Lead Score case study

By:

Geetakrishnasai Gunapati

Pramodini V Nayak

Abstract

Problem statement:

An education company named X Education sells online courses to industry professionals. Though X Education lands a lot of people on their website and gets them to provide their email address or phone number, their lead conversion rate(paying customer) is very poor.

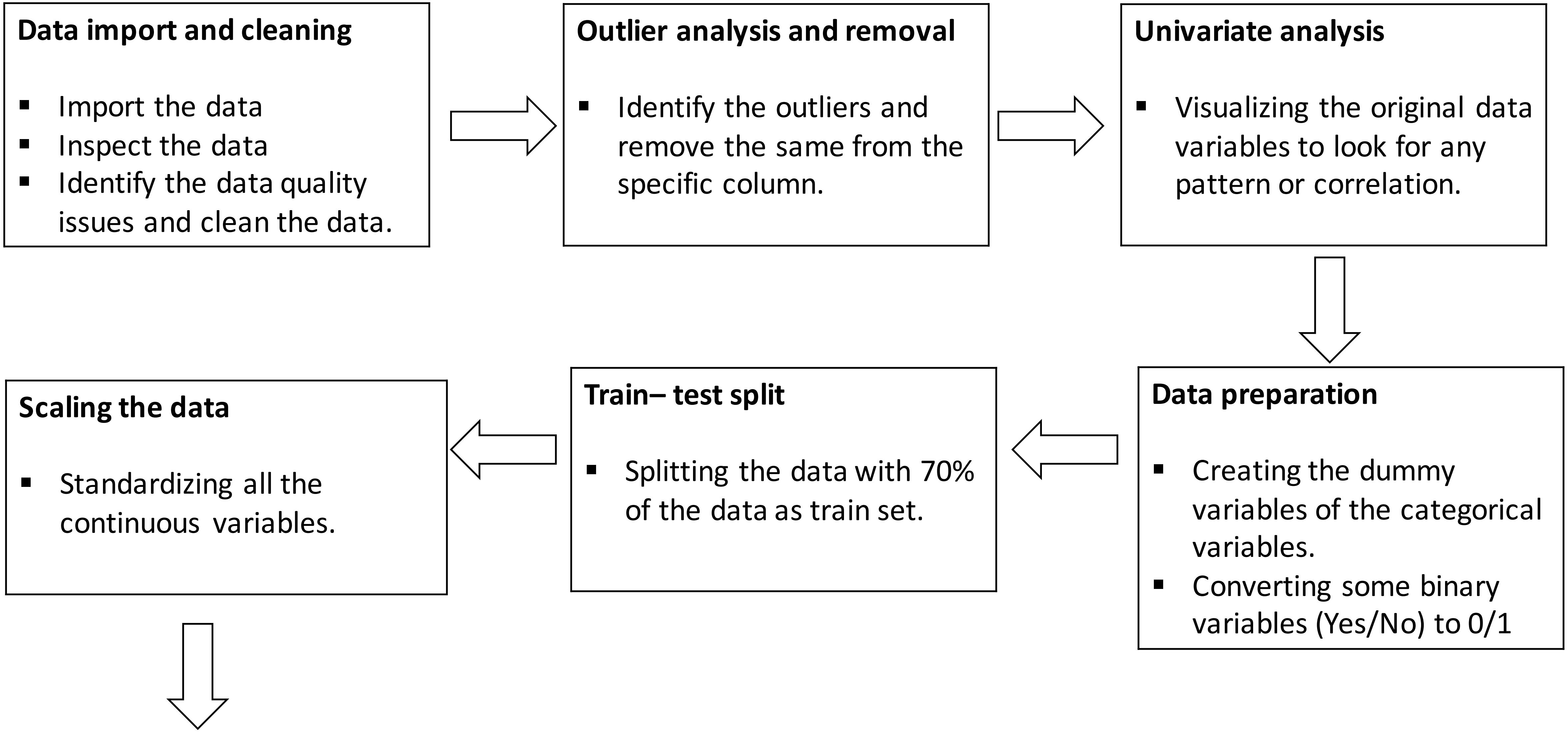
Objective:

To build a model such that a lead score is assigned to each of the leads in such a way that the customers with higher lead score have a higher conversion chance, and the customers with lower lead score have a lower conversion chance. And the target lead conversion rate is around 80%.

