

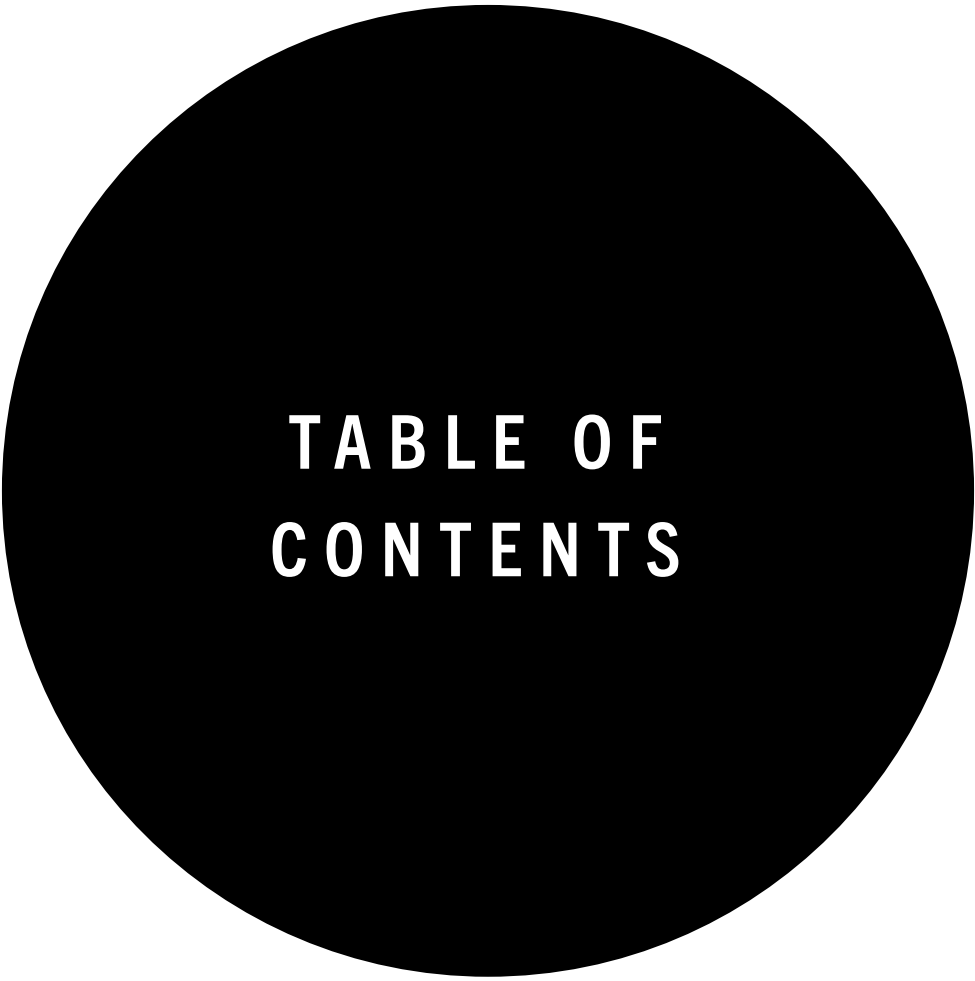
# LEAD SCORING CASE STUDY

---

Detection of Hot Leads to concentrate more of marketing efforts on them,  
improving conversion rates for an Education Company.

