LEAD SCORE CASE STUDY

Group Members:

1. Aakash S S

2. Arjun Bhardwaj

PROBLEM STATEMENT

- An education company named X Education sells online courses to industry professionals.

- Now, although 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'.

SOLUTION METHODOLOGY

- **Data Cleaning**: We dropped the variables that had high percentage of NULL values along row and column wise in them. This step also included imputing the missing values as and where required with median values in case of numerical variables and creation of new classification variables in case of categorical variables. The outliers were identified and removed.
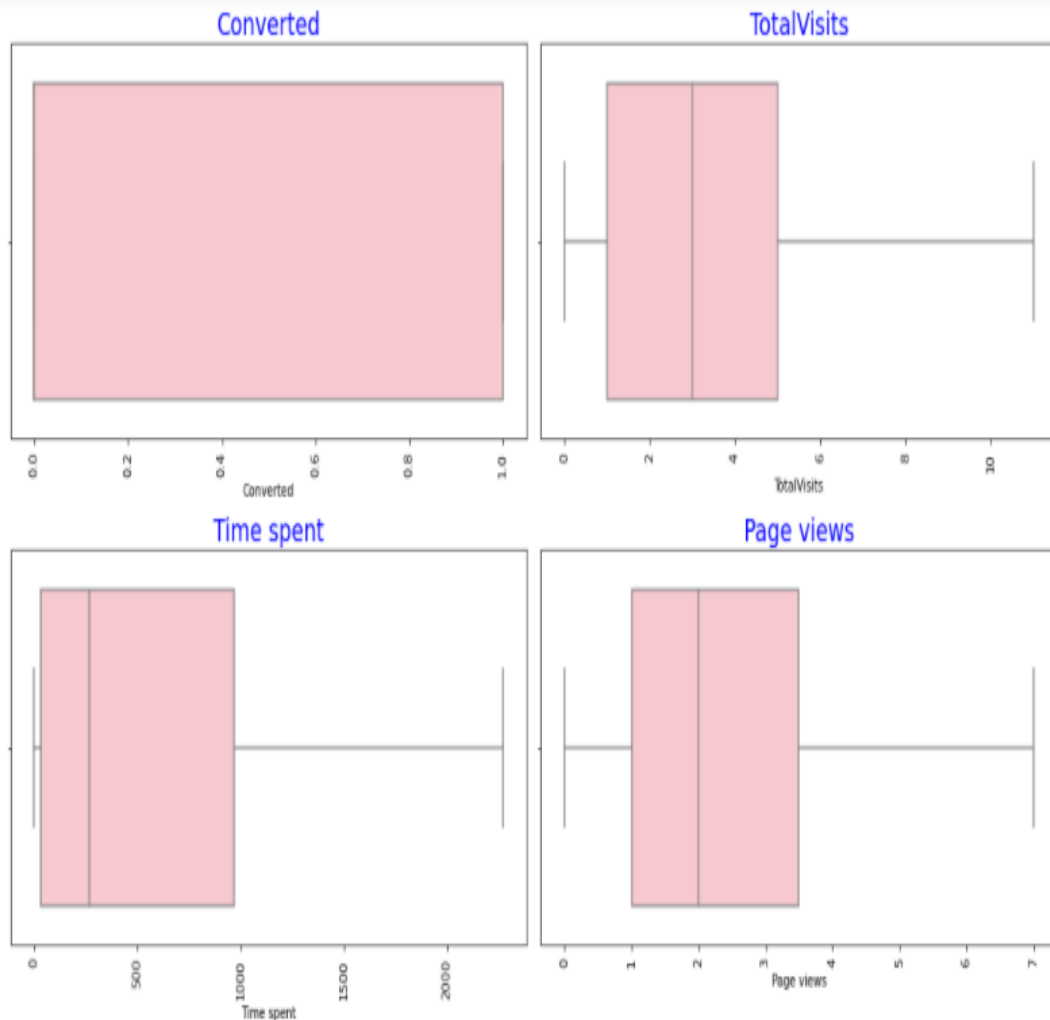
EDA "UNIVARIATE ANALYSIS"
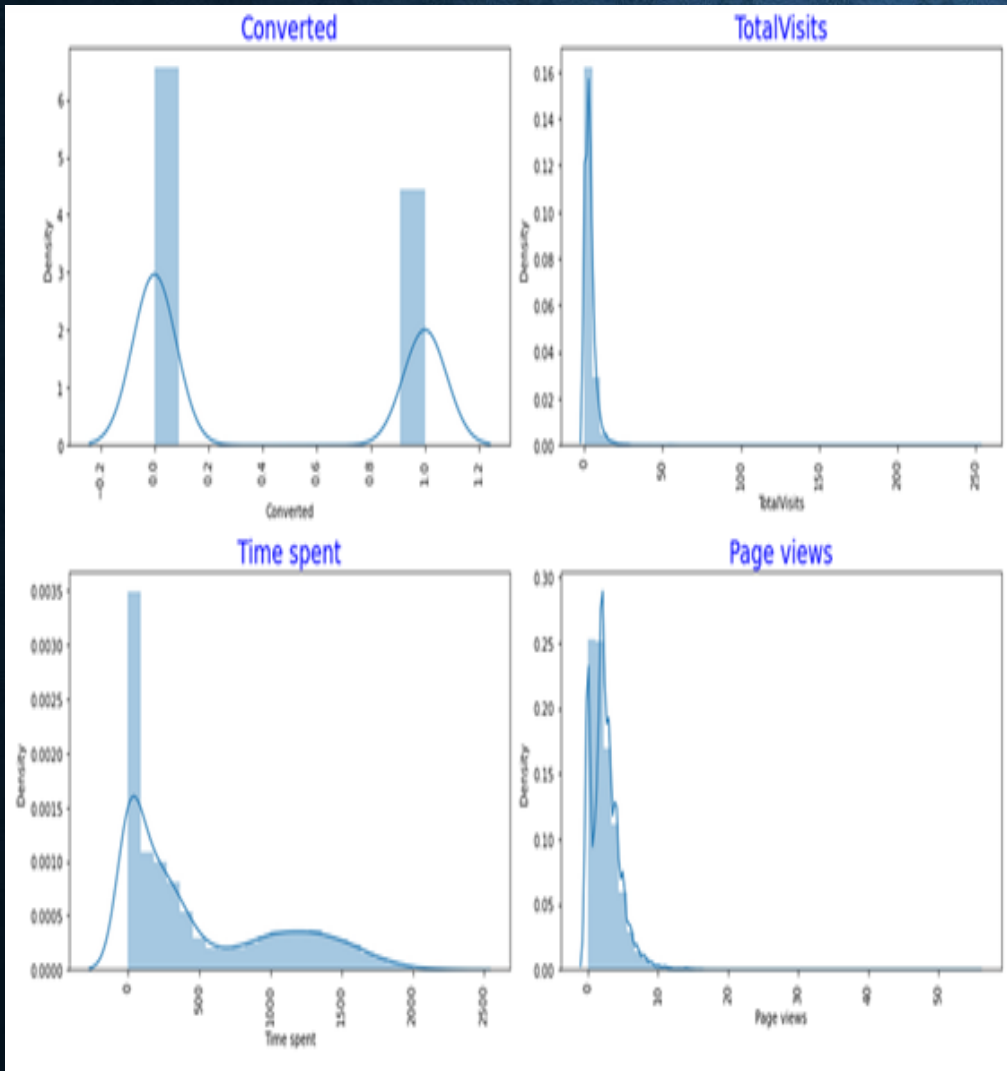
- ➤ Mostly all our variables are in normal distribution.
- ➤ But, here we have not treated our outliers and just vislausized it a step before.
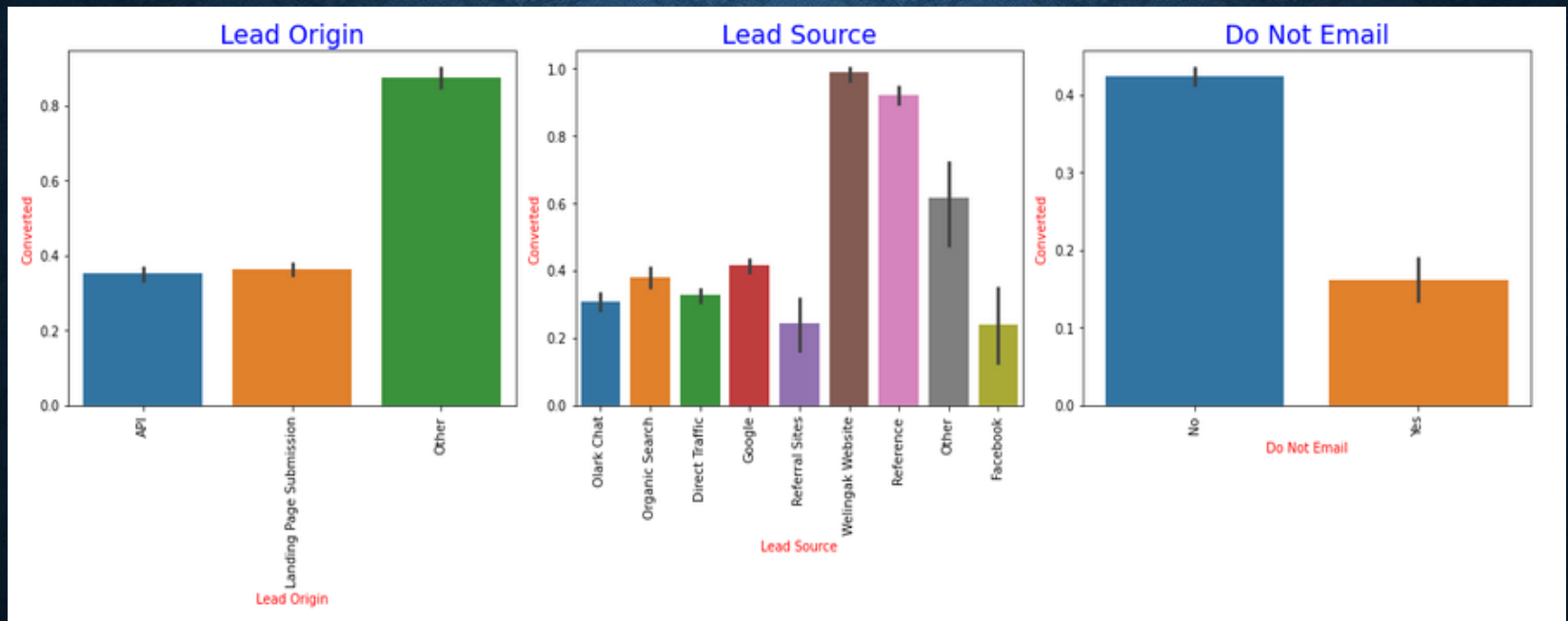
BOXPLOT VISUALISATION

