

Problem Statement

♦ An X Education need help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires us to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with 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%.



Goals and Objectives

♦ Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

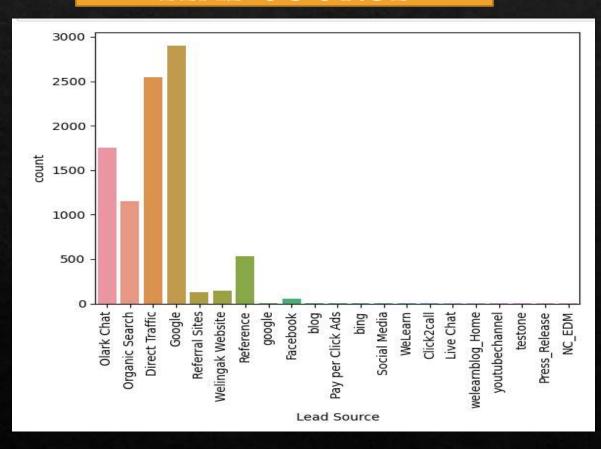


Steps performed for the analysis:

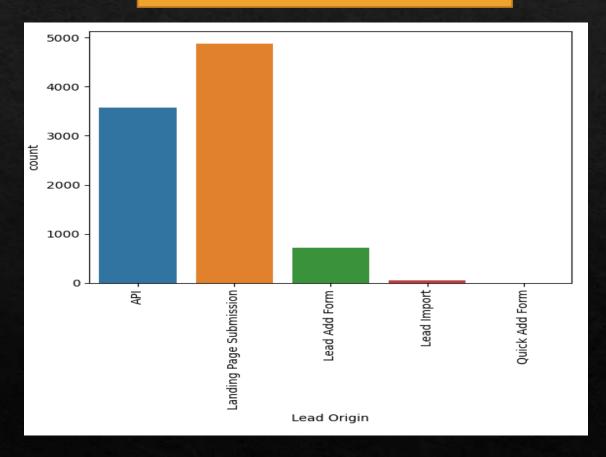
- Importing the required libraries and the dataset
- Inspecting the Data Frame
- Checking for missing values
- Analyzing columns individually and handling missing values
- Univariate Analysis and outlier treatment
- Dropping redundant columns
- Converting some binary variables (Yes/No) to 0/1
- Grouping column features
- Dummy creation
- Train Test Split
- Feature Scaling
- Model Building
- Feature Selection Using RFE
- Plotting the ROC Curve
- Finding Optimal Cut-Off Points
- Precision and Recall
- Making Predictions on the Test Set

