COMP5310 Project Stage 2 - Report

Name: Zhiyuan Cheng

ID: 500222766

UniKey: zche6038

Abstract

With the development of online learning industry, there is an increasing number of people now looking for good online learning platform. Learning outcome is one of the key factors that would attract customers. “PBS KIDS Measure Up!” is an APP that designed for early age child to learn STEM topics [1]. This project will focus on find and estimate factors that could potentially be used for interpreting learning outcome of app users from historical learning data using classification.

Setup

For stage 2 of this project, a research question is to visualise and quantify the relationships between learning outcome and features. The null hypothesis is that past assessment history will affect the final learning outcome, whereas the alternative hypothesis is that past assessment history does not affect the final learning outcome.

Also, another research question is to apply different classification models on the training set and test which one is better, by using paired McNemar’s test and compare f1-scores.

Approach

In state 1 of the project, a dataset contains 10382 rows and 112 features was extracted. In order to fit classification models, dummy variables are used for 3 categorical features, which result in 136 features. Among those features 9 of them are discarded from training set as they either contains device data, timestamp or data used for calculating true labels in stage 1. As the result, the dataset now has 128 features. Since grid search cross-validation will be used, the data is split to training and testing data with a ratio of 0.8. However, as there are a lot of hyper parameters to choose from, error-complexity plot was plotted prior to grid search for speeding up the grid search process.

To explore the importance of features, decision tree classifier was chosen, and it will also be used as the baseline model for the estimation. For the parameter, max depth gets tested against 1 to 9, criterion is chosen between ‘entropy’ and ‘gini’ and splitter method is chosen between ‘best’ and ‘random’.

For the classification problem, in addition to the decision tree classifier above, three more classifiers of different techniques are chosen: two ensemble methods: AdaBoost Classifier and Random Forest Classifier and a nearest neighbors method: K Neighbors Classifier.

Each classifier also uses grid search cross-validation to turn the hyper parameters. For AdaBoost classifier, use error-complexity plot to decide number of estimates and then choose learning rate from [10^{-3}, 10^{-2}, 10^{-1}, 1] using grid search.

For Random Forest classifier, use error-complexity plot to decide number of estimates and C (regularization parameter) are choosing from available choices. Criterion is chosen between 'entropy', maximum number features is chosen between square root and base 2 logarithm and class weight the chosen between 'balanced', 'balanced subsample' and 'None' (uniform).

Then for the K Neighbour classifier, first use error-complexity plot to decide number of neighbors, then use grid search to choose leaf size among [10, 20, 30, 40, 50] and distance function for Minkowski metric between Manhattan distance and Euclidean distance.

Lastly, for the comparison of different classification models, McNemar's test is used for pair-wised comparison between each pair of models.

Result

Decision Tree Classifier

The best parameter for decision tree classifier has a depth of 6, uses gini impurity as criterion and split on 'best' value. Figure 1 shows a compete graph of decision tree, and three subparts can be found in the appendix. The root of this decision tree classifier split on the average of historical accuracy, which indicates that this gives the lowest gini impurity, thus the most important factor in the classification.

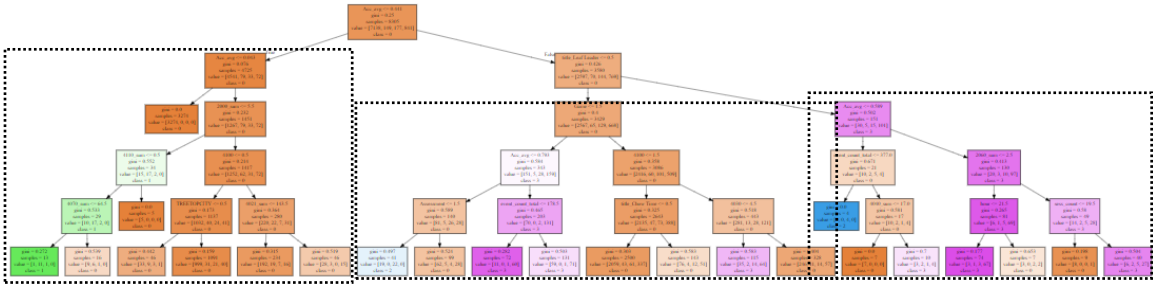

Figure 1 Full decision tree. See Figure 1(a), (b), (c) in Appendix for detailed subpart.

