Lead Scoring Case Study (X Education)

SUBMISSION

Group Name:

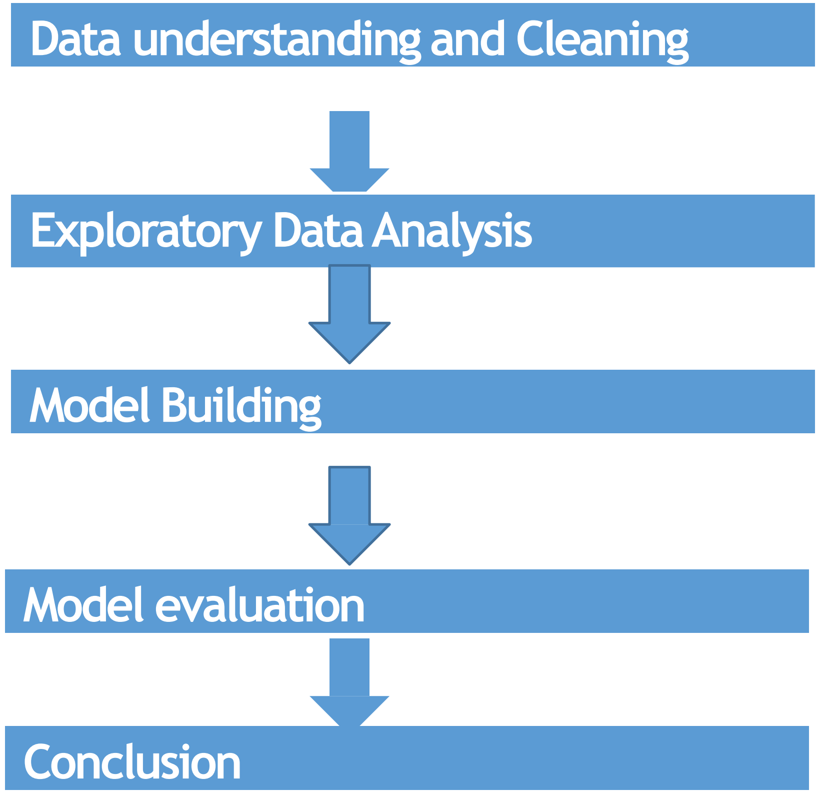
1. Sudheer Kumar Payyavula
2. Sravan
3. Vimalan
4. Praneeth

Abstract:

- This presentation contains the approach followed to build a logistic regression model for X Education.
- Our main intention here is to find the potential leads who actually converts as customers of X Education.
- Our intention is to assign a lead score to all the leads in a way that high score represents higher chances of lead conversion and low score represents lower chances of lead conversion.

Problem solving approach:

- ➢ Followed below approach for the analysis.



Data understanding and Cleaning

Exploratory Data Analysis

Model Building

Model evaluation

Conclusion

Data Understanding and Cleaning:

- Dataset contains 9240 rows and 37 columns.
- Dataset contains columns related to leads. like last notable activity, Lead origin, converted as customer or not etc.
- Few of the columns are last notable activity, Lead origin, converted as customer or not etc.
- As column names are too lengthy, They have been renamed to short names for the convinienience. Below are few of the columns which was renamed.
- Lead Origin - LO
  Lead Source -LS
  Do Not Email -DNE
  Do Not Call -DNC
  Total Time Spent on Website - TTSW
  Page Views Per Visit - PVPV
  Last Activity - LA

Data Understanding and Cleaning:

- Unique values in each column has been checked and we came to know few columns contains single level in them.
- Below are the few columns which contain single level.
- 'Update me on Supply Chain Content' column contains only one level 'NO'

  'Get updates on DM Content' column contains only one level 'NO'

  'I agree to pay the amount through cheque' column contains only level 'No'

  'Receive More Updates About Our Courses' column contains only level 'No'

  'Magazine' column contains only level 'No‘
- All the characters in the dataset have been changed to lower case for convinience.

