LEADS SCORING
CASE STUDY

# PROBLEM STATEMENT

► X Educationsells 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 knownas 'Hot Leads'.

► If they successfully identify this set of leads, the lead conversion rate should go upas 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.
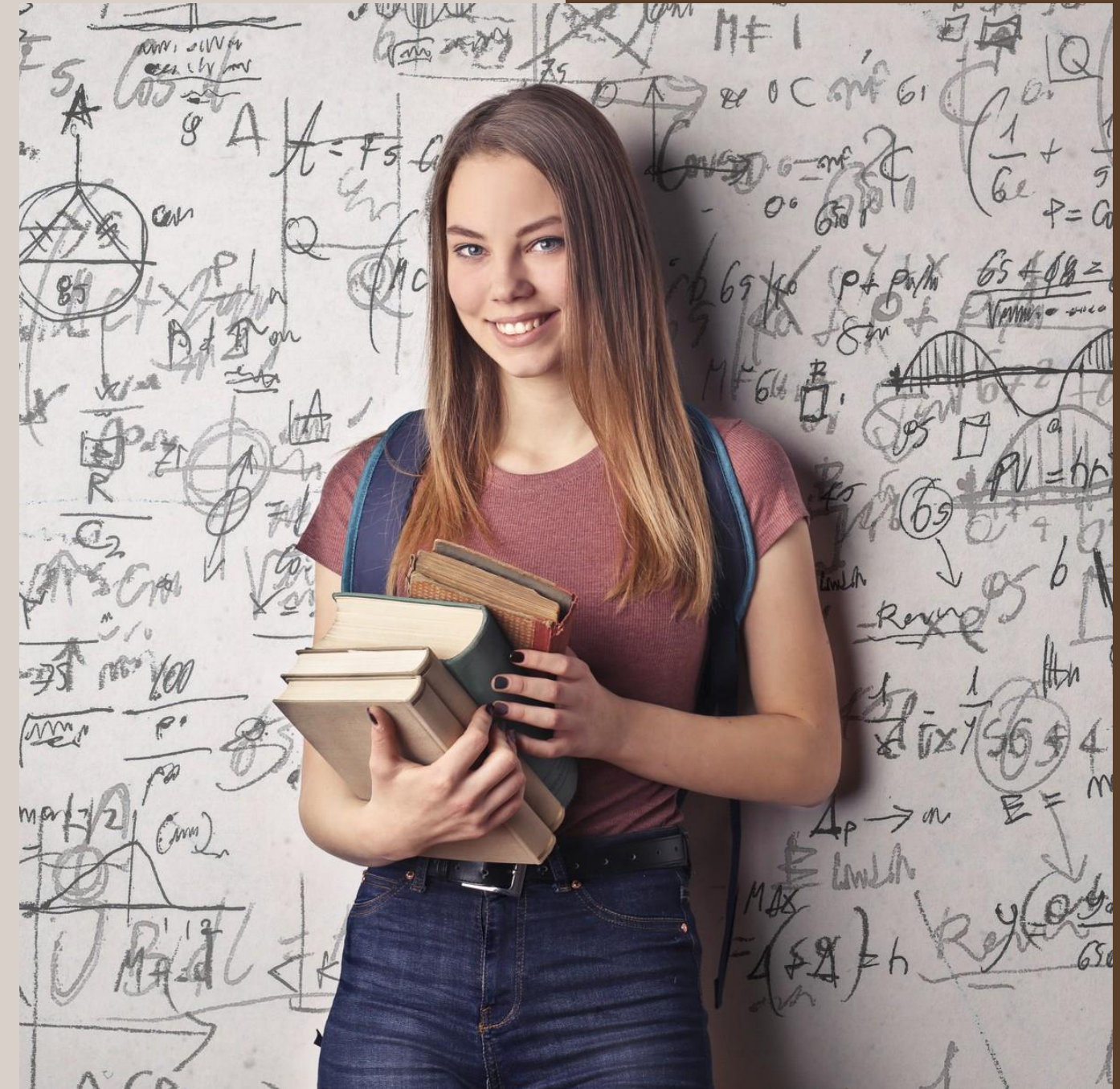
SOLUTION METHODOLOGY

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 a large number of missing values and are not useful for the analysis.

4. Imputation of the values, if necessary.

5. Check and handle outliers in data.

Exploratory Data Analysis (EDA)

1. Univariate data analysis: value count, distribution of variables, etc.

2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.

