



# X Education- Identifying Hot Leads (Case Study)


BY RITIKA SHARMA, RAVI RAJ KAMAL AND RITESH RAY

# Problem Statement-

- ▶ X education employs a variety of methods to attract potential customers, including advertising on websites and search engines such as Google. When individuals visit the company's website, they may browse available courses, watch videos, or fill out a form to express interest. These individuals are considered leads once they provide their contact information. Additionally, X education receives leads through referrals from past customers. The sales team then contacts these leads through phone calls and emails in an attempt to convert them into paying customers. While not all leads become customers, X education has a lead conversion rate of approximately 30%.
- ▶ The company has tasked you with creating a model that assigns a lead score to each potential customer. The goal is for customers with higher lead scores to have a greater chance of conversion, while those with lower lead scores have a lower chance. The CEO has set a target lead conversion rate of approximately 80%.
- ▶ 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.

# Solution Methodology-

- Data cleaning and data manipulation.
  1. Check and handle NA values and missing values.
  2. Drop columns, if it contains large amount of missing values and not useful for the analysis.
  3. Imputation of the values, if necessary.
  4. Check and handle outliers in data.
- EDA
  1. Univariate data analysis: value count, distribution of variable etc.
  2. Bivariate data analysis: correlation coefficients and pattern between the variables etc.

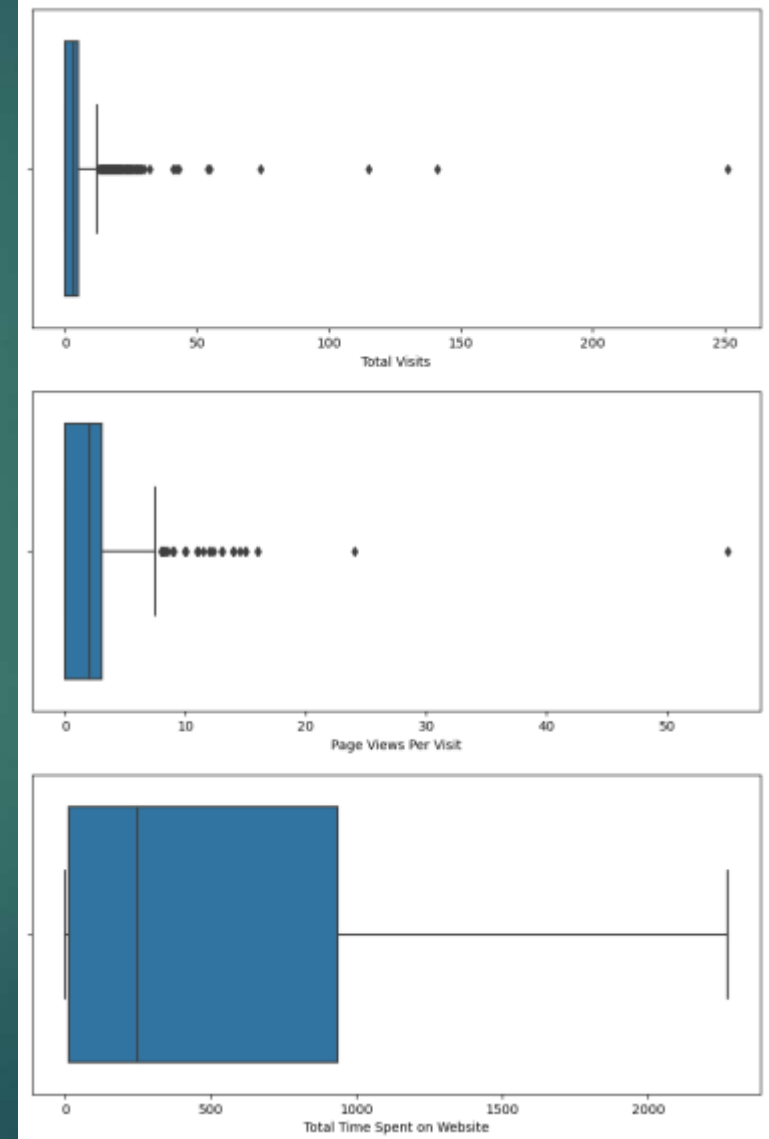
- 
- Feature Scaling & Dummy Variables and encoding of the data.
  - Classification technique: logistic regression used for the model making and prediction.
  - Validation of the model.
  - Model presentation.
  - Conclusions and recommendations.

