Lead Score Case Study

Submitted by:

Gaurav Kumar

Swapna Pemme

Contents:

- Problem Statement
- Assumptions
- Goal of the lead score
- Exploratory Data Analysis
- Model Building
- Model Evaluation
- Conclusion

Problem Statement :

- X Education offers online courses to industry professionals. The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses, fill up a course form, or watch some videos.
- The business wants us to develop a model in which each lead is given a lead score, with higher lead scores indicating a higher likelihood of conversion and lower lead scores indicating a lower likelihood of conversion.
- The desired lead conversion rate has been set by the CEO to be in the range of 80%.

Assumptions :

- Missing data threshold percentage has been assumed to be 35 %
- While splitting the data into train and test set;
  - Train set data : 70%
  - Test set data : 30%

Business Goal :

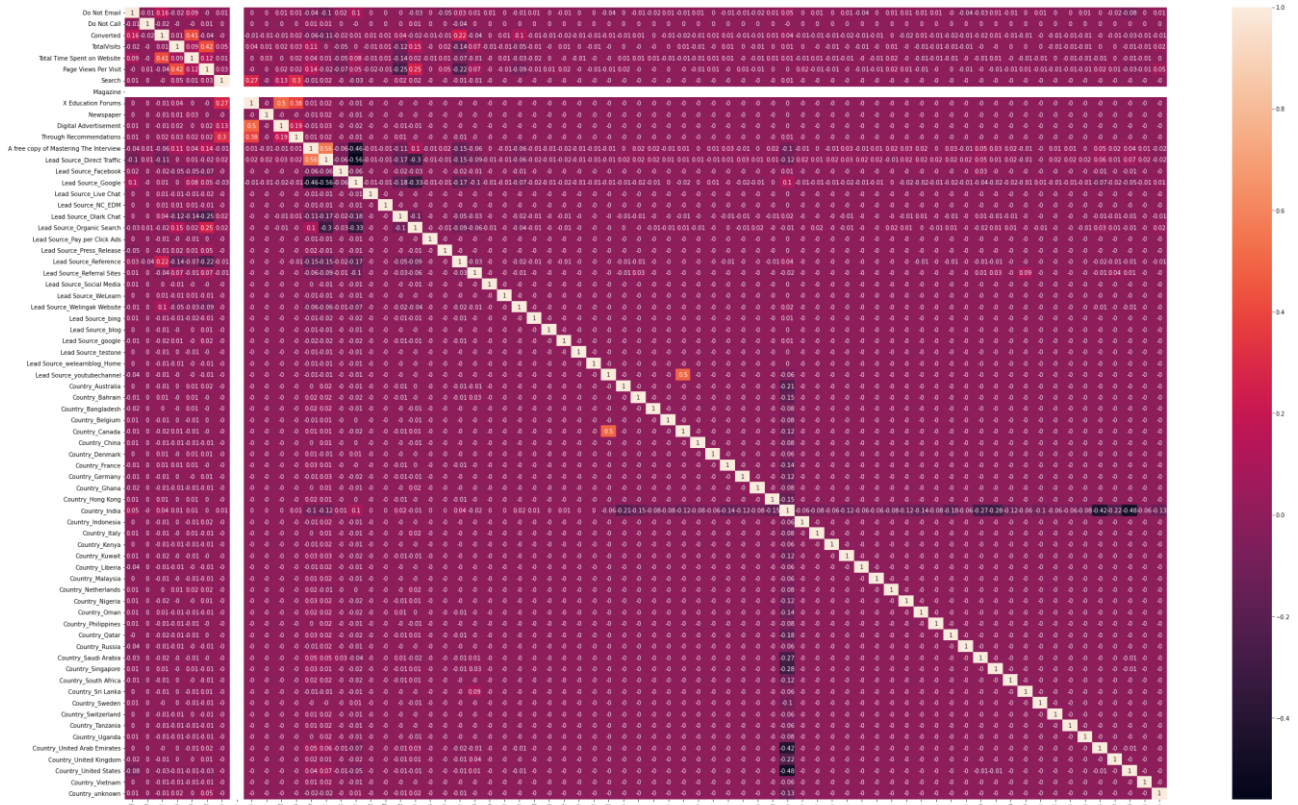
- The business needs a model to be created for choosing the most promising prospects.
- Create a logistic regression model to provide each lead a lead score between 0 and 100 that the business may use to target potential prospects.
- When a lead receives a higher score, it is more likely that it will convert, whereas when a lead receives a lower score, it is more likely that it won't convert.

