# Lead Score Case Study

Sindhu L
Shreevatsa Hegde
Shilpa Kamath

# Table of Contents

- Problem Statement
- Our Approach
- Data Cleaning
- Exploratory Data Analysis (EDA)
- Model Evaluation
- Results
- Recommendations
- Conclusion

# Problem Statement

- An education company named X Education sells online courses to industry professionals
- The company uses several marketing strategies for acquiring leads for their courses; Through their process, some of the leads get converted while most do not
- The typical lead conversion rate at X education **is around 30%**
- For improving their conversions they want to identify potential leads called "**HOT LEADS**"

# Business Objective

- Identify most promising leads, i.e, HOT LEADS
- Build a model to assign **Lead Score** to each leads; Higher the score ,higher the chance of conversion
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around **80%**.

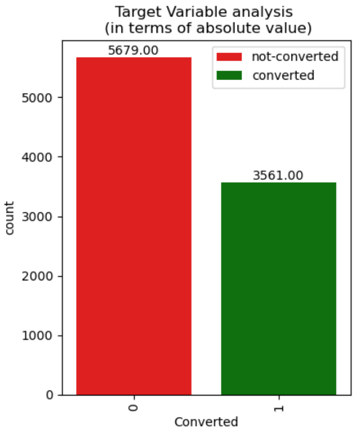
# Our Approach

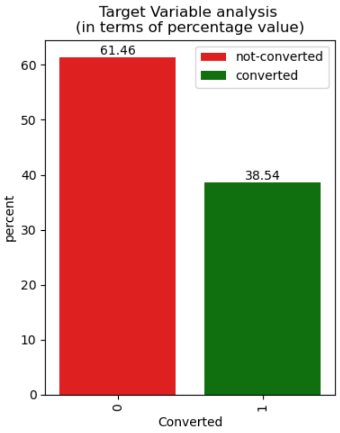
- Our Objective is to find whether a candidate is a potential lead or not; In this case our outcome is two i.e. yes/no ; So we are going with classification technique called **Logistic Regression**
- Data Understanding
- Data Cleaning
  - Data Quality Check
  - Handling Missing Values
  - Handling Outliers
- EDA
  - Univariate Analysis
  - Univariate Segmented Analysis
  - Bivariate Analysis
  - Multivariate Analysis
- Data Preparation for Logistic Regression
- Model Building
- Model Evaluation
- Model Prediction
- Lead Score generation

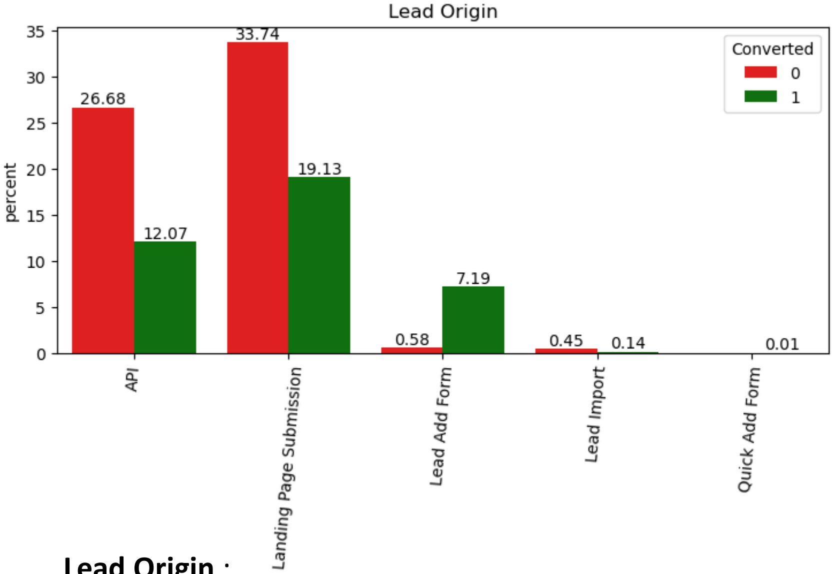
# Data Cleaning

- Given Data Set: Rows:9240,Columns:37
- Dropped Columns
  - With null values >30 %
  - Columns with unique values in all rows i.e. Prospect ID, Lead Number
  - Columns with only one unique categorical value
  - Skewed Columns
- Data Manipulation
  - Clubbed categorical values with lower frequencies in columns
  - Handling Misspelling in columns
  - Converted “Select” categorical value to NULL in columns
- Data Imputation
  - Categorical Columns: Replaced null values with Mode values
  - Numerical Columns: Replaced null values with Median values
- Outlier Handling
  - Capped the outliers above 99 percentile with 99th percentile value
- Data Set After Cleaning and standardization
  - Rows:9240,Columns:12

