Lead Scoring Case Study

Vikramjit Bora and Anjali Agarwal

Upgrad C28 January Batch

14th July 2021

Problem Statement

X Education sells online courses to industry professionals many interested professionals visit their website and browse for courses. The company generates leads through past referrals or visitors providing their contact details. Through the current sales process, only around 30% leads get converted. To improve this, ex education wishes to identify the most promising leads for the sales team.

The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

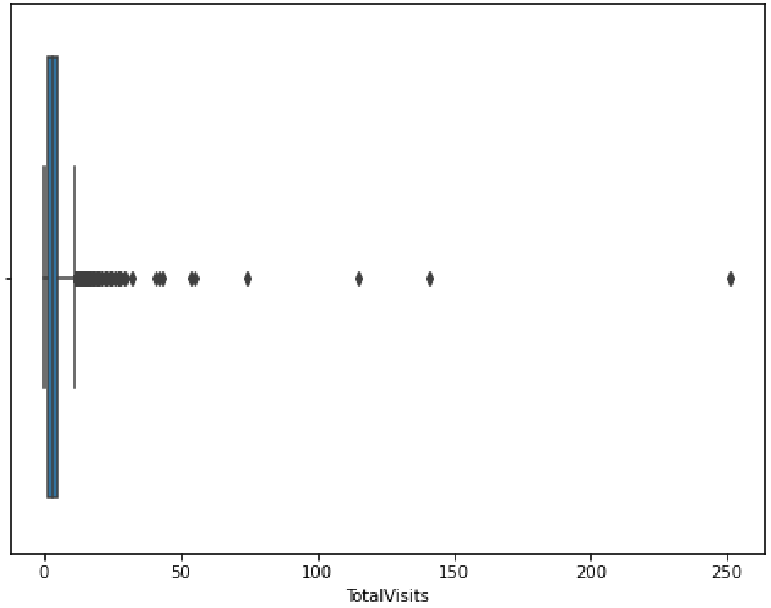
Analysis Approach

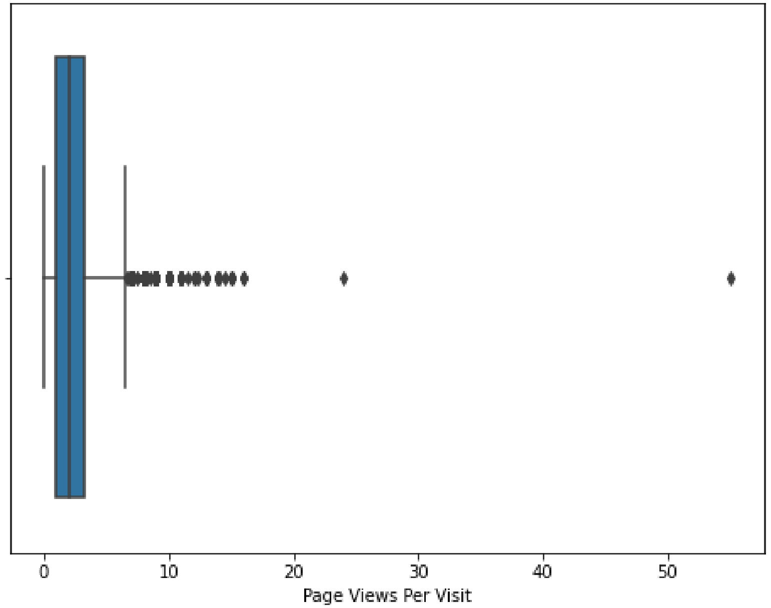
- Below are the steps followed for analysis of this case study:
- 1.Read the Leads dataset
- Performed a non-graphical EDA
- 2. Data Cleaning
  - Missing Value Analysis – removed columns with more than 40% data missing
  - Imputed missing values using Central Tendency
  - Outlier Analysis
  - 3. Checked data types
  - 4. Grouped lower values of categorical variables
  - 5. Checked the Target variable
  - 6. Performed Univariate, Bivariate and Multivariate Analysis(one variable at a time) on Categorical and Numerical variables
  - 7. Encoded categorical variables
  - 8. Checked Correlations
  - 9. Split data into train and test
  - 10. Performed feature scaling
  - 11. Performed feature selection using RFE
  - 12. Model building using Statsmodels
  - 13. Model evaluation by using metrics like Accuracy, Sensitivity, Specificity, False positive rate, Positive predictive value, Negative predictive value, Precision, Recall, ROC curve
  - 14, Finding the optimum cut-off
  - 15. Made predictions on the test set
  - 16. Feature Importance
  - 17. Conclusion
  - 18. Recommendations

Outlier Treatment

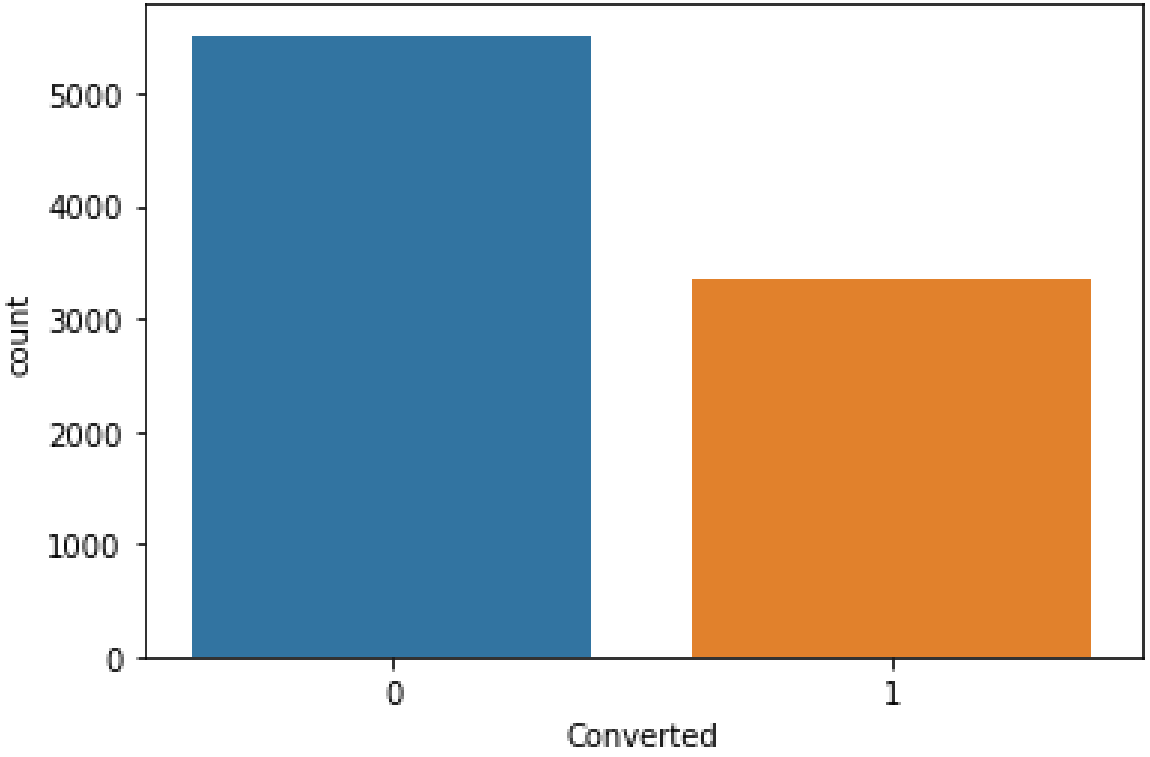
Insights:

- TotalVisits and page views per visit have outliers
- Remove the rows which have values greater than 99th percentile for these columns





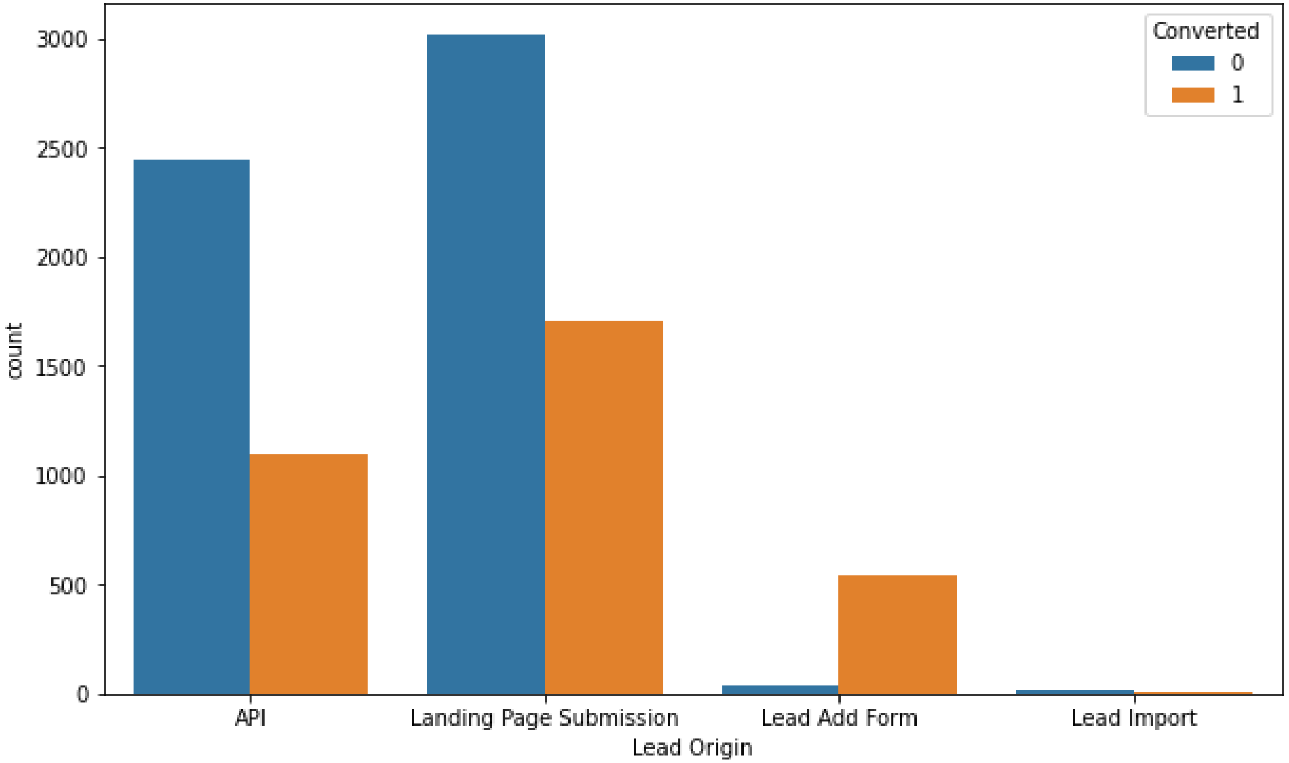
Checking the target variable



Insights:

- Only 37% of the prospects have converted
- We notice that the data is imbalanced

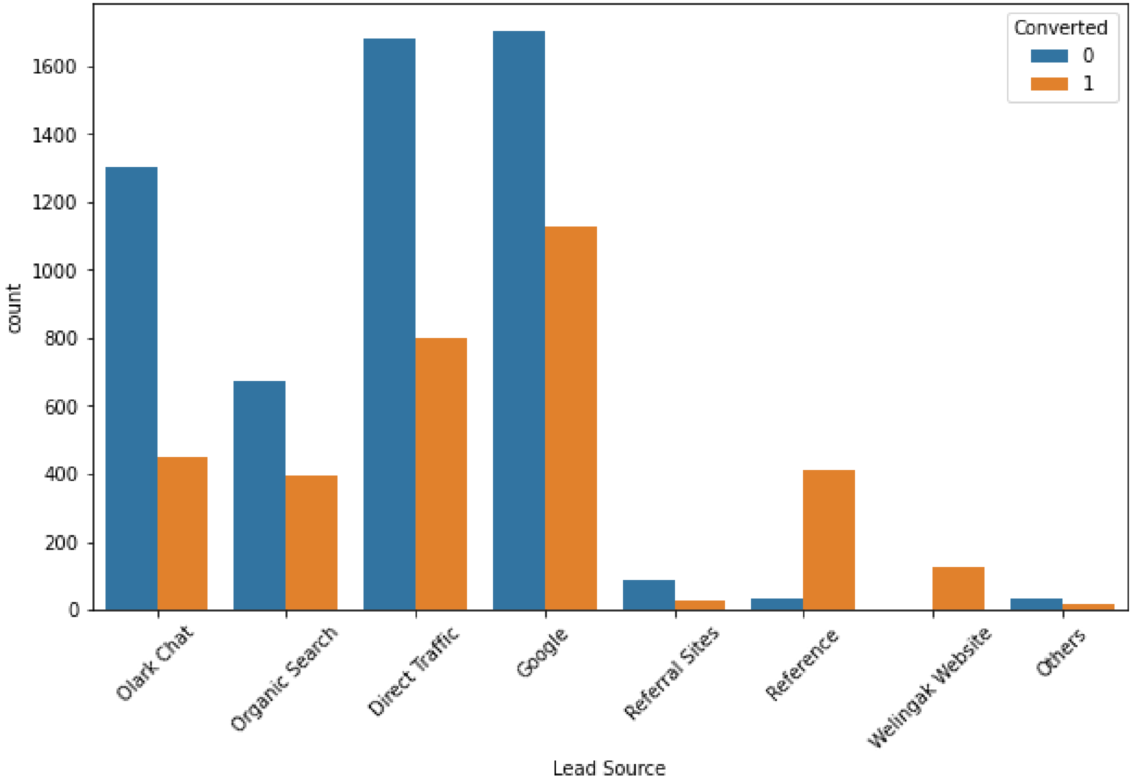
Comparing lead origin versus converted column



Insights:

- More leads are coming through prospects submitting on the landing page

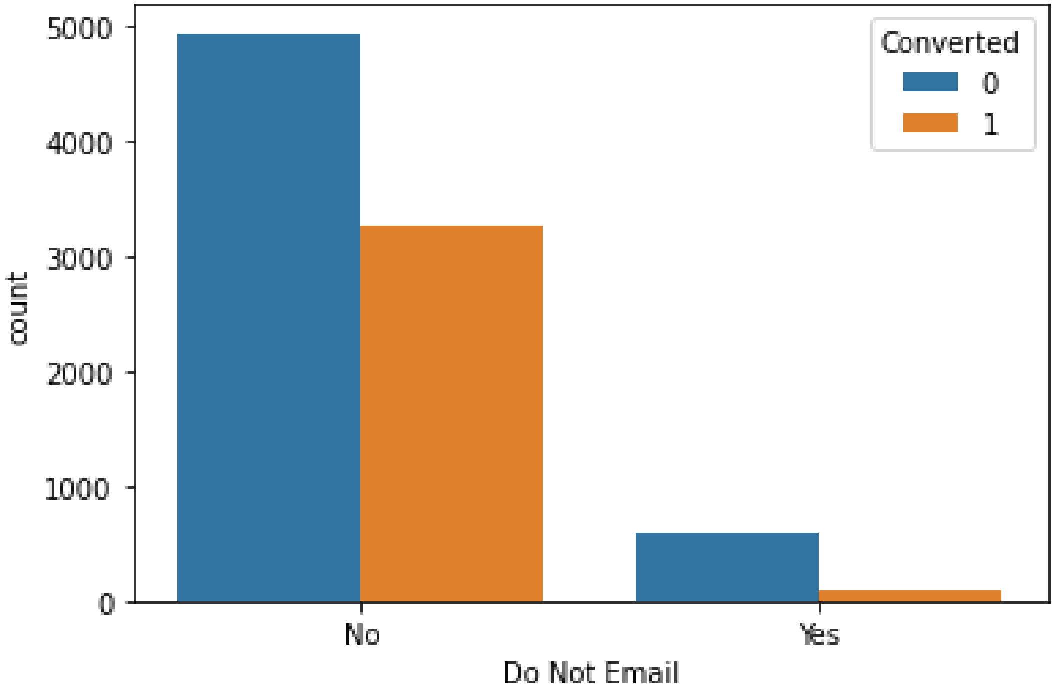
Comparing lead source versus converted column



Insights:

- most leads are coming through Google and direct traffic

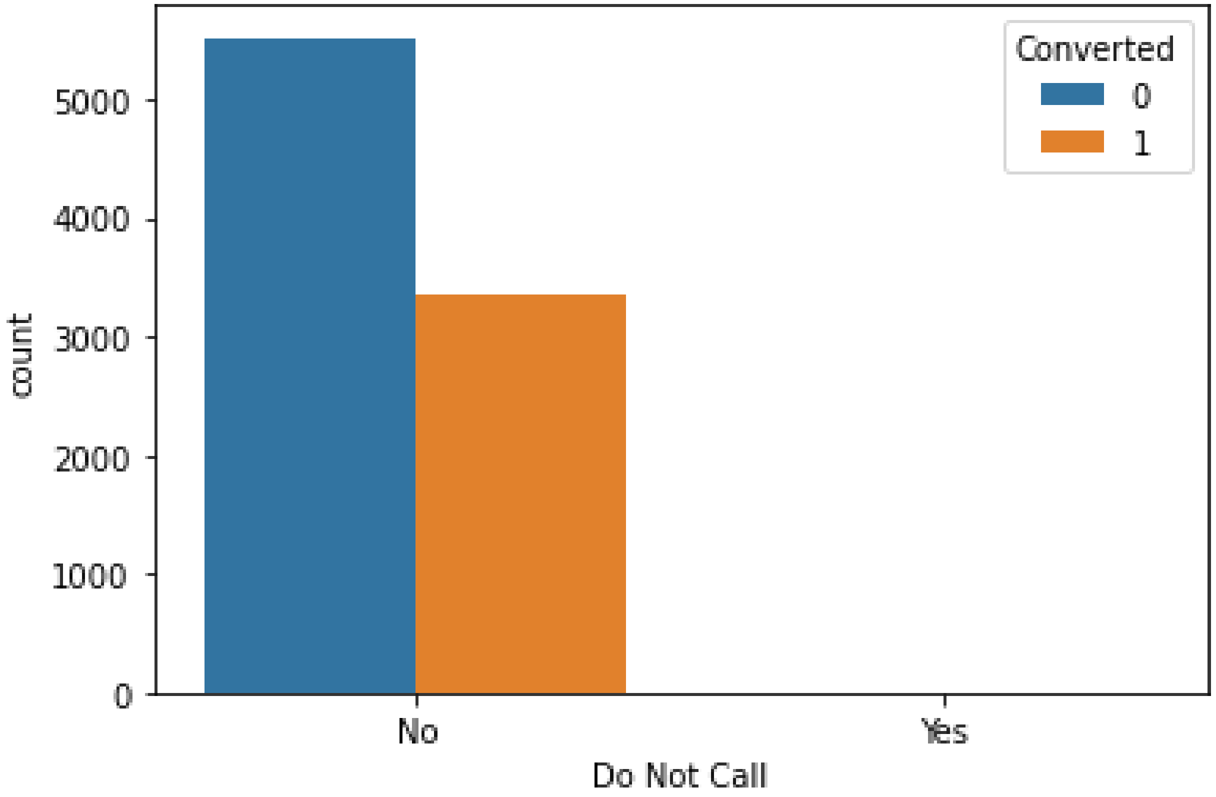
Comparing do not email versus converted column