Data Understanding and Cleaning:

- Below are the few columns which contain more than 3000 null values.
- 'Country‘

  'Specialization‘

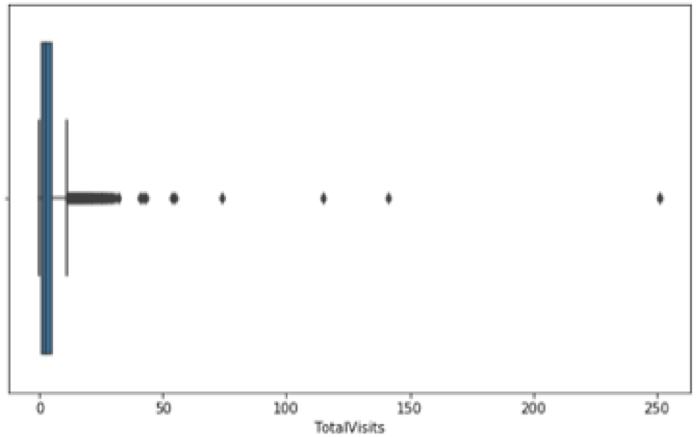
  'How do you know about X education‘

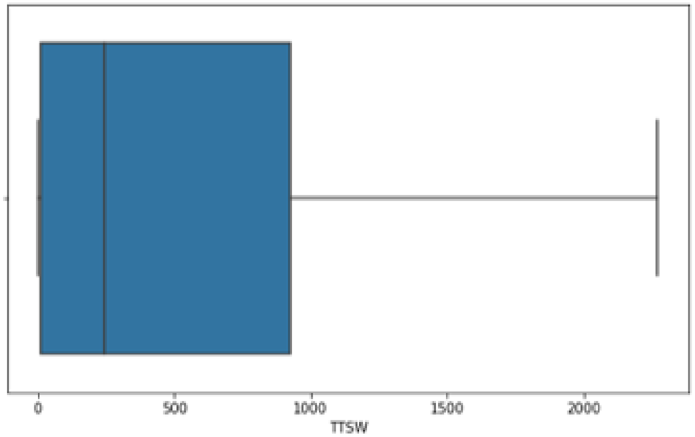
  'What is your career objective‘

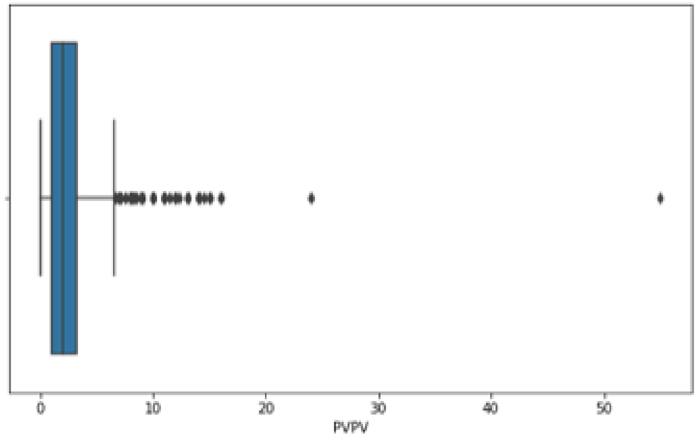
  'What matters most to you in choosing a course‘
- Above columns have been dropped.
- Few columns contain the level ‘SELECT’ which is insignificant. So these columns have been dropped.

Exploratory Data Analysis:

- Below are the Boxplots of numerical variables.
- We see good number of outliers in the data so we removed them.

