2017

LINKÖPING UNIVERSITY

Carles Sans Fuentes

[MAIN FEATURES AFFECTING EARNINGS ON POPULATION]

The project tries to identify the main features on the population to earn more than 50K dollars per year. Those features are identified using a very well-known rule based algorithm called a priori algorithm, based on confidence, support and lift

Index

1. MOTIVATION

Inequality in salaries due to inequality of opportunities and discrimination has always been one of the major issues that has concerned the population among different countries. There is always though (unless there is direct evidence from the discrimination) some background to reason upon the election of the best candidate for a position: experience, education and several other aspects (e.g. the physical appearance in an interview) define a lot of times why is one candidate chosen over another. Nevertheless, are there intrinsic variables that can make someone being discriminated indirectly or directly given some prejudices on the population? It is alleged and acknowledged by different countries, for instance, that there is clear evidence of discrimination for the female gender on their jobs; not only about the difference on salaries for a similar or equal position than a man, but rather also about the chances to get to a higher position given its capabilities and hence earning more money either in the public or private sector.

The usage of computers and its incremental power and speed during the last decades have been crucial in order to evaluate and assess objectively how large these issues are as well as several other ones (e.g. gender discrimination by origin or color). Thus, the evaluation of large amounts of data (called today Big Data) using algorithms that evaluate and assess information efficiently in almost no time has been able to show to the society the dimension of certain problems with clear proof to discuss and argument the subject that matters. Nevertheless, when the analysis of data is done, it is always crucial that when you run algorithms and establish models, all variables available should be included in the first instance for the purpose of seeking the maximum evidence and the true cause of the issue. Thus, it is important to be aware that information can be presented and shown from a lot of perspectives, and that some of them may lack of a good evaluation, the good data or a wrong model. Thus, it is always important to be able to show in a clear matter your findings making them reproducible. For this reason, I will be using a basic algorithm (the apriori) in this Data Mining project at the university of Linköping to establish rules that evaluate the main variables from the US census in 1994 that affect whether a person won more than $50K per year.

Background

The median salary (the income from the person who was at the 50% of the whole population by salary) in the US in 1994 was of $31,5221.

[1] Reference to this information can be found in http://www.davemanuel.com/median-household-income.php

Scope

The scope of this project is not about learning how the algorithm works in depth neither to find new revolutionary conclusions but understanding how to use algorithms in real data sets, learning how to use an algorithm (at my choice) in a software and derive conclusions from it.

2. DATASET EXPLANATION

The dataset used for this data mining a project is a well-known dataset called adults[2], implemented in several softwares and evaluated with different algorithms. The dataset, extracted by Barry Becker, contains information about the 1994 Census Boreau database in the US. More specifically, it contains 48,842 observations from people with 15 attributes to take into consideration with its following classes:

1. ***Age***: continuous variable related to the observations.
2. ***Workclass***: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
3. ***Fnlwgt***: The number of people the census takers believe that observation represents.
4. ***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.
5. ***Education***-num: continuous.
6. ***Marital-status***: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
7. ***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.
8. ***Relationship***: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
9. ***Race***: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
10. ***Sex***: Female, Male.
11. ***Capital-gain***: continuous.
12. ***Capital-loss***: continuous.
13. ***Hours-per-week***: continuous.
14. ***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.
15. ***Winning50K***: whether the observations win more than $50000 or not: >50K, <=50K.

The idea for this data set is to find relations and importance of variables to the variable of winning more than $50000 or not. Having that said, the dataset shows that the Probability for

[2] More information about the data set as well as the possibility of downloading it can be found in https://archive.ics.uci.edu/ml/datasets/adult

the label '>50K' is 24.78% and Probability for the label '<=50K': 75.22% (once unknowns are cleaned up on the data base provided[3])

3. THE A PRIORI ALGORITHM

The apriori algorithm is a frequent items set algorithm used to find association rules over transactional databases. It is really used in areas such as the food industry or the banking industry in order to understand costumer behavior and elaborate bundles and reccomed products to clients. The working procedure of the algorithm is starting with individual item sets and then enlarging them under some constraints, being the most typical ones the follwoing ones:

- *Minimum support*: Probability that a transaction contains product X and Y given all transactions.
- *Confidence*: Probablity of Y for all transactions containing X.
- *Lift*: measure that identifies the probability of a rule given the total possibility of that rule in the model. It is given by the formula lift (X) =Prob(X|Y) /Prob(X)
- *Target feature:* It is also typical to use some target features or products contained in the dataset as a constrain in order to assess rules based on this target feature.