Insights:

- most customers do not want to be contacted through email about the course
- This column has very low or no variability

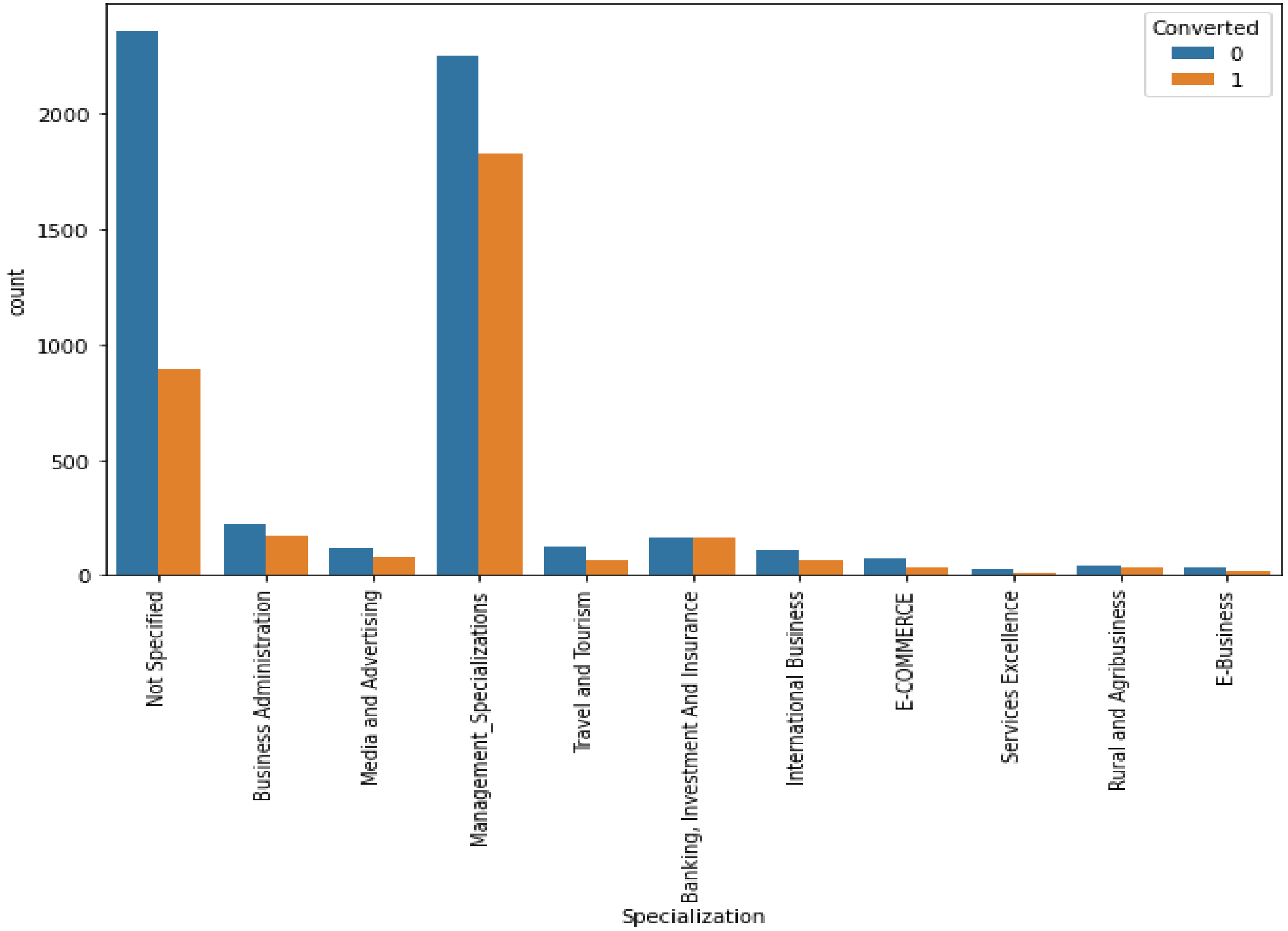
Comparing do not call versus converted column



Insights:

- what most customers do not want to be called about the course
- This column has very low or no variability

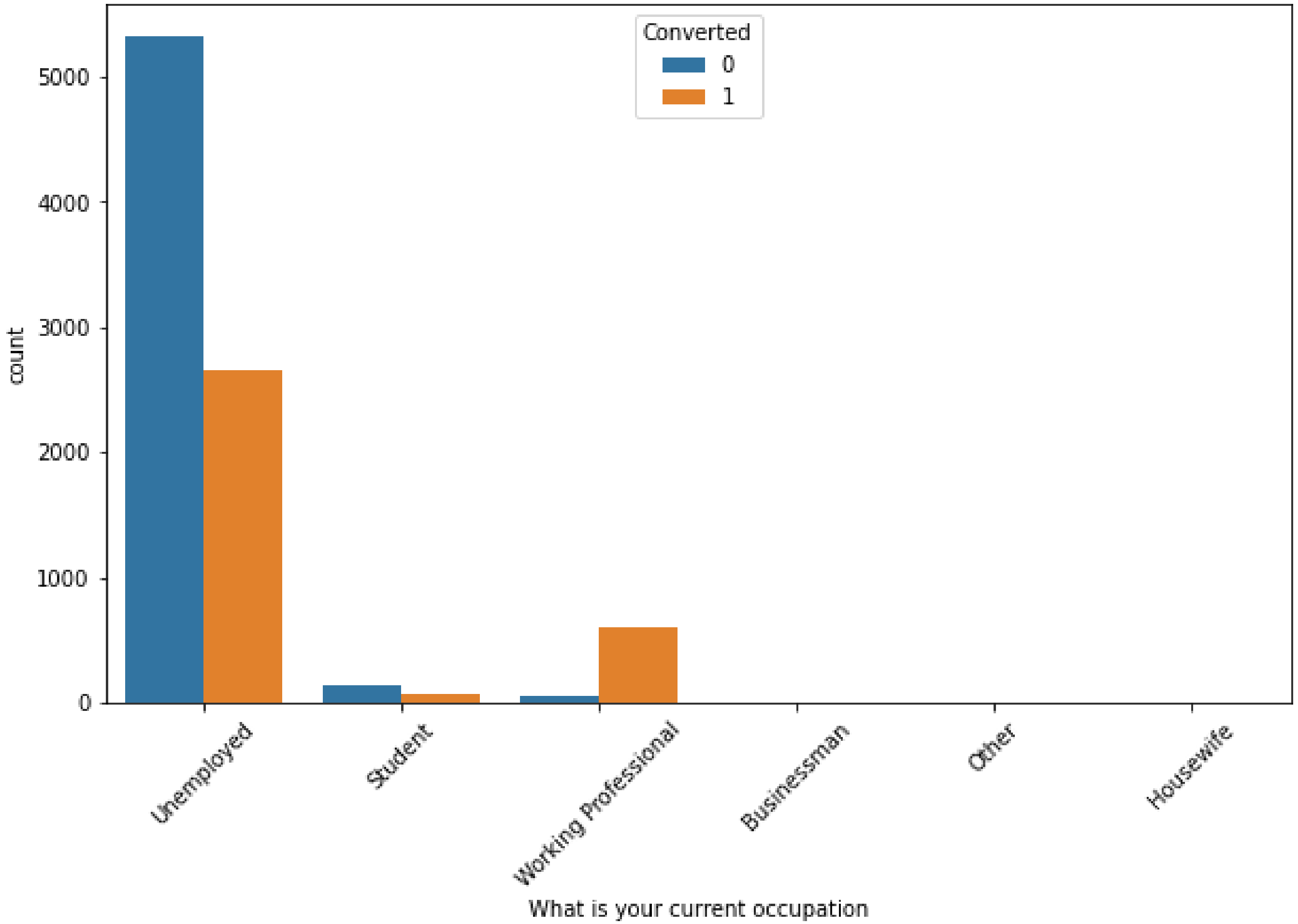
Comparing specialization versus converted column



Insights:

- a sizable number of customers have either chosen not to provide their specialization or they are from management

