ThoughtWorks Assignment

Q1- Descriptive analysis and insight with visualization is executed in the python notebook

Q2- As there are many outliers, It will good to develop tree based algorithm and then go for ensemble of some decision tree.

Q3. Built the model from scratch (optimization is required) and I achieved the following results from Random Forest models:

Mean Squared Error 2486021.574376074

ROOT Mean Squared Error 1576.7122674654606

R Squared 0.7239147595876901

I did the feature selection through Forward Selection and backward Selection and below are the top 15 features:

['annual\_income',

'house\_area',

'age',

'primary\_business',

'monthly\_expenses',

'social\_class',

'water\_availabity\_1.0',

'occupants\_count',

'young\_dependents',

'loan\_purpose\_education loan',

'secondary\_business\_none',

'loan\_tenure',

'sanitary\_availability\_1.0',

'sanitary\_availability\_0.0']

Q4. Yes, loan\_purpose is good predictor, but I combined multiple classes and then utilized the variable more accurately.

Q5. Fitness of model can be judged by following metrics:

MSE

RMSE

RSQUARED