# Lead Score Case Study

Group Members:
Jithin K
Kaushlendra Pandey

# Case study - Steps involved:

- Problem statement
- Data cleanup
- EDA - Univariate and Bivariate analysis
- Creating dummies/Train-test split.
- Model building/Evaluation.
- Conclusion/summary.

# Problem Statement:

- X education is an online education platform which provides online courses to industry professionals.
- The queries in their platforms are converted as leads and the lead conversion rate is 30% which is very bad as per the company.
- They are planning to separate the leads into hot leads and cold leads where in the conversion rate in hot leads is very high.
- We need to help them in finding out the proper hot leads so that the sales team can focus on the hot leads and the conversion rate will spike up.

# Business expectation:

- We are expected to identify the hot leads with the model building and evaluation techniques.
- We need to build a model to identify the potential leads and evaluate the prediction.
- We need to deploy the model do the predictions evaluate the results and analyze the end outcome then give a conclusion.

# Steps involved:

- Data scrub:
  - Did describe the data and check the info of the data set.
  - Checked the nulls and missing values.
  - Dropped the columns which had missing rows more than 3000.
  - Dropped the null rows and rows with value 'Select'.

- EDA:
  - Univariate analysis - Checked the data distribution, value counts etc.,
  - Bivariate analysis - Observed the results by plotting the categorical variables against the target variable.

- Creating Dummies - After the EDA did create dummies for the categorical variables and did feature scaling for the numeric variables.

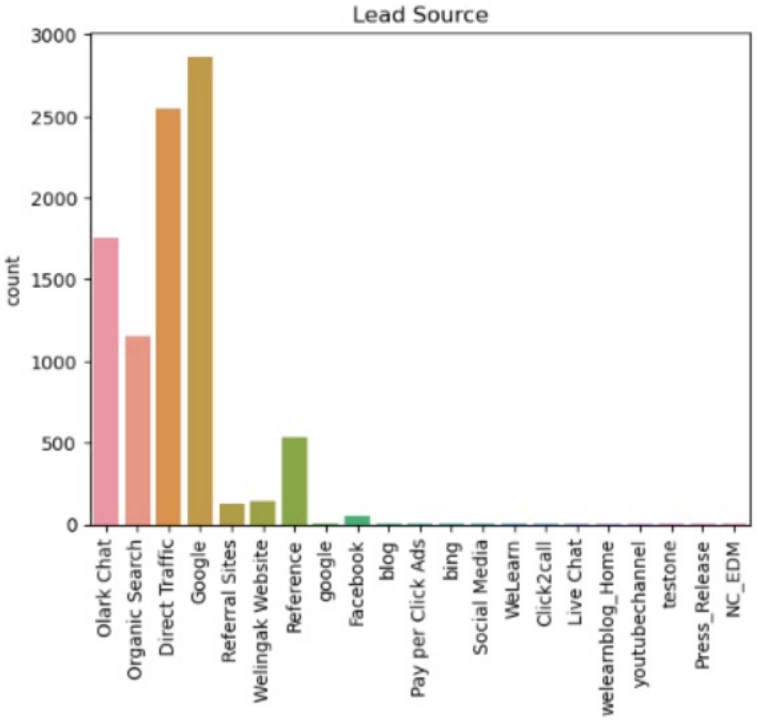
- Model building - Built a model using logistic regression technique and made the predictions on the dataset.

- After creating the model and making the predictions we evaluated the predictions with the help of accuracy metrics.
- Deployed and presented the model.
- Provided the summary and conclusions.

