Lead score case study

Group Members

Arjun R

Annu Chauhan

Aparna Nimbalkar

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.

Methodology

- Data cleaning and data manipulation

  Check and handle duplicate data.

  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

  Univariate data analysis: value count, distribution of variable etc.

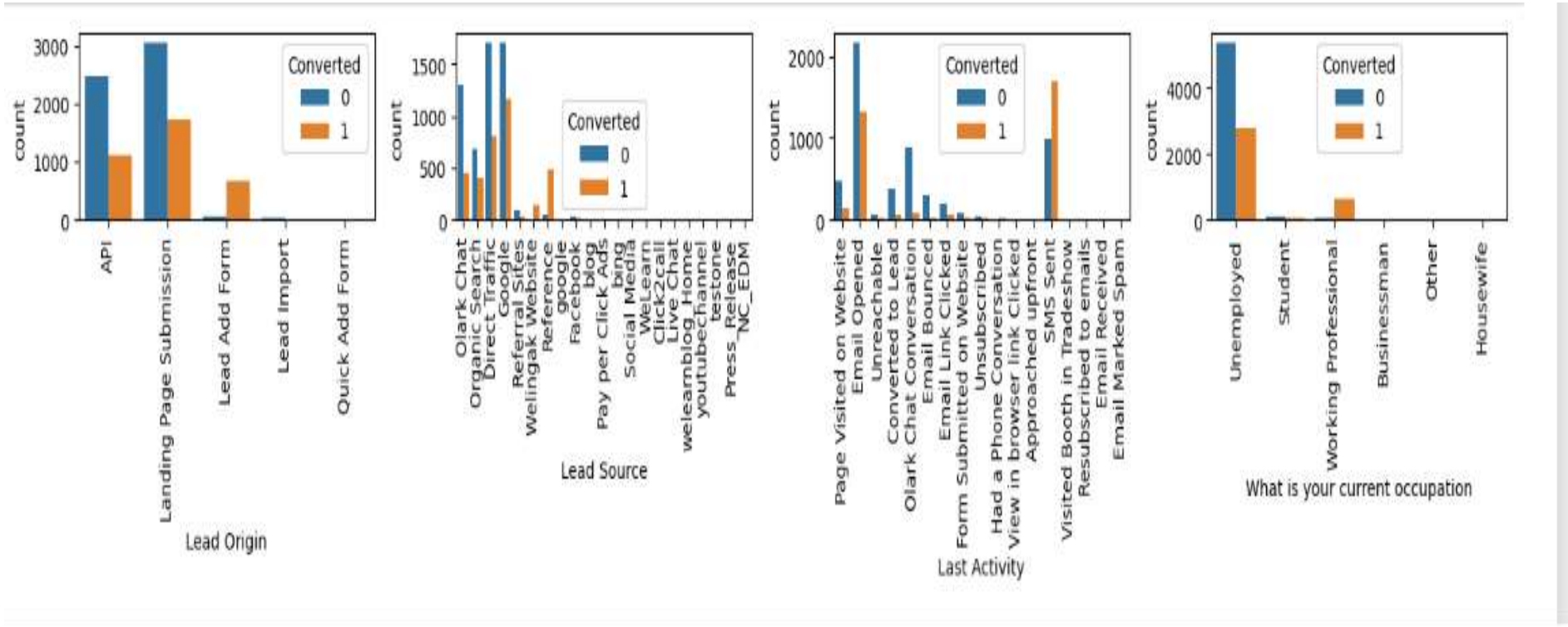
  Bivariate data analysis: correlation coefficients and pattern between the variables etc.

- 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.

