



# LEAD SCORING CASE STUDY

---

GARIMA VASWANI



# PROBLEM STATEMENT

---



- X Education sells online courses to industry professionals.
- X Education get lot of leads , its conversion rate is not very promising as out of 100 leads in a day only 30 of them get converted into potential course takers
- To make the process more efficient the company wishes to identify the most potential leads, also known as 'Hot leads'
- If they successfully identify the set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.





# BUSINESS OBJECTIVE

---



- X Education needs help in selecting the most promising leads i.e. Leads that are most likely to convert into paying customers
  - How did you identify this problem or need?



# STRATEGY



## Data Sourcing , Cleaning and Preparation

- Read the Data from Source
- Convert data into clean format suitable for analysis
- Remove duplicate data
- Outlier Treatment
- Exploratory Data Analysis
- Feature Standardization.



## Feature Scaling and Splitting Train and Test Sets

- Feature Scaling of Numeric data
- Splitting data into train and test set.



## Model Building

- Feature Selection using RFE
- Determine the optimal model using Logistic Regression
- Calculate various metrics like accuracy, sensitivity, specificity, precision and recall and evaluate the model.



## Result

- Determine the lead score and check if target final predictions amounts to 80% conversion rate.
- Evaluate the final prediction on the test set using cut off threshold from sensitivity and specificity metrics



