Household income prediction

based on the attributes of families

組員: 107590016 楊宗洛
107590020 朱柏霖
107590023 李承紘
107590037 應耀德

2020 / 12 / 28

CONTENT

- About the data
- Preprocessing
- Baseline (Logistic Regression)
- Models (Decision Tree, Random Forest, MDC, KNN, SVM, XGBoost, SOM + XGBoost)
- Evaluation
- Conclusion

ABOUT DATA

Type of file: A csv file
Source: Kaggle
Data provider: The Philippine Statistics Authority
Size:

- 21.61MB
- 60 columns
- 41544 rows

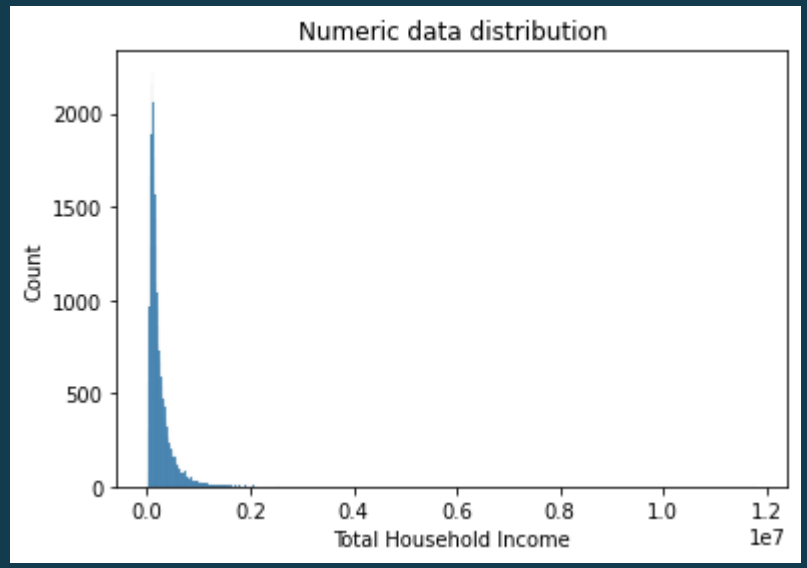
Each column represents:
one household attribute
Each row represents:
one household

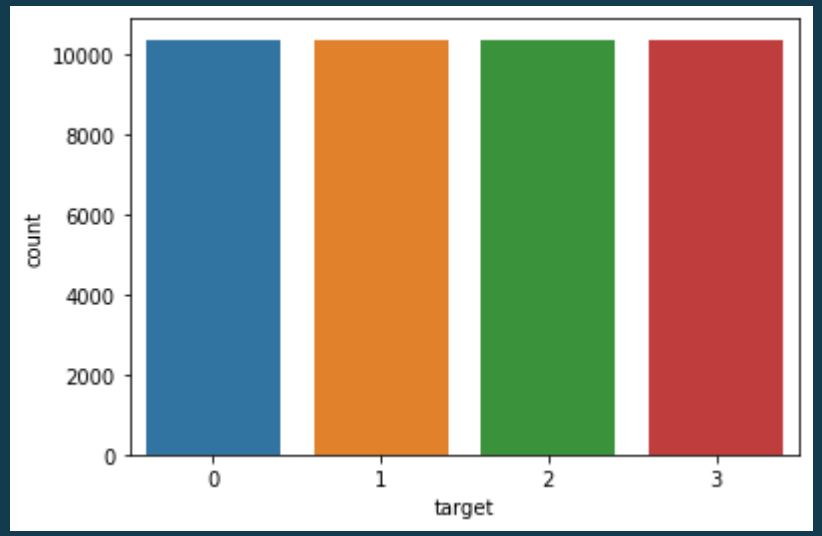
https://www.kaggle.com/grosvenpaul/family-income-and-expenditure

ABOUT DATA

Target: Split Total Household Income into 4 parts

1. Category 1
   (11284.999, 104895.0] - (< 25% of the distribution)
1. Category 2
   (104895.0, 164079.5] - (25%-50% of the distribution)
1. Category 3
   (164079.5, 291138.5] - (50%-75% of the distribution)
1. Category 4
   (291138.5, 11815988.0] - (> 75% of the distribution)

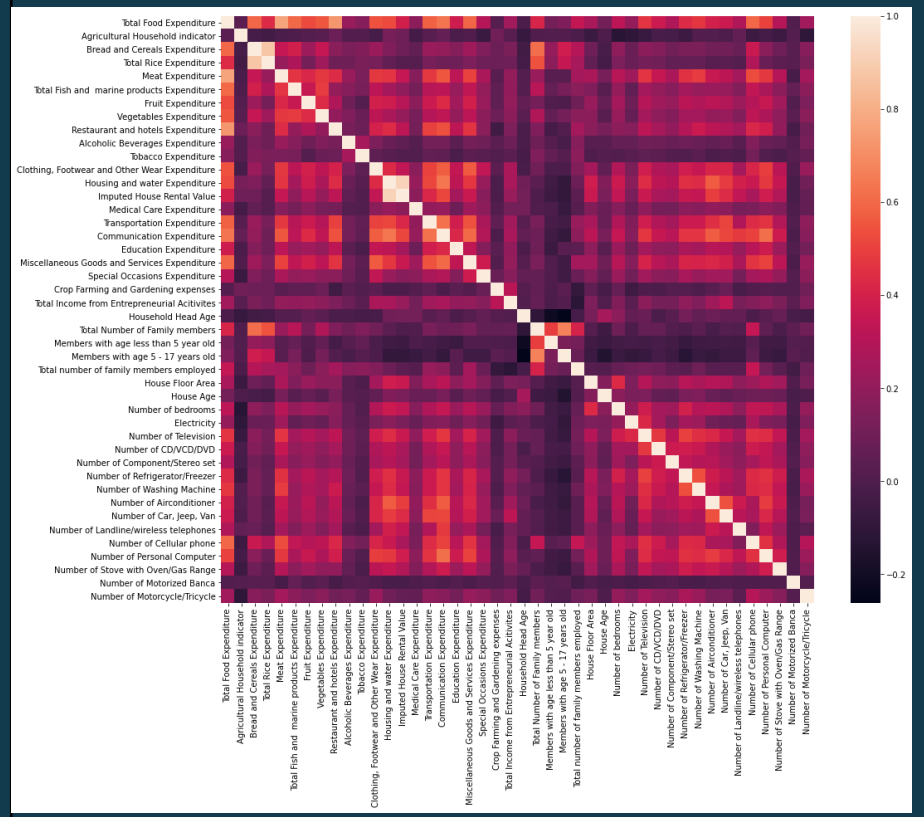




https://www.kaggle.com/grosvenpaul/family-income-and-expenditure

Features

- Continuous attributes
  - Total Household Income
  - Total Food Expenditure
  - Bread and Cereals Expenditure
  - Bread and Cereals Expenditure
  - Meat Expenditure
  - Total Fish and marine products Expenditure
  - Fruit Expenditure
  - Vegetables Expenditure
  - Restaurant and hotels Expenditure
  - Alcoholic Beverages Expenditure
  - Tobacco Expenditure
  - Clothing, Footwear and Other Wear Expenditure
- Housing and water Expenditure
- Imputed House Rental Value
- Medical Care Expenditure
- Transportation Expenditure
- Communication Expenditure
- Education Expenditure
- Miscellaneous Goods and Services Expenditure
- Special Occasions Expenditure
- Crop Farming and Gardening expenses
- Total Income from Entrepreneurial Activities
- House Floor Area



Features

- Discrete attributes
  - Agricultural Household indicator
  - Household Head Age
  - Total Number of Family members
  - Members with age less than 5 years old
  - Members with age 5 - 17 years old
  - Total number of family members employed
  - House Age
  - Number of bedrooms
  - Number of Television
  - Number of CD/VCD/DVD
  - Number of Component/Stereo set
  - Number of Refrigerator/Freezer
  - Number of Washing Machine
  - Number of Air conditioner
  - Number of Car, Jeep, Van
  - Number of Landline/wireless telephones
  - Number of Cellular phone
  - Number of Personal Computer
  - Number of Stove with Oven/Gas Range
  - Number of Motorized Banca
  - Number of Motorcycle/Tricycle
- Binary attributes
  - Household Head Job or Business Indicator
  - Electricity
- Nominal attributes
  - Main Source of Water Supply
  - Toilet Facilities
  - Tenure Status
  - Household Head Occupation
  - Household Head Marital Status
  - Region
  - Household Head Sex
  - Main Source of Income
  - Household Head Highest Grade Completed
  - Household Head Class of Worker
  - Type of Household
  - Type of Building/House
  - Type of Roof
  - Type of Walls

Preprocessing

- Data Cleaning
- Data Integration
  - Label-Encoder - Target
  - One-Hot Encoder
- Split data into train, validation, test
- Data Normalization – MinMaxScaler
- Feature selection

Preprocessing

Data Cleaning – Null Value & Fill N/A

```
Household Head Sex                          0
Household Head Age                          0
Household Head Marital Status               0
Household Head Highest Grade Completed      0
Household Head Job or Business Indicator    0
Household Head Occupation                7536
Household Head Class of Worker           7536
```

✓

```
dataset['Household Head Class of Worker'].value_counts(dropna=False)
```

```
Self-employed wihout any employee                          13766
Worked for private establishment                           13731
NaN                                                         7536
Worked for government/government corporation                2820
Employer in own family-operated farm or business            2581
Worked for private household                                 811
Worked without pay in own family-operated farm or business   285
Worked with pay in own family-operated farm or business       14
Name: Household Head Class of Worker, dtype: int64
```

✗

```
dataset['Household Head Occupation'].value_counts(dropna=False)
```

```
NaN                                                                    7536
Farmhands and laborers                                                 3478
Rice farmers                                                           2849
General managers/managing proprietors in wholesale and retail trade    2028
General managers/managing proprietors in transportation, storage and communications    1932
                                                                       ...
Hunters and trappers                                                      1
Musical instrument makers and tuners                                      1
Metal drawers and extruders                                               1
Personal care and related workers, n. e. c.                               1
Chemical processing plant operators n. e. c.                              1
Name: Household Head Occupation, Length: 379, dtype: int64
```

Preprocessing

Data Integration
Categorical attributes: 15
Numeric attributes: 43

Label Encoding - Target

Target: Split Total Household Income into 4 parts

1. Category 1 #0
   (11284.999, 104895.0] - (< 25% of the distribution)
1. Category 2 #1
   (104895.0, 164079.5] - (25%-50% of the distribution)
1. Category 3 #2
   (164079.5, 291138.5] - (50%-75% of the distribution)
1. Category 4 #3
   (291138.5, 11815988.0] - (> 75% of the distribution)

One Hot Encoding – Categorical attributes
(15 → 134)

- Main Source of Water Supply
- Toilet Facilities
- Tenure Status
- Household Head Occupation
- Household Head Marital Status
- Region
- Household Head Sex
- Main Source of Income
- Household Head Highest Grade Completed
- Household Head Class of Worker
- Type of Household
- Type of Building/House
- Type of Roof
- Type of Walls

Preprocessing

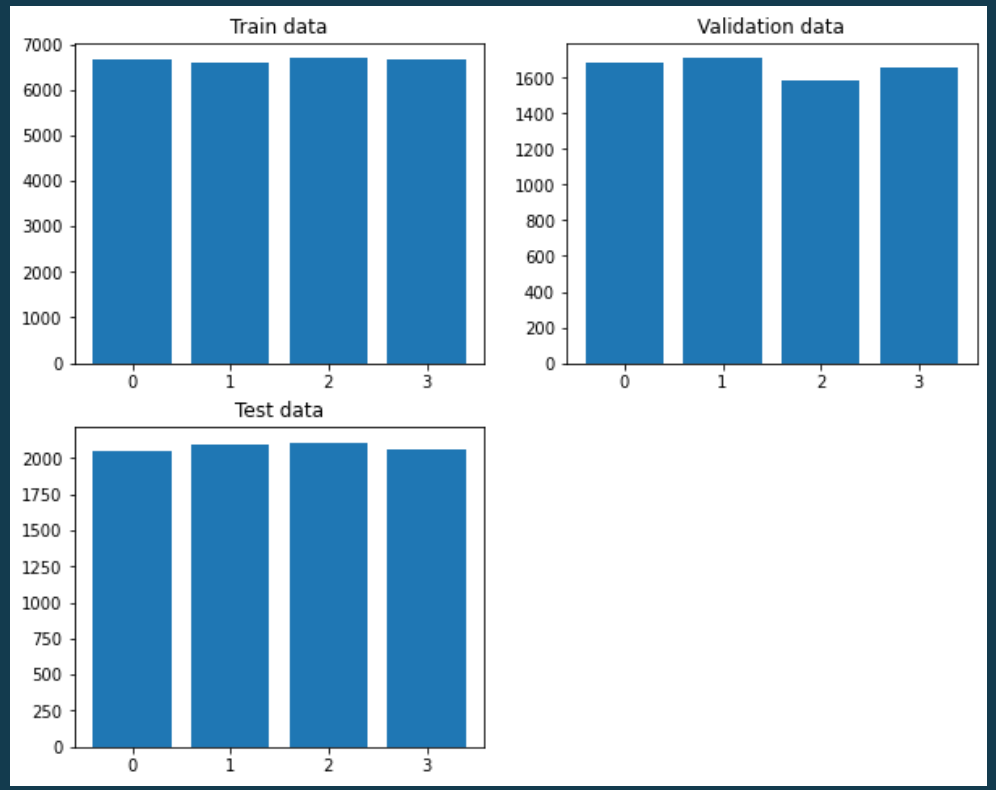
Split data into train, validation, test subset

