from pyspark.mllib.recommendation import ALS

data=sc.textFile('/FileStore/tables/ratings.dat')

def splitData(line):

data=line.strip().split('::')

return int(data[3]) % 10, (int(data[0]), int(data[1]), float(data[2]))

splitDataRDD=data.map(splitData)

numRatings=splitDataRDD.count()

totalUsers=splitDataRDD.mapValues(lambda x:x[0]).distinct().count()

totalMovies=splitDataRDD.mapValues(lambda x: x[1]).distinct().count()

trainingData=splitDataRDD.filter(lambda x:x[0]<6).values()

testData=splitDataRDD.filter(lambda x:x[0]>=6).values()

trainingCount=trainingData.count()

testCount=testData.count()

rank = 10

iterations = 10

lmbda=.1

model = ALS.train(trainingData, rank, iterations, lmbda)

testDataForPrediction=testData.map(lambda x: (x[0],x[1]))

predictions=model.predictAll(testDataForPrediction).map(lambda x :((x[0],x[1]),x[2]))

predictionAndRating=testData.map(lambda x: ((x[0],x[1]),x[2])).join(predictions).values()

MSE = predictionAndRating.map(lambda x: (x[0] - x[1])\*\*2).mean()

MSE