LEAD SCORING CASE STUDY

BY: NRK PAVAN

TABLE OF CONTENTS

PROBLEM STATEMENT

- Challenge: X Education struggling with a low lead conversion rate (~30%).
- Objective: Enhance lead conversion rate to the CEO's target of ~80%.
- Need: Develop a predictive model to score leads and prioritize follow-up.

SOLUTION METHODOLOGY

Approach: Develop a predictive model using logistic regression.

Methodology: Data preprocessing, feature engineering, model building, and evaluation.

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

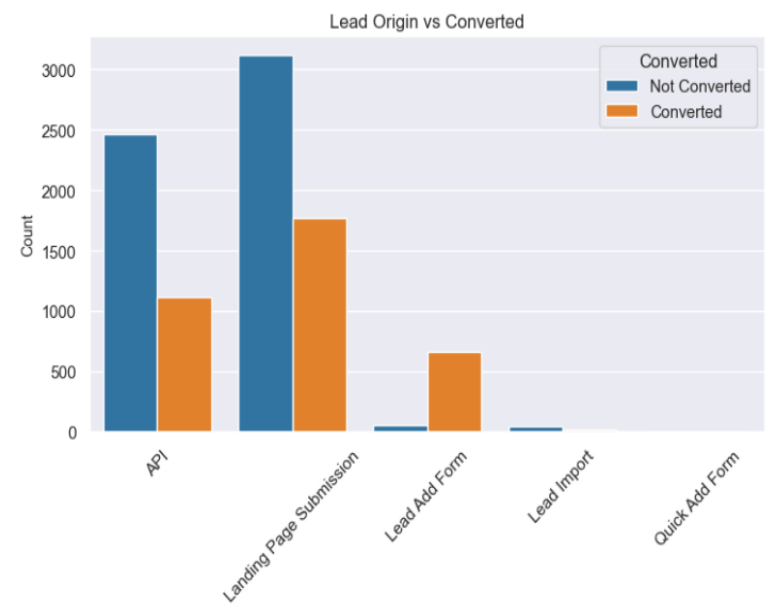
DATA MANIPULATION

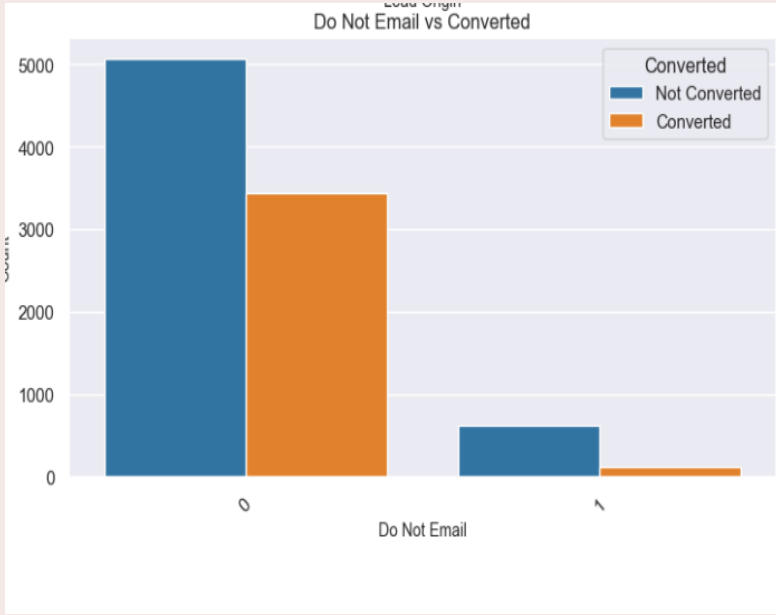
- Handling missing values, and dropping columns with excessive nulls.
- Imputing categorical variables based on value distributions
- Treating outliers, invalid data, and low-frequency values.
- Columns having a null value of more than 40% are How did you hear about X Education, Lead Quality, Lead Profile, Asymmetrique Activity Index, Asymmetrique Profile Index, Asymmetrique Activity Score, and Asymmetrique Profile Score.