The pseudocode for the algorithm can be found in the following picture[4]:

```
// Faster Algorithm
1) forall large k-itemsets l_k, k ≥ 2 do begin
2)     H_1 = { consequents of rules derived from l_k with one item in the consequent };
3)     call ap-genrules(l_k, H_1);
4) end

procedure ap-genrules(l_k: large k-itemset, H_m: set of m-item consequents)
    if (k > m + 1) then begin
        H_{m+1} = apriori-gen(H_m);
        forall h_{m+1} ∈ H_{m+1} do begin
            conf = support(l_k)/support(l_k − h_{m+1});
            if (conf ≥ minconf) then
                output the rule (l_k − h_{m+1}) ⟹ h_{m+1} with confidence = conf and support = support(l_k);
            else
                delete h_{m+1} from H_{m+1};
        end
        call ap-genrules(l_k, H_{m+1});
    end
```

Figure 1 PseudoCode from the apriori algorithm in Han - Data Mining Concepts and Techniques 3rd Edition – 2012, page 253.

[3] Evaluation and information just explained is found in http://www.cs.toronto.edu/~delve/data/adult/adultDetail.html

[4] Since it is an algorithm explained in class, only the main idea of it is given

One of the limitations of the algorithm is that it requires data to be discrete. Also, it can generate a huge amount of inefficient rules, so good constraints must be designed in order to be optimal.

4. DATASET CLEANING

Since the apriori algorithm requires all data to be classified as categorical some modifications and variable elimination has been performed:

Removals

The attribute *education-num* has been removed given that it provides the same information in then the categorical variable of *education*, which is already in bins.

Data modification

To find a better set of large rules, variables which were continuous have been simplified in bins in the following way.:

Age	***Fnlwgt***	***Capital gain***	***Capital loss***	***Hoursperweek***
<30s	<100M	No Gain	No Loss	<20 hours
30-40s	100-200M	Gain	Loss	20-34 hours
40-50s	200-300M			35-44 hours
50-65s	300-400M			45-60 hours
>65	>400M			>60 hours

Age has been simplified in decades mainly, being the last tram before retirement larger because I understand that larger changes on salary come before the 50s, so that the salary you have when you are 50 is already similar to your salary until you get retired.

Fnlwgt has been classified by how many people it concerns to, doing bins of 100.000 people.

Capital gain and *loss* has been classified as winnings, losses or no losses neither gains.

The *number of hours worked per week* has been classified as less than 20, considering this normally as almost sporadic work, 20-34 accounting for those people who work part time completely. From 35-44 it would stand for the amount of hours that a person does in an average job. From 45-60 counting for people who Works in more than one job or consultant, and then the variable more than 60.

5. TEST RESULTS

The apriori algorithm used for our different evaluations belongs to the arules library from the R[5] software. The methodology to find out rules has been as following:

- Separate the creation of rules from Winning more than 50K and less. In each case, it has been evaluated:
  - The first 10 rules by confidence given lift higher than 1
  - The first 10 rules by support given lift higher than 1
  - The number of variables that appear more times in the rules to see which are those having more effect given some support, lift and confidence.
  - Specific rules for Race = Black and Sex = Female

Winning more than $50K

The rules created have as a result >50K on the right side (since it is what it is evaluated). The support and confidence that has been used to create the first basic set of rules have been 0.02 and 0.5. This can be seen as a little but not even 25% of data containing the variable >50, so 0.02 stands for a 10% rule to it, so that it is not that low. The Figure below shows the 1,526 rules created by support and confidence:

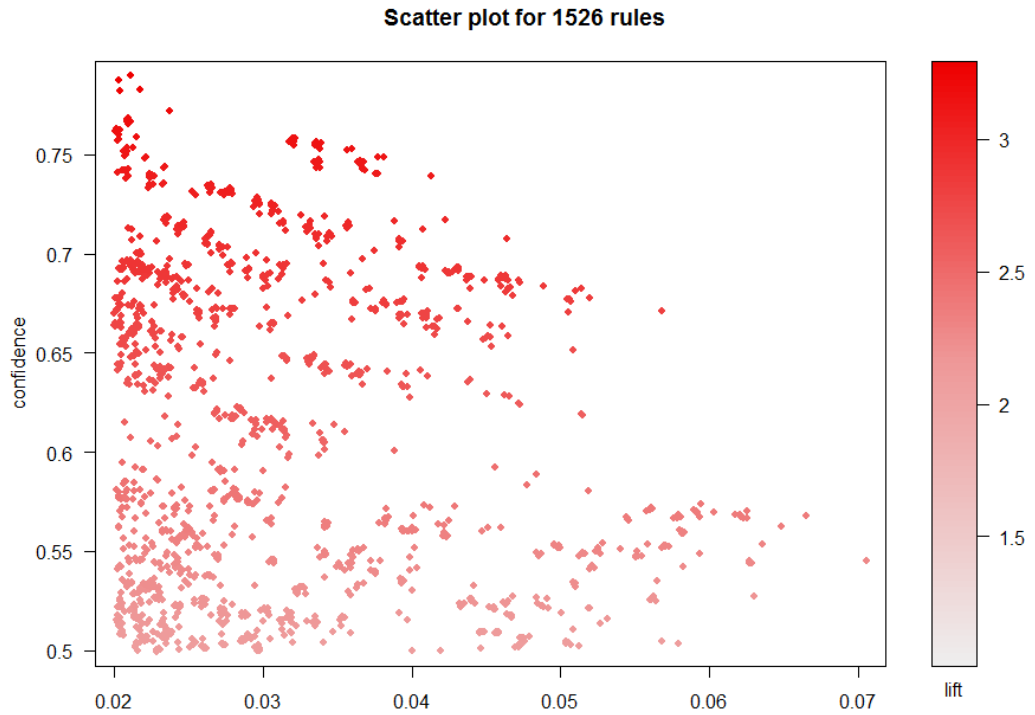

Figure 2 Scatterplot of 1526 rules created from the apriori algorithm with support and confidence equal to 0.02 and 0.5 respectively. Representation of the lift can also be seen in the right-hand side