Team — Adari Saikumar, Vishal Singh, Saloni Shah





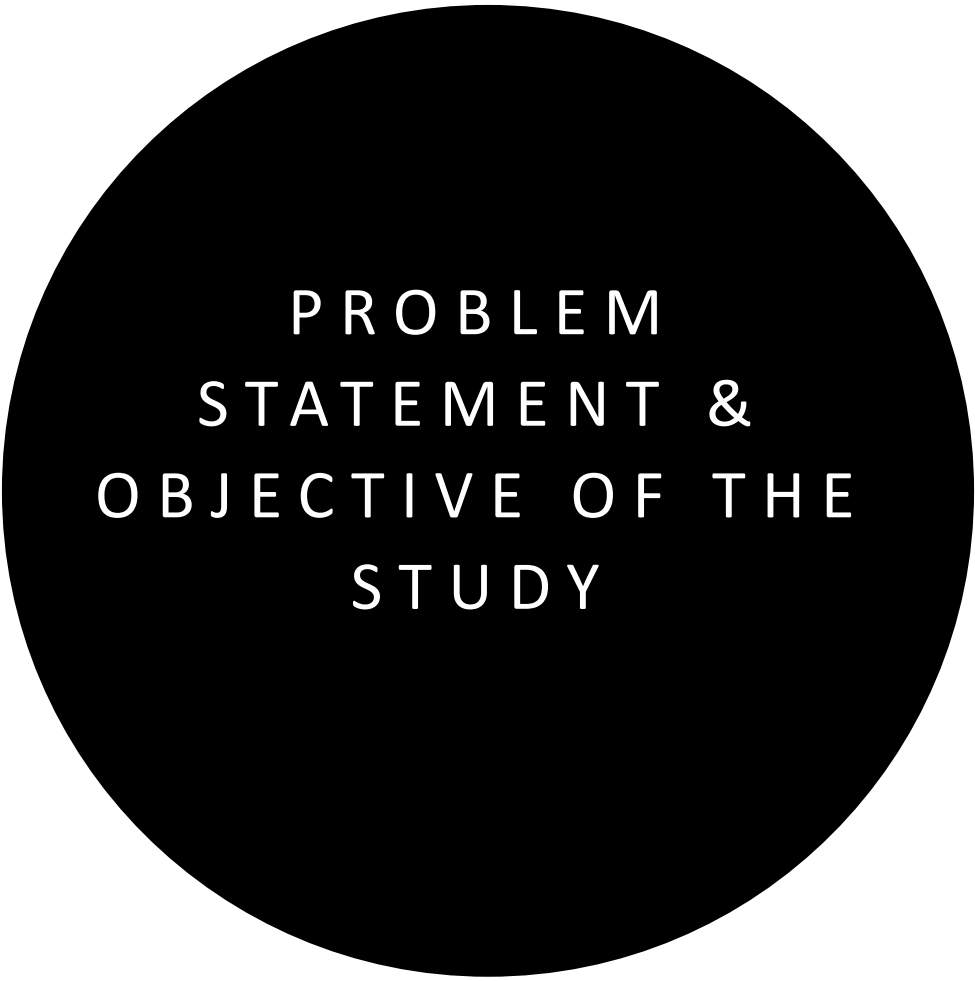
# TABLE OF CONTENTS

- Background of Education Company
- Problem Statement & Objective of the Study
- Suggested Ideas for Lead Conversion
- Analysis Approach
- Data Cleaning
- EDA
- Data Preparation
- Model Building (RFE & Manual fine tuning)
- Model Evaluation
- Recommendations



# BACKGROUND OF EDUCATION COMPANY

- An education company named 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

