Presentation on Lead Scoring Case Study

By Indrajeet Chaudhary

DCS 49 OCT-2022 BATCH

PURPOSE

- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. 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.
- There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

Solution Methodology

- 3.0 Cleaning Data
  - 3.1 Drop columns that are not useful for analysis
  - 3.2 Handle Missing values
- 4.0. Data Preparation for modelling
  - 4.1 Create dummy variables for all categorical variables
  - 4.2 Dummy variable creation
  - 4.3 Test-Train Split
  - 4.4 Scaling
  - 4.5 Looking at the correlations
- 5.0 Model Building
- 6.0 Model Evaluation
  - 6.1 Creating a data frame with the actual conversion flag and the predicted probabilities
  - 6.2 Finding the Optimal Cutoff
  - 6.3 Making Predictions on the Test Set
  - 6.4 Precision-Recall View
  - 6.5 Making Predictions on the Test Set

Problem Statement

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, 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

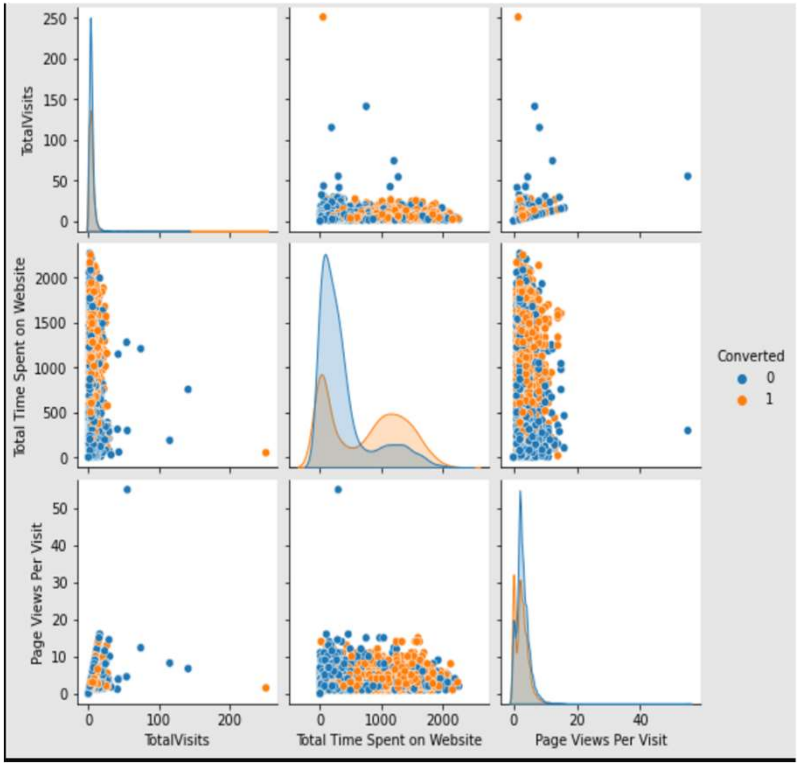
Business Objective:

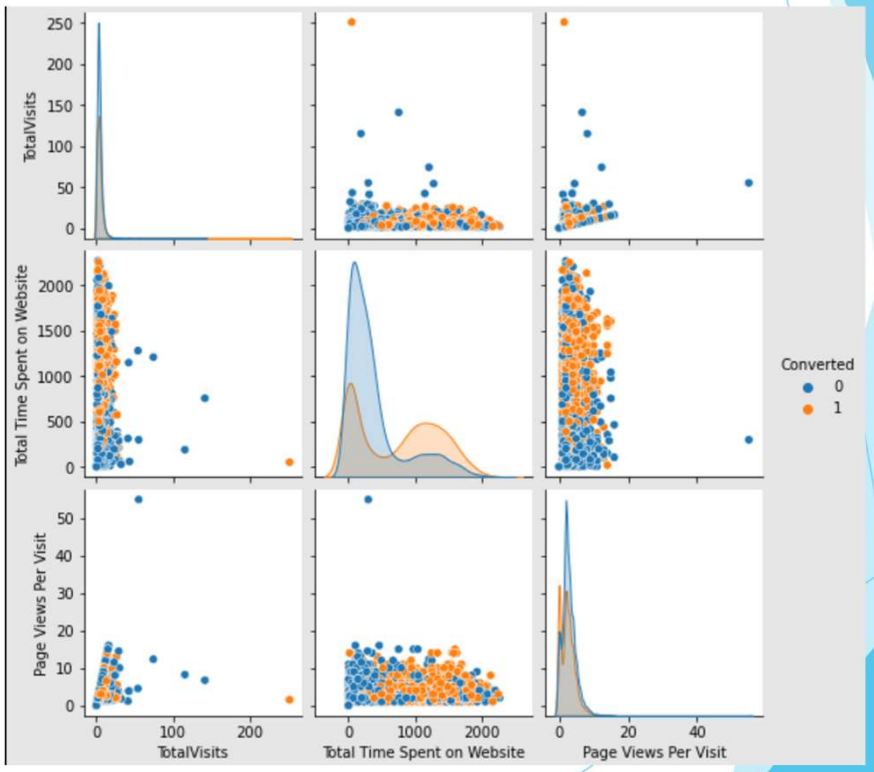
- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

