Lead Scoring Case Study

Using Logistic Regression

Authors

- Sairam Suravajhala
- Sainath Dekonda
- Pradeep Sajjan

Problem Statement

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
When these people fill up a form providing their email address or phone number, they are classified to be a lead which will be then passed to Sales team to start making calls or send emails to convert these leads. The typical lead conversion rate at X education is around 30%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. 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.

Goal

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.

Data Understanding & Data Cleaning

Insights

- Data has 9240 rows and 37 columns, and does not have duplicate records
- The features TotalVisits , Total Time Spent on Website , Page Views Per Visit have outliers
- There are 30 categorical and 7 numerical variables in dataframe
- 5 columns in dataframe has only same value across all records
- 17 columns have null values in the dataframe

Data Cleaning

- These are columns **How did you hear about X Education, Lead Quality**, **Lead Profile**, **Asymmetrique Activity Index, Asymmetrique Profile Index, Asymmetrique Activity Score, Asymmetrique Profile Score** are having more the 40% of null values.
- 7 columns have been removed from the dataframe , having more null values

Exploratory Data Analysis - EDA

Univariate Analysis - Categorical

- There are columns where the values are same for all records - Which would not contribute anything towards.

  [Model 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]

- There are columns where the values are Highly skewed towards single value - Which would not contribute anything towards Model.

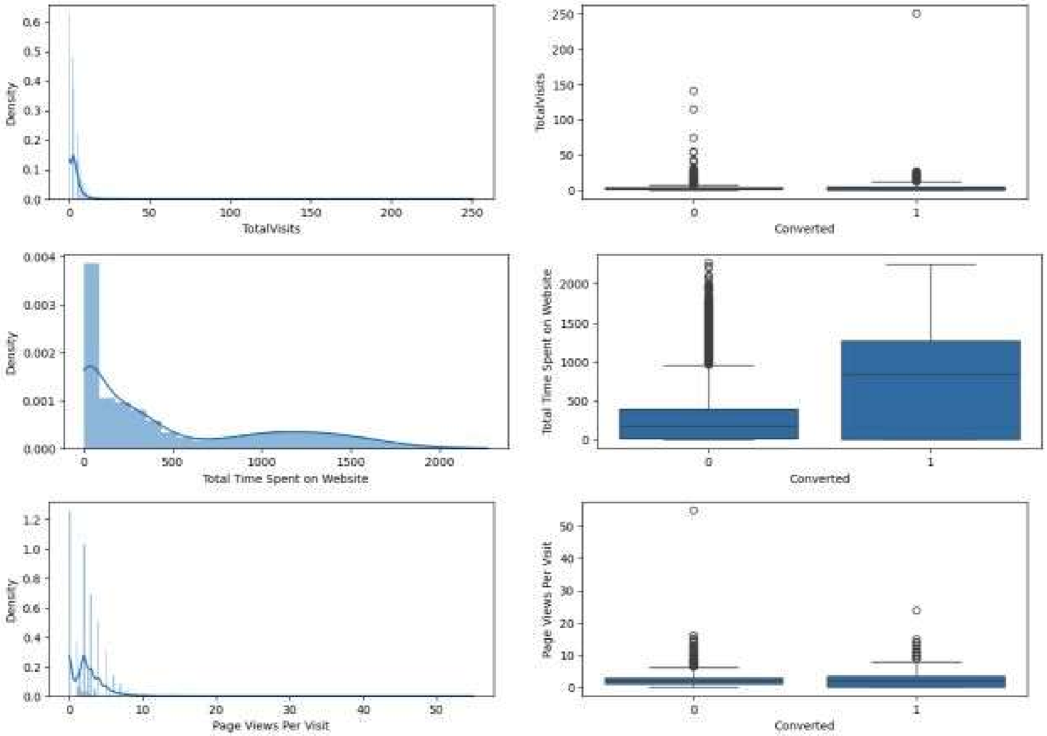
  [Do Not Call ,What matters most to you in choosing a course, Search, Newspaper Article, X Education Forums Newspaper Digital Advertisement Through Recommendations]

