# Lead Score Case study

## Group Members

1. Akanksha
2. Alen
3. Suyash

# Table of Contents

# Background of X Education Company

- 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.
- 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%.

# Problem Statement & Objective of the Study

- **Problem Statement:**
- ●X Education gets a lot of leads, its lead conversion rate is very poor at around 30%
- X Education wants to make lead conversion process more efficient by identifying the most potential
- leads, also known as Hot Leads
- Their sales team want to know these potential set of leads, which they will be focusing more on
- communicating rather than making calls to

- **Objective of the Study:**
- To help X Education select the most promising leads, i.e., the leads that are most likely to convert into
- paying customers.
- The company requires us to build a model wherein we need to assign a lead score to each of the leads
- such that the customers with a higher lead score have a higher conversion chance and the customers
- with a lower lead score have a lower conversion chance.
- The CEO has given a ballpark of the target lead conversion rate to be around 80%.

## Suggested Ideas for Lead Conversion

**Leads grouping**

.Leads grouping based on their propensity or likelihood to convert .this result in s focus group of hot leads

**Better Communication**

.we colud have to communicate with,which allow us to have a greater impact.

**Boost Conversion**

.We would have a greater conversion rate and be able to hit the 80% objective since we concentrated on hot leads that were more likely to that were more likely to convert

Since we have a target of 80% conversion rate, we would want to obtain a
high **sensitivity** in obtaining hot leads

# Analysis Approach

**Data Cleaning:** Loading Data Set, understanding & cleaning data

**EDA:** Check imbalance, Univariate & Bivariate analysis

**Data Preparation** Dummy variables, test-train split, feature scaling

**Building:** RFE for top 15 feature, Manual Feature Reduction & finalizing

**Model Evaluation:** Confusion matrix, Cutoff Selection, assigning Lead Score

**Predictions on Test Data:** Compare train vs test metrics, Assign Lead Score and get top features

**Recommendation:** Suggest top 3 features to focus for higher conversion & areas for improvement

# Data Cleaning

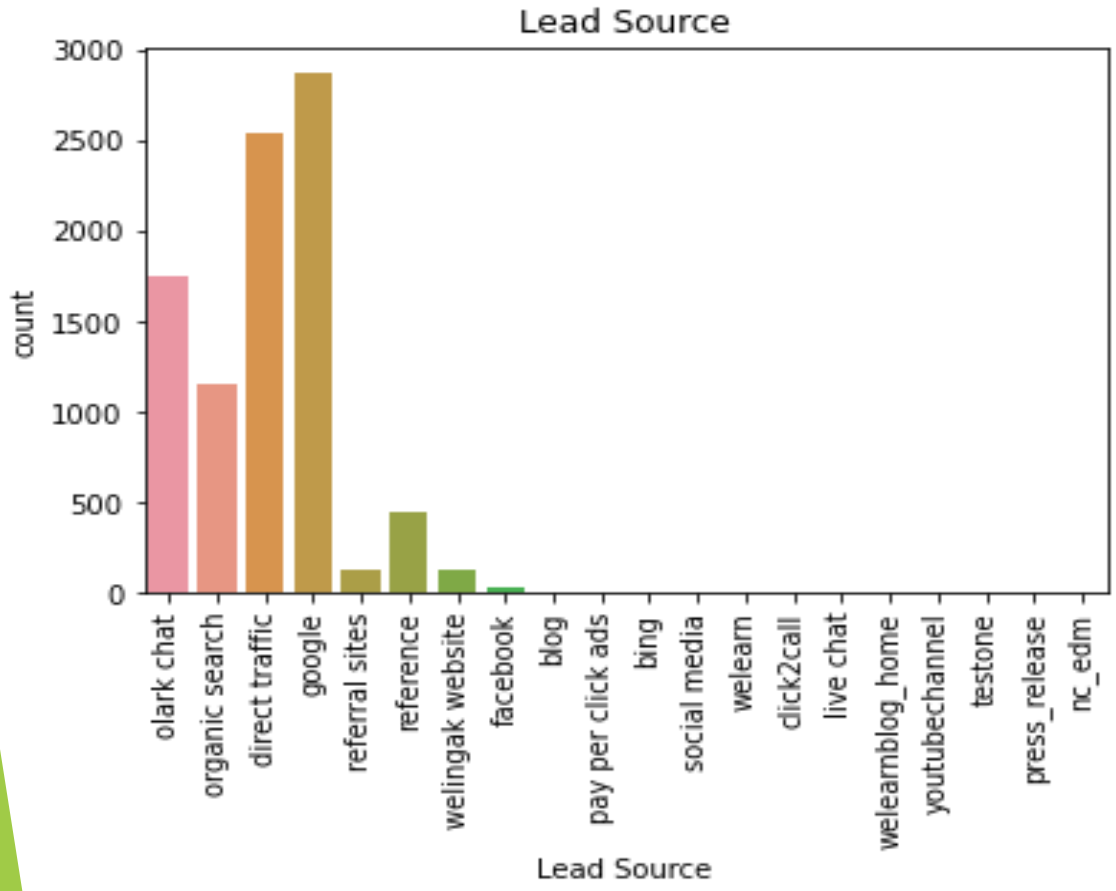
- **"Select"** level represents null values for some categorical variables, as customers did not choose any option from the list.
- Columns with over 40% null values were dropped.
- Missing values in categorical columns were handled based on value counts and certain considerations.
- Drop columns that don't add any insight or value to the study objective (tags, country)
- Imputation was used for some categorical variables.
- Additional categories were created for some variables.
- Columns with no use for modeling (Prospect ID, Lead Number) or only one category of response were dropped.
- Numerical data was imputed with mode after checking distribution

