Lead Score Case Study Assignment

UMAL BHOLE, UCHIT KUMAR & TEJAL PATEL

BATCH ID 4460, MARCH 2023

EXECUTIVE PROGRAM IN DATA SCIENCE

Step 1. Business & Data understanding

- This assignment is based on a case study of an education company named X Education which sells online courses to industry professionals.The company has published & markets its course on different websites along with Google. The people when hit these sites, they may or may not browse through these courses. Those interested in courses tend to fill up the forms where they provide their personal details. Accordingly, these people are considered as leads.
- The company then targets these leads & then approaches them through mails/calls.
- The current lead conversion rate is poor i.e around 30%. In order to increase the lead conversion rate, the identification of potential leads is important which are called ‘Hot leads’. Accordingly, the focus of the company shifts towards these hot leads instead of other leads.
- In the given data set of 9000 data points comprising various attributes, the target variable is given which is column "Converted’ which tells whether past lead was converted or not:

    a) for value=1, converted

    b) for value=0, not converted

- The objective of the case study is to build logistic regression model to assign a lead score between 0 & 100 to every lead which can be further used to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted. The model should also take the future requirements of company into consideration & accordingly should adjust.

Step 2. Data Cleaning

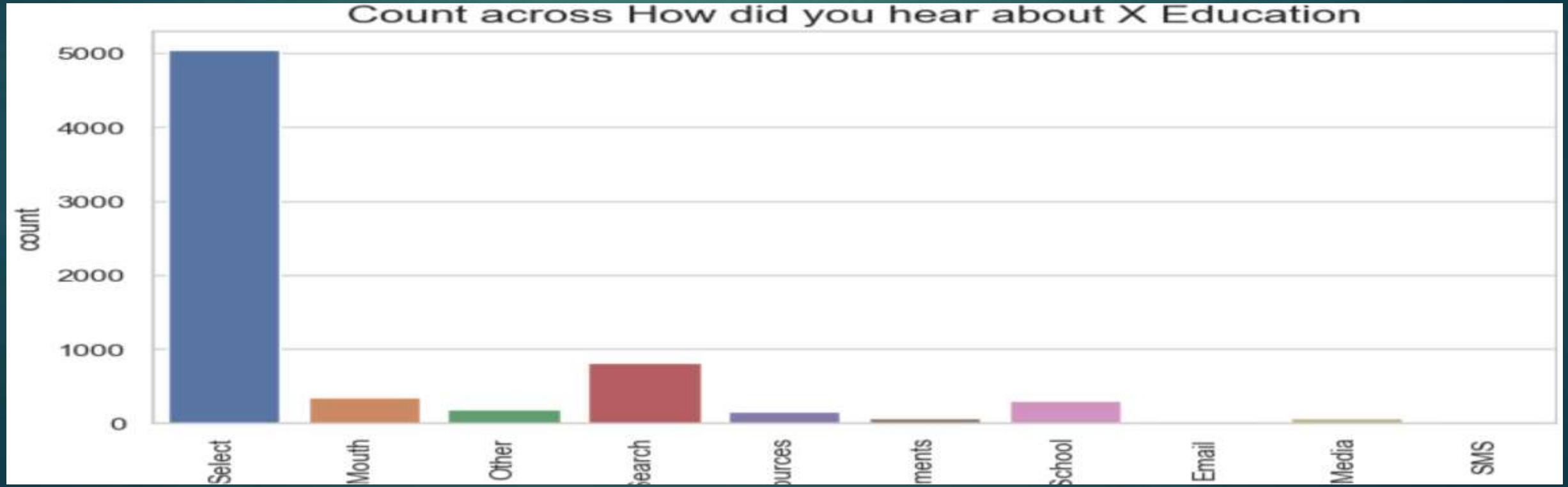
- First the Missing Values are identified in the data set. Also, the data types of the columns are checked to ensure that they are correct
- The columns are first identified on the basis of their type whether they are numeric or categoric
- For missing values, firstly the columns which have more than 50% missing values are identified & then dropped from data set as they have no significance

Step 3. Data Visualization

In order to understand the correlation/relation of different columns, the visualization is understood through count plots, pair plots

