# Lead Scoring Case Study

Presented By:

Mr. Rohan Bhosle

# Agenda

- Problem Statement
- Solution in Brief
- Approach
- Assumptions made.
- Outlier treatment.
- EDA.
- Model Building.
- Performance of Model
- Observations & Strategy/Recommendations

# Problem Statement

- X Education sells online courses to industry professionals. 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.
- The people who fill these forms are classified as a lead.
- As an analyst, it’s our responsibility is to create a model which would assign a lead score to each of these leads
- This should help the Company to increase their lead conversion rate from 30% to 80%.

# Solution in brief.

- We have historical data available with us along with the Target variables i.e. ‘Converted’ with values as Yes or No.
- It’s a classification problem, hence we will be using Logistic Regression to build a machine learning model.
- We will build our model in such a way that Recall (or Sensitivity) value would be between 78 to 82% as company wants to maximize their conversion rate to be around 80%.
- We will also require a Lead Score between the range of 0 to 100 where 100 refers to highest probability of lead getting converted and we can get this score through probability value which will be calculated using GLM (Generalized Linear Model).
- We will also need to provide few strategies to the company in the end based on the results of our model.

# Approach

I. **Reading Data. Read and understand the structure of data given.**

II. **Data Quality check and corrections.**

- Replacing Select Values with Nan.
- Dealing with Null / Missing values in df.
- Outlier Treatment.

III. **Visualizing the Data.**

IV. **Dummy Variable conversions.**

V. **Test-Train Split.**

VI. **Rescaling the numerical features.**

VII. **Feature selection using RFE.**

VIII. **Iteration of models.**

IX. **Prediction and Model Evaluation.**

X. **Plotting the ROC Curve.**

XI. **Optimizing the Cut-off value of Probability predicted.**

XII. **Predict the final conversion status of each lead using the optimized cut-off value**

XIII. **Model Evaluation on Train data.**

XIV. **Same model is applied on the test data set to check with the model performance.**

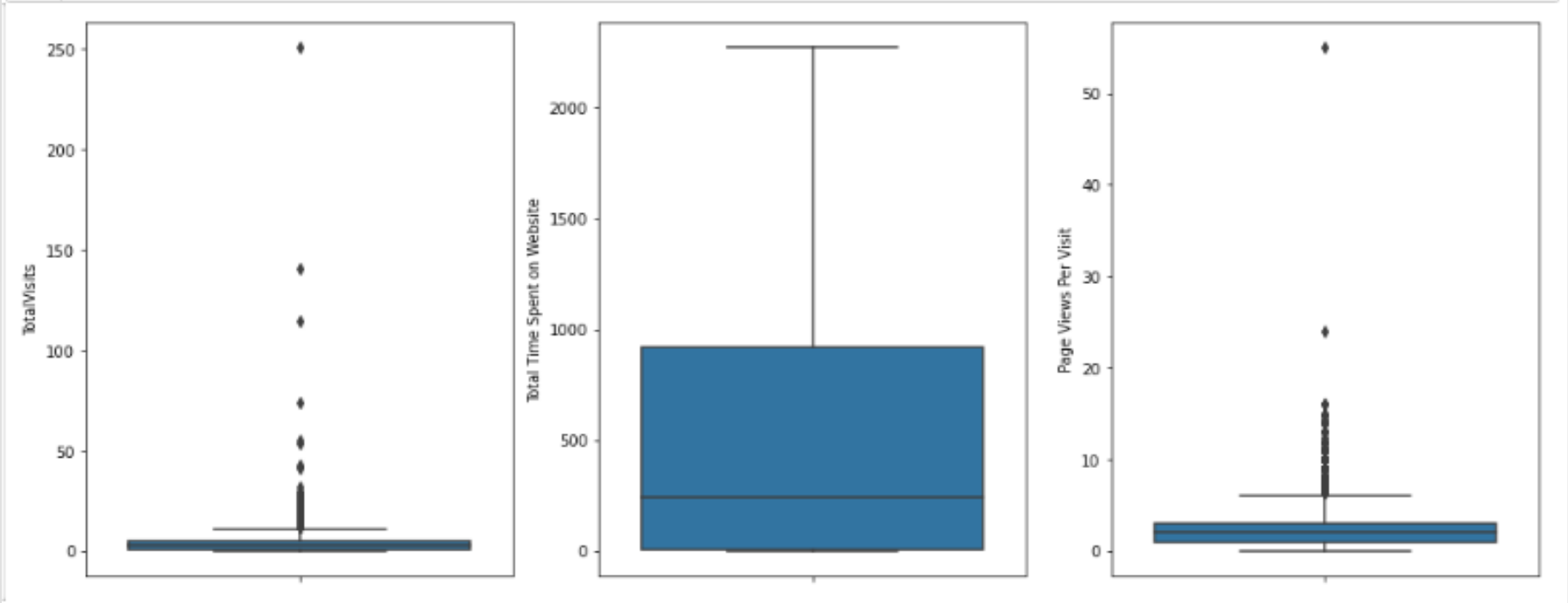
XV. **Lead score is calculated = Predicted probability * 100**

