LEAD SCORING CASE STUDY

DYUTIMAYA DAS & SHRUTI SRIVASTAVA

PROBLEM STATEMENT

X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google.

Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals.

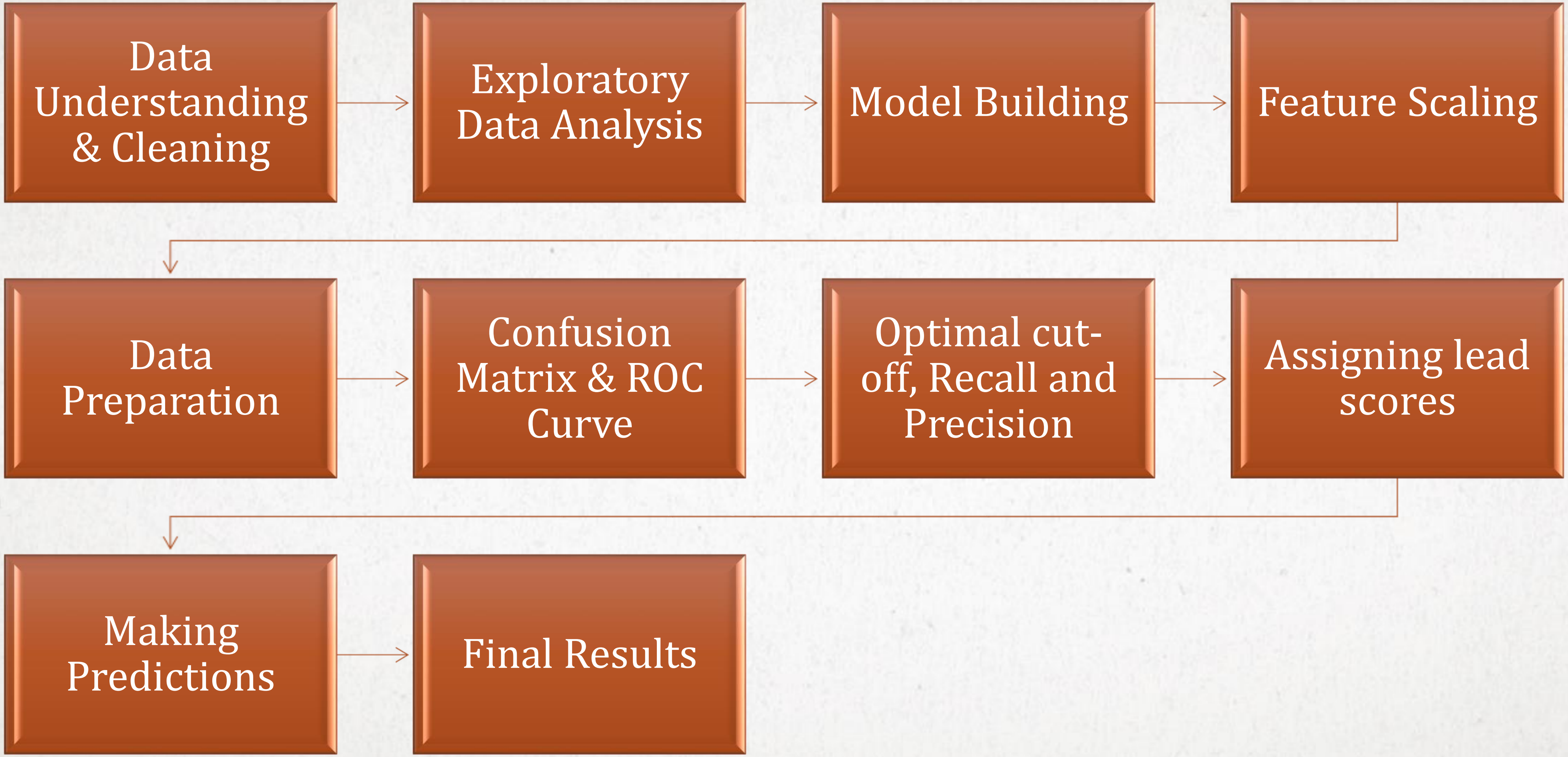
Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

BUSINESS OBJECTIVE

X Education needs help in selecting the most promising leads, i.e., the leads that are most likely to convert into paying customers.

