LEAD SCORING ASSIGNMENT

Nikesh Jain | Aanchal Pahwa | Aastha Gautam

DS C55

AGENDA

PROBLEM STATEMENT
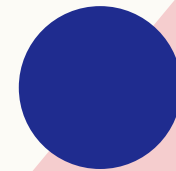
X Education offers online courses for industry professionals, promoting them on platforms such as Google.

The company aims to identify the most promising leads with a higher potential for conversion to paying customers. Despite generating numerous leads through various channels like email, website ads, and Google searches, the conversion rate stands at 30%.

The process of converting leads into customers, particularly targeting those interested in the courses, lacks efficiency.

BUSINESS GOAL

X Education seeks to identify promising leads and plans to create a predictive model for recognizing 'hot leads.'

The company intends to deploy this model for future use in lead identification and conversion processes.

METHODOLOGIES

Data Cleaning and Data Manipulation:
- Detect and address duplicate data.
- Identify and manage NA values and missing data.
- Remove columns with a significant number of missing values that are irrelevant for analysis.
- Impute missing values if required.

Exploratory Data Analysis (EDA):
- Conduct univariate data analysis, including value counts and variable distributions.
- Perform bivariate data analysis to examine correlation coefficients and relationships between variables.

Feature Scaling & Dummy Variables:
- Apply feature scaling to ensure all variables have the same scale.
- Create dummy variables to represent categorical data through encoding.

Classification Technique (Logistic Regression):
- Utilize logistic regression for model building and prediction. This method is suitable for binary classification tasks.

METHODOLOGIES

Model Validation:

- Validate the model to assess its performance and reliability. This may involve techniques such as cross-validation, splitting data into training and testing sets, and evaluating metrics like accuracy, precision, recall, and F1 score.

