Lead Scoring Case Study

Shyam Sunder Balaji
Vaishnavee Kulkarni
Vivek Cholkar

Problem Statement

- On any given day, several people end up on the website of X Education. Every person who lands on the website is a potential lead.

- It is important to identify which such leads may be accidental / not-so-interested and which ones might actually buy a course.

- Based on characteristics, such as number of visits to the website, number of hours spent on the website, and lead occupation, we must construct a logistic regression model.

Assumptions and Approach

- Load and read the Leads.csv file.

- Check the data types and missing values and remove columns with over 3000 null values.

- ‘Select’ value present in various columns can also be assumed to be a null value.

- Remove columns with singular values and index-like columns which are not useful for the analysis.

Assumptions and Approach

- Perform data visualization and exploratory data analysis.

- For all the categorical variables, create dummy variables for logistic regression modelling.

- Use MinMaxScaler to standardize the continuous variables.

- Using test_train_split from the sklearn library, split the dataset into 70% training and 30% testing sets.

Assumptions and Approach

- Perform logistic regression modelling and remove non-significant features by checking the p-value and VIF.

- Use a preliminary conversion probability (0.5) and compute the confusion matrix and the relevant metrics, i.e. accuracy, sensitivity, specificity, precision, and recall.

- Check the area under the ROC curve to check goodness of model predictions; area should be closer to 1.

Assumptions and Approach

- Using a range of conversion probabilities (0.0-0.9), find the optimal sensitivity-specificity and precision-recall values.

- Apply the best model on the test set, and for the optimal conversion probability, compute and compare the previously mentioned metrics.

