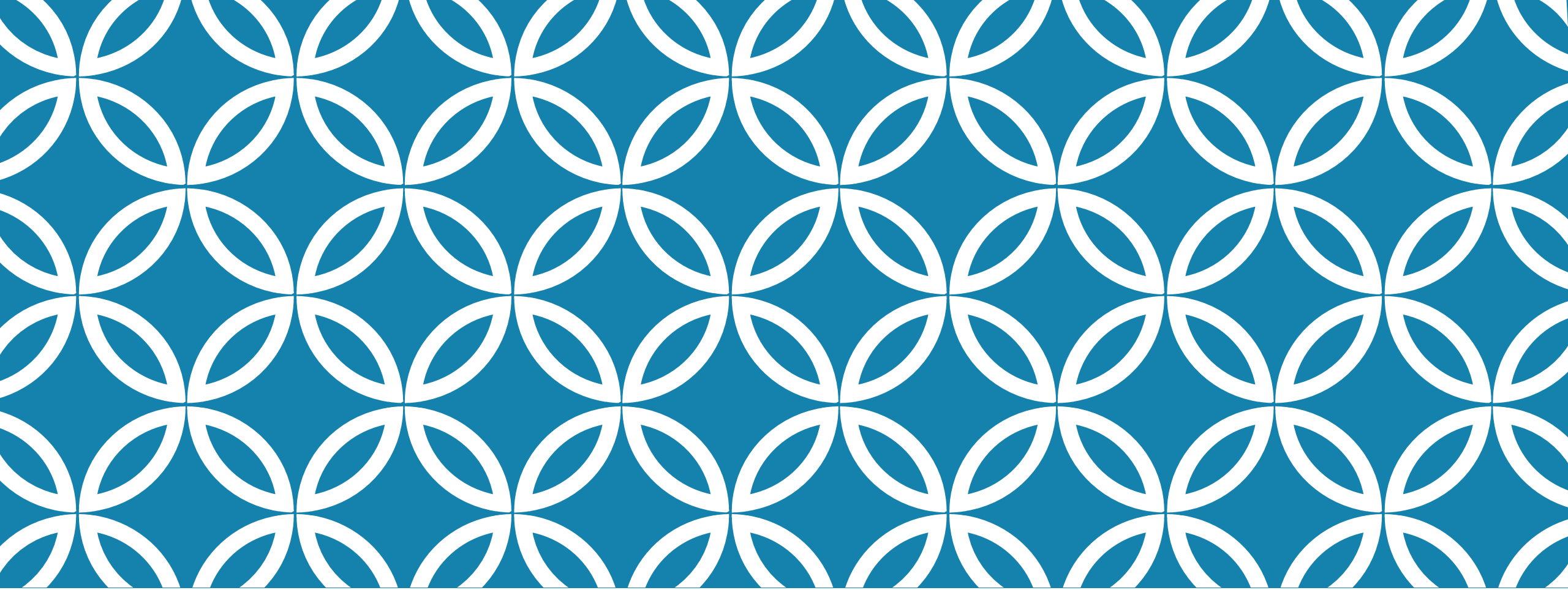


LEAD SCORING CASE STUDY

Group Members

- Anany Sharma
- Aswin Karthik P

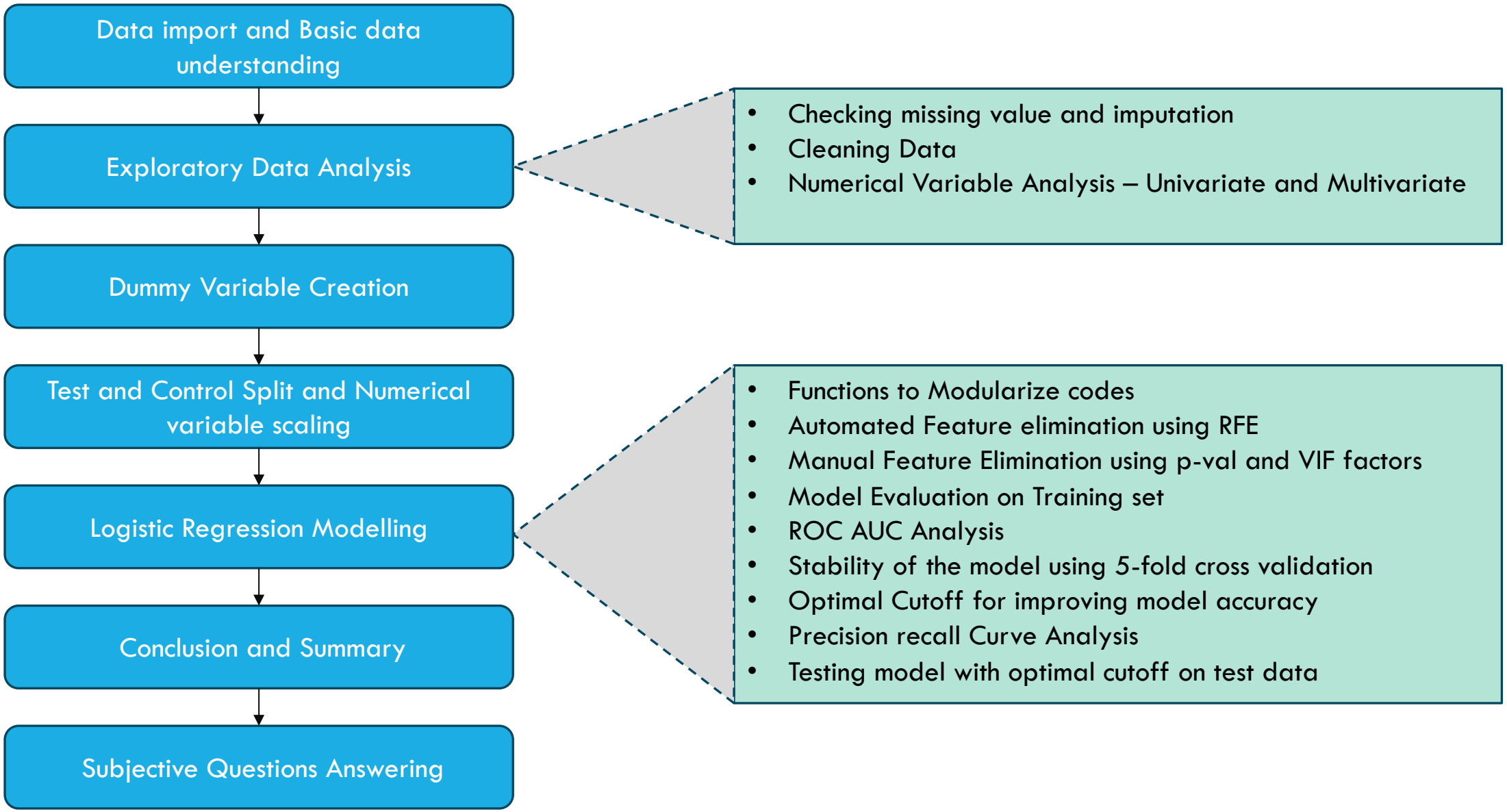
PROBLEM STATEMENT

- X Education sells online courses to industry professionals
- X Education gets a lot of leads (people who have shown interest to join the courses through online marketing), but the conversion is 30%
- To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'.
- Hot Leads can lead to increased conversion rates, since the sales team will focus on potential leads who have higher chances of joining the course

BUSINESS OBJECTIVE

- A ballpark of the target lead conversion rate of 80% has been set by the CEO. 50% more then current conversion rate.
- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads
- Understand various underlying factors affecting the conversion and how to control them to drive positive results
- Provision to handle future constraints and suggestion for the same

APPROACH


Data import and Basic data understanding
Exploratory Data Analysis
Dummy Variable Creation
Test and Control Split and Numerical variable scaling
Logistic Regression Modelling
Conclusion and Summary
Subjective Questions Answering
• Checking missing value and imputation
• Cleaning Data
• Numerical Variable Analysis – Univariate and Multivariate
• Functions to Modularize codes
• Automated Feature elimination using RFE
• Manual Feature Elimination using p-val and VIF factors
• Model Evaluation on Training set
• ROC AUC Analysis
• Stability of the model using 5-fold cross validation
• Optimal Cutoff for improving model accuracy
• Precision recall Curve Analysis
• Testing model with optimal cutoff on test data

DATA IMPORT AND BASIS DATA UNDERSTANDING

- The data has 9240 rows and 37 columns
- Out of 37 columns
  - 3 columns are of Integer type
  - 4 columns are of Float type
  - 30 columns are of Object type (String, which may contain categorical and ordinal variables)
- Prospect ID and Lead Number is an ID column, we need to verify if they are unique.
- Converted is Target column and has the value of 0 and 1.

EXPLORATORY DATA ANALYTICS – MISSING VALUES AND CLEANING DATA (1/2)

Column	Missing Value %
How did you hear about X Education	78.46
Lead Profile	74.19
Lead Quality	51.59
Asymmetrique Profile Score	45.65
Asymmetrique Activity Score	45.65
Asymmetrique Activity Index	45.65
Asymmetrique Profile Index	45.65
City	39.71
Specialization	36.58
Tags	36.29
What matters most to you in choosing a course	29.32
What is your current occupation	29.11
Country	26.63
Page Views Per Visit	1.48
TotalVisits	1.48
Last Activity	1.11
Lead Source	0.39

