Lead Scoring Case Study

Made by:

- Soumya Sengupta
- Dipanjan Barman
- Shahnawaz Mirza

Problem Statement

- Identify the set of leads of X Education so that the lead conversion rate should go up and the sales team of the company focus more on communication with the potential leads rather than making calls to every customer.
- To make the process more efficient, that company wishes to identify the most potential 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

Analysis approach

- Data understanding and exploring
- Data cleaning
  - Handling missing values
  - Outlier treatment
- Preparing the data for modeling
  - Handling data imbalance
  - Converting categorical variables
  - Data split into train and test set
  - Scaling the data
- Building model
  I. Train the model
  II. Test the model

I. **Train the model**

- Create the model with all the features.
- Removing the insignificant features by using RFE and manual approach (VIF and P-values).
- Predict the customers conversion with optimal probability cut off.
- Measure accuracy , sensitivity and specificity.
- Measure the Gini of the model.

II. **Test the model**

- Test the model using test data set with the final created model in train set.
- Measure accuracy, sensitivity and specificity in the set.
- Compare the above measure with the train set.

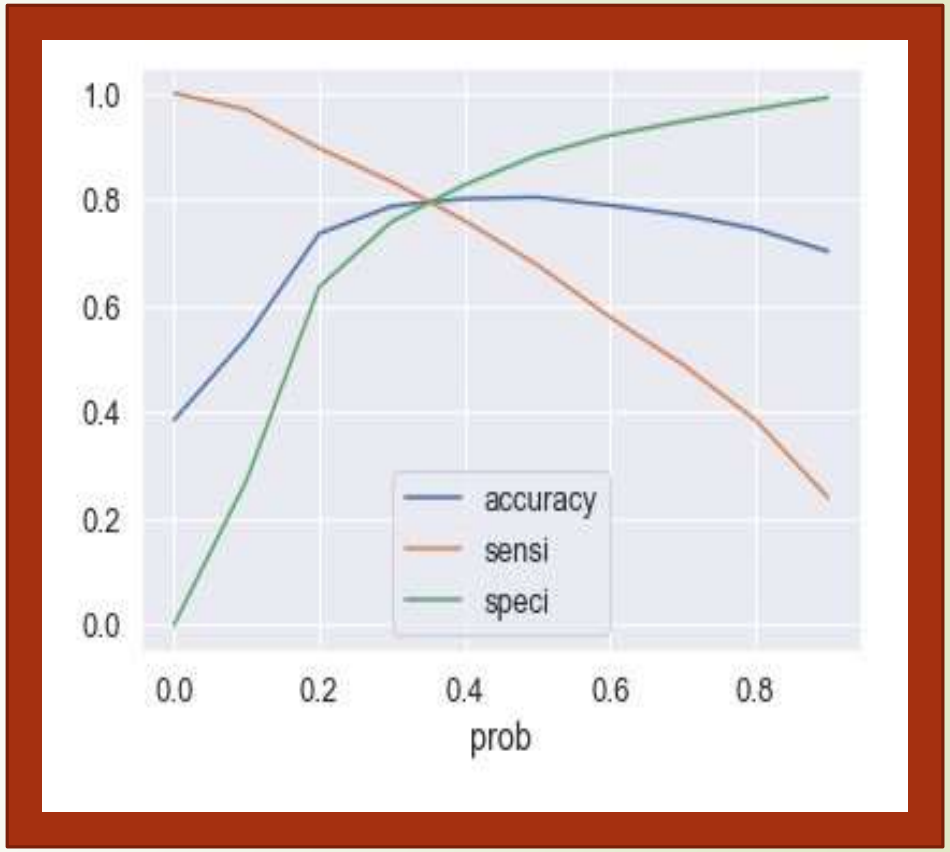
Model features

Features	Coefficient	Impacts (On decreasing order)
Total Time Spent on Website	0.9984	Positive
What is your current occupation	0.6162	Positive
Tags	0.5787	Positive
Lead Origin	0.4404	Positive
Last Notable Activity	0.4088	Positive
Lead Profile	0.4029	Positive
Lead Source	0.3169	Positive
City	0.2715	Positive
Last Activity	0.2433	Positive
TotalVisits	0.1728	Positive
Lead Quality	0.1643	Positive
Specialization	-0.0853	Negative
const	-0.2969	Negative
A free copy of Mastering The Interview	-0.4804	Negative
Page Views Per Visit	-0.5922	Negative
Do Not Email	-1.7634	Negative

