Case Study

“Lead Scoring”

Dated: 26th August 2019

Submitted By:

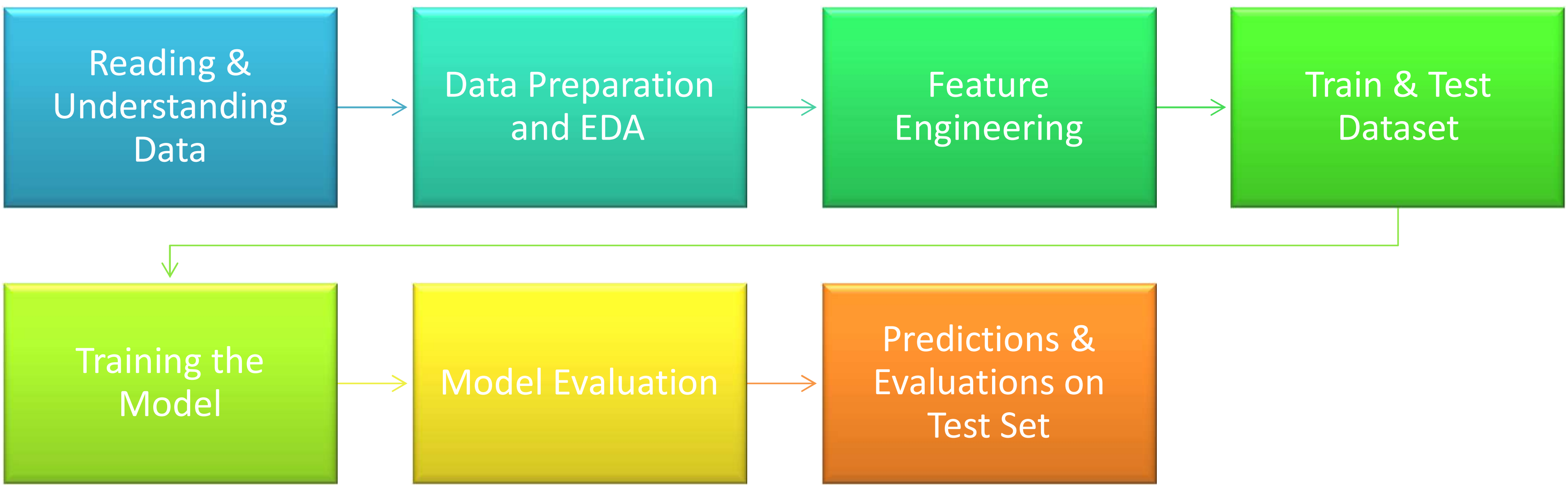
Sangeeta Kalra & Vikram mathur

Problem Statement & Overall Approach

Problem Statement

- Building a logistic regression model to assign a lead score of 0 to 100 to each lead 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.
- A ballpark of the target lead conversion rate to be around 80%.

