Lastfm Data Set

This data set contains playlists from the lastfm app which contain information about which artists users listen to.

The script then takes this data frame and creates a list of users which contains all the artists they listen to. The script then takes and makes a market basket model out of it predicting artists which users are more likely to listen to after their current preferences.

The final model of this set is sampled with the top 5 confident predictions:

inspect(sort(subset(musicrules, subset=lift > 5), by="confidence"))

lhs rhs support confidence lift count

[1] {led zeppelin,the doors} => {pink floyd} 0.01066667 0.5970149 5.689469 160

[2] {the pussycat dolls} => {rihanna} 0.01040000 0.5777778 13.415893 156

[3] {t.i.} => {kanye west} 0.01040000 0.5672727 8.854413 156

[4] {pink floyd,the doors} => {led zeppelin} 0.01066667 0.5387205 6.802027 160

[5] {sonata arctica} => {nightwish} 0.01346667 0.5101010 8.236292 202

[6] {judas priest} => {iron maiden} 0.01353333 0.5075000 8.562992 203

AdultUCI

This data set contains census style data on adults.

summary(AdultUCI)

age workclass fnlwgt education education-num

Min. :17.00 Private :33906 Min. : 12285 HS-grad :15784 Min. : 1.00

1st Qu.:28.00 Self-emp-not-inc: 3862 1st Qu.: 117551 Some-college:10878 1st Qu.: 9.00

Median :37.00 Local-gov : 3136 Median : 178145 Bachelors : 8025 Median :10.00

Mean :38.64 State-gov : 1981 Mean : 189664 Masters : 2657 Mean :10.08

3rd Qu.:48.00 Self-emp-inc : 1695 3rd Qu.: 237642 Assoc-voc : 2061 3rd Qu.:12.00

Max. :90.00 (Other) : 1463 Max. :1490400 11th : 1812 Max. :16.00

NA's : 2799 (Other) : 7625

marital-status occupation relationship race sex

Divorced : 6633 Prof-specialty : 6172 Husband :19716 Amer-Indian-Eskimo: 470 Female:16192

Married-AF-spouse : 37 Craft-repair : 6112 Not-in-family :12583 Asian-Pac-Islander: 1519 Male :32650

Married-civ-spouse :22379 Exec-managerial: 6086 Other-relative: 1506 Black : 4685

Married-spouse-absent: 628 Adm-clerical : 5611 Own-child : 7581 Other : 406

Never-married :16117 Sales : 5504 Unmarried : 5125 White :41762

Separated : 1530 (Other) :16548 Wife : 2331

Widowed : 1518 NA's : 2809

capital-gain capital-loss hours-per-week native-country income

Min. : 0 Min. : 0.0 Min. : 1.00 United-States:43832 small:24720

1st Qu.: 0 1st Qu.: 0.0 1st Qu.:40.00 Mexico : 951 large: 7841

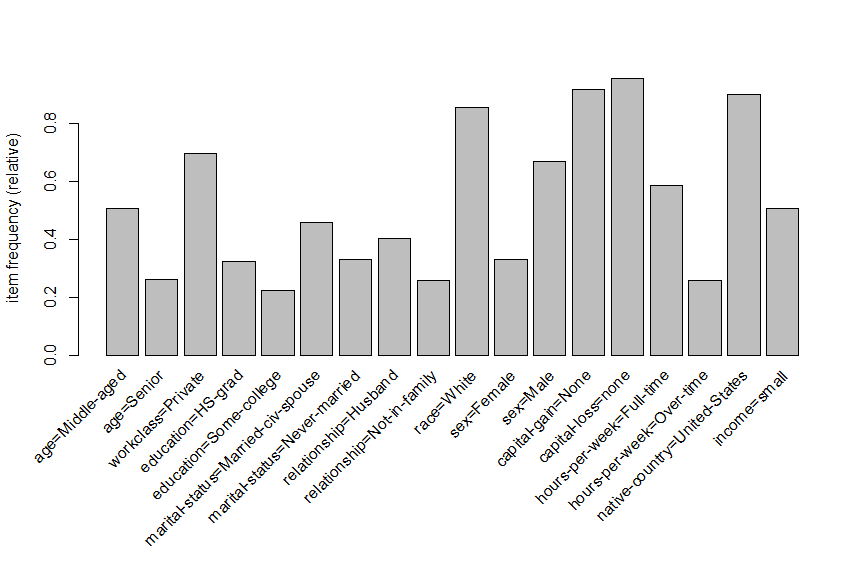
Median : 0 Median : 0.0 Median :40.00 Philippines : 295 NA's :16281