The company needs a model wherein you a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO has given a ballpark of the target lead conversion rate to be around 80%.

SOLUTION APPROACH



Data Understanding & Cleaning

Exploratory Data Analysis

Model Building

Feature Scaling

Data Preparation

Confusion Matrix & ROC Curve

Optimal cut-off, Recall and Precision

Assigning lead scores

Making Predictions

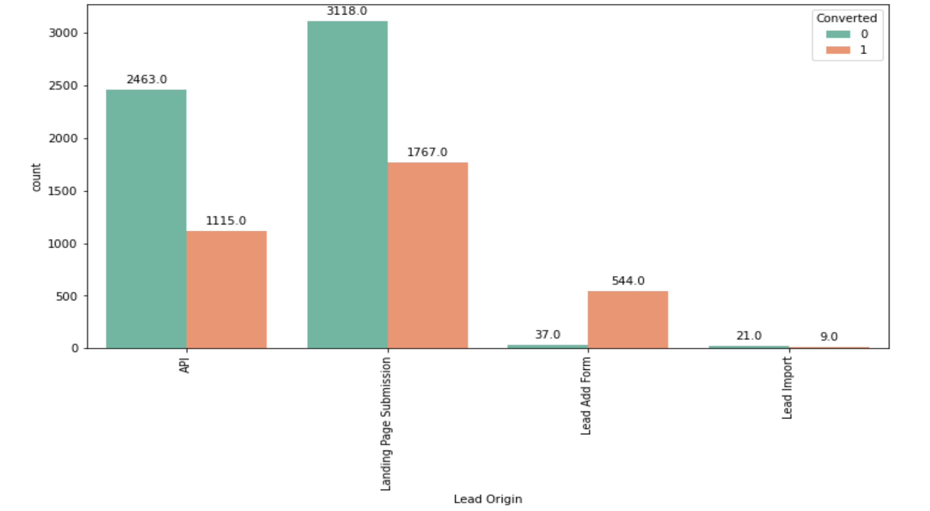
Final Results

Data Understanding & Cleaning

- The Dataset has 9240 rows and 37 columns.
- There are 7 numerical variables columns and remaining 30 columns seem to have categorical variables.
- Many of the categorical variables have a level called 'SELECT' which needs to be handled. We will replace this value with NaN as its as good as a null value. It only the user to select a value from the dropdown values.
- Dropping columns with more than 70% of missing data.
- Imputing the remaining null values for columns with lesser percentage of missing data.
- Assigned numerical variables to categories with 'Yes' to 1 and 'No' to 0.

