Feb.10

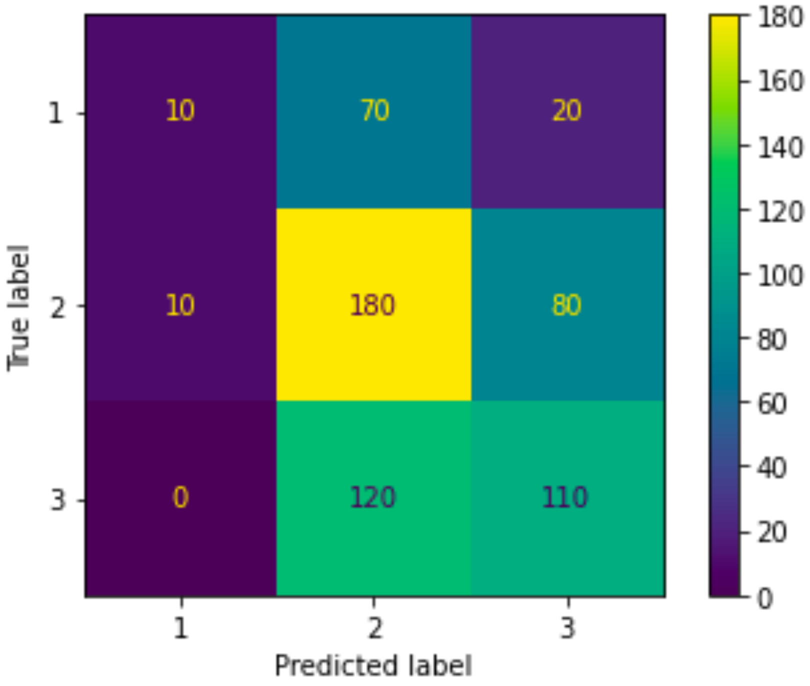
Meeting with Mengchen & Levin

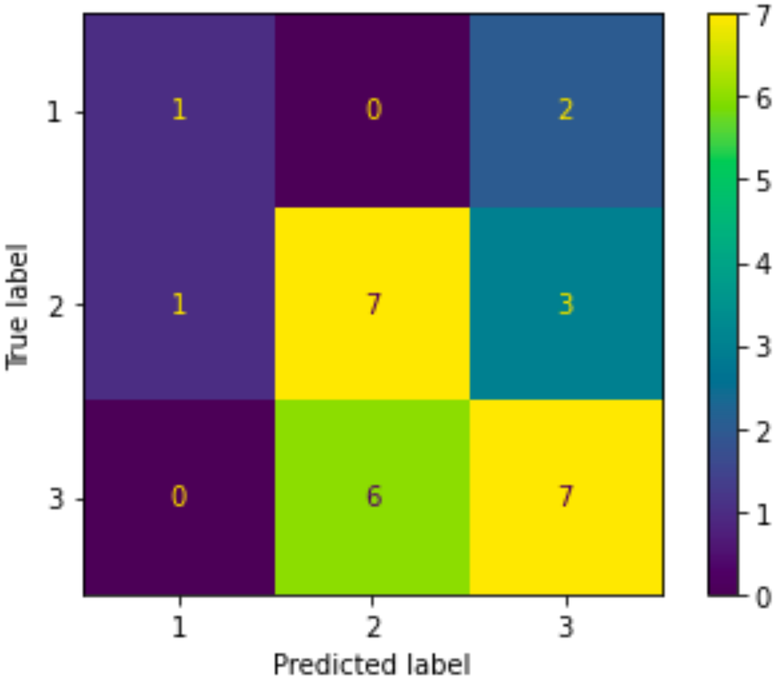
Retrain RF model

- Data
  - 87 episodes in total.
  - Train:test=7:3; train:dev=3:1 (4-folds cross validation)
  - Shuffle & repeat 10 times
- Param
  - max_depth=10, min_samples_leaf=2, max_features="log2", n_estimators=20
- Feature: 20 original + ai_eva + 3 probs (by the old RF model)
  - Also tried only use the 20 features, acc almost the same, but less level 1 prediction
- Results
  - Dev acc=0.5, test acc=0.56
  - Generated RF-2.model

Retrain RF model

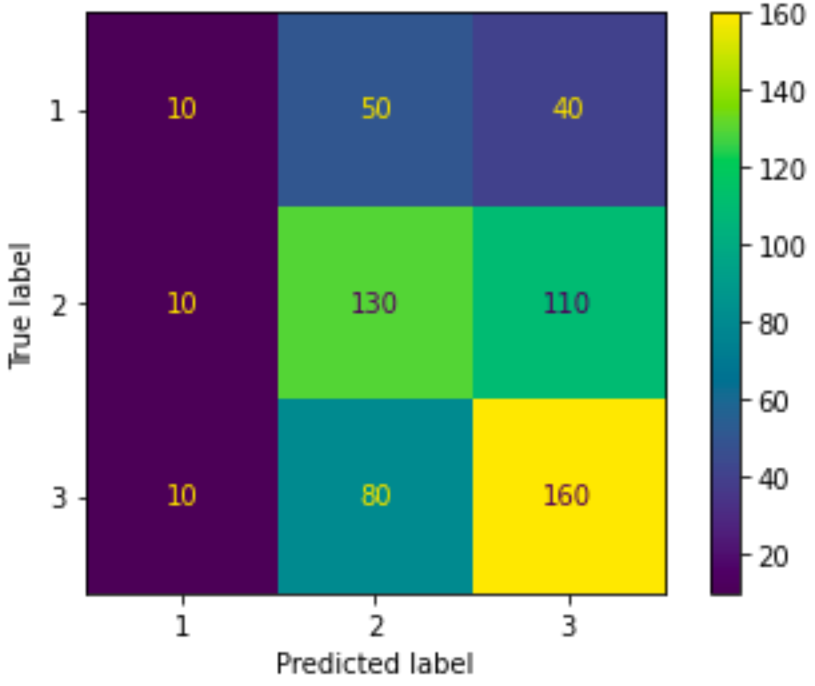
- 20+4 features
  - Train acc=0.5, test acc=0.56

