Neural FCA On Income Dataset

Muhammad Asad

Dataset

The goal is to predict whether an individual earns more than $50K based on various demographic and employment-related feature

Dataset Description

- Age: Age of the individual.
- Workclass: Employment sector of the individual.
- Fnlwgt: Final weight representing the number of people in the population that the sample represents.
- Education: Education level attained by the individual.
- Educational-num: Numeric representation of the education level.
- Marital-status: Marital status of the individual.
- Occupation: Type of work performed by the individual.
- Relationship: Family relationship role.

- Race: Race of the individual.
- Gender: Gender of the individual.
- Capital-gain: Income from investment sources (excluding salary).
- Capital-loss: Loss from investment sources.
- Hours-per-week: Number of hours worked per week.
- Native-country: Country of origin.
- Income: Income class ($50K or less, $50K or more). Y variable

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relationship	race	gender	hours-per-week	native-country
0	25	Private	226802	11th	7	Never-married	Machine-op-inspct	Own-child	Black	Male	40	United-States
1	38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	Husband	White	Male	50	United-States
2	28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	Husband	White	Male	40	United-States
3	44	Private	160323	Some-college	10	Married-civ-spouse	Machine-op-inspct	Husband	Black	Male	40	United-States
5	34	Private	198693	10th	6	Never-married	Other-service	Not-in-family	White	Male	30	United-States
...	...	...	...	...	...	...	...	...	...	...	...	...
48837	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	Wife	White	Female	38	United-States
48838	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	40	United-States
48839	58	Private	151910	HS-grad	9	Widowed	Adm-clerical	Unmarried	White	Female	40	United-States
48840	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-child	White	Male	20	United-States
48841	52	Self-emp-inc	287927	HS-grad	9	Married-civ-spouse	Exec-managerial	Wife	White	Female	40	United-States

Binarization

Age Group Ordinal

$$\text{Age Groups} = \begin{cases} \text{under 18} & \text{if age} < 18 \\ \text{18-24} & \text{if } 18 \leq \text{age} < 25 \\ \text{25-34} & \text{if } 25 \leq \text{age} < 35 \\ \text{35-44} & \text{if } 35 \leq \text{age} < 45 \\ \text{45-54} & \text{if } 45 \leq \text{age} < 55 \\ \text{55-64} & \text{if } 55 \leq \text{age} < 65 \\ \text{65+} & \text{if age} \geq 65 \end{cases}$$

2

Work Class and Final weight Column

- Having 7 unique value in workclass column.
- Apply Nominal Scale
- Creating binary variables for each unique category in the workclass column.
- Created Nominal Scaling for work class
- Created Interordinal Scaling for Final Weight Column

$$\text{fnlwgt\_quartiles} = \begin{cases} \text{Q1} & \text{if fnlwgt is in the first quartile} \\ \text{Q2} & \text{if fnlwgt is in the second quartile} \\ \text{Q3} & \text{if fnlwgt is in the third quartile} \\ \text{Q4} & \text{if fnlwgt is in the fourth quartile} \end{cases}$$

Education Level and Education Number Binarization

- Having 16 unique values in Education Level so did nominal scaling for education level
- Each column indicates whether an individual holds a specific education level
- Ordinal Scale for educational Stages

$$\text{Education Stages} = \begin{cases} \text{Less than High School} & \text{if educational-num} < 9 \\ \text{High School Graduate} & \text{if educational-num} = 9 \\ \text{Some College} & \text{if educational-num} \in \{10, 11\} \\ \text{Bachelors} & \text{if educational-num} = 13 \\ \text{Advanced Degree} & \text{if educational-num} > 13 \end{cases}$$

We then created binary variables for each of these educational stages.

Marital Status , Occupation And Relationship Column (Nominal)

- Marital Status having 7 unique value
- Created a binary column for each marital status category
- Same goes for Occupation, having 14 unique occupations in the data which is categorical
- Created a binary variable for each occupation category
- The relationship Column have 6 unique values
- Binarized by creating a binary variable for each unique relationship status

Race, Gender and Hours Worked per Week Column

- Race Column have 5 unique value so binarized by nominal scale
- Gender have 2 unique values in dataset so binarized by Dichotomous
- The feature hours-per-week was categorized into three groups based on work hours applying ordinal scale

$$\text{Hours Categories} = \begin{cases} \text{part-time} & \text{if hours} < 35 \\ \text{full-time} & \text{if } 35 \leq \text{hours} \leq 40 \\ \text{over-time} & \text{if hours} > 40 \end{cases}$$

For each category, a binary variable was created indicating the membership of an individual in that category.

