LEAD SCORE CASE STUDY

by:-

Divyanshu Pandey
Pooja Kumari
Poojitha

PROBLEM STATEMENT

▶ X Education, an online course provider for industry professionals, faces a challenge with a low lead conversion rate.

▶ Although they generate numerous leads, only a fraction convert into paying customers.

▶ To improve this, the company seeks to identify and prioritize 'Hot Leads' with a higher likelihood of conversion.

▶ The objective is to build a lead scoring model that assigns scores to leads, helping the sales team focus on potential customers.

▶ The CEO aims for an 80% lead conversion rate. A dataset with 9000 data points, including attributes like Lead Source, Total Time Spent on Website, and Last Activity, is provided.

▶ The target variable is 'Converted' (1 for converted, 0 for not converted). Handling 'Select' levels in categorical variables is also crucial.

Business Goals

- The primary business goal of this case study is to develop a logistic regression model that assigns a lead score ranging from 0 to 100 to each prospective lead, enabling the company to effectively target potential leads.

- A higher lead score indicates a 'hot' lead with a higher likelihood of conversion, while a lower score signifies a 'cold' lead less likely to convert.

- Additionally, the model should exhibit adaptability to address future changes in the company's requirements and challenges.

- These potential adjustments are documented separately and will be incorporated into the logistic regression model's recommendations, ensuring ongoing effectiveness and relevance in lead conversion strategies.

