Income Class Prediction through Expenditure Allocation

Genesis Adam D. Mendoza

About the data

	Total Household Income	Region	Total Food Expenditure	Main Source of Income	Agricultural Household indicator	Bread and Cereals Expenditure	Total Rice Expenditure	Meat Expenditure	Total Fish and marine products Expenditure	Fruit Expenditure	...
0	480332	CAR	117848	Wage/Salaries	0	42140	38300	24676	16806	3325	...
1	198235	CAR	67766	Wage/Salaries	0	17329	13008	17434	11073	2035	...
2	82785	CAR	61609	Wage/Salaries	1	34182	32001	7783	2590	1730	...
3	107589	CAR	78189	Wage/Salaries	0	34030	28659	10914	10812	690	...
4	189322	CAR	94625	Wage/Salaries	0	34820	30167	18391	11309	1395	...

- Family Income and Expenditure Survey (2015) from the Philippine Statistics Authority.
- Includes total annual income, sources of income, items of expenditure, household characteristics, etc.
- Includes $\sim 41{,}500$ households.

About the data

'Total Household Income', 'Region', 'Total Food Expenditure', 'Main Source of Income', 'Agricultural Household indicator', 'Bread and Cereals Expenditure', 'Total Rice 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 Acitivites', 'Household Head Sex', 'Household Head Age', 'Household Head Marital Status', 'Household Head Highest Grade Completed', 'Household Head Job or Business Indicator', 'Household Head Occupation', 'Household Head Class of Worker', 'Type of Household', 'Total Number of Family members', 'Members with age less than 5 year old', 'Members with age 5 - 17 years old', 'Total number of family members employed', 'Type of Building/House', 'Type of Roof', 'Type of Walls', 'House Floor Area', 'House Age', 'Number of bedrooms', 'Tenure Status', 'Toilet Facilities', 'Electricity', 'Main Source of Water Supply', '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 Motorized Banca', 'Number of Motorcycle/Tricycle'

Income classification

- Philippine Institute for Development Studies (PIDS) income classification scheme (yearly)
    - Poor: $[0, 131484)$
    - Low income: $[131484,\ 254328)$
    - Lower middle class: $[254328,\ 525936)$
    - Middle class: $[525936,\ 920028)$
    - Upper middle class: $[920028, 1577808)$
    - High income: $[1577808, 2629680)$
    - Rich: $[2629680, \infty)$
- Does not take family size into account

