LEAD SCORING CASE STUDY

PROBLEM STATEMENT

1.Challenge: X education struggling with low lead conversion rate.(30%)

2.Objective: Enhance lead conversion rate to CEO's target of 80%

3.Need: Develop a predictive model to score leads and prioritize follow-up.

SOLUTION METHODOLOGY

1.Approach: Develop a predictive model using logistic regression

2.Methodology: Data preprocessing, data engineering, model building and evaluation.

3.Emphasis: Achieve a balanced trade-off between accuracy, sensitivity and specificity.

DATA MANIPULATION

- Handling missing values, and dropping columns with excessive nulls.

- Imputing categorical variables based on value distribution.

- Treating outliers, invalid data and low frequency values.

- Columns having null values of more than 40% are how did you hear about X education, Lead quality, Lead profile, Asymmetrique activity index, Asymmetrique profile index, Asymmetrique activity score and Asymmetrique profile score.

EXPLORATORY DATA ANALYSIS

- Analyzing relationship between variables and lead conversion.
- Identifying key variables such as 'Lead origin', 'Current occupation', and 'Lead source'.

BIVARIATE ANALYSIS
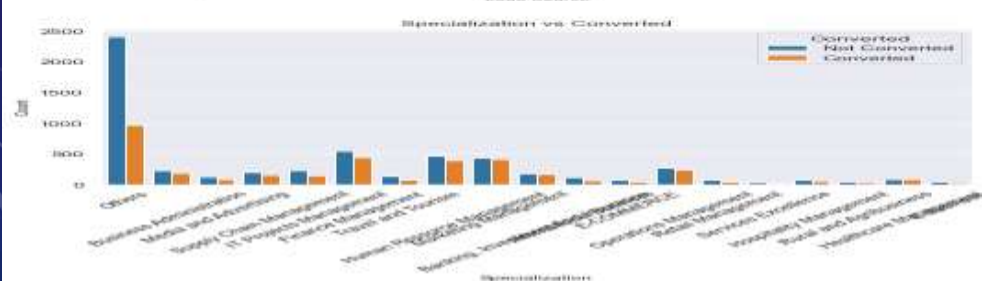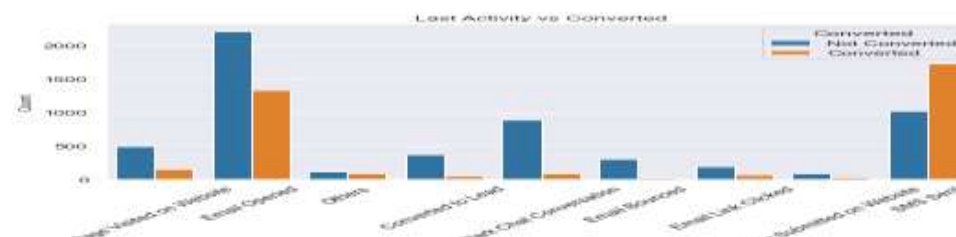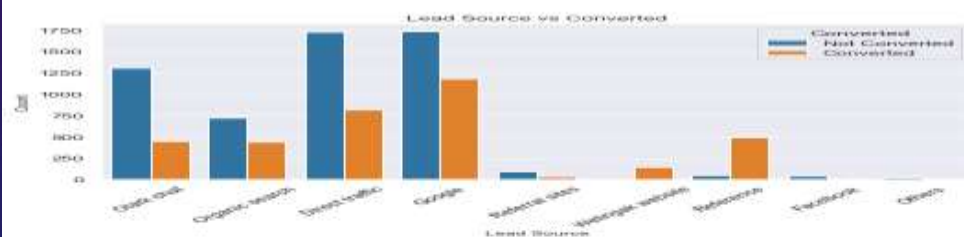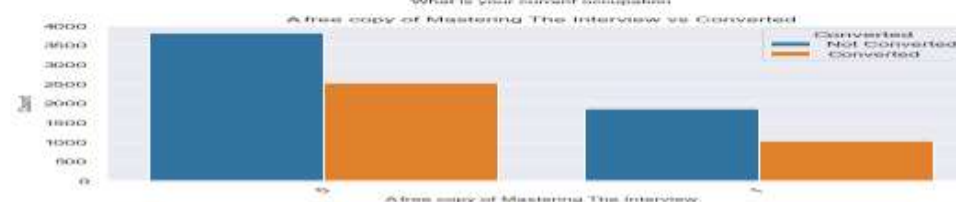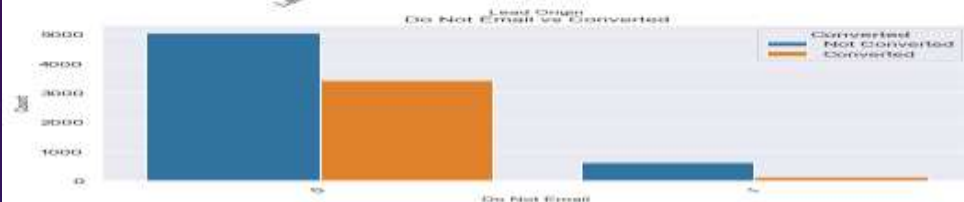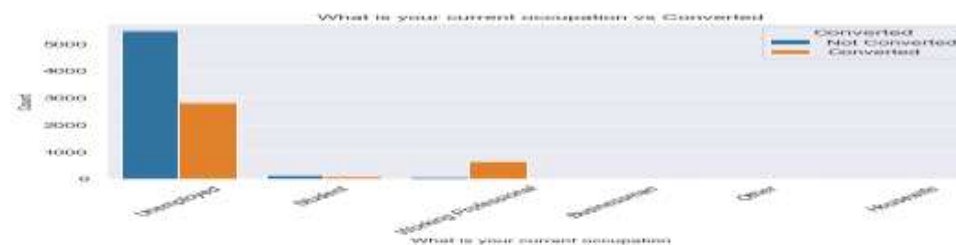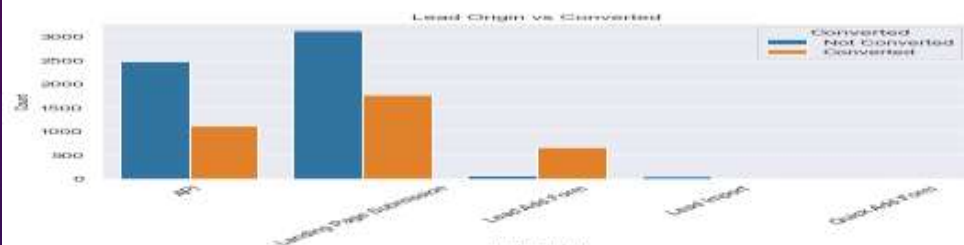
MULTIVARIATE ANALYSIS
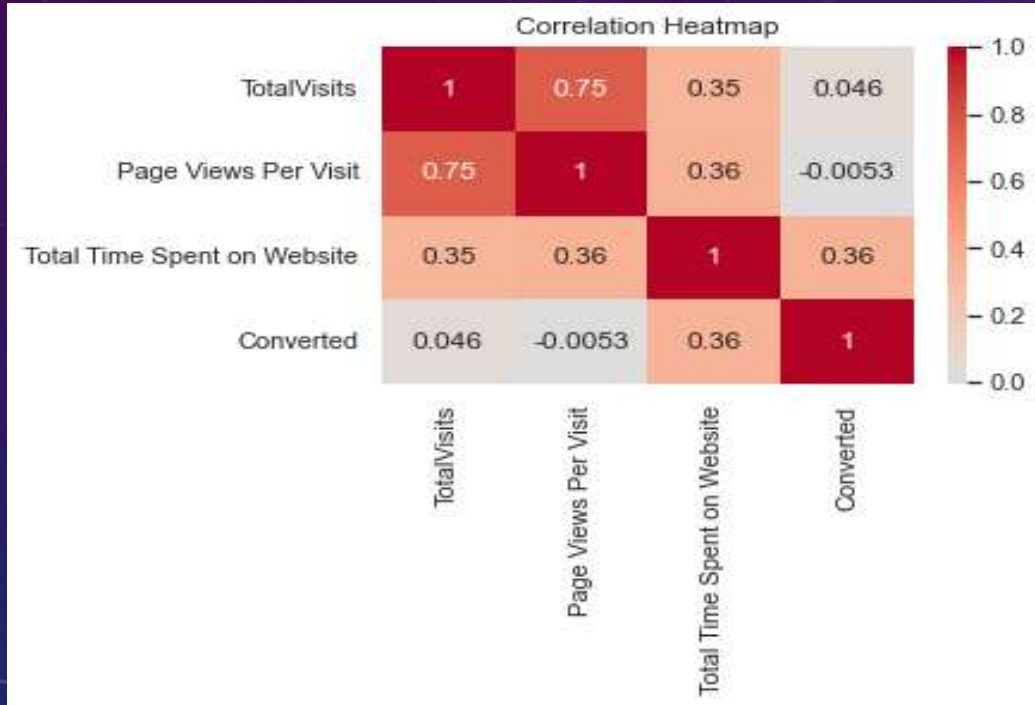
- Total visits vs page views per visit 0.75
- Total visits vs total time spent on website 0.35
- Total time spent on website vs converted 0.36
- Page views per visit vs converted -0.01