# Data Cleaning

- Skewed category columns were checked and dropped to avoid bias in logistic regression models.
- Outliers in **TotalVisits** and **Page Views Per Visit** were treated and capped.
- Invalid values were fixed and data was standardized in some columns, such as lead source.
- Low frequency values were grouped together to “Others”.
- Binary categorical variables were mapped.
- Other cleaning activities were performed to ensure data quality and accuracy.
  - Fixed Invalid values & Standardizing Data in columns by checking casing styles, etc. (lead
source has Google, google

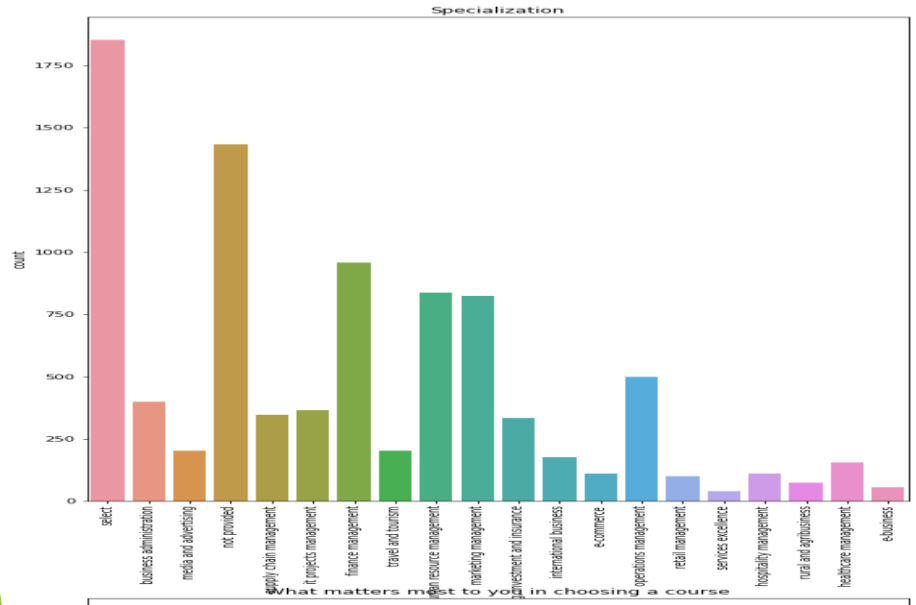
# EDA

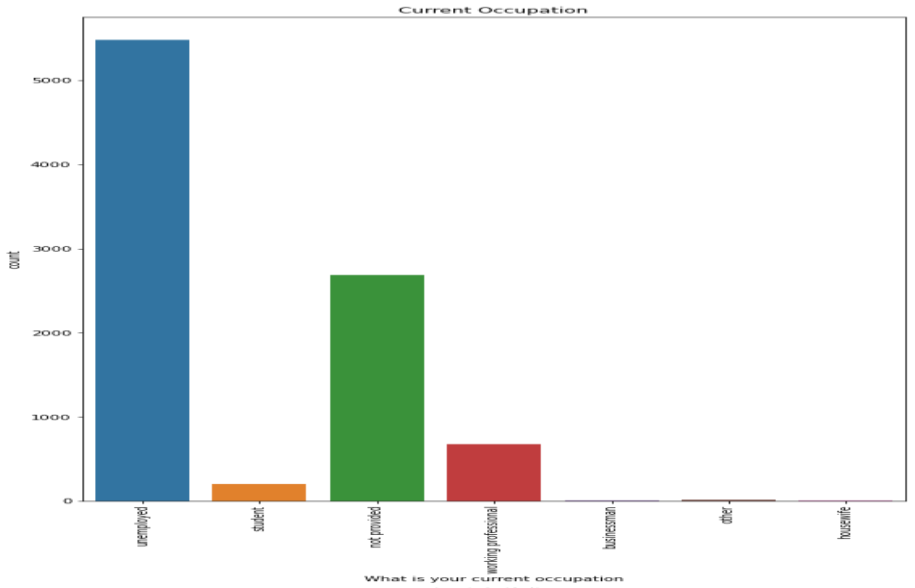
- Univariate Analysis – Categorical Variables

