Lead Scoring Case Study Presentation

By:
Sharath Chandra VC
Md Ibrahim Shariff

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

Data Cleaning and Data Manipulation.

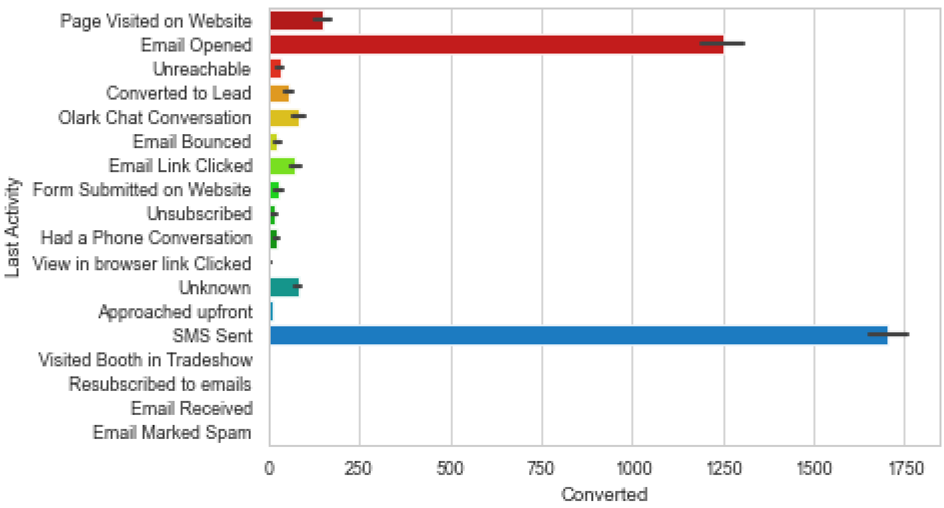
- Check and handle duplicate data.
- Check and handle NA values and missing values.
- Drop columns, if it contains large amount of missing values and not useful for the analysis.
- Imputation of the values, if necessary.
- Check and handle outliers in data.

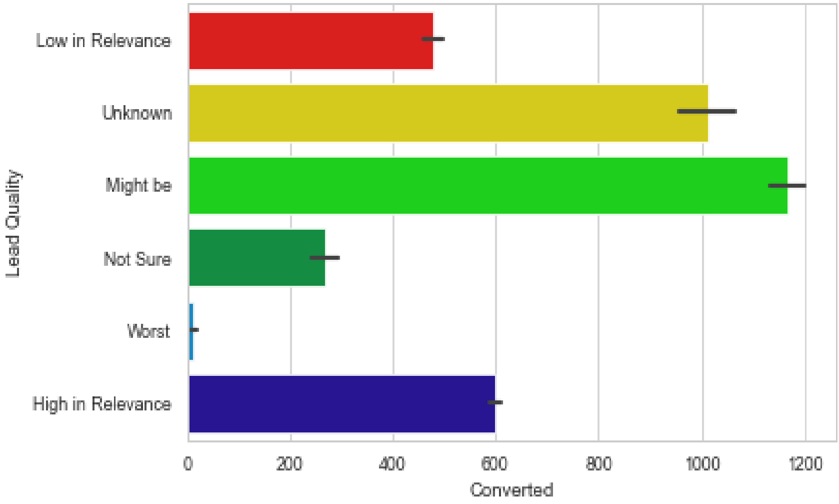
EDA:

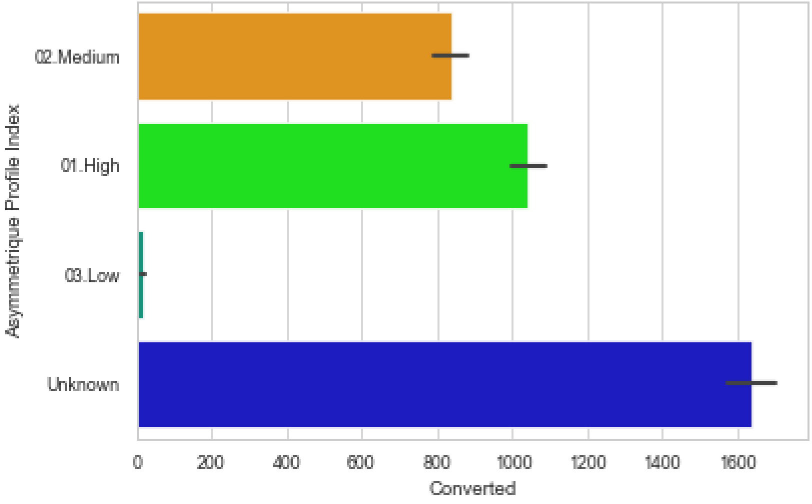
1. Univariate data analysis: value count, distribution of variable etc.
2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.

