

# LEAD SCORE CASE STUDY



## Group Members:

1. Radhika Mahajan
2. Aqib Jallal
3. Aishwarya Girhare



## Problem Statement:


- X Education sells 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 known as 'Hot Leads'.
- If they successfully identify this set of leads, the lead conversion rate should go up as 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 Methods:

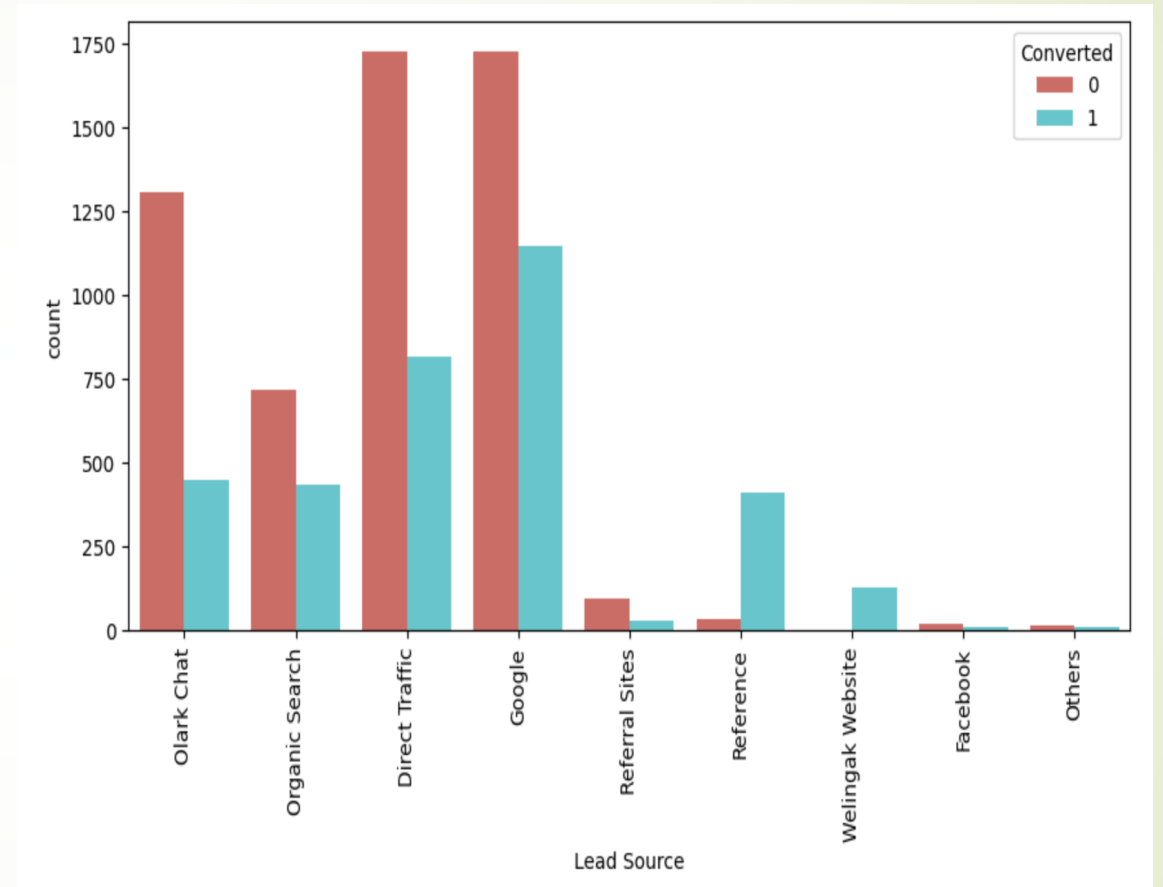
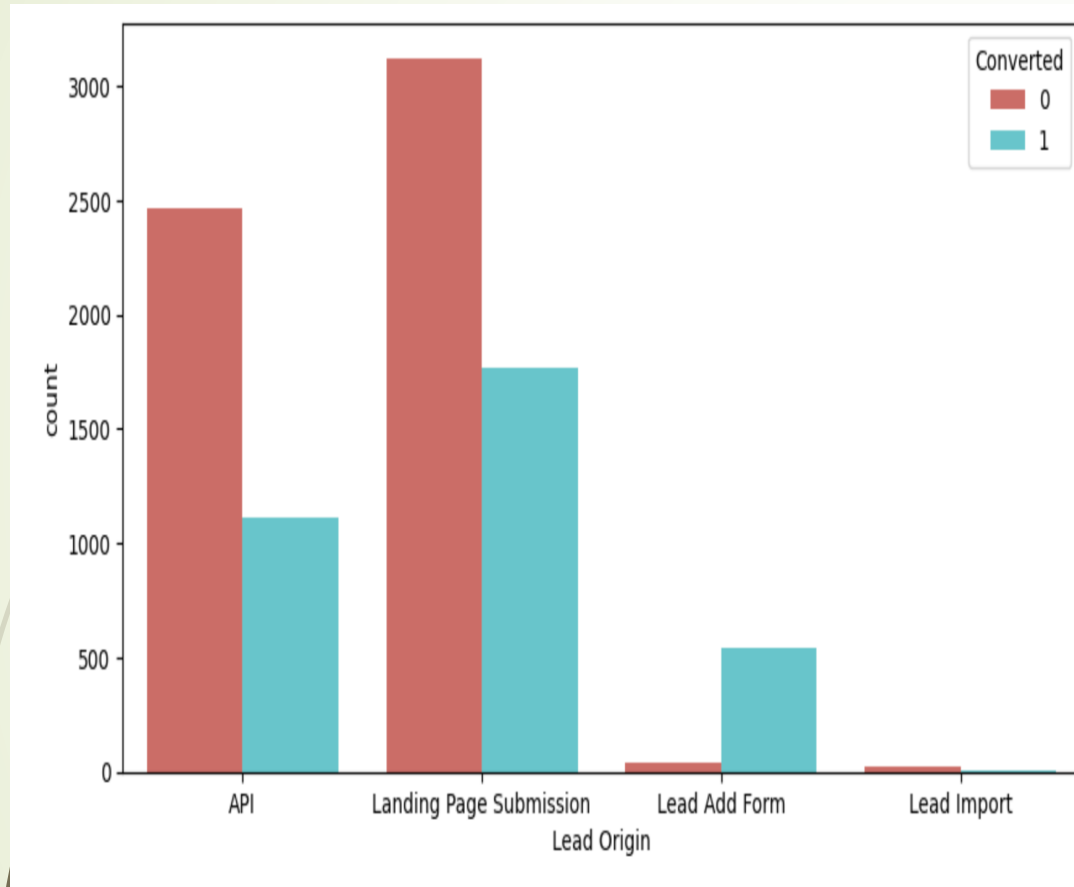
- Data cleaning and data manipulation.
    1. Treatment of missing values. 'Select' were treated as null.
    2. Drop columns, if it contains more than 40% missing values.
    3. Imputation of the values based on count plots.
    4. If very few rows with missing data, then those rows were removed.
  - **EDA** Univariate data analysis: count plots for categorical with target variable as hue, histogram for numerical.
  - Dummy Variables, Train-Test split, Feature Scaling.
  - Model building using logistic regression.
  - Generating predictions and Evaluation of the model.
  - Lead scores calculated and Hot leads identified.
  - Conclusions and recommendations.
- 