XVI. **Model Evaluation on Test data.**

# Assumptions made.

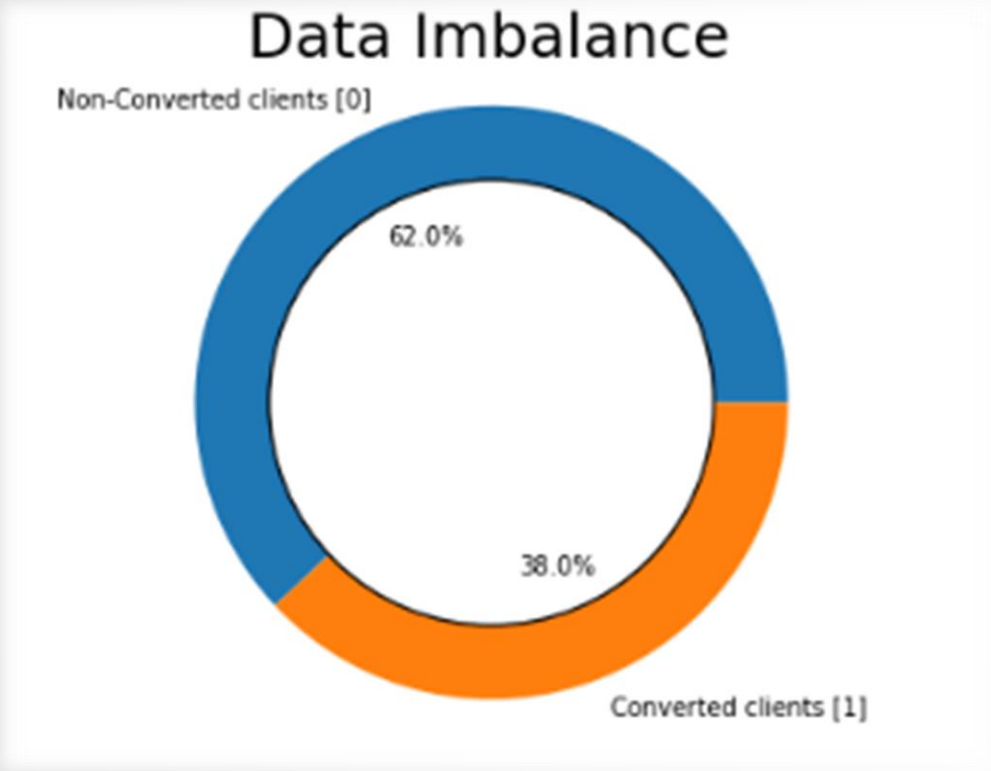
- The Select value in the data set is converted to Nan / null values
  - The people willing to enroll in the course wouldn't have opted for a drop down option or missed to give a value while filling the form.
  - Hence, these values can be considered as Null or missing values.
- Fields with missing value ratio > 40% are dropped.
- Variables with just a single values are also dropped.
  - `'Magazine'`
  - `'Receive More Updates About Our Courses'`
  - `'Update me on Supply Chain Content'`
  - `'Get updates on DM Content'`
  - `'I agree to pay the amount through cheque'`
- There exists a data imbalance in variables, these are dropped:
  - Newspaper
  - X Education Forums
  - Newspaper Article
- i.e. We can see only 1 or 2 lead count in any of the category, which will not contribute towards a good model.
- Records with ID are dropped : Prospect ID & Lead number as they wont be used in the regression.