Overall Approach



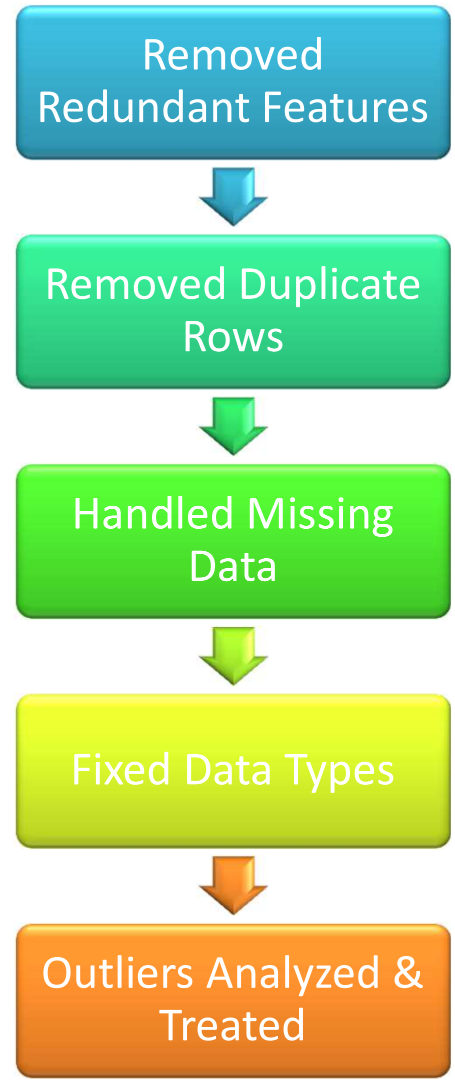
Overall Approach

- Step 1: Reading and Understanding Data
  - 1.1 Running Pandas Profiler
- Step 2: Data Preparation & EDA
  - 2.1 Removing Redundant Features
  - 2.2 Check for Duplicates
  - 2.3 Check & Treat Missing Values
  - 2.4 Check & Fix Datatypes
  - 2.5 Outlier Analysis & Treatment
  - 2.6 Numerical Features Analysis
  - 2.7 Categorical Features Analysis
- Step 3: Feature Engineering - Data Preparation
  - 3.1 Derived Metrics
  - 3.2 Dummy Encoding for Unordered Categorical Variables
- Step 4: Split Data into Training and Test Sets
- Step 5: Training the Model
  - 5.1 MinMax Scaling
  - 5.2 RFE
  - 5.3 Building model using statsmodel, for the detailed statistics
- Step 6: Model Evaluation
  - 6.1 Predictions
  - 6.2 ROC Curve
  - 6.3 Finding Optimal Cutoff Point
  - 6.4 Assign Lead Score
- Step 7: Predictions and Evaluation on the Test Set

- Reading & Understanding Data
- Data Preparation & EDA

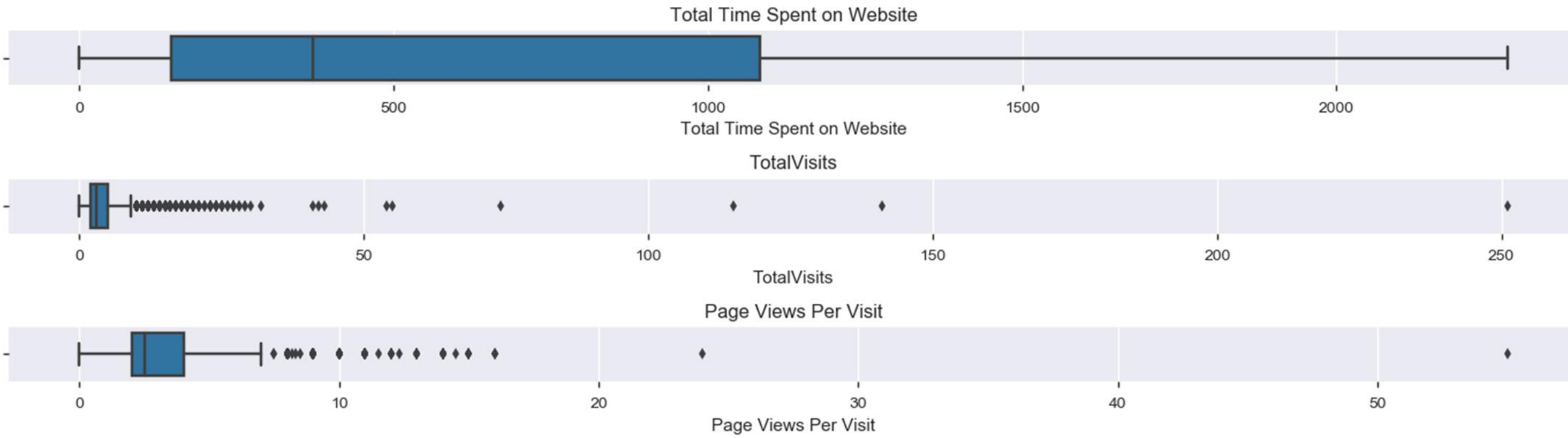
Data Preparation

- Removed Redundant Features
  - Removed Features with Constant Features
  - Removed Features with 95% Constant Value
  - Removed ID Feature having all Unique Values
  - Removed Features with High Missing Values
- Removed Duplicate Rows
- Handled Missing Data
  - Removed rows with high % missing values
  - Handling Missing data for Continuous Numerical Features
  - Handling Missing Data for Categorical Variables
    - Imputed with Mode
    - Imputed with an “unknown” value
- Fixed Data Types
- Outliers Analyzed & Treated