# Exploratory Data Analysis







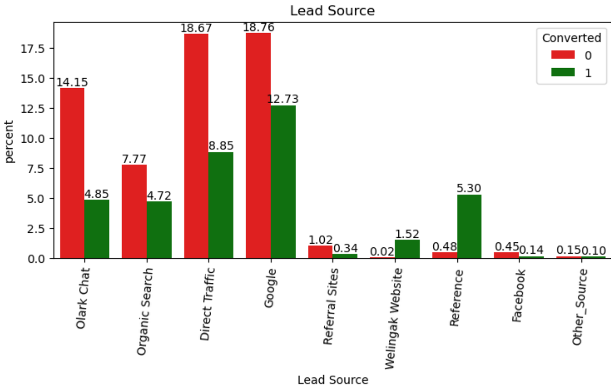
**Target Variable Analysis:**

- Converted – 38.54%
- Not Converted – 61.46%
- Data Imbalance Ratio – 1:1.59

**Lead Origin** :

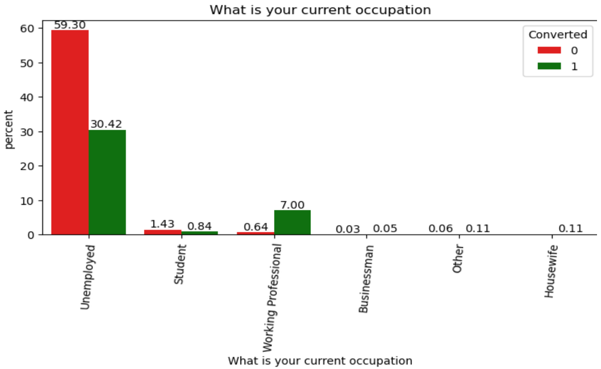
- Lead Add Form : Among this category 92% of them got converted; This could be a good indicator for analysis
- Landing Page Submission : Among this category 36% of them got converted
- API : Among this category 31% of them got converted

# Exploratory Data Analysis



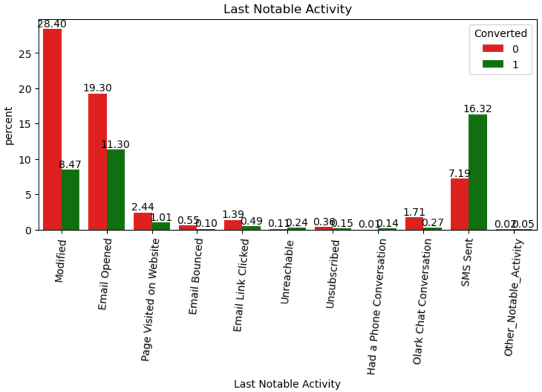
**Lead Source:**

- ➢ Wellingak website : Among this category 98% of them got converted; But it's overall contribution is low
- ➢ Reference : Among this category 92% of them got converted; This could be a good indicator for analysis
- ➢ Google : Among this category 40% of them got converted



**What is your current occupation**:

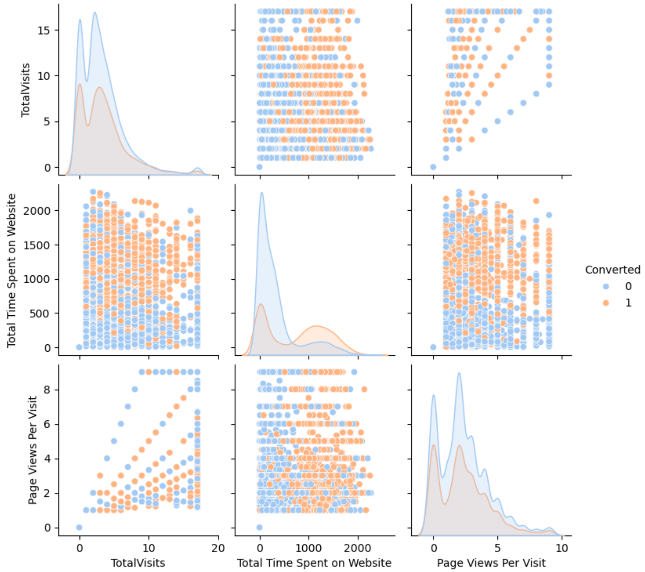
- ➢ Working professional: Among this category 91% of them got converted; This could be a good indicator for analysis
- ➢ Unemployed: Among this category 34% of them got converted
- ➢ Student: Among this category 37% of them got converted