DATA CONVERSION

- Numerical variables are normalized

- Creating dummy variables for categorical features using one-hot encoding

- Scaling numerical variables using standardization.

MODEL BUILDING

- Splitting the data into training and testing sets
- The first basic step for regression is performing a train test split, we have chosen 80:20 ratio.
- Using Recursive Feature Elimination(RFE) and manual feature reduction.
- Selecting model 7 based on significance and multicollinearity.
- Validating model using p-values, VIF and other diagnostic tests.

ROC CURVE

Receiver operating characteristic example
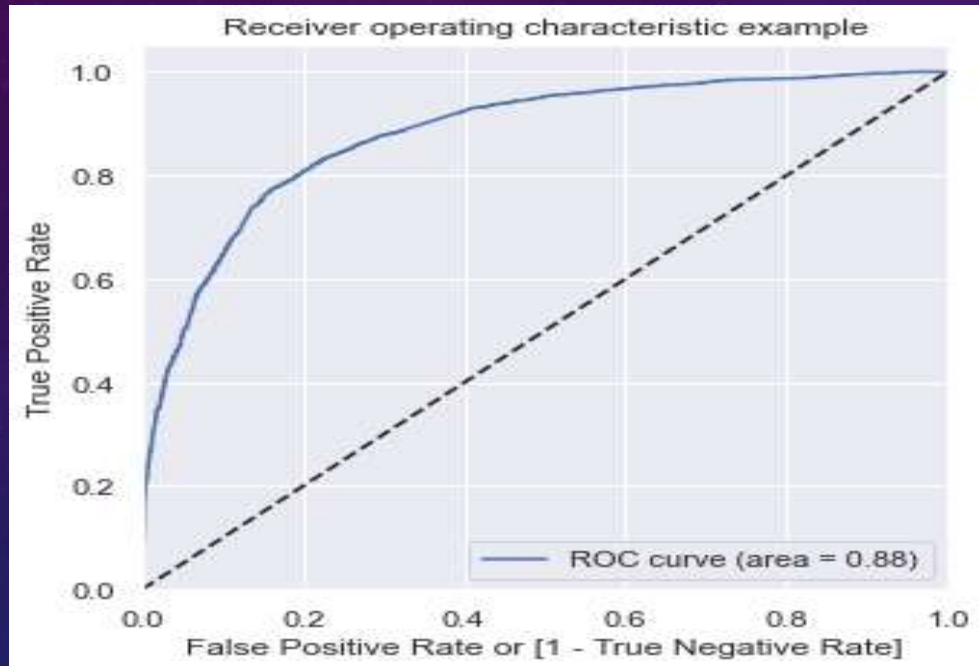
- Visual representation of the models, true positive rate against false positive rate.

- Interpretation: The ROC curve has an area of 0.88, indicating that the model performs well in differentiating between the 2 classes. The closer the AUC value is to 1, better the model's performance.

PREDICTION OF TEST DATA

- Standardise the test data.

- Applying the final model to predict lead conversion on the test data.

- Assigning lead scores using the optimal cut-off.

- The accuracy score we found was 0.816, precision 0.7173 and recall 8396.

- Lead score is created on test data set to identify hot leads - high the lead score higher the chance of conversion, low the lead score lower the chance of getting converted.

CONCLUSION

- Summary of project journey: Data cleaning, exploration, modelling and evaluation.

- Successful achievement of 80% conversion rate using data driven insights.

- High potential for optimizing lead conversion in the future.

- The model also achieved an accuracy of 81.60%, which is in line with the study's objectives.

THANK YOU