**KENNESAW STATE UNIVERSITY**

COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE

CS7050 Data Warehouse and mining

Section W02-Fall 2022

PROJECT

Aswani I Chandran

Priyanka Bagal

[achandr2@students.kennesaw.edu](mailto:achandr2@students.kennesaw.edu)

[pbagal@students.kennesaw.edu](mailto:pbagal@students.kennesaw.edu)

11/29/2022

**SOURCE-CODE**

**Decision Tree & Random Forest**

*//Covid-19 dataset*

import org.apache.spark.ml.linalg.\_

import org.apache.spark.ml.regression.\_

import org.apache.spark.ml.classification.{DecisionTreeClassifier,RandomForestClassifier, RandomForestClassificationModel}

import org.apache.spark.ml.feature.{VectorAssembler, VectorIndexer}

import org.apache.spark.ml.evaluation.MulticlassClassificationEvaluator

import org.apache.spark.mllib.evaluation.MulticlassMetrics

import org.apache.spark.ml.tuning.{ParamGridBuilder,TrainValidationSplit}

import org.apache.spark.ml.{PipelineModel, Pipeline}

import org.apache.spark.sql.types.{StructType,StructField,StringType,IntegerType};

*// Read the data//*

val data = spark.read.option("header","true").option("inferSchema","true").format("csv").load("C:/Spark/Cleaned-Data.csv")

*// check data//*

data.count

data.printSchema

*//drop insignificant columns*

val df=data.drop("Gender\_Female","Gender\_Male","Gender\_Transgender","Contact\_Dont-Know","Contact\_No","Contact\_Yes","Country")

df.printSchema

val data = df.toDF()

val Severity = Array("Severity\_Mild","Severity\_Moderate","Severity\_None","Severity\_Severe")

*// Assemble the four target variables*

val Severityassembler = new VectorAssembler(). setInputCols(Severity). setOutputCol("severity")

*//One-hot encoding of data*

val unhotUDF = udf((vec:Vector)=>if(vec.toArray.sum==0)0.0 else vec.toArray.indexOf(1).toDouble)

*//prepare the final data*

val withseverity = Severityassembler.transform(data).drop(Severity:\_\*). withColumn("severity", unhotUDF($"severity"))

*//Split in to 80% train, 10% cross validation,10% testing*

val Array(trainData, cvData, testData) = withseverity.randomSplit(Array(0.8,0.1, 0.1))

trainData.cache()

cvData.cache()

testData.cache()

val inputCols = trainData.columns.filter(\_ != "severity")

val assembler = new VectorAssembler().setInputCols(inputCols).setOutputCol("featureVector")

val assembledTrainData = assembler.transform(trainData)

assembledTrainData.select("featureVector").show(truncate = false)

val assembledCvData = assembler.transform(cvData)

assembledCvData.select("featureVector").show(truncate = false)

// first simple decision tree

//Target--> severity

val classifier = new DecisionTreeClassifier().setSeed(2345).setLabelCol("severity").setFeaturesCol("featureVector").setPredictionCol("prediction").setImpurity("gini")

*// train the model*

val model = classifier.fit(assembledTrainData)

*// show the tree*

println(model.toDebugString)

*// explore the model*

model.featureImportances.toArray.zip(inputCols).sorted.reverse.take(10).foreach(println)

*// make predictions*

val predictions = model.transform(assembledCvData)

predictions.select("severity", "prediction", "probability").show(false)

*//Evaluate the model*

val evaluator = new MulticlassClassificationEvaluator().setLabelCol("severity").setPredictionCol("prediction")

val accuracy = evaluator.setMetricName("accuracy").evaluate(predictions)

val f1 = evaluator.setMetricName("f1").evaluate(predictions)

println(accuracy)

println(f1)

val weightedPrecision = evaluator.setMetricName("weightedPrecision").evaluate(predictions)

println(weightedPrecision)

val predictionRDD = predictions.select("prediction", "severity").as[(Double,Double)].rdd

val multiclassMetrics = new MulticlassMetrics(predictionRDD)

println(multiclassMetrics.confusionMatrix)

*// Create confusion Matrix*

val confusionMatrix = predictions.groupBy("severity").pivot("prediction", (0 to 3)).count().na.fill(0.0).orderBy("severity")

confusionMatrix.show()

val assembler = new VectorAssembler().setInputCols(inputCols).setOutputCol("featureVector")

val indexer = new VectorIndexer().setMaxCategories(40).setInputCol("featureVector").setOutputCol("indexedVector")

// Random Forest Classifier//

val classifier = new RandomForestClassifier().setSeed(123456).setLabelCol("severity").setFeaturesCol("indexedVector").setPredictionCol("prediction").setImpurity("entropy").setMaxDepth(20).setMaxBins(300)

val pipeline = new Pipeline().setStages(Array(assembler, indexer, classifier))