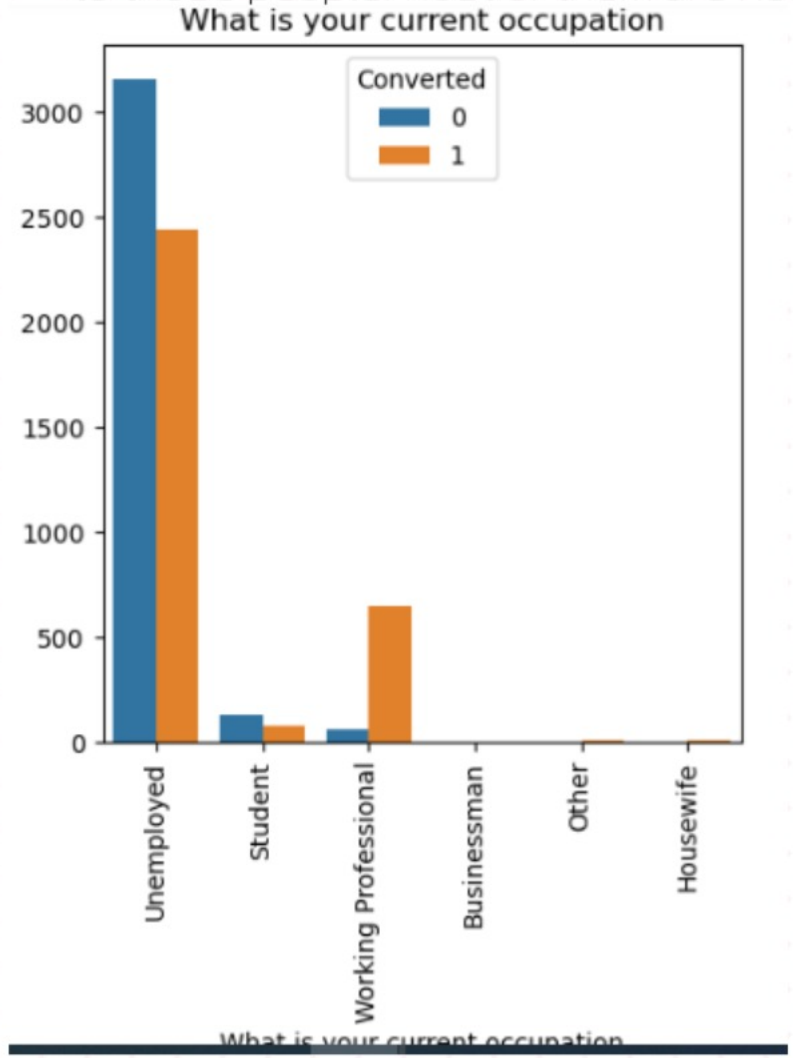
# EDA:

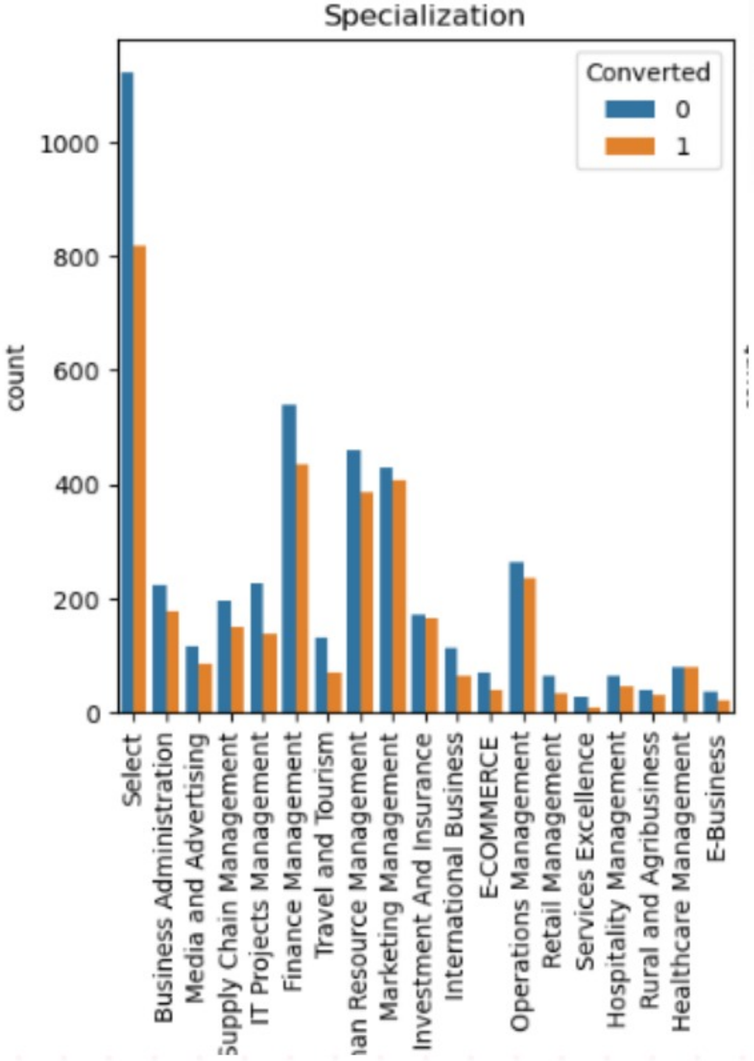
- The below graph shows that through google search the firm has got lot of leads, they can concentrate on Digital marketing and search engine optimization.

