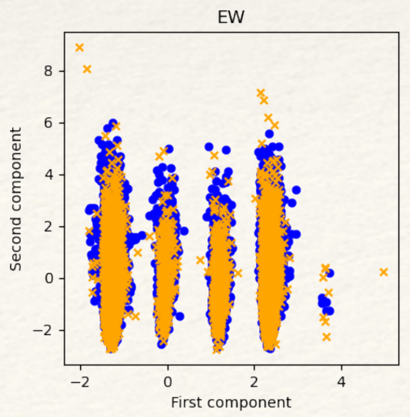
Bridging Accuracy and Interpretability:

A Rescaled Cluster-then-Predict Approach for Enhanced Credit Scoring

DIF
National Yang Ming Chiao Tung University
https://hackmd.io/@hwteng/HyKOPoA6d

Huei-Wen Teng, Ming-Hsuan Kang, I-Han Lee, Le-Chi Bai

20231022 Ten Kan Lee Bai rescaled clusetering.key



Motivations

- Machine learning is not utilized in practice (Son et al, 2019)
- Accuracy improvement not significant
- Results can not be interpretable

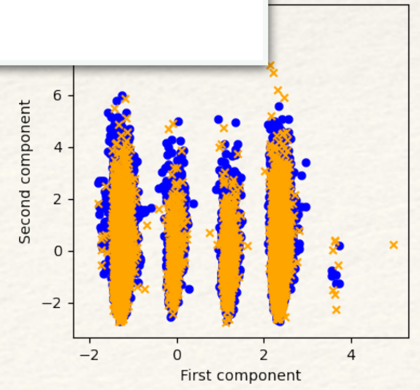
GDPR's Right to Explanation: the pros and the cons

Under the General Data Protection Regulation's Right to Explanation a user can ask for an explanation about an algorithmic decision made about them - but with it comes positives and negatives.

Written by Bill Brenner

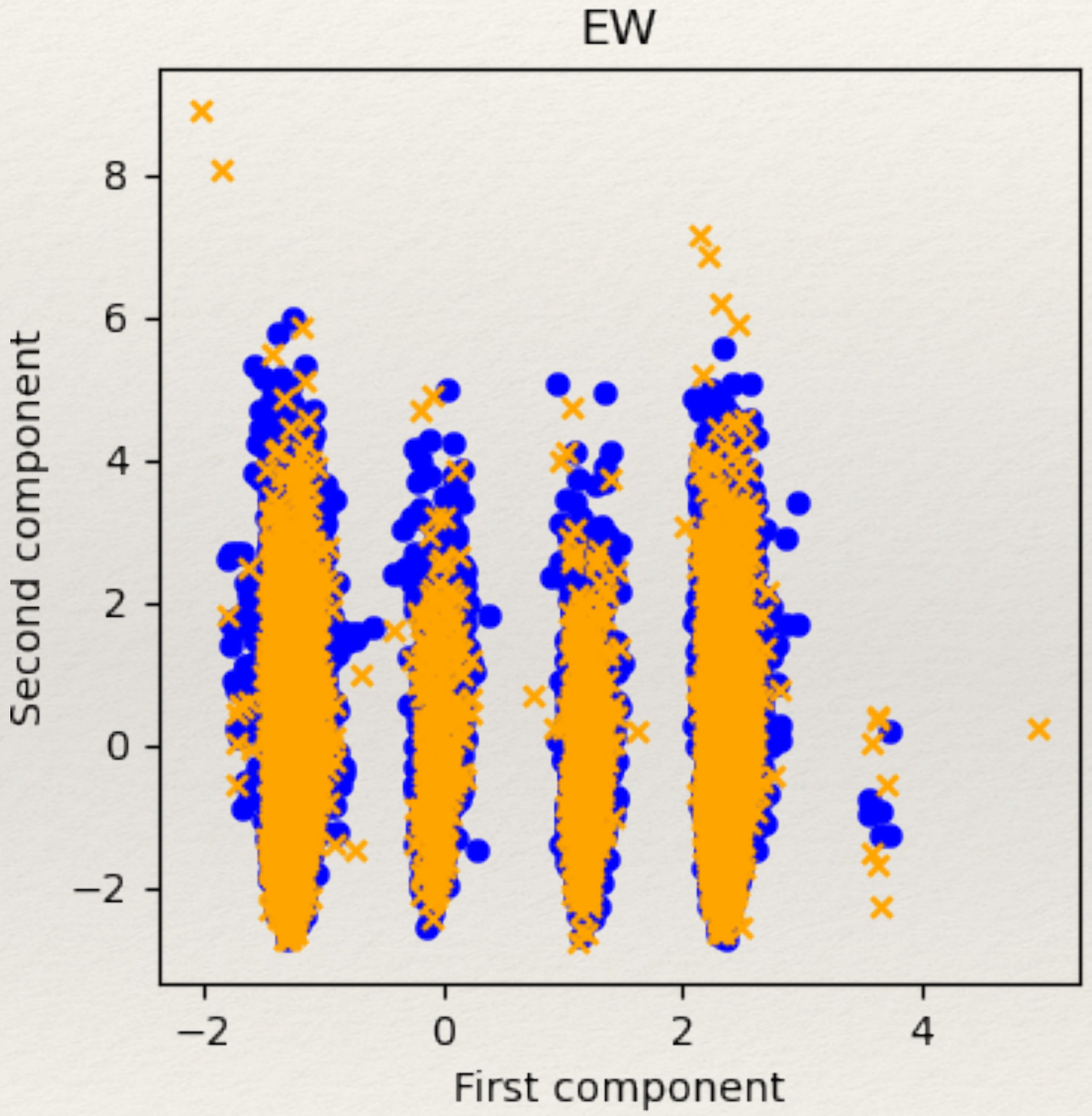
MAY 22, 2017

https://news.sophos.com/en-us/2017/05/22/gdprs-right-to-explanation-the-pros-and-the-cons/

