

Description

In this assignment, the task is to develop a parallel machine learning model in the Amazon AWS cloud environment. In this project, I have trained the “winequality” prediction model based on given training data and then utilized this model to test my model for the given validation dataset. I have utilized the Random Forest regression model for training and testing wine quality prediction applications and then calculated the accuracy of the model to determine how well the model performed. The task was to deploy our model with and without docker. The following are steps in the development of the Wine Quality prediction system.

Docker Link: <https://hub.docker.com/r/system25/winequalityprediction>

How to set up the cloud environment and run the model training and the application prediction without docker

Following are the steps to a setup cloud environment:

1. Navigate to the Amazon AWS website i.e. “<https://aws.amazon.com/>”
2. Create an Amazon AWS account and log in to the console by providing login credentials.
3. From the search for EC2 service and navigate to the EC2 dashboard.
4. Click on Launch Instance and follow all the steps by providing all the mandatory details.
5. Choose a number of instances. In this case, I have selected 5 instances because we are required to run our application on multiple instances.
6. Click on the launch button to launch your EC2 instance.
7. From the command line ssh into the ec2 machine from our local system
8. Copy application files and folders from the local system into the EC2 instance.
9. Run the following command to build the model
 1. Spark-submit test.py

How to set up the cloud environment and run the model training and the application prediction with docker

Following are the steps to run the application in docker in Amazon AWS:

- Login to your Amazon AWS account.
- SSH into EC2 instance from the local machine.
- Install docker into the EC2 instance by using the following commands:
 - `sudo yum install docker`
 - `sudo service docker start`
 - `sudo usermod -a -G docker ec2-user`
- Navigate to the directory with application files.
- Run the following command to build the image and run the container:
 - `docker build -t winequalityprediction`
 - `docker run -p 80:80 winequalityprediction`

Screenshots:

Docker build in ec2 instance

```
ec2-user@ip-172-31-33-167:~/winequalityprediction
Adding debian:Hellenic_Academic_and_Research_Institutions_ECC_RootCA_2015.pem
Adding debian:Hellenic_Academic_and_Research_Institutions_RootCA_2011.pem
Adding debian:Hellenic_Academic_and_Research_Institutions_RootCA_2015.pem
Adding debian:Hongkong_Post_Root_CA_1.pem
Adding debian:Hongkong_Post_Root_CA_3.pem
Adding debian:ISRG_Root_X1.pem
Adding debian:IdenTrust_Commercial_Root_CA_1.pem
Adding debian:IdenTrust_Public_Sector_Root_CA_1.pem
Adding debian:Izenpe.com.pem
Adding debian:Microsec_e-Szigno_Root_CA_2009.pem
Adding debian:Microsoft_ECC_Root_Certificate_Authority_2017.pem
Adding debian:Microsoft_RSA_Root_Certificate_Authority_2017.pem
Adding debian:NAVER_Global_Root_Certification_Authority.pem
Adding debian:NetLock_Arany_=Class_Gold= Főtanúsítvány.pem
Adding debian:Network_Solutions_Certificate_Authority.pem
Adding debian:OISTE_WISEKey_Global_Root_GB_CA.pem
Adding debian:OISTE_WISEKey_Global_Root_GC_CA.pem
Adding debian:QuoVadis_Root_CA.pem
Adding debian:QuoVadis_Root_CA_1_G3.pem
Adding debian:QuoVadis_Root_CA_2.pem
Adding debian:QuoVadis_Root_CA_2_G3.pem
Adding debian:QuoVadis_Root_CA_3.pem
Adding debian:QuoVadis_Root_CA_3_G3.pem
Adding debian:SSL.com_EV_Root_Certification_Authority_ECC.pem
```