Data used for analysis:

Leads dataset from the past with around 9240 data points.

Analysis methodology



Data import and cleaning

- Import the data
- Inspect the data
- Identify the data quality issues and clean the data.

Outlier analysis and removal

- Identify the outliers and remove the same from the specific column.

Univariate analysis

- Visualizing the original data variables to look for any pattern or correlation.

Data preparation

- Creating the dummy variables of the categorical variables.
- Converting some binary variables (Yes/No) to 0/1

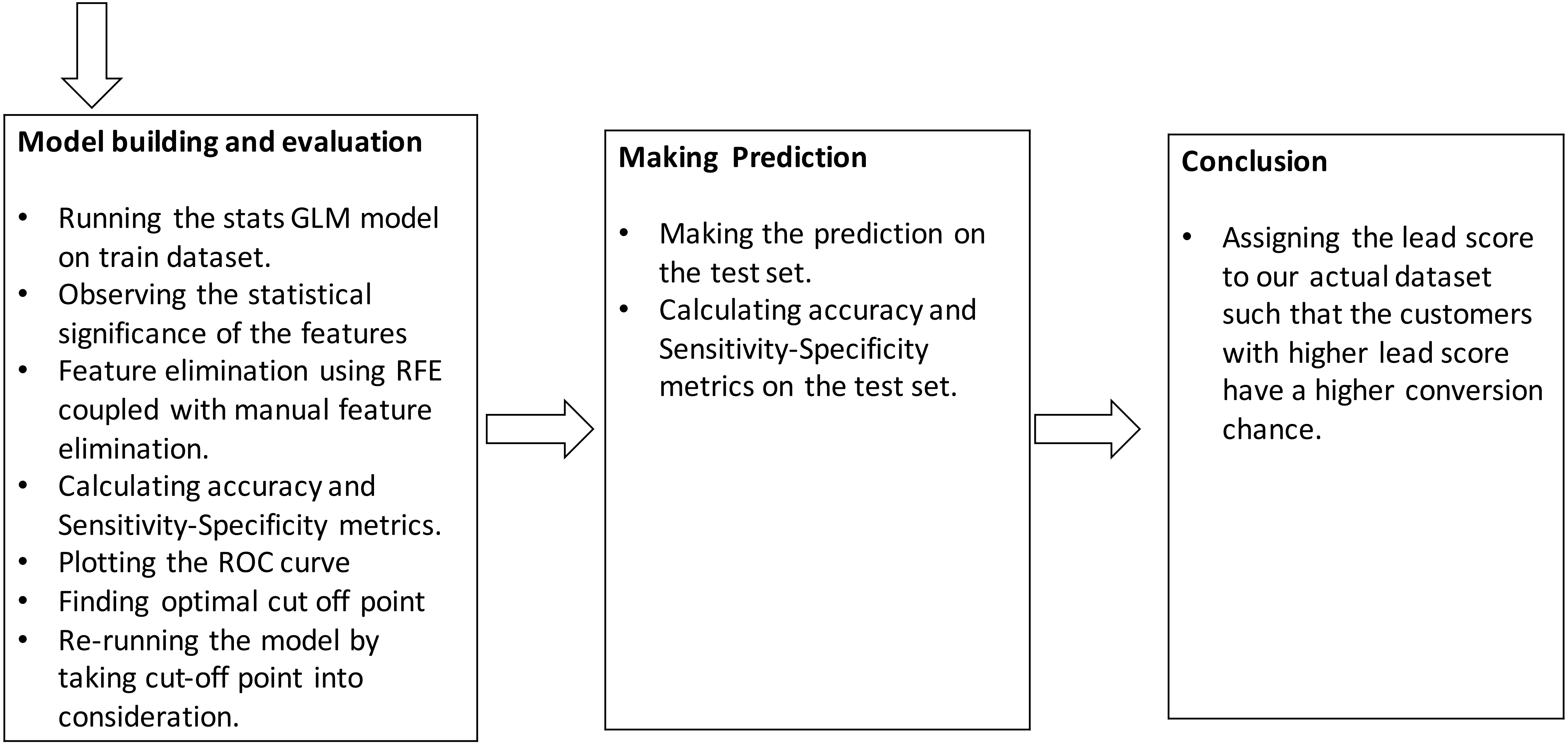
Train– test split

- Splitting the data with 70% of the data as train set.

