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.

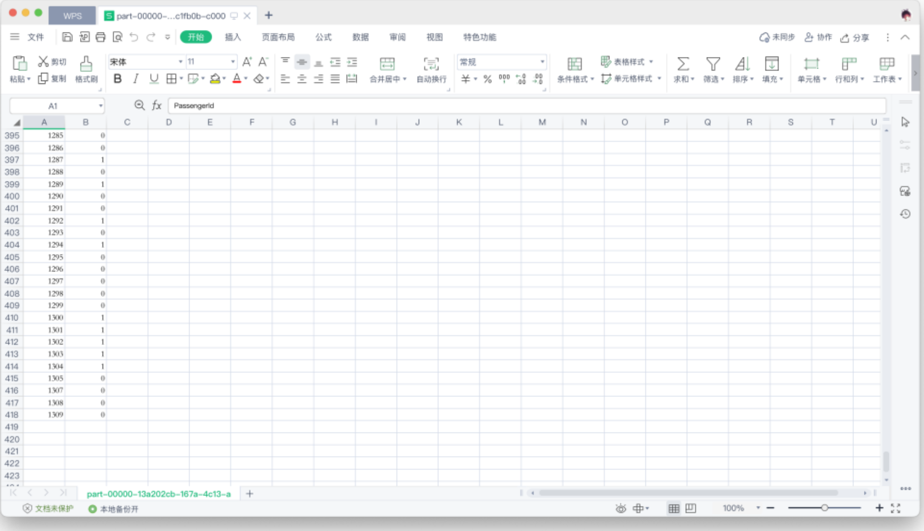
As a result, users have to add a row manually between the row 415 & row 416 in the output file: predictions.csv.

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.

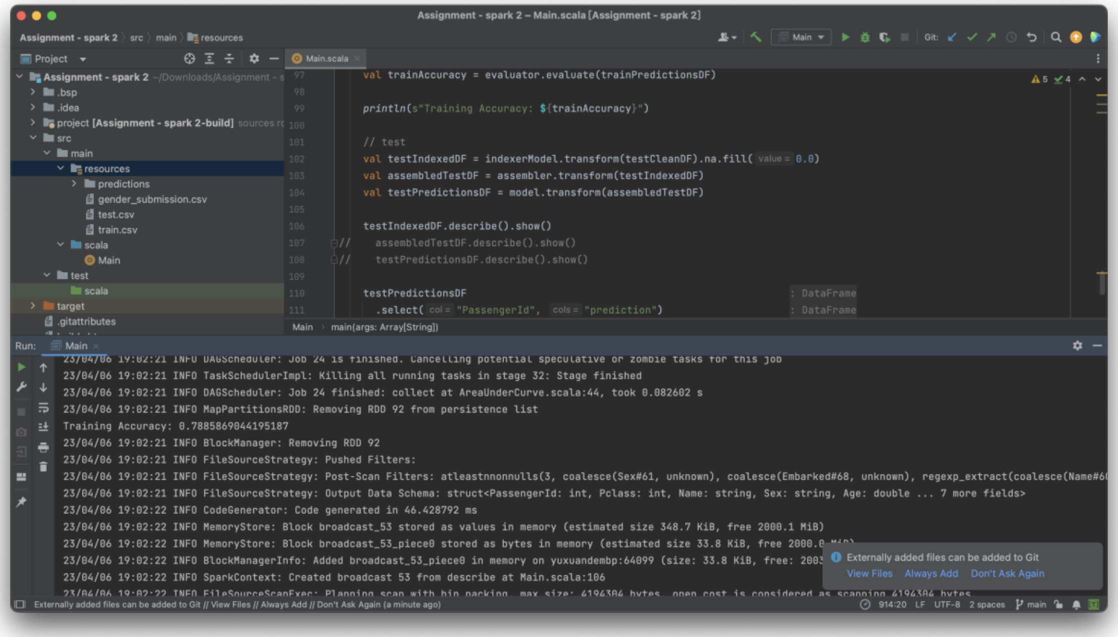
Results:

The output predictions.csv file that stores results:



	A	B
395	1285	0
396	1286	0
397	1287	1
398	1288	0
399	1289	1
400	1290	0
401	1291	0
402	1292	1
403	1293	0
404	1294	1
405	1295	0
406	1296	0
407	1297	0
408	1298	0
409	1299	0
410	1300	1
411	1301	1
412	1302	1
413	1303	1
414	1304	1
415	1305	0
416	1307	0
417	1308	0
418	1309	0
419		
420		
421		
422		
423		

The training accuracy is: 0.7885869044195187.