- Train: 64%
- Validation: 16%
- Test: 20%

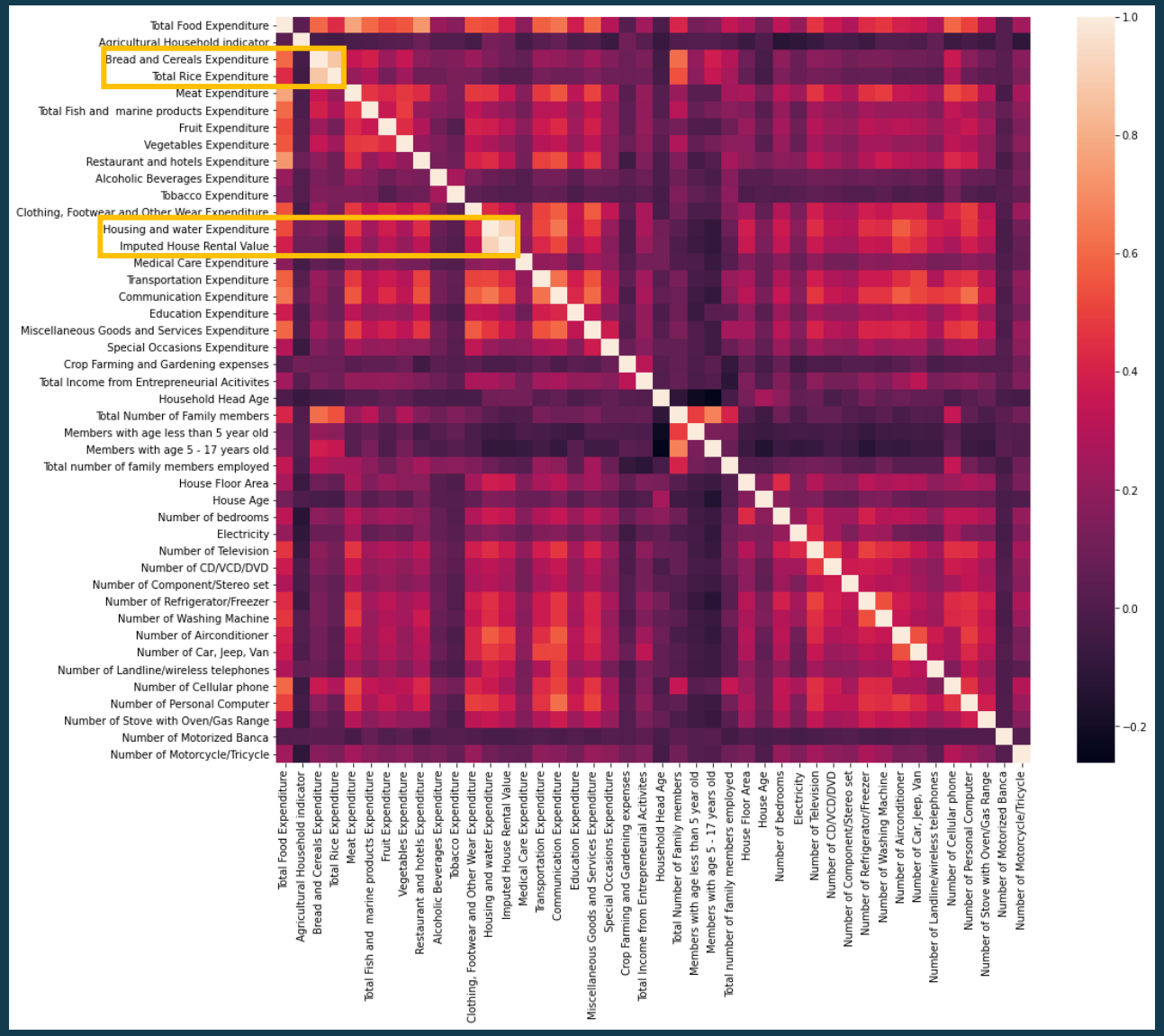
Data Normalization – MinMaxScaler

X_std = (X - X.min(axis=0)) / (X.max(axis=0) - X.min(axis=0))

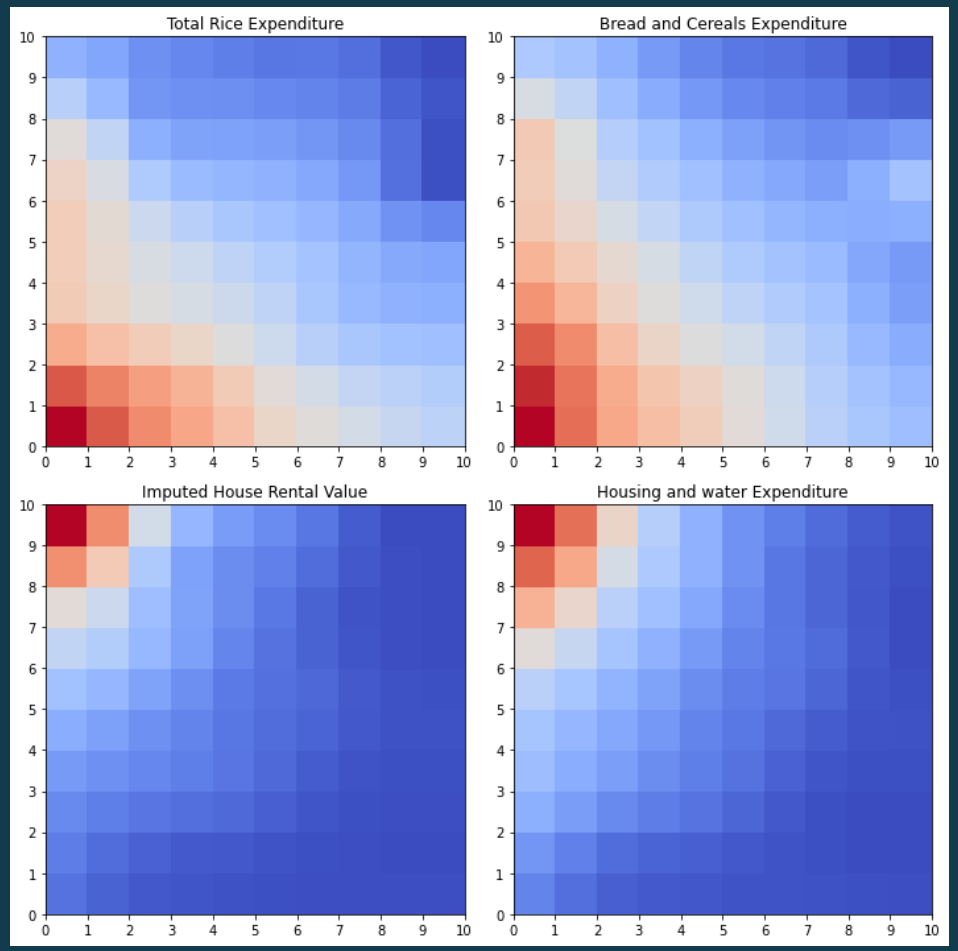
X_scaled = X_std * (max - min) + min



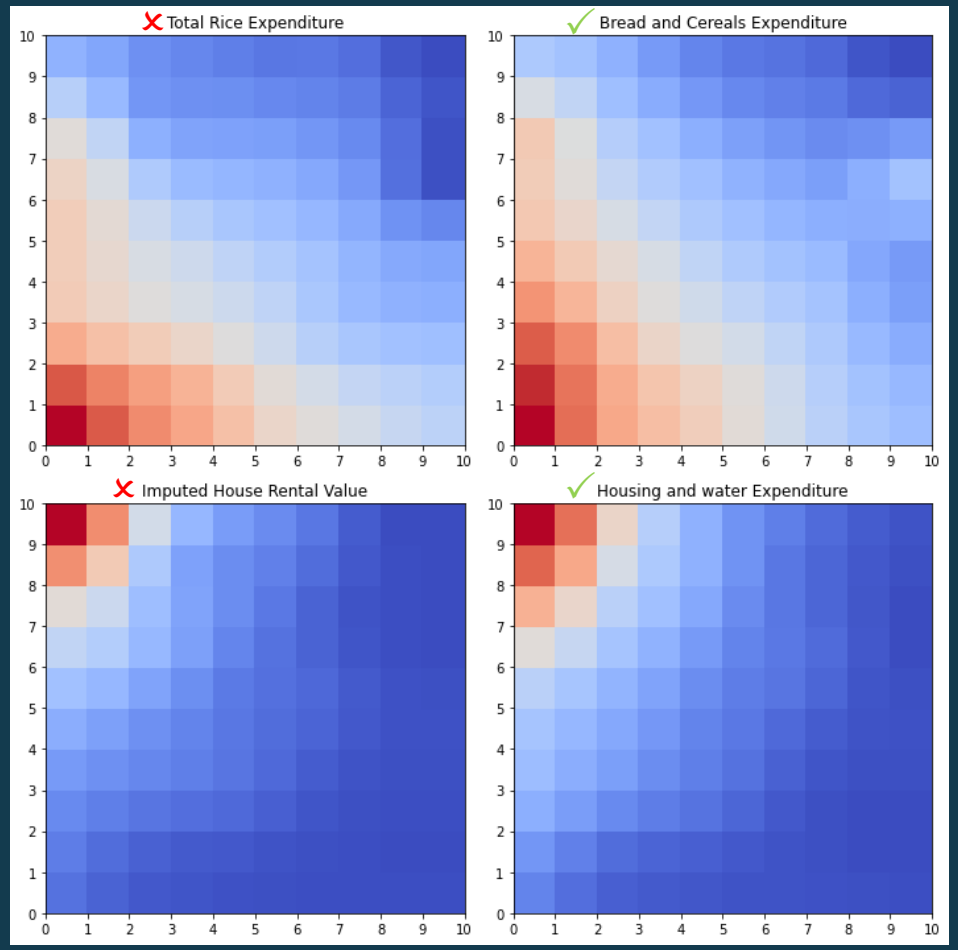
Feature Selection

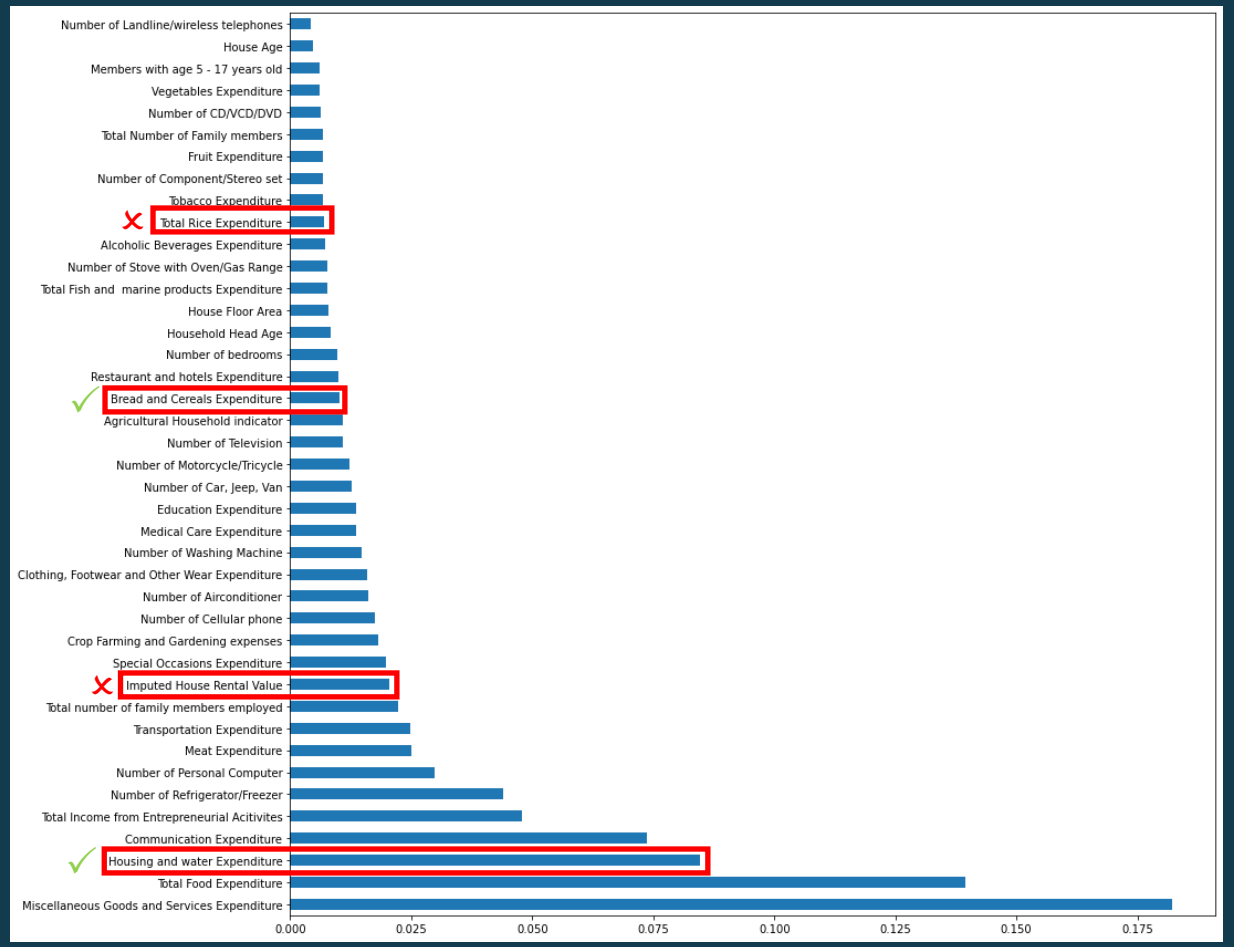


Visual-based feature selection



Feature Selection





(# of features: 177 – 2 = 175)

