databricksSmartFolio_v11_fit_multiplemodels_compare

```
# here we are mounting our S3 bucket with a specific MOUNT_NAME
try:
  import urllib
  ACCESS_KEY ="XX"
  SECRET_KEY ="XX"
  ENCODED_SECRET_KEY = urllib.parse.quote(SECRET_KEY,"")
  AWS_BUCKET_NAME = "sree-databricks-test1"
  MOUNT_NAME = "s3datav3"
  #dbutils.fs.unmount("/mnt/MOUNT_NAME")
  dbutils.fs.mount("s3n://%s:%s@%s" % (ACCESS_KEY, ENCODED_SECRET_KEY, AWS_BUCKET_NAME), "/mnt/%s"
% MOUNT_NAME)
  display(dbutils.fs.ls("/mnt/%s" % MOUNT_NAME))
except Exception as e:
  print("S3 already mounted")
```

```
S3 already mounted
```

```
MOUNT_NAME = "s3datav3"
#display(dbutils.fs.ls("/mnt/%s/alphav/" % MOUNT_NAME))
display(dbutils.fs.ls("dbfs:/mnt/%s/results/" % MOUNT_NAME))
```

path
dbfs:/mnt/s3datav3/results/r2results.csv
dbfs:/mnt/s3datav3/results/r2results_withfeatures.csv
dbfs:/mnt/s3datav3/results/test/

```
#"MMM", "AXP", "AAPL", "BA", "CAT", "CVX", "CSC", "KO", "DIS", "DOW", "XOM", "GS", "HD", "IBM",
"INT", "JNJ", "JPM", "MCD", "MRK", "MSF", "NKE", "PFE", "PG", "TRV", "UTX", "UNH", "VZ", "V",
"WMT", "WBA"
```

```
from datetime import datetime
```

```
# We moved this part to the ETL part. this code needs to be removed
def clean_dataframe(df_data1):
  df_data1 = df_data1.drop('_c0')
  df_data1 = df_data1.drop('1. open')
  df_data1 = df_data1.drop('2. high')
  df_data1 = df_data1.drop('3. low')
  df_data1 = df_data1.drop('5. volume')
  df_data1 = df_data1.drop('VWAP')
  df_data1 = df_data1.withColumnRenamed("4. close","close")
  df_data1.dropna()
  return df_data1
```

```
# HERE we are diving the data into train and test
from pyspark.sql.functions import unix_timestamp, lit
def get_train_testdata(df_data1):
  train_data = df_data1.filter(df_data1["DailyDate"] < unix_timestamp(lit('2016-06-01
00:00:00')).cast('timestamp'))
  test_data  = df_data1.filter(df_data1["DailyDate"] > unix_timestamp(lit('2016-06-01
00:00:00')).cast('timestamp'))
  print("Number of training records: " + str(train_data.count()))
  print("Number of testing records : " + str(test_data.count()))
  return train_data, test_data
```

```
from pyspark.ml.feature import VectorAssembler,RFormula
from pyspark.ml.regression import LinearRegression, GeneralizedLinearRegression,
DecisionTreeRegressor
from pyspark.ml.regression import LinearRegressionModel
from pyspark.ml import Pipeline, Model,PipelineModel
from pyspark.ml.evaluation import RegressionEvaluator
from pyspark.ml.tuning import CrossValidator, ParamGridBuilder
import sklearn.metrics
from pyspark.ml.evaluation import RegressionEvaluator
from pyspark.sql.functions import monotonically_increasing_id, lag
from pyspark.sql.window import Window
```

```
# this is a dataframe which holds the r2 value for each ticker and model associated with it
# in this, initialising
result_summary = spark.range(0).drop("id")
r2_columns = ['ticker', 'LR_R2', 'LR_RMSE','XG_R2','XG_RMSE']
result_summary = spark.createDataFrame([('Test',4.0,5.0,7.0,8.0)], r2_columns)
firstrow = spark.createDataFrame([('Test',4.0,5.0,7.0,8.0)], r2_columns)
result_summary = result_summary.union(firstrow )
result_summary.show()
```

```
+------+-----+-------+-----+-------+
|ticker|LR_R2|LR_RMSE|XG_R2|XG_RMSE|
+------+-----+-------+-----+-------+
|  Test|  4.0|    5.0|  7.0|    8.0|
|  Test|  4.0|    5.0|  7.0|    8.0|
+------+-----+-------+-----+-------+
```

```
# We are removing the columns that causes a dataleakage
def removecolumns(train_data):
  columns = train_data.columns
  columns.remove('close_lag7D')
  columns.remove('DailyDate')
  columns.remove('close_lag14D')
  columns.remove('close_lag28D')
  columns.remove('close_lag50D')
  columns.remove('close_lag200D')
  return columns
```

```
# HERE we are running the model for all the Dow 30 tickers.

#stockList = ["MMM", "AAPL"] #["MMM", "AXP", "AAPL", "BA", "CAT"]

stockList = ["MMM", "AXP", "AAPL", "BA", "CAT", "CVX", "KO", "DIS", "XOM", "GS", "HD", "IBM",
"INT", "JNJ", "JPM", "MCD", "MRK", "MSF", "NKE", "PFE", "PG", "TRV", "UTX", "UNH", "VZ", "V",
"WMT", "WBA"]

MOUNT_NAME = "s3datav3"

for stock in stockList:
  #stockList = ["MMM", "AAPL"] #["MMM", "AXP", "AAPL", "BA", "CAT"]
  df_data1 = spark.read\
      .format("csv")\
      .option('header', 'true')\
      .option('inferSchema', 'true')\
      .load("/mnt/"+MOUNT_NAME +"/alphaandfeatures/" + stock+ ".csv" )
  #df_data1.head(5)
  df_data2 = df_data1.drop('_c0').dropna()

  train_data, test_data = get_train_testdata(df_data2)
  columns = removecolumns(train_data)


  #Formula : close_lag7D ~ close + SMA + EMA + MACD_Signal + MACD_Hist + MACD + RSI + SlowD + SlowK
+ ADX + CCI + dateSMA7 + dateSMA14 + dateSMA28 + dateSMA50 + dateSMA200 + date_converted +
DateofMon + Month + Year + WeekSeq + SMA50 + SMA200 + SMA7 + SMA14 + SMA28 + Delta

  formula = "{} ~ {}".format("close_lag7D", " + ".join(columns))
  print("Formula : {}".format(formula))
  rformula = RFormula(formula = formula)
  lr = LinearRegression()
  pipeline = Pipeline(stages=[rformula, lr])
  # Parameter grid
  paramGrid = ParamGridBuilder()\
           .addGrid(lr.regParam,[0.01, .04])\
            .build()
  cv = CrossValidator()\
        .setEstimator(pipeline)\
        .setEvaluator(RegressionEvaluator()\
                        .setMetricName("r2"))\
        .setEstimatorParamMaps(paramGrid)\
        .setNumFolds(3)

  cvModel = cv.fit(train_data)
  cvModel.avgMetrics
  predictions = cvModel.transform(test_data)
  evaluator = RegressionEvaluator(labelCol="label",
                                  predictionCol="prediction",
                                  metricName="rmse")

  y_true = predictions.select('label').toPandas()
  y_pred  = predictions.select('prediction').toPandas()

  r2_score = sklearn.metrics.r2_score(y_true, y_pred)
  print('r2_score: {0}'.format(r2_score))

  rmse = evaluator.evaluate(predictions)
```

```
  print("Root Mean Squared Error (RMSE) on test data = %g" % rmse)

  print(cvModel.explainParams())
  firstrow =
spark.createDataFrame([(stock,format(r2_score),format(rmse),format(r2_score),format(rmse))],
r2_columns)
  result_summary = result_summary.union(firstrow )

#df_data.show(5)
#df2 = df_data1.toPandas()
#df2.isnull().sum()
#df2 = df_data1.toPandas()
#df2.size
#df2.dtypes
#df2.count()
```

```
Number of training records: 1402
Number of testing records : 214
Formula : close_lag7D ~ close + SMA + EMA + MACD_Signal + MACD_Hist + MACD + RSI + SlowD + SlowK +
ADX + CCI + dateSMA7 + dateSMA14 + dateSMA28 + dateSMA50 + dateSMA200 + date_converted + DateofMon
+ Month + Year + WeekSeq + SMA50 + SMA200 + SMA7 + SMA14 + SMA28 + Delta
/databricks/spark/python/pyspark/ml/util.py:791: UserWarning: Can not find mlflow. To enable mlflo
w logging, install MLflow library from PyPi.