STRATEGY

- Importing and reading data

- Cleaning and preparing data

- Exploratory data analysis

- Prepare for data modelling

- Dummy creation

- Train test spilt

- Scaling the data

- **Model building using stats model & RFE**

- **Model evaluation**

- Plotting ROC curve

- Test set prediction

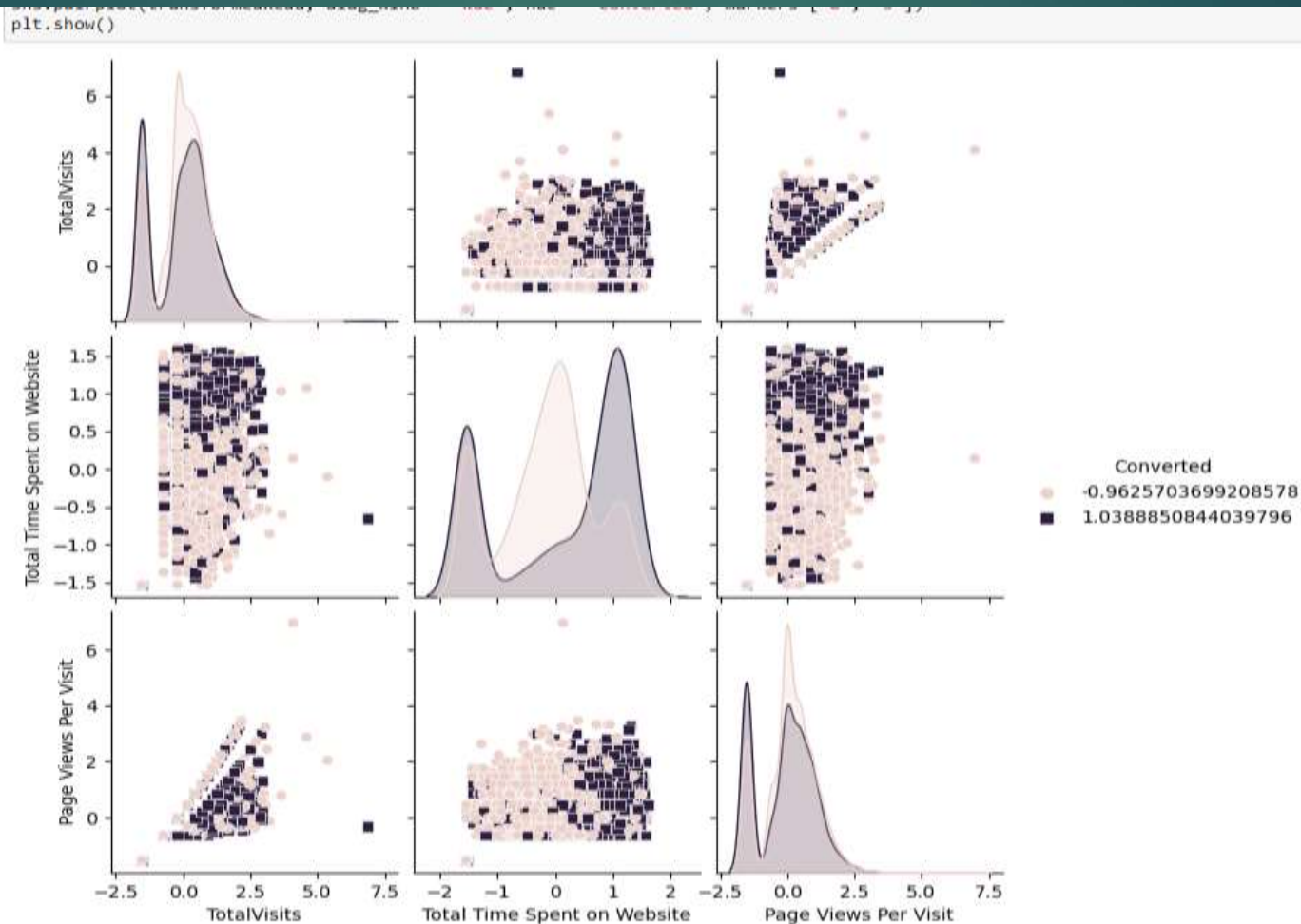- Final observation and recommendations

- Summary
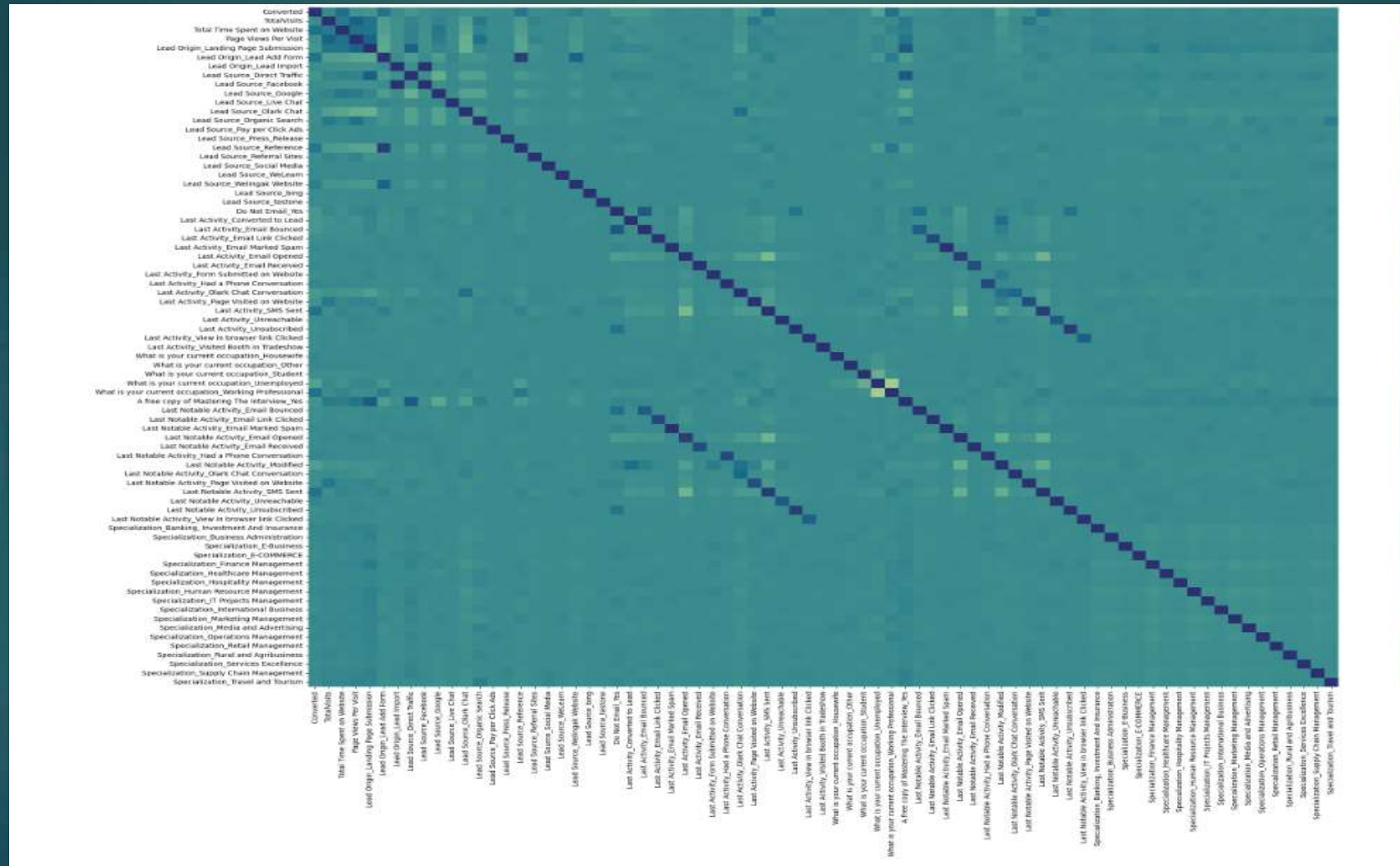
EDA And Data Modelling - I
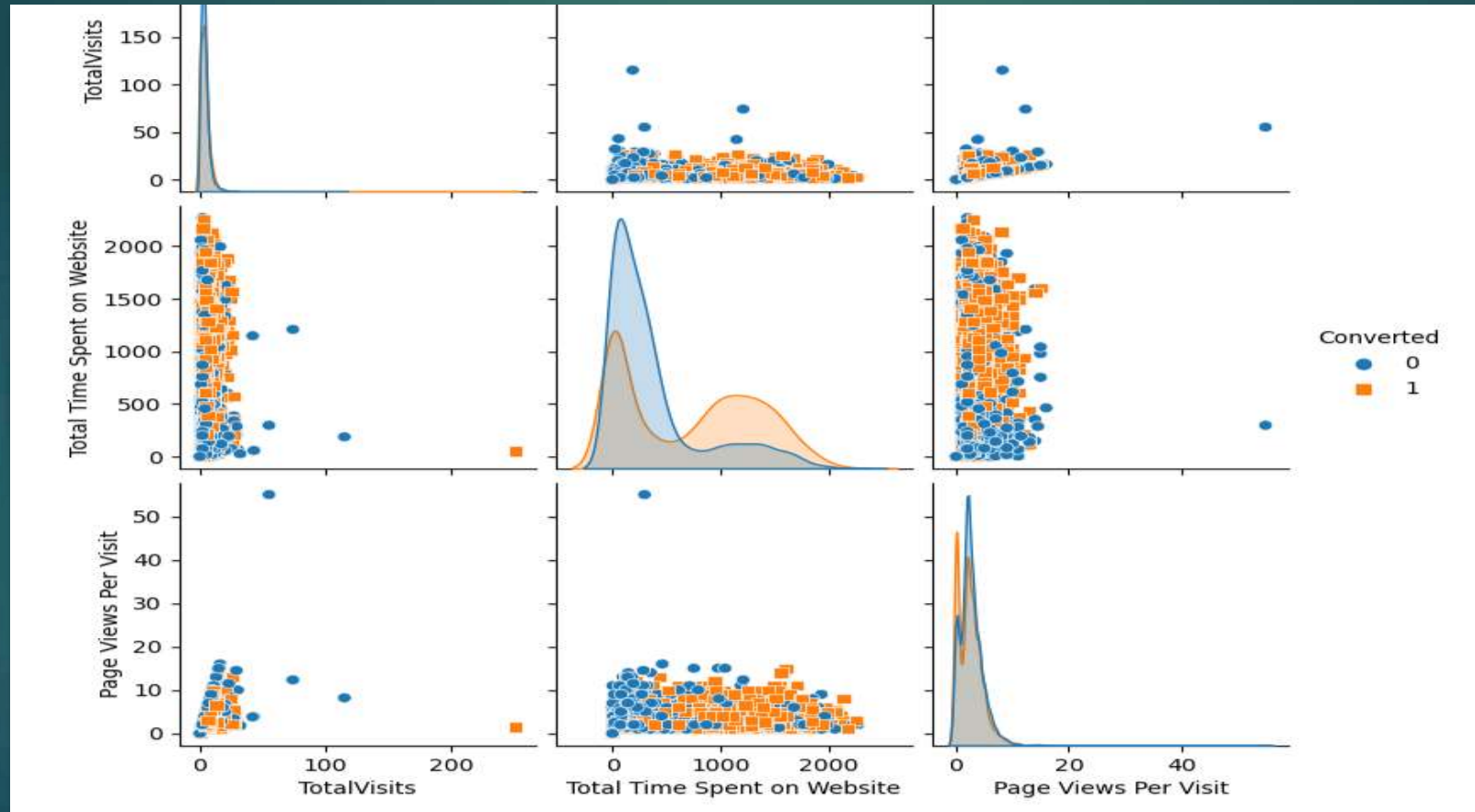
