Predication and Analysis of STEM and Non-STEM College Major Enrollment with ASSISTments

Ke Cao

Educational Data Mining

Introduction

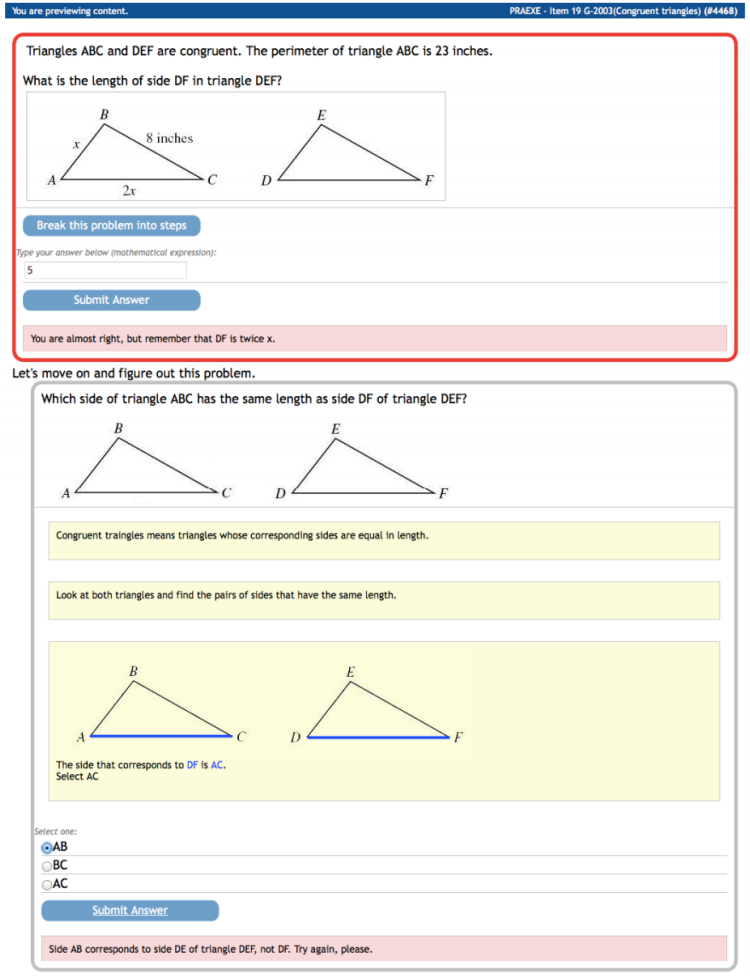
- Research shows that middle school is a crucial juncture for a student to start thinking about his or her academic achievement, college attendance, and future career.
- College and university degree programs in science, technology, engineering and mathematics (STEM) are considered STEM degrees, and they are in high demand across many industries.
- Educators and industry analysts detected a trend that indicated an academic deficiency in STEM areas for students entering college.
- It is important to let the students to be well-prepared with stem skills for stressful and rigorous college-level stem major courses

Previous Study

- Using such data, researchers did a large amount of statistical analysis. 66% accuracy of predicting whether students will choose STEM or non-STEM major was achieved by using logistic regression model
- Disengagement, which includes boredom, off-task, and frustration, has negative impact on attitude of higher education and causes low learning achievement

ASSISTments: Mathematics Educational Software

- ASSISTments is a free web-based mathematics tutoring system for middle-school mathematics, developed by Dr. Neil Hefferman
- ASSISTments evaluates a students knowledge level, detects and records a student's interaction, and assists students in understanding concepts



Motivation

- Logistic regression model has rigid assumption about multicollinearity, linearity of independent variables and log odds, and independent observations
- Better representation of samples will lead more accurate and more robust classification model
- Hyperparameter optimization to success higher accuracy was not discussed in previous studies
- Neglected the power of unsupervised learning

Data

- Generated from ASSISTments which tracks students from their use of the ASSISTments blended learning platform in middle school in 2004-2007, to their high school course-taking, college enrollment, and first job out of college.
- 87,8110 interactions stored in the log files with 76 features at ASSISTments from 517 students
- For each student we calculate the mean value for each feature in the sequence of interaction and dropped the sample without STEM or NON-STEM label
- Left with 492 samples

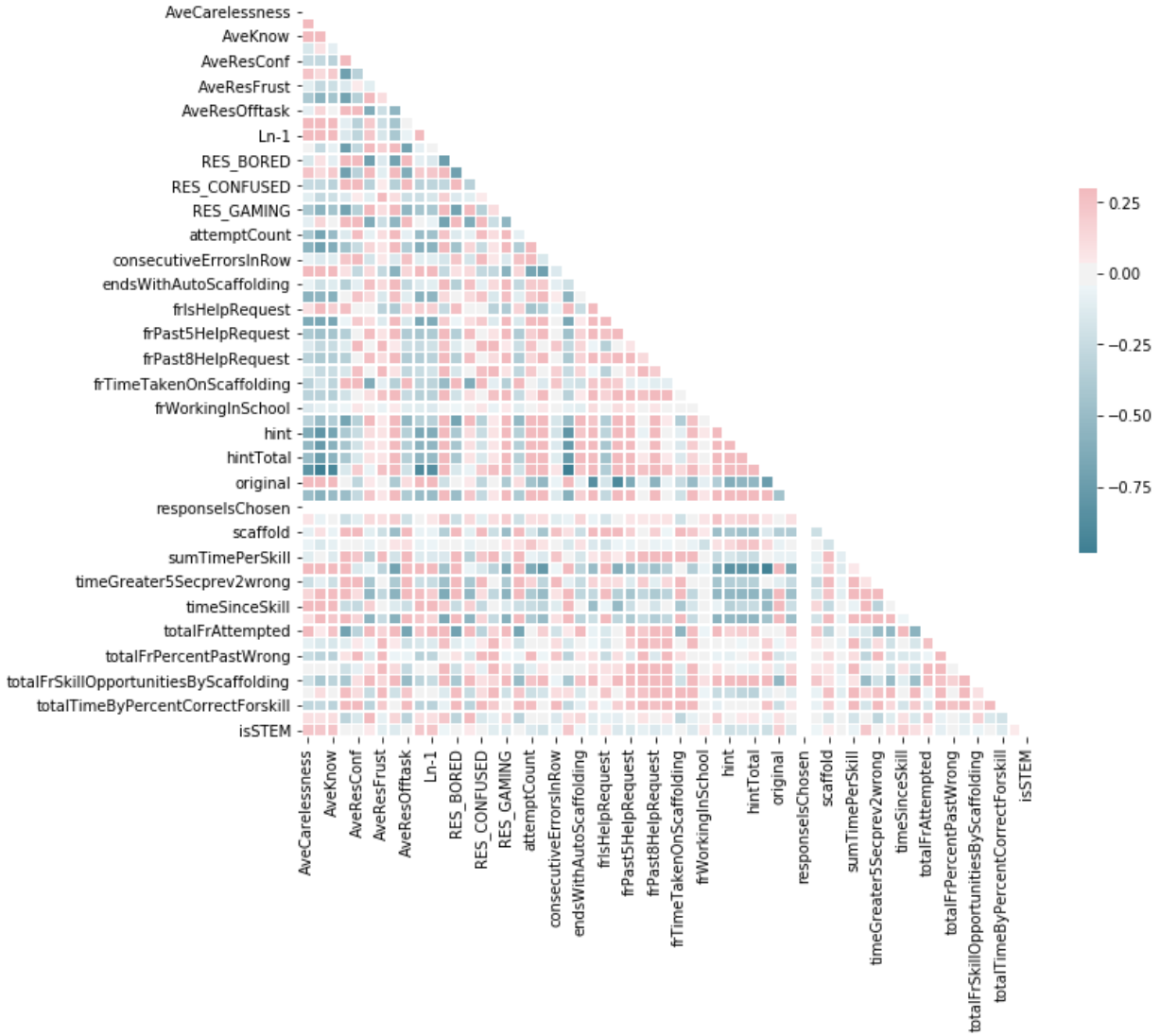
Missing Values

- In the feature MCAS, which is Massachusetts Comprehensive Assessment System, -999 indicates the missing value of the score.