Scaling the data

- Standardizing all the continuous variables.

Analysis methodology Cont.



Model building and evaluation

- Running the stats GLM model on train dataset.
- Observing the statistical significance of the features
- Feature elimination using RFE coupled with manual feature elimination.
- Calculating accuracy and Sensitivity-Specificity metrics.
- Plotting the ROC curve
- Finding optimal cut off point
- Re-running the model by taking cut-off point into consideration.

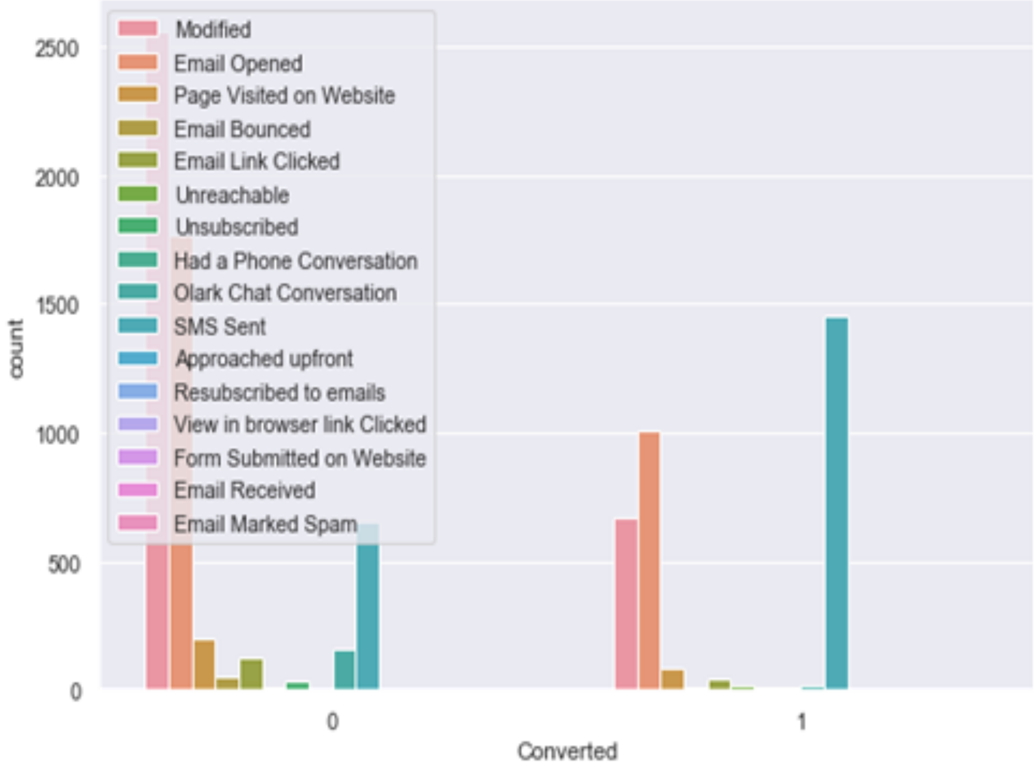
Making Prediction

- Making the prediction on the test set.
- Calculating accuracy and Sensitivity-Specificity metrics on the test set.