Evaluation matrix

Analysis of the curve

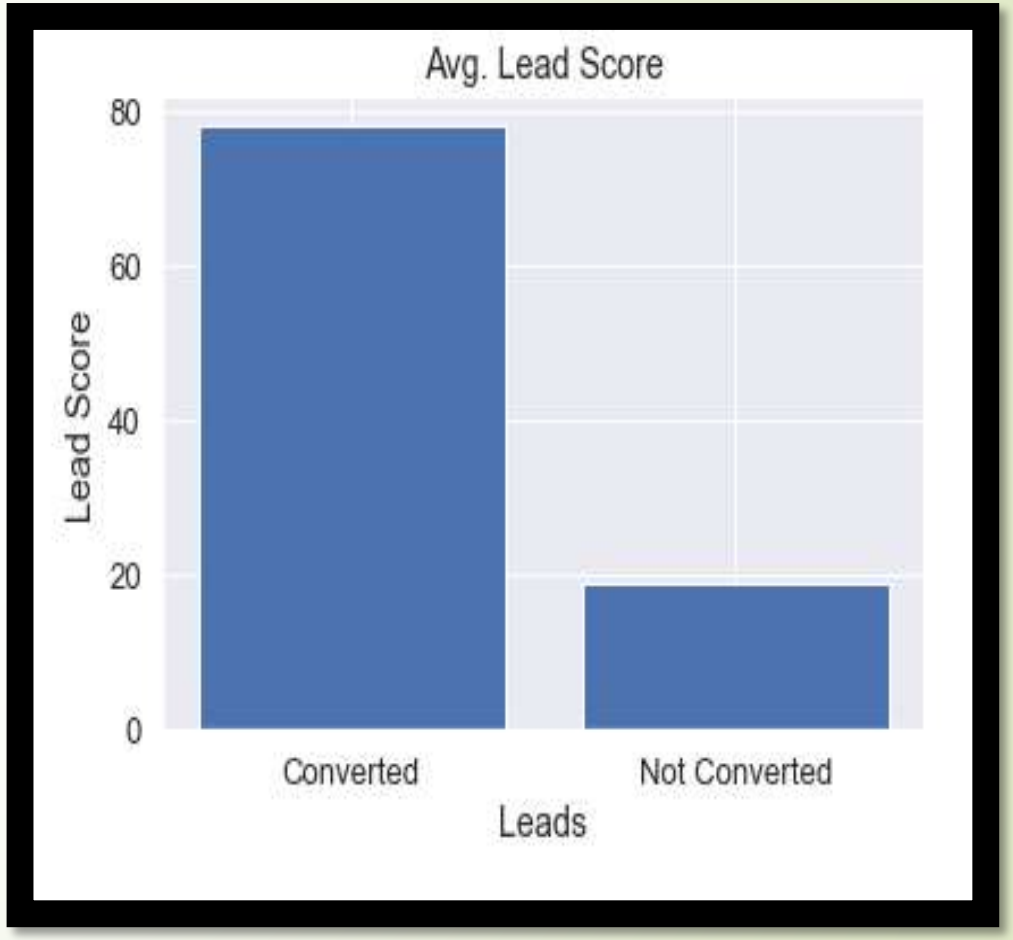
- *Accuracy* - Becomes stable after 0.35
- *Sensitivity* - Decreases with the increased probability.
- *Specificity* - Increases with the increasing probability.
- *At point 0.35* where the three parameters cut each other, we can see that there is a balance between sensitivity and specificity with a good accuracy.
- So, 0.35 may be the optimal cutoff.
- Predicted value with 0.35 probability cutoff