3. Feature Scaling & Dummy variables and encoding of the data.

4. Classification technique: logistic regression is used for model making and prediction.

5. Validation of the model.

6. Model presentation.

7. Conclusions and recommendations.

DATA MANIPULATION

- ► TOTAL NUMBER OF ROWS=37,TOTAL NUMBER OF COLUMNS =9240.

- ► SINGLE VALUE FEATURES LIKE"MAGAZINE", "RECEIVEMOREUPDATES ABOUT OUR COURSES", "UPDATE MY SUPPLY"

- ► CHAIN CONTENT", "GET UPDATES ON DM CONTENT", "I AGREE TO PAY THE AMOUNT THROUGH CHEQUE" ETC. HAVE BEEN DROPPED.

- ► REMOVING THE "PROSPECTID" AND "LEAD NUMBER" WHICH ARE 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 HAVE ENOUGH VARIANCE, WHICH HAVE DROPPED, THE FEATURES AS "DO NOT CALL", "WHAT MATTERS MOST TO YOU IN CHOOSING COURSE", "SEARCH", "NEWSPAPER, ARTICLE", "XEDUCATION FORUMS", "NEWSPAPER", "DIGITALADVERTISEMENT" ETC.

- ► DROPPING THE COLUMN SHAVING MORE THAN 35% AS MISSING VALUES SUCH AS 'HOW DID YOU HEAR ABOUT X EDUCATION' AND 'ASYMMETRIQUE PROFILE INDEX'.
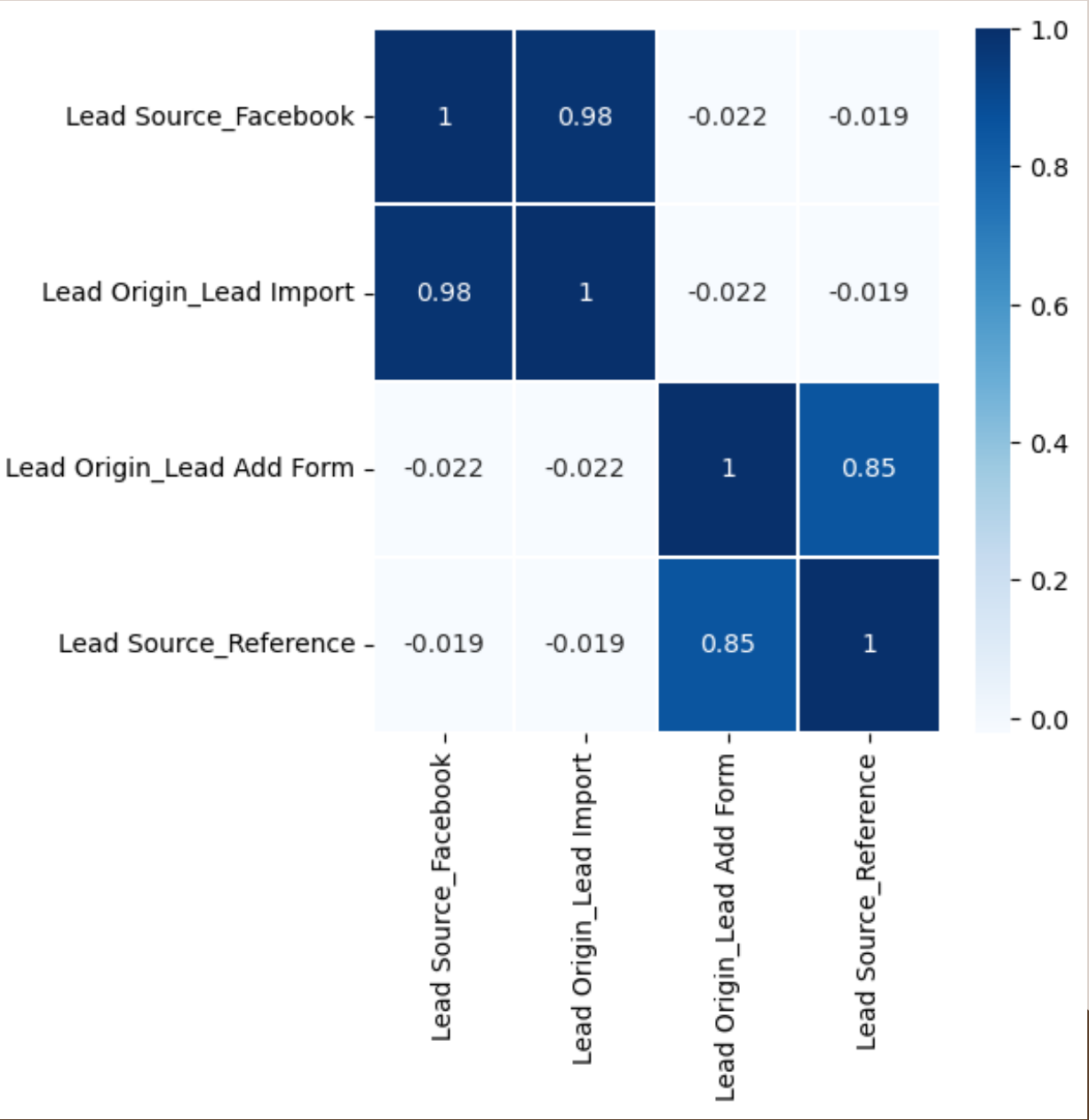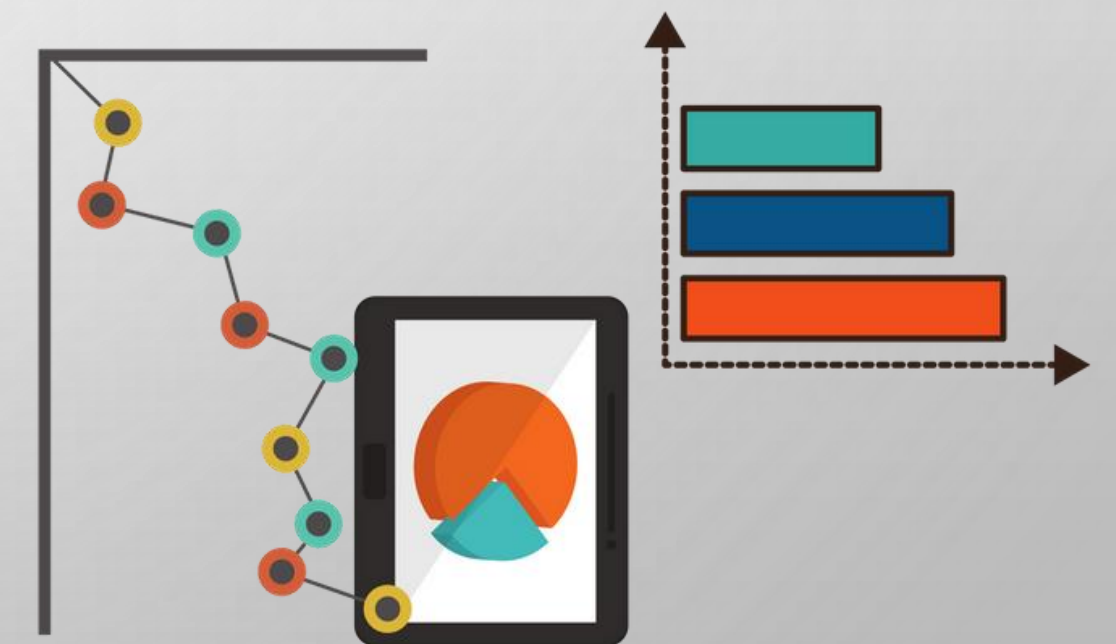
# EXPLORATORY DATA ANALYSIS (EDA)
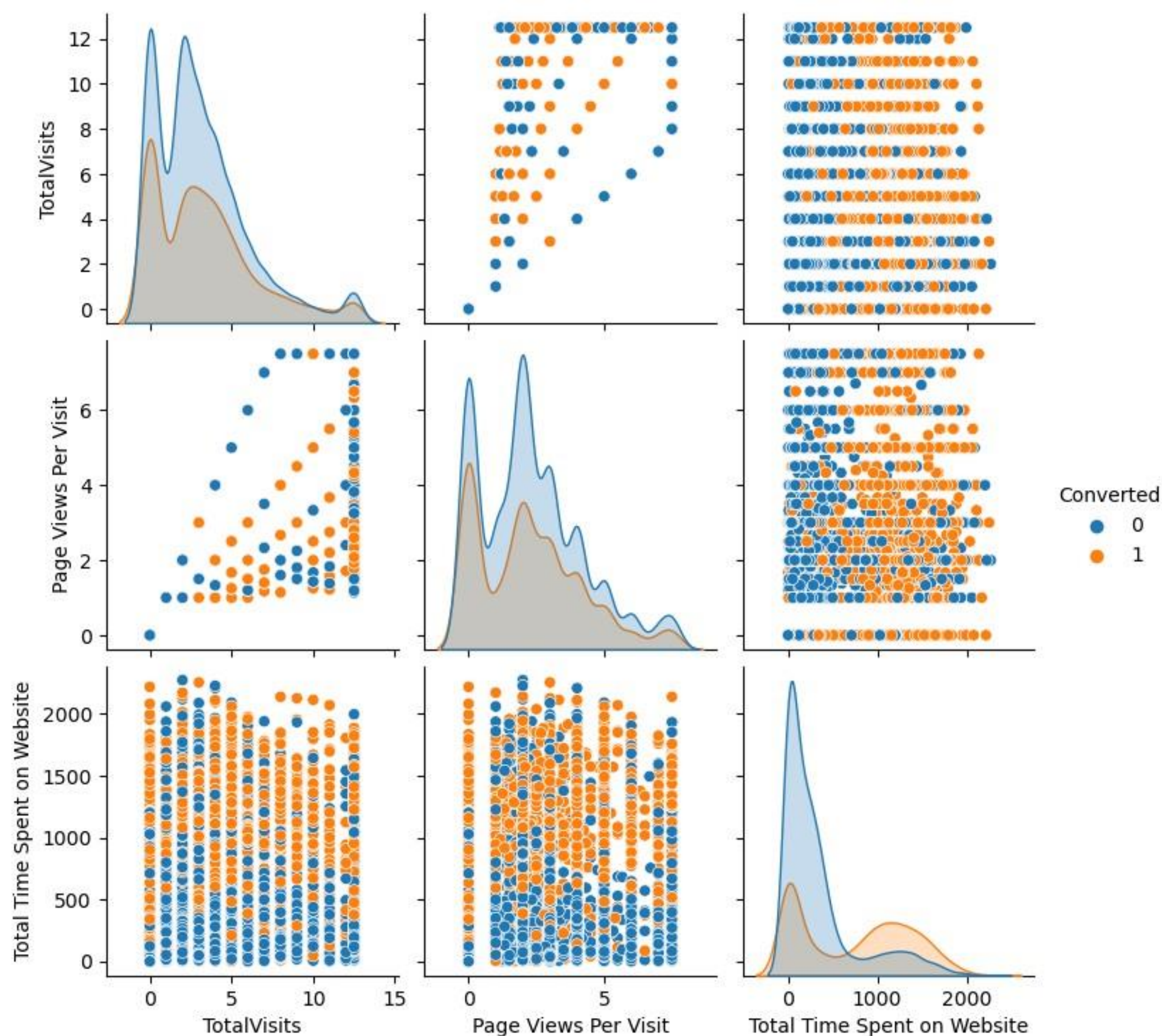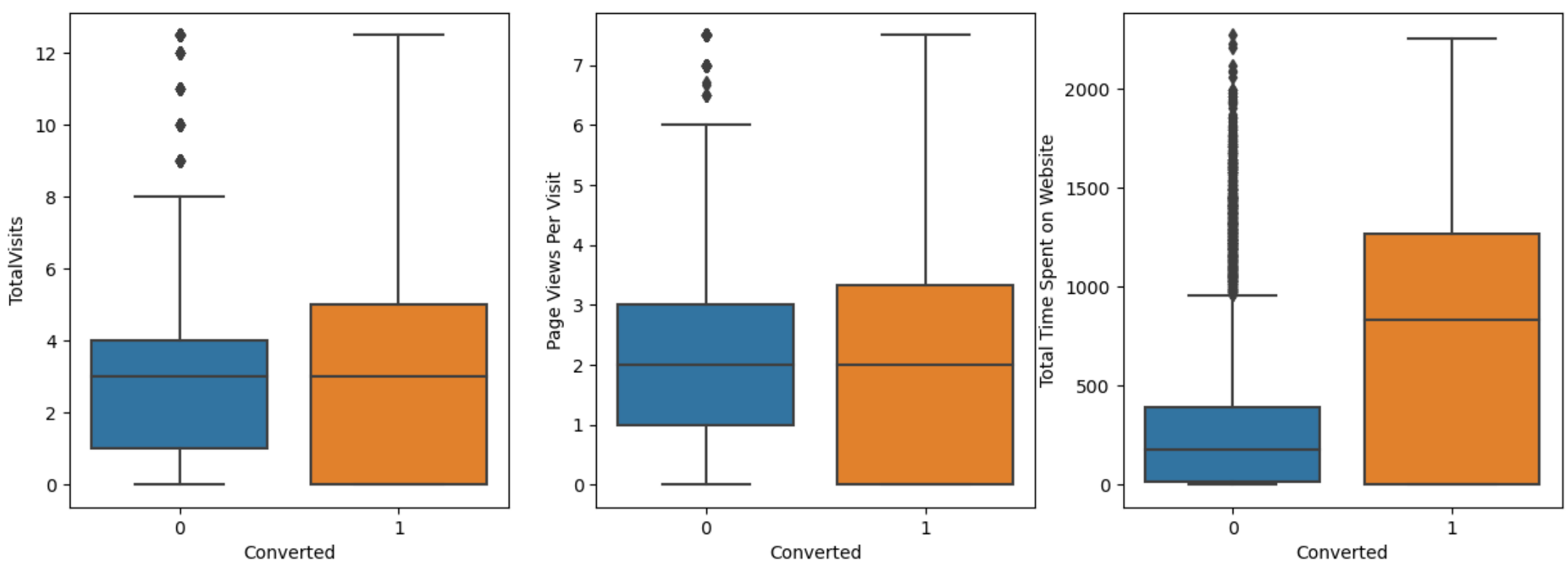
BOX PLOT