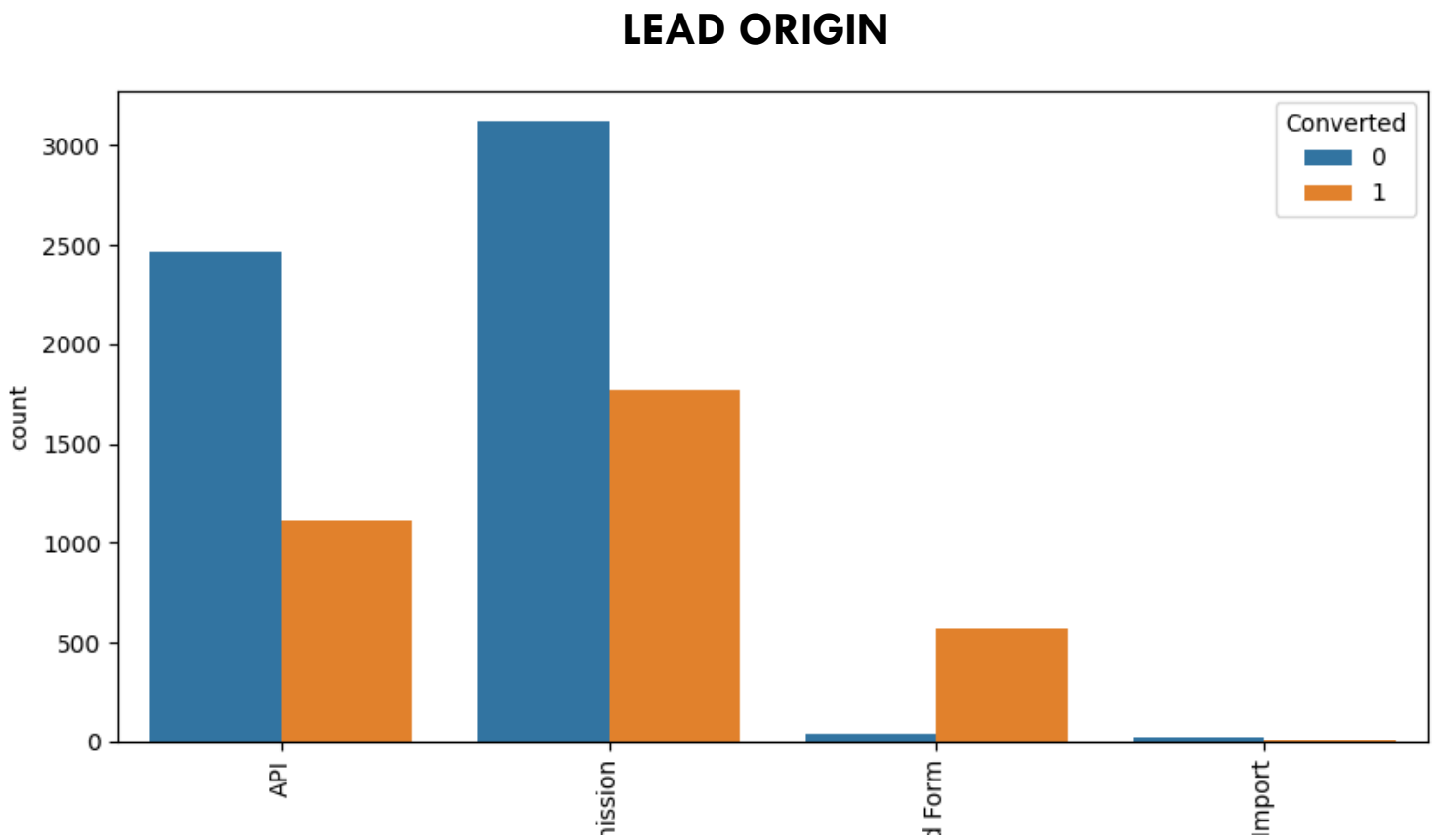
Column	Missing Value %
Receive More Updates About Our Courses	0
I agree to pay the amount through cheque	0
Get updates on DM Content	0
Update me on Supply Chain Content	0
A free copy of Mastering The Interview	0
Prospect ID	0
Newspaper Article	0
Through Recommendations	0
Digital Advertisement	0
Newspaper	0
X Education Forums	0
Lead Number	0
Magazine	0
Search	0
Total Time Spent on Website	0
Converted	0
Do Not Call	0
Do Not Email	0
Lead Origin	0
Last Notable Activity	0

EXPLORATORY DATA ANALYTICS – MISSING VALUES AND CLEANING DATA (2/2)

- We have 18 columns with missing value.
- Based on the missing value proportion and Data Dictionary we are planning to drop columns with more than 45% of missing values
- Imputation strategy is to fill NA values with mode (most common value) or creates a new category like “Not Given”
- Post Imputation, if the variable doesn’t offer any imputation then we drop it

Note: All observations and strategies are explained for each variable in the notebook

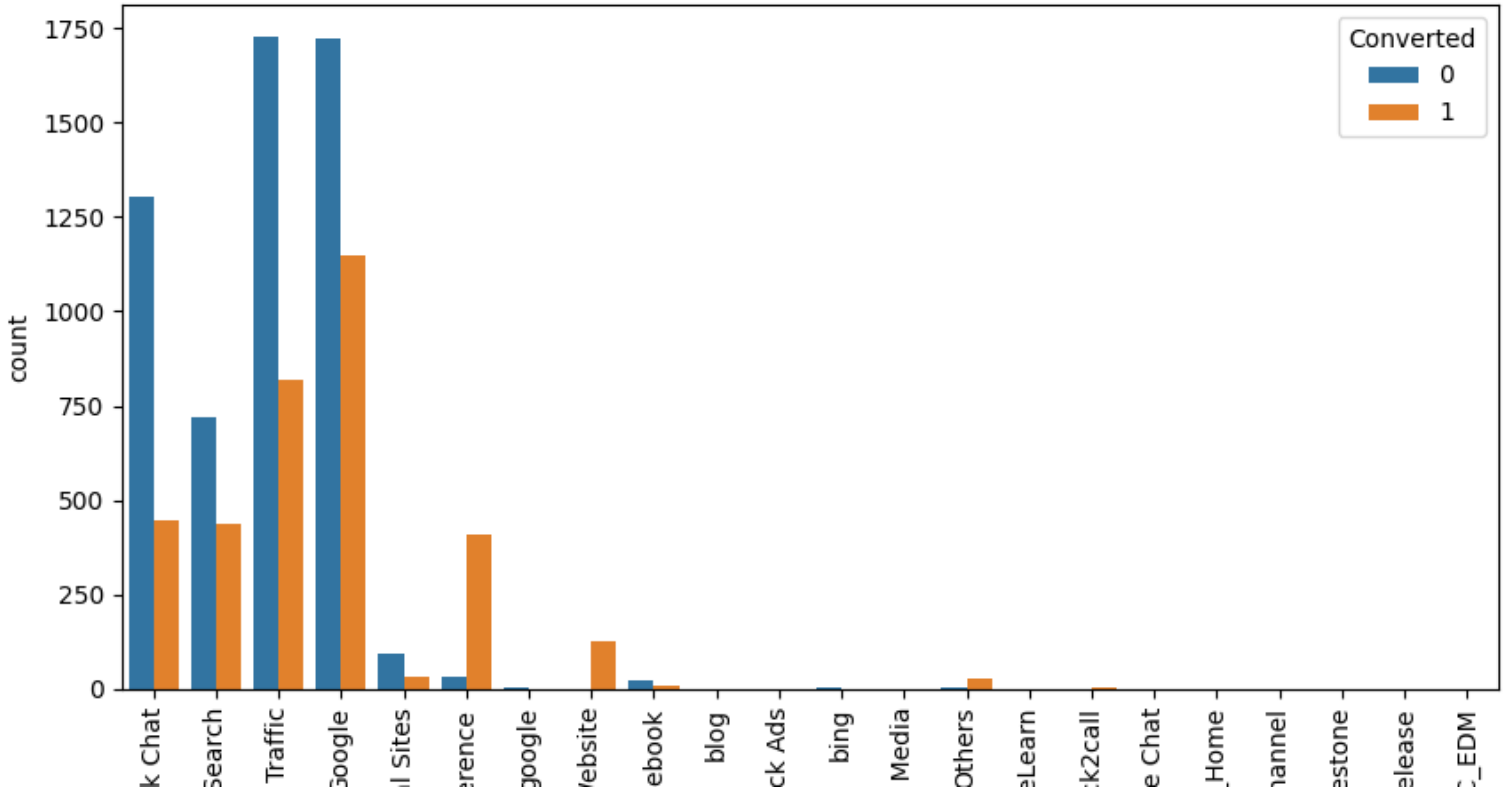
EXPLORATORY DATA ANALYTICS – CLEANING DATA (CATEGORICAL BIVARIATE ANALYSIS) (1/12)



- No change is required except for dummy creations.
- API and Landing Page submission generate the larger proportion of leads, but the conversion rate is low.

EXPLORATORY DATA ANALYTICS – CLEANING DATA (CATEGORICAL BIVARIATE ANALYSIS) (2/12)

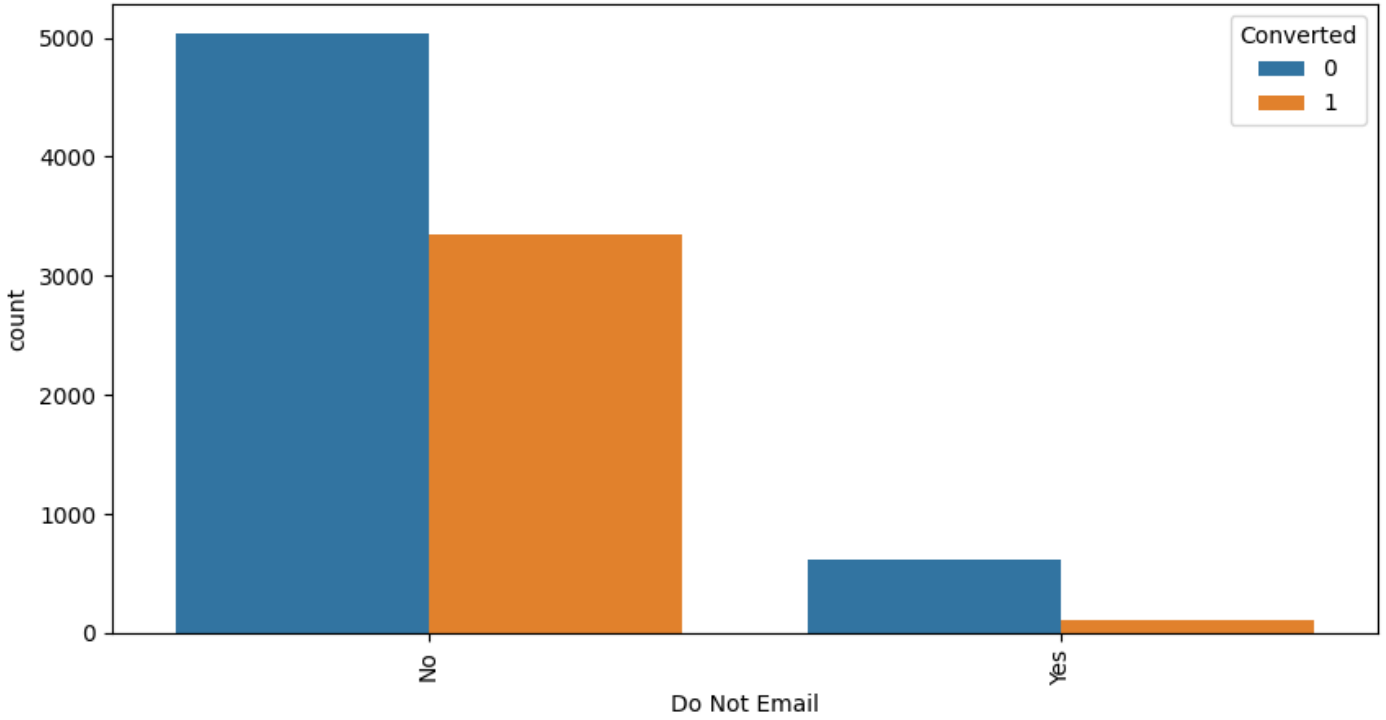
LEAD SOURCE



- There are some duplications like Google and google, Facebook and social media. Which can be clubbed together
- Also a lot of sources have very low frequency which can be clubbed under “Others”. (< 0.1)

EXPLORATORY DATA ANALYTICS – CLEANING DATA (CATEGORICAL BIVARIATE ANALYSIS) (3/12)