Feature Selection

Boruta Algorithm – based on Random Forest

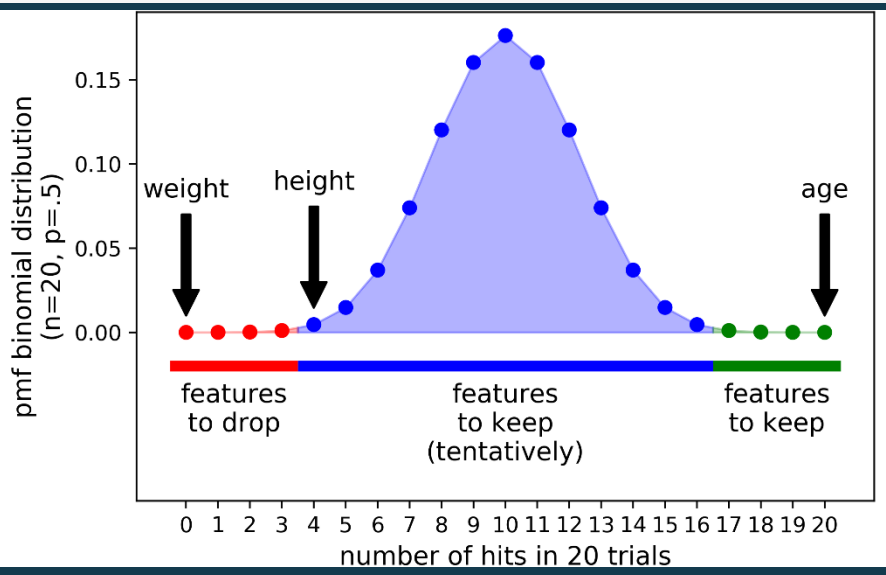
1. Shadow data

	age	height	weight	shadow_age	shadow_height	shadow_weight
0	25	182	75	51	176	75
1	32	176	71	32	182	71
2	47	174	78	47	168	78
3	51	168	72	25	181	72
4	62	181	86	62	174	86

	age	height	weight	shadow_age	shadow_height	shadow_weight
feature importance %	39	19	8	11	14	9
hits	1	1	0	-	-	-

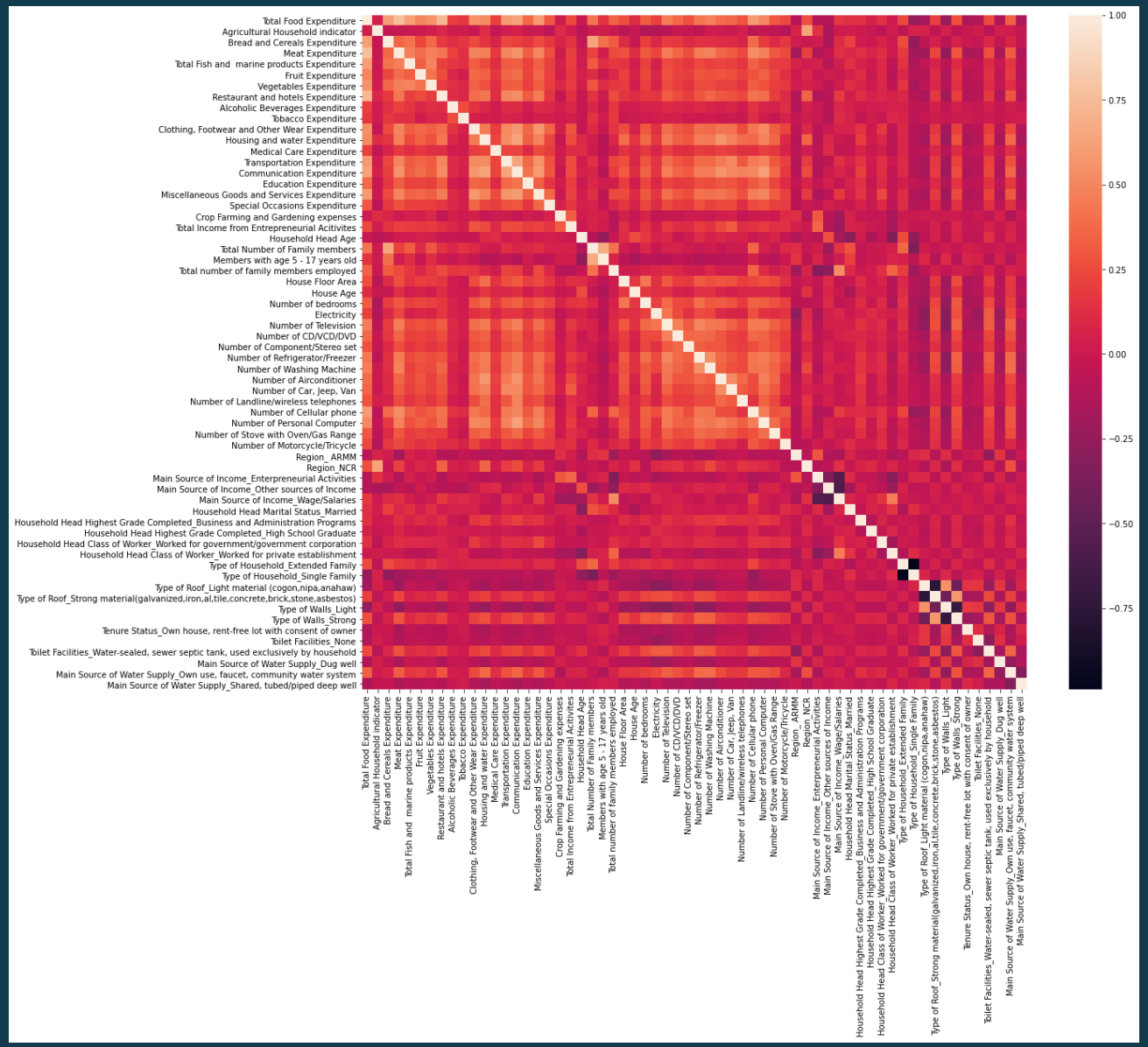
2. Binomial distribution

	age	height	weight
hits (in 20 trials)	20	4	0



Paper:
https://www.researchgate.net/publication/280138095_Feature_Selection_with_Boruta_Package

Feature Selection



# of features: 175 → 62 (max_depth=5)

Selected Features

- Total Food Expenditure
- Agricultural Household indicator
- Bread and Cereals Expenditure
- Meat Expenditure
- Total Fish and marine products Expenditure
- Fruit Expenditure
- Vegetables Expenditure
- Restaurant and hotels Expenditure
- Alcoholic Beverages Expenditure
- Tobacco Expenditure
- Clothing, Footwear and Other Wear Expenditure
- Housing and water Expenditure
- Medical Care Expenditure
- Transportation Expenditure
- Communication Expenditure
- Education Expenditure
- Miscellaneous Goods and Services Expenditure
- Special Occasions Expenditure
- Crop Farming and Gardening expenses
- Total Income from Entrepreneurial Acitivites
- Household Head Age
- Total Number of Family members
- Members with age 5 - 17 years old
- Total number of family members employed
- House Floor Area
- House Age
- Number of bedrooms
- Electricity
- Number of Television
- Number of CD/VCD/DVD
- Number of Component/Stereo set
- Number of Refrigerator/Freezer
- Number of Washing Machine
- Number of Airconditioner
- Number of Car, Jeep, Van
- Number of Landline/wireless telephones
- Number of Cellular phone
- Number of Personal Computer
- Number of Stove with Oven/Gas Range
- Number of Motorcycle/Tricycle
- Region_ ARMM
- Region_NCR
- Main Source of Income_Enterpreneurial Activities
- Main Source of Income_Other sources of Income
- Main Source of Income_Wage/Salaries
- Household Head Marital Status_Married
- Household Head Highest Grade Completed_Business and Administration Programs
- Household Head Highest Grade Completed_High School Graduate
- Household Head Class of Worker_Worked for government/government corporation
- Household Head Class of Worker_Worked for private establishment
- Type of Household_Extended Family
- Type of Household_Single Family
- Type of Roof_Light material (cogon,nipa,anahaw)
- Type of Roof_Strong material(galvanized,iron,al,tile,concrete,brick,stone,asbestos)
- Type of Walls_Light
- Type of Walls_Strong
- Tenure Status_Own house, rent-free lot with consent of owner
- Toilet Facilities_None
- Toilet Facilities_Water-sealed, sewer septic tank, used exclusively by household
- Main Source of Water Supply_Dug well
- Main Source of Water Supply_Own use, faucet, community water system
- Main Source of Water Supply_Shared, tubed/piped deep well

Baseline

LogisticRegression – All features
Train score: 0.7089288400782308
Cross-validation score: 0.6943352914502243

Validation score: 0.70091770723634 73

Test score:

	precision	recall	f1-score	support
0	0.74	0.80	0.77	2051
1	0.59	0.56	0.57	2090
2	0.62	0.63	0.63	2110
3	0.86	0.82	0.84	2058
accuracy			0.70	8309
macro avg	0.70	0.70	0.70	8309
weighted avg	0.70	0.70	0.70	8309

LogisticRegression – Selected features
Train score: 0.702572589137957
Cross-validation score: 0.6932449600198842

Validation score: 0.6982097186700768

Test score:

	precision	recall	f1-score	support
0	0.73	0.81	0.77	2051
1	0.58	0.55	0.57	2090
2	0.61	0.62	0.62	2110
3	0.86	0.81	0.84	2058
accuracy			0.70	8309
macro avg	0.70	0.70	0.70	8309
weighted avg	0.70	0.70	0.70	8309

KNN

Mix feature

	precision	recall	f1-score	support
0	0.52	0.64	0.58	2051
1	0.38	0.37	0.37	2090
2	0.38	0.37	0.38	2110
3	0.68	0.56	0.61	2058
accuracy			0.48	8309
macro avg	0.49	0.49	0.49	8309
weighted avg	0.49	0.48	0.48	8309

f_score: 0.483881 accuracy: 0.484294

==========

Only use numeric features

	precision	recall	f1-score	support
0	0.60	0.77	0.67	2051
1	0.42	0.43	0.43	2090
2	0.46	0.43	0.44	2110
3	0.83	0.61	0.70	2058
accuracy			0.56	8309
macro avg	0.57	0.56	0.56	8309
weighted avg	0.57	0.56	0.56	8309

f_score: 0.559320 accuracy: 0.559273

==========

Only use category features

	precision	recall	f1-score	support
0	0.49	0.56	0.52	2051
1	0.35	0.31	0.33	2090
2	0.33	0.33	0.33	2110
3	0.54	0.52	0.53	2058
accuracy			0.43	8309
macro avg	0.43	0.43	0.43	8309
weighted avg	0.43	0.43	0.43	8309

f_score: 0.426250 accuracy: 0.428451

==========

Selected mix feature

	precision	recall	f1-score	support
0	0.50	0.70	0.59	2051
1	0.39	0.37	0.38	2090
2	0.40	0.37	0.39	2110
3	0.76	0.54	0.63	2058
accuracy			0.50	8309
macro avg	0.51	0.50	0.50	8309
weighted avg	0.51	0.50	0.49	8309

f_score: 0.494422 accuracy: 0.495728

==========

Only use numeric features

	precision	recall	f1-score	support
0	0.60	0.77	0.67	2051
1	0.42	0.43	0.43	2090
2	0.46	0.43	0.44	2110
3	0.83	0.61	0.70	2058
accuracy			0.56	8309
macro avg	0.57	0.56	0.56	8309
weighted avg	0.57	0.56	0.56	8309

f_score: 0.559320 accuracy: 0.559273

==========

Only use category features

	precision	recall	f1-score	support
0	0.49	0.56	0.52	2051
1	0.35	0.31	0.33	2090
2	0.33	0.33	0.33	2110
3	0.54	0.52	0.53	2058
accuracy			0.43	8309
macro avg	0.43	0.43	0.43	8309
weighted avg	0.43	0.43	0.43	8309

f_score: 0.426250 accuracy: 0.428451

==========

MDC

Mix feature

	precision	recall	f1-score	support
0	0.56	0.55	0.56	2051
1	0.37	0.30	0.33	2090
2	0.38	0.35	0.37	2110
3	0.58	0.75	0.65	2058
accuracy			0.49	8309
macro avg	0.47	0.49	0.48	8309
weighted avg	0.47	0.49	0.48	8309

f_score: 0.475505 accuracy: 0.486340

Only use numeric features

	precision	recall	f1-score	support
0	0.57	0.66	0.62	2051
1	0.45	0.43	0.44	2090
2	0.49	0.46	0.48	2110
3				2058
accuracy				8309
macro avg	0.57	0.57	0.57	8309
weighted avg	0.57	0.57	0.57	8309

f_score: 0.571911 accuracy: 0.573113

Only use category features

	precision	recall	f1-score	support
0	0.54	0.54	0.54	2051
1	0.36	0.29	0.32	2090
2	0.36	0.31	0.33	2110
3	0.52	0.70	0.60	2058
accuracy			0.46	8309
macro avg	0.45	0.46	0.45	8309
weighted avg	0.45	0.46	0.45	8309

f_score: 0.447208 accuracy: 0.459502

Selected mix feature

	precision	recall	f1-score	support
0	0.55	0.57	0.56	2051
1	0.38	0.26	0.31	2090
2	0.37	0.36	0.36	2110
3	0.59	0.75	0.66	2058
accuracy			0.49	8309
macro avg	0.47	0.49	0.47	8309
weighted avg	0.47	0.49	0.47	8309

f_score: 0.471370 accuracy: 0.485257

Only use numeric features

	precision	recall	f1-score	support
0	0.57	0.66	0.62	2051
1	0.45	0.43	0.44	2090
2	0.49	0.46	0.48	2110
3				2058
accuracy				8309
macro avg	0.57	0.57	0.57	8309
weighted avg	0.57	0.57	0.57	8309

f_score: 0.571911 accuracy: 0.573113

Only use category features

	precision	recall	f1-score	support
0	0.54	0.54	0.54	2051
1	0.36	0.29	0.32	2090
2	0.36	0.31	0.33	2110
3	0.52	0.70	0.60	2058
accuracy			0.46	8309
macro avg	0.45	0.46	0.45	8309
weighted avg	0.45	0.46	0.45	8309

f_score: 0.447208 accuracy: 0.459502

Numeric features make more sense to the KNN and MDC model

DecisionTree& RandomForest

Mixed

	precision	recall	f1-score	support
0	0.78	0.78	0.78	2051
1	0.57	0.62	0.60	2090
2	0.61	0.60	0.61	2110
3	0.84	0.78	0.81	2058
accuracy			0.70	8309
macro avg	0.70	0.70	0.70	8309
weighted avg	0.70	0.70	0.70	8309

	precision	recall	f1-score	support
0	0.82	0.83	0.82	2051
1	0.64	0.68	0.66	2090
2	0.69	0.68	0.68	2110
3	0.88	0.84	0.86	2058
accuracy			0.76	8309
macro avg	0.76	0.76	0.76	8309
weighted avg	0.76	0.76	0.76	8309

