Machine Learning1

Lead Management- Group Assignment

IIITB Data Science Course 2

By: Hari Thapliyal & Papu Rai

Problem Statement

X Education sells professional courses. Lead Database is built using various marketing initiatives. Current conversion rate is 30%. We need to build a model using ML algorithms which can predict the conversion of a given lead.

We need to assign Lead Score to each lead.

Identify a probability cut-off level which will give accuracy of 80%.

All those leads which has probability above than this cut-off would be treated as hot-lead.

Prepare an approach which will optimize the time of sales team.

Analysis Approach

1. Duplicate Information Field Treatment
   - Identify fields which contains same information
   - Keep single field if any two fields has same information
2. Impute null values of numeric fields
3. Remove fields, with high % of null values and cannot be imputed
4. Remove Fields, which has 99+ % same information in the field.
5. Replace "Select" with np.nan wherever needed
6. Null Value Treatment for Categorical Fields (Detail in next page)
7. Visualize the distribution of data for categorical fields
8. Visualize the distribution of data for numerical fields

Analysis Approach Cont...

9. Visualize the distribution of data of categorical fields pre and post imputation
10. Prediction Using PCA
11. Prediction Using RFE
12. Prediction Using Statsmodel
13. Model Evaluation Using ROC
14. Choosing a Model
15. Metrics Using Selected Model
16. Identify Important Fields using VIF
17. Prepare a List of Leads along with Lead Score

Impute Categorical Variable

- Impute with with modes: For smaller null values percentage fields
  For every catergorical columns with less % of null values
  - Identify mode
  - Replace null values with mode
  - Perform numeric encoding
  - Create dummy fields
  - Merge all dummy fields to main dataset and remove corresponding original fields
- Impute using Logistic Regression method: For high null values percentage fields
  For every catergorical columns with high % of null values
  - Scale numberic fileds
  - For every categorical columns which has null values
    - Perform numeric encoding for the given columns, null value is assigned 0
    - Perform logistic regression using for non-null values
    - Predict categories for null categories using logistic regression

Results - In Business Terms

5 Important Fields from Dataset Which are important for Prediction are

1. Lead Profile
2. Lead Quality
3. Last Notable Activity
4. Occupation
5. Tags

VIF of Selected Variables

Dummy Variable	VIF
Tags_6	1.03
LeadPro_3	1.02
Tags_7	1.02
Tags_4	1.01
Tags_8	1
Tags_25	1
LeadQ_5	0.68
Occu_3	0.57
LeadPro_5	0.4
NotableAct_9	0.23
Tags_3	0.14
Tags_9	0.08
Tags_12	0.02

