Predicting Income with Lazy Formal Concept Analysis

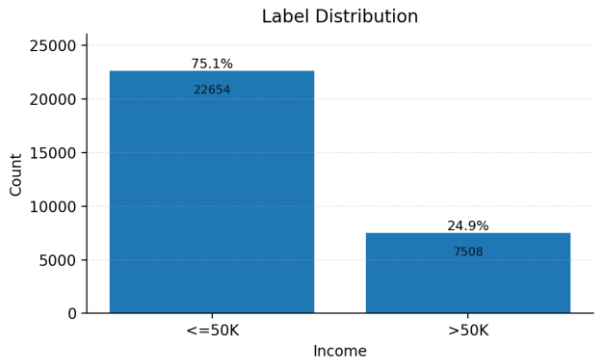
Presenter: XIA Yuqi

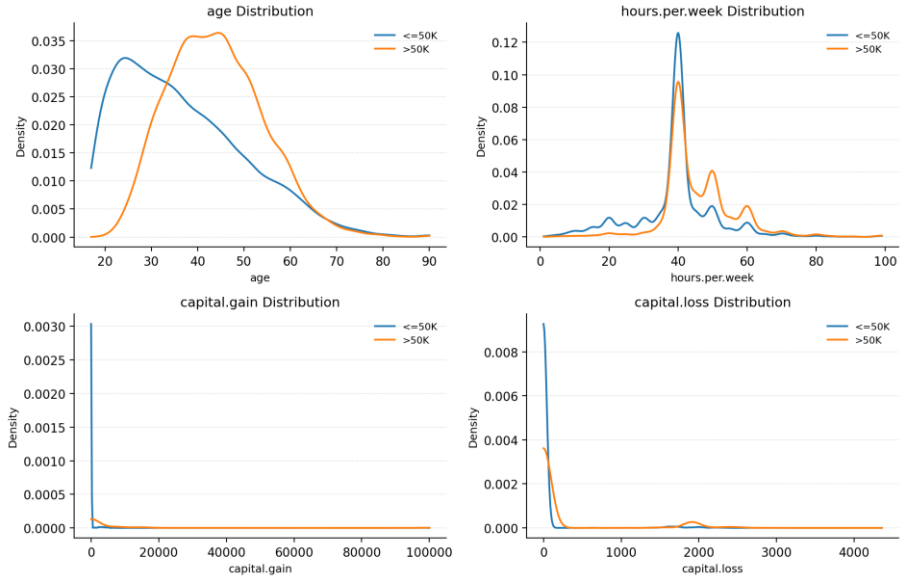
Date: December 2, 2025

Problem, Dataset & Exploratory Data Analysis

Task & Dataset:

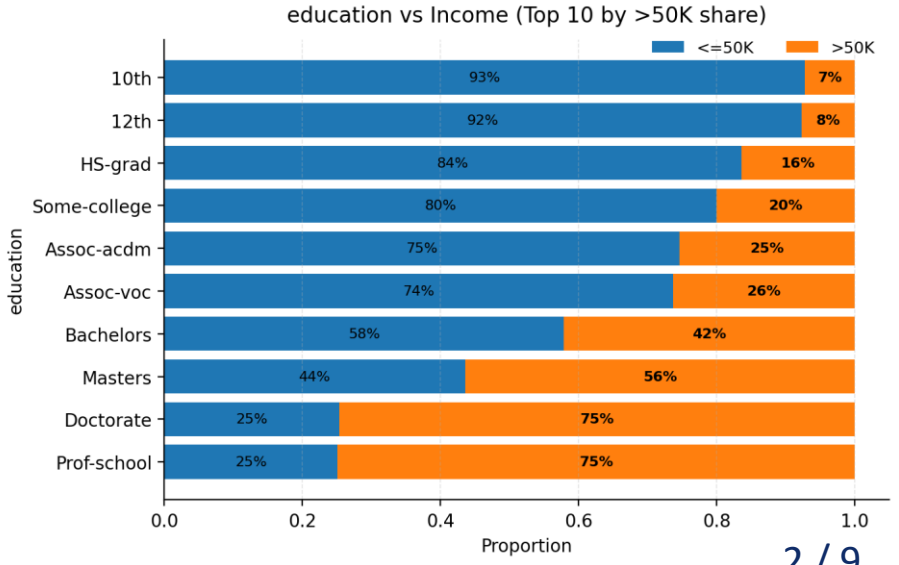
- Predict income >$50K/yr on the **Adult Income Dataset** (Kaggle).
- **Stratified subsample:** N = 1000 (train 800/ test 200).
- **Class imbalance:** ≤50K ≈ 75% vs.>50K ≈ 25%.