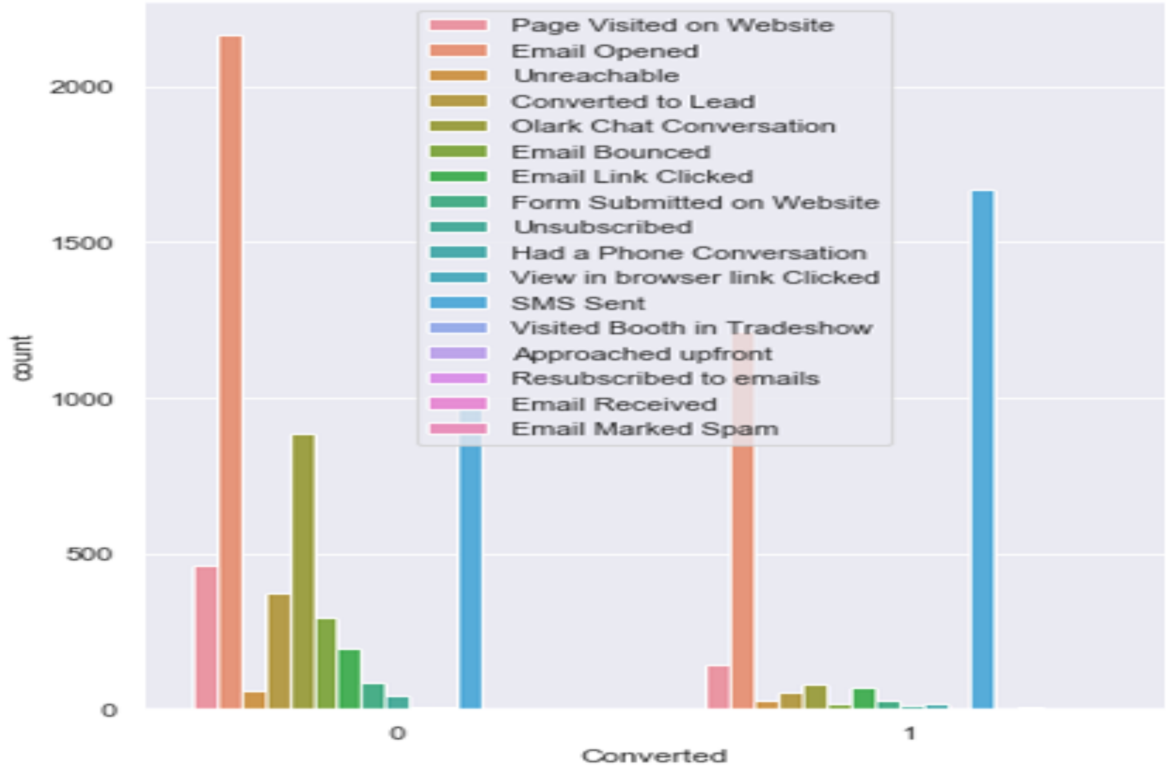
Conclusion

- Assigning the lead score to our actual dataset such that the customers with higher lead score have a higher conversion chance.

