CASE 2:

Determine Income based on characteristics

- Data Exploration
- Data Splitting
- Data Scaling
- Feature Reduction
- Model Selection
- Parameter Tuning
- Evaluation

CASE 2: Data Exploration

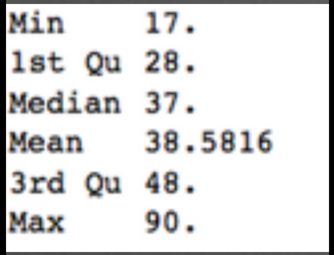
- Census Income Data Set
- https://archive.ics.uci.edu/ml/datasets/Census+Income

CASE 2: Data Exploration

- 15 Attributes:
- age: continuous.
- workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- fnlwgt: continuous.
- education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
- education-num: continuous.
- marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
- occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
- relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
- race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
- sex: Female, Male.
- capital-gain: continuous.
- capital-loss: continuous.
- hours-per-week: continuous.
- native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.
- class: >50K, <=50K

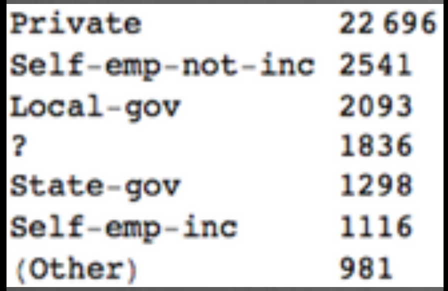
CASE 2: Data Exploration

Age



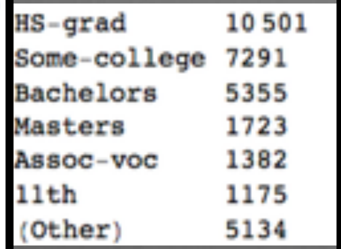
```
Min     17.
1st Qu  28.
Median  37.
Mean    38.5816
3rd Qu  48.
Max     90.
```

Workclass



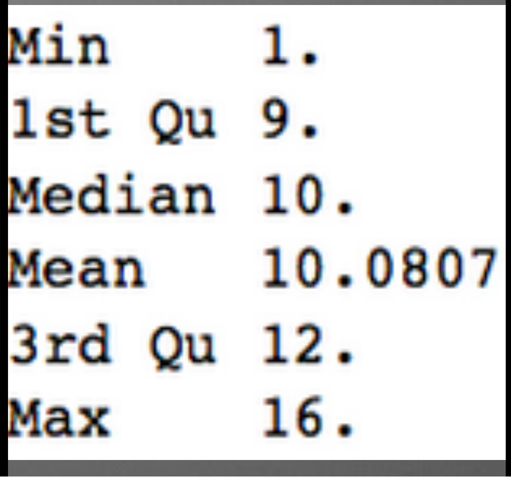
```
Private            22 696
Self-emp-not-inc   2541
Local-gov          2093
?                  1836
State-gov          1298
Self-emp-inc       1116
(Other)            981
```

Education



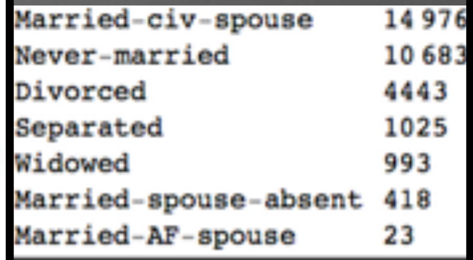
```
HS-grad       10 501
Some-college  7291
Bachelors     5355
Masters       1723
Assoc-voc     1382
11th          1175
(Other)       5134
```

Education-Num



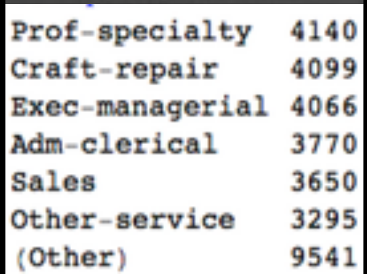
```
Min     1.
1st Qu  9.
Median  10.
Mean    10.0807
3rd Qu  12.
Max     16.
```

Marital-Status



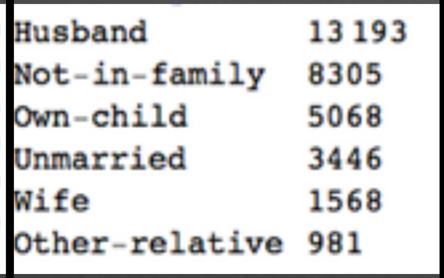
```
Married-civ-spouse     14 976
Never-married          10 683
Divorced               4443
Separated              1025
Widowed                993
Married-spouse-absent  418
Married-AF-spouse      23
```

