# LEAD SCORING CASE STUDY

Submitted By-
**JITAL ENGINEER**
**MUKUL YADAV**
**SIDDHI KADAM**

# Problem Statement

- An education company named X Education sells online courses to industry professionals who browse on their website for courses.

- Once they browse the courses, people fill up a form providing their details and hence are classified to be a lead. Moreover, the company also gets leads through past referrals. Through the process of mails and calls, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

- Now, although X Education gets a lot of leads, its lead conversion rate is very poor. Hence, 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.

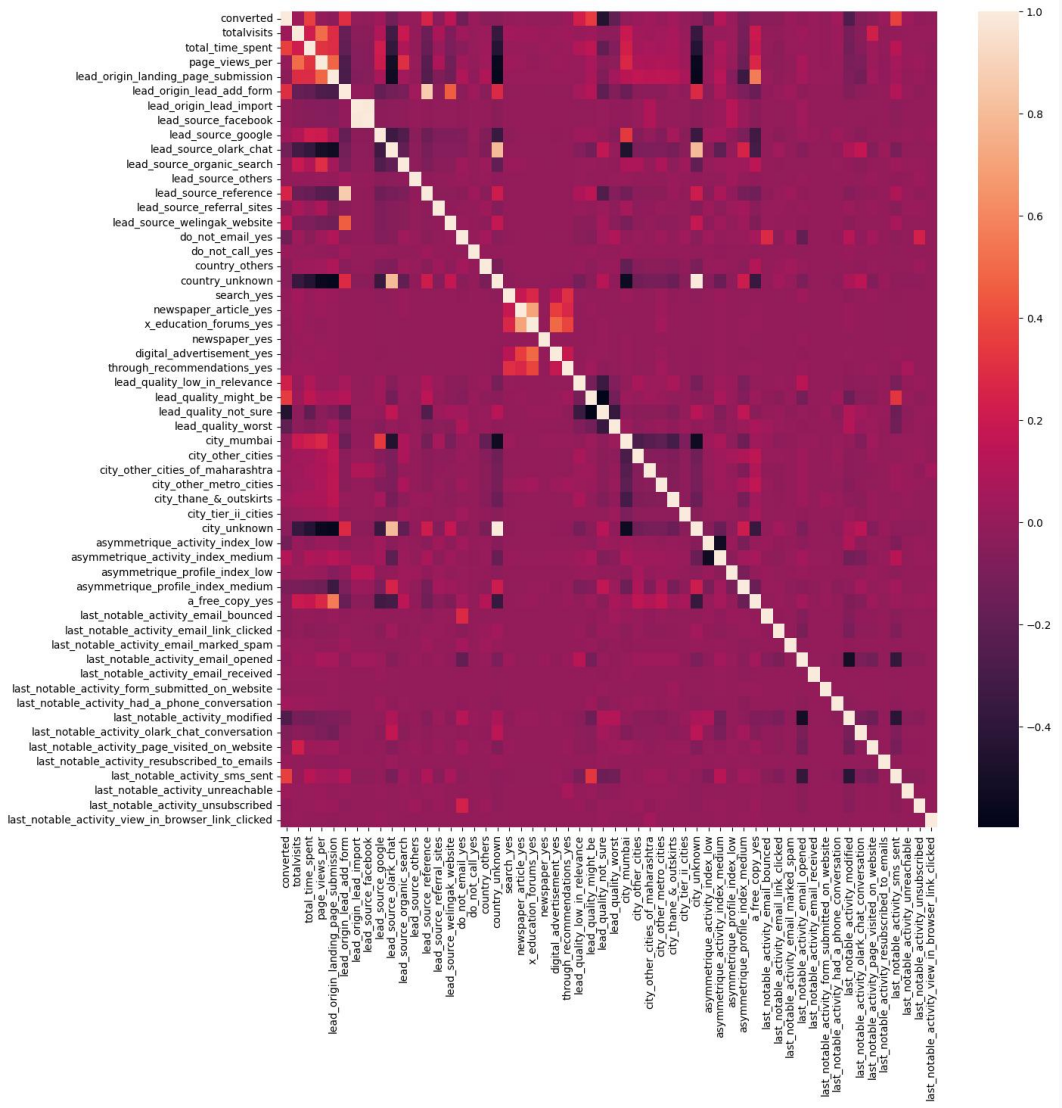
# Business goals

- To develop a model in order to select promising leads.
- Every lead should be assigned a lead score in order to indicate how promising the lead could be. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
- The CEO has given a ballpark of the target lead conversion rate to be around 80%.