## DATA MANIPULATION

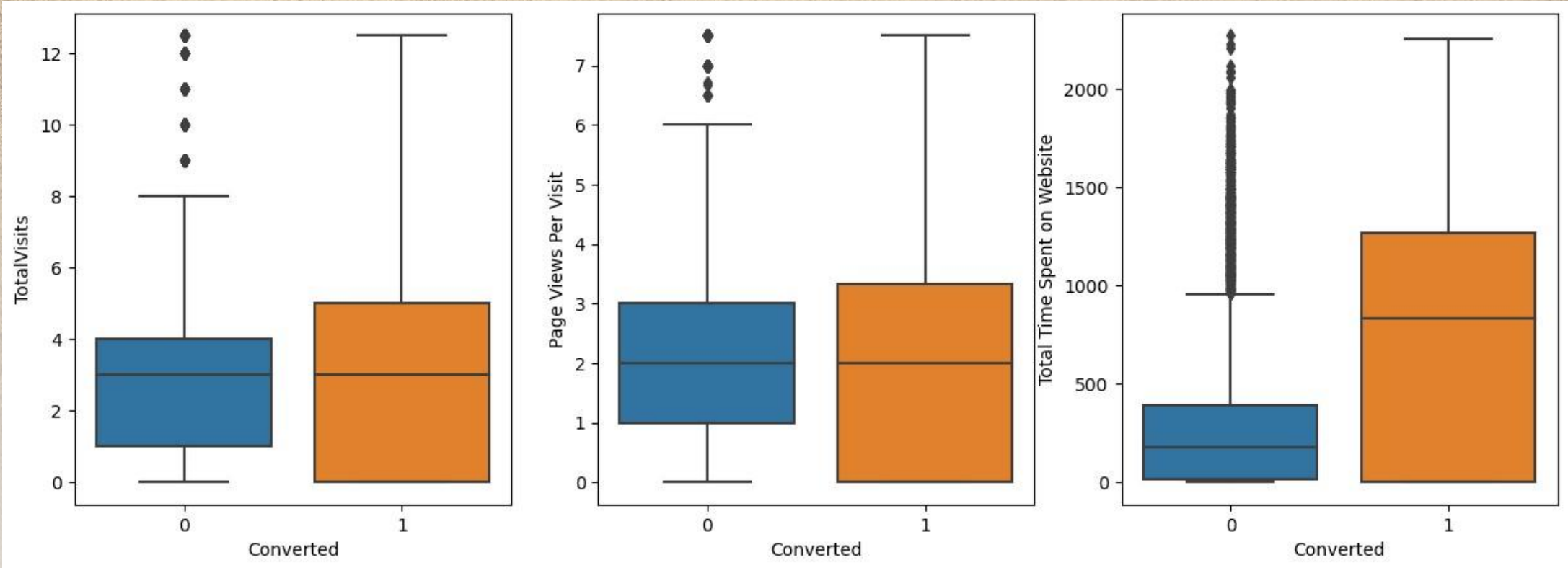
---

- ▶ Total Number of Rows=37, Total Number of Columns =9240.
- ▶ Single value features like “Magazine”, “ReceiveMoreUpdates About Our Courses”, “Update my supply”
- ▶ Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped.
- ▶ Removing the “ProspectID” and “Lead Number” which are not necessary for the analysis.
- ▶ After checking for the value counts for some of the object type variables, we find some of the features which have enough variance, which have dropped, the features are: “Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper, Article”, “XEducation Forums”, “Newspaper”, “DigitalAdvertisement” etc.
- ▶ Dropping the column shaving more than 35% as missing values such as ‘How did you hear about X Education’ and ‘Lead Profile’.



# BOX PLOTS

---

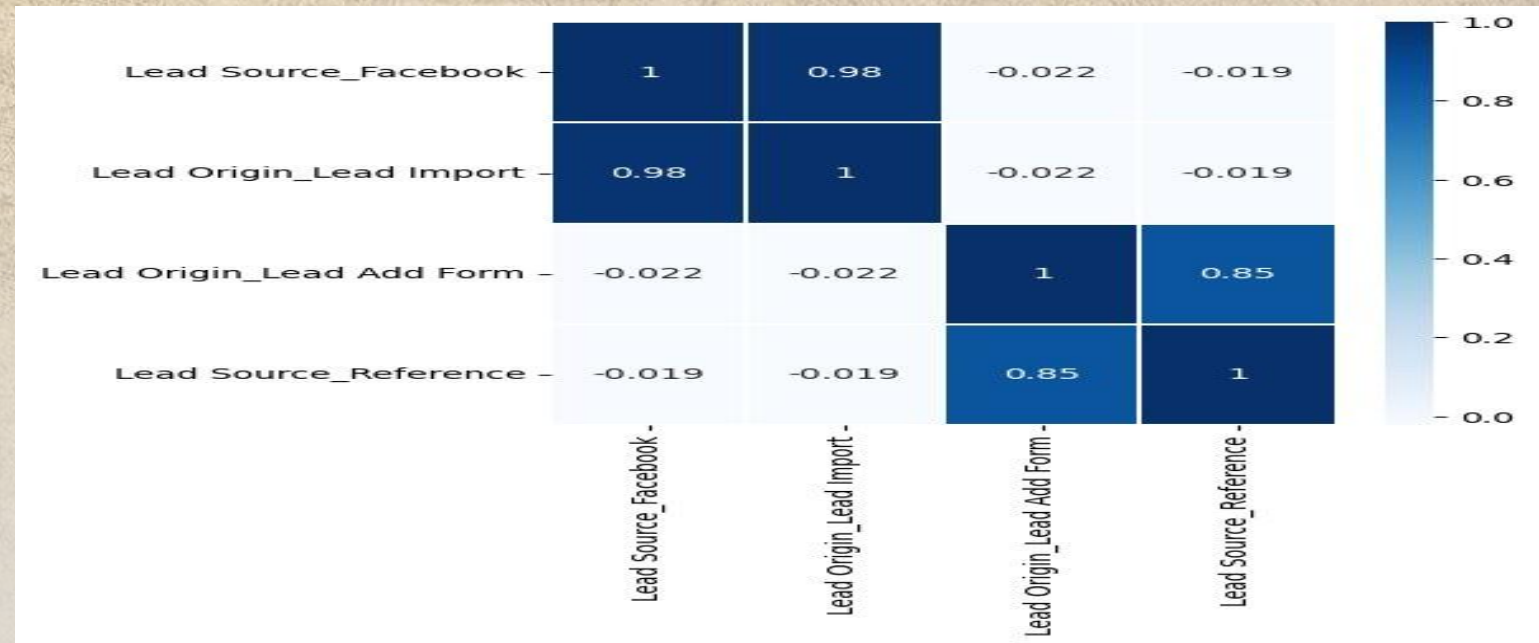




# HEAT MAPS

---

- Heat Maps





# NEXT STEPS

---

- Bivariate data analysis: correlation coefficients and pattern between the variables etc.
- Feature Scaling & Dummy variables and encoding of the data.
- Classification technique: logistic regression is used for model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.



# MODEL EVALUATION- SENSITIVITY AND SPECIFICITY ON THE TRAIN DATA SET

---

