LEAD SCORING CASE STUDY

LOGISTIC REGRESSION

PROBLEM STATEMENT

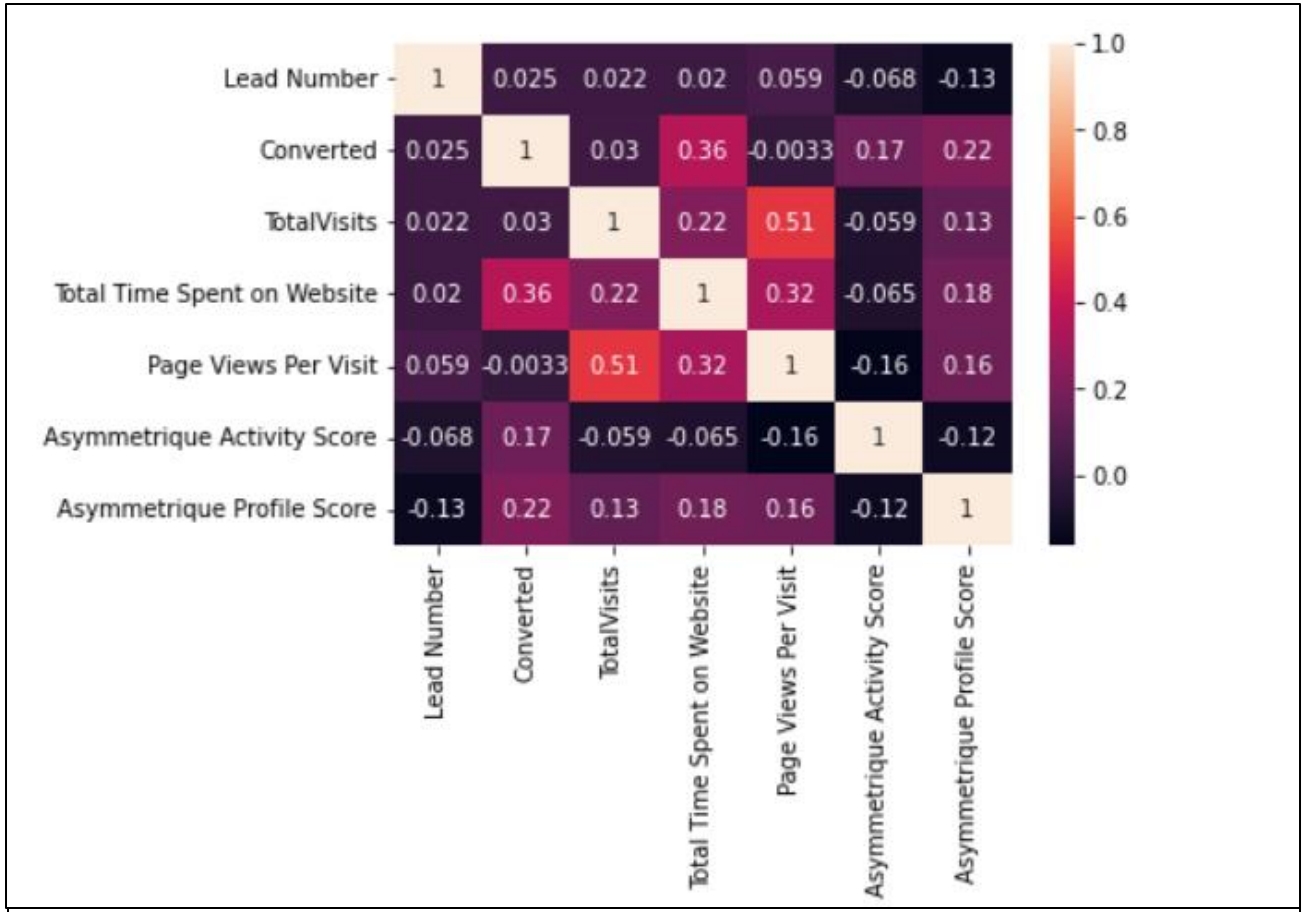
- The company requires us to build a model wherein we need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.
- This is classification machine learning problem. In this case study, we must determine **whether the given lead is highly likely to convert or not** based on given information.

