


# LEAD SCORING CASE STUDY

By - Dhruv Sharma  
Suyash Bakre

# PROBLEM STATEMENT

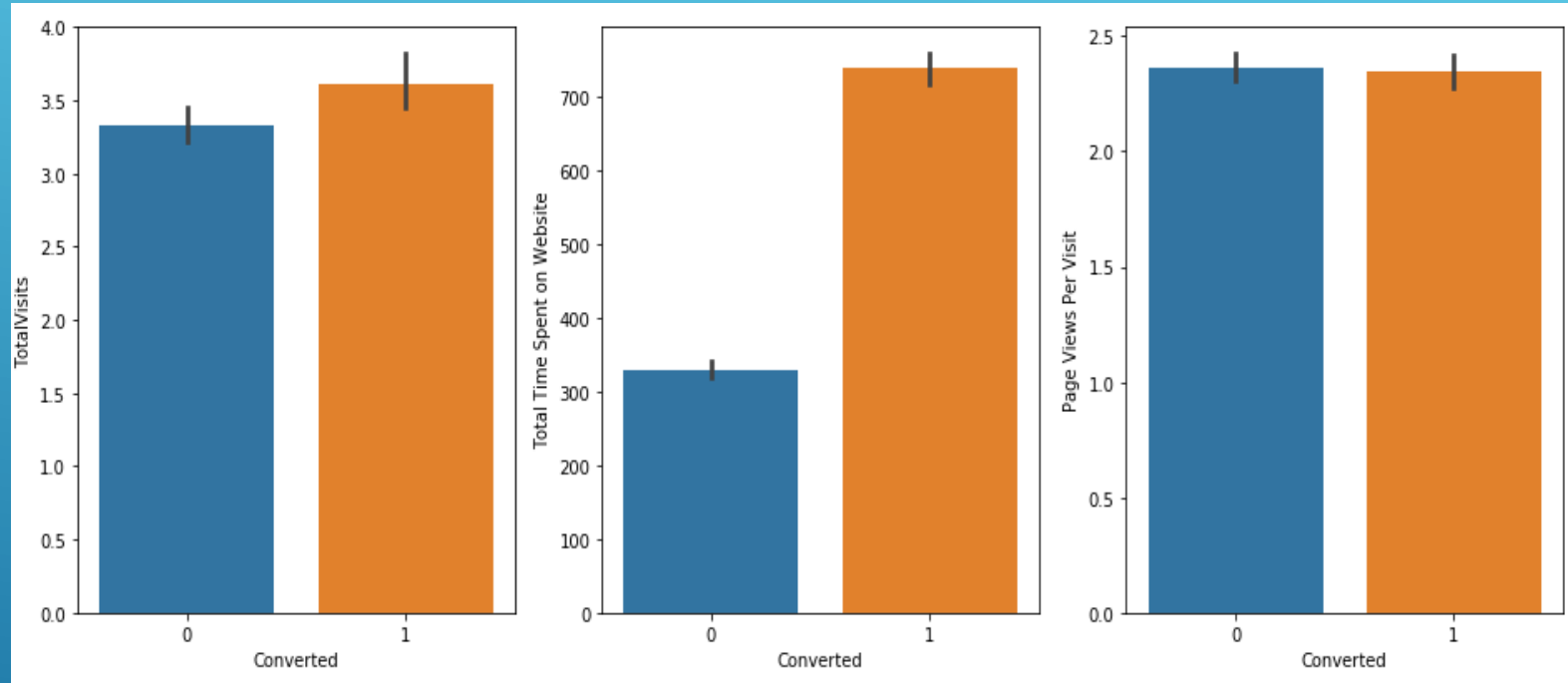
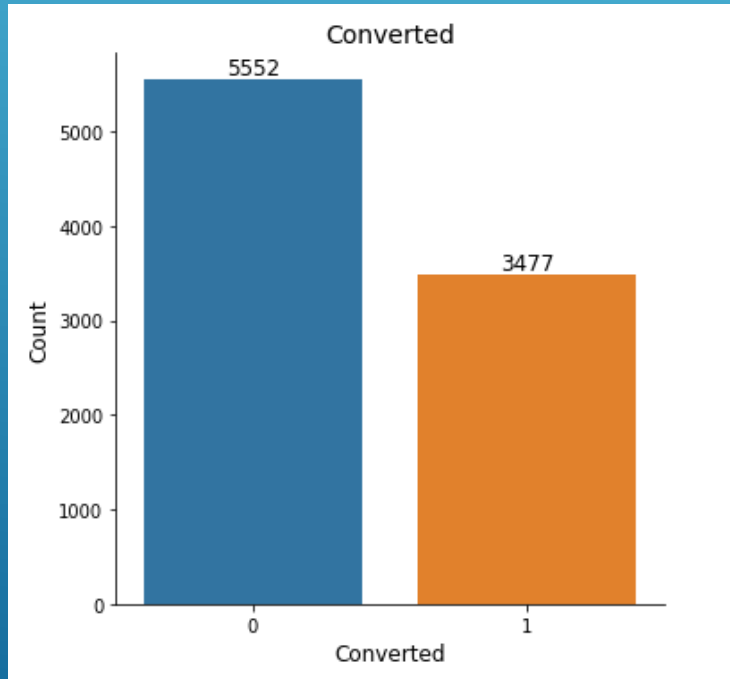
- X Education sells online courses to industry professionals
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'.
- If they successfully identify this 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