EDA

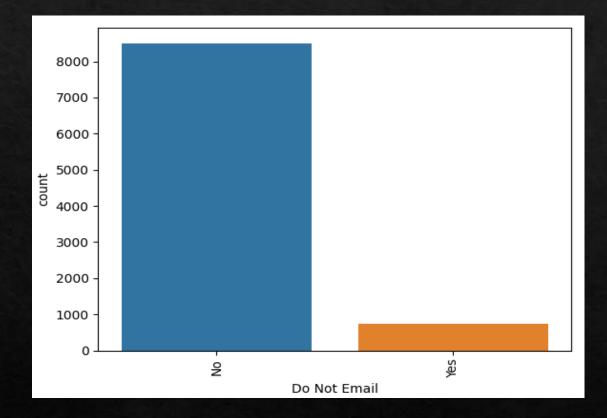
LEAD SOURCE



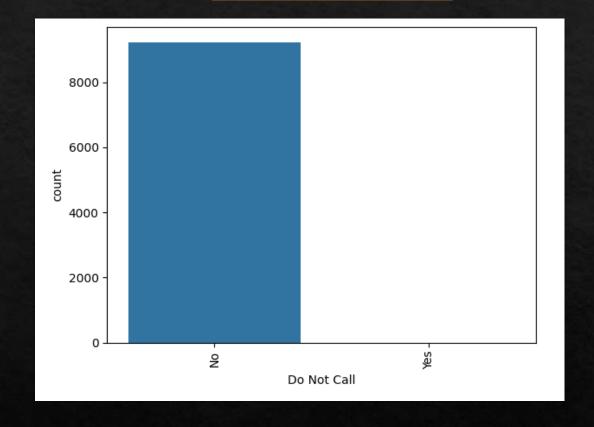
LEAD ORIGIN



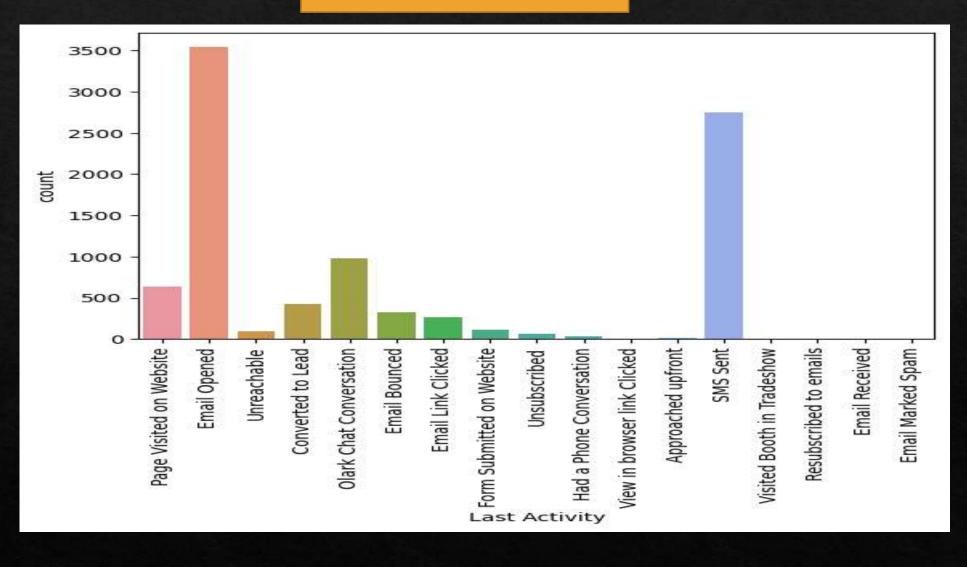
DO NOT EMAIL



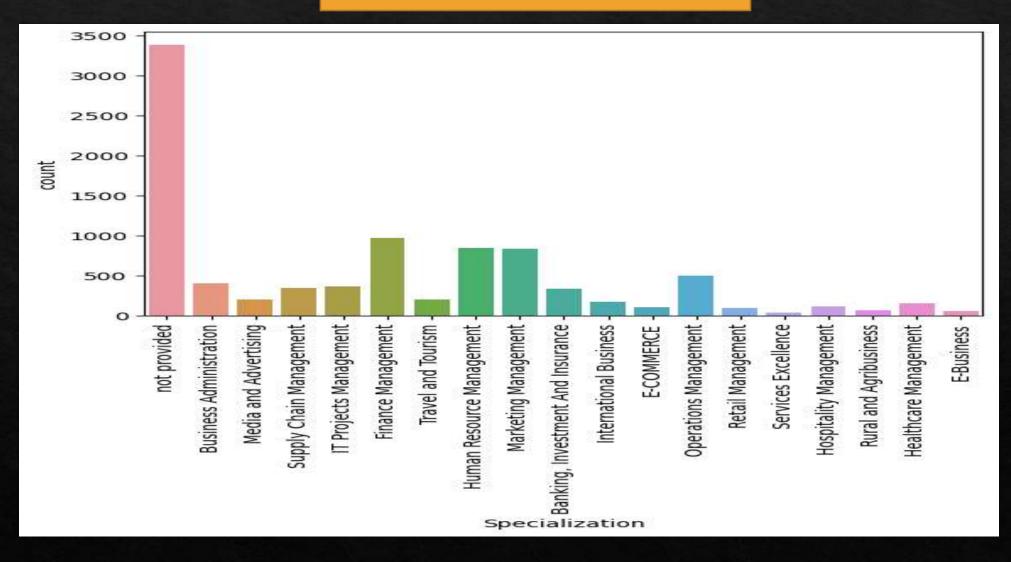