This is further supported by the looking at the feature importance of this model in Figure 2, when feature 'Acc_avg' has the highest feature importance value of 47.2%. However, despite that features 'title_Leaf Leader' rank the second on this list, its hard to give an explanation on its importance as 'title_Leaf Leader' is just a dummy variable represent one of many game titles. It is not seeming special compare to other titles by Appendix I, Figure 2(a) suggests that there could be some hidden correlation between features. On the other hand, the third feature on the list 'Game' is explainable as most of data is related to game event from stage 1's exploration. Also, among all of used features in the figure

```
[('Acc_avg', 0.47186434421601736),
 ('title_Leaf Leader', 0.14484520860975122),
 ('Game', 0.12163936254502887),
 ('4030', 0.056445128042438286),
 ('4100', 0.04341021935752289),
 ('title_Chow Time', 0.03524762853480693),
 ('2000_sum', 0.025529184622731772),
 ('event_count_total', 0.02314090246537647),
 ('Assessment', 0.0194193861525324),
 ('sess_count', 0.012218038468430507),
 ('4021_sum', 0.0081490091509724),
 ('2060_sum', 0.007245052323811948),
 ('hour', 0.0071725432427105626),
 ('4110_sum', 0.006252756050341388),
 ('4070_sum', 0.006193148207730218),
 ('TREETOPCITY', 0.005793735583659002),
 ('4040_sum', 0.005434352506137684)]
```

Figure 2 List of sorted important features

2, 9 out of 17 features are historical data, which in total contribute to about 60% of the importance.

The overall weighted average of f1-score for this decision tree classifier is 0.86 which is a rather high value.

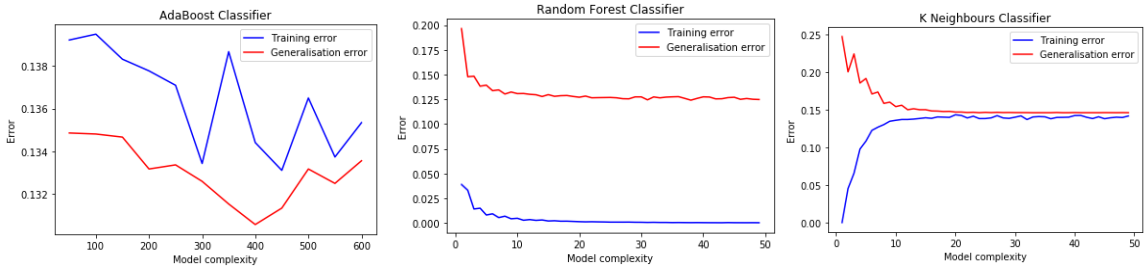

Figure 3 Error V.S. number of estimators

AdaBoost Classifier

An error-complexity plot was produced to find a desire number of estimators with learning rate of 0.1. The complexity is defined by number of estimators in a range between 50 and 600 and a step of 50 is chosen. The result in Figure 3 (a) suggests that 300 estimators is good enough. Then, perform grid search on AdaBoost classifier result in a learning rate of 0.1 and the overall weighted average of f1-score is 0.80.

Random Forest Classifier

An error-complexity plot was produced using number of estimators in a range between 1 and 50. After the grid search, the best parameter is uniform class weight, gini impurity criterion and square root for calculating maximum features. The overall weighted average of f1-score is 0.86.

K Neighbors Classifier

An error-complexity plot was produced using number of neighbours in a range between 1 and 50. From Figure 3 (b), we can see that 10 neighbours seems produce a balanced training and generalization error. After the grid search, the best parameter for K Neighbours classifier is 10 leaves and use Manhattan Distance for compute Minkowski matrix. The overall weighted average of f1-score is 0.80.

Compare different classifiers

P value / Test statistic	*Decision Tree*	*Ada Boosting*	*Random Forest*	*K Neighbors*
Decision Tree	/	0.01174338	0.82826254	0.00149112
Ada Boosting	6.34920635e+00	/	0.02144822	0.1138463
Random Forest	4.70588235e-02	5.29000000	/	0.00445953
K Neighbors	1.00895522e+01	2.50000000	8.08653846	/
f1-score	0.86	0.80	0.86	0.80

Table p-value result of paired-wise McNemar's Test and f1-score

The test statistics and p-values of pair wise McNemar's Test is given in the table. As we can see, decision tree classifier and random forest classifier is not significantly different, and ada boosting classifier and k neighbors classifier is also not significantly different. However, other pairs of classifiers are significantly different. This result shows high similarity to the f1-score as the two statically similar pair has the same f1-score.