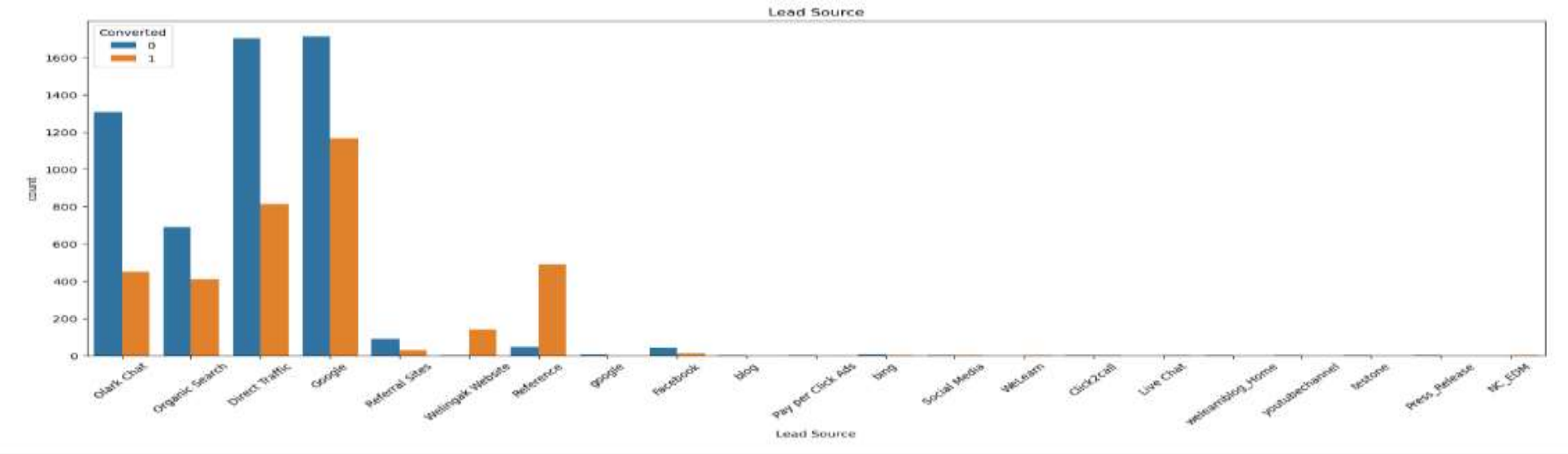
Data Manipulation

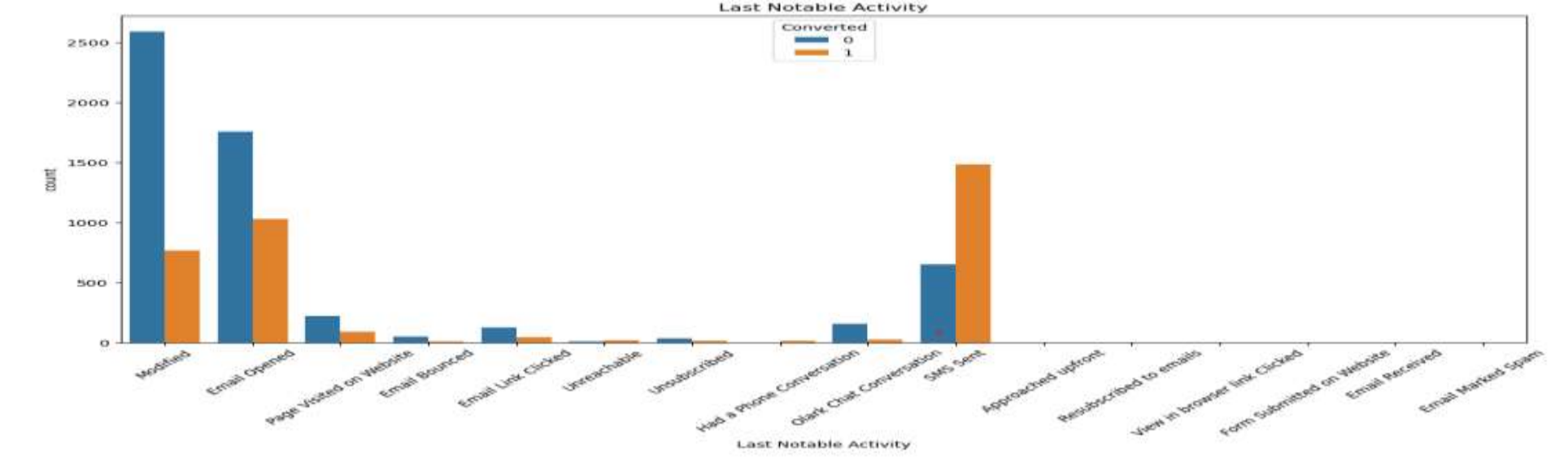
- Total Number of Rows =37, Total Number of Columns =9240.
- The level "Select" appearing in few columns. These are null values and won't be helpful to our investigation, we can impute NaN and ignore them
- Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis.
- Single value features like “predominated throughout the majority of the data points.These include Do Not Call, Search, Magazine, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement, Through Recommendations, 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 have been dropped.
- Dropping the columns having more than 45% as missing value such as ‘How did you hear about X Education’ and ‘Lead Profile’.