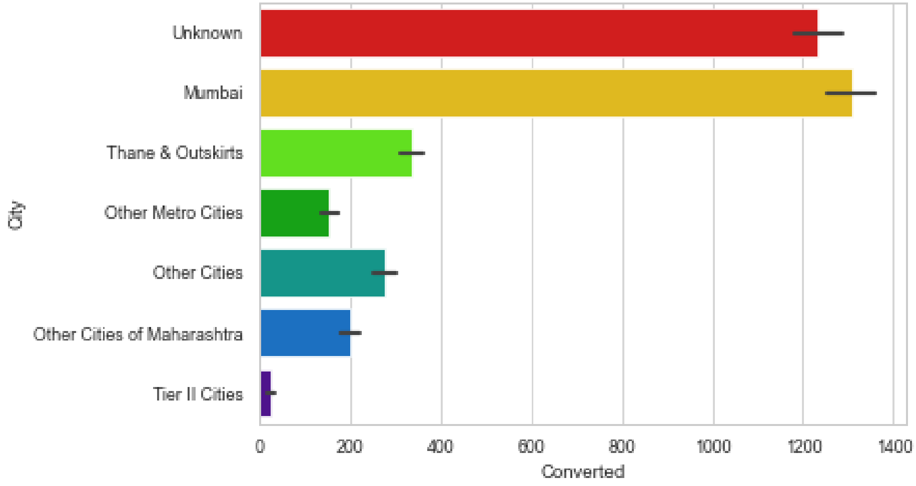
- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique: logistic regression used for the model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.