Conclusion

For the first research question, the aim is to determine if past assessment history is an important factor for estimate the learning outcome. Since the decision tree shows that the most important feature is historical data, and historical data contribute to 60% of importance in the classification, we can conclude that past assessment history is an important factor of estimate learning outcome.

On the other hand, in compare different classification method, we can conclude that decision tree classifier and random forest classifier are both good choice for this estimation with f1-score of 0.86. While ada boost classifier and k neighbours classifier perform similar to each other with a f1-score of 0.80, they are not the best choices compare to the other two classifiers. For the best classifier, decision tree classifier outperforms random forest classifier as it can train much faster.

However, looking at classification report of those classifiers in Figure 4 to 7 in the Appendix, its worth notice that data is not balanced. Most of users is in accuracy group 0 (this means they did not successfully complete an assessment in three trails) and all these classifiers can classify class 0 every well. This is not the same case for other classes. Only decision tree classifiers can identify all the other classes with a rather low quality (f1-score). Therefore, despite that the overall weighted f1-score of decision tree is a high value, the current classifier is still not ready as a solution for production.

Reference

1. https://www.kaggle.com/c/data-science-bowl-2019/overview

Appendix

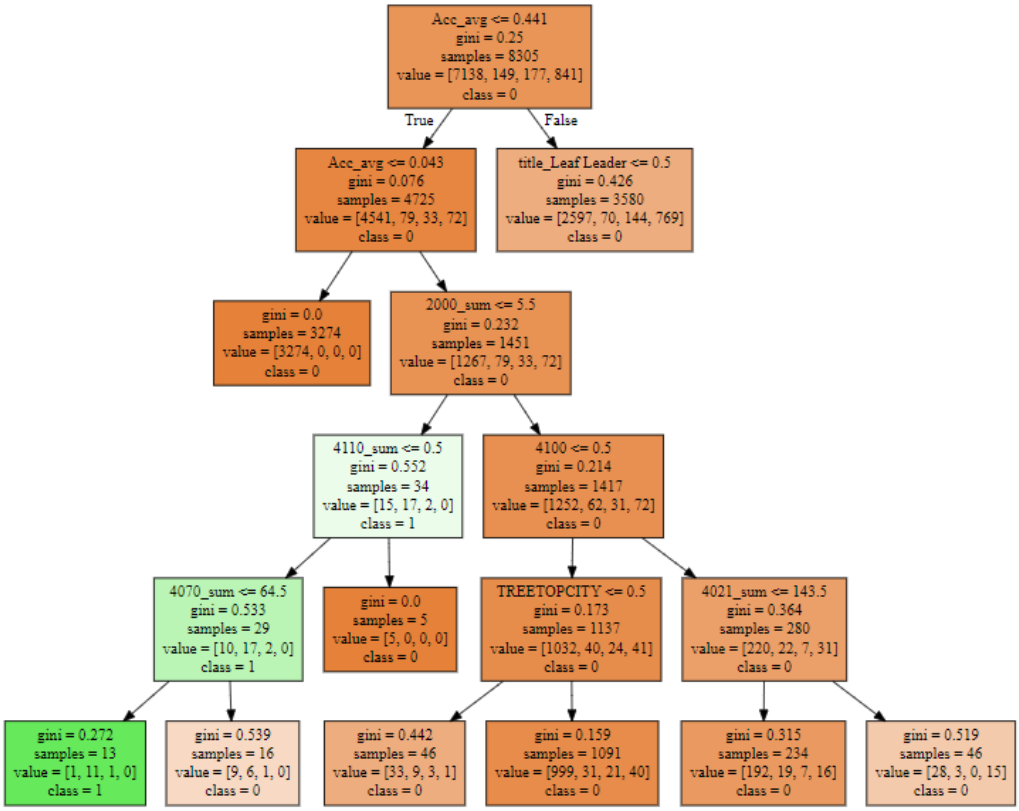

Figure 1 (a) Decision tree (left)

