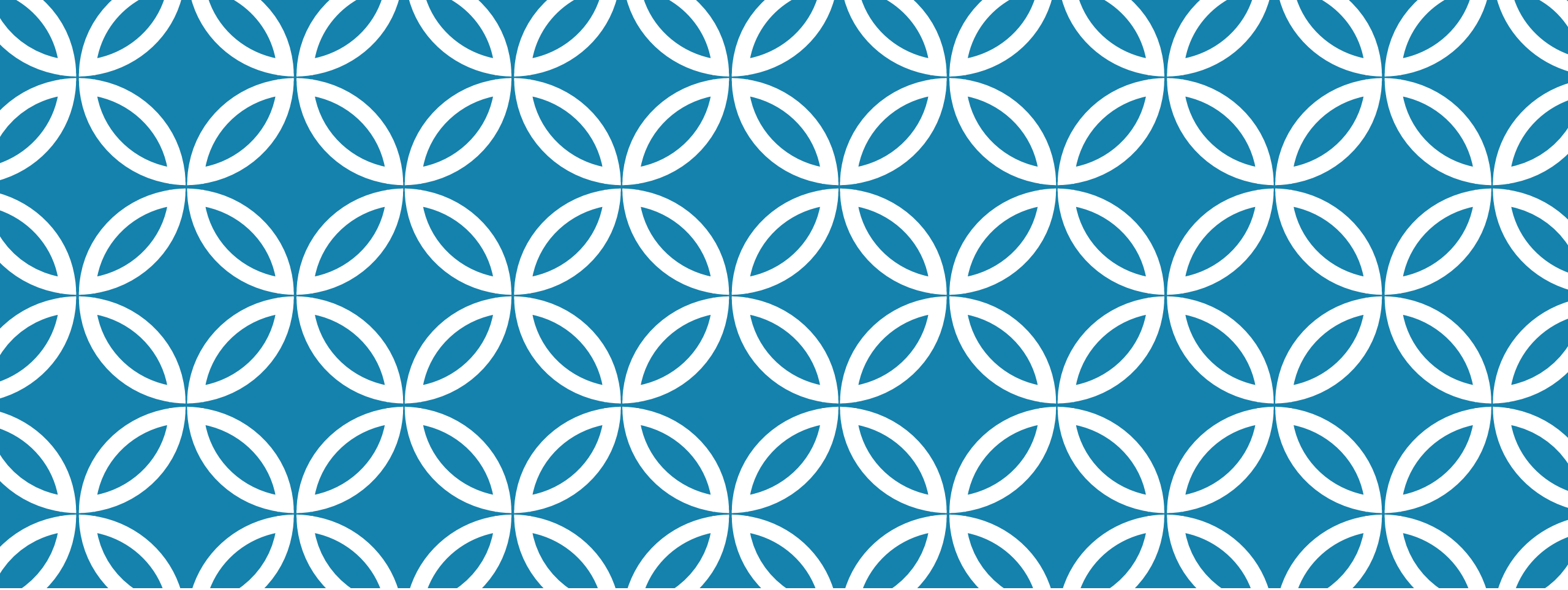


LEAD SCORING CASE STUDY

Group Members:

Arvind M

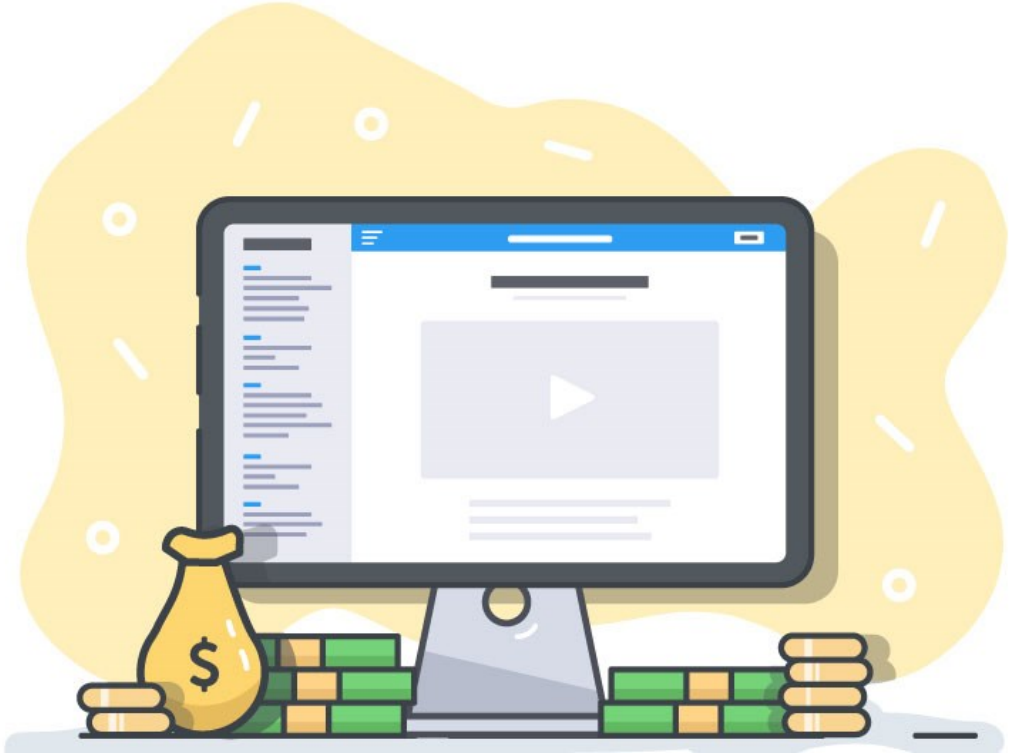
Bhakti K

Pragya B

PROBLEM STATEMENT

X Education sells online courses to industry professionals.

Through various marketing strategies, professionals are led to the company's website where they provide their contact details, hence, becoming a lead.



PROBLEM STATEMENT

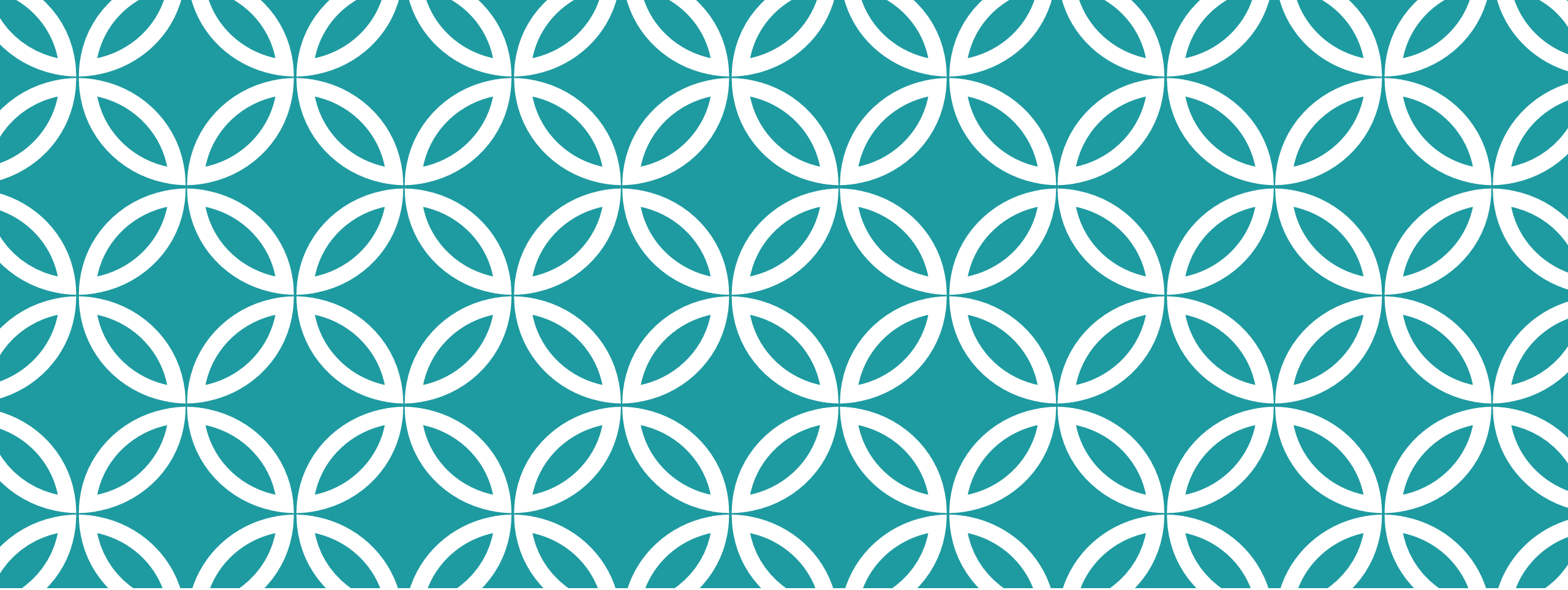
Currently the lead conversion rate for this company after sales communications is about 30%. To make efficient use of its resources, X Education wishes to identify the most potential leads – the "Hot Leads". They believe that doing so will lead to a rise in the lead conversion rate as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

