Lead Scoring Case Study

Using Logistic Regression

Presented by Eddie Amaitum

Problem Statement

- X Education sells online courses to professionals through their website and marketing efforts.
- The challenge: Low lead conversion rates, with only around 30% of acquired leads becoming paying customers. They need a lead scoring model to assign scores based on conversion potential.
- The CEO's target is to increase the conversion rate to around 80% , improve efficiency and boost overall conversion rates.
- A model is required to help prioritize communication and engagement with potential customers who are more likely to convert.

Objectives

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

There are some more problems presented by the company which our model should be able to adjust to if the company's requirement changes in the future so we will need to handle these as well.

Data

- A leads dataset from the past with around 9000 data points.
- This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc.
- The target variable, the column ‘Converted’ tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn’t converted.
- A data dictionary was provided

Solution Approach

Inspecting and understanding the data

Data Cleaning and some EDA

Exploratory Data Analysis

Data Preparation

Model Building

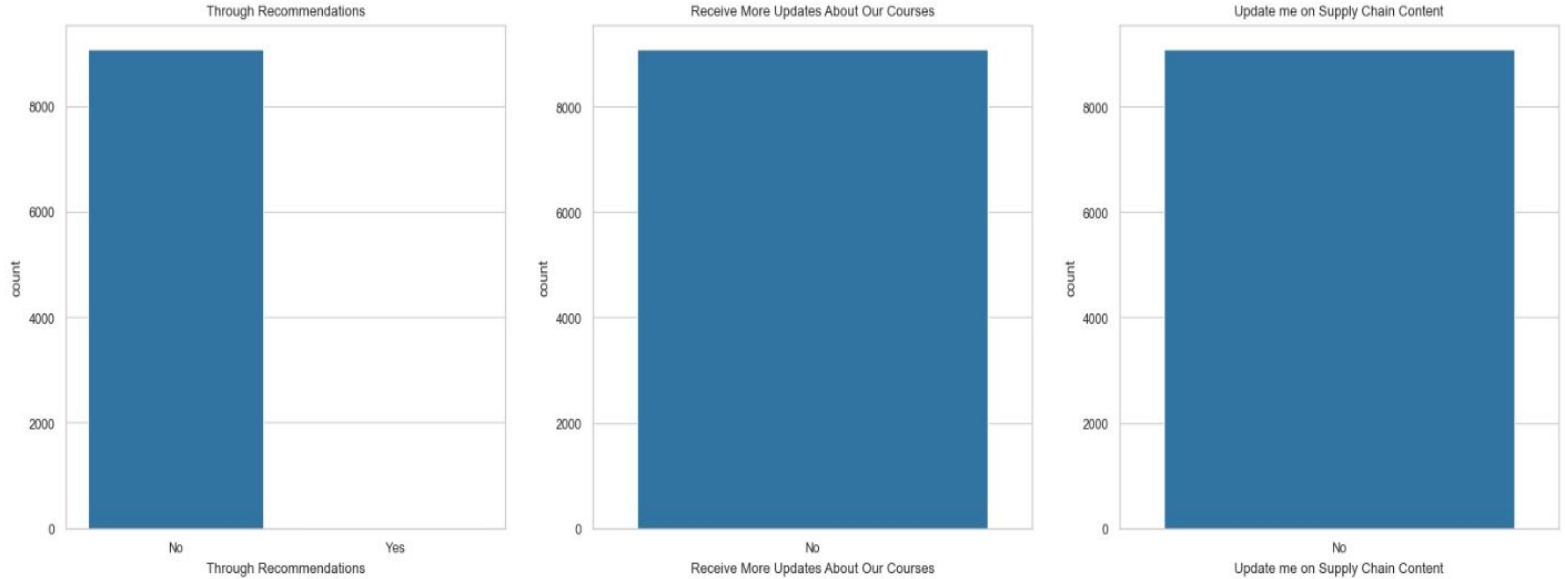
Model Evaluation

Making predictions on the test set

Calculating Lead Score

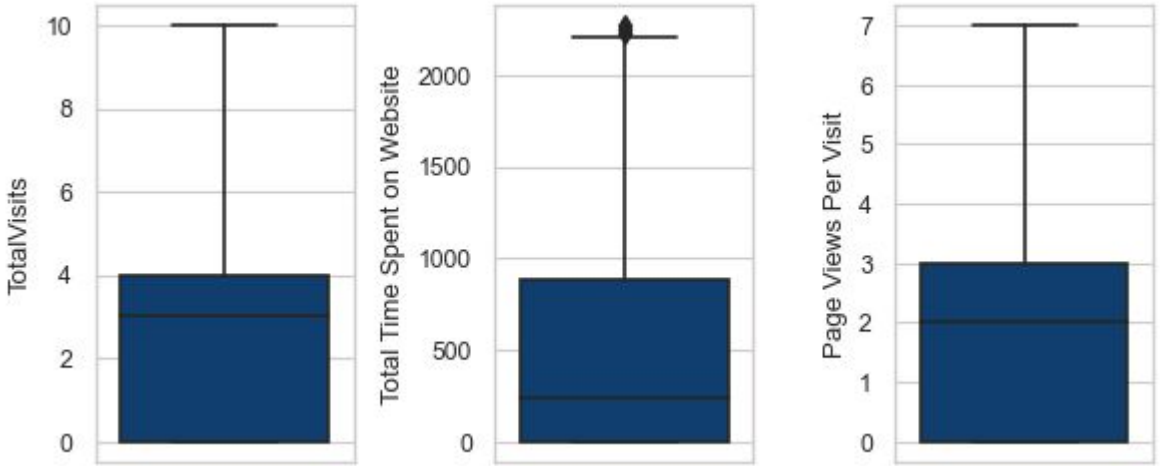
Determining Feature Importance

Select Visualizations



Skewed categorical features

Select Visualizations



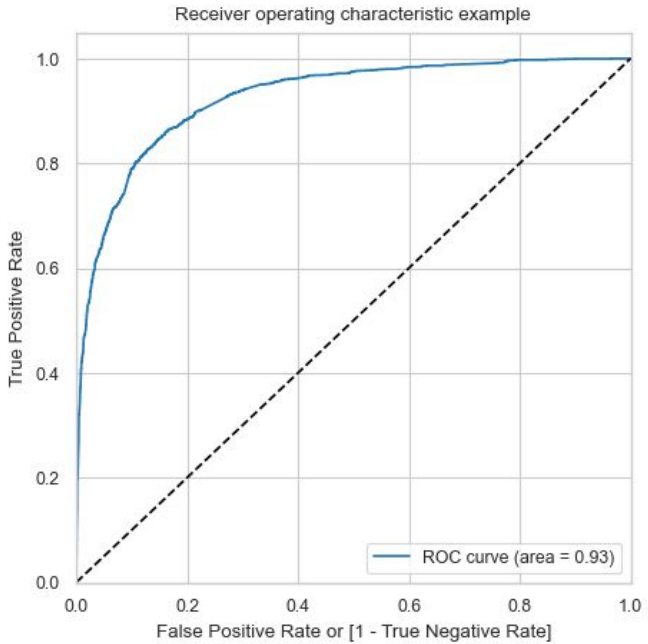
Numerical Variable Analysis

Insights on our recommended model

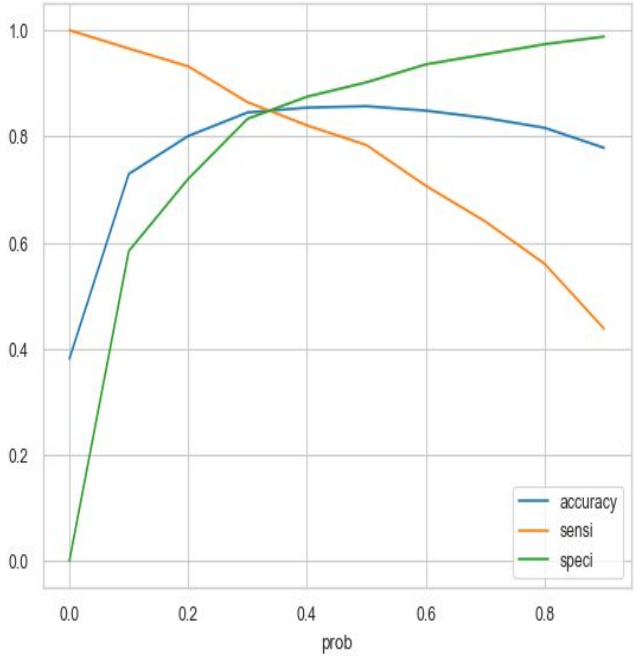
Having explored several models, our best model had the following characteristics:

- ➔ All p-values were less than 0.05 implying that our features were significant
- ➔ All VIFs were under 5 implying there was very low multicollinearity
- ➔ The model generalized very well on the test data set with test accuracy within less than 5% of the train accuracy

Model Evaluation

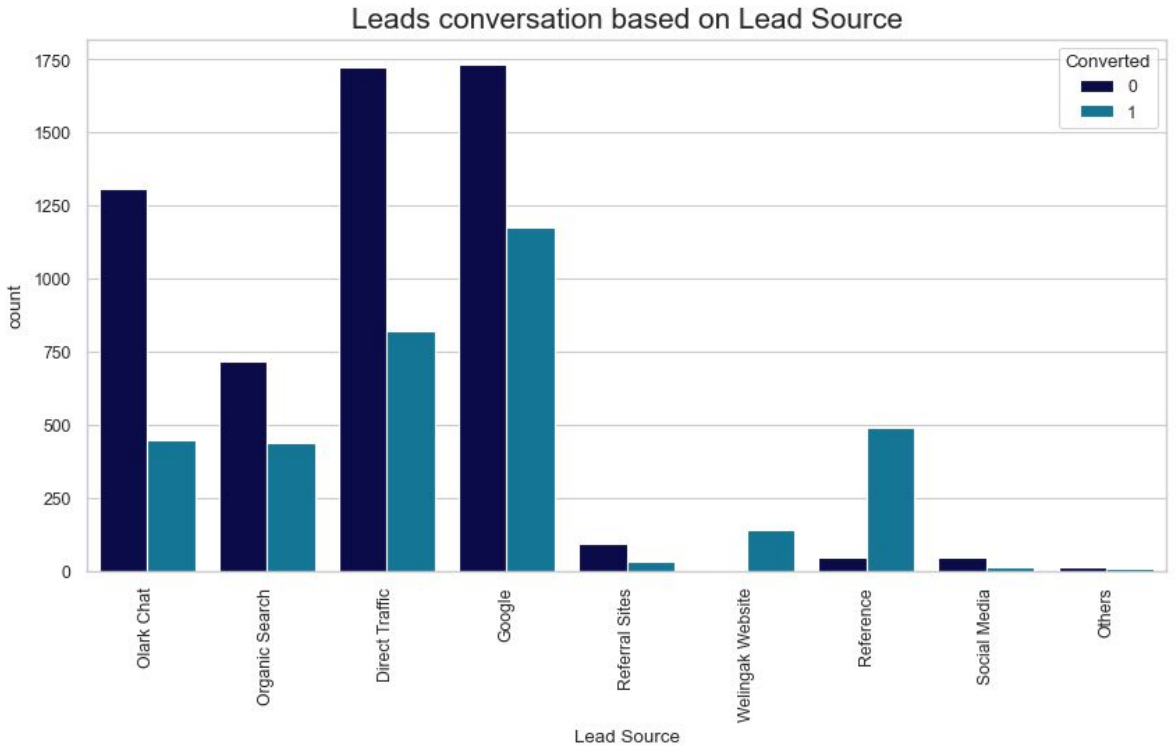


	precision	recall	f1-score	support
0	0.89	0.87	0.88	1636
1	0.78	0.82	0.80	944
accuracy			0.85	2580
macro avg	0.83	0.84	0.84	2580
weighted avg	0.85	0.85	0.85	2580



