

# LEAD SCORE ASSIGNMENT

By – Suryadevi Mohan & Binita Sagariya



# PROBLEM STATEMENT

- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- 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. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

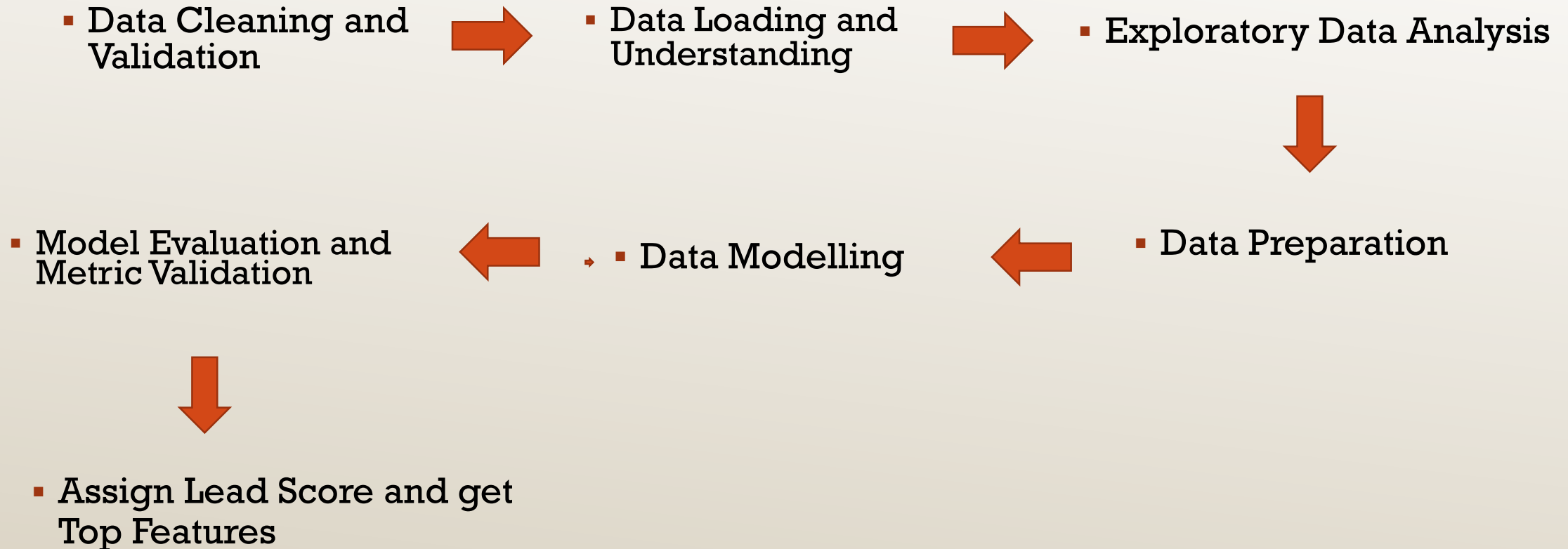


# OBJECTIVE

- X- Education wants to build a Logistic regression model to identify most potential Leads which they target in order to flourish their sales and for future use.
- They also wants to know the solution for their business problems which the logistic regression AI model infers it.



