Lead scoring case study

Submitted by :-

Sanchit Jain

Saniya rahil dhuka

Shashank Punde

Problem Statement

- An education company named 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. 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. 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, this means if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as Hot 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

Business Objective

- A model needs to be built using logistics regression such that we give every lead a score between 0-100, so that they can identify hot leads and increase their conversion rate.
- 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.
- The company has set the target lead conversion rate to be around 80%.

Solution Steps

1. Importing the data
2. Data Preparation
3. Exploratory data Analysis (EDA).
4. Dropping irrelevant columns and creating dummy variables.
5. Test-Train split
6. Feature Scaling
7. Model Building
8. Model Evaluation
9. Making Prediction on test set

Exploratory data analysis

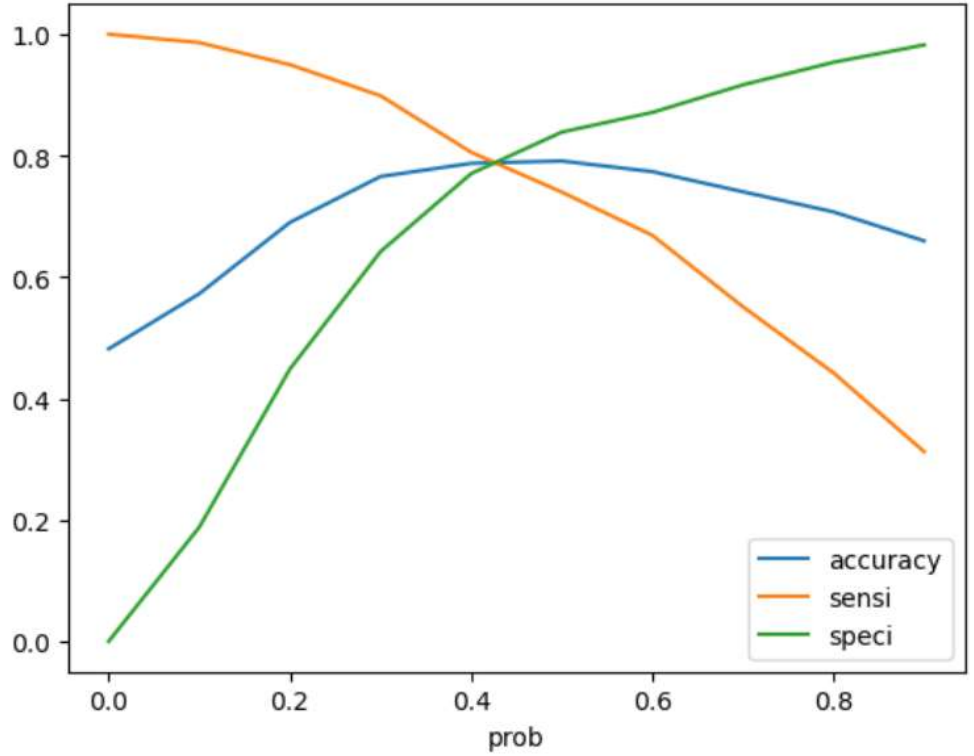
1. Dropping column ‘City’ and ‘Country’ as Majority of rows are from country ‘India’ and city “Mumbai’.
2. Dropping columns ‘Lead Profile’ and ‘How did you hear about X Education’ as they have ‘Select’ as value with very high number.
3. Dropping columns which has only value as NO in them – ‘Do Not Call', 'Search', 'Magazine', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations', 'Receive More Updates About Our Courses', 'Update me on Supply Chain Content', 'Get updates on DM Content', 'I agree to pay the amount through cheque’.
4. Creating Dummy variables for categorical values.

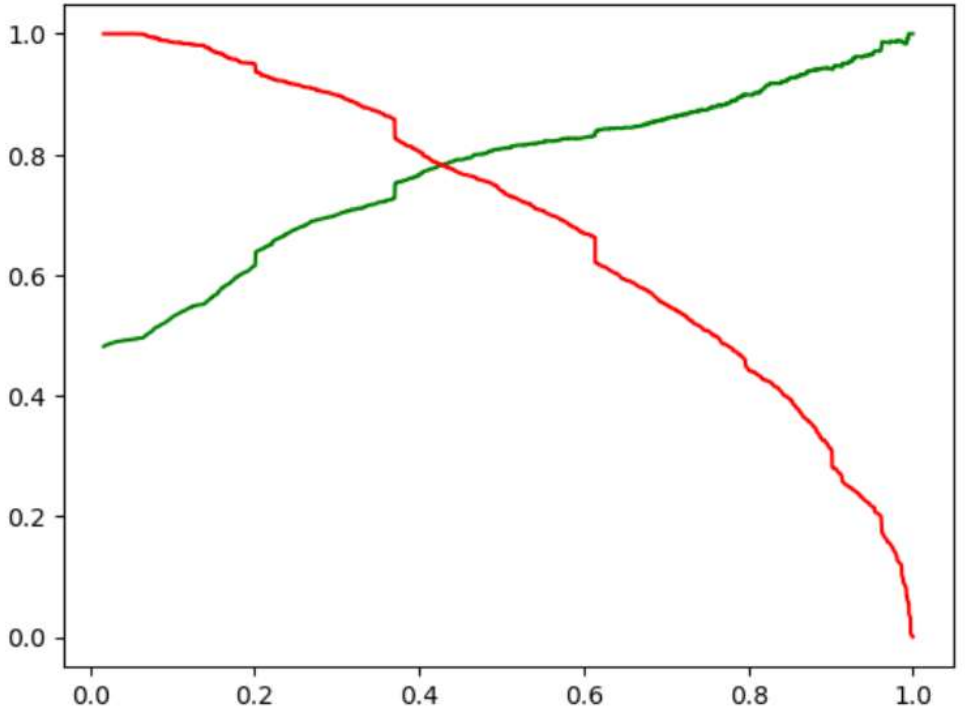
Correlation

- There is no correlation between the variables.

Model Evaluation

0.42 is the tradeoff between Precision and Recall therefore we can safely choose to consider any Prospect Lead with Conversion Probability higher than 42 % to be a hot Lead.





Model Observation

Train Data :-

Accuracy :- 79.0

Sensitivity :- 73.9

Specificity :- 83.8

Test Data :-

Accuracy :- 78.9

Sensitivity :- 78.8

Specificity :- 79.0

Final feature list :-

Total Time Spent on Website
Lead Origin_Lead Add Form
Lead Source_Olark Chat
Lead Source_Welingak Website
Do Not Email_Yes
Last Activity_Had a Phone Conversation
Last Activity_SMS Sent
What is your current occupation_Student
What is your current occupation_Unemployed
Last Notable Activity_Modified
Last Notable Activity_Unreachable

Conclusion

- Based on the final list of features that we can conclude that Leads who spent more time on website, more likely to convert.
- Lead Source is an important feature which should be focused upon of which 3 factors can be focused upon like add form, olark chat and welingak website.
- Leads which had a phone conversation are most likely to get converted into lead, therefore sales team should focus on making phone calls
- SMS sent also contribute in converting into a hot lead.
- Max conversion is with working professional therefore sales team should not focus on students or unemployed customers