Univariate Analysis

Handling Missing Values: we still have lots of missing rows will dropped

City	39.71
Specialization	36.58
Tags	36.29
What matters most to you in choosing a course	29.32
Country	26.63
TotalVisits	1.48

Handling Outliers

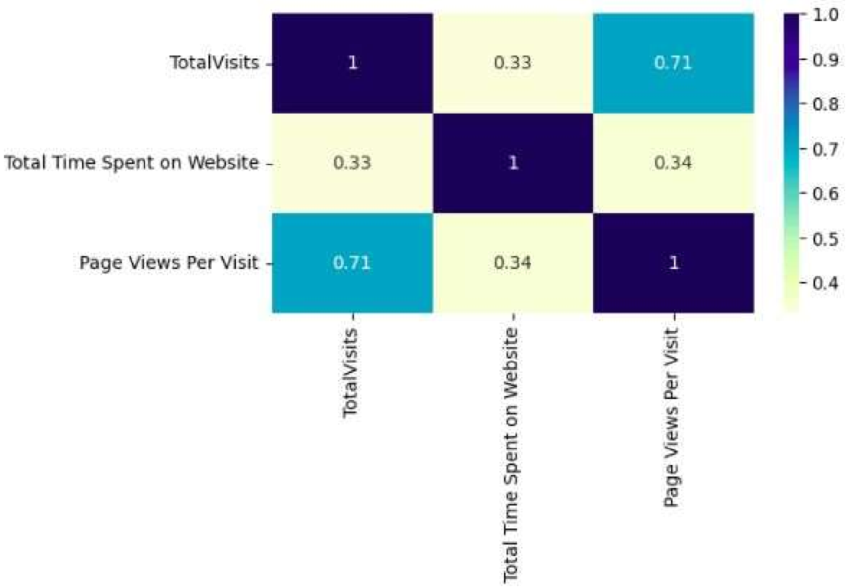


Handling Outliers

Insights

- We have found there are some outliers in
    - `TotalVisits`
    - `Total Time Spent on Website`
    - `Page Views Per Visit`

Bivariate Analysis - Numerical Values



Insights:

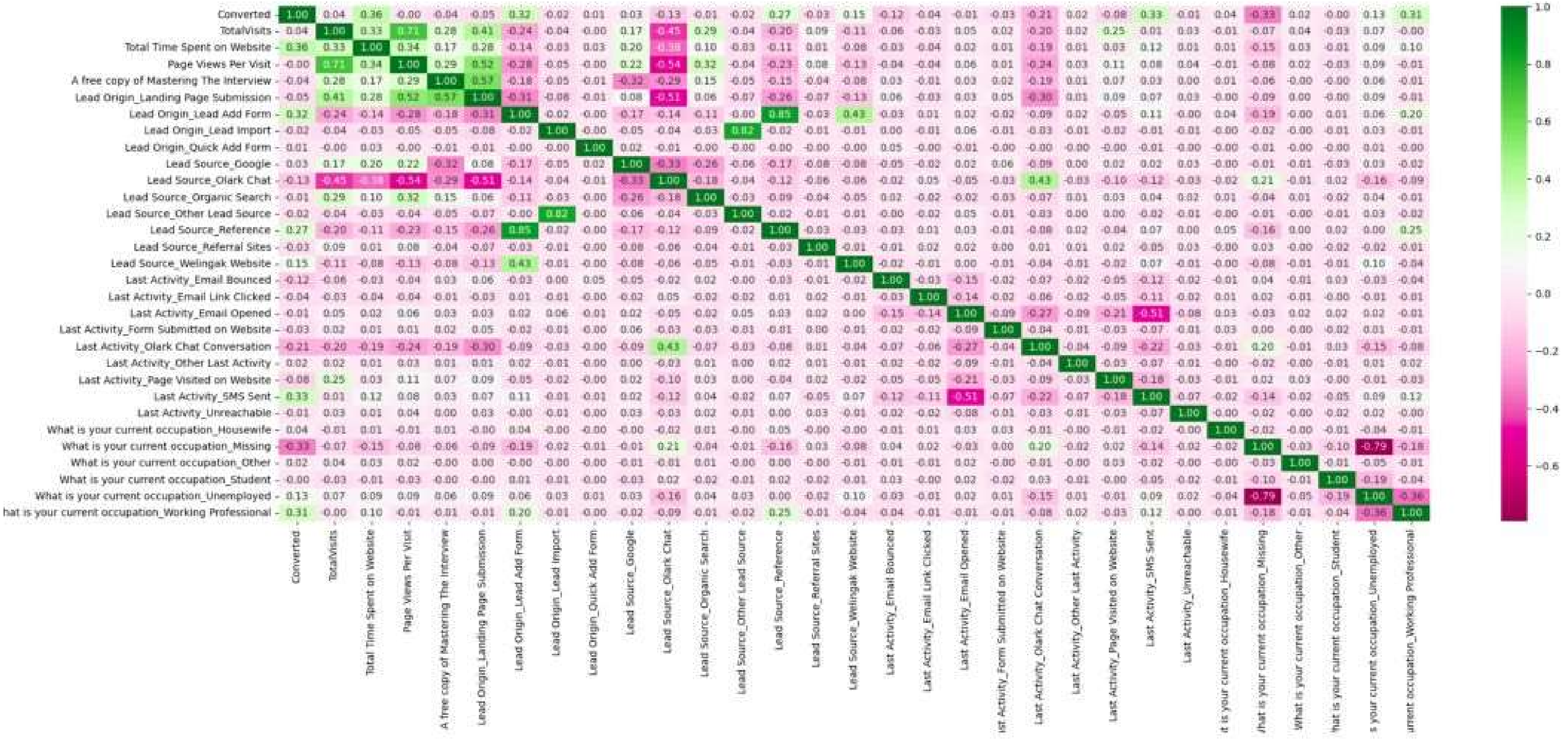
- There is a strong collinearity between both variables "Total Visits" and "Page views per visit" - We can also choose to drop one of the variable as keeping both of them does not significantly contribute to the model prediction.

Model Building - Data Preparation

Data Preparation

Dummy Variables - Create Dummy variables for categorical variables with multiple levels