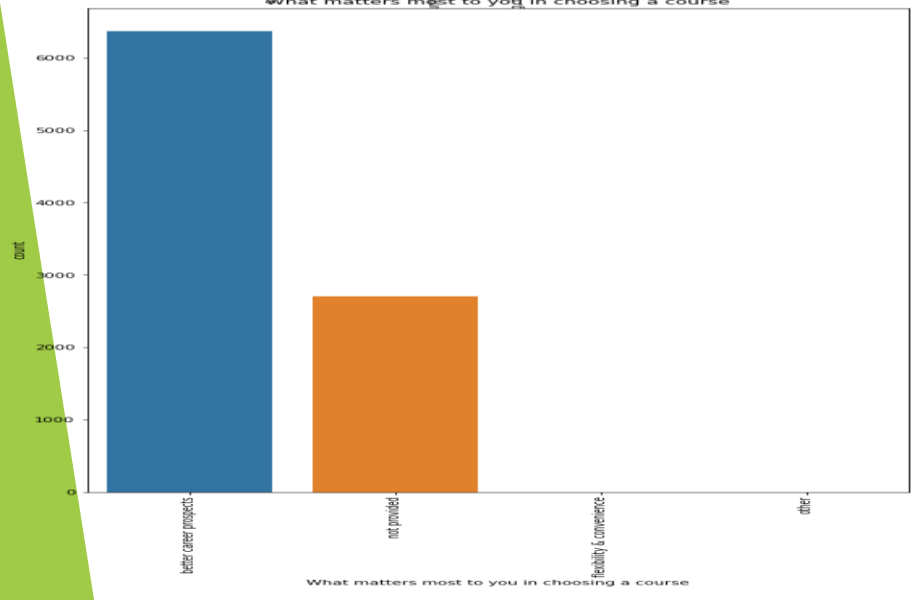


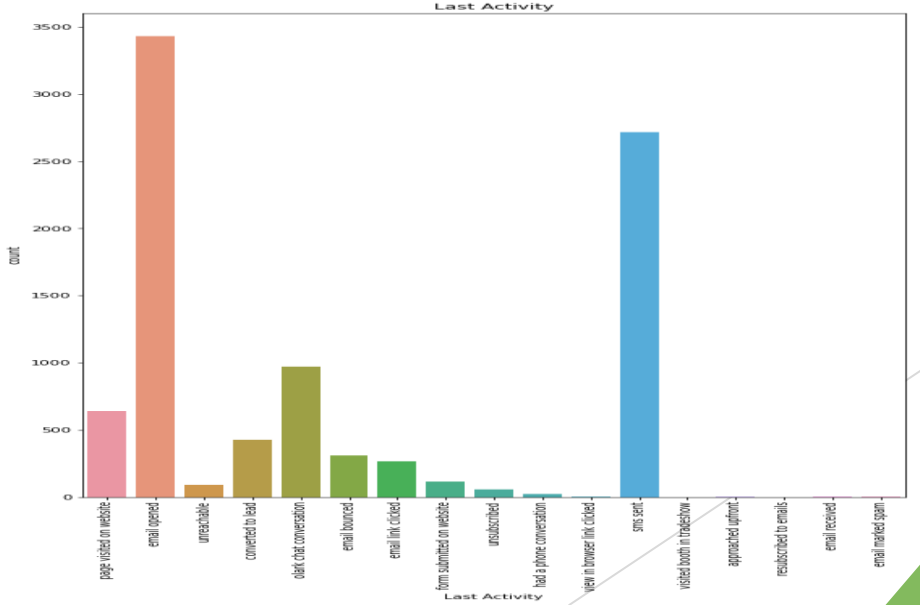
