LEAD SCORING CASE STUDY

SUBMISSION

Group Name:

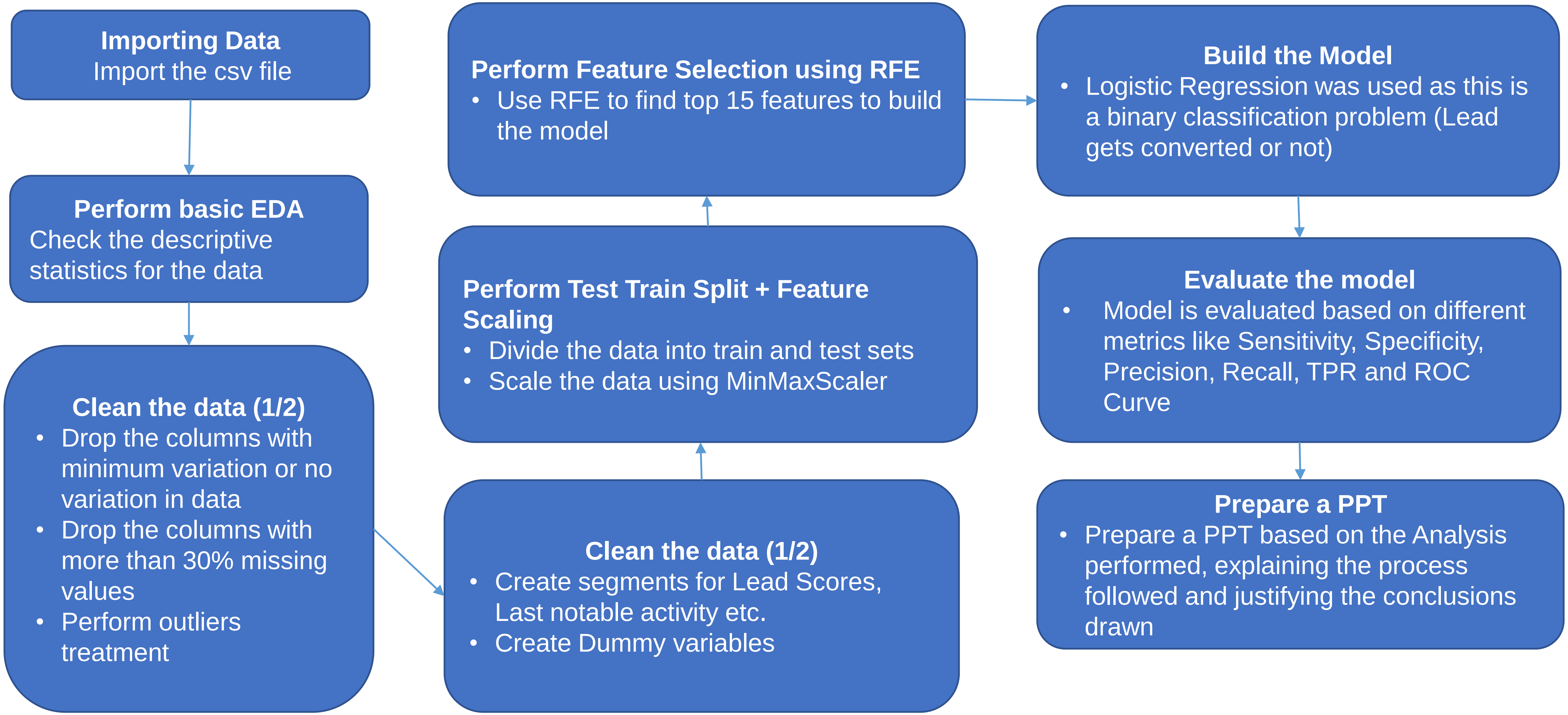
1. Chirag Maru (Group facilitator)
2. Avineet Kumar
3. Ritesh Kotian
4. Lokesh Sah

