Lead Score (case study)
Assignment

June Batch'2022

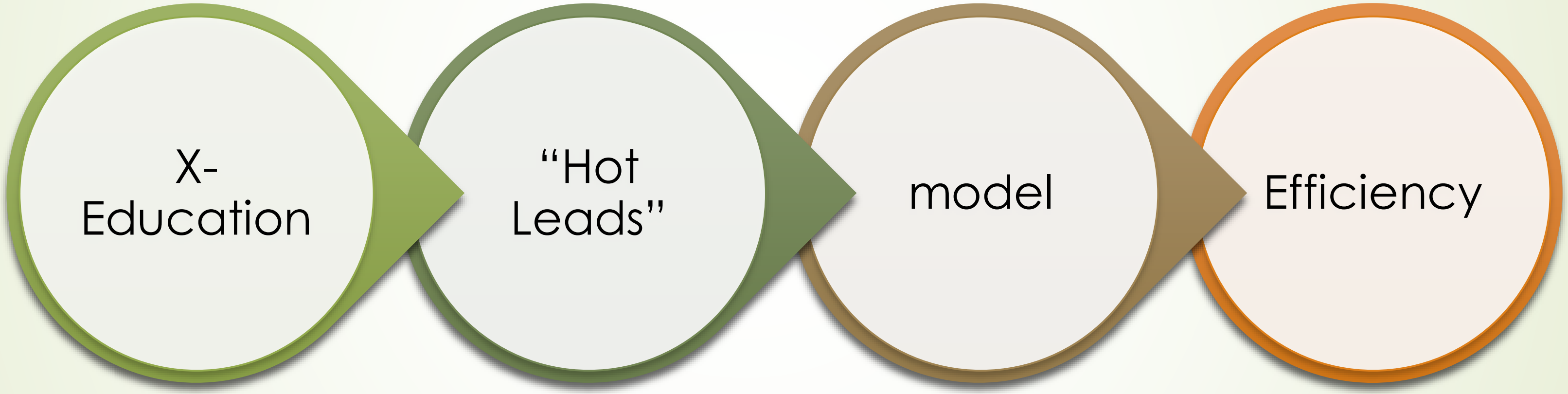
By:
Sourabh Kumar Soni
Sourabh.soni38@gmail.com, +919958444776

Table of Contents

- Problem of statement.
- Overall approach
- Understanding & cleaning the data.
- Exploratory Data Analysis.
- Model building.
- Model Evaluation.
- Conclusion.

Problem statement

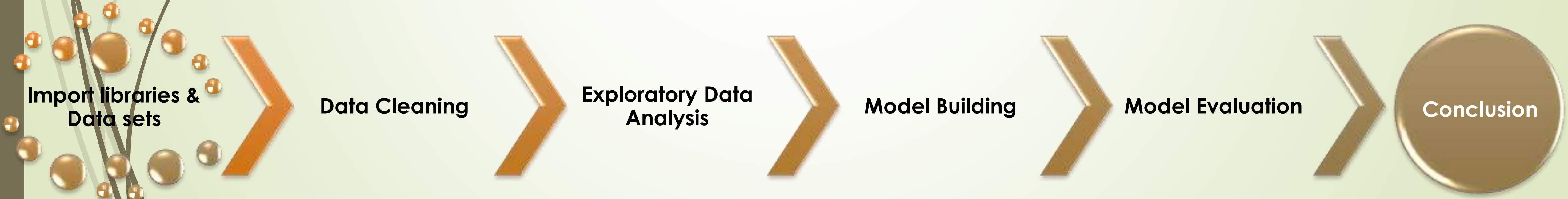
- An X Education needs assistance in choosing the leads that have the best chance of becoming paying clients. The business wants us to develop a model in which each lead is given a lead score, with higher lead scores indicating a higher likelihood of conversion and lower lead scores indicating a lower likelihood of conversion. The desired lead conversion rate has been estimated by the CEO to be in the range of 80%.



- The business wants to identify the most promising leads, often known as "Hot Leads," in order to increase the efficiency of this process. The lead conversion rate should increase if they are successful in locating this group of leads because the sales staff will be spending more time speaking with potential leads rather than calling everyone

Overall approach

- Import different kind of **libraries like numpy, pandas, matplotlib, seaborn, Sklearn** etc.
- Import **warnings & packages** etc.
- Read/extract two different set of data (i.e. leads data) as csv files after uploading in jupyter notebook.
- Checked data structures by using head(), describe(), shape, info() etc. functions.