  warnings.warn(_MLflowInstrumentation._NO_MLFLOW_WARNING)
r2_score: 0.778889708870703
Root Mean Squared Error (RMSE) on test data = 3.21909
estimator: estimator to be cross-validated (current: Pipeline_3b7239a3fbb8)
estimatorParamMaps: estimator param maps (current: [{Param(parent='LinearRegression_d7fa2b319ea3',
name='regParam', doc='regularization parameter (>= 0).'): 0.01}, {Param(parent='LinearRegression_d
7fa2b319ea3', name='regParam', doc='regularization parameter (>= 0).'): 0.04}])
evaluator: evaluator used to select hyper-parameters that maximize the validator metric (current:
RegressionEvaluator_24e8cab2ef22)
seed: random seed. (default: 7809051150349531440)
Number of training records: 1402
Number of testing records : 214
Formula : close_lag7D ~ close + SMA + EMA + MACD_Signal + MACD_Hist + MACD + RSI + SlowK + SlowD +
ADX + CCI + dateSMA7 + dateSMA14 + dateSMA28 + dateSMA50 + dateSMA200 + date_converted + DateofMon
+ Month + Year + WeekSeq + SMA50 + SMA200 + SMA7 + SMA14 + SMA28 + Delta
/databricks/spark/python/pyspark/ml/util.py:791: UserWarning: Can not find mlflow. To enable mlflo
w logging, install MLflow library from PyPi.
  warnings.warn(_MLflowInstrumentation._NO_MLFLOW_WARNING)
r2_score: 0.8961289527014993
Root Mean Squared Error (RMSE) on test data = 2.18302
estimator: estimator to be cross-validated (current: Pipeline_417d10aaf13b)
estimatorParamMaps: estimator param maps (current: [{Param(parent='LinearRegression_a96b5eb7b5ea',
name='regParam', doc='regularization parameter (>= 0).'): 0.01}, {Param(parent='LinearRegression_a
96b5eb7b5ea', name='regParam', doc='regularization parameter (>= 0).'): 0.04}])
evaluator: evaluator used to select hyper-parameters that maximize the validator metric (current:
RegressionEvaluator_b462878939c2)
seed: random seed. (default: 7809051150349531440)
Number of training records: 1402
Number of testing records : 214
Formula : close_lag7D ~ close + SMA + EMA + MACD + MACD_Hist + MACD_Signal + RSI + SlowK + SlowD +
ADX + CCI + dateSMA7 + dateSMA14 + dateSMA28 + dateSMA50 + dateSMA200 + date_converted + DateofMon
+ Month + Year + WeekSeq + SMA50 + SMA200 + SMA7 + SMA14 + SMA28 + Delta
/databricks/spark/python/pyspark/ml/util.py:791: UserWarning: Can not find mlflow. To enable mlflo
```

```
w logging, install MLflow library from PyPi.
  warnings.warn(_MLflowInstrumentation._NO_MLFLOW_WARNING)
r2_score: -0.77004963948451
Root Mean Squared Error (RMSE) on test data = 19.0569
estimator: estimator to be cross-validated (current: Pipeline_55bbc69b3f8f)
estimatorParamMaps: estimator param maps (current: [{Param(parent='LinearRegression_f9ee8232bfd7',
name='regParam', doc='regularization parameter (>= 0).'): 0.01}, {Param(parent='LinearRegression_f
9ee8232bfd7', name='regParam', doc='regularization parameter (>= 0).'): 0.04}])
evaluator: evaluator used to select hyper-parameters that maximize the validator metric (current:
RegressionEvaluator_9913b2a5994c)
seed: random seed. (default: 7809051150349531440)
Number of training records: 1402
Number of testing records : 214
Formula : close_lag7D ~ close + SMA + EMA + MACD_Signal + MACD_Hist + MACD + RSI + SlowD + SlowK +
ADX + CCI + dateSMA7 + dateSMA14 + dateSMA28 + dateSMA50 + dateSMA200 + date_converted + DateofMon
+ Month + Year + WeekSeq + SMA50 + SMA200 + SMA7 + SMA14 + SMA28 + Delta
/databricks/spark/python/pyspark/ml/util.py:791: UserWarning: Can not find mlflow. To enable mlflo
w logging, install MLflow library from PyPi.
  warnings.warn(_MLflowInstrumentation._NO_MLFLOW_WARNING)
r2_score: 0.960221141052438
Root Mean Squared Error (RMSE) on test data = 3.6545
estimator: estimator to be cross-validated (current: Pipeline_0aeb8fc35ccf)
estimatorParamMaps: estimator param maps (current: [{Param(parent='LinearRegression_9103f3608368',
name='regParam', doc='regularization parameter (>= 0).'): 0.01}, {Param(parent='LinearRegression_9
103f3608368', name='regParam', doc='regularization parameter (>= 0).'): 0.04}])
evaluator: evaluator used to select hyper-parameters that maximize the validator metric (current:
RegressionEvaluator_bb122872bde1)
seed: random seed. (default: 7809051150349531440)
Number of training records: 1402
Number of testing records : 214
Formula : close_lag7D ~ close + SMA + EMA + MACD_Hist + MACD + MACD_Signal + RSI + SlowK + SlowD +
ADX + CCI + dateSMA7 + dateSMA14 + dateSMA28 + dateSMA50 + dateSMA200 + date_converted + DateofMon
+ Month + Year + WeekSeq + SMA50 + SMA200 + SMA7 + SMA14 + SMA28 + Delta
/databricks/spark/python/pyspark/ml/util.py:791: UserWarning: Can not find mlflow. To enable mlflo
w logging, install MLflow library from PyPi.