Synopsis of the project

- **Project Brief**
  - X Education sells online courses to industry professionals.
- **Business and Data Understanding**
  - The company markets the courses via different websites and search engines. Once people land on X Education website they fill up a form and are classified as a lead. Once leads are acquired, sales team start contacting them. The typical lead conversion rate is 30% currently.
- **Business Objective and Strategy**
  - To increase the lead conversion rate to 80%.
- **Goals of Data Analysis**
  - To assign a lead score for each 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 wants to identify more than 80% "Hot Leads".
- **Base files used**

• Leads.csv
• Leads Data Dictionary.xlsx

Lead Scoring Case Study - Problem solving flowchart



Importing Data
Import the csv file

Perform basic EDA
Check the descriptive statistics for the data

Clean the data (1/2)

- Drop the columns with minimum variation or no variation in data
- Drop the columns with more than 30% missing values
- Perform outliers treatment

Clean the data (1/2)

- Create segments for Lead Scores, Last notable activity etc.
- Create Dummy variables

Perform Test Train Split + Feature Scaling

- Divide the data into train and test sets
- Scale the data using MinMaxScaler

Perform Feature Selection using RFE

- Use RFE to find top 15 features to build the model

Build the Model

- Logistic Regression was used as this is a binary classification problem (Lead gets converted or not)

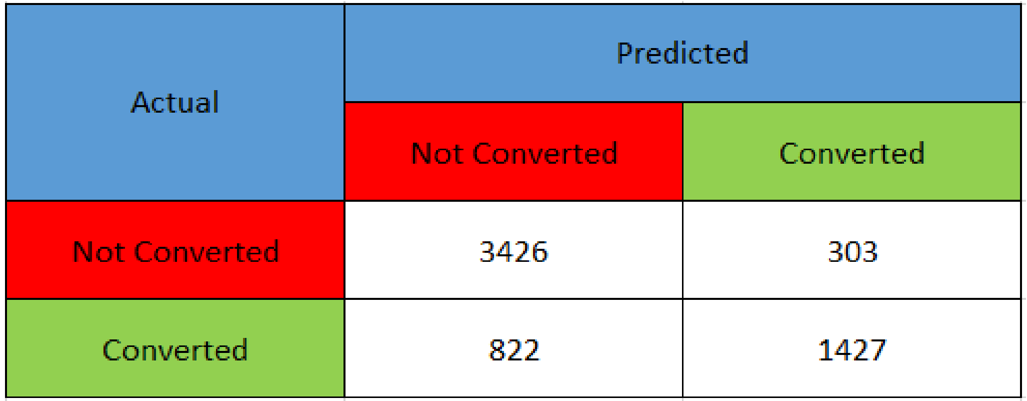
Evaluate the model

- Model is evaluated based on different metrics like Sensitivity, Specificity, Precision, Recall, TPR and ROC Curve

Prepare a PPT

- Prepare a PPT based on the Analysis performed, explaining the process followed and justifying the conclusions drawn

Confusion Matrix for the Logistic Regression Model



Actual	Predicted	
	Not Converted	Converted
Not Converted	3426	303
Converted	822	1427

