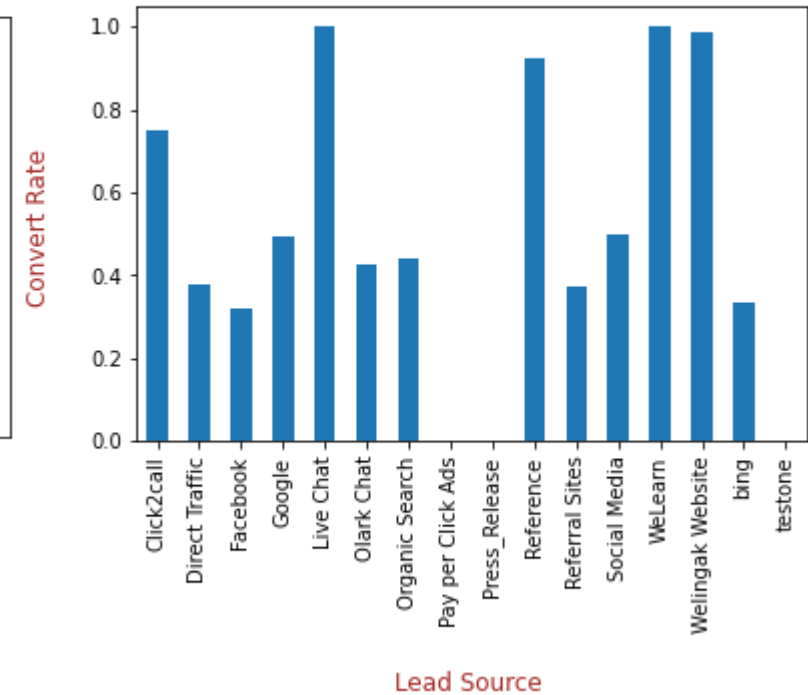
Lead Origin vs. Converted



Current Occupation vs. Converted



Lead Source vs. Converted



- ✓ 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

- ✓ Slitting the data into train and test split with 70:30 ratio
- ✓ Scale numerical feature using MinMax scaler
- ✓ Use Recursive feature Elimination (RFE ) to identify 15 most important feature
- ✓ Use p-value and Variance inflation factor to eliminate statistically insignificant features
- ✓ Finally, we ended up with 11 features for the model.
- ✓ We created a lead score (i.e.  $\text{Conversion probability} \times 100$ ) to give a score between 0 and 100. A higher score indicates a hot lead having a higher probability 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

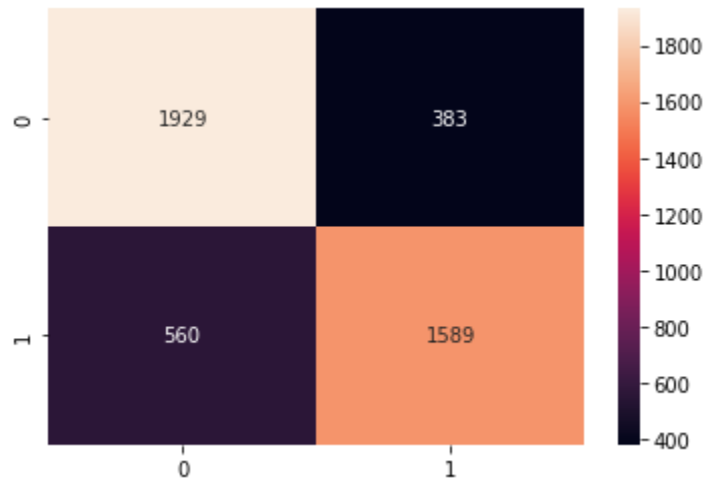
	coef	std err	z	P> z	[0.025	0.975]
const	0.2040	0.196	1.043	0.297	-0.179	0.587
TotalVisits	11.1489	2.665	4.184	0.000	5.926	16.371
Total Time Spent on Website	4.4223	0.185	23.899	0.000	4.060	4.785
Lead Origin_Lead Add Form	4.2051	0.258	16.275	0.000	3.699	4.712
Lead Source_Olark Chat	1.4526	0.122	11.934	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