Business Recommendations

- Looking at the feature Lead Source,the majority of generated leads come from Google and Direct traffic while the least number of leads originate from Others
- The Welingak website exhibits a very high conversion rate, hence, it is advisable to maximize leads from this website



Business Recommendations

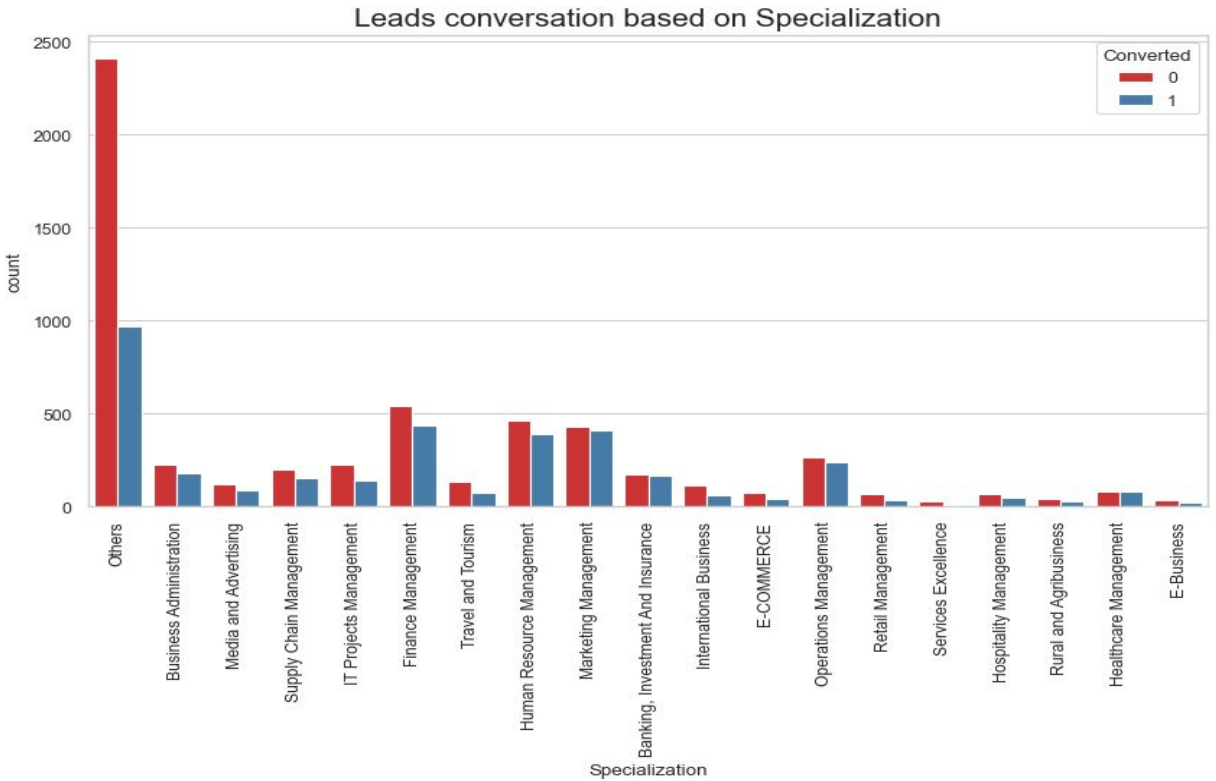
- The probability of lead conversion tends to increase as the values of the following features increase:
  - Lead Source_Welingak Website
  - Lead Origin_Lead Add Form
  - Lead Origin_Landing Page Submission
  - What is your current occupation_Working professional
  - Lead Source_Olark Chat
  - Lead Quality_High in Relevance
  - Total Time Spent on Website

Business Recommendations

- The probability of lead conversion tends to decrease as the values of the following features decrease:
  - Lead Quality_Worst
  - Lead Quality_Not Sure
  - Last Notable Activity_Email Link Clicked
  - Last Notable Activity_Modified
  - Last Notable Activity_Page Visited on Website
  - Last Notable Activity_Olark Chat Conversation
  - Last Notable Activity_Email Opened
  - Last Activity_Email Bounced
  - Do Not Email

