LEAD SCORING CASE STUDY
LOGISTIC REGRESSION

By-

YUGANDHARI BODAPATI
AMAN PANDEY

PROBLEM STATEMENT

- A company namely, “X Education” specializes in providing online courses to industry professionals.
- The company has listed the details of its course on various websites example Google.
- The people land on these websites and search for the courses in which they are interested.
- Thus leads are generated through Emails, site visit, google searches, advertisements.
- Although X Education generates a lot of leads but only 30% are converted.
- The company wants to identify HOT LEADS or Promising Leads so as to improves the conversion which in turn will improve the efficiency.

Business Agenda

- The company needs a model which can select the most promising leads out of the Leads pool.
- Every individual Lead should be allotted a score which indicates how promising it is to convert. The higher the lead score is the more promising it will be hence can be called as HOT LEAD.
- The new lead conversion rate should be around 80%.

Strategy

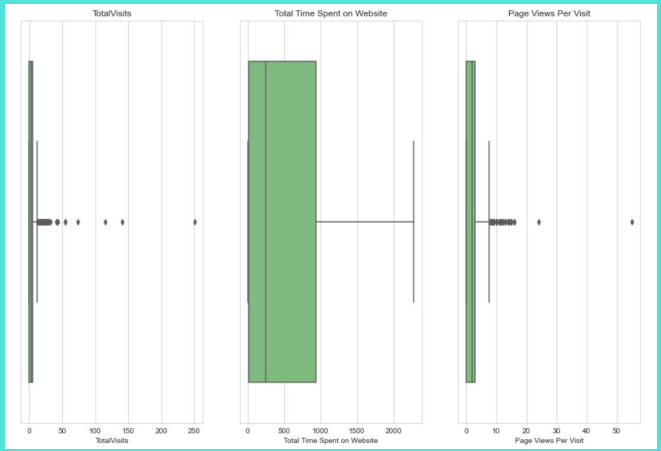
- Import Data
- Data cleaning and preparation for further analysis
- Exploratory Data Analysis
- Scaling features
- Preparing Data for modeling
- Building a logistic regression model
- Testing the model on Train set
- Evaluating the Model
- Testing the model on Test set
- Measuring the accuracy of the model.
- Assigning a Lead score
- Conclusion

Exploratory Data Analysis

- Data understanding - Checking for the categorical and numerical columns in data and checking for outliers.
- Data cleaning -Removing the unnecessary columns such as - 'Prospect ID', 'Lead Number', 'Country', 'I agree to pay the amount through cheque', 'A free copy of Mastering The Interview', 'City'.
- Replacing “Select” label in some columns with null values. Select label means that the responder hasn’t selected any option.
- Dropping columns having more than 35% null values.
- Now, Imputing the missing values in the columns with respective Mode values.

Outliers

- We can say that 'TotalVisits' & 'Page Views Per Visit' & 'Total Time Spent on Website' have outliers in them and we need to treat them to make our dataset fit for the analysis.
- Outlier Treatment: Remove top & bottom 1% of the Column Outlier values



Data Preparation for modeling

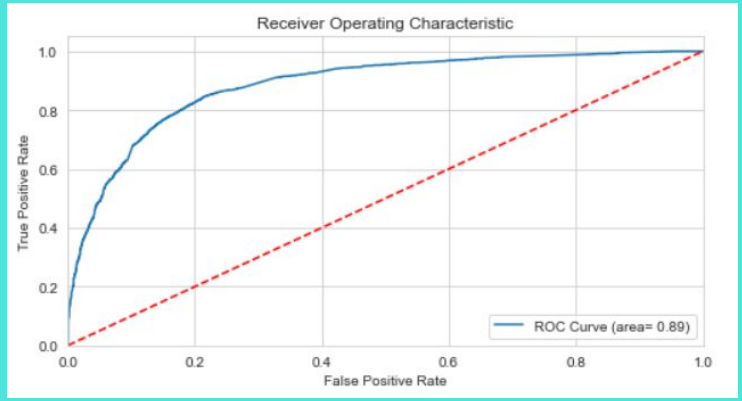
- Converting Yes/No to Binary variables and correcting datatype of Page Views Per Visit and TotalVisits.
- Creating dummy variables for the 8 categories and dropping the first level.
- Removing duplicate variables.
- Splitting the data in Train and Test sets.
- Feature Standardizations of numerical data.
- Initial conversion rate was found to be 38.45%.

Model Building

Eliminating insignificant features through RFE.
Determined optimal model using Logistic regression.
4th Model was finalised as optimal.

