X EDUCATION
LEAD SCORE CASE STUDY

Shenal Kothalawala

CASE STUDY BACKGROUND

- X Education is a company which provides online courses to various professionals in different industries.

- Those who are interested in pursuing studies will visit their website.

- The company markets its courses on several websites and search engines like Google.

- Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos.

- When these people fill up a form providing their email address or phone number, they are classified to be a lead.

- When these people fill up a form providing their email address or phone number, they are classified to be a lead.

- The typical lead conversion rate at X education is around 30%.

PROBLEM STATEMENT

- X Education gets a lot of leads but, its lead conversion rate is very poor.

- To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads.

- Once hot leads are identified, the marketing team can focus on them to convert.

- The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

PROPOSED SOLUTION

- A logistics regression model was used to handle this problem.

- There were 9240 data rows with 37 columns.

- Data were checked whether the types of data were in order or any amendment to be made. However, all data types were in correct format.

- Data cleaning was initiated to see what actions to be taken against missing values. All missing values more than 40% were removed for further analysis. Other missing values were imputed based on mode or created a new field as others depend on the case.

- For columns having multiple levels, the levels were reduced by clubbing together.
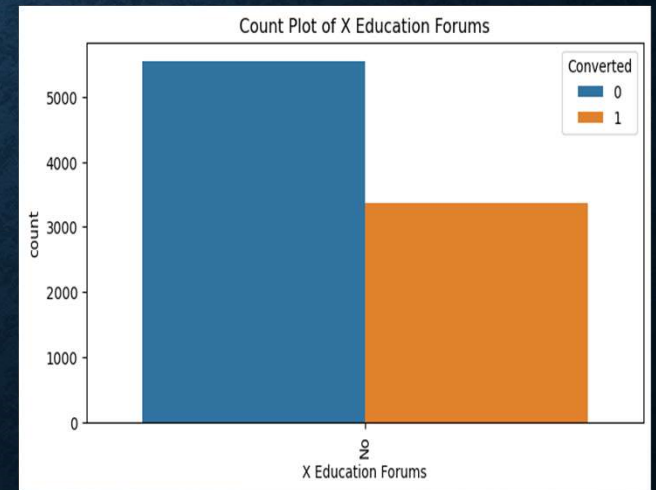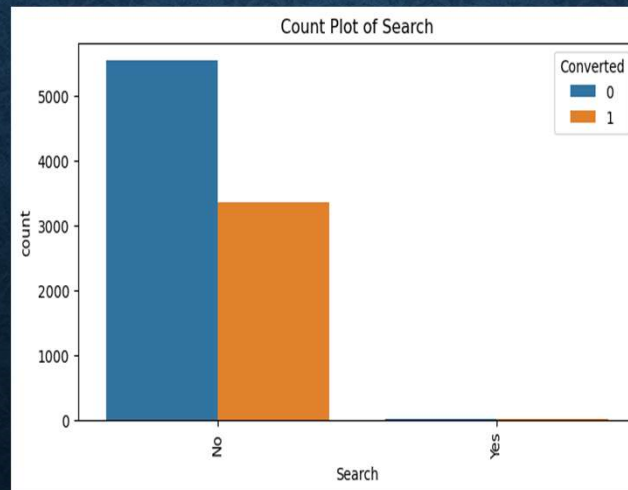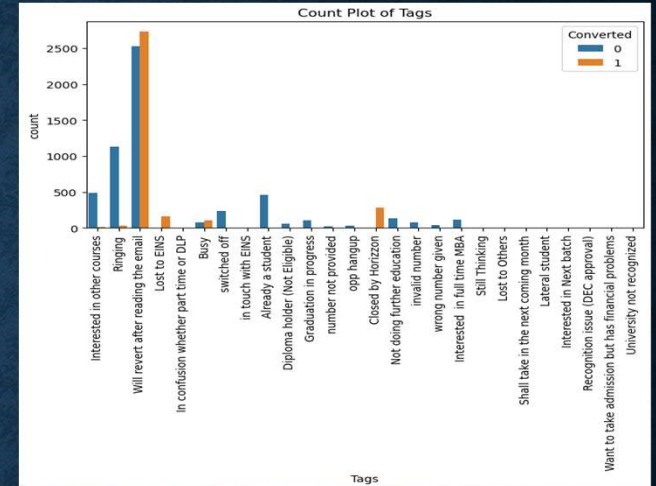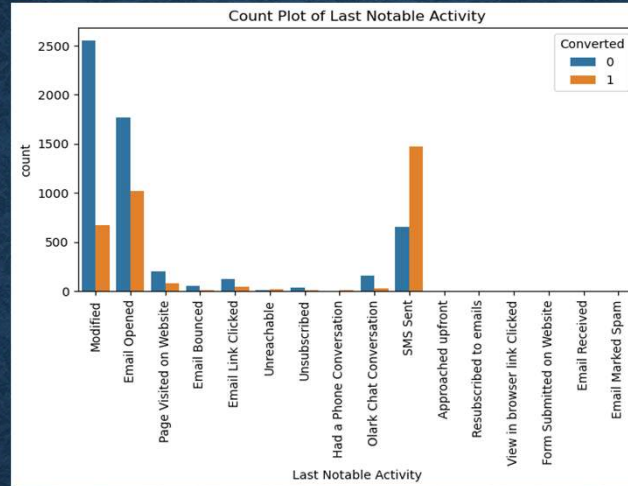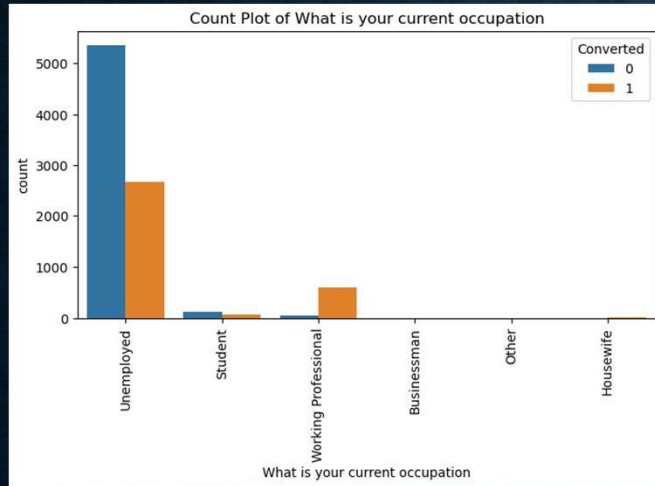
EDA – NUMERICAL DATA
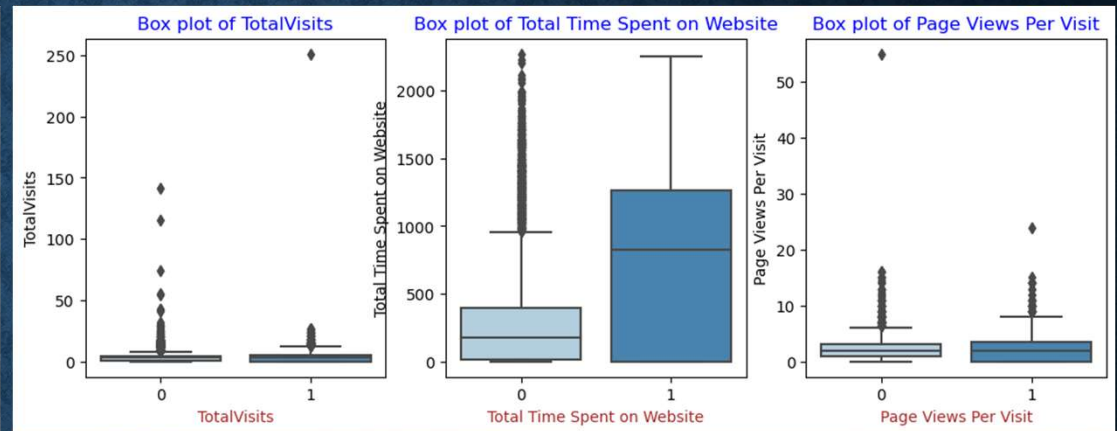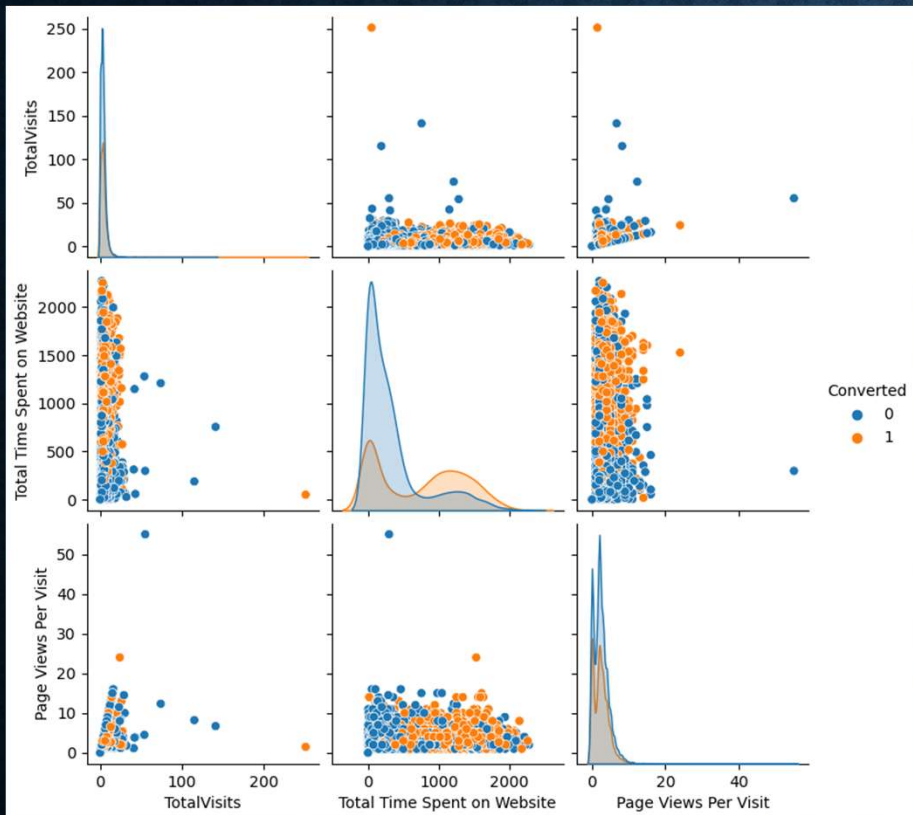
- Numerical data were represented along with converted variable to see the distribution.