- Impute missing MCAS by ridge regression with the adjusted R^2 of

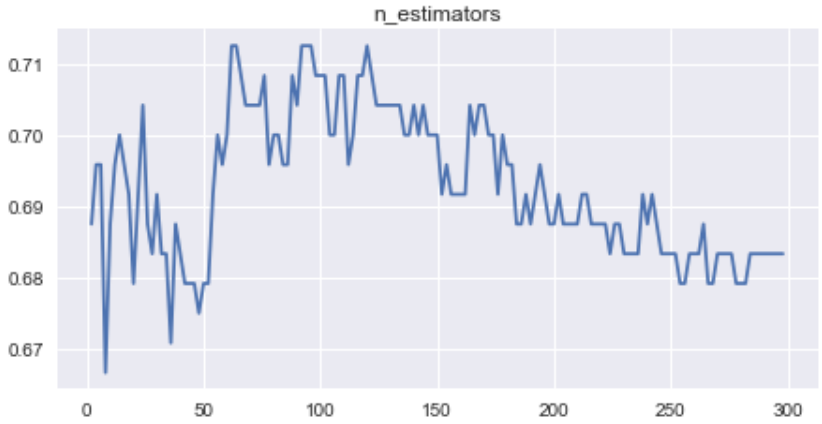
```
0.765328267971
```

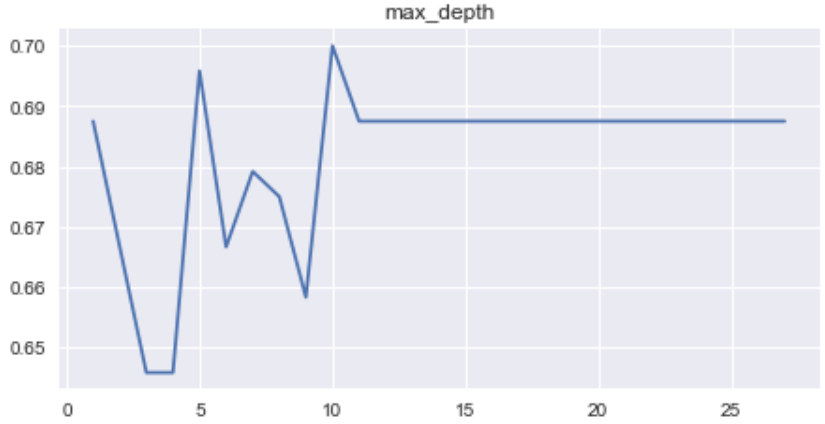
Correlation Matrix

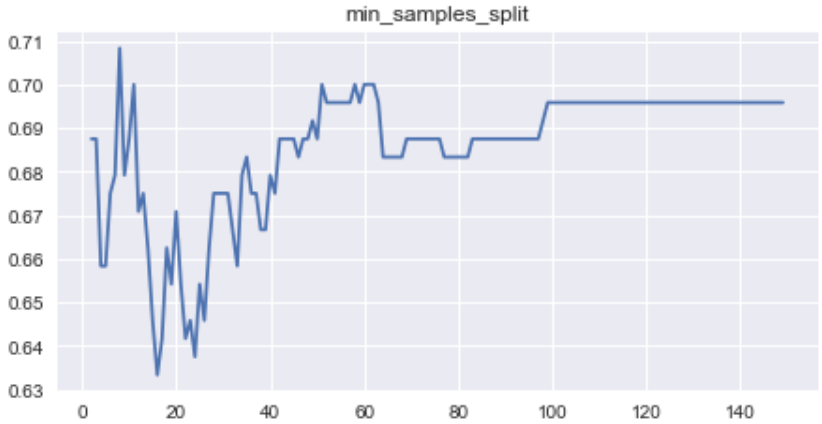


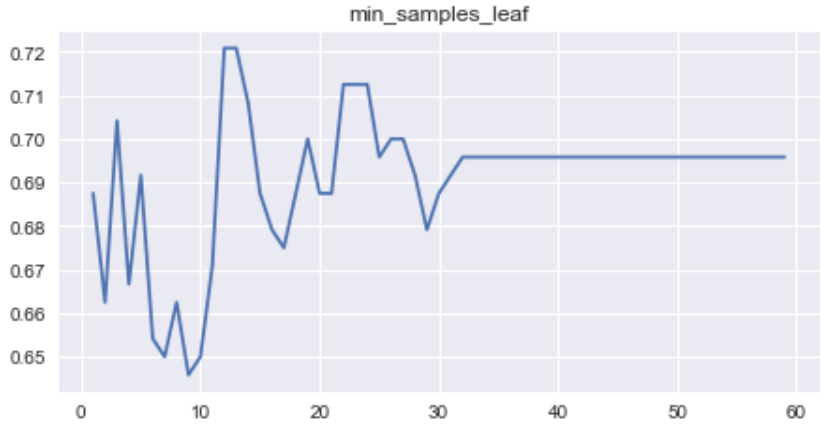
Random Forest without Feature Engineering

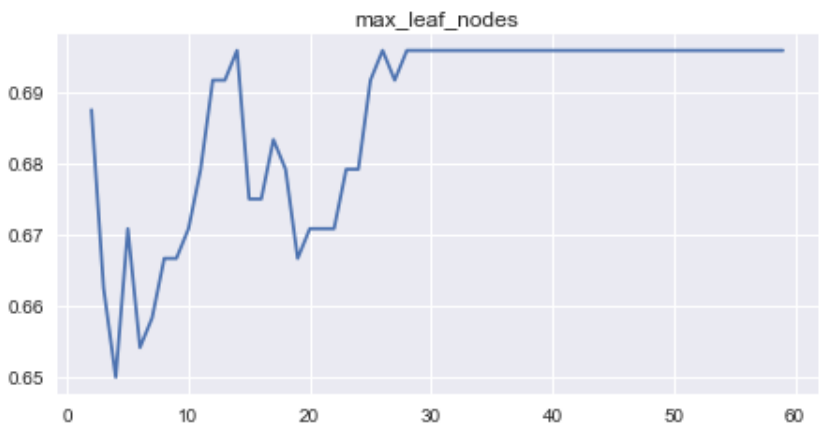
- Hyperparameter Oprimization

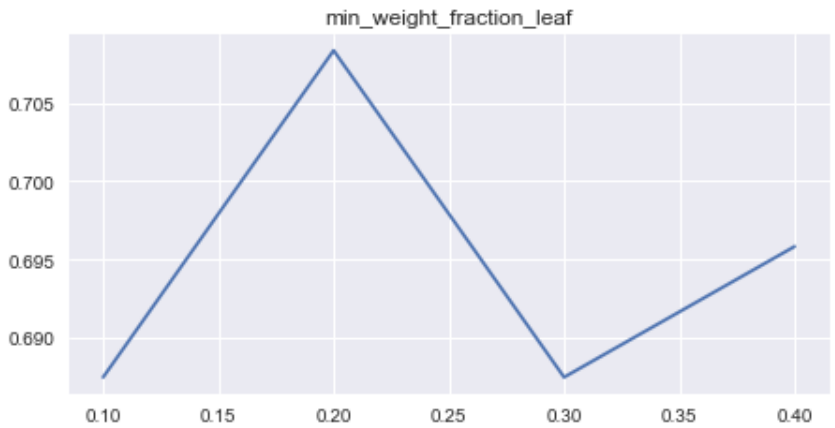

n_estimators
0.71
0.70
0.69
0.68
0.67
0
50
100
150
200
250
300


max_depth
0.70
0.69
0.68
0.67
0.66
0.65
0
5
10
15
20
25


min_samples_split
0.71
0.70
0.69
0.68
0.67
0.66
0.65
0.64
0.63
0
20
40
60
80
100
120
140


min_samples_leaf
0.72
0.71
0.70
0.69
0.68
0.67
0.66
0.65
0
10
20
30
40
50
60


max_leaf_nodes
0.69
0.68
0.67
0.66
0.65
0
10
20
30
40
50
60


min_weight_fraction_leaf
0.705
0.700
0.695
0.690
0.10
0.15
0.20
0.25
0.30
0.35
0.40