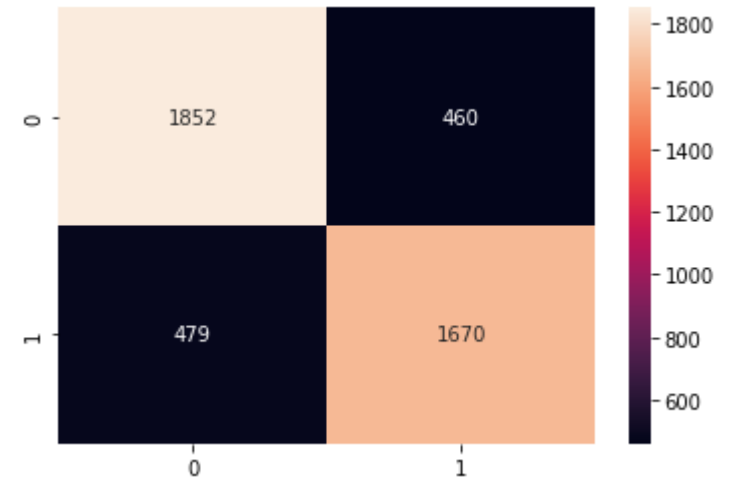
# MODEL EVALUATION

TRAINING SET



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

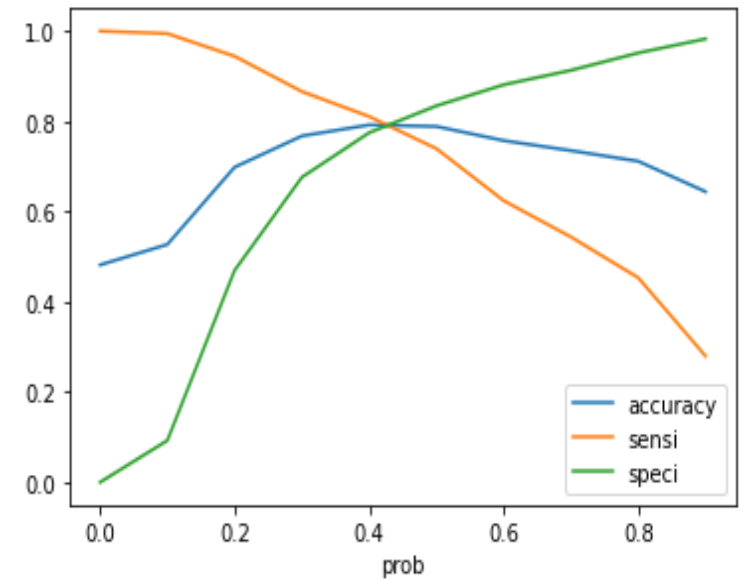
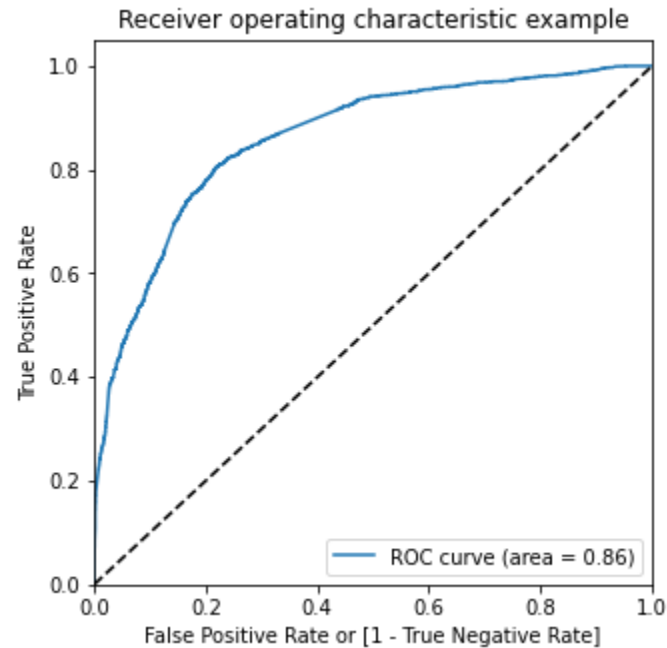
TEST SET



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 which contribute most towards the probability of a lead getting converted

- i. TotalVisits,*
- ii. Total Time Spent on Website,*
- iii. Lead Origin\_Lead Add Form*

Top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion

- i. Lead Origin\_Lead Add Form*
- ii. Last Activity\_Had a Phone Conversation*
- iii. Lead Source\_Welingak Website*

# RECOMMENDATION

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

**Scenario 1:**

So when the company has more interns we need have lower cutoff threshold so that our model can predict almost all leads. The flip side to this decrease in threshold will be that we will misclassify some non-conversions as conversions but this is a good tradeoff given we have more manpower to deal with it.

**Scenario 2:**

Typically, when the company has less people to call potential customers so it's good to have more accurate predictions in which case the model specificity should be much more higher. This would mean from the above graph we would have to choose a cutoff point which is much higher. The tradeoff of this is that we are going to miss some leads but given that the company has less manpower who can focus more on correctly predicted leads.

**Scenario 3:**

The company should focus on sending automated SMS and emails to potential leads during the time they have less manpower which allows for cost effective lead conversion without manual intervention.



*THANK YOU*