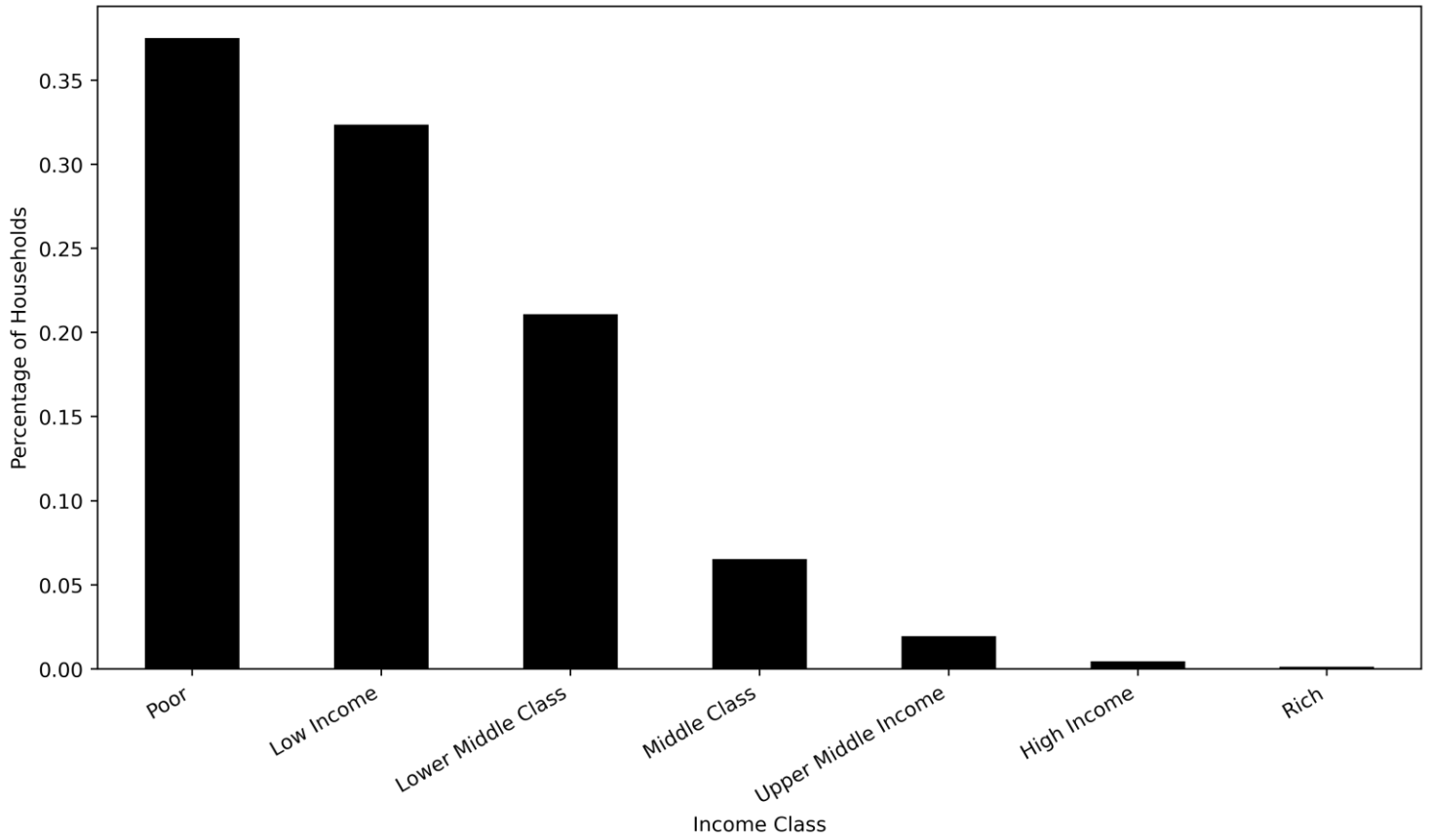
Income classification

	Total Household Income	Income Class
0	480332	Lower Middle Class
1	198235	Low Income
2	82785	Poor
3	107589	Poor
4	189322	Low Income
...	...	...
41539	119773	Poor
41540	137320	Low Income
41541	133171	Low Income
41542	129500	Poor
41543	128598	Poor

```
1  income_bins = [-float('inf'), 10957, 21194, 43828, 76669, 131484, 219140, float('inf')]
2  income_labels = ['Poor', 'Low Income', 'Lower Middle Class', 'Middle Class', 'Upper Middle Income', 'High Income', 'Rich']
3  raw_fies['Income Class'] = pd.cut(raw_fies['Total Household Income']/12, bins=income_bins, labels=income_labels, right=False)
```