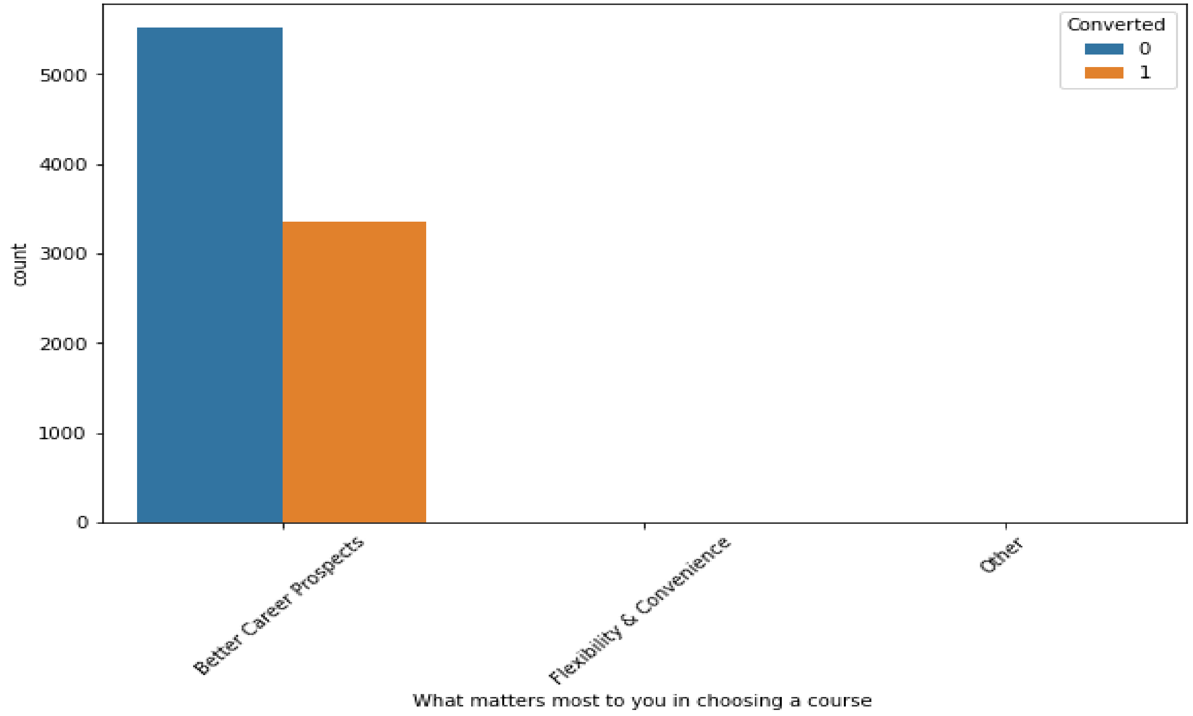
Comparing what is your current occupation versus converted column



Insights:

- a clear majority of customers are unemployed

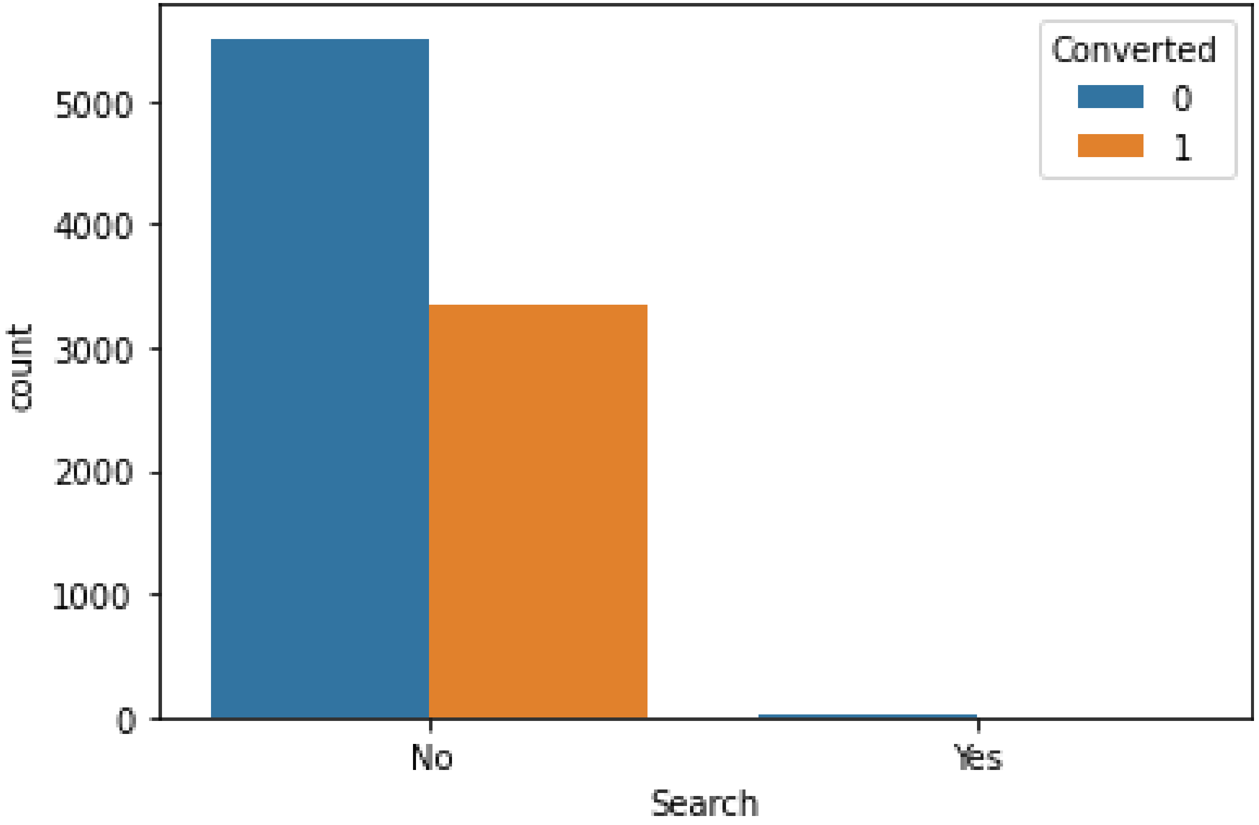
Comparing what matters most to you in choosing a course versus converted column



Insights:

- most of the customer's main motive behind joining a course is to look for better career prospects
- this column has very low or no variability

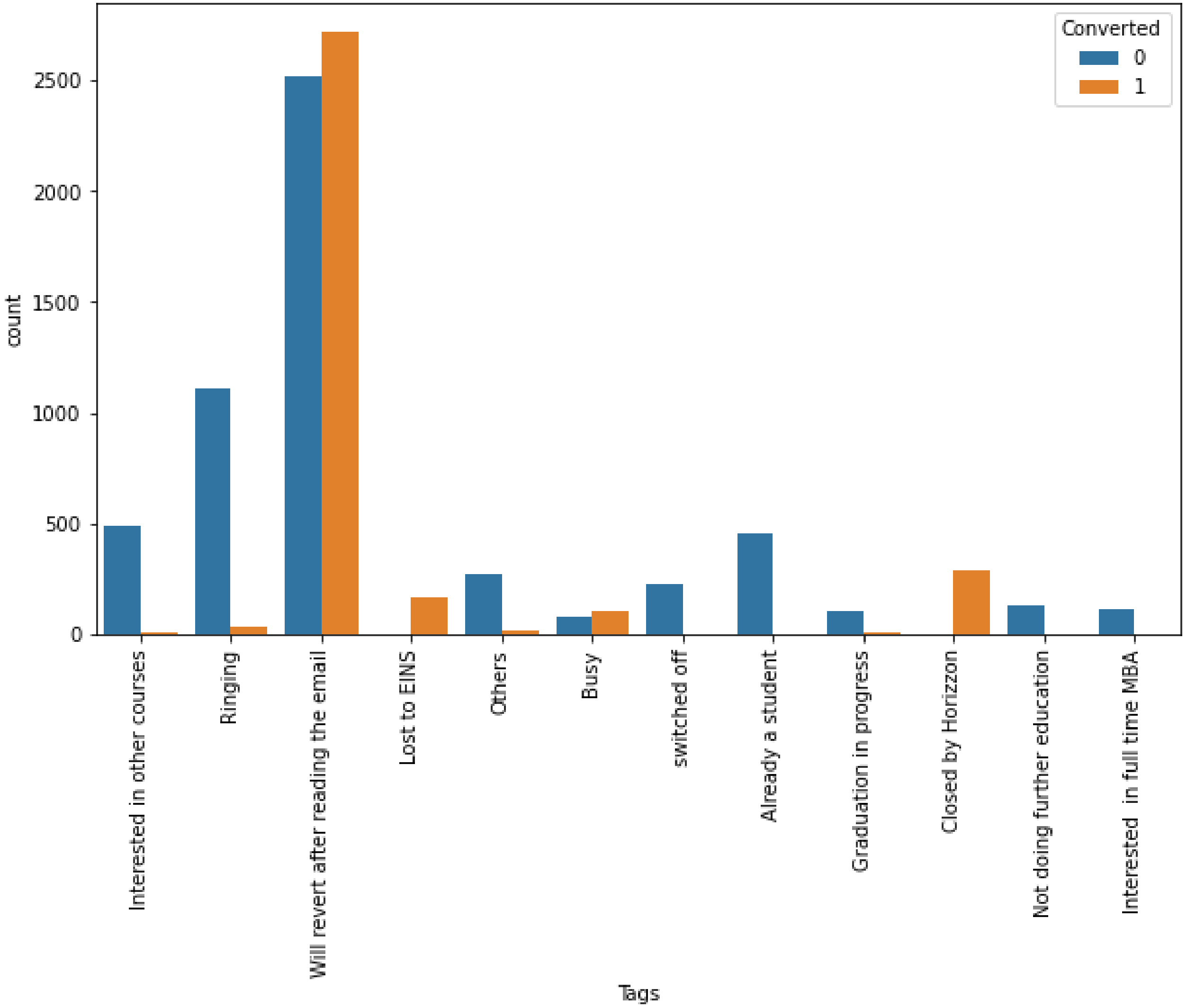
Comparing search versus converted column



Insights:

- most of the customers have not seen the ad of X education
- this column has very low or no variability