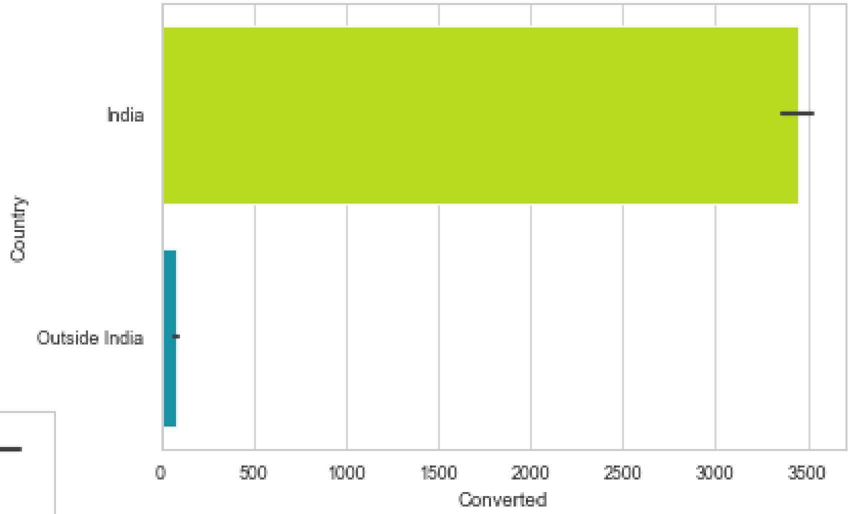
EDA

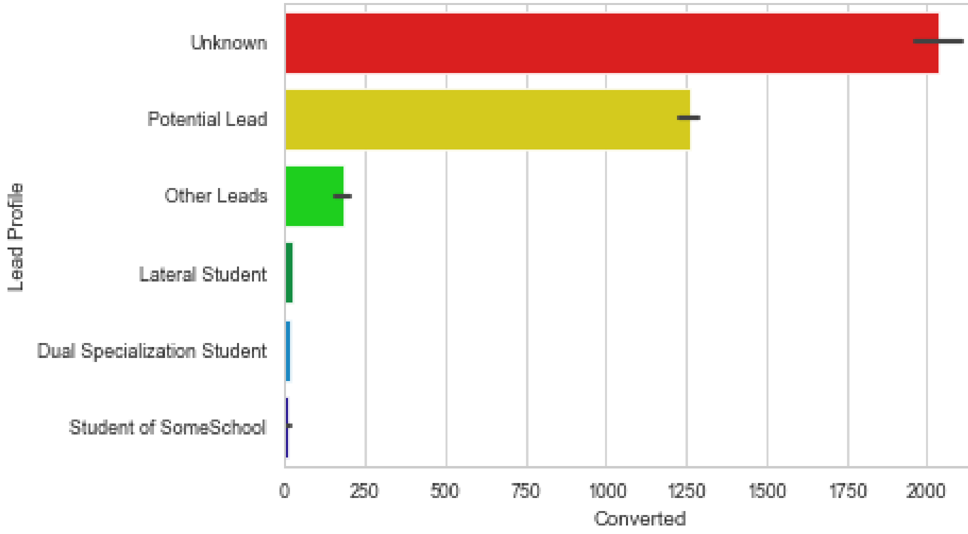

Last Activity
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Had a Phone Conversation
View in browser link Clicked
Unknown
Approached upfront
SMS Sent
Visited Booth in Tradeshow
Resubscribed to emails
Email Received
Email Marked Spam
0
250
500
750
1000
1250
1500
1750
Converted