ASSUMPTIONS

- ✓ *In this case study, columns with missing value percentages exceeding 30% ('Lead Quality,' 'Asymmetrique Activity Index,' 'Asymmetrique Profile Score,' 'Asymmetrique Activity Score,' 'Asymmetrique Profile Index') were removed. Imputing such extensive missing data could negatively affect the analysis.*
- ✓ *- 'SELECT' values in certain columns were replaced with null..*
- ✓ *- For numerical columns where the mean and median were identical but the maximum value was far outside the range, null values were imputed with 1.5 times the Interquartile Range (IQR) of the variable.*
- ✓ *- Unnecessary columns with heavily skewed data were dropped to prevent adverse impacts on overall model building.*
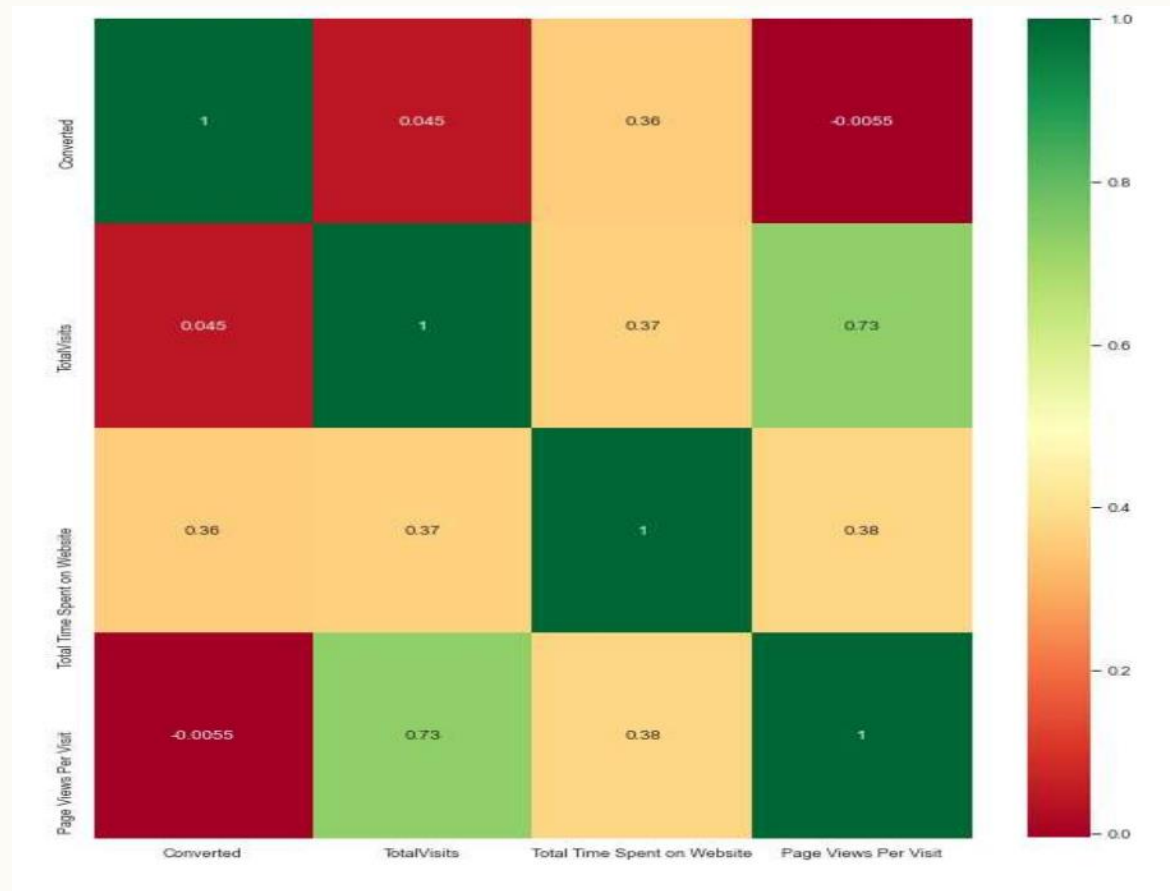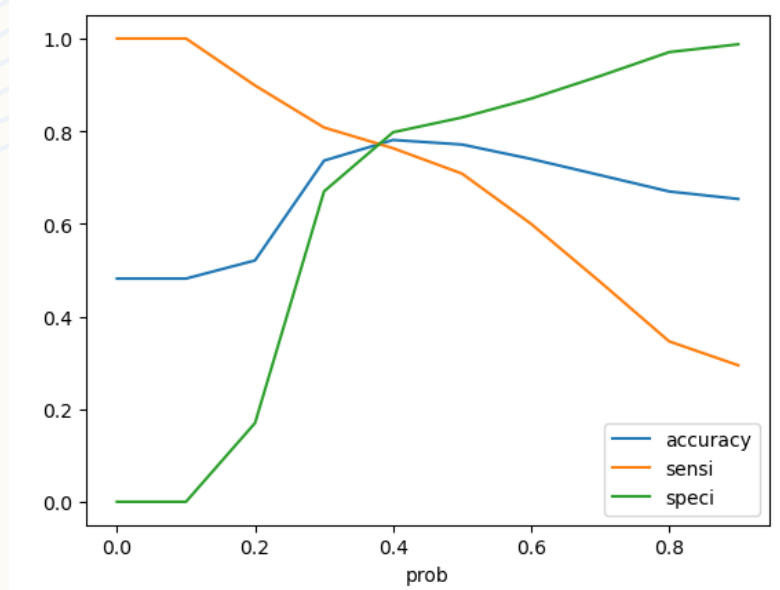
EXPLORATORY DATA ANALYSIS
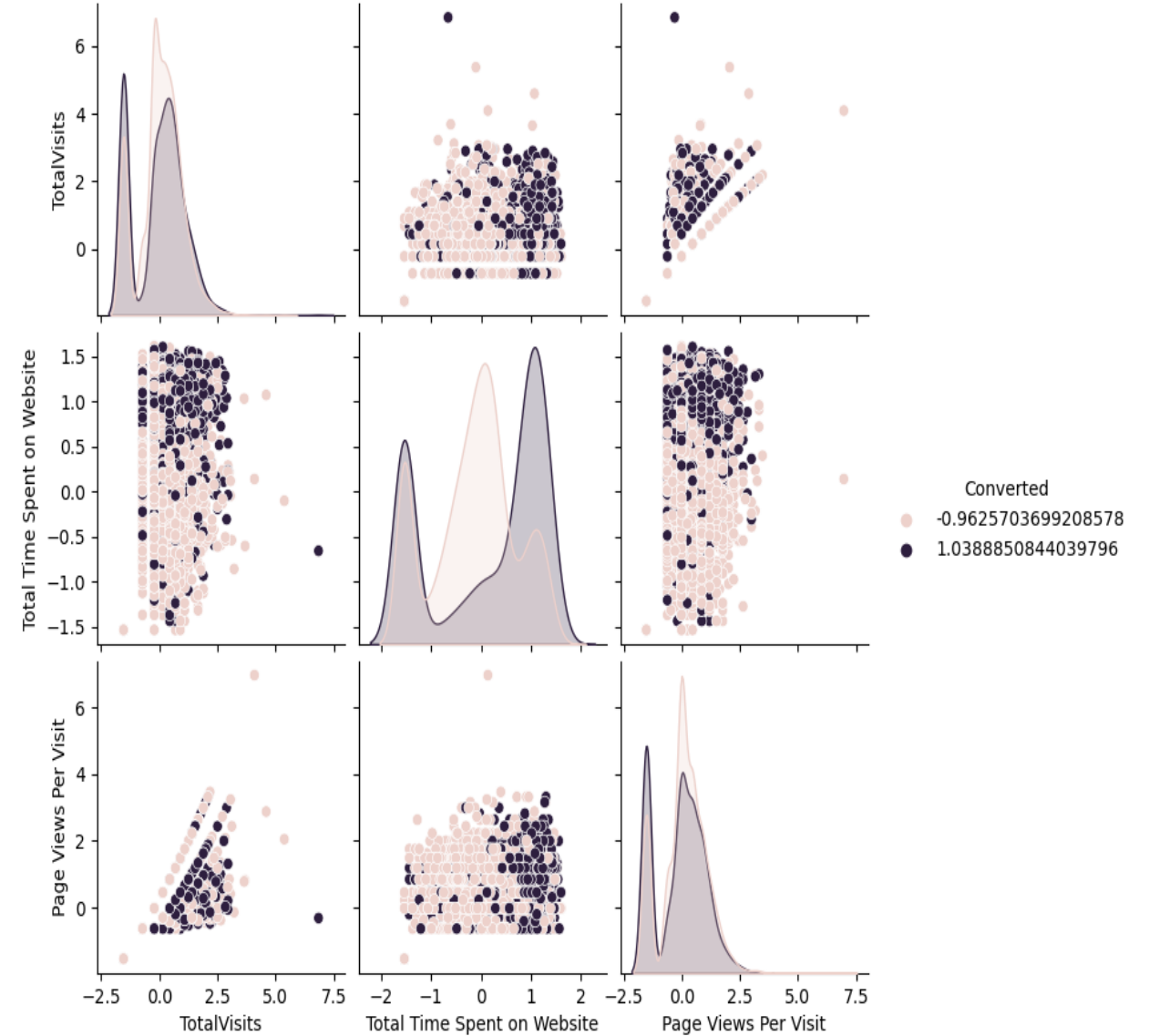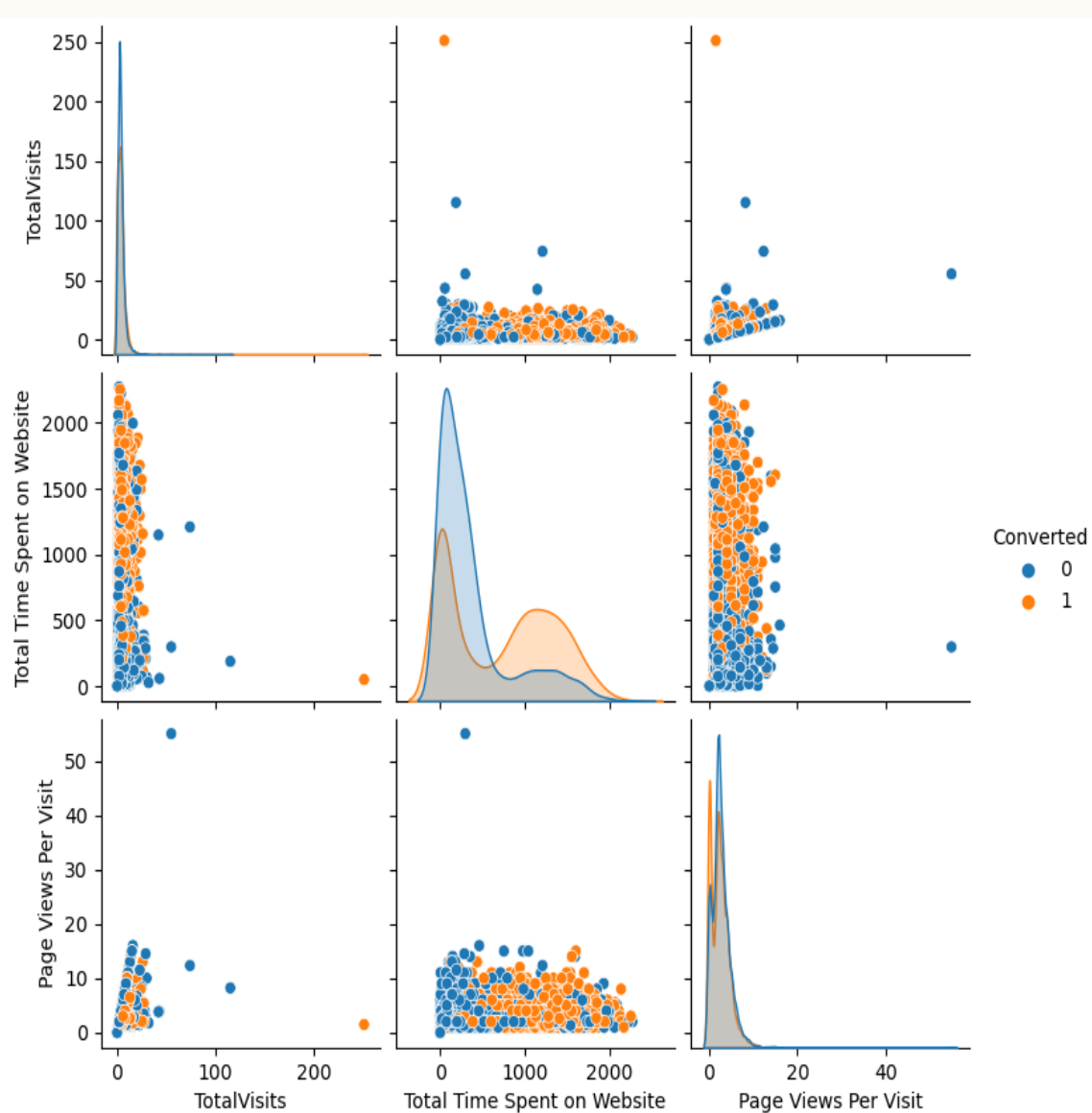
CORRELATION MATRIX