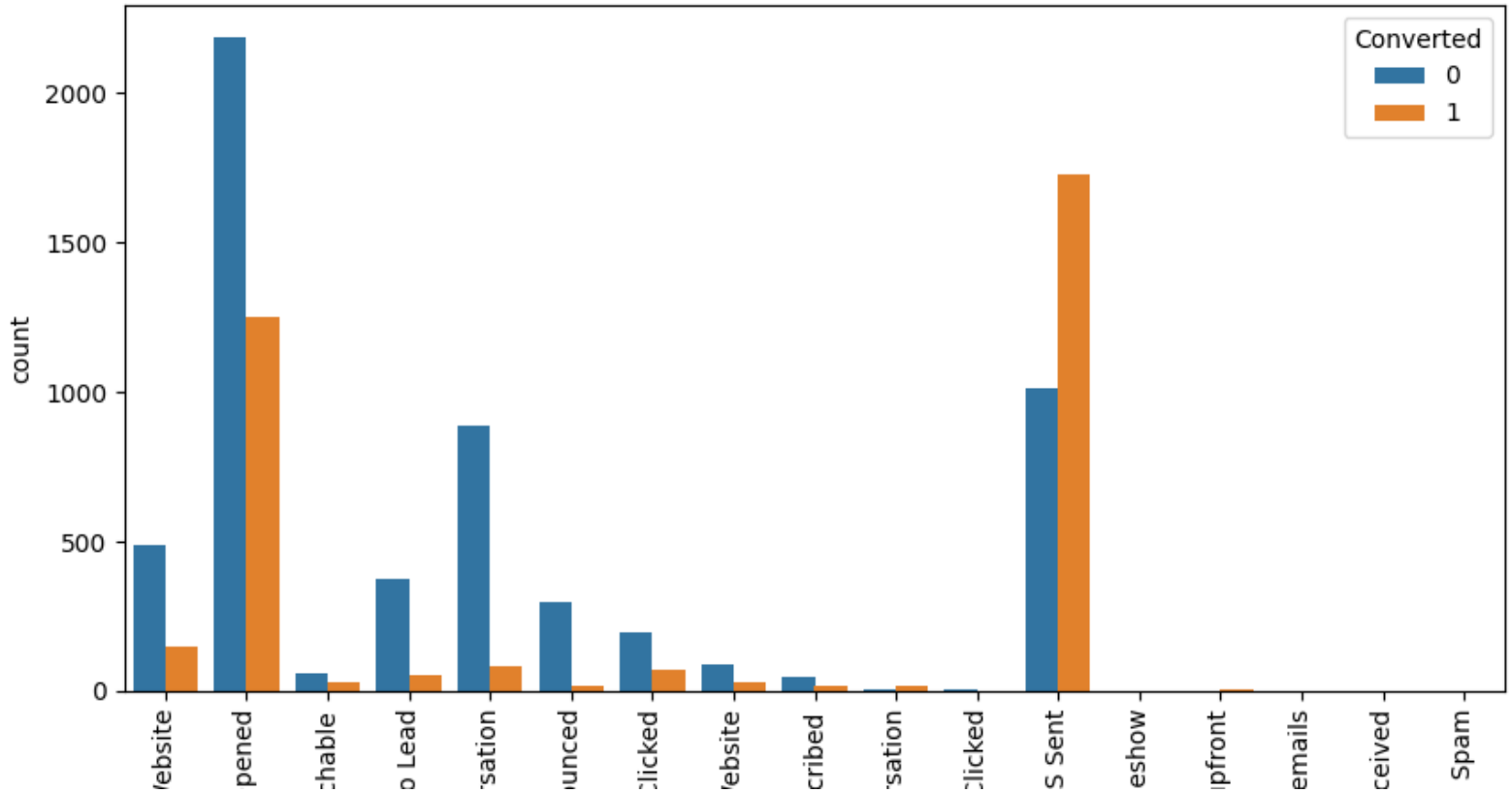
DO NOT EMAIL



- Doesn't need any cleaning

EXPLORATORY DATA ANALYTICS – CLEANING DATA (CATEGORICAL BIVARIATE ANALYSIS) (4/12)

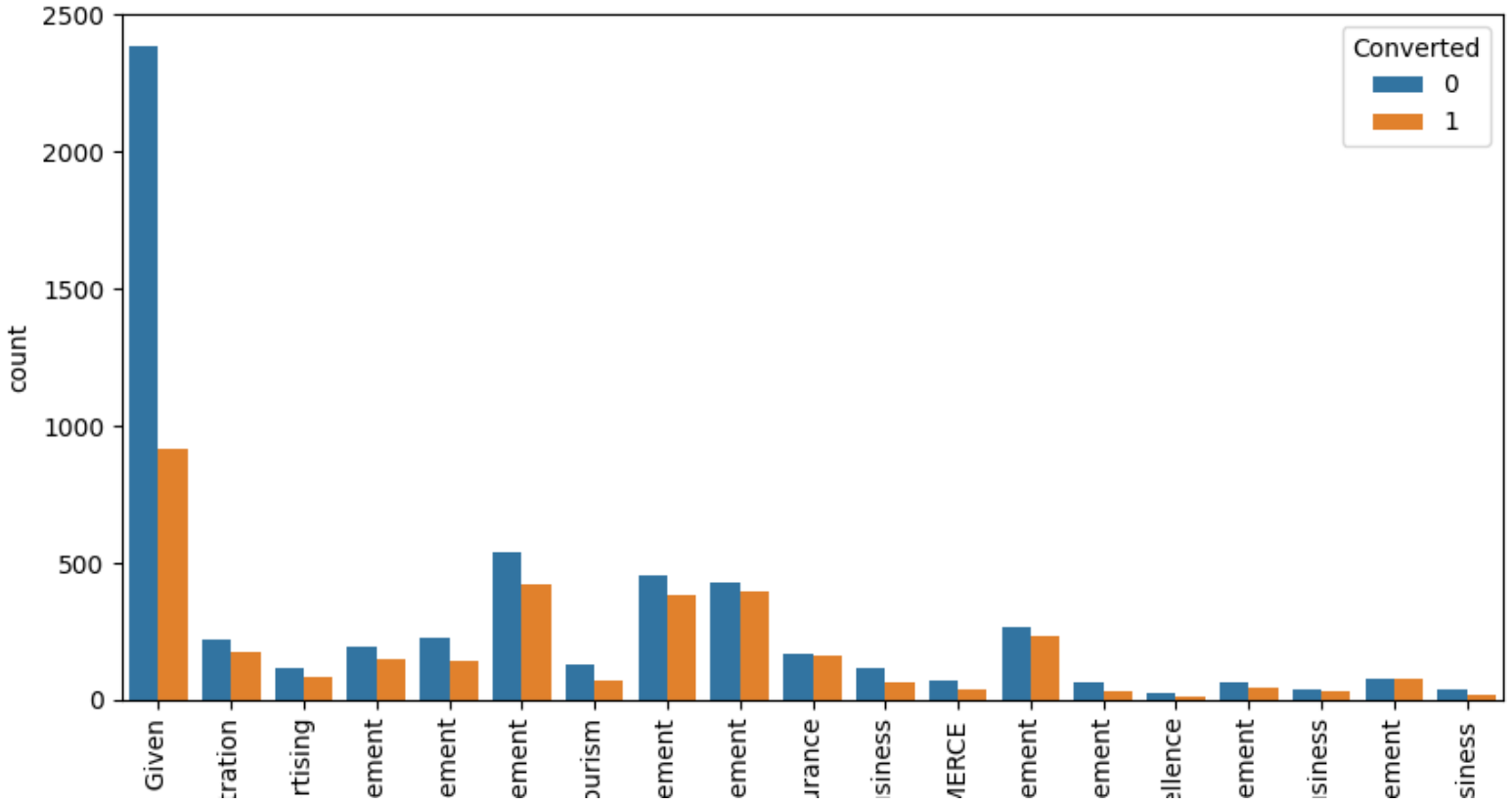
LAST ACTIVITY



- We can combine low occurring values into “Others”

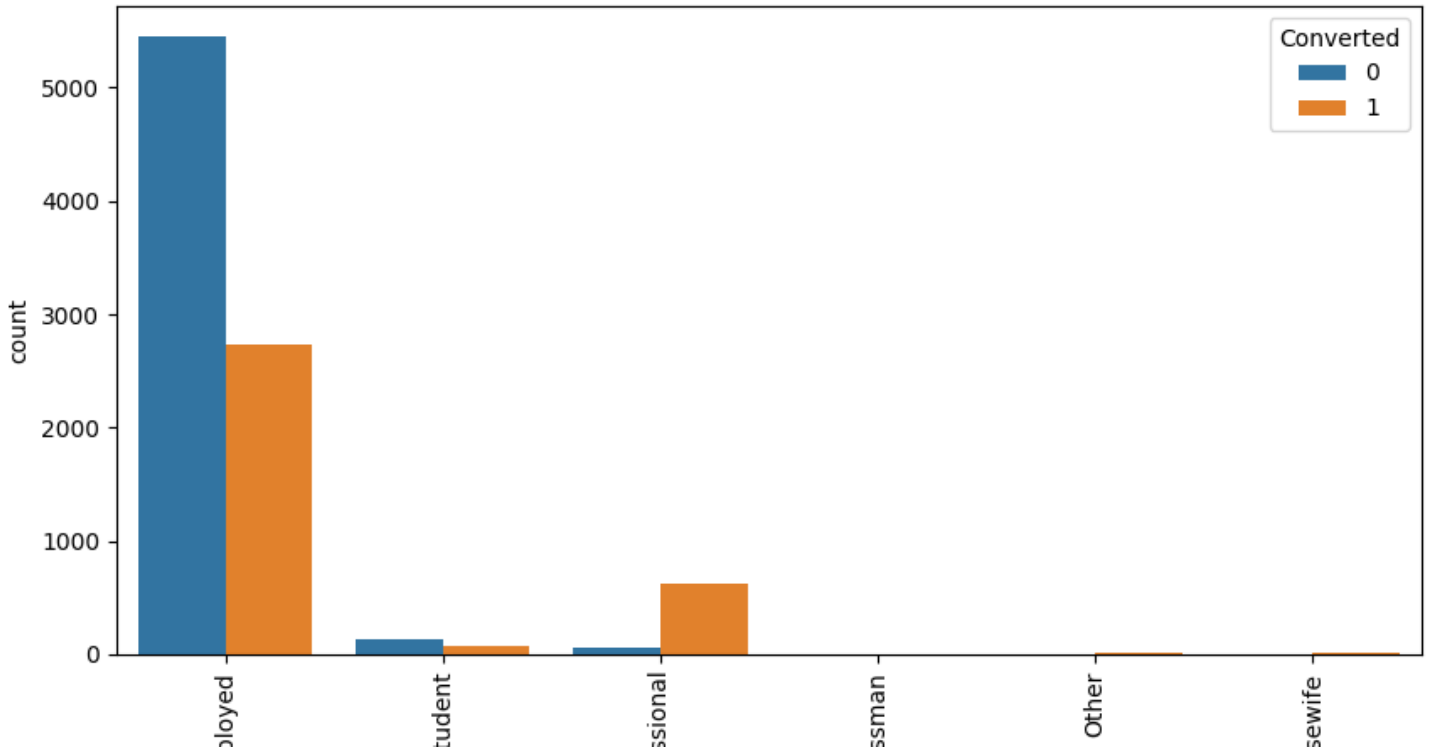
EXPLORATORY DATA ANALYTICS – CLEANING DATA (CATEGORICAL BIVARIATE ANALYSIS) (5/12)

