COMP5310 Project Stage 1 - Report

Name: Zhiyuan Cheng

ID: 500222766

UniKey: zche6038

Problem

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. This project will focus on analysis factors that could potentially be used for interpreting learning outcome of app users from historical learning data including activities during learning period, past learning experience, past assessment experience.

The result of this project would provide insight into understanding relationship between learning factors and its outcome. Those insight could be used for developing better learning technique or material, identify user’s learning efficiency in different time, identify user’s personalized learning strategies and so on. Note that the analysis in this project could be further extended to other online early learning applications, as well as all the other online learning platforms.

Data

The dataset used is a public dataset from 2019 Data Science Bowl Competition on Kaggle [1]. It can be acquired directly from its web page or using API [2]. The dataset is provided by Booz Allen Hamilton, and it contain data gathered from “PBS KIDS Measure Up!” APP while all personal information has been changed to random generated strings.

The dataset itself contain 5 csv files, and the training file contains 11.3 million rows. Because of the size of this project, only part of the training data will be used, and those data should be enough for training, validation and testing in stage 2. The training set contains 11 columns and each row in the training set is one event happened during a session. For this project, we are interested in data of each session as well as data of history sessions for each user. As the result, some cleaning and transformation are needed.

According to the dataset description [2], there are some users in the training set that did not participate any assessment. Therefore, the first step of cleaning is to remove rows with those users, since at least one assessment will be needed for compute the ground truth. This step removed about 3 million rows in the training set and reduce number of users from 17000 to 4242. In order to reduce the data size, 500 users are randomly selected from the total 4242 users for this project. After that, sessions that only contain one event are also removed from the data set. The reason is that every session has a welcome event indicate user launched a session, and a session with one event indicate this session has no valuable data. At this stage, the dataset contains 500 users, around 10 thousand sessions and around 1 million events. Since we are only interested in the game sessions, the number of dataset size should be sufficient.

After the data cleaning and reducing process, we need to extract features for each session and construct the final dataset for this project. There are 112 features extracted from each session: including 6 features about this session, 2 features about history of sessions of this user, 5 features about time, 6 features about the type and world of this session, 84 features about events happened during this session and history of sessions of this user and 9 features about the performance of this session if the session is an assessment. Those features can be view in the stage 1 code as the whole process of data cleaning and feature extraction are done using python with the help of the pandas library.

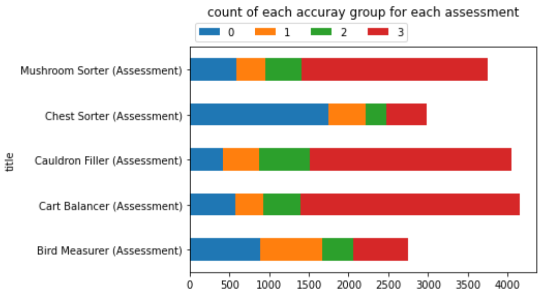

Figure 2: distribution of accuracy group by assessment

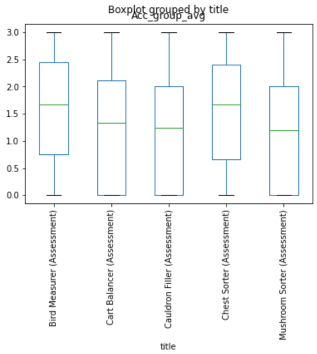

Figure 2: boxplot of accuracy group by assessment

Figure 1 gives the distribution of accuracy group by each type of assessment. It is the ground truth provided in the dataset, where 3 indicate the first attempt solved the assessment, 0 indicate the assessment was never solved and 1 and 2 lays in the middle. As we can see, the distribution varies in different assessment type. In figure 2, a boxplot of accuracy of the training data seems to showing a complete reverse distribution. The difference here is worth further analysis. In addition, some other graphs regarding other factors can be found in the code.

Proposal

For stage 2 of this project, the goal is to understand, visualise and quantify (if applicable) the relationships between learning outcome and features. If possible, set up machine learning models to estimate the learning outcome from extracted features.

Analysis effect of assessment type features would be prioritized. The hypothesis is that 'different assessment type effect the accuracy'. Also, compare past assessment history and occurrence of event to accuracy group may also be tested in stage 2. One possible hypothesis could be 'higher accuracy in the past assessment would help the child to have higher accuracy in future assessment'. In addition, feature selection method may be included to reduce number of features such that the training time can be minimized.

Reference

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