Native Country Column

- For the native-country feature, we first grouped countries into regions
- Native Country have 41 unique values
- Created Region for each Country
- Applying Nominal Scale

$$\text{native\_country\_region} = \begin{cases} 1 & \text{if native country is in a specific region} \\ 0 & \text{otherwise} \end{cases}$$

5

Results on Different Classification models

Model	Accuracy	ROC AUC	F1 Score (Class 1)
Random Forest	0.8173	0.8655	0.5984
Logistic Regression	0.8277	0.8823	0.6115
SVM	0.8284	0.8633	0.6017
K-Nearest Neighbors	0.8050	0.8237	0.5899
Gradient Boosting	0.8350	0.8881	0.6271

Table 1: Performance of different models on the Adult Income dataset.

Feature Importance for Standard Models

	Feature	Importance
26	marital-status_Married-civ-spouse	0.026784
0	age	0.014410
2	educational-num	0.009976
3	hours-per-week	0.008271
28	marital-status_Never-married	0.003365
7	workclass_Self-emp-not-inc	0.002147
33	occupation_Exec-managerial	0.001815
37	occupation_Other-service	0.001778
35	occupation_Handlers-cleaners	0.001358
34	occupation_Farming-fishing	0.001107
43	occupation_Transport-moving	0.000981
23	education_Prof-school	0.000915
44	relationship_Not-in-family	0.000767
48	relationship_Wife	0.000760
36	occupation_Machine-op-inspct	0.000583
46	relationship_Own-child	0.000465
19	education_Doctorate	0.000332
1	fnlwgt	0.000118
57	native-country_Cuba	0.000111
45	relationship_Other-relative	0.000103