Grid Search Cross Validation

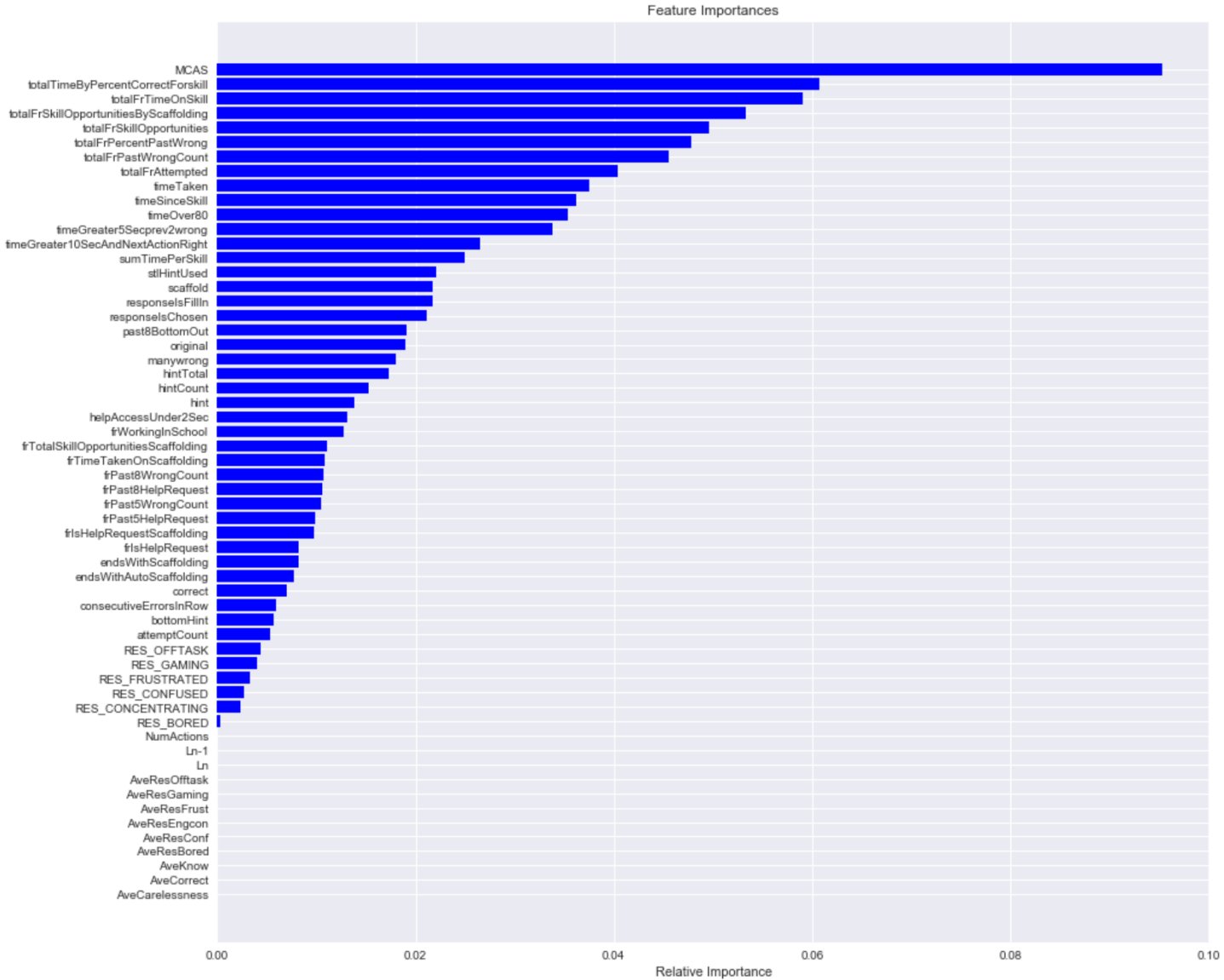
```
Model with rank: 1
Mean validation score: 0.7000)
Parameters: {'max_depth': 5, 'max_leaf_nodes': 10, 'min_samples_leaf': 10,
'min_samples_split': 60, 'min_weight_fraction_leaf': 0.2, 'n_estimators': 55}

Model with rank: 2
Mean validation score: 0.7000)
Parameters: {'max_depth': 5, 'max_leaf_nodes': 10, 'min_samples_leaf': 10,
'min_samples_split': 60, 'min_weight_fraction_leaf': 0.2, 'n_estimators': 60}

Model with rank: 3
Mean validation score: 0.7000)
Parameters: {'max_depth': 5, 'max_leaf_nodes': 10, 'min_samples_leaf': 10,
'min_samples_split': 60, 'min_weight_fraction_leaf': 0.2, 'n_estimators': 65}

Model with rank: 4
Mean validation score: 0.7000)
Parameters: {'max_depth': 5, 'max_leaf_nodes': 10, 'min_samples_leaf': 11,
'min_samples_split': 60, 'min_weight_fraction_leaf': 0.2, 'n_estimators': 55}
```

Feature Importance Plot



Feature Engineering

- Random Forest(RF), Linear Discriminant Analysis(LDA), logistic regression(LogReg), svm, and Recursive feature elimination with cross-validation(RFECV) by setting the response variable as STEM which is 1, or Non-STEM majors which is 0, and the mean of each coefficient generated from these five algorithms was collected.