APPROACH

- Exploratory data Analysis
- Cleaning of data
- Scaling of data
- Modelling of data
- Evaluation using various metrics
- Change of threshold
- Applying the technique to test set

EXPLORATORY DATA ANALYSIS

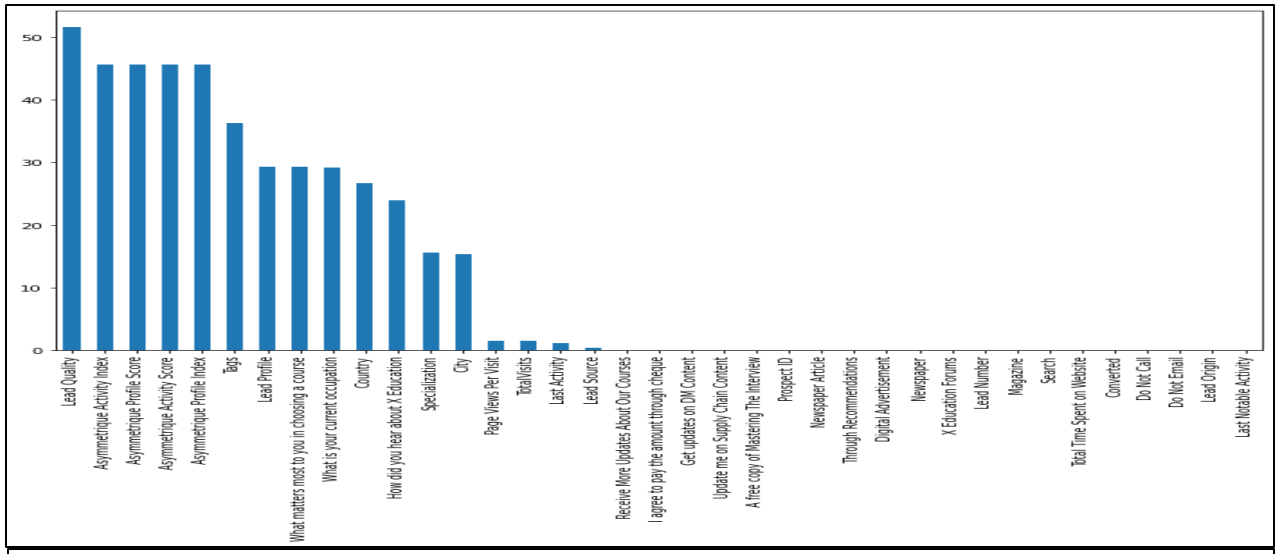
- We have found features are not highly correlated with each other, although there seems to negative correlation between some of the features.