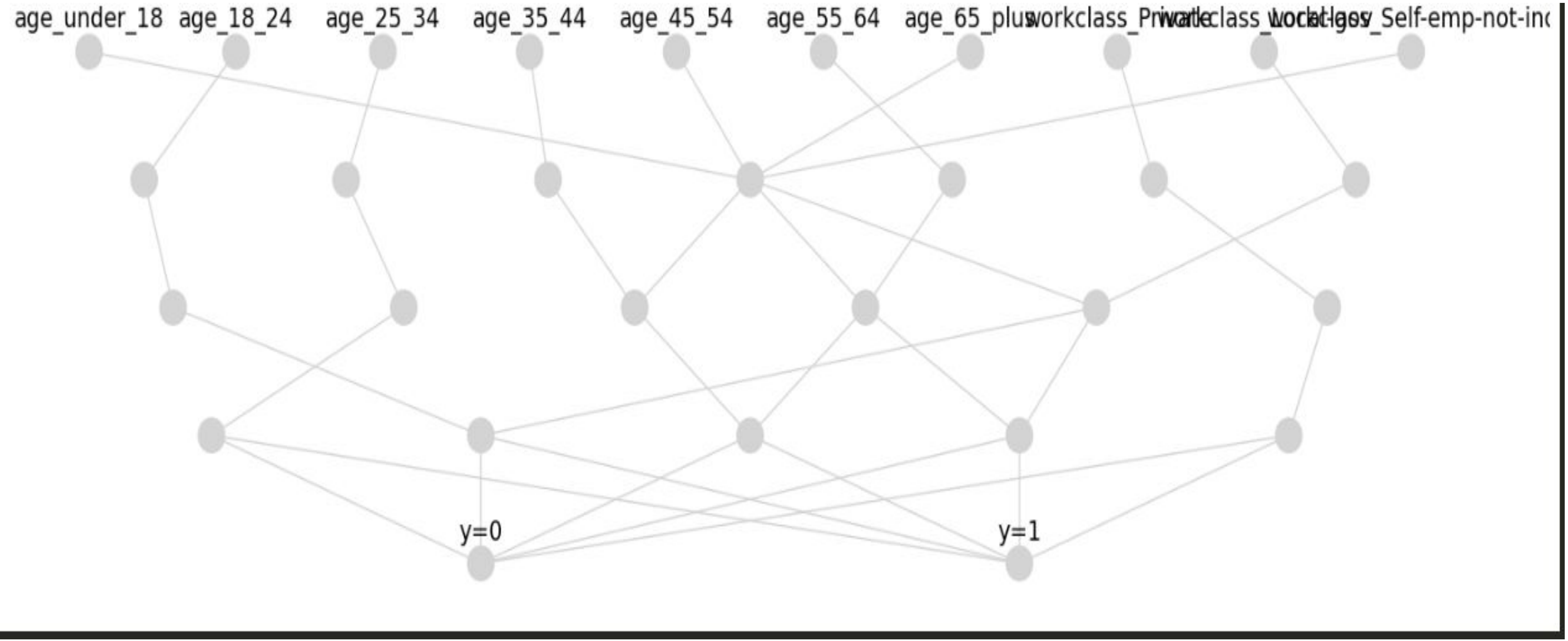
Neural FCA

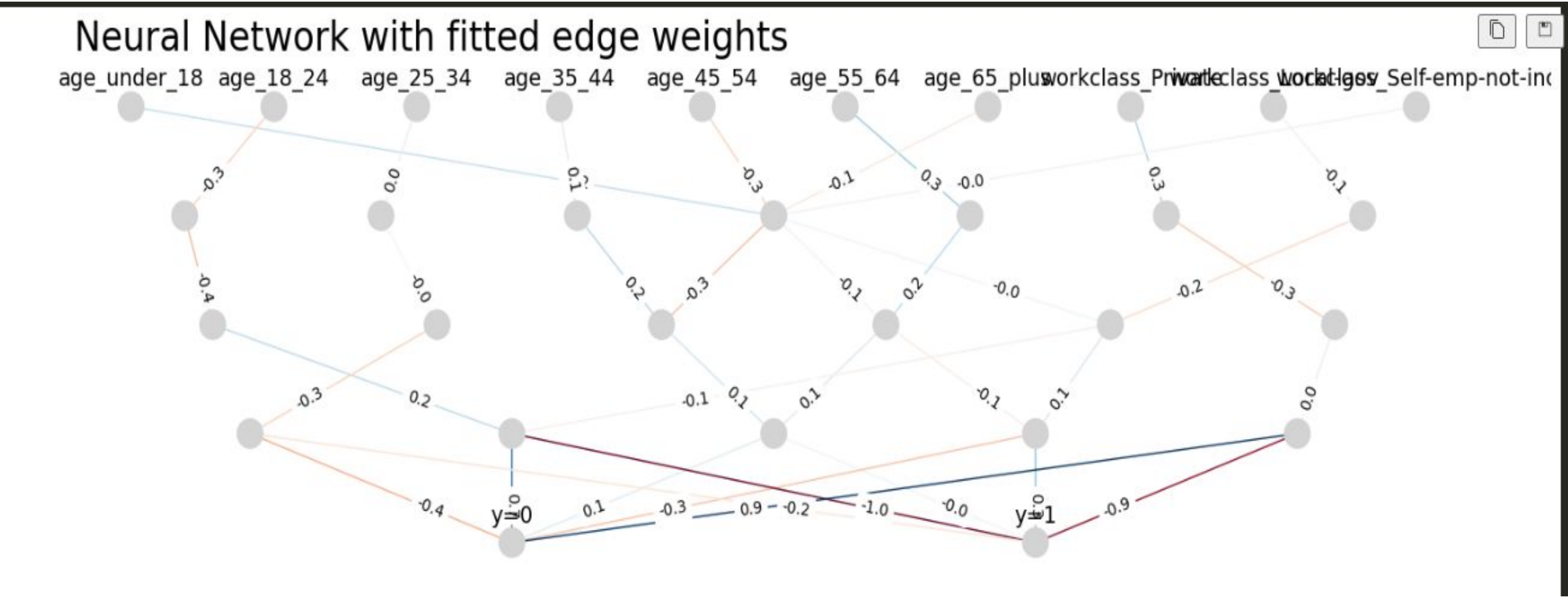
	age_under_18	age_18_24	age_25_34	age_35_44	age_45_54	age_55_64	age_65_plus	workclass_Private	workclass_Local-gov	workclass_Federal-gov	...	race_Black	race_Other
0	False	False	True	False	False	False	False	True	False	False	...	True	False
1	False	False	False	True	False	False	False	True	False	False	...	False	False
2	False	False	True	False	False	False	False	False	True	False	...	False	False
3	False	False	False	False	True	False	False	True	False	False	...	True	False
5	False	False	False	True	False	False	False	True	False	False	...	False	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...
3837	False	False	True	False	False	False	False	True	False	False	...	False	False
3838	False	False	False	True	False	False	False	True	False	False	...	False	False
3839	False	False	False	False	False	True	False	True	False	False	...	False	False
3840	False	True	False	False	False	False	False	True	False	False	...	False	False
3841	False	False	False	False	True	False	False	False	False	False	...	False	False

Model	Accuracy	F1-Score	Precision	Recall
Random Forest	0.8173	0.8655	0.5984	-
CBO	0.7700	0.7900	0.6500	0.7200