HEAT MAP

DATA CONVERSION

► **Numerical Variables are normalized**

► **Dummy Variables are created for object type variables**

► **Total Rows forAnalysis: 9240**

► **Total Columns for Analysis: 37**

Data Conversion

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 89%**

ROC CURVE
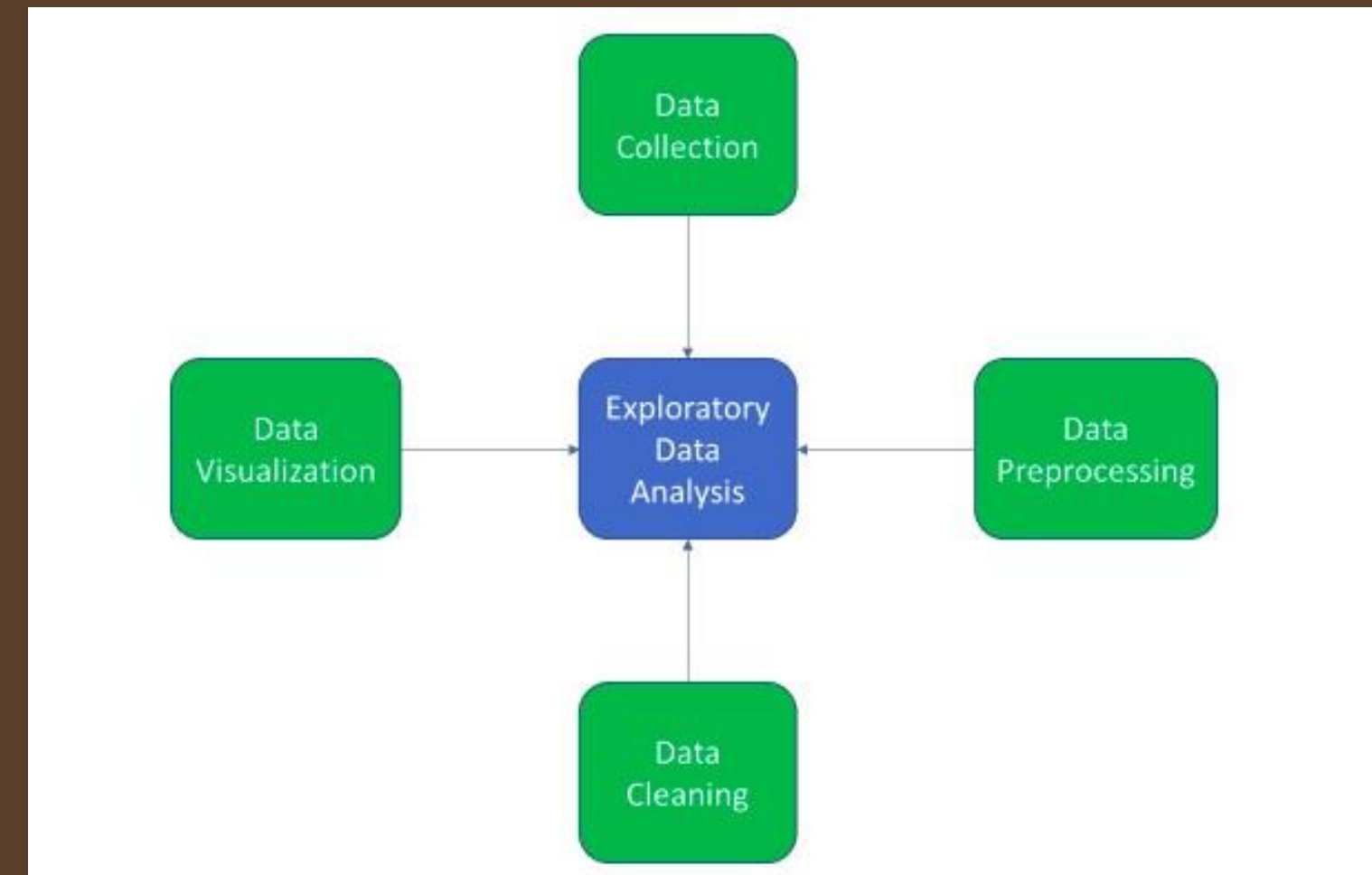
