



Lead Scoring Case Study

SUMMARY BY

BIJI KRISHNA , BAPPI BANIK,
SHANTANU BISWAS

Problem Statement

- ▶ X Education sells online courses to industry professionals.
- ▶ X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- ▶ To make this process more efficient, 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 Objective:

- ▶ X education wants to know most promising leads.
- ▶ For that they want to build a Model which identifies the hot leads.
- ▶ Deployment of the model for the future use.

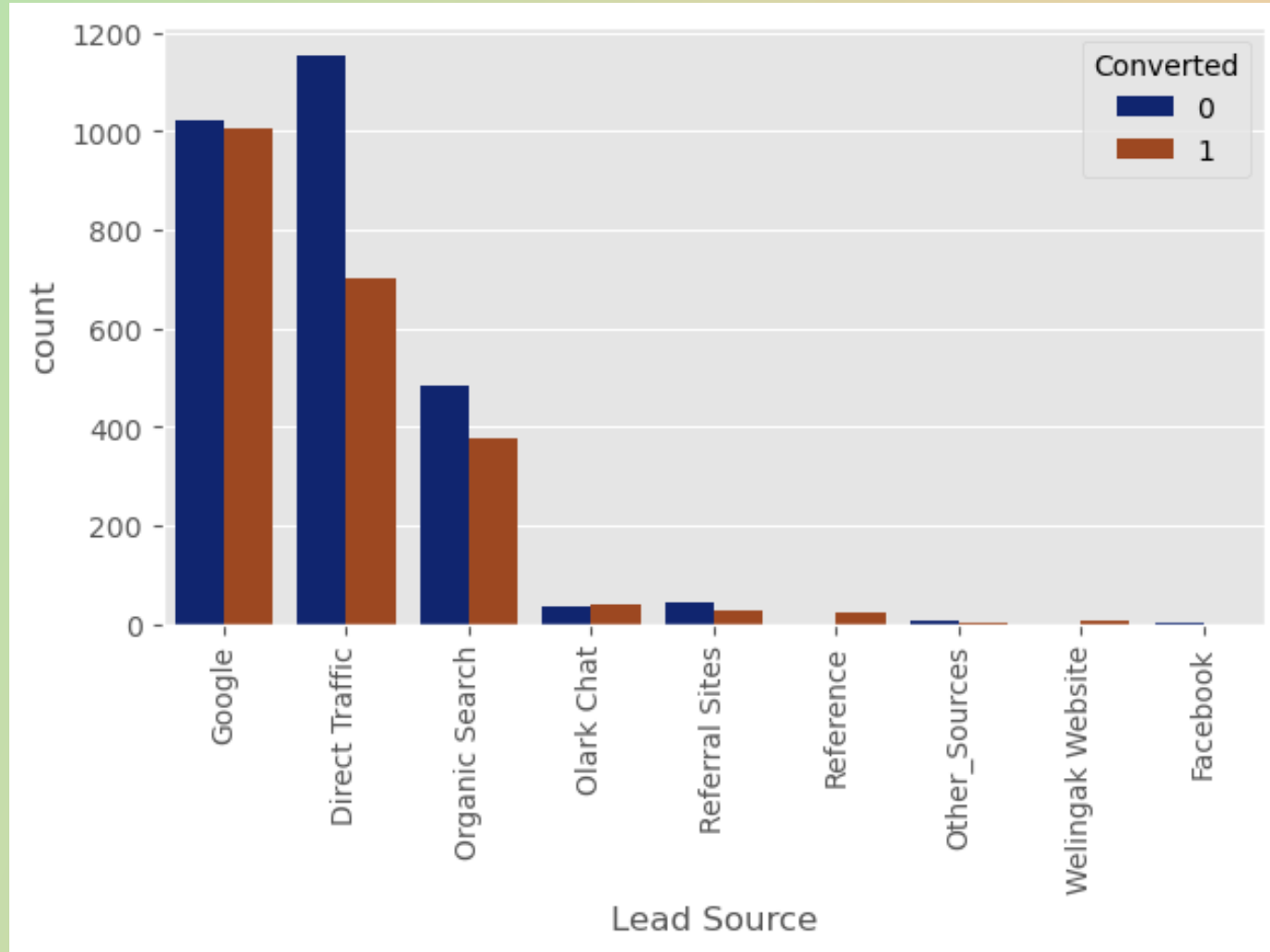
Solution Methodology

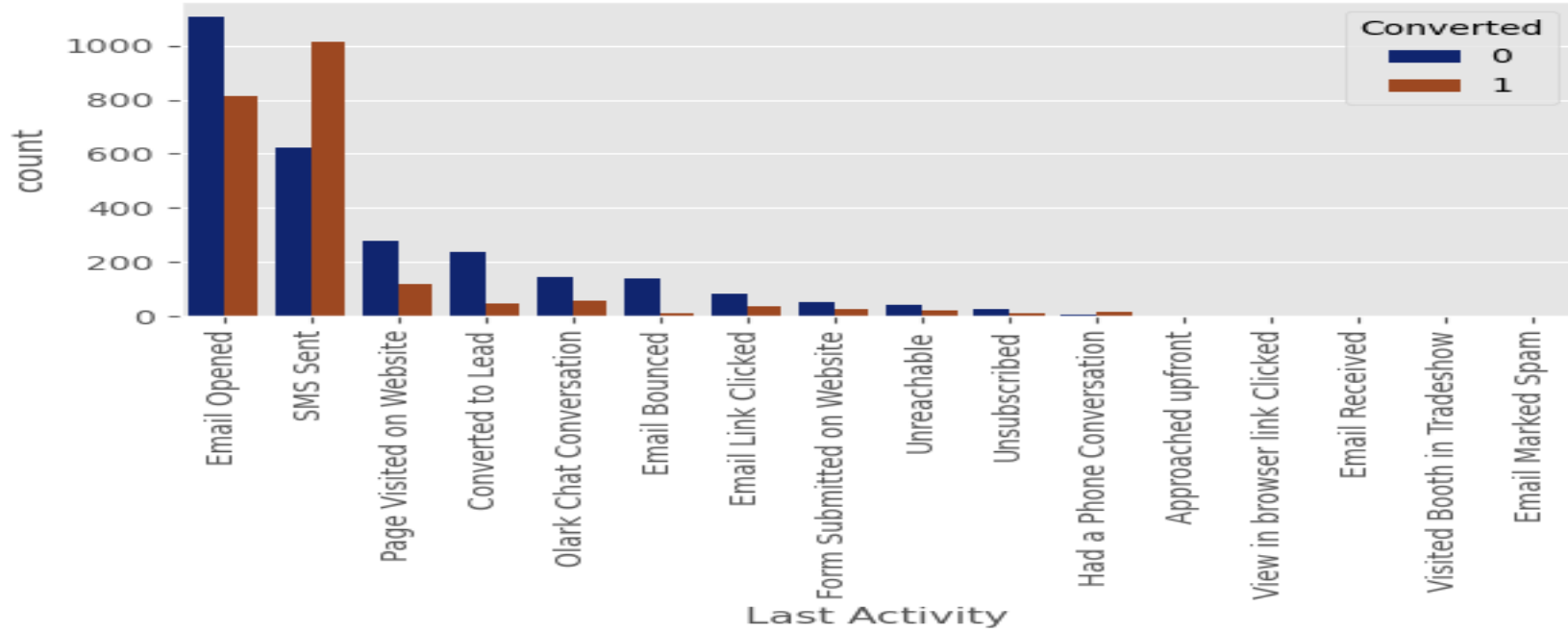
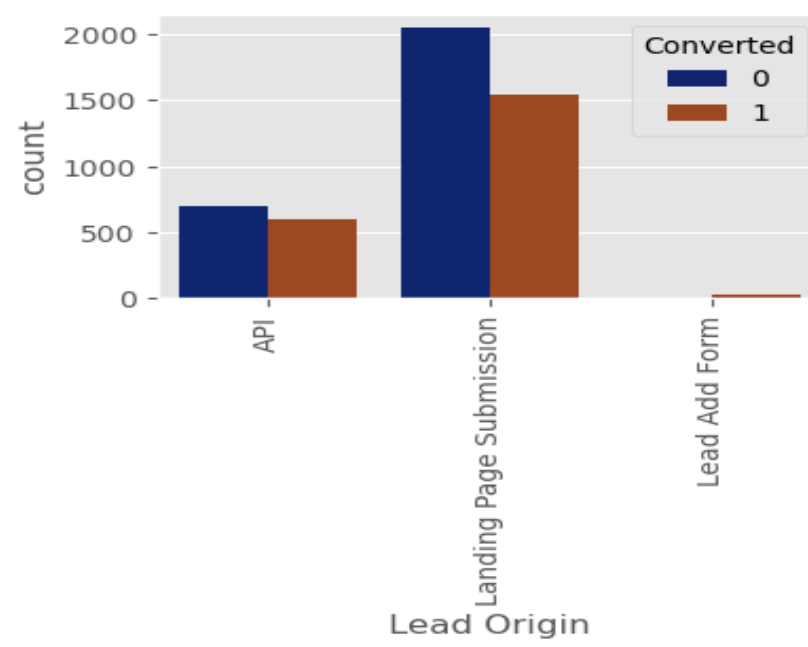
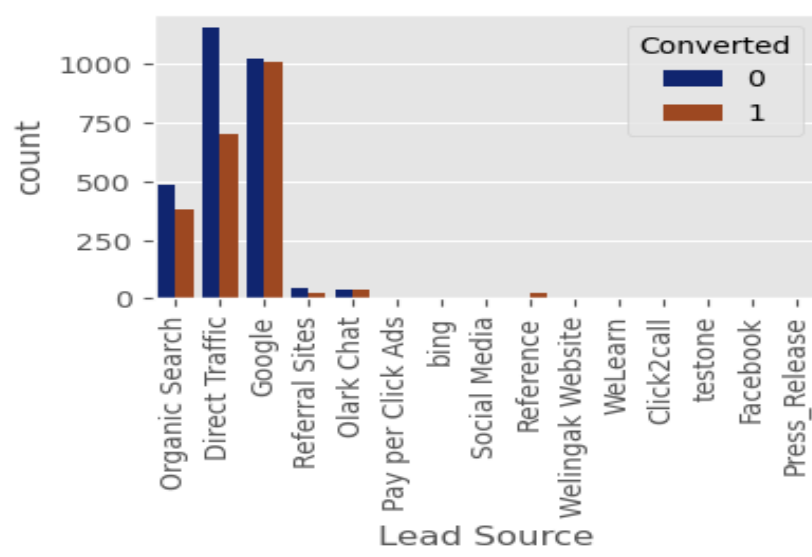
- ▶ Data cleaning and data manipulation.
 1. Check and handle duplicate data.
 2. Check and handle NA values and missing values.
 3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
 4. Imputation of the values, if necessary.
 5. Check and handle outliers in data.
- ▶ EDA
 1. Univariate data analysis: value count, distribution of variable etc.
 2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.
- ▶ Feature Scaling & Dummy Variables and encoding of the data.
- ▶ Classification technique: logistic regression used for the model making and prediction.
- ▶ Validation of the model.
- ▶ Model presentation.
- ▶ Conclusions and recommendations.