	RF	RFECV	lda	logReg	svc	Mean
AveCarelessness	1.0	0.0	0.48	0.0	0.0	0.3
AveCorrect	0.4	0.0	0.46	0.01	0.0	0.17
AveKnow	0.34	0.0	0.07	0.0	0.0	0.08
AveResBored	0.08	0.0	1.0	0.0	0.0	0.22
AveResConf	0.01	0.0	0.09	0.0	0.0	0.02
AveResEngcon	0.0	0.0	0.03	0.01	0.0	0.01
AveResFrust	0.34	0.0	0.16	0.0	0.0	0.1
AveResGaming	0.0	0.0	0.16	0.0	0.0	0.03
AveResOfftask	0.06	0.0	0.1	0.0	0.0	0.03
Ln	0.27	0.0	0.59	0.0	0.0	0.17
Ln-1	0.72	0.0	0.59	0.0	0.0	0.26
NumActions	0.01	0.6	0.0	0.56	0.41	0.32
RES_BORED	0.01	0.0	1.0	0.0	0.0	0.2
RES_CONCENTRATING	0.0	0.0	0.03	0.01	0.0	0.01
RES_CONFUSED	0.05	0.0	0.09	0.0	0.0	0.03
RES_FRUSTRATED	0.03	0.0	0.16	0.0	0.0	0.04
RES_GAMING	0.0	0.0	0.16	0.0	0.0	0.03
RES_OFFTASK	0.01	0.0	0.1	0.0	0.0	0.02
attemptCount	0.15	0.0	0.06	0.05	0.02	0.06
bottomHint	0.22	0.0	0.56	0.0	0.0	0.16
consecutiveErrorsInRow	0.04	0.0	0.13	0.0	0.02	0.04
correct	0.29	0.0	0.46	0.01	0.0	0.15
endsWithAutoScaffolding	0.0	0.87	0.77	0.0	0.0	0.33
endsWithScaffolding	0.0	0.0	0.21	0.01	0.01	0.05
frIsHelpRequest	0.04	0.0	0.32	0.01	0.0	0.07
frIsHelpRequestScaffolding	0.0	0.0	0.1	0.01	0.01	0.02
frPast5HelpRequest	0.38	0.0	0.24	0.03	0.0	0.13
frPast5WrongCount	0.16	0.0	0.05	0.01	0.01	0.05
frPast8HelpRequest	0.18	0.0	0.15	0.04	0.01	0.08
frPast8WrongCount	0.0	0.0	0.04	0.01	0.01	0.01
frTimeTakenOnScaffolding	0.05	0.27	0.0	0.24	0.55	0.22
frTotalSkillOpportunitiesScaffolding	0.04	0.0	0.0	0.05	0.0	0.02
frWorkingInSchool	0.0	0.0	0.07	0.02	0.01	0.02
helpAccessUnder2Sec	0.02	0.0	0.71	0.0	0.0	0.15
hint	0.28	0.0	0.49	0.01	0.0	0.16
hintCount	0.14	0.0	0.1	0.03	0.03	0.06
hintTotal	0.13	0.0	0.01	0.05	0.04	0.05
manywrong	0.27	0.07	0.36	0.02	0.0	0.14
original	0.32	0.0	0.12	0.01	0.0	0.09
past8BottomOut	0.07	0.0	0.02	0.01	0.01	0.02
responseIsChosen	0.0	1.0	0.0	0.0	0.0	0.2
responseIsFillIn	0.0	0.0	0.28	0.0	0.0	0.06
scaffold	0.0	0.2	0.63	0.0	0.01	0.17
stlHintUsed	0.0	0.8	0.45	0.0	0.0	0.25
sumTimePerSkill	0.05	0.53	0.0	0.47	0.46	0.3
timeGreater10SecAndNextActionRight	0.1	0.0	0.79	0.0	0.0	0.18
timeGreater5Secprev2wrong	0.0	0.93	0.18	0.0	0.0	0.22
timeOver80	0.04	0.0	0.02	0.0	0.0	0.01
timeSinceSkill	0.07	0.73	0.0	0.0	0.0	0.16
timeTaken	0.05	0.33	0.0	0.26	0.42	0.21
totalFrAttempted	0.02	0.4	0.0	0.72	1.0	0.43
totalFrPastWrongCount	0.06	0.0	0.02	0.02	0.03	0.03
totalFrPercentPastWrong	0.0	0.0	0.08	0.0	0.0	0.02
totalFrSkillOpportunities	0.03	0.0	0.0	0.1	0.04	0.03
totalFrSkillOpportunitiesByScaffolding	0.02	0.0	0.05	0.01	0.01	0.02
totalFrTimeOnSkill	0.04	0.47	0.0	1.0	0.62	0.43
totalTimeByPercentCorrectForskill	0.28	0.67	0.0	0.25	0.04	0.25
MCAS	0.39	0.13	0.0	0.23	0.52	0.25

Feature Selection

- The features with mean of performance metrics less than 0.05 were dropped empirically
- Concentration, confusion, frustration, off-task, and gaming, have small contribution to classify STEM or Non-STEM major chosen.
- Test skill effects including correctness, scaffolding, hint, MCAS, and first response time spent on knowledge component across all problems

Supervised Learning: Support Vector Machine

```
Grid scores on training set:

0.828 (+/-0.025) for {'C': 1, 'degree': 5, 'gamma': 1, 'kernel': 'rbf'}
0.828 (+/-0.025) for {'C': 1, 'degree': 5, 'gamma': 10, 'kernel': 'rbf'}
0.828 (+/-0.025) for {'C': 1, 'degree': 5, 'gamma': 100, 'kernel': 'rbf'}
0.828 (+/-0.025) for {'C': 1, 'degree': 10, 'gamma': 1, 'kernel': 'rbf'}
0.828 (+/-0.025) for {'C': 1, 'degree': 10, 'gamma': 10, 'kernel': 'rbf'}
0.828 (+/-0.025) for {'C': 1, 'degree': 10, 'gamma': 100, 'kernel': 'rbf'}
0.828 (+/-0.025) for {'C': 10, 'degree': 5, 'gamma': 1, 'kernel': 'rbf'}
0.828 (+/-0.025) for {'C': 10, 'degree': 5, 'gamma': 10, 'kernel': 'rbf'}
0.828 (+/-0.025) for {'C': 10, 'degree': 5, 'gamma': 100, 'kernel': 'rbf'}
0.828 (+/-0.025) for {'C': 10, 'degree': 10, 'gamma': 1, 'kernel': 'rbf'}
0.828 (+/-0.025) for {'C': 10, 'degree': 10, 'gamma': 10, 'kernel': 'rbf'}
0.828 (+/-0.025) for {'C': 10, 'degree': 10, 'gamma': 100, 'kernel': 'rbf'}
0.828 (+/-0.025) for {'C': 100, 'degree': 5, 'gamma': 1, 'kernel': 'rbf'}
0.828 (+/-0.025) for {'C': 100, 'degree': 5, 'gamma': 10, 'kernel': 'rbf'}
0.828 (+/-0.025) for {'C': 100, 'degree': 5, 'gamma': 100, 'kernel': 'rbf'}
0.828 (+/-0.025) for {'C': 100, 'degree': 10, 'gamma': 1, 'kernel': 'rbf'}
0.828 (+/-0.025) for {'C': 100, 'degree': 10, 'gamma': 10, 'kernel': 'rbf'}
0.828 (+/-0.025) for {'C': 100, 'degree': 10, 'gamma': 100, 'kernel': 'rbf'}
0.84
```