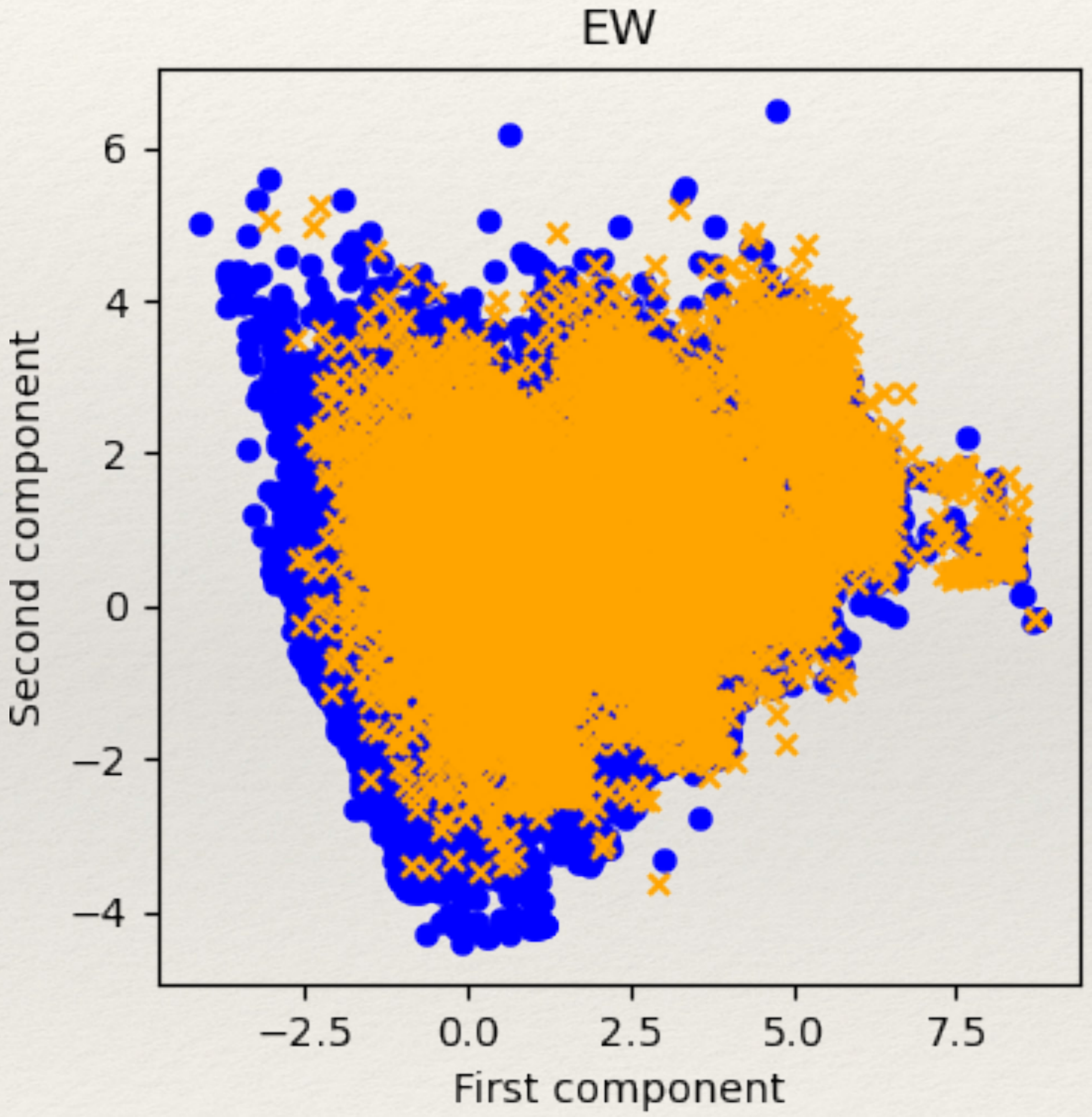


Clusters in the data

PAK: 26%



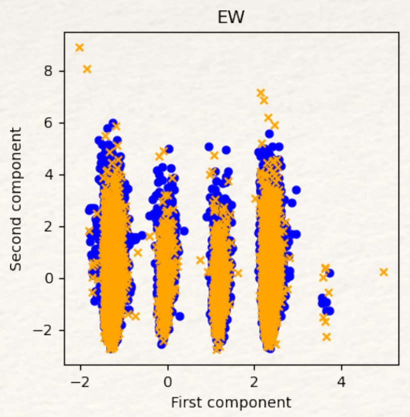
GMC: 6.7%



Rescaled clustering

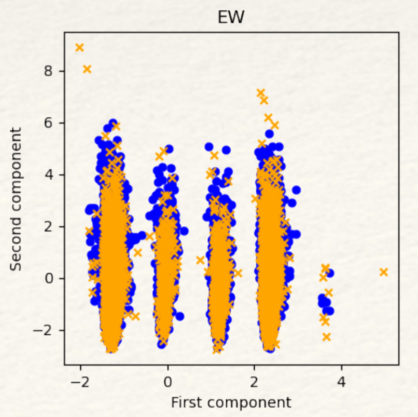
Outline

1. Motivations
2. A rescaled cluster-then-predict approach
3. The PAK dataset
4. The GMC dataset
5. Conclusion



Data preprocess

1. Logarithm Transformation
2. Z-score scaling
3. Missing value imputation



Missing data imputation

- A simple model: $y = \beta_0 + \beta_1 x + \varepsilon$ where some x's are missing.
- Define d and $\tilde{x}$:

$$d_i = \begin{cases} 1, & x_i \text{ missing} \\ 0, & x_i \text{ complete} \end{cases} \qquad \text{and} \qquad \tilde{x}_i = \begin{cases} 0, & x_i \text{ missing} \\ x, & x_i \text{ complete} \end{cases}$$

- Transformed model: $y = \beta_0 + \beta_1 \tilde{x} + \gamma d + \varepsilon$

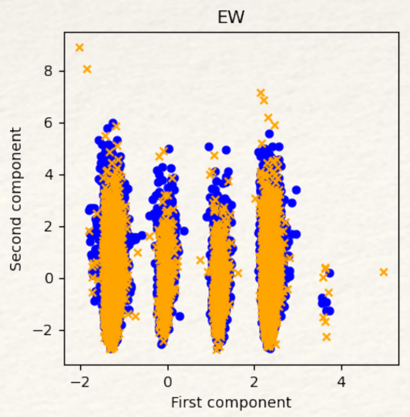
$$y = \begin{cases} \boxed{(\beta_0 + \gamma) + \varepsilon, \quad x_i \text{ missing}} \leftarrow \text{Irrelevant to } x! \\ \boxed{\beta_0 + \beta_1 x + \varepsilon, \quad x_i \text{ complete}} \leftarrow \text{The original model!} \end{cases}$$