[5] The library arules can be found in https://cran.r-project.org/web/packages/arules/arules.pdf

Rules vary from support 0.02-0.07 and confidence 0.5-0.8. Nevertheless, rules with 50% of confidence might be low to extract good general rules, so more search is done by subsetting from the first rules only those with lift bigger than 1.05 and confidence equal to 0.7:

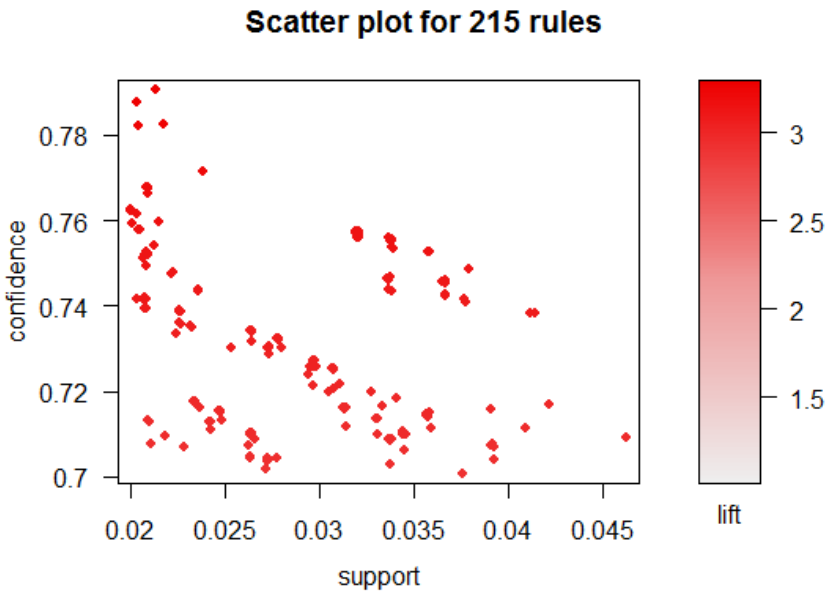

Figure 3 Scatterplot of 215 rules subsetted from the previous apriori algorithm general support and confidence constrained with lift and confidence equal to 1.05 and 0.7 respectively. Representation of the lift can also be seen in the right-hand side

From this subset rules, here we can see the top 10 rules by confidence and by support:

```
> inspect(head(rules_subsetWin50, n = 10, by= "support"))
      lhs                                                rhs                    support confidence     lift
[1]   {marital-status= Married-civ-spouse,
       occupation= Prof-specialty}                    => {Wining50K= >50K} 0.04628378  0.7088429 2.943493
[2]   {marital-status= Married-civ-spouse,
       occupation= Prof-specialty,
       race= White}                                   => {Wining50K= >50K} 0.04207617  0.7169021 2.976959
[3]   {marital-status= Married-civ-spouse,
       capitalgain= Gain}                             => {Wining50K= >50K} 0.04124693  0.7383178 3.065888
[4]   {marital-status= Married-civ-spouse,
       capitalgain= Gain,
       capitalloss=No Loss}                           => {Wining50K= >50K} 0.04124693  0.7383178 3.065888
[5]   {marital-status= Married-civ-spouse,
       occupation= Prof-specialty,
       nativeCountry= United-States}                  => {Wining50K= >50K} 0.04087838  0.7113843 2.954046
[6]   {occupation= Prof-specialty,
       relationship= Husband}                         => {Wining50K= >50K} 0.03918919  0.7073171 2.937156
[7]   {occupation= Prof-specialty,
       relationship= Husband,
       sex= Male}                                     => {Wining50K= >50K} 0.03918919  0.7073171 2.937156
[8]   {marital-status= Married-civ-spouse,
       occupation= Prof-specialty,
       sex= Male}                                     => {Wining50K= >50K} 0.03918919  0.7038058 2.922576
[9]   {marital-status= Married-civ-spouse,
       occupation= Prof-specialty,
       relationship= Husband}                         => {Wining50K= >50K} 0.03912776  0.7069922 2.935808
[10]  {marital-status= Married-civ-spouse,
       occupation= Prof-specialty,
       relationship= Husband,
       sex= Male}                                     => {Wining50K= >50K} 0.03912776  0.7069922 2.935808
```