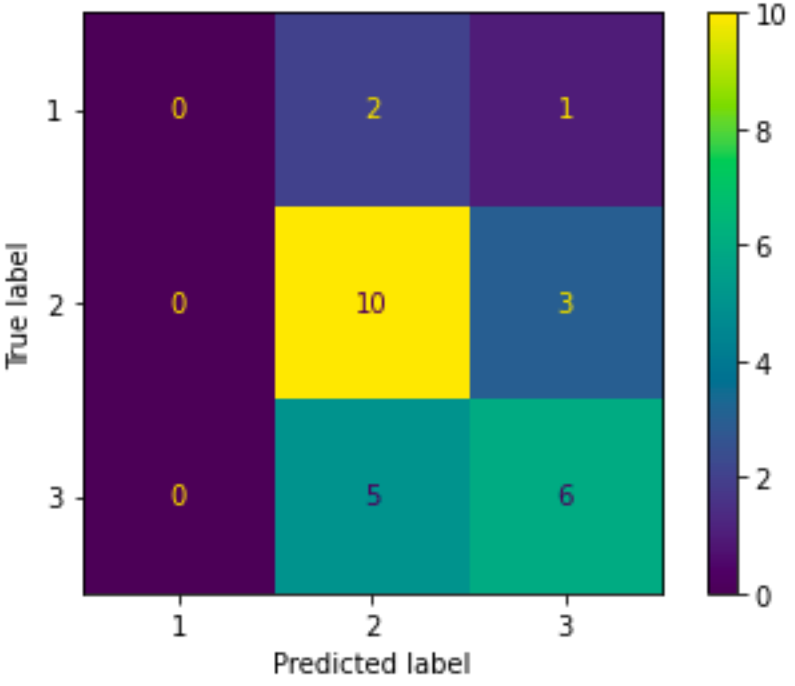




Retrain RF model

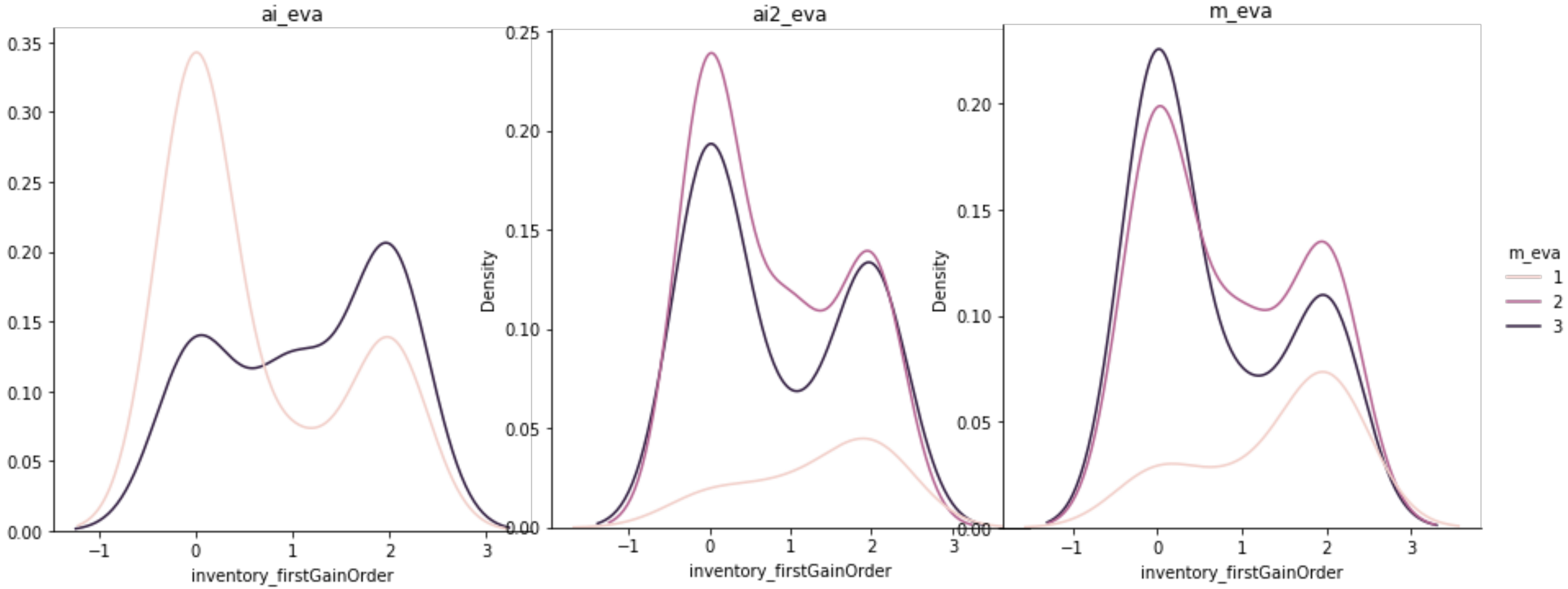
- 20 features
  - Train acc=0.5, test acc=0.59 (but precision of level1 gets worse)