Selected Mixed

	precision	recall	f1-score	support
0	0.78	0.80	0.79	2051
1	0.58	0.64	0.61	2090
2	0.65	0.60	0.62	2110
3	0.86	0.82	0.84	2058
accuracy			0.71	8309
macro avg	0.72	0.71	0.71	8309
weighted avg	0.72	0.71	0.71	8309

	precision	recall	f1-score	support
0	0.82	0.84	0.83	2051
1	0.65	0.68	0.66	2090
2	0.69	0.67	0.68	2110
3	0.88	0.84	0.86	2058
accuracy			0.76	8309
macro avg	0.76	0.76	0.76	8309
weighted avg	0.76	0.76	0.76	8309

DecisionTree& RandomForest

Numeric

	precision	recall	f1-score	support
0	0.78	0.79	0.78	2051
1	0.57	0.64	0.60	2090
2	0.62	0.60	0.61	2110
3	0.86	0.78	0.81	2058
accuracy			0.70	8309
macro avg	0.71	0.70	0.70	8309
weighted avg	0.71	0.70	0.70	8309

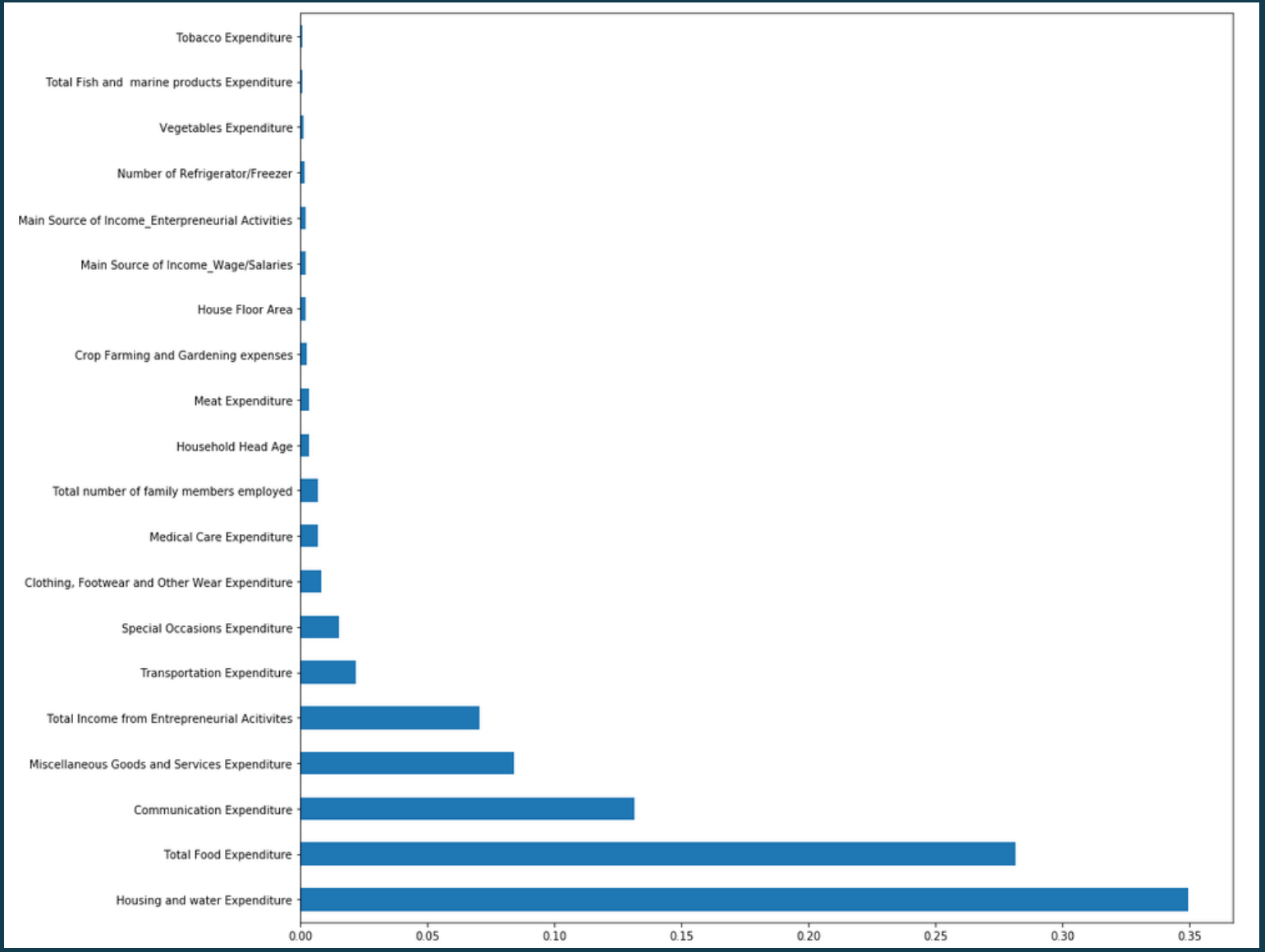
	precision	recall	f1-score	support
0	0.82	0.84	0.83	2051
1	0.65	0.68	0.66	2090
2	0.69	0.67	0.68	2110
3	0.88	0.85	0.86	2058
accuracy			0.76	8309
macro avg	0.76	0.76	0.76	8309
weighted avg	0.76	0.76	0.76	8309

Non-Num

	precision	recall	f1-score	support
0	0.49	0.56	0.52	2051
1	0.34	0.29	0.32	2090
2	0.34	0.29	0.32	2110
3	0.55	0.65	0.59	2058
accuracy			0.45	8309
macro avg	0.43	0.45	0.44	8309
weighted avg	0.43	0.45	0.44	8309

	precision	recall	f1-score	support
0	0.54	0.65	0.59	2051
1	0.38	0.24	0.29	2090
2	0.38	0.28	0.32	2110
3	0.53	0.78	0.63	2058
accuracy			0.48	8309
macro avg	0.46	0.49	0.46	8309
weighted avg	0.46	0.48	0.46	8309

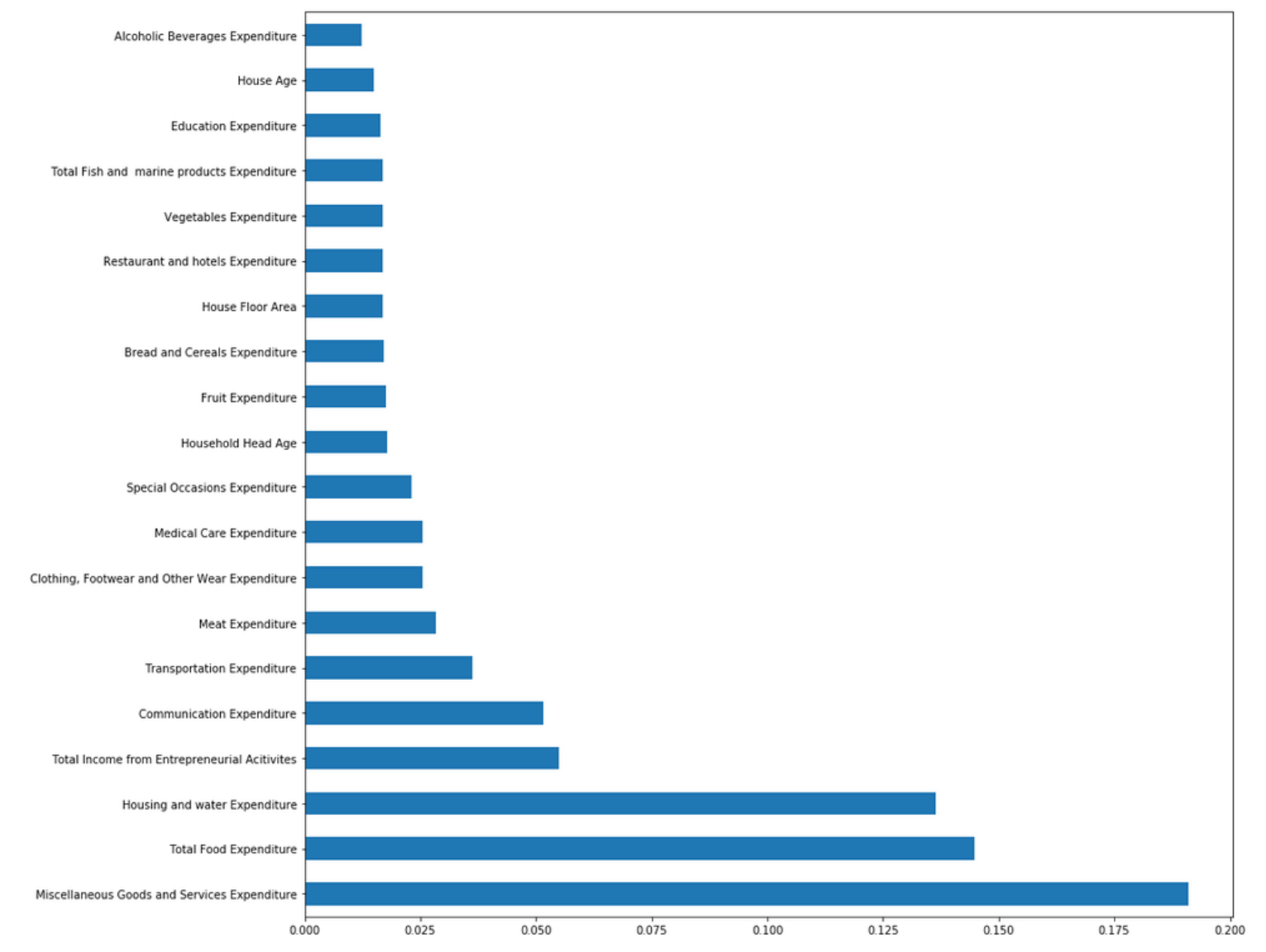
DecisionTree



Top-5 important features

1. Housing and water Expenditure
2. Total Food Expenditure
3. Communication Expenditure
4. Miscellaneous Goods and Services Expenditure
5. Total Income from Entrepreneurial Activities

RandomForest



Top-5 important features

1. Miscellaneous Goods and Services Expenditure
2. Total Food Expenditure
3. Housing and water Expenditure
4. Total Income from Entrepreneurial Activities
5. Communication Expenditure

SVM

Numeric only

	precision	recall	f1-score	support
0	0.78	0.80	0.79	2051
1	0.59	0.63	0.61	2090
2	0.65	0.66	0.65	2110
3	0.89	0.81	0.85	2058
accuracy			0.72	8309
macro avg	0.73	0.72	0.72	8309
weighted avg	0.73	0.72	0.72	8309

Categorical (Non-num)

	precision	recall	f1-score	support
0	0.58	0.62	0.60	2051
1	0.40	0.35	0.37	2090
2	0.40	0.38	0.39	2110
3	0.62	0.68	0.65	2058
accuracy			0.51	8309
macro avg	0.50	0.51	0.50	8309
weighted avg	0.50	0.51	0.50	8309

Mixed

	precision	recall	f1-score	support
0	0.70	0.75	0.73	2051
1	0.51	0.52	0.52	2090
2	0.56	0.56	0.56	2110
3	0.84	0.76	0.80	2058
accuracy			0.65	8309
macro avg	0.65	0.65	0.65	8309
weighted avg	0.65	0.65	0.65	8309