- Also drawn boxplots to see any outliers were present to handle the same.

- There were some outliers which were removed.

EDA – CATEGORICAL DATA
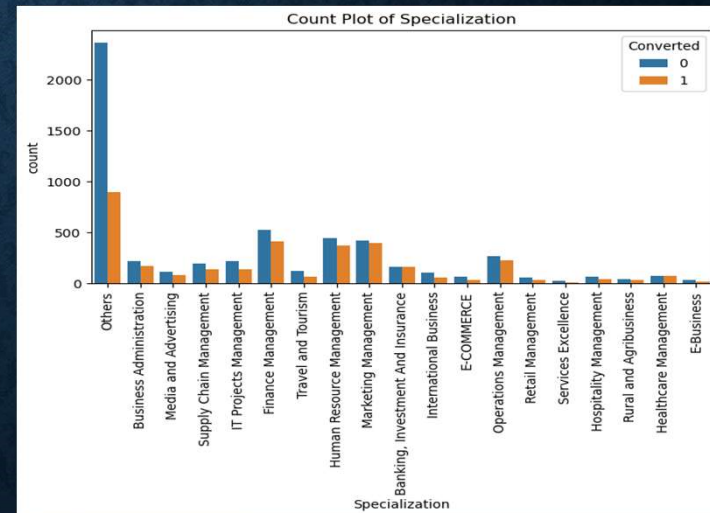