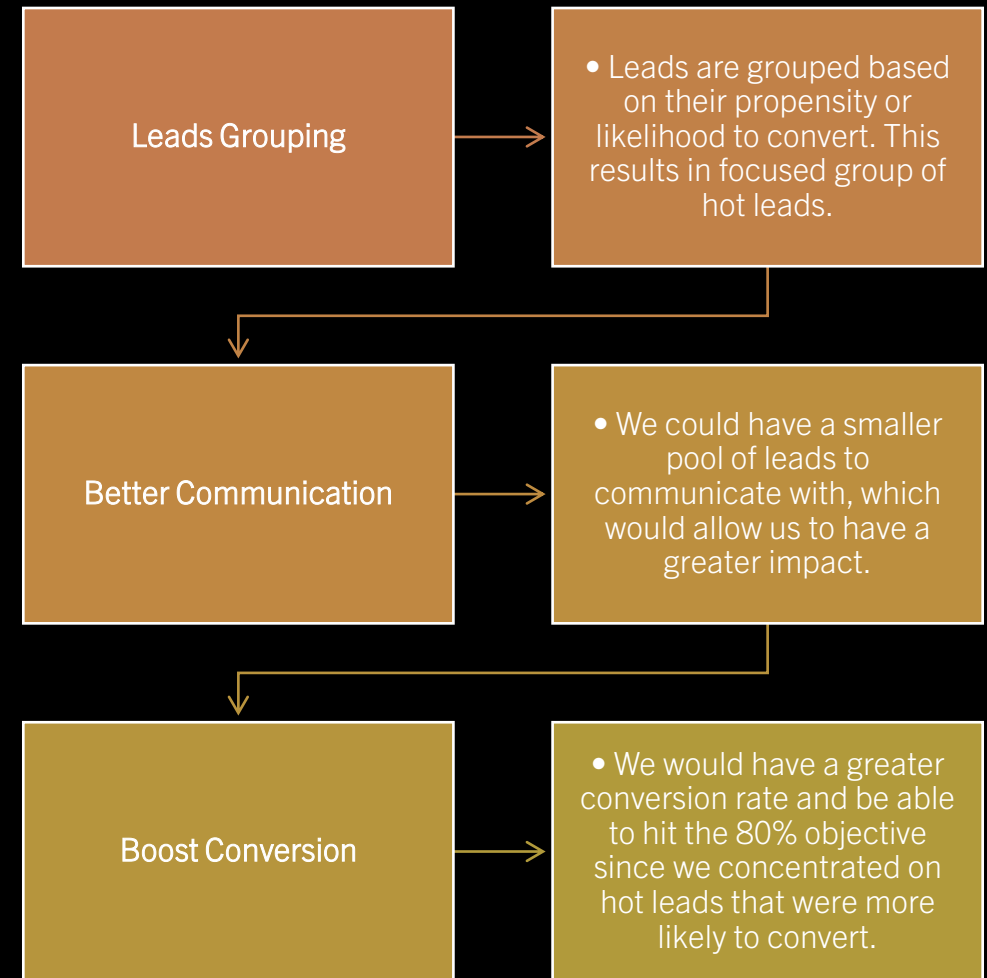
## **Problem Statement:**

- 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 everyone.