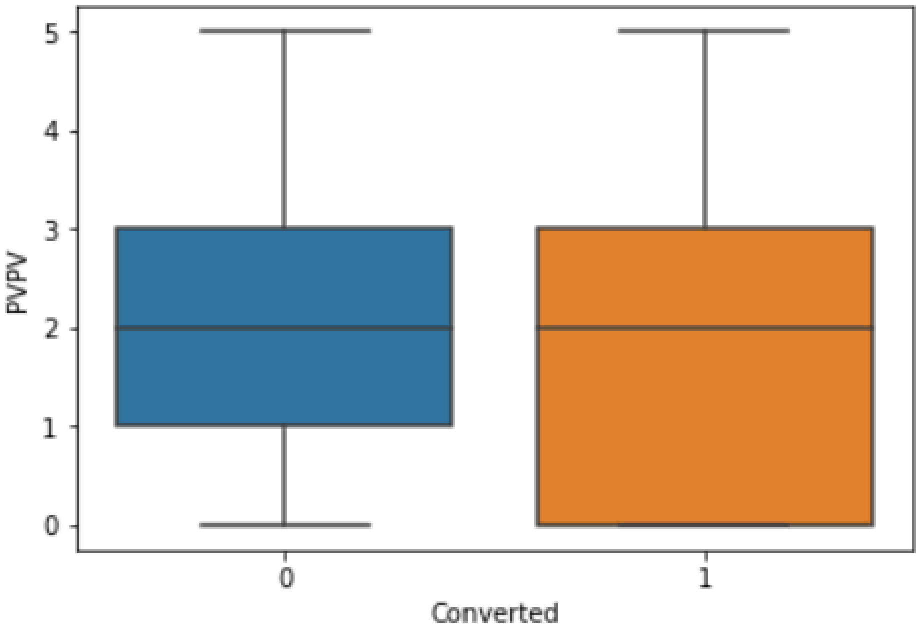


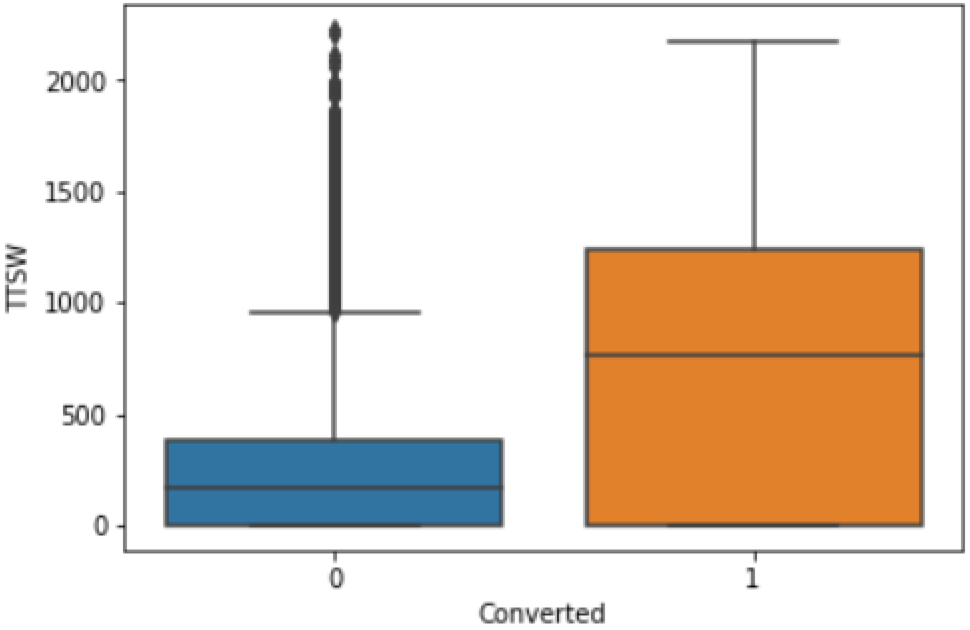




Exploratory Data Analysis:

- Below are the Boxplots of numerical variables by conversion status.
- We see ‘Total time spent on website’ is varying with conversion status.
- ‘Pages visited per visit’ is not varying much with conversion status.





Exploratory Data Analysis:

- Below are the observations after checking at the lead conversion rate in all the objective columns. The people who has 'lead origin' as 'lead add form' has 94% lead conversion rate.
- The people who has 'lead source' in 'nc_edm','live chat' has 100% lead conversion rate.
- The people who has chosen NO to 'do not email' option has better conversion rate than the people who chosen ‘YES'.
- The people who has 'last notable activity' in 'approached upfront','email marked spam' and 'resubscribed to emails' has 100% lead conversion rate.

Model Building:

- We have removed the columns with high multi collinearity.
- We have created Dummy variables for all the categorical variables.
- We have standardized the numerical columns data by using standard scaler.
- We have splitted the data into Train and Test sets.
- Initially we have selected top 20 features through RFE and then followed manual approach to eliminate featues with high VIF.

Model Building:

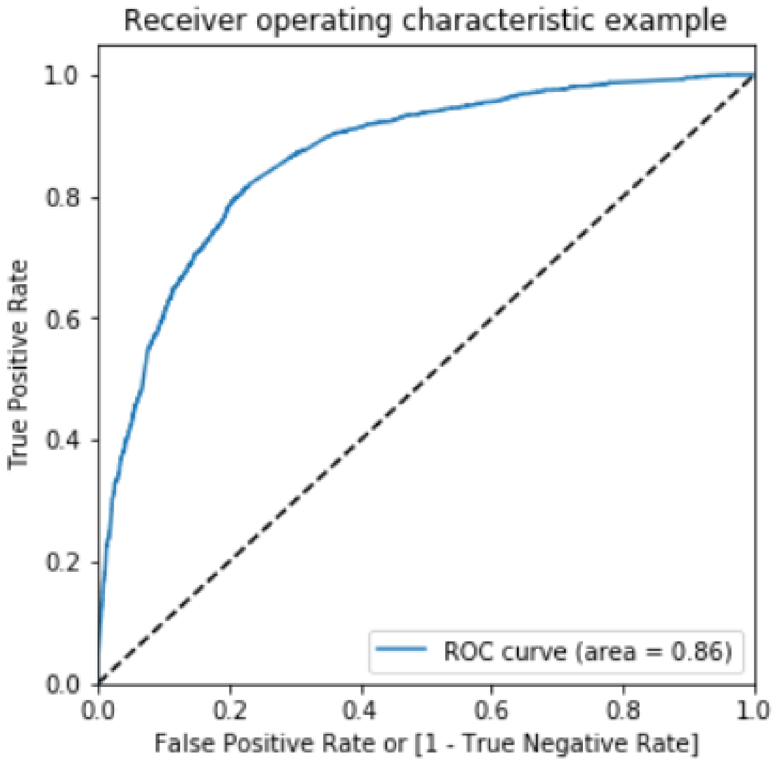
- Below are the features left in the final model.
- LNA_modified, TTSW
- DNE, LA_converted to lead
- LA_email bounced, LS_organic search
- LA_olark chat conversation, LS_reference
- LS_google, LNA_page visited on website
- LS_direct traffic, LNA_email link clicked
- LNA_email opened, LS_referral sites
- LNA_olark chat conversation, LS_welingak website