# Approach

- Importing The Dataset.
- Data Inspection.
- Data Cleaning.
- Exploratory Data Analysis.
- Prepare The Data For Model Building- Univariate Analysis And Outlier Detection, Bivariate Analysis.
- Model Building.
- Build a Logistic Regression model.
- Testing the model on train set.
- Evaluate model on different measures- precision, recall, sensitivity, specificity.
- Testing the model on test set.
- Evaluate model on different measures- precision, recall, sensitivity, specificity.
- Testing the model on test set.

# Exploratory data analysis

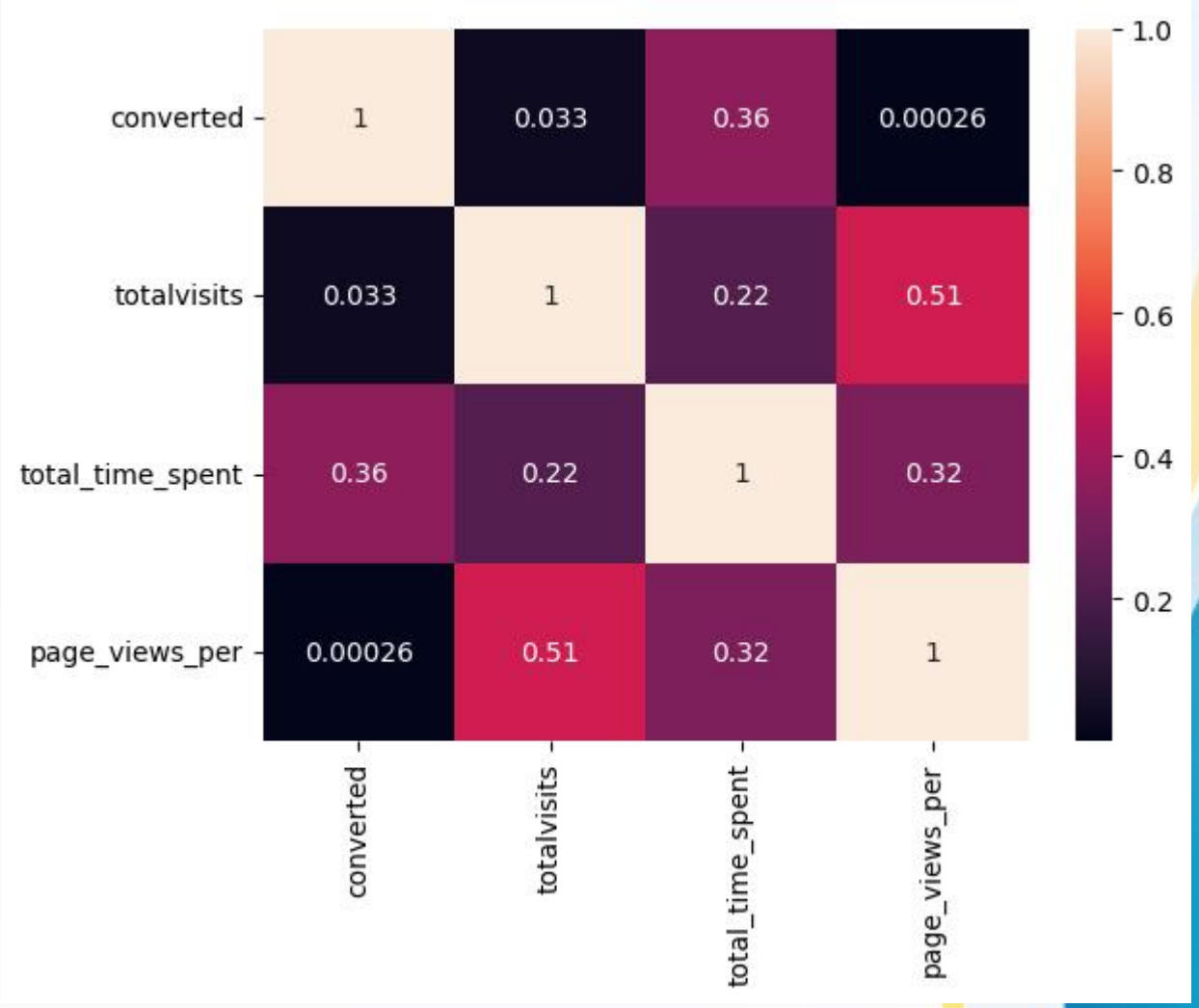



Heat map is shown in the figure.

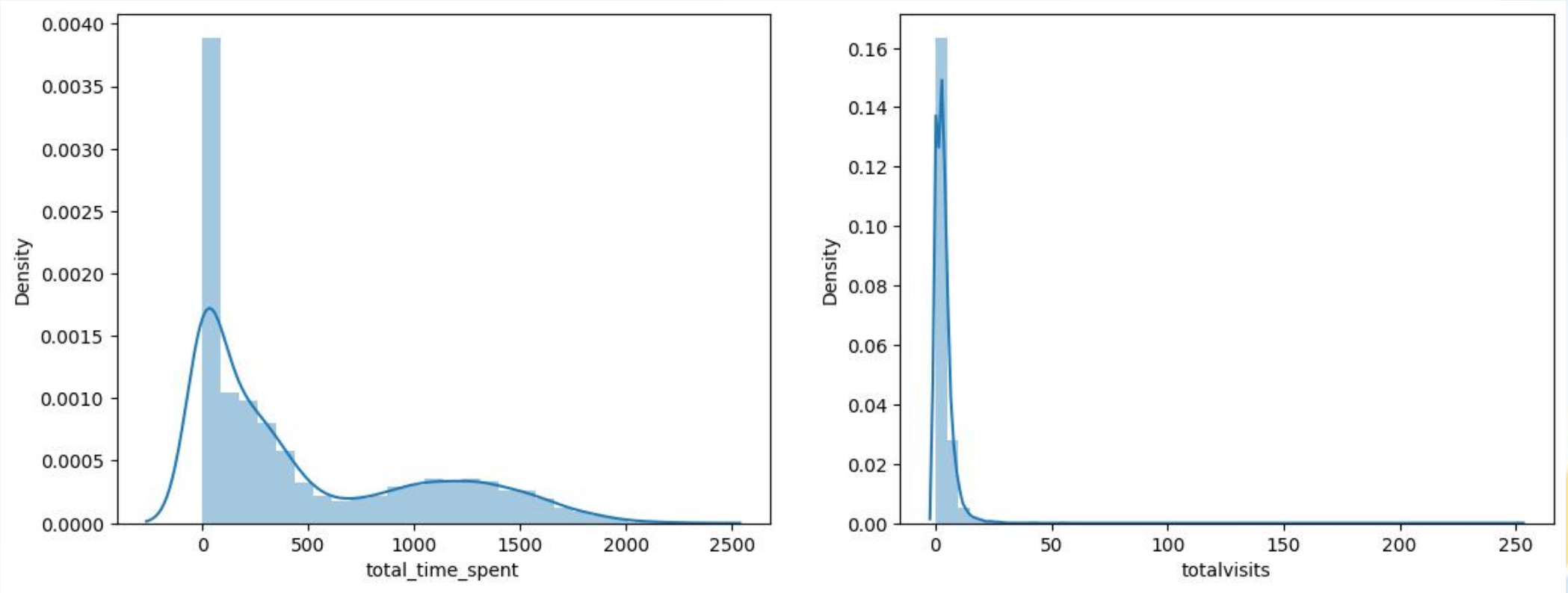
# Exploratory data analysis

- The continuous variables are:
✓ totalvisits
✓ total_time_spent
✓ page_views_per

- The heatmap for continuous variables is shown.