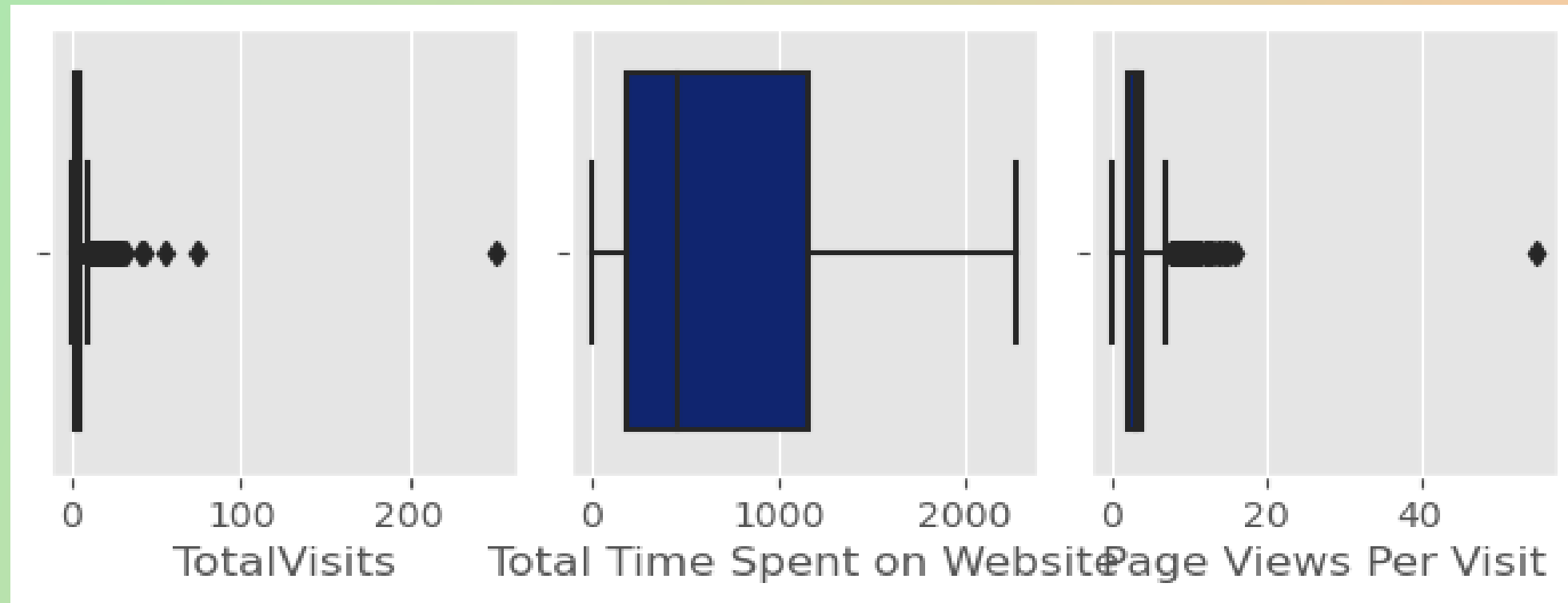
Data Manipulation

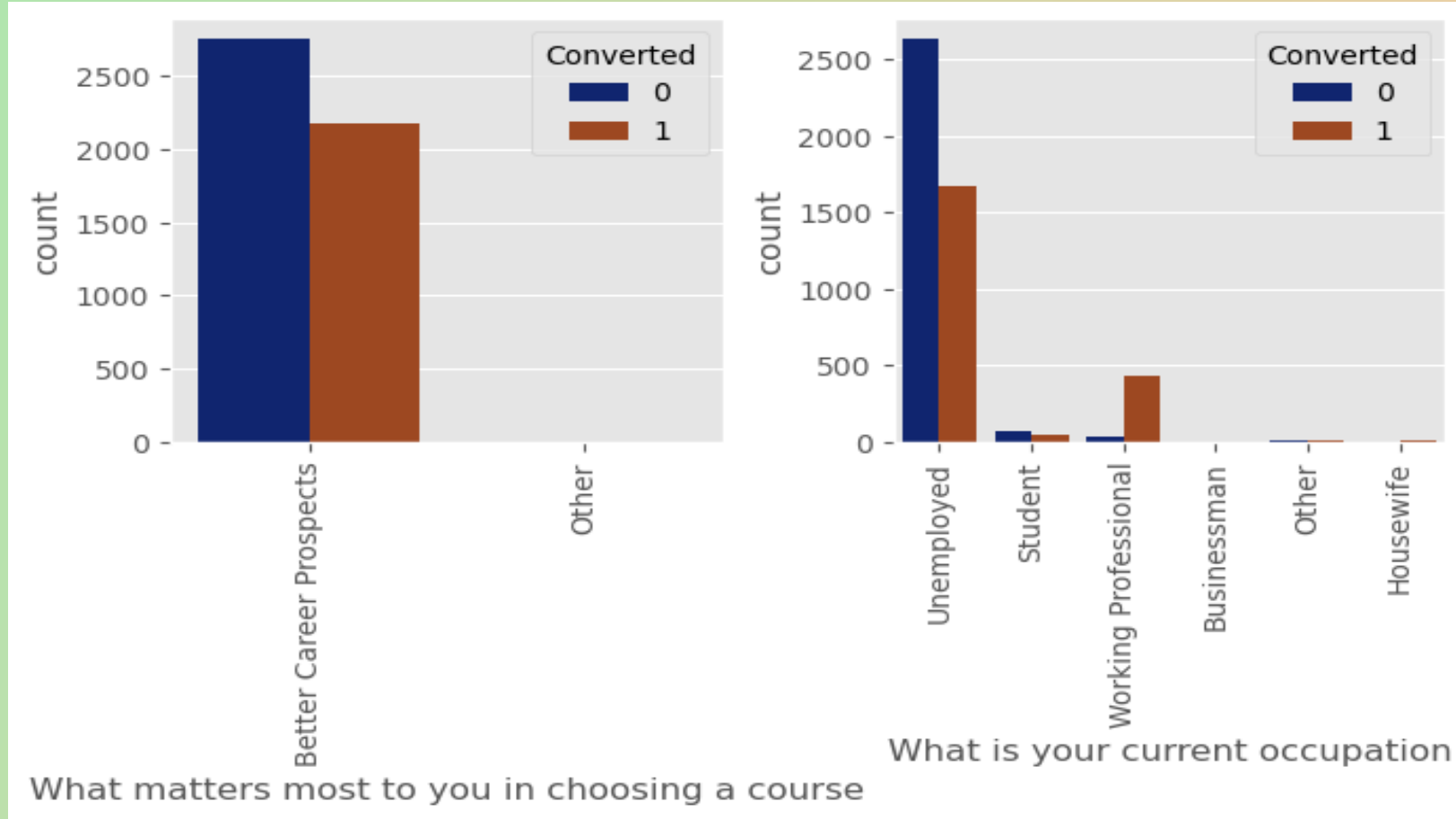
- ▶ Total Number of Rows =37, Total Number of Columns =9240.
- ▶ Single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply”
- ▶ Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped.
- ▶ Removing the “Prospect ID” and “Lead Number” which is not necessary for the analysis.
- ▶ After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped, the features are: “Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement” etc.
- ▶ Dropping the columns having more than 35% as missing value such as ‘How did you hear about X Education’ and ‘Lead Profile’.