Univariate Analysis



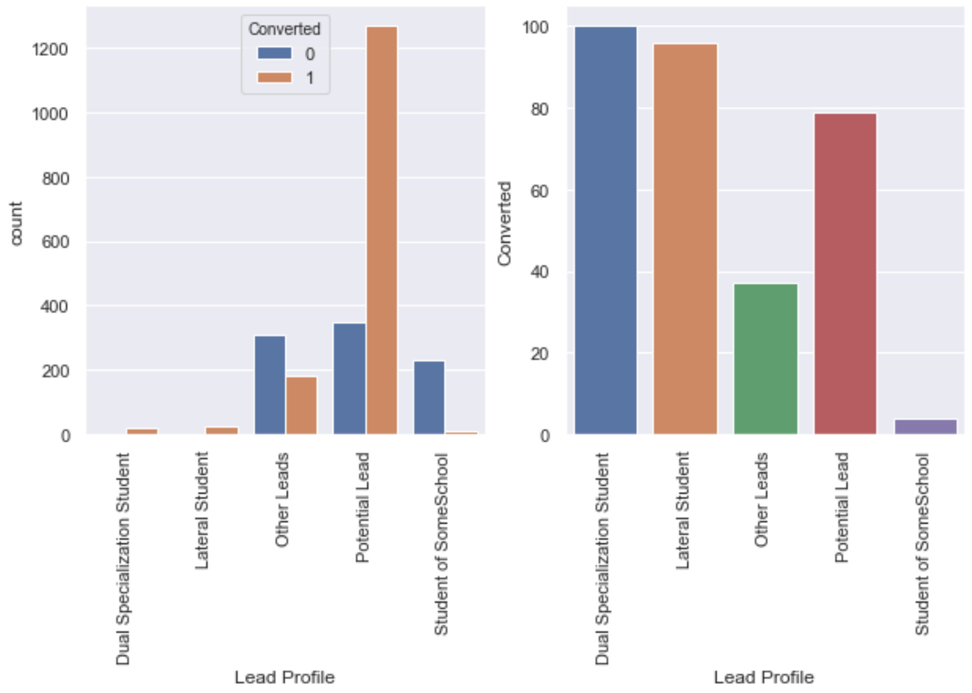
Converted versus Last Notable Activity



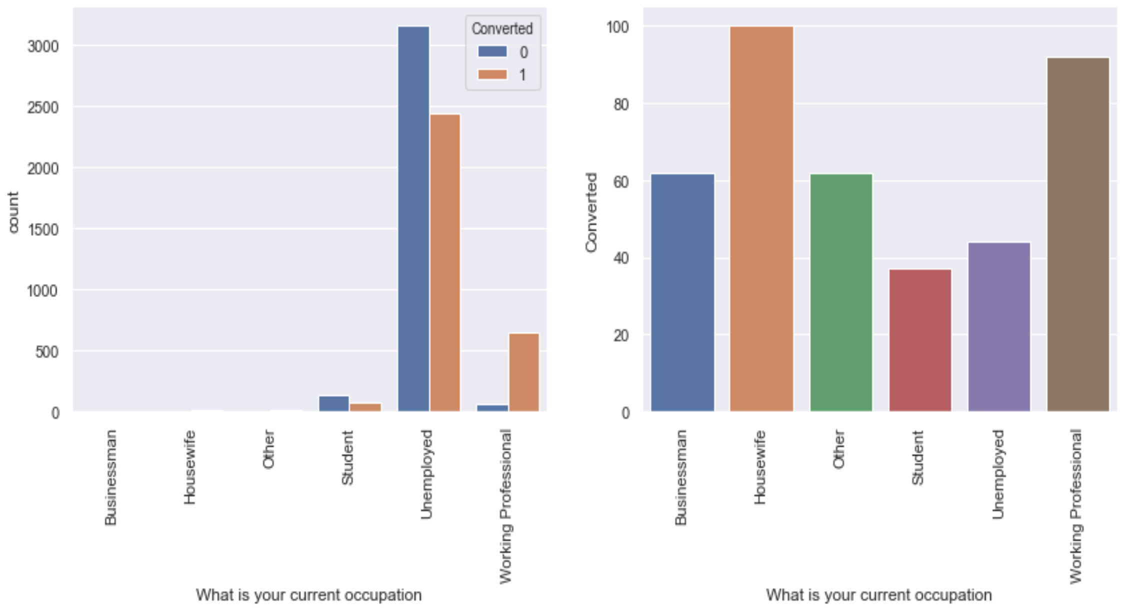
Converted versus Last Activity

- Converted 1 – indicates successful conversion.
- We see that 'Olark chat conversation' has the highest successful conversion count compared to other last notable activities.
- Similarly, SMS sent has higher successful conversion count compared to other last activities.

Univariate Analysis



Converted versus Lead Profile



Converted versus Profession

- Dual Specialization students and Lateral students have a very high conversion rate, hence emphasis should be made to acquire more such individuals
- Working Professionals have a higher conversion rate than the non working individuals therefore efforts should be made for reaching out to more working professionals