# SOLUTION APPROACH



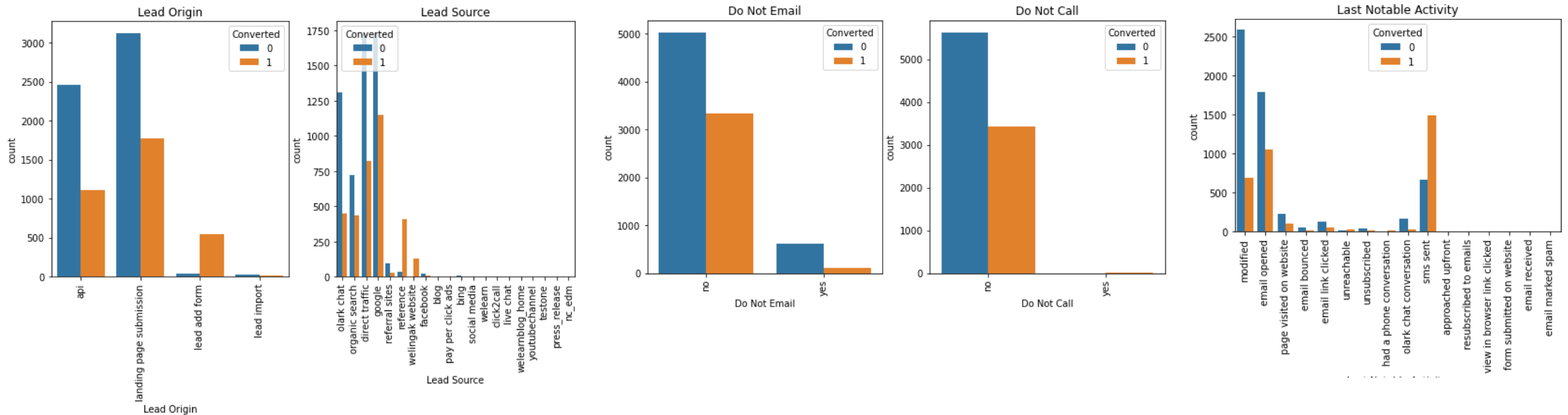
# DATA CLEANING

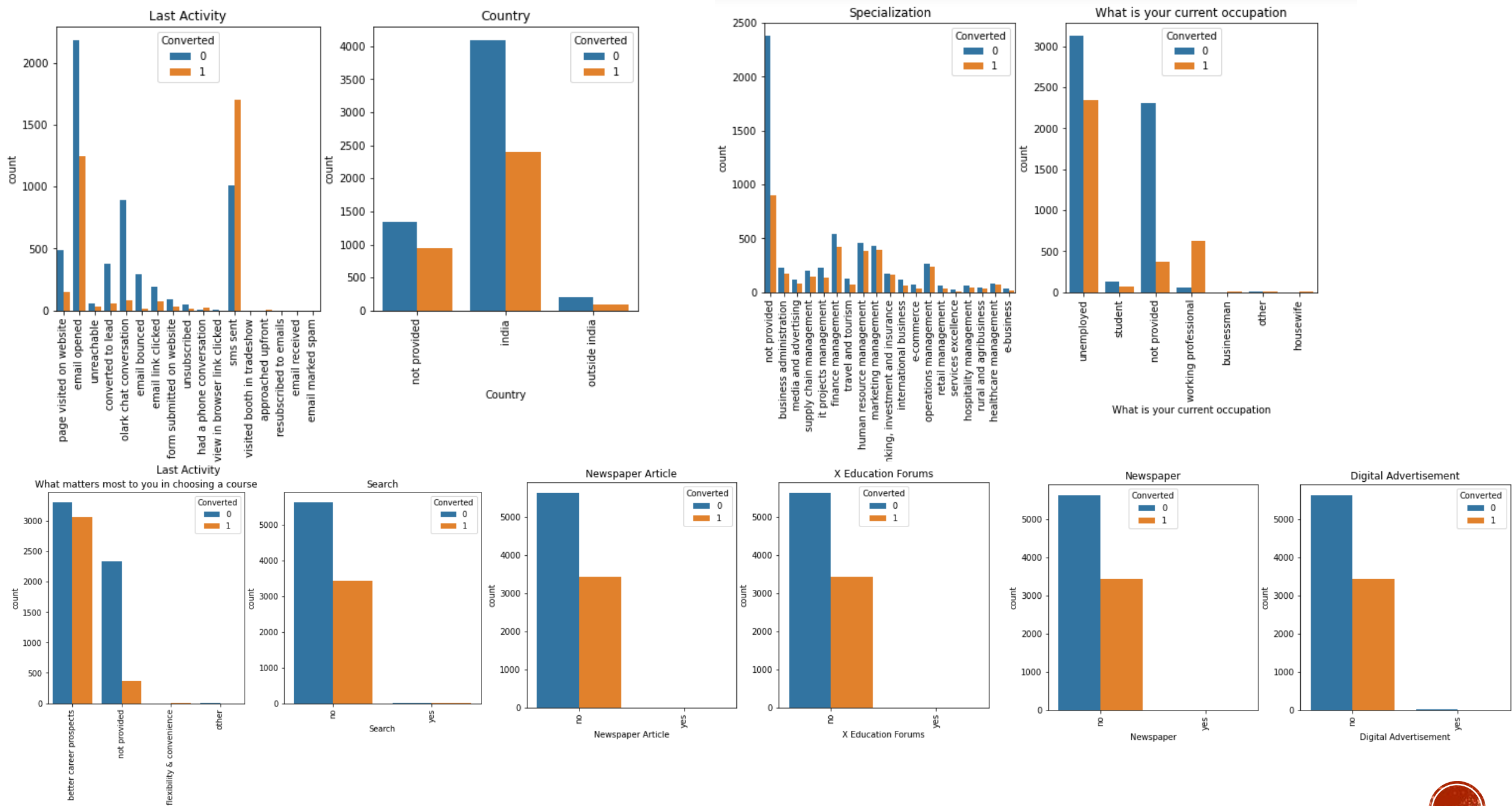
1. Select values are replaced by Null values.
2. Univariate categorical columns are dropped.
3. More than 35% null value columns are dropped.
4. Skewed columns such as Country column has been handled by assigning lower values to new value “outside India”.
5. Few columns considered though it has 35% null value as it gives some useful analysis  
For Ex: Specialization, What matters most to you in choosing a course.