  warnings.warn(_MLflowInstrumentation._NO_MLFLOW_WARNING)
r2_score: 0.7068957559801414
Root Mean Squared Error (RMSE) on test data = 3.80786
estimator: estimator to be cross-validated (current: Pipeline_ab3bfa76ca22)
estimatorParamMaps: estimator param maps (current: [{Param(parent='LinearRegression_ad4e775a1028',
name='regParam', doc='regularization parameter (>= 0).'): 0.01}, {Param(parent='LinearRegression_a
d4e775a1028', name='regParam', doc='regularization parameter (>= 0).'): 0.04}])
evaluator: evaluator used to select hyper-parameters that maximize the validator metric (current:
RegressionEvaluator_90229a539cee)
seed: random seed. (default: 7809051150349531440)
Number of training records: 1402
Number of testing records : 214
Formula : close_lag7D ~ close + SMA + EMA + MACD + MACD_Hist + MACD_Signal + RSI + SlowK + SlowD +
ADX + CCI + dateSMA7 + dateSMA14 + dateSMA28 + dateSMA50 + dateSMA200 + date_converted + DateofMon
+ Month + Year + WeekSeq + SMA50 + SMA200 + SMA7 + SMA14 + SMA28 + Delta
/databricks/spark/python/pyspark/ml/util.py:791: UserWarning: Can not find mlflow. To enable mlflo
w logging, install MLflow library from PyPi.
  warnings.warn(_MLflowInstrumentation._NO_MLFLOW_WARNING)
r2_score: 0.7993927475496361
Root Mean Squared Error (RMSE) on test data = 2.59339
estimator: estimator to be cross-validated (current: Pipeline_464081c75ffd)
estimatorParamMaps: estimator param maps (current: [{Param(parent='LinearRegression_490eddff1ada',
name='regParam', doc='regularization parameter (>= 0).'): 0.01}, {Param(parent='LinearRegression_4
```

```
90eddff1ada', name='regParam', doc='regularization parameter (>= 0).'): 0.04}])
evaluator: evaluator used to select hyper-parameters that maximize the validator metric (current:
RegressionEvaluator_f0155cd991fc)
seed: random seed. (default: 7809051150349531440)
Number of training records: 1402
Number of testing records : 214
Formula : close_lag7D ~ close + SMA + EMA + MACD + MACD_Hist + MACD_Signal + RSI + SlowK + SlowD +
ADX + CCI + dateSMA7 + dateSMA14 + dateSMA28 + dateSMA50 + dateSMA200 + date_converted + DateofMon
+ Month + Year + WeekSeq + SMA50 + SMA200 + SMA7 + SMA14 + SMA28 + Delta
/databricks/spark/python/pyspark/ml/util.py:791: UserWarning: Can not find mlflow. To enable mlflo
w logging, install MLflow library from PyPi.
  warnings.warn(_MLflowInstrumentation._NO_MLFLOW_WARNING)
r2_score: 0.5133430170554745
Root Mean Squared Error (RMSE) on test data = 0.951428
estimator: estimator to be cross-validated (current: Pipeline_f9874aa34046)
estimatorParamMaps: estimator param maps (current: [{Param(parent='LinearRegression_0ef456c404d0',
name='regParam', doc='regularization parameter (>= 0).'): 0.01}, {Param(parent='LinearRegression_0
ef456c404d0', name='regParam', doc='regularization parameter (>= 0).'): 0.04}])
evaluator: evaluator used to select hyper-parameters that maximize the validator metric (current:
RegressionEvaluator_2aa988634d13)
seed: random seed. (default: 7809051150349531440)
Number of training records: 1402
Number of testing records : 214
Formula : close_lag7D ~ close + SMA + EMA + MACD + MACD_Signal + MACD_Hist + RSI + SlowK + SlowD +
ADX + CCI + dateSMA7 + dateSMA14 + dateSMA28 + dateSMA50 + dateSMA200 + date_converted + DateofMon
+ Month + Year + WeekSeq + SMA50 + SMA200 + SMA7 + SMA14 + SMA28 + Delta
/databricks/spark/python/pyspark/ml/util.py:791: UserWarning: Can not find mlflow. To enable mlflo
w logging, install MLflow library from PyPi.
  warnings.warn(_MLflowInstrumentation._NO_MLFLOW_WARNING)
r2_score: 0.9047289043085577
Root Mean Squared Error (RMSE) on test data = 2.26868
estimator: estimator to be cross-validated (current: Pipeline_a77b120bb7ad)
estimatorParamMaps: estimator param maps (current: [{Param(parent='LinearRegression_93b03ece2a65',
name='regParam', doc='regularization parameter (>= 0).'): 0.01}, {Param(parent='LinearRegression_9
3b03ece2a65', name='regParam', doc='regularization parameter (>= 0).'): 0.04}])
evaluator: evaluator used to select hyper-parameters that maximize the validator metric (current:
RegressionEvaluator_5b9750aff3b0)
seed: random seed. (default: 7809051150349531440)
Number of training records: 1402
Number of testing records : 214
Formula : close_lag7D ~ close + SMA + EMA + MACD_Signal + MACD_Hist + MACD + RSI + SlowK + SlowD +
ADX + CCI + dateSMA7 + dateSMA14 + dateSMA28 + dateSMA50 + dateSMA200 + date_converted + DateofMon
+ Month + Year + WeekSeq + SMA50 + SMA200 + SMA7 + SMA14 + SMA28 + Delta
/databricks/spark/python/pyspark/ml/util.py:791: UserWarning: Can not find mlflow. To enable mlflo
w logging, install MLflow library from PyPi.
