

# **LEAD SCORING CASE STUDY**

By :Kanishka / Sanskriti / Kartik

## **Problem Statement**

- X Education sells online courses to industry professionals.
- Despite receiving a high volume of leads, X Education's conversion rate is low. For example, if they generate 100 leads in a day, only around 30 of them will convert.
- To improve efficiency, the organization aims to discover high-potential leads, or 'Hot Leads'.
- Identifying these leads can increase lead conversion rates since the sales staff can focus on connecting with them instead of calling everyone.

## **Business Objective**

- X Education seeks to identify the most promising leads and develop a model to do so.
- Developed the model for future use.

# Methodology for Solving

- Data cleanup and manipulation.
  - Check and manage duplicate data.
  - Check and handle NA and missing values.
  - Drop columns that contain a substantial number of missing values and are useless for the analysis.
  - If necessary, the values will be imputed.
  - Check for and handle outliers in data.
- Univariate data analysis includes value counts and variable distributions, among other things.
- Bivariate data analysis includes correlation coefficients

# EDA

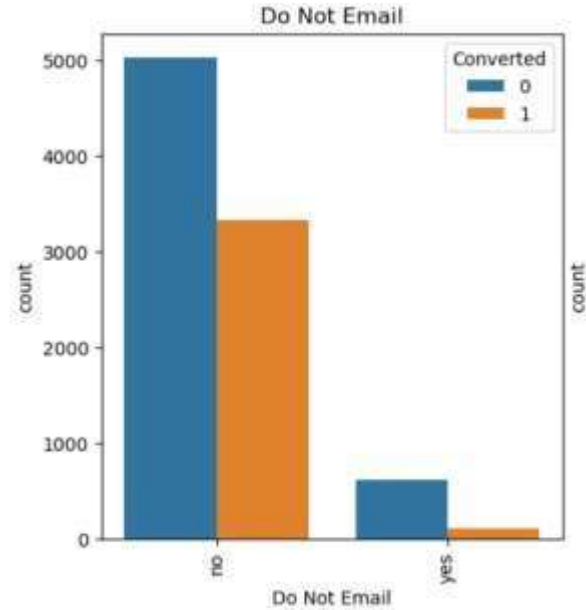
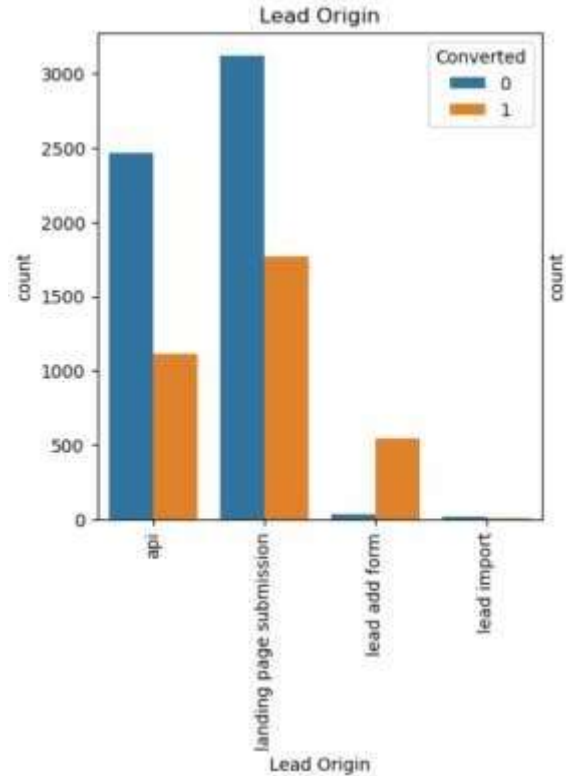
- Feature Scaling and Dummy variables and data encoding.
- Logistic regression is the classification approach used to create and predict models.
- Model validation, presentation, and recommendations.

