CIS400:600 ML

Graduate Project

Team members: Huahao Shang, Hongyang Chen, Ziyun Zhang, Yuxuan Hao

What we have done:

1. **First Look:** We simply looked at the whole data sets(data_set.csv, student_test.csv, sku_info_40.csv and sku_attr_40.csv). Understanding all those features columns, data rows, etc.

2. **Observation:** Then we focus on the data_set.csv, the main data. We did pre-processing and found out that there were large number of data that were missing 'discounts' and 'Original price'. Then, we also looked at the stduent_test.csv, and again, info missing also exists just like the main data.

3. **First Attempting**: The first attempt is only creating one model and drop the features 'discount' and 'original price'. The prediction contains large result that are not between 2-10. And the correction is pretty bad around 15%.

4. **Second Attempting:** To improve, we figured out another method that is to split the situations with the data that have missing features and the data that don't have missing features. And also filter the main data by sku-id from 2-10. Because the student_test.csv only contains sku_id between 2-10. This operation may increase the accuracy from our perspective. Then, making two model based on different situation, here we used RandomforestClassifier() and did the normal data split, scale and fitting. For student_test.csv, also split it based on whether have 'discount' &'original price' or not.

5. **Prediction Logistic**:For prediction, using Model 1 to predict the data that has missing data, using Model 2 to predict the data that doesn't have missing data. Then combine two prediction result together, and sort the dataset based on 'serial_number'

6. **Prediction Result**: the result is better than the first attempting, first the prediction dataset contains more result from 2-10 compare to the first attempting prediction. And the actual accuracy is also boosted to 51%.