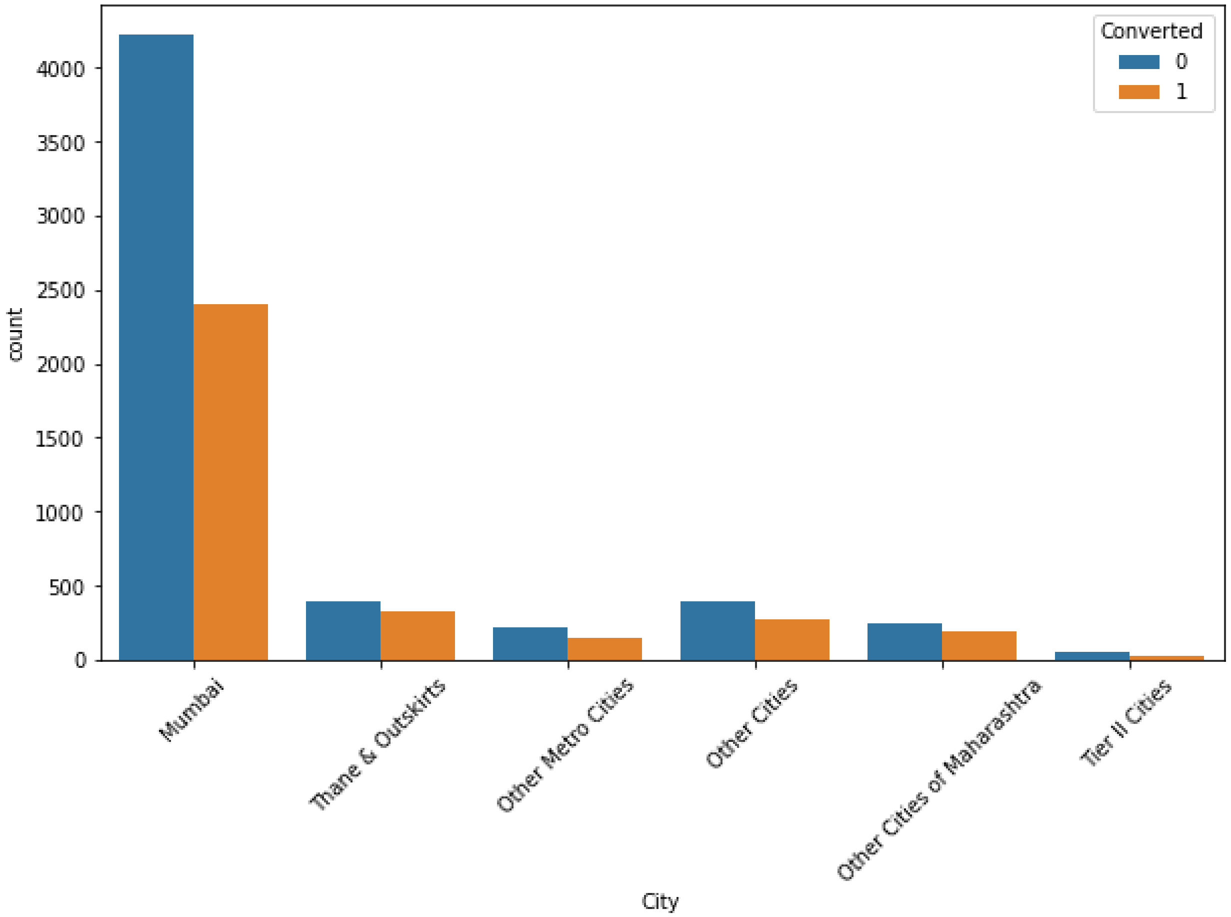
Comparing tags versus converted column



Insights:

- the current status of most of the customers is that they will return after reading the email

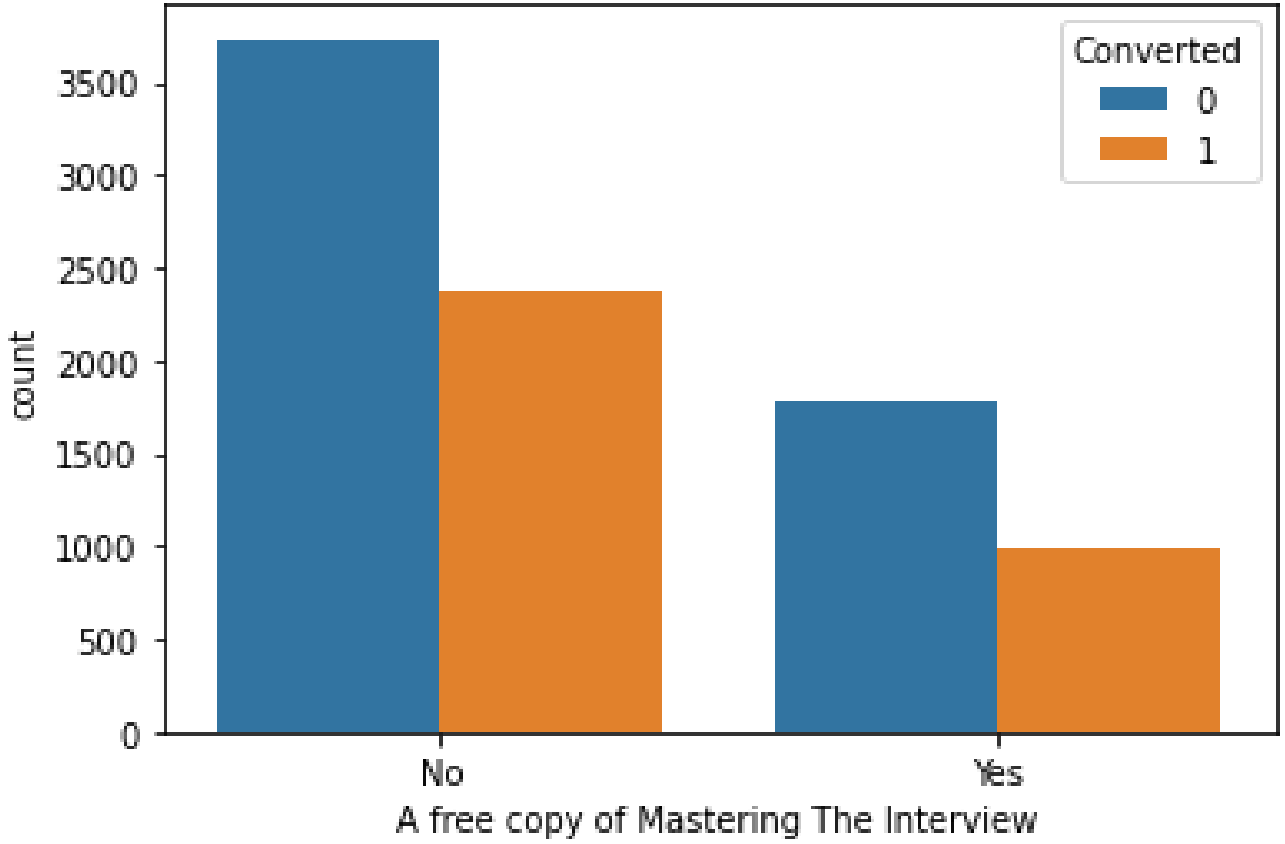
Comparing city versus converted column



Insights:

- most customers are from Mumbai

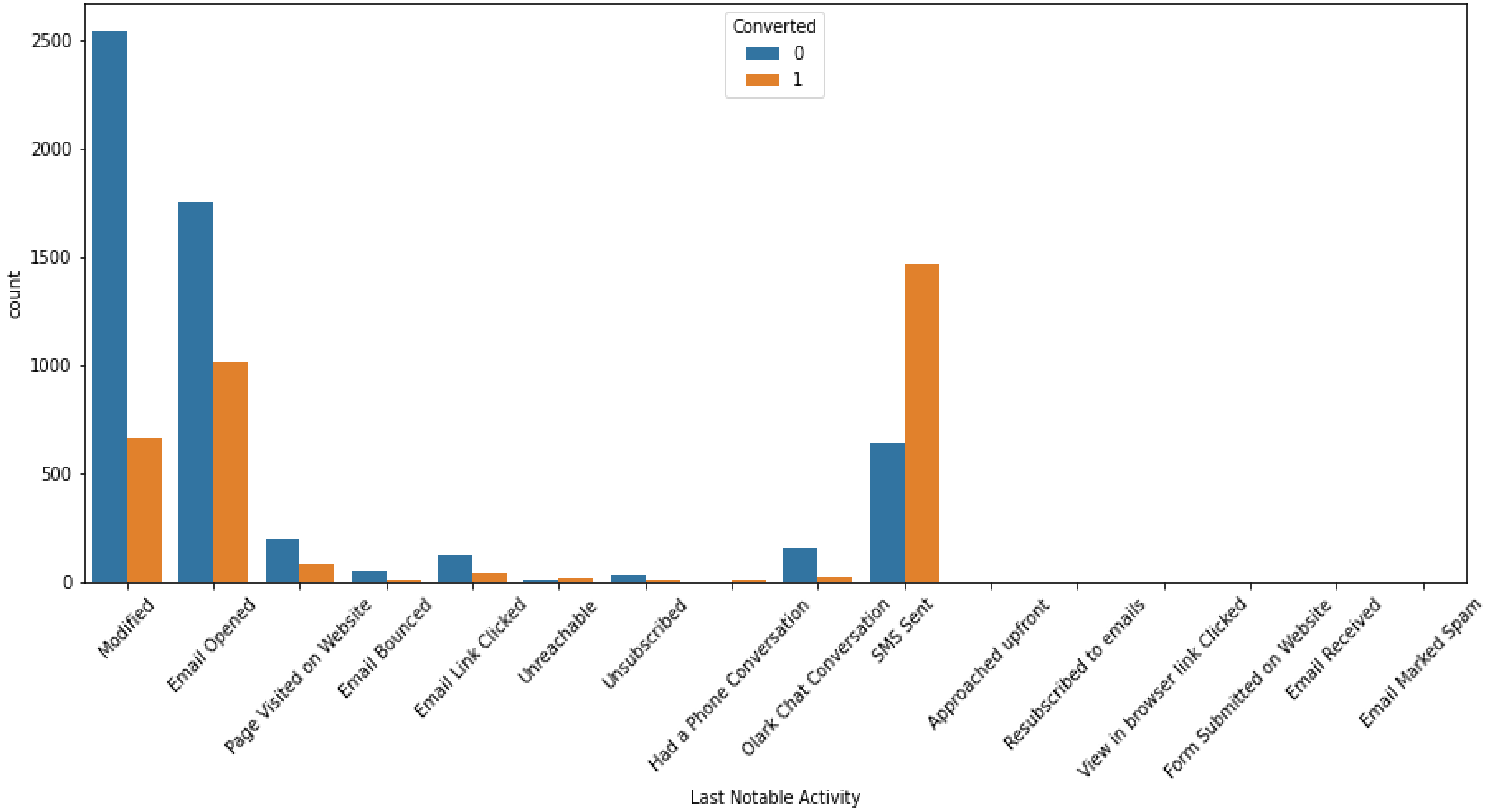
Comparing a free copy of mastering the interview versus converted column