EXPLORATORY DATA ANALYSIS

Lead Origin

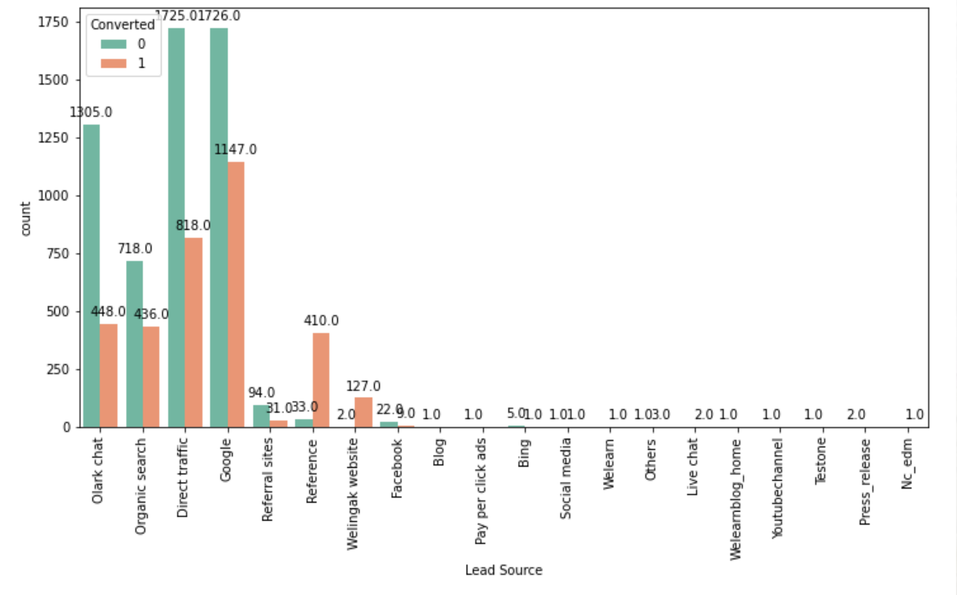


Inference

- API has around 31% (1097/3538) conversion rate.
- Landing Page Submission has around 36% (1702/4735) conversion rate.
- Lead Add Form has around 93% (543/580) conversion rate.
- Lead Import are very less in count.

To improve overall lead conversion rate, we need to focus more on improving lead conversions for API and Landing Page Submission origin. More leads should be generated from Lead Add Form.

Lead Source

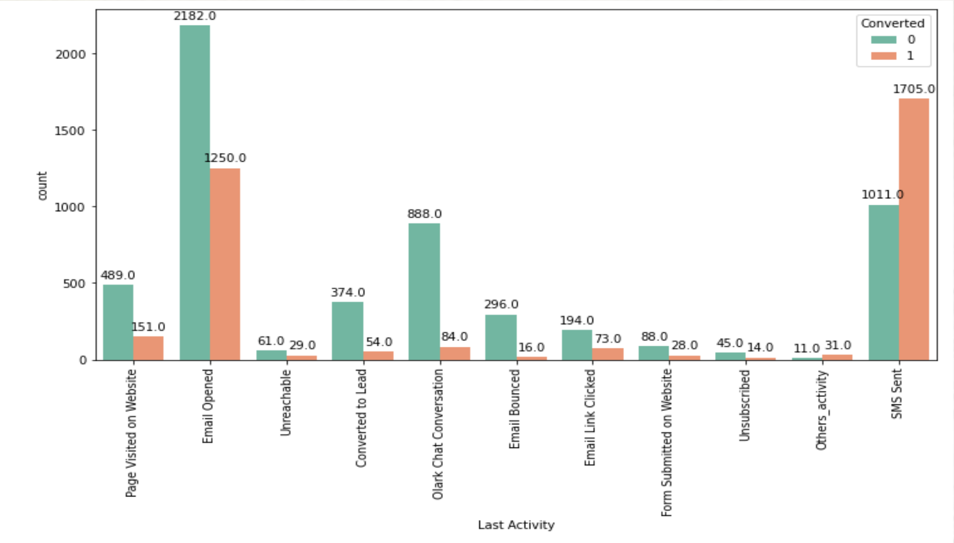


Inference

- We can conclude from the above graph that conversion rate is higher from 'Reference' and 'Welingak Website'.
- Google and Direct traffic are generating maximum number of leads.

EXPLORATORY DATA ANALYSIS (CONTD..)