Prediction Result on Test Data

30% of the Given data is Test Data

Cut off: 0.65 %

Accuracy: 0.91

Recall: 0.92

Precision: 0.86

Specificity: 0.90

Error Rate: 0.09

FPR 0.10

FNR: 0.08

Confusion Metrics Test Data

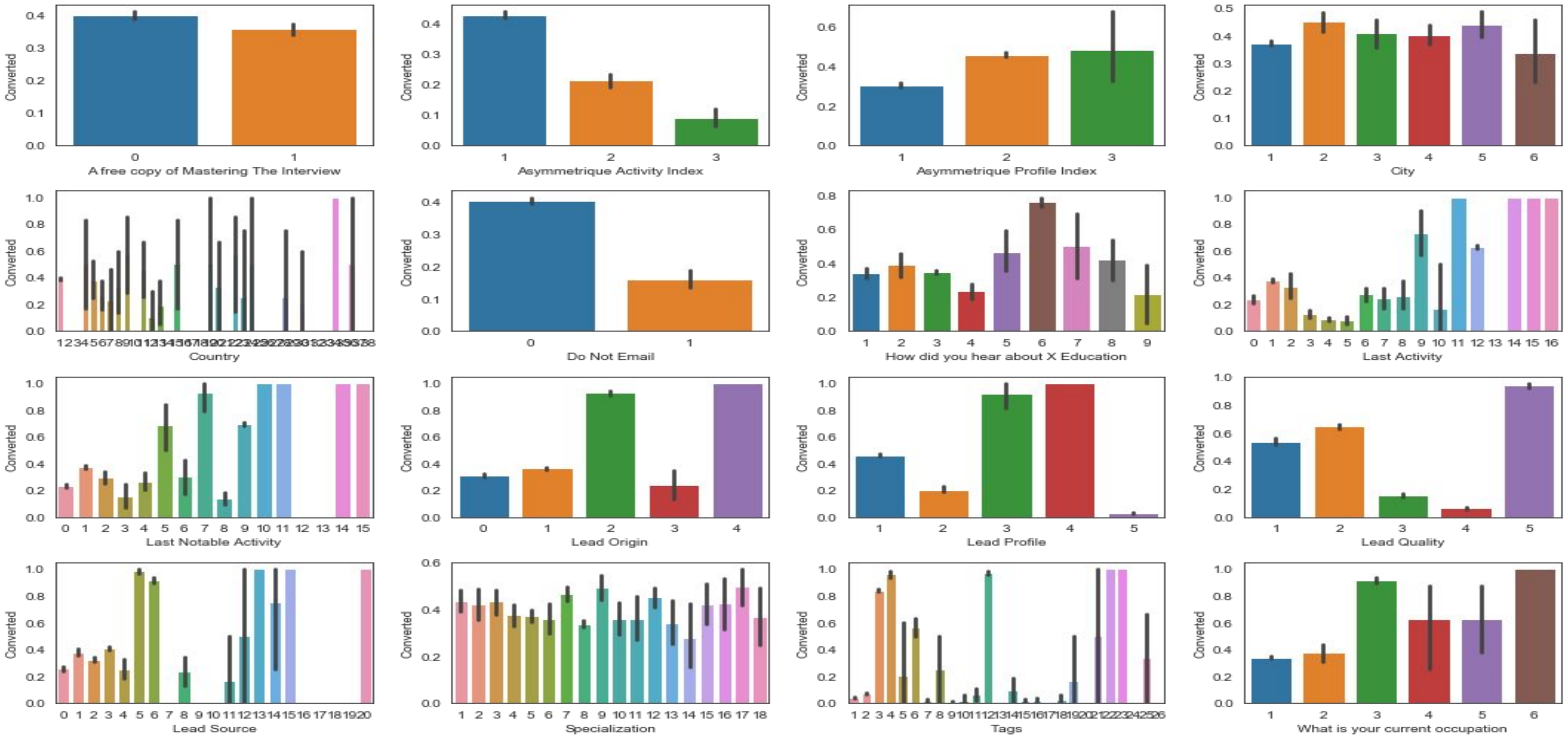
TN FP

FN TP

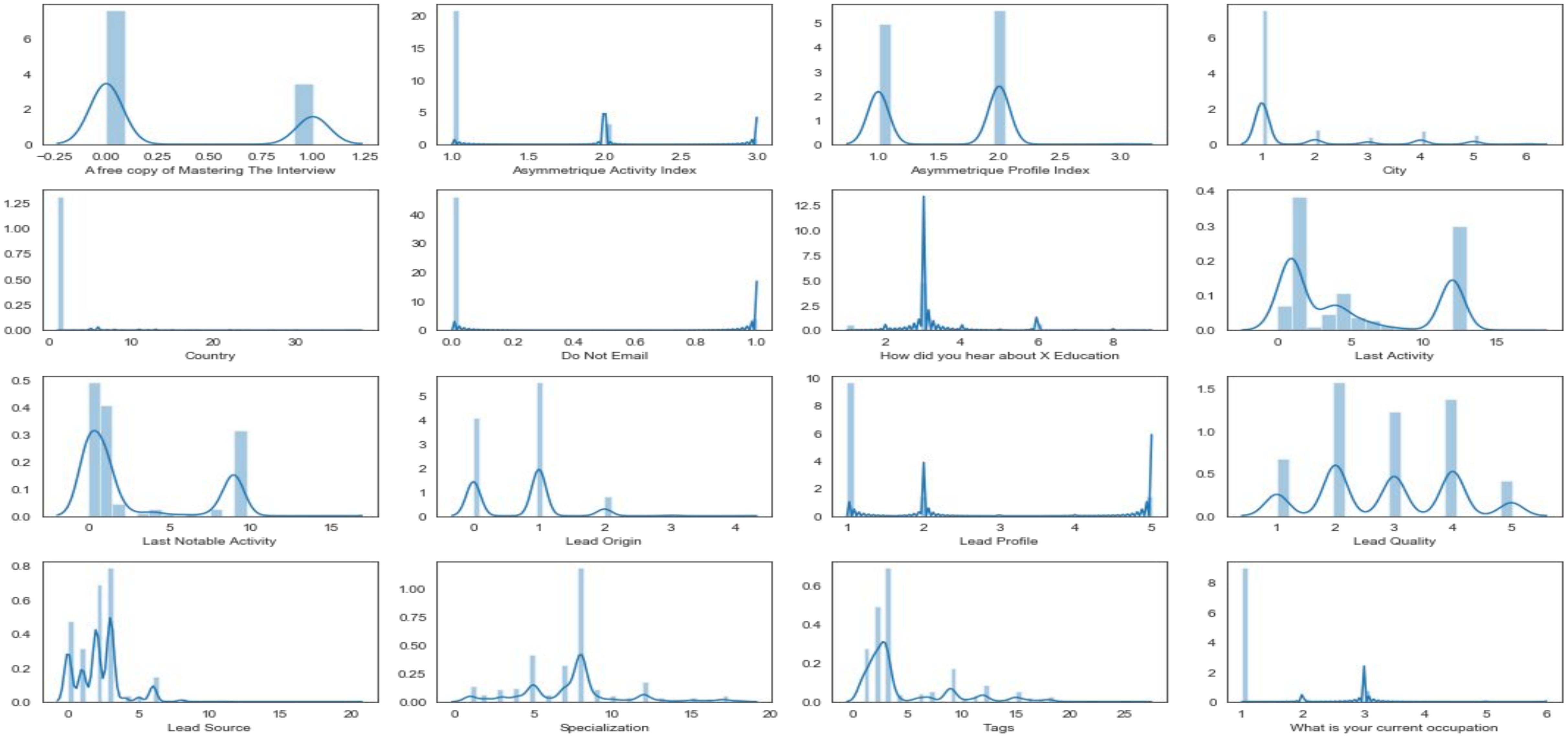