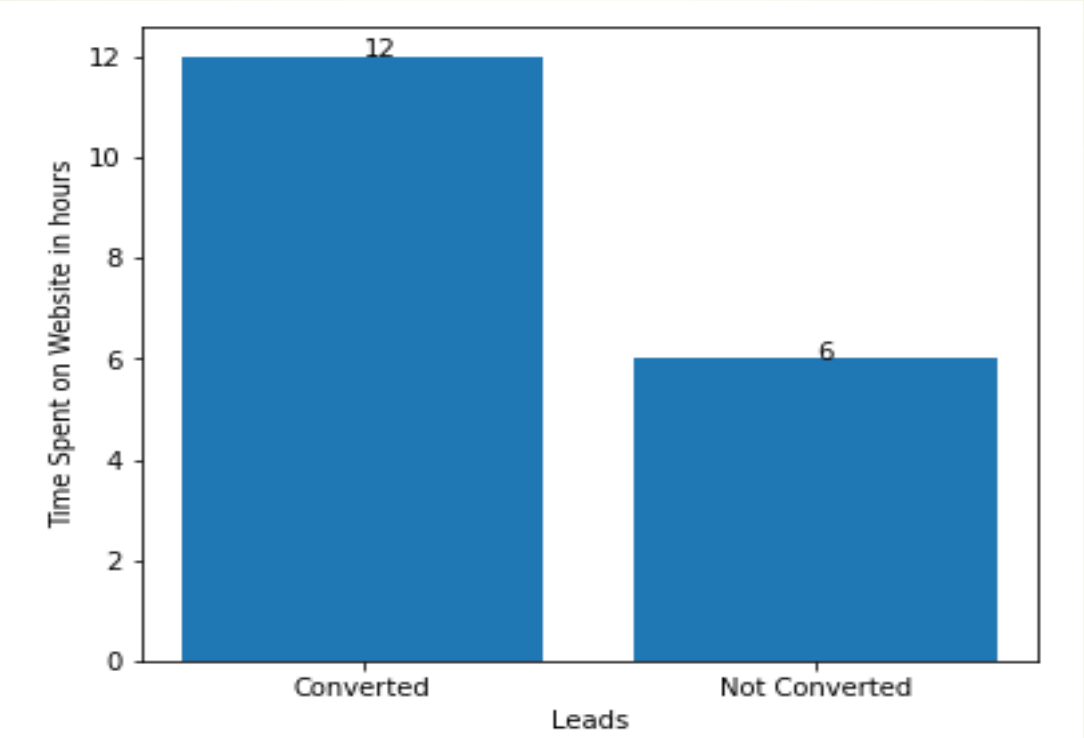
- **The optimal predicted probability for a customer being converted is 0.35.**
- **The model has good accuracy, sensitivity and specificity.**
- **There is a balance among accuracy, sensitivity and specificity.**

Lead Score

- The Lead Score, which ranges from 0 to 100, indicates the probability of the customer being converted or not.
- Higher the Lead Score, more the customers likely to be converted.
- The average Lead Score for the customers likely to be converted is 78.
- The average Lead Score for the customers not likely to be converted is 19.