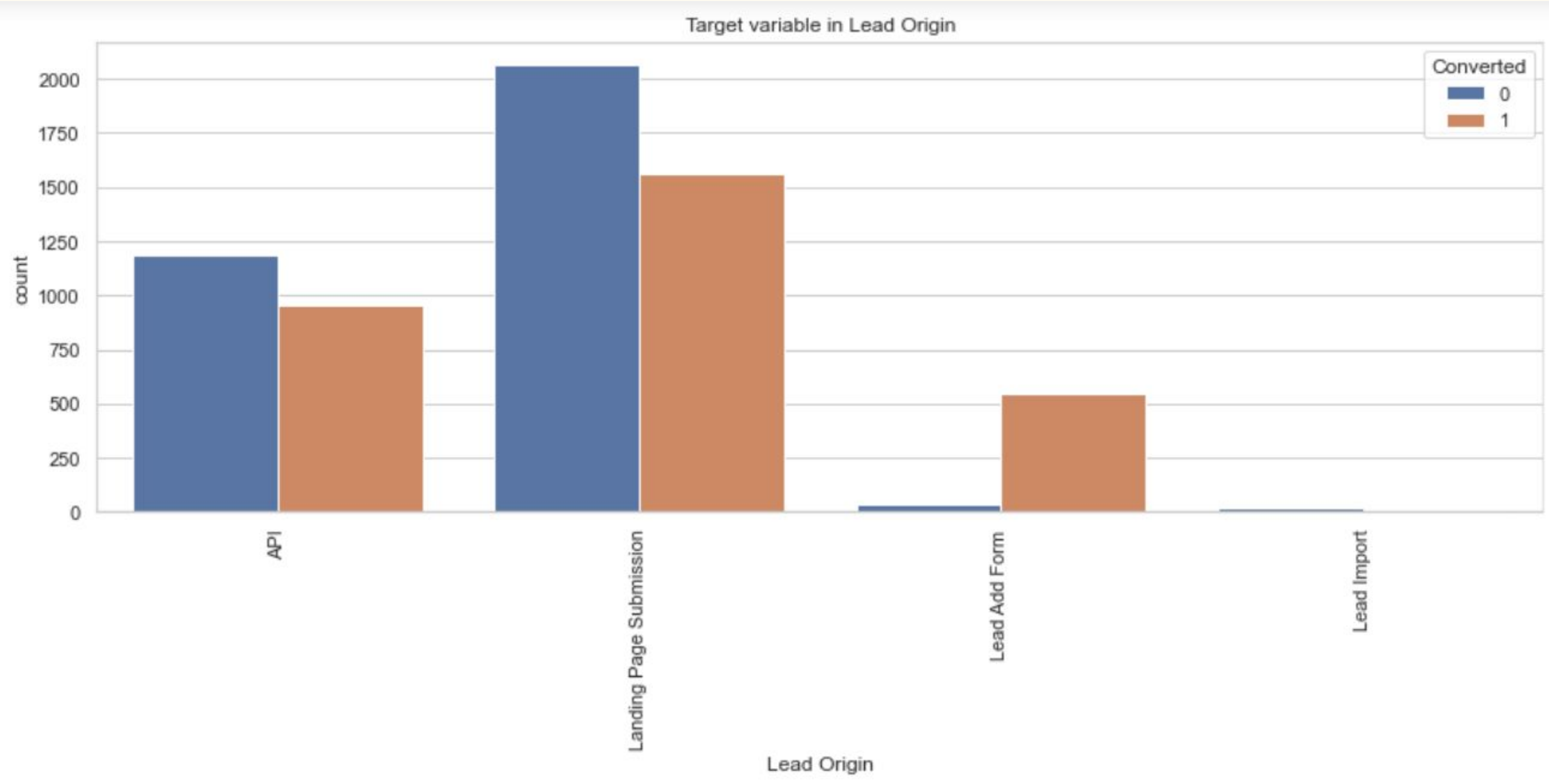
Data Analysis

Balance of Target Data

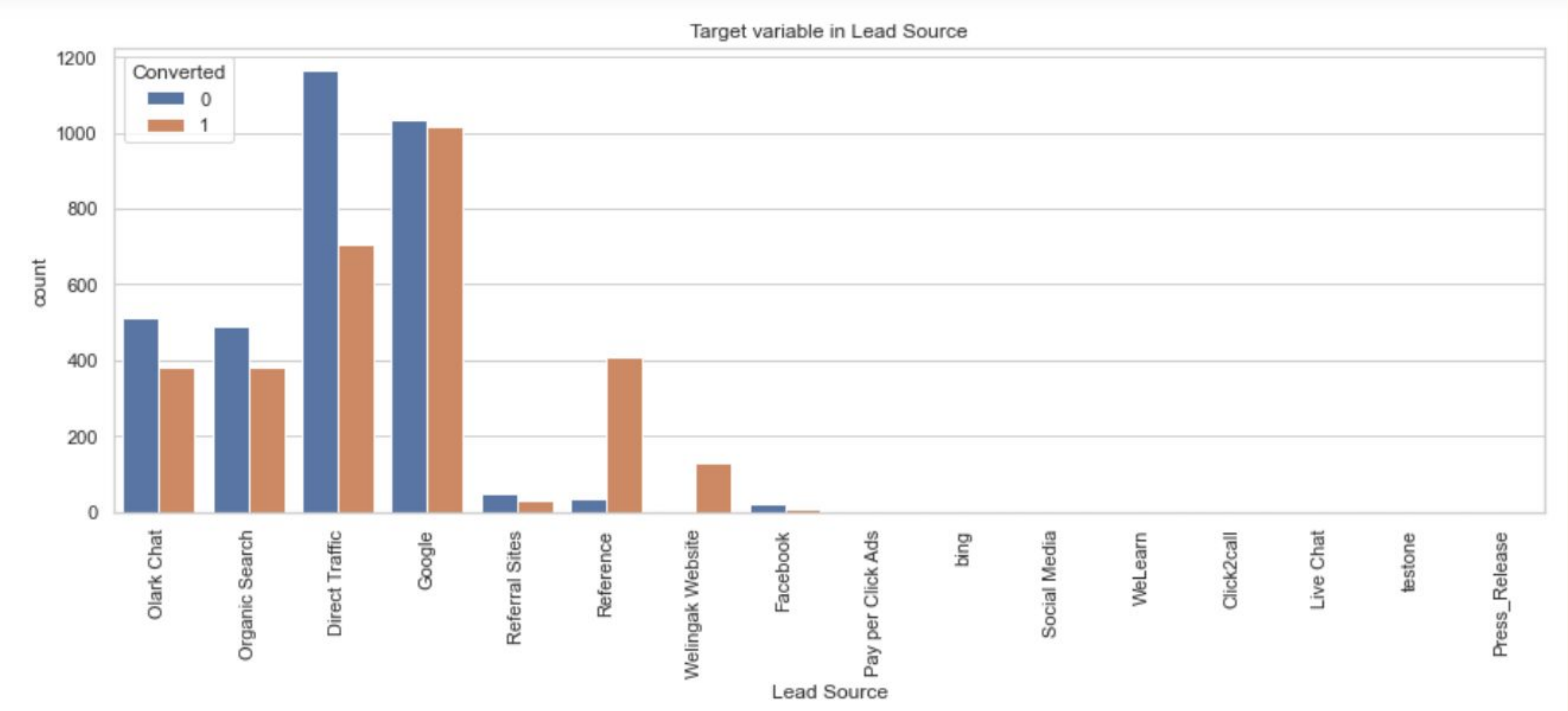
```
In [145]: #Check for data imbalance in the 'Converted' column

          leads.Converted.value_counts(normalize=True)

Out[145]: 0    0.519065
          1    0.480935
          Name: Converted, dtype: float64
```