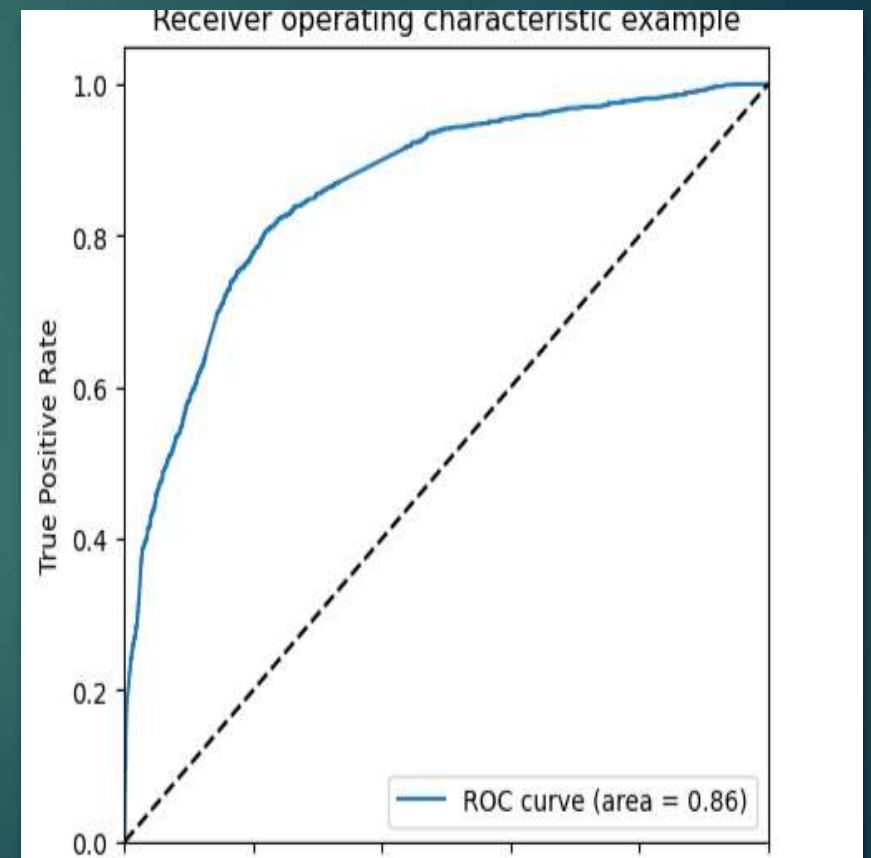# EDA And Data Modelling - II
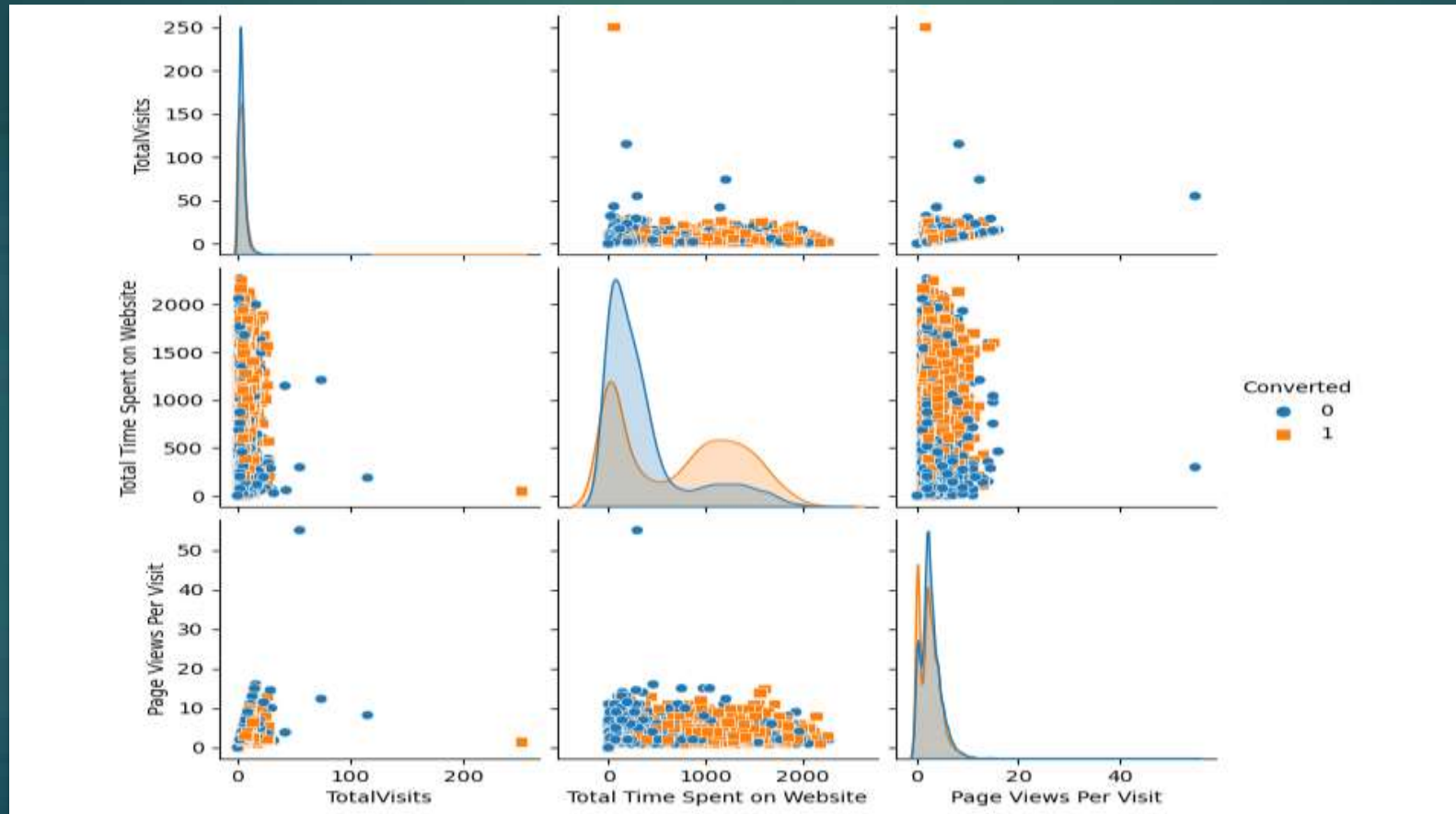
EDA And Data Modelling - III