DO NOT CALL

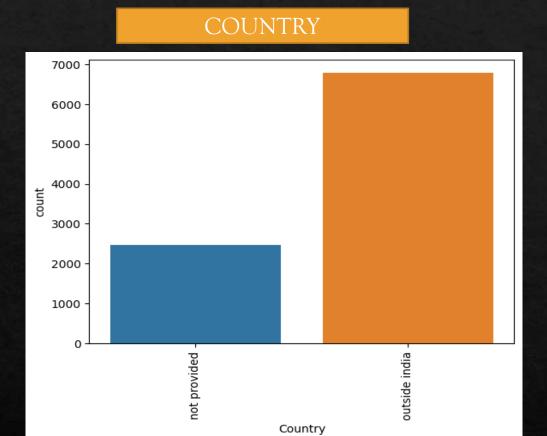


LAST ACTIVITY

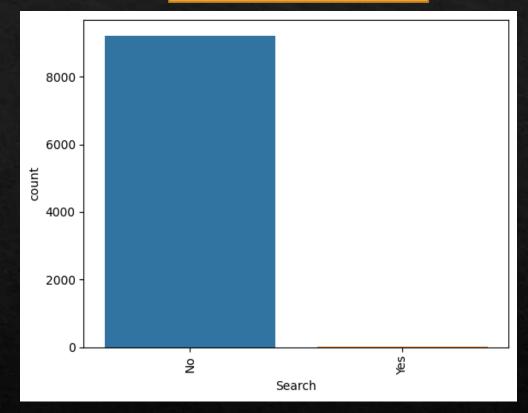


SPECIALIZATION

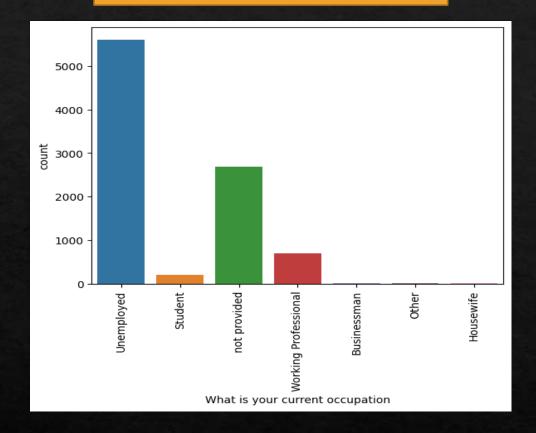




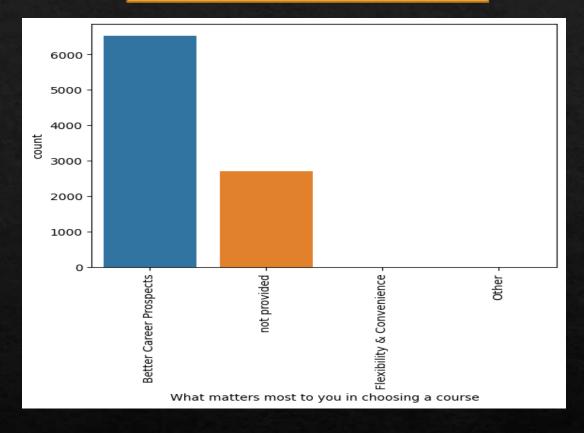
SEARCH