Receiver operating characteristic example
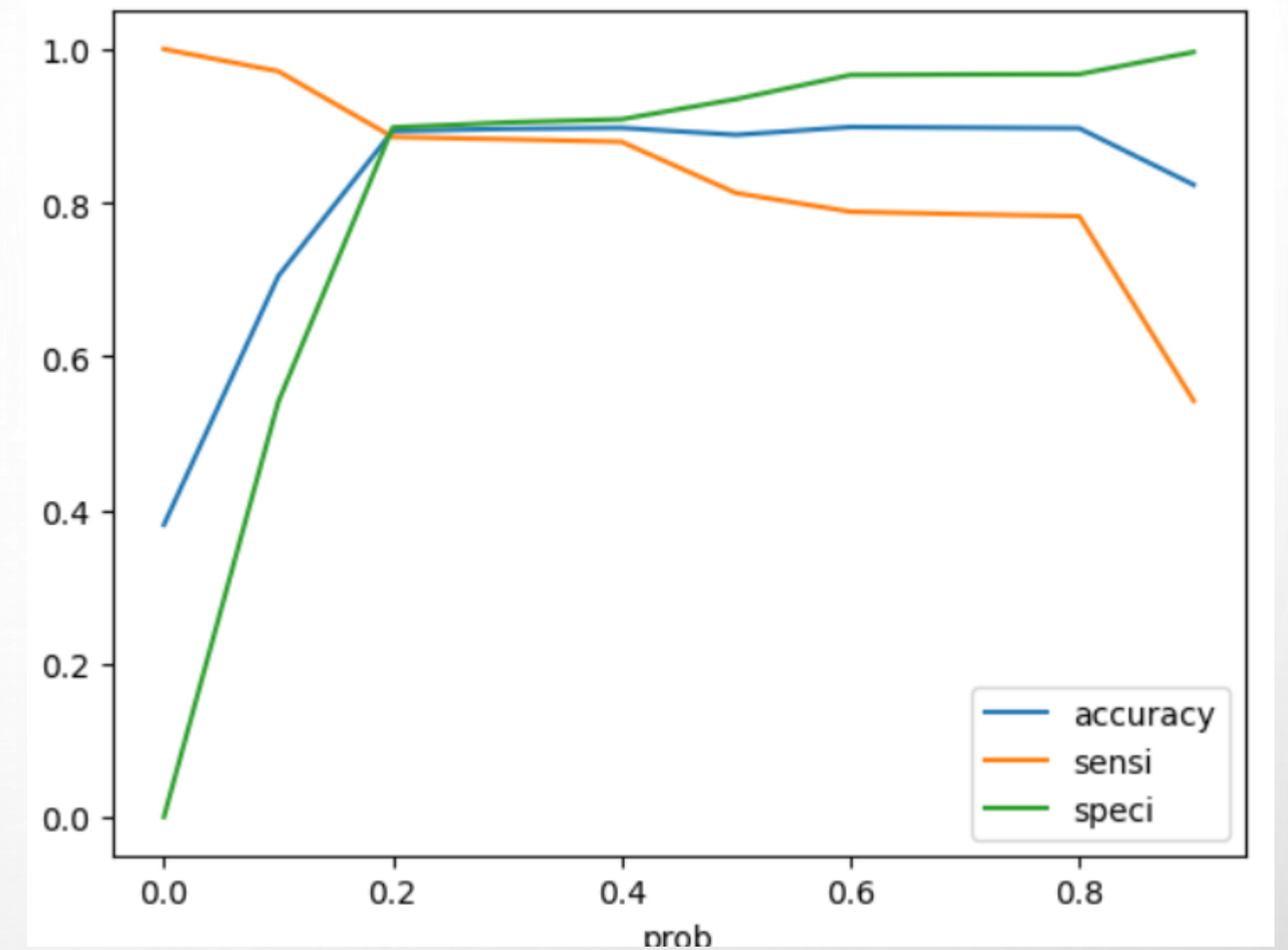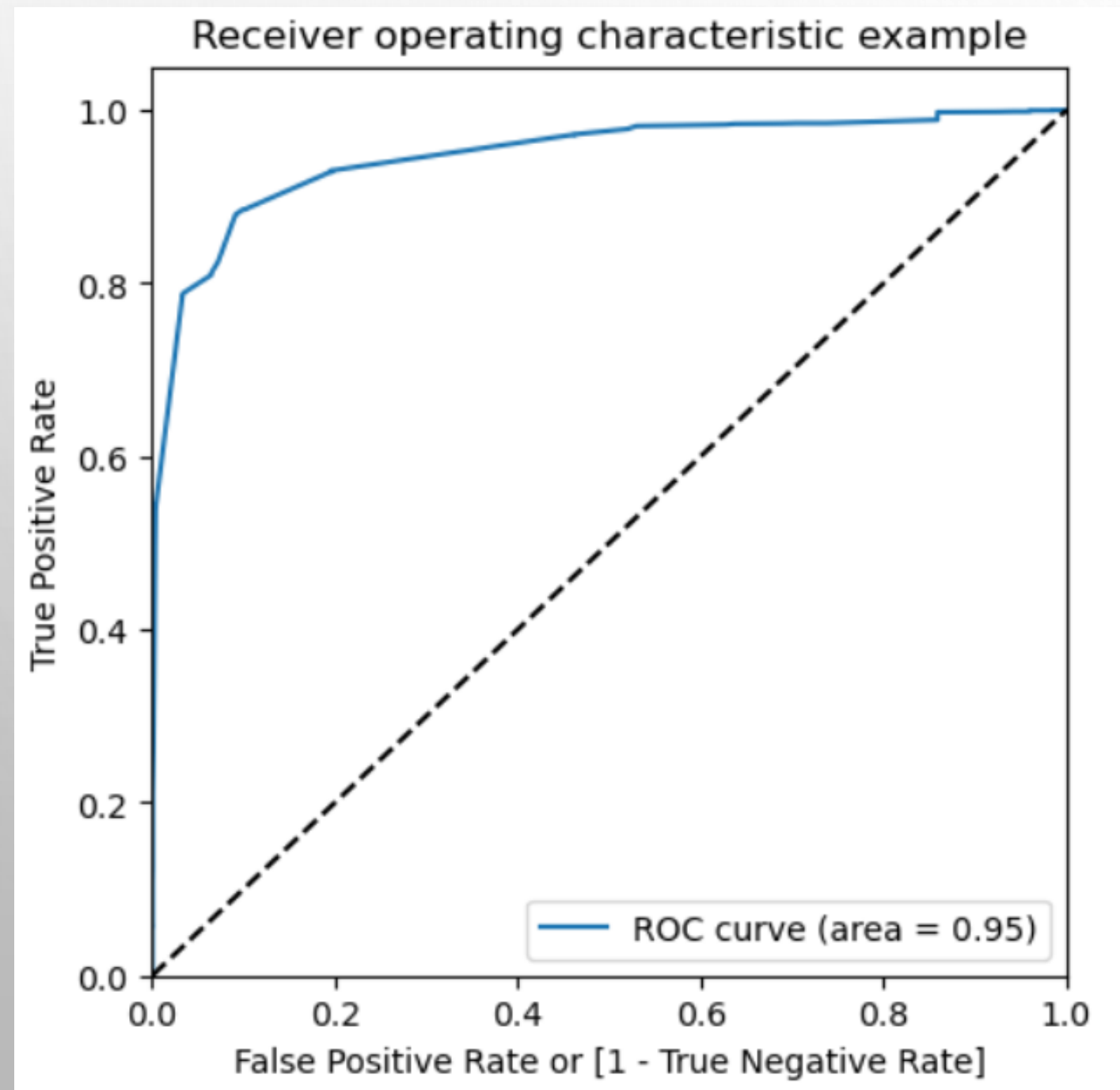
► Finding Optimal Cut off Point