Key EDA Findings:

- **Age & Hours:** Higher values correlate with high income.
- **Capital Gain:** Long-tail distribution is a strong signal for > 50K .
- **Education:** Higher levels show markedly higher > 50K share.



Method: Optimized Lazy FCA

Baseline Lazy FCA:

- **Binarization:** Numerical features quartile-binned; categorical features one-hot encoded.
- **Final formal context:** ≈ 102 binary attributes.
- **Prediction:** Compare positive vs negative evidence from train intersections.
- **Issue:** Sparse context + imbalance lowers minority recall.
- **F1 = 0.43, acc = 0.775**

Four Key Optimizations:

- **Counterexamples (x = 2):** Allows near-exclusive evidence, improving > 50K recall.
- **Min Cardinality (= 1):** Keeps useful intersections while filtering empty ones.
- **Threshold (τ = 0.40):** Rebalances decisions under class imbalance.
- **Top-k Evidence (k = 30):** Uses strongest evidence only; denoises and speeds up inference

Parameter Search

Finding the Best Parameter:

- Joint search over (x, min_cardinality) selects the best evidence robustness.
- **Best: x = 2, min_cardinality = 1.**

- Allowing a small number of counterexamples is beneficial, but too many counterexamples will weaken the discriminative power of the evidence.
- min_cardinality has no meaningful impact on performance here, so it can simply be set to the smallest value.

	x	min_cardinality	f1	accuracy
6	2	1	0.485714	0.820
7	2	2	0.485714	0.820
8	2	3	0.485714	0.820
3	1	1	0.450704	0.805
4	1	2	0.450704	0.805
5	1	3	0.450704	0.805
9	5	1	0.437500	0.820
10	5	2	0.437500	0.820
11	5	3	0.437500	0.820
0	0	1	0.430380	0.775
1	0	2	0.430380	0.775
2	0	3	0.430380	0.775
12	10	1	0.333333	0.800
13	10	2	0.333333	0.800
14	10	3	0.333333	0.800

Parameter Search

Finding the Best Parameter:

- Threshold search over tau rebalances minority-class decisions.
- **Best: τ = 0.40.**

	tau	f1	accuracy
0	0.40	0.533333	0.825
1	0.45	0.527778	0.830
2	0.50	0.485714	0.820
3	0.55	0.447761	0.815
4	0.60	0.375000	0.800
5	0.65	0.327869	0.795

- τ is the threshold that determines how strong positive evidence must be to predict the positive class (>50K).
- Because the Adult dataset is imbalanced with fewer positive samples, increasing tau (0.45→0.65) makes the model more conservative and less willing to predict positives, reducing positive-class recall, so F1 steadily drops.
- Therefore, in an imbalanced setting, a lower tau better protects the minority class, and tau = 0.40 provides the best balance here.

Parameter Search

Finding the Best Parameter:

- Top-k search over k finds the optimal evidence size.
- **Best: k = 30.**

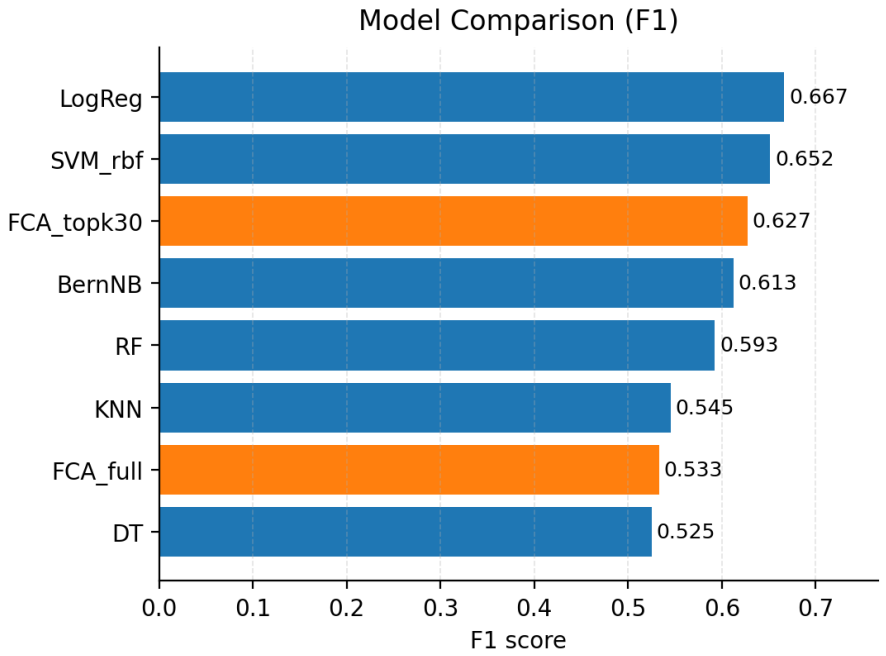
- Best k = 30, giving the highest F1 (0.627) with solid accuracy (0.810).
- When k is too small (e.g., 10 or 20), F1 is lower because there is not enough evidence to support reliable positive predictions, and accuracy can also suffer (notably at k=10).
- When k becomes too large (50, 80, 120, 160, 220), F1 gradually declines even though accuracy stays high. This indicates that adding many weaker intersections introduces noisy or less discriminative evidence, which hurts minority-class performance.

