Lead Scoring Case Study

19/09/2023

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.
- 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.
- If they successfully identify this set of leads, the lead conversion rate should go upas 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.

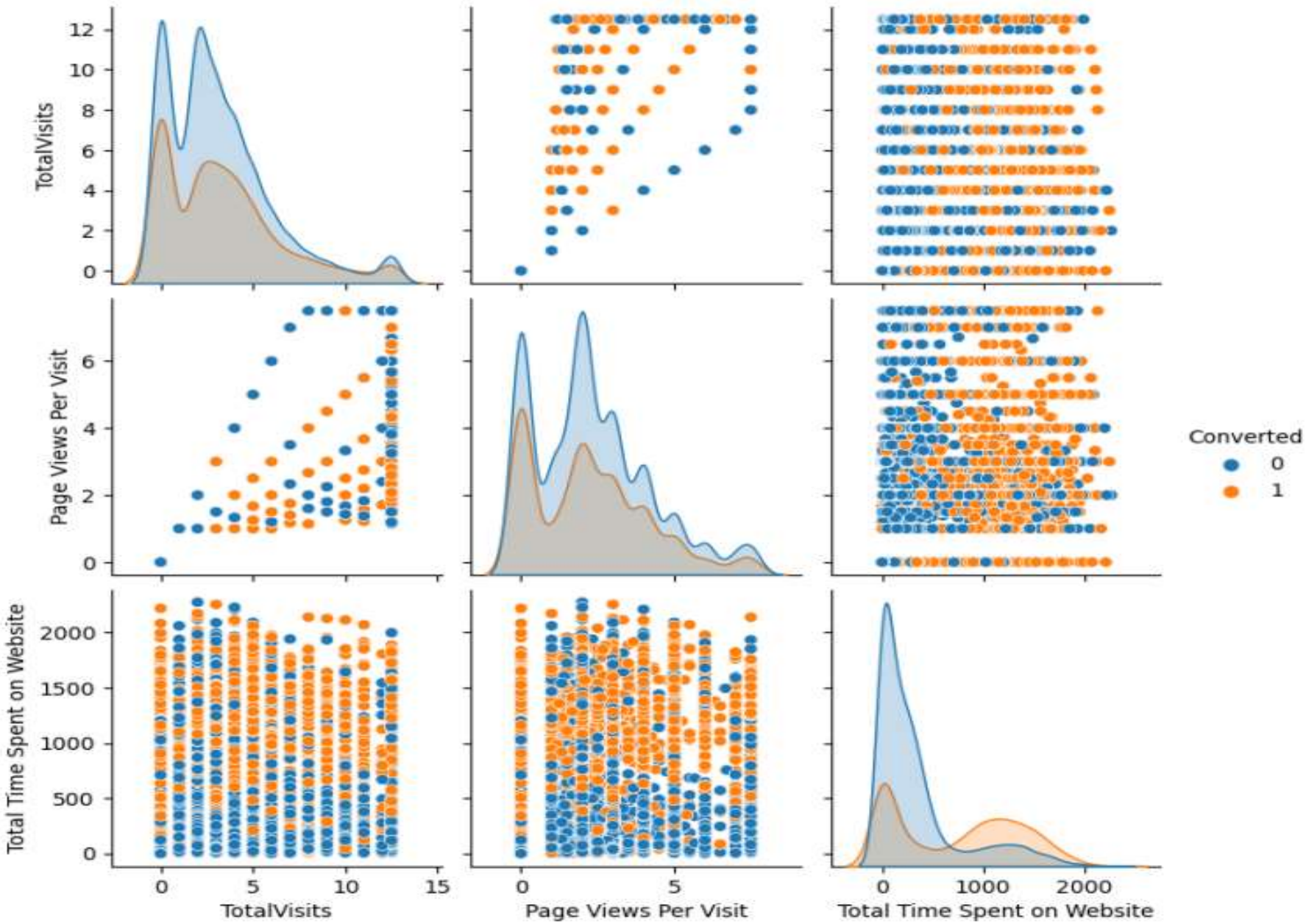
SOLUTION METHODOLOGY

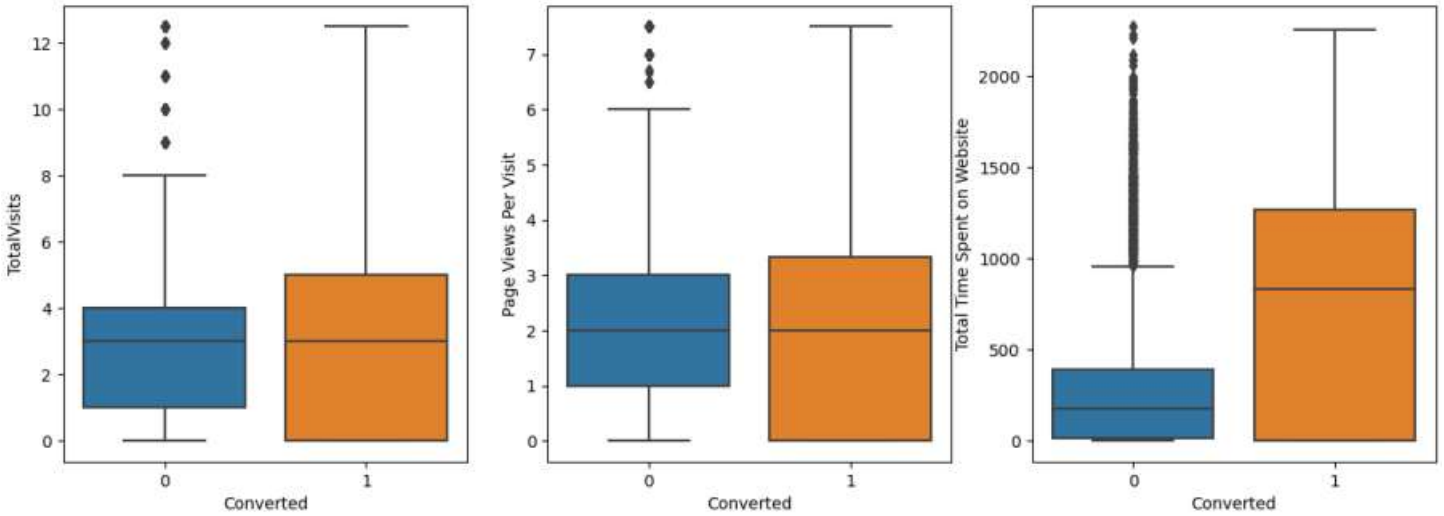
Data cleaning and data manipulation.	Exploratory Data Analysis (EDA)
Check and handle duplicate data.	Univariate data analysis: value count, distribution of variables, etc.
Check and handle NA values and missing values	Bivariate data analysis: correlation coefficients and pattern between the variables etc.
Drop columns, if it contains a large number of missing values and are not useful for the analysis.	Feature Scaling & Dummy variables and encoding of the data.
Imputation of the values, if necessary.	Classification technique: logistic regression is used for model making and prediction.
Check and handle outliers in data.	Validation of the model.

Data Manipulation

- Total Number of Rows=37,Total Number of Columns =9240
- Single value features like“Magazine” , “ReceiveMoreUpdates About Our Courses” , “Update my supply”
- Chain Content” , “Get updates on DM Content” , “I agree to pay the amount through cheque” etc. have been dropped.
- Removing the“ProspectID” and “Lead Number” which are 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 have enough variance, which have dropped, the features are: “Do Not Call” , “What matters most to you in choosing course” , “Search” , “Newspaper, Article” , “XEducation Forums” , “Newspaper” , “DigitalAdvertisement” etc.
- Dropping the column shaving more than 35% as missing values such as ‘How did you hear about X Education’ and‘Lead Profile’.