Occupation



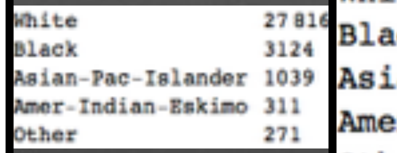
```
Prof-specialty   4140
Craft-repair     4099
Exec-managerial  4066
Adm-clerical     3770
Sales            3650
Other-service    3295
(Other)          9541
```

Relationship



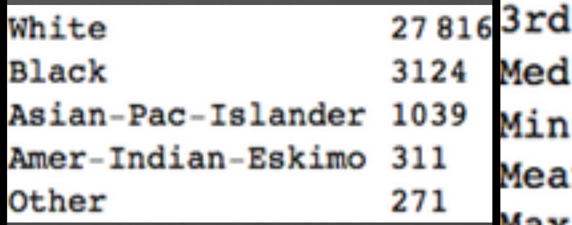
```
Husband         13 193
Not-in-family   8305
Own-child       5068
Unmarried       3446
Wife            1568
Other-relative  981
```

Race



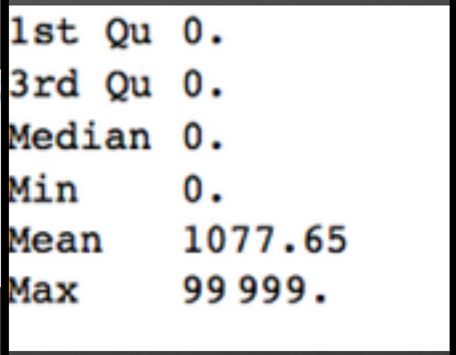
```
White               27 816
Black               3124
Asian-Pac-Islander  1039
Amer-Indian-Eskimo  311
Other               271
```

Sex



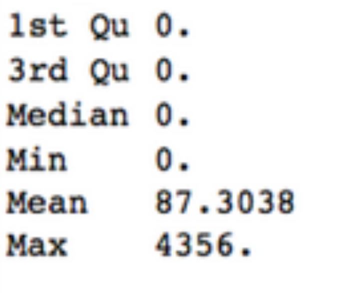
```
White               27 816
Black               3124
Asian-Pac-Islander  1039
Amer-Indian-Eskimo  311
Other               271
```

Capital-Gain



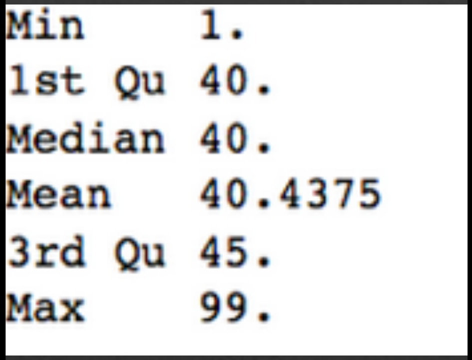
```
1st Qu  0.
3rd Qu  0.
Median  0.
Min     0.
Mean    1077.65
Max     99 999.
```

Capital-



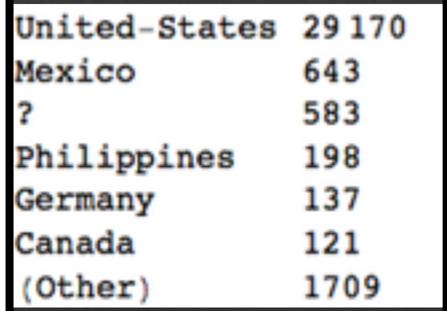
```
1st Qu  0.
3rd Qu  0.
Median  0.
Min     0.
Mean    87.3038
Max     4356.
```

Hours-Per-



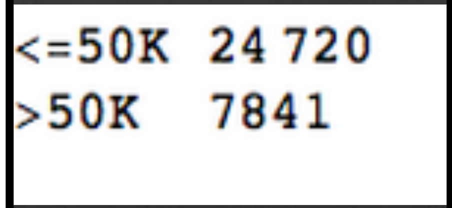
```
Min     1.
1st Qu  40.
Median  40.
Mean    40.4375
3rd Qu  45.
Max     99.
```

Native-



```
United-States  29 170
Mexico         643
?              583
Philippines    198
Germany        137
Canada         121
(Other)        1709
```

Income



```
<=50K  24 720
>50K   7841
```

CASE 2: Data Splitting

- Training Set: 32561
- Test Set: 16281

CASE 2: Data Scaling

- Convert categorical variables to numerical variables: workclass, education, marital-status, occupation, relationship, race, sex, native-country, class.