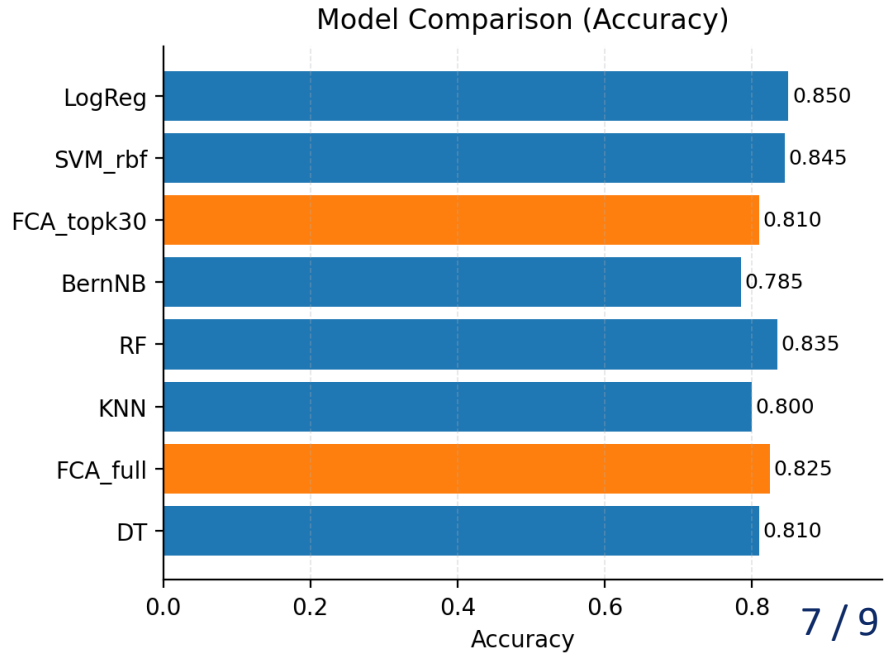
	k	f1	accuracy
2	30	0.627451	0.810
1	20	0.614035	0.780
4	80	0.606742	0.825
3	50	0.589474	0.805
5	120	0.588235	0.825
0	10	0.567376	0.695
7	220	0.556962	0.825
6	160	0.543210	0.815

Results Analysis

FCA Performance (updated):

- **FCA_full:** F1 = 0.533, Accuracy = 0.825.
- **FCA_topk30:** F1 = 0.627, Accuracy = 0.810.
- **Improvement:** +0.094 absolute F1 points (about +17.6% relative).
- **Speedup:** ≈ 9.6× faster inference than FCA full.





Interpretability & Conclusion

Interpretability Example (True Positive > 50K):

- Decision based on evidence ratio with **τ = 0.40**.
- **Strongest positive evidence pattern:** {Married-civ-spouse, Exec-managerial, White, Male}.

Conclusions:

- Optimizations (x = 2, min_cardinality = 1, τ = 0.40, k = 30) boost FCA robustness on imbalanced data.
- **FCA_topk30** is competitive with tuned ML baselines (F1 close to best models).
- Provides a strong combination **of performance, speed, and transparent evidence**.

Final Thought:

- Lazy FCA offers a transparent alternative where explaining why matters as much as predicting.

Thank you for your attention!

Спасибо за внимание!

Presenter: XIA Yuqi

Date: December 2, 2025