# EDA

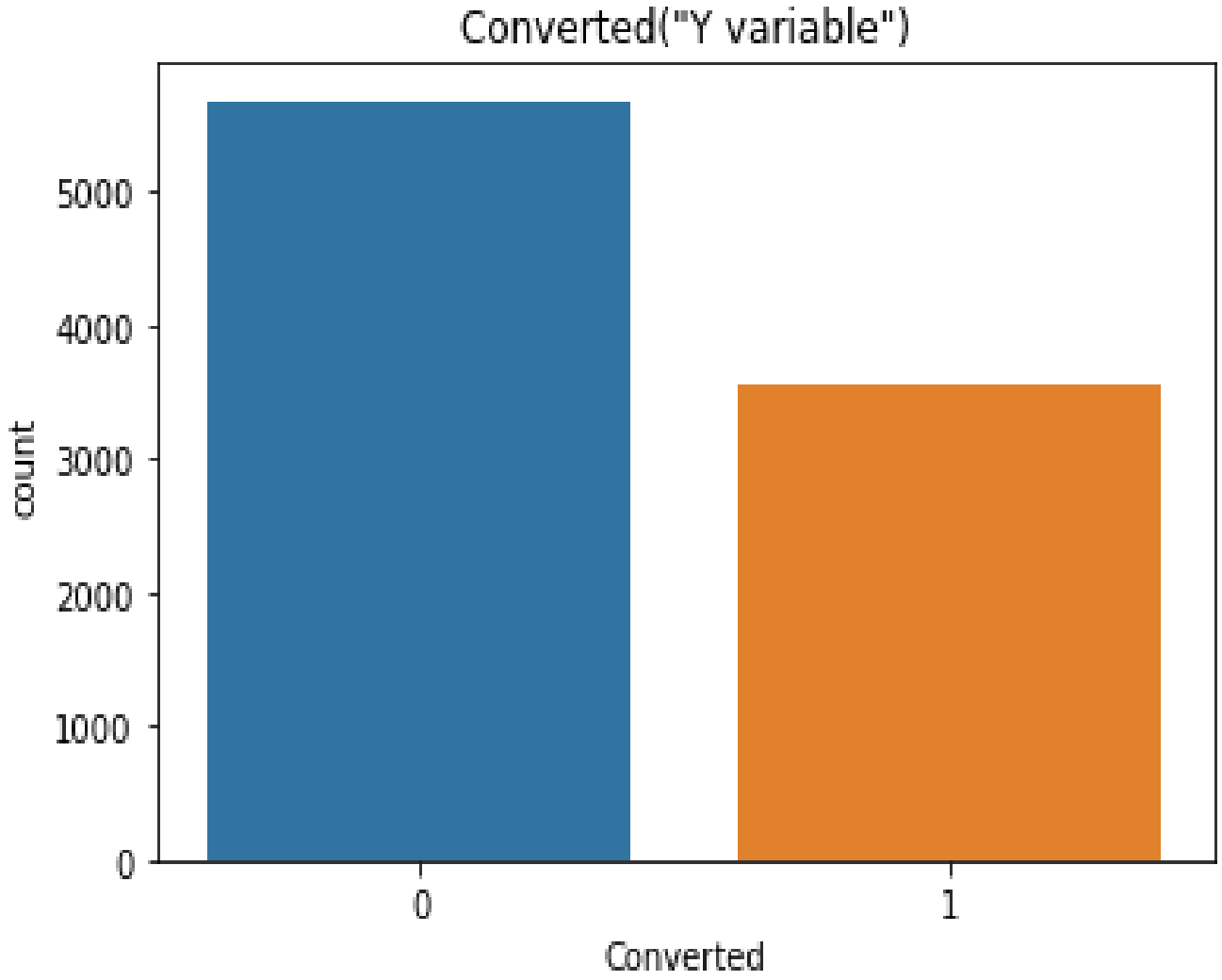
- Univariate Analysis – Categorical Variables

