Zeppelin

2017-02-02-Flight-D...

default

FINISHED

```
%sh

wget http://stat-computing.org/dataexpo/2009/2007.csv.bz2 -O /Users/Hamster/Desktop/Air_F
wget http://stat-computing.org/dataexpo/2009/2008.csv.bz2 -O /Users/Hamster/Desktop/Air_F
echo "download"
```

```
--2017-02-02 23:20:28--  http://stat-computing.org/dataexpo/2009/2007.csv.bz2
Resolving stat-computing.org (stat-computing.org)... 52.218.128.147
Connecting to stat-computing.org (stat-computing.org)|52.218.128.147|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 121249243 (116M) [application/x-bzip2]
```

Took 1 min 1 sec. Last updated by anonymous at February 02 2017, 11:21:28 PM.

FINISHED

```
%sh

wget ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/by_year/2007.csv.gz -O /Users/Hamster/De
wget ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/by_year/2008.csv.gz -O /Users/Hamster/De
echo "download"
```

```
--2017-02-02 23:21:51--  ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/by_year/2007.csv.gz
           => '/Users/Hamster/Desktop/Air_Flight_Data/weather_2007.csv.gz'
Resolving ftp.ncdc.noaa.gov (ftp.ncdc.noaa.gov)... 205.167.25.101, 2610:20:8040:2::101
Connecting to ftp.ncdc.noaa.gov (ftp.ncdc.noaa.gov)|205.167.25.101|:21... connected.
Logging in as anonymous ... Logged in!
==> SYST ... done.    ==> PWD ... done.
==> TYPE I ... done.  ==> CWD (1) /pub/data/ghcn/daily/by_year ... done.
==> SIZE 2007.csv.gz ... 197708335
==> PASV ... done.    ==> RETR 2007.csv.gz ... done.
Length: 197708335 (189M) (unauthoritative)
     0K .......... .......... .......... .......... ..........  0%  674K 4m46s
    50K .......... .......... .......... .......... ..........  0%  907K 4m9s
   100K .......... .......... .......... .......... ..........  0% 1.35M 3m33s
   150K .......... .......... .......... .......... ..........  0% 1.34M 3m15s
   200K .......... .......... .......... .......... ..........  0% 1.38M 3m3s
   250K .......... .......... .......... .......... ..........  0% 12.4M 2m35s
   300K .......... .......... .......... .......... ..........  0% 1.79M 2m28s
   350K                                                        0% 1.55M 2m24s
```

Took 1 min 0 sec. Last updated by anonymous at February 02 2017, 11:22:51 PM.

FINISHED

```
%dep

z.reset()
z.load("joda-time:joda-time:2.9.1")
```

```
DepInterpreter(%dep) deprecated. Remove dependencies and repositories through GUI interpreter
 menu instead.
DepInterpreter(%dep) deprecated. Load dependency through GUI interpreter menu instead.
res0: org.apache.zeppelin.dep.Dependency = org.apache.zeppelin.dep.Dependency@4eb3722a
```

Took 1 sec. Last updated by anonymous at February 02 2017, 11:06:19 PM.

FINISHED default

```
%spark
import org.apache.spark.rdd._
import scala.collection.JavaConverters._
import au.com.bytecode.opencsv.CSVReader
```

```
import org.apache.spark.rdd._
import scala.collection.JavaConverters._
import au.com.bytecode.opencsv.CSVReader
```

Took 5 sec. Last updated by anonymous at February 02 2017, 11:06:27 PM.

FINISHED

```
import java.io._
import org.joda.time._
import org.joda.time.format._
import org.joda.time.format.DateTimeFormat
import org.joda.time.DateTime
import org.joda.time.Days
```

```
import java.io._
import org.joda.time._
import org.joda.time.format._
import org.joda.time.format.DateTimeFormat
import org.joda.time.DateTime
import org.joda.time.Days
```

Took 1 sec. Last updated by anonymous at February 02 2017, 11:06:31 PM.