- ➢ Here, we have clearly treated the Outliers in all variables and visualised using box plot.
- ➢ The most important part used here to treaten the outlier is soft capping.
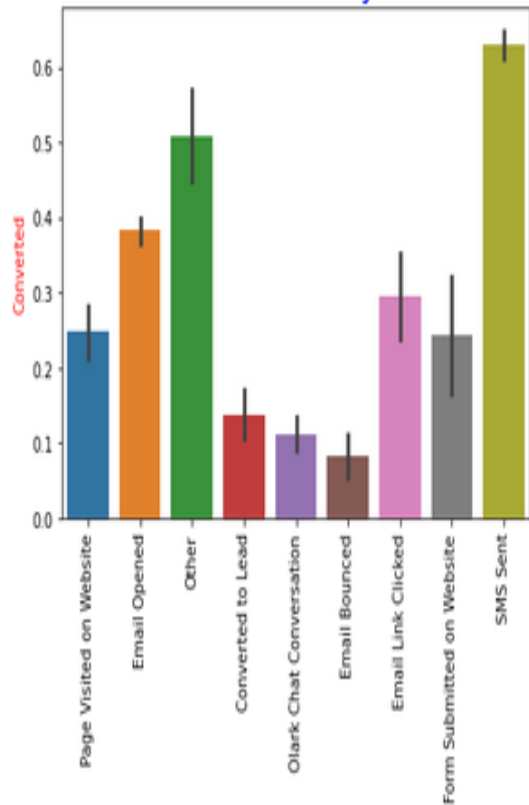- ➢ Now, it looks pretty good and ready for further analysis