CURRENT OCCUPATION



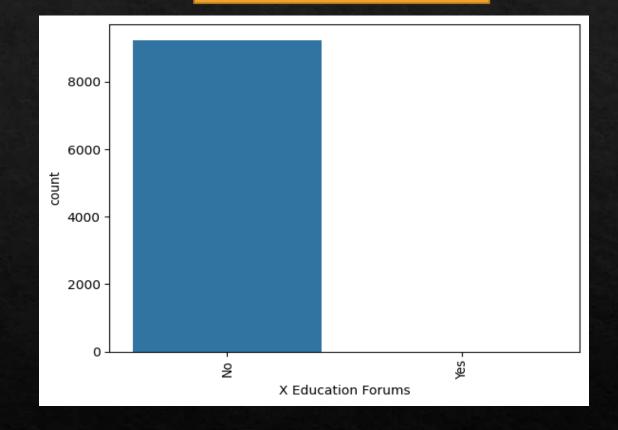
WHAT MATTERS THE MOST IN CHOOSING A COURSE?



NEWSPAPER ARTICLE

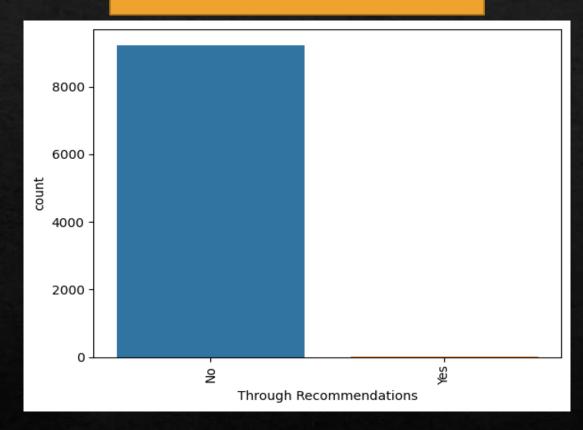
8000 -6000 count 4000 2000 -Jes Newspaper Article