FINISHED

```
case class DelayRec(year: String,
                    month: String,
                    dayOfMonth: String,
                    dayOfWeek: String,
                    crsDepTime: String,
                    depDelay: String,
                    origin: String,
                    distance: String,
                    cancelled: String) {

    val holidays = List("01/01/2007", "01/15/2007", "02/19/2007", "05/28/2007", "06/07/2
        "09/03/2007", "10/08/2007" ,"11/11/2007", "11/22/2007", "12/25/2007",
        "01/01/2008", "01/21/2008", "02/18/2008", "05/22/2008", "05/26/2008", "07/04/200
        "09/01/2008", "10/13/2008" ,"11/11/2008", "11/27/2008", "12/25/2008")

    def gen_features: (String, Array[Double]) = {
        val values = Array(
            depDelay.toDouble,
            month.toDouble,
            dayOfMonth.toDouble,
            dayOfWeek.toDouble,
            get_hour(crsDepTime).toDouble,
            distance.toDouble,
            days_from_nearest_holiday(year.toInt, month.toInt, dayOfMonth.toInt)
        )
        new Tuple2(to_date(year.toInt, month.toInt, dayOfMonth.toInt), values)
```

```
    }
    def get_hour(depTime: String) : String = "%04d".format(depTime.toInt).take(2)
    def to_date(year: Int, month: Int, day: Int) = "%04d%02d%02d".format(year, month, da

    def days_from_nearest_holiday(year:Int, month:Int, day:Int): Int = {
        val sampleDate = new org.joda.time.DateTime(year, month, day, 0, 0)

        holidays.foldLeft(3000) { (r, c) =>
        val holiday = org.joda.time.format.DateTimeFormat.forPattern("MM/dd/yyyy").parse
        val distance = Math.abs(org.joda.time.Days.daysBetween(holiday, sampleDate).getD
        math.min(r, distance)
     }
    }
```

```
defined class DelayRec
```

Took 1 sec. Last updated by anonymous at February 02 2017, 11:06:34 PM.

FINISHED

```
// function to do a preprocessing step for a given file
def prepFlightDelays(infile: String): RDD[DelayRec] = {
    val data = sc.textFile(infile)

    data.map { line =>
      val reader = new CSVReader(new StringReader(line))
      reader.readAll().asScala.toList.map(rec => DelayRec(rec(0),rec(1),rec(2),rec(3),re
    }.map(list => list(0))
    .filter(rec => rec.year != "Year")
    .filter(rec => rec.cancelled == "0")
    .filter(rec => rec.origin == "ORD")
}
```

```
prepFlightDelays: (infile: String)org.apache.spark.rdd.RDD[DelayRec]
```

Took 0 sec. Last updated by anonymous at February 02 2017, 11:06:37 PM.

FINISHED

```
val data_2007tmp = prepFlightDelays("/Users/Hamster/Desktop/Air_Flight_Data/flights_2007.
val data_2007 = data_2007tmp.map(rec => rec.gen_features._2)
val data_2008 = prepFlightDelays("/Users/Hamster/Desktop/Air_Flight_Data/flights_2008.csv

data_2007tmp.toDF().registerTempTable("data_2007tmp")

data_2007.take(5).map(x => x mkString ",").foreach(println)
```

```
data_2007tmp: org.apache.spark.rdd.RDD[DelayRec] = MapPartitionsRDD[6] at filter at <console>:
58
data_2007: org.apache.spark.rdd.RDD[Array[Double]] = MapPartitionsRDD[7] at map at <console>:5
2
data_2008: org.apache.spark.rdd.RDD[Array[Double]] = MapPartitionsRDD[15] at map at <console>:
50
warning: there was one deprecation warning; re-run with -deprecation for details
-8.0,1.0,25.0,4.0,11.0,719.0,10.0
41.0,1.0,28.0,7.0,15.0,925.0,13.0
45.0,1.0,29.0,1.0,20.0,316.0,14.0
-9.0,1.0,17.0,3.0,19.0,719.0,2.0
180.0,1.0,12.0,5.0,17.0,316.0,3.0
```

Took 8 sec. Last updated by anonymous at February 02 2017, 11:06:49 PM.

