Lead Scoring Case Study

By

Surya Prabha

Sharath Nair

Lalith Vamsi Bonthu

Contents:

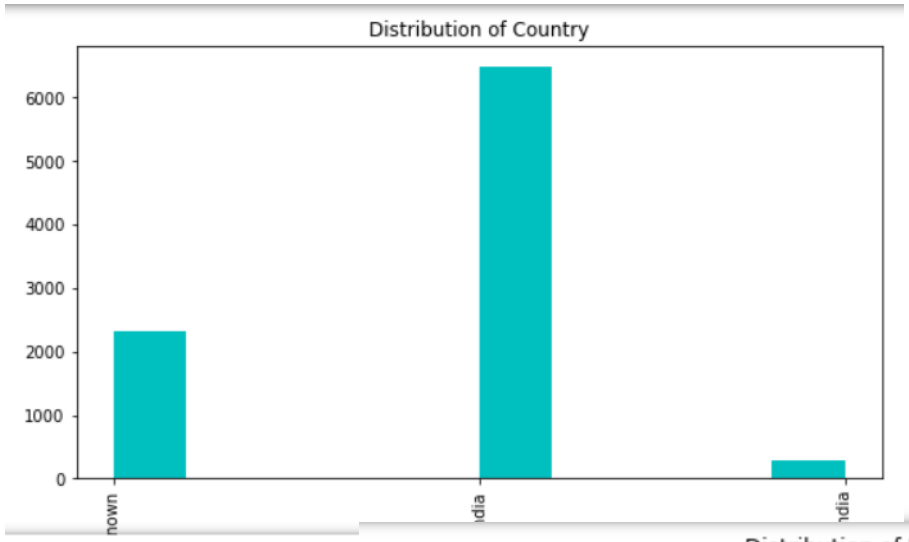
Problem Statement

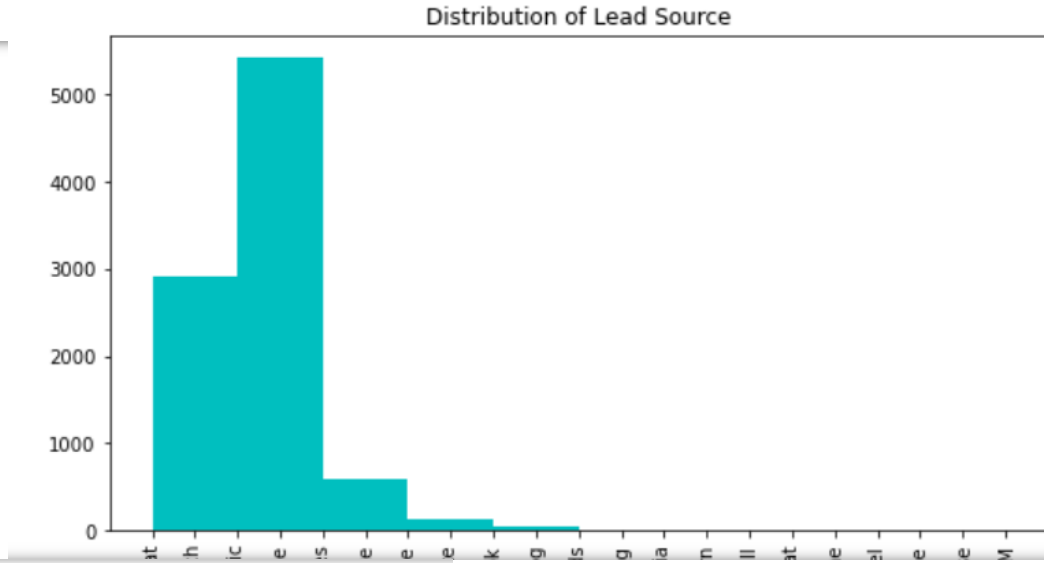
- The company X Education sells online courses to industry professionals, and markets its courses on several websites and search engines like Google.
- Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead.
- Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. The typical lead conversion rate at X education is around 30%.
- The company wants to build a model to improve the conversion rate. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