Income classification

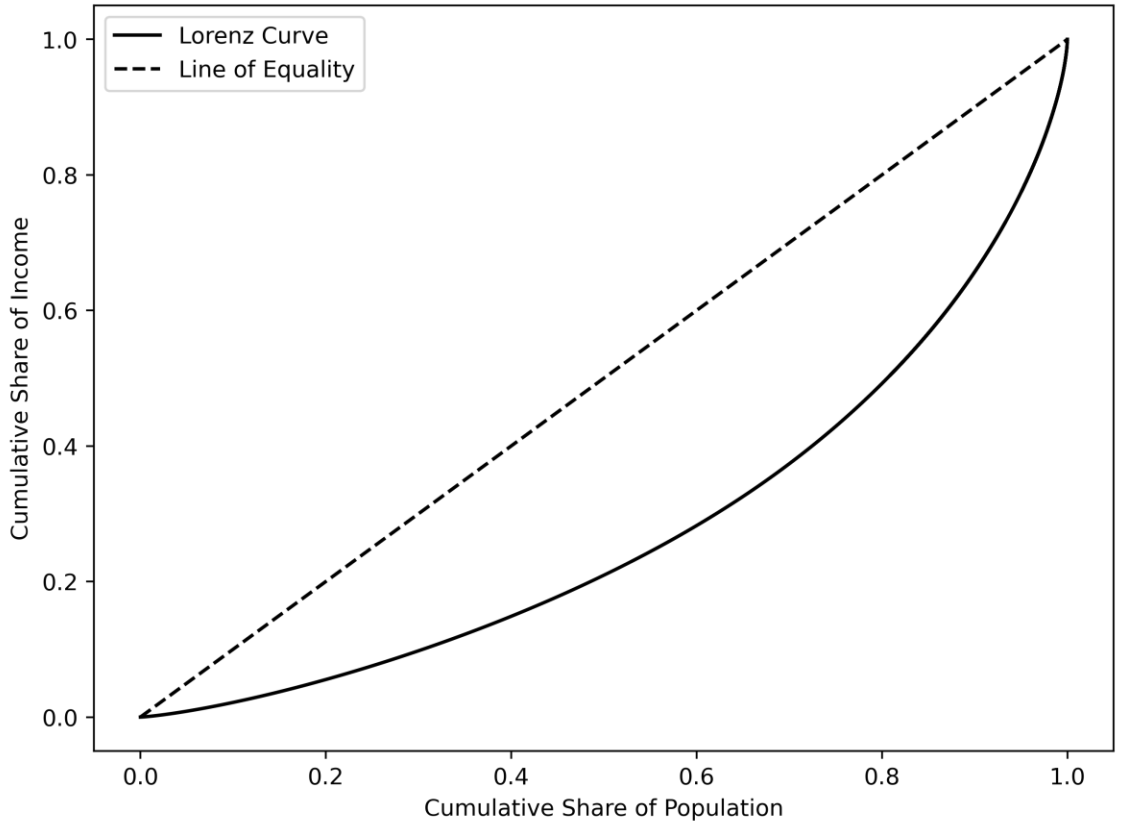
Majority of Filipinos are **poor**.