.

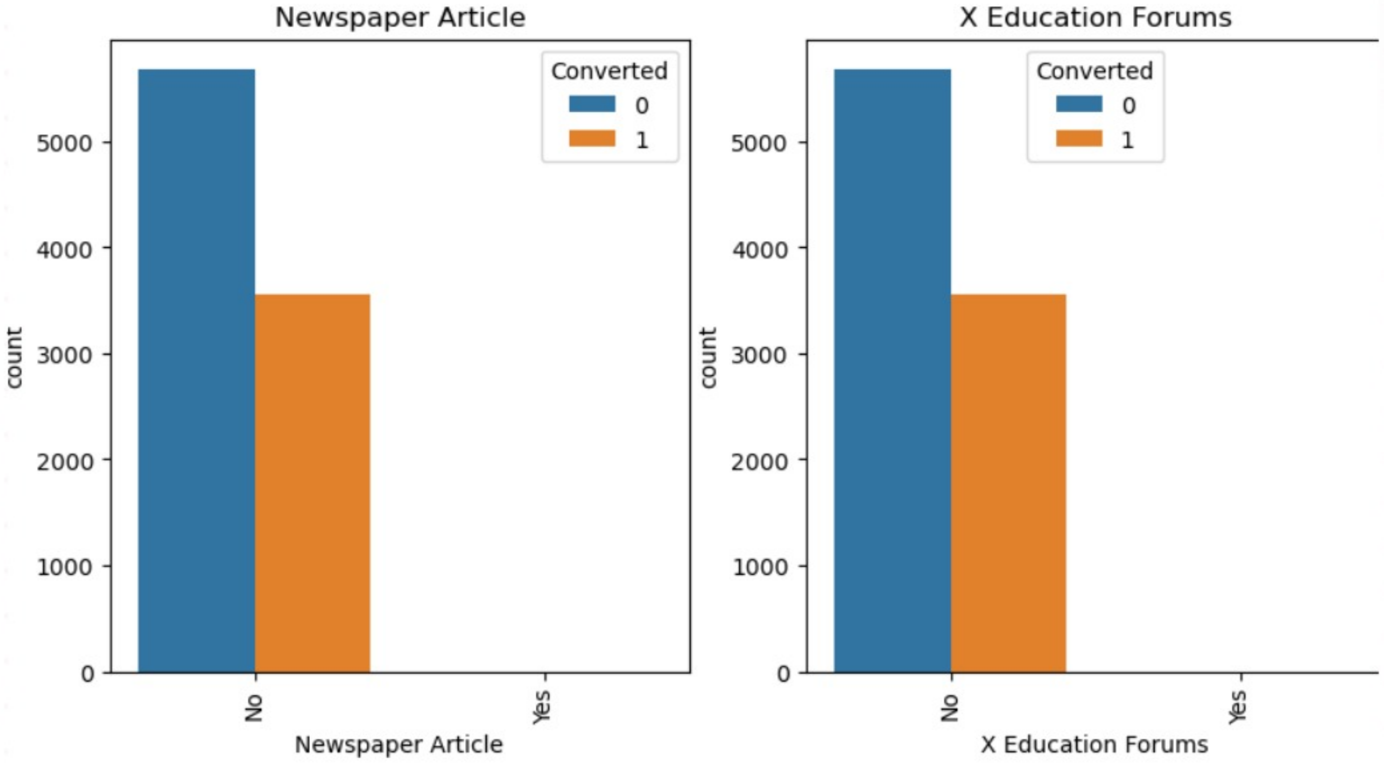


- Most of the Quereis are unemployed or working professionals the sales team can effectively pitch to those people. Most of them are not sure about the specialization.

