Lead Scoring Case Study using logistic regression

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. They have a process of form filling on their website after which the company identifies that individual as 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

- Lead X wants us to build a model to give every lead a lead score between 0-100 so that they can identify the Hot leads and increase their conversion rate as well.

  The CEO wants to achieve a lead conversion rate of 80%.

  They want the model to be able to handle future constraints as well, like peak time actions required, how to utilize full manpower, and what should be the approaches after achieving the target.

Problem Approach

- 1. Importing the data and inspecting the data frame
  2. Data preparation
  3. EDA
  4. Dummy variable creation
  5. Test-Train split
  6. Feature scaling
  7. Correlations
  8. Model Building (RFE, R-squared, VIF, and p-values)
  9. Model Evaluation
  10. Making predictions on the test set

EDA and Correlation

- Leads from HR, Finance & Marketing management specializations have a high probability of conversion.

  Lead Source & Lead origin: In lead source, leads through Google & direct traffic have a high probability of conversion. In Lead origin, most leads land on submission.

  Leads opening emails have a high probability of conversion. Similarly, sending SMS will also benefit.

  Leads that are unemployed are more interested in joining the course than others.

Model Evaluation

- Final Features list:
  - Lead Source_Olark Chat
  - Specialization_Others
  - Lead Origin_Lead Add Form
  - Lead Source_Welingak Website
  - Total Time Spent on Website
  - Lead Origin_Landing Page Submission
  - What is your current occupation_Working Professionals
  - Do Not Email

  Train Data: Accuracy: 80%, Sensitivity: 77%, Specificity: 80%

  Test Data: Accuracy: 80%, Sensitivity: 77%, Specificity: 80%

Observations

- 1. Conversion rate is 30-35% (close to average) for API and Landing page submission but very low for Lead Add form and Lead import. We need to focus more on leads originating from API and Landing page submission.

  2. The maximum number of leads is generated by Google/direct traffic. The maximum conversion ratio is by reference and Welingak website.

  3. Leads who spent more time on the website are more likely to convert.

  4. Most common last activity is email opened. Highest rate = SMS Sent. Max are unemployed. Max conversion with working professionals.

Conclusion

- We conclude that focusing on leads from API and landing page submissions, especially those who spend more time on the website, can improve conversion rates. Additionally, targeting unemployed individuals and those in HR, Finance & Marketing management specializations should be prioritized.