```
[1533  162]
 [ 85  992]
```

Visualizations of the most Important Results

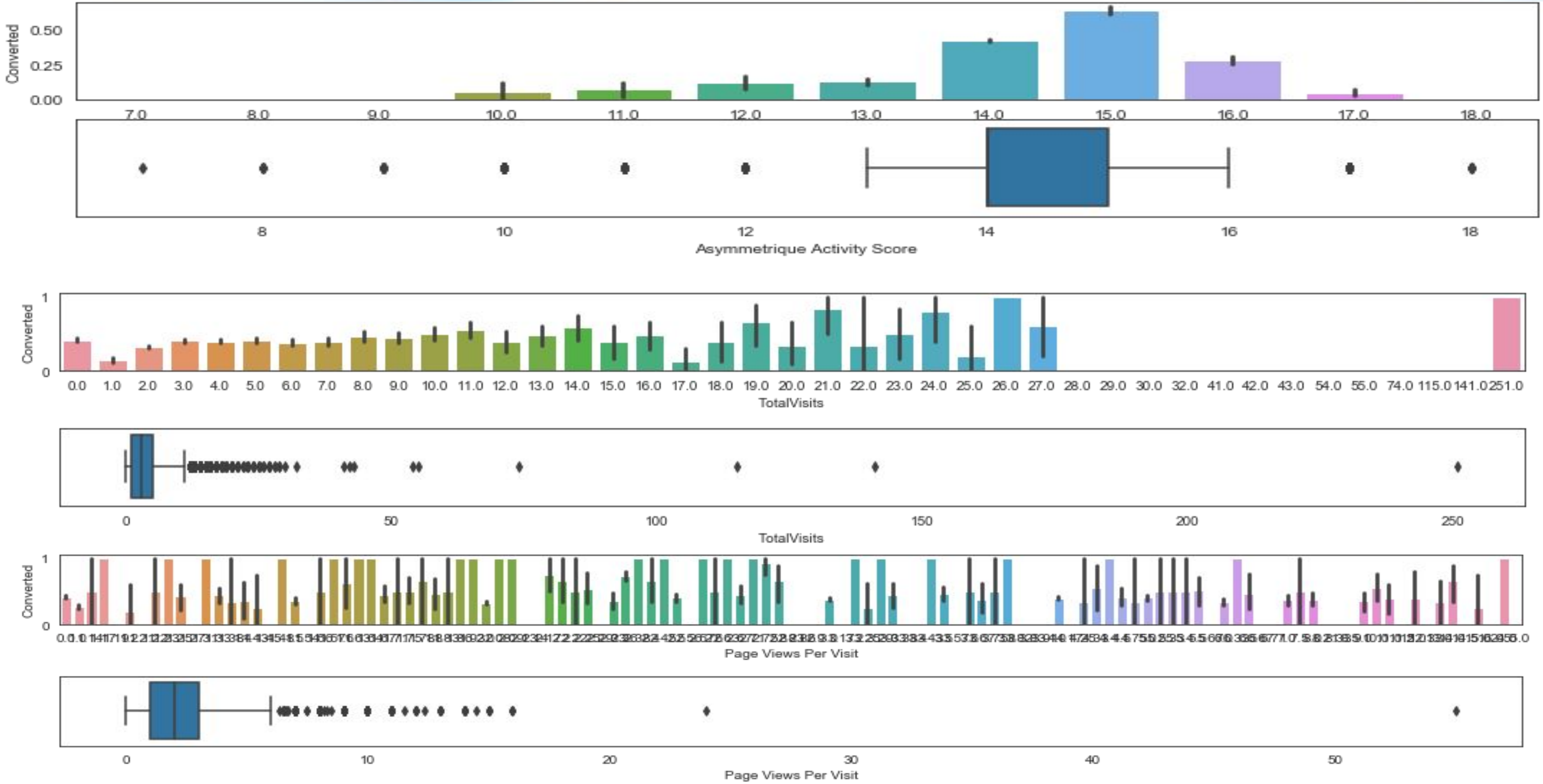