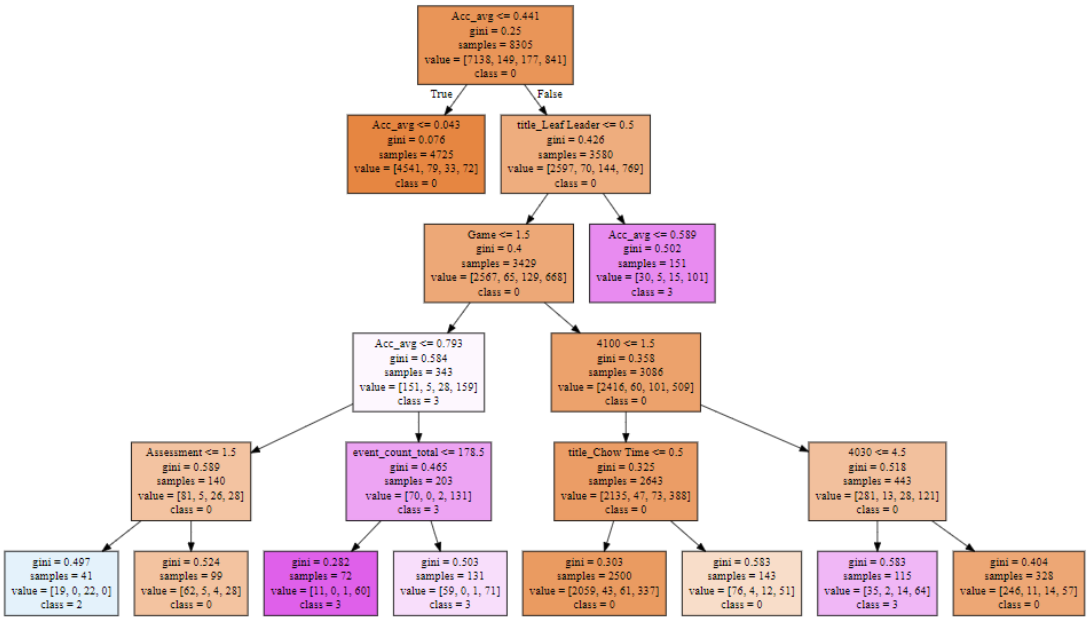

Figure 1 (b) Decision tree (middle)

