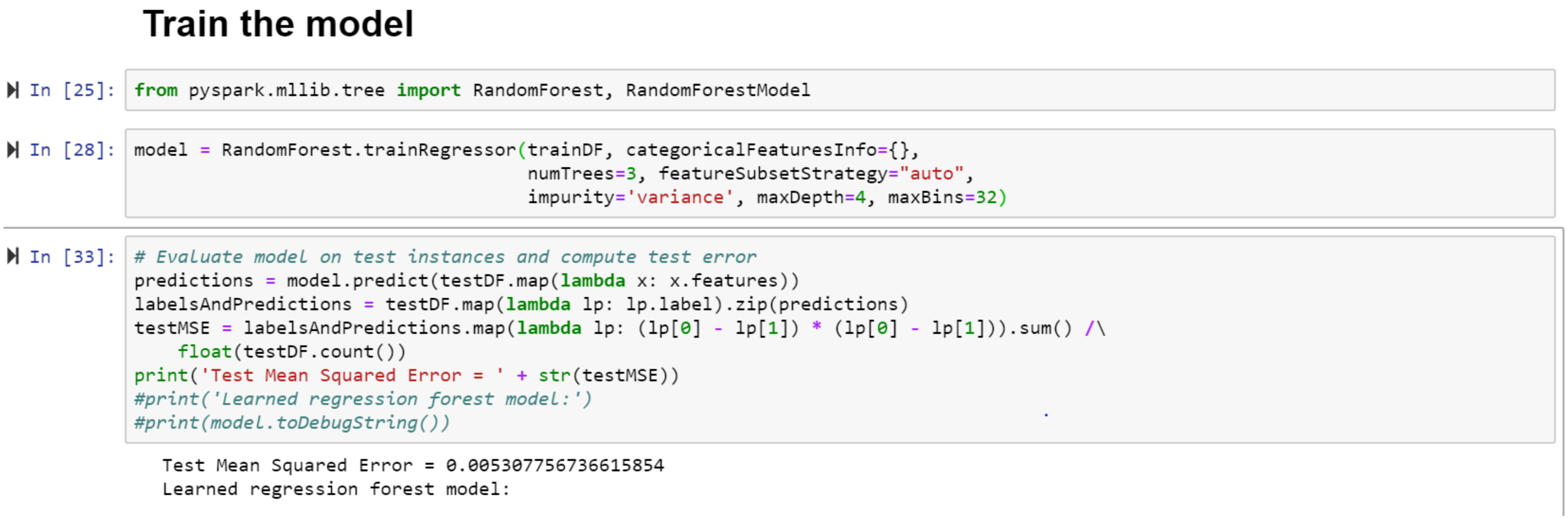
HW1

by8jj

motivation

Utilize random forest to train a model to make prediction on probability of students' admission