Distribution of Categorical Variables



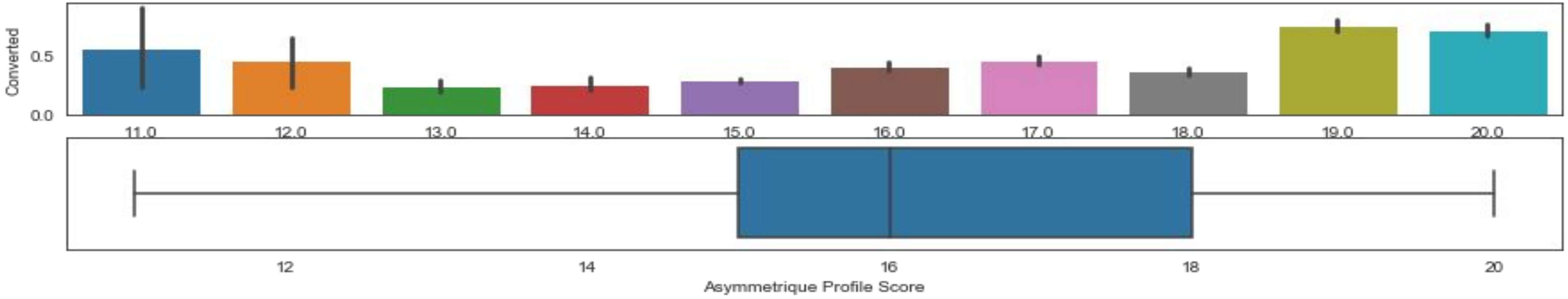
Distribution of Categorical Variables

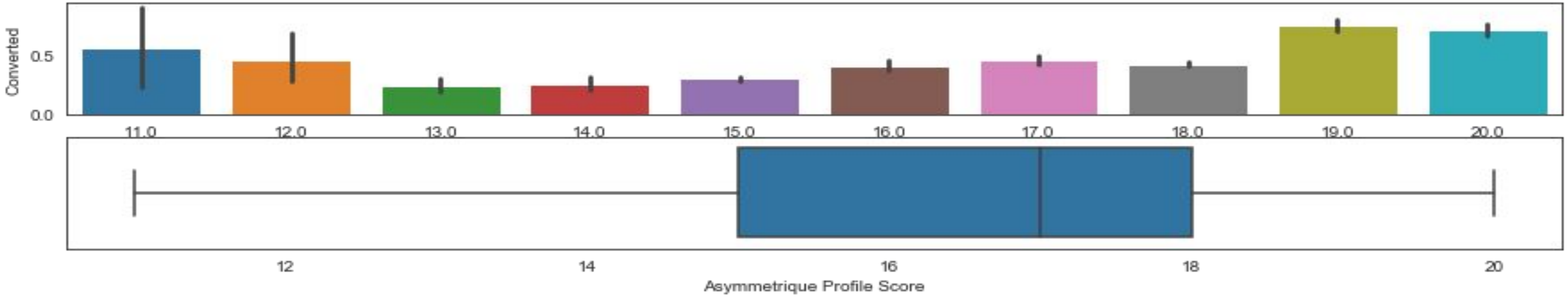