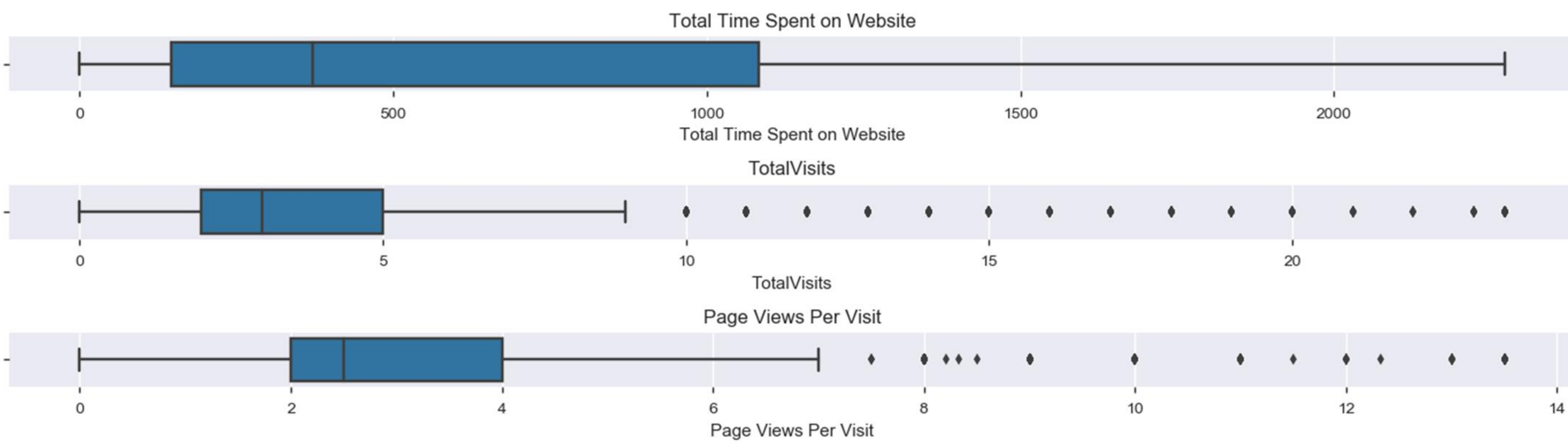


Prepared DataFrame Shape (7323, 13)

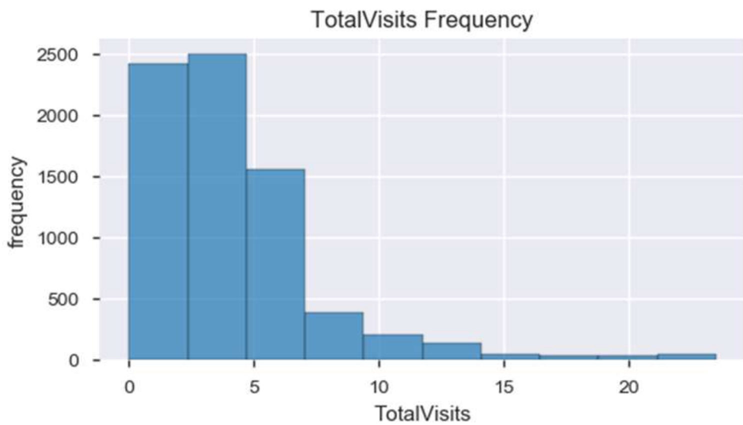
Outlier Analysis & Treatment



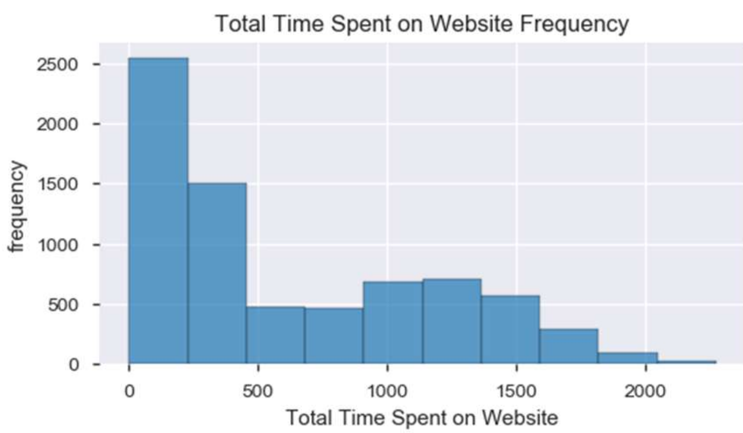
- Treated outliers based on IQR 0.5 to 0.95