Import libraries & Data sets

Data Cleaning

Exploratory Data Analysis

Model Building

Model Evaluation

Conclusion

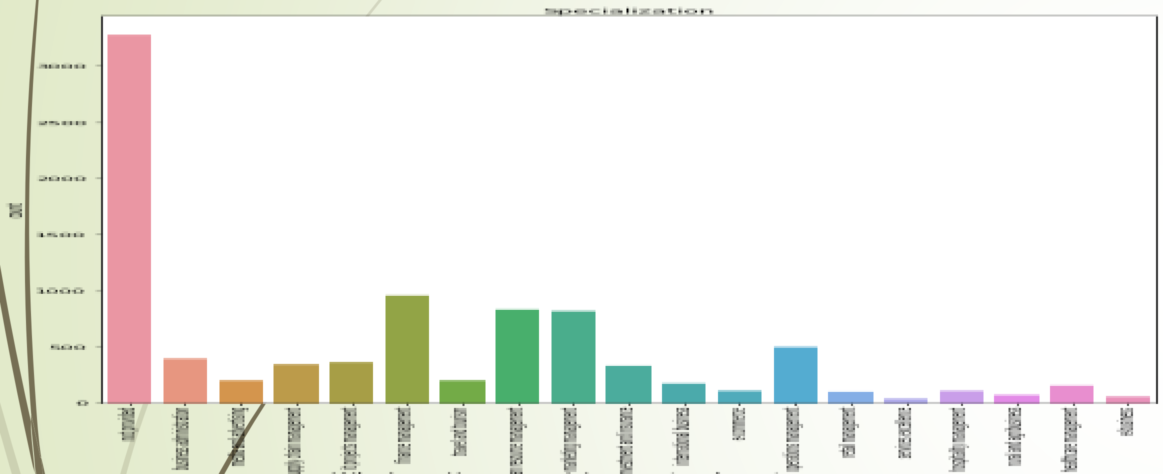
Understanding & cleaning the data frame

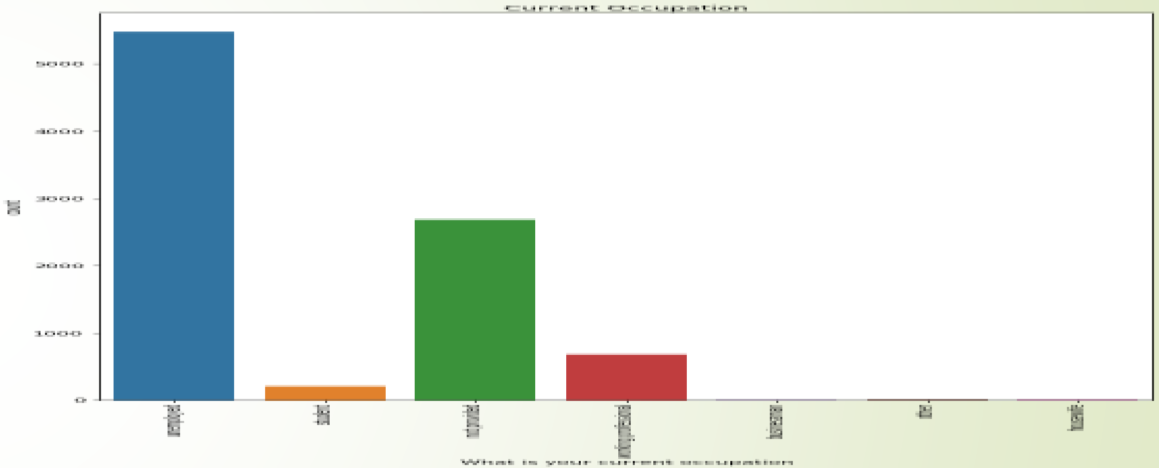
- Converting all the values to lower case.
- Replacing “select values with NaN.
- Identifying whether columns have unique values and dropped them.
- Identifying the missing values.
- there are a large number of null variables in 4 columns. However, since these are crucial columns, eliminating the rows with null values will cost us a lot of data. Therefore, we are going to substitute “not given" for the NaN values. In this manner, we have nearly no null values and all the data. If any of these are present in the model, it will be useless, and we can drop it off at that point.
- Eliminating columns having 35 percent of null values.