Figure 4 Top 10 rules by support

```
> inspect(head(rules_subsetWin50, n = 10, by= "confidence"))
     lhs                                                rhs                 support confidence     lift
[1]  {education= Masters,
      marital-status= Married-civ-spouse,
      nativeCountry= United-States}                  => {Wining50K= >50K} 0.02128378  0.7901938 3.281305
[2]  {education= Masters,
      marital-status= Married-civ-spouse,
      race= White,
      nativeCountry= United-States}                  => {Wining50K= >50K} 0.02020885  0.7880240 3.272294
[3]  {education= Masters,
      marital-status= Married-civ-spouse,
      race= White}                                   => {Wining50K= >50K} 0.02168305  0.7827051 3.250208
[4]  {education= Bachelors,
      marital-status= Married-civ-spouse,
      occupation= Exec-managerial}                   => {Wining50K= >50K} 0.02039312  0.7820966 3.247681
[5]  {education= Masters,
      marital-status= Married-civ-spouse}            => {Wining50K= >50K} 0.02377150  0.7716849 3.204446
[6]  {education= Masters,
      relationship= Husband}                         => {Wining50K= >50K} 0.02088452  0.7674944 3.187045
[7]  {education= Masters,
      marital-status= Married-civ-spouse,
      relationship= Husband}                         => {Wining50K= >50K} 0.02088452  0.7674944 3.187045
[8]  {education= Masters,
      relationship= Husband,
      sex= Male}                                     => {Wining50K= >50K} 0.02088452  0.7674944 3.187045
[9]  {education= Masters,
      marital-status= Married-civ-spouse,
      relationship= Husband,
      sex= Male}                                     => {Wining50K= >50K} 0.02088452  0.7674944 3.187045
[10] {education= Masters,
      marital-status= Married-civ-spouse,
      sex= Male}                                     => {Wining50K= >50K} 0.02094595  0.7662921 3.182052
```

Figure 5 Top 10 rules by confidence

Also, it can be seen below the number of times the variables are repeated in the previous general rule set and subset:

```
> featWin50_subset
              capitalgain=No Gain          hoursperweek=35-44 hours
                                1                                 1
                capitalloss= Loss                 fnlwgt=100-200M
                                2                                 4
       hoursperweek=45-60 hours                 education= Masters
                                6                                10
     occupation= Prof-specialty               education= Bachelors
                               29                                37
    occupation= Exec-managerial                  capitalloss=No Loss
                               50                                76
              capitalgain= Gain          nativeCountry= United-States
                               88                                97
                    race= White                          sex= Male
                               97                                98
             workclass= Private               relationship= Husband
                              119                               134
marital-status= Married-civ-spouse
                              147
```

Figure 7 Count of the variables repeated in the subset

```
> featWin50_global
                  Age=30-39s                  capitalloss= Loss
                           8                                11
       education= Masters                  occupation= Sales
                          25                                39
   hoursperweek=>60 hours                        Age=50-64s
                          48                                67
hoursperweek=35-44 hours          occupation= Prof-specialty
                          68                               131
         capitalgain= Gain         occupation= Exec-managerial
                         135                               259
       fnlwgt=100-200M                             Age=40-49s
                         270                               290
     education= Bachelors               hoursperweek=45-60 hours
                         298                               300
   capitalgain=No Gain                        workclass= Private
                         389                               439
   capitalloss=No Loss          nativeCountry= United-States
                         567                               724
           race= White                                sex= Male
                         727                               755
 relationship= Husband marital-status= Married-civ-spouse
                         846                               891
```

Figure 6 Count of the variables repeated in the set

Winning less than $50K

The rules have as a result the variable <50K on the right just (since it is what it is evaluated). The support and confidence that has been used to create the first basic set of rules have been 0.1 and 0.7. The Figure below shows 933 rules created by support and confidence:

Figure 8 Scatterplot of 933 rules created from the apriori algorithm with support and confidence equal to 0.1 and 0.7 respectively. Representation of the lift can also be seen in the right-hand side

Rules vary from support 0.1-0.7 and confidence 0.7-1. More search is done by subsetting from the first rules only those with lift bigger than 1.05 and confidence equal to 0.7:

Figure 9 Scatterplot of 703 rules subsetted from the previous apriori algorithm general support and confidence constrained with lift and confidence equal to 1.05 and 0.7 respectively. Representation of the lift can also be seen in the right-hand side