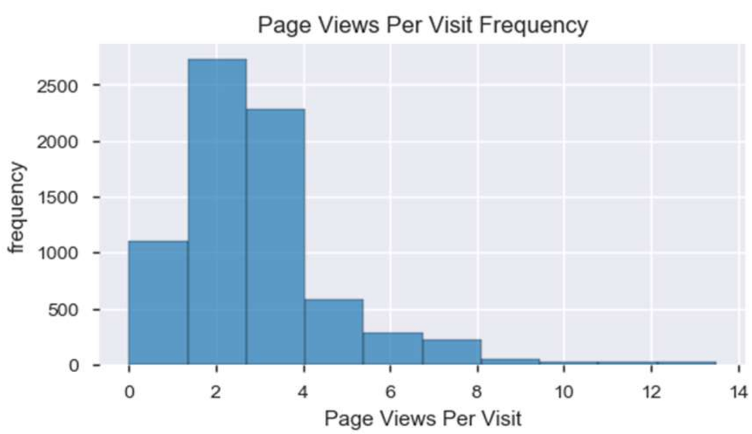
Numerical Feature Analysis



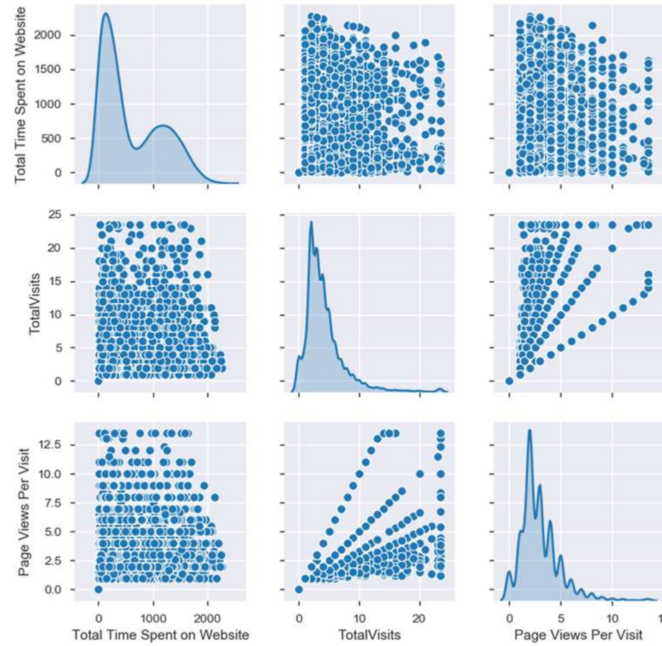
Range of 0 to 7 maximum density



0 to 500 range has highest density. Range 900 to 1600 has a healthy frequency.