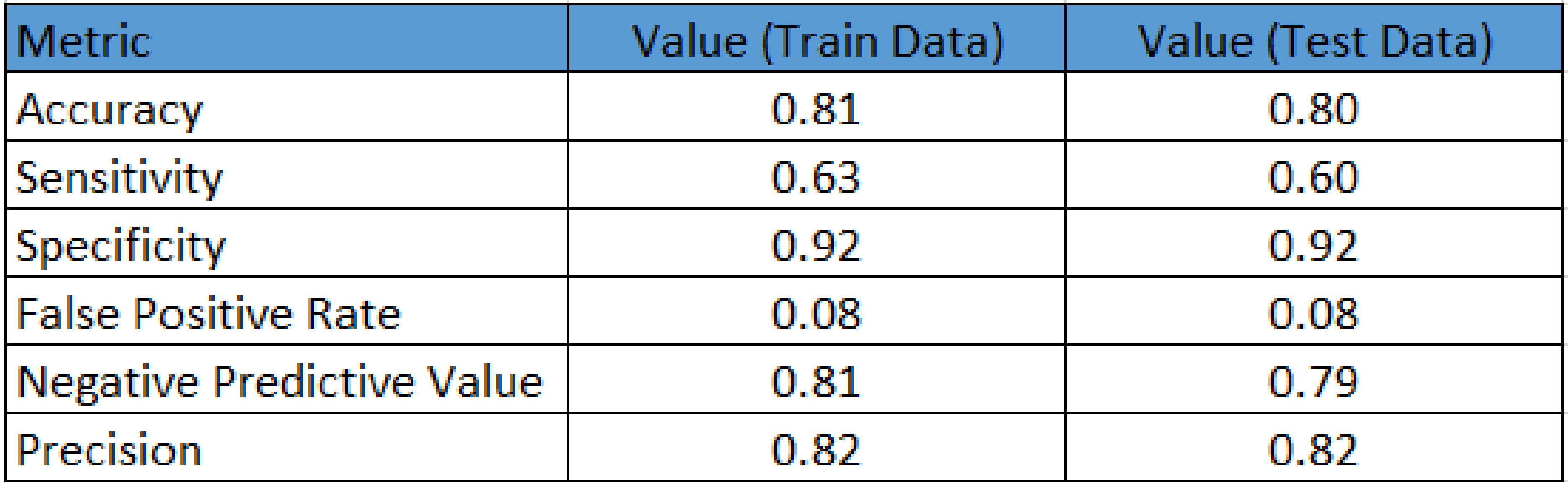
Training Data

Actual	Predicted	
	Not Converted	Converted
Not Converted	1446	134
Converted	390	593

Test Data

Confusion Matrix

Important Metrics for the Logistic Regression Model