  warnings.warn(_MLflowInstrumentation._NO_MLFLOW_WARNING)
r2_score: 0.6597778420287399
Root Mean Squared Error (RMSE) on test data = 2.12476
estimator: estimator to be cross-validated (current: Pipeline_5562382739f1)
estimatorParamMaps: estimator param maps (current: [{Param(parent='LinearRegression_abbbe1f39bf3',
name='regParam', doc='regularization parameter (>= 0).'): 0.01}, {Param(parent='LinearRegression_a
bbbe1f39bf3', name='regParam', doc='regularization parameter (>= 0).'): 0.04}])
evaluator: evaluator used to select hyper-parameters that maximize the validator metric (current:
RegressionEvaluator_e2c58e0a21a1)
seed: random seed. (default: 7809051150349531440)
Number of training records: 1402
Number of testing records : 214
Formula : close_lag7D ~ close + SMA + EMA + MACD_Signal + MACD_Hist + MACD + RSI + SlowD + SlowK +
```

```
ADX + CCI + dateSMA7 + dateSMA14 + dateSMA28 + dateSMA50 + dateSMA200 + date_converted + DateofMon
+ Month + Year + WeekSeq + SMA50 + SMA200 + SMA7 + SMA14 + SMA28 + Delta
/databricks/spark/python/pyspark/ml/util.py:791: UserWarning: Can not find mlflow. To enable mlflo
w logging, install MLflow library from PyPi.
  warnings.warn(_MLflowInstrumentation._NO_MLFLOW_WARNING)
r2_score: 0.9329992837687289
Root Mean Squared Error (RMSE) on test data = 9.7431
estimator: estimator to be cross-validated (current: Pipeline_91d7cbdd68c2)
estimatorParamMaps: estimator param maps (current: [{Param(parent='LinearRegression_d9655405a137',
name='regParam', doc='regularization parameter (>= 0).'): 0.01}, {Param(parent='LinearRegression_d
9655405a137', name='regParam', doc='regularization parameter (>= 0).'): 0.04}])
evaluator: evaluator used to select hyper-parameters that maximize the validator metric (current:
RegressionEvaluator_c75c4d2d4c2b)
seed: random seed. (default: 7809051150349531440)
Number of training records: 1402
Number of testing records : 214
Formula : close_lag7D ~ close + SMA + EMA + MACD_Hist + MACD + MACD_Signal + RSI + SlowK + SlowD +
ADX + CCI + dateSMA7 + dateSMA14 + dateSMA28 + dateSMA50 + dateSMA200 + date_converted + DateofMon
+ Month + Year + WeekSeq + SMA50 + SMA200 + SMA7 + SMA14 + SMA28 + Delta
/databricks/spark/python/pyspark/ml/util.py:791: UserWarning: Can not find mlflow. To enable mlflo
w logging, install MLflow library from PyPi.
  warnings.warn(_MLflowInstrumentation._NO_MLFLOW_WARNING)
r2_score: 0.8183854676783929
Root Mean Squared Error (RMSE) on test data = 3.17546
estimator: estimator to be cross-validated (current: Pipeline_dbfe633d070e)
estimatorParamMaps: estimator param maps (current: [{Param(parent='LinearRegression_2b57c5b9056a',
name='regParam', doc='regularization parameter (>= 0).'): 0.01}, {Param(parent='LinearRegression_2
b57c5b9056a', name='regParam', doc='regularization parameter (>= 0).'): 0.04}])
evaluator: evaluator used to select hyper-parameters that maximize the validator metric (current:
RegressionEvaluator_c4ee7265d9d3)
seed: random seed. (default: 7809051150349531440)
Number of training records: 1402
Number of testing records : 214
Formula : close_lag7D ~ close + SMA + EMA + MACD_Signal + MACD_Hist + MACD + RSI + SlowK + SlowD +
ADX + CCI + dateSMA7 + dateSMA14 + dateSMA28 + dateSMA50 + dateSMA200 + date_converted + DateofMon
+ Month + Year + WeekSeq + SMA50 + SMA200 + SMA7 + SMA14 + SMA28 + Delta
/databricks/spark/python/pyspark/ml/util.py:791: UserWarning: Can not find mlflow. To enable mlflo
w logging, install MLflow library from PyPi.
  warnings.warn(_MLflowInstrumentation._NO_MLFLOW_WARNING)
r2_score: 0.4441028304022917
Root Mean Squared Error (RMSE) on test data = 6.94138
estimator: estimator to be cross-validated (current: Pipeline_3b31d017da3b)
estimatorParamMaps: estimator param maps (current: [{Param(parent='LinearRegression_e800aafc55db',
name='regParam', doc='regularization parameter (>= 0).'): 0.01}, {Param(parent='LinearRegression_e
800aafc55db', name='regParam', doc='regularization parameter (>= 0).'): 0.04}])
evaluator: evaluator used to select hyper-parameters that maximize the validator metric (current:
RegressionEvaluator_3ca0ced1d10d)
seed: random seed. (default: 7809051150349531440)
Number of training records: 1402
Number of testing records : 214
Formula : close_lag7D ~ close + SMA + EMA + MACD_Signal + MACD_Hist + MACD + RSI + SlowK + SlowD +
ADX + CCI + dateSMA7 + dateSMA14 + dateSMA28 + dateSMA50 + dateSMA200 + date_converted + DateofMon
+ Month + Year + WeekSeq + SMA50 + SMA200 + SMA7 + SMA14 + SMA28 + Delta
/databricks/spark/python/pyspark/ml/util.py:791: UserWarning: Can not find mlflow. To enable mlflo
w logging, install MLflow library from PyPi.
  warnings.warn(_MLflowInstrumentation._NO_MLFLOW_WARNING)
r2_score: 0.6359993290565151
Root Mean Squared Error (RMSE) on test data = 2.33931
```

```
estimator: estimator to be cross-validated (current: Pipeline_f996360c0497)
estimatorParamMaps: estimator param maps (current: [{Param(parent='LinearRegression_7d47363e2f05',
name='regParam', doc='regularization parameter (>= 0).'): 0.01}, {Param(parent='LinearRegression_7
d47363e2f05', name='regParam', doc='regularization parameter (>= 0).'): 0.04}])
evaluator: evaluator used to select hyper-parameters that maximize the validator metric (current:
RegressionEvaluator_97c9fbd6ac30)
seed: random seed. (default: 7809051150349531440)
Number of training records: 1402
Number of testing records : 214
Formula : close_lag7D ~ close + SMA + EMA + MACD + MACD_Hist + MACD_Signal + RSI + SlowD + SlowK +
ADX + CCI + dateSMA7 + dateSMA14 + dateSMA28 + dateSMA50 + dateSMA200 + date_converted + DateofMon
+ Month + Year + WeekSeq + SMA50 + SMA200 + SMA7 + SMA14 + SMA28 + Delta
/databricks/spark/python/pyspark/ml/util.py:791: UserWarning: Can not find mlflow. To enable mlflo
w logging, install MLflow library from PyPi.
