LEAD SCORING CASE STUDY

Achuta Mukund Harsha
Preksha Binakia
Keval Dhodiya

AGENDA

Problem statement

Business Objective

Solution Methodology

Data Manipulation

Exploratory Data Analysis

Data Analysis through ML

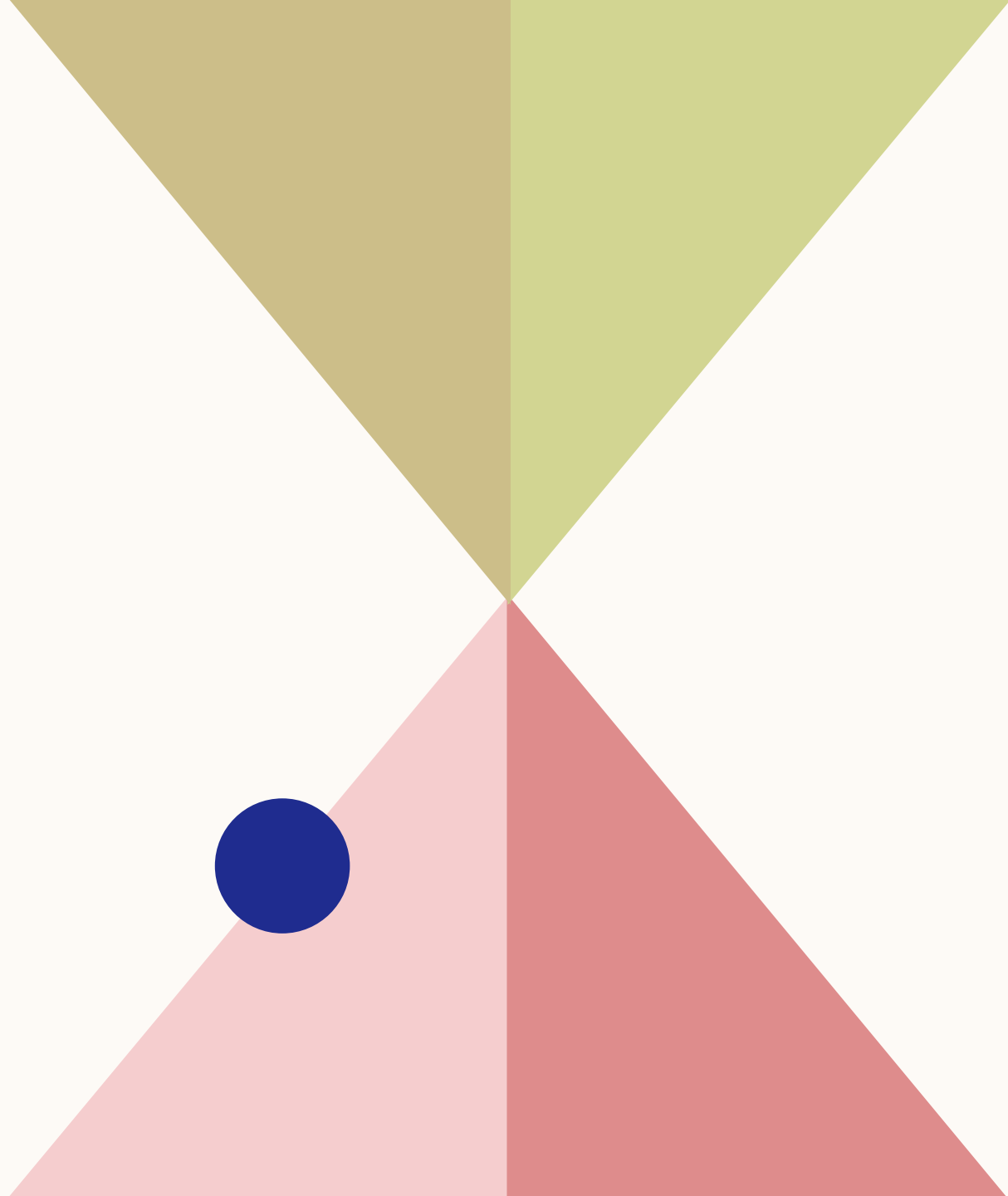
Building a correlation matrix and splitting data