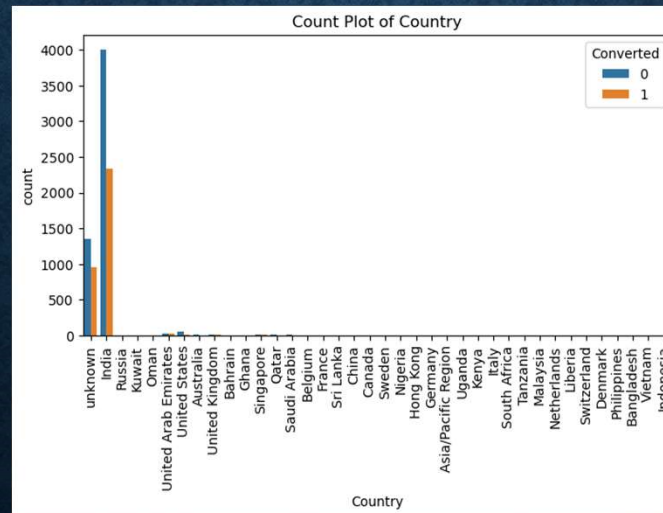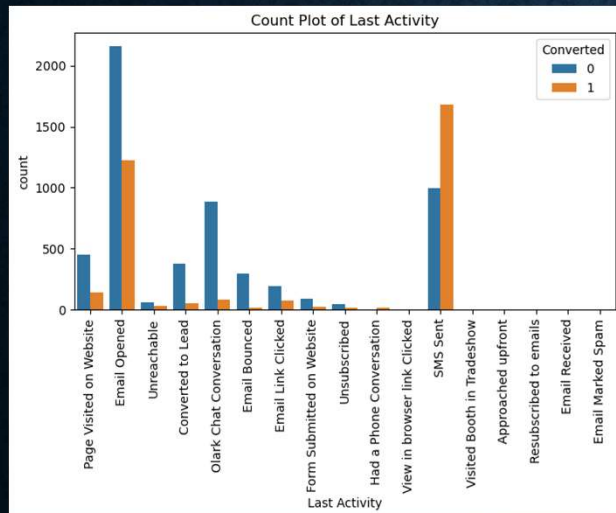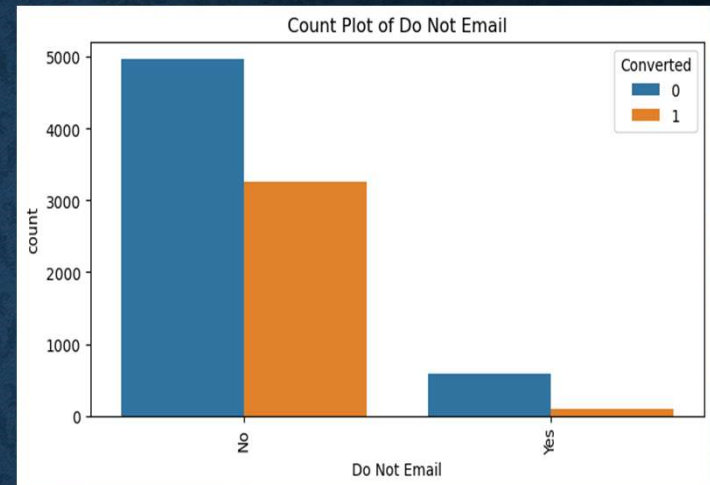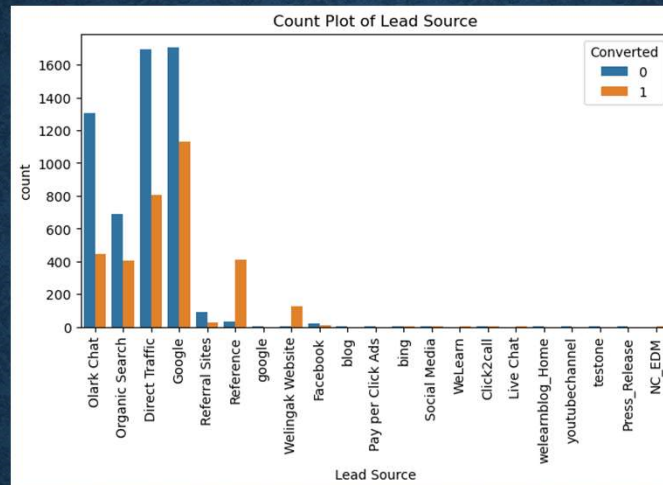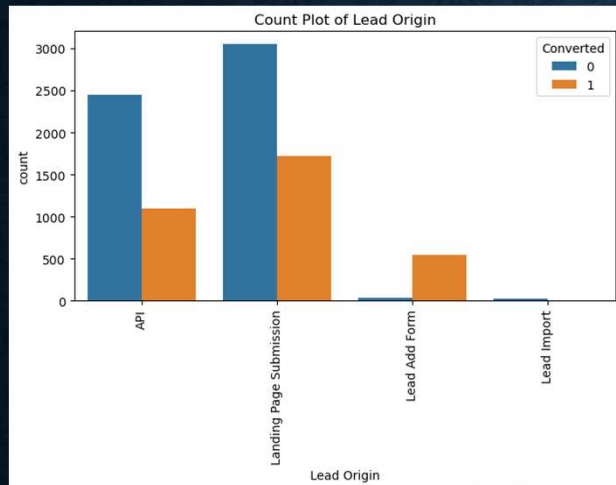
EDA – CATEGORICAL DATA