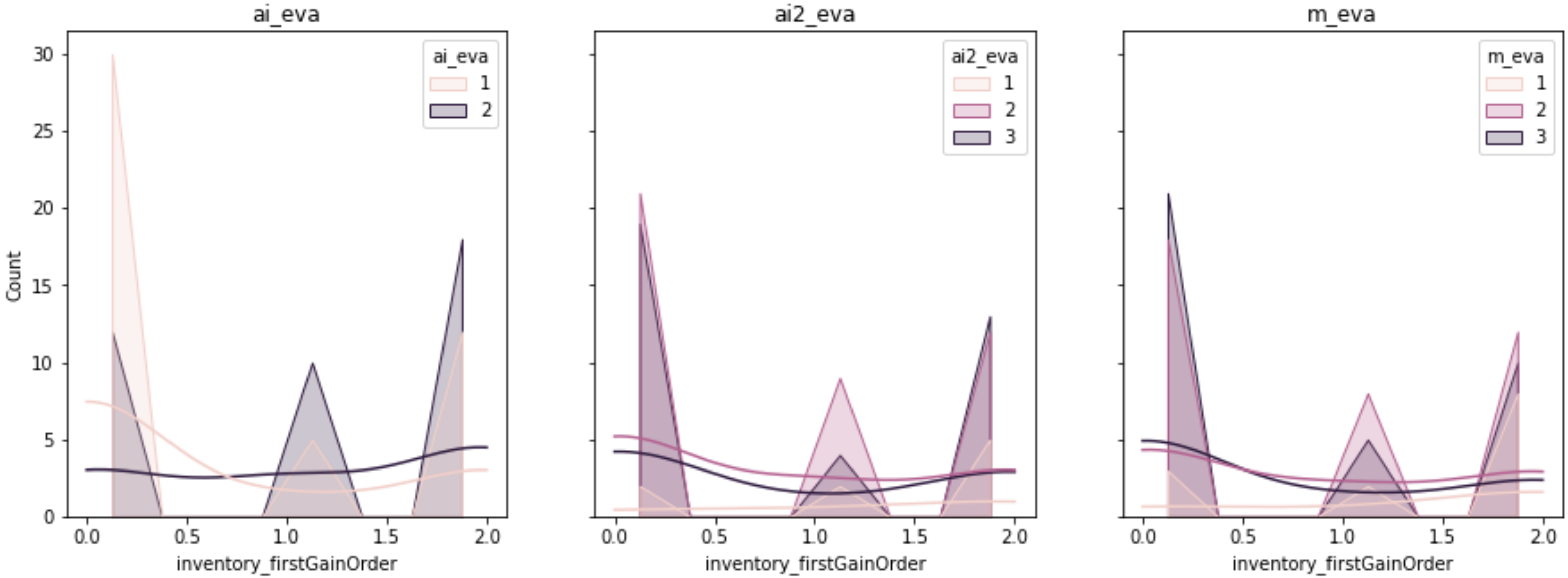
Feature Analysis

- inventory_firstGainOrder



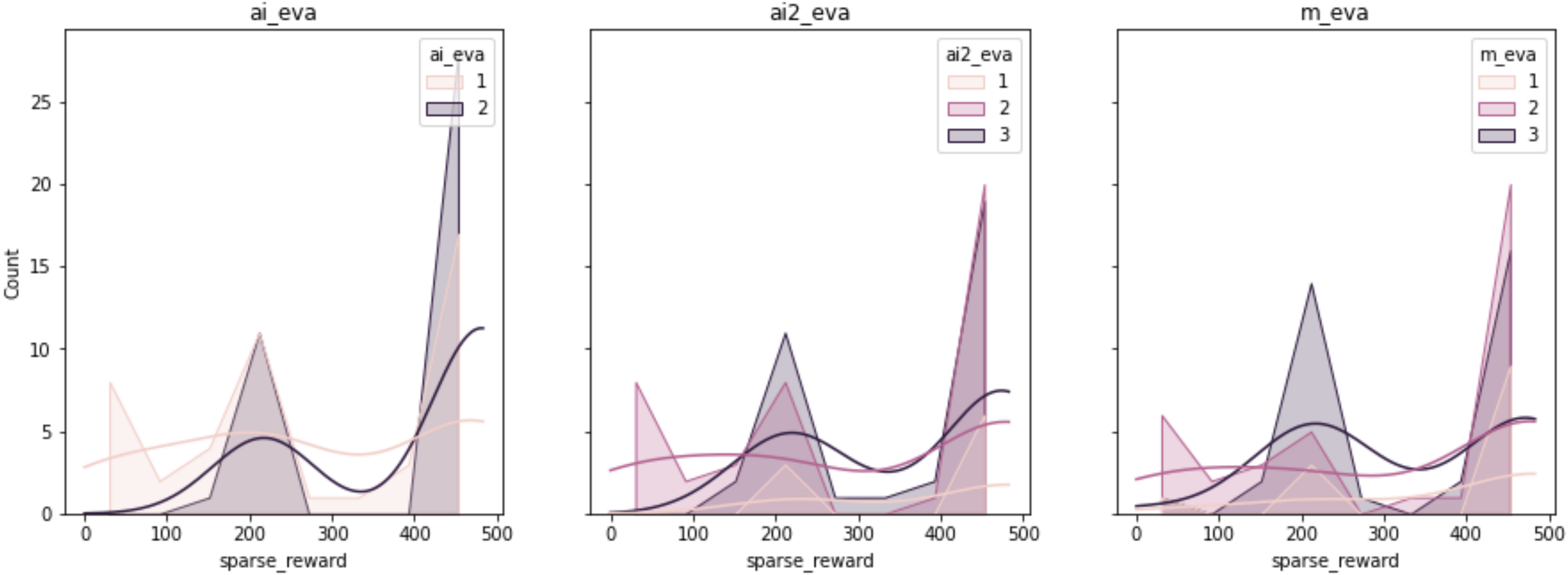
Feature Analysis

- inventory_firstGainOrder