- def parsePoint(line):

CASE 2: Feature Reduction

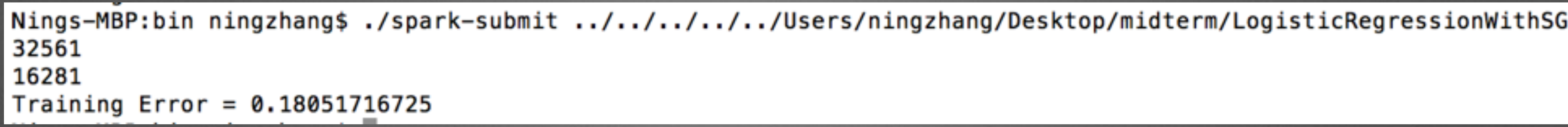
- fnlwgt
- capital-gain: continuous.
- capital-loss: continuous.

CASE 2: Model Selecting

- LogisticRegressionWithSGD
- Decision Tree Classification
- Bayes

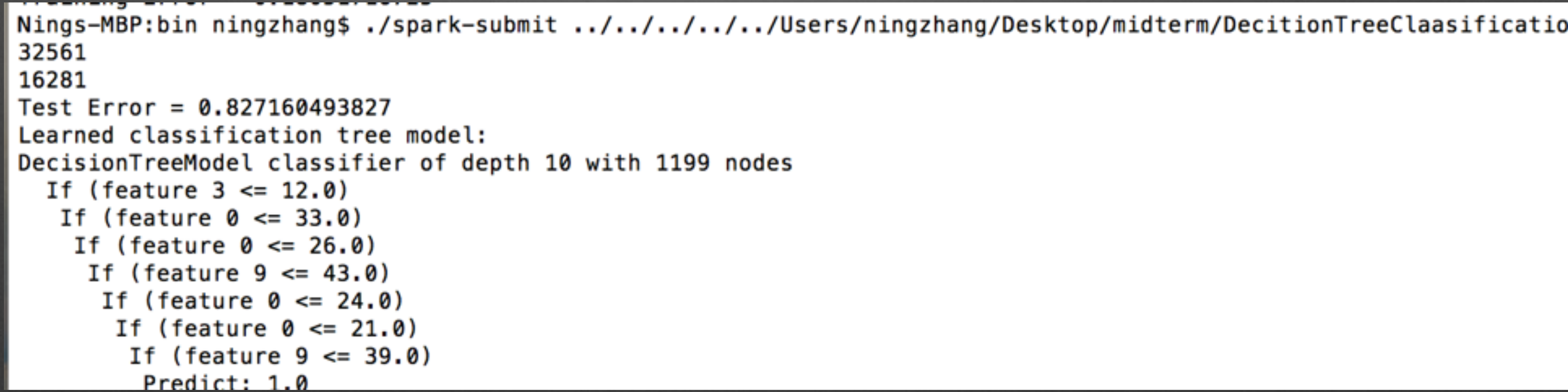
CASE 2: Evaluation

- LogisticRegressionWithSGD



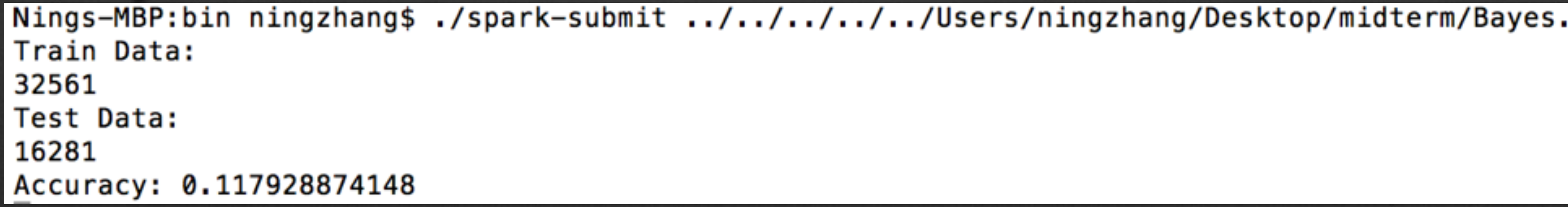
```
Nings-MBP:bin ningzhang$ ./spark-submit ../../../../../Users/ningzhang/Desktop/midterm/LogisticRegressionWithSG
32561
16281
Training Error = 0.18051716725
```