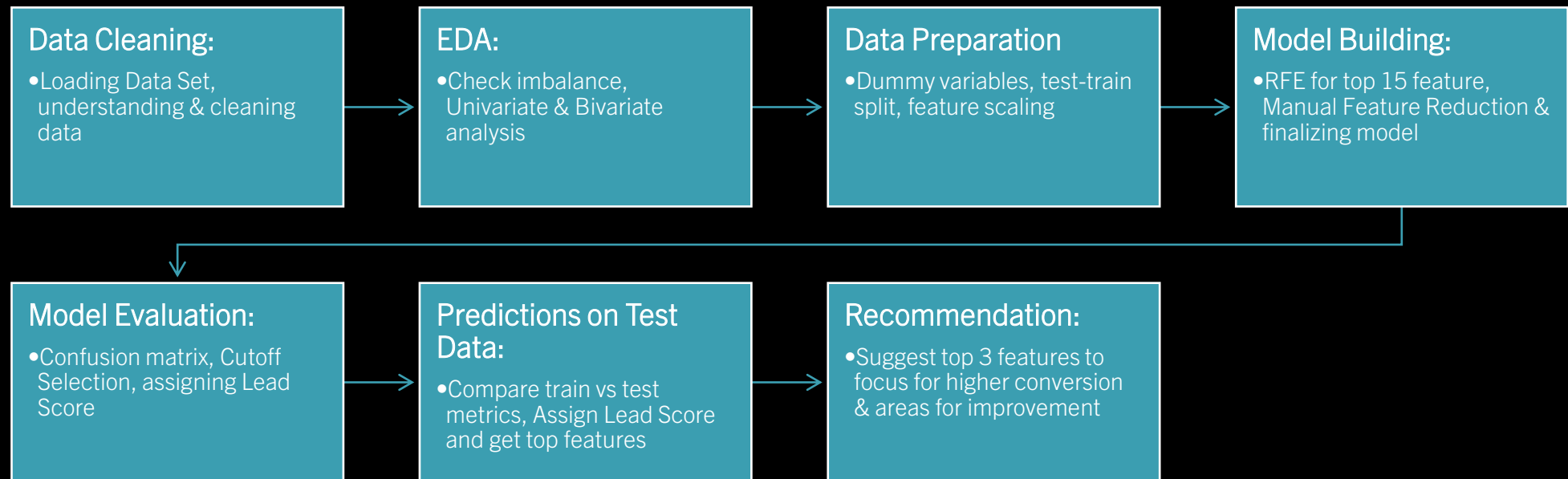
## **Objective of the Study:**

- To help Education company 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



# ANALYSIS APPROACH





# DATA CLEANING

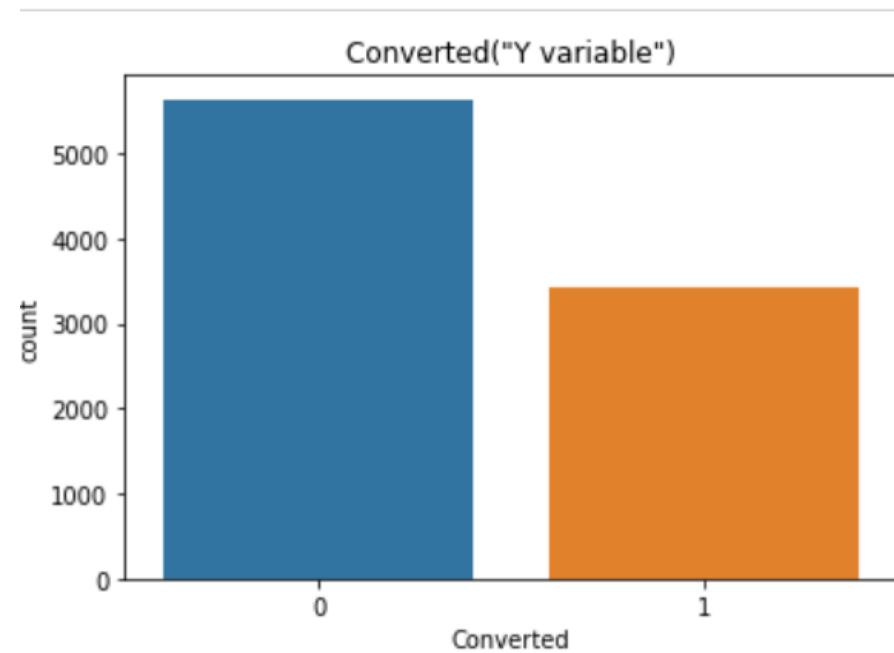
- **"Select"** level represents null values for some categorical variables, as customers did not choose any option from the list.
- Columns with over 40% null values were dropped.
- Missing values in categorical columns were handled based on value counts and certain considerations.
- Drop columns that 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.
- 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.