Building model

Plotting the different graphs for train data set

Plotting the different graphs for test data set

Summary



PROBLEM STATEMENT

- X Education sells online courses to industry professionals.
- Although X Education receives a lot of leads, it has a very low lead conversion rate. For instance, only about 30 of 100 leads they might acquire in a day might actually be converted.
- The company seeks to locate the most promising leads, also referred to as "Hot Leads," in order to increase the efficiency of this process.
- The lead conversion rate should increase if they are successful in locating this group of leads because the sales staff will be spending more time speaking with potential leads rather than calling everyone.

BUSINESS OBJECTIVE:

- X education wants to know most promising leads.
- They aim to create a model that detects hot leads for that purpose.
- They want to set up the model for 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.

❑ Exploratory Data Analysis

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

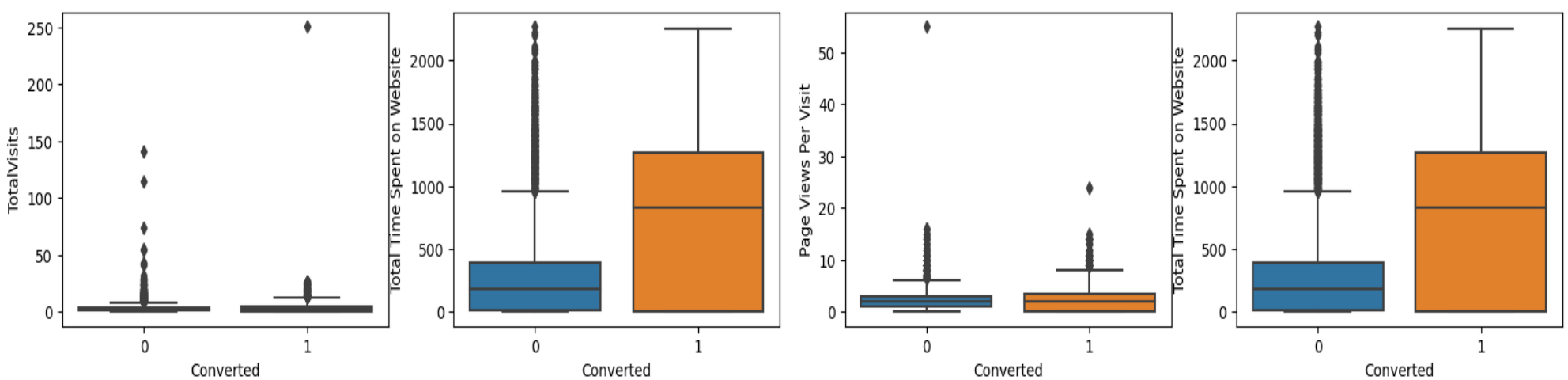
- After taking a look at the data,

  Total Number of Columns = 37, Total Number of Rows = 9240.
- Analysing the data we get to know that there are unnecessary values which surely needs to be dropped.
- Dropping the “Prospect ID” and “Lead Number” which is not necessary for the analysis.
- Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped.
- 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:

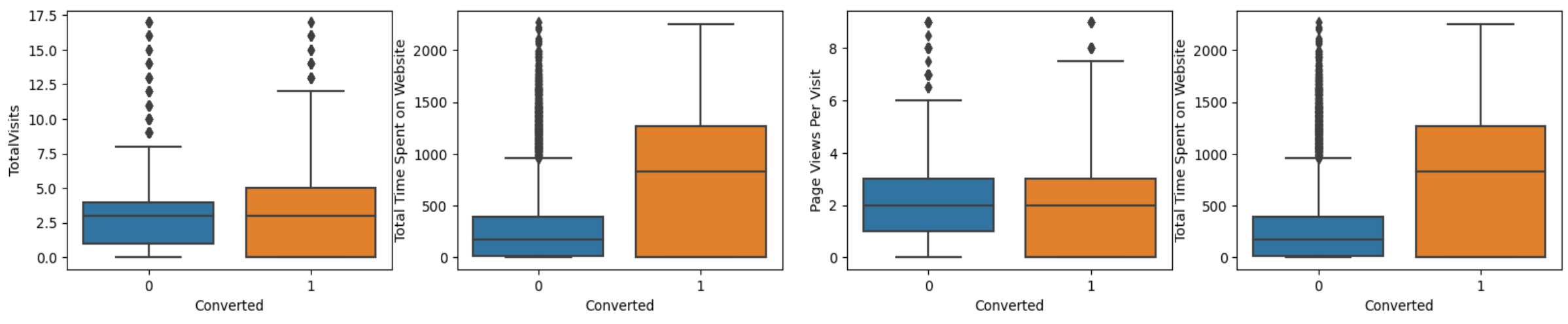
“Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement” etc.

EXPLORATORY DATA ANALYSIS

Visualizing our data using seaborn. We'll first make a pair plot of all the variables present to visualize which variables are most correlated to Leads.



- Understanding the Lead Conversion on TotalVisits, Total Time Spent on Website, Page Views Per Visit.