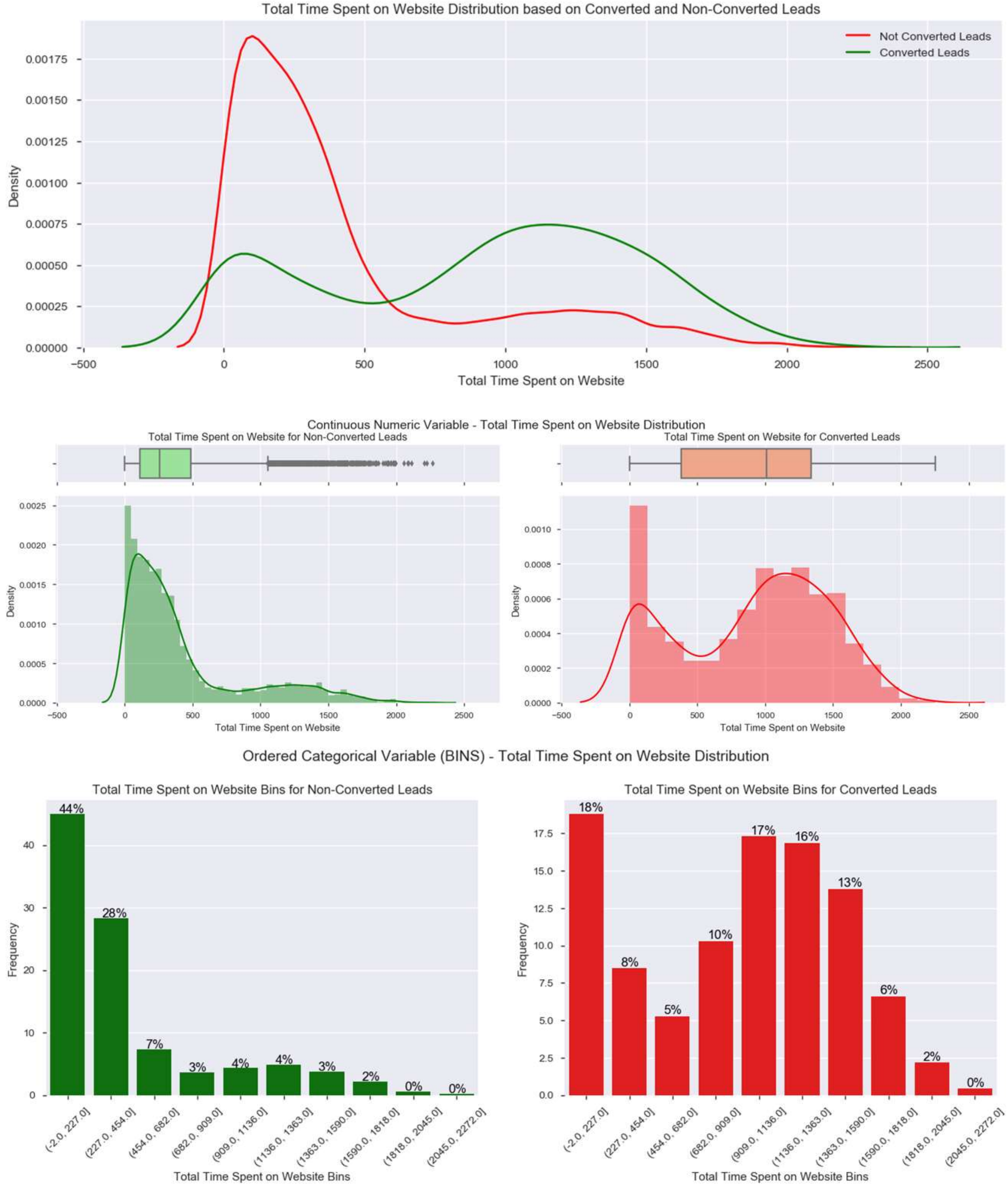


Range of 1 to 4 has the maximum density



"TotalVisits" and "Page Views Per Visit" have a positive correlation of 0.56

Numerical Feature Analysis



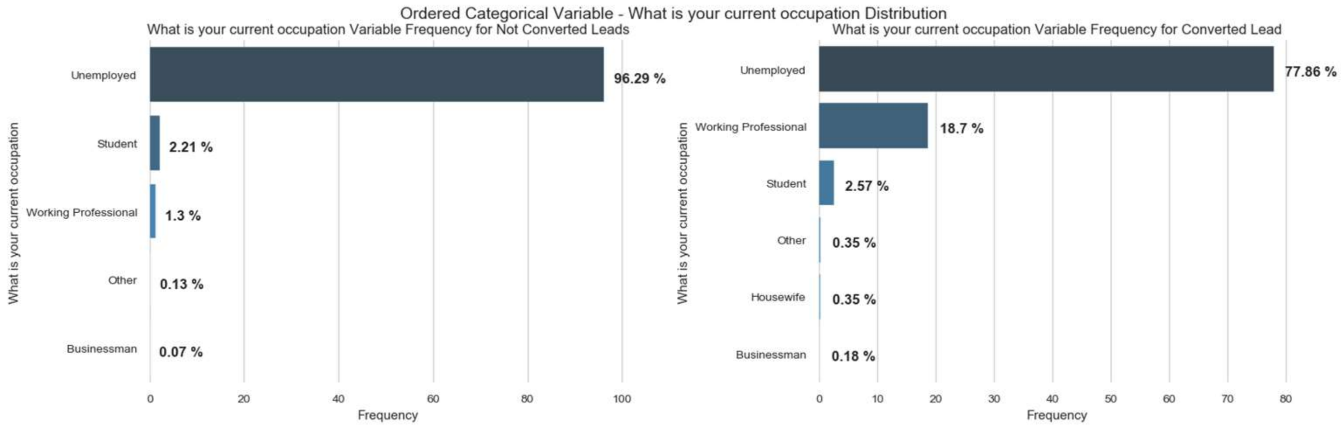
- If the Total Time Spent on Website is high then the conversion is also higher.
- If the Total Time Spent on Website is low then the conversion is very low.
- Clear visualization of correlation with conversion (y dependent variable)