Correlation Check for Dataframe

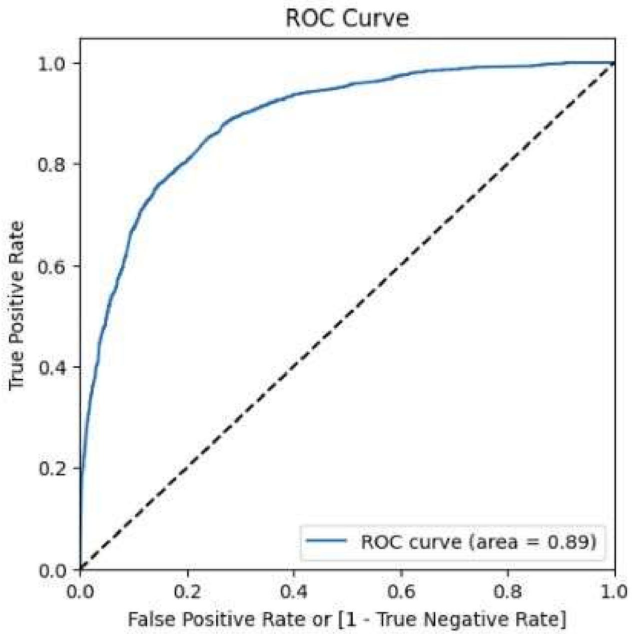


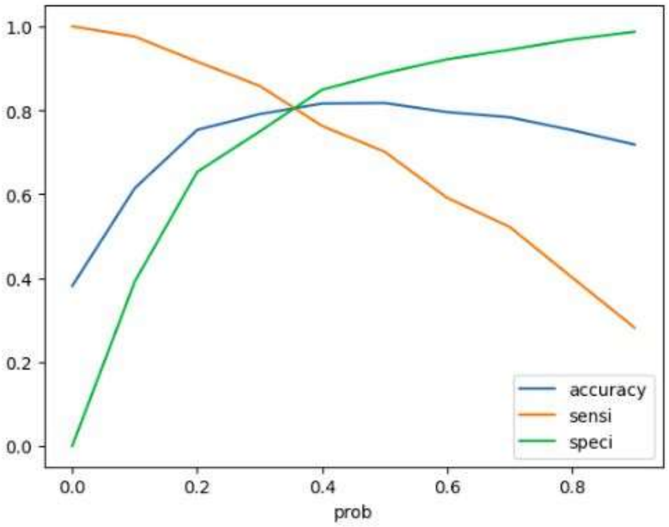
Model Building

Insights - Above Correlation img

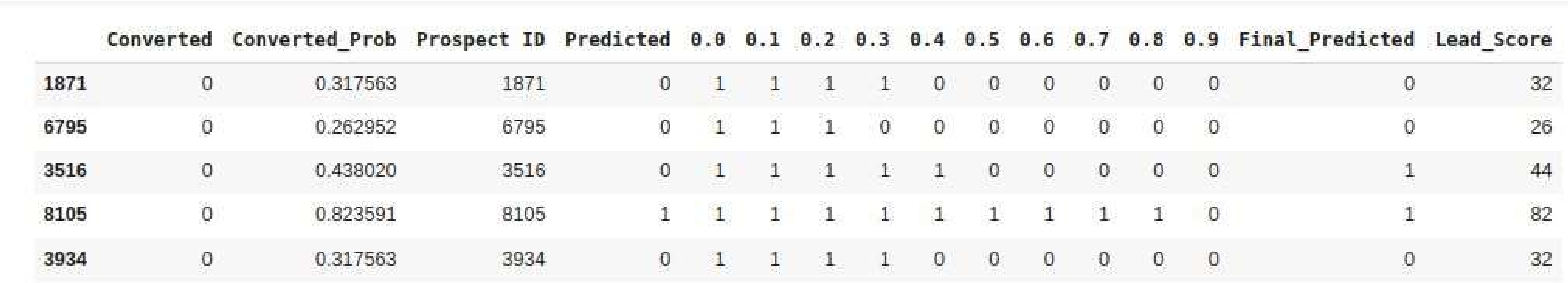
Since many of the features are already removed before EDA, not deleting one of the hightly correlated feature now. will handle further features deletion with RFE

Model Building - ROC Curve





Model Final Predictions



	Converted	Converted_Prob	Prospect ID	Predicted	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	Final_Predicted	Lead_Score
1871	0	0.317563	1871	0	1	1	1	1	0	0	0	0	0	0	0	32
6795	0	0.262952	6795	0	1	1	1	0	0	0	0	0	0	0	0	26
3516	0	0.438020	3516	0	1	1	1	1	1	0	0	0	0	0	1	44
8105	0	0.823591	8105	1	1	1	1	1	1	1	1	1	1	0	1	82
3934	0	0.317563	3934	0	1	1	1	1	0	0	0	0	0	0	0	32

Conclusion

- After trying few models, model with below characteristics is chosen
- All Features have p-value < 0.05, and have very low VIF Values ensuring no multicollinearity among features.
- With probability threshold @ 0.35, Model accuracy of 80.67 for train data , and 79.73% for test data is acceptable
- The Conversion probability of a lead increases with increase in values for following features in Descending Order:
    - Total Time Spent on Website
    - Lead Origin_Lead Add Form
    - What is your current occupation_Working Professional
    - Lead Source_Welingak Website
    - Last Activity_SMS Sent
    - Lead Source_Olark Chat
    - Last Activity_Other Last Activity
    - TotalVisits
    - Last Activity_Unreachable
    - Last Activity_Email Opened