1. Relationship of count of column'How did you hear about X education' mapped.

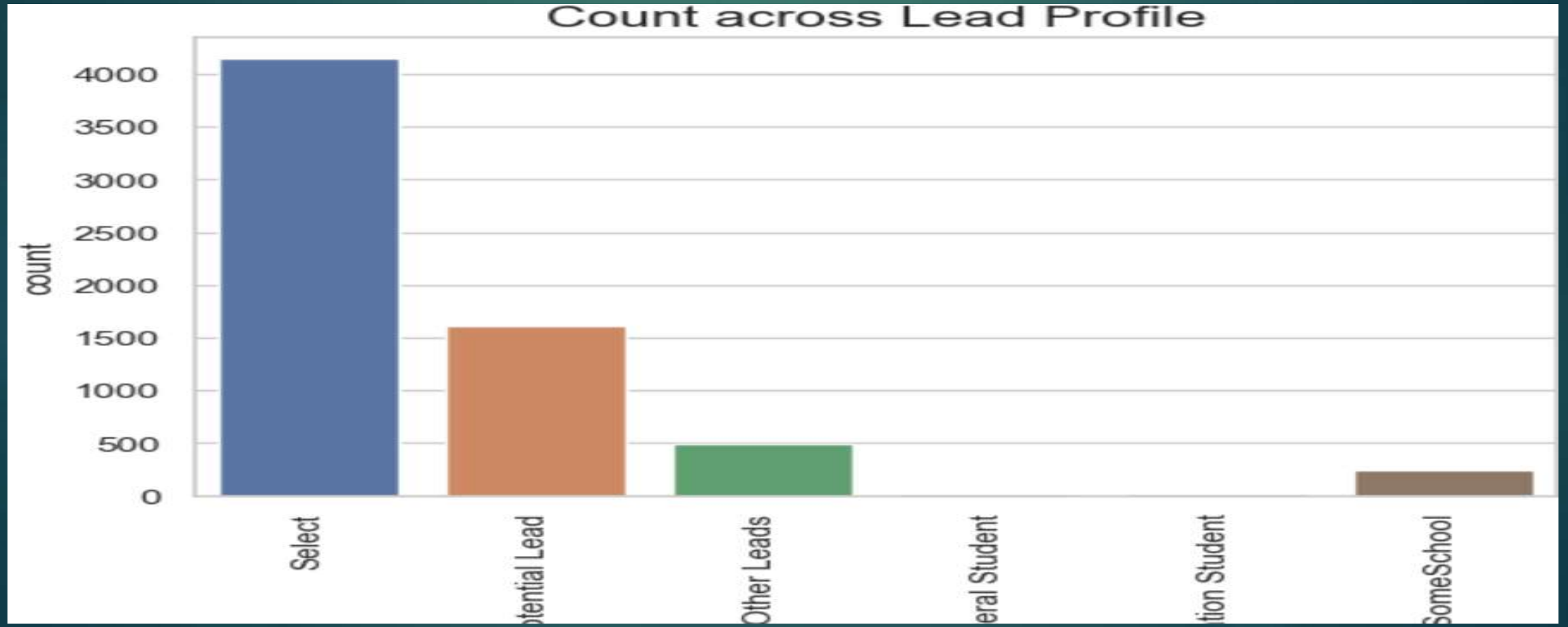
- Conclusion – Apart from Select category which is of no use, the parameter Online Search has the highest count.



Step 3. Data Visualization-contd..