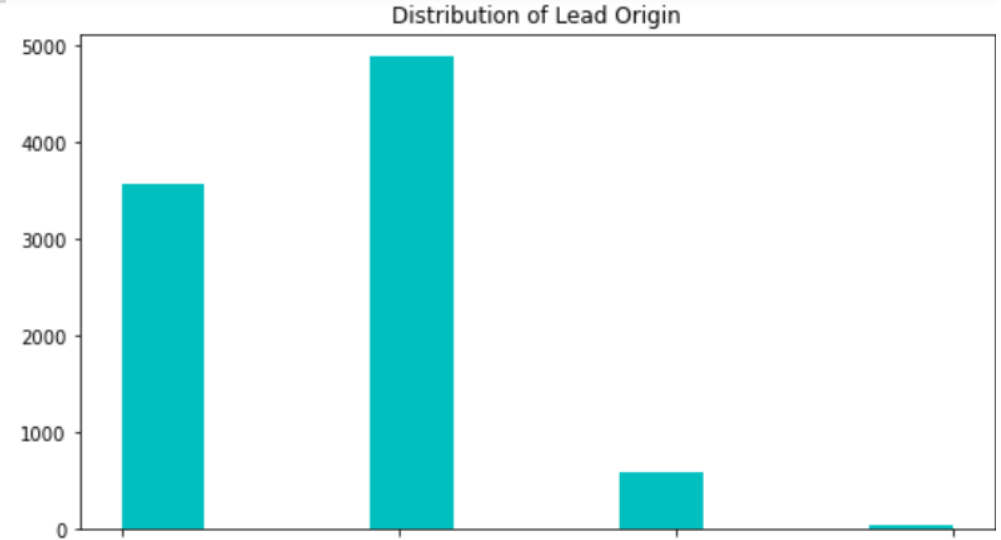
Process Overview

- The Process of building the logistic regression model include the following steps.
- Reading and Understand the data
- Data Cleansing
- Exploratory Data Analysis (EDA)
- Data Preparation
- Model Building
- Model Evaluation
- Prediction on Test set and calculating the lead scores

EDA(Exploratory Data Analysis)