From this subset rules, here we can see the top 10 rules by confidence and by support:

```
> inspect(head(rules_subsetNoWin50, n = 10, by= "support"))
     lhs                                    rhs                  support confidence     lift
[1]  {capitalgain=No Gain,
      capitalloss=No Loss}                => {Wining50K= <=50K} 0.7045147  0.8097070 1.066550
[2]  {capitalgain=No Gain,
      capitalloss=No Loss,
      nativeCountry= United-States}       => {Wining50K= <=50K} 0.6263514  0.8054502 1.060943
[3]  {workclass= Private,
      capitalgain=No Gain}                => {Wining50K= <=50K} 0.5232494  0.8126789 1.070465
[4]  {workclass= Private,
      capitalgain=No Gain,
      capitalloss=No Loss}                => {Wining50K= <=50K} 0.5076474  0.8271945 1.089585
[5]  {workclass= Private,
      capitalgain=No Gain,
      nativeCountry= United-States}       => {Wining50K= <=50K} 0.4599509  0.8069835 1.062963
[6]  {workclass= Private,
      capitalgain=No Gain,
      capitalloss=No Loss,
      nativeCountry= United-States}       => {Wining50K= <=50K} 0.4457617  0.8218573 1.082555
[7]  {workclass= Private,
      race= White,
      capitalgain=No Gain}                => {Wining50K= <=50K} 0.4380528  0.7984214 1.051685
[8]  {workclass= Private,
      race= White,
      capitalgain=No Gain,
      capitalloss=No Loss}                => {Wining50K= <=50K} 0.4244779  0.8135743 1.071644
[9]  {capitalloss=No Loss,
      hoursperweek=35-44 hours}           => {Wining50K= <=50K} 0.4233108  0.7999420 1.053688
[10] {capitalgain=No Gain,
      hoursperweek=35-44 hours}           => {Wining50K= <=50K} 0.4195025  0.8183452 1.077929
```

Figure 10 Top 10 rules by support

```
> inspect(head(rules_subsetNoWin50, n = 10, by= "confidence"))
     lhs                                    rhs                  support confidence     lift
[1]  {Age=<30s,marital-status= Never-married,
      relationship= Own-child, capitalgain=No Gain,
      capitalloss=No Loss,
      nativeCountry= United-States}       => {Wining50K= <=50K} 0.1052826  0.9979622 1.314521
[2]  {Age=<30s, relationship= Own-child,
      capitalgain=No Gain, capitalloss=No Loss,
      nativeCountry= United-States}       => {Wining50K= <=50K} 0.1103194  0.9977778 1.314278
[3]  {Age=<30s, marital-status= Never-married,
      relationship= Own-child, capitalgain=No Gain,
      capitalloss=No Loss}                => {Wining50K= <=50K} 0.1131757  0.9972936 1.313641
[4]  {Age=<30s, marital-status= Never-married,
      relationship= Own-child, capitalloss=No Loss,
      nativeCountry= United-States}       => {Wining50K= <=50K} 0.1076781  0.9971559 1.313459
[5]  {Age=<30s, marital-status= Never-married,
      relationship= Own-child,
      capitalgain=No Gain,
      nativeCountry= United-States}       => {Wining50K= <=50K} 0.1075246  0.9971518 1.313454
[6]  {Age=<30s, relationship= Own-child,
      capitalgain=No Gain,
      nativeCountry= United-States}       => {Wining50K= <=50K} 0.1127150  0.9970117 1.313269
[7]  {Age=<30s, relationship= Own-child,
      race= White, capitalgain=No Gain,
      capitalloss=No Loss}                => {Wining50K= <=50K} 0.1013514  0.9969789 1.313226
[8]  {Age=<30s, relationship= Own-child,
      capitalgain=No Gain,
      capitalloss=No Loss}                => {Wining50K= <=50K} 0.1190418  0.9969136 1.313140
[9]  {Age=<30s, relationship= Own-child,
      capitalloss=No Loss,
      nativeCountry= United-States}       => {Wining50K= <=50K} 0.1129300  0.9967471 1.312921
[10] {Age=<30s, marital-status= Never-married,
      relationship= Own-child,
      capitalgain=No Gain}                => {Wining50K= <=50K} 0.1155405  0.9965563 1.312669
```

Figure 11 Top 10 rules by confidence

Also, it can be seen below the number of times the variables are repeated in the previous general rules set and subset:

```
> featNowin50_subset
      education= Some-college}              fnlwgt=100-200M}              fnlwgt=200-300M}
                             1                             1                             1
      hoursperweek=20-34 hours          workclass= Private}          education= HS-grad}
                             1                             1                             2
     hoursperweek=20-34 hours}  relationship= Not-in-family}                   Age=30-39s
                             3                             3                             5
marital-status= Never-married}      relationship= Own-child}                fnlwgt=<100M
                             5                             5                             7
      marital-status= Divorced                  sex= Female}                    sex= Male}
                             9                            11                            14
              fnlwgt=200-300M                  race= White}       education= Some-college
                            15                            20                            26
      relationship= Own-child              fnlwgt=100-200M    relationship= Not-in-family
                            60                            68                            69
         capitalgain=No Gain}      hoursperweek=35-44 hours                      sex= Male
                            73                            74                            88
                  sex= Female           education= HS-grad      hoursperweek=35-44 hours}
                            91                            96                           103
         capitalloss=No Loss}                     Age=<30s         capitalloss=No Loss
                           144                           166                           203
marital-status= Never-married                  race= White                workclass= Private
                           203                           249                           254
          capitalgain=No Gain  nativeCountry= United-States}
                           291                           306
```