We are required to build a model where we assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

OBJECTIVE

To build a model that:

- Helps us identify the most promising leads
- That gives results after adjusting for company's future requirements



APPROACH

IMPORTING DATA AND DATA CLEANING

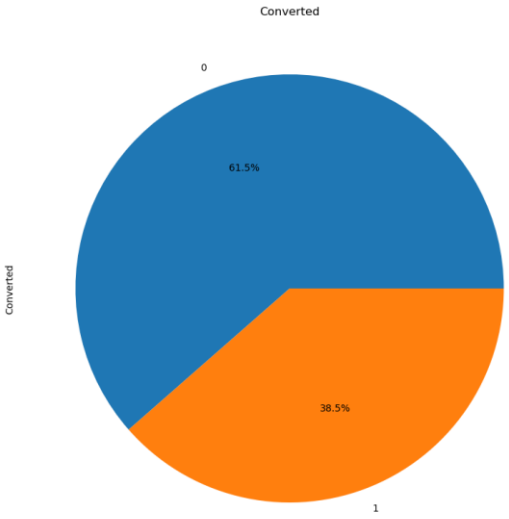
- Data description
- Identifying datatypes
- Handle missing values
  - Replaced numeric null values with mean
  - Found out null value percentage for string and replaced “Select” with null
- Classifying columns into Categorical, Numeric and Target
- Identifying necessary columns:

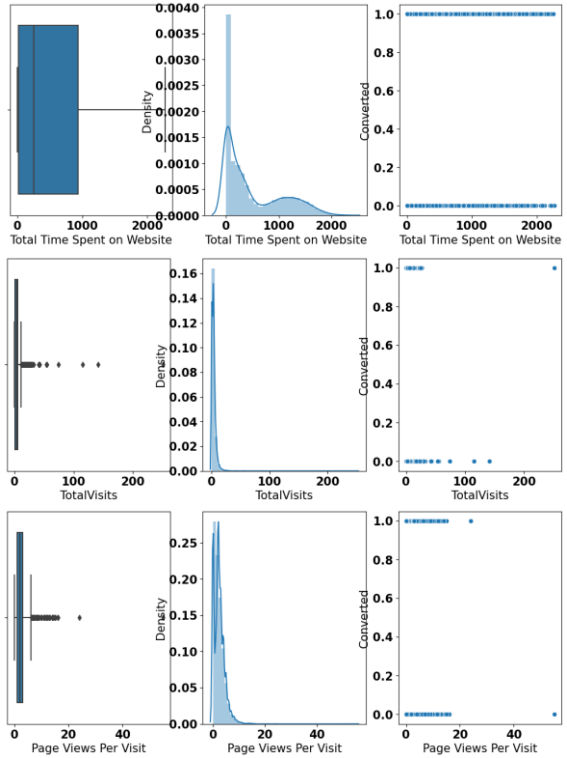
IMPORTING DATA AND DATA CLEANING (CONTINUED)

- Identifying necessary columns:
  - Do Not Email
  - A free copy of Mastering The Interview
  - Lead Source
  - Total Time Spent on Website
  - Page Views per Visit
  - Converted (Target variable)
  - Lead Origin
  - Last Activity
  - Total Visits
  - Last Notable Activity