# Outlier Treatment:




- We can clearly observe outliers within the variables TotalVisits & Page Views Per Visit.
- Treatment : soft capping on upper ends.
- There are no outliers in lower end as negative value is not possible in these variables and it can only start from 0.

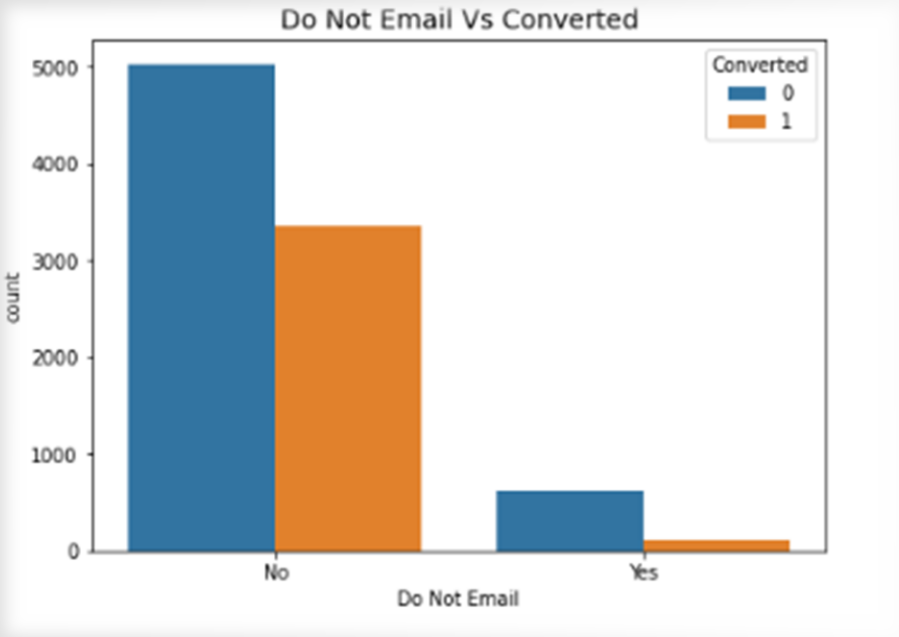
# EDA.