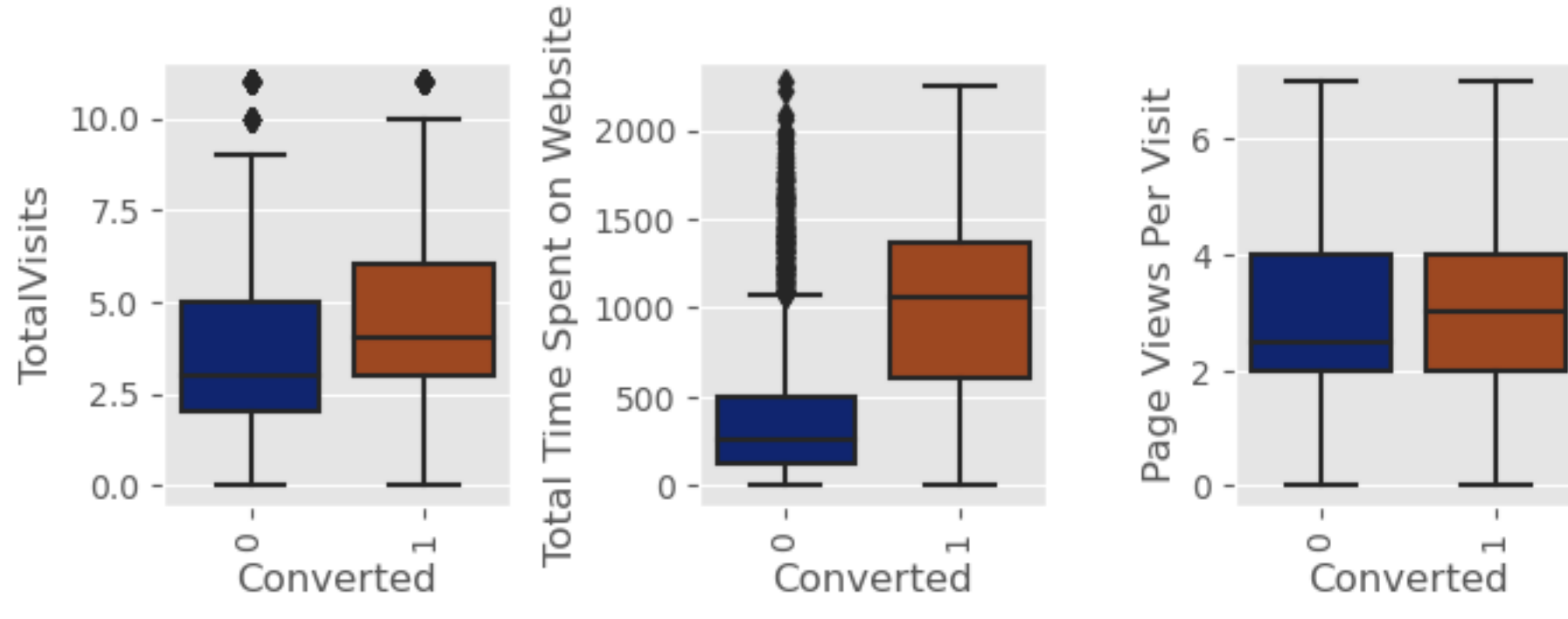
EDA











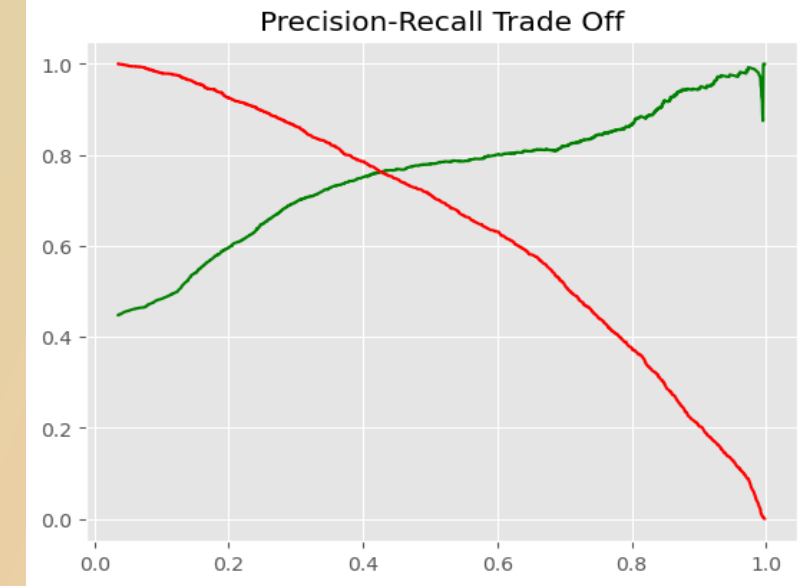