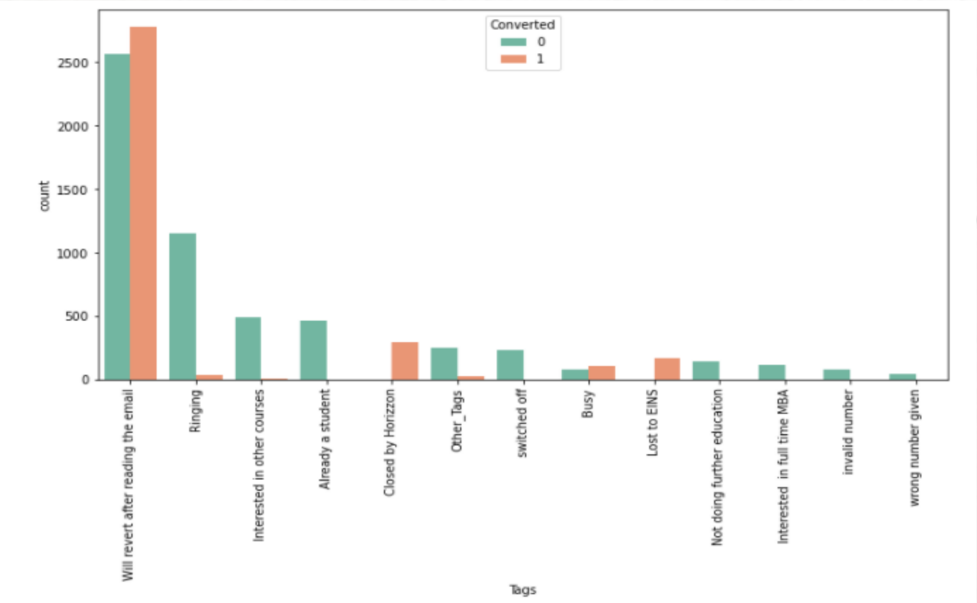
Last Activity



Inference

- Most number of people have opened the mail.
- Maximum number of conversions are coming from SMS sent.
- Only around 6-8% of emails are bounced

Tags



Inference

- The conversion rate for "Will revert after reading the email" is good
- The conversion for leads who are interested in the course is very low, the marketing strategies needs to be checked here.

DATA PREPARATION

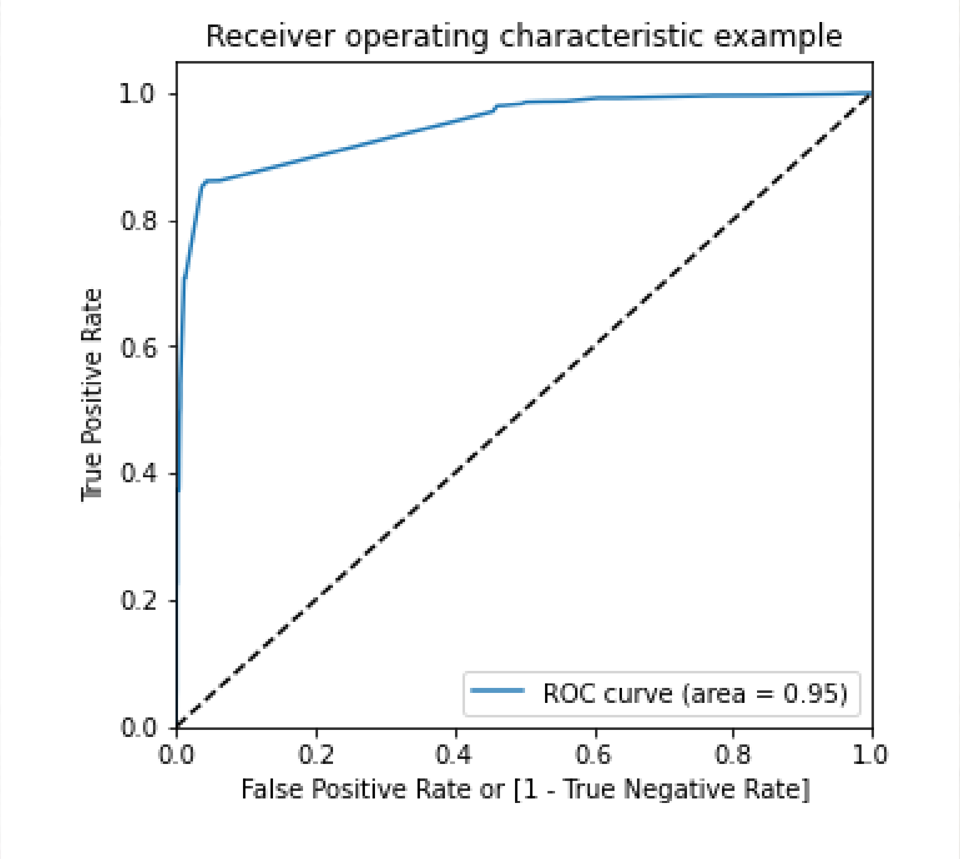
- Based on the exploratory data analysis we have seen that many columns are not adding any information to the model and will add biasness to the model hence we have dropped them.
- Some features are having skewed data so its better to drop those features as this may add skewness to our model and affect the p-value and distributions.
- Create dummy variables for categorical Features.
- Split the data set into train and test.
- Feature scaling by using StandardScaler.

