Lead Score Case Study

Divya Tyagi
Prakriti Mishra

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.

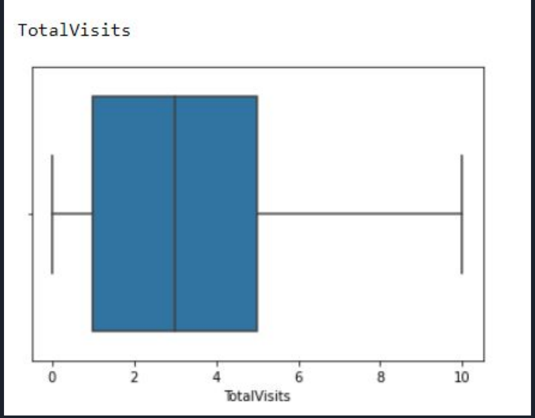
Solution Methodology

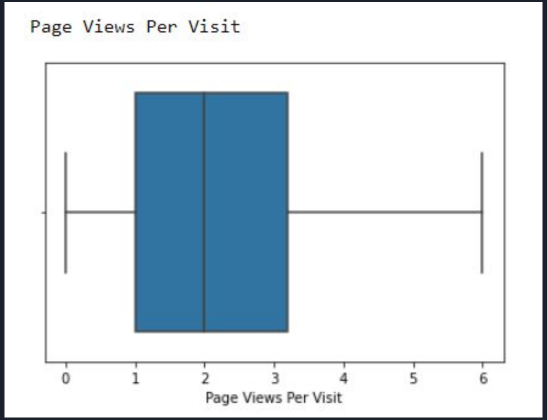
- Importing Data and Necessary Libraries
- Data cleaning and data manipulation.
  - Check and handle NA values and missing values.
  - Drop columns, if it contains large amount of missing values and not useful for the analysis.
  - Imputation of the values, if necessary.
  - Check and handle outliers in data.
- EDA and Data Visulization
  - 1.Univariate data analysis: value count, distribution of variable etc.
  - 2.Bivariate data analysis: correlation coefficients and pattern between the variables etc.
- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique: logistic regression used for the model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations

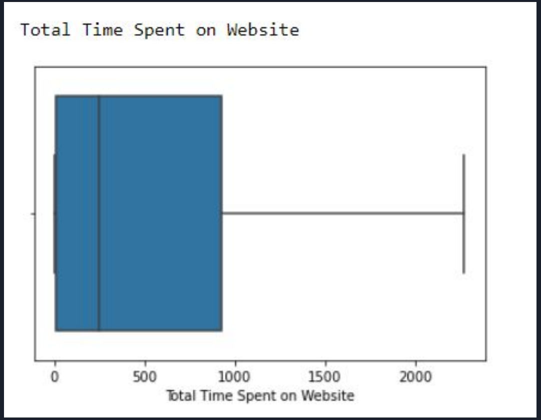
Data Cleaning

- Total Number of Columns = 37, Total Number of Rows =9240.
- Removing the “How did you hear about X Education”, “Lead Profile” and “Lead Quality” due to more than 50% missing values.
- Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis.
- Removing Asymmetrique columns as they have more than 45% missing value and are insignificant.
- “City” column is imputed with the category with highest frequency namely Mumbai.
- Similarly changes are made to other major categorical columns.
- For the columns with less than 2% NAN values, the rows with missing information can be dropped.
- “TotalVisits”, “Total Time Spent on Website”, “Page Views Per Visit” columns have outliers so they are capped at 95% value.