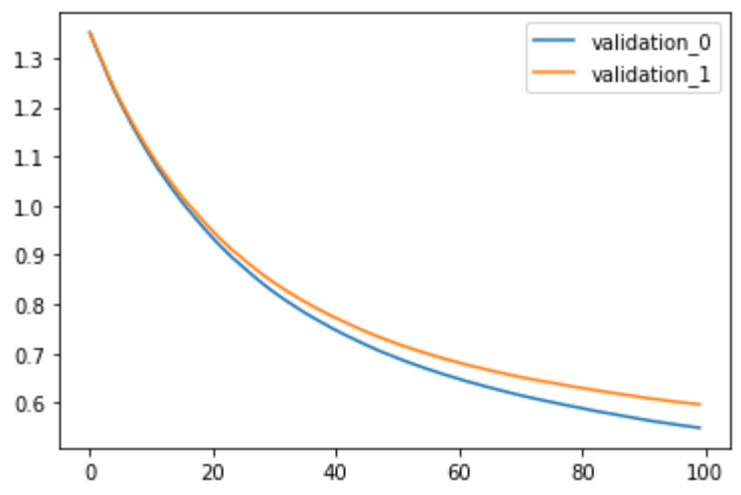
Selected Mixed

	precision	recall	f1-score	support
0	0.71	0.75	0.73	2051
1	0.51	0.53	0.52	2090
2	0.58	0.58	0.58	2110
3	0.86	0.77	0.81	2058
accuracy			0.66	8309
macro avg	0.67	0.66	0.66	8309
weighted avg	0.66	0.66	0.66	8309

XGBoost

XGBoost – All features

n_estimators=100,
max_depth=6,
learning_rate=0.04,
objective='multi:softmax',
booster='gbtree',
reg_alpha=2.0,
reg_lambda=2.5,
gamma=0.5,
min_child_weight=2.0,
colsample_bytree=0.7,
subsample=0.5



Train:

	precision	recall	f1-score	support
0	0.84	0.87	0.86	6648
1	0.71	0.72	0.71	6586
2	0.74	0.74	0.74	6688
3	0.91	0.86	0.89	6666
accuracy			0.80	26588
macro avg	0.80	0.80	0.80	26588
weighted avg	0.80	0.80	0.80	26588

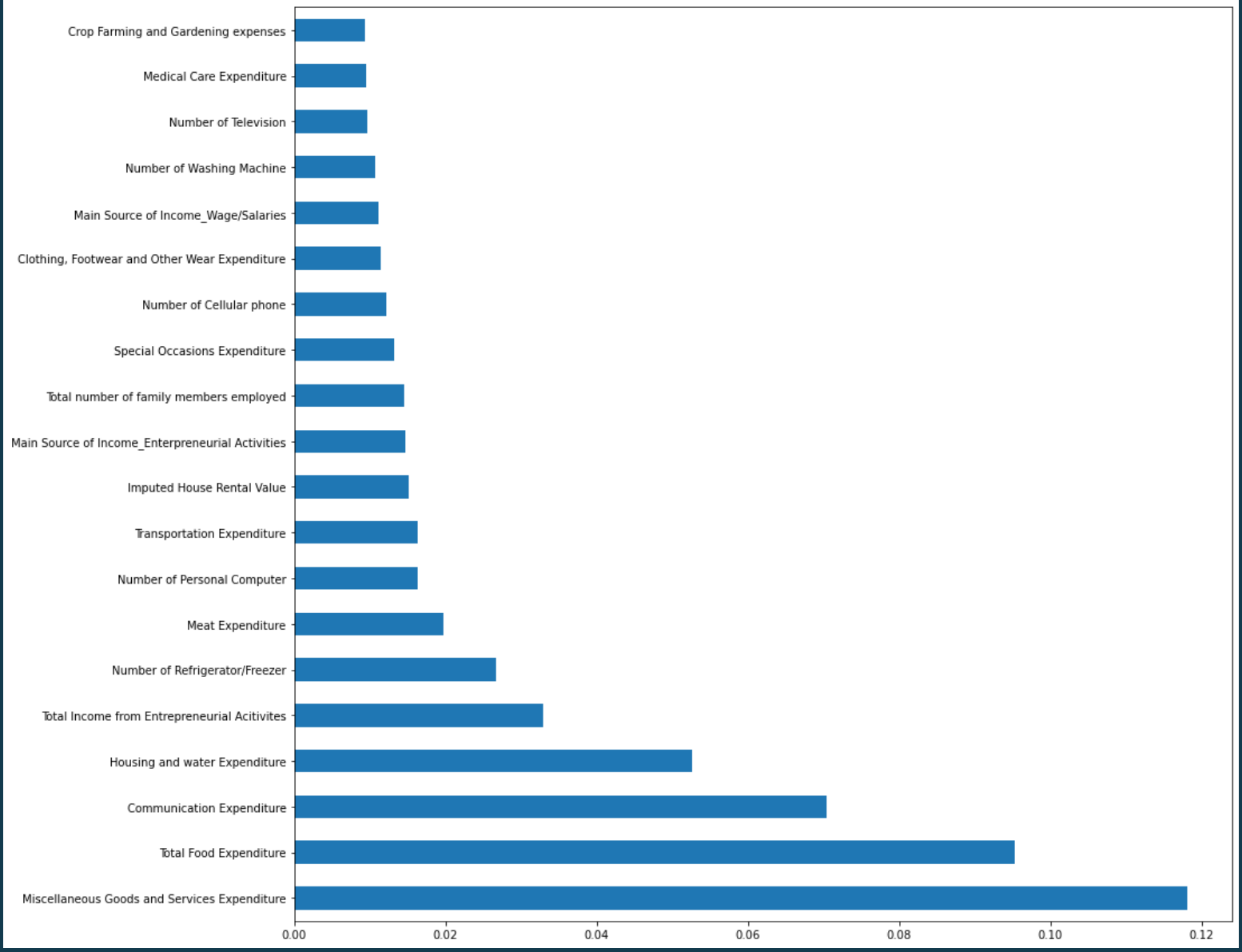
Validation:

	precision	recall	f1-score	support
0	0.82	0.86	0.84	1688
1	0.66	0.68	0.67	1709
2	0.67	0.67	0.67	1588
3	0.89	0.83	0.86	1662
accuracy			0.76	6647
macro avg	0.76	0.76	0.76	6647
weighted avg	0.76	0.76	0.76	6647

Test:

	precision	recall	f1-score	support
0	0.82	0.85	0.83	2051
1	0.65	0.67	0.66	2090
2	0.69	0.68	0.68	2110
3	0.89	0.83	0.86	2058
accuracy			0.76	8309
macro avg	0.76	0.76	0.76	8309
weighted avg	0.76	0.76	0.76	8309

XGBoost



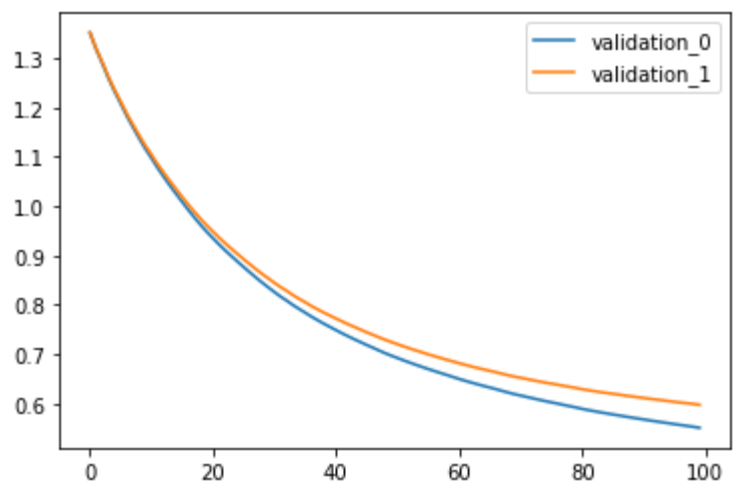
Top-5 important features

1. Miscellaneous Goods and Services Expenditure
2. Total Food Expenditure
3. Communication Expenditure
4. Housing and water Expenditure
5. Total Income from Entrepreneurial Activities

XGBoost

XGBoost – Selected features

n_estimators=100,
max_depth=6,
learning_rate=0.04,
objective='multi:softmax',
booster='gbtree',
reg_alpha=2.0,
reg_lambda=2.5,
gamma=0.5,
min_child_weight=2.0,
colsample_bytree=0.7,
subsample=0.5



Train:

	precision	recall	f1-score	support
0	0.85	0.87	0.86	6648
1	0.70	0.73	0.72	6586
2	0.74	0.73	0.74	6688
3	0.91	0.86	0.89	6666
accuracy			0.80	26588
macro avg	0.80	0.80	0.80	26588
weighted avg	0.80	0.80	0.80	26588

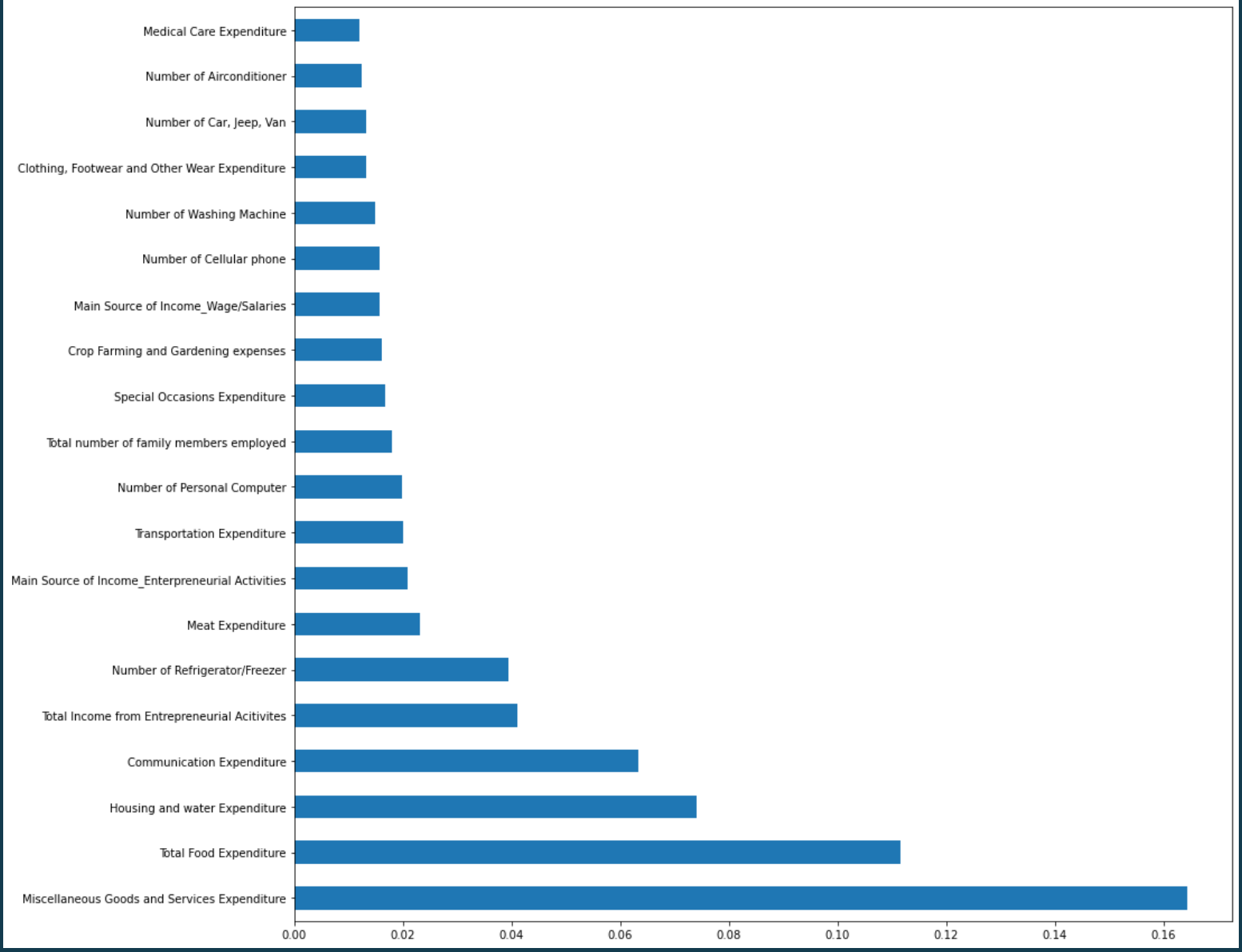
Validation:

	precision	recall	f1-score	support
0	0.82	0.86	0.84	1688
1	0.66	0.68	0.67	1709
2	0.67	0.67	0.67	1588
3	0.89	0.82	0.86	1662
accuracy			0.76	6647
macro avg	0.76	0.76	0.76	6647
weighted avg	0.76	0.76	0.76	6647

Test:

	precision	recall	f1-score	support
0	0.82	0.85	0.83	2051
1	0.64	0.68	0.66	2090
2	0.69	0.67	0.68	2110
3	0.89	0.84	0.86	2058
accuracy			0.76	8309
macro avg	0.76	0.76	0.76	8309
weighted avg	0.76	0.76	0.76	8309