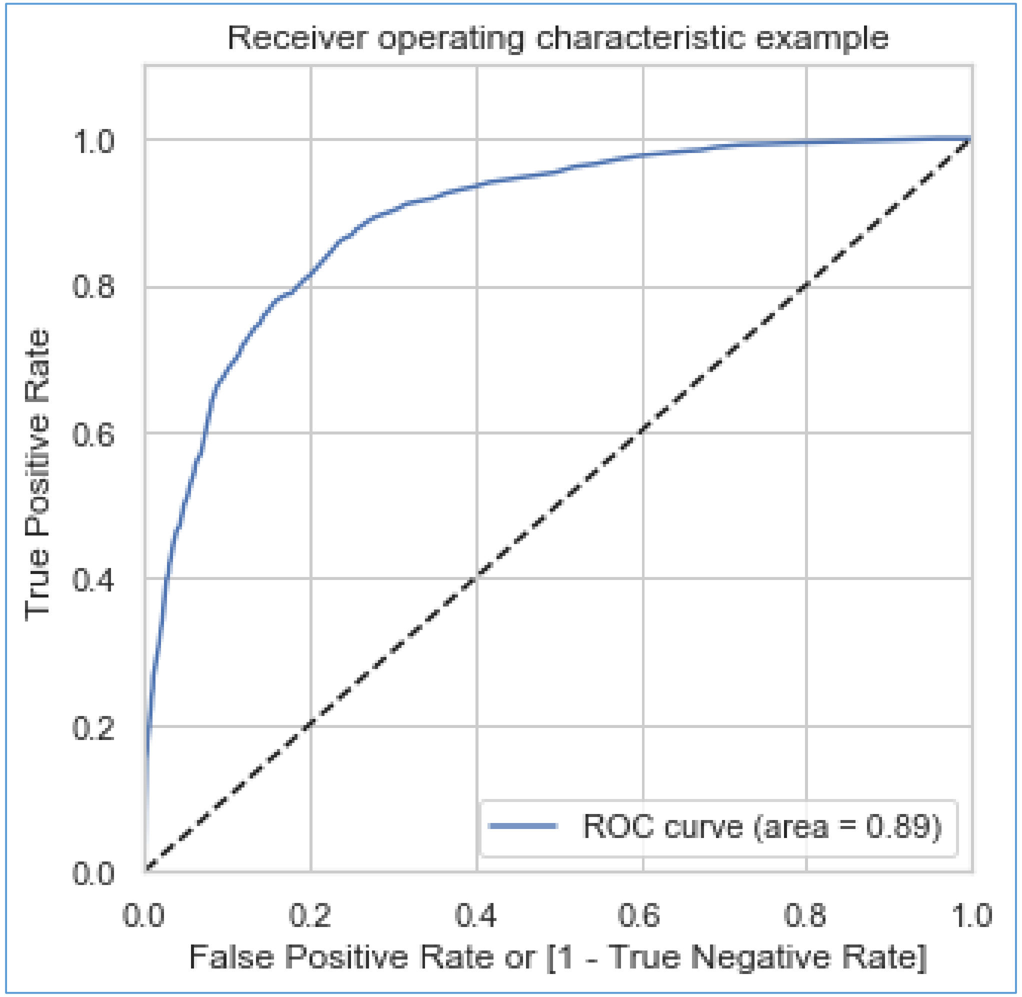
Metric	Value (Train Data)	Value (Test Data)
Accuracy	0.81	0.80
Sensitivity	0.63	0.60
Specificity	0.92	0.92
False Positive Rate	0.08	0.08
Negative Predictive Value	0.81	0.79
Precision	0.82	0.82