Mean : 1079 Mean : 87.5 Mean :40.42 Germany : 206

3rd Qu.: 0 3rd Qu.: 0.0 3rd Qu.:45.00 Puerto-Rico : 184

Max. :99999 Max. :4356.0 Max. :99.00 (Other) : 2517

NA's : 857

The following plot shows the frequency of categorical variables which are greater than 20% of the total variable:  


The top rules of this market basket where the income is small are defined at:

inspect(sort(rulesIncomeSmall, by = "confidence")[1:3])

lhs rhs support confidence lift count

[1] {workclass=Private,

marital-status=Never-married,

relationship=Own-child,

sex=Male,

hours-per-week=Part-time,

native-country=United-States} => {income=small} 0.01074895 0.7104195 1.403653 525

[2] {workclass=Private,

marital-status=Never-married,

relationship=Own-child,

sex=Male,

hours-per-week=Part-time} => {income=small} 0.01144507 0.7102922 1.403402 559

[3] {workclass=Private,

marital-status=Never-married,

relationship=Own-child,

sex=Male,

capital-gain=None,

hours-per-week=Part-time,

native-country=United-States} => {income=small} 0.01046231 0.7097222 1.402276 511

The market basket for large income is:

inspect(sort(rulesIncomeLarge, by = "confidence")[1:3])

lhs rhs support confidence lift count

[1] {marital-status=Married-civ-spouse,

capital-gain=High,

native-country=United-States} => {income=large} 0.01562180 0.6849192 4.266398 763

[2] {marital-status=Married-civ-spouse,

capital-gain=High,

capital-loss=none,

native-country=United-States} => {income=large} 0.01562180 0.6849192 4.266398 763

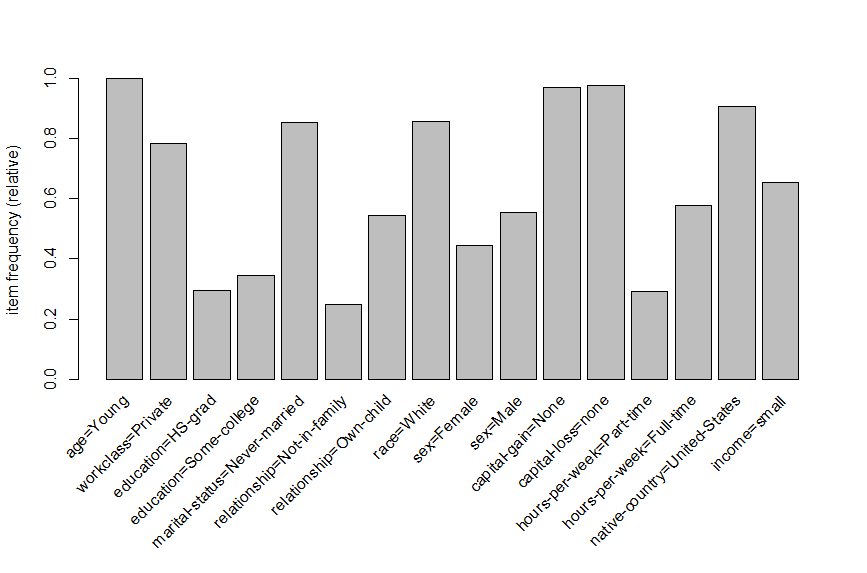
[3] {relationship=Husband,

race=White,

capital-gain=High,

native-country=United-States} => {income=large} 0.01302158 0.6846071 4.264454 636

One interesting thing to change in this analysis it to shift it towards those who are only classified as “Young”. This leads to the following plot:



This produces the top predictors for small and large income:

inspect(sort(rulesIncomeSmall, by = "confidence")[1])

lhs rhs support confidence lift count

[1] {workclass=Private,

marital-status=Never-married,

relationship=Own-child,

sex=Male,

hours-per-week=Part-time,

native-country=United-States} => {income=small} 0.01074895 0.7104195 1.403653 525

inspect(sort(rulesIncomeLarge, by = "confidence")[1])

lhs rhs support confidence lift count

[1] {marital-status=Married-civ-spouse,

capital-gain=High,

native-country=United-States} => {income=large} 0.0156218 0.6849192 4.266398 763