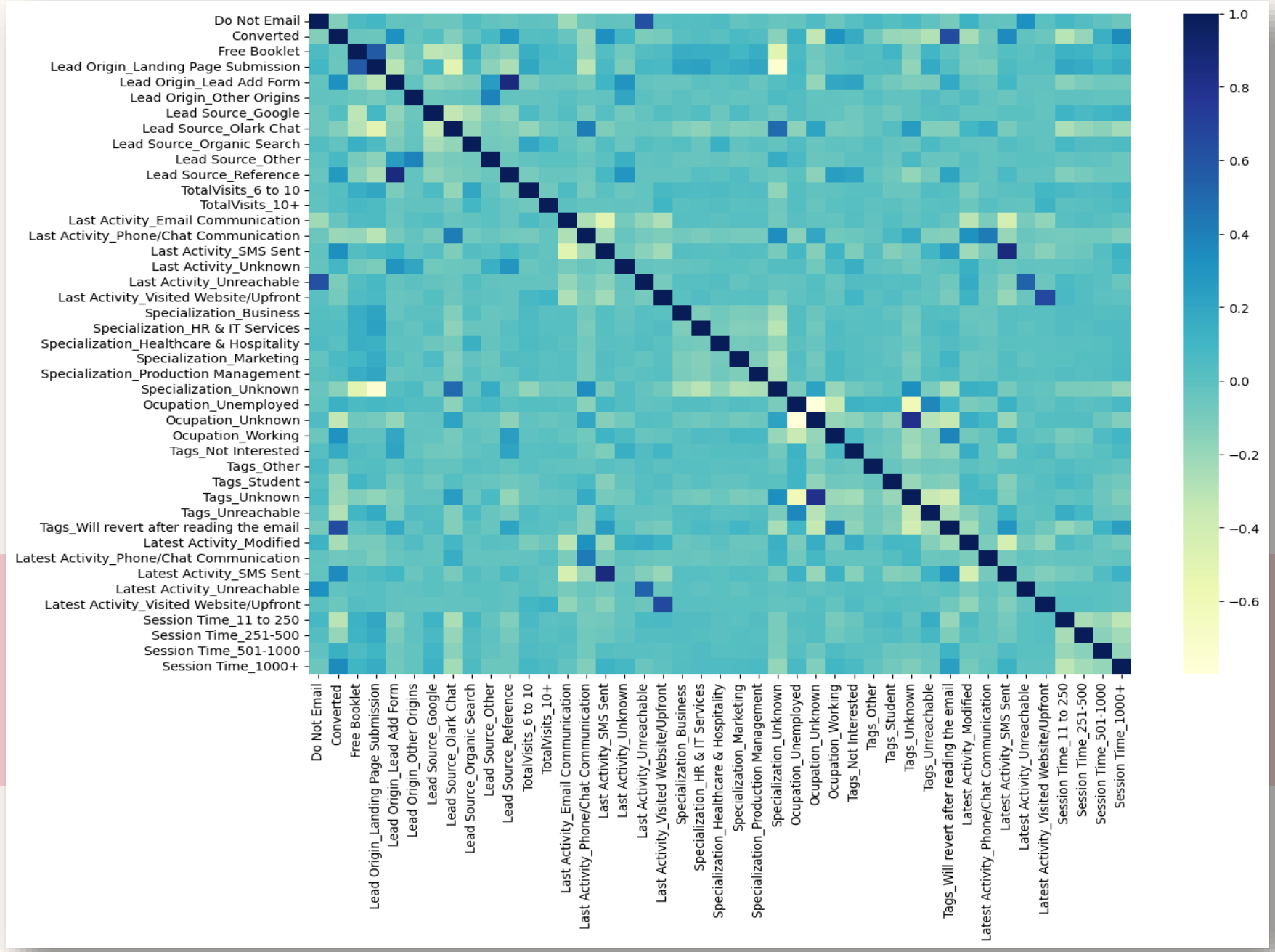


- Now we will bin the data, with the help of factor reduction.

- Factor reduction, also known as data binning or discretization, is a technique used to reduce the number of distinct values in a dataset by grouping them into bins or categories.

DATA ANALYSIS THROUGH MACHINE LEARNING

- Creating the Dummy Variables.
- Split the dataset into train and test dataset and scaled the datasets.
- Plot a heatmap to check for highly correlated features.
- The heatmap has shown on the next slide.


Do Not Email
Converted
Free Booklet
Lead Origin_Landing Page Submission
Lead Origin_Lead Add Form
Lead Origin_Other Origins
Lead Source_Google
Lead Source_Olark Chat
Lead Source_Organic Search
Lead Source_Other
Lead Source_Reference
TotalVisits_6 to 10
TotalVisits_10+
Last Activity_Email Communication
Last Activity_Phone/Chat Communication
Last Activity_SMS Sent
Last Activity_Unknown
Last Activity_Unreachable
Last Activity_Visited Website/Upfront
Specialization_Business
Specialization_HR & IT Services
Specialization_Healthcare & Hospitality
Specialization_Marketing
Specialization_Production Management
Specialization_Unknown
Ocupation_Unemployed
Ocupation_Unknown
Ocupation_Working
Tags_Not Interested
Tags_Other
Tags_Student
Tags_Unknown
Tags_Unreachable
Tags_Will revert after reading the email
Latest Activity_Modified
Latest Activity_Phone/Chat Communication
Latest Activity_SMS Sent
Latest Activity_Unreachable
Latest Activity_Visited Website/Upfront
Session Time_11 to 250
Session Time_251-500
Session Time_501-1000
Session Time_1000+
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

Building a correlation matrix and splitting data

- Creating a correlation matrix.

- A correlation matrix is a valuable tool in data analysis, providing insights into the relationships between variables and assisting in various tasks such as feature selection, model building, and understanding the underlying structure of the data.