# EDA

1 - Lead converted

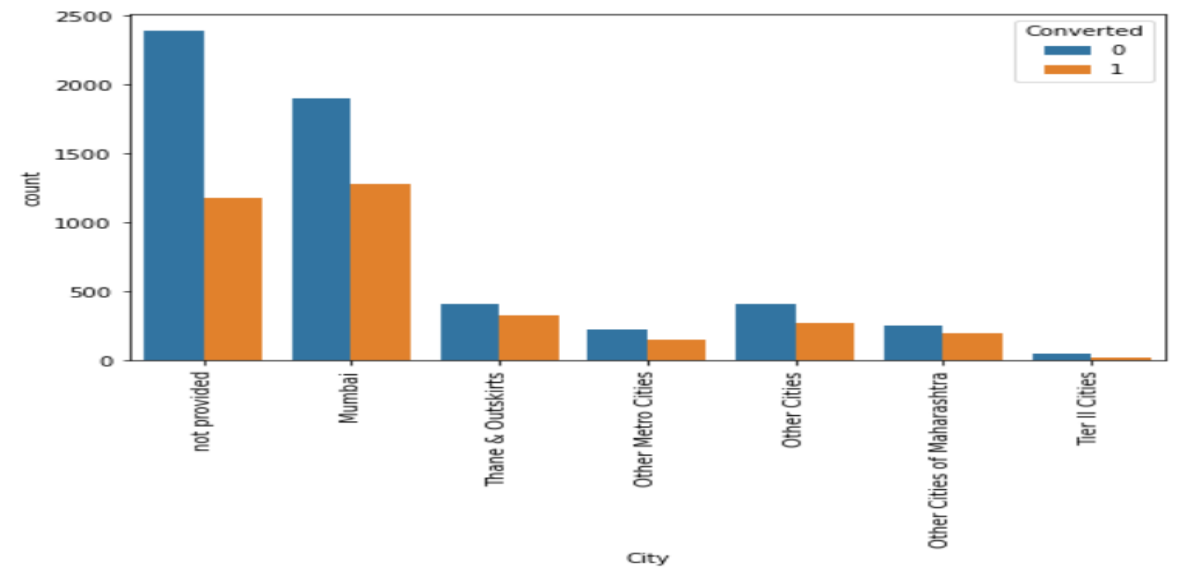
0 - Lead not converted

We can clearly see that, converted rate is quite less