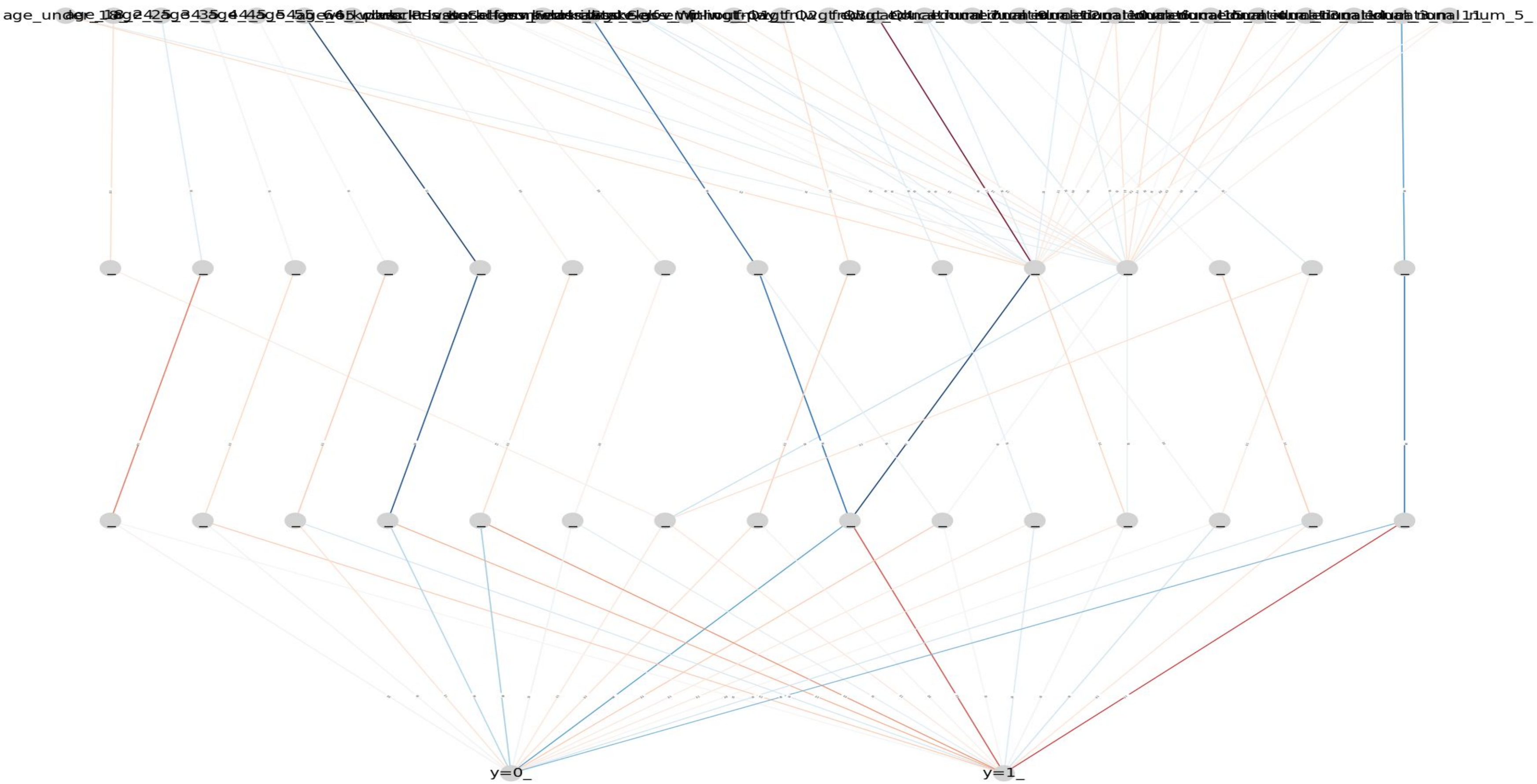
Table 2: Classification Results on Income Dataset using Neural FCA

Neural FCA


age_under_18
age_18_24
age_25_34
age_35_44
age_45_54
age_55_64
age_65_plus
y=0
y=1


Neural Network with fitted edge weights
age_under_18
age_18_24
age_25_34
age_35_44
age_45_54
age_55_64
y=0
y=1

Neural Network with fitted edge weights



Visualization: This lattice can be visualized to show the relationships between different concepts learned by the neural network, making the model's reasoning more transparent.

Reasoning:

By reading the network from top to bottom, you can form a narrative about the model's reasoning:

- Which input features (like “age_35_44” or “workclass_Private”) strongly influence the hidden nodes?
- How do those hidden nodes combine and interact to reinforce or counteract each other?
- Which sets of intermediate concepts correlate strongly with the final classification?

Improvements on Faster Calculation

```python
# Parallelize the process using joblib's Parallel
results = Parallel(n_jobs=n_jobs)(
    delayed(process_data_and_compute_f1)(X_sample, y_sample, sample_size, n_columns) for _ in range(n_jobs)
)

# Process the results
f1_scores = [result[0] for result in results]
best_concepts = [result[1] for result in results]
test_f1_scores = [result[2] for result in results]
y_preds_all = [result[3] for result in results]
y_probs_all = [result[4] for result in results]

# Display the best F1 scores for each parallel job
print(f"Best F1 Scores on Train: {f1_scores}")
print(f"Best Concepts: {best_concepts}")
print(f"Test F1 Scores: {test_f1_scores}")
print(f"Predictions: {y_preds_all}")
print(f"Prediction Probabilities: {y_probs_all}")
```