

Prepared by :-  
Rohitkumar Vora &  
Rajeev Ranjan Kashyap

# **Lead Scoring Case Study Analysis & finding**





# Agenda

1. Problem Statement
2. Data validation
3. Missing value treatment
4. Analysing with different Variables
5. Outliers handling
6. Dummy variable and train-test management
7. Model Building
8. Model Evaluation
9. Conclusion



# I. Problem Statement

- An education firm named X Education has appointed us to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company requires to build a model wherein we need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.



## 2. Data validation

- First of all we have imported all the libraries required to build the logistic regression model for assigning a score to every lead which are likely to be converted where we can assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.
- We further proceeded by importing the dataframe and inspected its size and shape.
- Using `.describe()` method we got the mean, median and standard deviation for the numerical variables.
- After this we have checked for the null values or missing values.



### 3. Missing value treatment (1 of 2)

- We observed that there is a value "select" in many columns which is where customer have not selected any option. So, here we have converted "select" to "NaN".
- After this we calculated the percentage of missing values and dropped those columns which have more than 50% missing values and having only one category value as these would not have any impact in further prediction.
- We also dropped all the sales related columns as they are not necessary for the model and also having high missing values.

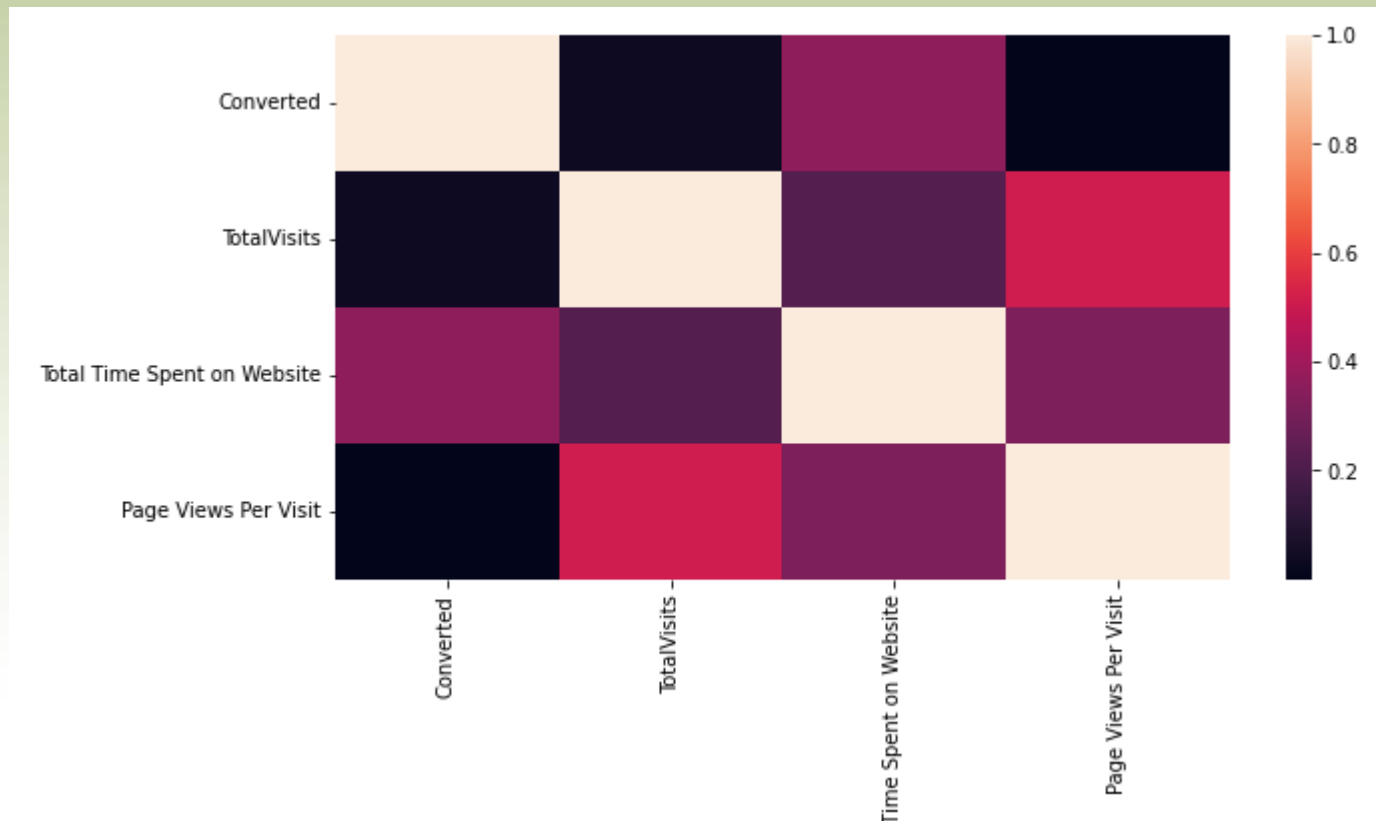


### 3. Missing value treatment (2 of 2)

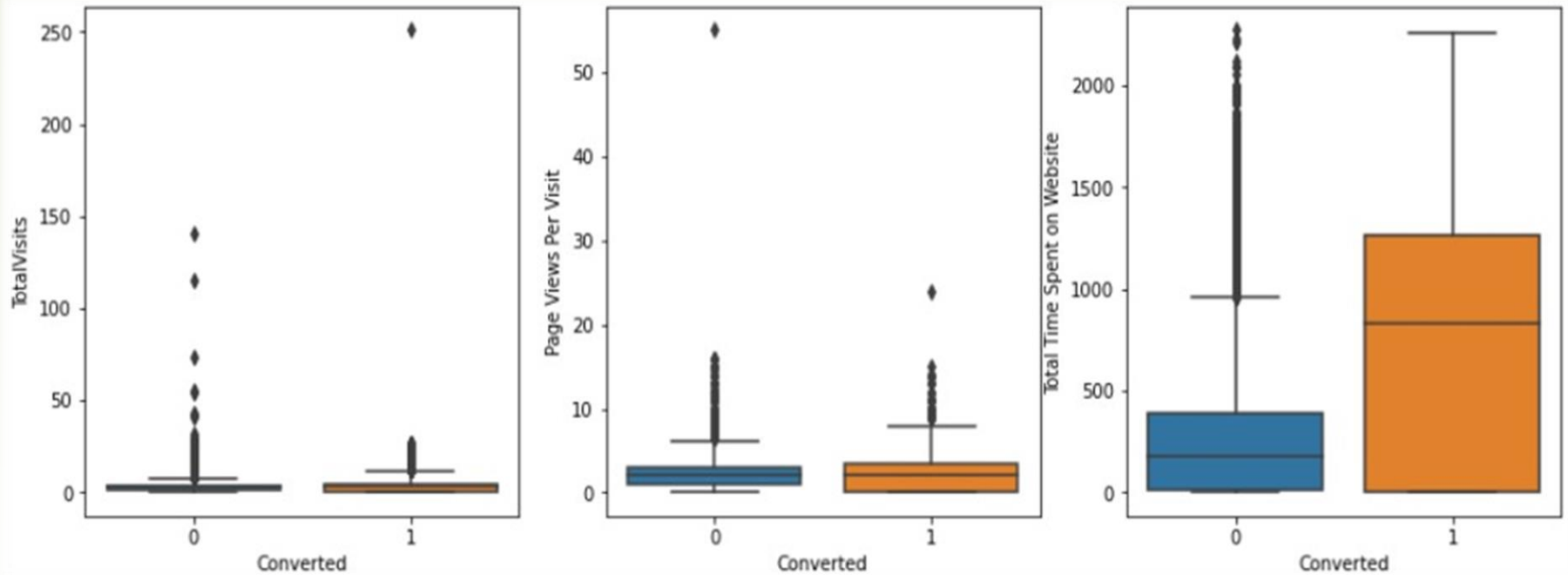
- We have dropped 'What matters most to you in choosing a course' column as it doesn't seem to be much significant after looking at its values.
- We checked the value counts of the categorical variables and also checked the skewness, and dropped the columns with high skewness.
- Again we checked for the missing values after dropping the unnecessary variables.
- Hereafter we imputed the missing value of the categorical columns with its mode and continuous variables with median.