SPECIALIZATION



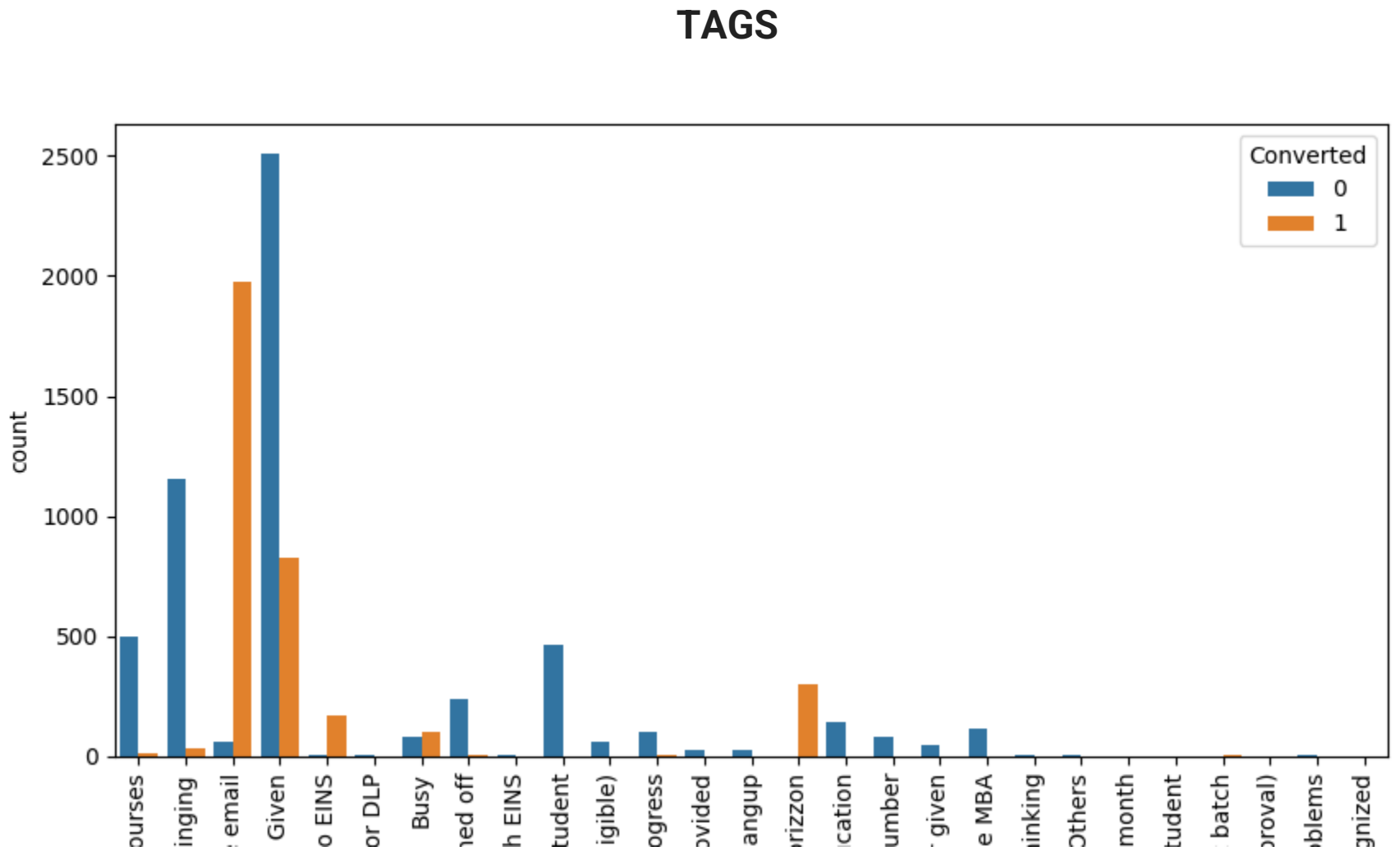
- No Change required

EXPLORATORY DATA ANALYTICS – CLEANING DATA (CATEGORICAL BIVARIATE ANALYSIS) (6/12)



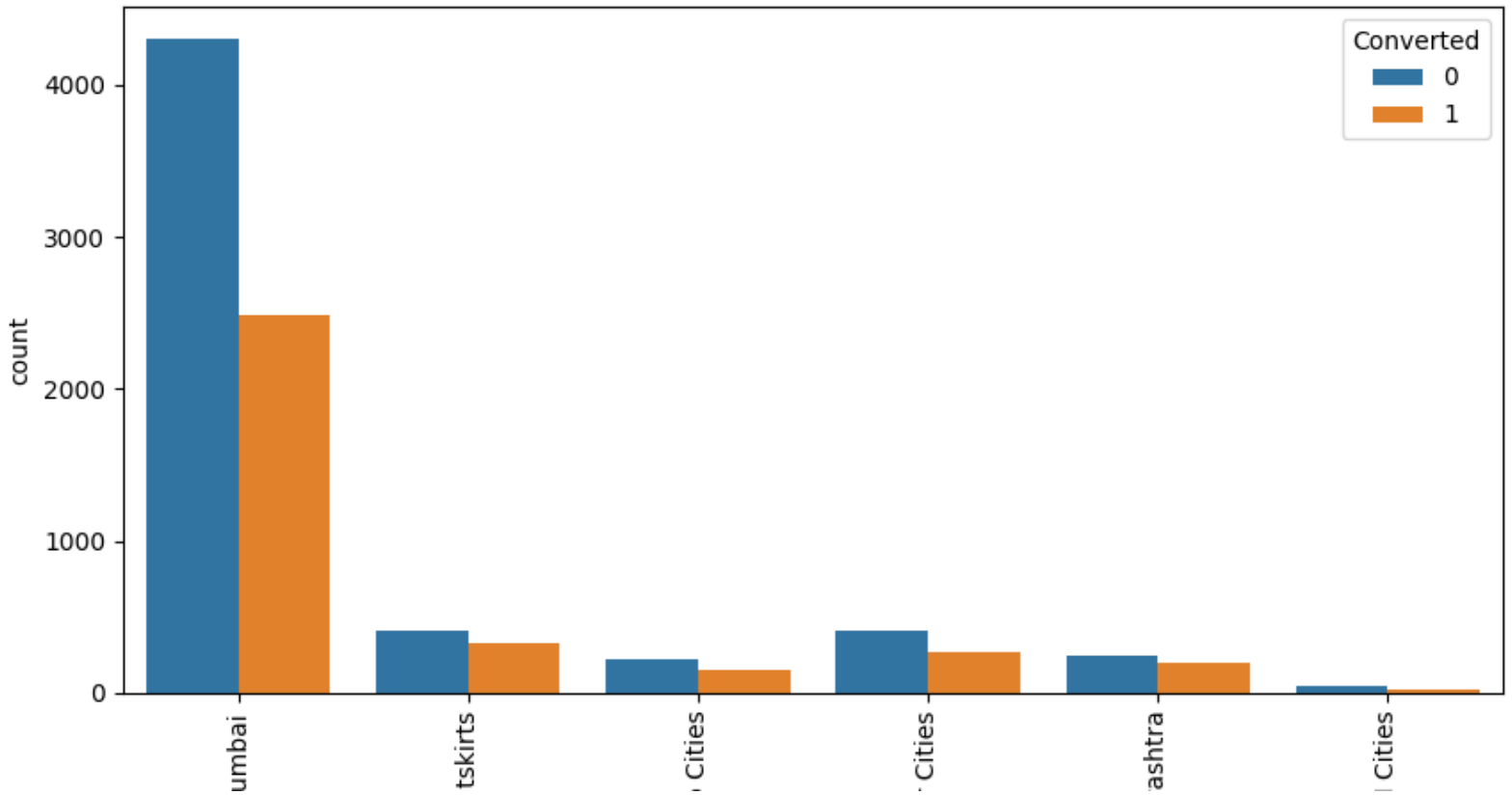
- No Change required

EXPLORATORY DATA ANALYTICS – CLEANING DATA (CATEGORICAL BIVARIATE ANALYSIS) (7/12)



- Replace low frequency values with “Others”

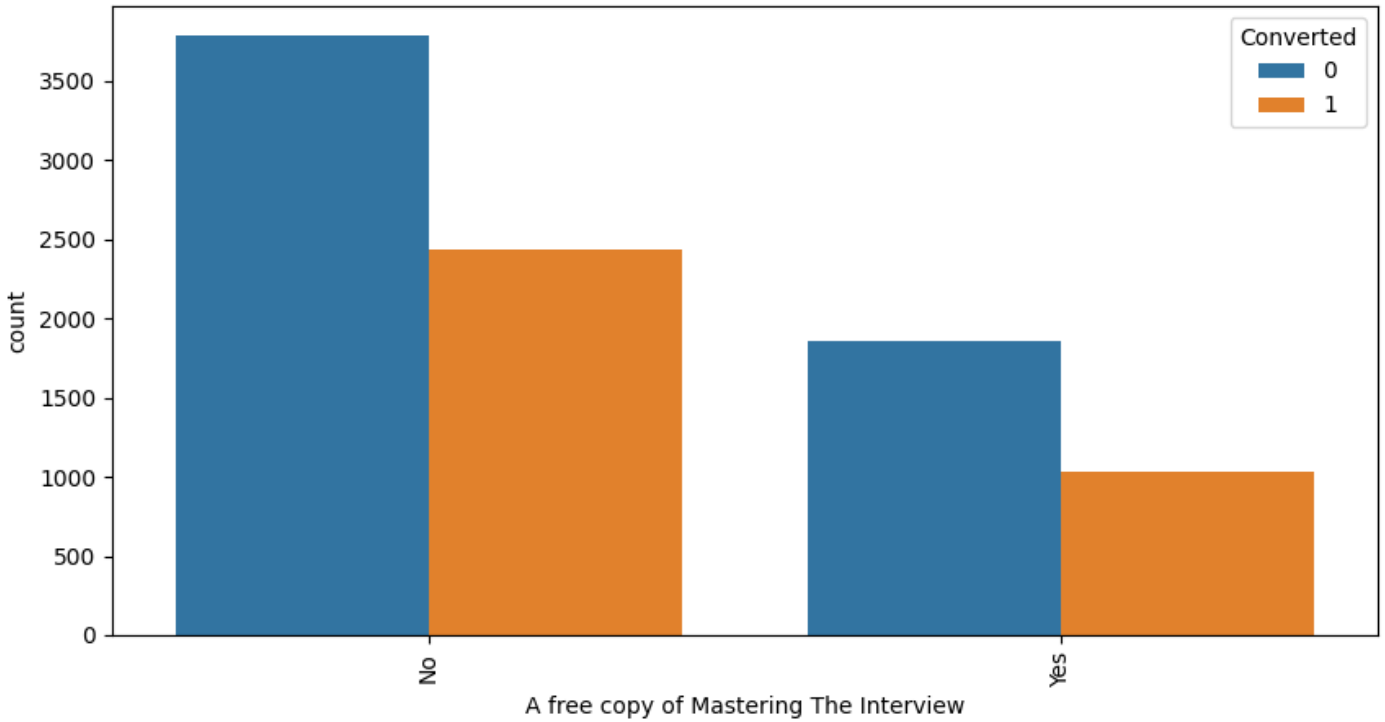
EXPLORATORY DATA ANALYTICS – CLEANING DATA (CATEGORICAL BIVARIATE ANALYSIS) (8/12)



