Carnegie Mellon University
Master of Computational Data Science

Carnegie Mellon University
Language Technologies Institute

Predicting Learning Outcomes

Nov 13 Standup

Mengjie Shen, Xiaoyu Zhang, Xinyu Gu, Yizhou Chen, Yuchen Wang
Advisor: Ken Koedinger

11-632 (Fall 2023)
MCDS Capstone Course

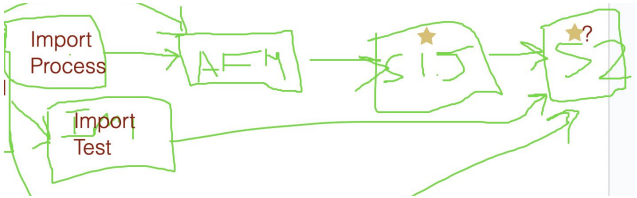
Sub Group 1-Mengjie, Xinyu

Last week:

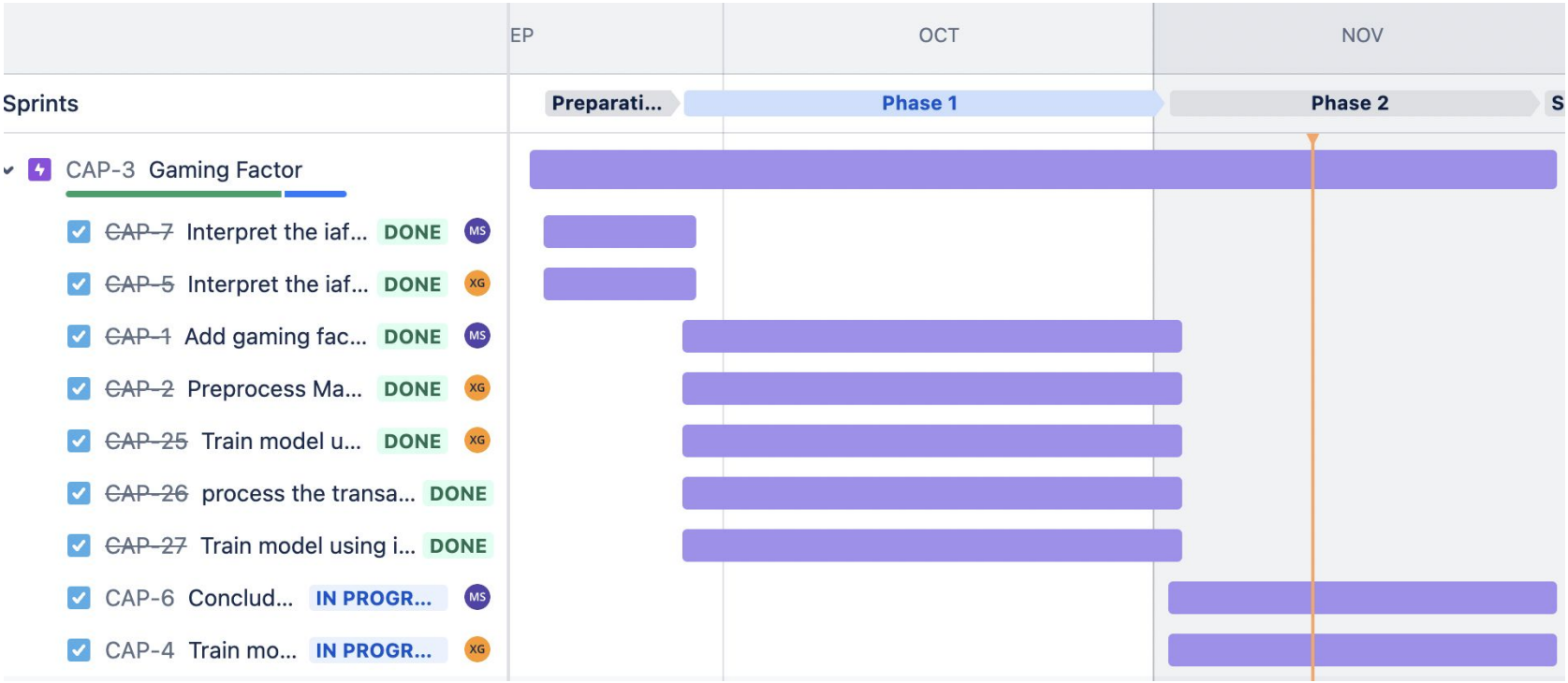
- Met with advisor to get familiar with LearnSphere Workflow
- Developed scripts for retrieving pre-test and post-test scores for workflow component

This week:

- Convert existing scripts to Workflow Components on LearnSphere
- Write draft report



Sub Group 1-Mengjie, Xinyu



Sub Group 2-Xiaoyu Zhang, Yizhou Chen, Yuchen Wang

- Using parameter derived from modeling the **learning process data** to predict students' **post-test score**

```
# Model 1: pretest only
test_scores %>%
  lm(Posttest ~ Pretest, data = .) %>%
  summ()

# Model 2: pretest + PredAvgiAFM
test_scores %>%
  lm(Posttest ~ PredAvgiAFM + Pretest, data = .) %>%
  summ()

# Model 3: pretest + int_iAFM
test_scores %>%
  lm(Posttest ~ int_iAFM + Pretest, data = .) %>%
  summ()

# Model 4: pretest + int_iAFM_reverse
test_scores %>%
  lm(Posttest ~ int_iAFM_reverse + Pretest, data = .) %>%
  summ()

# Model 5: pretest + int_iAFM + int_iAFM_reverse
test_scores %>%
  lm(Posttest ~ int_iAFM + int_iAFM_reverse + Pretest, data = .) %>%
  summ()
```