*// create param grid*

val paramGrid = new ParamGridBuilder().addGrid(classifier.impurity, Seq("gini", "entropy")).addGrid(classifier.maxDepth, Seq(1, 20)).addGrid(classifier.maxBins, Seq(40, 300)).addGrid(classifier.minInfoGain, Seq(0.0, 0.05)).build()

val multiclassEval = new MulticlassClassificationEvaluator().setLabelCol("severity").setPredictionCol("prediction").setMetricName("accuracy")

val validator = new TrainValidationSplit().setSeed(12345).setEstimator(pipeline).setEvaluator(multiclassEval).setEstimatorParamMaps(paramGrid).setTrainRatio(0.9)

val validatorModel = validator.fit(trainData)

val bestModel = validatorModel.bestModel

val forestModel = bestModel.asInstanceOf[PipelineModel].stages.last.asInstanceOf[RandomForestClassificationModel]

println(forestModel.extractParamMap)

println(forestModel.getNumTrees)

forestModel.featureImportances.toArray.zip(inputCols).sorted.reverse.foreach(println)

val testAccuracy = multiclassEval.evaluate(bestModel.transform(testData))

println(testAccuracy)

//91.7%

**//Logistic Regression//**

import org.apache.spark.ml.linalg.\_

import org.apache.spark.ml.regression.\_

import org.apache.spark.ml.classification.LogisticRegression

import org.apache.spark.ml.{PipelineModel, Pipeline}

import org.apache.spark.mllib.evaluation.MulticlassMetrics

import org.apache.spark.ml.tuning.{ParamGridBuilder, TrainValidationSplit}

import org.apache.spark.ml.evaluation.RegressionEvaluator

import org.apache.spark.ml.evaluation.MulticlassClassificationEvaluator

import org.apache.spark.sql.types.{StructType,StructField,StringType,IntegerType};

import org.apache.spark.ml.feature.{VectorAssembler, VectorIndexer}

import org.apache.spark.ml.evaluation.MulticlassClassificationEvaluator

*//Read data*

val data = spark.read.option("header","true").option("inferSchema","true").format("csv").load("C:/Spark/Cleaned-Data.csv")

val df=data.drop("Gender\_Female","Gender\_Male","Gender\_Transgender","Contact\_Dont-Know","Contact\_No","Contact\_Yes","Country")

val data = df.toDF()

val Severity = Array("Severity\_Mild","Severity\_Moderate","Severity\_None","Severity\_Severe")

*//Assemble the target variables*

val Severityassembler = new VectorAssembler(). setInputCols(Severity). setOutputCol("severity")

*//One-hot encoding*

val unhotUDF = udf((vec:Vector)=>if(vec.toArray.sum==0)0.0 else vec.toArray.indexOf(1).toDouble)

*//Set final data*

val withseverity = Severityassembler.transform(data).drop(Severity:\_\*). withColumn("severity", unhotUDF($"severity"))

val seed = 5043

*//Split in to 80% train,20% test*

val Array(trainData,testData) = withseverity.randomSplit(Array(0.8,0.2))

trainData.cache()

testData.cache()

val inputCols = trainData.columns.filter(\_ != "severity")

val assembler = new VectorAssembler().setInputCols(inputCols).setOutputCol("featureVector")

val indexer = new VectorIndexer().setMaxCategories(40).setInputCol("featureVector").setOutputCol("indexedVector")

**//Multinomial Logistic Regression Classifier with Elastic Net Parameter//**

val classifier = new LogisticRegression().setLabelCol("severity").setFeaturesCol("indexedVector").setPredictionCol("prediction").setMaxIter(100)

.setRegParam(0.02)

.setElasticNetParam(1)

val stages = Array(assembler, indexer, classifier)

val pipeline = new Pipeline().setStages(stages)

val multiclassEval = new MulticlassClassificationEvaluator().setLabelCol("severity").setPredictionCol("prediction").setMetricName("accuracy")

*//Tuning hyper parameter*

*//We use a ParamGridBuilder to construct a grid of parameters to search over//*

val paramGrid = new ParamGridBuilder().addGrid(classifier.regParam, Seq(0.1, 0.01)).addGrid(classifier.elasticNetParam, Seq(0.0,0.5,0.8,1.0)).addGrid(classifier.fitIntercept).build()

*// TrainValidationSplit will try all combinations of values and determine best model using the evaluator//*

val validator = new TrainValidationSplit()

.setEstimator(pipeline)

.setEvaluator(multiclassEval)

.setEstimatorParamMaps(paramGrid)

.setTrainRatio(0.8)

val validatorModel = validator.fit(trainData)

val paramsAndMetrics = validatorModel.validationMetrics.zip(validatorModel.getEstimatorParamMaps).sortBy(-\_.\_1)

paramsAndMetrics.foreach { case (metric, params) =>

println(metric)

println(params)

println()}

val bestModel = validatorModel.bestModel

println(bestModel.asInstanceOf[PipelineModel].stages.last.extractParamMap)

println(validatorModel.validationMetrics.max)

val cvAccuracy = multiclassEval.evaluate(bestModel.transform(cvData))

println(cvAccuracy)

val testAccuracy = multiclassEval.evaluate(bestModel.transform(testData))

println(testAccuracy)

val trainAccuracy = multiclassEval.evaluate(bestModel.transform(trainData))

println(trainAccuracy)