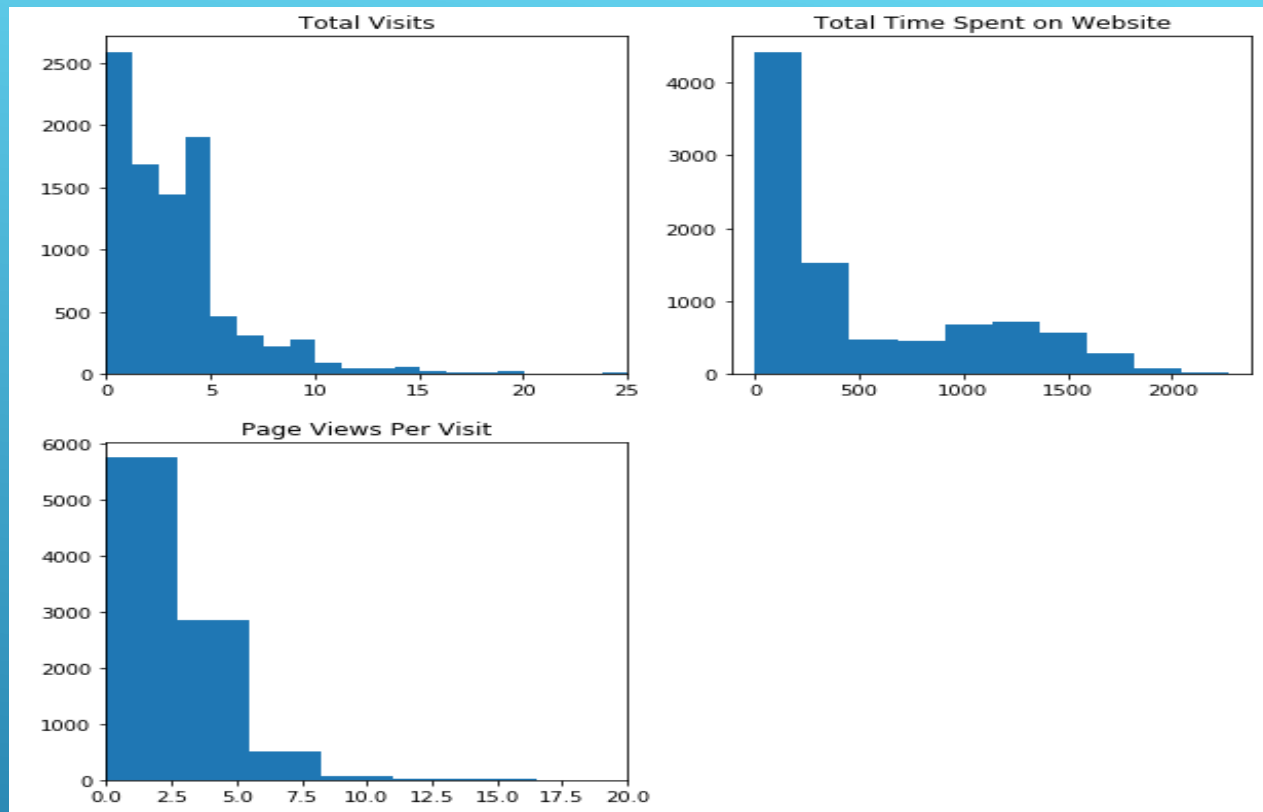
- X education wants to know most promising leads
  - For that they want to build a Model which identifies the hot leads.
  - Target is to get lead conversion rate around 80%
- 
- A series of white lines of varying lengths and slopes are positioned in the bottom right corner of the slide, creating a modern, abstract graphic element.