From above Correlation Matrix, we can see that
- ✓ Converted is having positive correlation with Total Time Spent on Website
- ✓ And negative relationship with Page Views Per Visit

MODEL EVALUATION(TRAIN)

Accuracy Sensitivity & Specificity

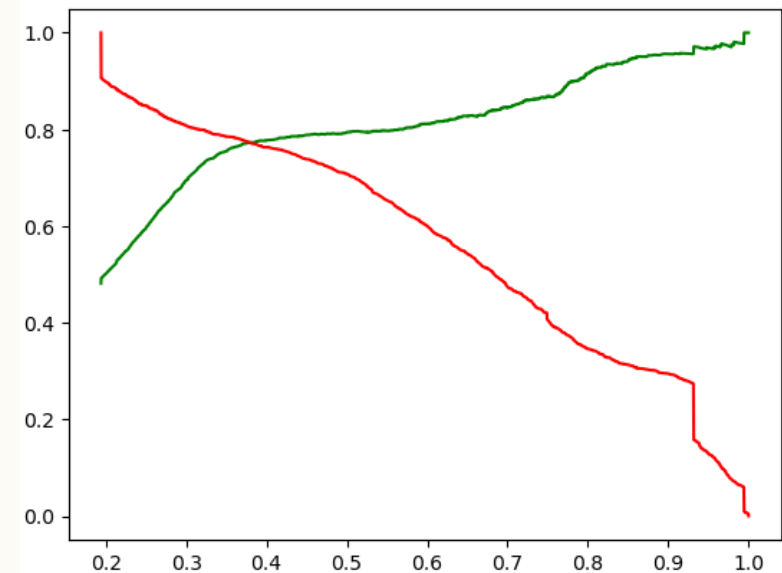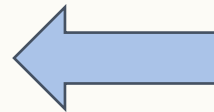|	
1860	452
523	1626