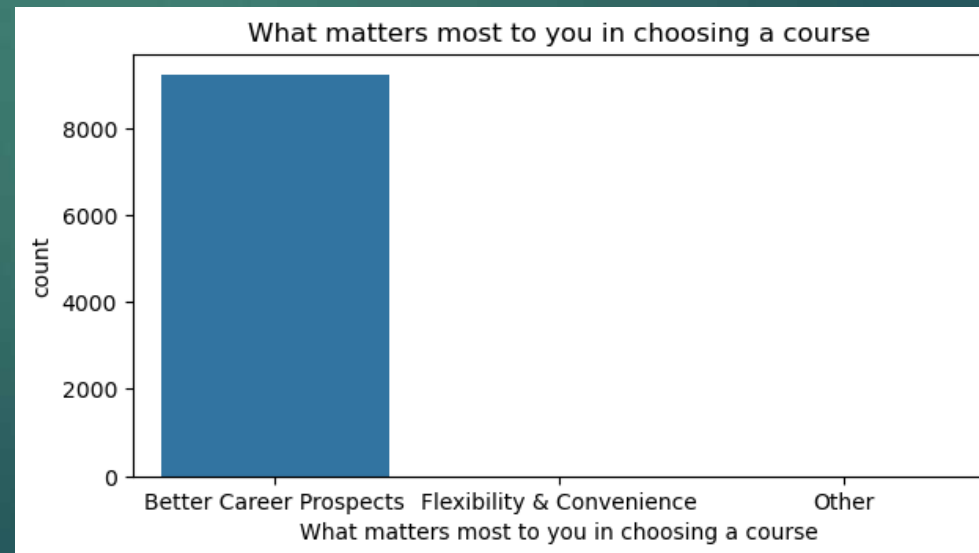
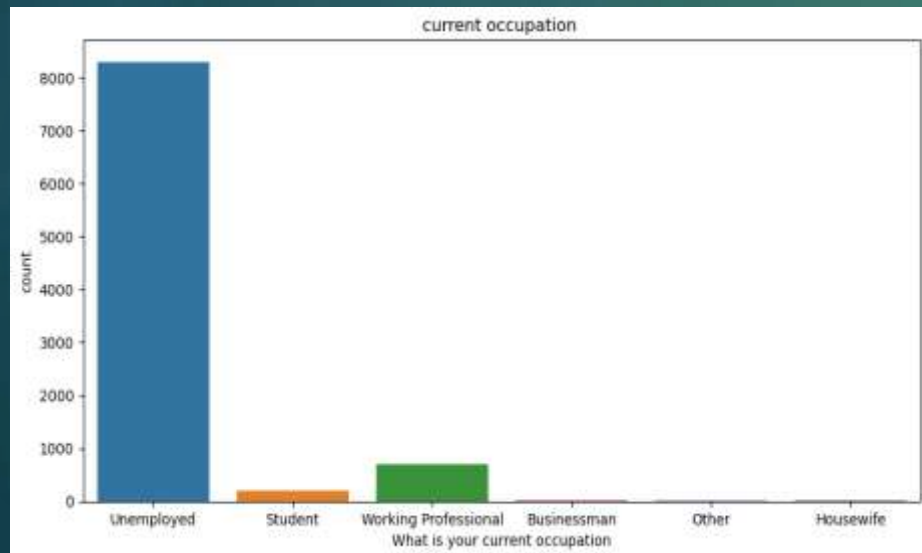