- Decision Tree Classification



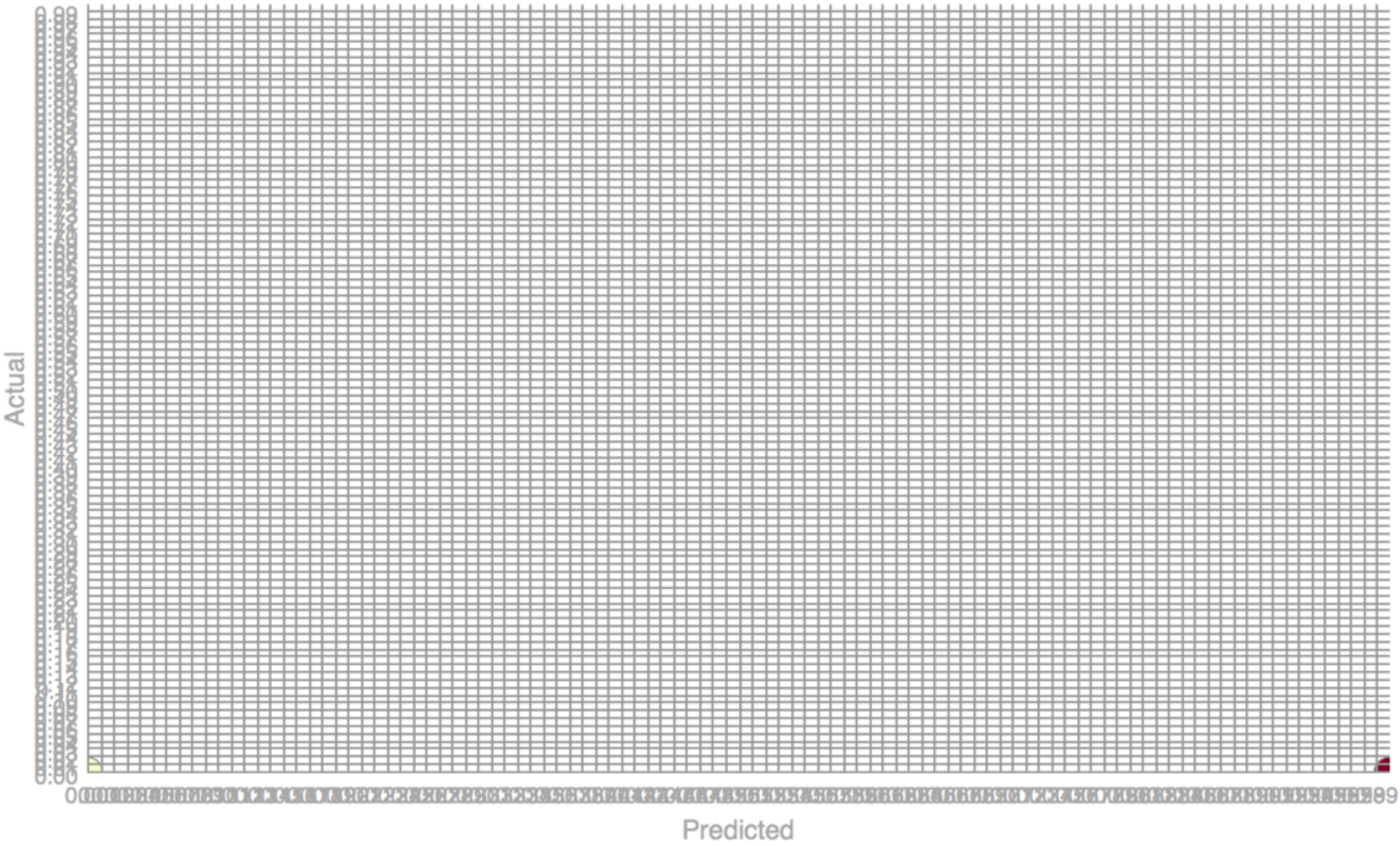
```
Nings-MBP:bin ningzhang$ ./spark-submit ../../../../../Users/ningzhang/Desktop/midterm/DecitionTreeClaasificatio
32561
16281
Test Error = 0.827160493827
Learned classification tree model:
DecisionTreeModel classifier of depth 10 with 1199 nodes
  If (feature 3 <= 12.0)
   If (feature 0 <= 33.0)
    If (feature 0 <= 26.0)
     If (feature 9 <= 43.0)
      If (feature 0 <= 24.0)
       If (feature 0 <= 21.0)
        If (feature 9 <= 39.0)
         Predict: 1.0
```

- Bayes



```
Nings-MBP:bin ningzhang$ ./spark-submit ../../../../../Users/ningzhang/Desktop/midterm/Bayes.
Train Data:
32561
Test Data:
16281
Accuracy: 0.117928874148
```

CASE 2: Evaluation


Actual
Predicted