Insights:

- majority of customers do not want a free copy of the mastering the interview

Comparing last notable activity versus converted column



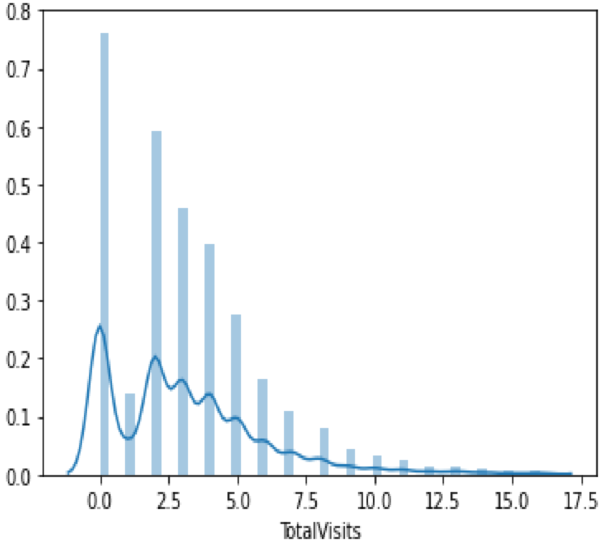
Insights:

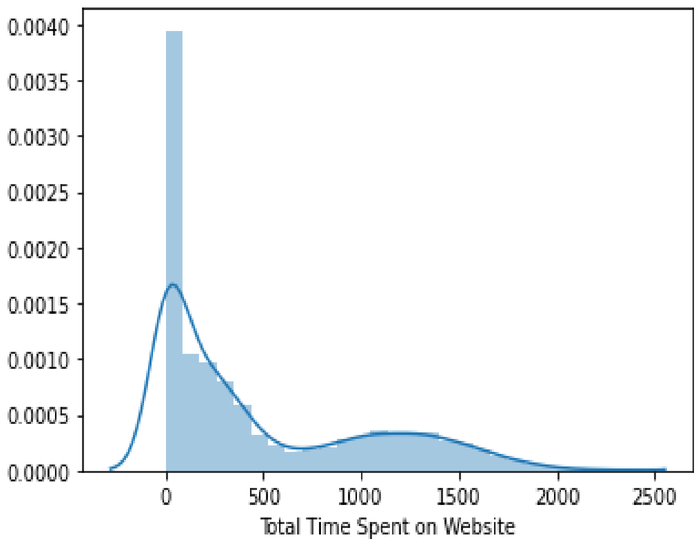
- last notable activity of most of the students is either modified, email opened or SMS sent

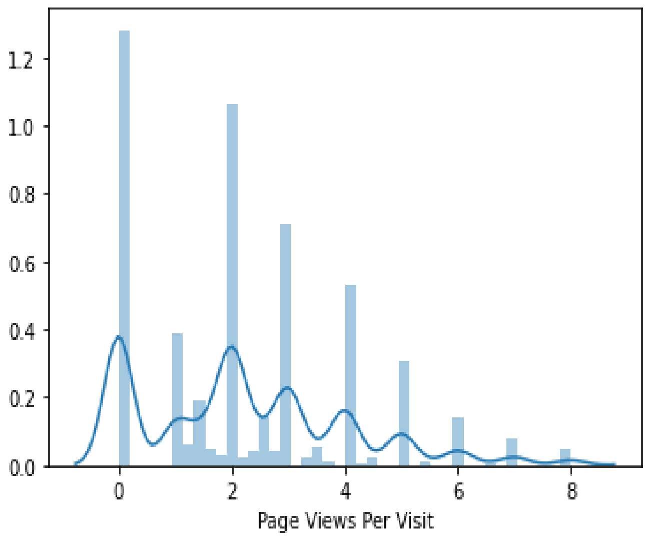
Visualizing the numerical distributions

Insights:

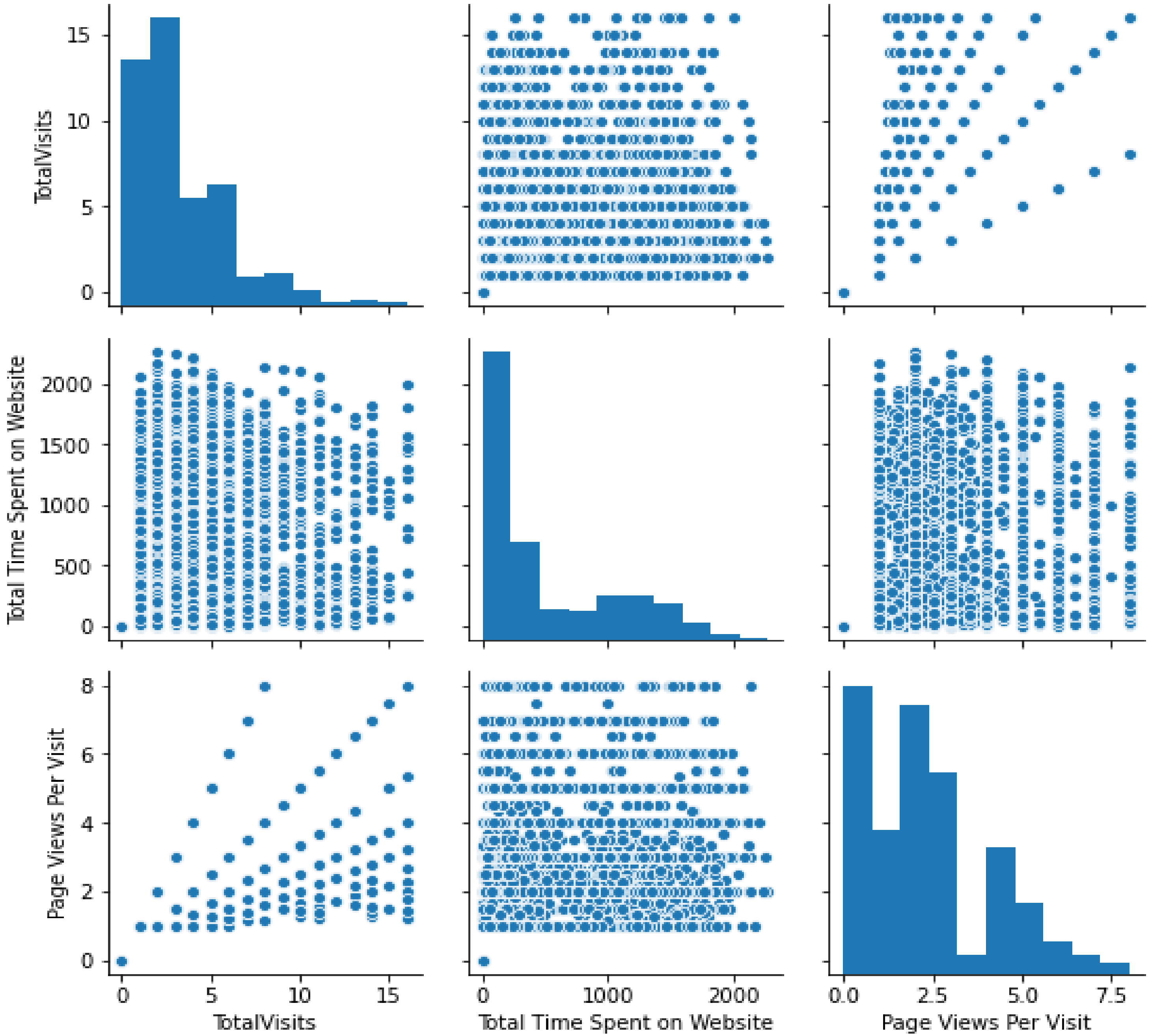
- all three distributions are right skewed that is their mean is greater than mode