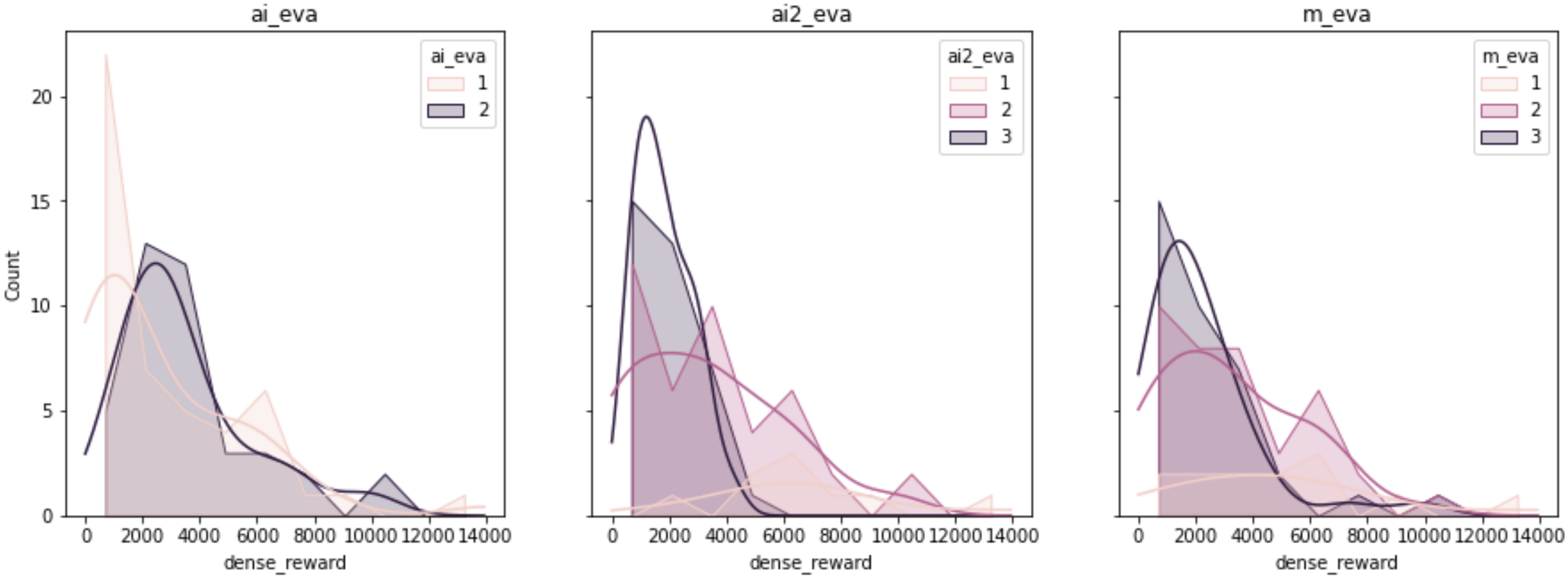
Feature Analysis

- sparse_reward



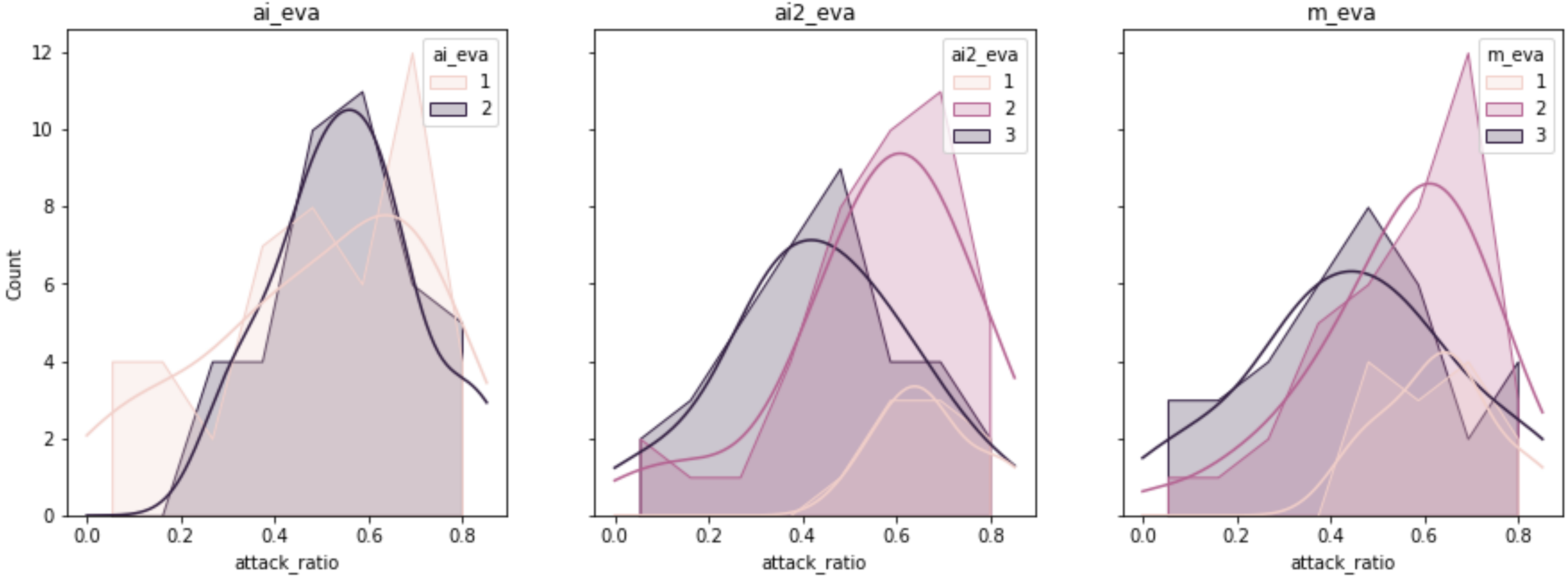
Feature Analysis

- dense_reward