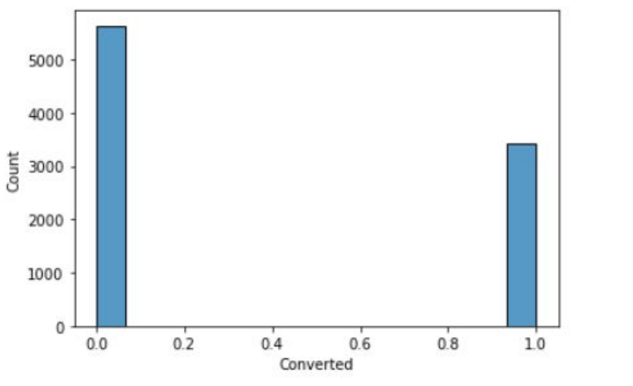
EDA and Data Visualization



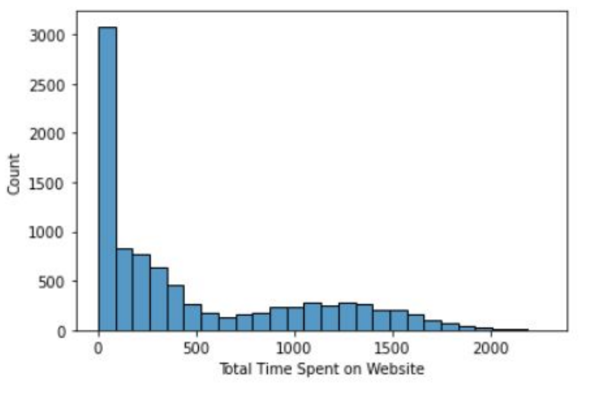




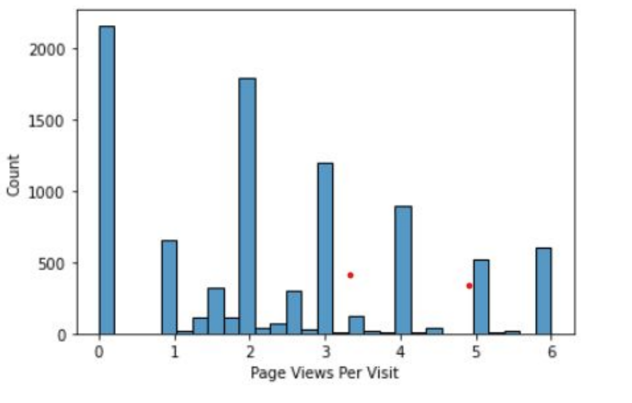
Converted



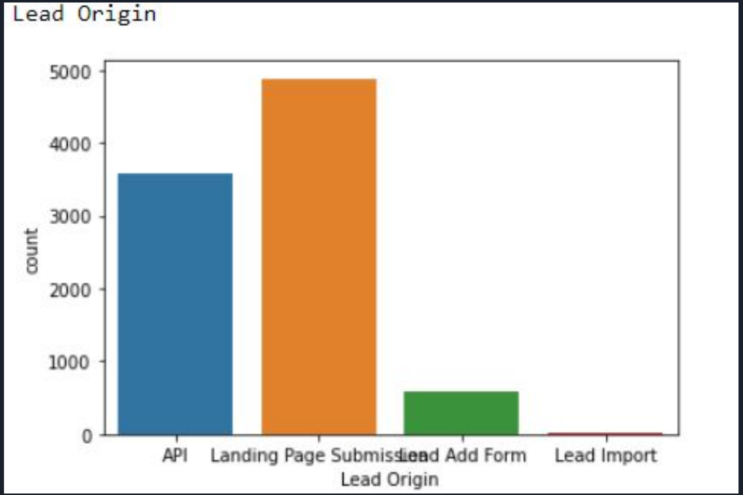
Total Time Spent on Website

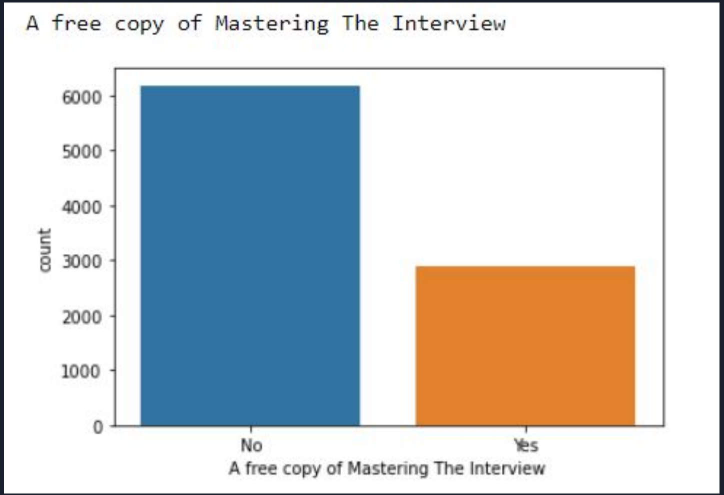


Page Views Per Visit

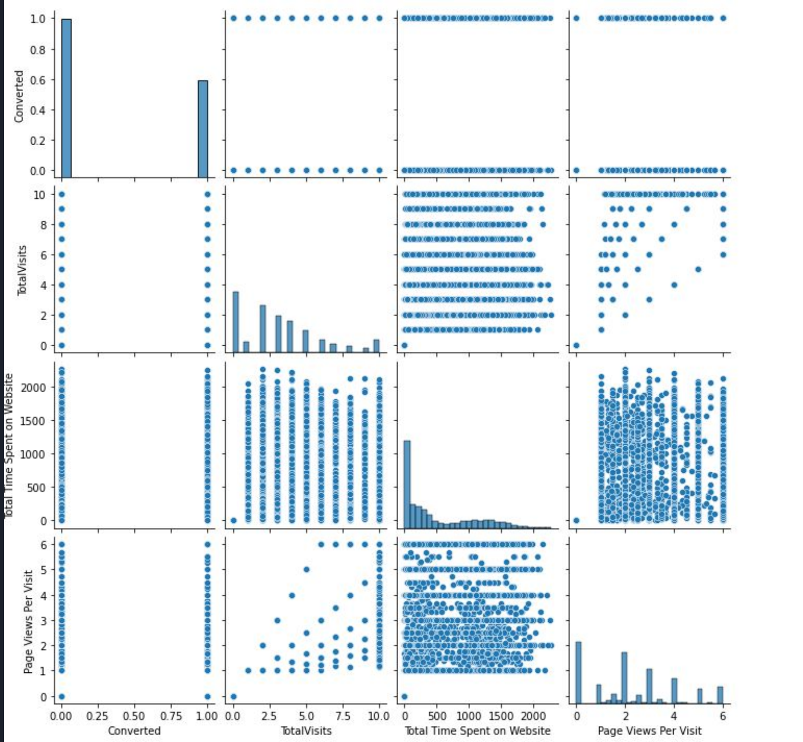


Bivariate Analysis





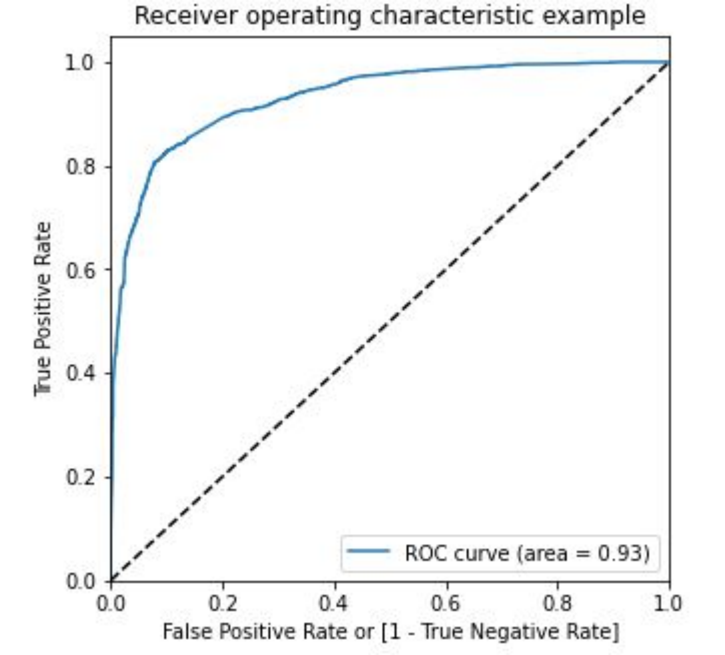
Multivariate Analysis

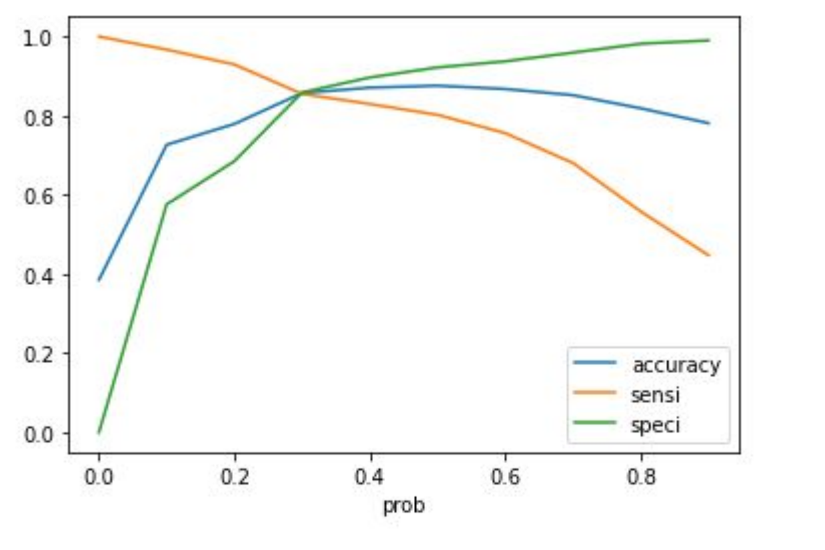


Model Building

- 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 vifvalue is greater than 5
- Predictions on test data set
- Overall accuracy 81%

ROC Curve





From the curve above, 0.37 is the optimum point to take as a cutoff probability

Conclusion

Following are the conclusions from the final model we have made:

- The P-values of all the features in the final model is zero, which shows high significance of features present.