Confusion Matrix of SVM

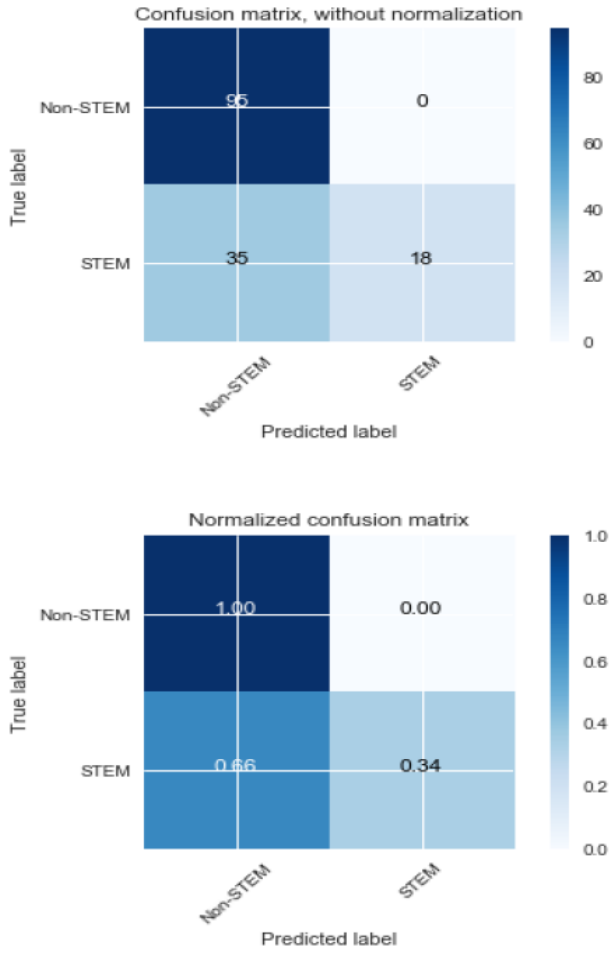


Fig. 3. Confusion Matrix of Trained SVM

Multi Layer Perceptron

```
{'activation': 'relu', 'hidden_layer_sizes': (35, 80, 1), 'learning_rate': 'adaptive', 'max_iter':
1000, 'momentum': 0.5, 'solver': 'adam'}

Grid scores on training set:

0.497 (+/-0.295) for {'activation': 'relu', 'hidden_layer_sizes': (35, 80, 1), 'learning_rate':
'invscaling', 'max_iter': 800, 'momentum': 0.3, 'solver': 'sgd'}
0.666 (+/-0.152) for {'activation': 'relu', 'hidden_layer_sizes': (35, 80, 1), 'learning_rate':
'invscaling', 'max_iter': 800, 'momentum': 0.3, 'solver': 'adam'}
0.613 (+/-0.263) for {'activation': 'relu', 'hidden_layer_sizes': (35, 80, 1), 'learning_rate':
'invscaling', 'max_iter': 800, 'momentum': 0.5, 'solver': 'sgd'}
0.570 (+/-0.266) for {'activation': 'relu', 'hidden_layer_sizes': (35, 80, 1), 'learning_rate':
'invscaling', 'max_iter': 800, 'momentum': 0.5, 'solver': 'adam'}
0.552 (+/-0.308) for {'activation': 'relu', 'hidden_layer_sizes': (35, 80, 1), 'learning_rate':
'invscaling', 'max_iter': 800, 'momentum': 0.8, 'solver': 'sgd'}
0.663 (+/-0.116) for {'activation': 'relu', 'hidden_layer_sizes': (35, 80, 1), 'learning_rate':
'invscaling', 'max_iter': 800, 'momentum': 0.8, 'solver': 'adam'}
0.503 (+/-0.338) for {'activation': 'relu', 'hidden_layer_sizes': (35, 80, 1), 'learning_rate':
'invscaling', 'max_iter': 1000, 'momentum': 0.3, 'solver': 'sgd'}
0.686 (+/-0.093) for {'activation': 'relu', 'hidden_layer_sizes': (35, 80, 1), 'learning_rate':
'invscaling', 'max_iter': 1000, 'momentum': 0.3, 'solver': 'adam'}
0.529 (+/-0.333) for {'activation': 'relu', 'hidden_layer_sizes': (35, 80, 1), 'learning_rate':
'invscaling', 'max_iter': 1000, 'momentum': 0.5, 'solver': 'sgd'}
0.686 (+/-0.106) for {'activation': 'relu', 'hidden_layer_sizes': (35, 80, 1), 'learning_rate':
'invscaling', 'max_iter': 1000, 'momentum': 0.5, 'solver': 'adam'}
```

Confusion Matrix of MLP

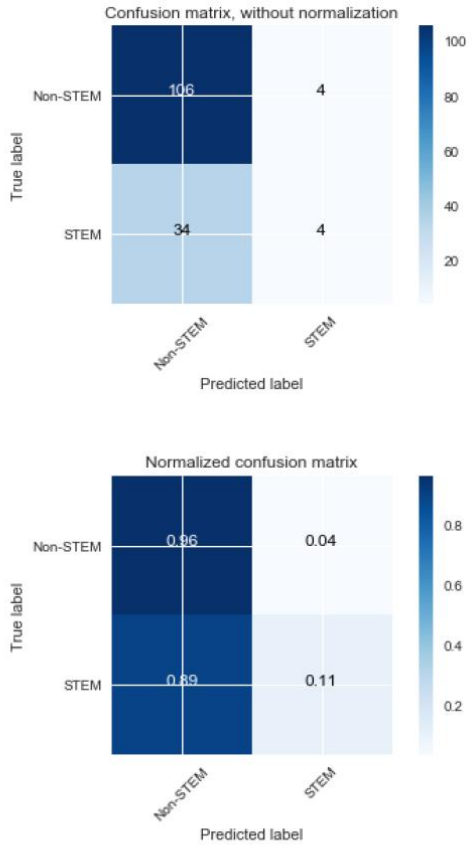


Fig. 4. Confusion Matrix of Trained MLP

In SVM and MLP

- Trained SVM to predict the 30% test data set, finally we got 76.35% accuracy
- In the 30% test data set, 74.32% accuracy with MLP model

Limitation in Supervised Learning

- First of all, we have imbalanced classes in our data set, where total number of isSTEM is 151 and total number of nonSTEM is 341.
- The number of our samples is still small.
- Secondly, as the original log records of students obtained the sequence interaction between students and ASSISTments, there should be much more insight can be discovered.

Future Study in supervised learning

- Deal with imbalanced data, sophisticated resampling methods can be executed to improve prediction accuracy, such as Modified synthetic minority oversampling technique(MSMOTE) and algorithmic Ensemble Techniques, such as Bootstrap Aggregating.
- Extract large amount of samples from web based version of ASSISTments with the help of big data techniques to improve our models.
- Further more, with the sequence interaction records which have specific structural architecture, we could combine time series and deep learning algorithm to discover the variation of knowledge level of students in the process of interaction with ASSISTments by using convolution neutral network or recurrent neural network and other more interesting implication which can be inferred from the records

Unsupervised Learning

- Different from supervised learning, in unsupervised learning part, we use the same set of features as in previous study (for predicting STEM major):