EDA


count
Converted
0
1
API
Landing Page Submission
Lead Add Form
Lead Import
Quick Add Form
Lead Origin
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Welingak Website
Reference
google
Facebook
blog
Pay per Click Ads
bing
Social Media
WeLearn
Click2call
Live Chat
welearnblog_Home
youtubechannel
testone
Press_Release
NC_EDM
Lead Source
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Had a Phone Conversation
View in browser link Clicked
Approached upfront
SMS Sent
Visited Booth in Tradeshow
Resubscribed to emails
Email Received
Email Marked Spam
Last Activity
Unemployed
Student
Working Professional
Businessman
Other
Housewife
What is your current occupation

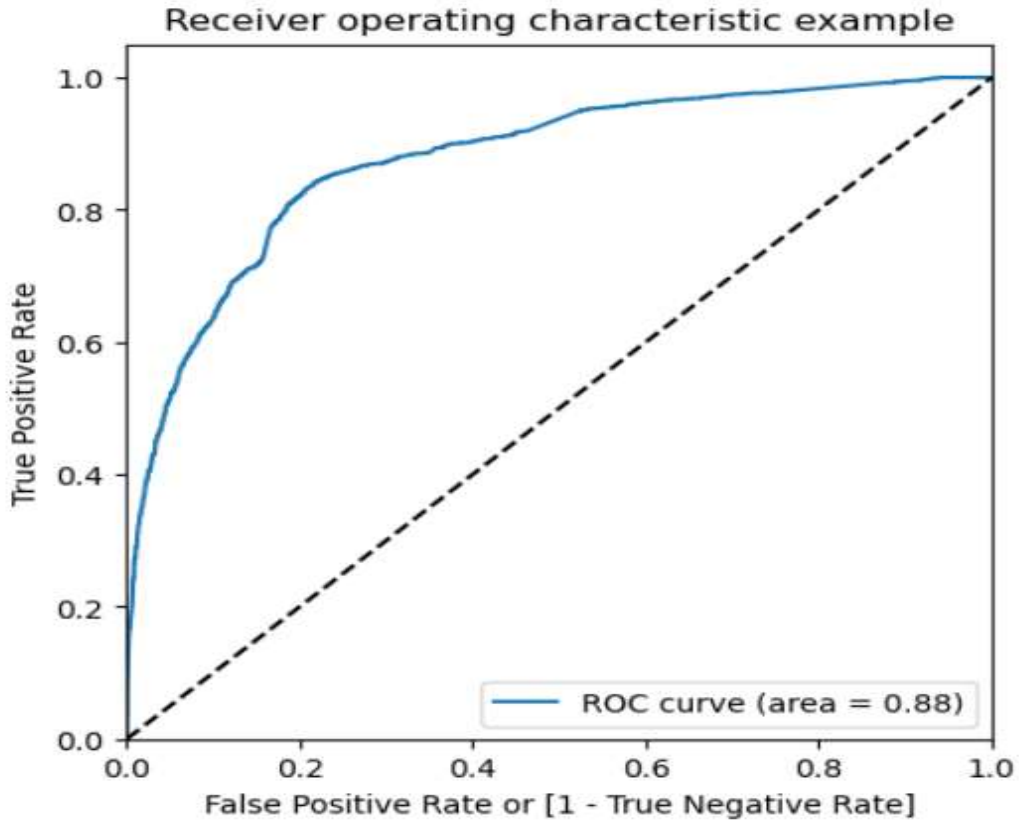
Categorical Variable Relation