## 4. Analysing with different Variables

- After missing value treatment we have calculated the correlation between different variables which has given the following result:



## 5. Checking outlier







## 6. Dummy variable and train-test management

- Once the null values are treated and we found that there are no missing values anymore, we then have created the dummy variables for the variables having two or more distinct categories in order to correctly analyze attribute variables.
- After this we have split the data into train and test data set, where train dataset consists of 70% of the data and test data contains 30% of the data, and we have then used the min-max scaler to scale the numerical variables of the training dataset.
- Once scaling is done, we will proceed to build the model.

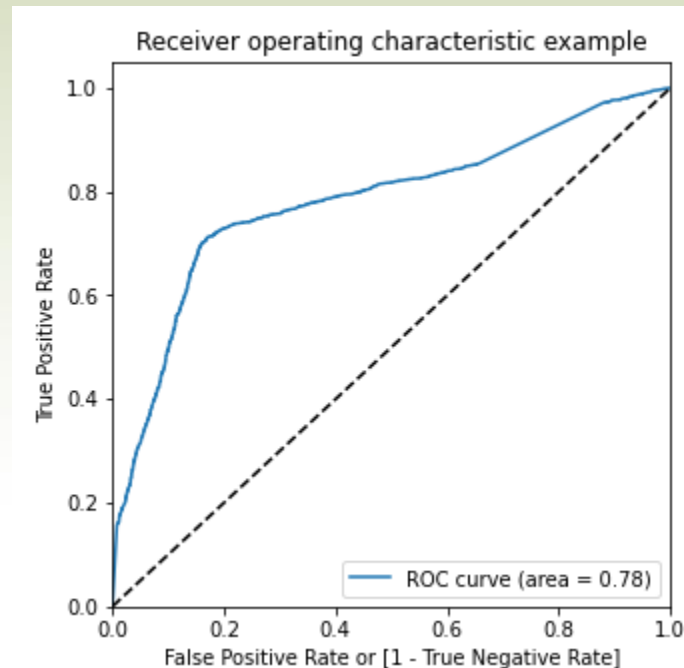


## 7. MODEL BUILDING

- We have used statsmodel and sklearn library here in order to look at the statistics part and to be able to select the features better.
- As statsmodel does not add a constant or an intercept by default we have added a constant and have provided the family as binomial.
- Then we will fit the model and will check its summary to get the number of observations and features.
- We checked for the P-values and coefficients, then doing some course tuning, we have eliminated the variables and selected the best 18 variables.
- After this we will be using recursive feature elimination to eliminate the variables one by one, this we call the shortlisting of the variables.

## 8. MODEL EVALUATION

- In evaluating any model we check how good our model is. For this we have created a confusion matrix first and checked the overall accuracy.
- After this we have calculated the true positive, true negative, false positive and false negative values to calculate the Sensitivity and Specificity and also drew the ROC curve.



A background image showing a collage of diverse people's faces, including a man and a woman, with a focus on a smiling woman in the foreground.

## 9. CONCLUSION

- From the final model developed we are able to generate a lead score which shows how likely the lead is getting converted.
- From the final predictions it is evident that the people with lead score more than 30 are likely to get converted and sales team can concentrate people with lead score. more than 30 to convert them to a lead.



- Thank you