  warnings.warn(_MLflowInstrumentation._NO_MLFLOW_WARNING)
r2_score: 0.6018796087264624
Root Mean Squared Error (RMSE) on test data = 2.78219
estimator: estimator to be cross-validated (current: Pipeline_8529c7ce4f15)
estimatorParamMaps: estimator param maps (current: [{Param(parent='LinearRegression_dad010fc853c',
name='regParam', doc='regularization parameter (>= 0).'): 0.01}, {Param(parent='LinearRegression_d
ad010fc853c', name='regParam', doc='regularization parameter (>= 0).'): 0.04}])
evaluator: evaluator used to select hyper-parameters that maximize the validator metric (current:
RegressionEvaluator_5fd4337ab172)
seed: random seed. (default: 7809051150349531440)
Number of training records: 1402
Number of testing records : 214
Formula : close_lag7D ~ close + SMA + EMA + MACD_Hist + MACD + MACD_Signal + RSI + SlowK + SlowD +
ADX + CCI + dateSMA7 + dateSMA14 + dateSMA28 + dateSMA50 + dateSMA200 + date_converted + DateofMon
+ Month + Year + WeekSeq + SMA50 + SMA200 + SMA7 + SMA14 + SMA28 + Delta
/databricks/spark/python/pyspark/ml/util.py:791: UserWarning: Can not find mlflow. To enable mlflo
w logging, install MLflow library from PyPi.
  warnings.warn(_MLflowInstrumentation._NO_MLFLOW_WARNING)
r2_score: 0.9376712622080364
Root Mean Squared Error (RMSE) on test data = 2.69347
estimator: estimator to be cross-validated (current: Pipeline_5323fd1bc7b8)
estimatorParamMaps: estimator param maps (current: [{Param(parent='LinearRegression_3a41c240113b',
name='regParam', doc='regularization parameter (>= 0).'): 0.01}, {Param(parent='LinearRegression_3
a41c240113b', name='regParam', doc='regularization parameter (>= 0).'): 0.04}])
evaluator: evaluator used to select hyper-parameters that maximize the validator metric (current:
RegressionEvaluator_c9dd22c05d65)
seed: random seed. (default: 7809051150349531440)
Number of training records: 1402
Number of testing records : 214
Formula : close_lag7D ~ close + SMA + EMA + MACD_Signal + MACD_Hist + MACD + RSI + SlowK + SlowD +
ADX + CCI + dateSMA7 + dateSMA14 + dateSMA28 + dateSMA50 + dateSMA200 + date_converted + DateofMon
+ Month + Year + WeekSeq + SMA50 + SMA200 + SMA7 + SMA14 + SMA28 + Delta
/databricks/spark/python/pyspark/ml/util.py:791: UserWarning: Can not find mlflow. To enable mlflo
w logging, install MLflow library from PyPi.
  warnings.warn(_MLflowInstrumentation._NO_MLFLOW_WARNING)
r2_score: 0.7349223888299787
Root Mean Squared Error (RMSE) on test data = 2.76177
estimator: estimator to be cross-validated (current: Pipeline_5f7d6d3b1855)
estimatorParamMaps: estimator param maps (current: [{Param(parent='LinearRegression_e9c11e2d4dbe',
name='regParam', doc='regularization parameter (>= 0).'): 0.01}, {Param(parent='LinearRegression_e
9c11e2d4dbe', name='regParam', doc='regularization parameter (>= 0).'): 0.04}])
evaluator: evaluator used to select hyper-parameters that maximize the validator metric (current:
RegressionEvaluator_1703920a83cd)
seed: random seed. (default: 7809051150349531440)
```

```
Number of training records: 1402
Number of testing records : 214
Formula : close_lag7D ~ close + SMA + EMA + MACD + MACD_Hist + MACD_Signal + RSI + SlowK + SlowD +
ADX + CCI + dateSMA7 + dateSMA14 + dateSMA28 + dateSMA50 + dateSMA200 + date_converted + DateofMon
+ Month + Year + WeekSeq + SMA50 + SMA200 + SMA7 + SMA14 + SMA28 + Delta
/databricks/spark/python/pyspark/ml/util.py:791: UserWarning: Can not find mlflow. To enable mlflo
w logging, install MLflow library from PyPi.
  warnings.warn(_MLflowInstrumentation._NO_MLFLOW_WARNING)
r2_score: 0.4749928875102779
Root Mean Squared Error (RMSE) on test data = 1.78031
estimator: estimator to be cross-validated (current: Pipeline_54385c696cbd)
estimatorParamMaps: estimator param maps (current: [{Param(parent='LinearRegression_e1f9184e4dab',
name='regParam', doc='regularization parameter (>= 0).'): 0.01}, {Param(parent='LinearRegression_e
1f9184e4dab', name='regParam', doc='regularization parameter (>= 0).'): 0.04}])
evaluator: evaluator used to select hyper-parameters that maximize the validator metric (current:
RegressionEvaluator_7d1a8d46ce20)
seed: random seed. (default: 7809051150349531440)
Number of training records: 1402
Number of testing records : 214
Formula : close_lag7D ~ close + SMA + EMA_x + EMA_y + MACD + MACD_Hist + MACD_Signal + RSI + SlowK
+ SlowD + ADX + CCI + dateSMA7 + dateSMA14 + dateSMA28 + dateSMA50 + dateSMA200 + date_converted +
DateofMon + Month + Year + WeekSeq + SMA50 + SMA200 + SMA7 + SMA14 + SMA28 + Delta
/databricks/spark/python/pyspark/ml/util.py:791: UserWarning: Can not find mlflow. To enable mlflo
w logging, install MLflow library from PyPi.
  warnings.warn(_MLflowInstrumentation._NO_MLFLOW_WARNING)
r2_score: 0.6684602426174241
Root Mean Squared Error (RMSE) on test data = 0.335184
estimator: estimator to be cross-validated (current: Pipeline_bc6db4983e9b)
estimatorParamMaps: estimator param maps (current: [{Param(parent='LinearRegression_b46eb99a6211',
name='regParam', doc='regularization parameter (>= 0).'): 0.01}, {Param(parent='LinearRegression_b
46eb99a6211', name='regParam', doc='regularization parameter (>= 0).'): 0.04}])
evaluator: evaluator used to select hyper-parameters that maximize the validator metric (current:
RegressionEvaluator_4ec32941b467)
seed: random seed. (default: 7809051150349531440)
Number of training records: 1402
Number of testing records : 214
Formula : close_lag7D ~ close + SMA + EMA + MACD_Hist + MACD + MACD_Signal + RSI + SlowK + SlowD +
ADX + CCI + dateSMA7 + dateSMA14 + dateSMA28 + dateSMA50 + dateSMA200 + date_converted + DateofMon
+ Month + Year + WeekSeq + SMA50 + SMA200 + SMA7 + SMA14 + SMA28 + Delta
/databricks/spark/python/pyspark/ml/util.py:791: UserWarning: Can not find mlflow. To enable mlflo
w logging, install MLflow library from PyPi.