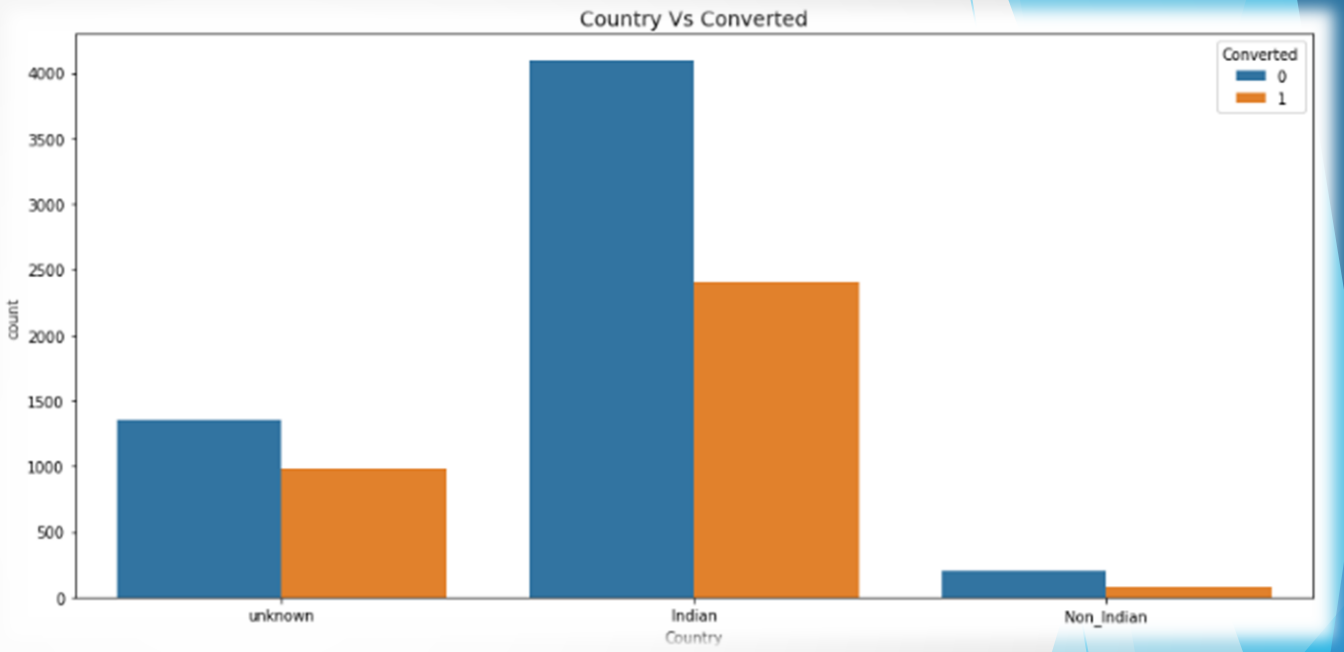



We can see that the Non-Converted lead within the given data set is high when compared