Wen Sheng

Taylor Arnold

February 18, 2016

Problem Set 02 Data Analysis

Data Mining and Machine Learning

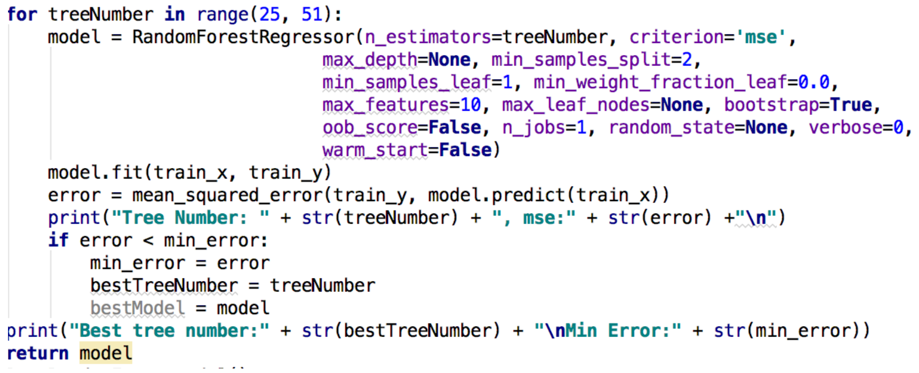
1. Describe how you decided the ultimate number of trees to use.

Solution：Using for-loop to try running tree number from 25 - 50 to find the tree number with minimum Mean Square Error. The result is using 50 will get the best result.

Best Tree Number: 50

Minimum Mean Square Error: 0.0319992743516

The running script looks like this:



```
for treeNumber in range(25, 51):
    model = RandomForestRegressor(n_estimators=treeNumber, criterion='mse',
                                  max_depth=None, min_samples_split=2,
                                  min_samples_leaf=1, min_weight_fraction_leaf=0.0,
                                  max_features=10, max_leaf_nodes=None, bootstrap=True,
                                  oob_score=False, n_jobs=1, random_state=None, verbose=0,
                                  warm_start=False)
    model.fit(train_x, train_y)
    error = mean_squared_error(train_y, model.predict(train_x))
    print("Tree Number: " + str(treeNumber) + ", mse:" + str(error) +"\n")
    if error < min_error:
        min_error = error
        bestTreeNumber = treeNumber
        bestModel = model
print("Best tree number:" + str(bestTreeNumber) + "\nMin Error:" + str(min_error))
return model
```

2. Compare the ‘out-of-bag’ mean squared error. Describe the patterns.

Running the Random Forest Model model, I got following result:

A. Mean Squared Error of each states:

- CT: 0.031999
- NY: 0.049216
- MT: 0.095416

- CA: 0.038667

B. Describe what patterns or surprising features arise.

Although New York is more distance closer to state Connecticut, the random forest model (using Connecticut data to fit) is more compatible to situation in state California, while applies worst to state Montana. Since the data is about the employment statistics, it shows that the Origin-Destination Employment patterns of Connecticut and California is more similar. Using Connecticut data to predict New York is relatively fit with mean squared error of 0.05, but using Connecticut data to predict Montana is inappropriate as the mean squared error rise to 0.1

---

3. Linear Regression Model and OSE. Describe the patterns.

Running the linear regression model, I got following result:

A. Mean Squared Error of each states:

- CT: 0.0289036617567
- NY: 0.0467652027483
- MT: 0.088323152621
- CA: 0.0350336440766

B. Describe what patterns or surprising features arise.

In overview surprising, the overall mean squared error using Linear Regression is lower than using Random Forest Regression. But the state data pattern result is consistent with both models. The statistics in Connecticut and California is more similar with each other, while New York is relatively fit and Montana having worst prediction result.

---

4. What are the 3 worst counties in terms of Mean Squared Error. Explain any patterns you see and suggest a possible solution.

A. Worst 3 counties in terms of Mean Squared Error:

Worst counties in New York:

- ny-041:0.139018573758 Hamilton County
- ny-095:0.114171686704 Schoharie County
- ny-003:0.111952013297 Allegany County

Worst Counties in California:

- ca-049:0.129004175942 Modoc County
- ca-093:0.112016077362 Siskiyou County
- ca-003:0.100567615465 Alpine County