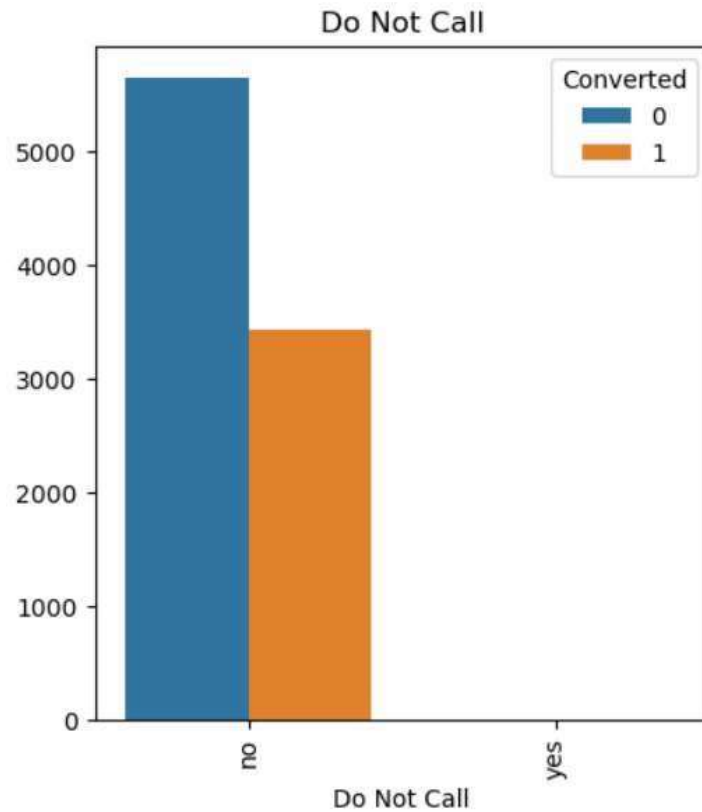
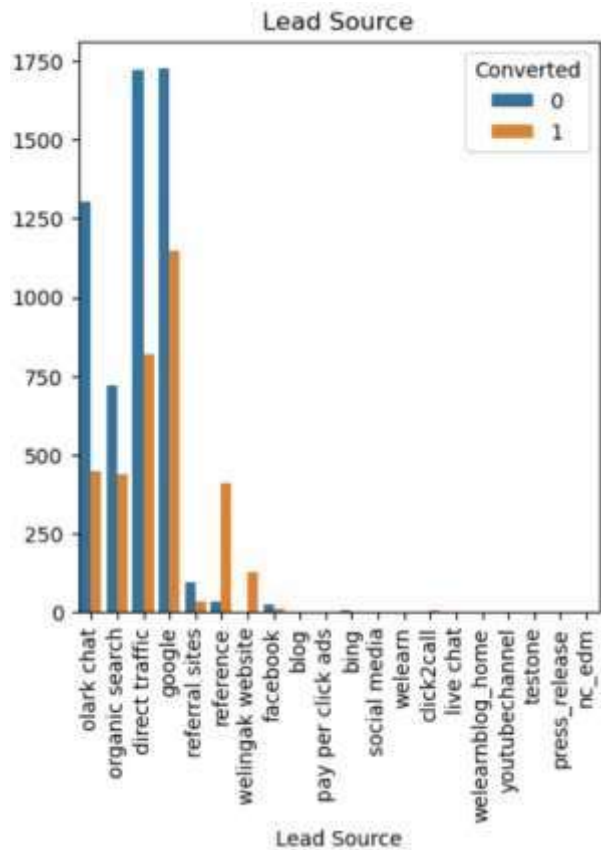
# Data Manipulation

- Total rows = 37, total columns = 9240.
- Single value features include "Magazine," "Receive More Updates About Our Courses," and "Update me on Supply."
- "Chain Content," "Get updates on DM Content," and "I agree to pay the amount through cheque" have been removed.
- Removed unnecessary "Prospect ID" and "Lead Number" fields for analysis.
- Object type variables with insufficient variance were removed, including "Do Not Call," "What matters most to you in choosing a course," "Search," "Newspaper Article," "X Education Forums," "Newspaper," and "Digital Advertisement."
- Dropping columns with more than 35% missing values, such as 'How did you hear about X Education' and 'Lead Profile'.