	Features	VIF
15	Last Notable Activity_Modified	2.44
1	Total Time Spent on Website	2.24
4	Lead Source_Google	2.20
3	Lead Source_Direct traffic	2.11
9	Last Activity_Olark Chat Conversation	1.81
13	Last Notable Activity_Email Opened	1.69
2	Lead Origin_Lead Add Form	1.54
5	Lead Source_Organic search	1.47
16	Last Notable Activity_Olark Chat Conversation	1.40
7	Lead Source_Welingak website	1.30
8	Last Activity_Converted to Lead	1.26
0	Do Not Email	1.18
11	What is your current occupation_Working Profes...	1.18
17	Last Notable Activity_Page Visited on Website	1.09
6	Lead Source_Referral sites	1.05
12	Last Notable Activity_Email Link Clicked	1.03
14	Last Notable Activity_Had a Phone Conversation	1.00
10	Last Activity_View in browser link Clicked	1.00

Prediction Train dataset



ROC Curve Plotting -

- ROC curve shows the trade off between sensitivity and specificity - means if sensitivity increases specificity will decrease.
- The curve closer to the left side border then right side of the border is more accurate.
- The curve closer to the 45-degree diagonal of the ROC space is less accurate.

Points to be noted from the ROC Curve

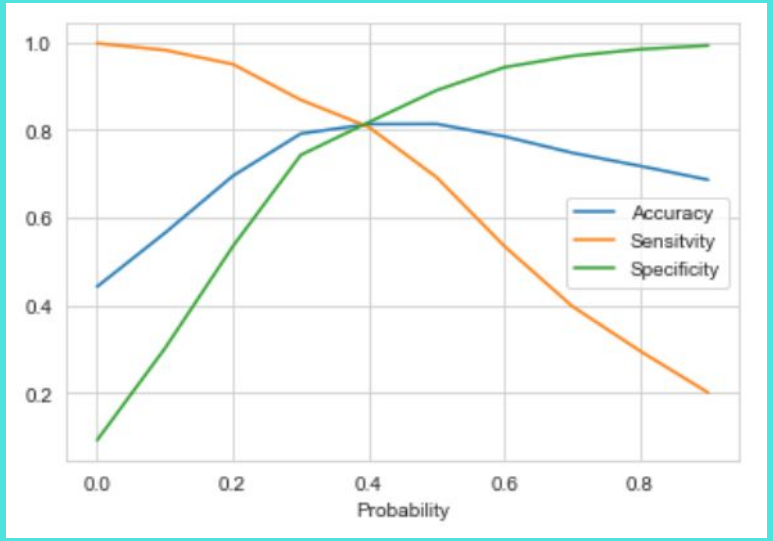
- The curve is closer to the left border than to the right border hence our model is having great accuracy.
- The curve area is 88% of the total area.

Model Evaluation

It was found that 0.4 is perfect for the probability cutoff.

The cutoff is based on -

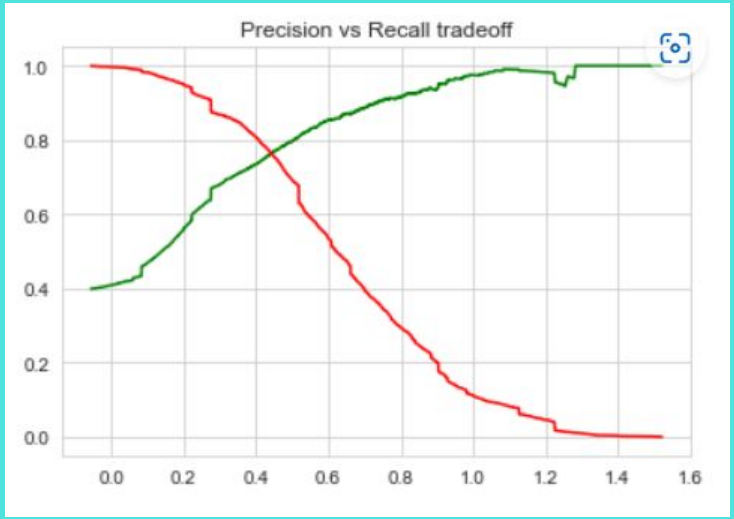
- **Accuracy Score = 81%**
- **Sensitivity Score = 80%**
- **Specificity Score = 82%**



Model Prediction

Calculated Accuracy, Sensitivity and Specificity.
Found Precision and Recall score;

- **> Train datasets ~ 73% and 79% respectively.**
- **> Test datastets ~ 71% and 79% respectively.**



Conclusion

- **The Accuracy, Precision and Recall score we got from the test data are in the acceptable region.**
- **Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are:**
  - **Last Notable Activity_Modified**
  - **Total Time Spent on Website**
  - **Lead Source_Google.**