Exploratory Data Analysis – I :

- Reading and Understanding the input data
- Data Cleaning- Handling Null values and removing higher null values data
- Imputation of the data based upon a mode value of Total Visits, Page views per visit, and Last Activity.
- Converted Binary variables into 0 and 1.
- Converted categorical variables into dummy variables
- Checked for duplicated data
- Assumption : Threshold percentage of missing data - 35 %
- Imputation and dropping of values having “Select” in their columns
- Performed Train-Test split of derived data after completion of Data cleaning and Data preparation
- Performed scaling of derived data

Exploratory Data Analysis – II :

- Created co-relation metrics of all the derived variables



Note : Please refer the python script file for details of co-relation metrics

Model Building:

- Build the First Model
- Feature Selection Using RFE
- Assessing the model with Stats Models
- Checking VIFs[variance inflation factor] and calculating model accuracy
- Metrics beyond simply accuracy
- Calculated the other performance measures like
  specificity, sensitivity, false positive rate, positive predictive value, Negative predictive value
- Finding Optimal Cutoff Point
- Finding Precision and Recall
- Making predictions on the test set and checking the performance measures of the model
- Generating the score variable

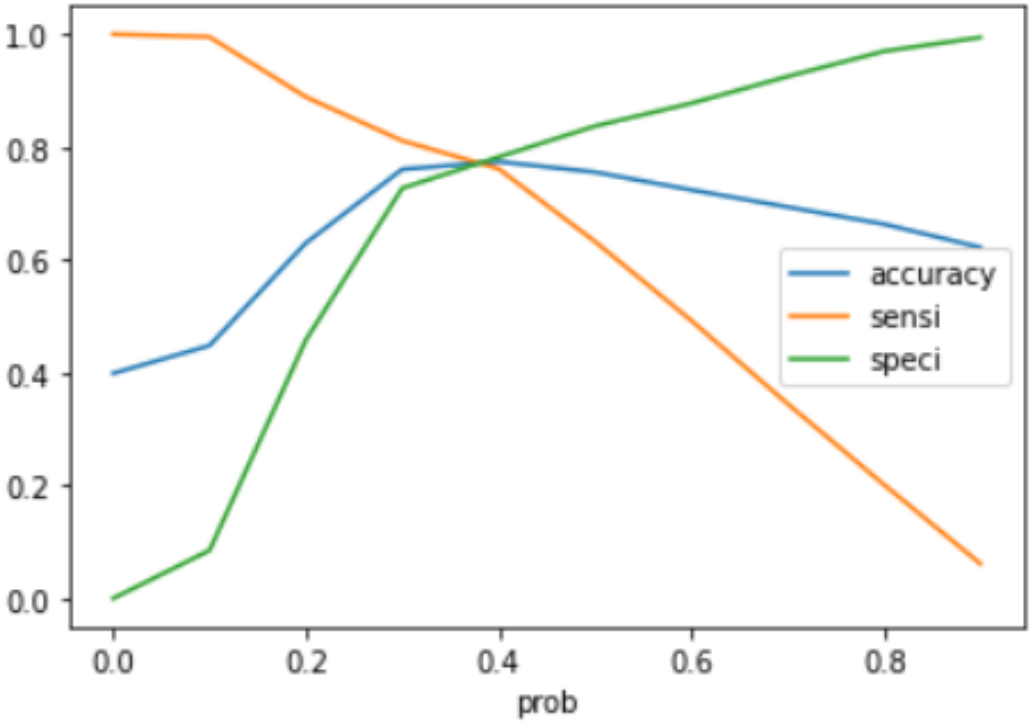
The names of all the feature variables and their respective **VIFs**

	Features	VIF
0	Do Not Email	1.12
3	Lead Source_Olark Chat	1.07
4	Lead Source_Reference	1.06
1	Total Time Spent on Website	1.03
5	Lead Source_Welingak Website	1.01
2	Search	1.00
6	Country_Kuwait	1.00
7	Country_Nigeria	1.00
8	Country_Qatar	1.00
9	Country_Saudi Arabia	1.00

Model Evaluation:

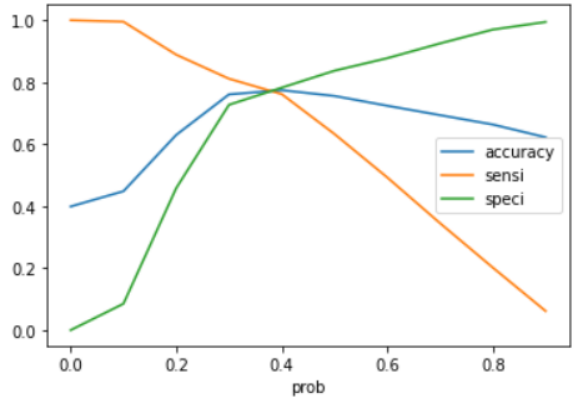
Accuracy, Sensitivity, and Specificity:

The graph depicts an optimal cut-off of 0.4 based on Accuracy, Sensitivity, and Specificity



Model Evaluation – Sensitivity, Specificity and Precision, Recall on Test Dataset

- Accuracy - 77%
- Sensitivity - 76 %
- Specificity - 78 %
- False Positive Rate - 21%
- Positive Predictive Value - 70 %
- Positive Predictive Value – 83%
- Precision- 72%
- Recall- 62%



Conclusion:

Logistic regression Model:

- ✓ The model performs well, with an accuracy close to 77%.
- ✓ The Accuracy, Sensitivity, Specificity, and precision, recall curves were used to choose the threshold values.
- ✓ The model has a sensitivity of 83% and a specificity of 62%.

Recommendation:

- ✓ Majority of the leads can be found from variable “Country” and followed by other variable like “Lead Source” and way of communication.
- ✓ As per the Logistic regression model, it is highly recommended that X Education should focus on factors;
  - ➢ Communication – Email
  - ➢ Total Time Spent on Website
  - ➢ Through Recommendations
  - ➢ Country – Italy, Kuwait, Qatar and Saudi Arabia
  - ➢ Lead Source – Google search and referral sites