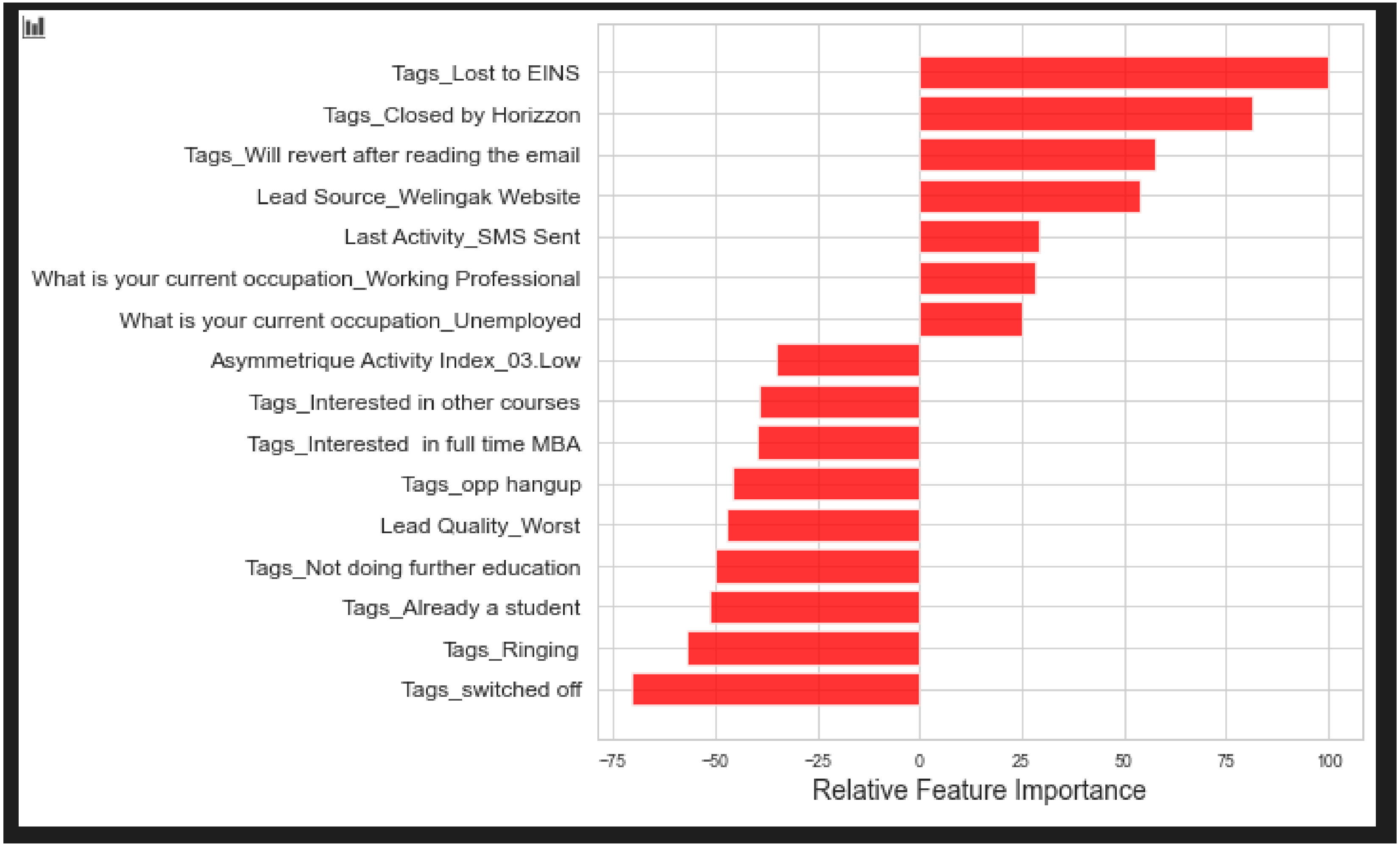

Lead Quality
Low in Relevance
Unknown
Might be
Not Sure
Worst
High in Relevance
0
200
400
600
800
1000
1200
Converted


Asymmetrique Profile Index
02.Medium
01.High
03.Low
Unknown
0
200
400
600
800
1000
1200
1400
1600
Converted


City
Unknown
Mumbai
Thane & Outskirts
Other Metro Cities
Other Cities
Other Cities of Maharashtra
Tier II Cities
0
200
400
600
800
1000
1200
1400
Converted


Country
India
Outside India
0
500
1000
1500
2000
2500
3000
3500
Converted


Lead Profile
Unknown
Potential Lead
Other Leads
Lateral Student
Dual Specialization Student
Student of SomeSchool
0
250
500
750
1000
1250
1500
1750
2000
Converted



