Lead Scoring Case Study

Submitted by
Muskan Chaudhary

Introduction

In today’s competitive market, businesses receive a high volume of leads, but not all leads convert into customers. Efficiently identifying high-potential leads is crucial for optimizing sales efforts and improving conversion rates.

This case study explores how we implemented a **data-driven lead scoring model** to prioritize leads based on their likelihood to convert. By leveraging historical data, behavioral patterns, and predictive analytics, we developed a structured approach to categorize and rank leads, enabling the sales team to focus on the most promising opportunities.

Problem Statement

An X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company 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. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’. The company requires you to build a model wherein need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%

Data set

Leads.csv

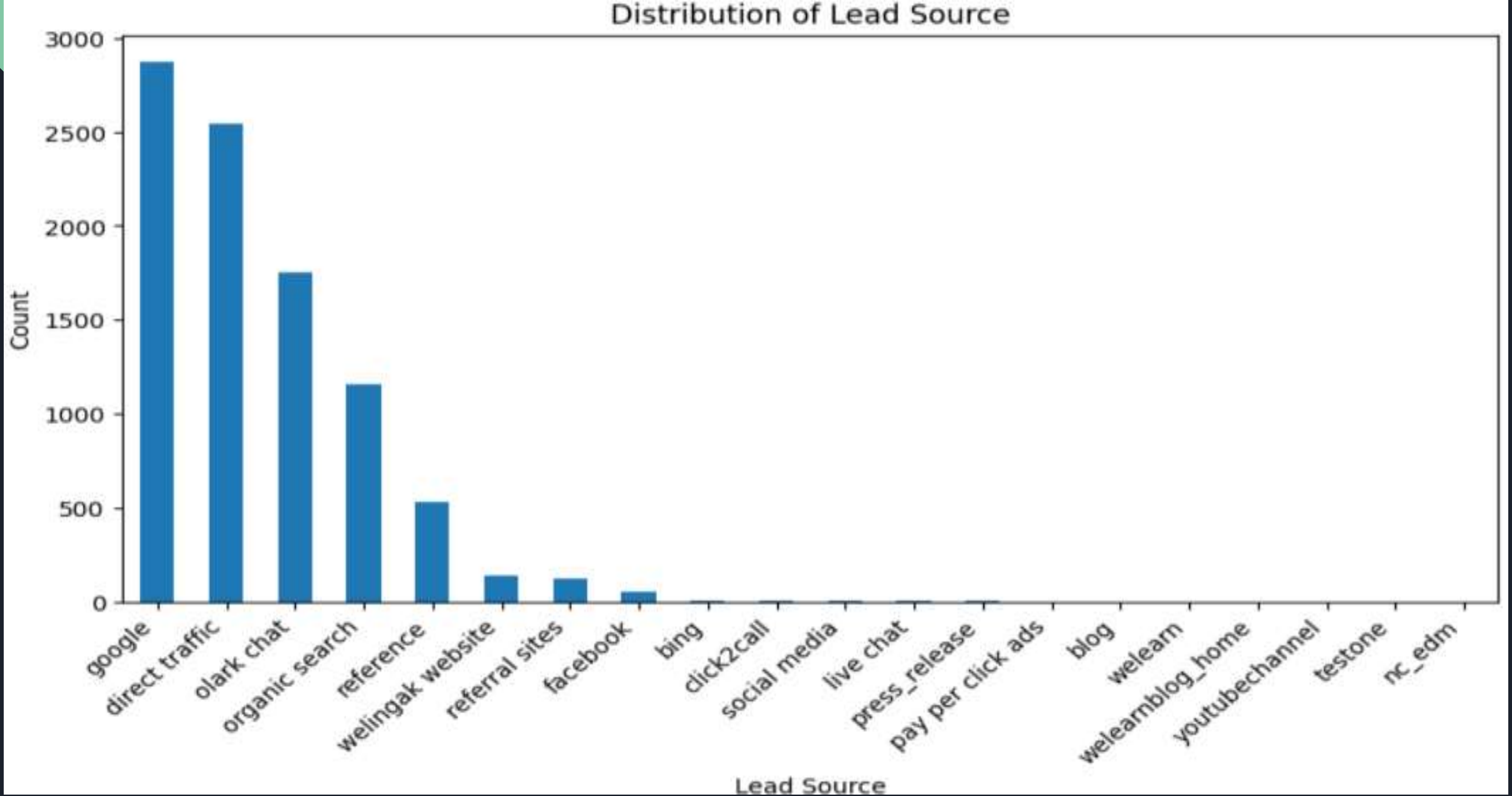
Libraries

1. Numpy
2. Pandas
3. Seaborn
4. Matplotlib
5. Sklearn
6. Statsmodel

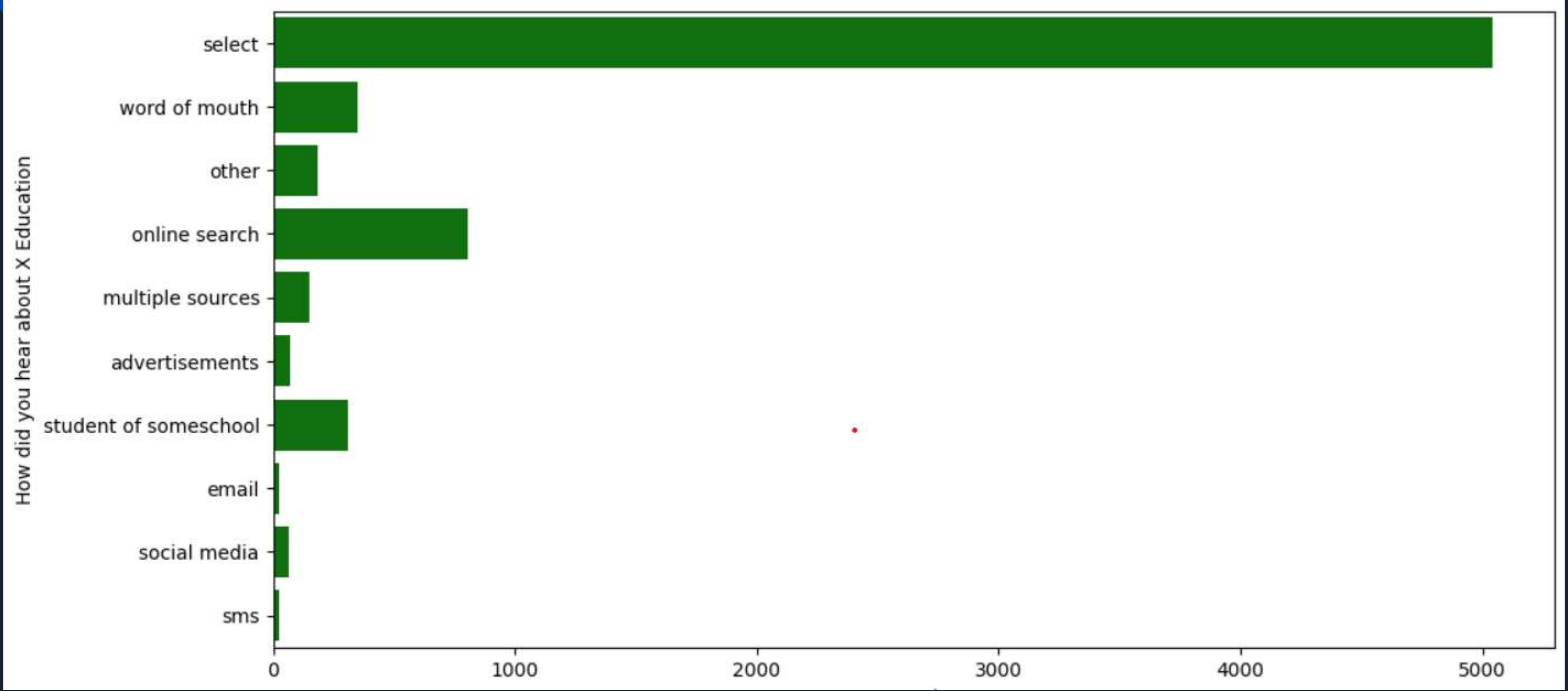
Problem Solving Methodology

1. Data Sourcing
2. Data Cleaning and Preparation
3. Data Visualization
4. Univariate / Bivariate Analysis
5. Feature Engineering
6. Splitting the data into Test and Train dataset
7. Building a Logistic Regression model and calculate lead score
8. Evaluating a model by using different metrics
9. Apply Best model in Test dataset