to Converted leads:

- Non Converted total% = 62.0%
- Converted total% = 38.0%

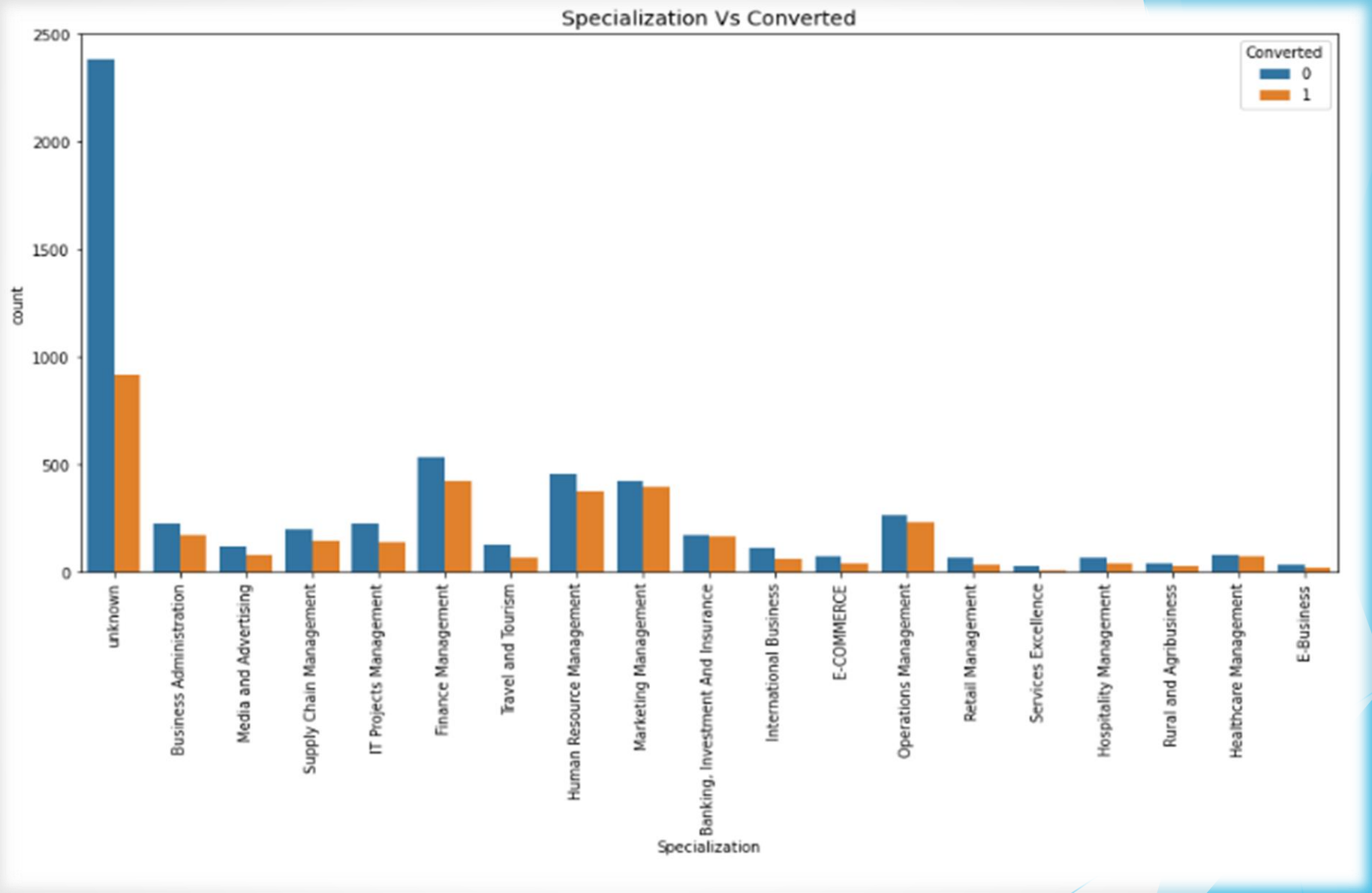
# EDA.