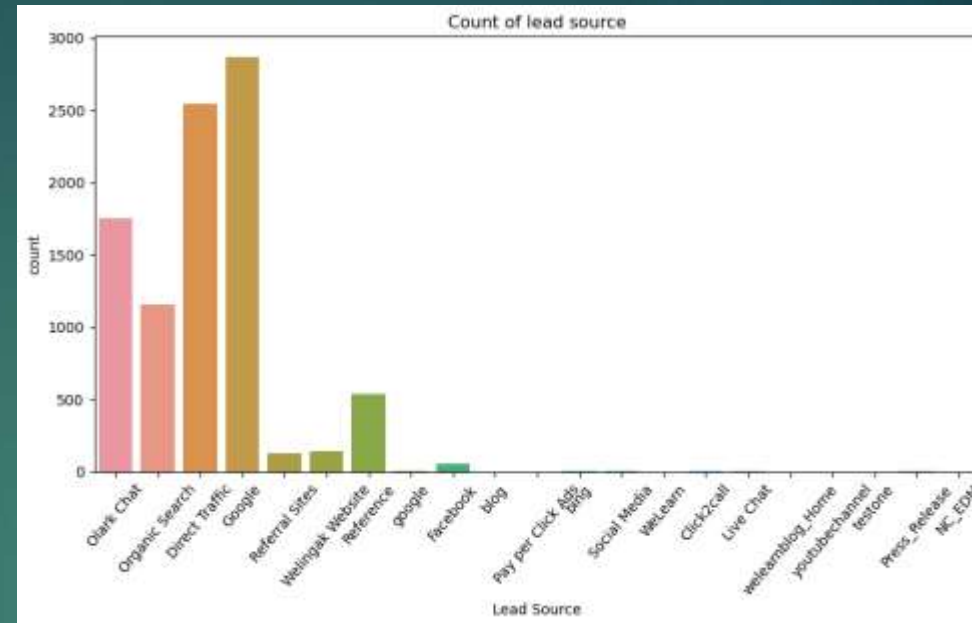
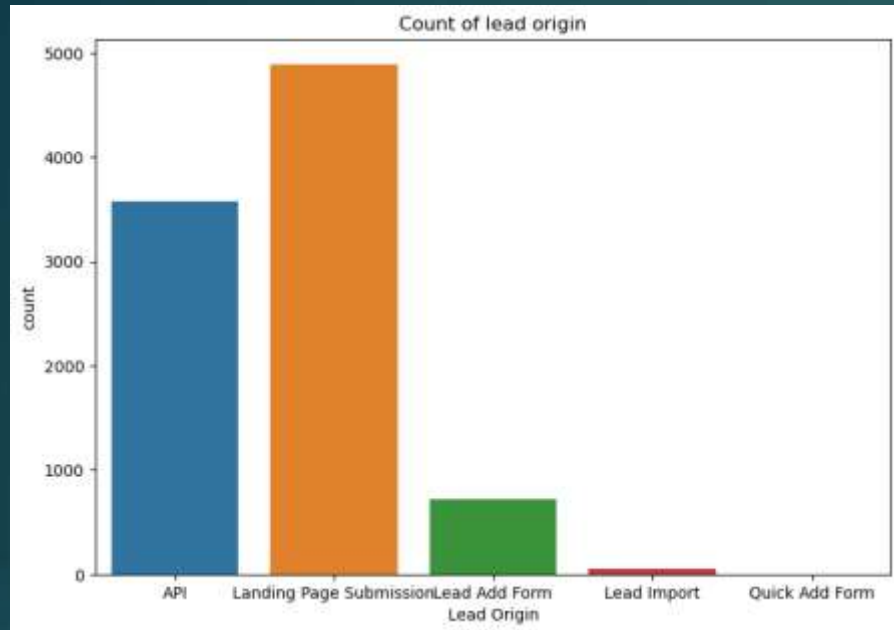
# Data Manipulation