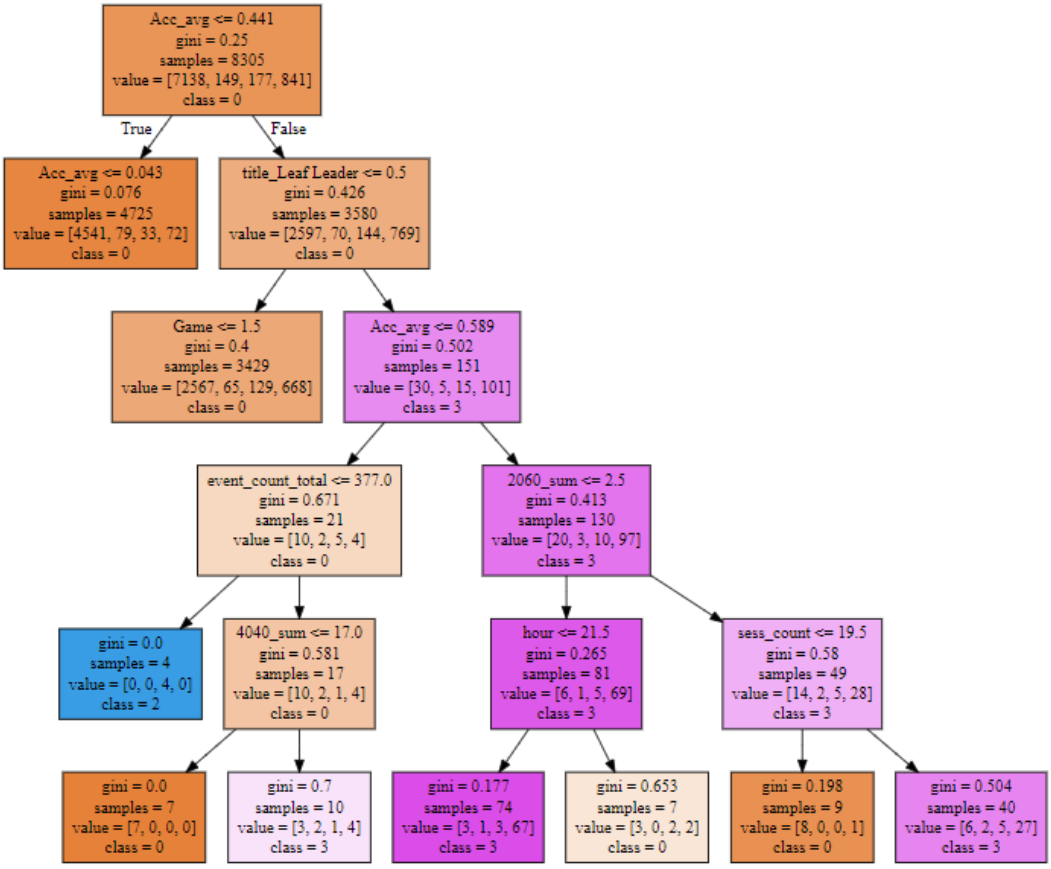

Figure 1 (c) Decision tree (right)

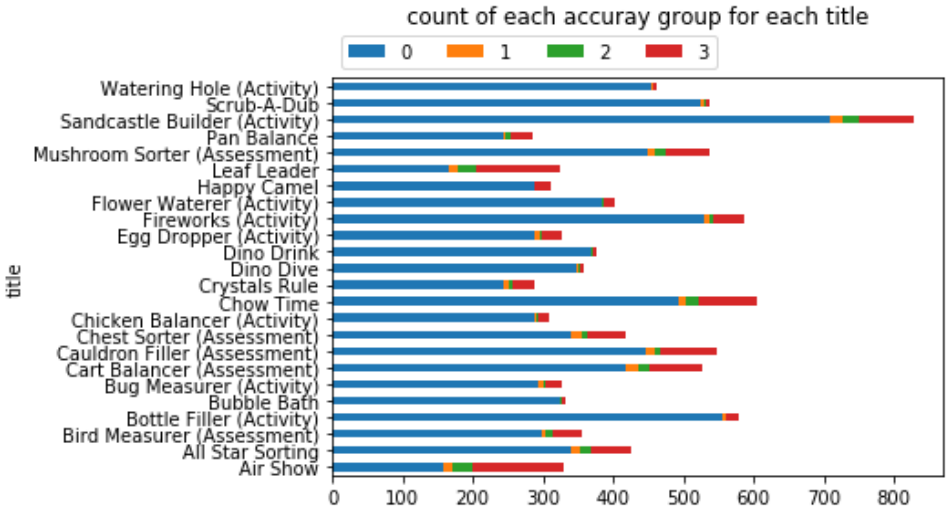

Figure 2 (a) Stats of accuracy group by title

```
Best parameters: {'criterion': 'gini', 'max_depth': 6, 'splitter': 'best'}
              precision    recall  f1-score   support

           0       0.90      0.98      0.94      1786
           1       0.11      0.03      0.05        35
           2       0.42      0.14      0.20        37
           3       0.73      0.34      0.46       219

    accuracy                           0.88      2077
   macro avg       0.54      0.37      0.41      2077
weighted avg       0.86      0.88      0.86      2077
```

Figure 4 classification report of decision tree classifier

```
Best parameters: {'learning_rate': 0.1}
              precision    recall  f1-score   support

           0       0.86      1.00      0.93      1786
           1       0.00      0.00      0.00        35
           2       1.00      0.03      0.05        37
           3       0.88      0.03      0.06       219

    accuracy                           0.86      2077
   macro avg       0.68      0.26      0.26      2077
weighted avg       0.85      0.86      0.80      2077
```

Figure 5 classification report of ada boost classifier

```
Best parameters: {'class_weight': None, 'criterion': 'gini', 'max_features': 'sqrt'}
              precision    recall  f1-score   support

           0       0.90      0.98      0.94      1786
           1       0.00      0.00      0.00        35
           2       0.46      0.16      0.24        37
           3       0.62      0.32      0.42       219

    accuracy                           0.88      2077
   macro avg       0.50      0.37      0.40      2077
weighted avg       0.85      0.88      0.86      2077
```

Figure 6 classification report of random forest classifier

```
Best parameters: {'leaf_size': 10, 'p': 2}
              precision    recall  f1-score   support

           0       0.86      1.00      0.92      1786
           1       0.00      0.00      0.00        35
           2       0.00      0.00      0.00        37
           3       0.00      0.00      0.00       219

    accuracy                           0.86      2077
   macro avg       0.21      0.25      0.23      2077
weighted avg       0.74      0.86      0.80      2077
```

Figure 7 classification report of knn classifier