Correlation

ROC Curve

▶The area under the curve of the ROC is 0.86 which is quite good. So it seem to have a good model.

Receiver operating characteristic example

ROC curve (area = 0.86)

Optimal Cutoff

As you can see that around 0.42, we get the optimal values of the three metrics.

▶ Cutoff value:- 0.42

▶ Accuracy :- 79.08%

▶ Sensitivity:- 79.33%

▶ Specificity :- 78.84%
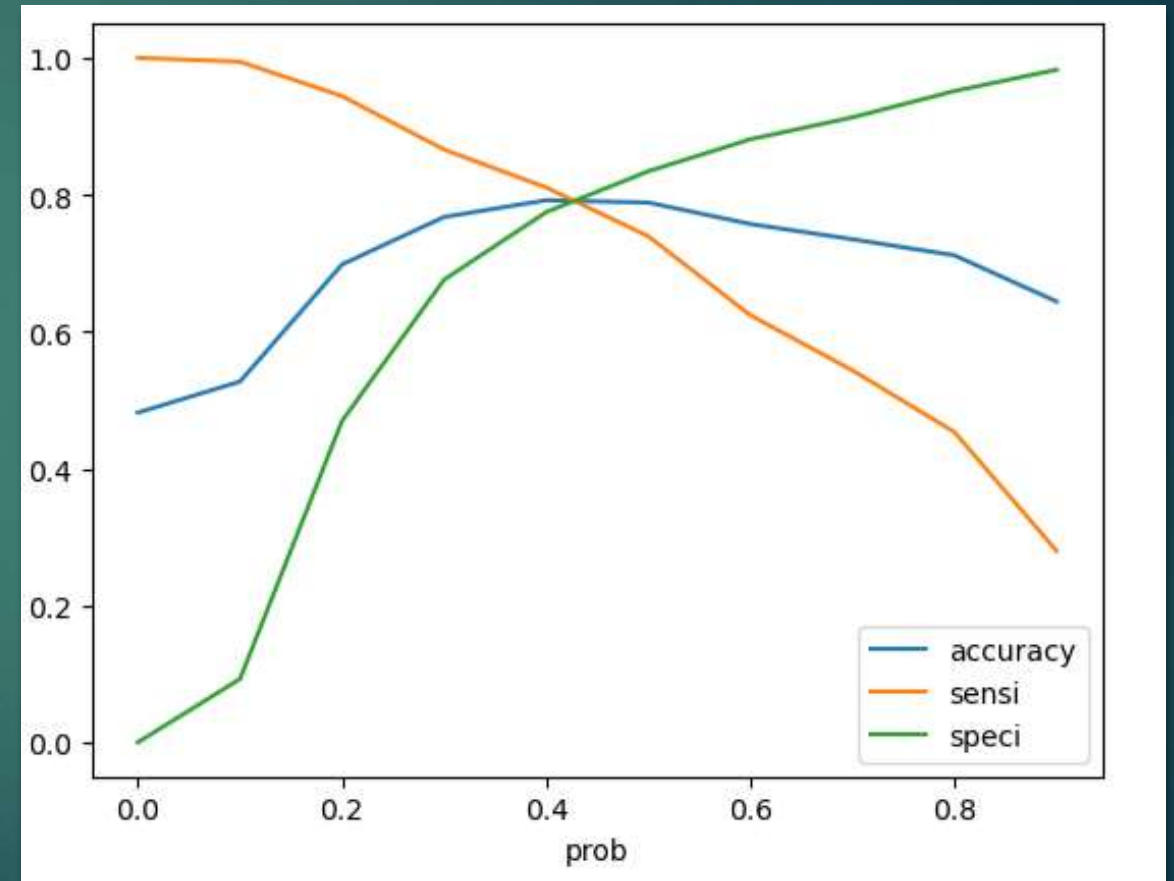
Threshold

▶The intersection point of the curve is the threshold value where the model achieves a balance between precision and recall. It can be used to optimize the performance of the model based on business requirement, Here our probability threshold is 0.42 approx. from above curve.
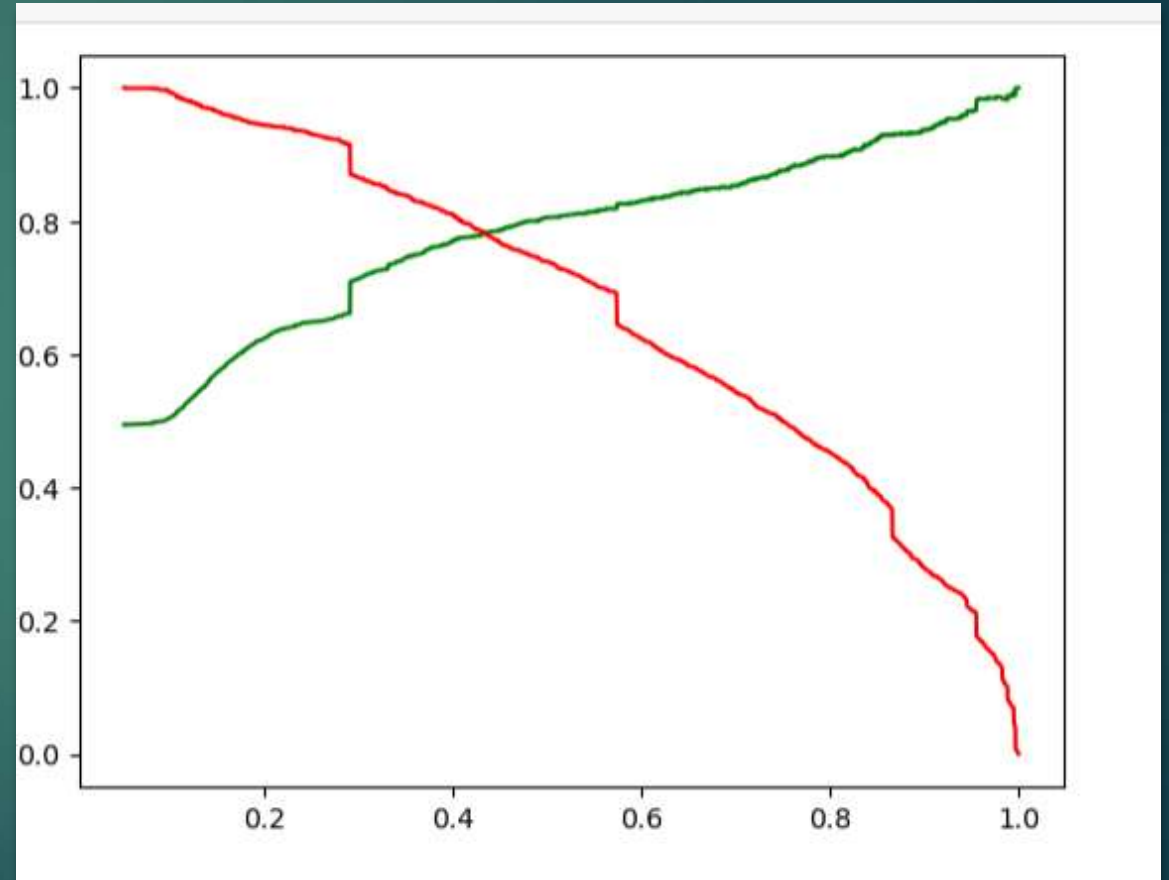
Conclusion

- The Customer/Leads who fills the form are the potential leads

- We should target on the working professional

- We should target leads for which the Tag value is "Will revert after reading the email"

- We should target such leads for whom Last Activity is SMS Sent

- We can go for a lower cut-off threshold value so that we can target more and more "Hot Leads" and adopt a more aggressive strategy

Recommendation

▶ Its good to collect often and run the model and get updated with the potential leads. There is a belief that the best time to call your potential leads is within few hours after the lead shows interest in the courses.

▶ Concentrate efforts on leads tagged as "Will revert after reading the email.

▶ Give preference to leads with the last activity recorded as SMS Sent.

▶ Explore the option of using a lower cut-off threshold to aggressively target and capture more "Hot Leads.

▶ Engage working professionals with personalized messaging.

▶ Offer incentives/discounts for successful referrals to encourage more references

Thank you