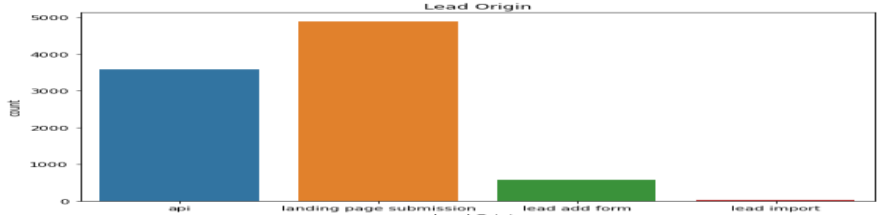


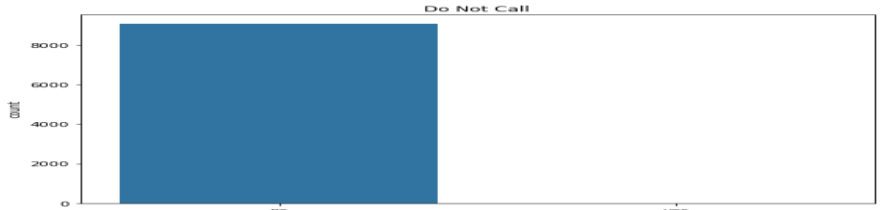



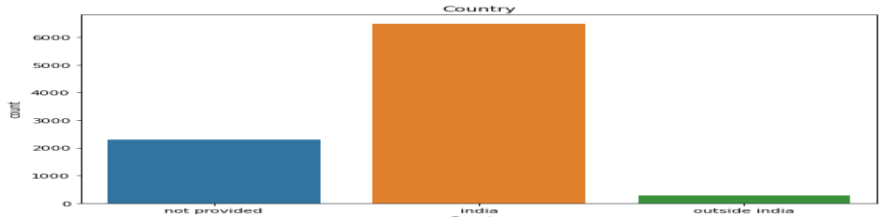



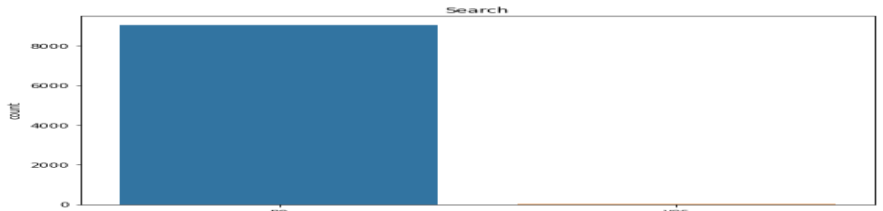

Converted("Y variable")
count
5000
4000
3000
2000
1000
0
0
1
Converted

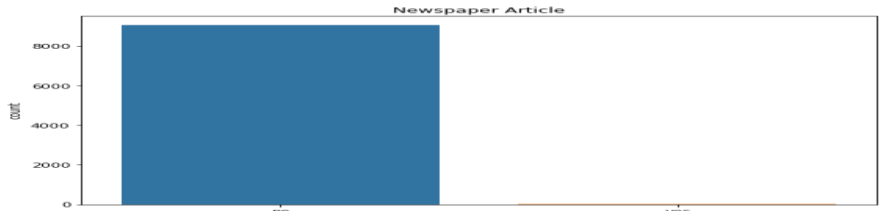

Lead Origin
count
api
landing page submission
lead add form
lead import
Lead Origin