Philippine Institute for Development Studies (PIDS) income classification scheme

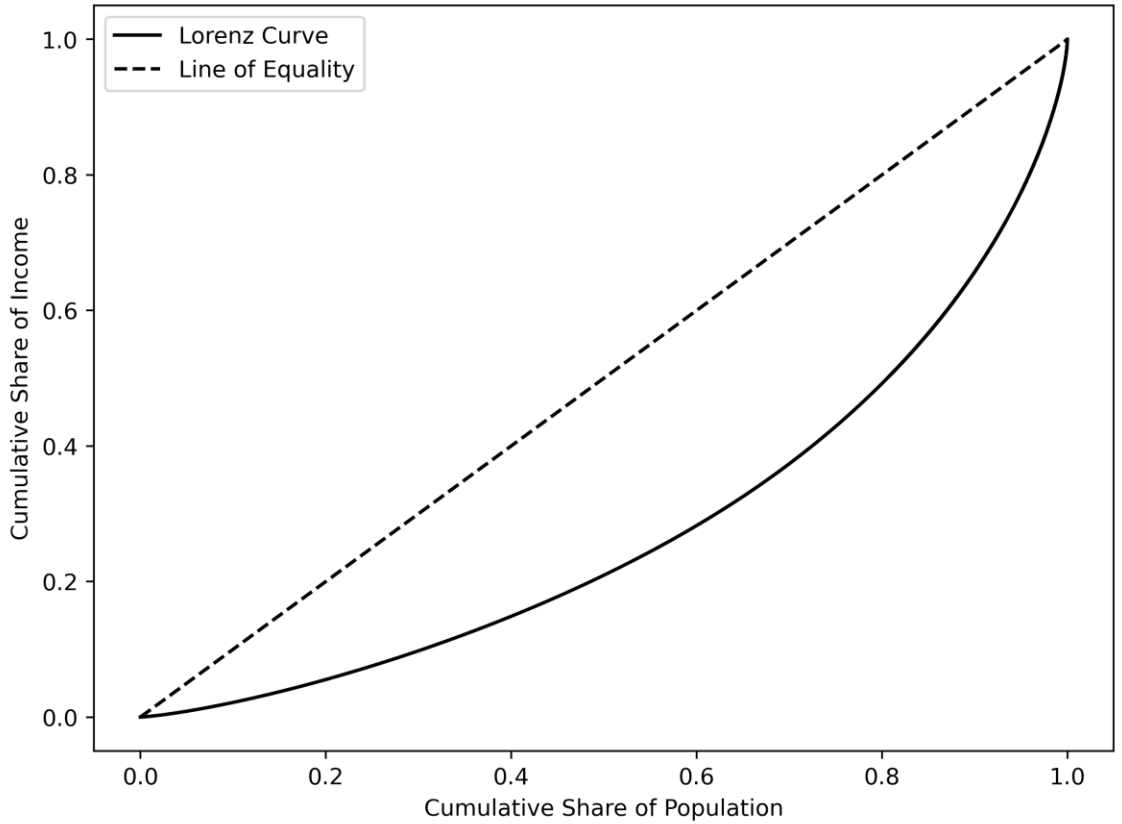
Income inequality

- If we sort the total household income of each family and we take their cumulative sum, we will get its **Lorenz curve**.
- For example, 50% of the population holds just around 20% of the total income.



Income inequality

- The Gini coefficient is twice the area between the line of equality and the Lorenz curve.
- 0 means perfect income equality and 1 means perfect inequality.
- In the FIES (2015) dataset, we have calculated a Gini coefficient of 0.4438.



Income inequality

Caveat

We are **not** proposing a solution to the problem of income inequality.

- Income inequality is complex (Estudillo, 1997):
  - Rising proportion of urban households
  - Age distribution
  - Increasing number of highly educated households
  - Wage rate inequality

- Main question:
  - How does income inequality affect the spending behavior of different income classes?

Estudillo, J. P. (1997). Income inequality in the Philippines, 1961–91. The Developing Economies, 35(1), 68-95.

Expenditure allocation

	Total Food Expenditure	Bread and Cereals Expenditure	Total Rice Expenditure	Meat Expenditure	Total Fish and marine products Expenditure	Fruit Expenditure	Vegetables Expenditure	Restaurant and hotels Expenditure	Alcoholic Beverages Expenditure
0	117848	42140	38300	24676	16806	3325	13460	3000	0
1	67766	17329	13008	17434	11073	2035	7833	2360	960
2	61609	34182	32001	7783	2590	1730	3795	4545	270
3	78189	34030	28659	10914	10812	690	7887	6280	480
4	94625	34820	30167	18391	11309	1395	11260	6400	1040
...	...	...	...	...	...	...	...	...	...
41539	44875	23675	21542	1476	6120	1632	3882	1805	0
41540	31157	2691	1273	1886	4386	1840	3110	9090	0
41541	45882	28646	27339	480	4796	1232	3025	3330	0
41542	81416	29996	26655	2359	17730	2923	7951	13660	0
41543	78195	43485	41205	1985	7735	2062	7114	5750	0

Expenditure allocation

```
Index(['Total Food Expenditure', 'Bread and Cereals Expenditure',
       'Total Rice 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'],
      dtype='object')
```