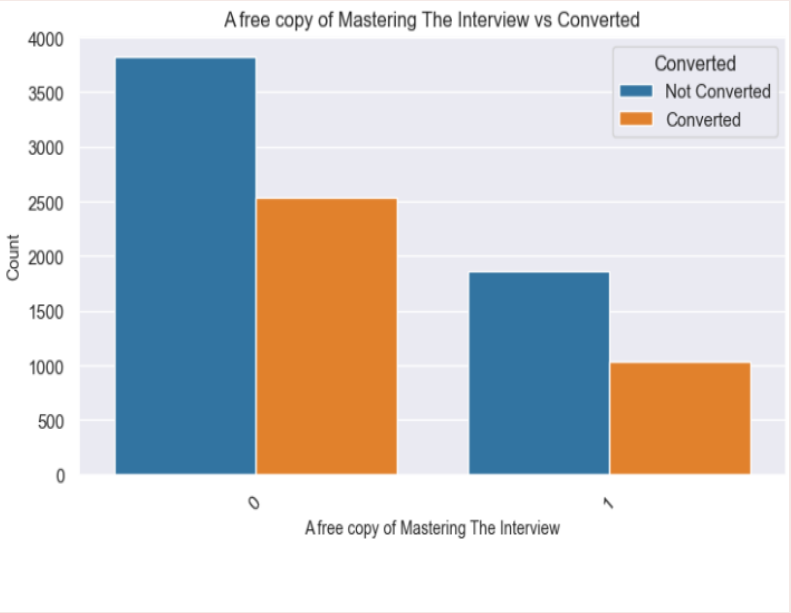
EXPLORATORY DATA ANALYSIS

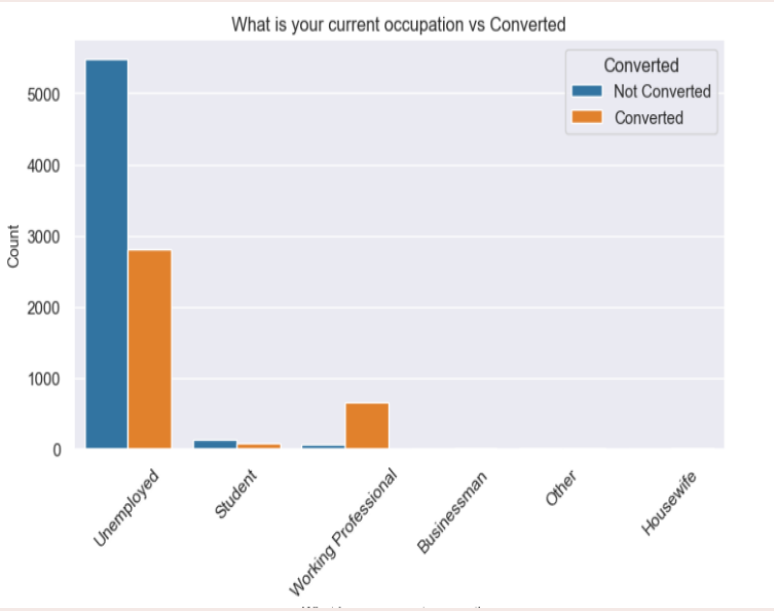
- Analyzing relationships between variables and lead conversion.
- Identifying key variables such as 'Lead Origin', 'Current Occupation', and 'Lead Source'.