2. Relationship of count of column'Lead Profile' mapped.

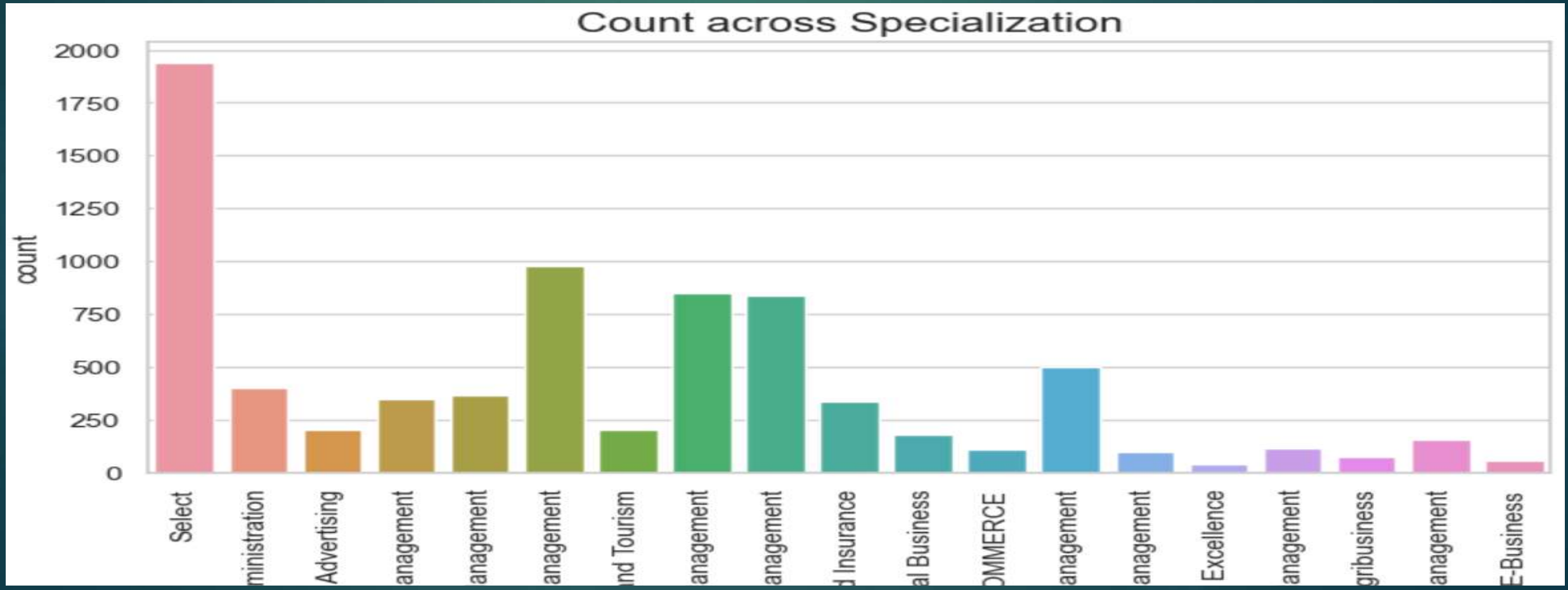
- Conclusion – Apart from Select category which is of no use, the parameter Potential Lead has the highest count.



Step 3. Data Visualization-contd..

3. Relationship of count of column 'Specialization' mapped.

- Conclusion – Apart from Select category which is of no use, the parameter IT Projects Management has the highest count.



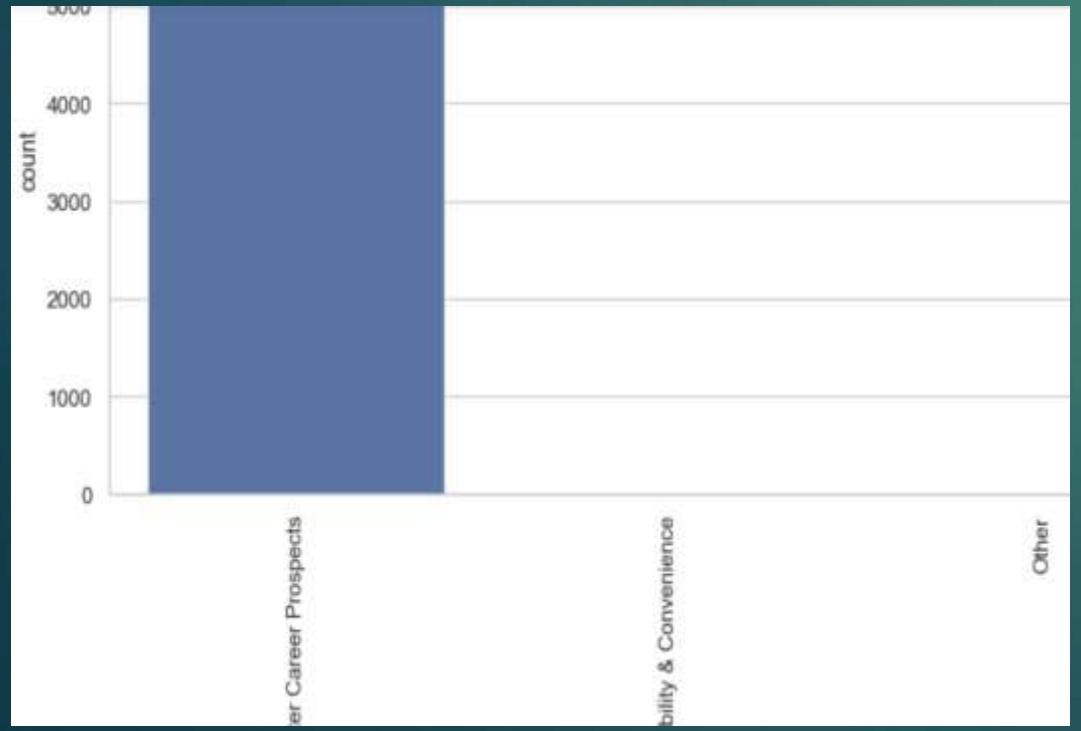
Step 3. Data Visualization-contd..