FINISHED

```
1 %sql
2 select dayofWeek, case when depDelay > 15 then 'delayed' else 'ok' end , count(1)
3 from data_2007tmp group by dayofweek , case when depDelay > 15 then 'delayed' else 'ok' e
```

default

dayofWeek	CASE WHEN (CAST(depDelay AS DOUBLE) > CAST(15 AS DOUBLE)) THEN c
1	delayed
7	ok
1	ok
6	delayed
2	delayed
3	ok
4	delayed
3	delayed
5	ok

Took 25 sec. Last updated by anonymous at February 02 2017, 11:07:32 PM.

FINISHED

```
1 %sql
2 select cast( cast(crsDepTime as int) / 100 as int) as hour,  case when depDelay > 15 ther
  when depDelay > 15 then 'delayed' else 'ok' end
```

hour	delay
12	ok
13	ok
20	delayed
10	ok
19	ok
15	ok
15	delayed
21	ok
8	ok

Took 23 sec. Last updated by anonymous at February 02 2017, 11:09:18 PM.

FINISHED

```
1 %spark
```

```
import org.apache.spark.mllib.regression.LabeledPoint
import org.apache.spark.mllib.linalg.Vectors
import org.apache.spark.mllib.feature.StandardScaler
```

```
import org.apache.spark.mllib.regression.LabeledPoint
import org.apache.spark.mllib.linalg.Vectors
import org.apache.spark.mllib.feature.StandardScaler
```

Took 0 sec. Last updated by anonymous at February 02 2017, 11:09:25 PM.

FINISHED

```
def parseData(vals: Array[Double]): LabeledPoint = {
    LabeledPoint(if (vals(0)>=15) 1.0 else 0.0, Vectors.dense(vals.drop(1)))
  }
```

```
parseData: (vals: Array[Double])org.apache.spark.mllib.regression.LabeledPoint
```

Took 0 sec. Last updated by anonymous at February 02 2017, 11:09:28 PM.

FINISHED

```
// Prepare training set
val parsedTrainData = data_2007.map(parseData)
parsedTrainData.cache
val scaler = new StandardScaler(withMean = true, withStd = true).fit(parsedTrainData.map(
val scaledTrainData = parsedTrainData.map(x => LabeledPoint(x.label, scaler.transform(Vec
scaledTrainData.cache
```

```
parsedTrainData: org.apache.spark.rdd.RDD[org.apache.spark.mllib.regression.LabeledPoint] = Ma
pPartitionsRDD[48] at map at <console>:60
res11: parsedTrainData.type = MapPartitionsRDD[48] at map at <console>:60
scaler: org.apache.spark.mllib.feature.StandardScalerModel = org.apache.spark.mllib.feature.St
andardScalerModel@4b73aaf8
scaledTrainData: org.apache.spark.rdd.RDD[org.apache.spark.mllib.regression.LabeledPoint] = Ma
pPartitionsRDD[51] at map at <console>:63
res12: scaledTrainData.type = MapPartitionsRDD[51] at map at <console>:63
```

Took 28 sec. Last updated by anonymous at February 02 2017, 11:09:58 PM.

FINISHED

```
// Prepare test/validation set
val parsedTestData = data_2008.map(parseData)
parsedTestData.cache
val scaledTestData = parsedTestData.map(x => LabeledPoint(x.label, scaler.transform(Vecto
scaledTestData.cache
```

```
parsedTestData: org.apache.spark.rdd.RDD[org.apache.spark.mllib.regression.LabeledPoint] = Map
PartitionsRDD[52] at map at <console>:58
res15: parsedTestData.type = MapPartitionsRDD[52] at map at <console>:58
scaledTestData: org.apache.spark.rdd.RDD[org.apache.spark.mllib.regression.LabeledPoint] = Map
PartitionsRDD[53] at map at <console>:67
res16: scaledTestData.type = MapPartitionsRDD[53] at map at <console>:67
```

Took 1 sec. Last updated by anonymous at February 02 2017, 11:10:12 PM.