- The people who opt for a email service have more Conversion rate than the people who haven't opted for Email service.
- Which means that the customer gets convinced better when called by a sales representative.
- Enrollment is high from India compared to any other countries.

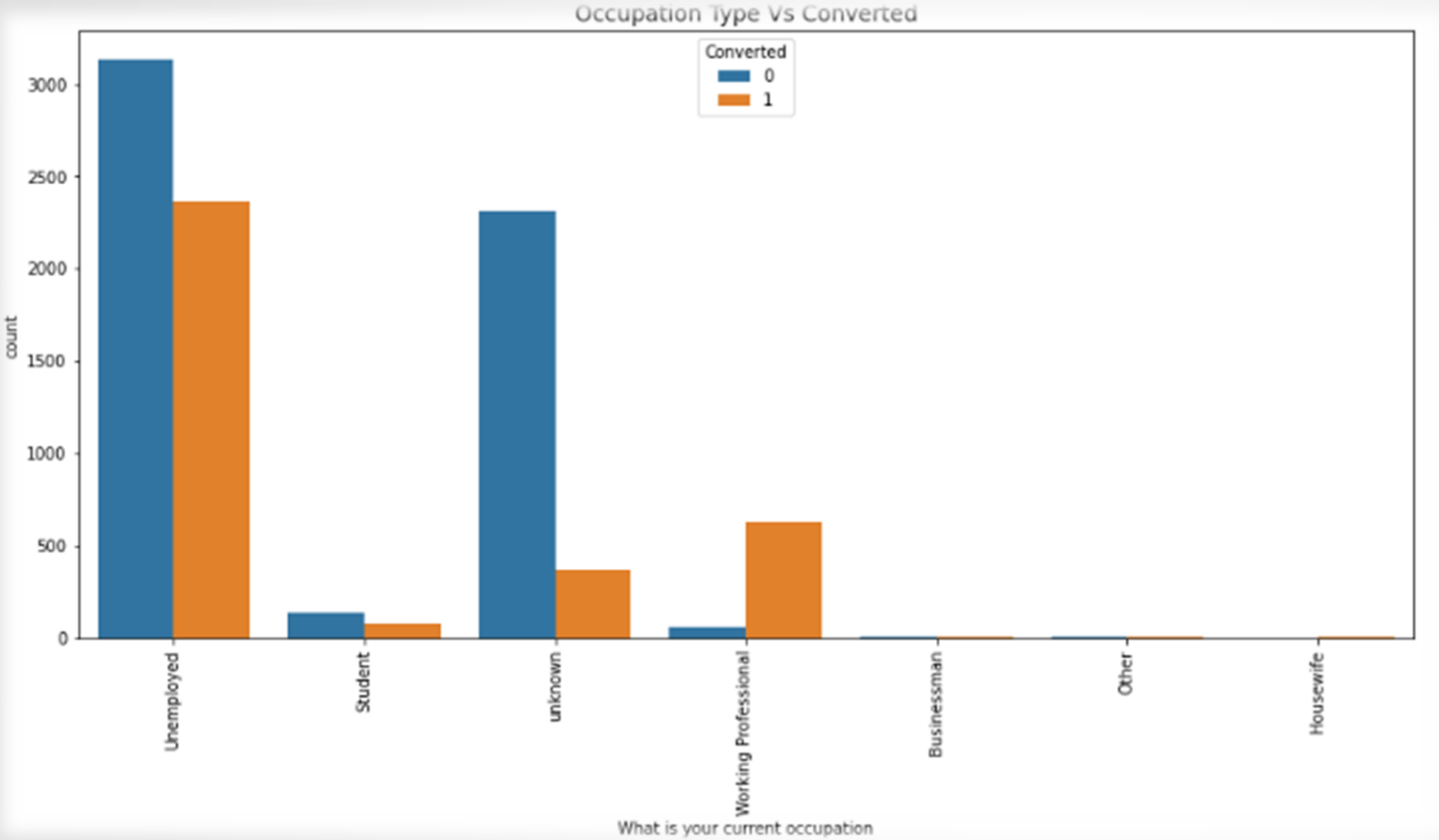
# EDA.