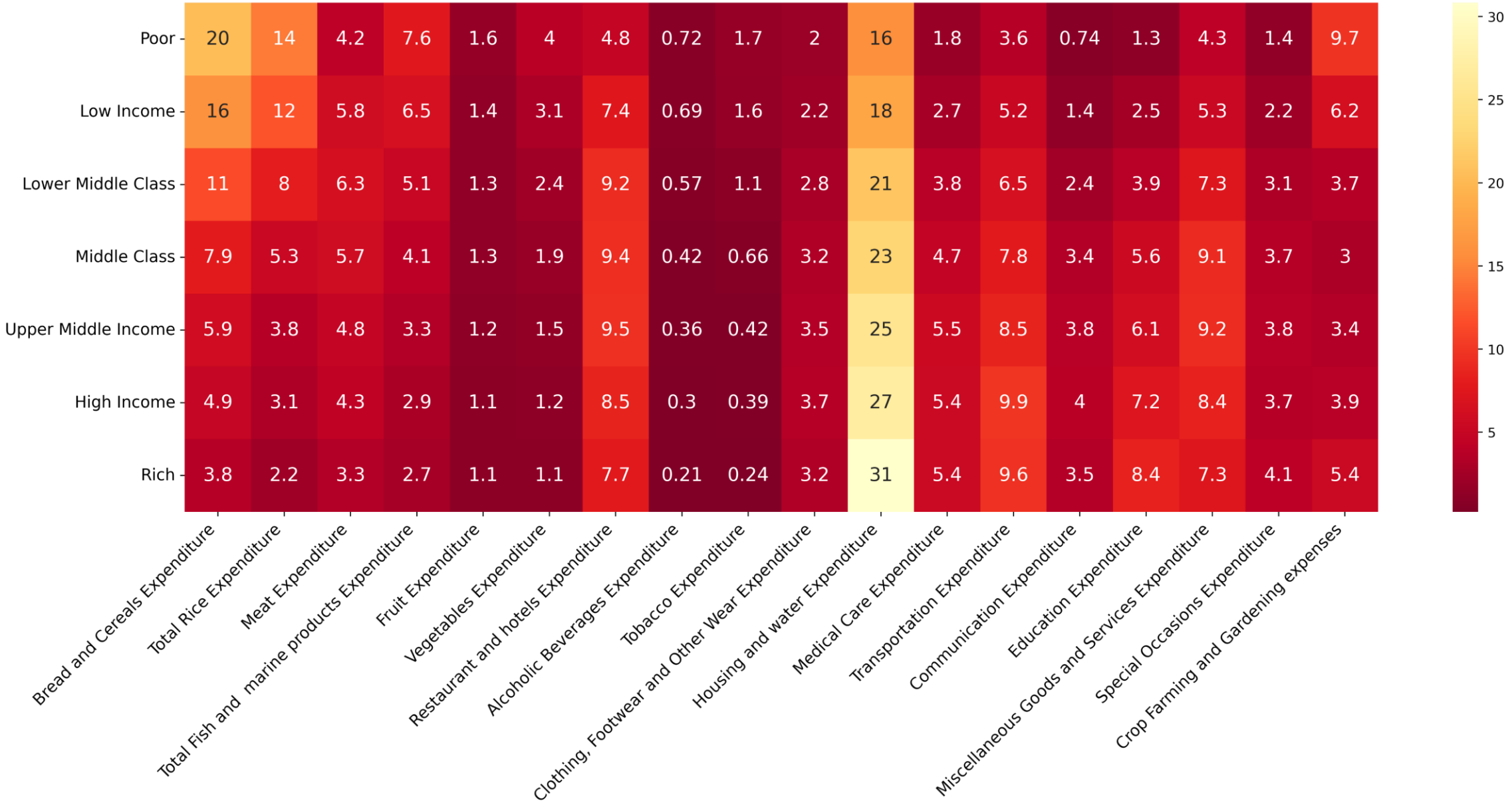
	Total Food	Bread and	Total Rice	Meat	Total Fish and marine	Fruit	Vegetables	Restaurant	Alcoholic
									es ure
									0
									60
									70
									80
									40
									...
415									0
415									0
415									0
41542	81416	29996	26655	2359	17730	2923	7951	13660	0
41543	78195	43485	41205	1985	7735	2062	7114	5750	0