- No changes required

EXPLORATORY DATA ANALYTICS – CLEANING DATA (CATEGORICAL BIVARIATE ANALYSIS) (9/12)

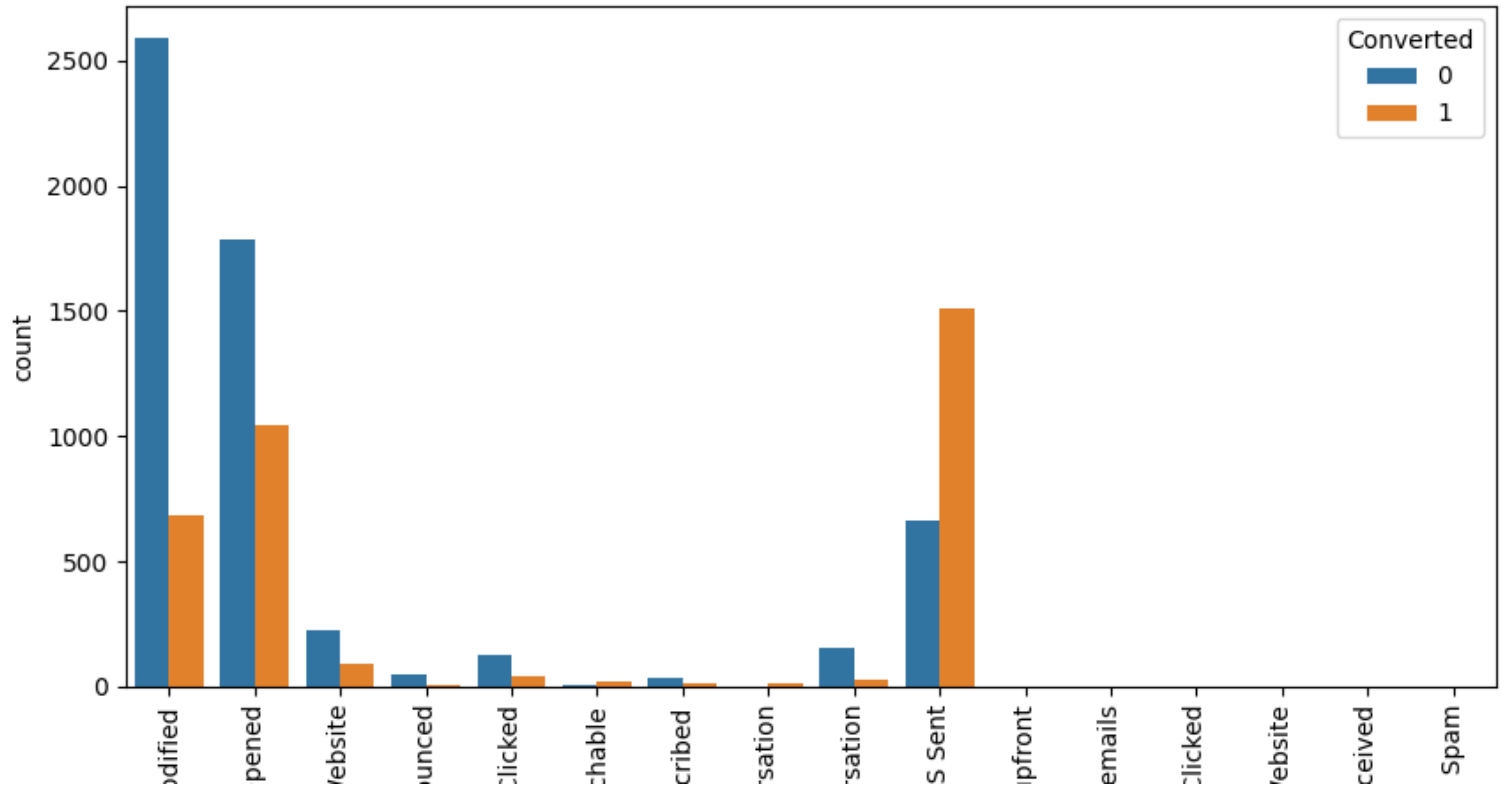
A FREE COPY OF MASTERING THE INTERVIEW



- No changes required

EXPLORATORY DATA ANALYTICS – CLEANING DATA (CATEGORICAL BIVARIATE ANALYSIS) (10/12)

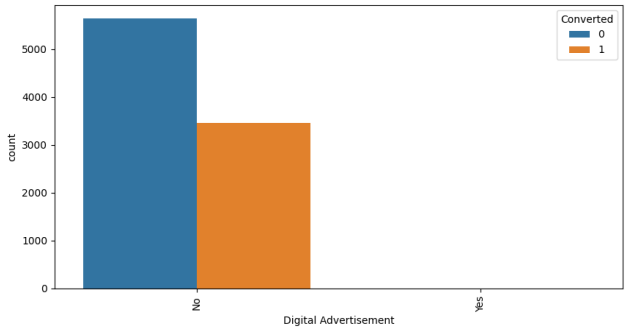
LAST NOTABLE ACTIVITY



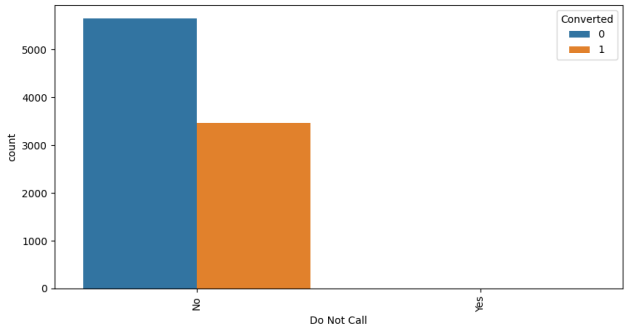
- Replace less frequent values with Others

EXPLORATORY DATA ANALYTICS – CLEANING DATA (CATEGORICAL BIVARIATE ANALYSIS) (11/12)

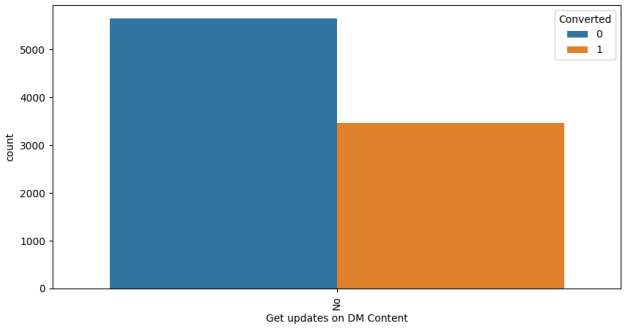
DIGITAL ADVERTISEMENT



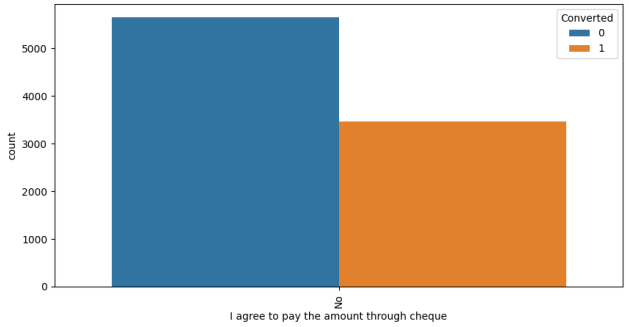
DO NOT CALL



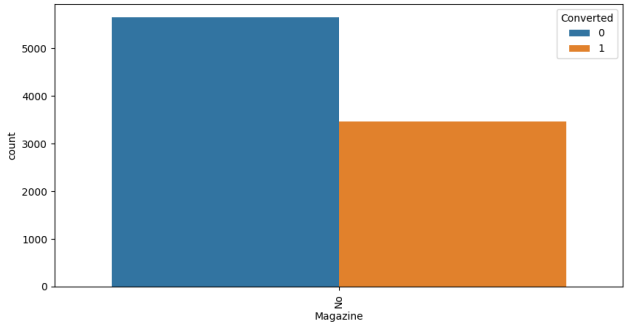
GET UPDATES ON DM CONTENT



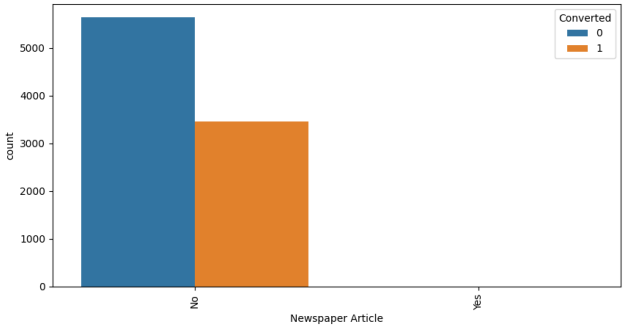
I AGREE TO PAY THE AMOUNT THROUGH CHEQUE



MAGZINE



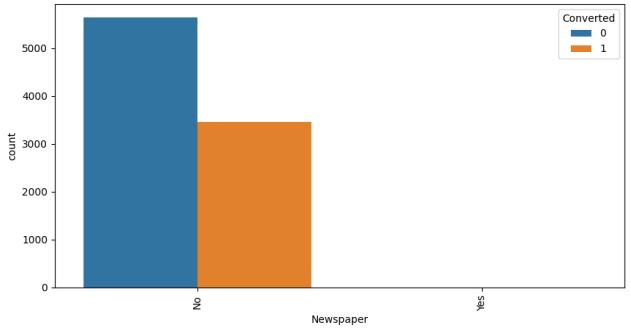
NEWSPAPAER ARTICLE



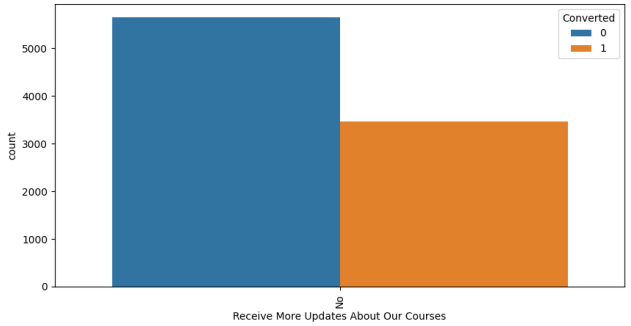
- More than 99% of the values are No, hence there is not variance from which the model can learn.
- We can drop the variable