Plot showing the feature variables based on their relative coefficient values

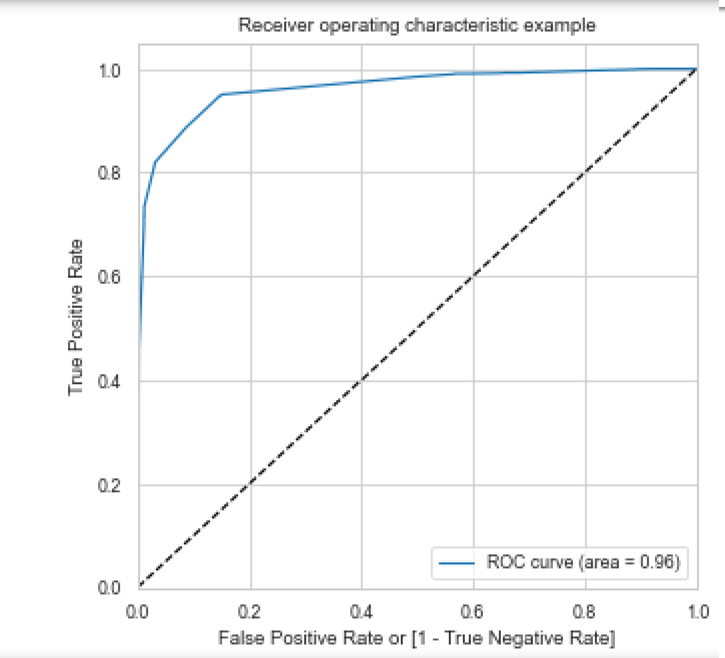
Data Conversion

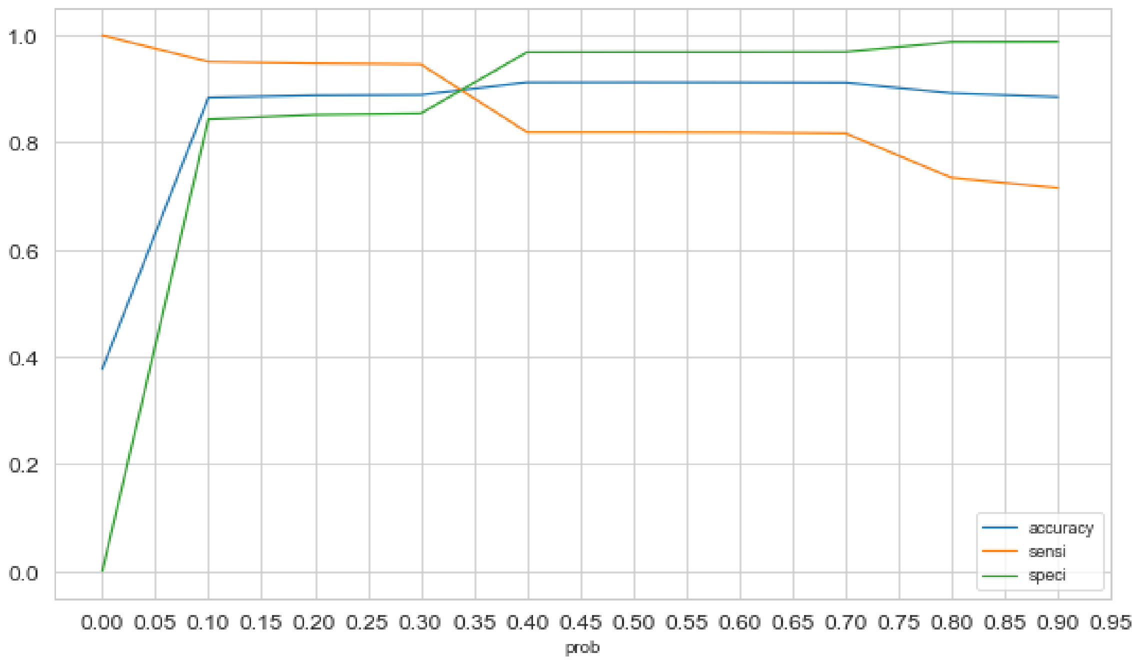
- Numerical Variables are Normalized
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 8792
- Total Columns for Analysis: 43

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



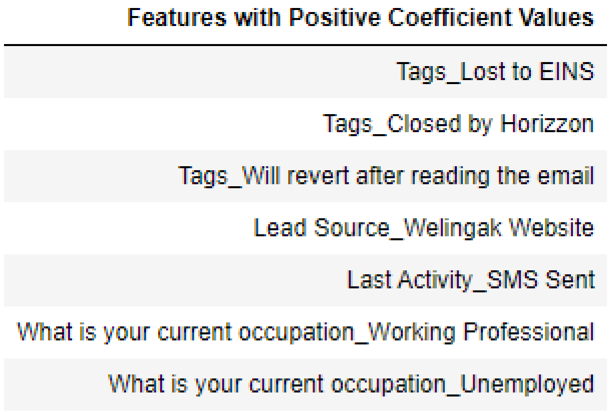


Finding Optimal Cut off Point

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

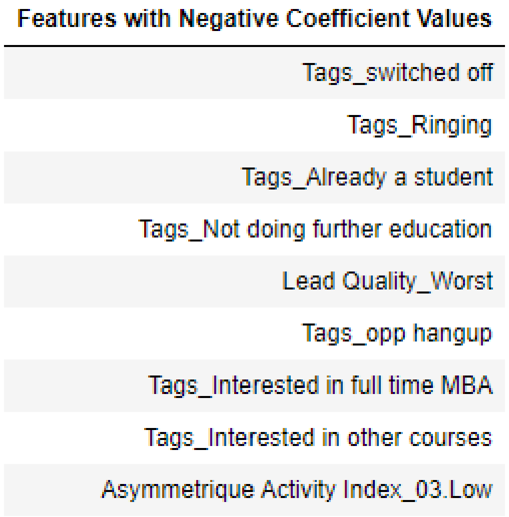
Conclusion

The conversion probability of a lead increases with increase in values of the following features in descending order:



Features with Positive Coefficient Values
Tags_Lost to EINS
Tags_Closed by Horizzon
Tags_Will revert after reading the email
Lead Source_Welingak Website
Last Activity_SMS Sent
What is your current occupation_Working Professional
What is your current occupation_Unemployed

The conversion probability of a lead increases with decrease in values of the following features in descending order:



Features with Negative Coefficient Values
Tags_switched off
Tags_Ringing
Tags_Already a student
Tags_Not doing further education
Lead Quality_Worst
Tags_opp hangup
Tags_Interested in full time MBA
Tags_Interested in other courses
Asymmetrique Activity Index_03.Low

- **Another point to note here is that, depending on the business requirement, we can increase or decrease the probability threshold value with in turn will decrease or increase the Sensitivity and increase or decrease the Specificity of the model.**

- **High Sensitivity will ensure that almost all leads who are likely to Convert are correctly predicted where as high Specificity will ensure that leads that are on the brink of the probability of getting Converted or not are not selected.**

Thank you ☺