EXPLORATORY DATA ANALYSIS

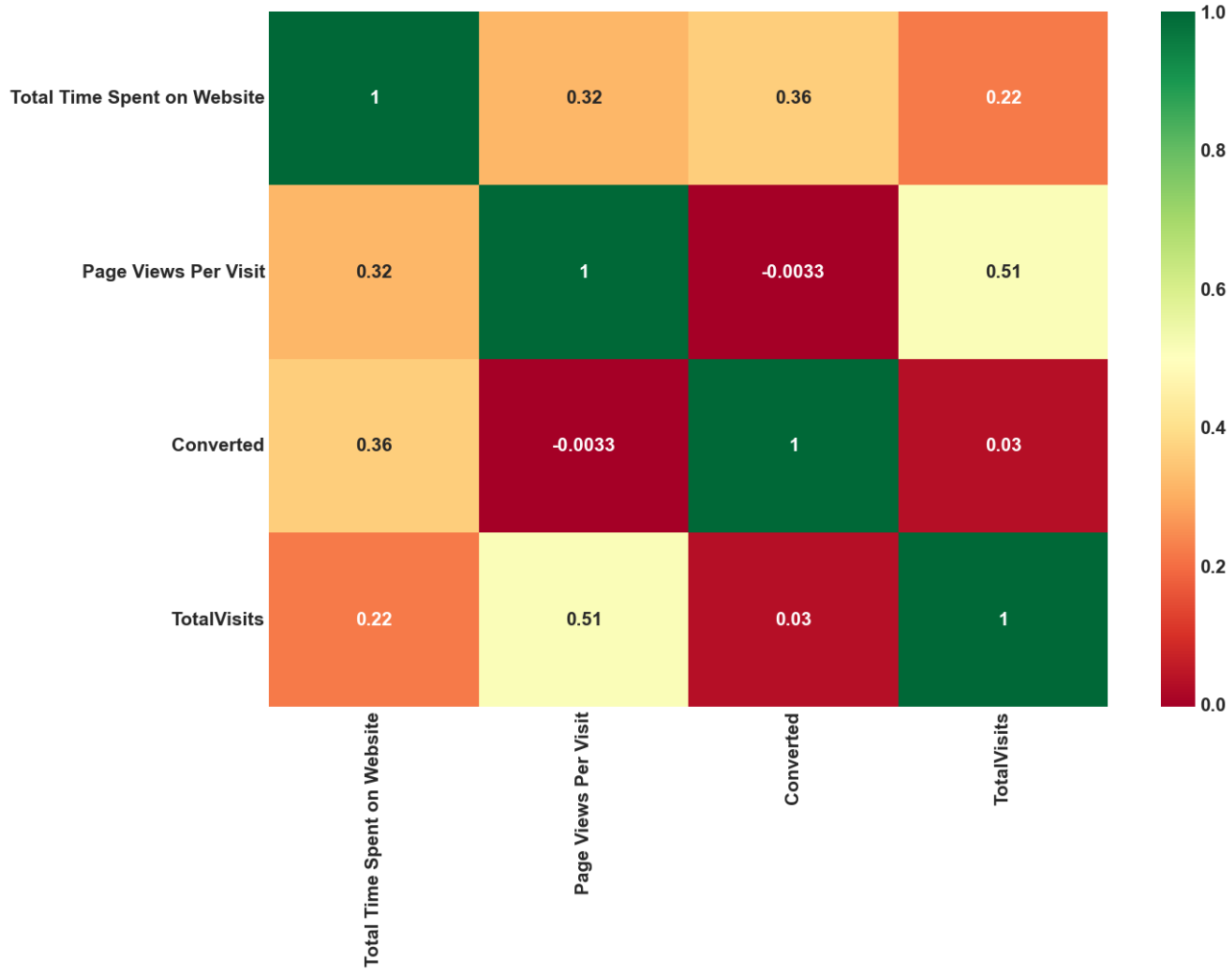
- Visualisation
  - Pi Chart to indicate Converted and Non-converted leads
  - Charts for numerical columns
  - Bar Charts for categorical columns