EXPLORATORY DATA ANALYTICS – CLEANING DATA (CATEGORICAL BIVARIATE ANALYSIS) (12/12)

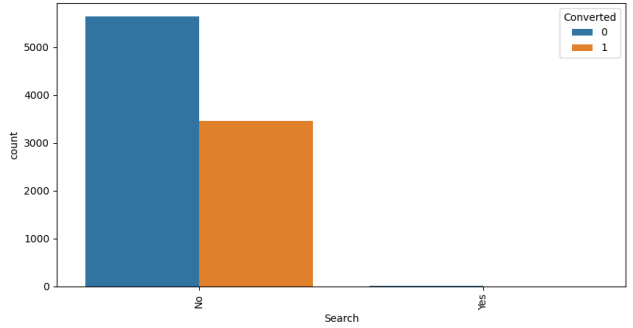
NEWSPAPER



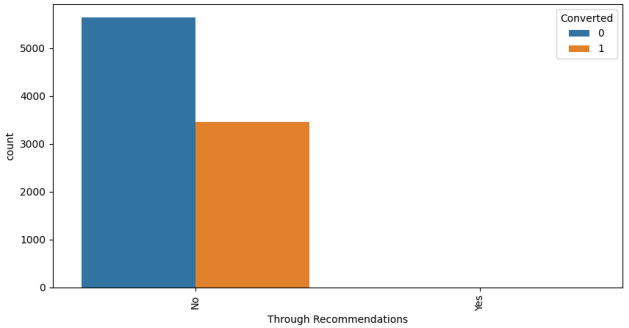
RECEIVE MORE UPDATES ABOUT OUR COURSES



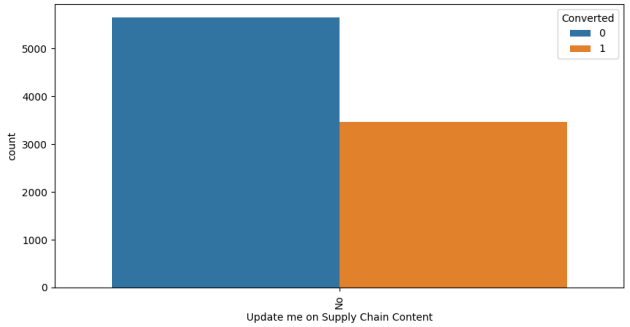
SEARCH



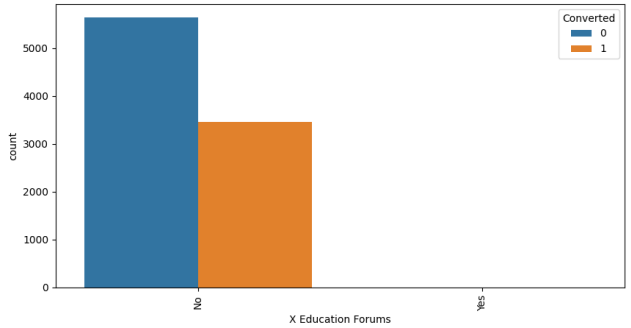
THROUGH RECOMMENDATIONS



UPDATE ME ON SUPPLY CHAIN CONTENT

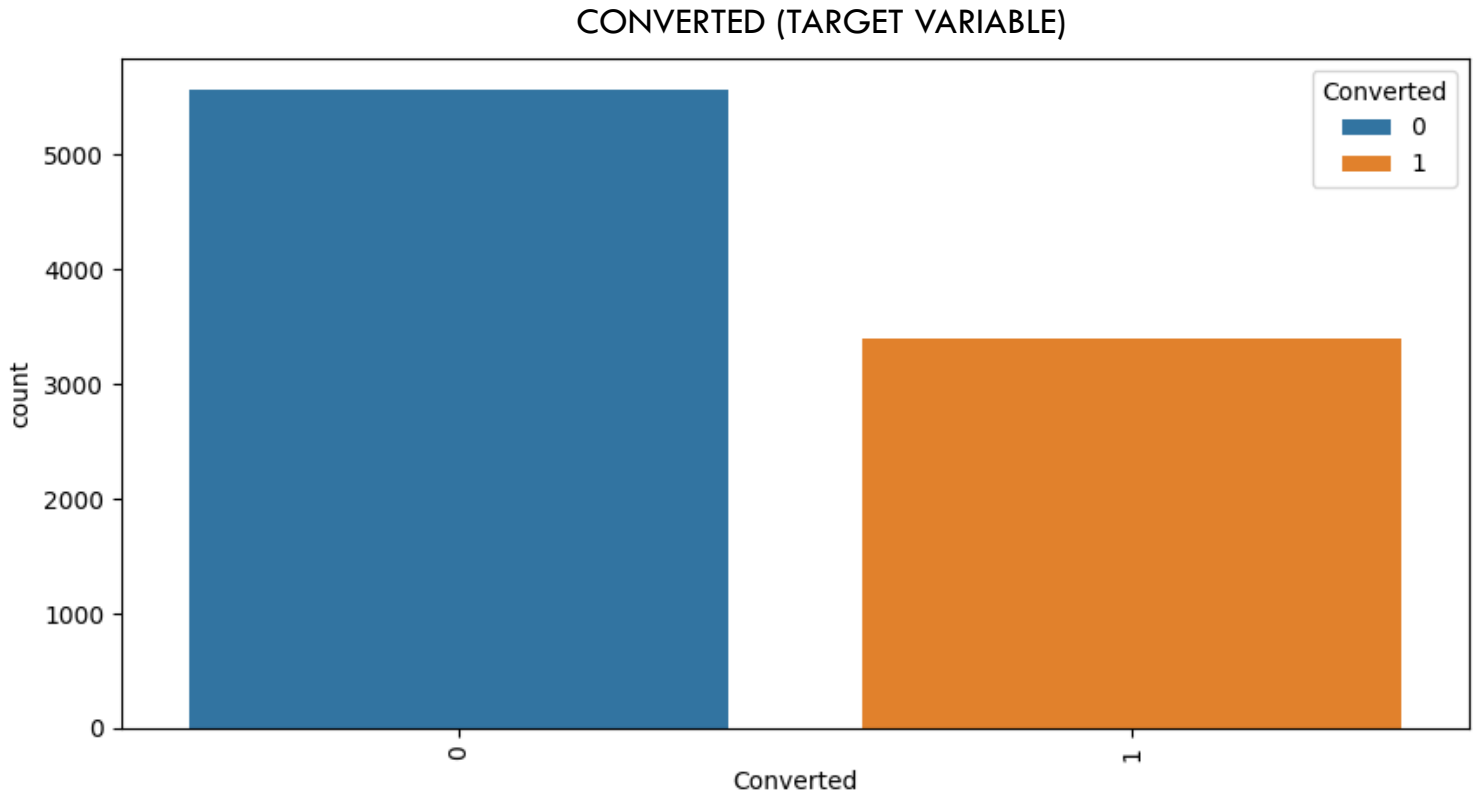


X EDUCATION FORUMS



- More than 99% of the values are No, hence there is not variance from which the model can learn.
- We can drop the variable

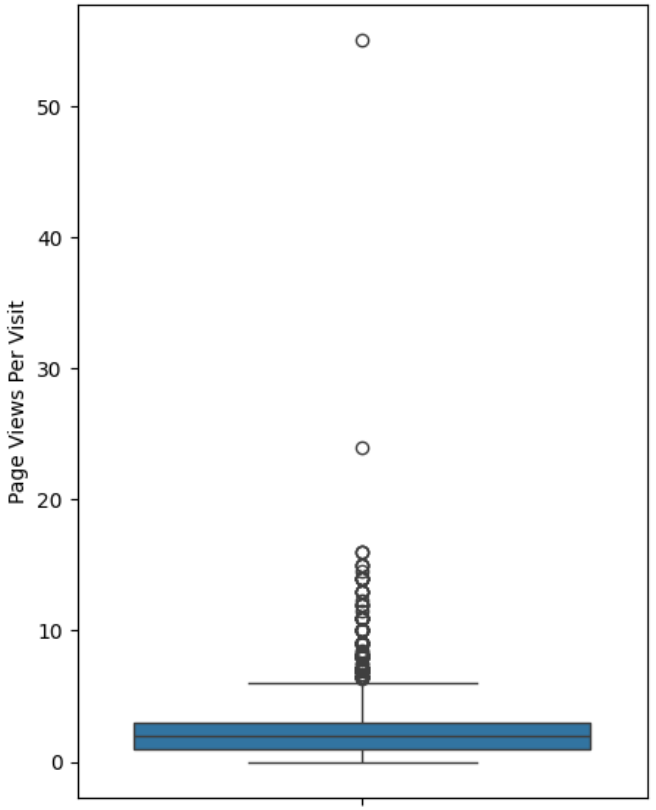
EXPLORATORY DATA ANALYTICS – NUMERICAL ANALYSIS (UNIVARIATE) (1/2)



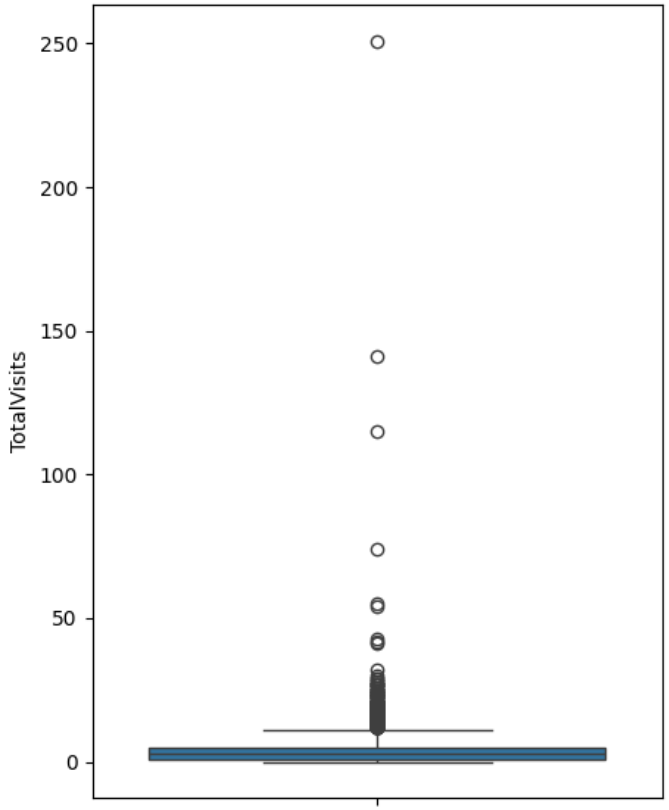
- Converted is a binary variable with 0 or 1 flag
- We don't see a huge class imbalance (62% vs 38%)

EXPLORATORY DATA ANALYTICS – NUMERICAL ANALYSIS (UNIVARIATE) (2/2)