Expenditure allocation

What percentage of the total expenditure is allocated for each expenditure item?

```
#Extract the expenditure columns
exp_cols = [col for col in raw_fies.columns if 'Expenditure' in col or 'expense' in col]
tent_feat_cols = exp_cols
fies_tent = raw_fies[tent_feat_cols]

#Remove the redundant column
fies_tent = fies_tent.drop(columns= ['Total Food Expenditure'])
norm_fies = fies_tent.sum(axis = 1)
fies = 100*fies_tent.div(norm_fies, axis = 0)

full = pd.DataFrame(raw_fies['Income Class']).join(fies)
grouped_df = full.groupby('Income Class')[fies.columns].mean()

plt.figure(figsize=(20, 7))
sns.heatmap(grouped_df, annot=True, cmap='YlOrRd_r', linewidths=.5, linecolor='black')
plt.ylabel('Income Class')
plt.xlabel('Expenditure Category')
plt.xticks(rotation=45, ha='right')
plt.savefig('ExpenditureAllocation.png', bbox_inches = 'tight', dpi = 400)
plt.show()
```

What percentage of the total expenditure is allocated for each expenditure item?



Prediction Model

```python
forest_model = RandomForestClassifier(random_state=0)

max_leaf_nodes = [node_val for node_val in range(1, 900, 1)]
max_depth = [depth for depth in range(1, 300, 1)]

random_grid = {'max_leaf_nodes': max_leaf_nodes, 'max_depth': max_depth}
n_estimators = [estim for estim in range(1, 100, 1)]

model_random = RandomizedSearchCV(
    estimator=forest_model,
    param_distributions={**random_grid, 'n_estimators': n_estimators},
    n_iter=30, cv=3, verbose=3, random_state=0, n_jobs=-1
)

pipeline = Pipeline(steps = [('preprocessor', preprocess), ('model', model_random)])
pipeline.fit(x_train, y_train.values.ravel())

opt_estim = model_random.best_params_['n_estimators']
opt_nodes = model_random.best_params_['max_leaf_nodes']
opt_depth = model_random.best_params_['max_depth']
```

```
Fitting 3 folds for each of 30 candidates, totalling 90 fits
The optimal n_estimator is: 99
The optimal max_leaf_nodes is: 747
The optimal max_depth is: 32
The best score for the training data given the optimum parameters is 74.90%
```