EXPLORATORY DATA ANALYSIS

- Correlation Matrix

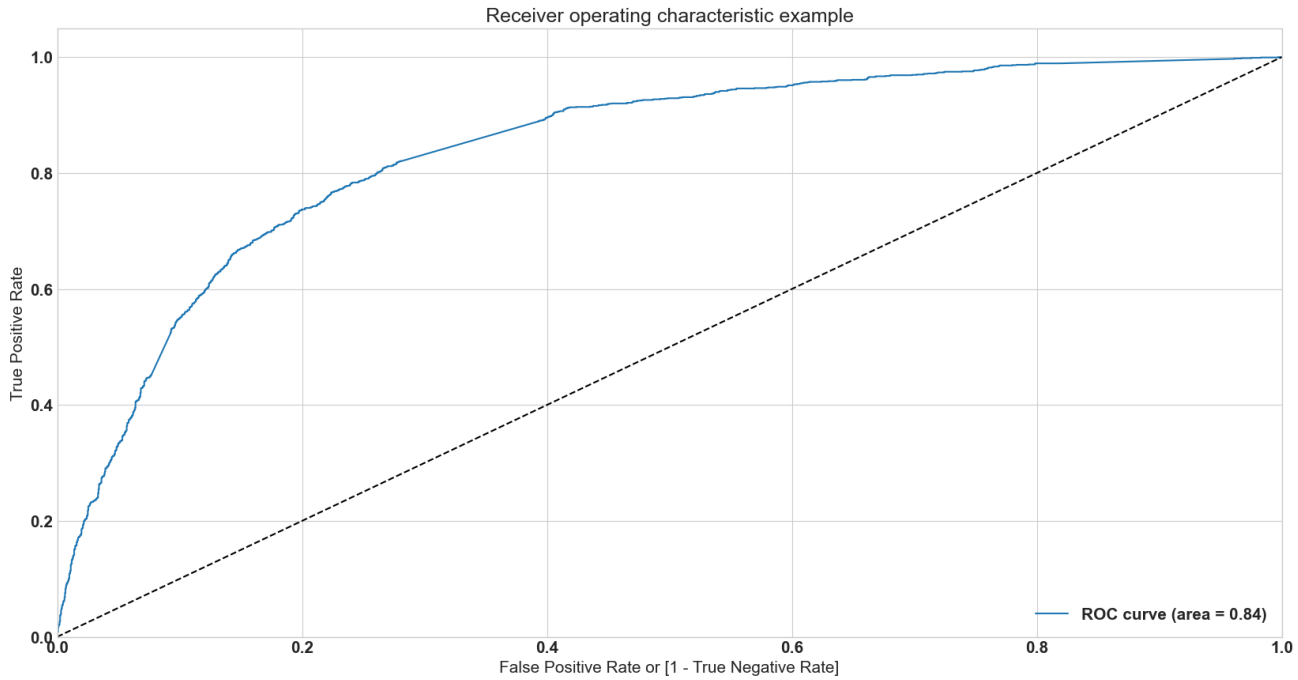


EXPLORATORY DATA ANALYSIS

- Creating Dummy variables for categorical columns with non-binary values
- Outlier Treatment
- Normalising of continuous variables:
  - Total Visits
  - Total Time Spent on Website
  - Page Views Per Visit