► 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 0.2.

PREDICTION ON TEST SET

► Before predicting on the test set, we need to standardize the test set and need to have exact same columns present in our final train dataset.

► After doing the above step, we started predicting the test set, and the new prediction values were saved in a new data frame.

► After this we did model evaluation i.e. finding the accuracy, precision, and recall.

► The accuracy score we found was 0.89, precision 0.83, and recall 0.88 approximately.

► This shows that our test prediction is having accuracy, precision, and recall scores in an acceptable range.

► This also shows that our model is stable with good accuracy and recall/sensitivity.

► 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 & RECOMMENDATIONS

Prioritize Leads Closed by Horizzon
Leads previously closed by Horizzon show the highest conversion potential. Focus on these leads by replicating successful strategies, targeting them in campaigns, and assigning them to experienced sales reps.

Emphasize High-Relevance Lead Quality
Leads marked as "High in Relevance" are more likely to convert. Prioritize these in your qualification process, refine the lead scoring system accordingly, and allocate more resources to nurturing these leads.
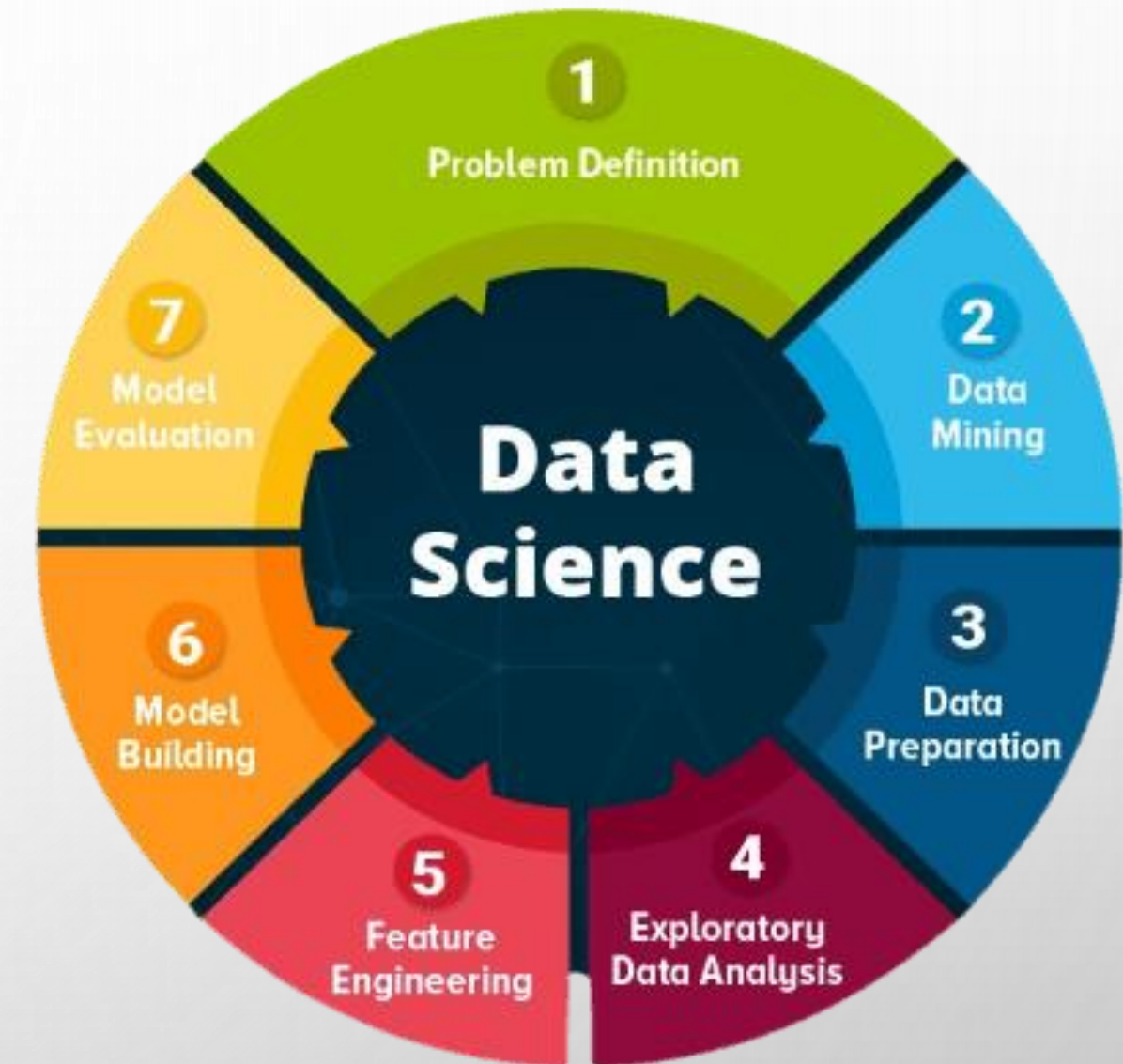
Optimize Lead Add Form Origin
Leads from the 'Lead Add Form' are strong conversion predictors. Enhance and promote this form, ensuring it's user-friendly and effective. Consider A/B testing and UI improvements to further boost conversion rates.

Revise Approach for Negative Coefficient Features
Features like Tags_Ringing, Interested in other courses, and Email Bounced negatively impact conversion. Adjust strategies for these leads, such as using alternative communication channels or excluding them from certain campaigns, to improve overall conversion rates.

Keeping these in mind X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.

THANK YOU