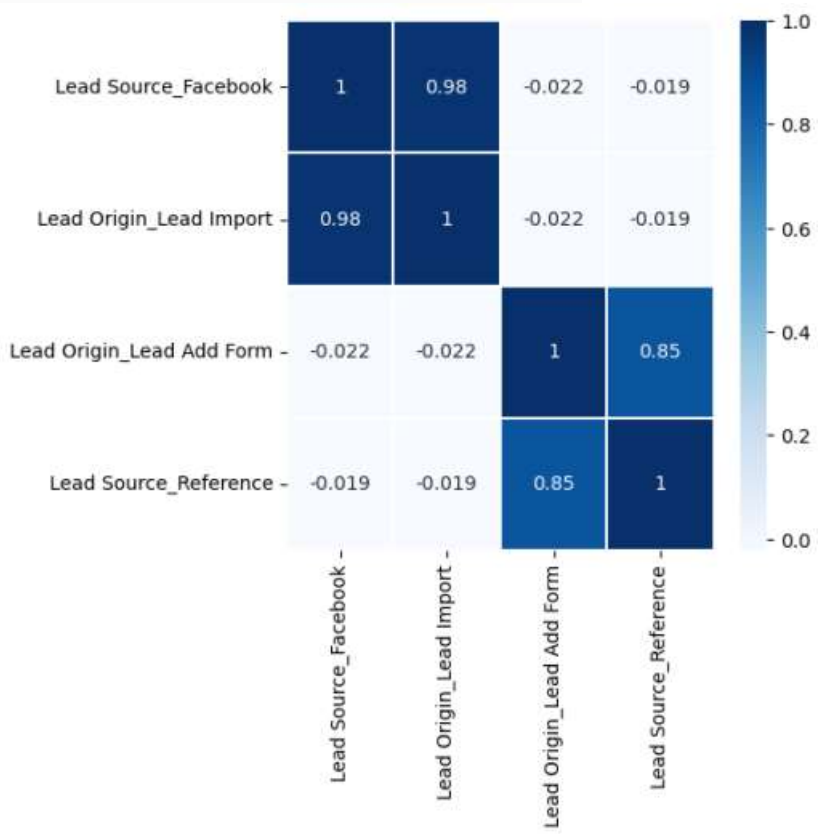
EDA


TotalVisits
Page Views Per Visit
Total Time Spent on Website
TotalVisits
Page Views Per Visit
Total Time Spent on Website
Converted
0
1