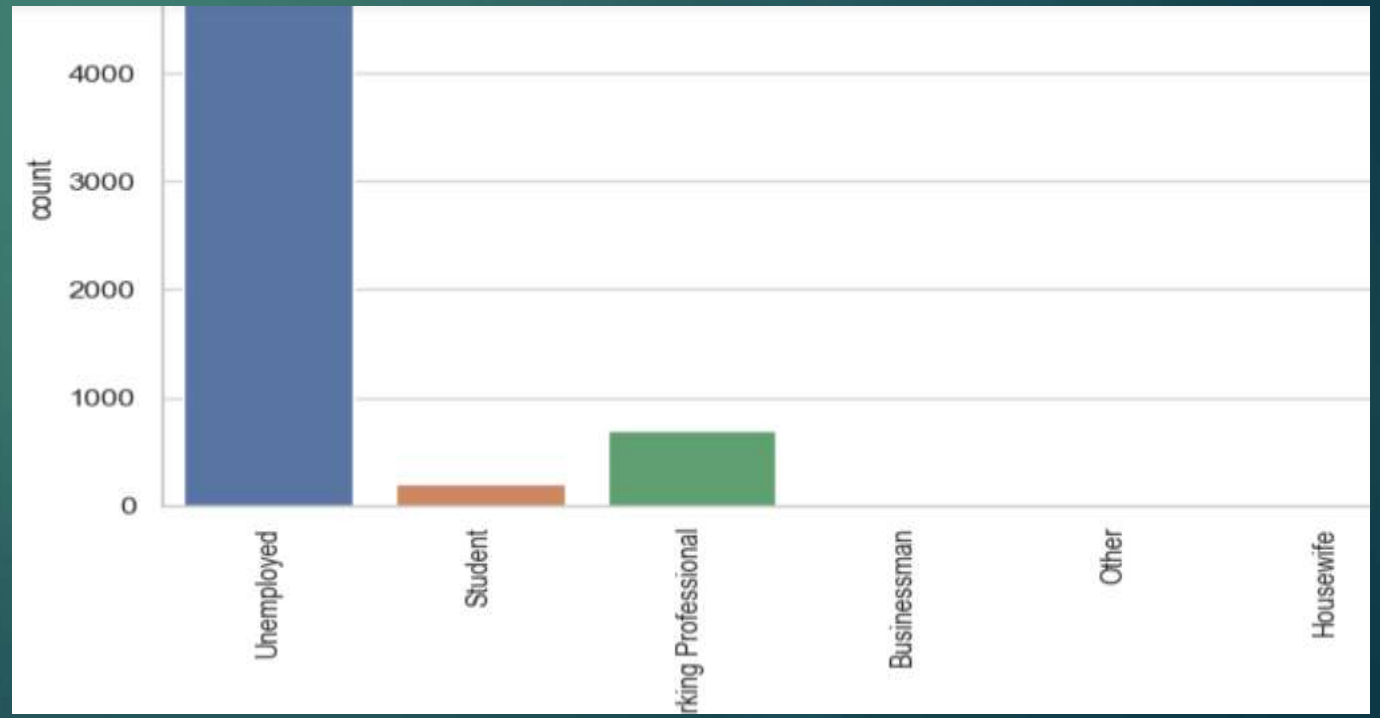
4. Relationship of count of column ’ What matters most to you in choosing a course’ mapped.

- Conclusion – The parameter Better Career Prospects has the highest count.