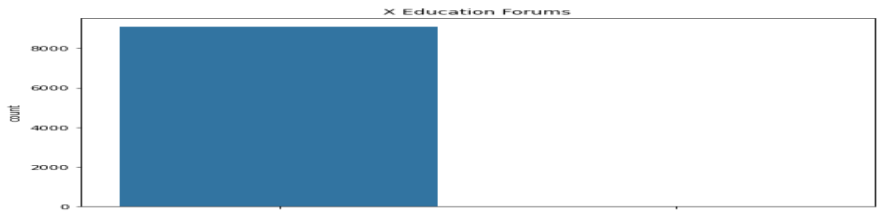

Do Not Email
count
no
yes
Do Not Email

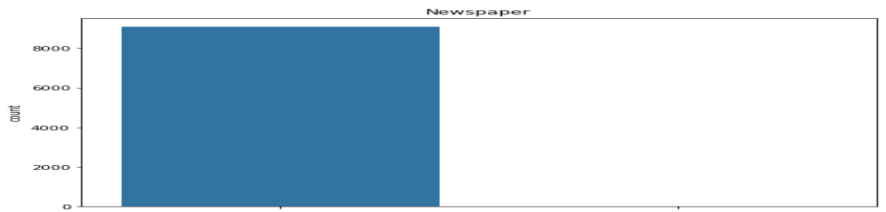

Do Not Call
count
no
yes
Do Not Call

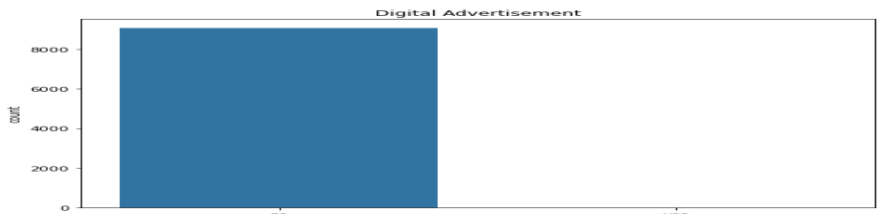

Country
count
not provided
India
outside India
Country

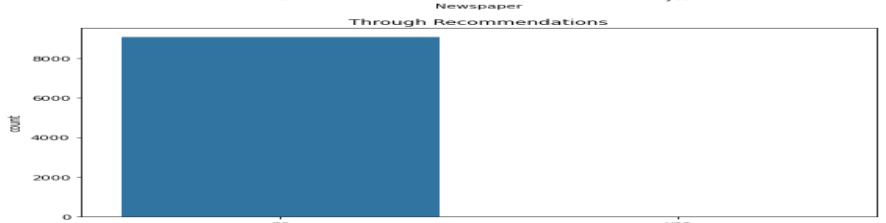

Search
count
no
yes
Search


Newspaper Article
count
no
yes
Newspaper Article

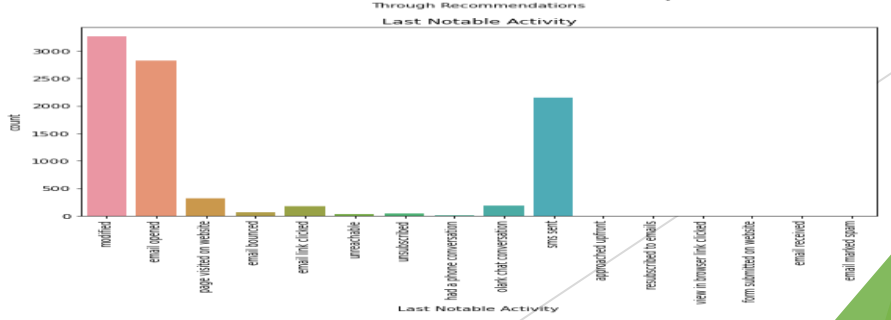

X Education Forums
count
no
yes
X Education Forums


Newspaper
count
no
yes
Newspaper

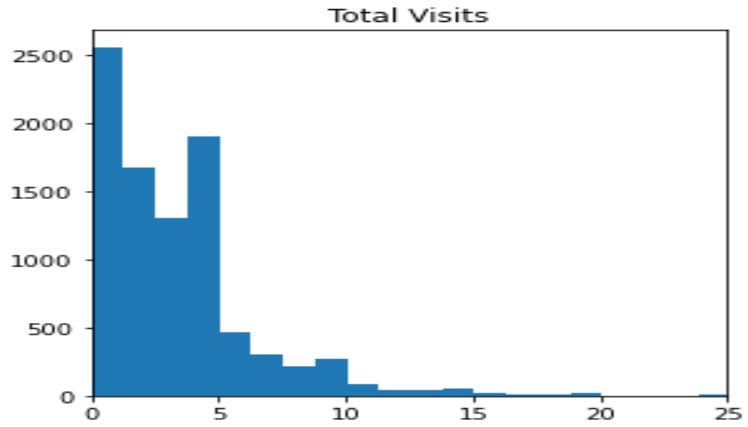

Digital Advertisement
count
no
yes
Digital Advertisement

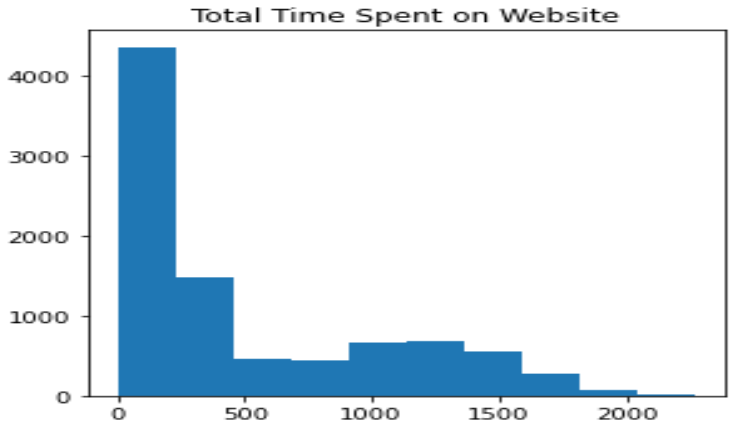

Through Recommendations
count
no
yes
Through Recommendations

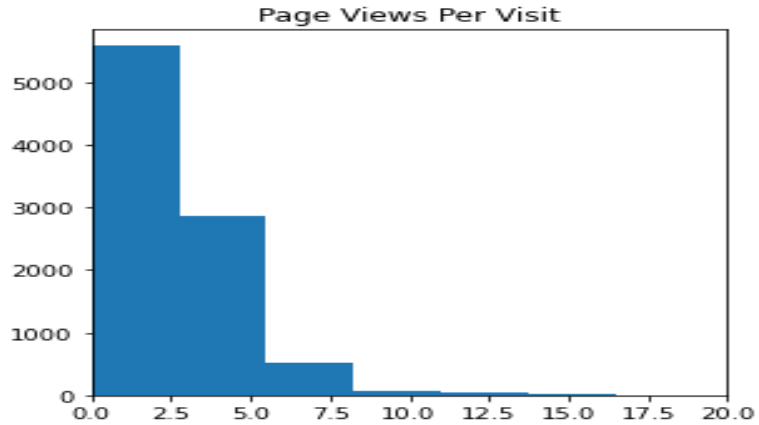

A free copy of Mastering The Interview
count
no
yes
A free copy of Mastering The Interview