Box Plot

Heat Map



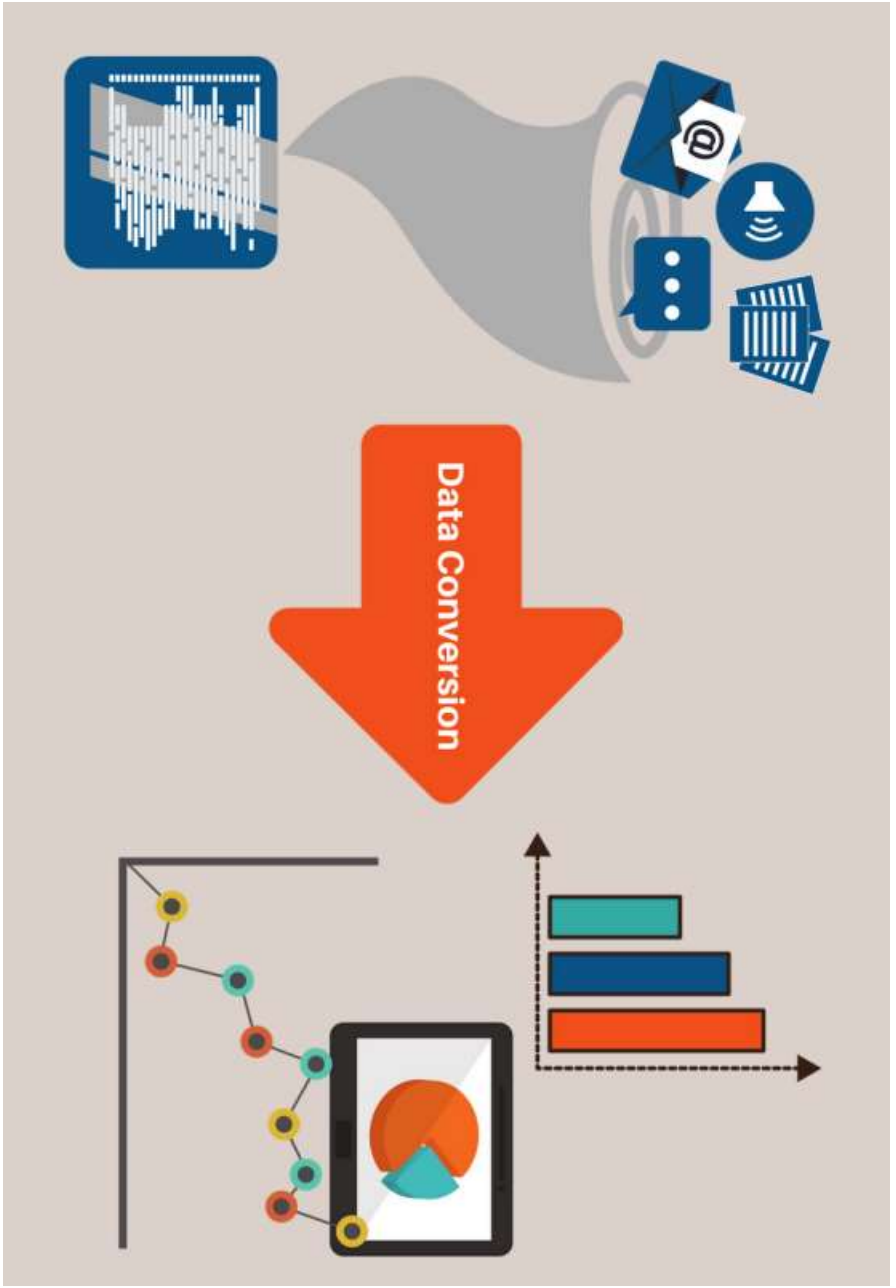
Data Conversion

Numerical Variables are normalized

Dummy Variables are created for object type variables

Total Rows for Analysis: 9240

Total Columns for Analysis: 37

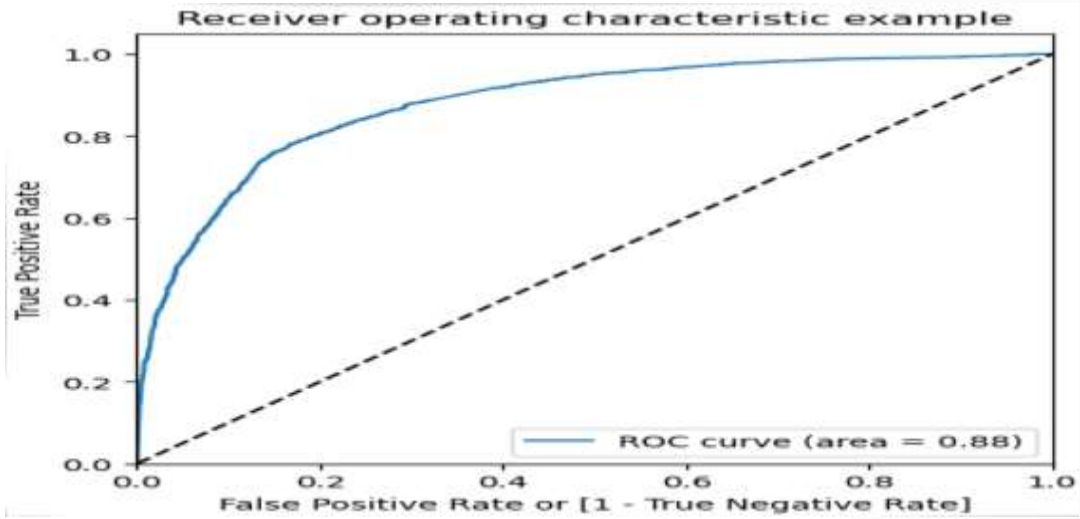


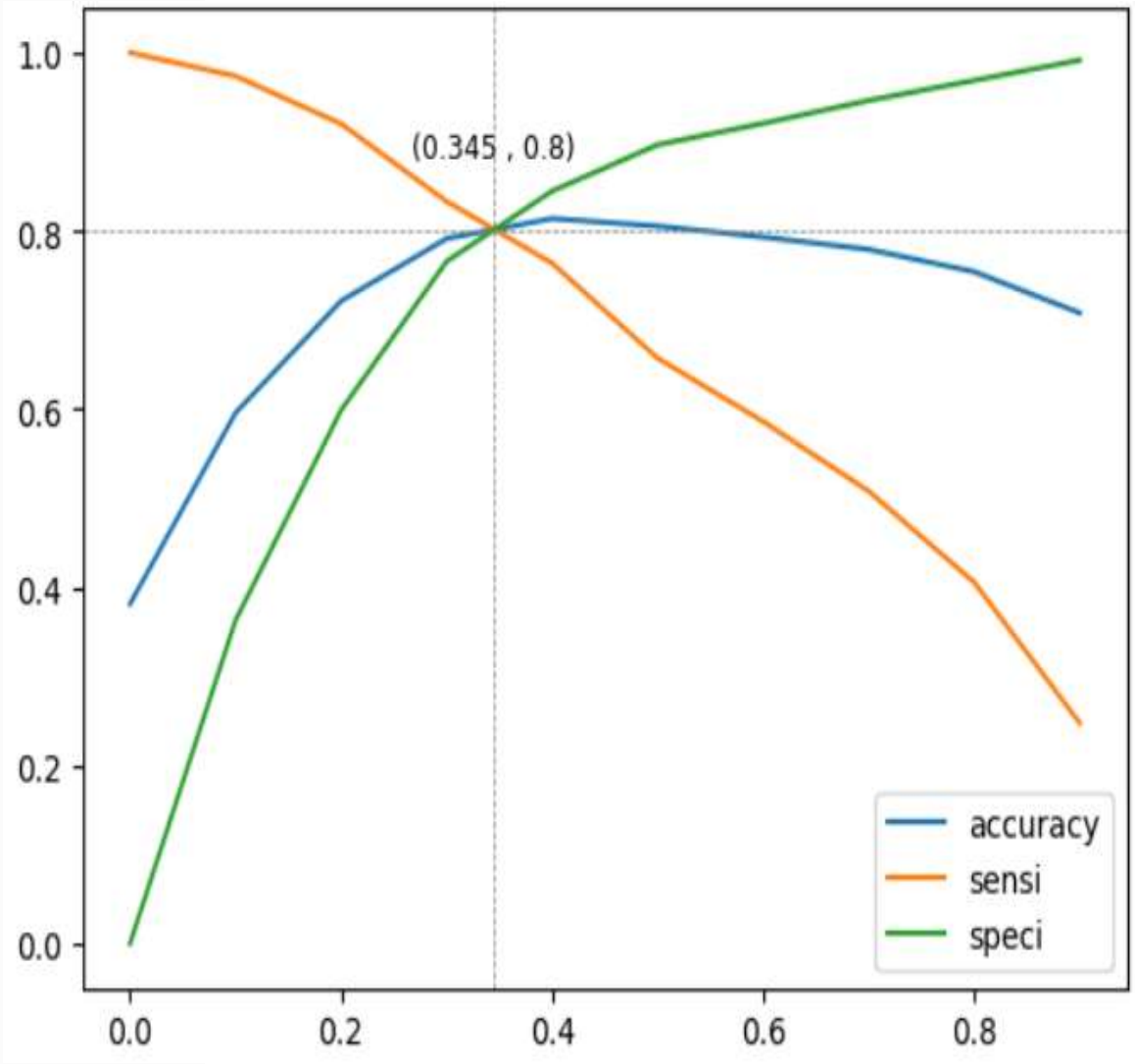
MODEL BUILDING

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

ROC Curve

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





PREDICTION ON TEST SET

- Before predicting on the test set, we need to standardize the test set and need to have exact same columns present in our final train dataset.
- After doing the above step, we started predicting the test set, and the new prediction values were saved in a new data frame.
- After this we did model evaluation i.e. finding the accuracy, precision, and recall.
- The accuracy score we found was 0.82, precision 0.75, and recall 0.75 approximately.
- This shows that our test prediction is having accuracy, precision, and recall scores in an acceptable range.
- This also shows that our model is stable with good accuracy and recall/sensitivity.
- Lead score is created on test dataset to identify hot leads – high the lead score higher the chance of conversion, low the lead score lower the chance of getting converted.

CONCLUSION

- The total time spent 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.

Thank you

- By Uchit Kumar
  Umal Kumar
  Tejal Patel