Data Manipulation

- Total Number of Rows =37, Total Number of Columns =9240.
- Single value features like “Magazine”, “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” etc. have been dropped.
- Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis.
- After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are: “Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper
- Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement” etc.
- Dropping the columns having more than 35% as missing value such as ‘How did you hear about X Education’ and ‘Lead Profile’.

Data Preparation for modelling

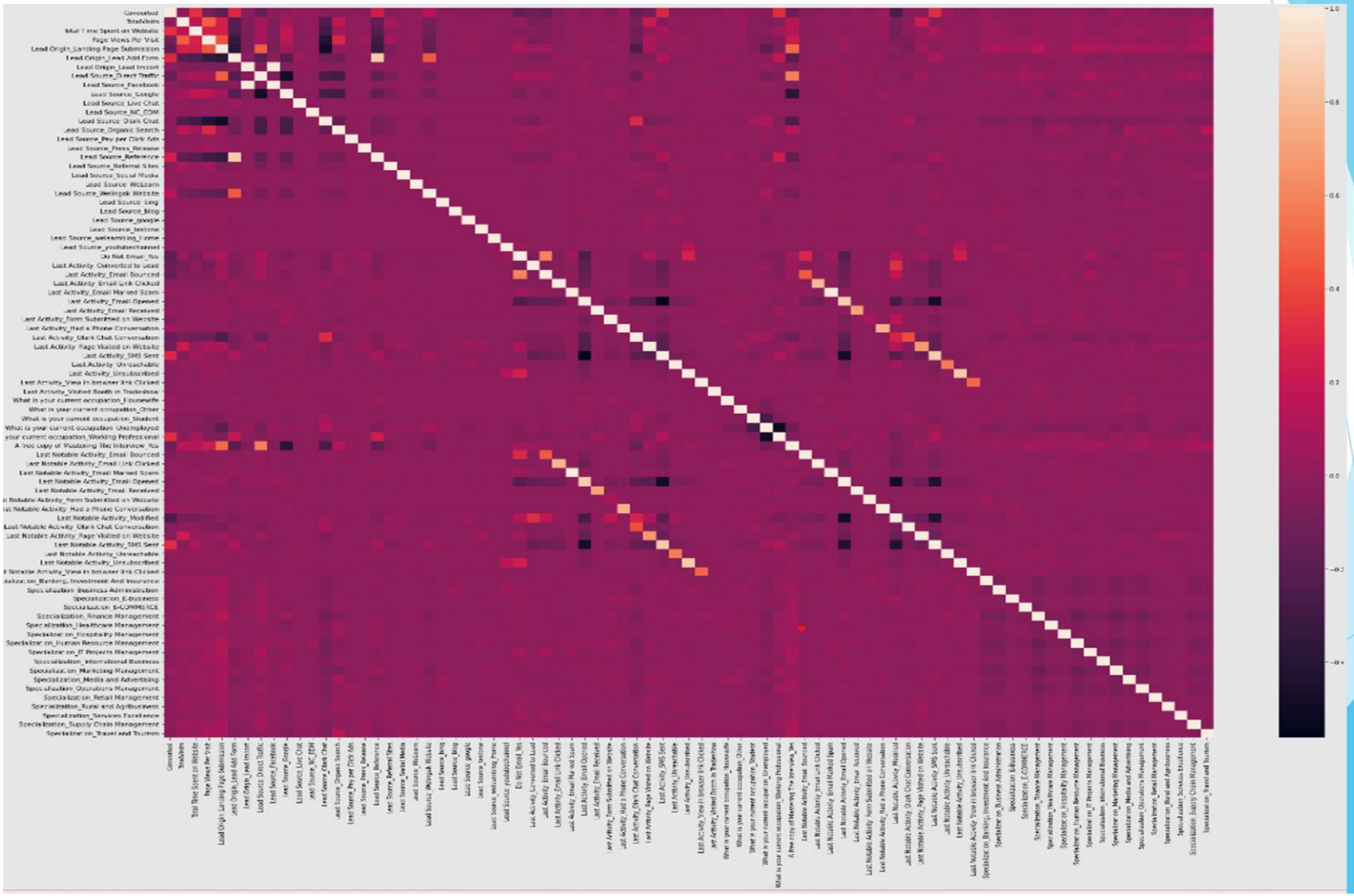




Variable Impacting the conversion rate

	Features	VIF
10	What is your current occupation_Unemployed	2.83
6	Last Activity_Had a Phone Conversation	2.58
12	Last Notable Activity_Had a Phone Conversation	2.57
1	Total Time Spent on Website	2.13
0	TotalVisits	1.80
2	Lead Origin_Lead Add Form	1.64
7	Last Activity_SMS Sent	1.53
11	What is your current occupation_Working Profes...	1.39
3	Lead Source_Olark Chat	1.36
4	Lead Source_Welingak Website	1.34
5	Do Not Email_Yes	1.07
9	What is your current occupation_Student	1.07
8	What is your current occupation_Other	1.01