FINISHED

```
scaledTrainData.take(3).map(x => (x.label, x.features)).foreach(println)
```

```
(0.0,[-1.6160463330366632,1.0549272994666004,0.03217026353736743,-0.518924417544128,0.03408393
342430724,-0.28016830994663705])
```

2017-02-02-Flight-Data

```
(1.0,[-1.6160463330366632,1.5098311893165115,-1.4710902487728246,1.4641277433573872,-0.7436888
225169944,0.062357586732433884])
```

Took 1 sec. Last updated by anonymous at February 02 2017, 11:10:16 PM.

default

FINISHED

```
// Function to compute evaluation metrics
def eval_metrics(labelsAndPreds: RDD[(Double, Double)]) : Tuple2[Array[Double], Array[Dc
    val tp = labelsAndPreds.filter(r => r._1==1 && r._2==1).count.toDouble
    val tn = labelsAndPreds.filter(r => r._1==0 && r._2==0).count.toDouble
    val fp = labelsAndPreds.filter(r => r._1==1 && r._2==0).count.toDouble
    val fn = labelsAndPreds.filter(r => r._1==0 && r._2==1).count.toDouble

    val precision = tp / (tp+fp)
    val recall = tp / (tp+fn)
    val F_measure = 2*precision*recall / (precision+recall)
    val accuracy = (tp+tn) / (tp+tn+fp+fn)
    new Tuple2(Array(tp, tn, fp, fn), Array(precision, recall, F_measure, accuracy))

}
```

```
eval_metrics: (labelsAndPreds: org.apache.spark.rdd.RDD[(Double, Double)])(Array[Double], Arra
y[Double])
```

Took 1 sec. Last updated by anonymous at February 02 2017, 11:10:20 PM.

FINISHED

```
import org.apache.spark.rdd._
import org.apache.spark.rdd.RDD
```

```
import org.apache.spark.rdd._
import org.apache.spark.rdd.RDD
```

Took 0 sec. Last updated by anonymous at February 02 2017, 11:10:23 PM.

FINISHED

```
class Metrics(labelsAndPreds: org.apache.spark.rdd.RDD[(Double, Double)]) extends java.i

    private def filterCount(lftBnd:Int,rtBnd:Int):Double = labelsAndPreds
                                                           .map(x => (x._1.toInt, x._2.t
                                                           .filter(_ == (lftBnd,rtBnd)).

    lazy val tp = filterCount(1,1)  // true positives
    lazy val tn = filterCount(0,0)  // true negatives
    lazy val fp = filterCount(0,1)  // false positives
    lazy val fn = filterCount(1,0)  // false negatives

    lazy val precision = tp / (tp+fp)
    lazy val recall = tp / (tp+fn)
    lazy val F1 = 2*precision*recall / (precision+recall)
    lazy val accuracy = (tp+tn) / (tp+tn+fp+fn)
}
```

```
defined class Metrics
```

Took 0 sec. Last updated by anonymous at February 02 2017, 11:10:26 PM.