XGBoost

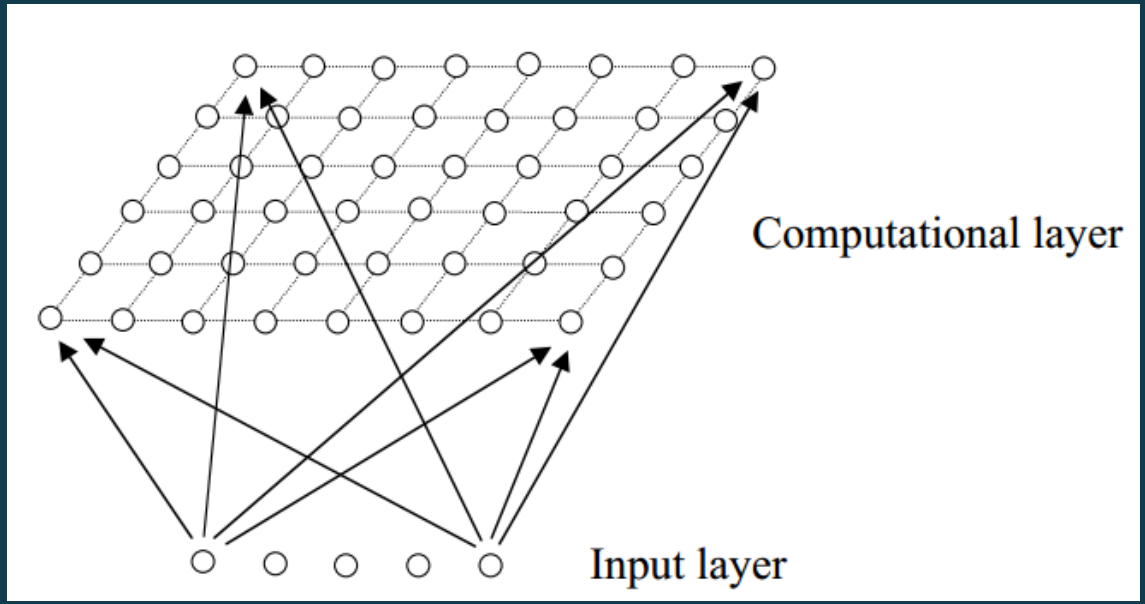


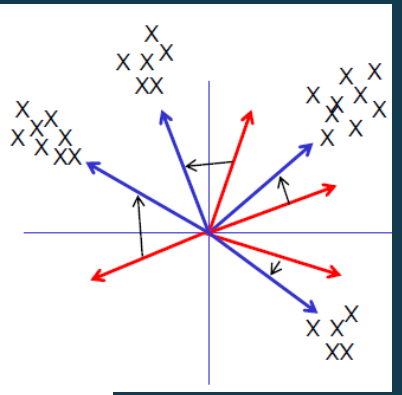
Top-5 important features

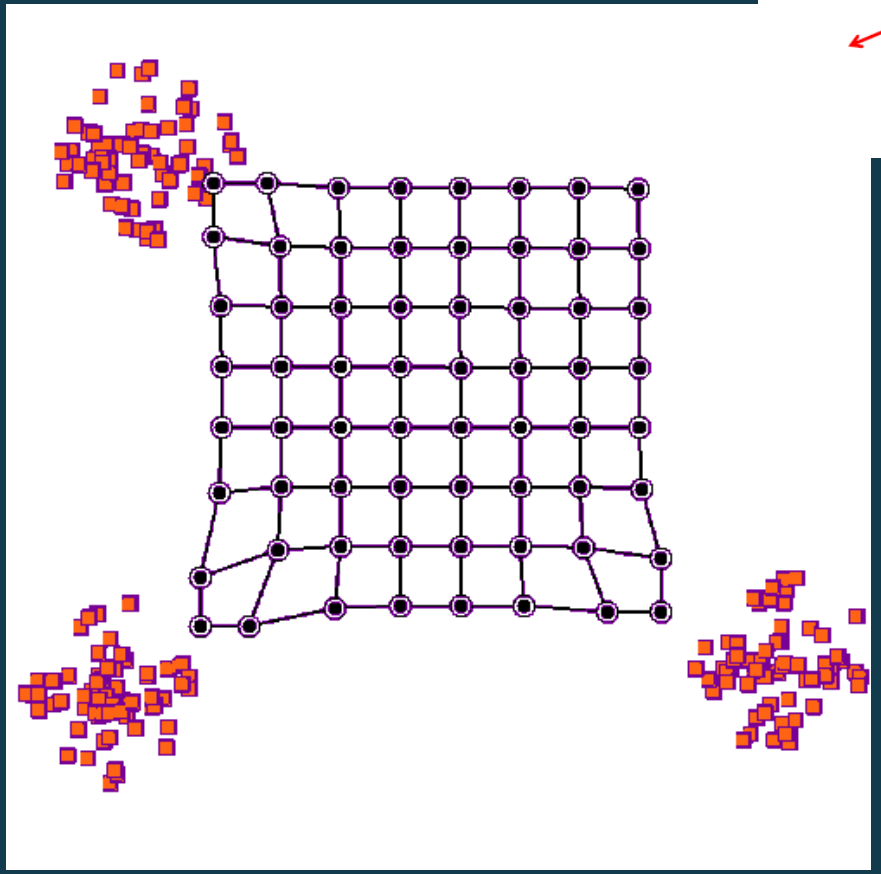
1. Miscellaneous Goods and Services Expenditure
2. Total Food Expenditure
3. Housing and water Expenditure
4. Communication Expenditure
5. Total Income from Entrepreneurial Activities

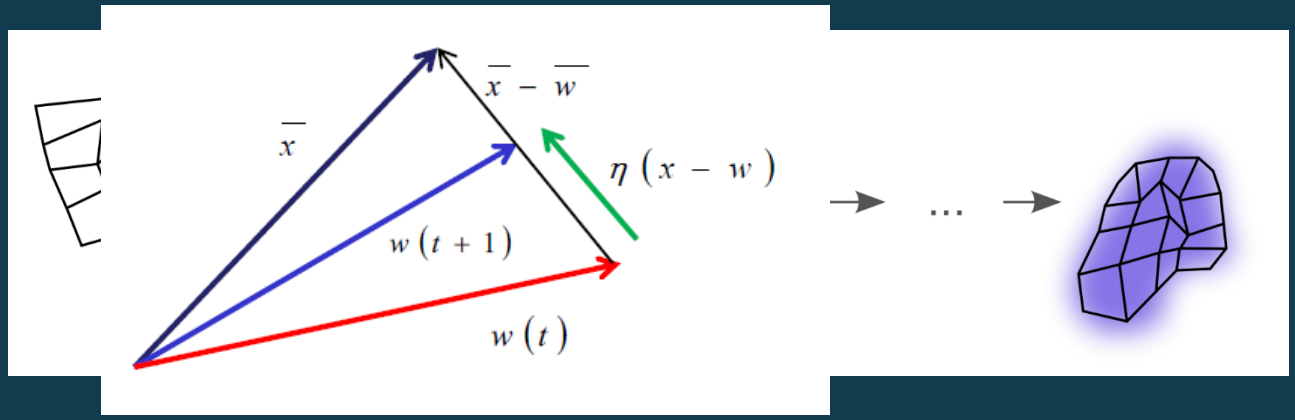
SOM+XGBoost

What is SOM ? Gamma, Learning_rate



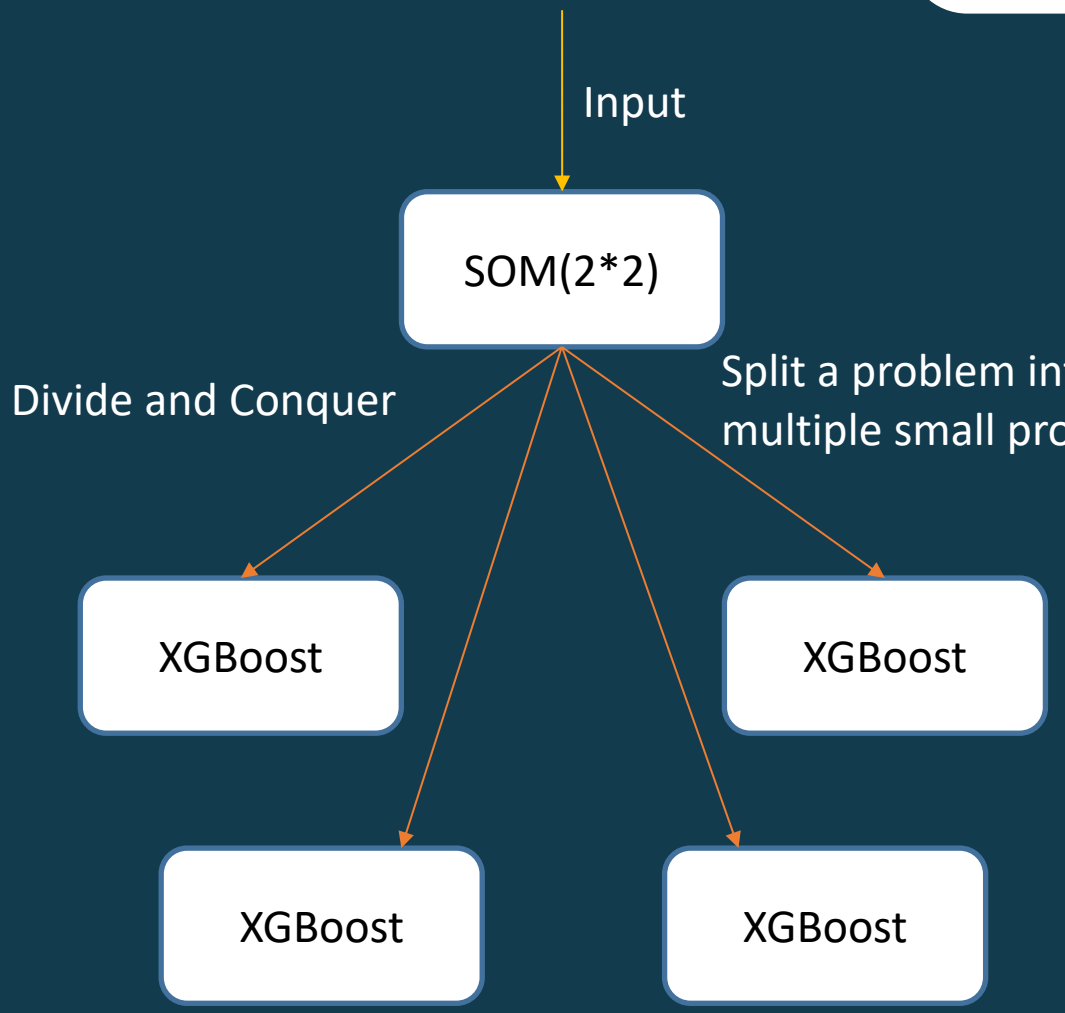






$$\Delta w_{ji} = \eta(t)\ .\ T_{j,I(\mathbf{x})}(t)\ .\ (x_i - w_{ji})$$

SOM+XGBoost



Calculate weighted average f1_score and accuracy based on (# of data in cluster / total # of data)

SOM model
size_x=2,
size_y=2,
len(features),
sigma=1, learning_rate=1.2,
neighborhood_function='triangle'

XGBoost
n_estimators=100,
max_depth=6,
learning_rate=0.04,
objective='multi:softmax',
booster='gbtree',
reg_alpha=2.0,
reg_lambda=2.5,
gamma=0.5,
min_child_weight=2.0,
colsample_bytree=0.7,
subsample=0.5

SOM+XGBoost

All features

Train

weighted avg f_score: 0.8362633198800108
weighted avg accuracy: 0.8362419136452535

	precision	recall	f1-score	support
0	0.88	0.90	0.89	1778
1	0.77	0.80	0.78	1837
2	0.81	0.81	0.81	1927
3	0.94	0.88	0.91	1687
accuracy			0.85	7229
macro avg	0.85	0.85	0.85	7229
weighted avg	0.85	0.85	0.85	7229

	precision	recall	f1-score	support
0	0.89	0.92	0.90	1043
1	0.75	0.71	0.73	829
2	0.77	0.79	0.78	1164
3	0.93	0.91	0.92	1673
accuracy			0.85	4709
macro avg	0.83	0.83	0.83	4709
weighted avg	0.85	0.85	0.85	4709

	precision	recall	f1-score	support
0	0.88	0.92	0.90	2759
1	0.77	0.81	0.79	1959
2	0.84	0.69	0.75	930
3	0.94	0.64	0.76	213
accuracy			0.84	5861
macro avg	0.86	0.76	0.80	5861
weighted avg	0.84	0.84	0.83	5861

	precision	recall	f1-score	support
0	0.83	0.76	0.79	1068
1	0.73	0.76	0.75	1961
2	0.77	0.81	0.79	2667
3	0.93	0.89	0.91	3093
accuracy			0.82	8789
macro avg	0.81	0.80	0.81	8789
weighted avg	0.82	0.82	0.82	8789

Validation

weighted avg f_score: 0.7616251247579151
weighted avg accuracy: 0.7622987814051453

	precision	recall	f1-score	support
0	0.83	0.84	0.83	449
1	0.66	0.72	0.69	458
2	0.69	0.69	0.69	417
3	0.90	0.81	0.85	407
accuracy			0.76	1731
macro avg	0.77	0.76	0.77	1731
weighted avg	0.77	0.76	0.77	1731

	precision	recall	f1-score	support
0	0.85	0.90	0.87	282
1	0.62	0.55	0.58	216
2	0.60	0.61	0.61	285
3	0.83	0.84	0.84	392
accuracy			0.75	1175
macro avg	0.73	0.73	0.72	1175
weighted avg	0.74	0.75	0.74	1175

	precision	recall	f1-score	support
0	0.85	0.88	0.86	690
1	0.69	0.74	0.72	517
2	0.69	0.54	0.61	211
3	0.78	0.50	0.61	62
accuracy			0.77	1480
macro avg	0.75	0.67	0.70	1480
weighted avg	0.77	0.77	0.77	1480

	precision	recall	f1-score	support
0	0.73	0.73	0.73	267
1	0.67	0.65	0.66	518
2	0.70	0.75	0.72	675
3	0.90	0.87	0.89	801
accuracy			0.77	2261
macro avg	0.75	0.75	0.75	2261
weighted avg	0.77	0.77	0.77	2261