B. Patterns and possible solution:

Searching the document, the 3 worst counties in each states share a common property: all of them are having less population compare to other counties inside the same state. Hamilton has the lowest population in New York and Alpine has the lowest population in California. It seems also compatible with common sense: area with less population have smaller job markets.

```
States ny
ny-041:0.139018573758
ny-095:0.114171686704
ny-003:0.111952013297
ny-017:0.110439867881
ny-077:0.107841026586
ny-049:0.107185236149
ny-025:0.107037057688
ny-123:0.106934101989
ny-009:0.0962813388081
ny-097:0.0949091686752
ny-035:0.0930315676307
ny-057:0.0922400801161
ny-089:0.0868361478409
ny-043:0.0863868758764
ny-045:0.0856134542993
ny-013:0.0852681756528
ny-031:0.0848371467958
ny-105:0.0836446672403
ny-101:0.0831418823114
ny-021:0.0812919723477
ny-099:0.0804814797746
ny-033:0.0797449282958
ny-051:0.0779977140327
ny-115:0.0775996145302
ny-023:0.0774190195752
ny-107:0.0735671920995
ny-121:0.0731556087088
ny-113:0.0730842209495
ny-053:0.0730309732566
ny-065:0.0716660530627
ny-039:0.0683490724516
ny-117:0.0671794998097
ny-075:0.0671746240119
ny-011:0.0650232051137
ny-019:0.0637134639344
ny-111:0.0636674822571
ny-073:0.0635035073798
ny-007:0.0596715074983
ny-015:0.0591336287584
ny-037:0.0583856899289
ny-069:0.0563321997254
ny-083:0.0518757333407
ny-079:0.0497599248543
ny-109:0.0488836004145
ny-071:0.0483285239021
ny-063:0.0446377755032
ny-027:0.0426759691978
ny-091:0.0409643342122
ny-001:0.0389028589714
ny-067:0.0376041010732
ny-087:0.0332628372989
ny-093:0.0326337945276
ny-119:0.031989400829
ny-029:0.0310647103067
ny-103:0.0301581065047
ny-055:0.0298520929966
ny-059:0.0238303068257
ny-085:0.02202558358
ny-081:0.018214892142
ny-047:0.0164041862428
ny-005:0.0156284025764
ny-061:0.0122777550392
States ca
ca-049:0.129004175942
ca-093:0.112016077362
ca-003:0.100567615465
ca-063:0.100346154892
ca-033:0.0927486997717
ca-035:0.0922271686807
ca-023:0.0884622516137
ca-045:0.088243721951
ca-105:0.0876794919957
ca-027:0.0835750794501
ca-011:0.0787258287162
ca-021:0.0763434202117
ca-089:0.0759565635985
ca-103:0.0748714879434
ca-015:0.0724630507126
ca-007:0.0706211670197
ca-109:0.069939308302
ca-043:0.0698450613079
ca-091:0.069398870751
ca-025:0.0688193144022
ca-057:0.0675186872246
ca-051:0.0674049685089
ca-009:0.0613579178696
ca-115:0.0602915801016
ca-005:0.0585343756599
ca-107:0.0575629574606
ca-017:0.0575312341685
ca-039:0.0568680393798
ca-047:0.0541328716682
ca-031:0.0536353524195
ca-029:0.0519407138066
ca-071:0.0506546630302
ca-101:0.0502307713547
ca-069:0.0501473222899
ca-079:0.0497685115054
ca-053:0.0477650846684
ca-061:0.0462766796319
ca-097:0.0446750902915
ca-065:0.0425121515215
ca-019:0.0414228536366
ca-077:0.0406258798828
ca-087:0.0369235711754
ca-095:0.0368672307971
ca-099:0.0364235715602
ca-083:0.0363564689644
ca-041:0.033731449155
ca-113:0.0321019844537
ca-055:0.0313784814941
ca-073:0.0312143995193
ca-013:0.0303765156861
ca-067:0.028611639036
ca-111:0.0282786262581
ca-037:0.0278198245196
```

ca-081:0.0234742378615
ca-001:0.0234459537235
ca-059:0.0226320114066
ca-085:0.0210927379376
ca-075:0.0190977244745

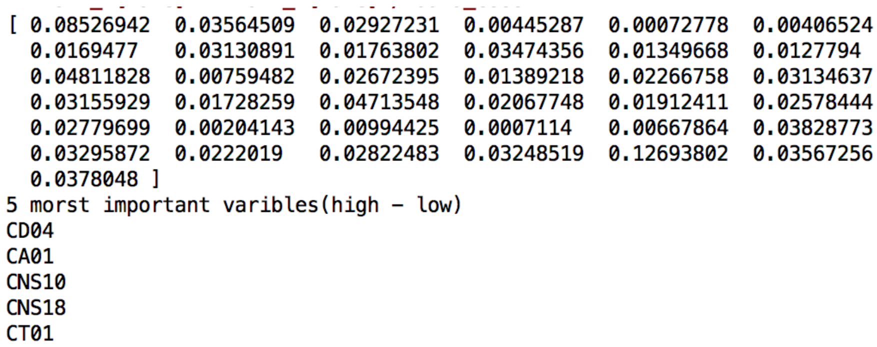
---

5. Calculate variables importance score and analyze the result

A. 5 most important variables in Random Forest Model:

Sorting the variable importance list, I found following 5 variables have highest importance in the Random Forest Regression model:

- **CD04**: Number of jobs for workers with Educational Attainment: Bachelor's degree or advanced degree
- **CA01**: Number of jobs for workers age 29 or younger
- **CNS10**: Number of jobs in NAICS sector 52 (Finance and Insurance)
- **CNS18**: Number of jobs in NAICS sector 72 (Accommodation and Food Services)
- **CT01**: Number of jobs for workers with Ethnicity: Not Hispanic or Latino[10]



```
[ 0.08526942  0.03564509  0.02927231  0.00445287  0.00072778  0.00406524
  0.0169477   0.03130891  0.01763802  0.03474356  0.01349668  0.0127794
  0.04811828  0.00759482  0.02672395  0.01389218  0.02266758  0.03134637
  0.03155929  0.01728259  0.04713548  0.02067748  0.01912411  0.02578444
  0.02779699  0.00204143  0.00994425  0.0007114   0.00667864  0.03828773
  0.03295872  0.0222019   0.02822483  0.03248519  0.12693802  0.03567256
  0.0378048 ]
5 morst important varibles(high - low)
CD04
CA01
CNS10
CNS18
CT01
```

B. Doest it make sense that these would help identify areas with a high promotion of high income earners?

Yes. The result of this 5 most importance variables is compatible with reality. People with higher education degrees and working in areas of Finance/ Insurance or Accommodation/Food Services tend to have higher education. Also interestedly, the figure shows that people's ethnicity and age(working history) also influences income a little bit: people not Hispanic and not Latino with number of jobs for workers under or equal to 29 tend to have higher income.