Last Notable Activity
count
modified
email opened
page visited on website
email bounced
email link clicked
unreachable
unsubscribed
had a phone conversation
olark chat conversation
sms sent
approached upfront
resubscribed to emails
view in browser link clicked
form submitted on website
email received
email marked spam
Last Notable Activity

# Numerical variables



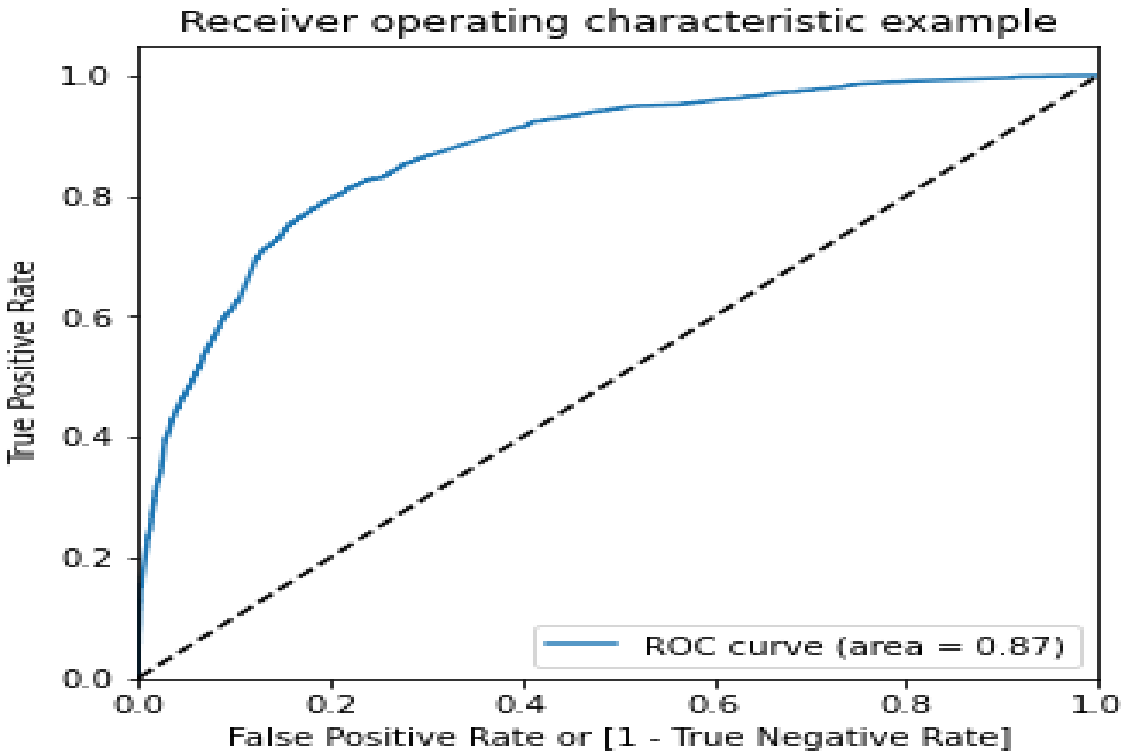



# Model Building

- 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 test data set
- Overall accuracy 81%