Test

weighted avg f_score: 0.7584503520165833
weighted avg accuracy: 0.7588157419665424

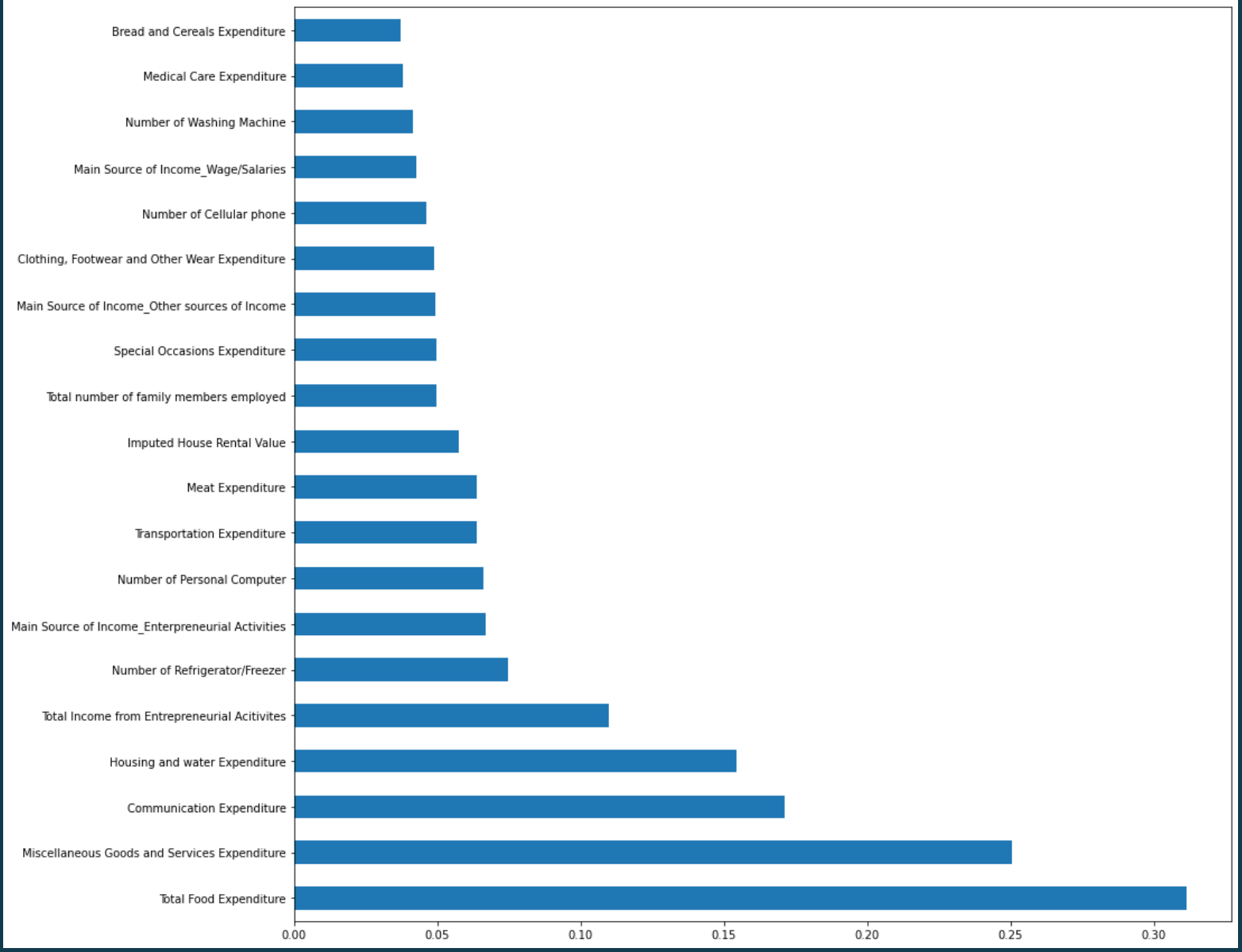
	precision	recall	f1-score	support
0	0.84	0.83	0.83	540
1	0.67	0.70	0.68	557
2	0.71	0.75	0.73	573
3	0.89	0.80	0.84	480
accuracy			0.77	2150
macro avg	0.78	0.77	0.77	2150
weighted avg	0.77	0.77	0.77	2150

	precision	recall	f1-score	support
0	0.84	0.87	0.85	341
1	0.54	0.50	0.52	256
2	0.65	0.64	0.64	387
3	0.86	0.87	0.87	502
accuracy			0.75	1486
macro avg	0.72	0.72	0.72	1486
weighted avg	0.74	0.75	0.74	1486

	precision	recall	f1-score	support
0	0.84	0.88	0.86	843
1	0.68	0.74	0.71	626
2	0.72	0.57	0.64	315
3	0.86	0.44	0.58	70
accuracy			0.76	1854
macro avg	0.78	0.66	0.70	1854
weighted avg	0.77	0.76	0.76	1854

	precision	recall	f1-score	support
0	0.74	0.71	0.72	327
1	0.64	0.67	0.66	651
2	0.68	0.71	0.69	835
3	0.90	0.86	0.88	1006
accuracy			0.75	2819
macro avg	0.74	0.74	0.74	2819
weighted avg	0.76	0.75	0.76	2819

SOM+XGBoost



Top-5 important features

1. Total Food Expenditure
2. Miscellaneous Goods and Services Expenditure
3. Communication Expenditure
4. Housing and water Expenditure
5. Total Income from Entrepreneurial Activities

SOM+XGBoost

Selected features

Train

weighted avg f_score: 0.8353064361614327
weighted avg accuracy: 0.8353016398375207

	precision	recall	f1-score	support
0	0.90	0.94	0.92	3159
1	0.79	0.82	0.81	1957
2	0.87	0.66	0.75	770
3	0.99	0.60	0.75	127
accuracy			0.86	6013
macro avg	0.89	0.76	0.81	6013
weighted avg	0.86	0.86	0.86	6013

	precision	recall	f1-score	support
0	0.82	0.78	0.80	1175
1	0.74	0.77	0.75	1950
2	0.78	0.81	0.79	2216
3	0.93	0.88	0.91	2113
accuracy			0.82	7454
macro avg	0.82	0.81	0.81	7454
weighted avg	0.82	0.82	0.82	7454

	precision	recall	f1-score	support
0	0.84	0.72	0.77	303
1	0.75	0.78	0.77	667
2	0.80	0.83	0.82	1179
3	0.94	0.92	0.93	1711
accuracy			0.85	3860
macro avg	0.83	0.81	0.82	3860
weighted avg	0.86	0.85	0.85	3860

	precision	recall	f1-score	support
0	0.87	0.88	0.88	2011
1	0.74	0.74	0.74	2012
2	0.77	0.79	0.78	2523
3	0.92	0.88	0.90	2715
accuracy			0.83	9261
macro avg	0.82	0.83	0.82	9261
weighted avg	0.83	0.83	0.83	9261

Validation

weighted avg f_score: 0.7634193207879316
weighted avg accuracy: 0.763953663306755

	precision	recall	f1-score	support
0	0.87	0.91	0.89	811
1	0.73	0.77	0.75	528
2	0.70	0.50	0.58	162
3	0.90	0.42	0.58	45
accuracy			0.81	1546
macro avg	0.80	0.65	0.70	1546
weighted avg	0.80	0.81	0.80	1546

	precision	recall	f1-score	support
0	0.75	0.75	0.75	284
1	0.67	0.66	0.66	491
2	0.68	0.73	0.70	536
3	0.89	0.84	0.86	511
accuracy			0.74	1822
macro avg	0.75	0.74	0.74	1822
weighted avg	0.75	0.74	0.74	1822

	precision	recall	f1-score	support
0	0.78	0.65	0.71	88
1	0.63	0.63	0.63	183
2	0.69	0.75	0.72	319
3	0.91	0.88	0.89	433
accuracy			0.78	1023
macro avg	0.75	0.73	0.74	1023
weighted avg	0.78	0.78	0.78	1023

	precision	recall	f1-score	support
0	0.82	0.84	0.83	505
1	0.63	0.63	0.63	507
2	0.65	0.67	0.66	571
3	0.87	0.82	0.85	673
accuracy			0.75	2256
macro avg	0.74	0.74	0.74	2256
weighted avg	0.75	0.75	0.75	2256

Test

weighted avg f_score: 0.765380433404814
weighted avg accuracy: 0.7655554218317489

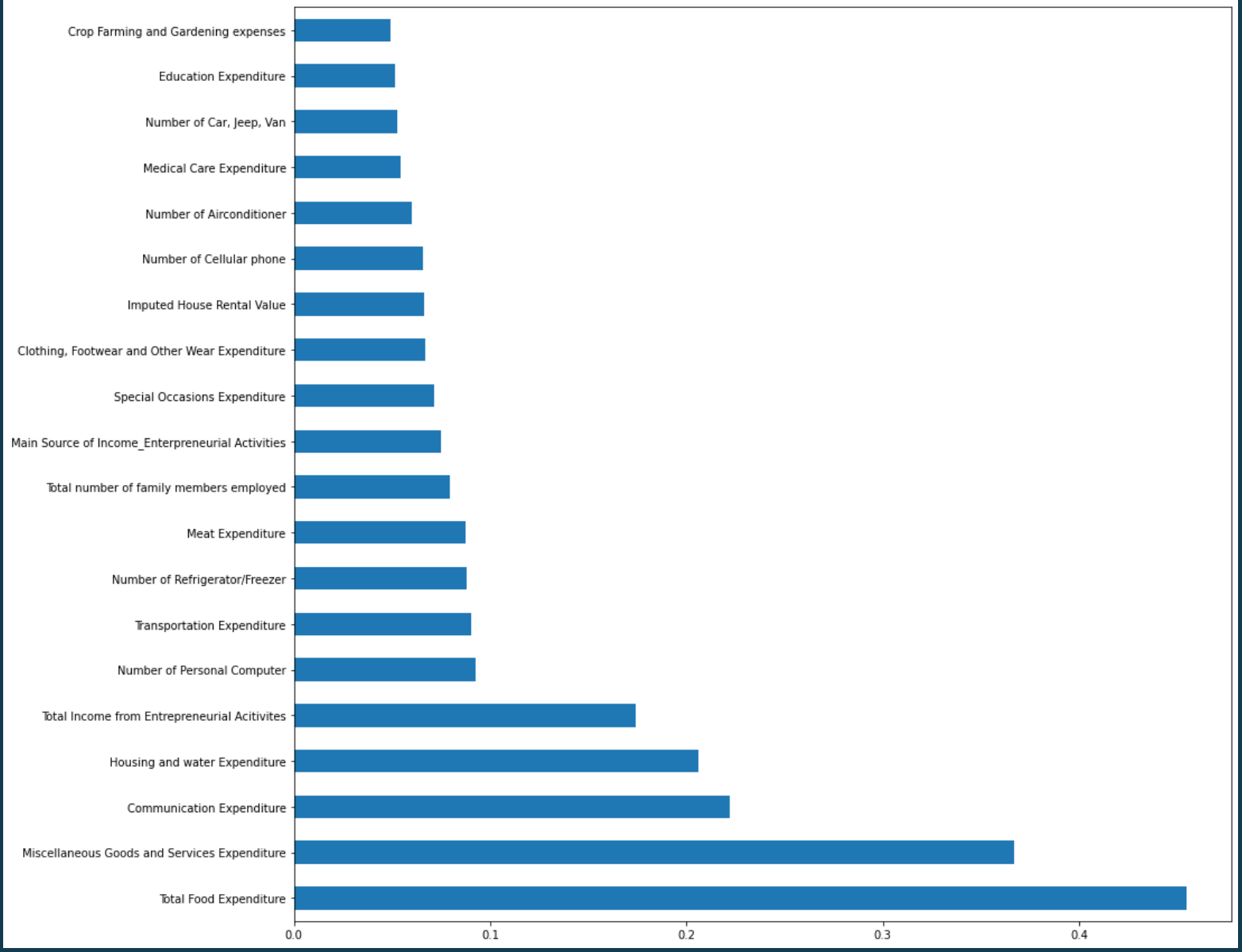
	precision	recall	f1-score	support
0	0.85	0.91	0.88	958
1	0.70	0.74	0.72	641
2	0.76	0.56	0.64	278
3	1.00	0.34	0.51	50
accuracy			0.79	1927
macro avg	0.83	0.64	0.69	1927
weighted avg	0.79	0.79	0.78	1927

	precision	recall	f1-score	support
0	0.76	0.75	0.76	346
1	0.66	0.70	0.68	587
2	0.70	0.70	0.70	662
3	0.89	0.85	0.87	643
accuracy			0.75	2238
macro avg	0.75	0.75	0.75	2238
weighted avg	0.75	0.75	0.75	2238

	precision	recall	f1-score	support
0	0.80	0.60	0.68	87
1	0.62	0.66	0.64	212
2	0.69	0.74	0.71	384
3	0.91	0.87	0.89	566
accuracy			0.78	1249
macro avg	0.75	0.72	0.73	1249
weighted avg	0.79	0.78	0.78	1249

	precision	recall	f1-score	support
0	0.83	0.81	0.82	660
1	0.62	0.63	0.63	650
2	0.68	0.74	0.71	786
3	0.90	0.83	0.86	799
accuracy			0.76	2895
macro avg	0.76	0.75	0.75	2895
weighted avg	0.76	0.76	0.76	2895