Metrics for the Model (Training Data vs Test Data)

Important Metrics for the Logistic Regression Model

ROC curve shows the tradeoff between the True Positive Rate (TPR) and the False Positive Rate (FPR)

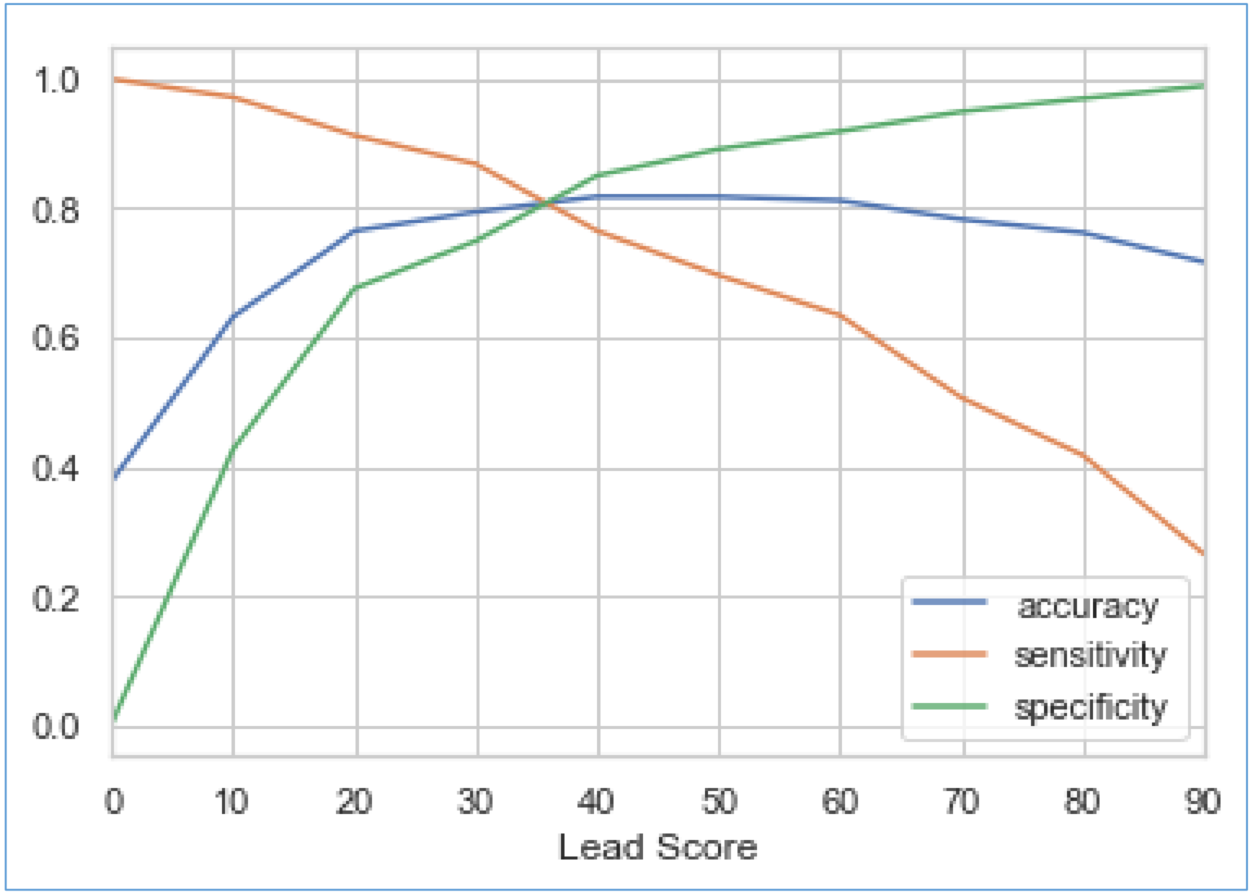
- Area under the curve = 0.89



ROC Curve

Important Metrics for the Logistic Regression Model

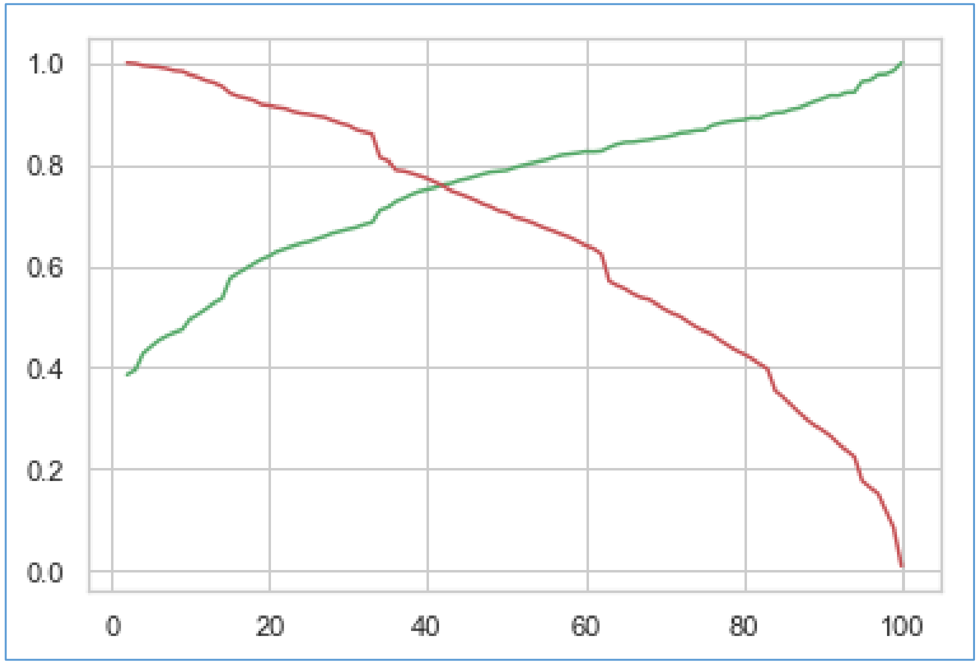
We plot curve between accuracy, sensitivity and specificity.