- 80 – 20 train-test division
- We use a random forest classifier

Prediction Model

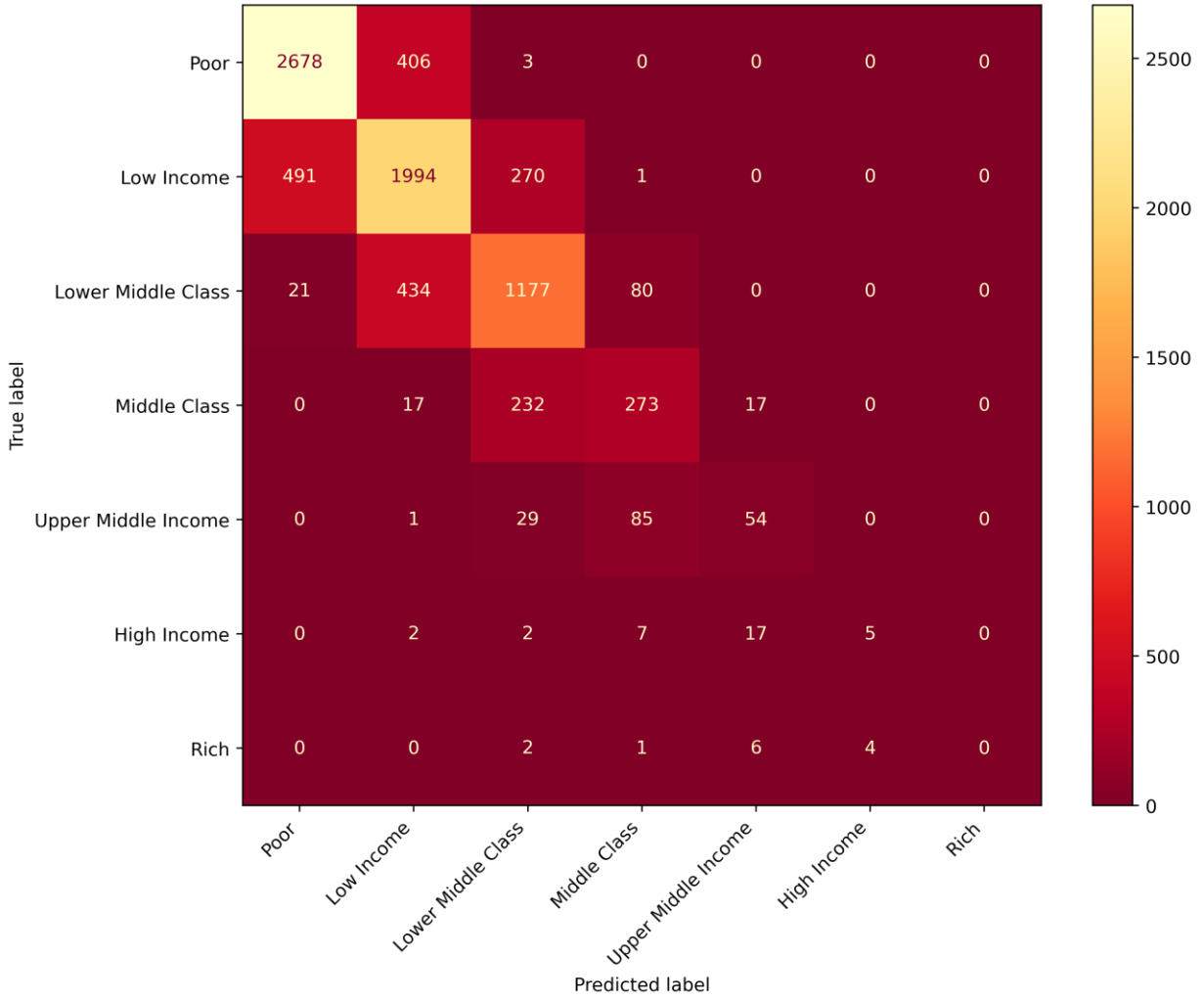
```
final_model = RandomForestClassifier(random_state=0, n_estimators=opt_estim, max_leaf_nodes=opt_nodes, max_depth=opt_depth)
pipeline = Pipeline(steps = [('preprocessor', preprocess), ('model', final_model)])
pipeline.fit(x_train, y_train.values.ravel())
y_predict = pipeline.predict(x_test)

correct = np.array(y_test == y_predict).sum()/y_test.count()
print('The accuracy of prediction for spending habits is {:.2f}%'.format(correct*100))
```