FINISHED

2017-02-02-Flight-Data

```
%spark

import org.apache.spark.mllib.classification.LogisticRegressionWithSGD

// Build the LogisticRegression model
val model_lr = LogisticRegressionWithSGD.train(scaledTrainData, numIterations=100)

// Predict
val labelsAndPreds_lr = scaledTestData.map { point =>
    val pred = model_lr.predict(point.features)
    (pred, point.label)
}
val m_lr = eval_metrics(labelsAndPreds_lr)._2
```

```
import org.apache.spark.mllib.classification.LogisticRegressionWithSGD
warning: there was one deprecation warning; re-run with -deprecation for details
model_lr: org.apache.spark.mllib.classification.LogisticRegressionModel = org.apache.spark.mll
ib.classification.LogisticRegressionModel: intercept = 0.0, numFeatures = 6, numClasses = 2, t
hreshold = 0.5
labelsAndPreds_lr: org.apache.spark.rdd.RDD[(Double, Double)] = MapPartitionsRDD[139] at map a
t <console>:78
m_lr: Array[Double] = Array(0.3735363068960268, 0.6427763108261033, 0.47249298123322336, 0.591
5277487847792)
precision = 0.37, recall = 0.64, F1 = 0.47, accuracy = 0.59
```

Took 30 sec. Last updated by anonymous at February 02 2017, 11:11:02 PM.

FINISHED

```
println(model_lr.weights)
```

```
[-0.05519239973775391,0.005877388355994204,-0.03625359858318008,0.3903949271784436,0.049943146
70964247,7.940537333813864E-4]
```

Took 0 sec. Last updated by anonymous at February 02 2017, 11:11:22 PM.

FINISHED

```
%spark

import org.apache.spark.mllib.tree.DecisionTree

// Build the Decision Tree model
val numClasses = 2
val categoricalFeaturesInfo = Map[Int, Int]()
val impurity = "gini"
val maxDepth = 10
val maxBins = 100
val model_dt = DecisionTree.trainClassifier(parsedTrainData, numClasses, categoricalFeat

// Predict
val labelsAndPreds_dt = parsedTestData.map { point =>
    val pred = model_dt.predict(point.features)
    (pred, point.label)
}
val m_dt = eval_metrics(labelsAndPreds_dt)._2
println("precision = %.2f, recall = %.2f, F1 = %.2f, accuracy = %.2f".format(m_dt(0), m_
```

```
import org.apache.spark.mllib.tree.DecisionTree
numClasses: Int = 2
categoricalFeaturesInfo: scala.collection.immutable.Map[Int,Int] = Map()
impurity: String = gini
```

```
model_dt: org.apache.spark.mllib.tree.model.DecisionTreeModel = DecisionTreeModel classifier o
f depth 10 with 1845 nodes
labelsAndPreds_dt: org.apache.spark.rdd.RDD[(Double, Double)] = MapPartitionsRDD[183] at map a
t <console>:83
m_dt: Array[Double] = Array(0.40568143388569494, 0.25139360409069955, 0.31042335161991513, 0.6
821280529627531)
precision = 0.41, recall = 0.25, F1 = 0.31, accuracy = 0.68
```

Took 3 sec. Last updated by anonymous at February 02 2017, 11:11:31 PM.

FINISHED

```
%spark

import org.apache.spark.mllib.tree.RandomForest
import org.apache.spark.mllib.tree.configuration.Strategy

val treeStrategy = Strategy.defaultStrategy("Classification")
val numTrees = 40  // The original numTrees was 100, but decreased it to 40. ALB
val featureSubsetStrategy = "auto" // Let the algorithm choose
val model_rf = RandomForest.trainClassifier(parsedTrainData, treeStrategy, numTrees, fea


```

```
import org.apache.spark.mllib.tree.RandomForest
import org.apache.spark.mllib.tree.configuration.Strategy
treeStrategy: org.apache.spark.mllib.tree.configuration.Strategy = org.apache.spark.mllib.tre
e.configuration.Strategy@6b04339e
numTrees: Int = 40
featureSubsetStrategy: String = auto
model_rf: org.apache.spark.mllib.tree.model.RandomForestModel =
TreeEnsembleModel classifier with 40 trees
```

Took 17 sec. Last updated by anonymous at February 02 2017, 11:12:03 PM.