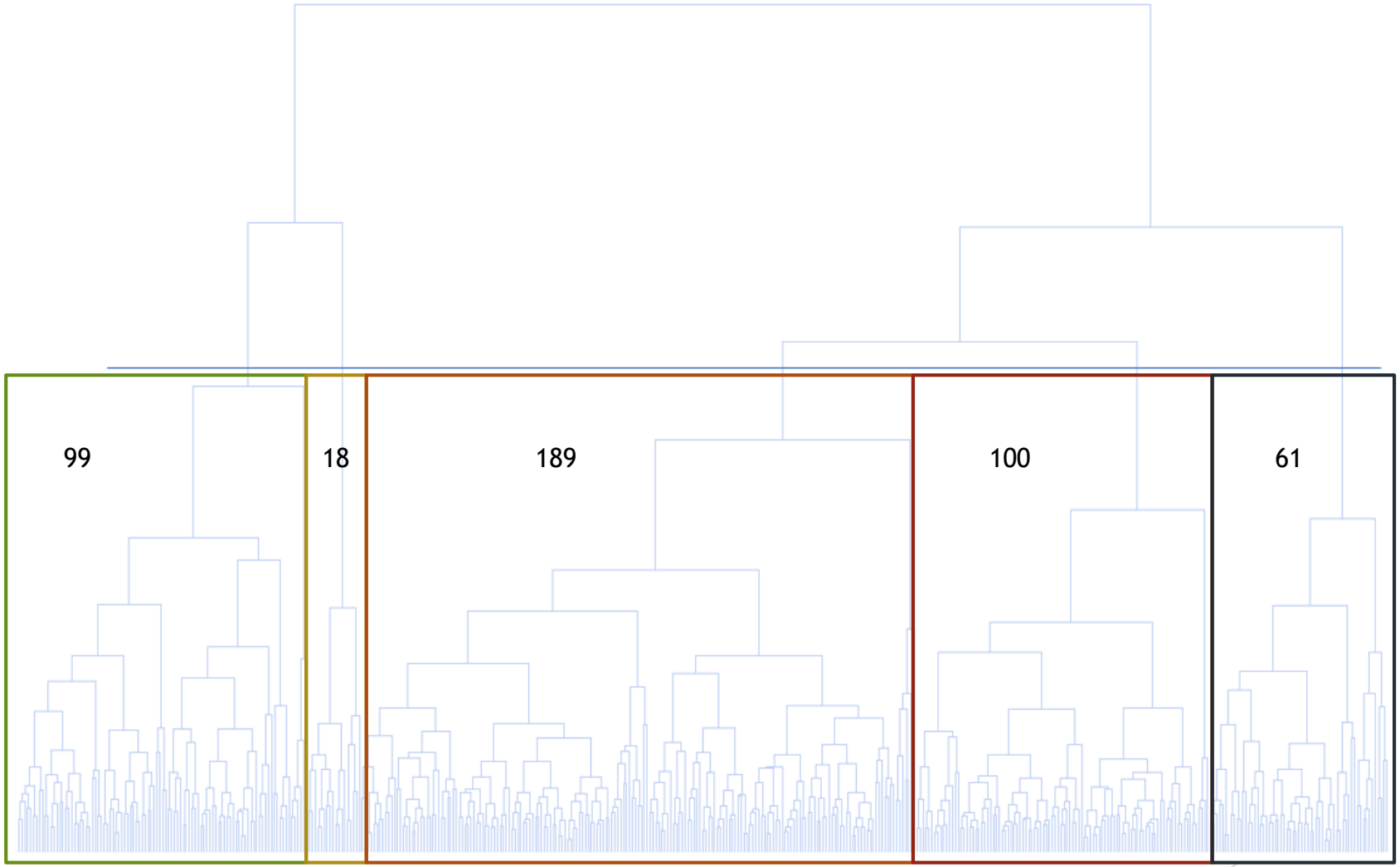
1	knowledge level	6	confusion
2	correctness	7	frustration
3	number of actions	8	off-task behavior
4	boredom	9	gaming
5	engaged concentration	10	carelessness

Unweighted Pair Group Method with Arithmetic Mean (UPGMA):

- Clustering model using the RapidMiner agglomerate clustering operator with measure type as numerical values on Euclidean distances.

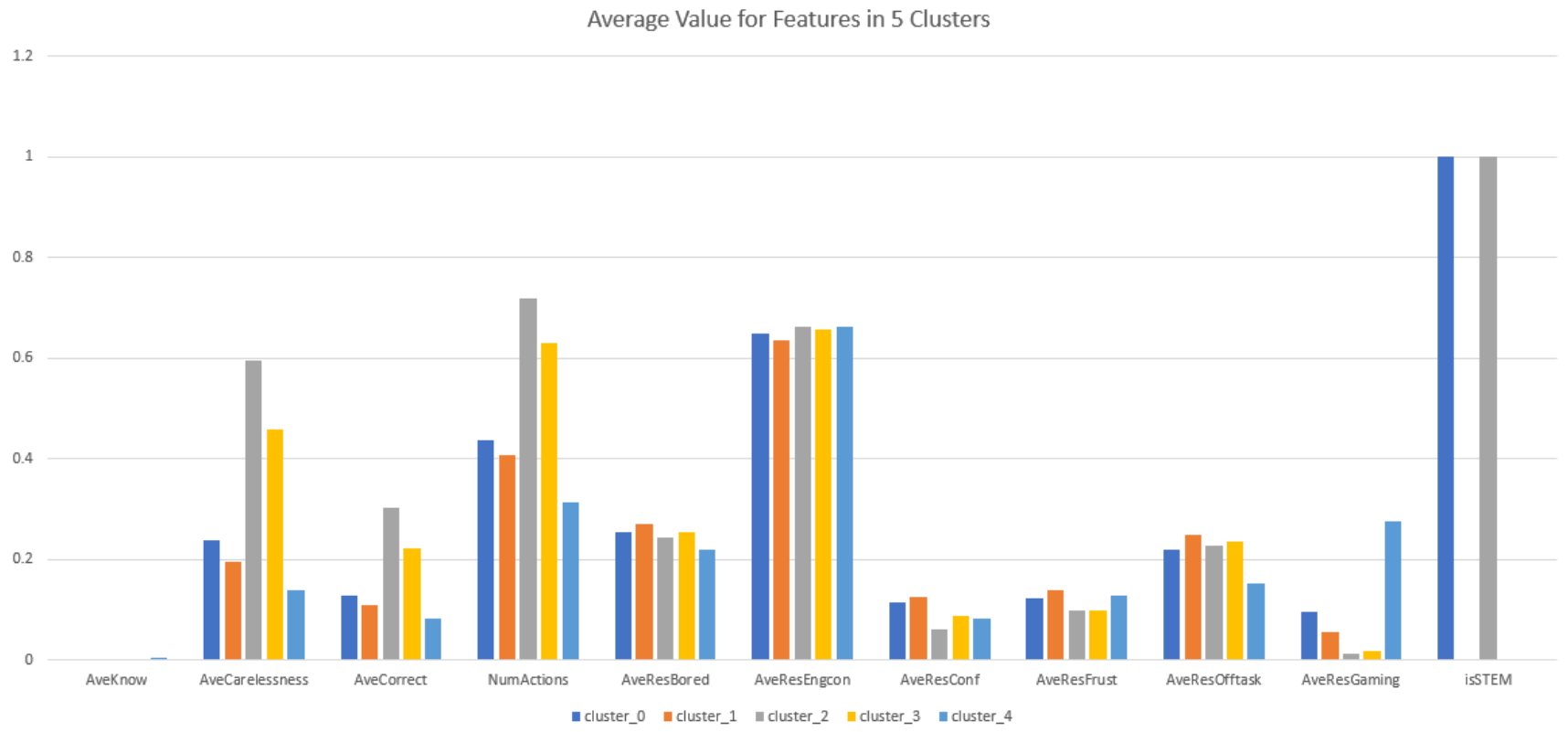
- To get the actual clusters and their corresponding data points, we need to estimate and choose a proper level in the dendrogram to split the hierarchical data structure into smaller clusters.