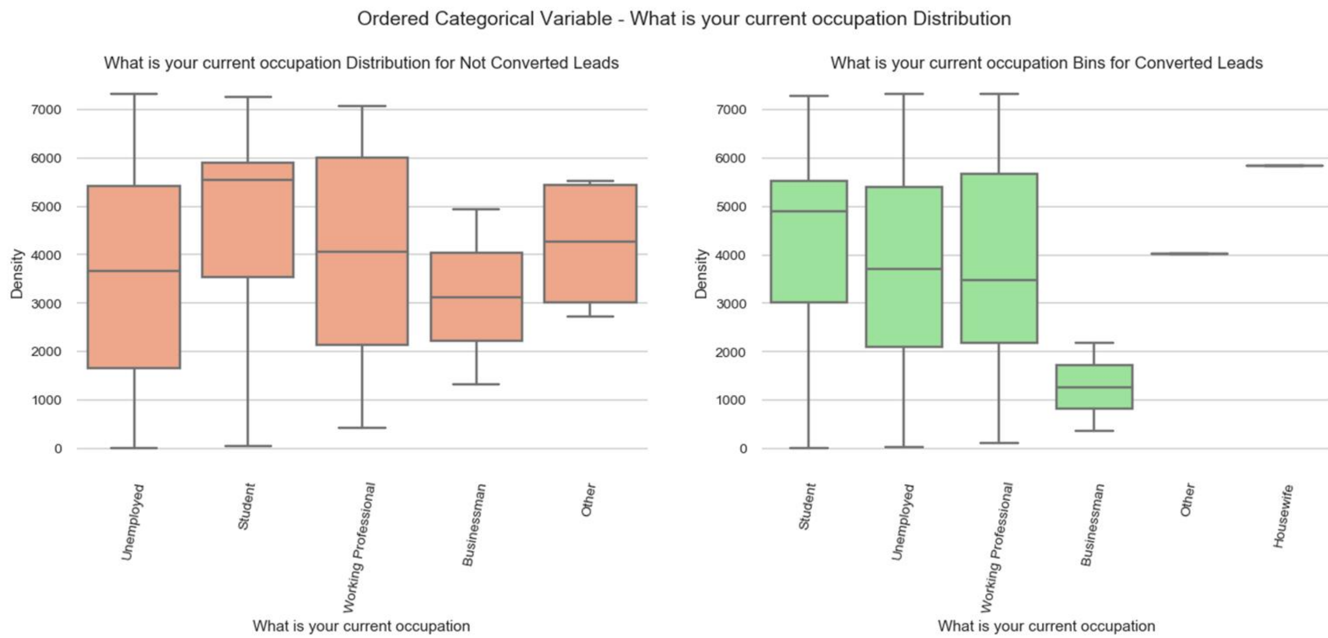
Similar Analysis done on all Numerical Variables – Not included in presentation

Categorical Feature Analysis

- Working Professional Occupation has a very high conversion. This value seems to have the highest positive correlation. This is followed by Unemployed.

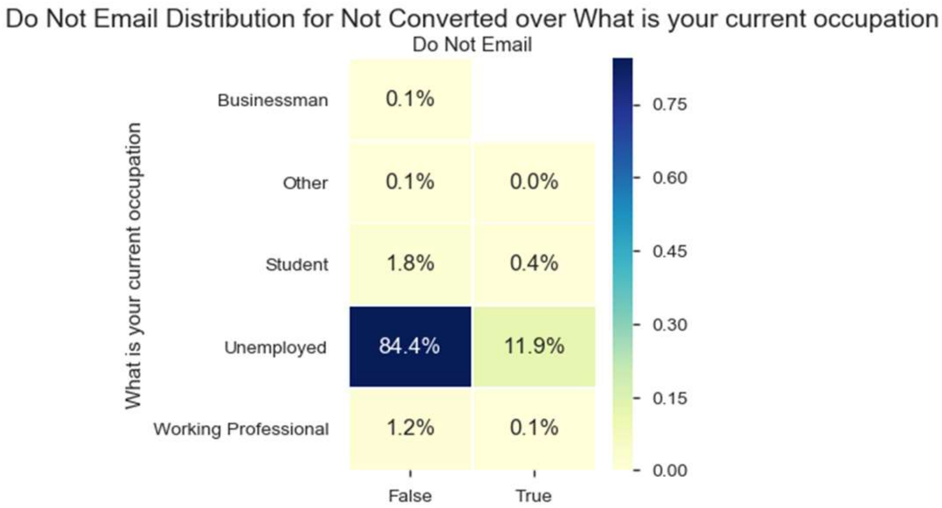
- Businessman occupation has very low conversion.