FINISHED

```
%spark

// Predict
val labelsAndPreds_rf = parsedTestData.map { point =>
    val pred = model_rf.predict(point.features)
    (point.label, pred)
}

val m_rf = new Metrics(labelsAndPreds_rf)
println("precision = %.2f, recall = %.2f, F1 = %.2f, accuracy = %.2f"
        .format(m_rf.precision, m_rf.recall, m_rf.F1, m_rf.accuracy))

```

```
labelsAndPreds_rf: org.apache.spark.rdd.RDD[(Double, Double)] = MapPartitionsRDD[227] at map a
t <console>:82
m_rf: Metrics = Metrics@27086c56
precision = 0.48, recall = 0.14, F1 = 0.22, accuracy = 0.71
```

Took 14 sec. Last updated by anonymous at February 02 2017, 11:12:21 PM.

FINISHED

```
%spark

```

```
import org.apache.spark.SparkContext._
import scala.collection.JavaConverters._
import au.com.bytecode.opencsv.CSVReader
import java.io._

// function to do a preprocessing step for a given file

def preprocess_spark(delay_file: String, weather_file: String): RDD[Array[Double]] = {
  // Read wether data
  val delayRecs = prepFlightDelays(delay_file).map{ rec =>
        val features = rec.gen_features
        (features._1, features._2)
  }

  // Read weather data into RDDs
  val station_inx = 0
  val date_inx = 1
  val metric_inx = 2
  val value_inx = 3

  def filterMap(wdata:RDD[Array[String]], metric:String):RDD[(String,Double)] = {
    wdata.filter(vals => vals(metric_inx) == metric).map(vals => (vals(date_inx), vals(v
  }

  val wdata = sc.textFile(weather_file).map(line => line.split(","))
                    .filter(vals => vals(station_inx) == "USW00094846")
  val w_tmin = filterMap(wdata,"TMIN")
  val w_tmax = filterMap(wdata,"TMAX")
  val w_prcp = filterMap(wdata,"PRCP")
  val w_snow = filterMap(wdata,"SNOW")
  val w_awnd = filterMap(wdata,"AWND")

  delayRecs.join(w_tmin).map(vals => (vals._1, vals._2._1 ++ Array(vals._2._2)))
           .join(w_tmax).map(vals => (vals._1, vals._2._1 ++ Array(vals._2._2)))
           .join(w_prcp).map(vals => (vals._1, vals._2._1 ++ Array(vals._2._2)))
           .join(w_snow).map(vals => (vals._1, vals._2._1 ++ Array(vals._2._2)))
           .join(w_awnd).map(vals => vals._2._1 ++ Array(vals._2._2))
}


val data_2007 = preprocess_spark("/Users/Hamster/Desktop/Air_Flight_Data/flights_2007.cs
val data_2008 = preprocess_spark("/Users/Hamster/Desktop/Air_Flight_Data/flights_2008.cs

data_2007.take(5).map(x => x mkString ",").foreach(println)
```

```
import org.apache.spark.SparkContext._
import scala.collection.JavaConverters._
import au.com.bytecode.opencsv.CSVReader
import java.io._
preprocess_spark: (delay_file: String, weather_file: String)org.apache.spark.rdd.RDD[Array[Dou
ble]]
data_2007: org.apache.spark.rdd.RDD[Array[Double]] = MapPartitionsRDD[277] at map at <console
>:100
data_2008: org.apache.spark.rdd.RDD[Array[Double]] = MapPartitionsRDD[319] at map at <console
>:100
42.0,11.0,25.0,7.0,20.0,316.0,3.0,-22.0,33.0,0.0,0.0,46.0
-4.0,11.0,25.0,7.0,6.0,316.0,3.0,-22.0,33.0,0.0,0.0,46.0
0.0,11.0,25.0,7.0,12.0,925.0,3.0,-22.0,33.0,0.0,0.0,46.0
32.0,11.0,25.0,7.0,17.0,316.0,3.0,-22.0,33.0,0.0,0.0,46.0
-3.0,11.0,25.0,7.0,13.0,654.0,3.0,-22.0,33.0,0.0,0.0,46.0
```

Zeppelin

2017-02-02-Flight-D...