Distribution of Important Numeric Variables

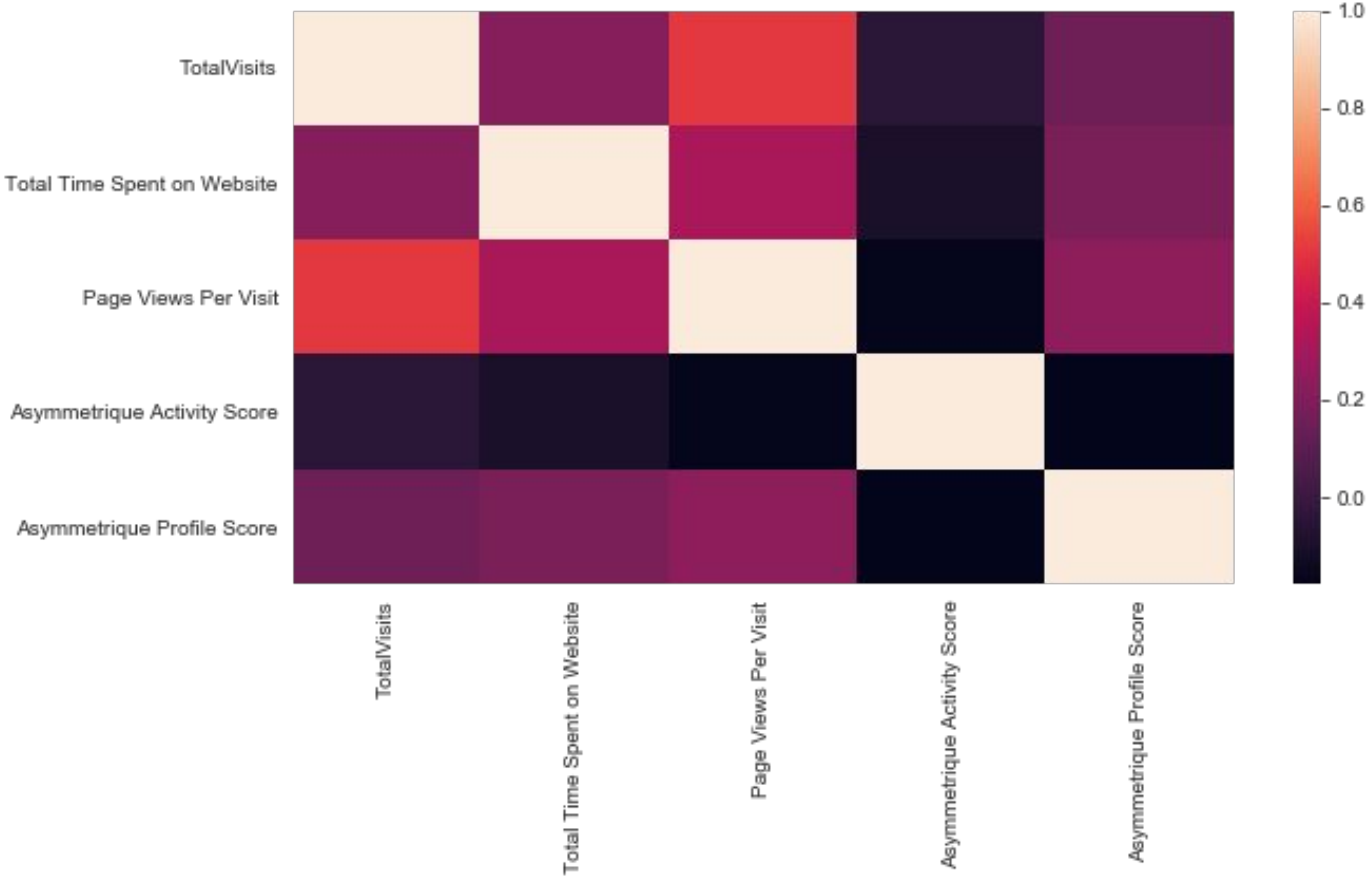


Distribution of Asymetrique Profile Score Before and After Null Value Treament