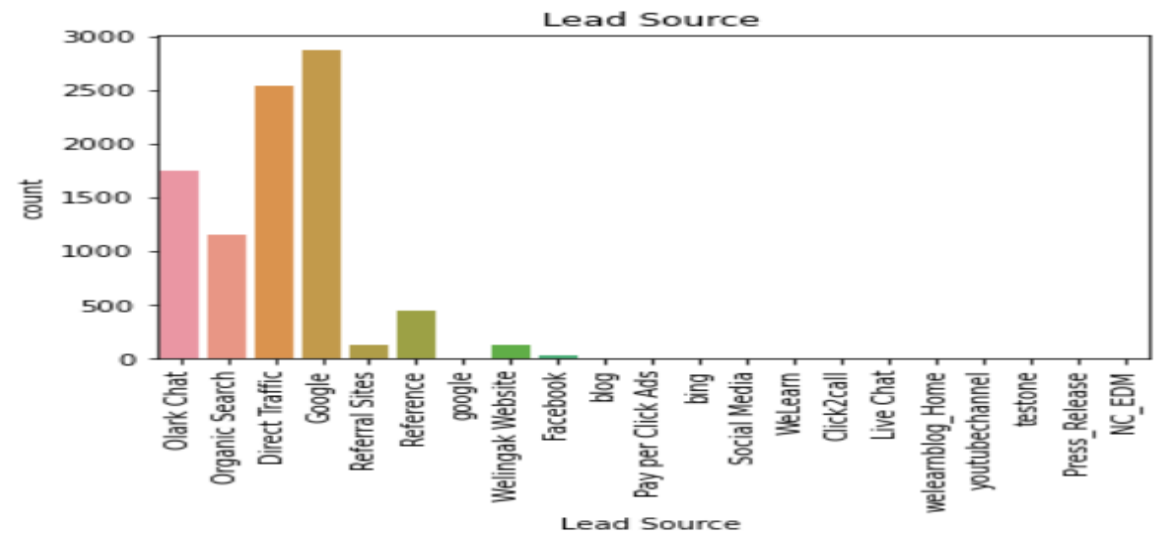




# UNIVARIATE ANALYSIS – CATEGORICAL VARIABLES

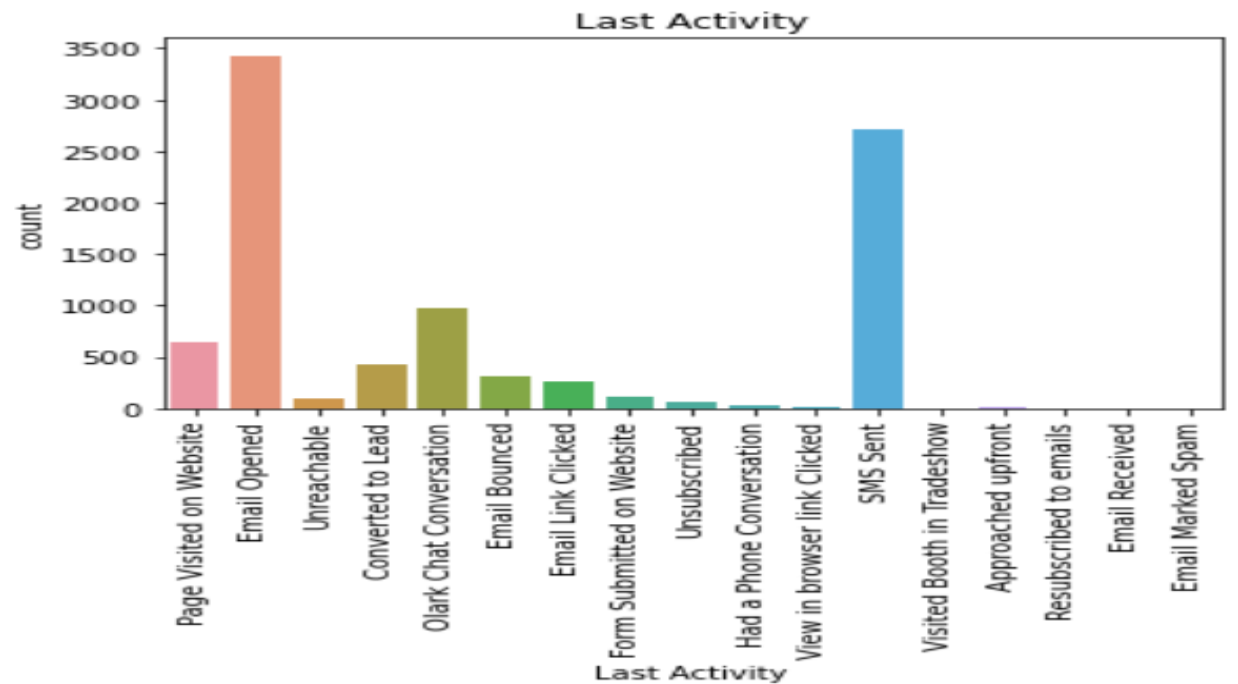
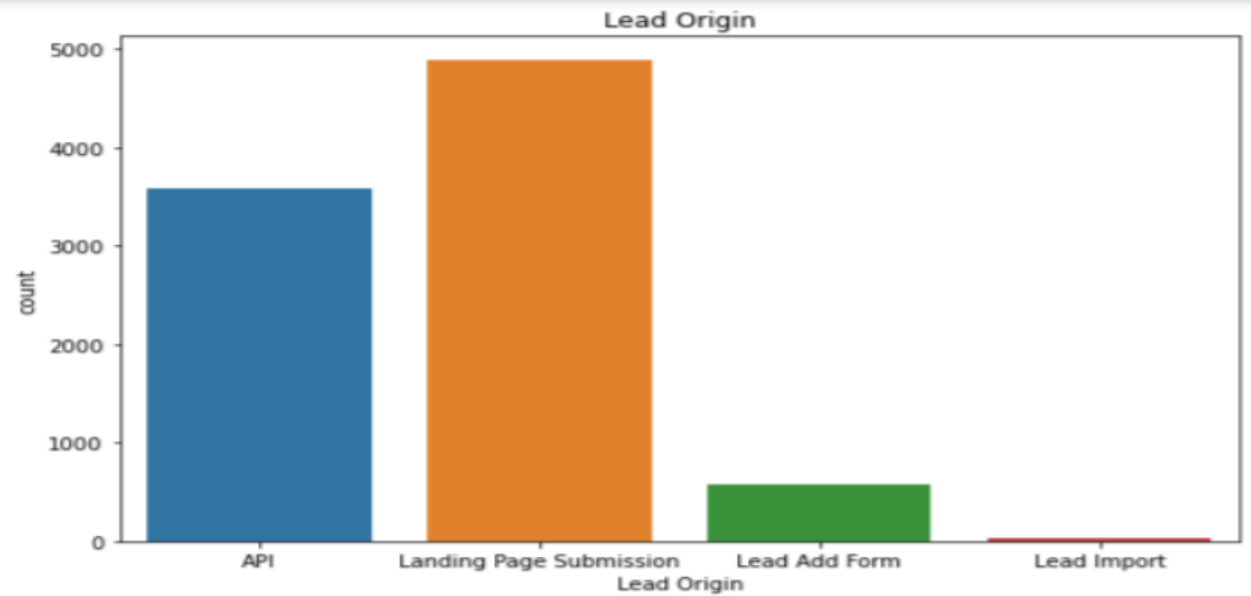