Dendrogram



Clusters' Statistics

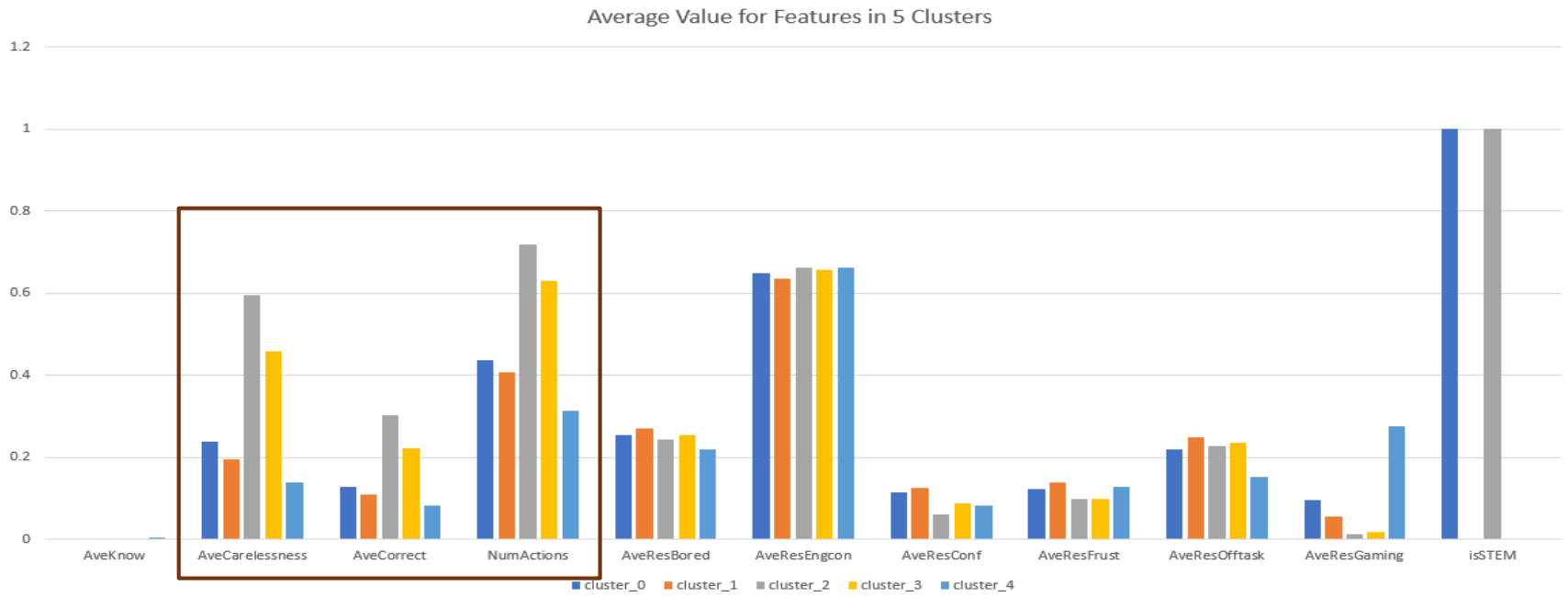
group	0	1	2	3	4
# students	99	189	18	61	100



Clusters' Statistics

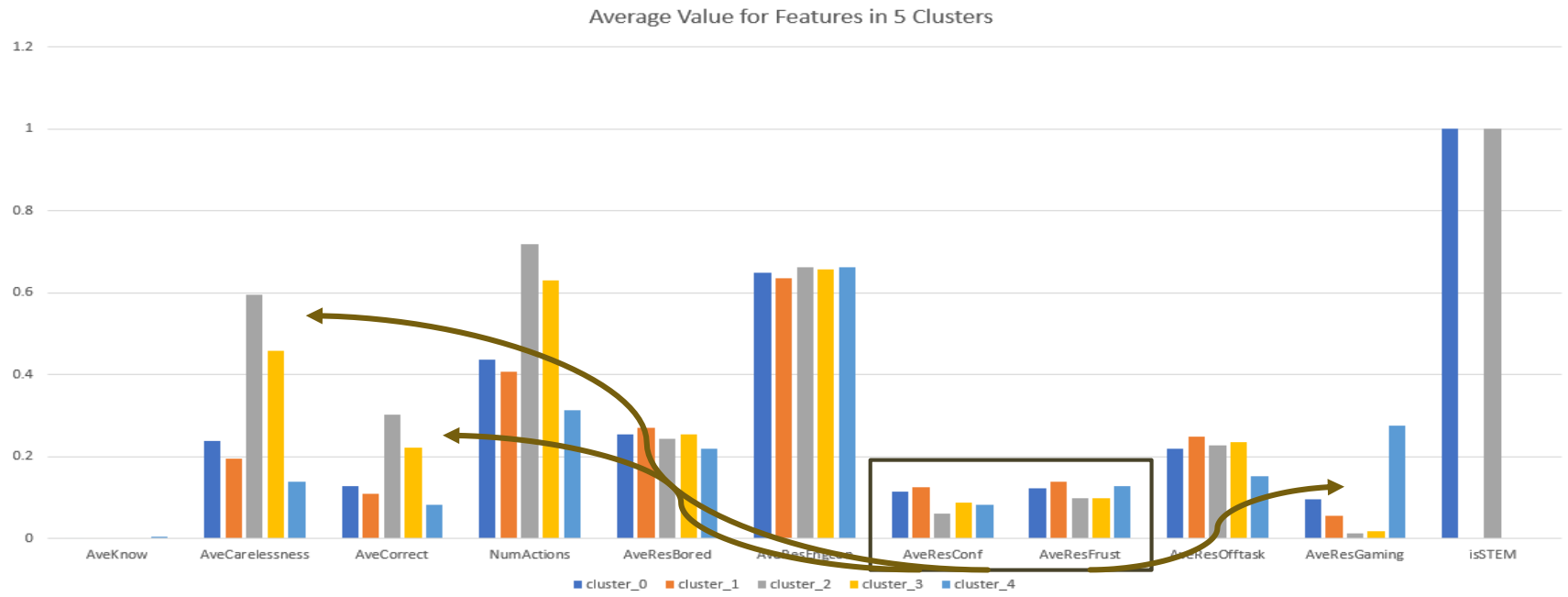
group	# students	characteristic
0	99	careful learners
1	189	confused and frustrated learners
2	18	experimenters choosing STEM majors
3	61	experimenters choosing NON-STEM majors
4	100	gamers

Cluster 0



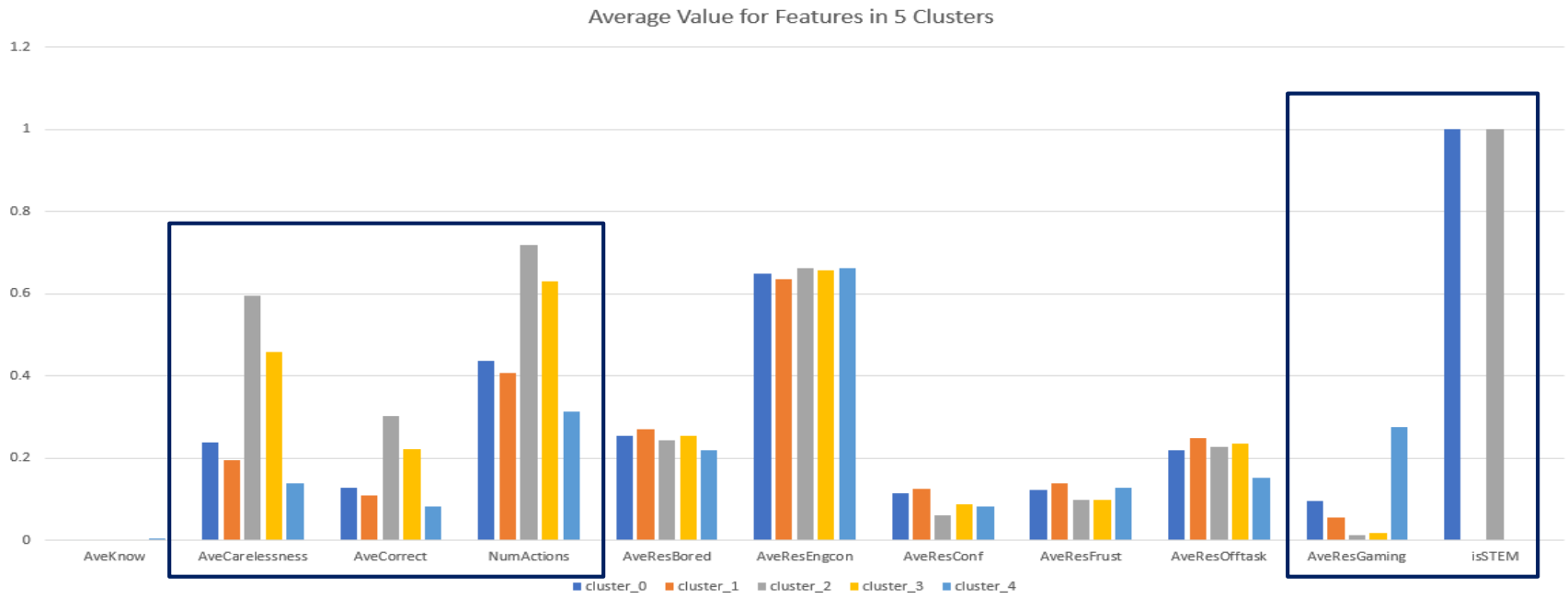
careful learners: careful, not many actions, not many correct problems, chose STEM major