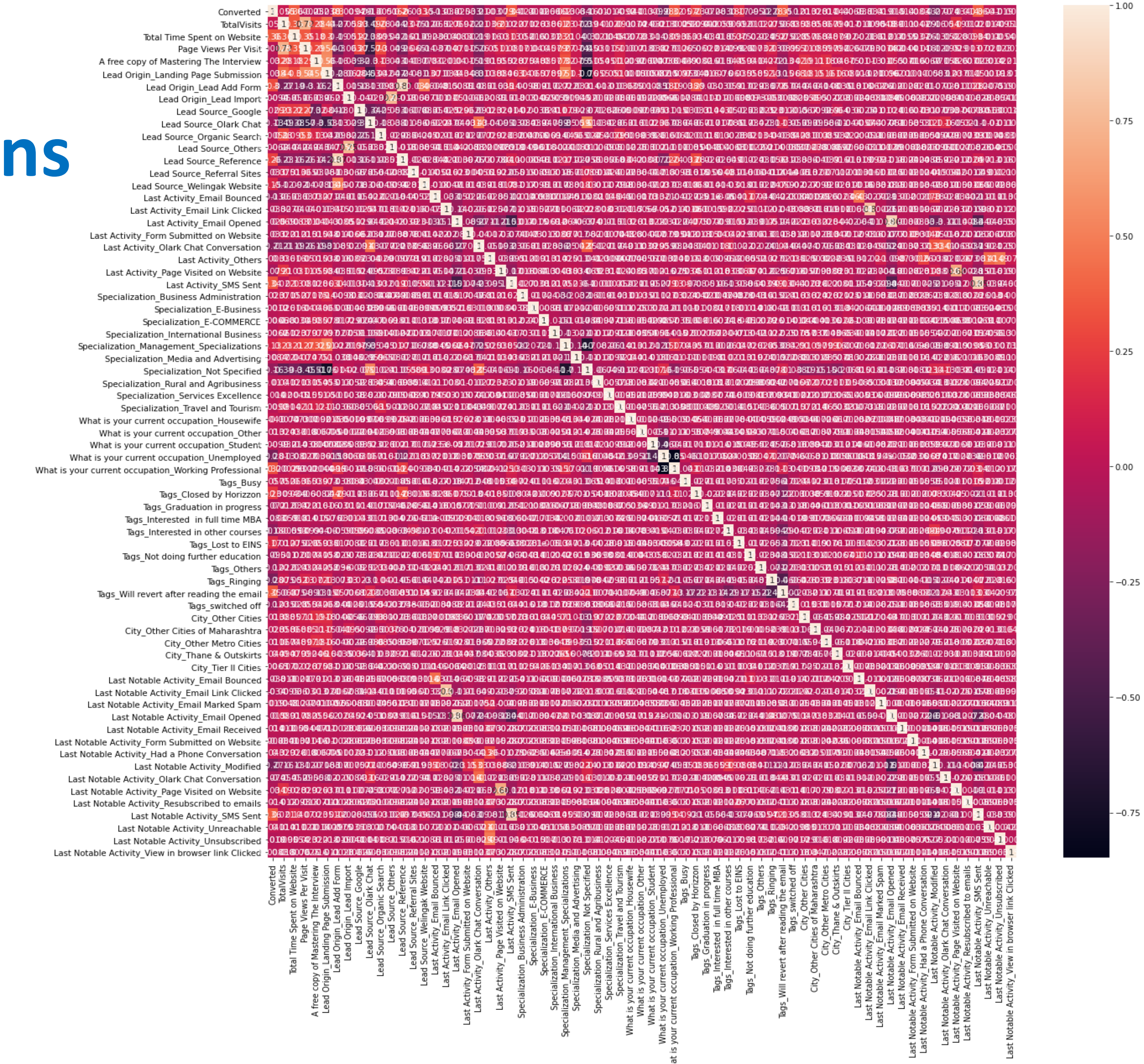
Visualizing the pair plots of numerical distributions



Insights:

- there doesn't seem to be much direct linear relationship between these variables. But there seems to be some relationship between total visits and pageviews per visit

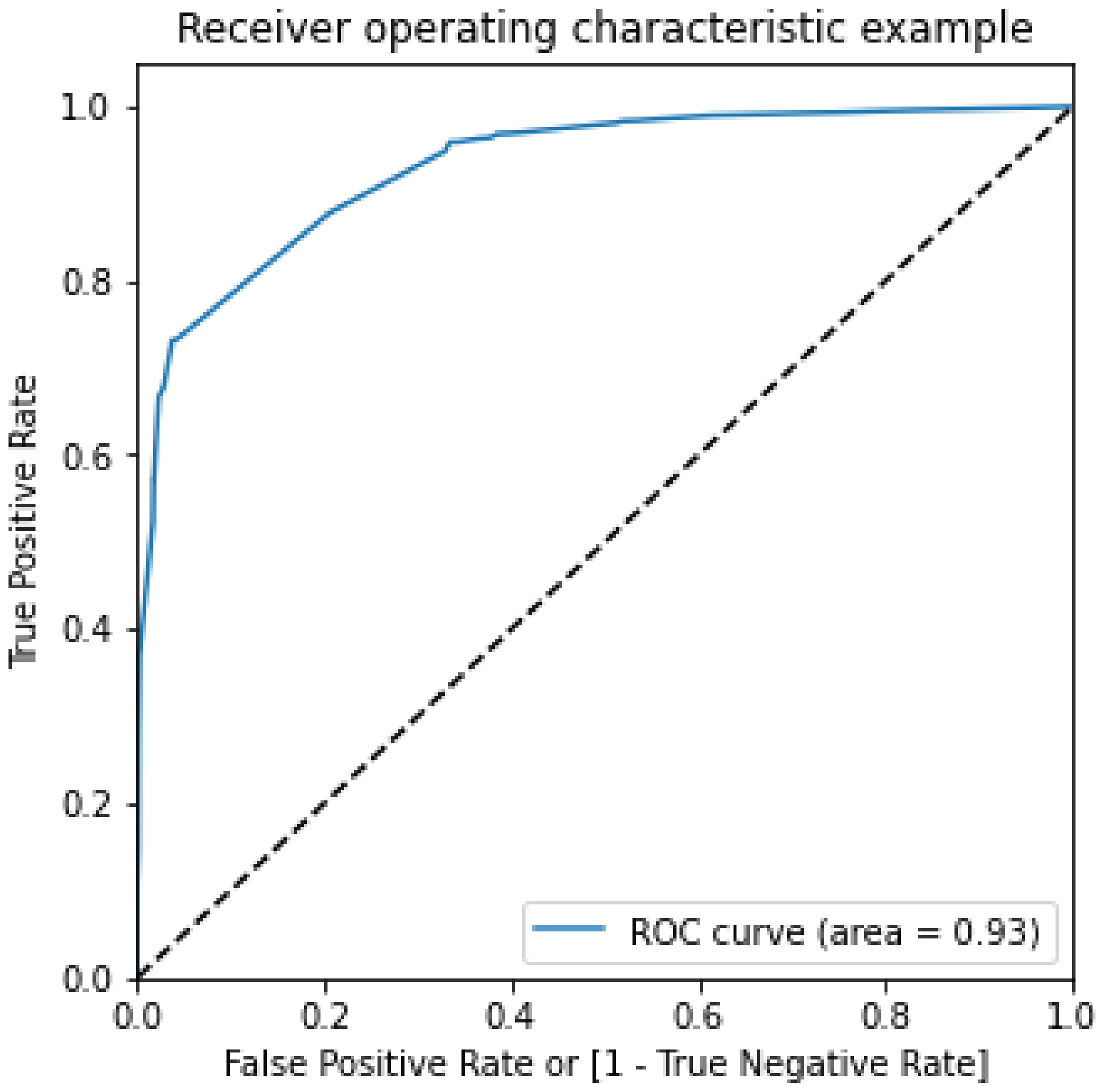
Visualizing the correlations using heatmap



Insights:

- other than a few others only in the top left corner we see high correlations

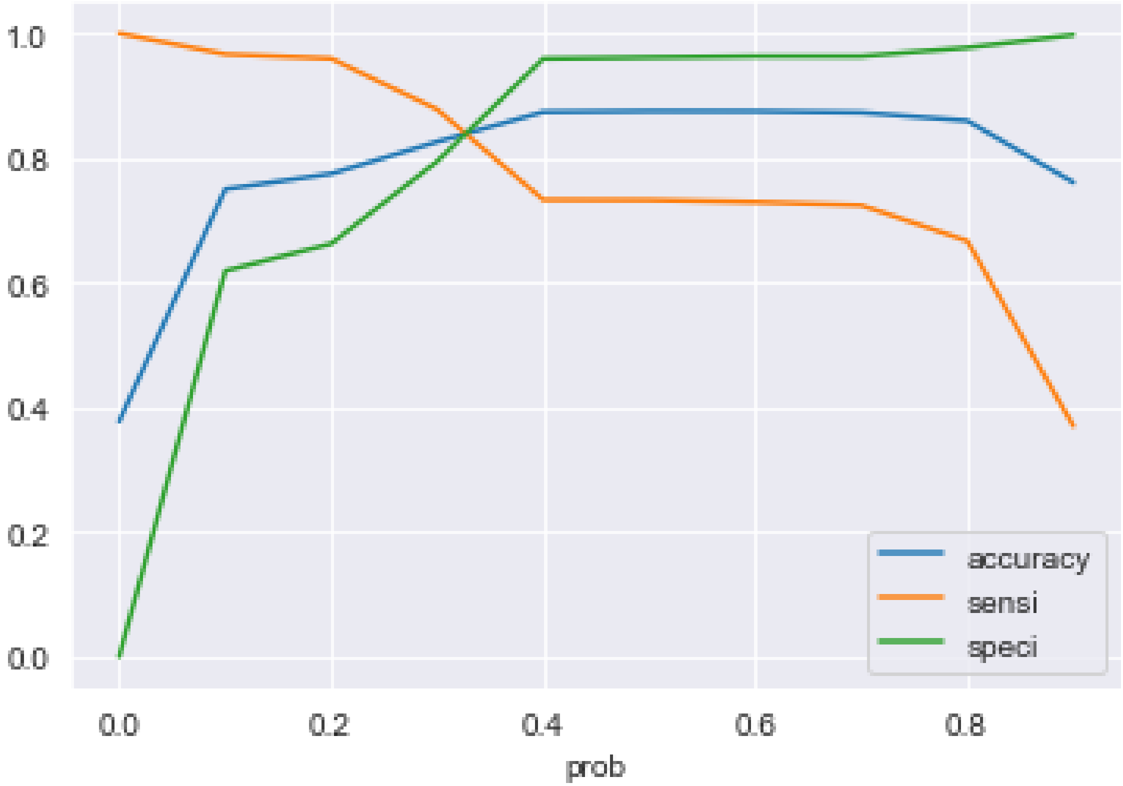
Visualizing the ROC to find the optimum cut off