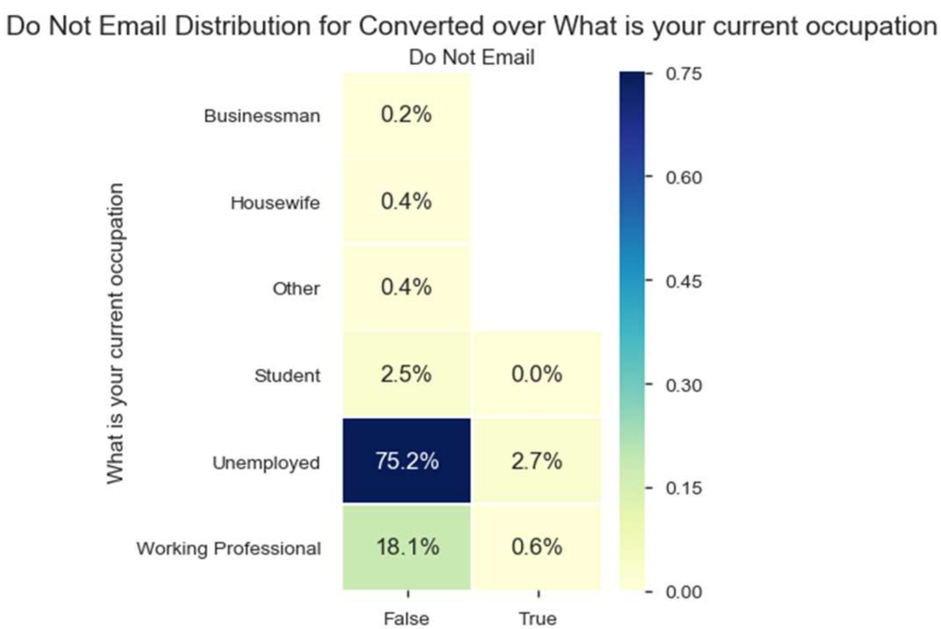


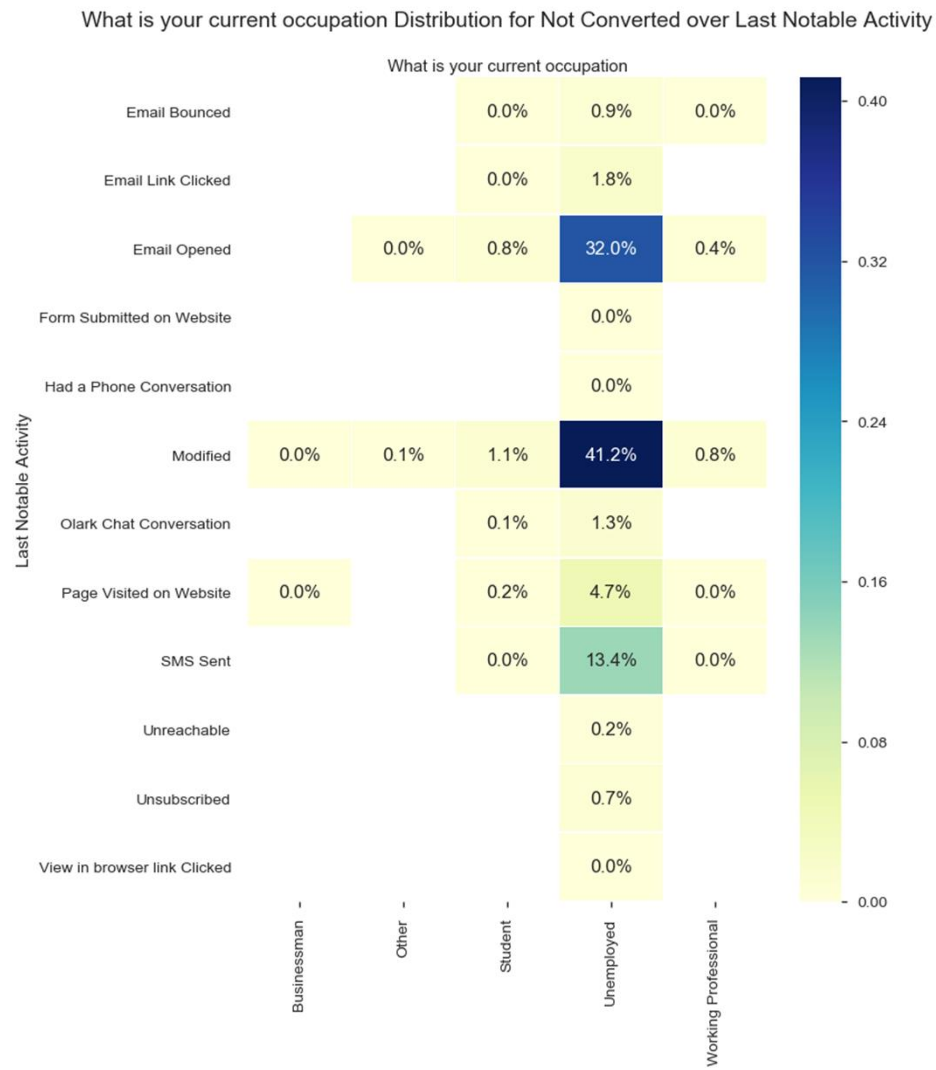


Similar Analysis done on all Categorical Features – Not included in presentation

Categorical Feature - Bivariate Analysis (Graphs)







- Feature Engineering
- Split into Test & Train
- Training the Model
- Model Evaluation

Feature Engineering

- Derived Metrics Possibilities
  - Combine the three numerical variables by multiplying them into a single derived column. This will yield the overall time spent online in one feature
  - The numerical features could even be binned into
  - Not creating any derived metrics as the above 2 points are not compelling enough and don't seem to add a lot of value.