  warnings.warn(_MLflowInstrumentation._NO_MLFLOW_WARNING)
r2_score: -0.21719471758512254
Root Mean Squared Error (RMSE) on test data = 2.81502
estimator: estimator to be cross-validated (current: Pipeline_7a8c1328984c)
estimatorParamMaps: estimator param maps (current: [{Param(parent='LinearRegression_9f2cc3686d53',
name='regParam', doc='regularization parameter (>= 0).'): 0.01}, {Param(parent='LinearRegression_9
f2cc3686d53', name='regParam', doc='regularization parameter (>= 0).'): 0.04}])
evaluator: evaluator used to select hyper-parameters that maximize the validator metric (current:
RegressionEvaluator_c5d04647b055)
seed: random seed. (default: 7809051150349531440)
Number of training records: 1402
Number of testing records : 214
Formula : close_lag7D ~ close + SMA + EMA + MACD + MACD_Hist + MACD_Signal + RSI + SlowK + SlowD +
ADX + CCI + dateSMA7 + dateSMA14 + dateSMA28 + dateSMA50 + dateSMA200 + date_converted + DateofMon
+ Month + Year + WeekSeq + SMA50 + SMA200 + SMA7 + SMA14 + SMA28 + Delta
/databricks/spark/python/pyspark/ml/util.py:791: UserWarning: Can not find mlflow. To enable mlflo
w logging, install MLflow library from PyPi.
```

```
  warnings.warn(_MLflowInstrumentation._NO_MLFLOW_WARNING)
r2_score: 0.5825555716073947
Root Mean Squared Error (RMSE) on test data = 1.00699
estimator: estimator to be cross-validated (current: Pipeline_b9061bf53532)
estimatorParamMaps: estimator param maps (current: [{Param(parent='LinearRegression_25b27e7177c1',
name='regParam', doc='regularization parameter (>= 0).'): 0.01}, {Param(parent='LinearRegression_2
5b27e7177c1', name='regParam', doc='regularization parameter (>= 0).'): 0.04}])
evaluator: evaluator used to select hyper-parameters that maximize the validator metric (current:
RegressionEvaluator_ea55b7c76edf)
seed: random seed. (default: 7809051150349531440)
Number of training records: 1402
Number of testing records : 214
Formula : close_lag7D ~ close + SMA + EMA + MACD_Signal + MACD + MACD_Hist + RSI + SlowD + SlowK +
ADX + CCI + dateSMA7 + dateSMA14 + dateSMA28 + dateSMA50 + dateSMA200 + date_converted + DateofMon
+ Month + Year + WeekSeq + SMA50 + SMA200 + SMA7 + SMA14 + SMA28 + Delta
/databricks/spark/python/pyspark/ml/util.py:791: UserWarning: Can not find mlflow. To enable mlflo
w logging, install MLflow library from PyPi.
  warnings.warn(_MLflowInstrumentation._NO_MLFLOW_WARNING)
r2_score: 0.6651533543787265
Root Mean Squared Error (RMSE) on test data = 1.51953
estimator: estimator to be cross-validated (current: Pipeline_2daa230392a9)
estimatorParamMaps: estimator param maps (current: [{Param(parent='LinearRegression_9ea7261feee6',
name='regParam', doc='regularization parameter (>= 0).'): 0.01}, {Param(parent='LinearRegression_9
ea7261feee6', name='regParam', doc='regularization parameter (>= 0).'): 0.04}])
evaluator: evaluator used to select hyper-parameters that maximize the validator metric (current:
RegressionEvaluator_7422bec41906)
seed: random seed. (default: 7809051150349531440)
Number of training records: 1402
Number of testing records : 214
Formula : close_lag7D ~ close + SMA + EMA + MACD_Signal + MACD_Hist + MACD + RSI + SlowK + SlowD +
ADX + CCI + dateSMA7 + dateSMA14 + dateSMA28 + dateSMA50 + dateSMA200 + date_converted + DateofMon
+ Month + Year + WeekSeq + SMA50 + SMA200 + SMA7 + SMA14 + SMA28 + Delta
/databricks/spark/python/pyspark/ml/util.py:791: UserWarning: Can not find mlflow. To enable mlflo
w logging, install MLflow library from PyPi.
  warnings.warn(_MLflowInstrumentation._NO_MLFLOW_WARNING)
r2_score: 0.5635839918886696
Root Mean Squared Error (RMSE) on test data = 2.75411
estimator: estimator to be cross-validated (current: Pipeline_49667434acf6)
estimatorParamMaps: estimator param maps (current: [{Param(parent='LinearRegression_4c9676e2753d',
name='regParam', doc='regularization parameter (>= 0).'): 0.01}, {Param(parent='LinearRegression_4
c9676e2753d', name='regParam', doc='regularization parameter (>= 0).'): 0.04}])
evaluator: evaluator used to select hyper-parameters that maximize the validator metric (current:
RegressionEvaluator_0ff432f5b61b)
seed: random seed. (default: 7809051150349531440)
Number of training records: 1402
Number of testing records : 214
Formula : close_lag7D ~ close + SMA + EMA_x + EMA_y + MACD_Signal + MACD_Hist + MACD + RSI + SlowK
+ SlowD + ADX + CCI + dateSMA7 + dateSMA14 + dateSMA28 + dateSMA50 + dateSMA200 + date_converted +
DateofMon + Month + Year + WeekSeq + SMA50 + SMA200 + SMA7 + SMA14 + SMA28 + Delta
/databricks/spark/python/pyspark/ml/util.py:791: UserWarning: Can not find mlflow. To enable mlflo
w logging, install MLflow library from PyPi.
  warnings.warn(_MLflowInstrumentation._NO_MLFLOW_WARNING)
r2_score: 0.790991836947174
Root Mean Squared Error (RMSE) on test data = 2.0179
estimator: estimator to be cross-validated (current: Pipeline_badb0d18691b)
estimatorParamMaps: estimator param maps (current: [{Param(parent='LinearRegression_94710cfbd9a2',
name='regParam', doc='regularization parameter (>= 0).'): 0.01}, {Param(parent='LinearRegression_9
4710cfbd9a2', name='regParam', doc='regularization parameter (>= 0).'): 0.04}])
```

```
evaluator: evaluator used to select hyper-parameters that maximize the validator metric (current:
RegressionEvaluator_aacdd242f4e2)
seed: random seed. (default: 7809051150349531440)
Number of training records: 1402
Number of testing records : 214
Formula : close_lag7D ~ close + SMA + EMA + MACD_Signal + MACD + MACD_Hist + RSI + SlowK + SlowD +
ADX + CCI + dateSMA7 + dateSMA14 + dateSMA28 + dateSMA50 + dateSMA200 + date_converted + DateofMon
+ Month + Year + WeekSeq + SMA50 + SMA200 + SMA7 + SMA14 + SMA28 + Delta
/databricks/spark/python/pyspark/ml/util.py:791: UserWarning: Can not find mlflow. To enable mlflo
w logging, install MLflow library from PyPi.