EDA BIVARIATE ANALYSIS

➢ From the plot, its clear that most of the convertion happens through the detailed profile read from the email.
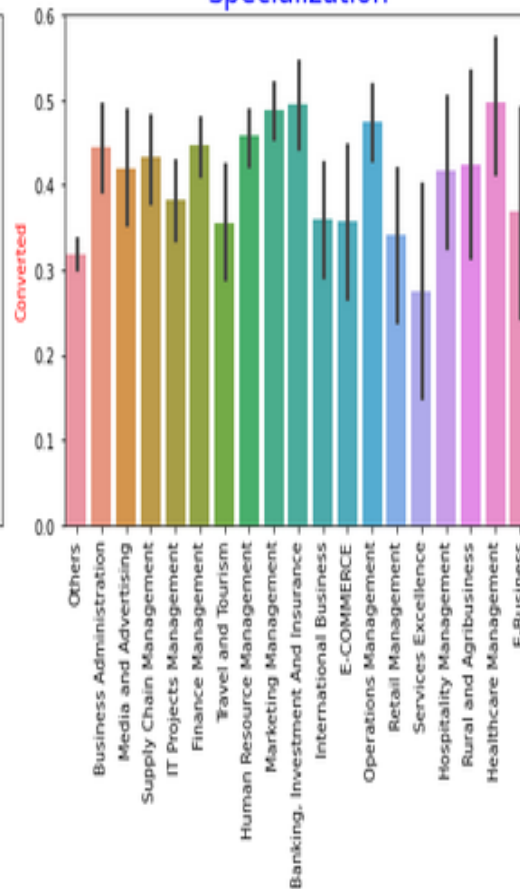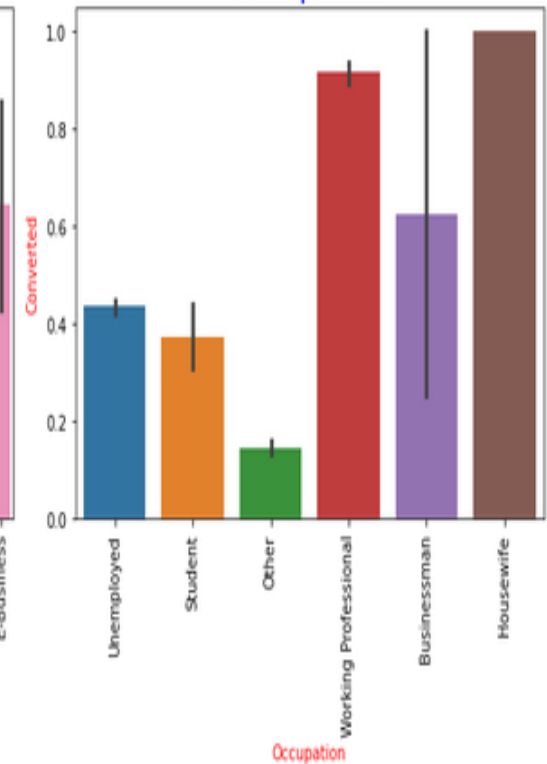
EDA BIVARIATE ANALYSIS

# SUMMARY FROM ANALYSIS

- Large amount of customers are converted into the course through their references.