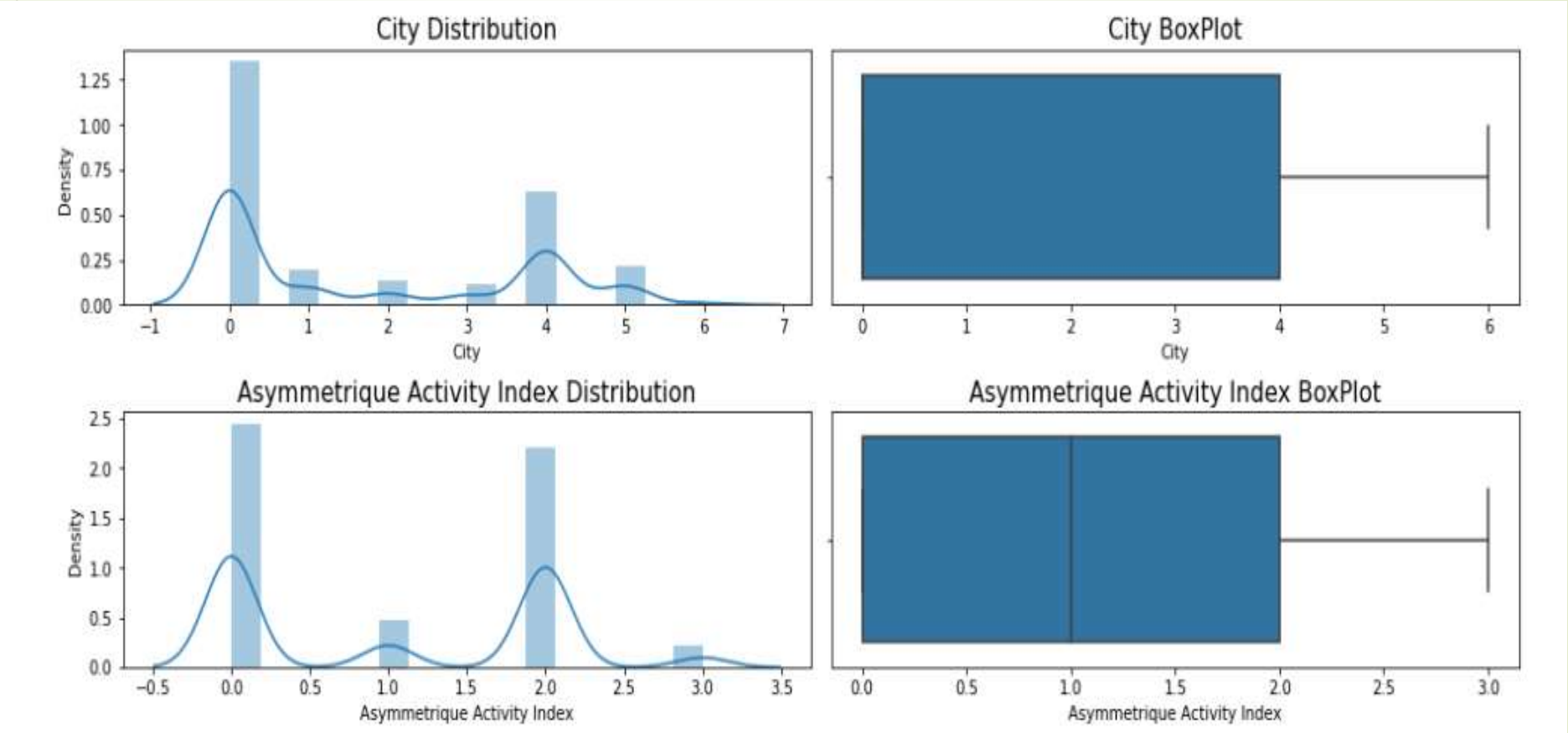


Average Total Time Spent on Website

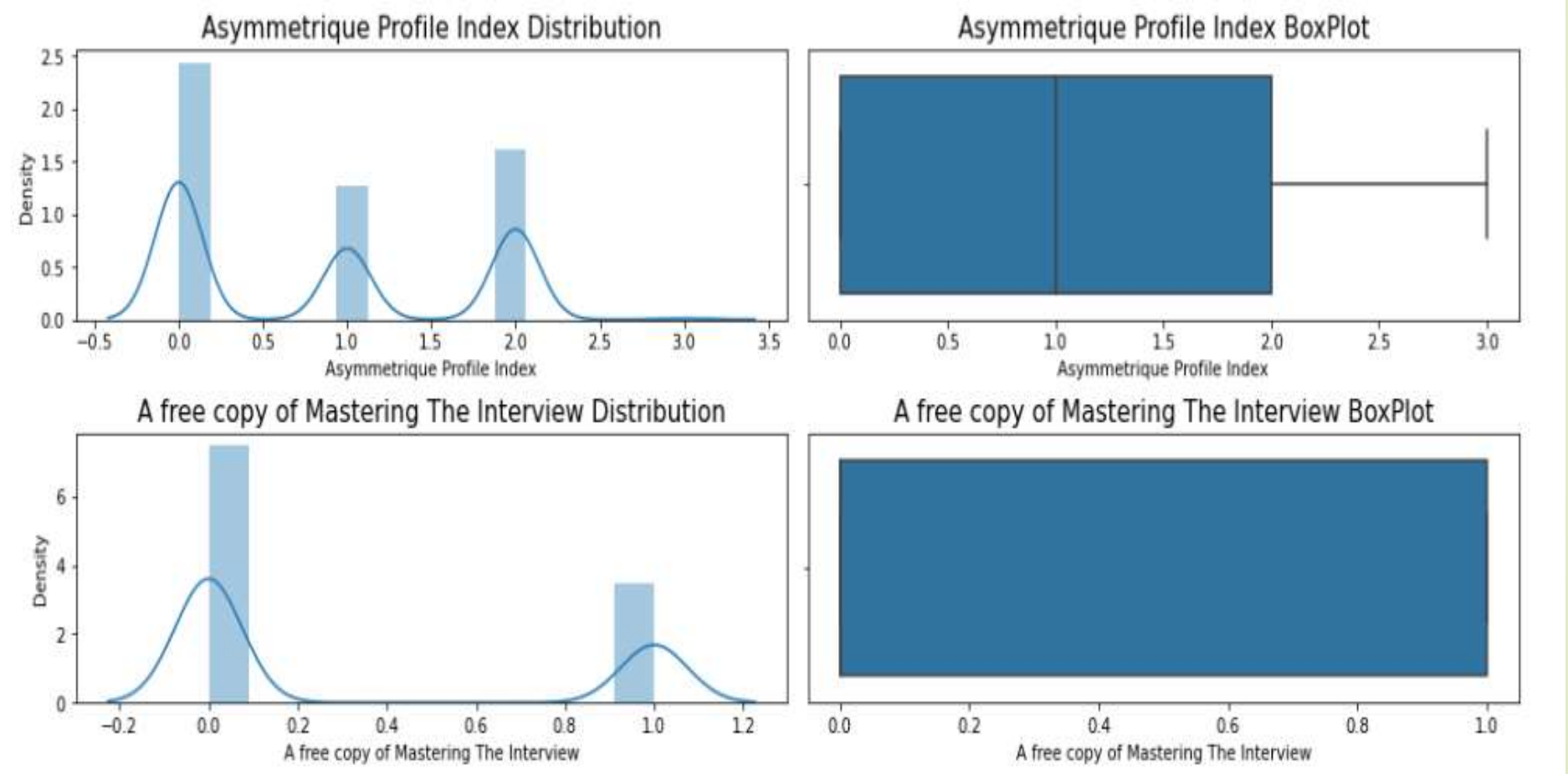


- The average total time spent on the website for the converted customers is 12 hrs.
- The average total time spent on the website for the not converted customers is 6 hrs.

Univariate Analysis



Univariate Analysis



Business recommendation for higher conversion rate

- Highly likely to be converted leads:-
  I. Total time spent on website more than 12 hrs.
  II. Lead score more than 78
  III. Lead source Welingak Website and Reference.
- Very less likely to be converted leads :-
  I. Customers opted for 'Do not email' option.
  II. Total time spent on website less than 6 hrs.
  III. Lead score less than 19.
  IV. Lead source Direct Traffic, Referral Sites, Organic Search and Google.
  V. Last activity of the customers is any of 'Olark chat conversation', 'page visited on the website', 'Email bounced', 'Form submitted on website', 'Email link clicked'.

Thank you!