  warnings.warn(_MLflowInstrumentation._NO_MLFLOW_WARNING)
r2_score: 0.8604935355917289
Root Mean Squared Error (RMSE) on test data = 4.40985
estimator: estimator to be cross-validated (current: Pipeline_2a1d0c0a3f69)
estimatorParamMaps: estimator param maps (current: [{Param(parent='LinearRegression_a46bdd8f026b',
name='regParam', doc='regularization parameter (>= 0).'): 0.01}, {Param(parent='LinearRegression_a
46bdd8f026b', name='regParam', doc='regularization parameter (>= 0).'): 0.04}])
evaluator: evaluator used to select hyper-parameters that maximize the validator metric (current:
RegressionEvaluator_5567c13eb859)
seed: random seed. (default: 7809051150349531440)
Number of training records: 1402
Number of testing records : 214
Formula : close_lag7D ~ close + SMA + EMA + MACD_Signal + MACD + MACD_Hist + RSI + SlowD + SlowK +
ADX + CCI + dateSMA7 + dateSMA14 + dateSMA28 + dateSMA50 + dateSMA200 + date_converted + DateofMon
+ Month + Year + WeekSeq + SMA50 + SMA200 + SMA7 + SMA14 + SMA28 + Delta
/databricks/spark/python/pyspark/ml/util.py:791: UserWarning: Can not find mlflow. To enable mlflo
w logging, install MLflow library from PyPi.
  warnings.warn(_MLflowInstrumentation._NO_MLFLOW_WARNING)
r2_score: 0.625990841710762
Root Mean Squared Error (RMSE) on test data = 1.53299
estimator: estimator to be cross-validated (current: Pipeline_195638bab67d)
estimatorParamMaps: estimator param maps (current: [{Param(parent='LinearRegression_779986ed67ff',
name='regParam', doc='regularization parameter (>= 0).'): 0.01}, {Param(parent='LinearRegression_7
79986ed67ff', name='regParam', doc='regularization parameter (>= 0).'): 0.04}])
evaluator: evaluator used to select hyper-parameters that maximize the validator metric (current:
RegressionEvaluator_3e073345f115)
seed: random seed. (default: 7809051150349531440)
Number of training records: 1402
Number of testing records : 214
Formula : close_lag7D ~ close + SMA + EMA + MACD_Hist + MACD + MACD_Signal + RSI + SlowK + SlowD +
ADX + CCI + dateSMA7 + dateSMA14 + dateSMA28 + dateSMA50 + dateSMA200 + date_converted + DateofMon
+ Month + Year + WeekSeq + SMA50 + SMA200 + SMA7 + SMA14 + SMA28 + Delta
/databricks/spark/python/pyspark/ml/util.py:791: UserWarning: Can not find mlflow. To enable mlflo
w logging, install MLflow library from PyPi.
  warnings.warn(_MLflowInstrumentation._NO_MLFLOW_WARNING)
r2_score: 0.4043532032384801
Root Mean Squared Error (RMSE) on test data = 3.17542
estimator: estimator to be cross-validated (current: Pipeline_6fdd2afc1034)
estimatorParamMaps: estimator param maps (current: [{Param(parent='LinearRegression_84b1fbf28d28',
name='regParam', doc='regularization parameter (>= 0).'): 0.01}, {Param(parent='LinearRegression_8
4b1fbf28d28', name='regParam', doc='regularization parameter (>= 0).'): 0.04}])
evaluator: evaluator used to select hyper-parameters that maximize the validator metric (current:
RegressionEvaluator_f49245125b66)
seed: random seed. (default: 7809051150349531440)
Number of training records: 1402
Number of testing records : 214
Formula : close_lag7D ~ close + SMA + EMA + MACD + MACD_Hist + MACD_Signal + RSI + SlowK + SlowD +
ADX + CCI + dateSMA7 + dateSMA14 + dateSMA28 + dateSMA50 + dateSMA200 + date_converted + DateofMon
```

```
+ Month + Year + WeekSeq + SMA50 + SMA200 + SMA7 + SMA14 + SMA28 + Delta
/databricks/spark/python/pyspark/ml/util.py:791: UserWarning: Can not find mlflow. To enable mlflo
w logging, install MLflow library from PyPi.
  warnings.warn(_MLflowInstrumentation._NO_MLFLOW_WARNING)
r2_score: 0.3526133722884067
Root Mean Squared Error (RMSE) on test data = 1.61553
estimator: estimator to be cross-validated (current: Pipeline_b0c0c2dd8101)
estimatorParamMaps: estimator param maps (current: [{Param(parent='LinearRegression_211baacba61d',
name='regParam', doc='regularization parameter (>= 0).'): 0.01}, {Param(parent='LinearRegression_2
11baacba61d', name='regParam', doc='regularization parameter (>= 0).'): 0.04}])
evaluator: evaluator used to select hyper-parameters that maximize the validator metric (current:
RegressionEvaluator_250aaedeff08)
seed: random seed. (default: 7809051150349531440)
Number of training records: 1402
Number of testing records : 214
Formula : close_lag7D ~ close + SMA + EMA + MACD_Signal + MACD_Hist + MACD + RSI + SlowK + SlowD +
ADX + CCI + dateSMA7 + dateSMA14 + dateSMA28 + dateSMA50 + dateSMA200 + date_converted + DateofMon
+ Month + Year + WeekSeq + SMA50 + SMA200 + SMA7 + SMA14 + SMA28 + Delta
/databricks/spark/python/pyspark/ml/util.py:791: UserWarning: Can not find mlflow. To enable mlflo
w logging, install MLflow library from PyPi.
  warnings.warn(_MLflowInstrumentation._NO_MLFLOW_WARNING)
r2_score: -0.04517011915054803
Root Mean Squared Error (RMSE) on test data = 2.14958
estimator: estimator to be cross-validated (current: Pipeline_df3e63033087)
estimatorParamMaps: estimator param maps (current: [{Param(parent='LinearRegression_1f4464a616b8',
name='regParam', doc='regularization parameter (>= 0).'): 0.01}, {Param(parent='LinearRegression_1
f4464a616b8', name='regParam', doc='regularization parameter (>= 0).'): 0.04}])
evaluator: evaluator used to select hyper-parameters that maximize the validator metric (current:
RegressionEvaluator_540bac171e09)
seed: random seed. (default: 7809051150349531440)
```

```
pdata = df_data2.toPandas()
pdata.tail(5)
```

Out[11]:

	close	DailyDate	SMA	EMA	MACD_Signal	MACD_Hist	MACD	RSI	SlowK	SlowD	ADX	CCI
1612	83.05	2017-03-31	84.5305	84.1665	-0.2333	-0.2192	-0.4525	41.9830	56.4248	65.8836	18.2914	-78.3132
1613	82.95	2017-04-03	84.3840	84.0506	-0.2843	-0.2040	-0.4883	41.4489	32.7816	53.7774	18.6798	-75.4789
1614	82.50	2017-04-04	84.2095	83.9029	-0.3368	-0.2099	-0.5467	39.0933	17.9678	35.7247	19.2433	-87.6173
1615	81.17	2017-04-05	84.0100	83.6426	-0.4079	-0.2845	-0.6924	33.2197	14.9022	21.8838	20.5496	-151.8662
1616	81.66	2017-04-06	83.8275	83.4538	-0.4782	-0.2813	-0.7595	36.8965	23.9270	18.9323	21.7906	-140.4961

```
formula = "{} ~ {}".format("close_lag7D", " + ".join(columns))
print("Formula : {}".format(formula))
```

```
Formula : close_lag7D ~ close + SMA + EMA + MACD_Signal + MACD_Hist + MACD + RSI + SlowK + SlowD +
ADX + CCI + dateSMA7 + dateSMA14 + dateSMA28 + dateSMA50 + dateSMA200 + date_converted + DateofMon
+ Month + Year + WeekSeq + SMA50 + SMA200 + SMA7 + SMA14 + SMA28 + Delta
```

```
#pdata = train_data.toPandas()
#pdata.head(5)