Insights:

- the AUC curve should have a value close to 1. We are getting a value of .93 indicating a good predictive model

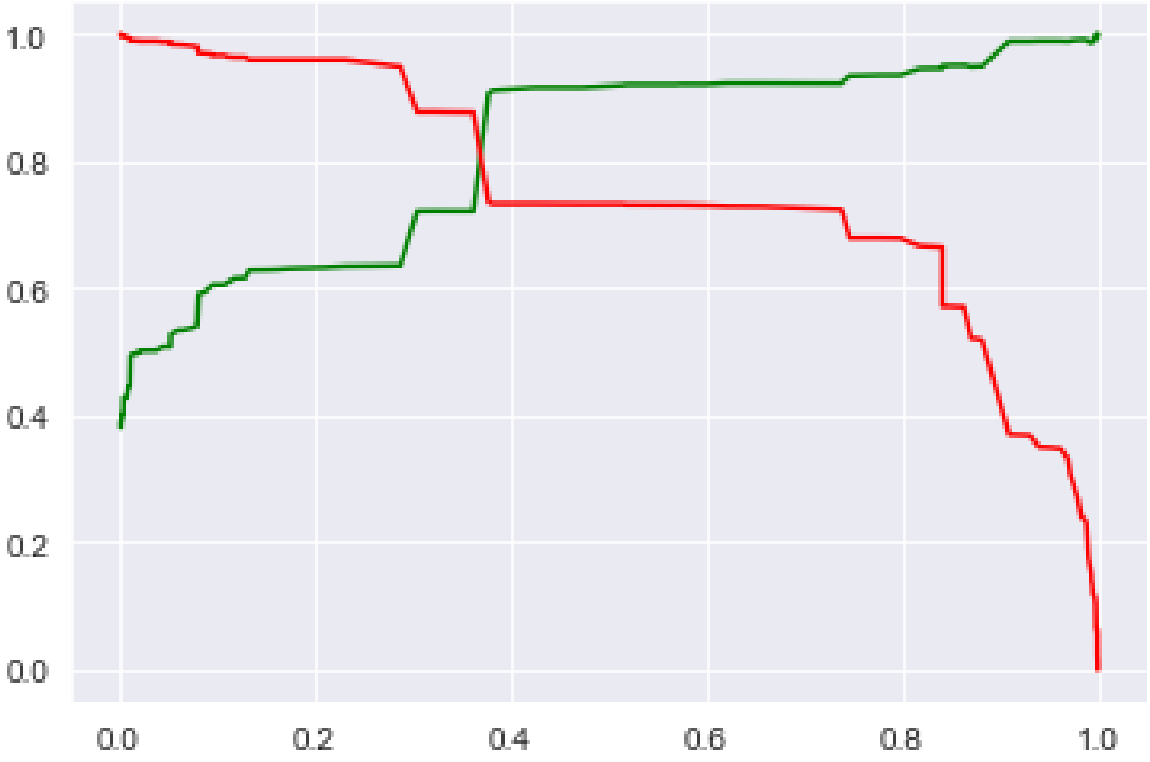
Noting accuracy, sensitivity and specificity for various probabilities



Insights:

- cutoff of 0.3 is the optimum point such that we can maximize sensitivity without reducing specificity too much

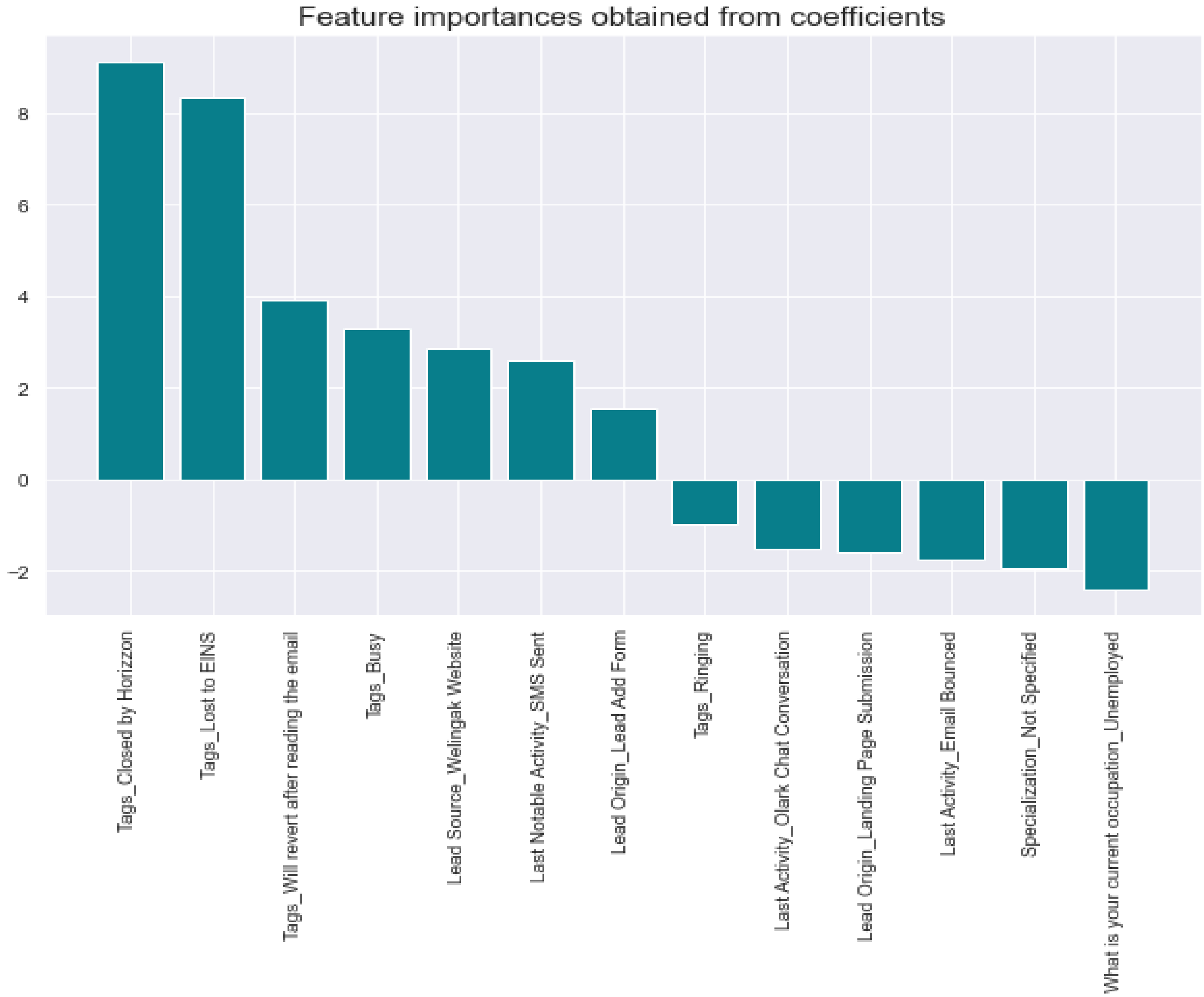
Plotting precision and recall for various probabilities



Insights:

- a cut off closer to 0.4 would have given a balanced precision and recall. However, in our model we would want to increase recall as much as possible without compromising the precision too much. Hence a cut off of 0.3 seems to work fine.

Feature Importance



Insights:

- larger the coefficient, greater the impact on the prediction.
- Positive coefficient means that the variable is impacting the prediction positively and vice versa for negative.
- Zero coefficient means that the variable has no impact on the prediction.

Conclusion

Evaluation metrics for the Train Data:

Accuracy : 82.59%

Sensitivity : 87.78%

Specificity : 79.44%

False positive rate : 20.55%

Positive predictive value : 72.13%

Negative predictive value : 91.47%

Precision : 72.13%

Recall : 87.78%

Evaluation metrics for the Test Data:

Accuracy : 82.77%

Sensitivity : 87.22%

Specificity : 80.04%

False positive rate : 19.95%

Positive predictive value : 72.80%

Negative predictive value : 91.09%

Precision : 72.81%

Recall : 87.22%

We can see that the model performs reasonably well on the test set signifying that the model is generalizing well. It has a sensitivity value of 87.22%. This means out of all converted leads; model has correctly predicted 87.22% of them which is a very good measure.

We have also assigned Lead scores to each customer in the range of 0-100 which can be used by X Education to target potential leads. A higher score means that the lead is hot, i.e., is most likely to convert whereas a lower score means that the lead is cold and will mostly not get converted.

Recommendations

Parameters **positively** impacting the lead conversion:

- Tag assigned to a customer is any of the following
    - Closed by Horizzon
    - Lost to EINS
    - Will revert after reading the email
    - Busy
- Lead Source is Welingak Website
- Last Notable Activity performed by the student is SMS Sent
- Lead Origin is Lead Add Form

This means that any customer with the above parameters is a very promising lead and is most likely to convert into paying customer. Hence, X Education should concentrate on such customers to increase the conversion rate of their leads.

Parameters **negatively** impacting the lead conversion:

- Tags assigned to a customer is Ringing
- Last Activity performed by the customer is any of the following
    - Olark Chat Conversation,
    - Email Bounced
- Lead Origin is Landing Page Submission
- Specialization is Not Specified by the customer
- Customer is Unemployed

This means that any customer with the above parameters is less likely to convert into paying customer

In conclusion, X Education can look at the lead scores assigned to each customer, and along with the above parameters can focus their efforts on such a way as to maximize the lead conversion ratio.

Thank you