Model Evaluation:

- We have predicted probabilities of conversion of train dataset.
- We have chosen initial cutoff probability as 0.5 and assigned predicted conversion status to each row.
- Below is the metrics obtained for cut off of 0.5
- Accuracy is 0.79
- Sensitivity is 0.66
- Precision is 0.76

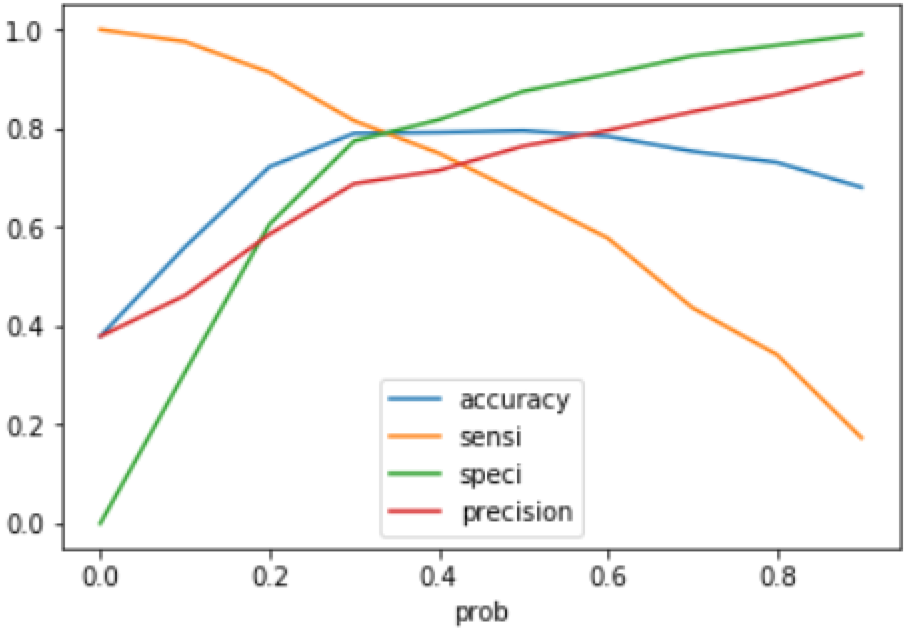
Model Evaluation:

- Below is the ROC Curve of our model.
- ROC curve is inclined towards and Y axis and area under curve is 0.86 which is pretty good.



Model Evaluation:

- Below is the plot between Accuracy, Sensitivity, Specificity and Precision at different probability cutoff's.
- We see Accuracy, Sensitivity and Specificity meet at 0.33.
- But precision is meeting these metrics at different points.



Model Evaluation:

- Our business objective is to have Precision to be atleast 0.8.
- From above plot we can say that at probability cutoff of 0.62 Precision is 0.8.
- So we chosen cutoff at 0.62 and predicted for the train dataset and below are the updated metrics now.
- Accuracy is 0.79
- Sensitivity is less which is 0.56
- Specificity is high which is 0.91
- Precision is 0.8 which is good

Model Evaluation:

- Now we made predictions on Testset.
- Below are the metrics obtained for Testset.
- Accuracy is 0.80
- Sensitivity is less which is 0.58
- Specificity is high which is 0.93
- Precision is 0.84 which is good
- We had to compromise for Sensitivity as our business objective is to maintain precision grater than 0.8.

Conclusion:

- We have achieved our business objective of having precision greater than or equal to 0.8.
- We have built a logistic regression model which predicts the potential leads by using few diver variables.
- Out of 100 leads predicted by our model atleast 80 will be potential leads.