Docker Run

```
ec2-user@ip-172-31-38-174:~/winequalityprediction
at org.sparkproject.guava.cache.LocalCache$Segment.loadSync(LocalCache.java:2379)
... 58 more
22/08/05 11:01:56 INFO FileScanRDD: Reading File path: file:///winequalitypredict/ValidationDataset.csv, range: 0-8762,
partition values: [empty row]
22/08/05 11:01:56 INFO BlockManagerInfo: Removed broadcast_136_piece0 on a3524101733a:44055 in memory (size: 27.0 KiB, f
ree: 413.7 MiB)
22/08/05 11:01:56 INFO PythonRunner: Times: total = 210, boot = 9, init = 200, finish = 1
22/08/05 11:01:56 INFO Executor: Finished task 0.0 in stage 68.0 (TID 68). 2718 bytes result sent to driver
22/08/05 11:01:56 INFO TaskSetManager: Finished task 0.0 in stage 68.0 (TID 68) in 285 ms on a3524101733a (executor driv
er) (1/1)
22/08/05 11:01:56 INFO TaskSchedulerImpl: Removed TaskSet 68.0, whose tasks have all completed, from pool
22/08/05 11:01:56 INFO DAGScheduler: ResultStage 68 (toPandas at /test.py:30) finished in 0.295 s
22/08/05 11:01:56 INFO DAGScheduler: Job 68 is finished. Cancelling potential speculative or zombie tasks for this job
22/08/05 11:01:56 INFO TaskSchedulerImpl: Killing all running tasks in stage 68: Stage finished
22/08/05 11:01:56 INFO DAGScheduler: Job 68 finished: toPandas at /test.py:30, took 0.300492 s
F1 score: 0.55625
22/08/05 11:01:56 INFO SparkContext: Invoking stop() from shutdown hook
22/08/05 11:01:56 INFO SparkUI: Stopped Spark web UI at http://a3524101733a:4040
22/08/05 11:01:56 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
22/08/05 11:01:56 INFO MemoryStore: MemoryStore cleared
22/08/05 11:01:56 INFO BlockManager: BlockManager stopped
22/08/05 11:01:56 INFO BlockManagerMaster: BlockManagerMaster stopped
22/08/05 11:01:56 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
22/08/05 11:01:56 INFO SparkContext: Successfully stopped SparkContext
22/08/05 11:01:56 INFO ShutdownHookManager: Shutdown hook called
22/08/05 11:01:56 INFO ShutdownHookManager: Deleting directory /tmp/spark-a43b3d49-371f-4c6b-99f5-06c180614b58
22/08/05 11:01:56 INFO ShutdownHookManager: Deleting directory /tmp/spark-a43b3d49-371f-4c6b-99f5-06c180614b58/pyspark-8
3735332-3b37-4bf8-aa27-3ec37c8614e4
22/08/05 11:01:56 INFO ShutdownHookManager: Deleting directory /tmp/spark-9f13ab34-0f04-47c3-a6fe-00e0cfbf2ee7
[ec2-user@ip-172-31-38-174 winequalityprediction]$
```

Docker run

```
ec2-user@ip-172-31-38-174:~/winequalityprediction
at org.sparkproject.guava.cache.LocalCache$Segment.loadSync(LocalCache.java:2379)
... 58 more
22/08/05 11:01:56 INFO FileScanRDD: Reading File path: file:///winequalitypredict/ValidationDataset.csv, range: 0-8762,
partition values: [empty row]
22/08/05 11:01:56 INFO BlockManagerInfo: Removed broadcast_136_piece0 on a3524101733a:44055 in memory (size: 27.0 KiB, f
ree: 413.7 MiB)
22/08/05 11:01:56 INFO PythonRunner: Times: total = 210, boot = 9, init = 200, finish = 1
22/08/05 11:01:56 INFO Executor: Finished task 0.0 in stage 68.0 (TID 68). 2718 bytes result sent to driver
22/08/05 11:01:56 INFO TaskSetManager: Finished task 0.0 in stage 68.0 (TID 68) in 285 ms on a3524101733a (executor driv
er) (1/1)
22/08/05 11:01:56 INFO TaskSchedulerImpl: Removed TaskSet 68.0, whose tasks have all completed, from pool
22/08/05 11:01:56 INFO DAGScheduler: ResultStage 68 (toPandas at /test.py:30) finished in 0.295 s
22/08/05 11:01:56 INFO DAGScheduler: Job 68 is finished. Cancelling potential speculative or zombie tasks for this job
22/08/05 11:01:56 INFO TaskSchedulerImpl: Killing all running tasks in stage 68: Stage finished
22/08/05 11:01:56 INFO DAGScheduler: Job 68 finished: toPandas at /test.py:30, took 0.300492 s
F1- score: 0.55625
22/08/05 11:01:56 INFO SparkContext: Invoking stop() from shutdown hook
22/08/05 11:01:56 INFO SparkUI: Stopped Spark web UI at http://a3524101733a:4040
22/08/05 11:01:56 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
22/08/05 11:01:56 INFO MemoryStore: MemoryStore cleared
22/08/05 11:01:56 INFO BlockManager: BlockManager stopped
22/08/05 11:01:56 INFO BlockManagerMaster: BlockManagerMaster stopped
22/08/05 11:01:56 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
22/08/05 11:01:56 INFO SparkContext: Successfully stopped SparkContext
22/08/05 11:01:56 INFO ShutdownHookManager: Shutdown hook called
22/08/05 11:01:56 INFO ShutdownHookManager: Deleting directory /tmp/spark-a43b3d49-371f-4c6b-99f5-06c180614b58
22/08/05 11:01:56 INFO ShutdownHookManager: Deleting directory /tmp/spark-a43b3d49-371f-4c6b-99f5-06c180614b58/pyspark-8
3735332-3b37-4bf8-aa27-3ec37c8614e4
22/08/05 11:01:56 INFO ShutdownHookManager: Deleting directory /tmp/spark-9f13ab34-0f04-47c3-a6fe-00e0cfbf2ee7
[ec2-user@ip-172-31-38-174 winequalityprediction]$
```