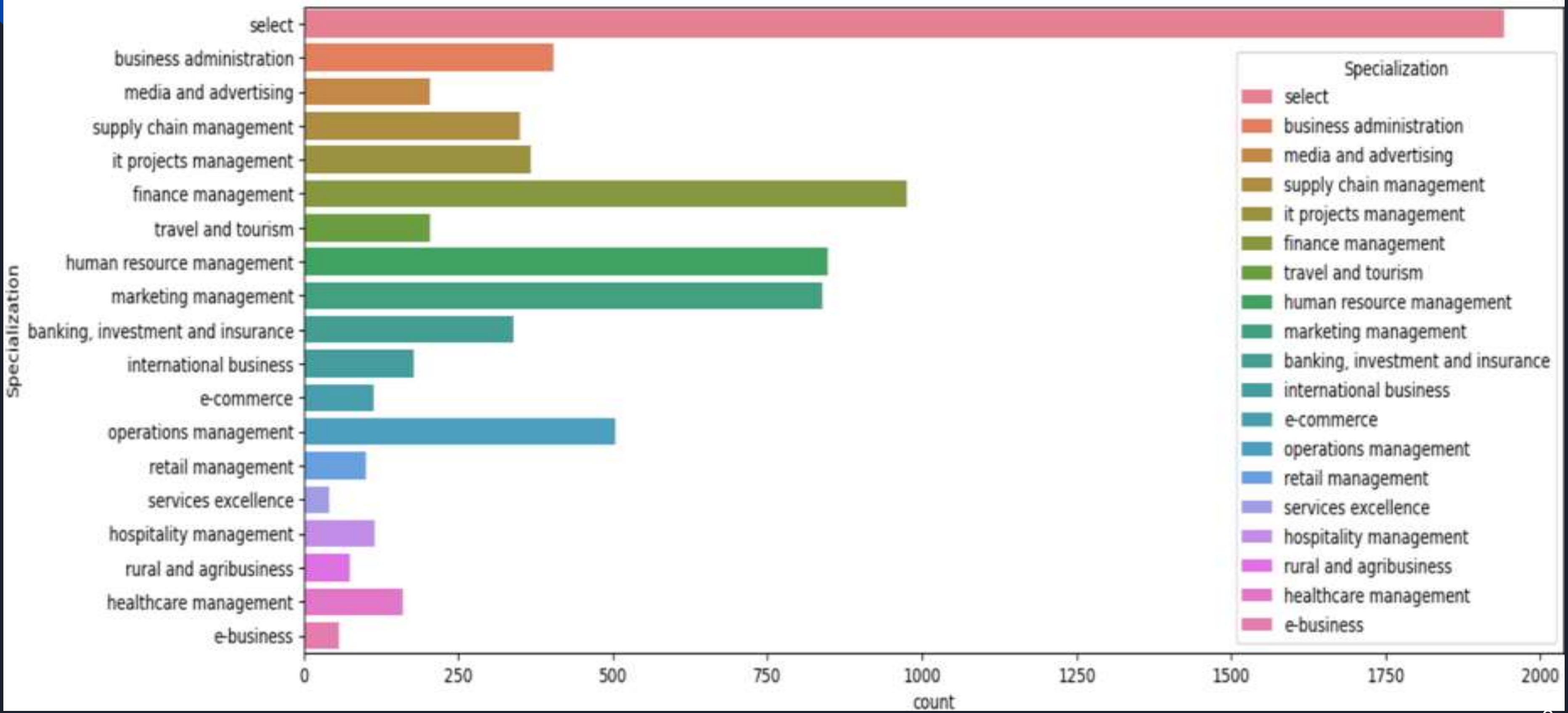
Data Vizualization



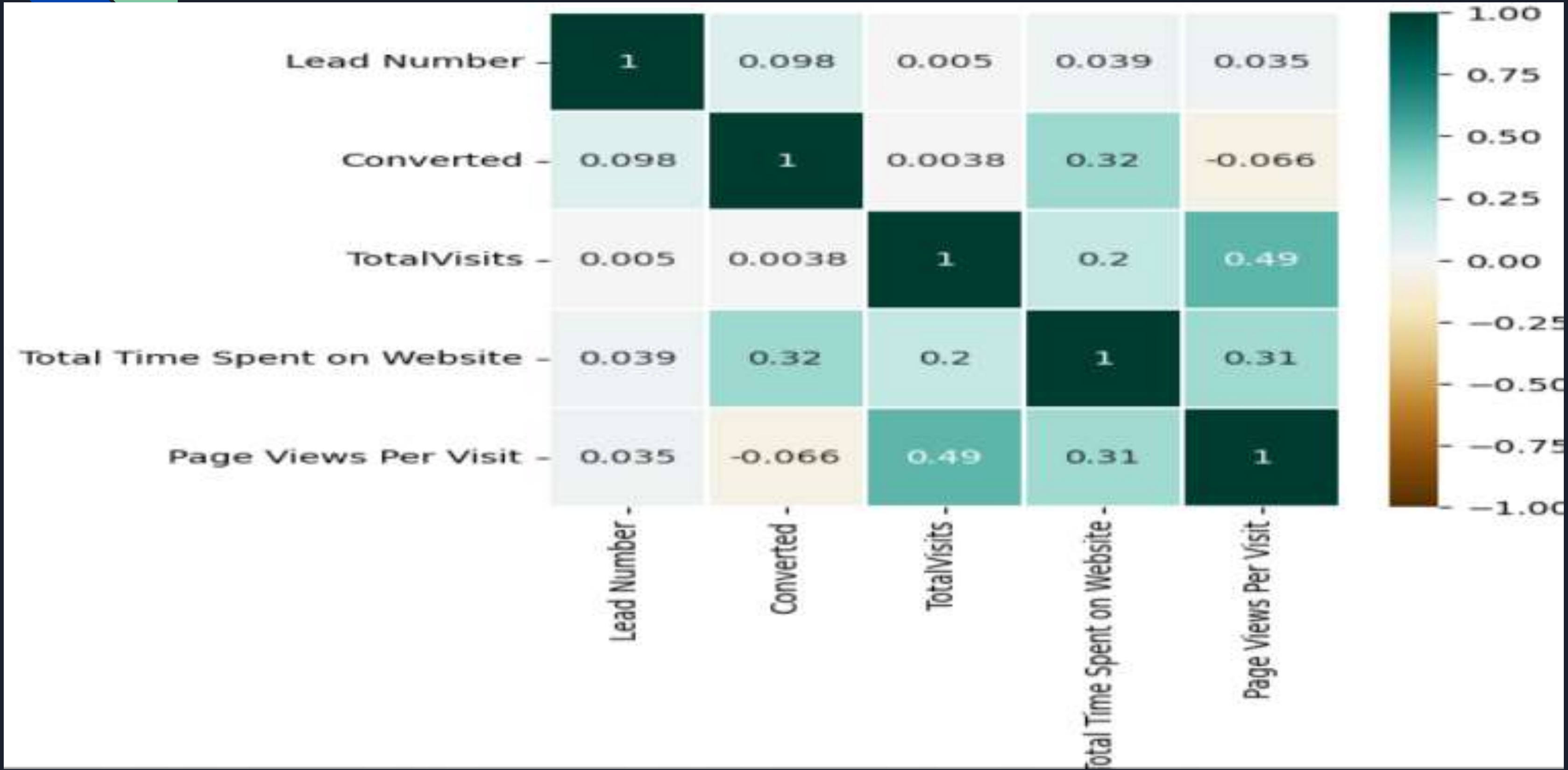
How did you hear about X_Education



Specialization



Correlation Matrix



	Lead Number	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit
Lead Number	1	0.098	0.005	0.039	0.035
Converted	0.098	1	0.0038	0.32	-0.066
TotalVisits	0.005	0.0038	1	0.2	0.49
Total Time Spent on Website	0.039	0.32	0.2	1	0.31
Page Views Per Visit	0.035	-0.066	0.49	0.31	1

Conclusion

1. While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
2. Accuracy, Sensitivity and Specificity values of test set are around 91.91%, 83.4% and 97.07% which are approximately closer to the respective values calculated using trained set.
3. lead score calculated shows the conversion rate on the final predicted model is around 92.05% (in train set) and 91.97% in test set
4. The top variables that contribute for lead getting converted in the model are :

   1. Total time spent on website

   2. What is your current occupation

   3. Lead Add Form from Lead Origin

   4. Had a Phone Conversation from Last Notable Activity

1. Hence overall this model seems to be good.

Thank You