✓ 9.2s

```
The accuracy of prediction for spending habits is 74.39%
```

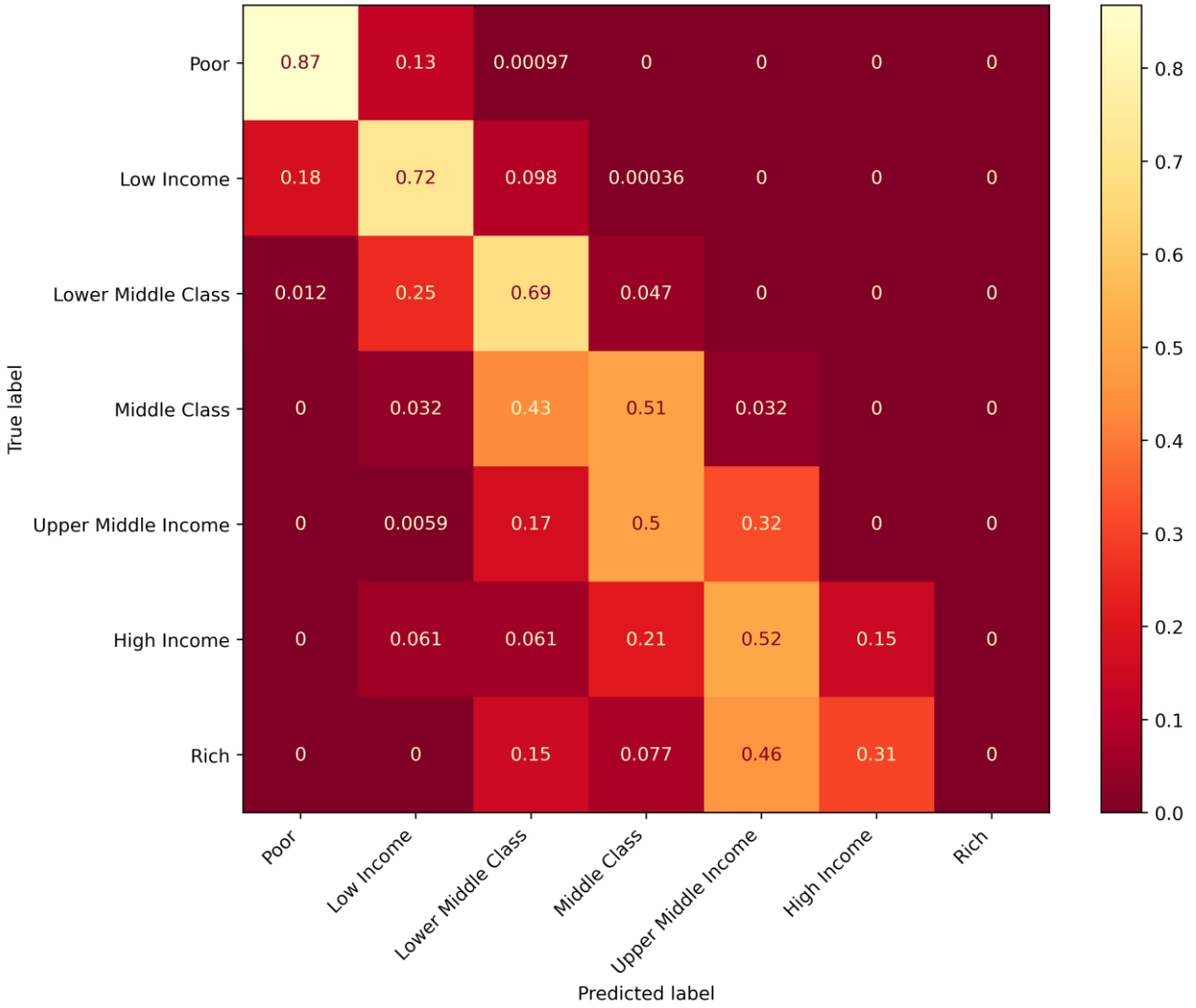
Prediction Model

The model fails to classify higher classes accurately!



Prediction Model

The model fails to classify higher classes accurately!



Synthesis

- How does income inequality affect the spending behavior of different income classes?
  - Lower income classes have more constrained budget, forcing them to prioritize one expenditure more than the other.
  - Higher income classes have more freedom in budgeting.
- This implies that expenditure proportions are sufficient determinants on the key areas to focus on in addressing the effects of income inequality on lower income classes.

Bread and Cereals Expenditure	20.369395
Total Rice Expenditure	14.378163
Meat Expenditure	4.232902
Total Fish and marine products Expenditure	7.620689
Fruit Expenditure	1.559692
Vegetables Expenditure	3.954199
Restaurant and hotels Expenditure	4.834301
Alcoholic Beverages Expenditure	0.717937
Tobacco Expenditure	1.715629
Clothing, Footwear and Other Wear Expenditure	1.961966
Housing and water Expenditure	15.861505
Medical Care Expenditure	1.847219
Transportation Expenditure	3.584264
Communication Expenditure	0.744097
Education Expenditure	1.257522
Miscellaneous Goods and Services Expenditure	4.270640
Special Occasions Expenditure	1.402415
Crop Farming and Gardening expenses	9.687467

Recommendation

Explore behavioral patterns encoded within FIES that determines higher income classes.

End

Region?

Accuracy: 73.93%

Different Training Data Size