MODELING WITH LOGISTIC REGRESSION

- After the adjustments with respect to missing values, outlier treatments etc. the final data set was 8924 rows with 66 columns.

- The 66 columns include the dummy variables created for categorical variables.

- Numerical data were standardized using minmax scaler.

- A RFE model was used to select the best 15 predictors.

- Once the 15 predictors were taken, then using statsmodels based on the P-value and VIF values several models were built removing insignificant predictors.

- The final model had 11 predictors and this was then evaluated based on accuracy, sensitivity and specificity.
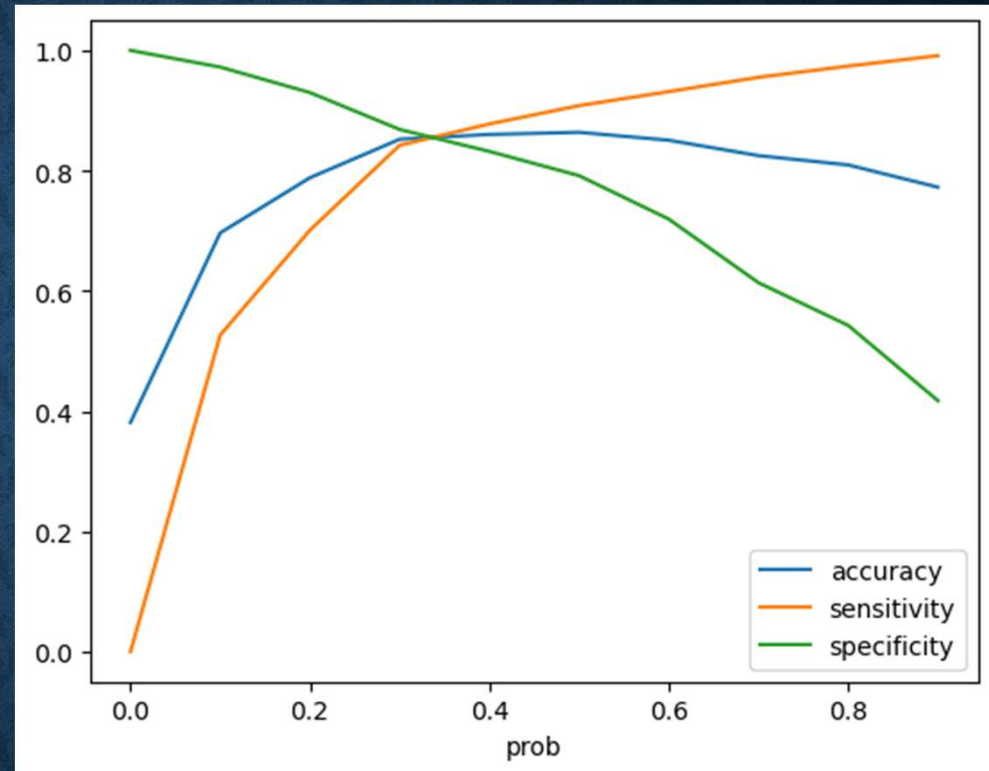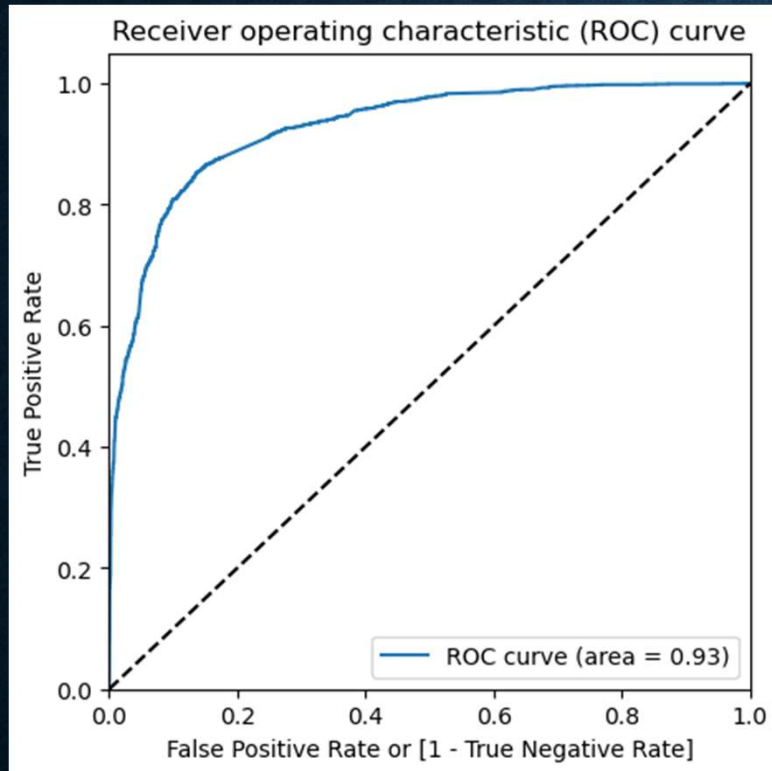
- The ROC curve shows AUC is 0.93 which is a pretty decent figure and we can say the model is good. The cut-off values for predictions were done with 0.5. To check the optimal value cross interception was selected based on the line graphs of accuracy, sensitivity and specificity. This value came to 0.34.

MODEL EVALUATION

Train data:

Accuracy - 0.86 Sensitivity: 0.79 Specificity: 0.91

Test data:

Accuracy - 0.86 Sensitivity: 0.84 Specificity: 0.87

- The test data were transformed using the learnings of training data.

- The test data seems to have solid results. Thus, the model is really good.

RECOMMENDATIONS

- The Company should use model 5 as the basis for any future predictions

- X education should focus on leads who spent a lot of time on website

- X education should continuosly follow-up with leads marked as Will revert after reading the email as they are potential leads for conversion.

- X education should also consider leads originated from Lead Add Form

- Working professionals seems to have a higher interest hence high chance of converting them.

- X Education should make fields mandatory to capture more data about leads. For eg- Country unknown is a potential lead and this is a level where we imputed with unknown as these fields were filled with 'Select' data. If this (and all others) made mandatory then we can draw meaningful conclusions.

- X Education should not focus on leads originated from Landing page submission, Specialization others and last activity Olak Chat convesation as these are highly unlikely to get converted