- Total Number of Rows = 9240, Total Number of Columns = 37
- Dropped columns having more than 40% null value like, “Asymmetrique Activity Index”, “Lead Profile”.
- Dropped unnecessary columns like, “Prospect ID”, “Lead Number”, “Last Notable Activity”.
- Replaced null values with mode/zero.

# EDA Analysis

- Most of the students learn about courses through online sources.
- 'Total visits' and 'page views per visit' have outliers but outliers are considered as hot leads as they are visiting website a lot.
- Many of them are looking for better career opportunities in finance management

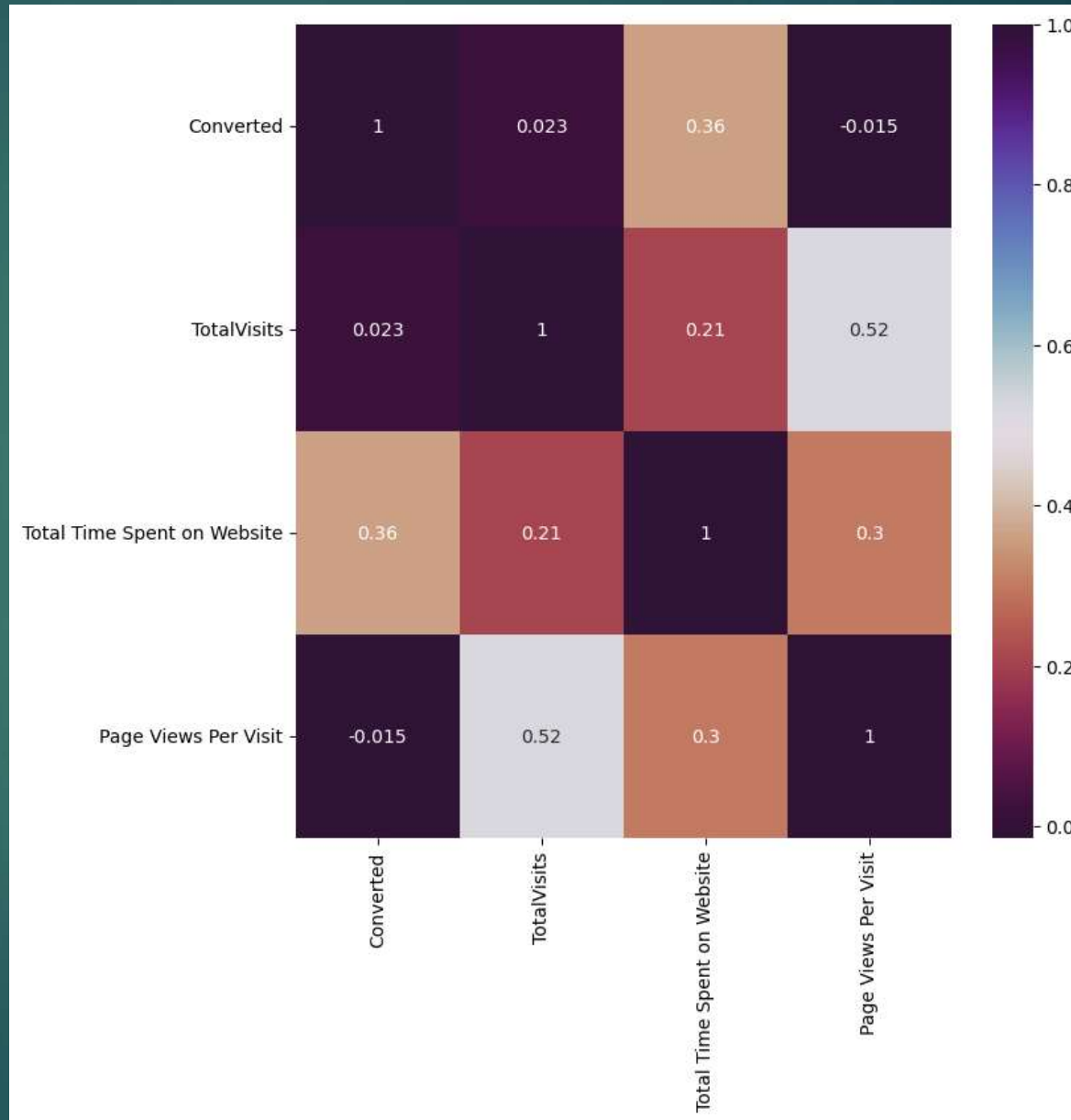




- 
- Most of the count of current lead comes from landing page submission and then API.
  - Maximum count of lead sources come in from Google and then Direct Traffic.
  - Current Occupation of most of the students are “Unemployed”.
  - Students choose courses based on better career prospects while choosing a course.