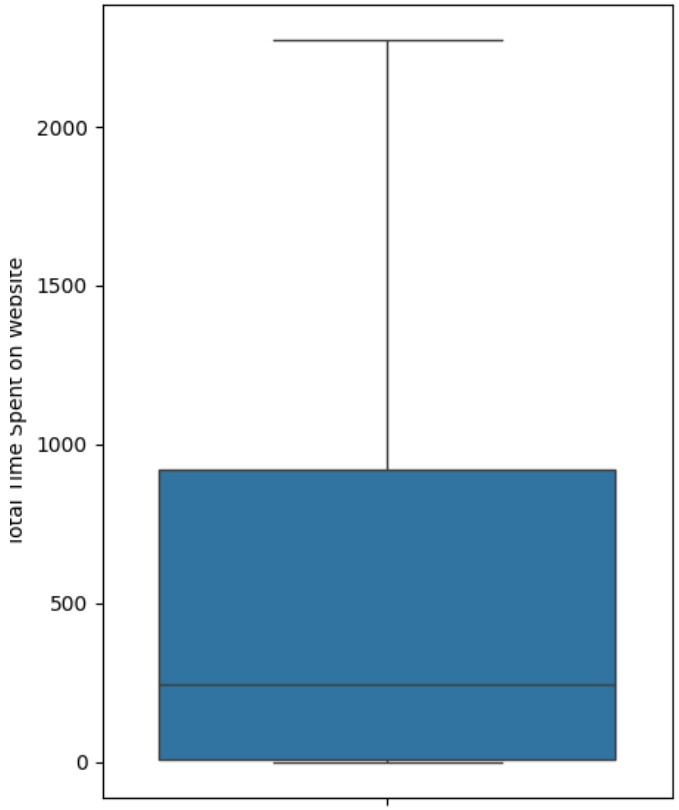
PAGE VIEWS PER VISIT



TOTALVISITS



TOTAL TIME SPENT ON WEBSITE

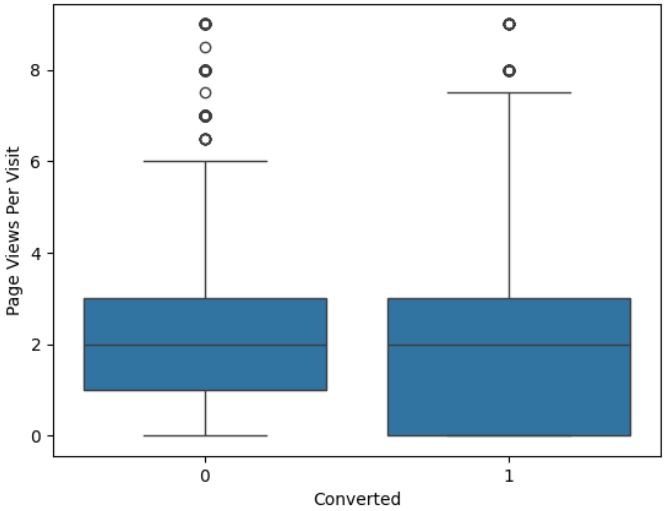


- Total Time Spent on Website doesn't have any outliers.
- TotalVisits and Page Views Per Visit is left skewed and the values above 99th percentile causes the issues. Hence lets cap the higher value at 99th percentile

EXPLORATORY DATA ANALYTICS – NUMERICAL ANALYSIS (MULTIVARIATE) (1/2)

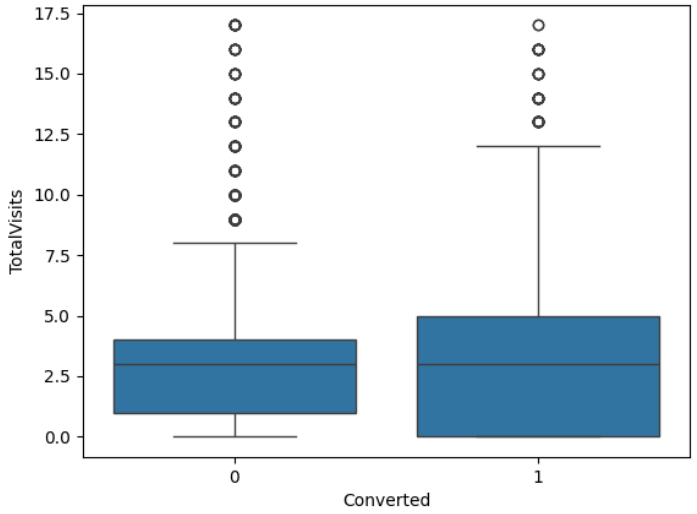
CONVERTED
VS

PAGE VIEWS PER VISIT



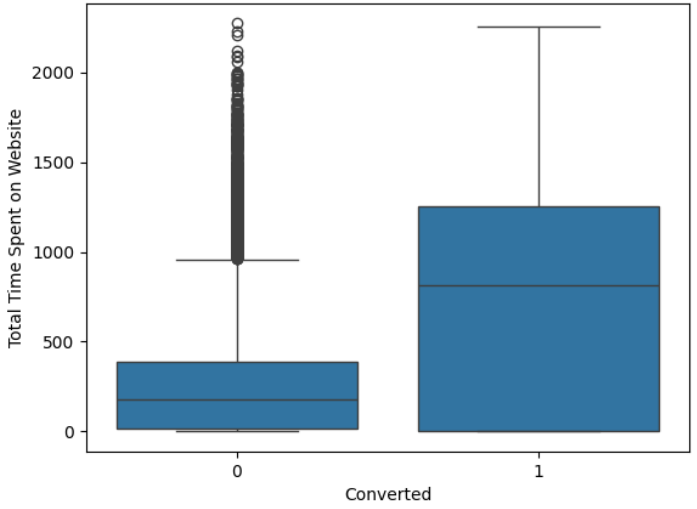
CONVERTED
VS

TOTALVISITS



CONVERTED
VS

TOTAL TIME SPENT ON WEBSITE



- Nothing conclusive can be said based on Total Visits, since the median of 1 and 0 are very close
- Nothing conclusive can be said based on Page Views Per Visit, since the median of 1 and 0 are very close
- More time spent on website, leads to more conversion of leads, since the median of 1 is much larger than 0 for Total Time Spent on Website

EXPLORATORY DATA ANALYTICS – NUMERICAL ANALYSIS (MULTIVARIATE) (2/2)

HEATMAP

- Reaffirms the above observation, since the correlation between Total Time Spent on Website is the highest wrt Converted
- TotalVisits and Page Views Per Visit are highly correlated. We can use anyone of this variable, but lets wait for RFE results to be sure

DUMMY VARIABLE CREATION

- For all the categorical variables, dummy variables are created before starting the modelling activity

- For columns without “Others”, “Not given” etc. cuts, we create dummy and drop first
  - Lead Origin, City, A free copy of Mastering The Interview, Do Not Email

- For columns with “Others”, “Not given” etc. cuts, we consciously drop the “Others” or “Not given” dummy variables in maintain the meaning of the variables
  - Lead Source, Last Activity, Specialization, Tags, Last Notable Activity, What is your current occupation

- Post dummification we have 73 variables in total

TRAIN TEST SPLIT AND NUMERICAL VARIABLES SCALING FOR MODELLING

- We create a 80%-20% random train and test split of the data
- All the numerical values in the train are standardized using Standard Scaler – Making the variables mean close to 0 and standard deviation close to 1
- Test data set will be standardized on the go, while scoring

LINEAR REGRESSION MODELLING

- 15 important features are selected using RFE on logistic regression models. All the numerical features are eliminated as a part of RFE
- Post this manual feature elimination was done based on p-val of the variables and VIF factor
- We have a final set of 13 features

Features	P-val	VIF	Variable importance	Effect
Tags_Closed by Horizzon	0	1.05	8.0130	Positive
Tags_Lost to EINS	0	1.03	6.7208	Positive
Tags_Will revert after reading the email	0	0.04	4.9413	Positive
Tags_Already a student	0	0.31	-4.0044	Negative
Tags_switched off	0	1.02	-3.9723	Negative
Tags_invalid number	0	1.01	-3.9129	Negative
Lead Source_Welingak Website	0	1.03	4.4460	Positive
Tags_Ringing	0	0.18	-3.1518	Negative
Tags_Not doing further education	0.007	1.03	-2.7571	Negative
Last Activity_SMS Sent	0	0.06	2.1861	Positive
Tags_Interested in other courses	0	0.42	-1.7319	Negative
Last Notable Activity_Modified	0	0.01	-1.6720	Negative
Tags_Interested in full time MBA	0.017	1.02	-1.4234	Negative
Lead Origin_Lead Add Form	0	0.75	0.8907	Positive

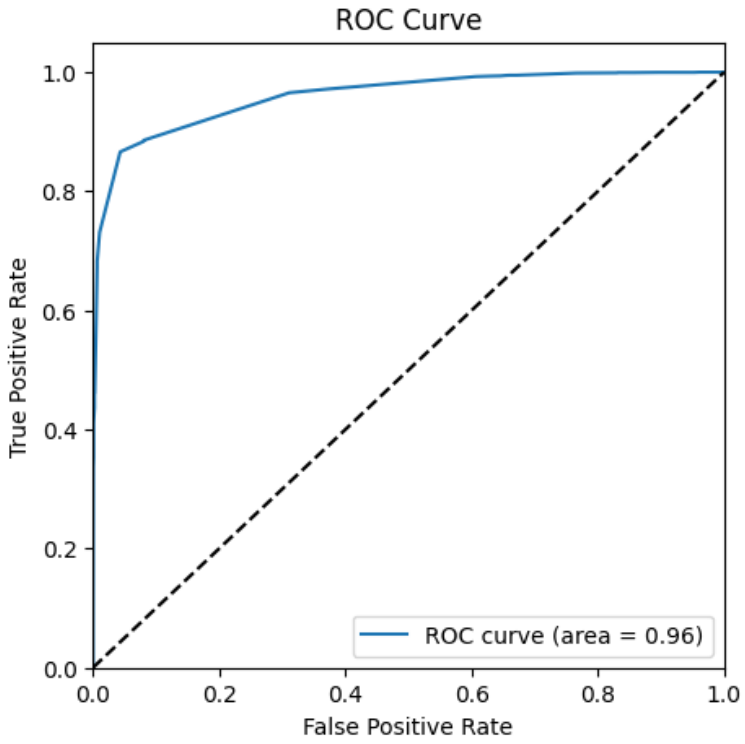
MODEL PERFORMANCE ON TRAINING DATA (0.5 CUTOFF)

CONFUSION MATRIX

Actual/Predicted	0	1
0	4250	192
1	364	2356

PERFORMANCE METRICS

Metrics	Performance
Accuracy	92.2%
Precision	92.4%
Recall/Sensitivity	86.6%
Specificity	95%



- The performance metrics are stable across all runs.
- Recall can be improved. And as expected, it shows the conversion rate is above 80% as per CEO's vision
- The ROC AUC is 0.96 which suggests that the model is a good one