The rescaling methods

- The ith feature: $x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \dots, x_p^{(i)})$ and the target $y^{(i)}$ for $i = 1, \dots, n$
- Introduce a rescaor vector, $w = (w_1, \dots, w_p)$, to rescale the feature

$$\tilde{x}_j^{(i)} = w_j x_j^{(i)}$$

1. The Equal Weight (EW) rescalor sets $w = (1, \dots, 1)$



Rescaled clustering

2. The regression (REG) rescalor:

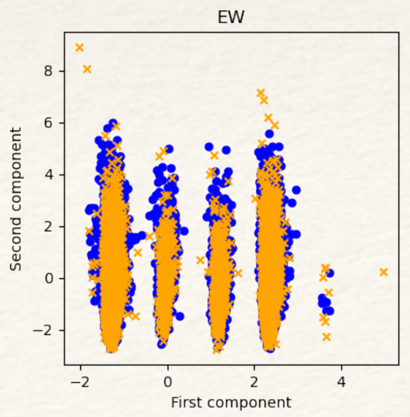
$$y^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \beta_2 x_2^{(i)} + \cdots + \beta_p x_p^{(i)} + \varepsilon_i, \quad \varepsilon_i \sim N(0,\sigma^2)$$

sets w as the estimated value of $\beta = (\beta_1, \ldots, \beta_p)$

3. The logistic regression (LR) rescalor

$$y^{(i)} \,|\, x^{(i)} \sim Bernoulli\left(\frac{1}{1 + e^{-\left(\beta_0 + \beta_1 x_1^{(i)} + \beta_2 x_2^{(i)} \cdots + \beta_p x_p^{(i)}\right)}}\right),$$

The LR rescalor set w as the estimated β.

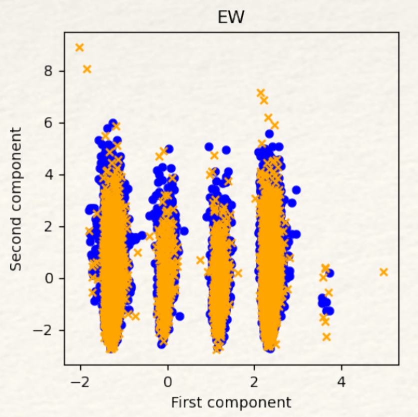


Rescaled clustering

4. Let $P_{x_j}(a)$ be sample marginal probability of x_j taking on the value a. Define $P_y(b)$ and $P_{x_j,y}(a, b)$ similarly.

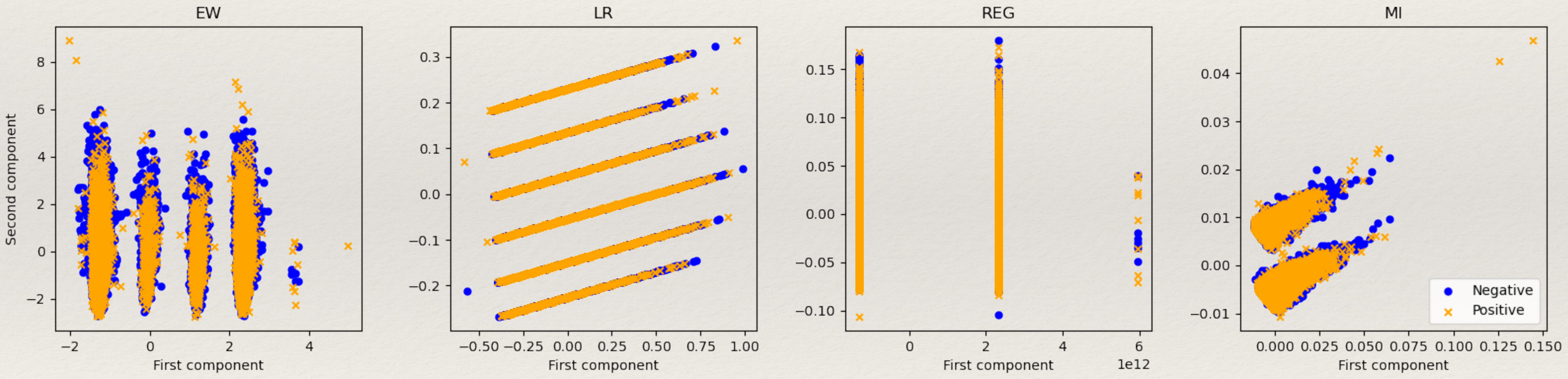
The Multual Information (MI) rescalor

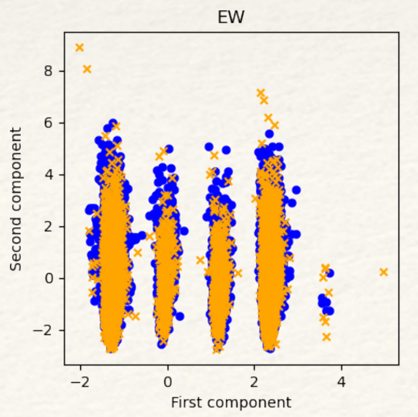
$$MI(x_j, y) = \sum_{(a,b)} P_{x_j,y}(a, b) \log \frac{P_{x_j,y}(a, b)}{P_{x_j}(a) P_y(b)}$$