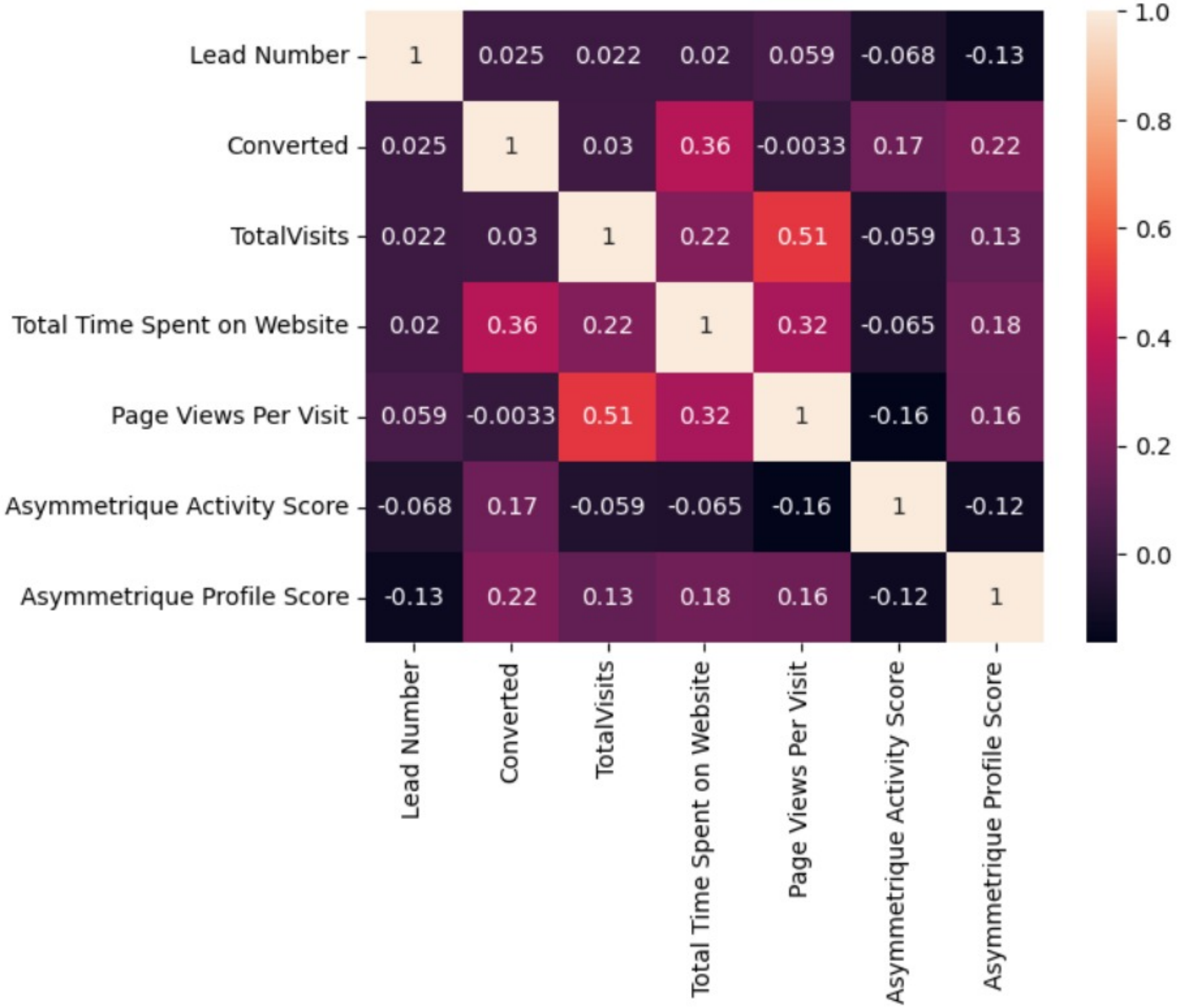




- It's better not spending much on the newspaper ads as it doesn't add much value to the lead conversion.



✧ Finding the correlation between the columns.



# Dummies Creation/Train-test split:

- Created dummies for the below categorical columns and dropped the original columns.
  - Leads origin, Lead source, Do not email, Last activity
  - What is your occupation, A free copy of mastering the interview,
  - Last notable activity.
- Dropped the above columns after creating the dummies.
- Perform train test split with 70% for training and 30% for testing.
- Performed the Min-max scalar for the following numerical columns
  - Total visits, Total time spent on website, page views per visit.

