OULAD Dataset Analysis

1. Link & Usage

Original dataset: https://analyse.kmi.open.ac.uk/open_dataset
GitHub repository: https://github.com/ZhuoyueWang/OULAD_Analysis
Preprocessed & Sequencing files: https://drive.google.com/drive/folders/1l431ud1P-VrXPbHdpwn_rcbmqR5oT3FF?usp=sharing

2. Project Description

The main goal is to have a workable unsupervised learning model to predict students' next action based on the dataset. To be more specifically, we want to evaluate how accurate the model, that was trained by all data except last actions of students, can predict the next action.
However, because of my misunderstanding on the project goal, I mistakenly got accuracy by not using sequence learning. After realizing this problem, I decided to follow the semisup example and use Keras to get the MSE value for the model.
The main process of this project is like following:

1. Concatenating several csv files to get the expected format we need for unsupervised learning.
2. Generate sequencing files for the preprocessed dataset.
3. Run the unsupervised learning model and get the final mse value.

3. Data Description

I have attached my preprocessed data and sequencing data in the link above. For the sequencing files, I used the built-in make_sequence function (refers to your semisup-example). For the preprocessed data, my idea is like below:
To get the expected information, I made concatenation from studentVle.csv, vle.csv and studentInfo.csv. Eventually, I got students' demographic attributes, clicking record and their actions. For all categorical variables, I transform them using one hot encoding. My understanding for this project is to use all features in the preprocessed data to predict the activity_type entry (the values contain resource, oucontent, url and etc).
In total, the files contain 69 columns.

4. Result

I didn't get the MSE value for the data file around 2500000 rows because it is too large. The Keras couldn't complete that. (My terminal killed this process). However, I got 229.0347 under 100000 rows in the third epoch. I will raise my questions in the section below.

5. Confusion & Trouble

1. I am not sure whether I am approaching the model fitting correctly. After reviewing the example, I think the y-value of the unsupervised learning model on the UCI_EPM dataset is activity_editor, which is one of the activity types. Its value are either 1 or 0. The reason why my model failed could be that it can only get really large values on this entry, that are far away from 1 or 0. However, I think this problem should be easy to be solved, because as long as I receive some clarification about the y-value, I can just change the selected values for y-value entry for the model.fit function.
2. Because the data files are really large, and my AWS is expired. I only got and uploaded 1/4 completed data files (around 2500000 rows).
3. Due to the large data size, my local machine killed my process which was running the model or generating the sequencing files sometimes.

6. Further extension

1. It would be much better if we could have the supervised learning model to perform the prediction.
2. We can try different ways to fit the unsupervised learning model such as LSTM autoencoders.