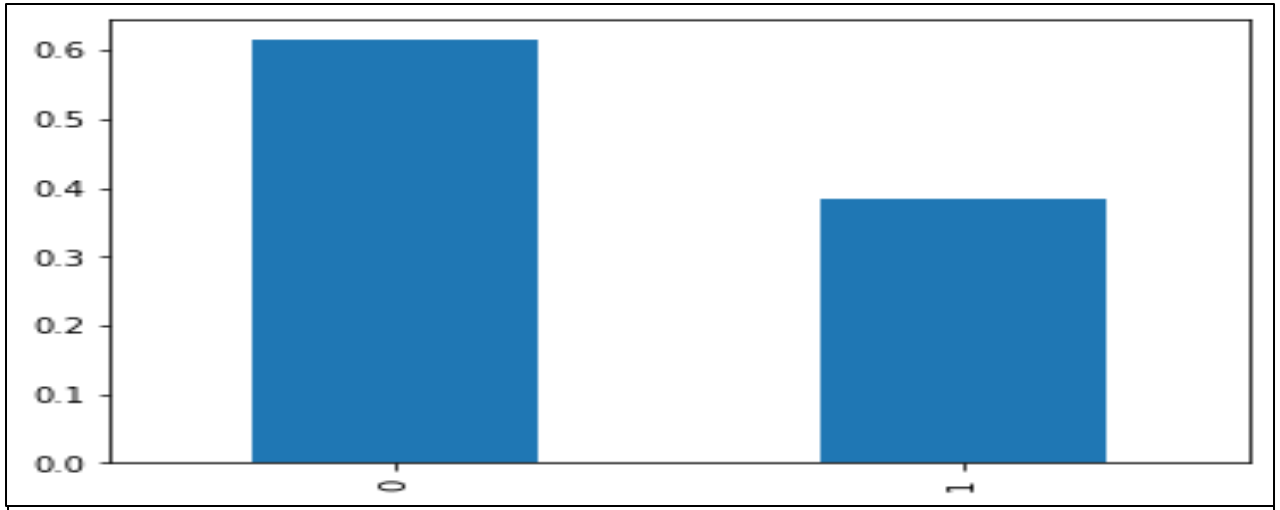
Correlation matrix

- There seems to be high number of nan values.
- Features which have less number of nan values are being imputed with median or most frequent values.
- Features which have high number of nan values are being imputed with ‘missing’ and -1.

- Also, we have around 39% of lead as converted.