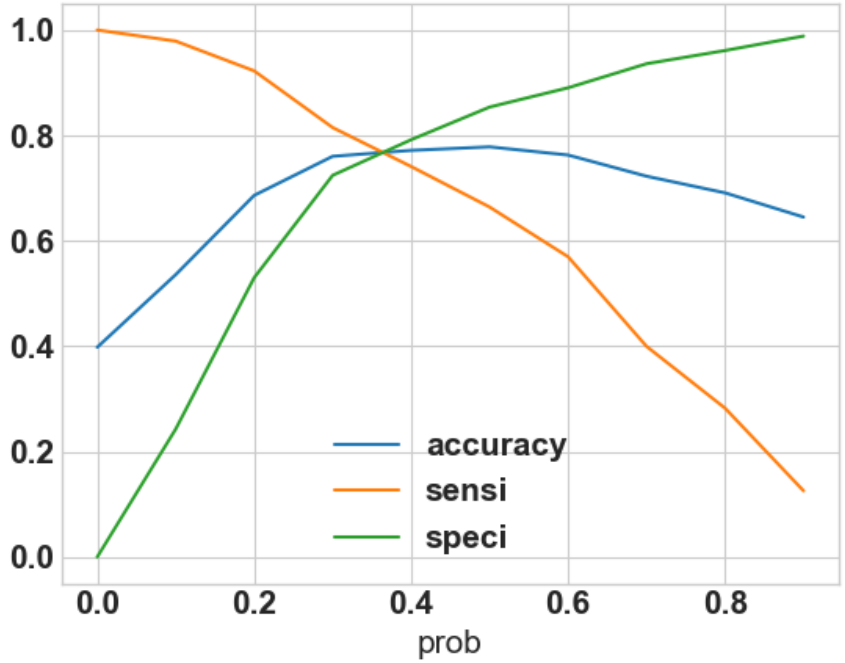
BUILDING THE MODEL

- Splitting data into Train and Test
- Feature Selection by RFE and then, by checking VIF
- After dropping 9 columns, we find the ROC for Model#10



MODEL EVALUATION

- Finding Optimal Cutoff Point
  - We do this to identify the probability where sensitivity and specificity is balanced.
  - Optimum Cutoff Point for our model comes out to be 0.42