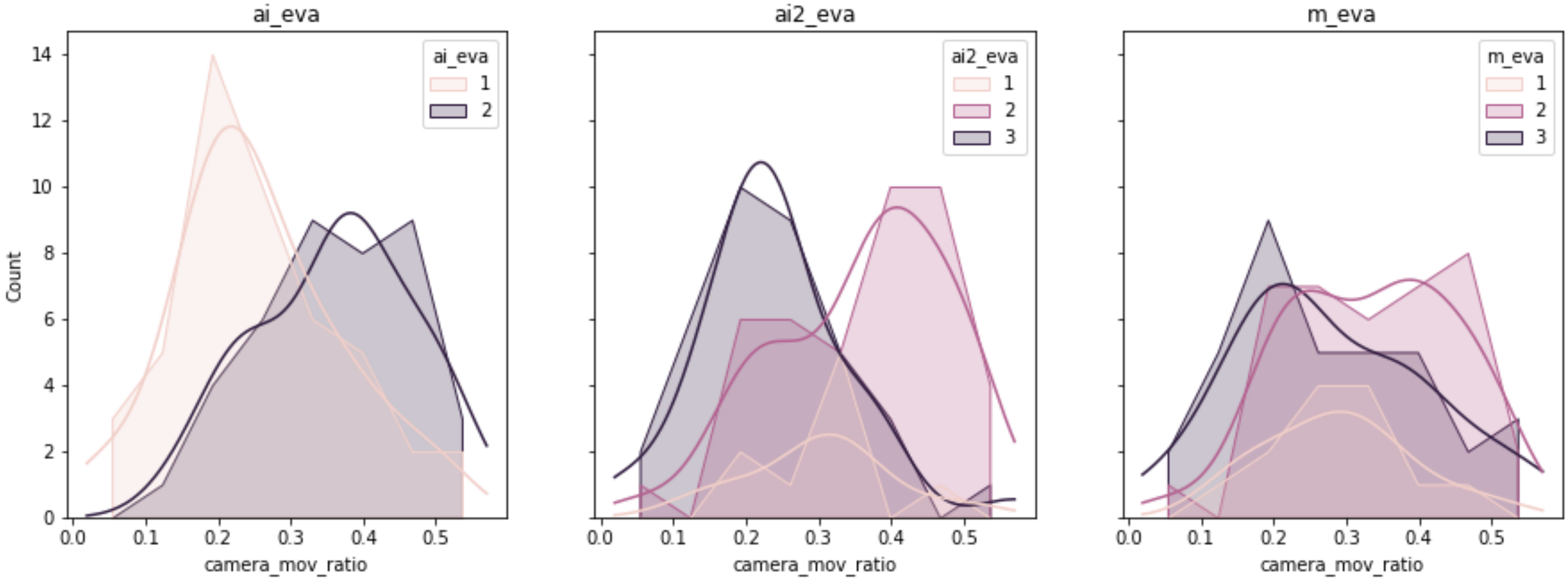
Feature Analysis

- attack_ratio



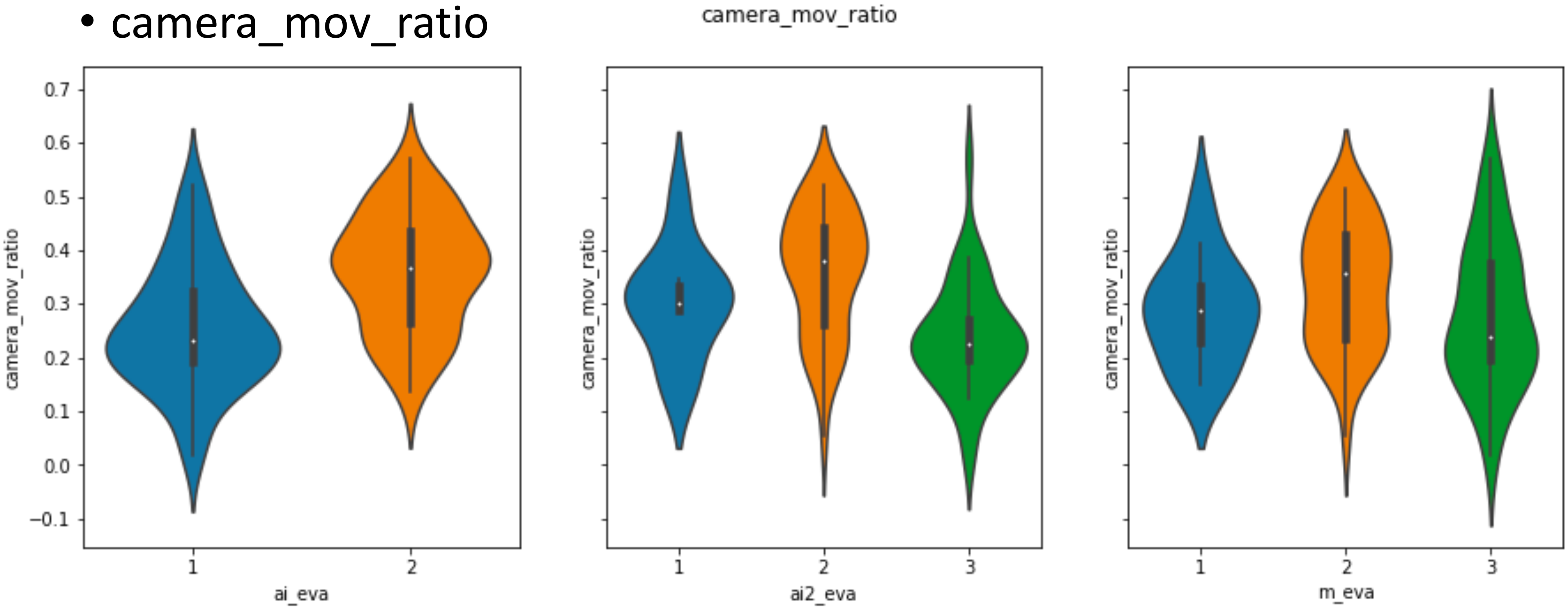
Feature Analysis

- camera_mov_ratio



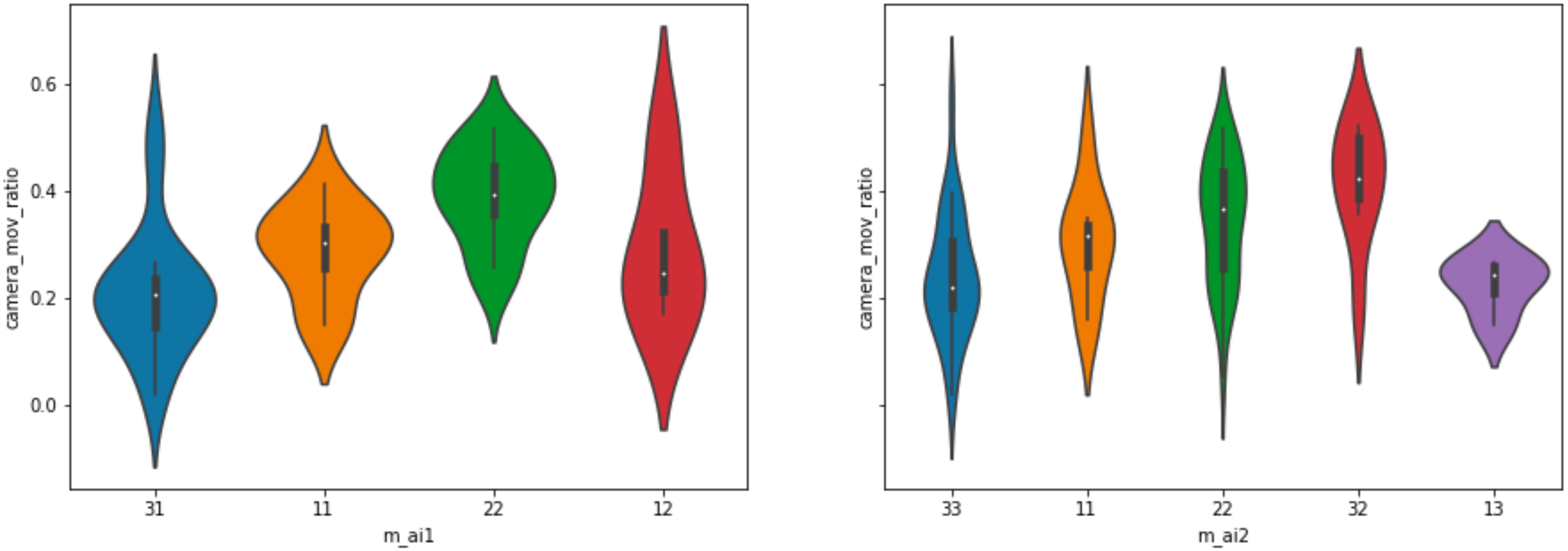
Feature Analysis

- camera_mov_ratio



human label-AI label pairs comparision

- e.g. “31” means human gave label 3 (advanced) while ai only gave label 1(junior).