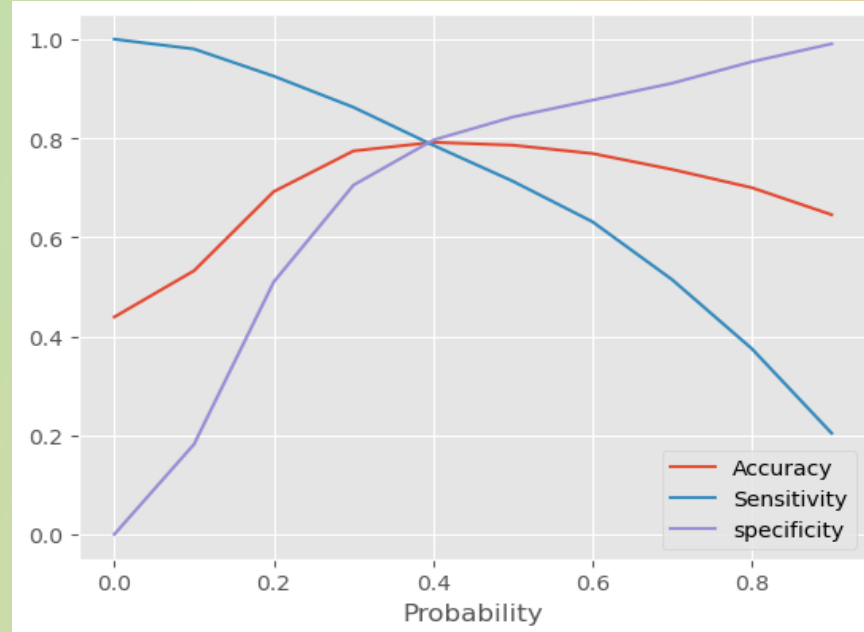
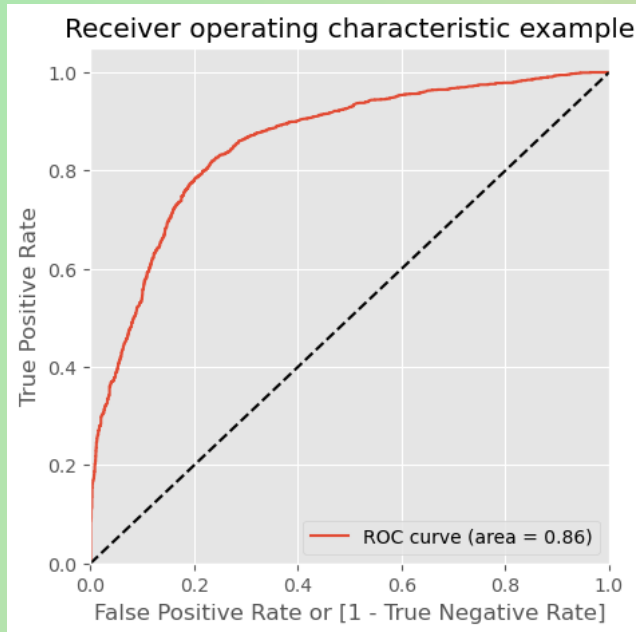
Data Conversion