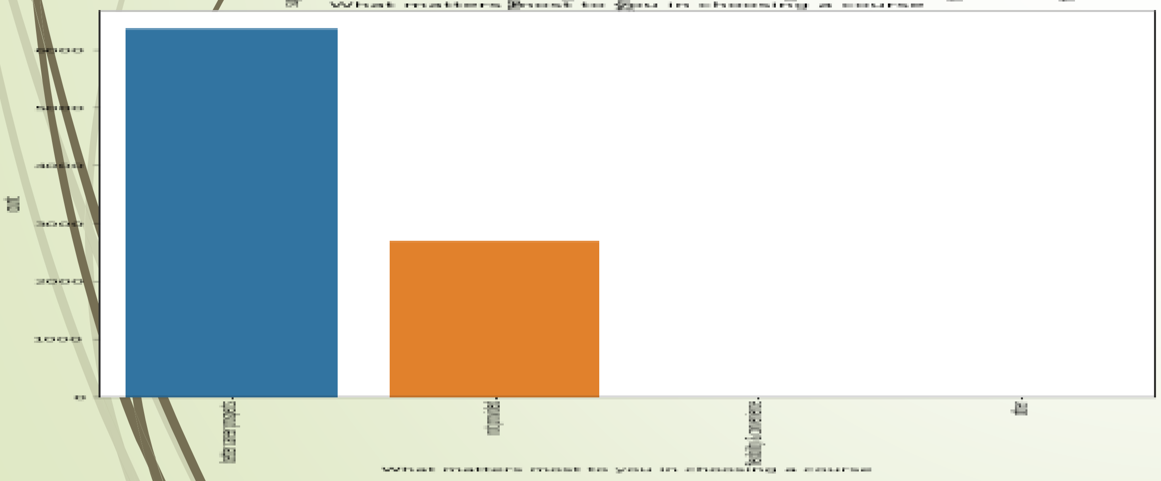
Exploratory Data Analysis (EDA)

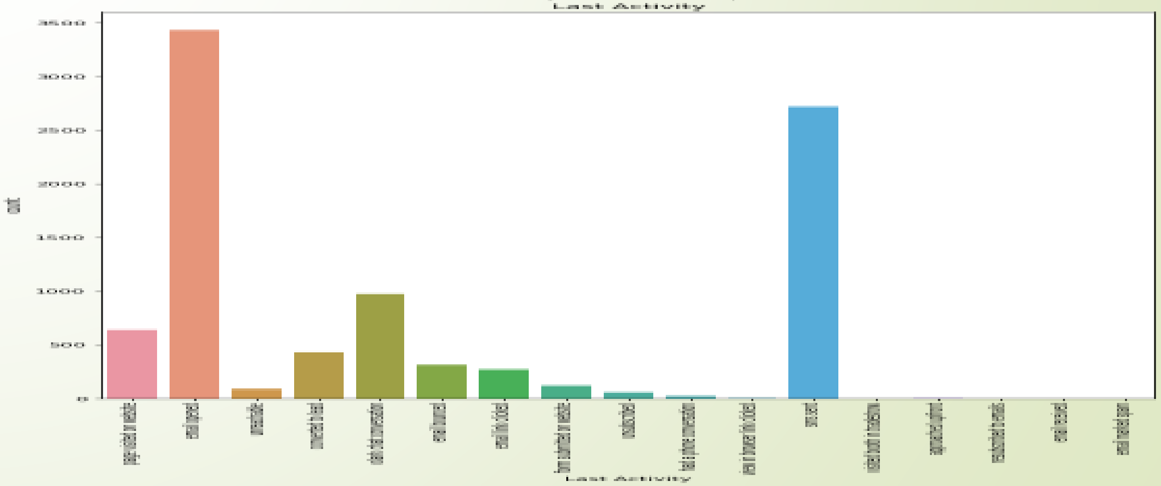
- Univariate analysis

- Categorical variable:





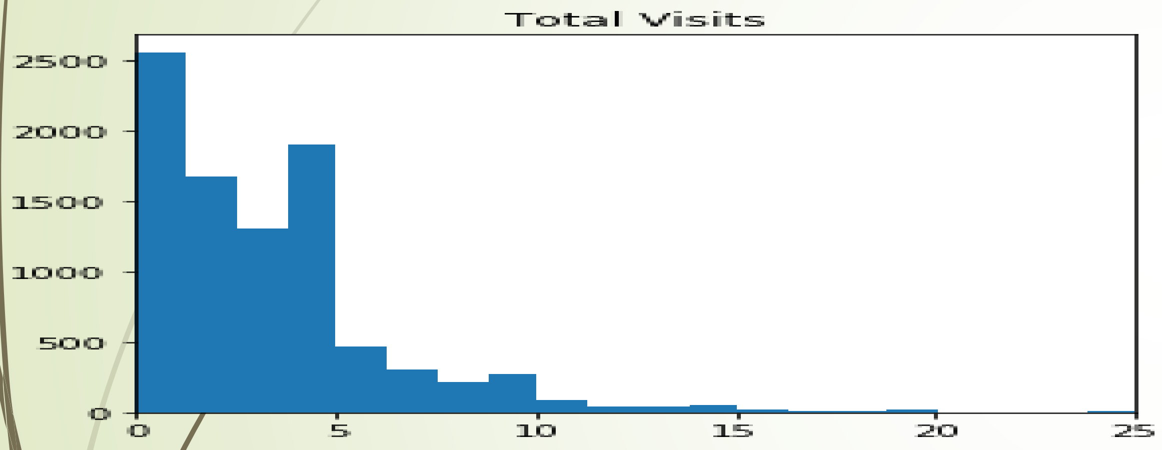




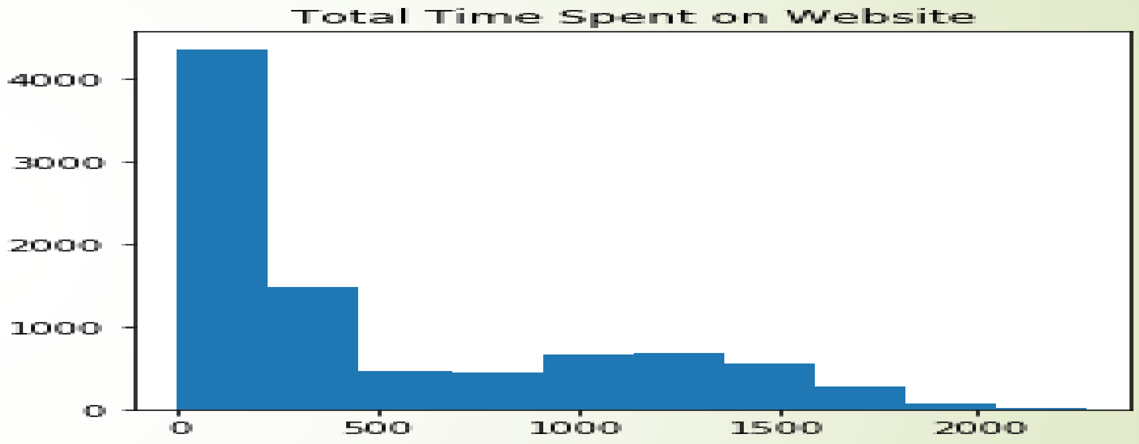
Exploratory Data Analysis (EDA)

- Univariate analysis

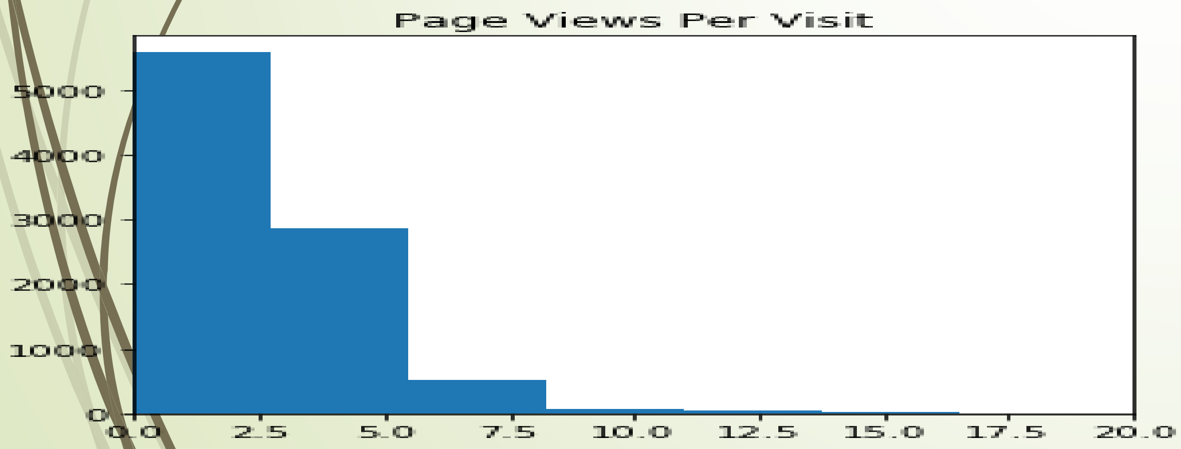
- Numerical variable:



Total Visits



Total Time Spent on Website

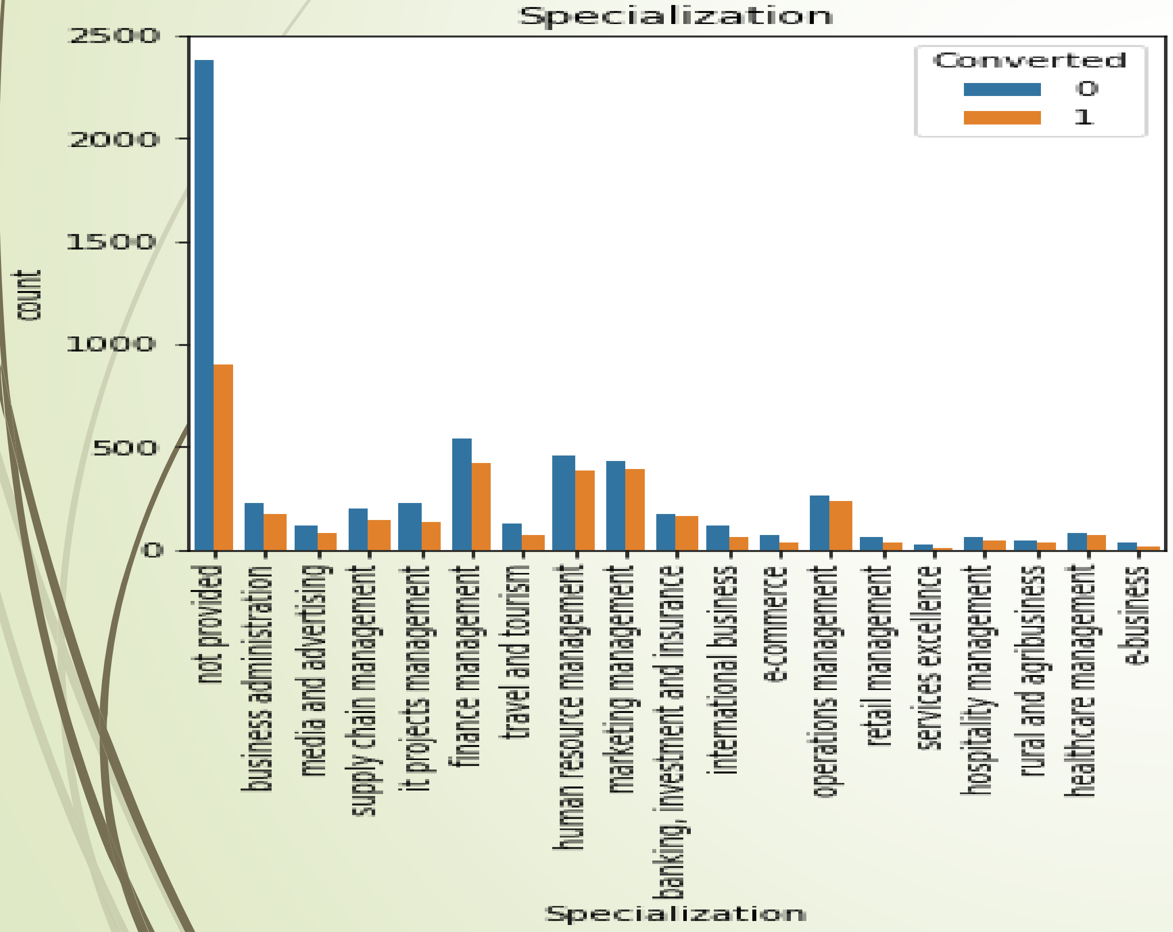


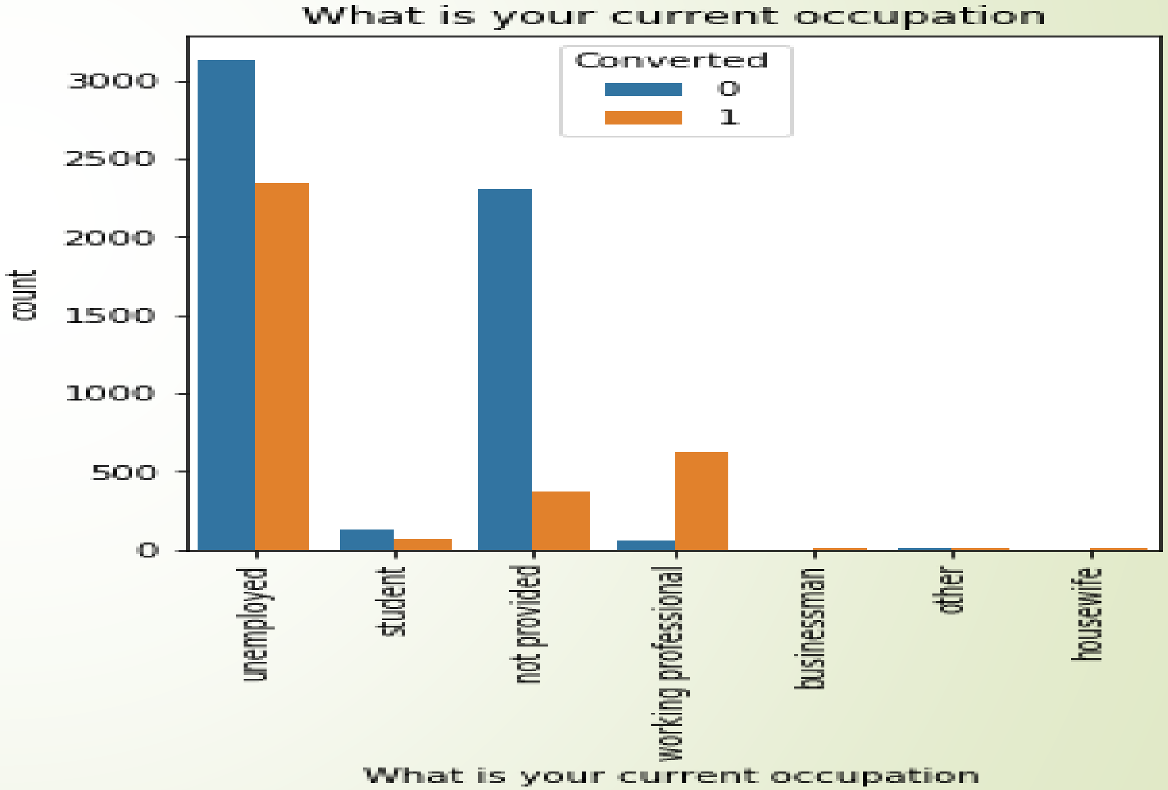
Page Views Per Visit