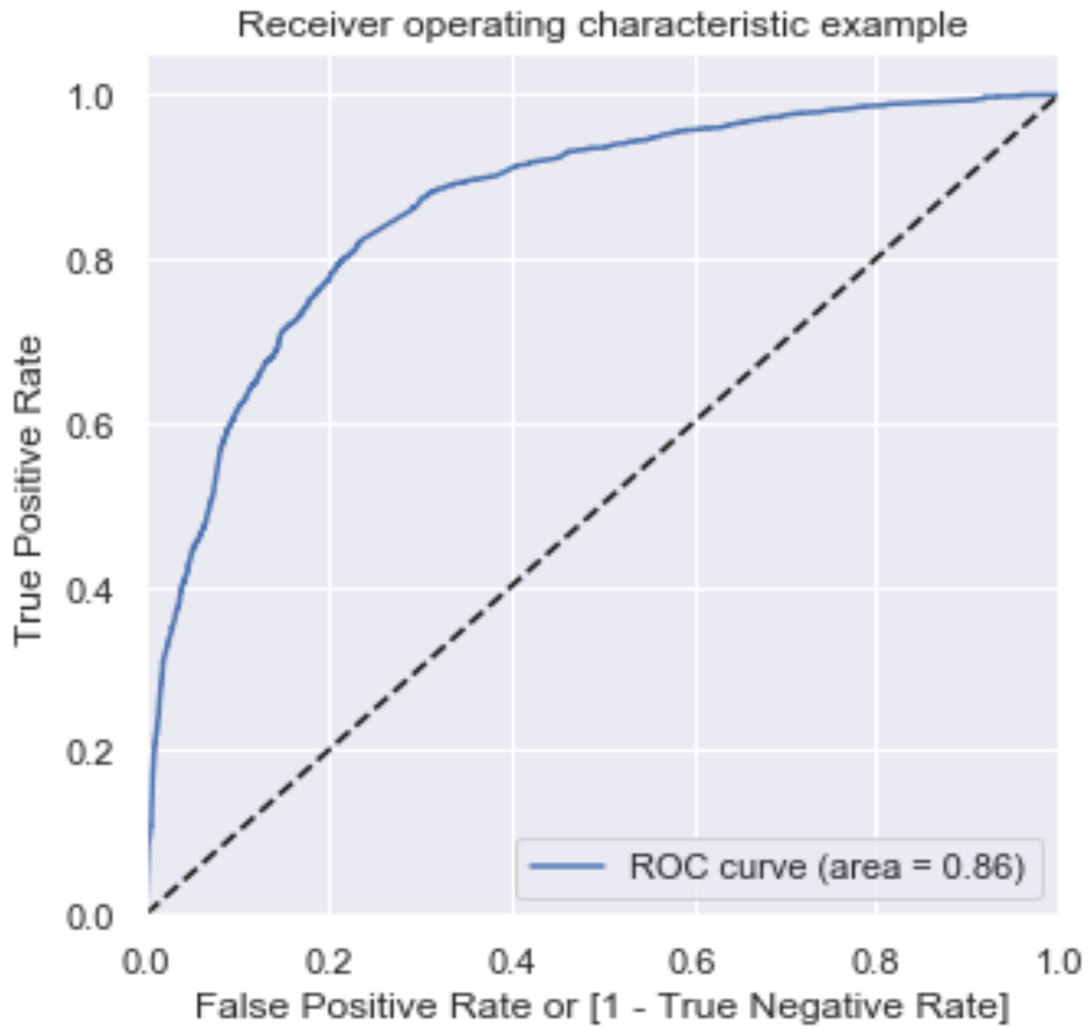
RFE Model

Equation for **'Converted'** from our RFE model building coupled with manual feature elimination is:

-0.2543 + 1.1457 x Total Time Spent on Website + 4.5445 x Lead Origin_Lead Add Form + 1.3501 x Lead Origin_Lead Import + 1.2437 X Lead Source_Olark Chat -2.0781 x Last Activity_Email Bounced + 1.9347 x Last Activity_Had a Phone Conversation -1.135 x Last Activity_Olark Chat Conversation + 0.4736 x Last Activity_SMS Sent -1.2083 x Last Activity_Unsubscribed – 1.4526x Last Notable Activity_Email Link Clicked -0.8938 x Last Notable Activity_Email Opened -1.4998 x Last Notable Activity_Modified -1.4098 x Last Notable Activity_Olark Chat Conversation -1.3276 x Last Notable Activity_Page Visited on Website

- ✓ **Lead Origin_Lead Add Form, Last Activity_Had a Phone Conversation and Lead Origin Lead Import** are the variables in the model which should be focused the most on in order to increase the probability of lead conversion.