Code



Train the model

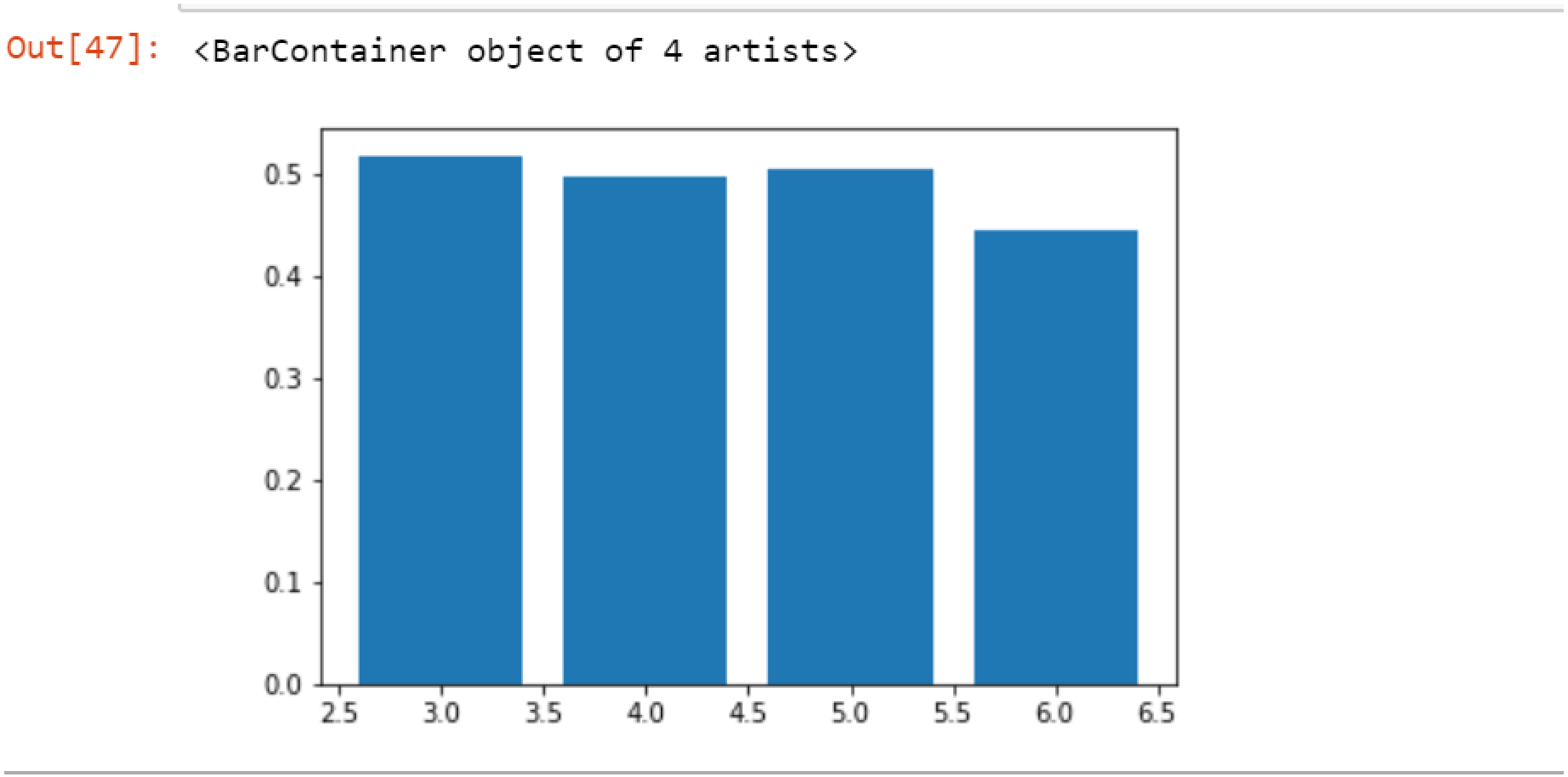
```
In [25]: from pyspark.mllib.tree import RandomForest, RandomForestModel
```

```
In [28]: model = RandomForest.trainRegressor(trainDF, categoricalFeaturesInfo={},
                                             numTrees=3, featureSubsetStrategy="auto",
                                             impurity='variance', maxDepth=4, maxBins=32)
```

```
In [33]: # Evaluate model on test instances and compute test error
         predictions = model.predict(testDF.map(lambda x: x.features))
         labelsAndPredictions = testDF.map(lambda lp: lp.label).zip(predictions)
         testMSE = labelsAndPredictions.map(lambda lp: (lp[0] - lp[1]) * (lp[0] - lp[1])).sum() /\
             float(testDF.count())
         print('Test Mean Squared Error = ' + str(testMSE))
         #print('Learned regression forest model:')
         #print(model.toDebugString())
```

```
Test Mean Squared Error = 0.005307756736615854
Learned regression forest model:
```

Visualization-parameter optimization



```
Out[47]: <BarContainer object of 4 artists>
```