Splitting the Data into Training and Testing Sets

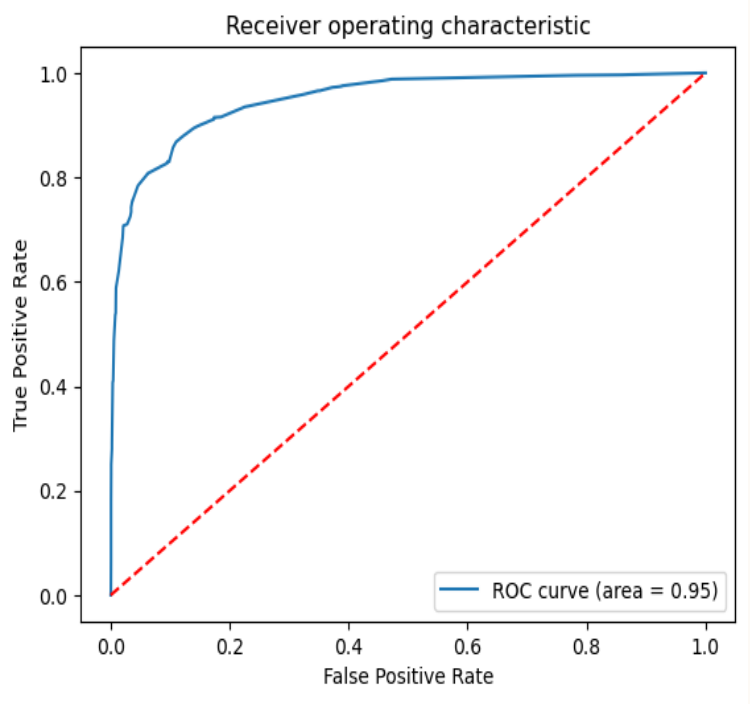
- Splitting the data for model evaluation.

- It helps ensure that your model performs well on unseen data and provides reliable insights for decision-making.

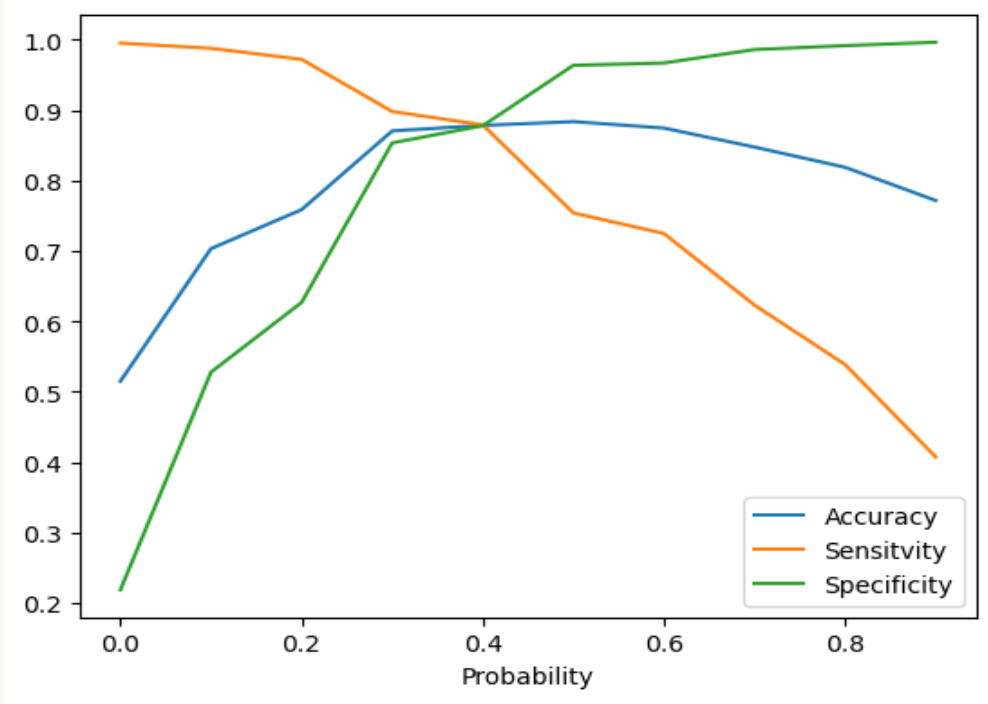
BUILDING MODEL

- Building model to check the data is sufficient and correct or not.
- It use to Prediction and decision-making, Performance improvement , Automation and Efficiency etc.
- We have made several models such as model with 20 features, model with 15 features and model with 10 features(Model 3).
- Out of which, model with 10 features is the most accurate.
- It has 95% AUC