Accuracy, Sensitivity, and Specificity tradeoff

Important Metrics for the Logistic Regression Model

Since the target is to achieve 80% or higher conversion, we’ve taken cutoff as 0.60 to achieve precision of 0.80.

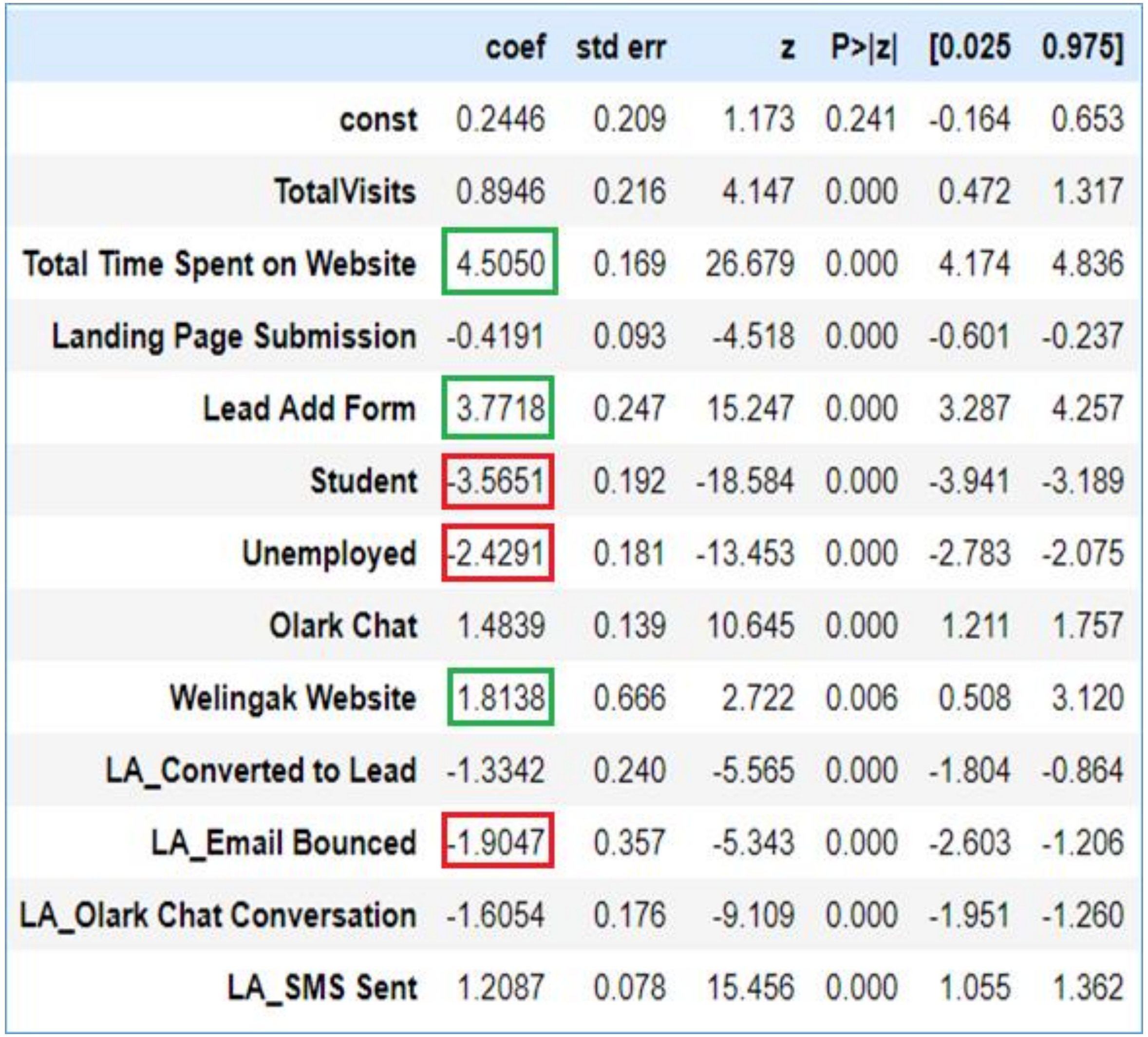


Precision – Recall Curve

Conclusions from the Logistic Regression Model