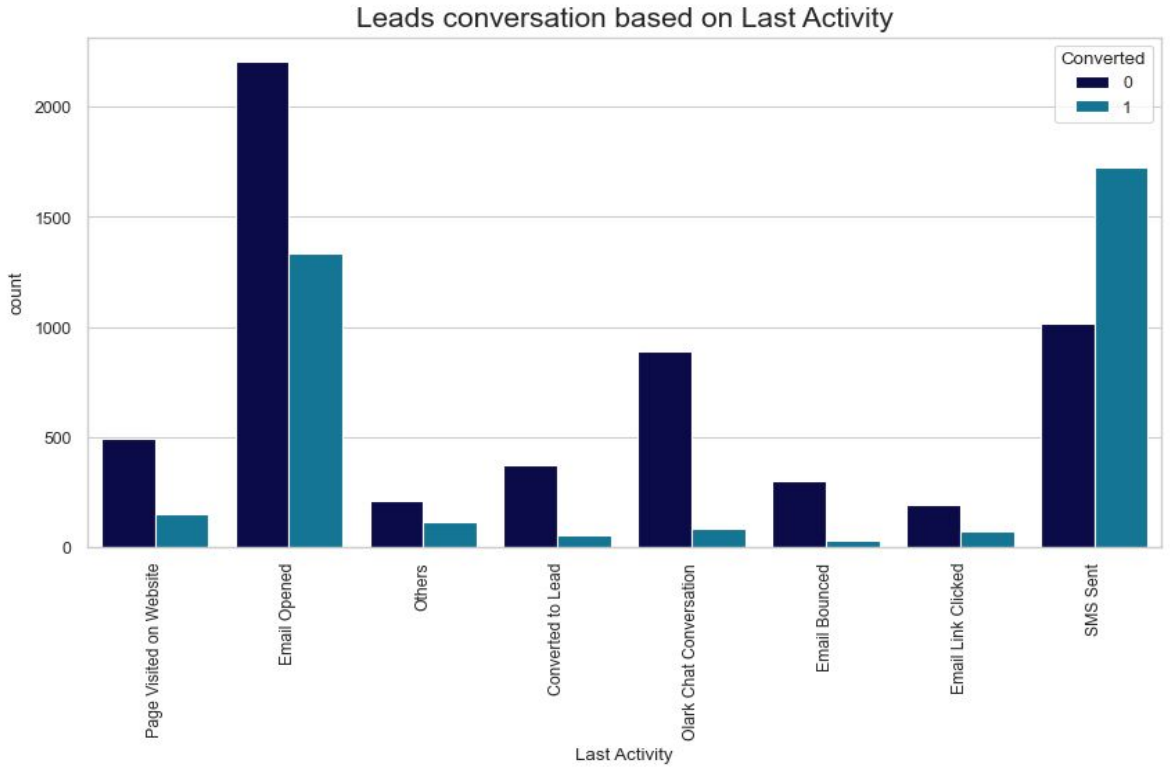
Business Recommendations

Management specialization is important as managers seem to be likely to convert favorably