- Accurracy
  78%
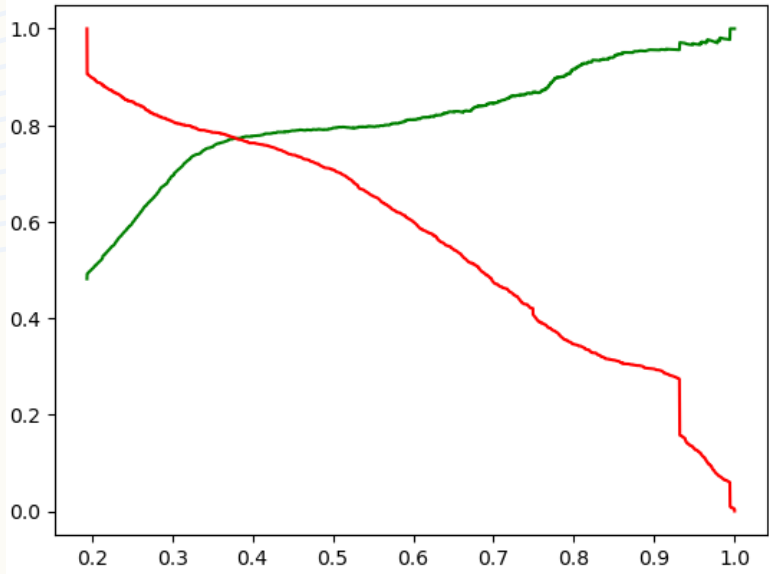- Sensitivity
  75%
- Specificity
  80%

- Precision
  79%
- Recall
  75%

MODEL EVALUATION(TEST)

Accuracy Sensitivity & Specificity

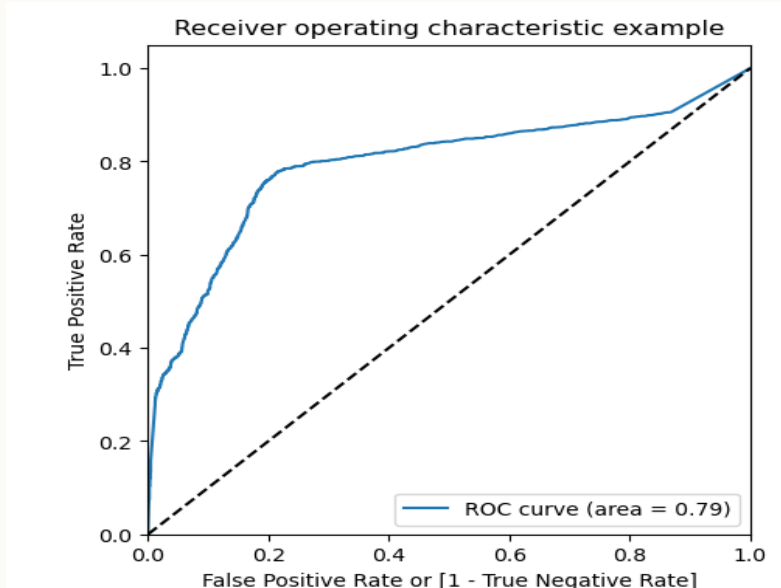|	
804	192
236	680

- Precison
  79%
- Recall
  74%

ROC CURVE(Reciever Operating Characteristic)
The area under the curve of the ROC is 0.79 which is quite good. So we seem to have a good model.

AREAS OF FOCUS

✓ **Prospects who spend more time on the website** are more likely to become Hot Leads, suggesting that the Sales team should prioritize reaching out to them.

✓ **Lead Scores associated with Welingak websites and referrals** exhibit the highest conversion rates, indicating the potential for additional marketing efforts on these websites. The Sales team can also send course details and promotional offers to existing users to generate more Hot Leads.

✓ **Leads contacted through email or SMS communication** channels have a higher likelihood of conversion.

✓ **Targeting the Unemployed and Working Professionals** within the Occupation category can generate more leads by providing information about the available courses.

SUMMARY

Logistic Regression Model Implementation

- The final model achieved an accuracy of 78.5%, with a Recall of 74.8% and Precision of 77.9%.

- The optimal cut-off point was determined by balancing Precision and Recall scores.

- When applied to the test dataset, the model performed well, yielding a Recall of 74.4% and a Precision of 77.9%.

- In summary, the model appears to be effective in identifying potential leads with a high likelihood of conversion using the Lead Score as a predictive metric.

THANK YOU