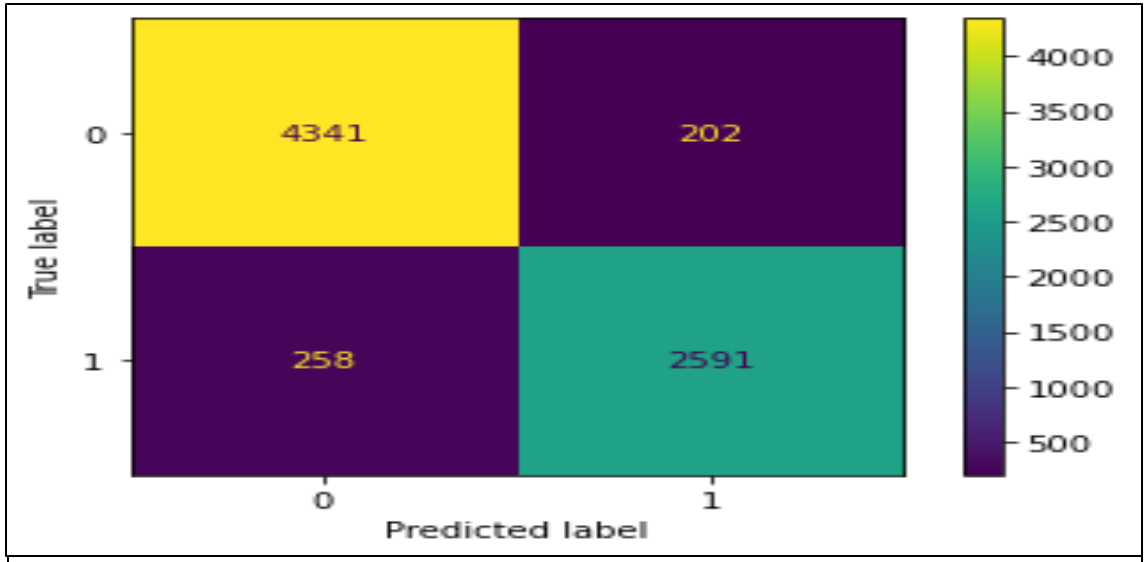
Bar plot of nan values



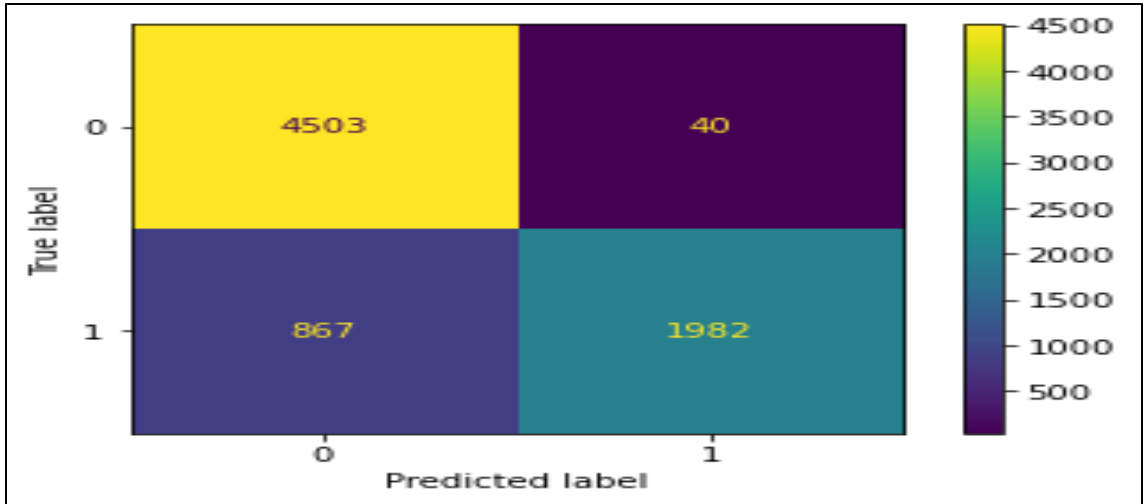
Bar plot of Target feature ‘Converted’

MODEL EVALUATION ON TRAIN SET

- Training set have high number of False positives as we can see in first confusion matrix.
- Our objective being to flag most potential leads, model's precision must be high.
- After changing threshold, False positives have become drastically down.