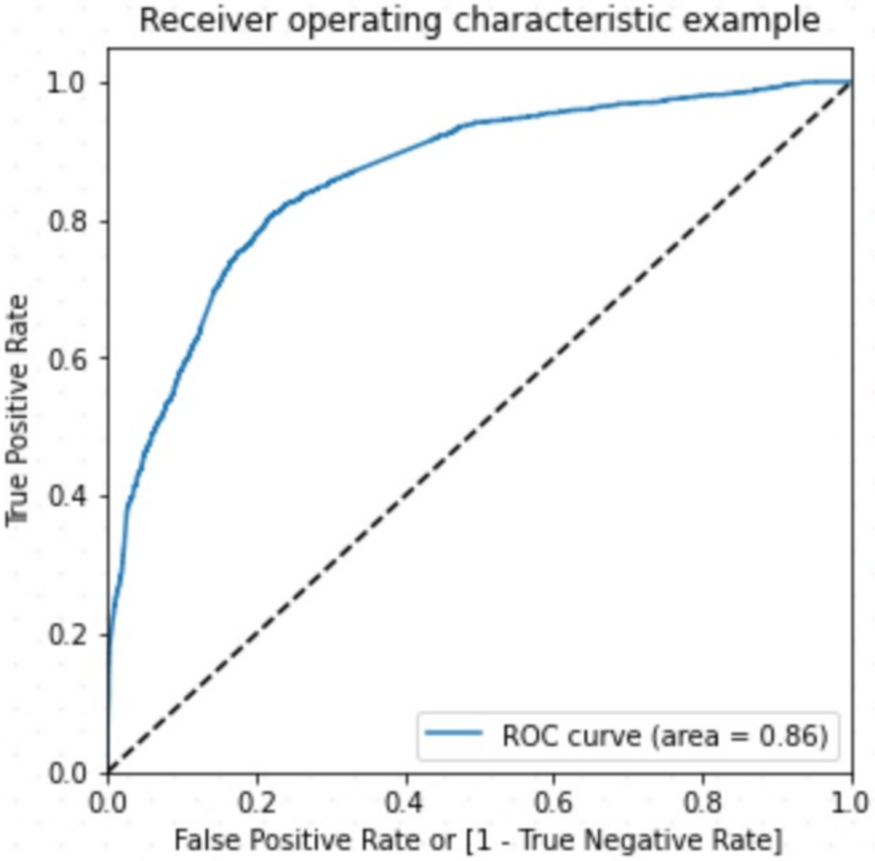
# Model building:

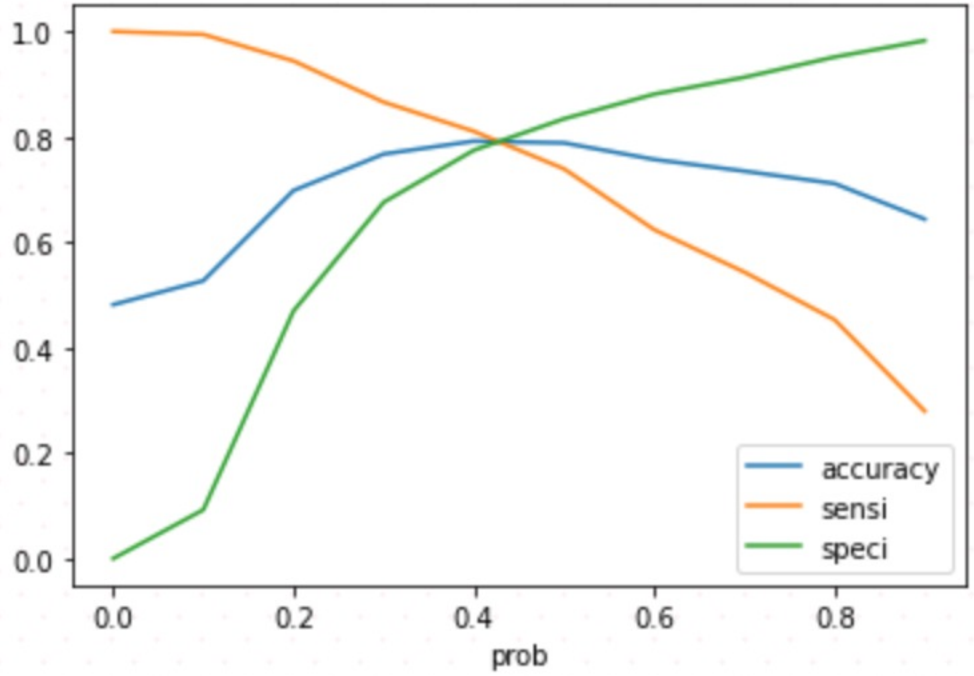
- After assigning the target column 'converted' to y_train and rest of the columns to x_train we will perform the RFE with no of variable selected as 15.
- With the resultant columns checked the p-value and VIF values.
- Dropped the VIFs greater than 5 and p-values greater than 0.05 one at a time.
- After having the VIFs and p-values in a promising range finalized the dataset and proceeded to the next step.