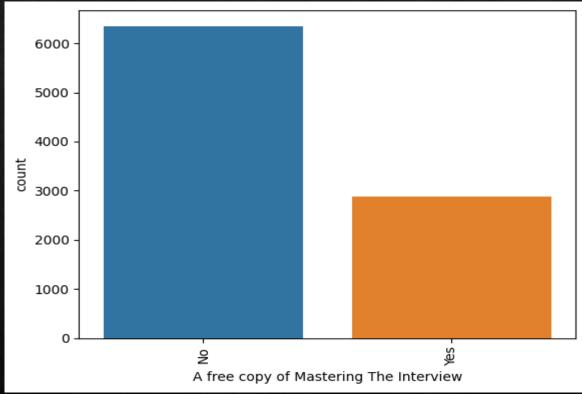
X EDUCATION FORUMS



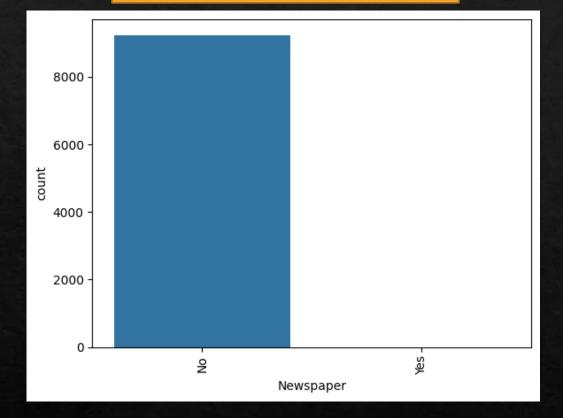
THROUGH RECOMMENDATIONS

A FREE COPY OF MASTERING THE INTERVIEW

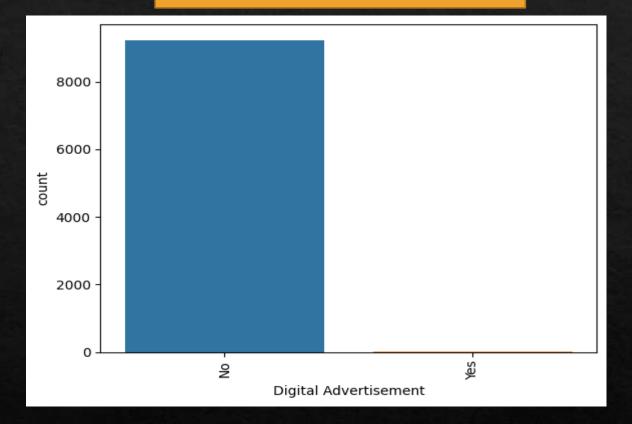




NEWSPAPER



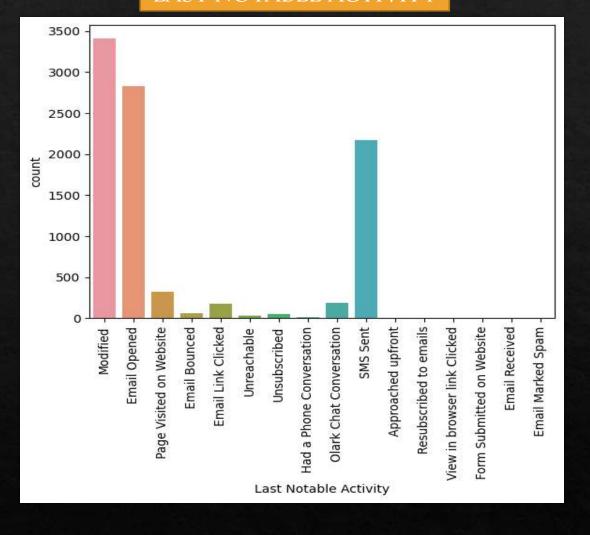
DIGITAL ADVERTISEMENT

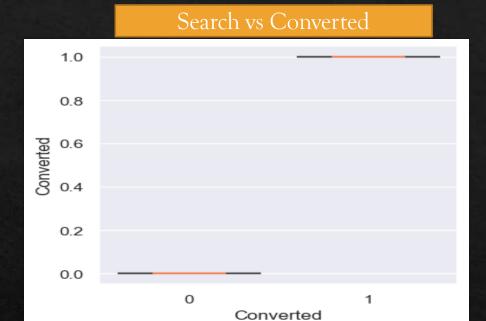


TAGS

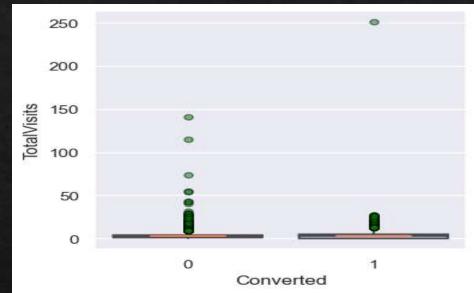
3500 3000 2500 2000 count 1500 1000 500 Ringing not provided Other Tags Already a student Closed by Horizzon Interested in Next batch Interested in other courses Will revert after reading the email Still Thinking Tags