1. Top three variables which contribute most towards the probability of a lead being converted:
   a. Total Time Spent on Website
   b. Lead Add Form
   c. Welingak Website

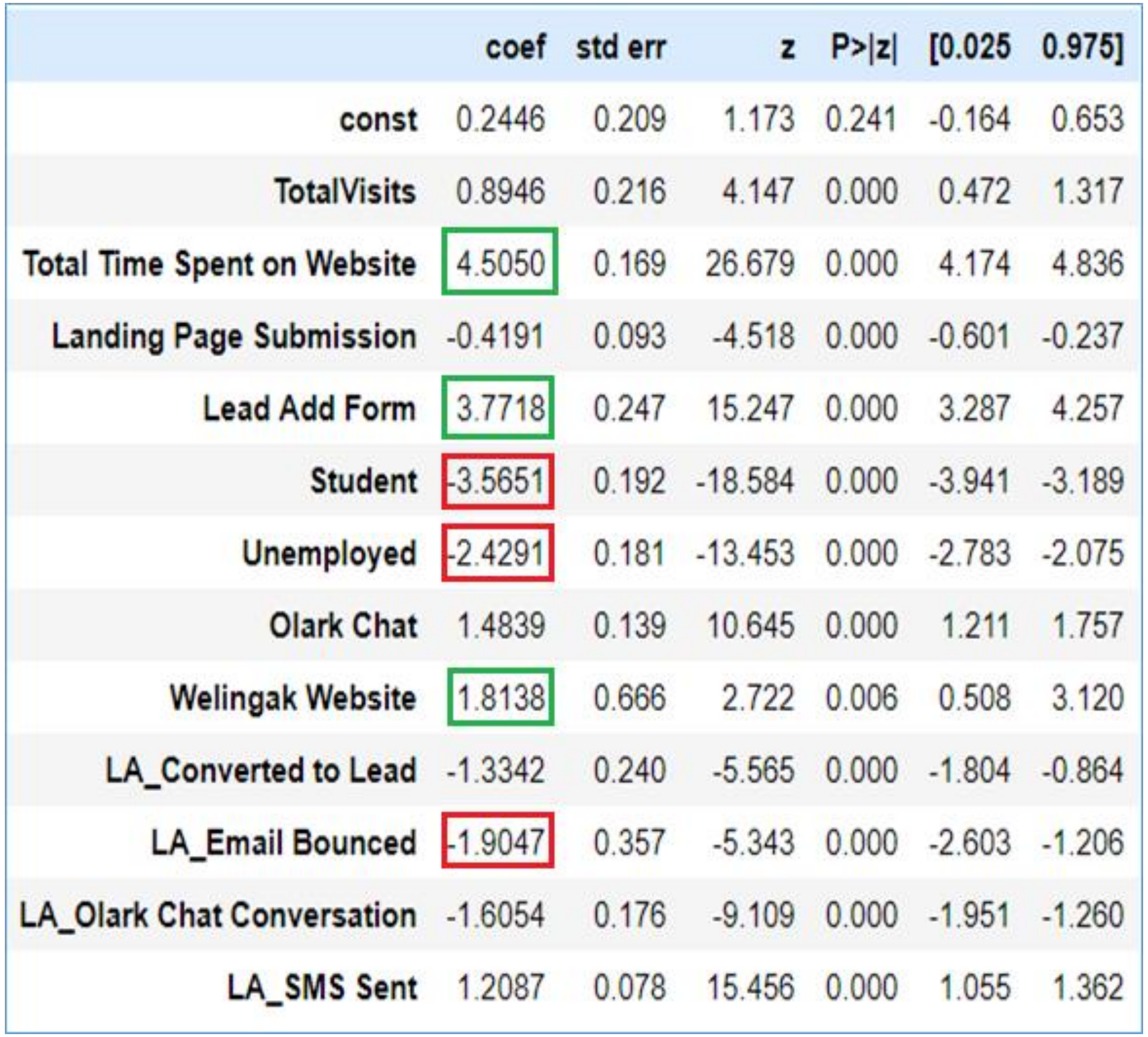
2. Top 3 categorical/ dummy variables which should be focused the most on in order to increase the probability of lead conversion are:
   a. Lead Add Form
   b. Welingak Website
   c. Olark Chat