Most converted leads are from Mumbai



Lead source is from Google & Direct Traffic combined.

# UNIVARIATE ANALYSIS – CATEGORICAL VARIABLES







# DATA PREPARATION

- **Binary level categorical columns were already mapped to 1 / 0 in previous steps**
- **Created dummy features (one-hot encoded) for categorical variables – Lead Origin, Lead Source, Last Activity, Specialization, Current\_occupation**
- **Splitting Train & Test Sets** - 70:30 % ratio was chosen for the split
- **Feature scaling** - Standardization method was used to scale the features
- **Checking the correlations** - Predictor variables which were highly correlated with each other were dropped.



# MODEL BUILDING

## Feature Selection

- The data set has lots of dimension and large number of features.
- This will reduce model performance and might take high computation time.
- Hence it is important to perform **Recursive Feature Elimination** (RFE) and to select only the
- important columns.
- Then we can manually fine tune the model.
- Manual Feature Reduction process was used to build models by dropping variables with p – value greater than 0.05.
- Model 3 looks stable after 3 iteration with significant p-values within the threshold (p-values < 0.05) and
- No sign of multicollinearity with VIFs less than 5
- Hence, model 3 will be our final model, and we will use it for Model Evaluation which further will be used to make predictions.

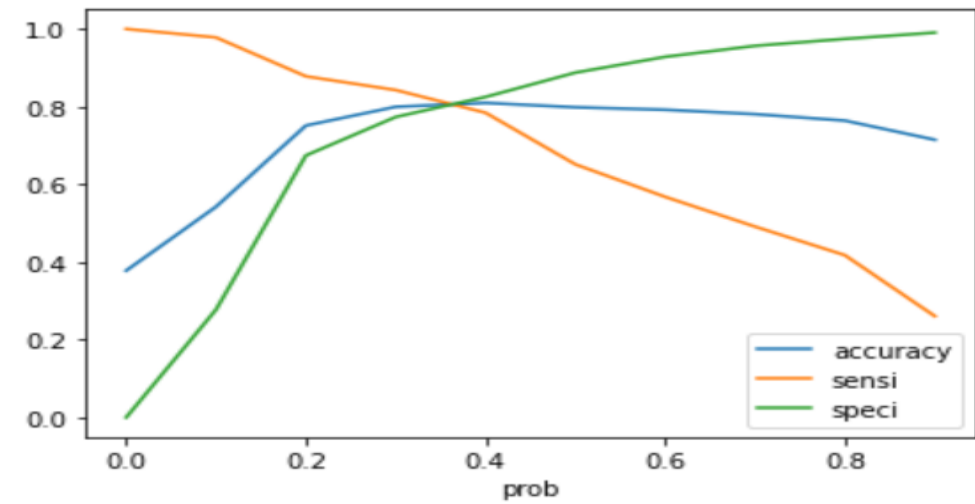
# MODEL EVALUATION

- Confusion Matrix:

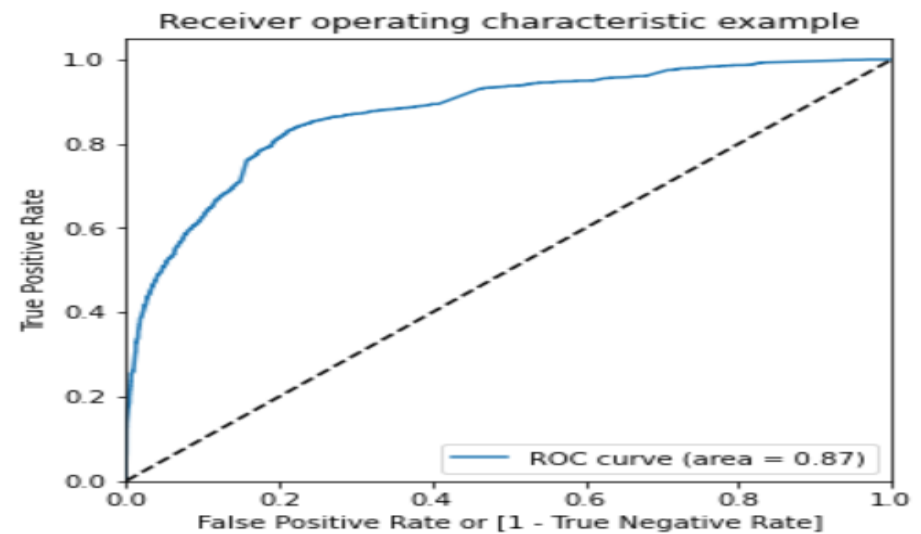
```
: # Creating confusion matrix
confusion = metrics.confusion_matrix(y_train_pred_final.Converted, y_train_pred_final.Predicted )
confusion

: array([[3475,  440],
        [ 829, 1549]], dtype=int64)
```

# ROC CURVE



From the graph it is visible that the optimal cut off is at 0.35.



The area under ROC curve is 0.88 which is a very good value



# RECOMMENDATIONS

- As per the problem statement, increasing lead conversion is crucial for the growth and success of Education company. To achieve this, we have developed a regression model that can help us identify the most significant factors that impact lead conversion.
- We have determined the following features that have the highest positive coefficients, and these features should be given priority in our marketing and sales efforts to increase lead conversion.
  - Lead Source\_Welingak Website: 5.39
  - Lead Source\_Reference: 2.93
  - Current\_occupation\_Working Professional: 2.67
  - Last Activity\_SMS Sent: 2.05
  - Total Time Spent on Website: 1.05
  - Last Activity\_Email Opened: 0.94
  - Lead Source\_Olark Chat: 0.91
- We have also identified features with negative coefficients that may indicate potential areas for improvement. These include:
  - Specialization in Hospitality Management: -1.09
  - Specialization in Others: -1.20
  - Lead Origin of Landing Page Submission: -1.26





# RECOMMENDATIONS

## To increase our Lead Conversion Rates:

- Focus on features with positive coefficients for targeted marketing strategies.
- Develop strategies to attract high-quality leads from top-performing lead sources.
- Optimize communication channels based on lead engagement impact.
- Engage **working professionals** with tailored messaging.
- More budget/spend can be done on **Welingak Website** in terms of advertising, etc.
- Incentives/discounts for providing reference that convert to lead, encourage providing more references.
- Working professionals to be aggressively targeted as they have high conversion rate and will
- have better financial situation to pay higher fees too.

## To identify areas of improvement

- Analyze negative coefficients in specialization offerings.
- Review landing page submission process for areas of improvement.

Thank you