FINISHED default

```
%spark

import org.apache.spark.mllib.regression.LabeledPoint
import org.apache.spark.mllib.linalg.Vectors
import org.apache.spark.mllib.feature.StandardScaler

def parseData(vals: Array[Double]): LabeledPoint = {
  LabeledPoint(if (vals(0)>=15) 1.0 else 0.0, Vectors.dense(vals.drop(1)))
}

// Prepare training set
val parsedTrainData = data_2007.map(parseData)
val scaler = new StandardScaler(withMean = true, withStd = true).fit(parsedTrainData.map
val scaledTrainData = parsedTrainData.map(x => LabeledPoint(x.label, scaler.transform(Ve
parsedTrainData.cache
scaledTrainData.cache

// Prepare test/validation set
val parsedTestData = data_2008.map(parseData)
val scaledTestData = parsedTestData.map(x => LabeledPoint(x.label, scaler.transform(Vect
parsedTestData.cache
scaledTestData.cache

scaledTrainData.take(5).map(x => (x.label, x.features)).foreach(println)
```

```
import org.apache.spark.mllib.regression.LabeledPoint
import org.apache.spark.mllib.linalg.Vectors
import org.apache.spark.mllib.feature.StandardScaler
parseData: (vals: Array[Double])org.apache.spark.mllib.regression.LabeledPoint
parsedTrainData: org.apache.spark.rdd.RDD[org.apache.spark.mllib.regression.LabeledPoint] = Ma
pPartitionsRDD[320] at map at <console>:80
scaler: org.apache.spark.mllib.feature.StandardScalerModel = org.apache.spark.mllib.feature.St
andardScalerModel@75704fdf
scaledTrainData: org.apache.spark.rdd.RDD[org.apache.spark.mllib.regression.LabeledPoint] = Ma
pPartitionsRDD[323] at map at <console>:84
res36: parsedTrainData.type = MapPartitionsRDD[320] at map at <console>:80
res37: scaledTrainData.type = MapPartitionsRDD[323] at map at <console>:84
parsedTestData: org.apache.spark.rdd.RDD[org.apache.spark.mllib.regression.LabeledPoint] = Map
PartitionsRDD[324] at map at <console>:80
scaledTestData: org.apache.spark.rdd.RDD[org.apache.spark.mllib.regression.LabeledPoint] = Map
PartitionsRDD[325] at map at <console>:88
res38: parsedTestData.type = MapPartitionsRDD[324] at map at <console>:80
res39: scaledTestData.type = MapPartitionsRDD[325] at map at <console>:88
```

Took 4 sec. Last updated by anonymous at February 02 2017, 11:13:52 PM.

FINISHED

```
%spark

import org.apache.spark.mllib.classification.LogisticRegressionWithSGD

// Build the Logistic Regression model
val model_lr = LogisticRegressionWithSGD.train(scaledTrainData, numIterations=100)

// Predict
val labelsAndPreds_lr = scaledTestData.map { point =>
```

```
    val pred = model_lr.predict(point.features)
    (point.label, pred)
}
val m_lr = new Metrics(labelsAndPreds_lr)
println("precision = %.2f, recall = %.2f, F1 = %.2f, accuracy = %.2f"
        .format(m_lr.precision, m_lr.recall, m_lr.F1, m_lr.accuracy))
```

```
import org.apache.spark.mllib.classification.LogisticRegressionWithSGD
warning: there was one deprecation warning; re-run with -deprecation for details
model_lr: org.apache.spark.mllib.classification.LogisticRegressionModel = org.apache.spark.mll
ib.classification.LogisticRegressionModel: intercept = 0.0, numFeatures = 11, numClasses = 2,
 threshold = 0.5
labelsAndPreds_lr: org.apache.spark.rdd.RDD[(Double, Double)] = MapPartitionsRDD[475] at map a
t <console>:95
m_lr: Metrics = Metrics@452ddebd
precision = 0.40, recall = 0.68, F1 = 0.50, accuracy = 0.62
```

Took 1 min 3 sec. Last updated by anonymous at February 02 2017, 11:15:00 PM.