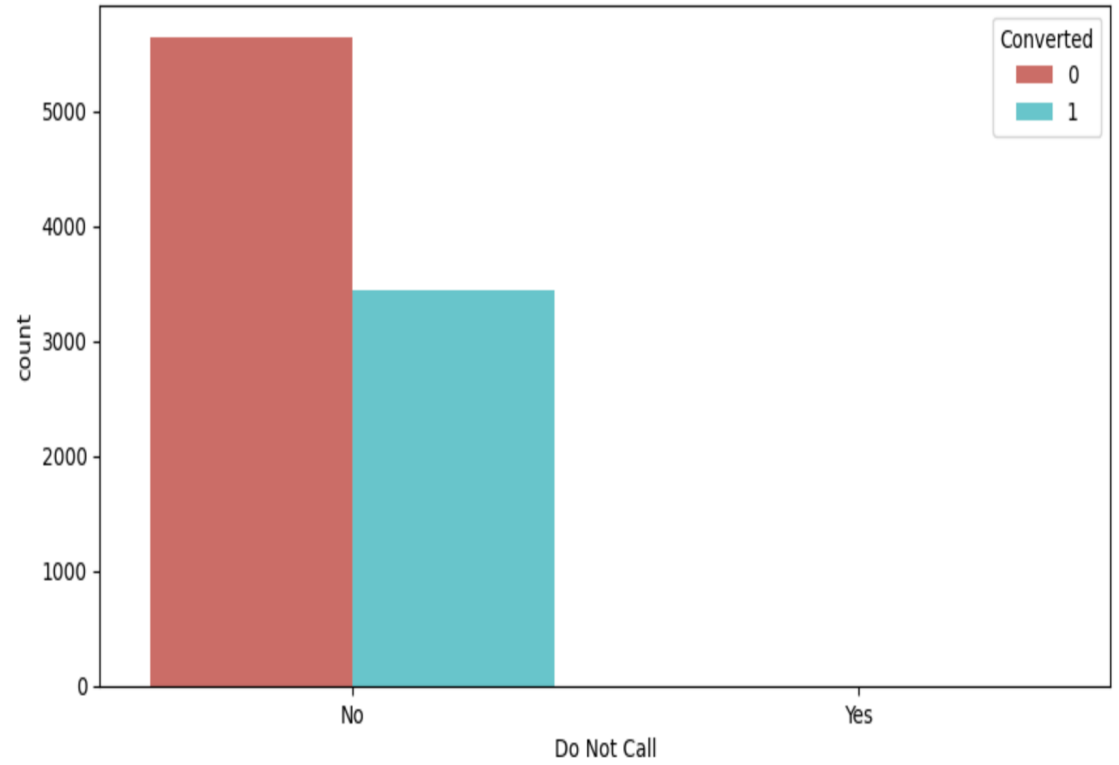
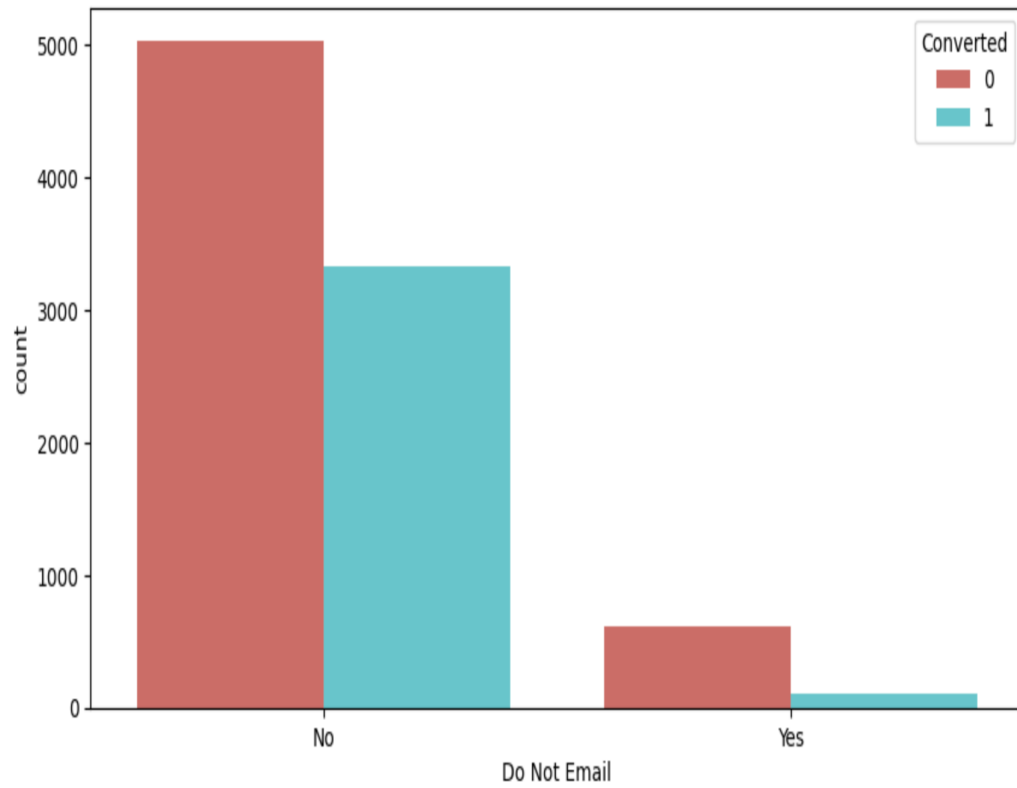
## Data Manipulation:

- Total Number of Rows =9240, Total Number of Columns =37 initially.
- Single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update me on Supply Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped.
- Removing the “Prospect ID” and “Lead Number” which is 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 has no enough variance, which we have dropped, the features are “What matters most to you in choosing course”, “Search”, “Newspaper Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement” etc.
- Dropping the columns having more than 40% as missing value such as ‘How did you hear about X Education’ and ‘Lead Profile’.

# EDA



# EDA





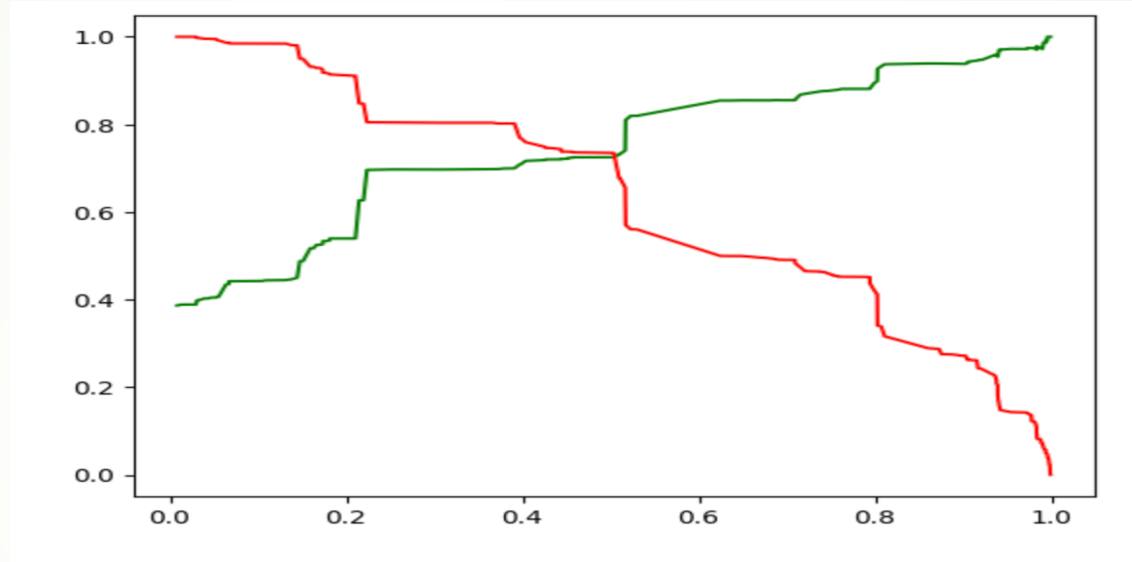
## Data Conversion

- Numerical Variables are scaled using Standard scaler.
- Dummy Variables are created for object type variables.

## 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 train and test data set
- Overall accuracy approx. 79% on train as well as test data.

# Trade-off curve between precision and recall



## Optimal Cut off Point

- Optimal cut off probability is that probability where we get balanced precision and recall.
- From the graph the optimal cut off is 0.5.





# Recommendation:

- The company should make calls to the following leads as these are more likely to get converted:
  - The leads with the lead source as "Welingak Website"
  - The leads who are "working professionals"
  - The leads with origin as 'Lead Add Form'
  - The leads who spent "more time on the website"
  
- The company should NOT spend too much time on the following type of leads as these are less likely to get converted:
  - The leads with last activity as "Olark Chat conversation"
  - The leads whose Specialization was "Others"
  - The leads with lead origin as 'Landing page submission'
  - The leads whose last notable activity was 'Email opened'
  - The leads whose last notable activity was 'Page visited on website'
  - The leads with last notable activity as "Olark Chat conversation"
  - The leads who chose the option of "Do not Email" as "yes"
  - The leads with last notable activity as "Modified"
  - The leads with last notable activity as "Email link clicked"