Model	# students	F-statistic	R-squared	Adjusted R-squared	p
pretest	129	71.18	0.36	0.35	0.00
pretest + PredAvgiAFM	129	49.19	0.44	0.43	0.00
pretest + int_iAFM	129	84.03	0.57	0.56	0.00
pretest + int_iAFM_reverse	129	75.66	0.55	0.54	0.00
pretest + int_iAFM + int_iAFM_reverse	129	55.73	0.57	0.56	0.00

Sub Group 2-Xiaoyu Zhang, Yizhou Chen, Yuchen Wang

- Pairwise ANOVA to analyze the effect of student estimates
- Model 1 vs. Model X (other models)

  High F-statistic: variable “*pretest”* is significantly related to “posttest”

  Low R-squared: “*pretest”* does not explain a significant proportion of the variance in the dependent variable, we should incorporate more variables
- Model 3 vs. Model 4

  “Intercepts” is a better-fitting variable compared to “predAvgIafm” and is more effective in explaining the variation in the posttest scores.

Sub Group 2-Xiaoyu Zhang, Yizhou Chen, Yuchen Wang

- Model 3 vs. Model 5

  Same R-squared: **int_iAFM_reverse** doesn't seem to significantly enhance the model's ability to explain the variation in Posttest scores

  Possible reasons:

  1. **int_iAFM_reverse** has limited contribution to the posttest score

  2. **int_iAFM** and **int_iAFM_reverse** might be highly correlated or redundant, which could lead to multicollinearity issues in Model 5

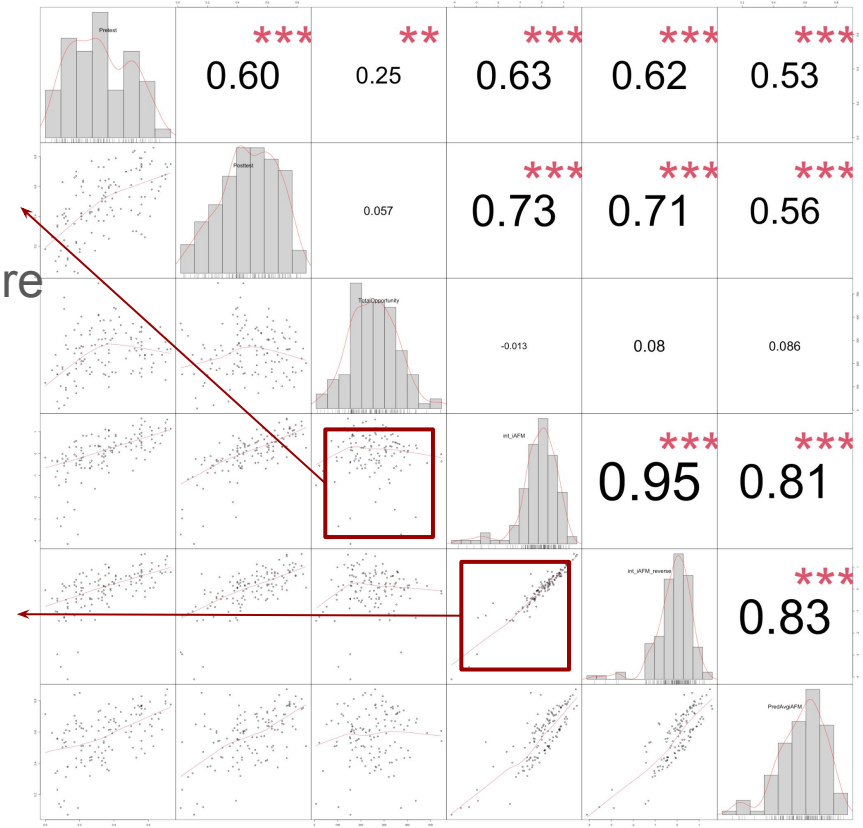
  3. New variable increase the model complexities

Sub Group 2-Xiaoyu Zhang, Yizhou Chen, Yuchen Wang

- Correlation study

Observation 1: corr(total_opportunity, intercept_reverse) = -0.013
Interpretation: The tutor system gives more practice to students who need it.

Observation 2: corr(intercept, intercept_reverse) = 0.95
Interpretation: students with good initial scores have better final scores



Sub Group 2-Xiaoyu Zhang, Yizhou Chen, Yuchen Wang

Step 2 Alternative - 1 parameter fit

Create a table with both pre and post in separate rows for each student

Student	Test-Time	Test-Score	Process-Model-Prediction1	Process-Model-Prediction2
S1	Pre	.4	prob(-1.1) [intercept_iAFM]	prob(-1.1) [intercept_iAFM]
S1	Post	.6	prob(.4) [intercept_iAFM_reverse]	prob(.34) [max-Opp-iAFM-pred]

S2 …

Run analyses

Two parameter version:

Model1: Test-Score ~ Process-Model-Prediction1 [+ Intercept]

Model2: Test-Score ~ Process-Model-Prediction2 [+ Intercept]

One parameter version:

Model3: Test-Score ~ 1* Process-Model-Prediction1 [+ Intercept]

Model4: Test-Score ~ 1* Process-Model-Prediction2 [+ Intercept]

Sub Group 2-Xiaoyu Zhang, Yizhou Chen, Yuchen Wang

Next week:

- Besides F-statistics, report more statistics (e.g., AIC, BIC) for the 5 models in Step 2
- Correlation matrix with pretest, posttest, total_opp, intercept_iAFM, intercept_iAFM_reverse, PredAvgiAFM
- Further analyze the differences between 5 models
- Make scatter plots based on results from step 4