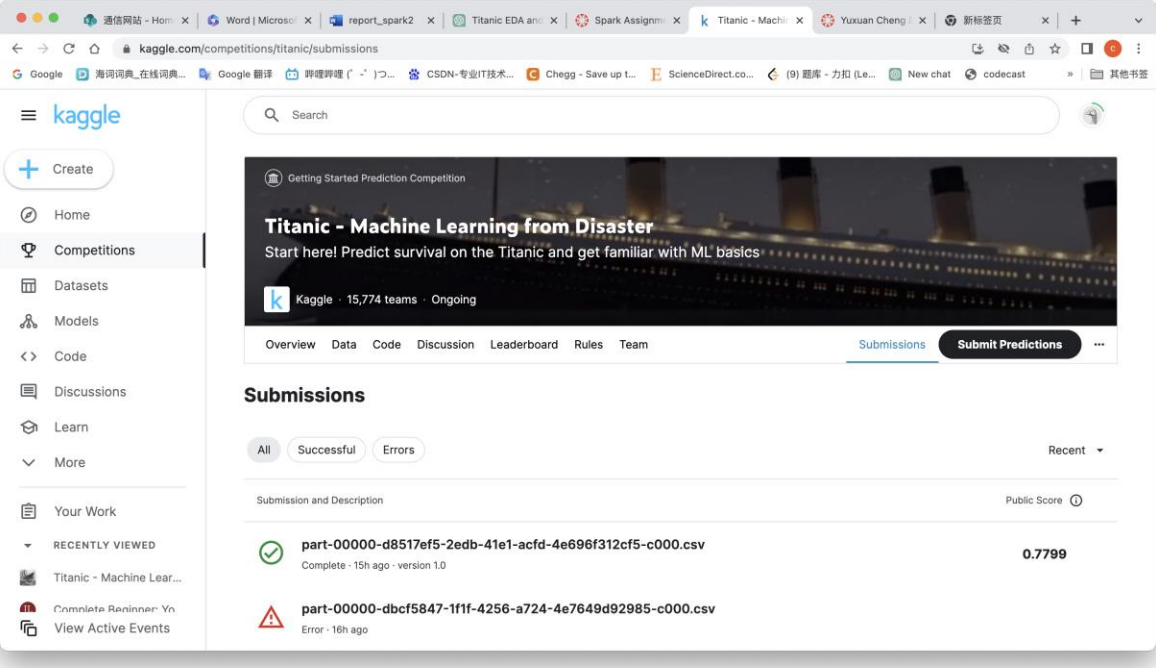


```scala
val trainAccuracy = evaluator.evaluate(trainPredictionsDF)

println(s"Training Accuracy: ${trainAccuracy}")

// test
val testIndexedDF = indexerModel.transform(testCleanDF).na.fill( value = 0.0)
val assembledTestDF = assembler.transform(testIndexedDF)
val testPredictionsDF = model.transform(assembledTestDF)

testIndexedDF.describe().show()
//   assembledTestDF.describe().show()
//   testPredictionsDF.describe().show()

testPredictionsDF
  .select( col = "PassengerId",  cols = "prediction")
```

```
23/04/06 19:02:21 INFO DAGScheduler: Job 24 is finished. Cancelling potential speculative or zombie tasks for this job
23/04/06 19:02:21 INFO TaskSchedulerImpl: Killing all running tasks in stage 32: Stage finished
23/04/06 19:02:21 INFO DAGScheduler: Job 24 finished: collect at AreaUnderCurve.scala:44, took 0.082602 s
23/04/06 19:02:21 INFO MapPartitionsRDD: Removing RDD 92 from persistence list
Training Accuracy: 0.7885869044195187
23/04/06 19:02:21 INFO BlockManager: Removing RDD 92
23/04/06 19:02:21 INFO FileSourceStrategy: Pushed Filters:
23/04/06 19:02:21 INFO FileSourceStrategy: Post-Scan Filters: atleastnnonnulls(3, coalesce(Sex#61, unknown), coalesce(Embarked#68, unknown), regexp_extract(coalesce(Name#6
23/04/06 19:02:21 INFO FileSourceStrategy: Output Data Schema: struct<PassengerId: int, Pclass: int, Name: string, Sex: string, Age: double ... 7 more fields>
23/04/06 19:02:22 INFO CodeGenerator: Code generated in 46.428792 ms
23/04/06 19:02:22 INFO MemoryStore: Block broadcast_53 stored as values in memory (estimated size 348.7 KiB, free 2000.1 MiB)
23/04/06 19:02:22 INFO MemoryStore: Block broadcast_53_piece0 stored as bytes in memory (estimated size 33.8 KiB, free 2000.0
23/04/06 19:02:22 INFO BlockManagerInfo: Added broadcast_53_piece0 in memory on yuxuandembp:64099 (size: 33.8 KiB, free: 2003
23/04/06 19:02:22 INFO SparkContext: Created broadcast 53 from describe at Main.scala:106
23/04/06 19:02:22 INFO FileSourceScanExec: Planning scan with bin packing, max size: 4194304 bytes, open cost is considered as scanning 4194304 bytes
```

Externally added files can be added to Git // View Files // Always Add // Don't Ask Again (a minute ago)

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



Getting Started Prediction Competition

Titanic - Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics

Kaggle · 15,774 teams · Ongoing

Submissions

Submission and Description	Public Score
part-00000-d8517ef5-2edb-41e1-acfd-4e696f312cf5-c000.csv Complete · 15h ago · version 1.0	0.7799
part-00000-dbcf5847-1f1f-4256-a724-4e7649d92985-c000.csv Error · 16h ago	