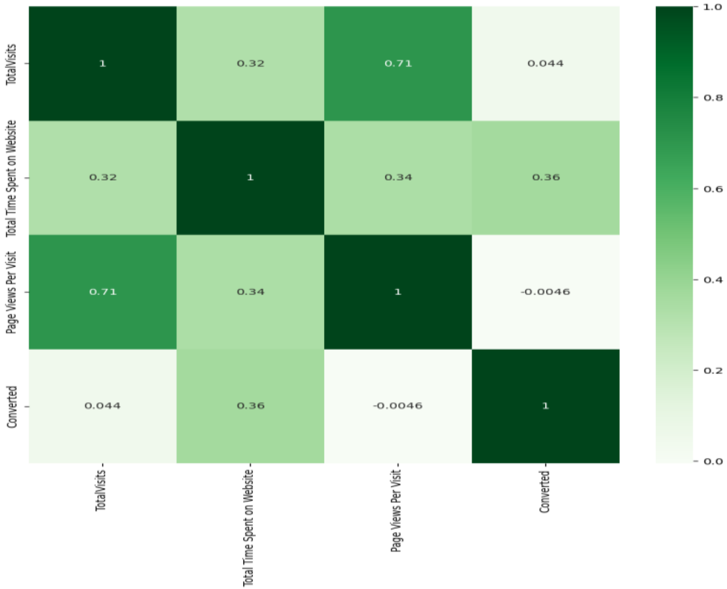


**Last Notable Activity** :

- ➢ SMS-Sent: 69% of people to whom SMS has been sent has been converted ;This could be a potential indicator in our analysis
- ➢ Email Opened- 63% of people who opened mail have been converted ;This could be a potential indicator in our analysis

# Exploratory Data Analysis





- TotalVisits and Page Views Per Visit have positive correlation; Multicollinearity exists between them
- **Total Time Spent on Website** have positive relationship with target variable and it has fairly good relationship (value=0.36); This could be a good indicator for our analysis

# Model Evaluation

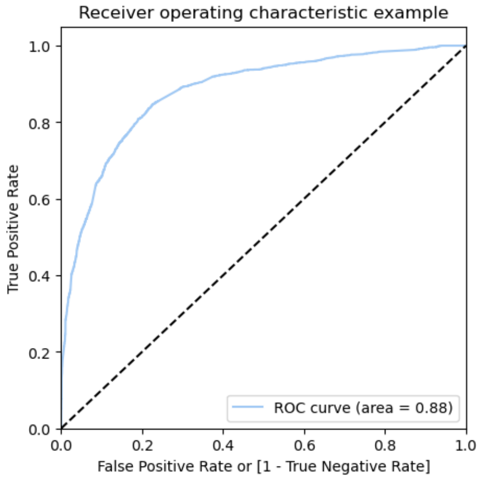
Data regarding Model:

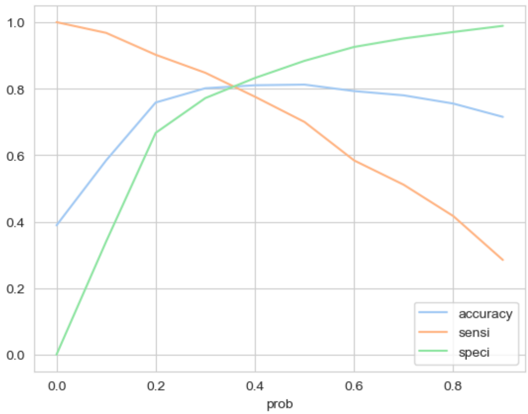
- Train – Test Split: 70:30
- Used Recursive Feature Elimination (RFE) to select 15 variables
- Final model : 13 variable with p-value of variables less than 0.05 and vif less than 5

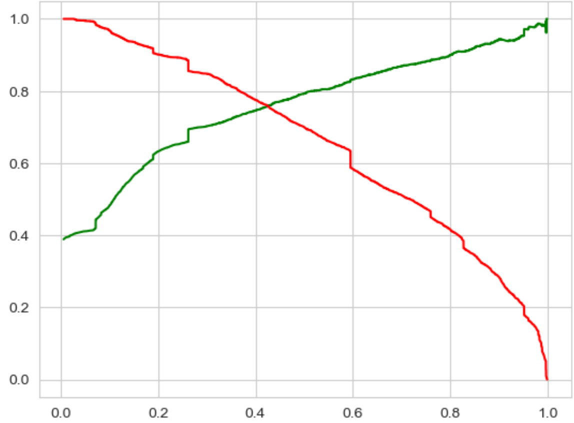
- ROC=0.88: It indicates that the model is good in predicting values
- The optimal cut off point is approx.0.35
- The Precision and Recall cut-off is near 0.41.
- Using optimal cut off point as 0.35, We got