Cluster 1



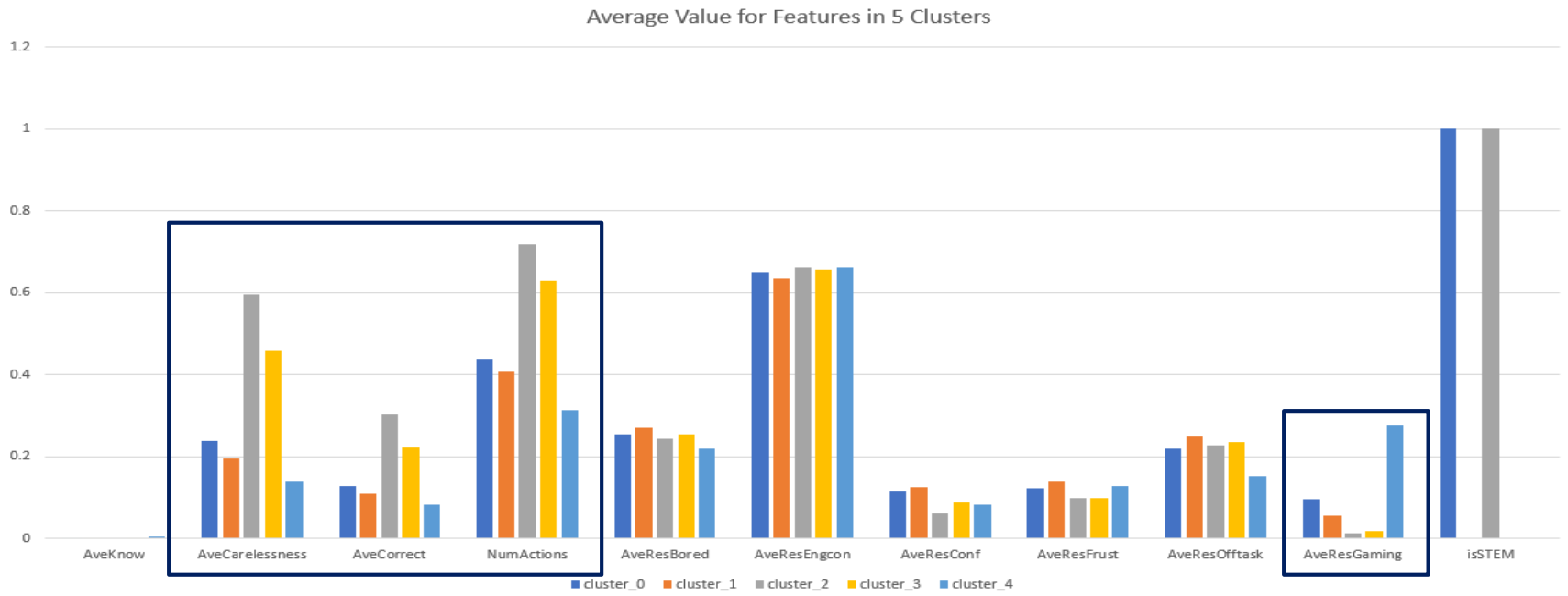
confused and frustrated learners: highest frustration and confusion

Cluster 2



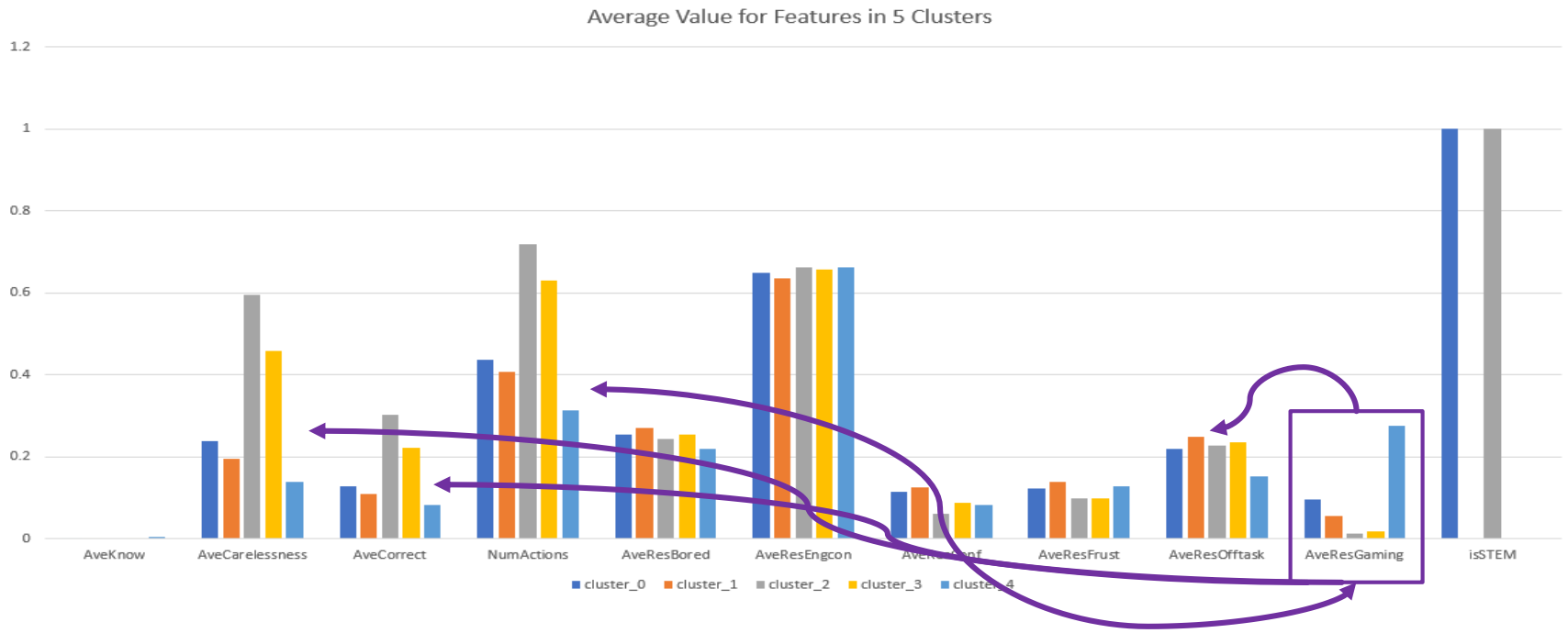
experimenters choosing STEM majors: careless, most actions and correct problems, least gaming, chose STEM major

Cluster 3



experimenters choosing NON-STEM majors: careless, lots of actions and correct problems, almost no gaming

Cluster 4



gamers: careful yet not interested in the tutoring system, often off-task

Thank you