5. Relationship of count of column ’ What is your current occupation’ mapped

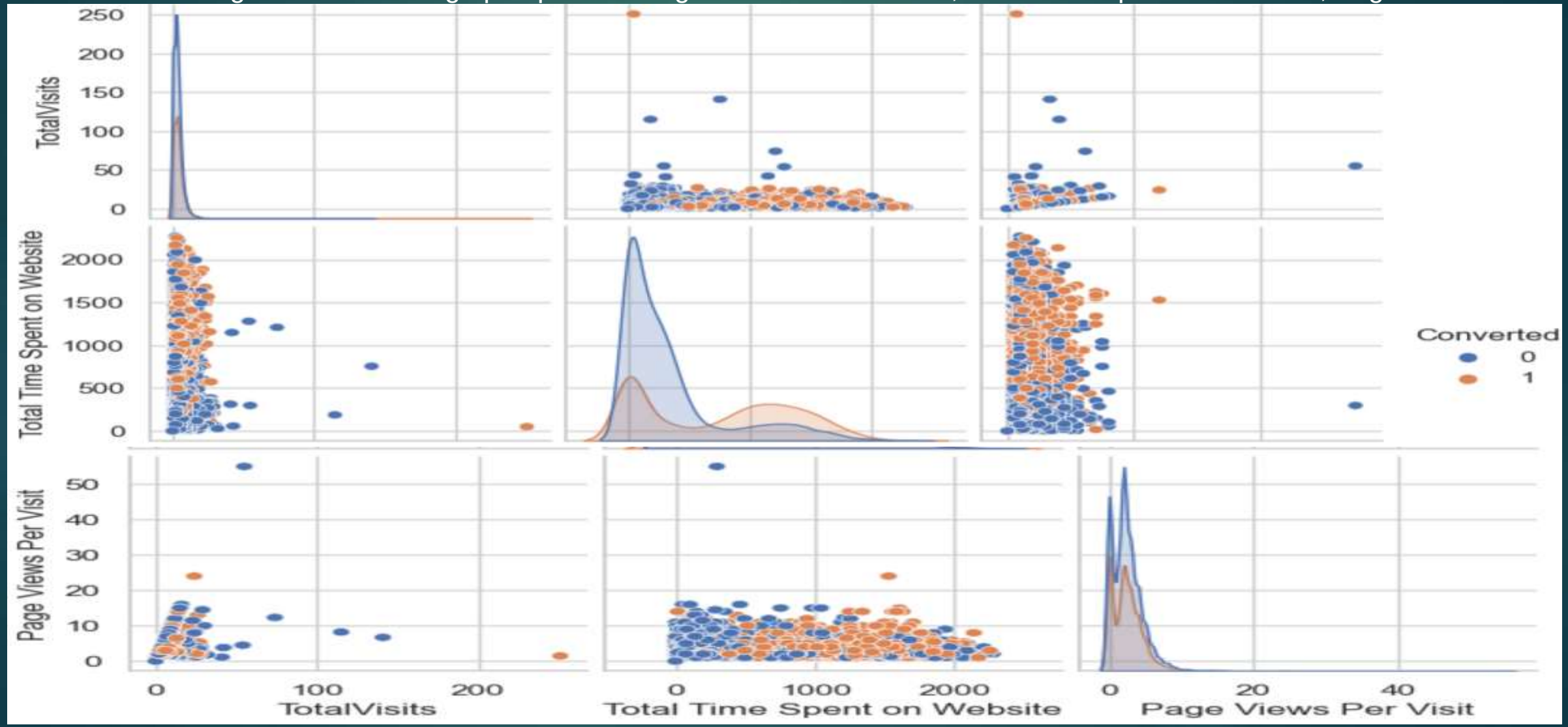
- Conclusion – Apart from Select category which is of no use, the parameter Unemployed has the highest count





Step 3. Data Visualization-contd..

6. Correlation wrt target variable through pair plots among columns- Total Visits, Total Time Spent on Website, Page Views Per Visit



Step 4. Model building

- Dummy Variable : Creation of Dummy variables for dealing with categorical data. Accordingly, columns have been selected & then first column level dropped.
- Train Test split : Splitting the complete dataset in 70% TRAIN AND 30% TEST
- Scaling : Through Min Max Scaler, the scaling is performed by scaler fit transform
- Model Building : The Model building is done with help of RFE (Recursive Feature Elimination). Eventually 4 different models have been built and columns with high P value & VIF have been removed. So 4^{th} model has selected as final model.
- Model evaluation - Calculating Confusion Matrix, Sensitivity & Specificity
- Calculating ROC curve:

  Conclusion – area under curve is 0.86 which is good
- Calculating Precision & Recall

Step 5. Final conclusion

- 0.42 has been taken as final cut off threshold value. The overall accuracy came out to be 80%.

- More budget/spend can be done on Welingak Website in terms of advertisement

- Incentives/discounts for providing the reference that convert to lead,encourage clients to provide more references

- Working professionals to be aggresiveky targeted as have high conversion rate and will have better financial situation to pay higher fees

Thank you