	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2446	0.209	1.173	0.241	-0.164	0.653
TotalVisits	0.8946	0.216	4.147	0.000	0.472	1.317
Total Time Spent on Website	4.5050	0.169	26.679	0.000	4.174	4.836
Landing Page Submission	-0.4191	0.093	-4.518	0.000	-0.601	-0.237
Lead Add Form	3.7718	0.247	15.247	0.000	3.287	4.257
Student	-3.5651	0.192	-18.584	0.000	-3.941	-3.189
Unemployed	-2.4291	0.181	-13.453	0.000	-2.783	-2.075
Olark Chat	1.4839	0.139	10.645	0.000	1.211	1.757
Welingak Website	1.8138	0.666	2.722	0.006	0.508	3.120
LA_Converted to Lead	-1.3342	0.240	-5.565	0.000	-1.804	-0.864
LA_Email Bounced	-1.9047	0.357	-5.343	0.000	-2.603	-1.206
LA_Olark Chat Conversation	-1.6054	0.176	-9.109	0.000	-1.951	-1.260
LA_SMS Sent	1.2087	0.078	15.456	0.000	1.055	1.362

Conclusions from the Logistic Regression Model

3. Top three variables which decrease the probability of conversion of lead are:
   a. Student
   b. Unemployed
   c. Last Activity - Email Bounced



	coef	std err	z	P>|z|	[0.025	0.975]
const	0.2446	0.209	1.173	0.241	-0.164	0.653
TotalVisits	0.8946	0.216	4.147	0.000	0.472	1.317
Total Time Spent on Website	4.5050	0.169	26.679	0.000	4.174	4.836
Landing Page Submission	-0.4191	0.093	-4.518	0.000	-0.601	-0.237
Lead Add Form	3.7718	0.247	15.247	0.000	3.287	4.257
Student	-3.5651	0.192	-18.584	0.000	-3.941	-3.189
Unemployed	-2.4291	0.181	-13.453	0.000	-2.783	-2.075
Olark Chat	1.4839	0.139	10.645	0.000	1.211	1.757
Welingak Website	1.8138	0.666	2.722	0.006	0.508	3.120
LA_Converted to Lead	-1.3342	0.240	-5.565	0.000	-1.804	-0.864
LA_Email Bounced	-1.9047	0.357	-5.343	0.000	-2.603	-1.206
LA_Olark Chat Conversation	-1.6054	0.176	-9.109	0.000	-1.951	-1.260
LA_SMS Sent	1.2087	0.078	15.456	0.000	1.055	1.362

Conclusions from the Logistic Regression Model

As per our analysis we can conclude on the following:

1. We should focus more on the leads spending more time on our website as they seem to be more interested in our courses and hence have higher conversion rate
2. Also leads coming through the Welingak Website have a higher conversion rate so we should give top priority to these
3. Unemployed and Students seem to be less likely to be converted and should be followed with low priority
4. Some leads are providing fake email ids and should be ignored for saving time and better lead conversion rates