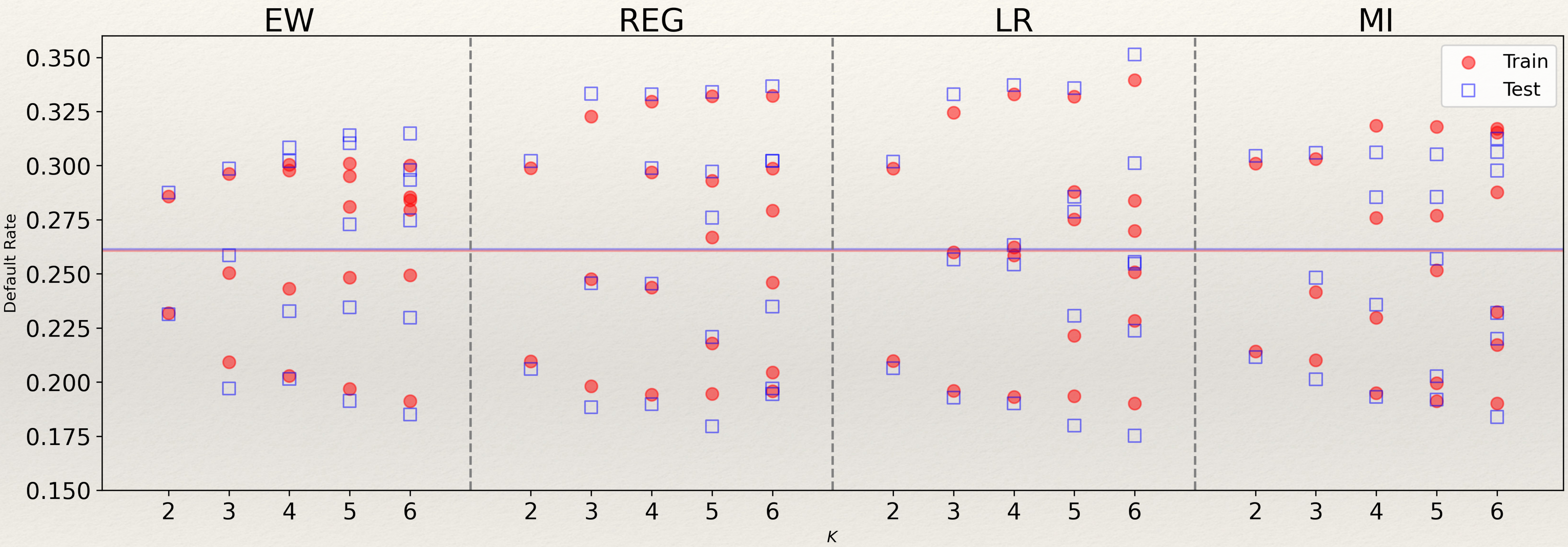
The PAK dataset

- The overall default rate 26%, 50,000 cases, $p = 12$.