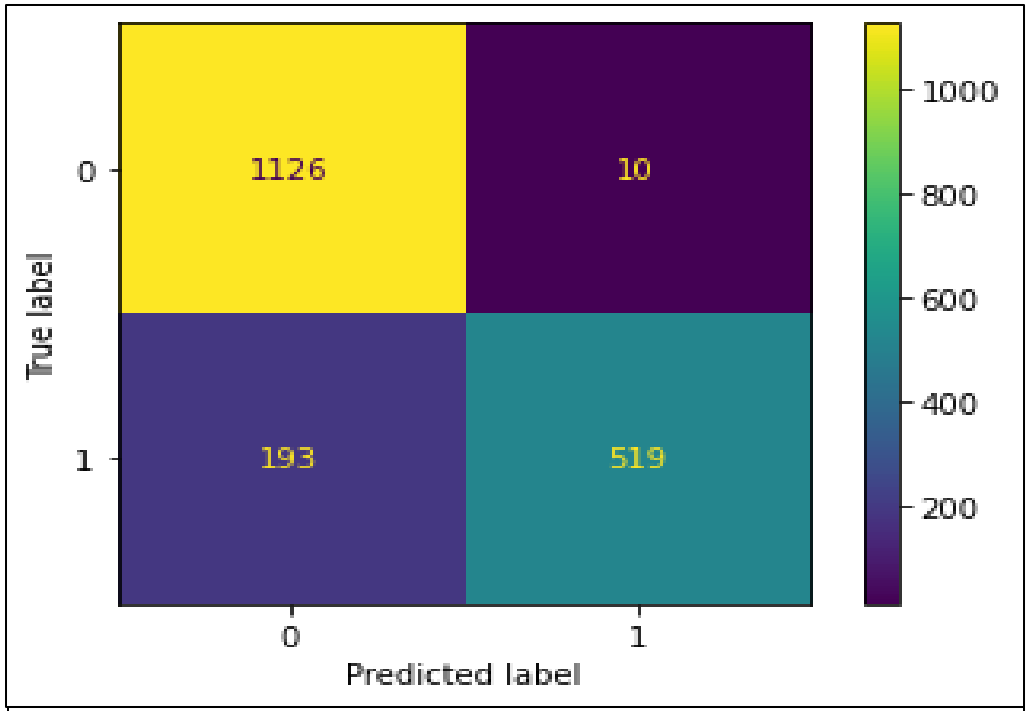
Confusion matrix of training set



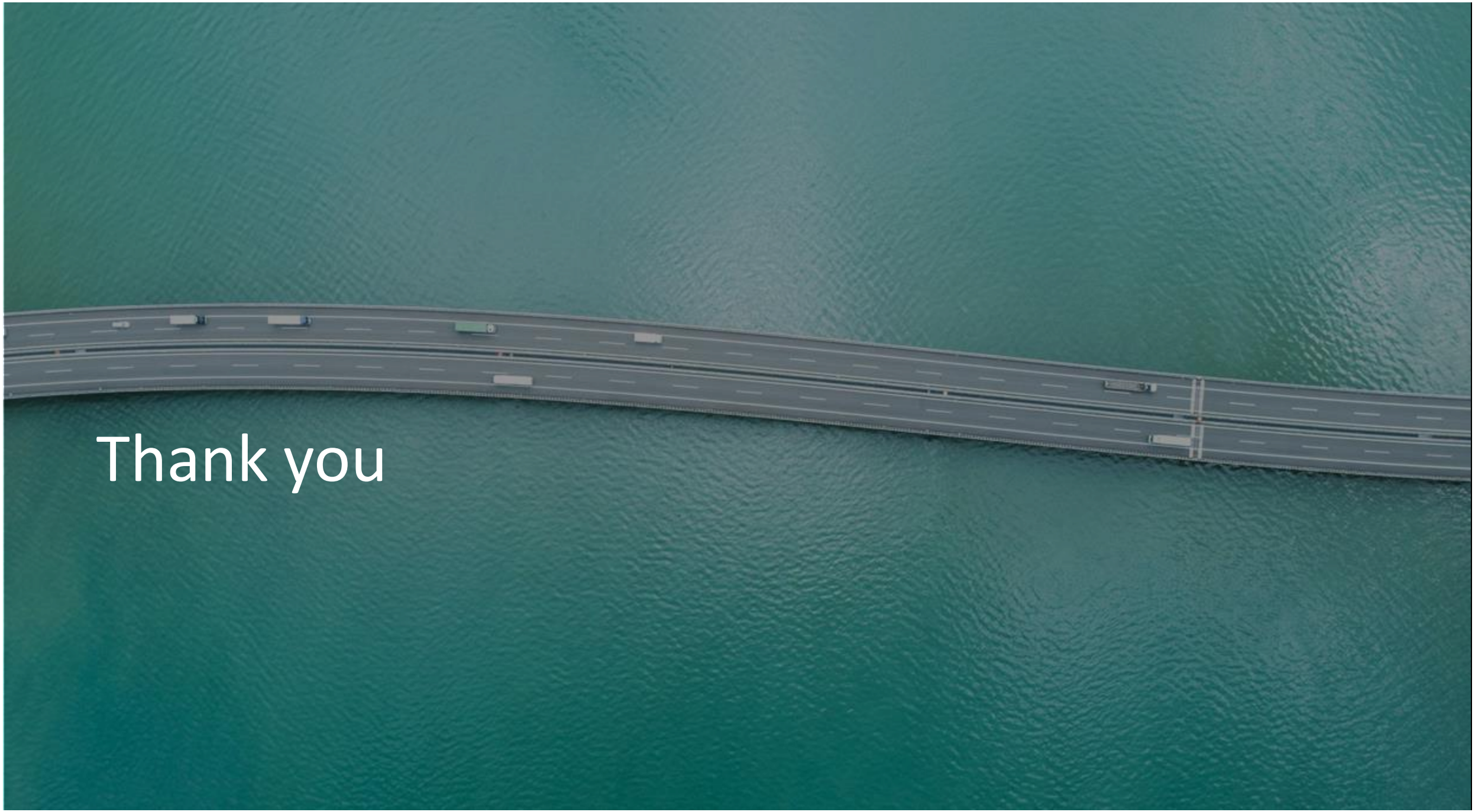
Confusion matrix of training set after increased threshold

FINAL EVALUATION ON TEST DATA

- As we can see our model is doing well on test set, the False positives are very low in number.
- Also, our model is giving around 92% accuracy.



Confusion matrix of test set after increased threshold


Thank you