Exploratory Data Analysis (EDA)

- Univariate analysis

- Relating all the variables:

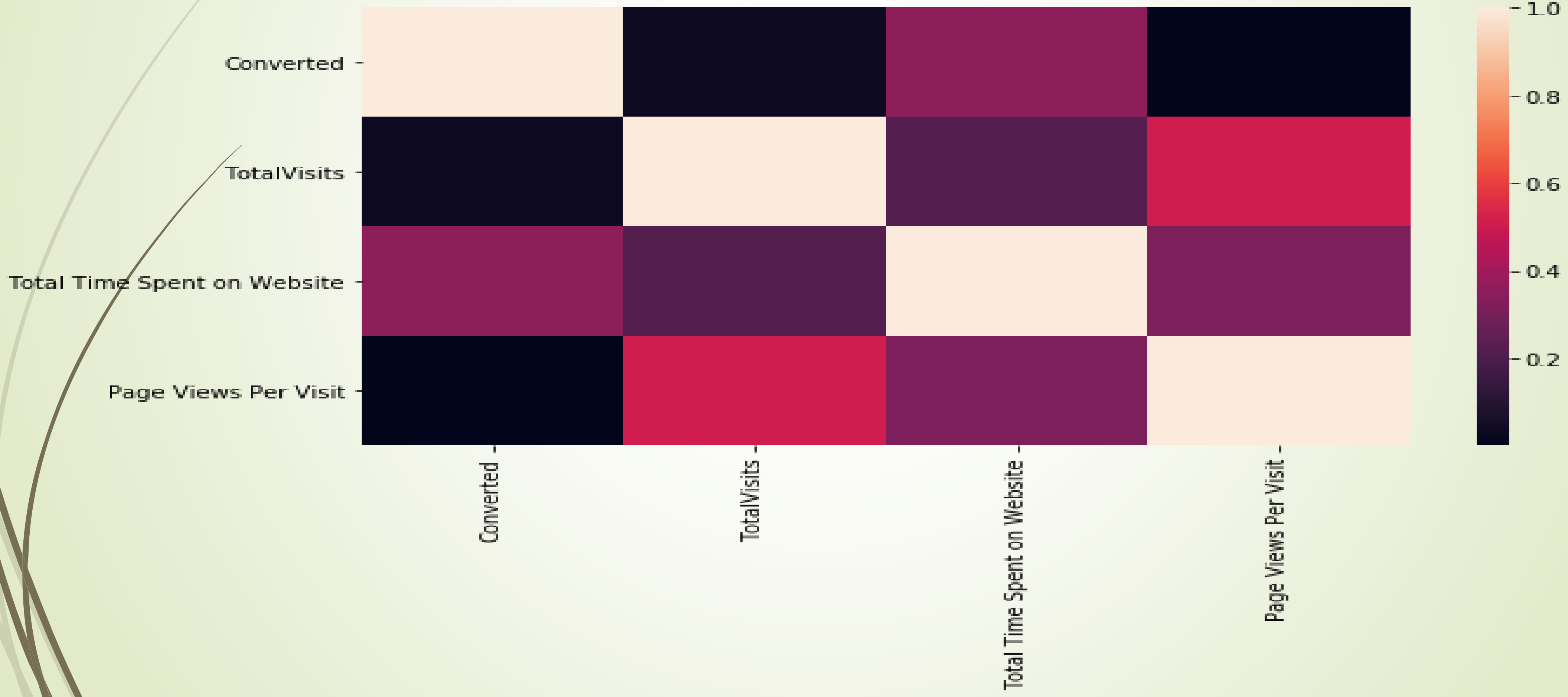




Exploratory Data Analysis (EDA)