Figure 12 Count of the variables repeated in the subset

```
> featNowin50_global
                  Age=50-64s       education= Some-college}              fnlwgt=200-300M}
                           1                             1                             1
      hoursperweek=20-34 hours          education= HS-grad}              fnlwgt=100-200M}
                           1                             2                             2
          workclass= Private}  hoursperweek=20-34 hours}  relationship= Not-in-family}
                           2                             3                             3
marital-status= Never-married}      relationship= Own-child}      marital-status= Divorced
                           5                             5                             9
                 sex= Female}                fnlwgt=<100M                    sex= Male}
                          11                            15                            21
       education= Some-college                  race= White}              fnlwgt=200-300M
                          28                            28                            37
                  Age=30-39s    relationship= Own-child      relationship= Not-in-family
                          42                            60                            69
                  sex= Female         capitalgain=No Gain}           education= HS-grad
                          91                            94                           103
      hoursperweek=35-44 hours              fnlwgt=100-200M      hoursperweek=35-44 hours}
                         105                           129                           132
                     Age=<30s                  sex= Male          capitalloss=No Loss}
                         166                           177                           183
 marital-status= Never-married         capitalloss=No Loss          workclass= Private
                         203                           267                           339
                   race= White         capitalgain=No Gain  nativeCountry= United-States}
                         355                           365                           419
```

Figure 13 Count of the variables repeated in the set

```
> inspect(head(rules_subsetWoman, n = 10, by= "confidence"))
     lhs                                                  rhs                support confidence     lift
[1]  {workclass= Private, marital-status= Never-married,
      sex= Female, capitalgain=No Gain,
      capitalloss=No Loss}                          => {Wining50K= <=50K} 0.1025799  0.9820641 1.293580
[2]  {marital-status= Never-married,
      sex= Female,
      capitalgain=No Gain,
      capitalloss=No Loss}                          => {Wining50K= <=50K} 0.1335074  0.9792746 1.289906
[3]  {workclass= Private, marital-status= Never-married,
      sex= Female,
      capitalgain=No Gain}                          => {Wining50K= <=50K} 0.1053440  0.9791607 1.289756
[4]  {marital-status= Never-married,
      sex= Female,
      capitalgain=No Gain,
      capitalloss=No Loss,
      nativeCountry= United-States}                 => {Wining50K= <=50K} 0.1210074  0.9791252 1.289709
[5]  {marital-status= Never-married,
      race= White,sex= Female,
      capitalgain=No Gain,
      capitalloss=No Loss}                          => {Wining50K= <=50K} 0.1059275  0.9776077 1.287710
[6]  {marital-status= Never-married,
      sex= Female,
      capitalgain=No Gain}                          => {Wining50K= <=50K} 0.1371929  0.9759668 1.285549
[7]  {Age=<30s, sex= Female,
      capitalgain=No Gain,
      capitalloss=No Loss,
      nativeCountry= United-States}                 => {Wining50K= <=50K} 0.1016585  0.9755379 1.284984
[8]  {marital-status= Never-married,
      sex= Female, capitalgain=No Gain,
      nativeCountry= United-States}                 => {Wining50K= <=50K} 0.1243857  0.9754335 1.284846
[9]  {Age=<30s,sex= Female, capitalgain=No Gain,
      capitalloss=No Loss}                          => {Wining50K= <=50K} 0.1123771  0.9752132 1.284556
[10] {marital-status= Never-married,
      race= White, sex= Female,
```

Figure 14 Top 10 rules for woman winning <50K by confidence

```
> inspect(head(rules_subsetwoman, n = 10, by= "support"))
     lhs                                                  rhs                support confidence     lift
[1]  {sex= Female}                                    => {Wining50K= <=50K} 0.2945946  0.8905394 1.173023
[2]  {sex= Female,
      capitalloss=No Loss}                          => {Wining50K= <=50K} 0.2863636  0.8963661 1.180698
[3]  {sex= Female,
      capitalgain=No Gain}                          => {Wining50K= <=50K} 0.2839066  0.9109184 1.199867
[4]  {sex= Female,
      capitalgain=No Gain,
      capitalloss=No Loss}                          => {Wining50K= <=50K} 0.2756757  0.9178853 1.209043
[5]  {sex= Female,
      nativeCountry= United-States}                 => {Wining50K= <=50K} 0.2644349  0.8892791 1.171363
[6]  {sex= Female,
      capitalloss=No Loss,
      nativeCountry= United-States}                 => {Wining50K= <=50K} 0.2570025  0.8956438 1.179747
[7]  {sex= Female,
      capitalgain=No Gain,
      nativeCountry= United-States}                 => {Wining50K= <=50K} 0.2547604  0.9102381 1.198971
[8]  {sex= Female,
      capitalgain=No Gain,
      capitalloss=No Loss,
      nativeCountry= United-States}                 => {Wining50K= <=50K} 0.2473280  0.9178254 1.208965
[9]  {race= White,
      sex= Female}                                  => {Wining50K= <=50K} 0.2338452  0.8810461 1.160519
[10] {race= White,
      sex= Female,
      capitalloss=No Loss}                          => {Wining50K= <=50K} 0.2270885  0.8878482 1.169478
```