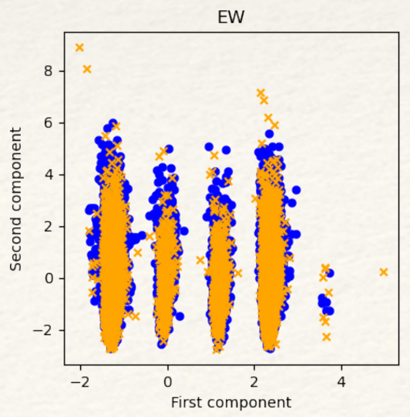




Default rates

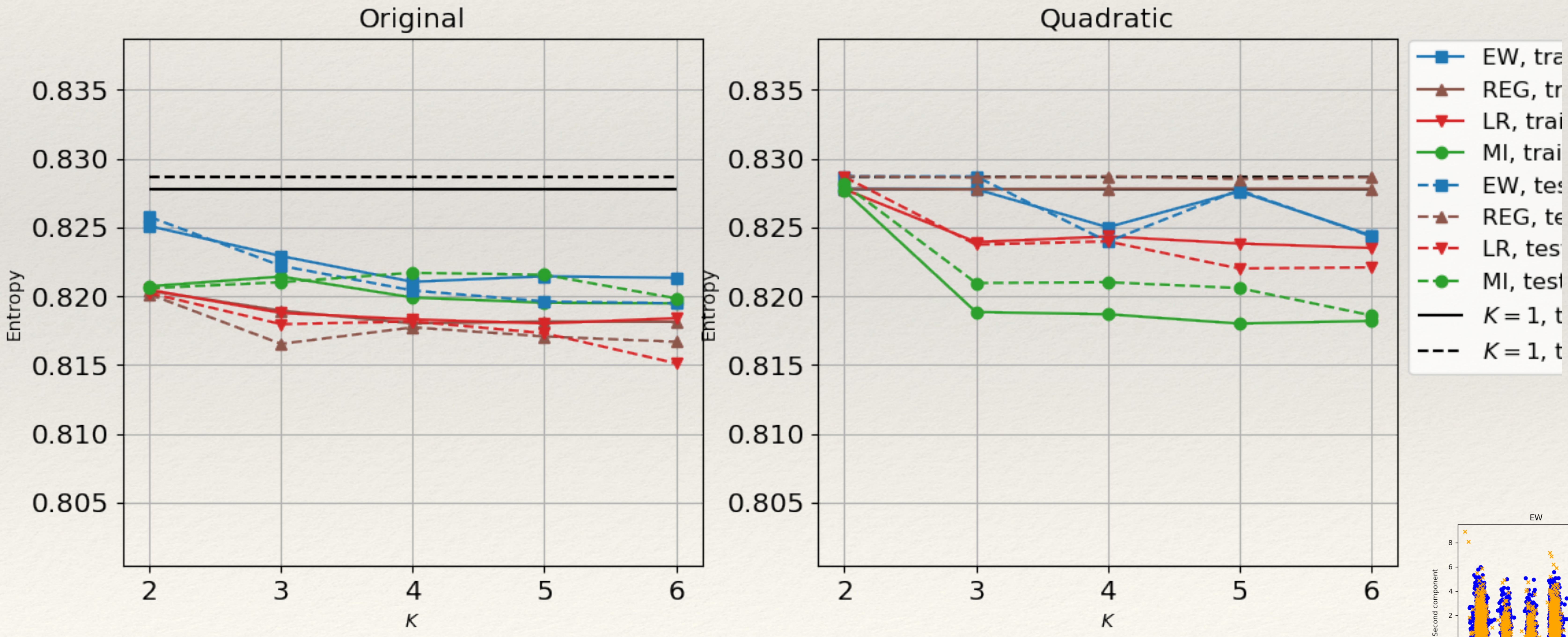


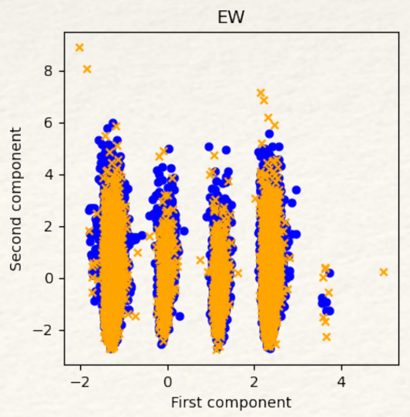
Rescaled clustering



Entropy

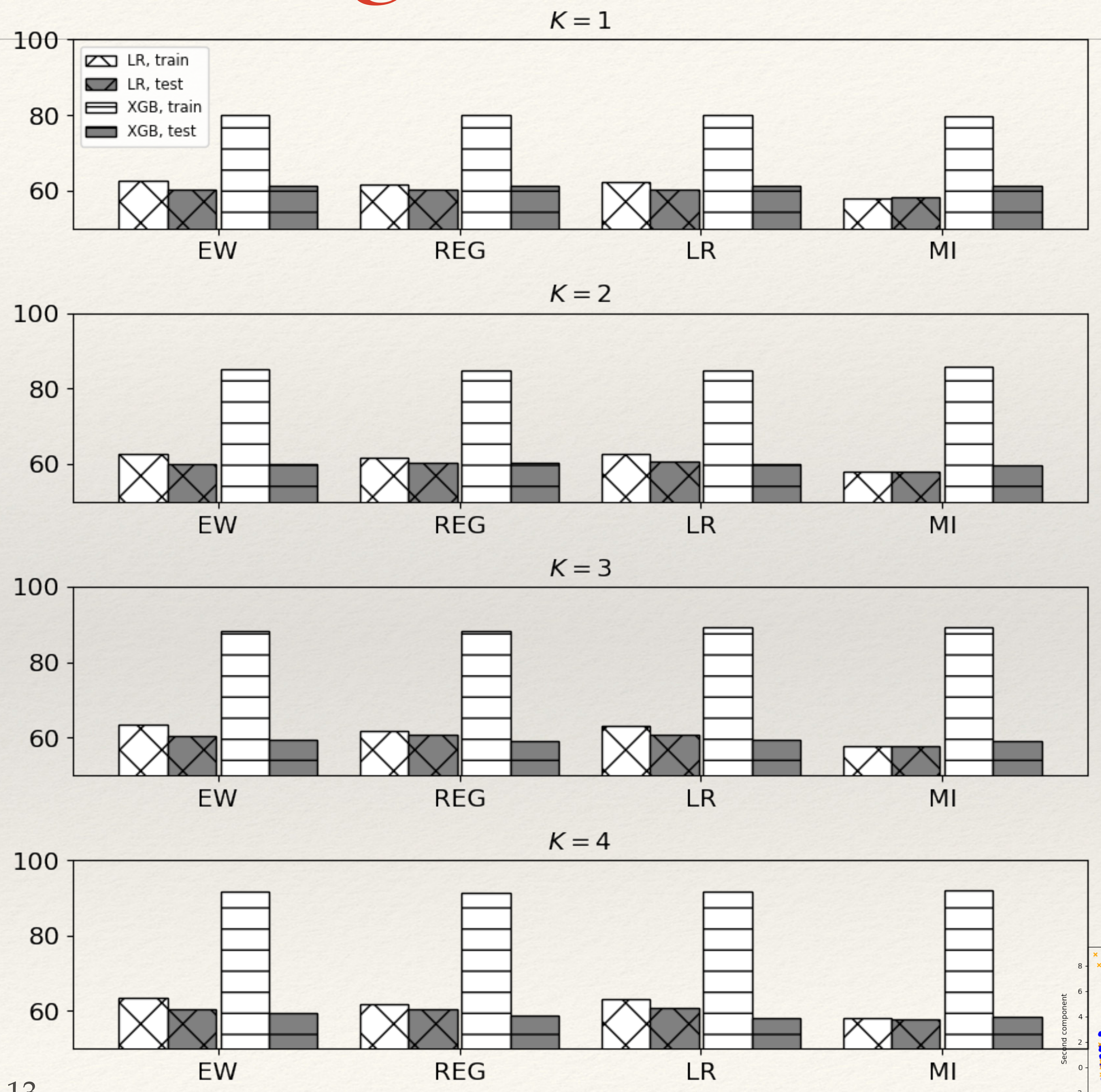
- Entropy: $h_i = -\left(p_i \log(p_i) + (1-p_i)\log(1-p_i)\right)$, p_i the default rate for the ith cluster
- Total Entropy: $h = \sum_{i=1}^{K} q_i h_i$, q_i the fraction of points in the ith cluster.
- Rescaled clustering reduces randomness!

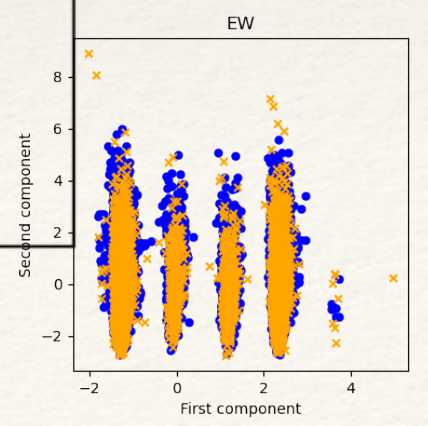




AUC in the training set

- The original feature set
- XGBoost: overfitting
- Logistic Regression: robust





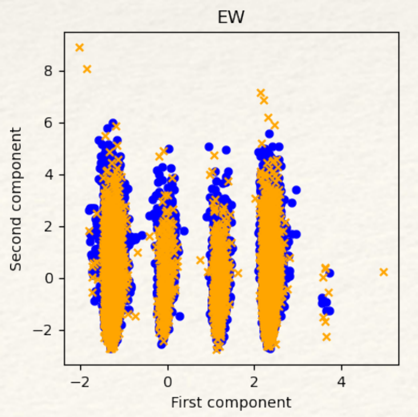
Rescaled clustering

AUC in the testing set

- XGBoost ~> Logistic Regression
- XGBoost not improved
- Logistic Regression > XGBoost
    - Quadratic feature set
    - *K* increases

	XGBoost				Logistic Regression			
K	EW	REG	LR	MI	EW	REG	LR	MI
The original feature set								
1	60.59	61.20	61.31	**61.32**	60.16	58.22	60.14	60.19
2	60.59	59.57	60.07	60.15	60.16	57.90	60.56	60.29
3	60.59	59.16	59.40	59.19	60.16	57.81	60.86	60.75
4	60.59	58.63	57.98	58.76	60.16	57.84	60.96	60.32
The quadratic feature set								
1	59.94	59.94	59.94	60.38	61.45*	**62.00***	61.99*	59.96
2	59.07		59.07	59.25	60.60		61.70*	59.69
3	58.81		58.57	59.03	60.01		61.93*	59.49
4	58.50		58.30	58.43	58.96		61.81*	59.09