# EXPLORATORY DATA ANALYSIS

The conversion rate is around 39%

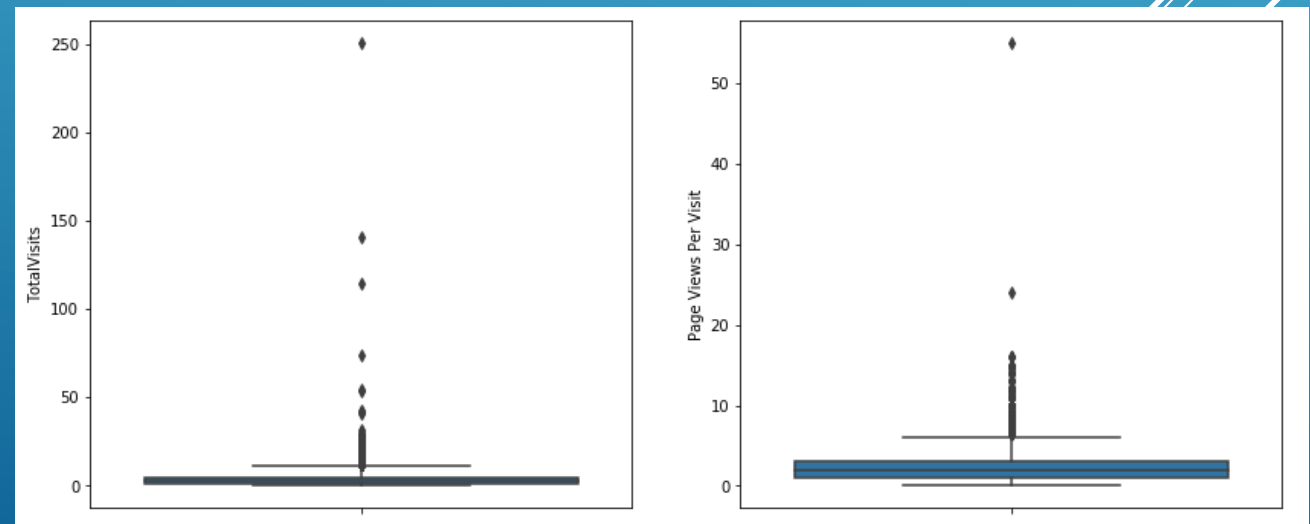


TotalVisits, Total Time spent on website and Page views per visit seem to be having the highest conversion rates.