- Majority of the clients are converted by seeing the profile or details provided by the company through Email only.

- Banking and Healthcare management domain experienced customers are mostly get converted here.

- The surprise here is most of the housewife's are willing to join the courses.

AN ROC CURVE DEMONSTRATES SEVERAL THINGS:
IT SHOWS THE TRADEOFF BETWEEN SENSITIVITY AND SPECIFICITY (ANY INCREASE IN SENSITIVITY WILL BE ACCOMPANIED BY A DECREASE IN SPECIFICITY).
THE CLOSER THE CURVE FOLLOWS THE LEFT-HAND BORDER AND THEN THE TOP BORDER OF THE ROC SPACE, THE MORE ACCURATE WILL BE THE MODEL.
THE CLOSER THE CURVE COMES TO THE 45-DEGREE DIAGONAL OF THE ROC SPACE, THE LESS ACCURATE WILL BE THE MODEL.
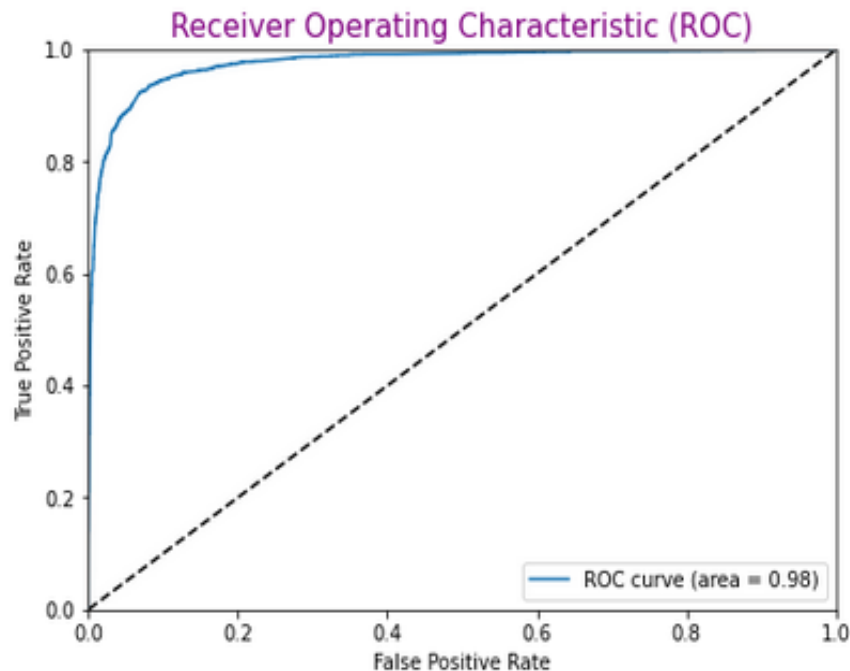
Now, as per our ROC curve, the curve is extremely top left handed towards the y axis. It implies the model is more accurate..

SUMMARY

- The variables were found to be good. We then created the data frame having the converted probability values and we had an initial assumption that a probability value of more than 0.5 means 1 else 0.

- Based on the above assumption, we derived the Confusion Metrics and calculated the overall Accuracy of the model.

- We also calculated the '**Sensitivity**' and the '**Specificity**' matrices to understand how reliable the model is.

- learnings show that to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be 92 %, Sensitivity=93%, Specificity= 92% (Approx).

CONCLUSION

- Top three variables that contribute most towards the probability of a lead getting converted :

  - **Lost to EINS (Tags)**

  - **Closed by Horizzon (Tags)**

  - **Welingak Website (lead_source)**

- 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 92%, 93% and 92% 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 more than 80%

- Hence overall this model seems to be excellent.