# Univariate analysis

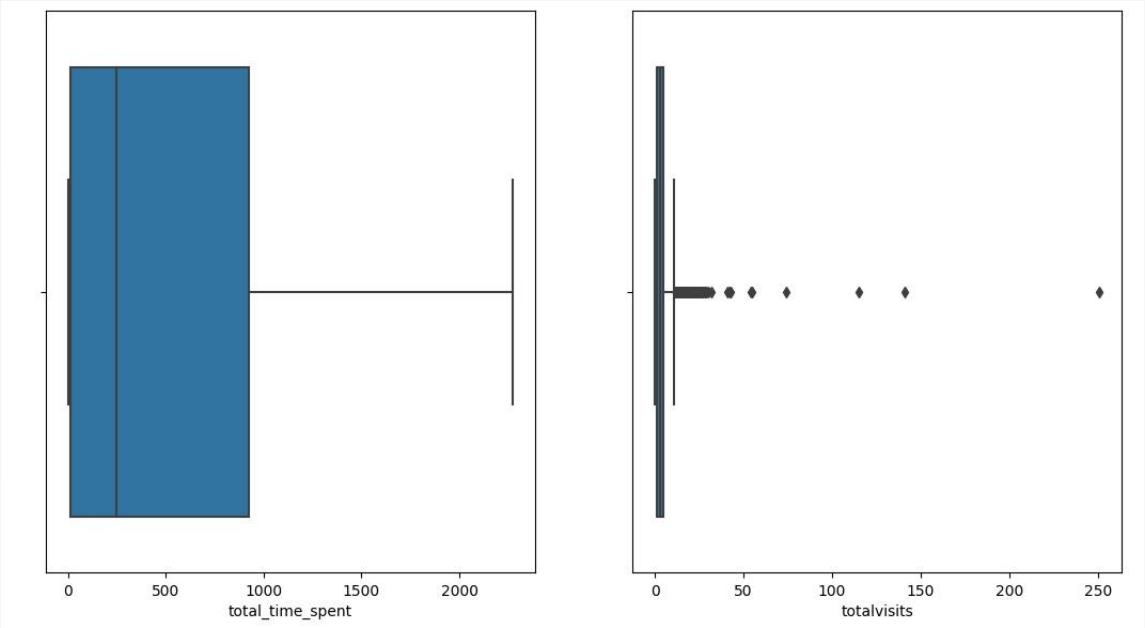



- The above figure shows the Univariate analysis of:

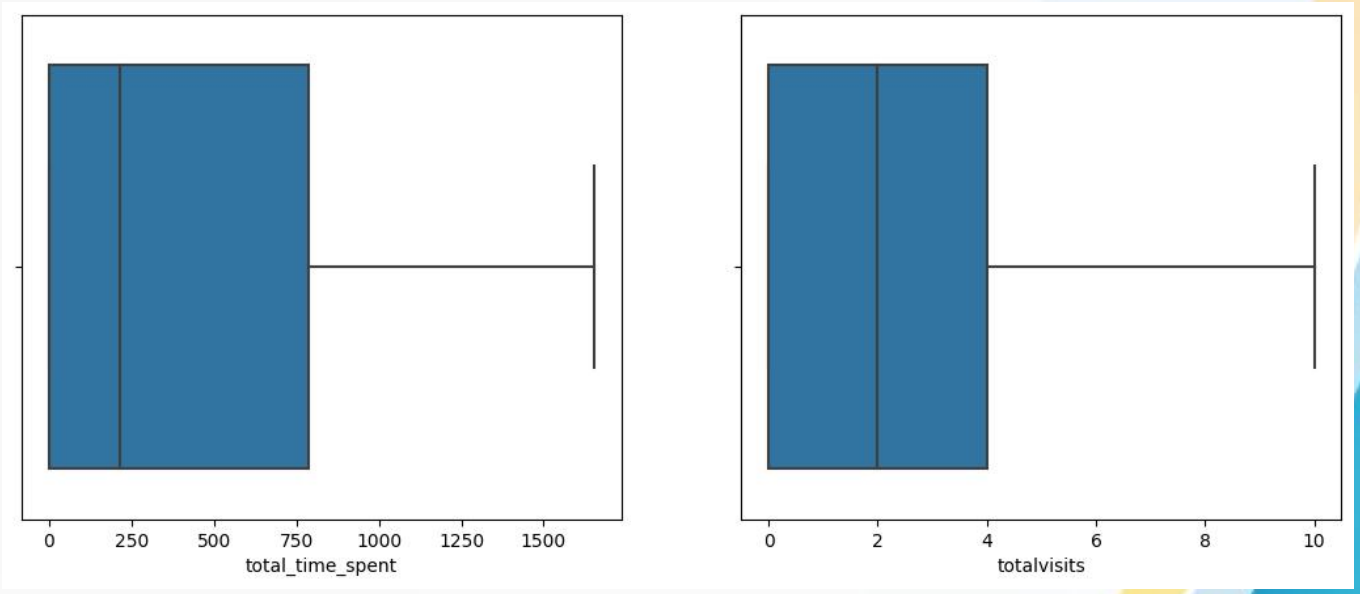
✓ total_time_spent vs Density.