- With a ~ 52-48 split, the converted column did not have a huge data imbalance.

Lead origin vs Converted

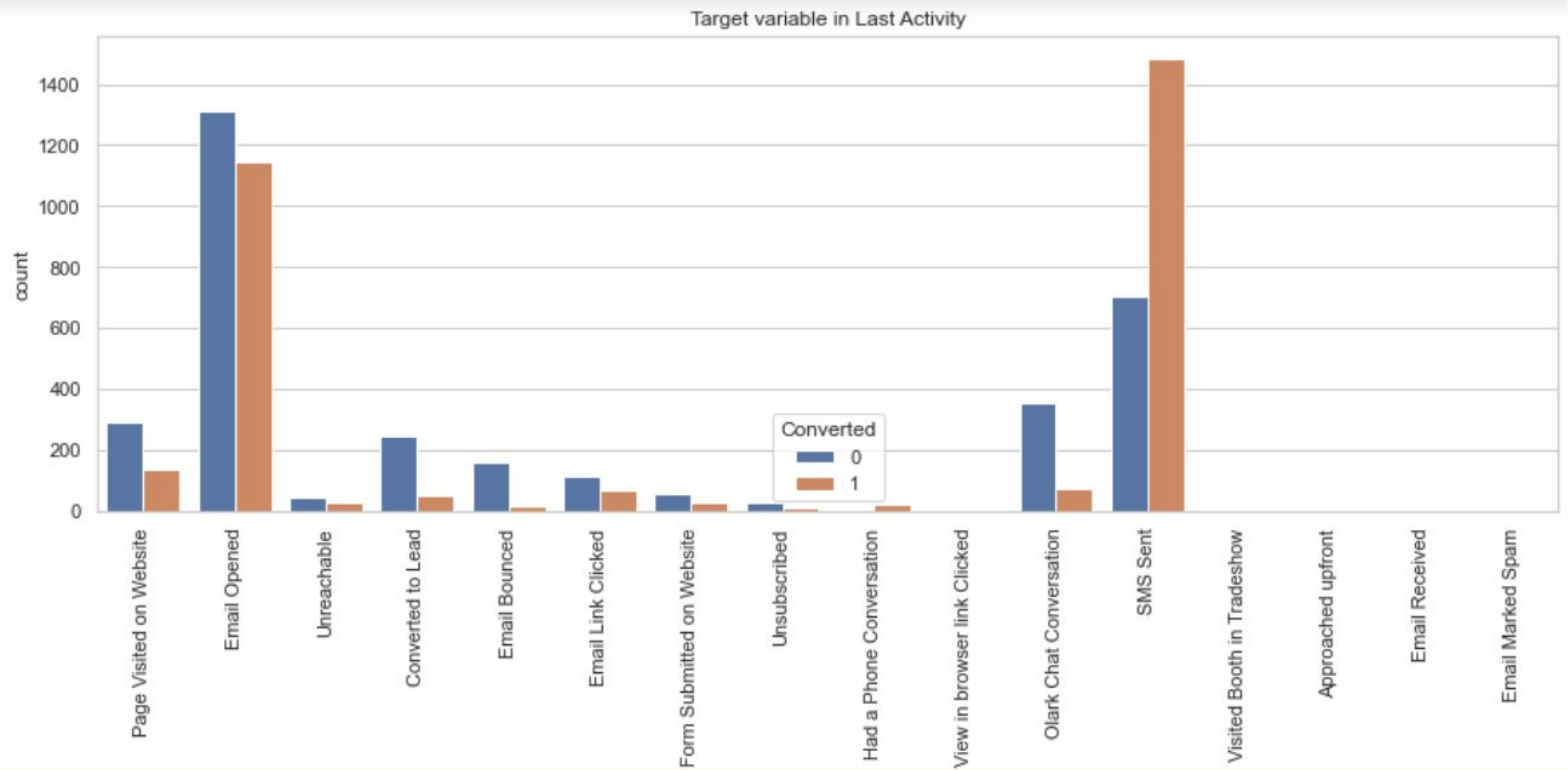


Lead source vs Converted



Lead source vs Converted

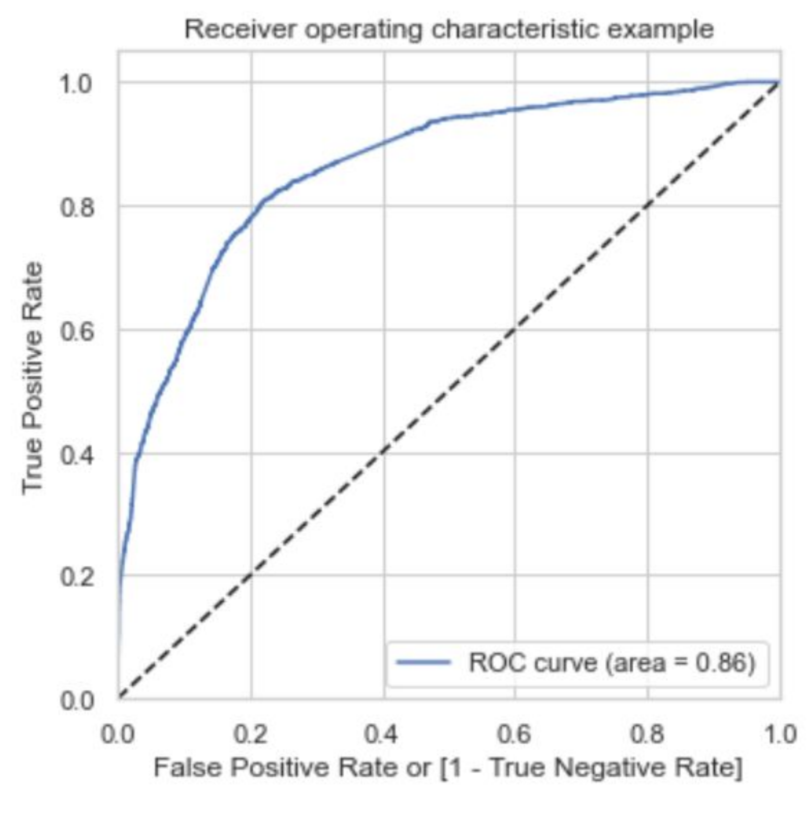
Last activity vs Converted



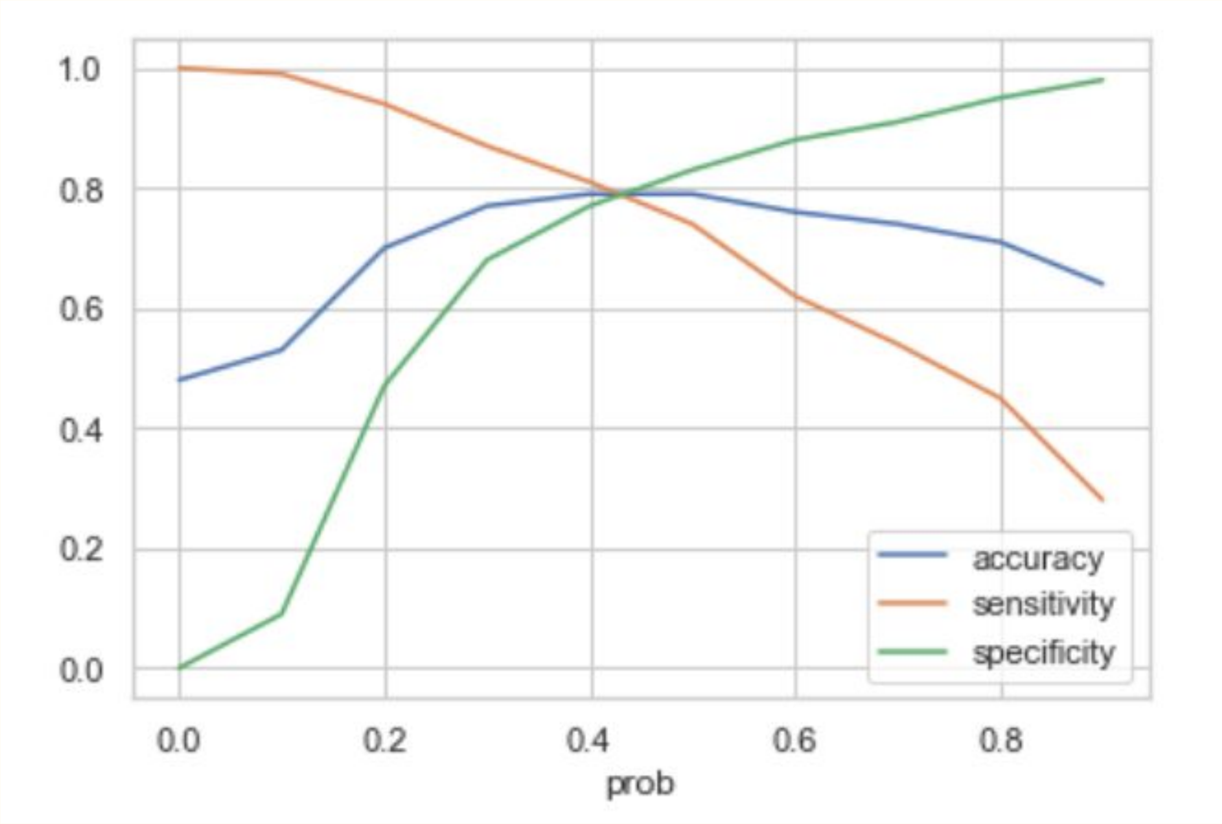
Final Feature Set

	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

ROC curve


Receiver operating characteristic example
True Positive Rate
False Positive Rate or [1 - True Negative Rate]
ROC curve (area = 0.86)
0.0
0.2
0.4
0.6
0.8
1.0

Precision-Recall curve



Conclusions

- We noted that the total time spent and the total number of visits to the X Education website are the two most important features.
- 1-3 visits may account accidental clicks but people visiting the website more than thrice and interact with the chatbot are more likely to be convertible leads.
- Such leads must be targeted appropriately through phone calls and emails.
- They should be notified of lean options, discounts, and referral bonuses.
- If they lead needs more convincing, they can be offered trial classes and short calls with successful alums instead of just marketing personnel.
- Unemployed leads and people looking for career transitions should also be focussed upon.
- Spending should be focussed on Google ads as they seem to bring in the most convertible leads.
- Follow-up calls should only be made after sufficient time in order to not alienate the potential customer.