Figure 15 Top 10 rules for woman winning <50K by support

```
> inspect(head(rules_subsetBlack, n = 10, by= "confidence"))
     lhs                                rhs                  support confidence     lift
[1]  {marital-status= Never-married,
      race= Black,
      capitalgain=No Gain,
      capitalloss=No Loss}          => {Wining50K= <=50K} 0.03799140  0.9872306 1.300386
[2]  {marital-status= Never-married,
      race= Black,
      capitalgain=No Gain,
      capitalloss=No Loss,
      nativeCountry= United-States} => {Wining50K= <=50K} 0.03418305  0.9858282 1.298538
[3]  {marital-status= Never-married,
      race= Black,
      capitalgain=No Gain}          => {Wining50K= <=50K} 0.03906634  0.9852827 1.297820
[4]  {marital-status= Never-married,
      race= Black,
      capitalgain=No Gain,
      nativeCountry= United-States} => {Wining50K= <=50K} 0.03513514  0.9845095 1.296801
[5]  {marital-status= Never-married,
      race= Black,
      capitalloss=No Loss}          => {Wining50K= <=50K} 0.03921990  0.9762997 1.285987
[6]  {marital-status= Never-married,
      race= Black,
      capitalloss=No Loss,
      nativeCountry= United-States} => {Wining50K= <=50K} 0.03535012  0.9754237 1.284833
[7]  {marital-status= Never-married,
      race= Black}                  => {Wining50K= <=50K} 0.04029484  0.9747400 1.283933
[8]  {marital-status= Never-married,
      race= Black,
      nativeCountry= United-States} => {Wining50K= <=50K} 0.03630221  0.9744435 1.283542
[9]  {workclass= Private,
      race= Black,
      sex= Female,
      capitalgain=No Gain}          => {Wining50K= <=50K} 0.03049754  0.9650146 1.271122
[10] {race= Black,
```

Figure 16 Top 10 rules for race = Black winning <50K by confidence

```
> inspect(head(rules_subsetBlack, n = 10, by= "support"))
     lhs                                rhs                  support confidence     lift
[1]  {race= Black}                   => {Wining50K= <=50K} 0.08406020  0.8761204 1.154030
[2]  {race= Black,
      capitalloss=No Loss}           => {Wining50K= <=50K} 0.08194103  0.8831513 1.163292
[3]  {race= Black,
      capitalgain=No Gain}           => {Wining50K= <=50K} 0.08132678  0.9000680 1.185574
[4]  {race= Black,
      capitalgain=No Gain,
      capitalloss=No Loss}           => {Wining50K= <=50K} 0.07920762  0.9084185 1.196574
[5]  {race= Black,
      nativeCountry= United-States}  => {Wining50K= <=50K} 0.07619779  0.8760593 1.153950
[6]  {race= Black,
      capitalloss=No Loss,
      nativeCountry= United-States}  => {Wining50K= <=50K} 0.07432432  0.8825675 1.162523
[7]  {race= Black,
      capitalgain=No Gain,
      nativeCountry= United-States}  => {Wining50K= <=50K} 0.07358722  0.8997371 1.185139
[8]  {race= Black,
      capitalgain=No Gain,
      capitalloss=No Loss,
      nativeCountry= United-States}  => {Wining50K= <=50K} 0.07171376  0.9075010 1.195365
[9]  {workclass= Private,
      race= Black}                   => {Wining50K= <=50K} 0.05992015  0.8965993 1.181005
[10] {workclass= Private,
      race= Black,
      capitalloss=No Loss}           => {Wining50K= <=50K} 0.05863022  0.9038826 1.190599
```

Figure 17 Top 10 rules for race = Black winning <50K by support

6. CONCLUSIONS

Evaluation

After having cleaned the top 10 rules for winning more and less than $50K, the Figure below summarizes the best rules among the top 10 ones for the >$50K:

```
> inspect(head(rules_subsetWin50, n = 10, by= "support"))
     lhs                                rhs                  support confidence     lift
[2]  {marital-status= Married-civ-spouse,
      occupation= Prof-specialty,
      race= White}                  => {Wining50K= >50K} 0.04207617  0.7169021 2.976959
[4]  {marital-status= Married-civ-spouse,
      capitalgain= Gain,
      capitalloss=No Loss}          => {Wining50K= >50K} 0.04124693  0.7383178 3.065888
[5]  {marital-status= Married-civ-spouse,
      occupation= Prof-specialty,
      nativeCountry= United-States} => {Wining50K= >50K} 0.04087838  0.7113843 2.954046
[10] {marital-status= Married-civ-spouse,
      occupation= Prof-specialty,
      relationship= Husband,
      sex= Male}                    => {Wining50K= >50K} 0.03912776  0.7069922 2.935808
```

```
> inspect(head(rules_subsetWin50, n = 10, by= "confidence"))
     lhs                                rhs                  support confidence     lift
[2]  {education= Masters,
      marital-status= Married-civ-spouse,
      race= White,
      nativeCountry= United-States} => {Wining50K= >50K} 0.02020885  0.7880240 3.272294
[4]  {education= Bachelors,
      marital-status= Married-civ-spouse,
      occupation= Exec-managerial}  => {Wining50K= >50K} 0.02039312  0.7820966 3.247681
[9]  {education= Masters,
      marital-status= Married-civ-spouse,
      relationship= Husband,
      sex= Male}                    => {Wining50K= >50K} 0.02088452  0.7674944 3.187045
```