ROC Curve

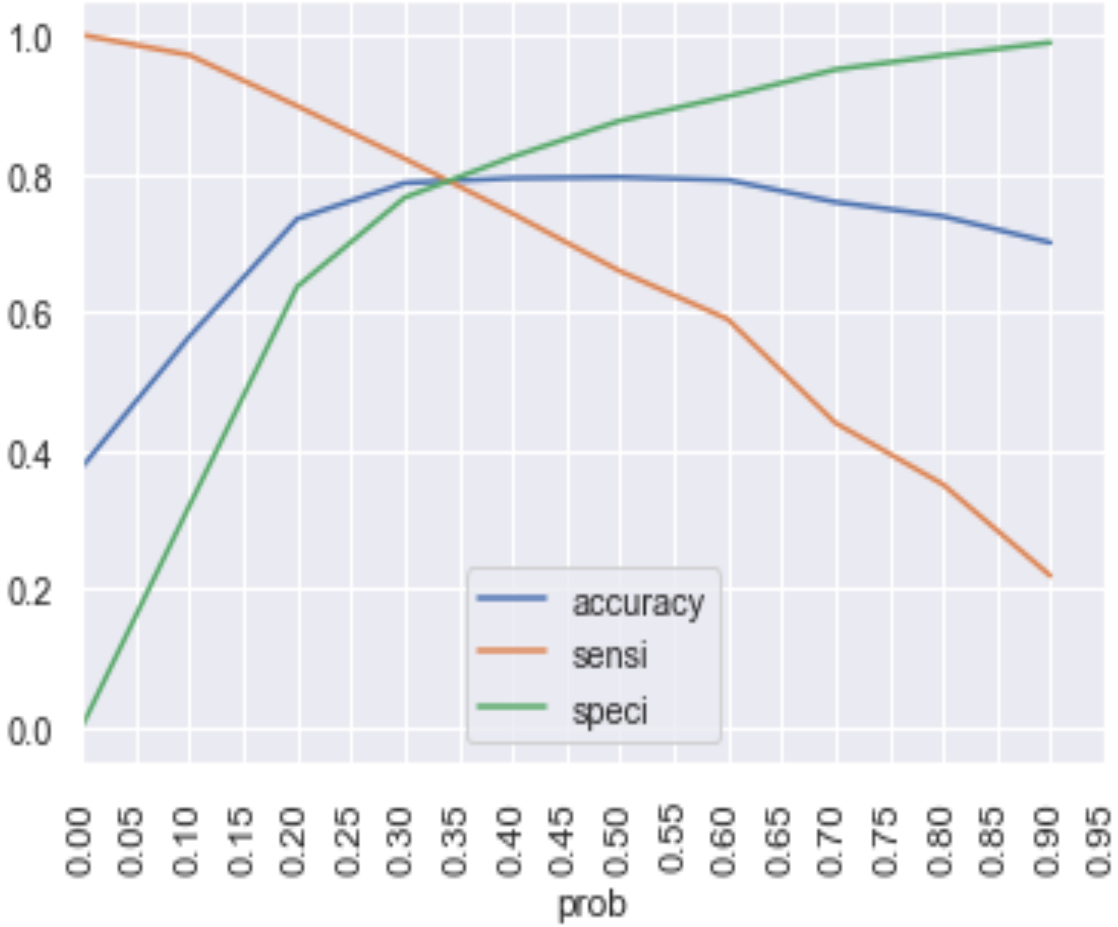


ROC curve shows trade off between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).

As we see, the curve is closer to the left-hand border and then the top border of the ROC space. This shows the accuracy of our model.

And the area under our ROC curve is 0.86.

Optimal Cut-off Point



As we can see, when the probability thresholds are very low, the sensitivity is very high and specificity is very low. Similarly, for larger probability thresholds, the sensitivity values are very low but the specificity values are very high. And at about 0.34, the three metrics seem to be almost equal with decent values and hence, we choose 0.34 as the optimal cut-off point.

Model Evaluation Metrics

Metrics	Train set	Test set	Final Model
Sensitivity	83.83	79.04	79.04
Specificity	78.43	78.66	78.66
False positive rate	21.5	21.33	21.33
Positive predictive value	0.7	68.98	68.98
Negative predictive value	88.98	86.2	86.2

Metrics across train set, test set and our final model has consistent value.

Recommendations

- ✓ Emphasis should be made on targeting working professionals as they have higher rate of conversion
- ✓ Lead Origin from Lead Add Form and Lead Import also has higher positive impact on conversion hence more efforts should be made to focus on individuals from these origin points
- ✓ Olark Chat seems to be a pretty good lead source hence more focus should be put toward this source for acquiring potential leads
- ✓ Individual who had a phone conversation seems to join the course more hence efforts should me made to reach out to potential leads in order to increase the conversion rate
- ✓ It can be seen that the more time individuals spend on the website the more is the chance of conversion, hence more emphasis should be made in targeting the individuals with last notable activity as visited website