- The VIFs of all the features of the final model is well below the threshold of 5 , which is a good thing as it shows no multi collinearity is present between the features.
- Accuracy of the final model is 92%, which is quiet decent.
- The values we got at 0.5 threshold were decent but we got better metrics at a threshold of 0.37. which is again a better value, but it is subject and can be increased or decreased as per the business needs.
- The final model has 13 features.

TRAIN Vs TEST DATASET

TRAIN DATASET (Threshold = 0.37)

- Accuracy = 0.8669500866005353
- Recall = 0.8344235486508585
- Precision = 0.822652156388553
- AUC = 93%

TEST DATASET

- Accuracy = 0.8663239074550129
- Recall = 0.8301314459049545
- Precision = 0.8072763028515241
- AUC=92%

Top 13 features of the final model are:

- Do Not Email
- Total Time Spent on Website
- Lead Origin_Lead Import
- Lead Source_Olark Chat
- Lead Source_Reference
- Lead Source_Welingak Website
- Last Activity_Olark Chat Conversation
- What is your current occupation_Working Professional
- Tags_Interested in other courses
- Tags_Lost to EINS
- Tags_Ringing
- Tags_Will revert after reading the email
- Last Notable Activity_Others

Features with Negative Coefficient:

- Do Not Email
- Last Activity_Olark Chat Conversation
- Tags_Interested in other courses
- Tags_Ringing

Features with Positive Coefficient:

- Total Time Spent on Website
- Lead Origin_Lead Import
- Lead Source_Olark Chat
- Lead Source_Reference
- Lead Source_Welingak Website
- What is your current occupation_Working Professional
- Tags_Lost to EINS
- Tags_Will revert after reading the email
- Last Notable Activity_Others

Recommendation

The following mattered most for a lead conversion:

- When the lead source was:
  - Welingak website
  - Reference
  - Olark Chat
- Current occupation: Working Professional
- Last Notable Activity: Others
- Lead Origin: Lead Import
- Total time spent on website