Business Recommendations

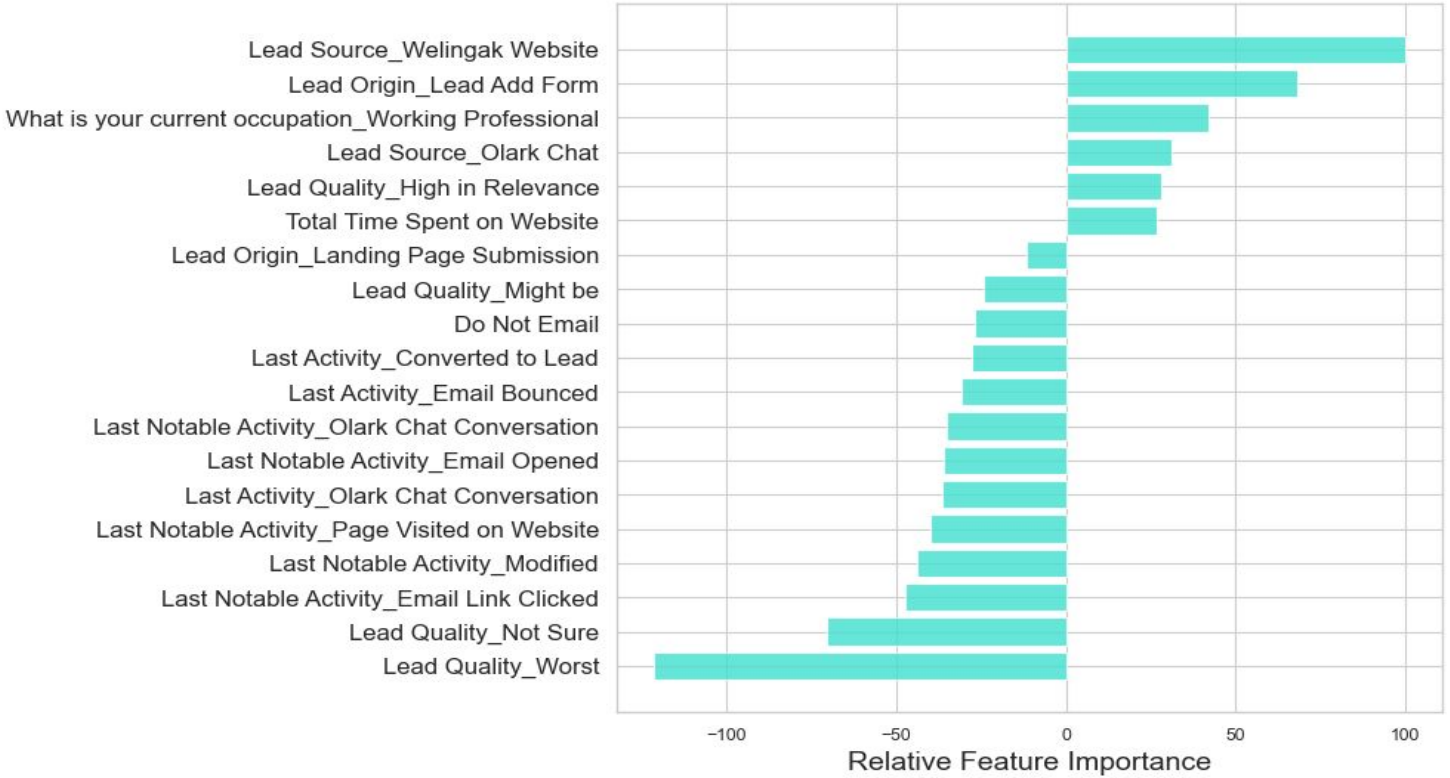
Looking at the feature Last Activity, SMS Sent has the highest success conversion rate followed by Email Opened



Lead Score

LeadID	Converted	Conversion_Prob	final_predicted	Lead_Score
0	0	0.76	1	76
1	0	0.23	0	23
2	1	0.86	1	86
3	0	0.06	0	6
4	1	0.55	1	55

Feature Importance by Magnitude



Looking Forward

It is important to note that, based on the business requirements, we have the flexibility to adjust the probability threshold value. Modifying this threshold value allows us to control the trade-off between sensitivity and specificity in the model. Increasing the threshold will decrease sensitivity but increase specificity, while decreasing the threshold will have the opposite effect, increasing sensitivity but decreasing specificity. This adjustment allows us to tailor the model's behavior to align with specific business needs and priorities.

Looking Forward

A high sensitivity value ensures that most leads who are likely to convert are correctly predicted as such. On the other hand, a high specificity value ensures that leads with borderline conversion probabilities are not falsely selected. In other words, high sensitivity focuses on minimizing false negatives (leads who should have been identified as likely to convert but were missed), while high specificity aims to minimize false positives (leads who are incorrectly identified as likely to convert). Balancing these two measures is crucial to achieve an optimal prediction outcome for lead conversion.

THANK YOU!