Rescaled clustering

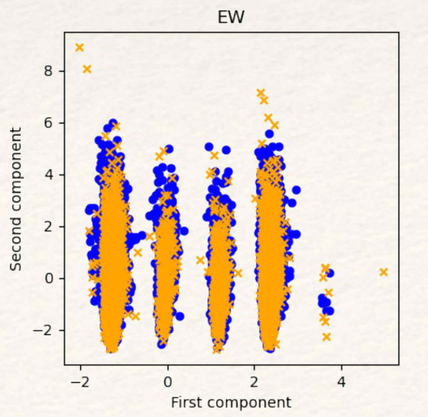


Positive-case clustering

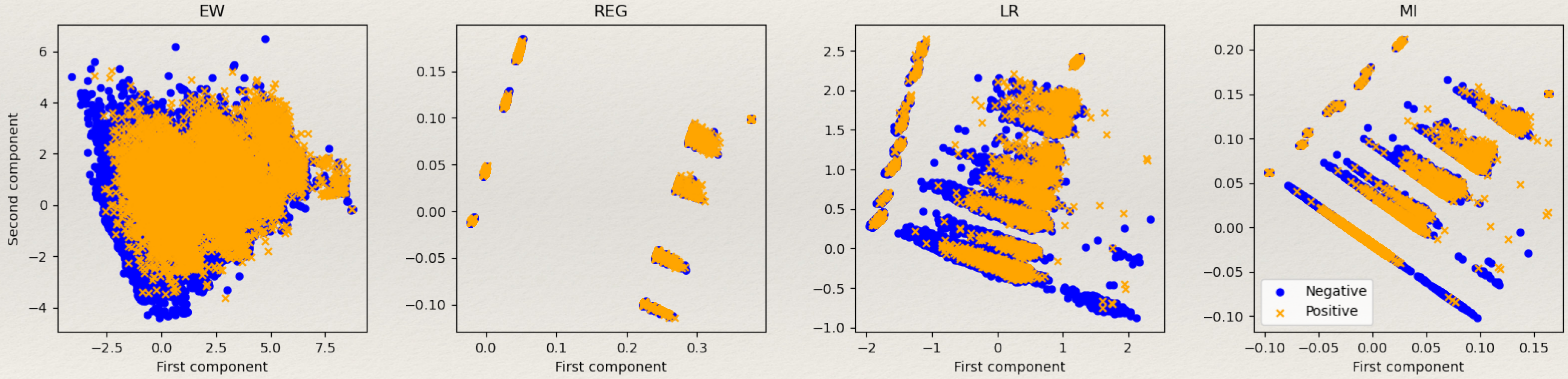
- Clustering only postive cases
- Comparative to clustering all cases
- Same conclusion
  - XGBoost does not benifit
  - Logistic Regression is improved

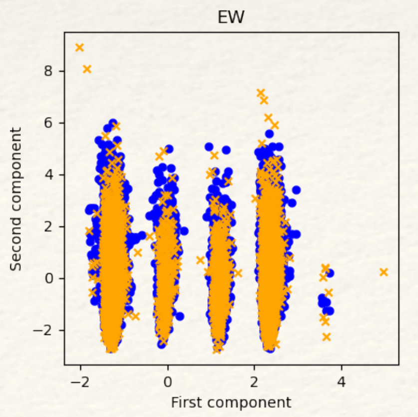
	XGBoost				Logistic Regression			
K	EW	REG	LR	MI	EW	REG	LR	MI
The original feature set								
1	60.59	61.20	61.31	**61.32**	60.16	58.22	60.14	60.19
2	60.59	59.33	59.48	60.00	60.16	58.26	60.56	60.22
3	60.59	59.90	59.24	59.78	60.16	57.32	60.50	60.16
4	60.59	59.18	58.81	59.01	60.16	57.79	61.07	60.44
The quadratic feature set								
1	59.94	59.94	59.94	60.38	61.45*	62.00*	**61.99***	59.96
2	58.84		59.07	59.33	60.08		61.78*	59.77
3	59.11		58.04	58.77	59.58		61.89*	59.34
4	57.95		58.00	57.55	59.42		61.77*	58.79

Rescaled clustering

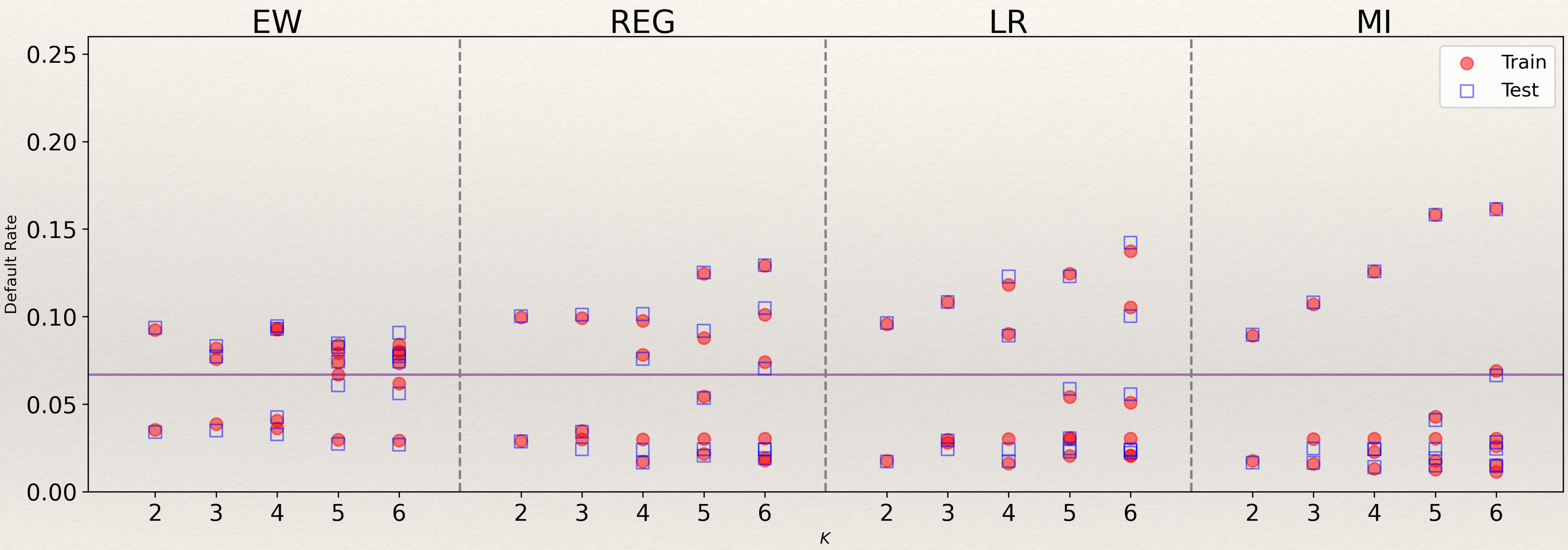


The GMC dataset

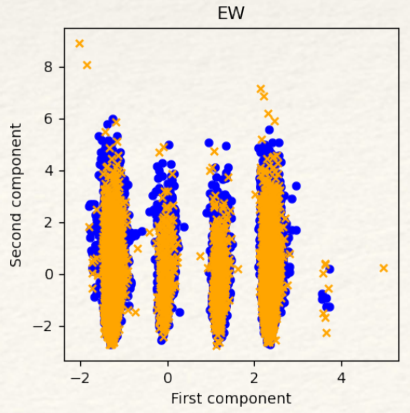
- The overall default rate 6.7%, 150,000 cases, $p = 12$.
- 