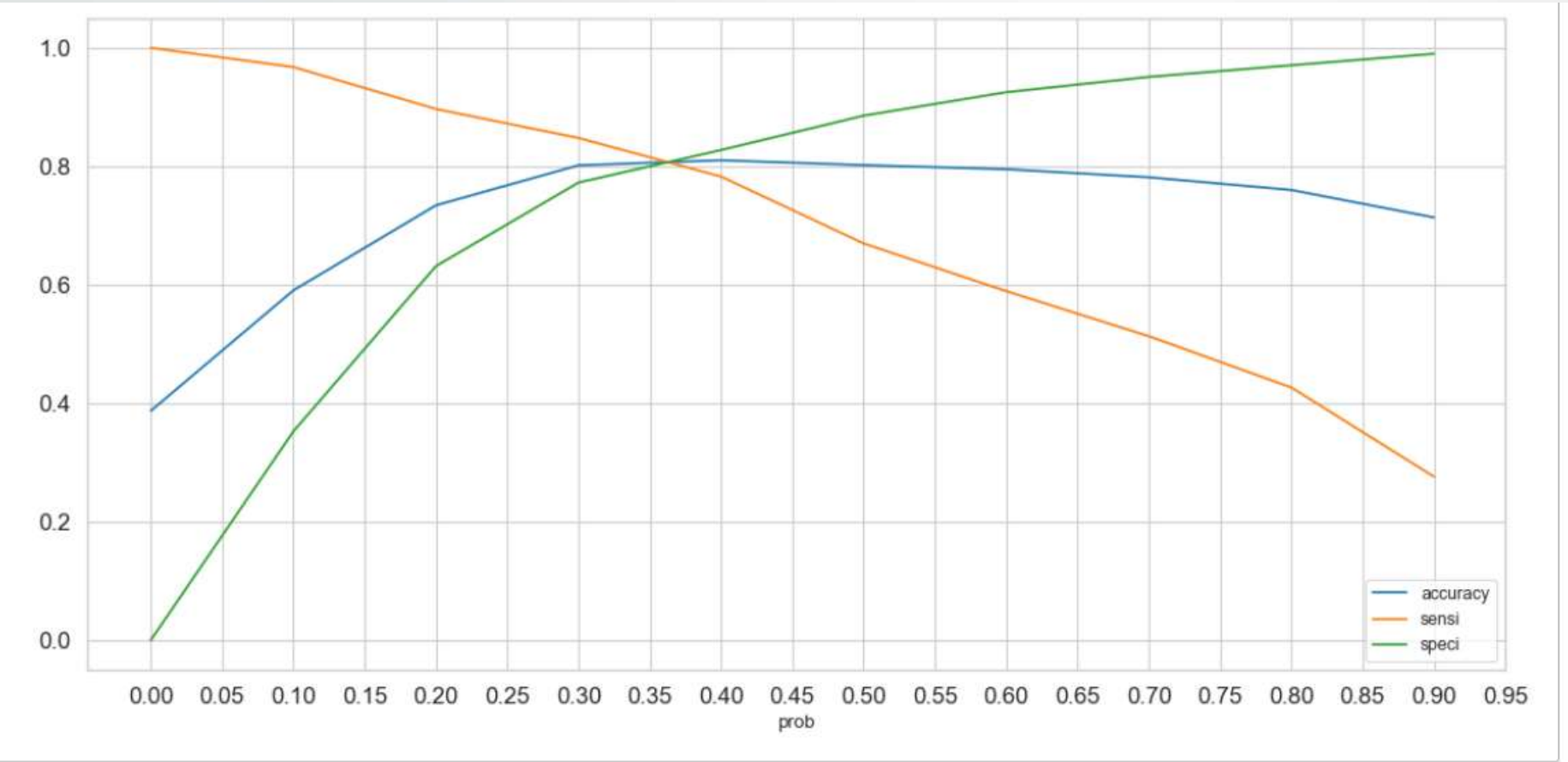


Model Building

- 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 remaining variables as output
- Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5
- Predictions on test data set
- Overall accuracy 80%



The area under the curve of the ROC is nearly equal to 1 which is quite good. so we seem to have a good model.



Evaluating the model with optimal probability cutoff as 0.36

Conclusion

- It was found that the variables that mattered the most in the potential buyers are :
  - The total time spend on the Website.
  - Total number of visits.
  - Page Views Per Visit
- 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.

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.