Docker build

```
File Edit Selection View Go Run Terminal Help train.py - winequalityprediction - Visual Studio Code
EXPLORER
WINEQ...
Dockerfile
test.py
train.py
TrainingDataset.csv
ValidationDataset.csv
train.py
1 from pyspark.sql import SparkSession
2 from pyspark import SparkContext, SparkConf
3 from pyspark.ml.classification import LogisticRegression
4 from pyspark.ml.feature import VectorAssembler
5 import numpy as np
6 from pyspark.mllib.evaluation import MulticlassMetrics
7 from pyspark.mllib.regression import LabeledPoint
8 conf = SparkConf().setAppName('random').setMaster('master')
9 sc = SparkContext.getOrCreate()
10 spark = SparkSession(sc)
11 loaddata = spark.read.format("csv").load("TrainingDataset.csv",delimiter=";", inferSchema=True, header=True)
12 loaddata.show()
13 vector = VectorAssembler(inputCols=loaddata.columns[:11], outputCol="features")
14 data_x = vector.transform(loaddata)
15 data_x.show()
16 data_y = data_x.select("features", "quality")
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL JUPYTER
[+] Building 164.0s (6/12)
=> => extracting sha256:7f162c881e4f4672d8c7093a578c2dffa9d249bc73fe24db52f9eb05c0575fa38 0.3s
=> => extracting sha256:e3f48ccb2876ae27ffef0b6d25a19592a3753e336d789239f8ae90c12f9a98 0.7s
=> => extracting sha256:315a1520c10e286803d5dabe4d67d96aed42a9fda4c8e6f1bb7655ee3ee5e7 0.0s
=> => extracting sha256:65ec519192cbf8c17cf0c3fdd6f5b282d4c1393a505282bc176dbad363bfc92 0.2s
=> [stage-1 2/7] RUN set -ex && pip install numpy && pip install pandas && pip install sklearn && pip install pysp 22.2s
=> => # pip command
=> => # pip install pandas
=> => # Collecting pandas
=> => # Downloading pandas-1.4.3-cp310-cp310-man
=> => # ylinux_2_17_x86_64.manylinux2014_x86_64.wh
=> => # 1 (11.6 MB)
OUTLINE
TIMELINE
Discovering Python Interpreters Ln 1, Col 1 Spaces: 4 UTF-8 CRLF Python
```

Image Creation

```
ec2-user@ip-172-31-33-167:~/winequalityprediction
Removing intermediate container 1b83b5fff8aa
--> 727a549c6220
Step 8/14 : COPY test.py /winequality/
--> 5b71fedc0744
Step 9/14 : COPY ValidationDataset.csv /winequality/
--> 5a7dca484e5f
Step 10/14 : COPY lg1model /winequality/
--> 6117b8c22126
Step 11/14 : ENV PROG_NAME test.py
--> Running in afc46b3cd4fd
Removing intermediate container afc46b3cd4fd
--> 3f4633b62577
Step 12/14 : ADD ${PROG_NAME} .
--> ca2e91441c6d
Step 13/14 : ENTRYPOINT ["spark-submit","test.py"]
--> Running in efeb50a6c385
Removing intermediate container efeb50a6c385
--> 9c0a9bfc8ac6
Step 14/14 : CMD ["ValidationDatasetwcsv"]
--> Running in 1fc80d0fecbb
Removing intermediate container 1fc80d0fecbb
--> 24e72277d855
Successfully built 24e72277d855
Successfully tagged winequalitypredict:latest
[ec2-user@ip-172-31-33-167 winequalityprediction]$
```

Model

```
ec2-user@ip-172-31-33-167:~/winequalityprediction
22/08/04 21:16:49 INFO DAGScheduler: Job 3 finished: showString at NativeMethodAccessorImpl.java:0, took 0.650012 s
22/08/04 21:16:49 INFO CodeGenerator: Code generated in 43.385794 ms
```

fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	quality
0.9978	3.51	0.56	0.7	0.0	1.9	0.076	11
0.9968	3.2	0.68	0.88	0.0	2.6	0.098	25
0.997	3.26	0.65	0.76	0.04	2.3	0.092	15
0.998	3.16	0.58	0.28	0.8	1.9	0.075	17
0.9978	3.51	0.56	0.7	0.0	1.9	0.076	11
0.9978	3.51	0.56	0.66	0.4	1.8	0.075	13
0.9964	3.3	0.46	0.4	0.06	1.6	0.069	15
0.9946	3.39	0.47	10.0	0.0	1.2	0.065	15
0.9968	3.36	0.57	9.5	0.02	2.0	0.073	9
0.9978	3.35	0.8	10.5	0.36	6.1	0.071	17
0.9959	3.28	0.54	0.58	0.08	1.8	0.097	15
0.9978	3.35	0.8	10.5	0.36	6.1	0.071	17
0.9943	3.58	0.52	0.615	0.0	1.6	0.089	16
0.9974	3.26	1.56	0.62	0.1	1.6	0.114	9
0.9986	3.16	0.88	0.62	0.2	3.8	0.176	52
0.9986	3.17	0.93	0.2	0.19	3.9	0.17	51