Default rates

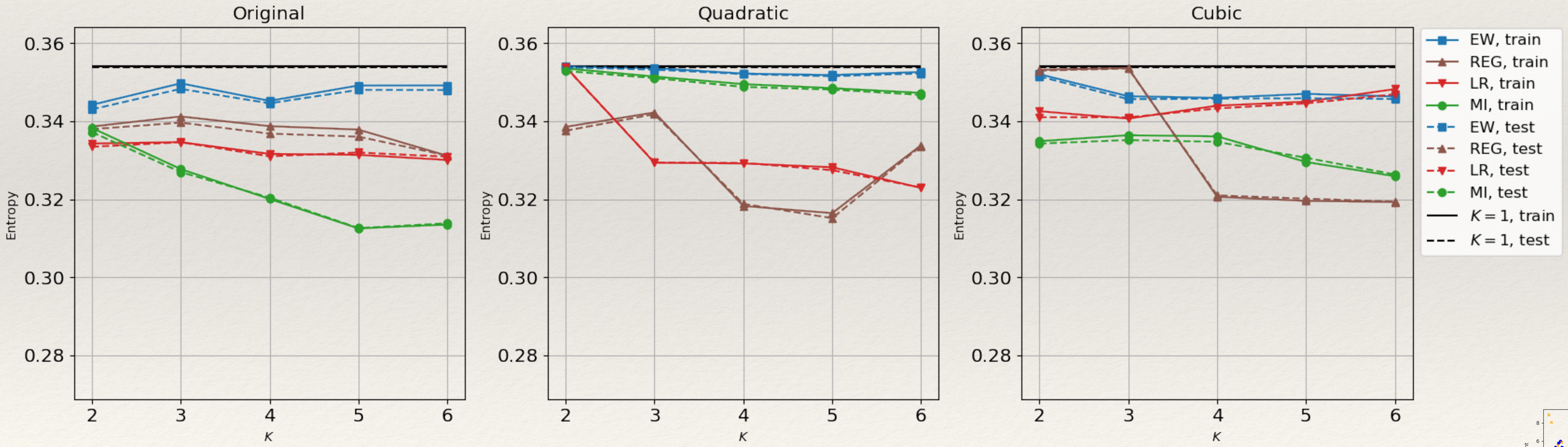


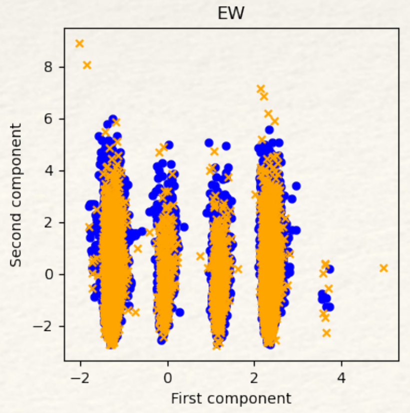
Rescaled clustering



Entropy

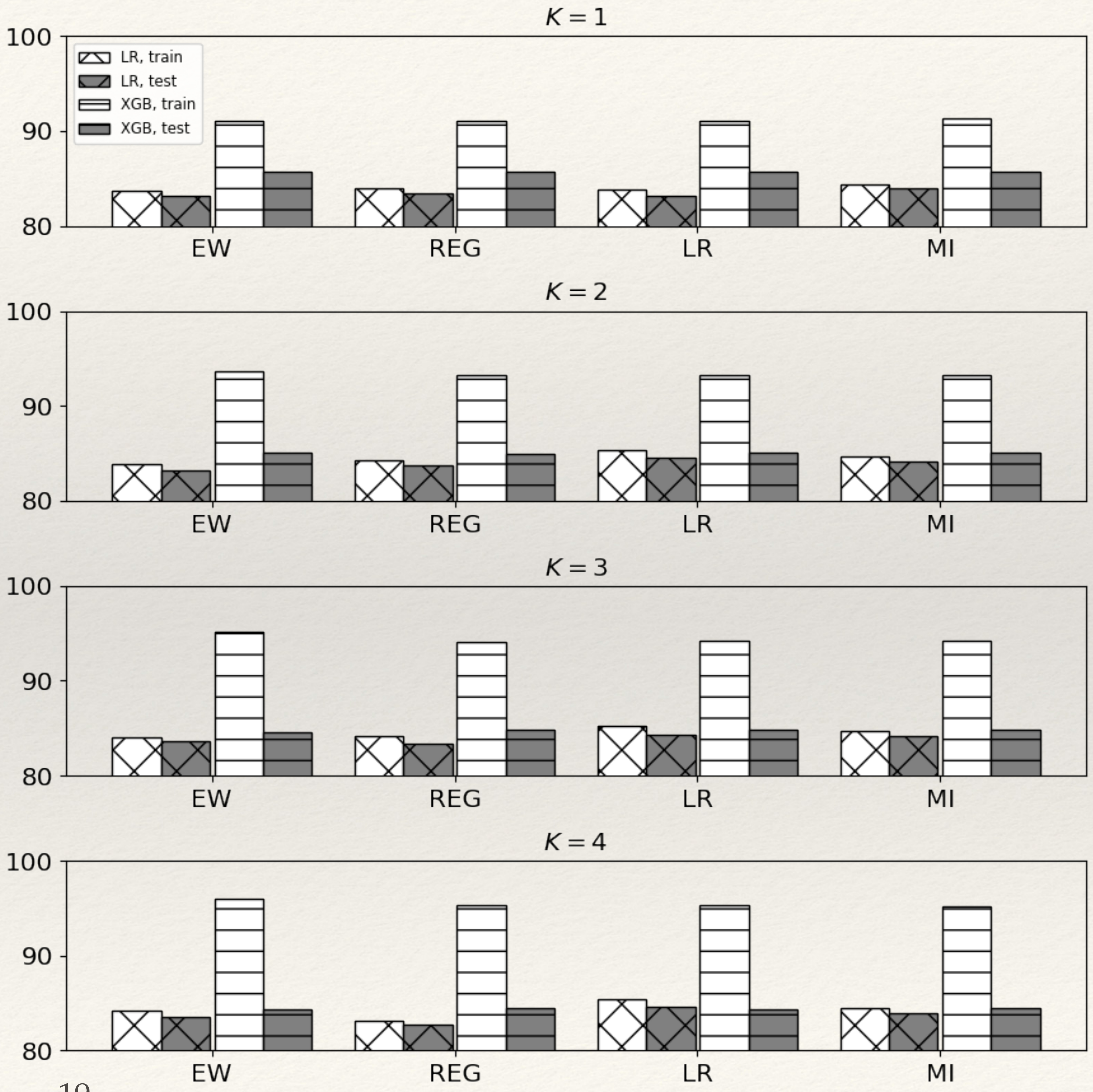
- Rescaled clustering reduces randomness!