LAST NOTABLE ACTIVITY

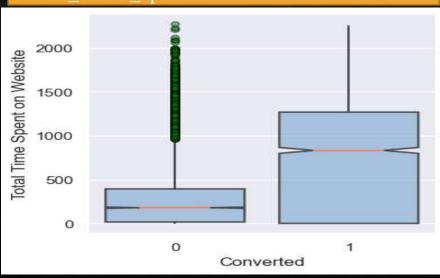




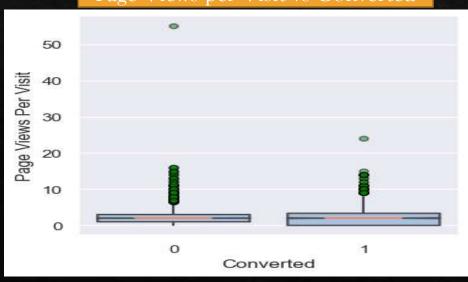




Total_Time_spent on website vs Converted

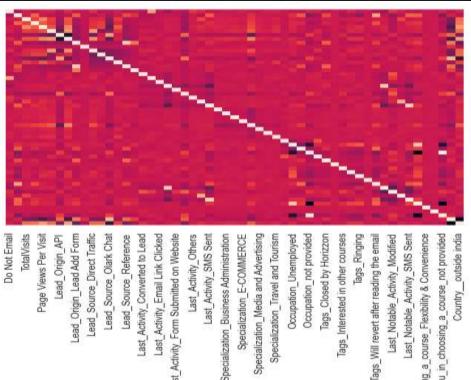


Page Views per Visit vs Converted



CORRELATION HEATMAP

Do Not Email Total Time Spent on Website Lead_Origin_API Lead_Origin_Lead Import Lead_Source_Olark Chat Lead_Source_Welingak Website Last_Activity_Email Link Clicked Last_Activity_Olark Chat Conversation Last_Activity_SMS Sent Specialization_Business Specializations Specialization_Media and Advertising Occupation_Student Occupation_not provided Tags_Interested in Next batch Tags_Ringing Last_Notable_Activity_Email Opened Last_Notable_Activity_SMS Sent What_matters_most_to_you_in_choosing_a_course_Other Country outside india



Form Submitted on We Last_Activity_Email Link Cl

Last Activity

What matters most to you in choosing a course not provided Last Notable Activity Modifie Last Notable Activity SMS Ser s most to you in choosing a course Flexibility & Conven Country

Interested in other co

- 1.00 - 0.75 -0.50-0.250.00

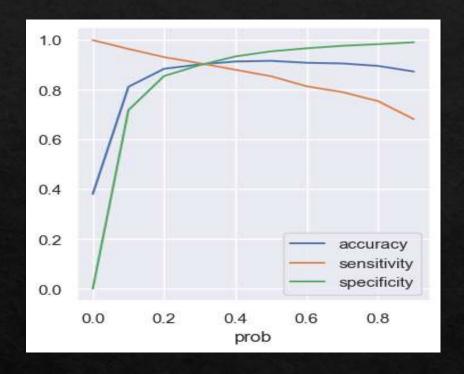
DATA CONVERSION:

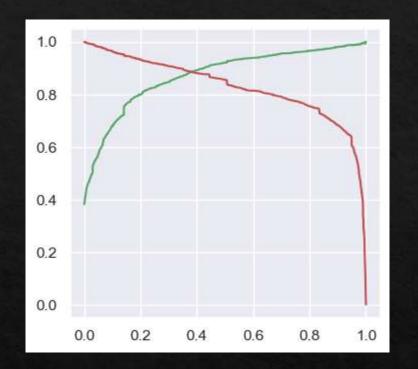
- ♦ Numerical variables are normalised.
- Dummy variables are created for objective type variables.
- ♦ Total Rows for Analysis = 9240
- ♦ Total columns for Analysis = 37

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 20 variables as output.
- ♦ Building Model by removing the variables whose p-value is greater than 0.05 and vif value is greater than 5.
- ♦ Predictions on test data set.
- ♦ Overall accuracy is 85%.

ROC CURVE





- Finding Optimal Cut off point.
- Optimal cut off probability is that where we get balanced sensitivity and specificity.
- ♦ From the first graph it is visible that the optimal cut-off is at 0.3.

Conclusion:

- ♦ It was found that the variables that mattered the most in the potential buyers are (In descending order):
- 1. What_matters_most_to_you_in_choosing_a_course
- 2. Tags_Will revert after reading the email
- 3. Last_Notable_Activity_Modified
- 4. a.Tags_Other Tags b.Tags_Ringing
- 5. a. Last_Activity_SMS Sent b. Last_Activity_Olark Chat Conversation
- 6. Occupation_Working Professional
- 7. Lead_Origin_Lead Import
- ♦ So it is evident that our model worked pretty well. The conversion rate before model building was 38.53% while after model building it went upto 82.09%. We were successful in target lead conversion rate to be even higher than 80% as demanded by the CEO of X Education.
- * Keeping these in mind the 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.