Shuyan Bian - SEC01 (NUID 002309442)

Big Data System Engineering with Scala Spring 2024 Assignment No. 7



- GitHub Repo URL -

- List of Tasks Implemented

- Code

FileName.Scala (Either text or screenshot can be added below of your code snippet, don't add the whole code just the part you completed)

- Unit tests / Results

1. Exploratory Data Analysis

First class tickets are far more expensive than others.

The survival rate of first-class passengers is significantly higher than that of second-class passengers, and the survival rate of second-class passengers is significantly higher than that of third-class passengers.

```
titanic.withColumn("Group", ceil(col("Age")/10)).groupBy(
"Group").agg(avg("Fare"), (sum("Survived")/count("*")).as("Surviv
al_rate")).orderBv("Group").show();
                avg(Fare)
                                 Survival_rate
 Group
  NULL 22.158566666666673
                            0.2937853107344633
     1 30 . 434439062500008
                                       0.59375
     2 | 29 . 529531304347838 |
                            0.3826086956521739
     3 | 28 . 306718695652194 |
                            0.3652173913043478
     4 42.49610000000002 0.44516129032258067
        41.16318139534884 0.38372093023255816
     6 44.77480238095238 0.40476190476190477
     7 45.91078235294117 0.23529411764705882
     8 | 25 . 936680000000003 |
                                           0.2
```

The survival rate of teenagers is significantly higher than that of other age groups.

2. Feature Engineering

There are only two entries with Embarked = NULL, so I remove them.

Because the Cabin column has too many missing values, I remove it.

The Passengerld column is not used in the process of training, but it needs to be preserved for producing submission.

I drop the Ticket column, since I think it is hard to use.

SibSp and Parch can be combined to produce a FamilySize column. (Of course, Sibsp and Parch columns can be removed.)

There are many missing values in the Age column. I fill them with the average.

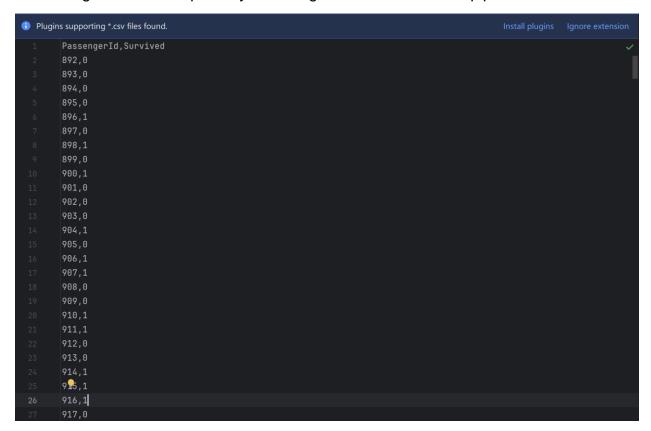
Finally, I find that the test data also has to be preprocessed in a slightly different manner. I choose to fill the missing Fare value with average. (It is integrated with the pipeline.)

Maybe there is a more right and modular way to organize the preprocess procedure, but now it works for this certain data. I'll not bother with it.

3. Prediction

```
val embarkedIndexer = new StringIndexer().setInputCol("Embarked").setOutputCol("EmbarkedIndex")
val embarkedEncoder = new OneHotEncoder().setInputCol("EmbarkedIndex").setOutputCol("EmbarkedVec")
val sexIndexer = new StringIndexer().setInputCol("Sex").setOutputCol("SexIndex")
val sexEncoder = new OneHotEncoder().setInputCol("SexIndex").setOutputCol("SexVec")
val assembler = new VectorAssembler()
    .setInputCols(Array("Pclass", "Age", "FamilySize", "Fare", "SexVec", "EmbarkedVec"))
    .setOutputCol("features")
val lr = new LogisticRegression().setLabelCol("Survived").setFeaturesCol("features")
val pipeline = new Pipeline().setStages(Array(
    preTransformer,
    sexIndexer, sexEncoder,
    embarkedIndexer, embarkedEncoder,
    assembler,
    lr))
```

I use logistic regression for training and predicting. The details are obvious from the code. I organize them separately and integrate them as a whole pipeline.



I produce my prediction also as a csv file.

Titanic - Machine Learning from Disaster





Overview Data Code Models Discussion Leaderboard Rules Team Submissions

Submissions