camera_mov_ratio

Feb.17

Meeting with Levin

Retrain RF model

- We would like to move on with the combined model (24 features)

- Computing the marginal distributions of the confusion matrix of predicted and actual (from humans) evaluations. Are both distribution significantly different (chi-square test)?

- The test performance being higher than the training performance?
  - A mistake, it’s the final test higher than mean of dev.
- 10-fold cross validation (all 87 episodes)
  - mean accuracy : 0.58
  - standard deviation : 0.12 (high fluctuation)
  - Train acc: 0.89 ± 0.018

Feb.24

Meeting with Levin

Retrain RF model

- Computing the marginal distributions of the confusion matrix of predicted and actual (from humans) evaluations. Are both distribution significantly different (chi-square test)

- Fine-tune the 10-fold cross validation (all 87 episodes)
    - Influence: n_estimators > max_depth > min_samples_leaf > max_features

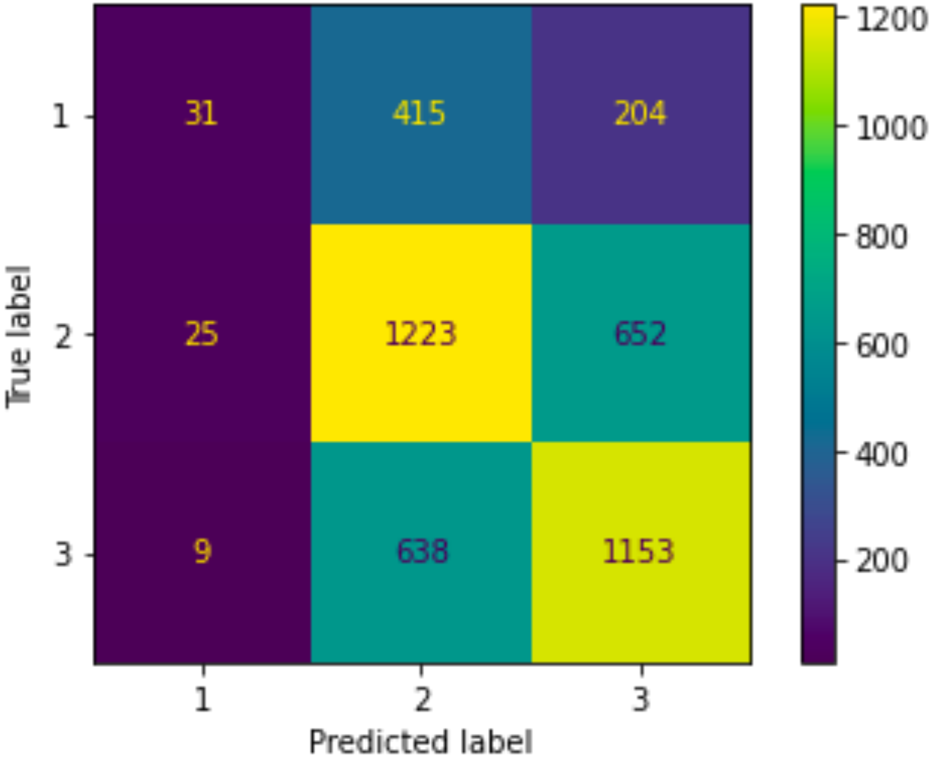
n_estimators	max_depth	min_samples_leaf	max_features	Dev acc	Train acc	Ave acc
10	10	2	"log2"	0.57 ±0.18	0.90 ±0.024	0.86
		3	"log2"	0.54 ±0.14	0.84 ±0.029	0.83
		4	"log2"	0.547 ±0.185	0.798 ±0.026	0.793
	5	2	"log2"	0.52 ±0.16	0.86 ±0.026	0.85
		3	"log2"	0.55 ±0.17	0.83 ±0.021	0.83
		4	"log2"	0.549 ±0.198	0.801 ±0.021	0.793
5	10	4	"log2"	0.561 ±0.174	0.764 ±0.048	0.782
20	5	2	"log2"	0.51 ±0.11	0.88 ±0.022	0.82
50	5	2	"log2"	0.52 ±0.17	0.91 ±0.018	0.93
10	15	2	"log2"	0.56 ±0.17	0.91 ±0.025	0.86
10	20	2	"log2"	0.56 ±0.17	0.91 ±0.025	0.86
10	3	2	"log2"	0.48 ±0.19	0.76 ±0.031	0.76

March 7

Fix the parameters & test

- max_depth=5, n_estimators=50, min_samples_leaf=4
- Repeat n=50 times

- 10-folds dev:
  - acc= 0.5535 ± 0.154
  - train_acc= 0.8413 ± 0.0253
- 50 times shuffle&repeat:
  - acc= **0.5535 ± 0.0236**
  - train_acc= 0.8413 ± 0.0101