FINISHED

```
%spark

println(model_lr.weights)
```

```
[-0.002641910361922112,0.016515122840474603,-0.021125675242088107,0.42541032171918614,0.047846
74910826409,0.010015678133278754,0.02539098796205592,-0.15003098865096928,0.2829929186564463,
0.2365535217050361,0.15029474134967363]
```

Took 1 sec. Last updated by anonymous at February 02 2017, 11:15:09 PM.

FINISHED

```
%spark

import org.apache.spark.mllib.tree.DecisionTree

// Build the Decision Tree model
val numClasses = 2
val categoricalFeaturesInfo = Map[Int, Int]()
val impurity = "gini"
val maxDepth = 10
val maxBins = 100
val model_dt = DecisionTree.trainClassifier(parsedTrainData, numClasses, categoricalFeat

// Predict
val labelsAndPreds_dt = parsedTestData.map { point =>
    val pred = model_dt.predict(point.features)
    (point.label, pred)
}
val m_dt = new Metrics(labelsAndPreds_dt)
println("precision = %.2f, recall = %.2f, F1 = %.2f, accuracy = %.2f"
        .format(m_dt.precision, m_dt.recall, m_dt.F1, m_dt.accuracy))
```

```
import org.apache.spark.mllib.tree.DecisionTree
numClasses: Int = 2
categoricalFeaturesInfo: scala.collection.immutable.Map[Int,Int] = Map()
impurity: String = gini
maxDepth: Int = 10
maxBins: Int = 100
model_dt: org.apache.spark.mllib.tree.model.DecisionTreeModel = DecisionTreeModel classifier o
f depth 10 with 1861 nodes
```

```
labelsAndPreds_dt: org.apache.spark.rdd.RDD[(Double, Double)] = MapPartitionsRDD[523] at map a
t <console>:98
m_dt: Metrics = Metrics@5496fa95
precision = 0.51, recall = 0.32, F1 = 0.40, accuracy = 0.72
```

Took 7 sec. Last updated by anonymous at February 02 2017, 11:15:20 PM.

default

FINISHED

```
%spark

import org.apache.spark.mllib.tree.RandomForest
import org.apache.spark.mllib.tree.configuration.Strategy

val treeStrategy = Strategy.defaultStrategy("Classification")
val model_rf = RandomForest.trainClassifier(parsedTrainData,
                                            treeStrategy,
                                            numTrees = 40, // Original numTrees was 100,
                                            featureSubsetStrategy = "auto", seed = 125)

// Predict
val labelsAndPreds_rf = parsedTestData.map { point =>
    val pred = model_rf.predict(point.features)
    (point.label, pred)
}
val m_rf = new Metrics(labelsAndPreds_rf)
println("precision = %.2f, recall = %.2f, F1 = %.2f, accuracy = %.2f"
        .format(m_rf.precision, m_rf.recall, m_rf.F1, m_rf.accuracy))
```

```
import org.apache.spark.mllib.tree.RandomForest
import org.apache.spark.mllib.tree.configuration.Strategy
treeStrategy: org.apache.spark.mllib.tree.configuration.Strategy = org.apache.spark.mllib.tre
e.configuration.Strategy@3f6794d5
model_rf: org.apache.spark.mllib.tree.model.RandomForestModel =
TreeEnsembleModel classifier with 40 trees
labelsAndPreds_rf: org.apache.spark.rdd.RDD[(Double, Double)] = MapPartitionsRDD[571] at map a
t <console>:98
m_rf: Metrics = Metrics@39edf839
precision = 0.59, recall = 0.34, F1 = 0.43, accuracy = 0.74
```

Took 49 sec. Last updated by anonymous at February 02 2017, 11:16:48 PM.

READY