SOM+XGBoost



Top-5 important features

1. Total Food Expenditure
2. Miscellaneous Goods and Services Expenditure
3. Communication Expenditure
4. Housing and water Expenditure
5. Total Income from Entrepreneurial Activities

F1 - score

Evaluation

	Decision Tree	Random Forest	SVM	KNN	MDC	SOM	SOM + XGBoost
Non-Selected	0.7	0.76	0.65	0.4827	0.4724	0.76	0.7588
Selected	0.7	0.76	0.66	0.4918	0.4696	0.76	0.7653

Accuracy

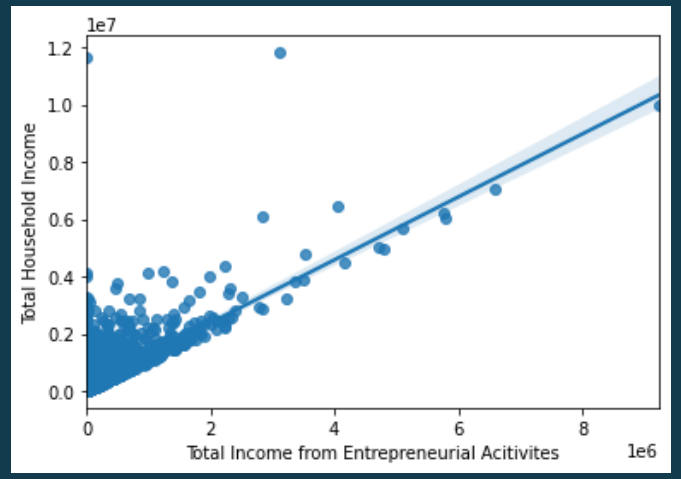
Evaluation

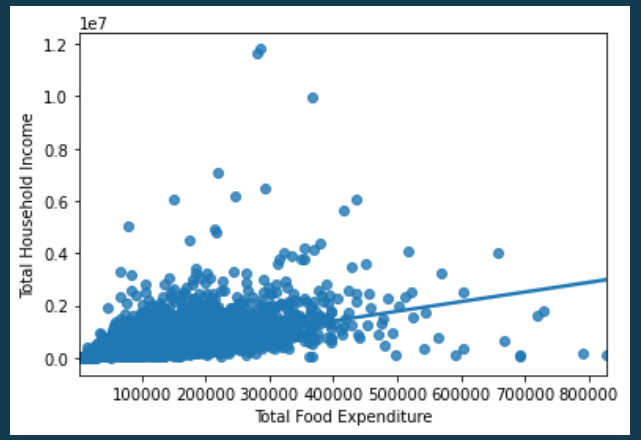
	Decision Tree	Random Forest	SVM	KNN	MDC	SOM	SOM + XGBoost
Non-Selected	0.7	0.76	0.65	0.4833	0.4830	0.76	0.7584
Selected	0.71	0.76	0.66	0.4936	0.4832	0.76	0.7653

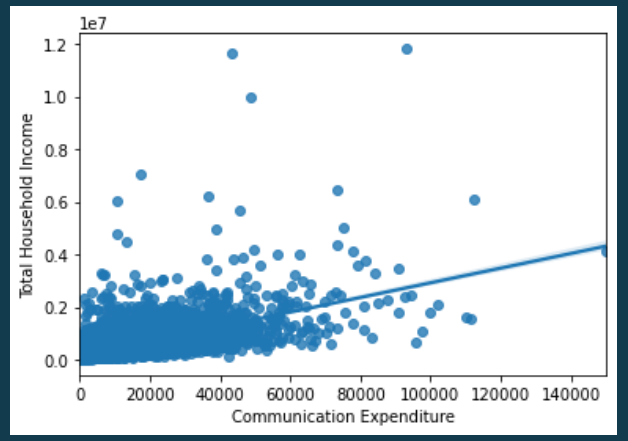
Conclusion

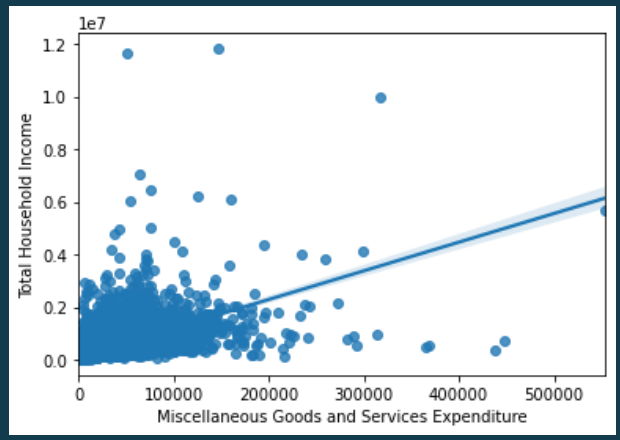
Top-5 important features

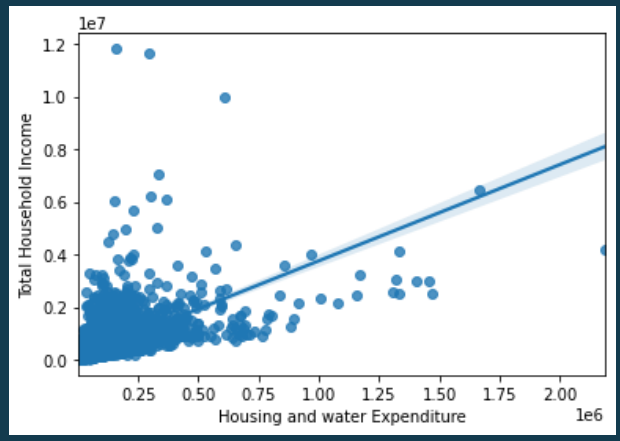
1. Total Food Expenditure
2. Miscellaneous Goods and Services Expenditure
3. Communication Expenditure
4. Housing and water Expenditure
5. Total Income from Entrepreneurial Activities



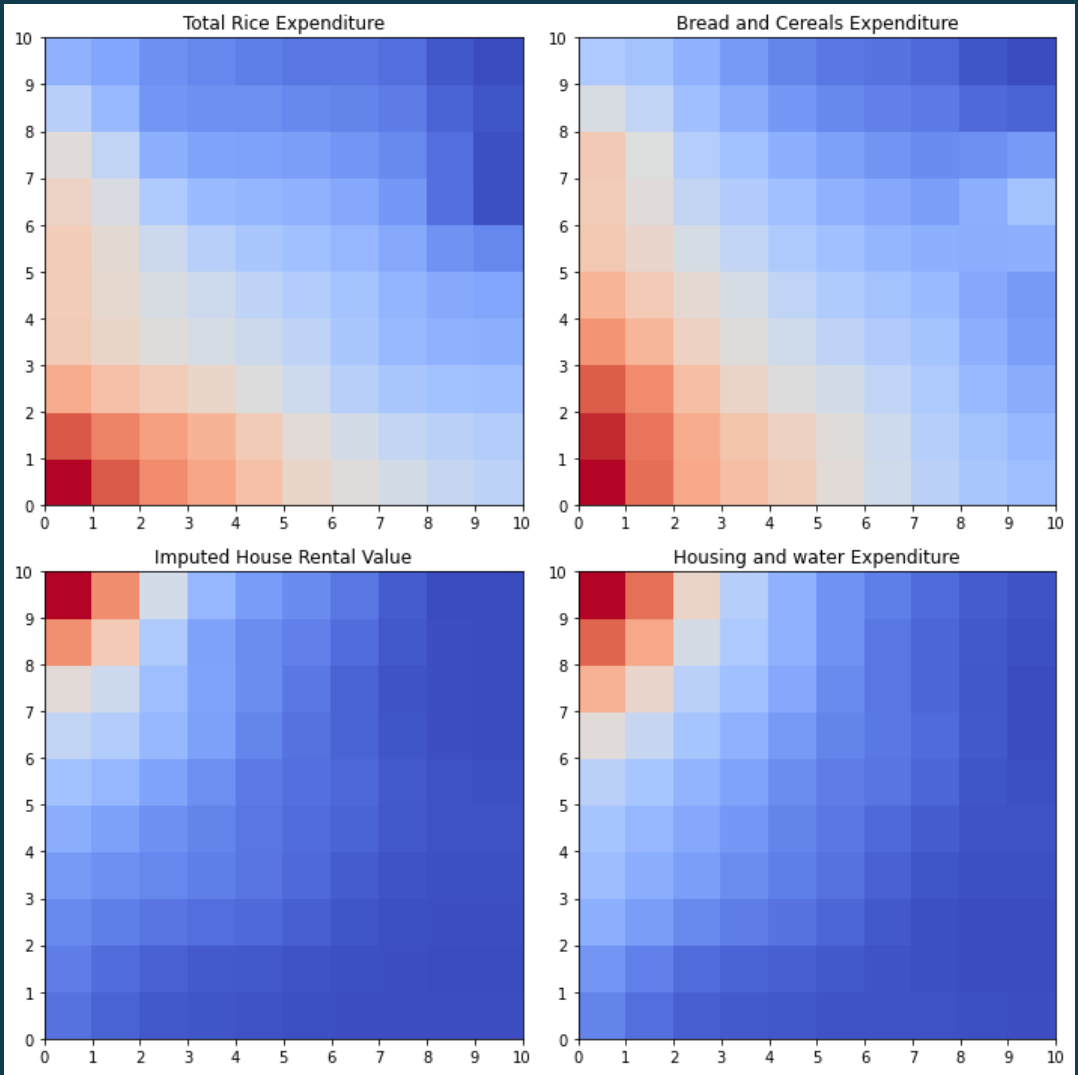


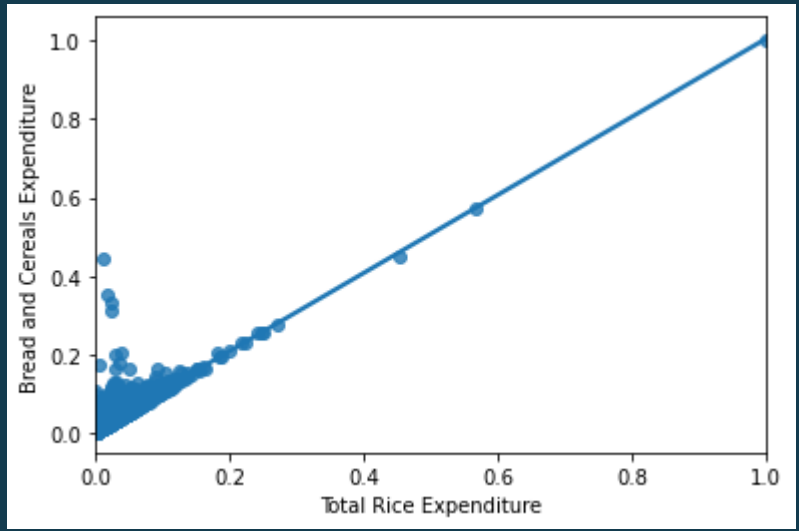


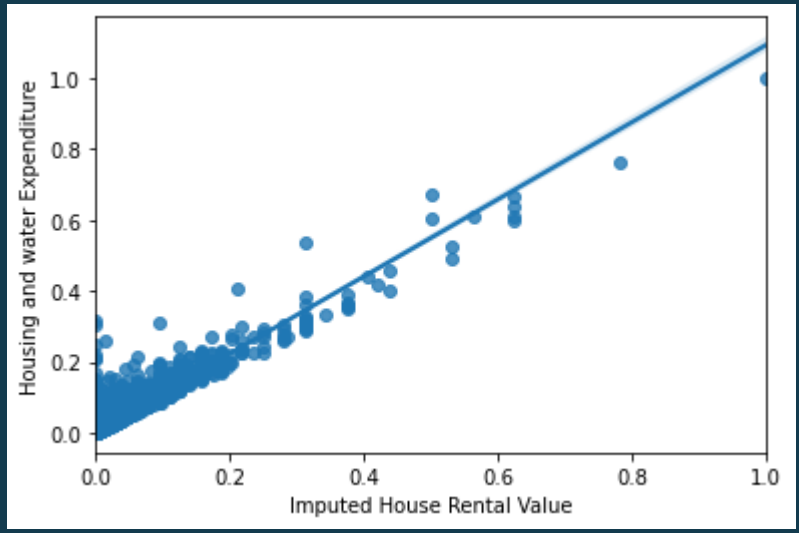


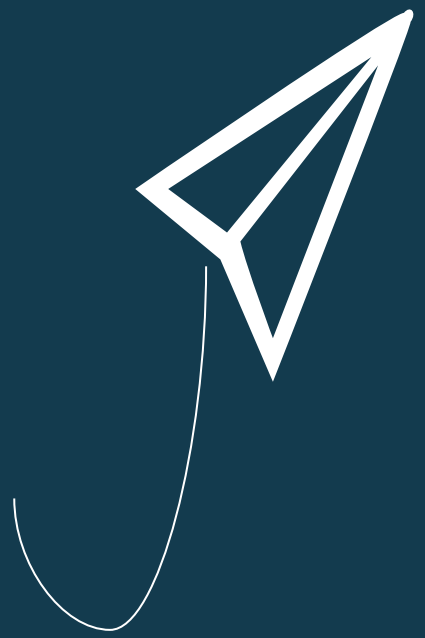


Conclusion


Total Rice Expenditure
Bread and Cereals Expenditure
Imputed House Rental Value
Housing and water Expenditure


Bread and Cereals Expenditure
Total Rice Expenditure


Housing and water Expenditure
Imputed House Rental Value



Have A Nice Day