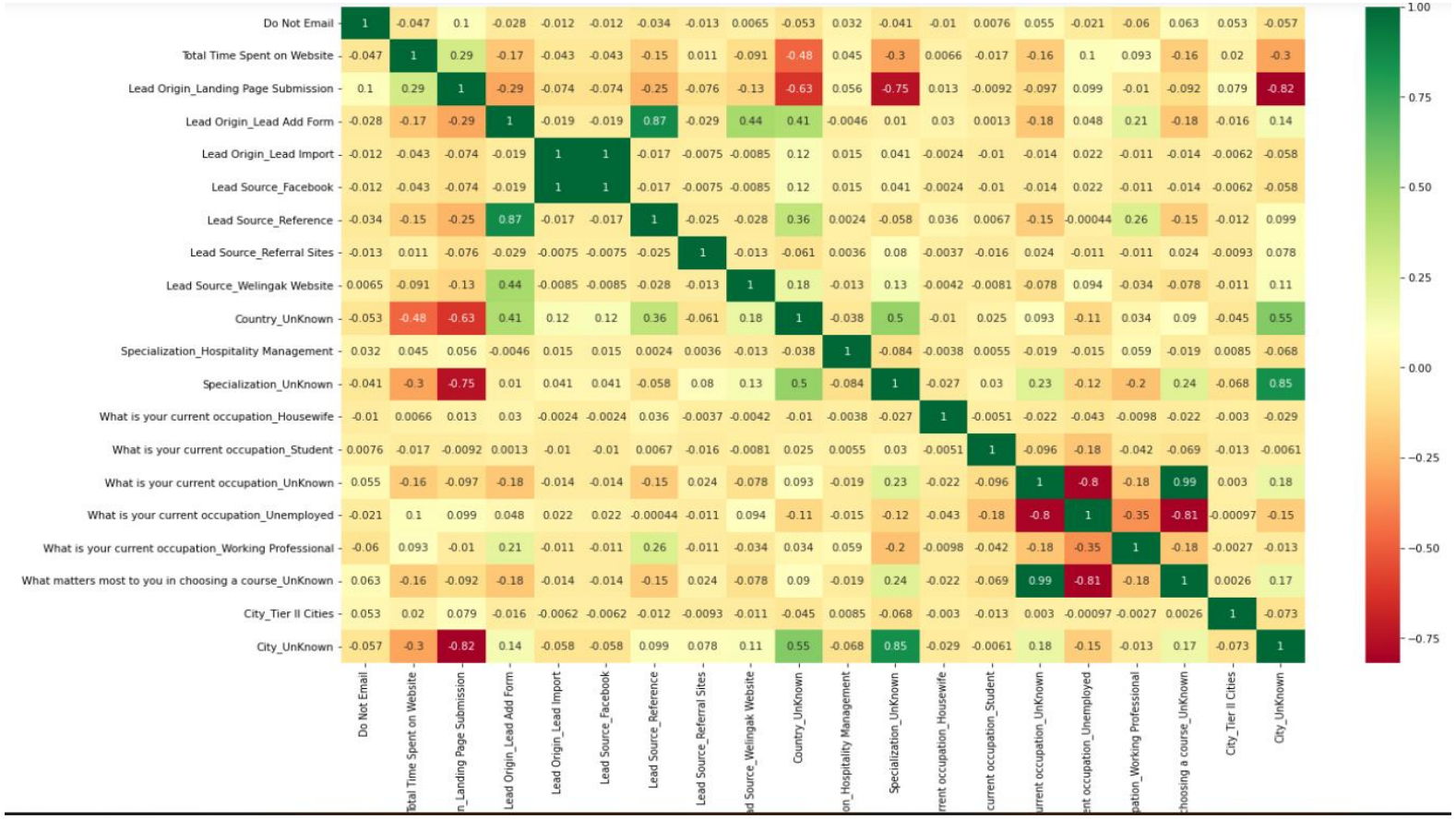






- Majority of the leads are coming from India, and very few are from outside India
- Google is being the major source of the leads
- People from IT and Finance specializations are more interested in taking courses, however people are unlike to provide their specialization while filling the form.
- Majority of the leads are coming in are from "Unemployed" as occupation, followed by working professionals
- Almost all the leads coming in are looking for "Better Career Prospects"
- Mumbai city is the major contributor of the leads

Logistic Regression Model:



	coef	std err	z	P>|z|	[0.025	0.975]
const	0.0087	0.119	0.073	0.942	-0.225	0.243
Do Not Email	-1.2538	0.167	-7.500	0.000	-1.581	-0.926
Total Time Spent on Website	1.1301	0.041	27.316	0.000	1.049	1.211
Lead Origin_Landing Page Submission	-0.8777	0.125	-7.035	0.000	-1.122	-0.633
Lead Source_Reference	2.2301	0.235	9.487	0.000	1.769	2.691
Lead Source_Welingak Website	5.3135	1.013	5.247	0.000	3.329	7.298
Country_UnKnown	1.2449	0.117	10.605	0.000	1.015	1.475
Specialization_UnKnown	-1.0377	0.123	-8.404	0.000	-1.280	-0.796
What is your current occupation_UnKnown	-1.2852	0.089	-14.450	0.000	-1.460	-1.111
What is your current occupation_Working Professional	2.0826	0.180	11.568	0.000	1.730	2.435

In [107]:
```
#Checking VIFs
VIF(col)
```

Out[107]:

	Features	VIF
5	Country_UnKnown	2.66
6	Specialization_UnKnown	2.18
7	What is your current occupation_UnKnown	1.58
3	Lead Source_Reference	1.40
2	Lead Origin_Landing Page Submission	1.36
1	Total Time Spent on Website	1.30
8	What is your current occupation_Working Profes...	1.19
0	Do Not Email	1.11

Evaluation Metrics

- At the optimal cut off of 0.28 we have the following metrics from the test set
  - accuracy = 76.6%
  - sensitivity = 86.2%
  - specificity = 70.8%
- The Conversion rates at the optimal cut-0ff point of 0.28 are as follows
  - Train set = 84.5%
  - Test set = 86.2%

Conclusions

- The top most 3 important features in the final model are,
  1. Lead Source_Welingak Website
  2. Lead Origin_Lead Add Form
  3. What is your current occupation_Working Professional
- The Sensitivity, Specificity scores on test set are very good, and are closer to the train set scores at optimal cut-off point.
- Hence the model is a very good model and in business terms, this model has an ability to adjust with the company's requirements in coming future.
- The conversion rate on the final predicted model is greater than 80% on both train and test sets, which is the target given by the CEO of the company