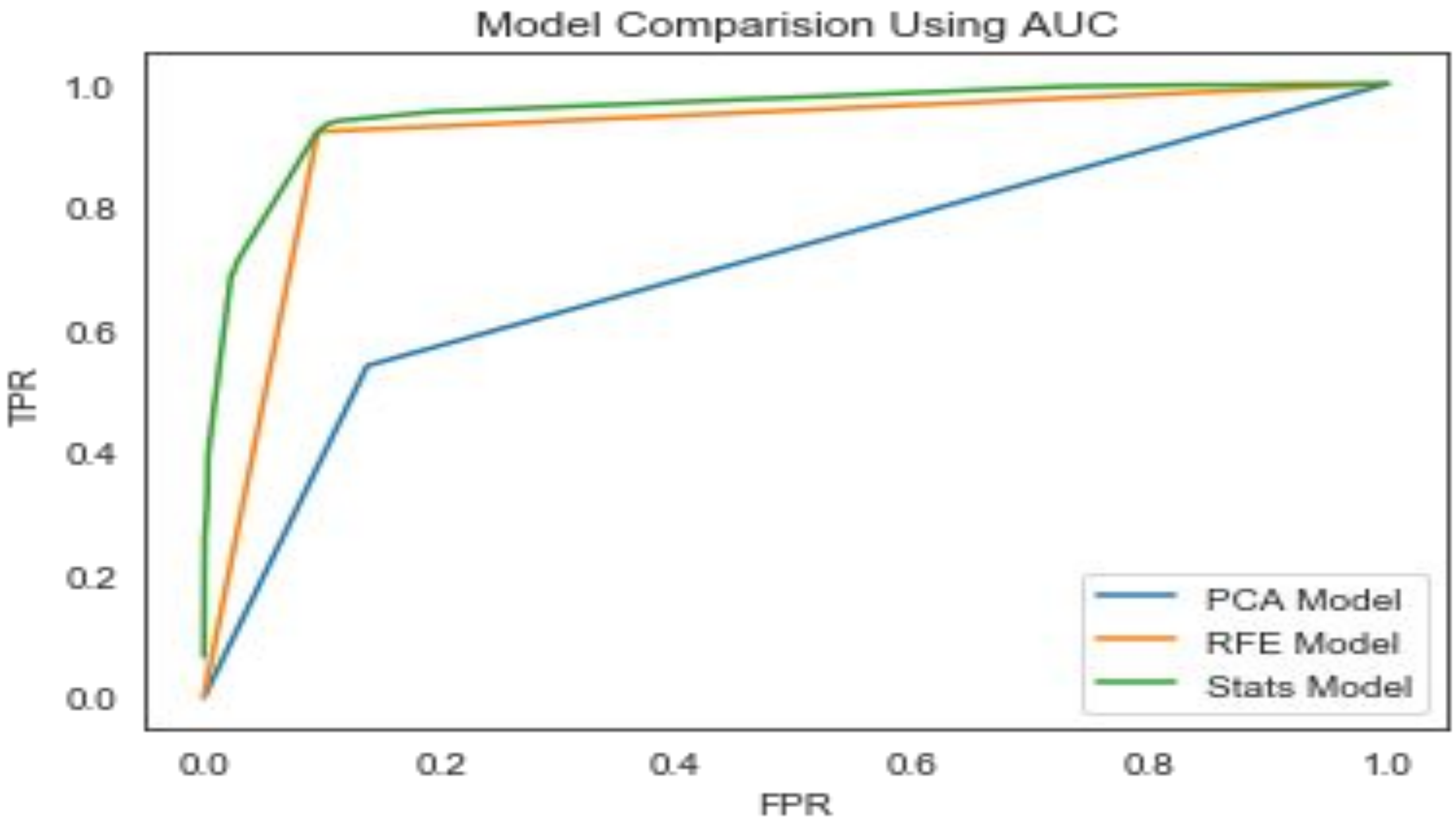




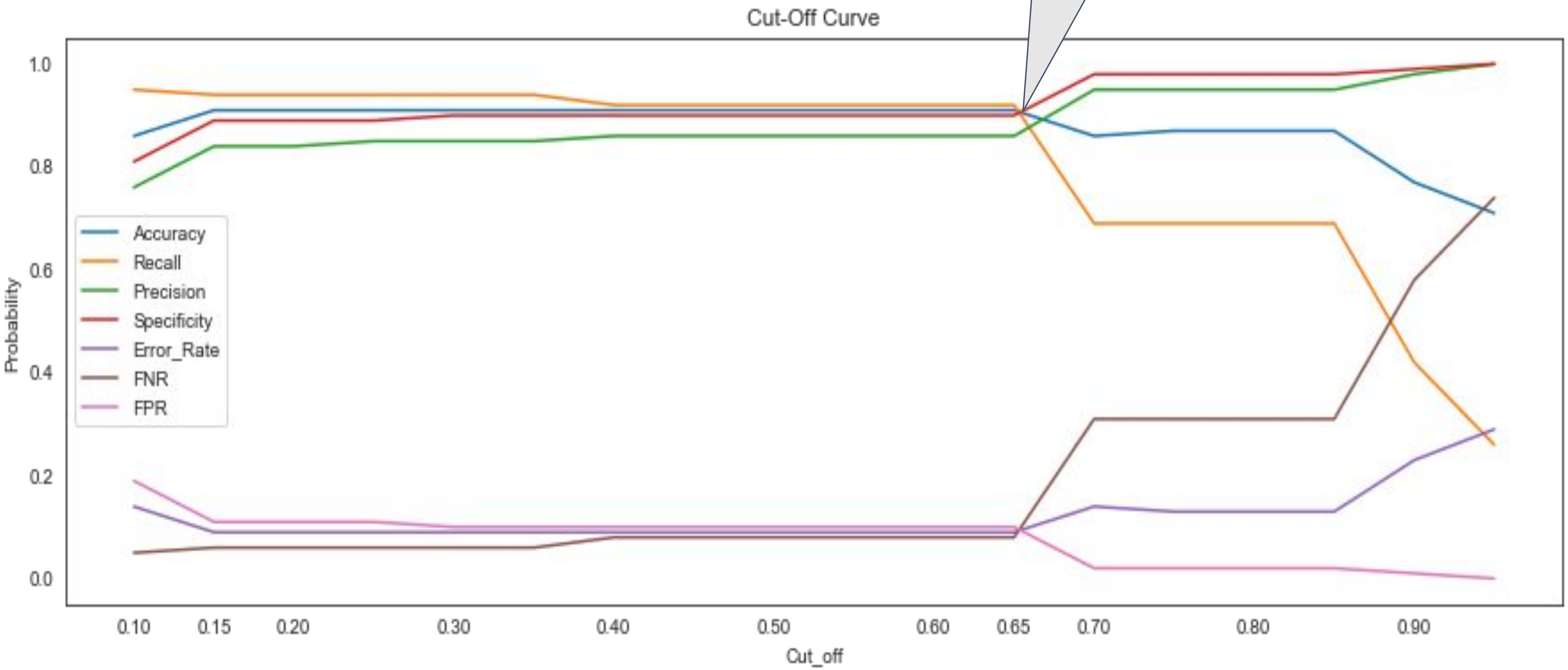
Relationship Between Numerical Variables