# HERE THE collected data is written to a csv file.
# This csv file is uploaded to S3

pandas_result_summary = result_summary.toPandas()
pandas_result_summary = pandas_result_summary.iloc[2:]
pandas_result_summary.to_csv("GBr2results_withfeatures.csv")
dbutils.fs.cp('file:/databricks/driver/r2results_withfeatures.csv','dbfs:/mnt/s3datav3/results
/r2results_withfeatures.csv')
pandas_result_summary.head(5)
```

```
  java.io.FileNotFoundException: File file:/databricks/driver/r2results_withfeatures.csv does not
  exist
```

```
#df_data2.show(10)
df_data3 = df_data2['close','DailyDate','close_lag7D']
df_data3.show(10)
```

```
+-----+-------------------+-----------+
|close|          DailyDate|close_lag7D|
+-----+-------------------+-----------+
|35.18|2010-11-03 00:00:00|      34.83|
|35.95|2010-11-04 00:00:00|      34.51|
|35.14|2010-11-05 00:00:00|      34.01|
|35.09|2010-11-08 00:00:00|      34.01|
|35.24|2010-11-09 00:00:00|      34.42|
|35.11|2010-11-10 00:00:00|      34.76|
|35.21|2010-11-11 00:00:00|      34.89|
|34.83|2010-11-12 00:00:00|      33.98|
|34.51|2010-11-15 00:00:00|      34.31|
|34.01|2010-11-16 00:00:00|      33.68|
+-----+-------------------+-----------+
only showing top 10 rows
```

```
import matplotlib.pyplot as plt
from pylab import rcParams
import seaborn as sns

pdata = df_data3.toPandas()
rcParams['figure.figsize'] = 25, 10
fig = plt.figure()

ax1 = sns.lineplot(data=pdata, x="DailyDate", y="close", markers=True, dashes=False,label ='raw
close')
#ax1 = sns.lineplot(data=pdata, x="date", y="avg7", markers=True, dashes=False)
ax1 = sns.lineplot(data=pdata, x="DailyDate", y="close_lag7D", markers=True, dashes=False,label
='close_lag7D')
#ax1 = sns.lineplot(data=pdata, x="date", y="avg28", markers=True, dashes=False)
#ax1 = sns.lineplot(data=pdata, x="date", y="avgof14D_lag7D", markers=True, dashes=False,label
='avgof14D_lag7D')
display(fig)
```

```
  Command skipped
```

```
#firstrow =
spark.createDataFrame([(stock,format(r2_score),format(rmse),format(r2_score),format(rmse))],
r2_columns)
#result_summary = result_summary.union(firstrow )
```

Command skipped

```
train_data.show(5)
```

Command skipped

```
#from pyspark.sql.functions import monotonically_increasing_id, lag
#from pyspark.sql.window import Window

#write_train_data = train_data['DailyDate','close']
#write_test_data = test_data['DailyDate']
#write_test_data = test_data.withColumn('id', monotonically_increasing_id())
#temp = y_true['label']

#write_test_data = write_test_data.withColumn('y_true', temp)
#write_test_data = write_test_data.withColumn('y_pred', y_pred)
```

Command skipped

```
import matplotlib.pyplot as plt
from pylab import rcParams
import seaborn as sns

data_zoom = write_train_data #df_stg2[(df_stg2['date'] > '2017-01-01' )]
pdata = data_zoom.toPandas()

rcParams['figure.figsize'] = 25, 10
fig = plt.figure()

ax1 = sns.lineplot(data=write_train_data, x="DailyDate", y="close", markers=True,
dashes=False,label ='raw open')
#ax1 = sns.lineplot(data=pdata, x="date", y="avg7", markers=True, dashes=False)
#ax1 = sns.lineplot(data=pdata, x="date", y="avgof14D_lag0D", markers=True, dashes=False,label
='avgof14D_lag0D')
#ax1 = sns.lineplot(data=pdata, x="date", y="avg28", markers=True, dashes=False)
#ax1 = sns.lineplot(data=pdata, x="date", y="avgof14D_lag7D", markers=True, dashes=False,label
='avgof14D_lag7D')
display(fig)
```

Command skipped

```
test_data
columns.remove('DailyDate')
```

```
  Command skipped

print(result_summary)

  Command skipped

bestModel = cvModel.bestModel
bestModel.stages[-1]._java_obj.parent().getRegParam()
#bestModel.getParam


  Command skipped

bestModel.save("/tmp/rf20200501a")

  Command skipped

rfPath = "/tmp/rf20200501a"
sameRFModel = PipelineModel.load(rfPath)

  Command skipped

sameRFModel
predictions = sameRFModel.transform(test_data)
evaluator = RegressionEvaluator(labelCol="label",
                                predictionCol="prediction",
                                metricName="rmse")

  Command skipped

import pandas as pd
import numpy as np
result_summary  = pd.DataFrame([])
for i in np.arange(0, 4):
    if i % 2 == 0:
        result_summary.append(pd.DataFrame({'A': i, 'B': i + 1}, index=[0]), ignore_index=True)
result_summary.head()

  Command skipped

print('done')
#%sh
#ls
#%sh
#ls -lrt

  Command skipped

cvModel.save('file:/databricks/driver/'+stock+'.csv')

  Command skipped

dbutils.fs.mkdirs("/mnt/"+MOUNT_NAME +"/allmodels/")

  Command skipped

best_lr = cvModel.bestModel
best_lr.save("/mnt/"+MOUNT_NAME +"/allmodels/testcvmodel/")
```

```
  Command skipped
```

```
rfPath = "/mnt/"+MOUNT_NAME +"/allmodels/testcvmodel/"
sameRFModel = PipelineModel.load(rfPath)
```

```
  Command skipped
```

```
  Command skipped
```