MODEL BUILDING

- Using the statsmodel library, a logistic regression model has been built in python using the GLM() function.
- Recursive feature elimination technique has been used to remove the weakest features. (number of features chosen = 15).
- Based on the p-value and VIF values some more features were dropped which had either high p value(grater than 0.05) or VIF value more than 5.
- The final model has 13 variables with p value almost equal to zero and VIF < 2.

ROC CURVE

- ROC Curve demonstrates:
  - Trade-off between sensitivity and FPR(1-specificity)
  - Closer the curve follows the left-hand border and then the top border of ROC space, the more accurate the test
  - Closer the curve comes to 45° diagonal of the ROC space, the less accurate the test
- For our model, ROC curve is towards the upper left corner, and area under the curve is 0.95 Thus, our model is an optimal choice to move forward with the analysis



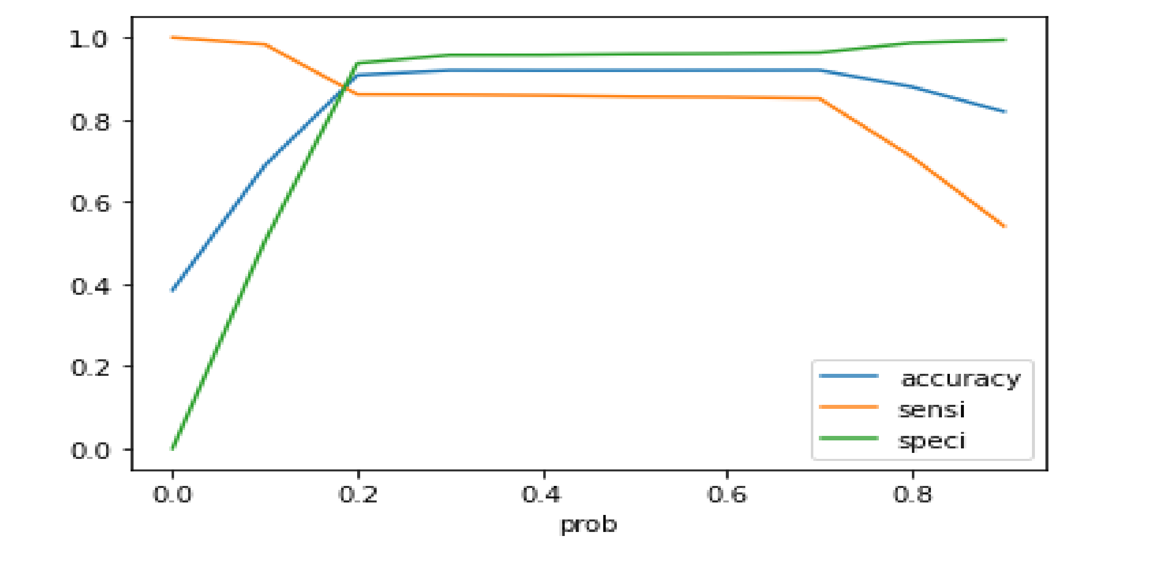
OPTIMAL CUT-OFF POINT

Plotting accuracy, sensitivity and specificity for various probabilities.