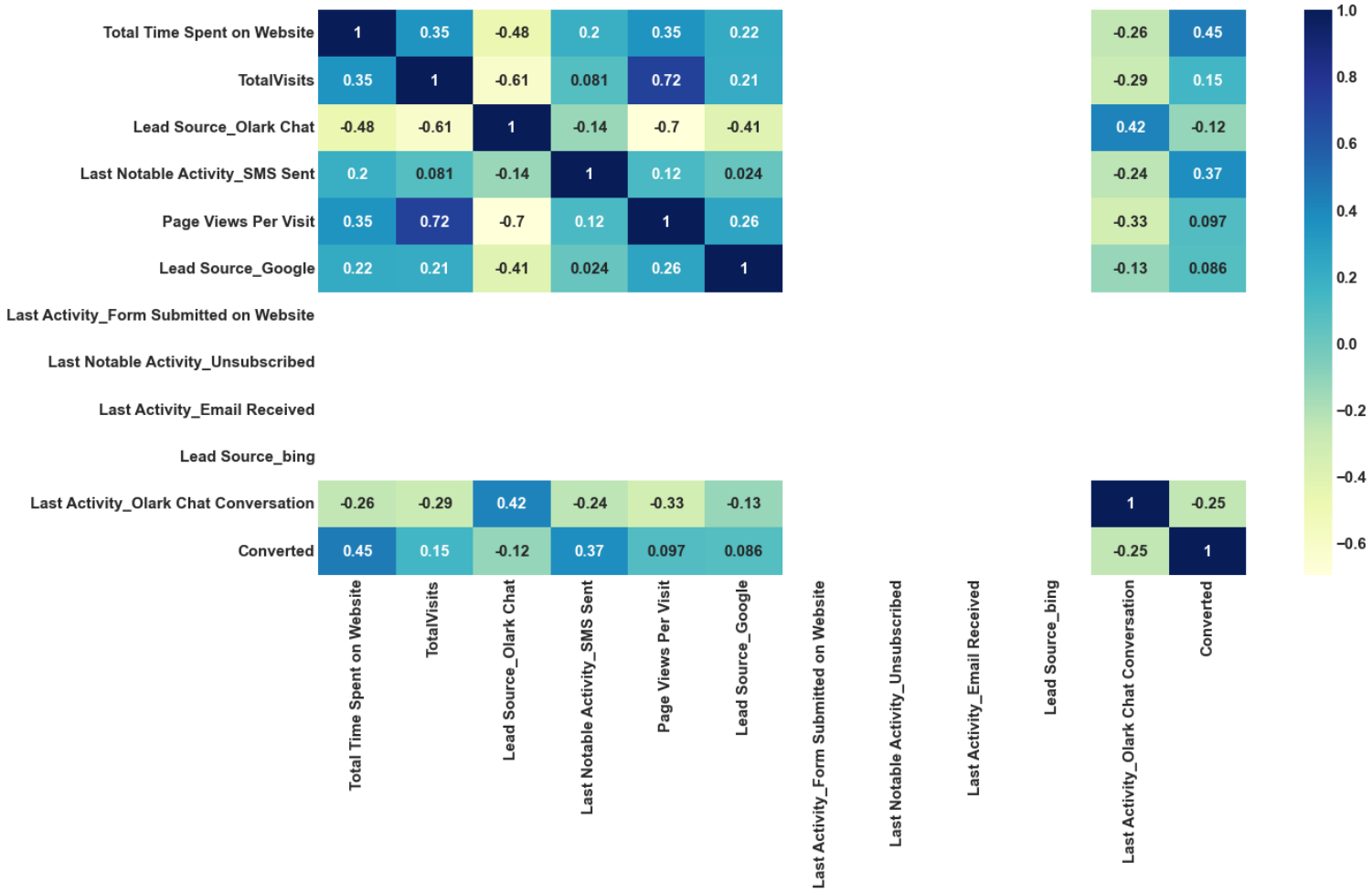


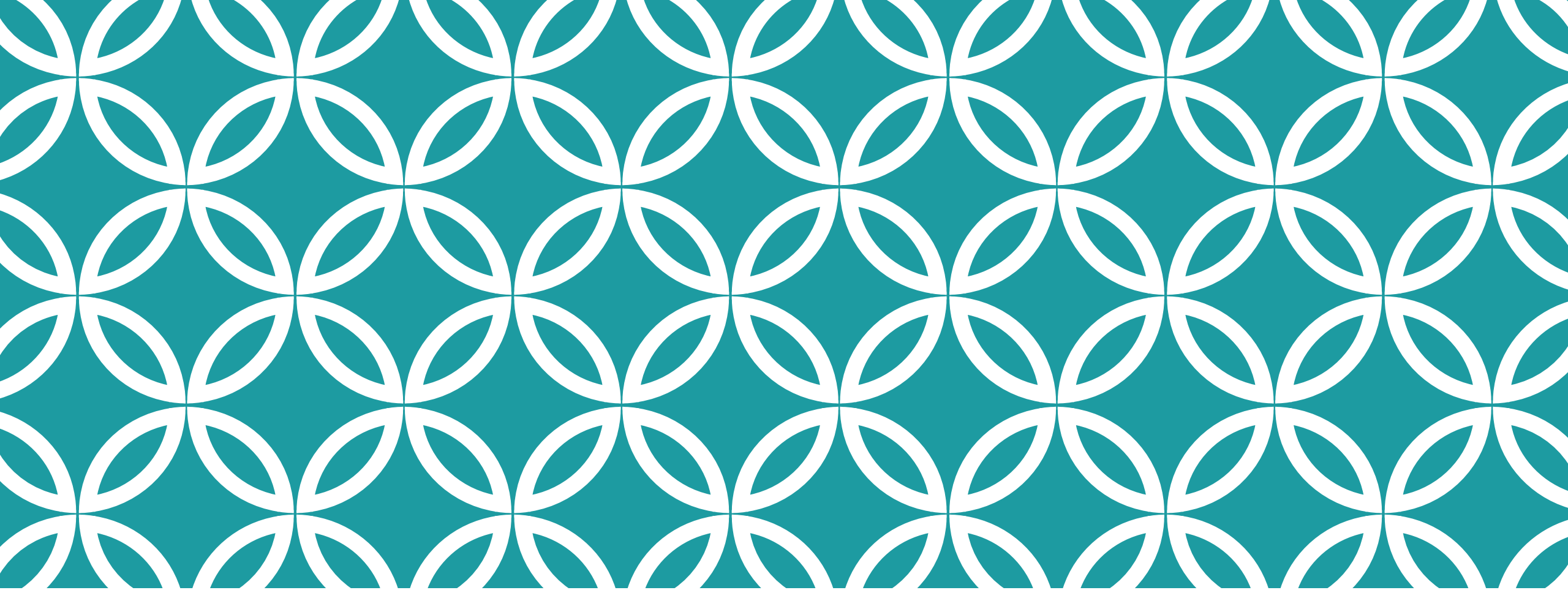
MODEL EVALUATION

- Check overall accuracy using confusion matrix
  - Result:
    - sensitivity: 73.0%
    - specificity: 81.0%
    - precision: 71.0%
    - recall: 73.0%
    - f1 score: 72.0%
- Make predictions on test set
- Check Correlation Matrix