Model Evaluation

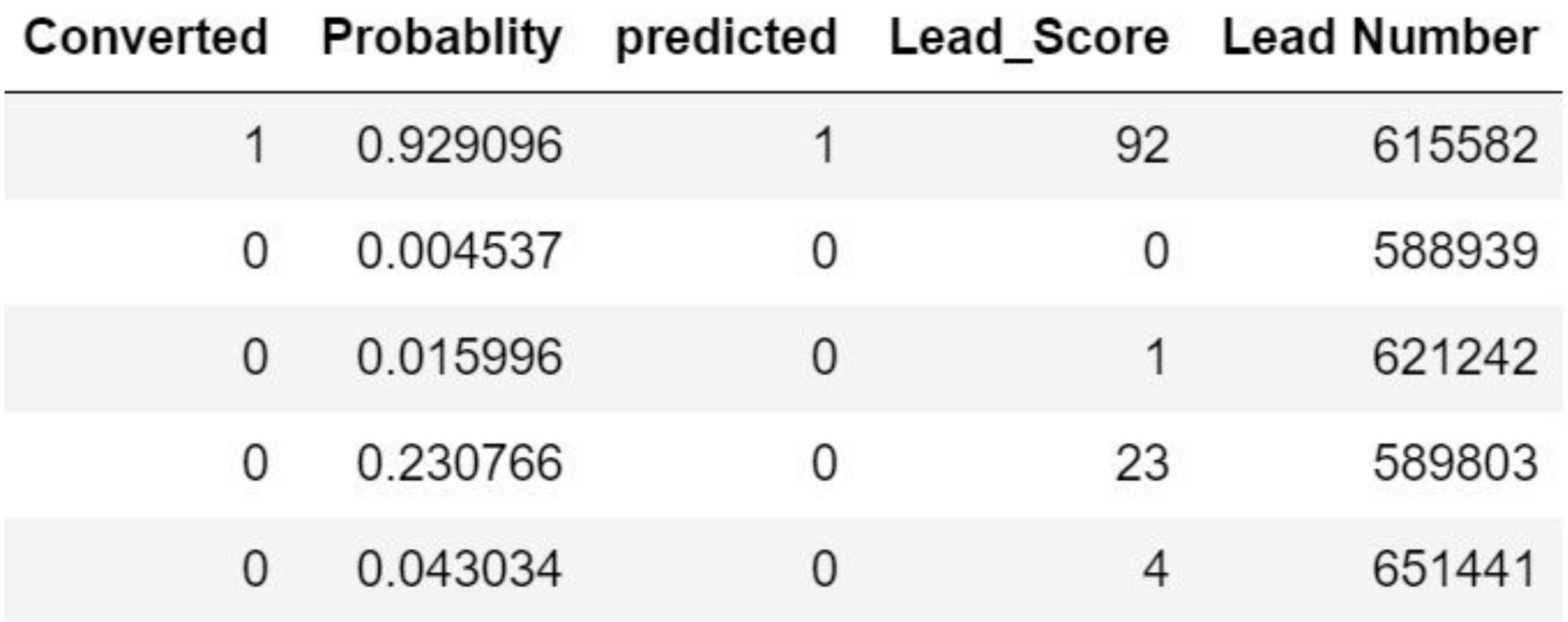


Cut-Off Curve



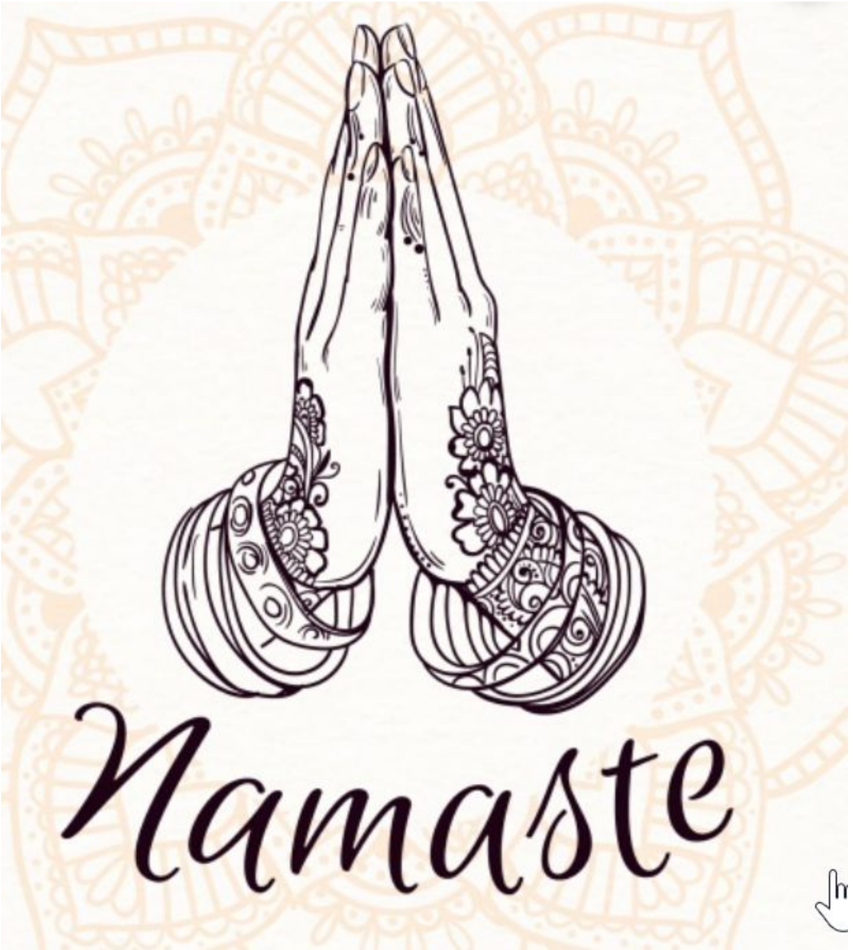
Cut-off at 65% is ideal

Header Rows from Output



Converted	Probablity	predicted	Lead_Score	Lead Number
1	0.929096	1	92	615582
0	0.004537	0	0	588939
0	0.015996	0	1	621242
0	0.230766	0	23	589803
0	0.043034	0	4	651441

Thank you


Namaste