Histogram of TotalVisits, Total time spent and page views per visit

## OUTLIER DETECTION



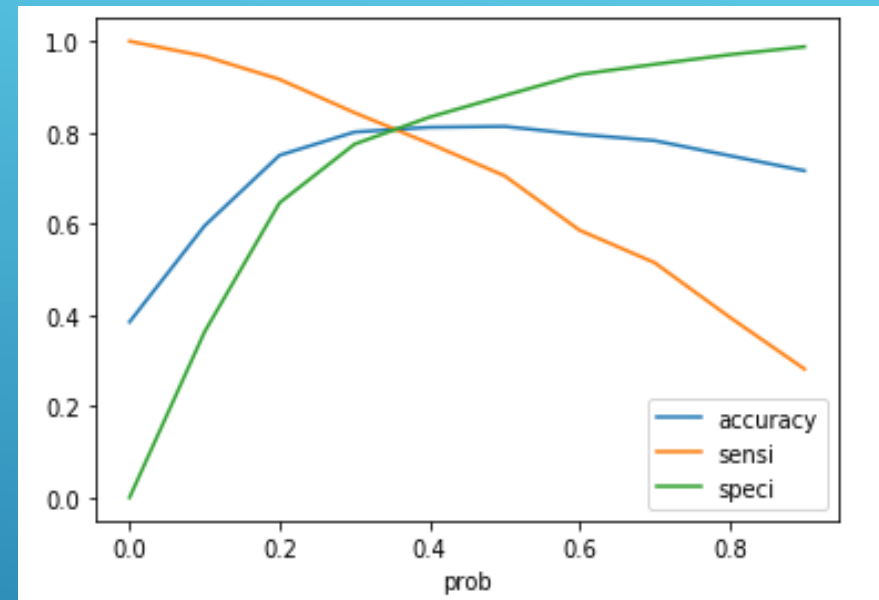
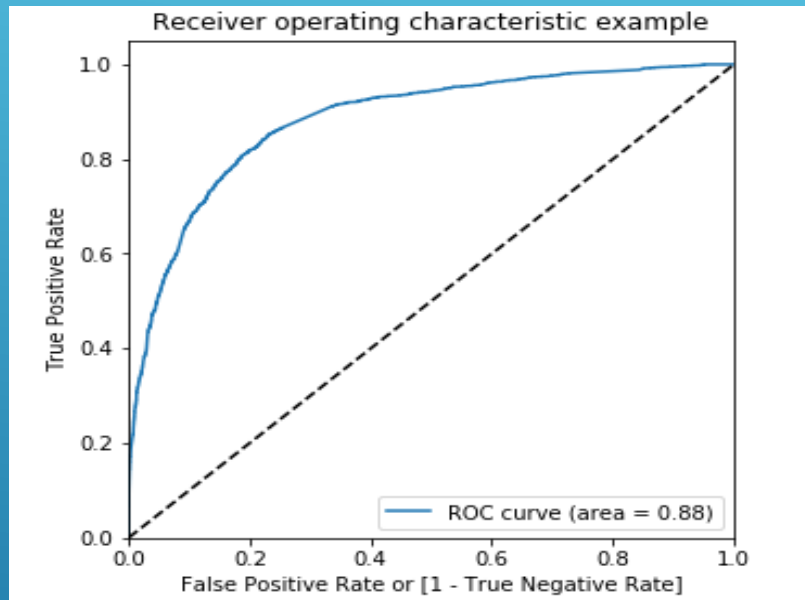
# DATA PREPARATION

- The binary variables i.e. variables having Yes/No as a category were converted into 0's and 1's
  - Dummy variables were created for the features having more than 2 categories
  - Outliers in TotalVisits and PageViewsPerVisit were treated
  - The data was split into testing and training data based on a 30-70 ratio
  - MinMax Scaler was used to scale down the numerical variables and bring them onto the same scale
- 
- A series of three parallel white diagonal lines extending from the bottom right towards the top right of the slide.

# DATA MODELLING

- Used statsmodels to build a logistic regression model based on the data which we preprocessed in the earlier step
- Using RFE, selected the 20 best features out of all the features
- We build model with these 20 features & eliminated the features having insignificant p value and high VIF by removing the variable whose p- value is greater than 0.05 and VIF value is greater than 5
- After training the model, we made predictions on the test data set
- Plotted the confusion matrix and the overall accuracy was around 81% on the training set while the accuracy was 80% on the test set

# ROC CURVE



- After analyzing the graph above, we can assume that the optimal cut-off point is around 0.35
- We can say that our model is a good one as the plotted line is pulling more towards the top left of the plot, maximizing the area under the curve and the line is rising faster as well



# EVALUATION METRICS FOR TRAIN AND TEST DATASETS

For Training Dataset :-

- Accuracy - 0.8132
- Sensitivity - 0.7519
- Specificity – 0.8809
- Confusion Matrix – 

3425	463
717	1715
- Precision – 0.7874
- Recall – 0.7051

For Testing Dataset :-

- Accuracy – 0.8073
- Sensitivity – 0.8047
- Specificity – 0.8088
- Confusion Matrix – 

1346	318
204	841
- Precision – 0.7256
- Recall – 0.8047



# CONCLUSION

- While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
- Accuracy, Sensitivity and Specificity values of test set are around 80%, 81% and 80% which are approximately closer to the respective values calculated using trained set.
- Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 80%
- The top 3 variables that contribute to lead conversion rate are:
  - Total Time spent on Website
  - Lead Origin\_Lead Add form
  - What is your current occupation\_ Working Professional
- Hence through the overall process we derive a model where the objective of solving the business problem is successfully fulfilled