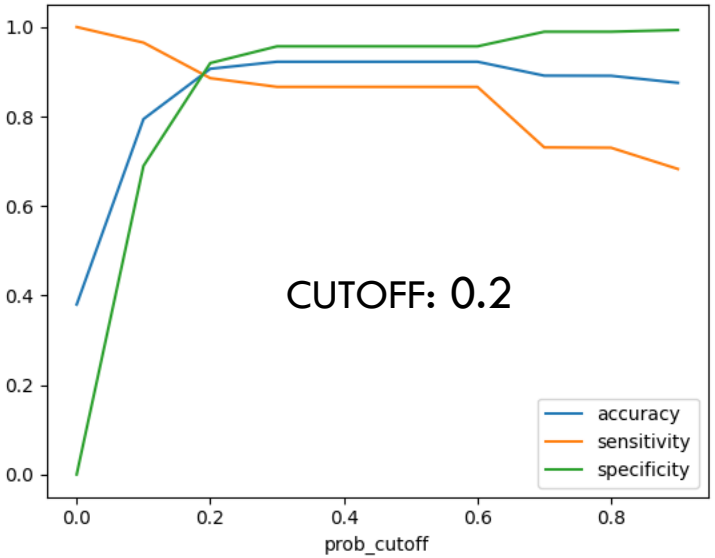
MODEL STABILITY – 5-FOLD CROSS VALIDATION ON TRAINING DATA

CV Run	Accuracy	Precision	Recall
1	93.2%	93.4%	88.4%
2	93%	92.2%	89.1%
3	91%	91.1%	84.7%
4	91.8%	92.8%	85.3%
5	91.9%	92.8%	85.5%
mean	92.2%	92.5%	86.6%
std	0.9%	0.8%	2%

- In the 5-fold cross validation on the training set, the stability of the model is pretty good
- We can see that the standard deviation for accuracy and precision is close to 1% and but recall is bit higher at 2% which is acceptable
- The mean of the performance metrics are in the acceptable range and shows the model performance is good throughout

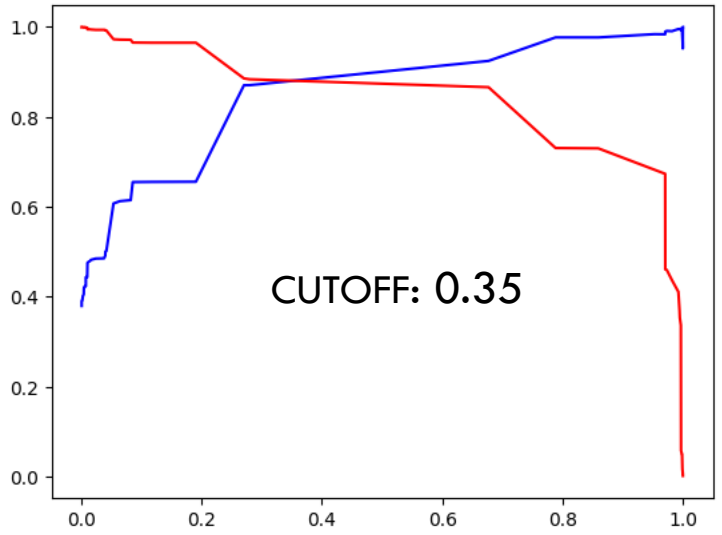
CUT-OFF ANALYSIS

ACCURACY, SENSITIVITY, SPECIFICITY ANALYSIS



Metrics	Performance
Accuracy	90%
Precision	87%
Recall/Sensitivity	88.5%
Specificity	91%

PRECISION VS RECALL



Metrics	Performance
Accuracy	92.2%
Precision	92.5%
Recall/Sensitivity	86.6%
Specificity	95.6%

- The performance of 0.35 cut-off is marginally good since it had better accuracy, precision, specificity than 0.2 cut-off model.
- It is slightly off in recall, but that's fine. Since, counting a lead as potential conversion case even if may not is better than leaving a potential conversion case out.

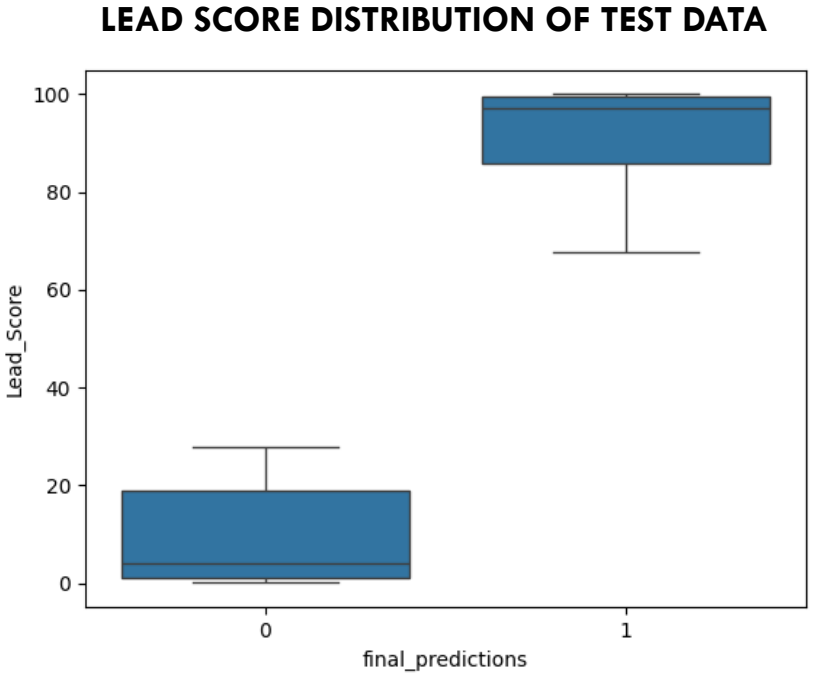
TEST DATA PERFORMANCE (CUTOFF – 0.35) AND LEAD SCORE

CONFUSION MATRIX

Actual/Predicted	0	1
0	1060	56
1	88	587

PERFORMANCE METRICS

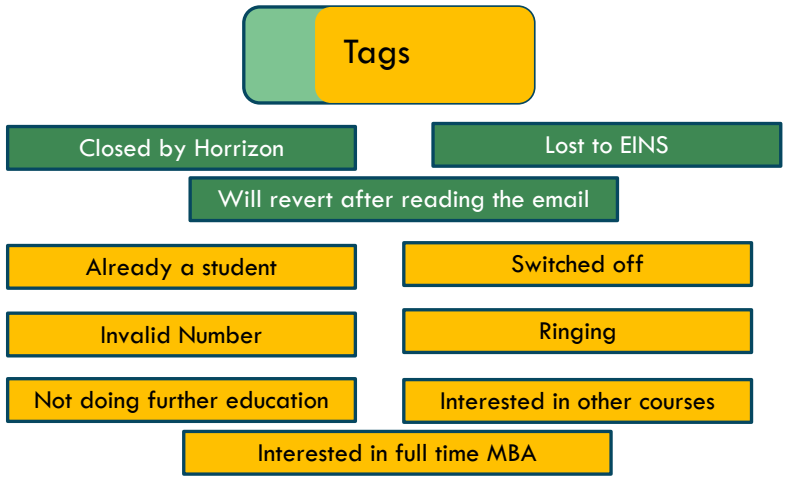
Metrics	Performance
Accuracy	92%
Precision	91.2%
Recall/Sensitivity	87%
Specificity	95%

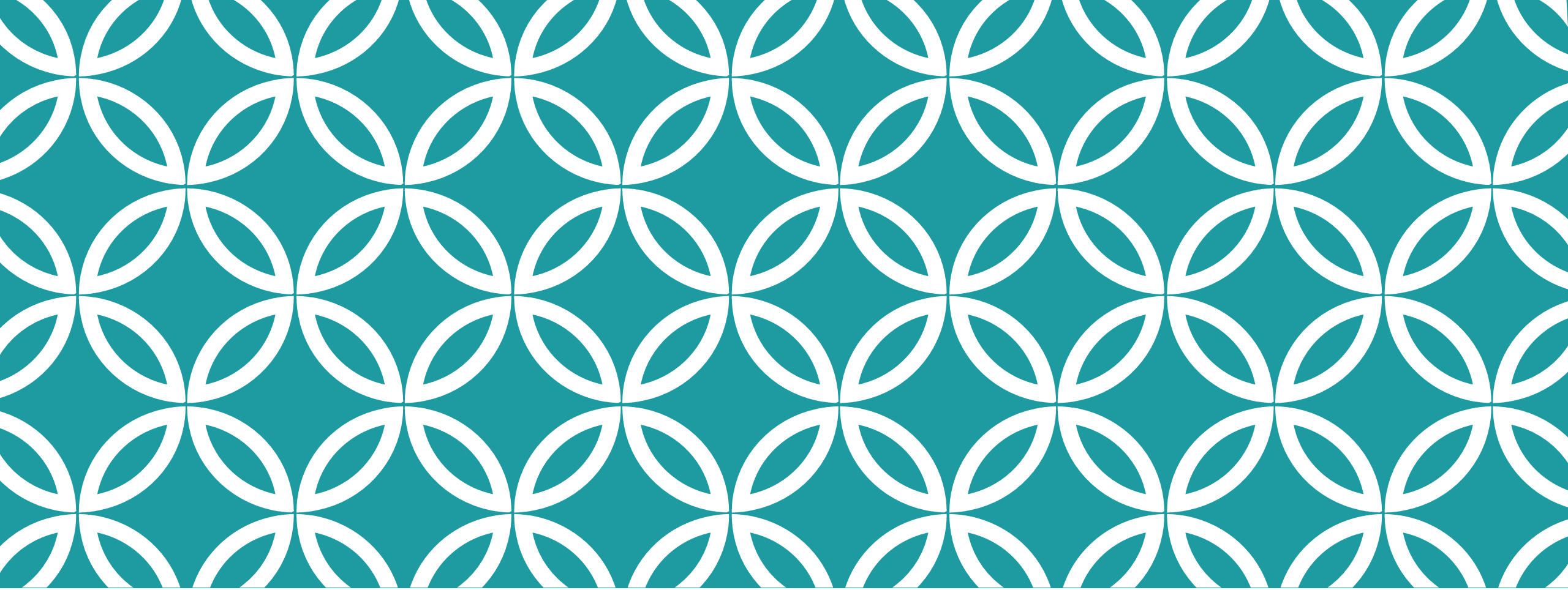


- The performance on the test data is very good and is consistent with training data
- Lead score the 100*Prediction Probability
- As expected in the problem statement, we get higher scores for hot leads which have a greater potential to get converted.

CONCLUSION

- The current status of the leads, denoted by the Tags places a very crucial part in scoring the leads
  - 10 out of 13 important variables are based on Tags
  - Tags like closed by Horizzon, Lost to EINS and Will revert after reading email has higher positive impact towards conversion
  - Tags like Already a student, switched off and invalid number has a high negative impact towards conversion. These tags makes sense logically.
- Last Notable Activity with Modified values has a negative impact towards conversions.
- Last Activity of SMS sent has a good positive impact on lead conversions too.
- Welingak Website as a Lead Source has a reasonable positive impact towards conversions
- Leads originating from Lead Add forms have a low but positive impact towards conversions
- Tags such as Ringing, Not doing further education, Interested in other courses and Interested in full time MBA has a reasonable negative impact on towards the conversion





THANK YOU !