- Univariate analysis

- Relating all the variables:

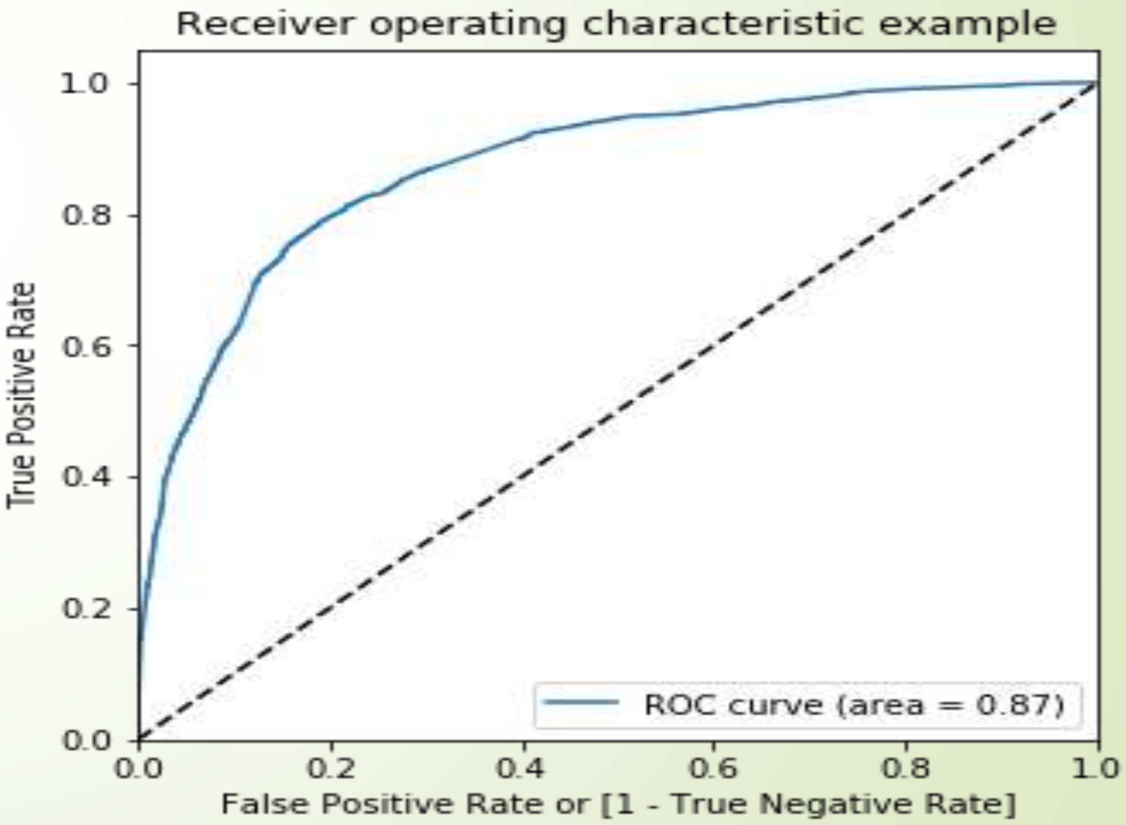


Model Building

- Import logistic regression model.
- Import (Recursive Feature Elimination) RFE.
- Applied Generalized linear regression model.
- Created a Variance Inflation Factor (VIF) data frame with all the variables where p-values don't appear right, but the VIF values do. So eliminate "Last Notable Activity had a phone conversation" as a result. All of the VIF values are favourable, and each p-value is less than 0.05 to fix the model.
- Then we make prediction or predicted probabilities in trained model, where data frame containing the conversion rate and the likelihood of the projected values.