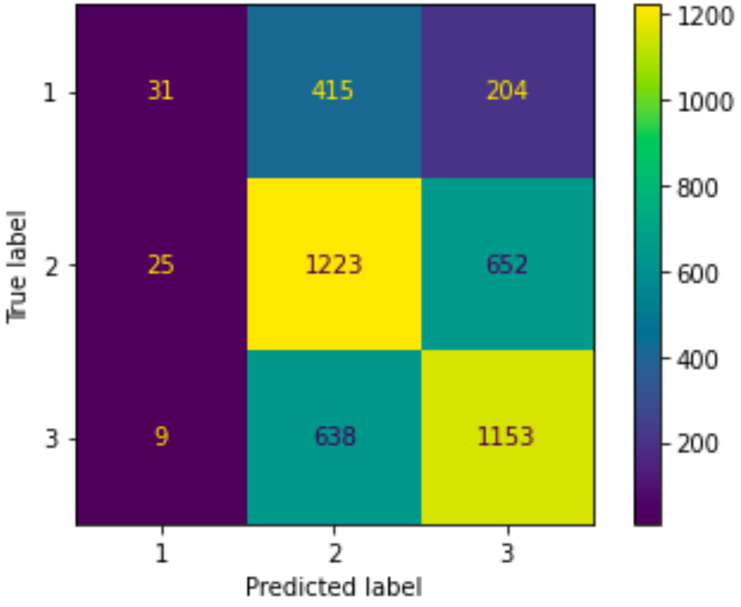
Marginal distributions

- Truths
  - Sum of 50 repeats: [650, 1900, 1800]
  - Mean: **[13, 38, 36]**
- Preds
  - Sum of 50 repeats: [65, 2276, 2009]
  - Mean: **[1.3 ± 0.9 , 45.52 ± 3.16, 40.18 ± 3.06]**

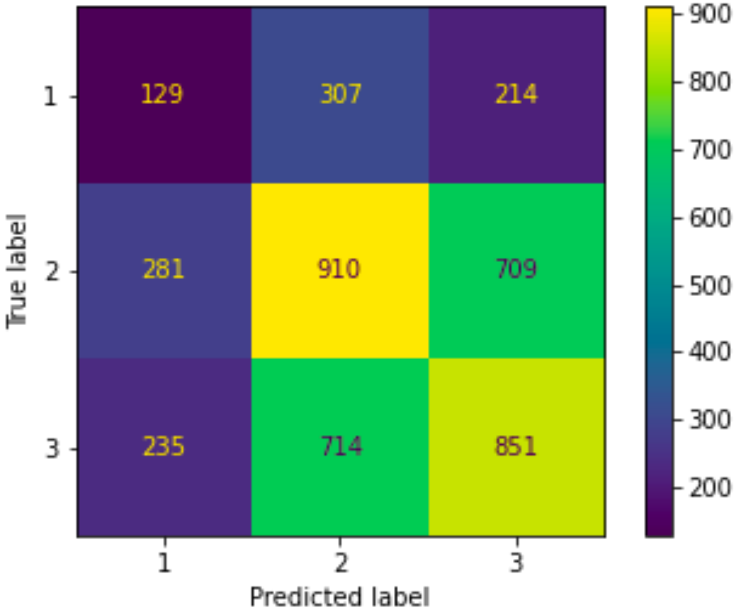
	Proportion %	Recall %(tp/tp+fn)	Precision% (tp/tp+fp)	F1 %
Class 1 (junior)	14.94	4.77 (31/650)	47.69 (31/65)	8.67
Class 2 (medium)	43.68	64.37	53.73	58.57
Class 3 (advanced)	41.38	64.06	57.39	60.54

Sample the proba results of RF
(same parameters)

- Original
  - 10-folds dev:
    - acc= 0.5535 ± 0.154
  - 50 times shuffle&repeat:
    - acc= **0.5535 ± 0.0236**



- Sampled
  - 10-folds dev:
    - acc= 0.4348 ± 0.1621
  - 50 times shuffle&repeat:
    - **acc= 0.4348 ± 0.0544**



Sampled Marginal distributions

- Truths: **[13, 38, 36]**
- Original Preds : [1.3 ± 0.9, 45.52 ± 3.16, 40.18 ± 3.06]
- Sampled Preds: **[13.02 ±** 3.30**, 38.08 ±** 4.21**, 35.9 ±** 4.05**]**
  - Sum of 50 repeats: [651, 1904, 1795]

Original Preds	Proportion %	Recall %(tp/tp+fn)	Precision% (tp/tp+fp)	F1 %
Class 1 (junior)	14.94	4.77 (31/650)	47.69 (31/65)	8.67
Class 2 (medium)	43.68	64.37	53.73	58.57
Class 3 (advanced)	41.38	64.06	57.39	60.54

Sampled Preds	Proportion %	Recall %(tp/tp+fn)	Precision% (tp/tp+fp)	F1 %
Class 1 (junior)	14.94	19.85 (129/650)	19.82 (129/651)	19.83
Class 2 (medium)	43.68	47.89 (910/1900)	47.79 (910/1904)	47.84
Class 3 (advanced)	41.38	47.28 (851/1800)	47.41 (851/1795)	47.34