# Evaluation Training set:

- Here we predicted the conversion by adding the conversion probability.
- Then we did the evaluation using the following evaluation techniques confusion matrix, accuracy, sensitivity and specificity.
- Then we drew a ROC curve to validate the predicted values. The ROC value is 0.86.
- The evaluated metrics values are as follows
  - Accuracy - 78%
  - Sensitivity - 73%
  - Specificity - 83%
  - Optimal cutoff point - 0.42

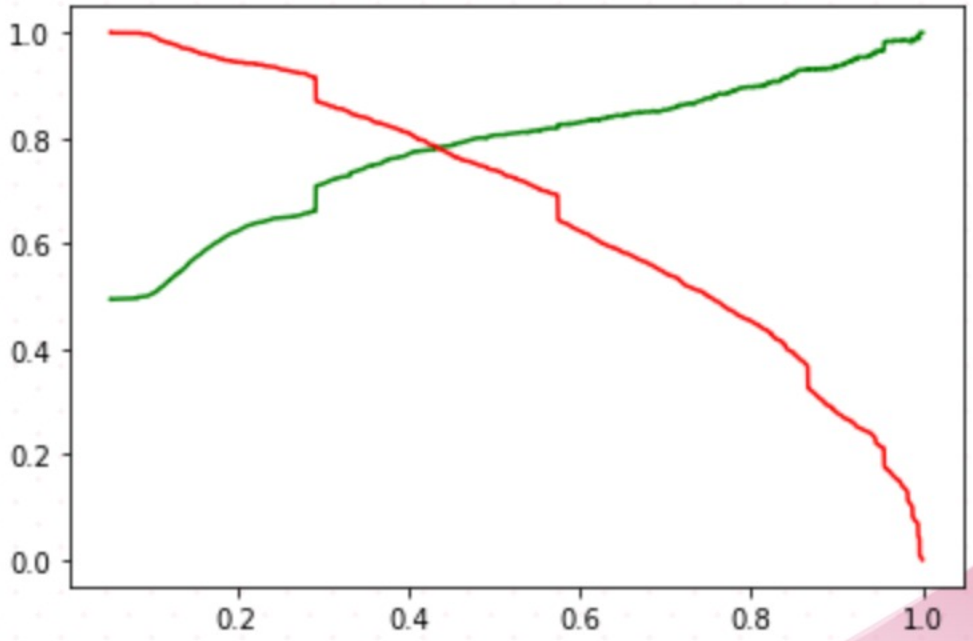
- Below we can find the ROC curve and the optimal cutoff point graph.





# Prediction and model evaluation - Test data set:

- Following are the resultant from the final evaluation.
  - Overall accuracy - 78%
  - Sensitivity - 77%
  - Specificity - 78%
  - Optimal cutoff point - 0.44



# Conclusion and recommendations:

- Visitors who spend more on the website tend to enroll for the courses mos of the time.

- Visitors who visit multiple times to the website and making queries are interested in course enrollment.

- Most of the converted leads are from Google, direct traffic and organic search. As recommended the firm can invest on digital marketing/google ads.

- Most of the persons enrolled themselves after getting SMS and through Olark chat conversation.

- Mostly the working professionals and people who are unemployed are making the queries so that they can get a job and upscale themselves. So the company can effectively pitch to working professionals and unemployed people to increase the lead conversion rate.