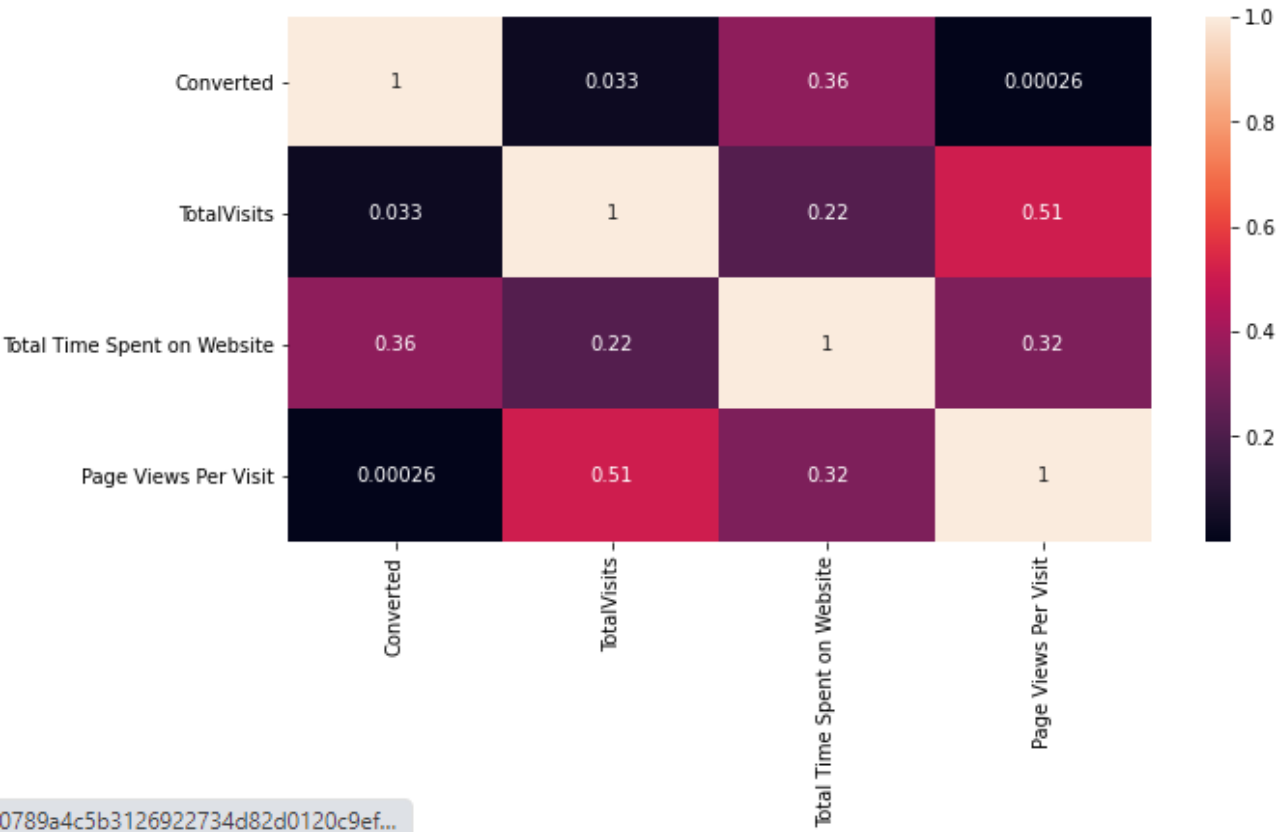
# EXPLORATORY DATA ANALYSIS

- EDA performed for both Categorical and numeric variables
- Univariate and Bivariate analysis has been done below are the some plots of feature variable that are converted/not converted





Correlation among the variable has been checked using correlation metrics



0789a4c5b3126922734d82d0120c9ef...