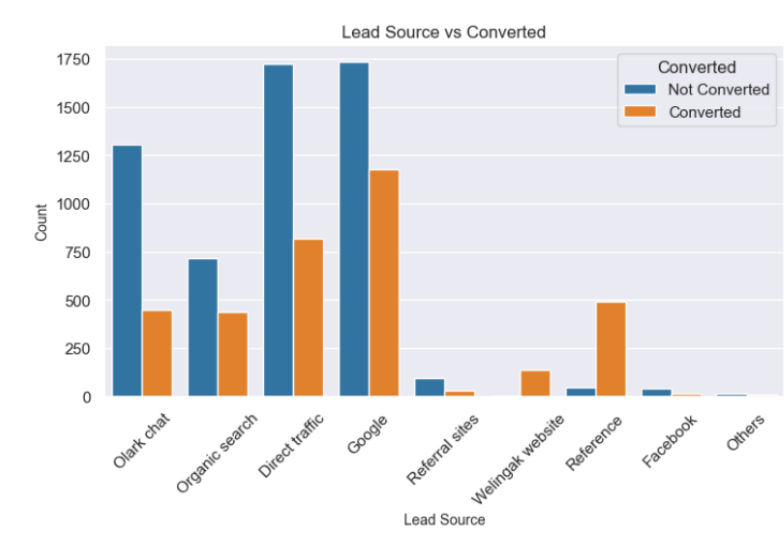
BIVARIATE ANALYSIS

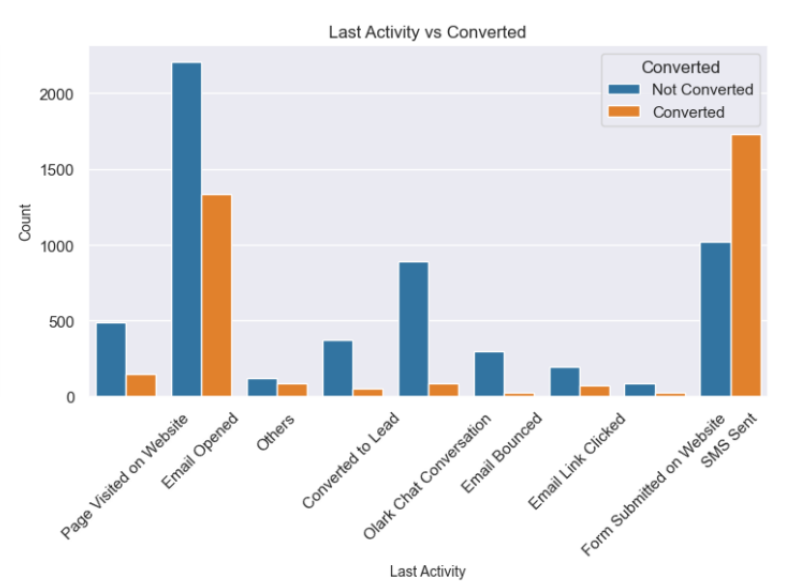






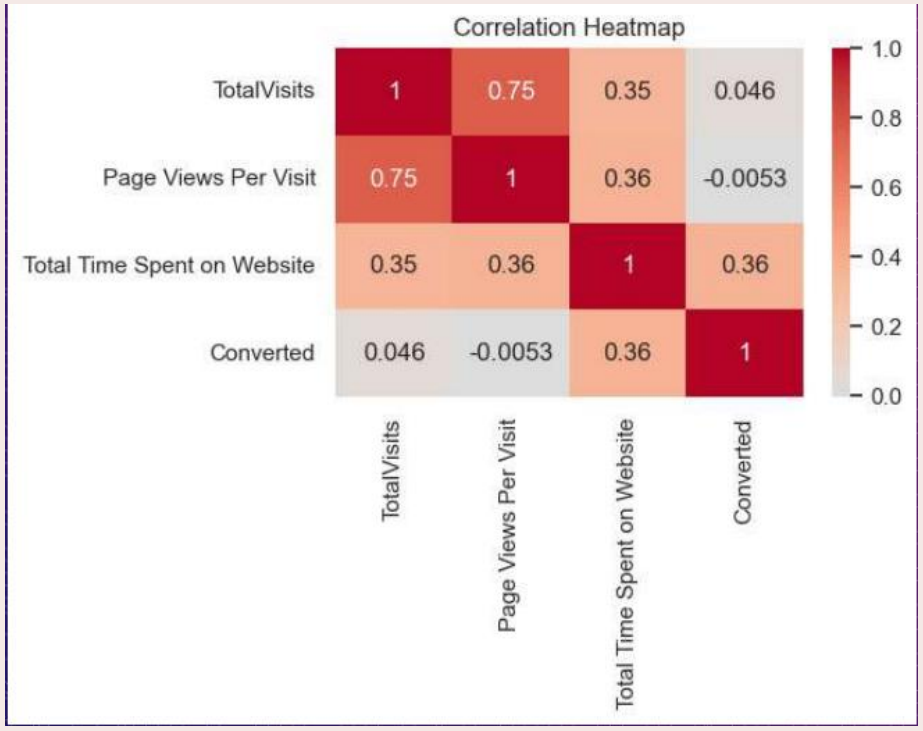






MULTIVARIATE ANALYSIS

- 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

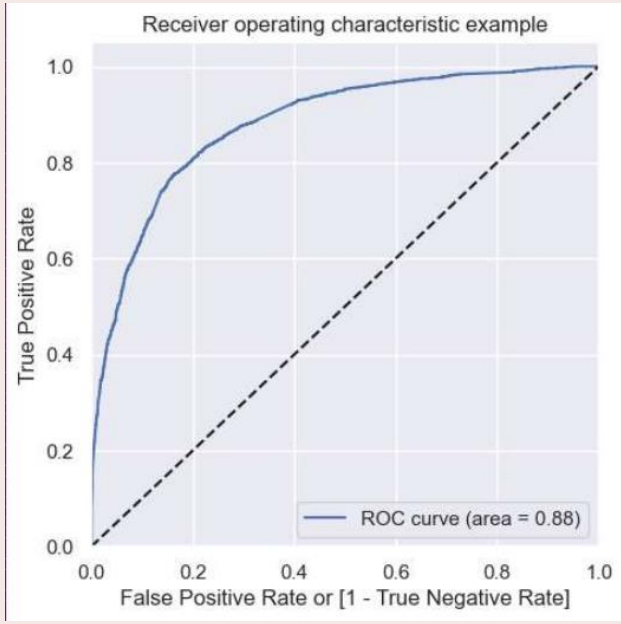
- Numerical Variables are normalized
- Creating dummy variables for categorical features using one-hot encoding.
- Scaling numerical variables using standardization.

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

- Visual representation of the model's true positive rate (sensitivity) against false positive rate.
- Interpretation: The ROC curve has an area of 0.88, indicating that the model performs well in differentiating between the two classes. The closer the AUC value is to 1, the better the model’s performance.



PREDICTION OF TEST DATA

- standardize the test set
- Applying the final model to predict lead conversion on 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 dataset 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, modeling, 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.

RECOMMENDATIONS

- Allocating additional budget to promote the Welingak Website for improved lead acquisition.
- Implementing incentives or discounts for lead referrals to encourage conversions.
- Targeted marketing campaigns for working professionals to capitalize on higher conversion potential.

THANK YOU