- Dummy Encoding
  - Creating Dummy Variables for Categorical Variables
  - Removed redundant features having “unknown” value
  - Analyzed Correlation

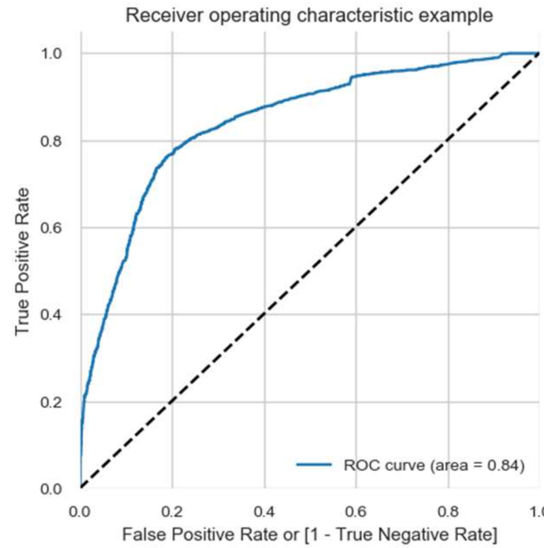
Featured Engineered DataFrame shape (9323, 85)

Training The Model

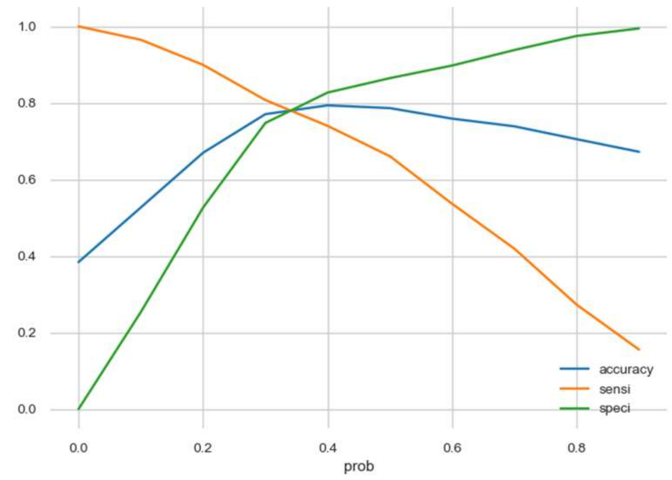
- Train DataFrame Shape: (5126, 84) Test DataFrame Shape: (2197, 84)
- Scaling Numerical Features using MinMaxScaler
- RFE (Recursive Feature Elimination)
  - Reduced Dimentionality by extracing ranked features
  - Reduced the feature list to 10 to model
- Modelling (using GLM & VIF)
  - Model with following features chosen
    - Total Time Spent on Website
    - Lead Origin_Lead Add Form
    - Do Not Email_True
    - Last Notable Activity_Modified
    - Last Notable Activity_Olark Chat Conversation
    - Last Notable Activity_Page Visited on Website
    - Current Occupation_Working Professional
  - Top 3 Features contributing most towards highest probabilities
    - Total Time Spent on Website
    - Lead Origin_Lead Add Form
    - Current Occupation_Working Professional

Model Evaluation

- Predicted values on Train dataset
- ROC Curve
  - The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.
  - The curve is good for our model
- Finding Optimal Cut-off Point
  - Cut Off Point 0.34
  - Confusion Matrix



Actual / Predicted	Not Converted	Converted
Not Converted	2505	652
Converted	438	1531



  - Metrics at ballpark 80%
    - Accuracy 79%
    - Sensitivity 77%
    - Specificity 79%

Evaluation on Test Data Set

Metrics on Test Data Set

Accuracy of 81%
Sensitivity of 79%
Specificity of 81%

Assigned Lead Score 0-100

	Lead Number	Converted	Converted_Prob	predicted	**Lead Score**
0	597640	True	0.783231	1	**78**
1	606086	True	0.892178	1	**89**
2	641652	True	0.716747	1	**72**
3	609351	False	0.104436	0	**10**
4	607845	False	0.177670	0	**18**