13	Last Notable Activity_Unreachable	1.01

Looking at the correlations



Data Conversion

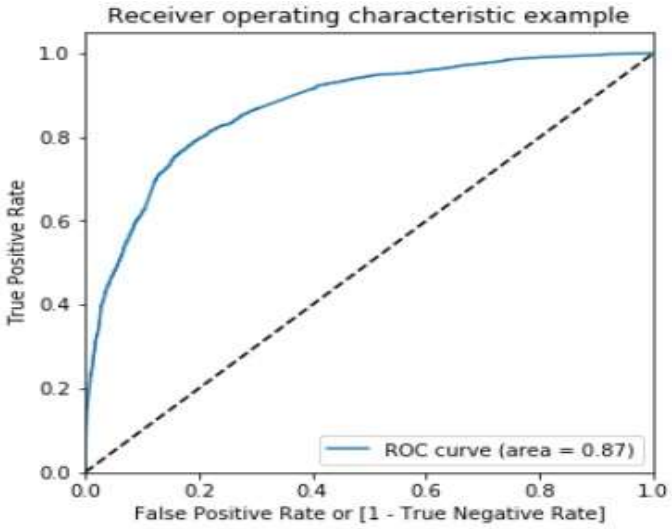
- Numerical Variables are Normalised
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 8792
- Total Columns for Analysis: 43

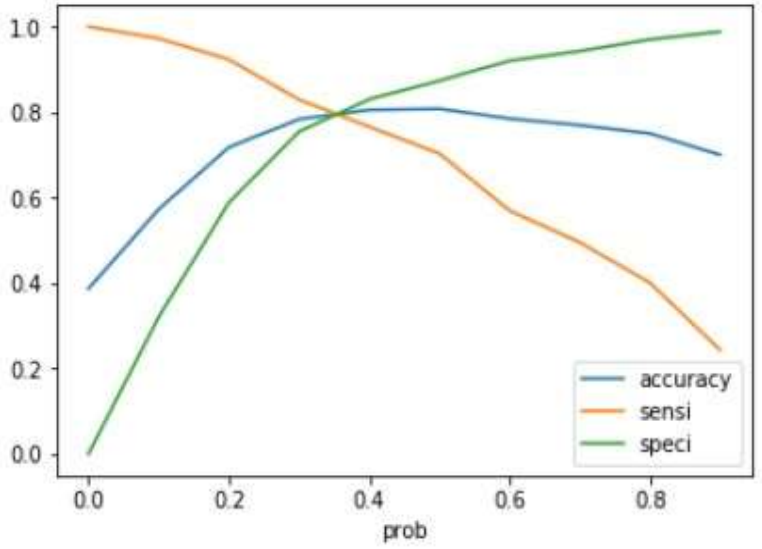
Model Building

- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p- value is greater than 0.05 and vif
- value is greater than 5
- Predictions on test data set
- Overall accuracy 81%

ROC Curve

Finding the Optimal Cut-off





Finding Optimal Cut off Point

- Optimal cut off probability is that
- probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut off is at 0.35.

Conclusion

It was found that the variables that mattered the most in the potential buyers are (In descending order) :

- The total time spend on the Website.
- Total number of visits.
- When the lead source was:
  - Google
  - Direct traffic
  - Organic search
  - Welingak website
- When the last activity was:
  - SMS
  - Olark chat conversation
- When the lead origin is Lead add format.
  - When their current occupation is as a working professional.

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.