We can observe a comparatively better counts of conversion rate in :

- Finance Management, Human Resource Management, Marketing Management & Operations Management

# EDA.

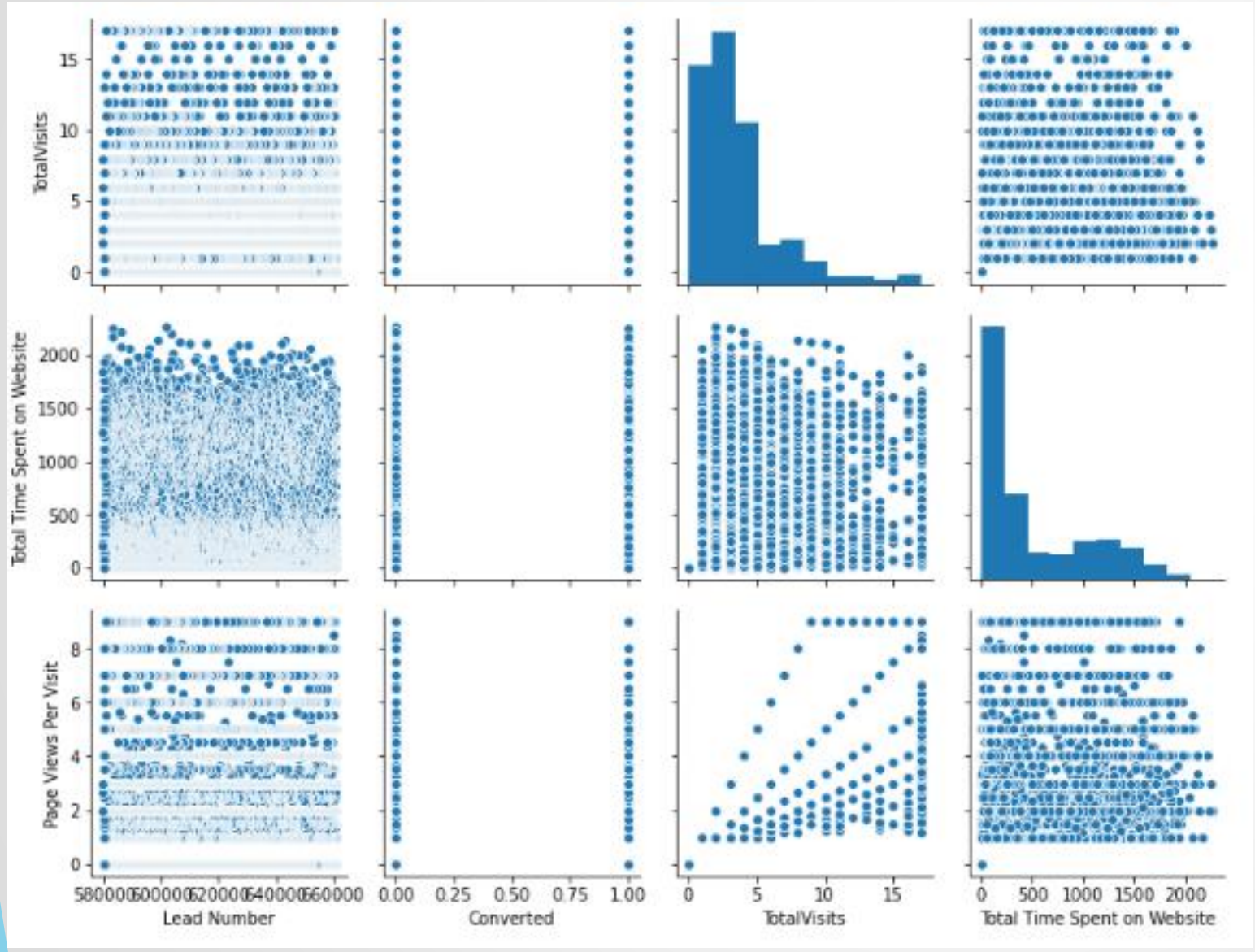



| Converted | 0 | 1 |
|---|---|---|
| What is your current occupation | | |
| Businessman | 3.0 | 5.0 |
| Housewife | NaN | 9.0 |
| Other | 6.0 | 9.0 |
| Student | 132.0 | 75.0 |
| Unemployed | 3133.0 | 2369.0 |
| Working Professional | 55.0 | 624.0 |
| unknown | 2313.0 | 370.0 |

We can see a positive conversion rate for below applicants with Occupations:

- Working Professional,
- Other,
- Housewife and
- Businessman.

# EDA : Pair-Plot



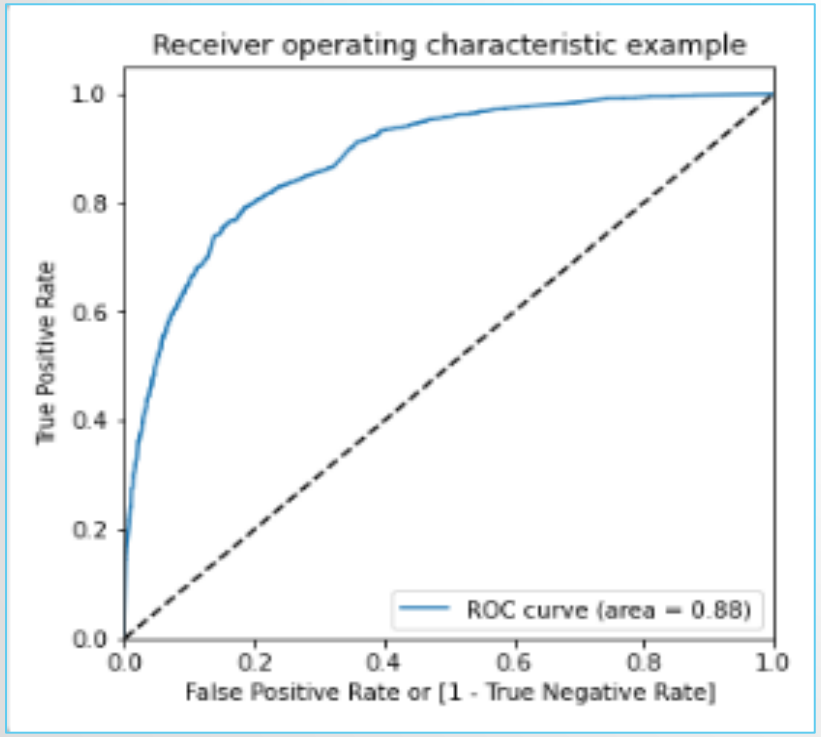

- We can see a linear relationship between Page Views per visit & TotalVisits. So lets remove 'Page Views per visit'.
- Total Spent in website is comparitively less.
- For TotalVisits data is left skewed but normally distributed.
- Let us drop either of 3 variables, let us drop Page Views Per Visit