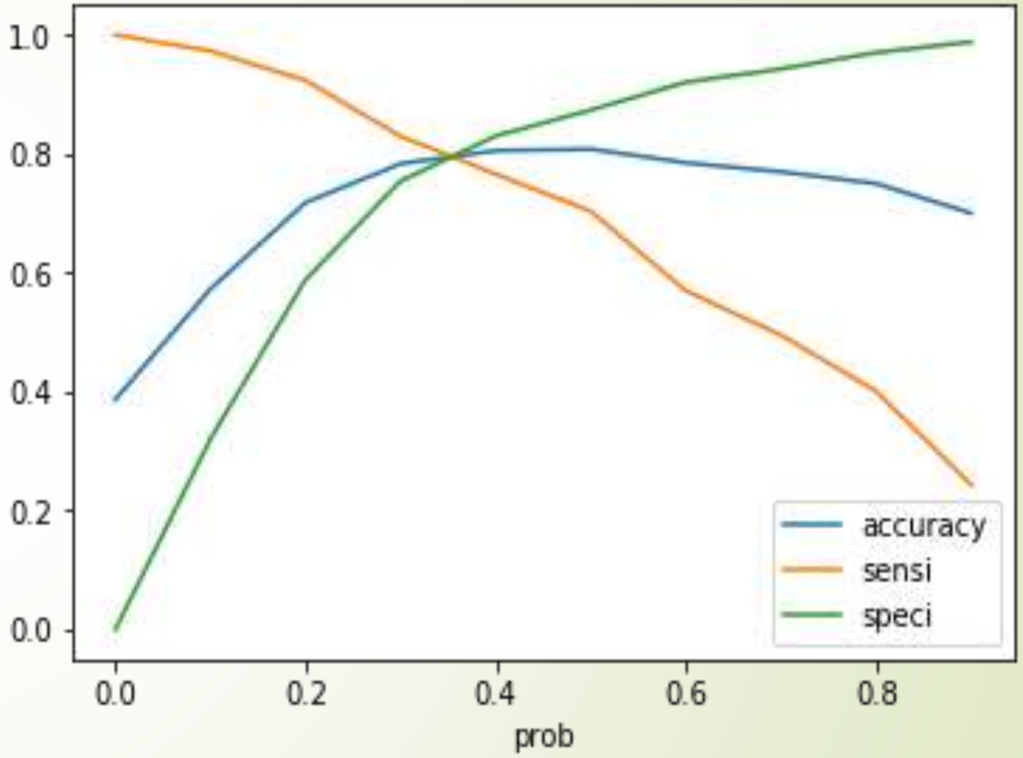
Model Evaluation

- Creating a confusion matrix and finding overall accuracy.
- With the present cutoff set at 0.5, we have approximately 81% accuracy, 70% sensitivity, and 87% specificity.
- ROC curve to optimize, where the prior cutoff was chosen at random.
- The ROC curve's area under it is 0.87, which is an extremely good value.
- # creating a data frame to see the values of accuracy, sensitivity, and specificity at different values of probability cutoffs and making confusing matrix to find values of sensitivity, accuracy and specificity for each level of probability.

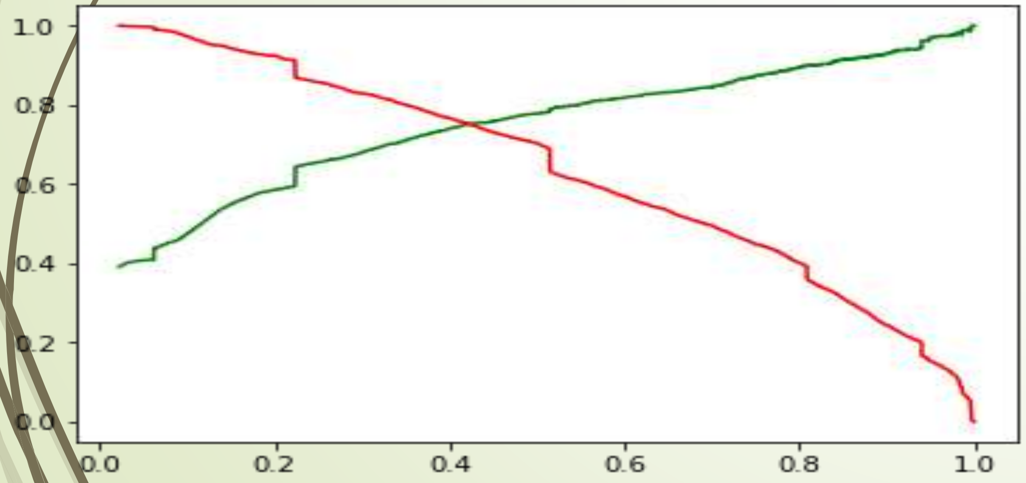


Model Evaluation

- The graph makes clear that 0.35 is the ideal cutoff.
- We have accuracy, sensitivity, and specificity of roughly 80% with the current cut off of 0.35.



- After making prediction, we evaluated Precision is now at 78% and Recall is currently around 70% with the current cut off of 0.35.



Conclusion

- According to study, the characteristics that had the most impact on prospective buyers were, in descending order: the average amount of time spent on the website, number of visitors overall, The last activity was either an SMS or an Olark chat, if the format of the lead add is the lead origin, and if they are a working professional at the time, when Google, direct traffic, organic search, or the Welingak website was the lead source.
- With these in mind, X-Education has a high chance of convincing nearly all potential clients to change their thoughts and buy their courses, which will allow them to prosper.

Thank you