- ▶ Numerical Variables are Normalised
- ▶ Dummy Variables are created for object type variables
- ▶ Total Rows for Analysis: 8792
- ▶ Total Columns for Analysis: 43

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 70:30 ratio.
- ▶ Use RFE for Feature Selection
- ▶ Running RFE with 15 variables as output
- ▶ Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5
- ▶ Predictions on test data set
- ▶ Overall accuracy 79%

ROC Curve



- **Finding Optimal Cut off Point**
- Optimal cut off probability is that
- probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut off is at 0.38.

Conclusion

- The Logistic regression model predicts the probability of the target variable having certain value. The cut off value is used to obtain predicted value of target variable.
- Optimum cut off value is selected at 0.38 , any lead with probability greater than 0.38 can be considered as Hot Lead and any lead with less than 0.38 probability can be considered as a Cold Lead.
- Final model has 10 features :- 'Total Time Spent on Website', 'Last Activity Converted to Lead', 'Last Activity Email Bounced', 'Last Activity_Had a Phone Conversation', 'Last Activity_SMS Sent', 'What is your current occupation Student', 'What is your current occupation_Unemployed', 'What is your current occupation_Working Professional', 'Last Notable Activity Modified', 'Last Notable Activity Unreachable'
- Top 3 Factors are
 1. Last Notable Activity_Unreachable with coefficient 2.346566
 2. Last Activity_Had a Phone Conversation with coefficient 2.312539
 3. What is your current occupation_Working Professional with coefficient 1.258886
- Train Accuracy Score: 0.789962286045837 Sensitivity : 0.799074686054197 Specificity : 0.7828335056876939 Precision Score : 0.7421731123388582 Recall Score : 0.799074686054197
- Test Accuracy Score: 0.7821380243572396 Sensitivity : 0.7954887218045112 Specificity : 0.7712177121771218 Precision Score : 0.7398601398601399 Recall Score : 0.7954887218045112
- Accuracy, Sensitivity and Specificity values of test set are around 78%, 79% and 77% which are approximately closer to the respective values calculated using trained set.
- Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 0.799074686054197 (train) / 0.7954887218045112 (test)
- Hence overall this model seems to be good.
- The final model has Precision of 0.74, this means 74% of predicted hot leads are True Hot Leads.
- Final Prediction conversion on both train and test set is around 80%+ which is in line with the target
- Thank you!!! Bappi Banik ,Biji Krishna, Shantanu Biswas.

Thank You