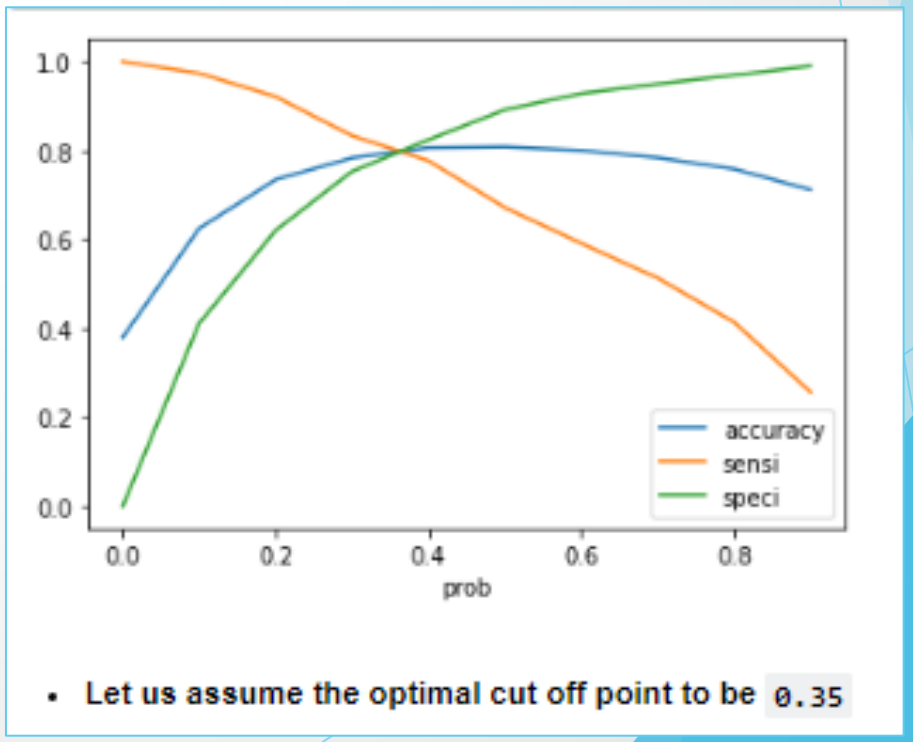
# Model Building:

- **Dummy Variables:**
  - Fields with Yes/No categories are converted to 1/0
  - Other categorical variables are modified to dummy variables by dropping respective dummy column.
- **Test-Train Split:**
  - The given data set is divided into Train Data & Test Data at 7:3 ratio.
- **Rescaling the numerical features:**
  - The numerical fields like: TotalVisits, Total Time Spent on Website are rescaled using Standardization.
- **Feature selection using RFE:**
  - Recursive elimination method is used to select the best 30 variables contributing to the conversion rate.
- **Iteration of models:**
  - A pre-defined function is used to generate a model and VIF.
  - Depending upon the p-values & VIF we can reduce down the variables.
  - We have conducted 13 iterations to finalize with the best fitting model.
- **Prediction and Model Evaluation:**
  - After optimizing the VIF and p-values we can go ahead and predict the new conversion status of each lead by taking the cut off as 50% probability.
  - Verify with the model statistics, i.e. Accuracy **80.8%**, Sensitivity **67.2%** & Specificity **89.1%**.

# Model Building (Cont.)

- **Optimizing the Cut-off value of Probability predicted.**
- **Plotting the ROC Curve:**
  - An ROC curve demonstrates several things:
  - It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
  - The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.
  - The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.



# Performance of Model

- Accuracy, Sensitivity (Recall) & Specificity from Train Data set:
  - **Accuracy: 79.7%**
  - **Sensitivity: 80.8%**
  - **Specificity: 79.1%**
- Accuracy, Sensitivity (Recall) & Specificity from Test Data set:
  - **Accuracy: 80.4%**
  - **Sensitivity: 81.1%**
  - **Specificity: 80.0%**
- We can see that the final model is fetching almost similar model parameters which show us that the model is good to go!
- Also, the Sensitivity or Recall is nothing but the conversion rate predicted using the model.
- With the model we are able to predict a conversion rate of 81.1% on test data.
- We have also added 'Lead Score' column the train/test dataset
  - its value will be between 0 to 100 specifying the probability of lead getting converted into a customer, being 100 as highest probability

# Observations

We have got below few important variables which has high impact of lead conversion:

- **Occupation:** From the final model we ran, we can clearly see that 'Working Professionals' has more probability being converted to customer.
- **Lead Origin:** During the identification step of the lead, people who fills a form have more chances of being converted to customer.
- **Last Activity:** Professionals who already had been called up by the sales team have more chances of being converted.

# Strategy/Recommendations

**# 1 -** From the final model this has been seen that leads with high probability have more chances to get converted when they already had a conversation/call with the sales team. So, sales team should reach out to such leads again and ensure they purchase or enroll for the course. Apart from this, sales team should also focus on working professionals as they have more chances of getting converted

**# 2 -** Probably sales team can start making plans for next quarter and they can spend some time on identifying potential leads/customers for the next batch. Also, they can ask existing leads/customers to refer their friends and avail some discounts through it, and this can be communicated via sending an email to existing leads who already enrolled for the course.

# Thank You!