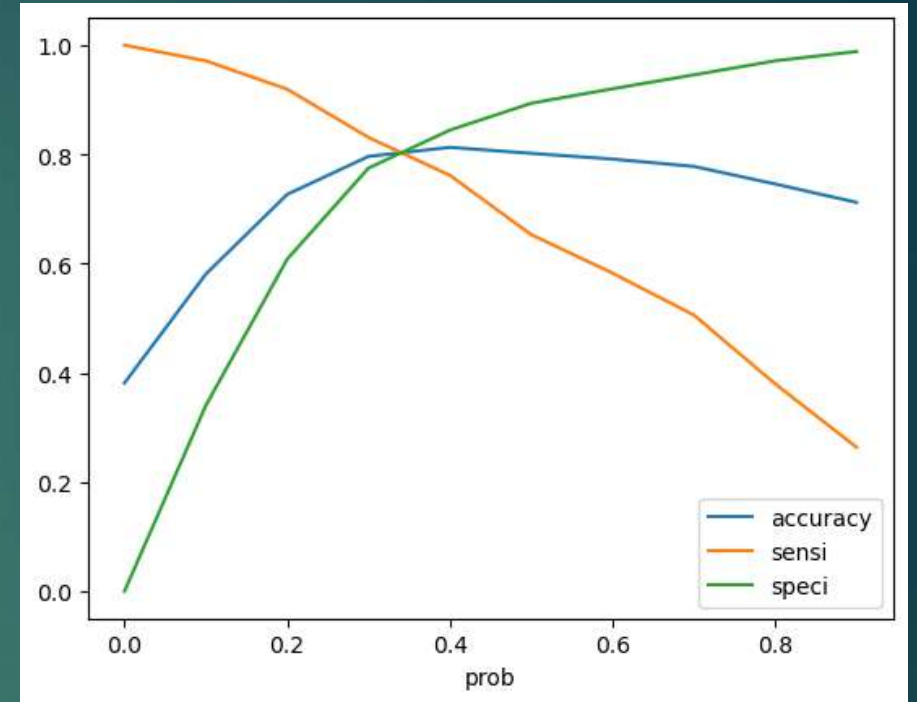
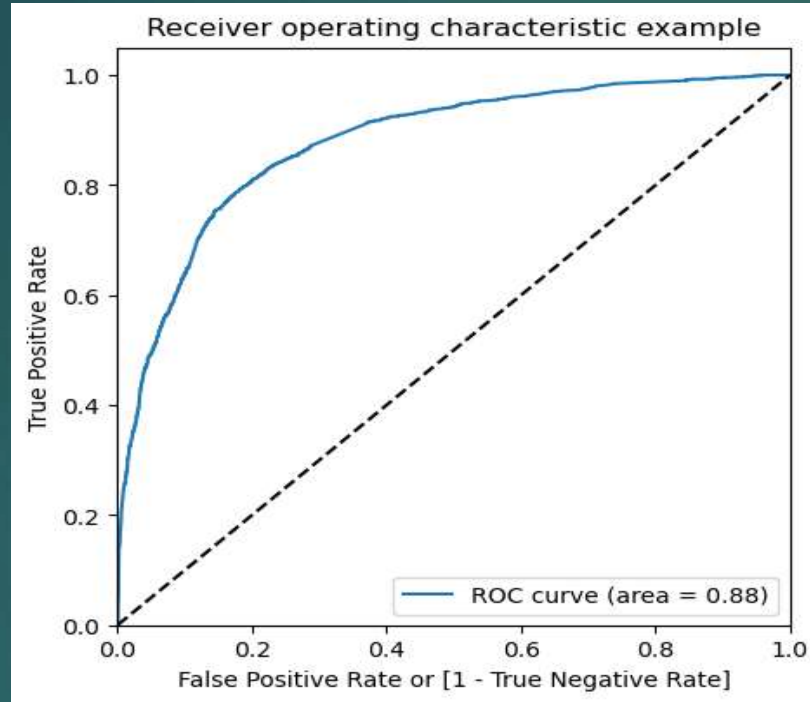
Total visits have positive correlation with Page views per visit and Total time spend have positive correlation with Converted.



# 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 variables, “Newspaper”, ”Lead Source\_google”, “What is your current occupation\_Housewife”.
- Predictions on test data set
- Overall accuracy 81%

# ROC Curve



- Finding Optimal Cut off Point
- Optimal cut off probability is that probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut off is at 0.35.

# Conclusion

It was found that the variables that mattered the most in the potential buyers are (In descending order) :

- The total time spend on the Website.
- Total number of visits.
- When the lead source was:
  - a. Google
  - b. Direct traffic
  - c. Organic search
  - d. Welingak website
- When the last activity was:
  - a. SMS
  - b. Olark chat conversation
- When the lead origin is Landing Page Submission.

*Thank You.*

