Big Data Laboratory

Week 5

Divya K Raman

EE15B085

In this assignment, we learn to build a custom transformer and estimator using sparkML. The goal of this assignment also includes understanding how spark-submit works.

Question 1 requires us to build a custom Estimator for normalizing a column in the house_price.csv and include it in a Pipeline to train a model. The trained model is then used to generate predictions for house_price. The code for the same can be found in mainCode.py. All features except the ID were used. Output column is SalePrice. Categorical variables are converted using StringIndexer.

Missing numerical values were imputed with the mean of the respective columns using imputer. A custom estimator and transformer are used to normalised the data. Linear regularisation along with L1 and L2 regularisation is used.

Question 2 requires us to run the code type the following command on command line. Use: ‘./spark-submit mainCode.py /path/to/house_prices.csv’. The RMSE and R-squared metrics will be printed upon running the code.

```
2019-03-13 16:35:24 INFO  ContextHandler:781 - Started o.s.j.s.ServletContextHandler@53cf550d{/stages/stage/json,null,AVAILABLE,@Spark}
2019-03-13 16:35:24 INFO  ContextHandler:781 - Started o.s.j.s.ServletContextHandler@6ebb08a2{/stages/pool,null,AVAILABLE,@Spark}
2019-03-13 16:35:24 INFO  ContextHandler:781 - Started o.s.j.s.ServletContextHandler@3454d73f{/stages/pool/json,null,AVAILABLE,@Spark}
2019-03-13 16:35:24 INFO  ContextHandler:781 - Started o.s.j.s.ServletContextHandler@128b4fa9{/storage,null,AVAILABLE,@Spark}
2019-03-13 16:35:24 INFO  ContextHandler:781 - Started o.s.j.s.ServletContextHandler@3bc517c8{/storage/json,null,AVAILABLE,@Spark}
2019-03-13 16:35:24 INFO  ContextHandler:781 - Started o.s.j.s.ServletContextHandler@d8df60a{/storage/rdd,null,AVAILABLE,@Spark}
2019-03-13 16:35:24 INFO  ContextHandler:781 - Started o.s.j.s.ServletContextHandler@1ff225ab{/storage/rdd/json,null,AVAILABLE,@Spark}
2019-03-13 16:35:24 INFO  ContextHandler:781 - Started o.s.j.s.ServletContextHandler@574cd7fb{/environment,null,AVAILABLE,@Spark}
2019-03-13 16:35:24 INFO  ContextHandler:781 - Started o.s.j.s.ServletContextHandler@4424f18d{/environment/json,null,AVAILABLE,@Spark}
2019-03-13 16:35:24 INFO  ContextHandler:781 - Started o.s.j.s.ServletContextHandler@26fb62a5{/executors,null,AVAILABLE,@Spark}
2019-03-13 16:35:24 INFO  ContextHandler:781 - Started o.s.j.s.ServletContextHandler@28aad4ac{/executors/json,null,AVAILABLE,@Spark}
2019-03-13 16:35:24 INFO  ContextHandler:781 - Started o.s.j.s.ServletContextHandler@19682d4d{/executors/threadDump,null,AVAILABLE,@Spark}
2019-03-13 16:35:24 INFO  ContextHandler:781 - Started o.s.j.s.ServletContextHandler@7aa31f74{/executors/threadDump/json,null,AVAILABLE,@Spark}
2019-03-13 16:35:24 INFO  ContextHandler:781 - Started o.s.j.s.ServletContextHandler@730b4cf{/static,null,AVAILABLE,@Spark}
2019-03-13 16:35:24 INFO  ContextHandler:781 - Started o.s.j.s.ServletContextHandler@7aae5816{/,null,AVAILABLE,@Spark}
2019-03-13 16:35:24 INFO  ContextHandler:781 - Started o.s.j.s.ServletContextHandler@4d3bddd2{/api,null,AVAILABLE,@Spark}
2019-03-13 16:35:24 INFO  ContextHandler:781 - Started o.s.j.s.ServletContextHandler@4c146da{/jobs/job/kill,null,AVAILABLE,@Spark}
2019-03-13 16:35:24 INFO  ContextHandler:781 - Started o.s.j.s.ServletContextHandler@68c78585{/stages/stage/kill,null,AVAILABLE,@Spark}
2019-03-13 16:35:24 INFO  SparkUI:54 - Bound SparkUI to 0.0.0.0, and started at http://192.168.0.10:4040
2019-03-13 16:35:24 INFO  Executor:54 - Starting executor ID driver on host localhost
2019-03-13 16:35:24 INFO  Utils:54 - Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 36101.
2019-03-13 16:35:24 INFO  NettyBlockTransferService:54 - Server created on 192.168.0.10:36101
2019-03-13 16:35:24 INFO  BlockManager:54 - Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
2019-03-13 16:35:24 INFO  BlockManagerMaster:54 - Registering BlockManager BlockManagerId(driver, 192.168.0.10, 36101, None)
2019-03-13 16:35:24 INFO  BlockManagerMasterEndpoint:54 - Registering block manager 192.168.0.10:36101 with 366.3 MB RAM, BlockManagerId(driver, 19
2.168.0.10, 36101, None)
2019-03-13 16:35:24 INFO  BlockManagerMaster:54 - Registered BlockManager BlockManagerId(driver, 192.168.0.10, 36101, None)
2019-03-13 16:35:24 INFO  BlockManager:54 - Initialized BlockManager: BlockManagerId(driver, 192.168.0.10, 36101, None)
2019-03-13 16:35:25 INFO  ContextHandler:781 - Started o.s.j.s.ServletContextHandler@2fdadac3{/metrics/json,null,AVAILABLE,@Spark}



Steps: Normalisation, Imputation and Linear Regression
Results after regression :

RMSE: 35027.6433974
R-squared: 0.829931164246
```

References:

1. https://github.com/mkbehbehani/spark-advanced-regression-kaggle/blob/b48e87699be02fb8edae3408e8429272fd79b0c0/src/main/scala/DataExtractor.scala
2. https://shkurnykov.site/house-prices-kaggle/
3. https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/19095846306138/3526927814287595/8071950455163429/latest.html