Confusion Matrix: [1389, 337], [ 202, 844]

Accuracy=0.81,Sensitivity=0.81,Specificity=0.80,Recall=0.81

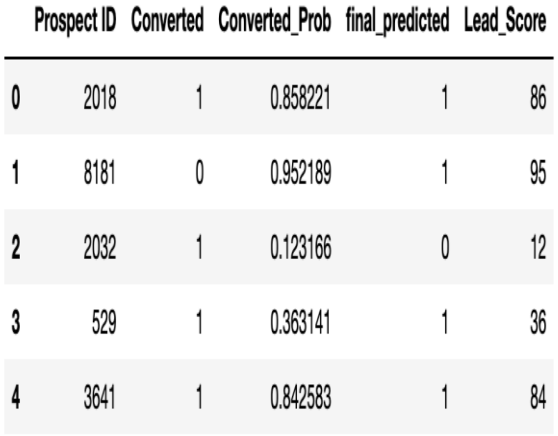




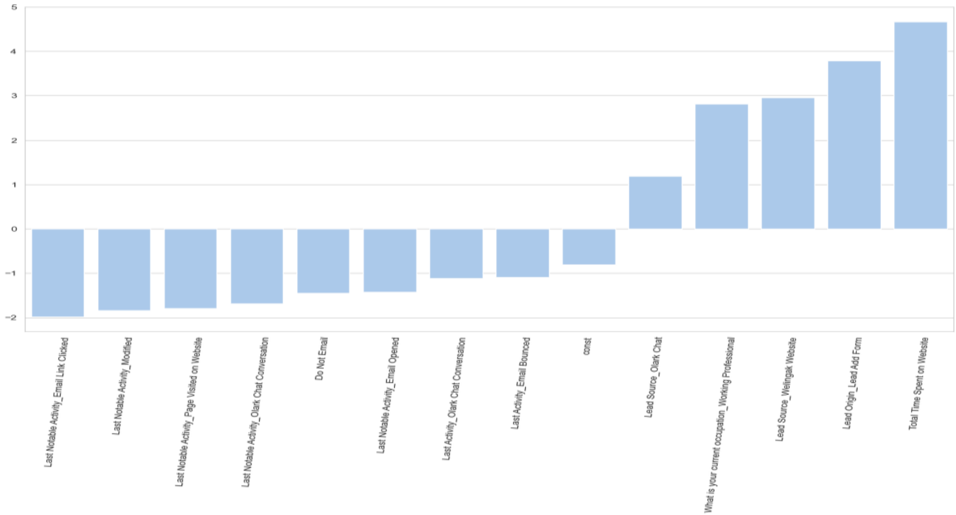


# Results

- Customers having higher lead scores are considered as ‘hot leads’ who are likely to get converted.
- Top Variables influencing the Conversion is "Total Time Spent on Website","Lead Origin_Lead Add Form","Lead Source_Welingak Website"



| | Prospect ID | Converted | Converted_Prob | final_predicted | Lead_Score |
|---|---|---|---|---|---|
| 0 | 2018 | 1 | 0.858221 | 1 | 86 |
| 1 | 8181 | 0 | 0.952189 | 1 | 95 |
| 2 | 2032 | 1 | 0.123166 | 0 | 12 |
| 3 | 529 | 1 | 0.363141 | 1 | 36 |
| 4 | 3641 | 1 | 0.842583 | 1 | 84 |



# Recommendations

- The model built has good scores in terms of Accuracy, Sensitivity, Specificity and Precision. Hence use this model to predict the lead score and concentrate on ‘hot leads’.
- Sales team should concentrate on  leads depending on total time spent on website, lead source generated through Welingak website and lead originated from lead add form.
- Leads contacted through SMS and email have better conversion rate and hence sales team should share promotional campaign like new courses or discount offers on courses to leads who have higher lead score through SMS and email.
- The target audience should be Indian working professionals as conversion rate is higher in this category.
- By providing job placement assistance and educational loan offers, unemployed leads conversion can be increased.

# Conclusion

- The Model looks promising  in predicting leads
- Using Logistic Regression Technique, we could able to achieve our goal of predicting the “HOT LEAD” with an accuracy of 80%
- Overall this model proves to be accurate

# Thank you