NUID: 002773368

Name: Yuxuan Cheng

This is the assignment for assignment – spark 2.

Important Notes:

This project uses scala 2.13 and spark 3.3.2. The code should be run with JDK 8. (Otherwise, it aborts tasks and raise errors).

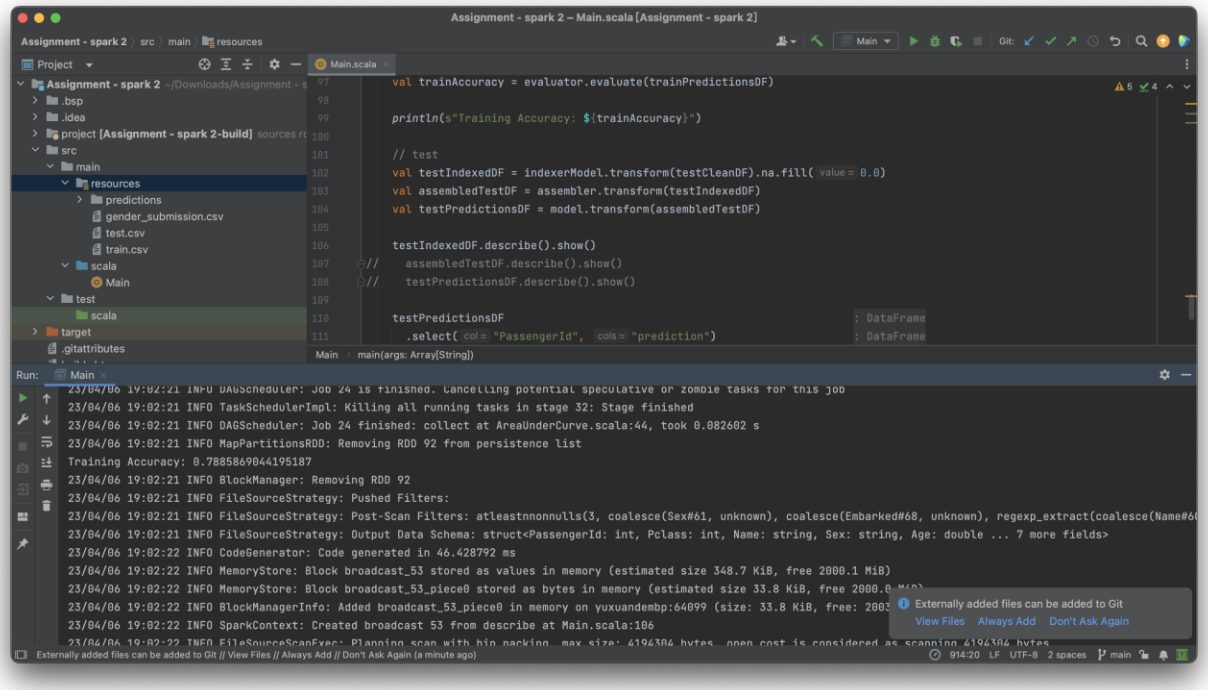
One problem in this project that is not solved is: after testing the output data frame has 1 less rows than the original test.csv(which has 419 rows). I have checked the null values carefully to ensure that no rows should be automatically deleted by the trasform method. But that does not work.

As a result, users have to add a row manually between the row 415 & row 416 in the output file: predictions.csv.

Methodology:

As an overview of the whole project, the main function starts with reading “train.csv” and “test.csv” files under directory: “src/main/resources/”. Then we get 2 data frames and deal with them by identifying useful properties, setting the features, filling empty values with default values, and dropping the useless columns. After that we define vector assembler that include all required columns, and use the linear regression to start training. Afterwards we apply the model to the test.csv and output our results to a csv file.

Result: the training accuracy is: 0.7885869044195187.



After filling the 1306 column in the test case, the output csv file got accuracy of around 0.78 on Kaggle test case.