Cut-Off point is 0.20 where all three coincide, resulting:

- Accuracy:90.82%
- Sensitivity: 86.14%
- Specificity : 93.75%
- False Positive rate :6.24%
- Positive Predictive Value: 89.62%
- Negative Predictive Value:91.52%
- Precision: 89.62%
- Recall: 86.14%

Actual/Predicted	Converted	Non-Converted
Converted	3661	244
Non-Converted	339	2107

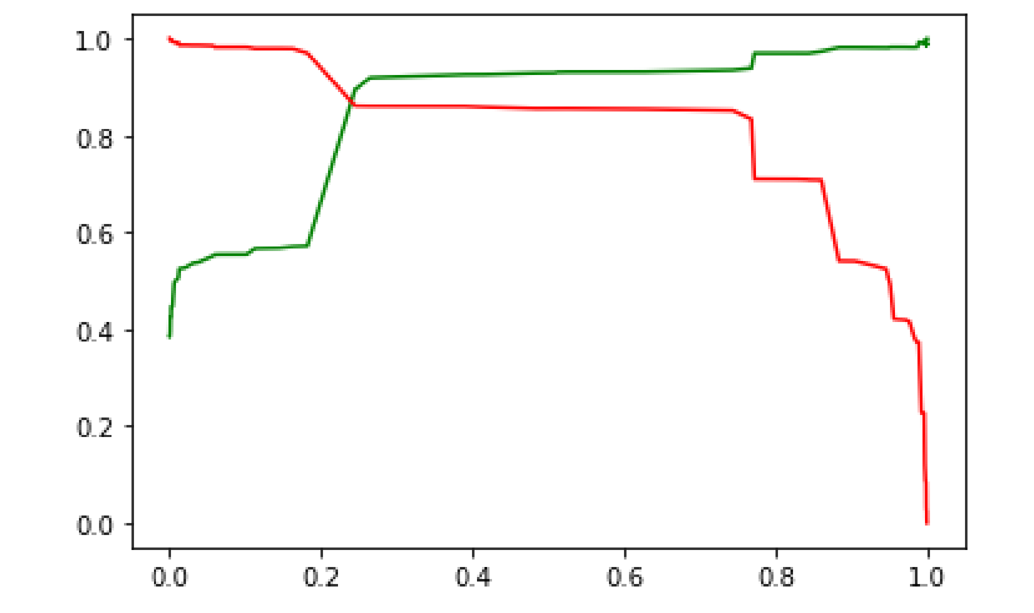


MODEL EVALUATION: PRECISION & RECALL(TRAIN DATA)

As per business requirement, we have chosen 0.20 as a Cut-Off value, which gives better results for both accuracy and precision

- **Accuracy: 90.82%**
- **Precision: 89.62%**
- **Recall: 86.14%**

The graph shows a trade-off between Precision and Recall



PREDICTION ON TEST DATA

With a chosen cut-off value of **0.20** ; we get below results

- Accuracy:90.41%
- Sensitivity: 84.52%
- Specificity : 93.77%
- False Positive rate :6.22%
- Positive Predictive Value: 88.55%
- Negative Predictive Value:91.39%
- Precision: 88.55%
- Recall: 84.52%

Actual/Predicted	Converted	Non-Converted
Converted	1626	108
Non-Converted	153	836

As Precision (positive predicted value) > 80%, we can use the same model for achieving our objective of increasing the conversion rate

Top Contributing Variables

- Lead Origin
  - Lead Origin Lead Add Form
  - Lead Origin is the one of the important parameters to keep flowing the funnel for sales team
- Lead Source
  - Lead Source_Welingak Website
  - This is again important as we understand which source has given us more conversion, so we need to constantly monitor & increase the flow from that source
- Tags:-
  - Tags_Lost to EINS
  - This one is Important feature which is having highly effective to convert lead.

Top 3 Categorical Dummy Variables

- Tags_Lost to EINS
  - This actually help us to identify the action should be taken for this as its important for lead conversion prediction.
- Tags_Closed by Horizzon
- Tags_Will revert after reading the email.
  - This action shows the interest of the student

TOP   Features

CONCLUSION

- The model is prepared for prediction of the conversion of the leads. The probability values are generated by the model. The cutoff decided for the model is 0.2. All leads whose probability is generated above this threshold value can be classified as Hot Lead.
- Our focus should be very high on
  - Tags
  - Lead Source
  - Lead Origin
- We should plan some additional marketing activity on these variables
- Our focus is less on reference program we need to build strategy to bring more candidate through reference as acquisition cost from this model will be less than all other model