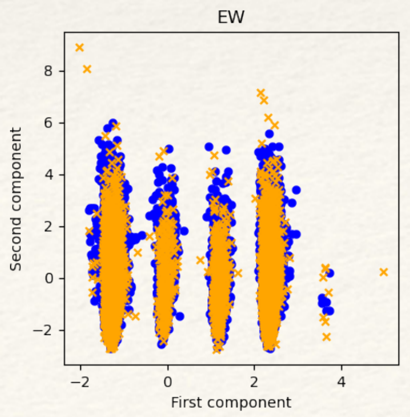


AUC in the training set

- The original feature set
- XGBoost: overfitting
- Logistic Regression: robust



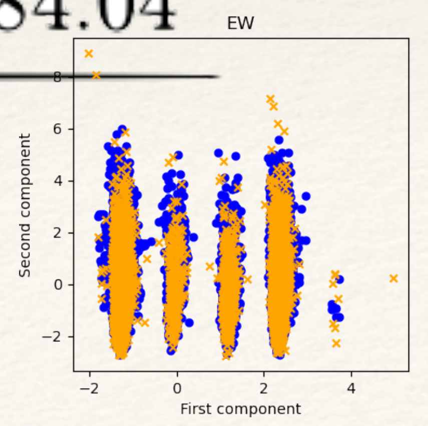
Rescaled clustering



AUC in the testing set

- XGBoost ~> Logistic Regression
- XGBoost not improved
- Logistic Regression > XGBoost
  - Quadratic feature set
  - K increases

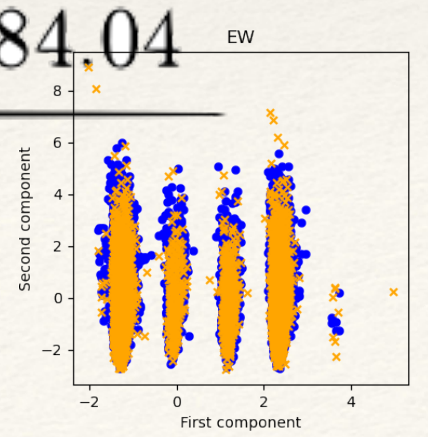
	XGBoost				Logistic Regression			
K	EW	REG	LR	MI	EW	REG	LR	MI
The original feature set								
1	85.66	**85.73**	85.66	85.66	83.17	**84.02**	83.19	83.48
2	85.07	85.28	85.01	84.87	83.12	84.14	84.54	83.68
3	84.56	84.89	84.87	84.82	83.61	84.14	84.27	83.42
4	84.32	84.48	84.35	84.51	83.48	83.92	84.56	82.66
The quadratic feature set								
1	85.35	85.24	85.24	85.23	84.36	84.43	84.51	84.46
2	84.36	84.68	84.43	84.53	84.50	84.30	85.11	84.50
3	83.51	84.09	84.43	83.83	84.54	84.48	85.23	84.37
4	83.51	83.56	84.02	83.82	84.39	84.54	85.40	84.34
The cubic feature set								
1	84.87	84.87	84.77	84.87	83.44	84.72	85.26	82.51
2	84.05	84.17	83.78	84.24	83.67	84.78	85.51	82.55
3	83.59	83.24	83.15	83.49	83.44	84.75	85.52^{+}	82.63
4	82.65	82.85	82.60	82.86	83.48	84.81	85.42	84.04



Positive-case clustering

- Clustering only postive cases
- Comparative to clustering all cases
- The same conclusion holds!
  - XGBoost does not benifit
  - Logistic Regression is improved

	XGBoost				Logistic Regression			
K	EW	REG	LR	MI	EW	REG	LR	MI
The original feature set								
1	85.66	**85.73**	85.66	85.66	83.17	**84.02**	83.19	83.48
2	85.07	85.28	85.01	84.87	83.12	84.14	84.54	83.68
3	84.56	84.89	84.87	84.82	83.61	84.14	84.27	83.42
4	84.32	84.48	84.35	84.51	83.48	83.92	84.56	82.66
The quadratic feature set								
1	85.35	85.24	85.24	85.23	84.36	84.43	84.51	84.46
2	84.36	84.68	84.43	84.53	84.50	84.30	85.11	84.50
3	83.51	84.09	84.43	83.83	84.54	84.48	85.23	84.37
4	83.51	83.56	84.02	83.82	84.39	84.54	85.40	84.34
The cubic feature set								
1	84.87	84.87	84.77	84.87	83.44	84.72	85.26	82.51
2	84.05	84.17	83.78	84.24	83.67	84.78	85.51	82.55
3	83.59	83.24	83.15	83.49	83.44	84.75	85.52^{+}	82.63
4	82.65	82.85	82.60	82.86	83.48	84.81	85.42	84.04



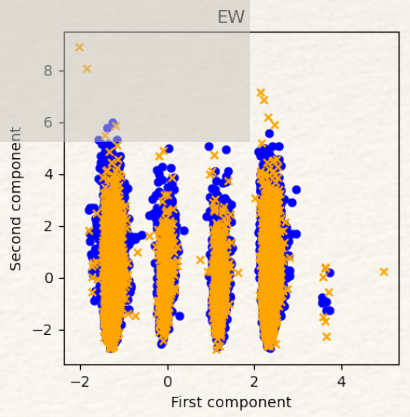
Rescaled clustering

Conclusion

- **XGBoost**
  - The state-of-the-art machine learning algorithm
  - Overfitting
  - Can not be improved
- **Logisitic regression!**
  - Interpretable
  - No overfitting
  - Could be improved
- Positive-case clustering ~ All-case clustering

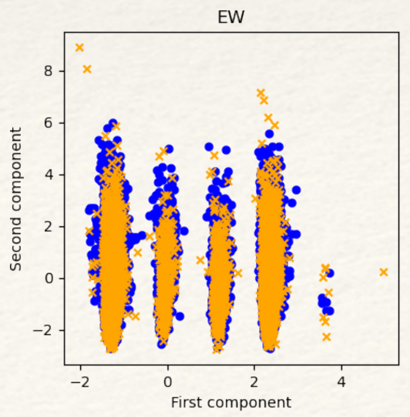
Rescaled clustering

	Logisitc Regression	XGBoost
Robust	✓	✗
Adding Polynomial transformed features	✓	✗
EW rescaling	✗	✗
REG rescaling	✓	✗
LG rescaling	✗	✗
MI rescaling	✗	✗



Outlook

- Other domain applications: credit reform, fraud detection
- Other rescaling: iScore
- More sophisticated clustering methods: spectral clustering



TALK HISTORY

20230623 Humboldt University PRI
20231022 NYCU 專班