# MODEL BUILDING AND EVALUATION

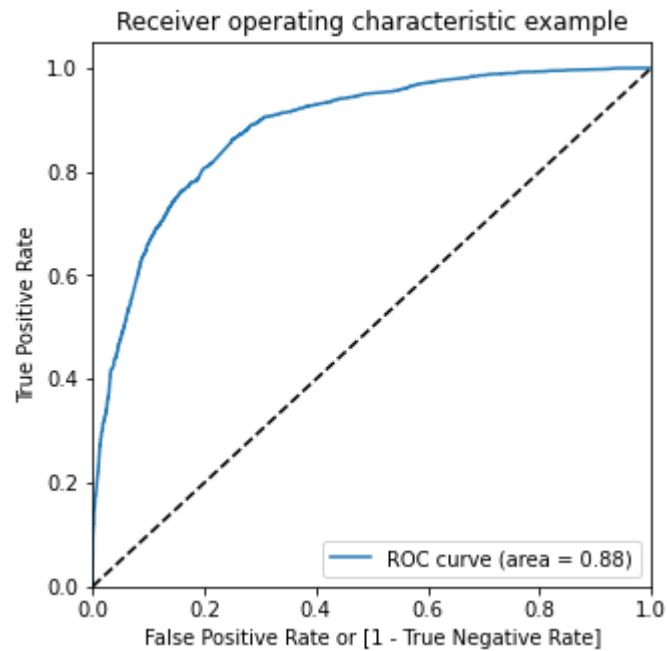
- The Train and Test set split was done at 70 – 30 percentage from the data frame
- RFE was done to achieve the best 15 variables and the rest of the variables were eliminated manually depending upon the P –values and VIF values. We ensure at the end of model building the VIF is less than 2.5 and P value is less than 0.05.
- From Confusion Matrix Overall accuracy , Specificity, Sensitivity is calculated



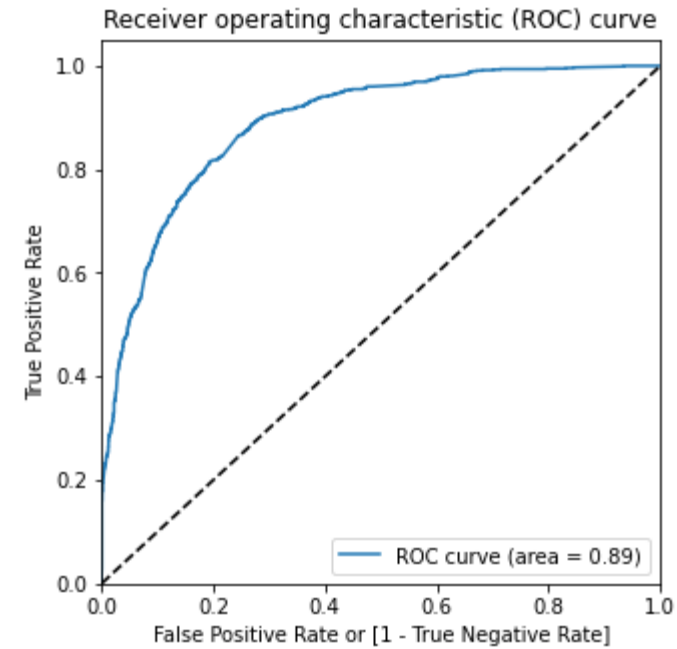
# MODEL EVALUATION

	Overall Model Accuracy	Specificity	Sensitivity	Precision	Recall	AUC
Train Set	0.80	0.80	0.80	0.72	0.80	0.88
Test Set	0.81	0.80	0.81	0.70	0.81	0.89

ROC Curve on Train Data



ROC Curve on Test Data



# CONCLUSION

- Based on our model, some features are identified which contribute most to a Lead getting converted successfully.
- The variables given below in descending:
  - 1. Last activity is Olark Chat Conversation
  - 2. Current occupation\_student
  - 3. Lead Origin\_lead add form
  - 4. Lead sources are
    - i). Olark Chat
    - ii). Welingak Website
  - 5. Total time spent on website is more
  - 6. Total Visit on website is more
  - 7. When the last activity is SMS sent and Do not Email is Yes
  - 8. When their current occupations are Unemployed and a working professional.
- Keeping these variables in mind the X Education can develop their Sales by focusing these potential buyers to buy their courses.



**THANK YOU**