MODEL EVALUATION

- Check Correlation Matrix





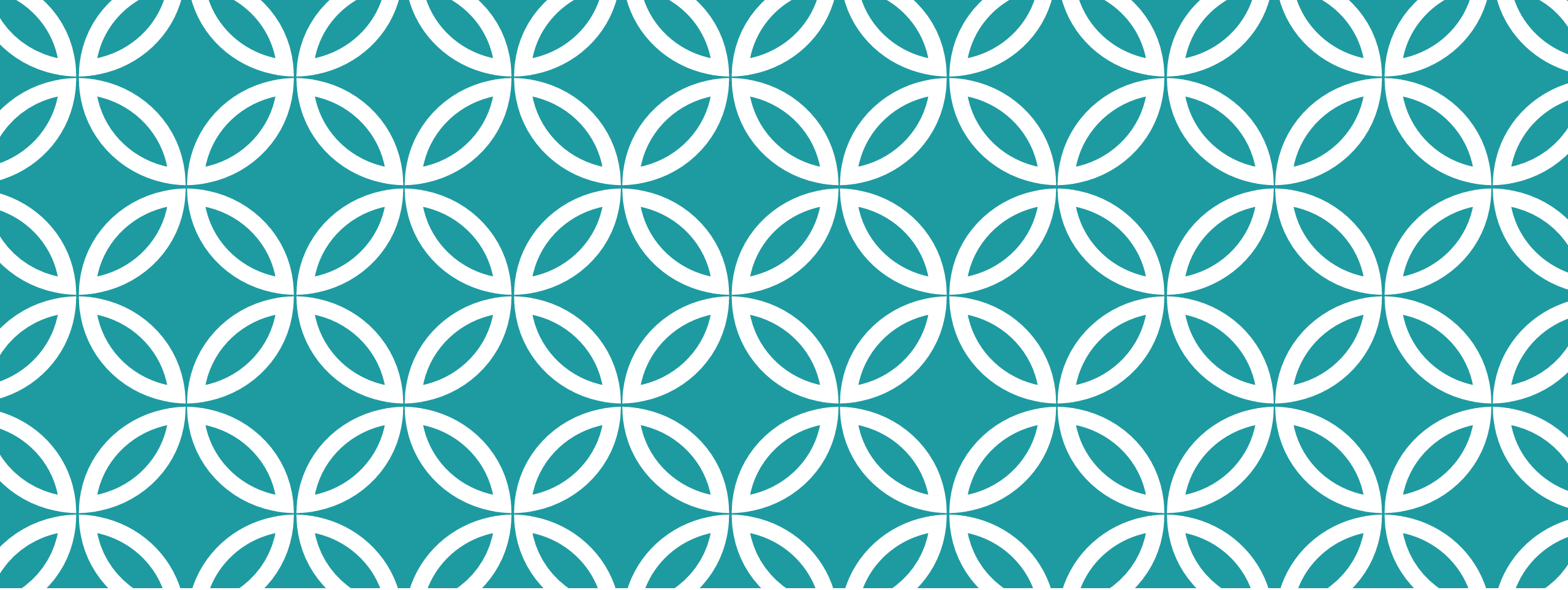
CONCLUSIONS AND IMPLICATIONS

Final Model Summary:

- Accuracy : 76%
- Other statistics:
  - Sensitivity: 73.0%
  - Specificity: 81.0%
  - Precision: 71.0%
  - Recall: 73.0%
  - f1 score: 72.0%

Top 3 indicators are:

- Total Time Spent on Website
- Last Activity_Olark Chat Conversation
- Last Notable Activity_SMS Sent



RECOMMENDATIONS

- They should target people whose current occupation is unemployed / working professionals, as it might help in filtering if the call to be made or not.
- People who are not only visiting the site but the time they are spending should also be observed
- Referrals should be given a high priority, as are low hanging fruits
- Anyone visiting the website repeatedly should also top the list as it clearly shows their interest
- Leads through olark chat are also having high probability of conversion