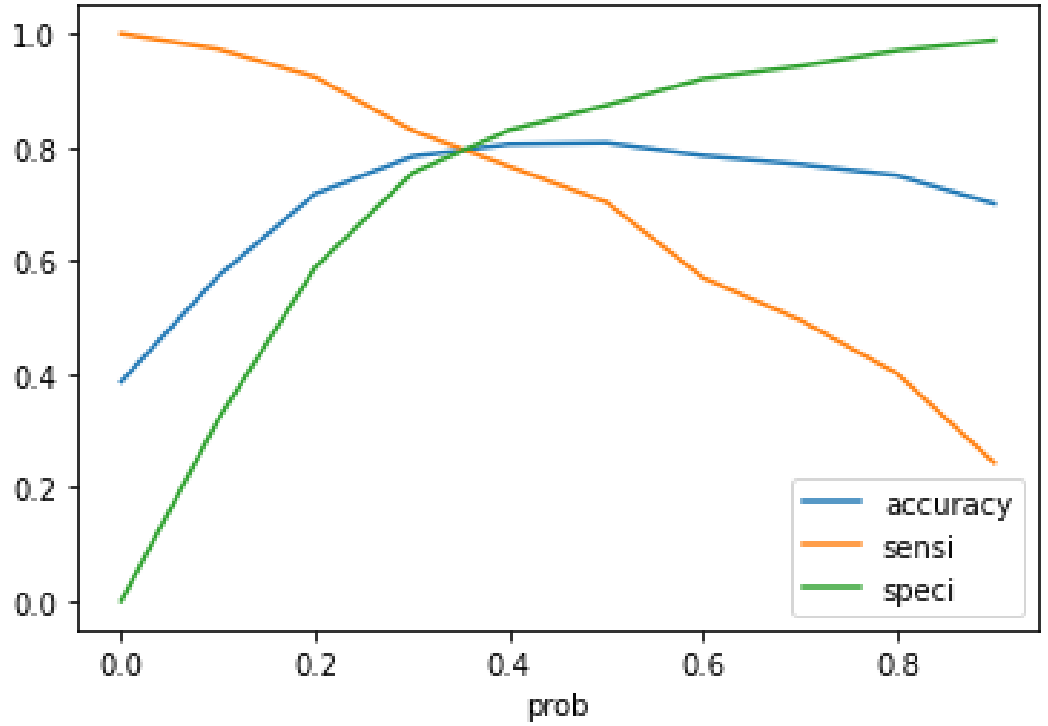
# Optimise Cut off (ROC Curve)



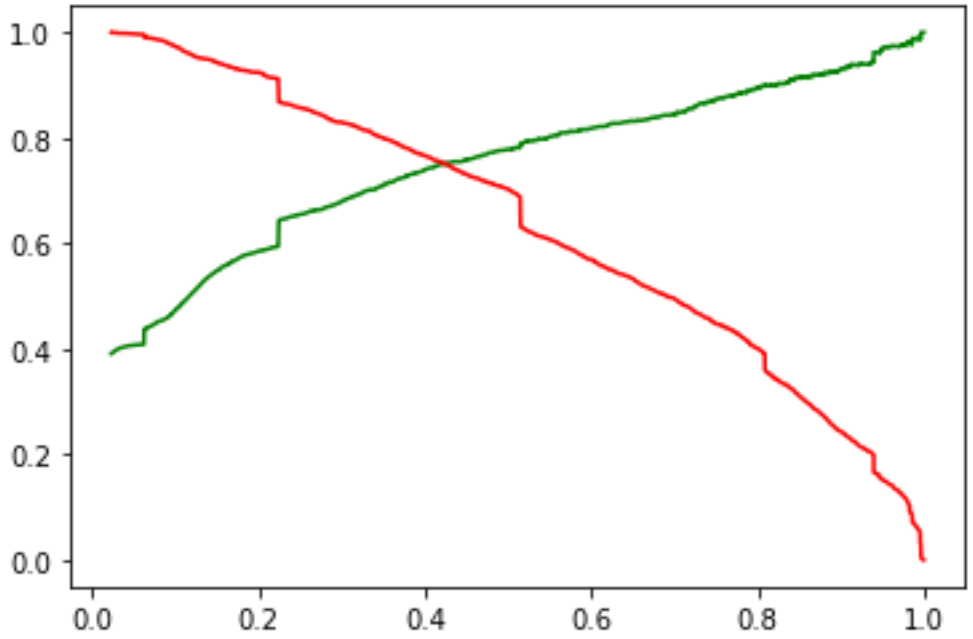

**The area under ROC curve is 0.87 which is a very good value.**

# Finding optimoff pointal cut

1.optimal cut off probability is that
2.probability where we get balanced sensitivity and specificity
3.from the second graph it is visible that the optimal cut off is at 0.35



# Precision and recall tradeoff



# 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 Lead add format.
- When their current occupation is as a working professional.
  Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.

# Assigning Lead Score

**So there are 979 leads which can be contacted and have a high chance of getting converted.¶**

## Recommendations:
- The company should make calls to the leads coming from the lead sources "Total Visits" and "Total Time Spent on Website " as these are more likely to get converted.
- company should make calls to the leads who are the "working professionals" as they are more likely to get converted.
- should make calls to the leads coming from the lead sources "google" as these are more likely to get converted.
- should not make calls to the leads whose last activity was "Olark Chat Conversation" as they are not likely to get converted.
- should not make calls to the leads who chose the option of "Do not Email" as "yes" as they are not likely to get converted.