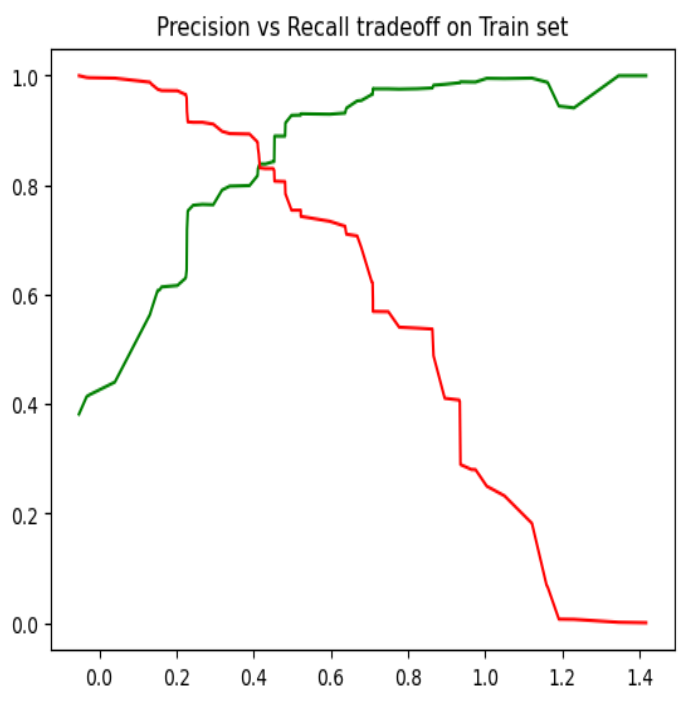
PLOTTING THE DIFFERENT GRAPHS FOR TRAIN DATA SET



ROC for Train set

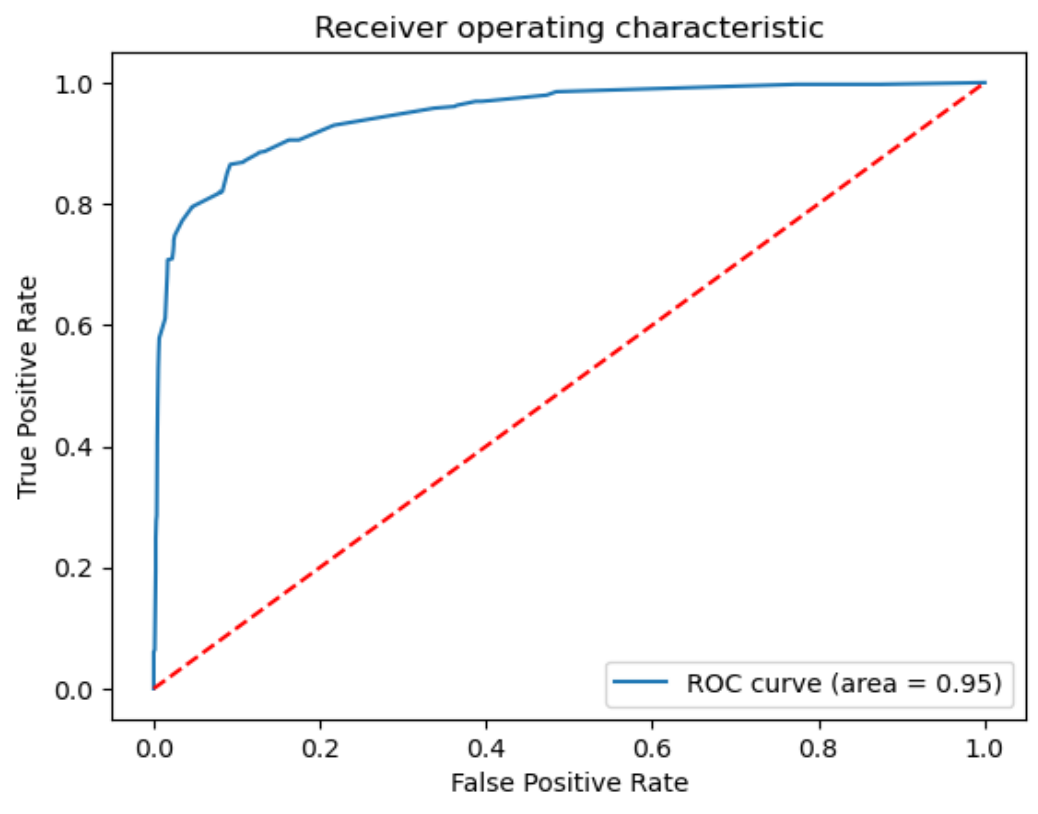


Accuracy, Sensitivity, Specificity

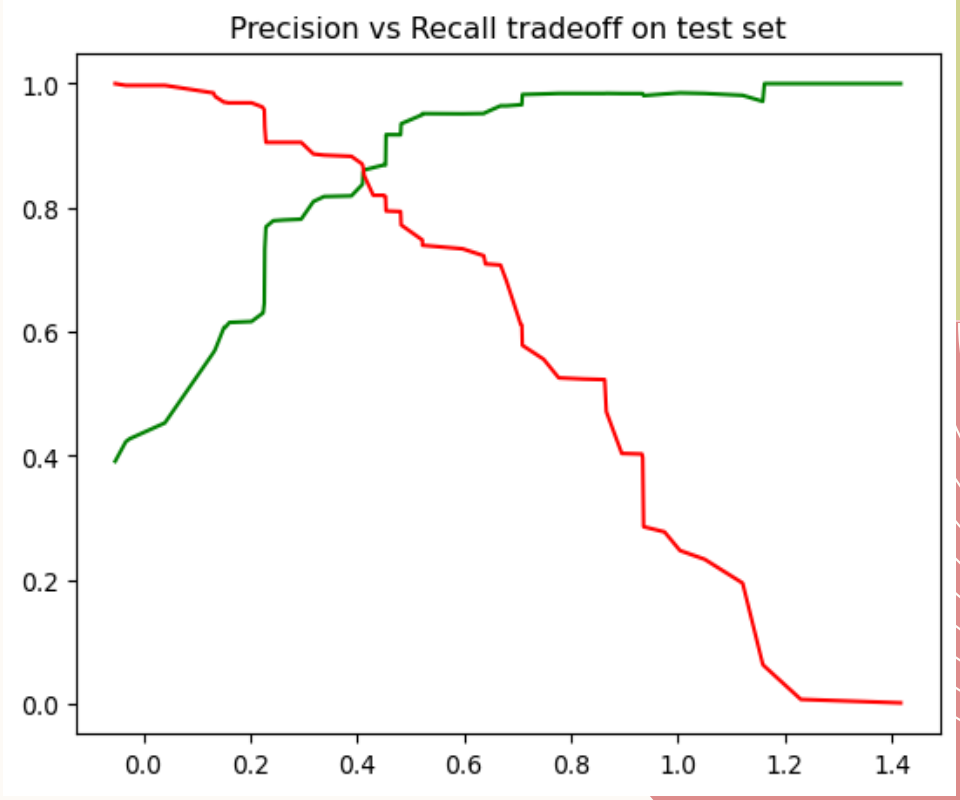


Precision Vs recall tradeoff

PLOTTING THE DIFFERENT GRAPHS FOR TEST DATA SET



ROC for Test set



Precision Vs recall tradeoff

SUMMARY

1. Top Three Important features responsible for conversion rate are :

- Tags
- Lead Origin
- Session Time (Total Time Spent on Website)

2. Top 3 categorical/dummy variables which should be focused the most are:

- Tags: Will revert after reading the email
- Lead Origin: Lead Add Form
- Session Time_1000+ (Total Time Spent on Website)

3. To make the lead conversion more aggressive focus on the leads with high Lead Score.

- Contact Leads with high Lead Score first
- Then the ones with moderately High Lead score and so on
- Avoid the Leads with Less lead score