✓ totalvisits vs Density.

# Outlier Analysis

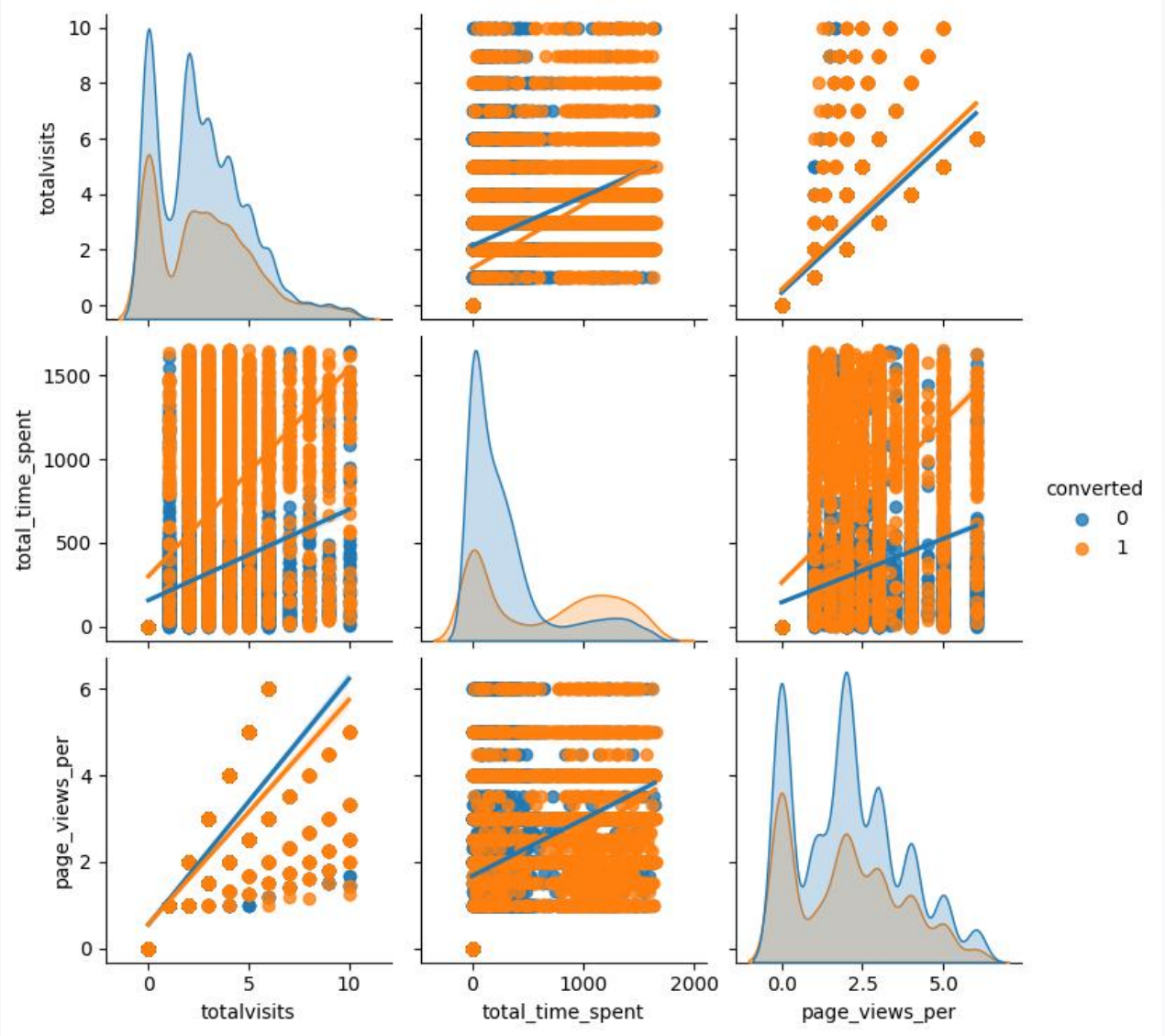



- ✓ The above figure shows the presence of Outliers in total_time_spent and totalvisits.




- ✓ The above figure shows the boxplot after removal of Outliers in total_time_spent and totalvisits.