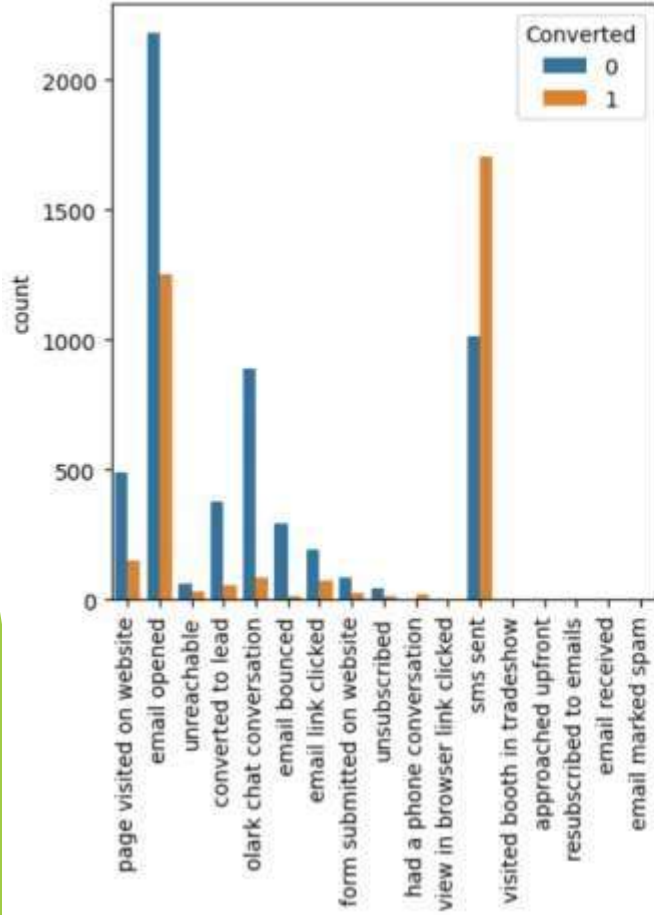
# EDA

## Categorical Variable Relation

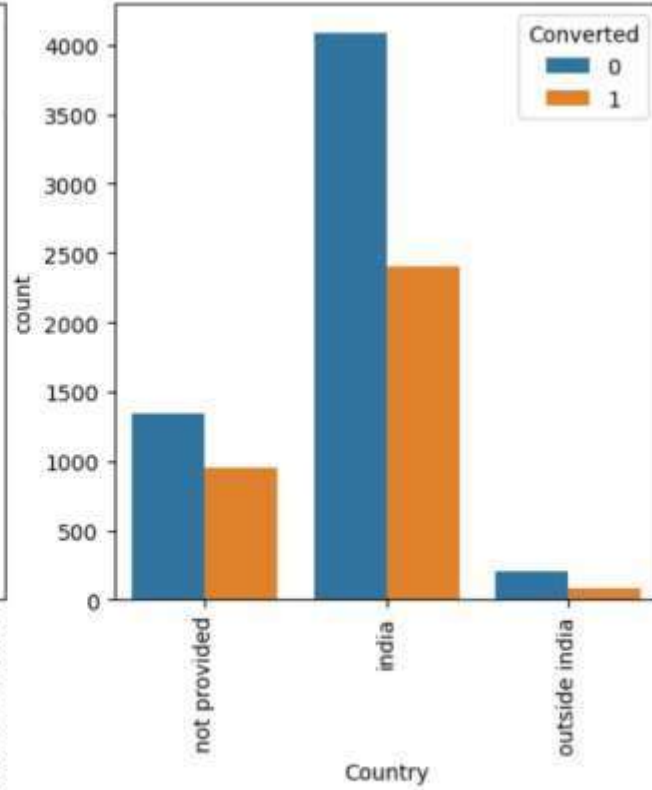




### Last Activity



### Country





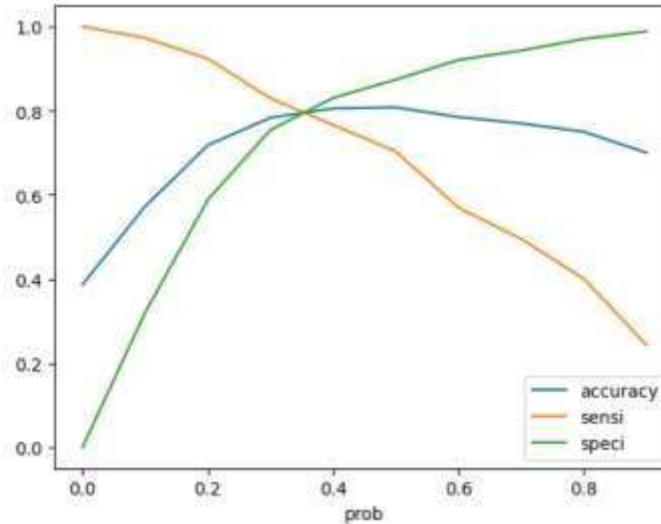
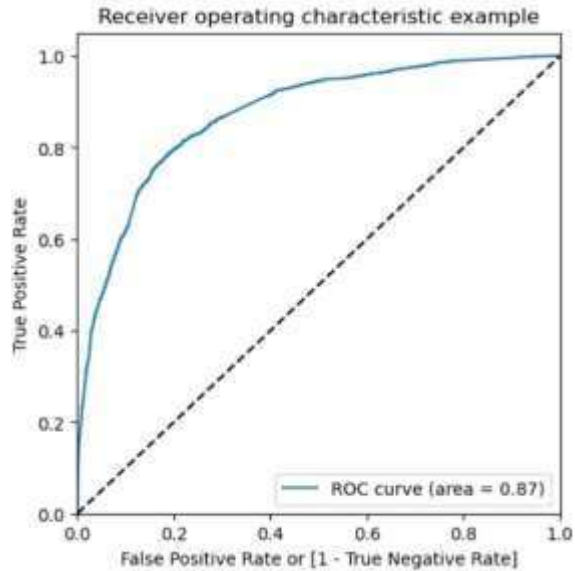
# Data Conversion

- Numerical Variables are Normalised
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 8792
- Total Columns for Analysis: 43

# Model Building

- The first stage in regression is to partition the data into training and testing sets. We used a train-test ratio of 70:30.
- Use RFE for feature selection.
- Run RFE with 15 variables as output.
- Build model by deleting variables with  $p\text{-value} > 0.05$  and  $vif > 5$ .
- Made predictions on test data set with 81% accuracy.

# ROC Curve



## Finding the Optimal Cut-off Point

The ideal cut off probability is 0.35, which provides balanced sensitivity and specificity, as shown in the second graph.

# Conclusion

It was found that the variables that mattered the most among the potential buyers are:

- The total time spend on the Website.
- Total number of visits.
- When the lead source was:
  - Google
  - Direct traffic
  - Organic search
  - Welingak website
- When the last activity was:
  - SMS
  - Olark chat conversation
- When their current occupation is as a working professional.

With this in mind, X Education can grow since they have a great possibility of persuading nearly every prospective customer to change their mind and enroll in their courses.