# Bivariate analysis

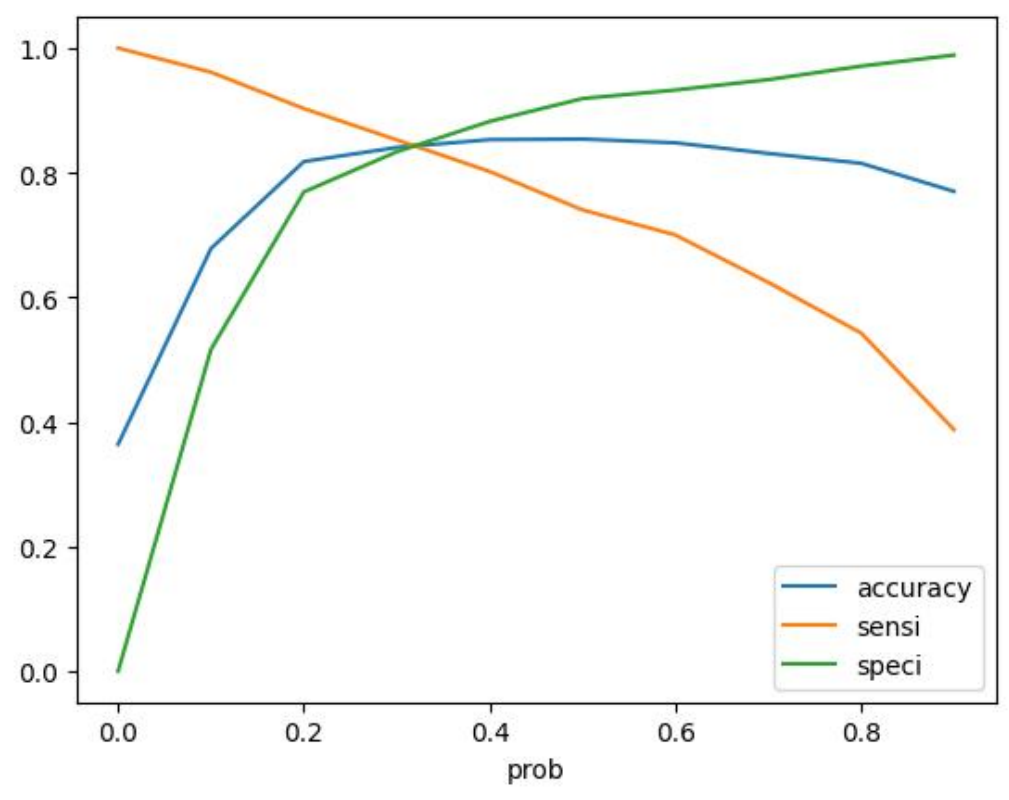


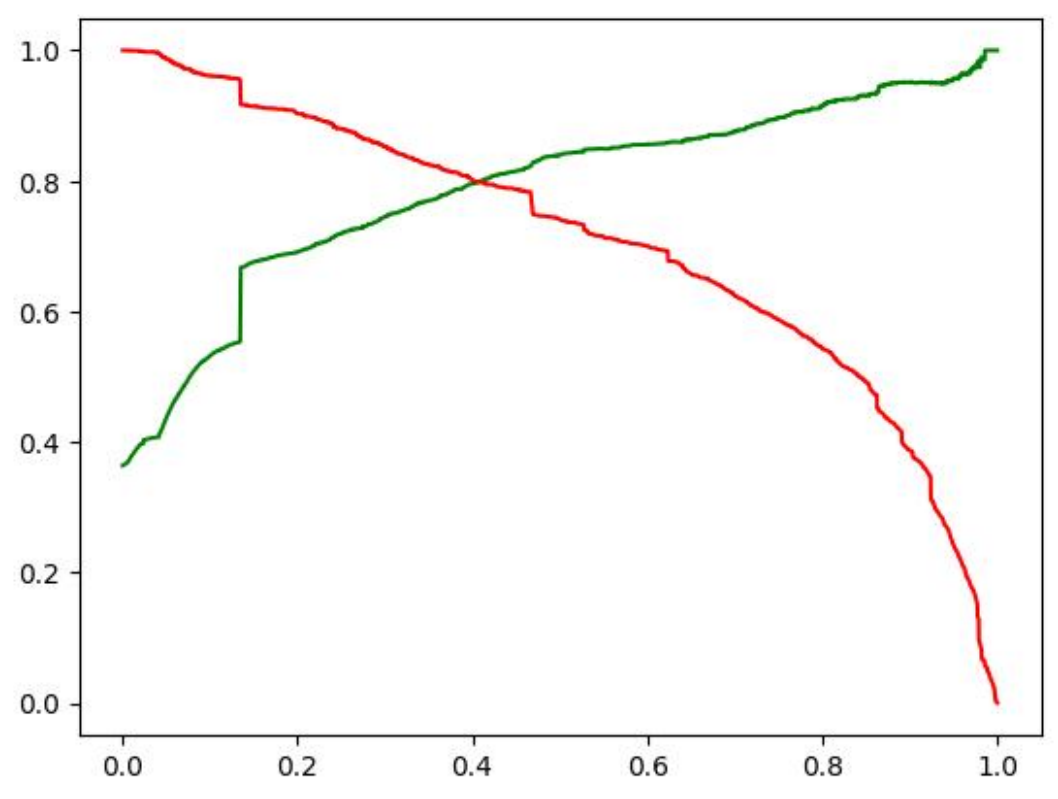

- The above figure shows the Bivariate analysis of totalvisits , total_time_spent and page_views_per.

# Model building

- Splitting the data into train set and test set.
- MinMax Scaling.
- Build the first model.
- Feature selection using RFE to eliminate less relevant variables.
- Build the next model.
- Eliminate variables based on high p-values.
- Check VIF.
- Get the predicted values on the train set.
- Calculate accuracy, sensitivity, specificity, precision and other metrics.
- Predict using test set.
- Calculate accuracy, sensitivity, specificity, precision and other metrics.

# Model evaluation- train set




Confusion Matrix:

| | |
|---|---|
| 3355 | 296 |
| 543 | 1547 |

- Accuracy- 0.8465
- Sensitivity- 0.8248
- Specificity- 0.8589
- Precision- 0.7699

# Model evaluation- test set

Confusion Matrix:

| | |
|---|---|
| 1345 | 205 |
| 156 | 755 |

- Accuracy- 0.8533
- Sensitivity- 0.8487
- Specificity- 0.8677
- Precision- 0.7864

# Results

- The model evaluation has been done on the train dataset and the test dataset and Sensitivity-Specificity and Precision-Recall metrics have been calculated.

- For train set:

✓ Accuracy- 84%
✓ Sensitivity- 82%
✓ Specificity- 85%
✓ Precision- 77%

- For test set:

✓ Accuracy- 85%
✓ Sensitivity- 84%
✓ Specificity- 86%
✓ Precision- 78%

# Overall Observations

- 'TotalVisits' , 'Total Time Spent on Website' , 'Page Views Per Visit' are features which contribute most towards the probability of a lead getting converted.
- The lead score calculated shows the conversion rate on the final predicted model is approximately 82% in train set and 84% in test set.
- Building our model helps us in monitoring and tailoring the information sent to the leads.
- More focus should be laid on converted leads by holding question-answer sessions to determine their intention to join the online courses.