Figure 18 Most important rules in the top 10 for winning >50K

If you have a profesional specialty and you are husband, male or white, you are probable to win more than $50K. Moreover, if your education is master and you are white, husband or male, or if you have a bachelor's diploma and work as an ex-managerial being married, your probabilities of also winning more than $50K are quite high. Also, the fact about having capital gains affect your probability of winning more than 50K.

The attributes that are more repeated for the rules on the set and subset are: being husband, married-civ-spouse, male, white, native-country equal to US, having capital gains, working on the private sector, being ex-managerial, or having at least bachelor and prof specialty and age 40-49.

```
> inspect(head(rules_subsetNoWin50, n = 10, by= "support"))
     lhs                                  rhs                 support confidence     lift
[2]  {capitalgain=No Gain,
      capitalloss=No Loss,
      nativeCountry= United-States} => {Wining50K= <=50K} 0.6263514  0.8054502 1.060943
[6]  {workclass= Private,
      capitalgain=No Gain,
      capitalloss=No Loss,
      nativeCountry= United-States} => {Wining50K= <=50K} 0.4457617  0.8218573 1.082555
[8]  {workclass= Private,
      race= White,
      capitalgain=No Gain,
      capitalloss=No Loss}          => {Wining50K= <=50K} 0.4244779  0.8135743 1.071644
[9]  {capitalloss=No Loss,
      hoursperweek=35-44 hours}     => {Wining50K= <=50K} 0.4233108  0.7999420 1.053688
[10] {capitalgain=No Gain,
      hoursperweek=35-44 hours}     => {Wining50K= <=50K} 0.4195025  0.8183452 1.077929
```

```
> inspect(head(rules_subsetNoWin50, n = 10, by= "confidence"))
     lhs                                  rhs                 support confidence     lift
[1]  {Age=<30s,
      marital-status= Never-married,
      relationship= Own-child,
      capitalgain=No Gain,
      capitalloss=No Loss,
      nativeCountry= United-States}  => {Wining50K= <=50K} 0.1052826  0.9979622 1.314521
[7]  {Age=<30s,
      relationship= Own-child,
      race= White,
      capitalgain=No Gain,
      capitalloss=No Loss}           => {Wining50K= <=50K} 0.1013514  0.9969789 1.313226
```

Figure 19 Most important rules in the top 10 for winning <50K

On the other hand, , if you don't have capital gain or loss, if your working class is private and your working time per week is between 35-44 hours the probabilities of not winning $50K are higher. Also, if you have never been married and you are under 30 years old with a kid without capital gains or losses, then your probabilities of not winning $50K are almost 100%.

The attributes that are more repeated for the rules on the set and subset are: being never-married or not in family, age<30, white, native country equal to US, no capital gains, sex = female, sex = male, education = HS-grad, working 35-44 hoursperweek.

Given that woman are 0.33 of the cases, having 0.29 is about 88% of the woman

```
> inspect(head(rules_subsetWoman, n = 10, by= "support"))
     lhs                                  rhs                 support confidence     lift
[1]  {sex= Female}                     => {Wining50K= <=50K} 0.2945946  0.8905394 1.173023
[4]  {sex= Female,
      capitalgain=No Gain,
      capitalloss=No Loss}           => {Wining50K= <=50K} 0.2756757  0.9178853 1.209043
[5]  {sex= Female,
      nativeCountry= United-States}  => {Wining50K= <=50K} 0.2644349  0.8892791 1.171363
[8]  {sex= Female,
      capitalgain=No Gain,
      capitalloss=No Loss,
      nativeCountry= United-States}  => {Wining50K= <=50K} 0.2473280  0.9178254 1.208965
[9]  {race= White,
      sex= Female}                   => {Wining50K= <=50K} 0.2338452  0.8810461 1.160519
```

```
> inspect(head(rules_subsetWoman, n = 10, by= "confidence"))
     lhs                                  rhs                 support confidence     lift
[3]  {workclass= Private,
      marital-status= Never-married,
      sex= Female,
      capitalgain=No Gain}           => {Wining50K= <=50K} 0.1053440  0.9791607 1.289756
[5]  {marital-status= Never-married,
      race= White,
      sex= Female,
      capitalgain=No Gain,
      capitalloss=No Loss}           => {Wining50K= <=50K} 0.1059275  0.9776077 1.287710
[6]  {marital-status= Never-married,
      sex= Female,
      capitalgain=No Gain}           => {Wining50K= <=50K} 0.1371929  0.9759668 1.285549
[9]  {Age=<30s,
      sex= Female,
      capitalgain=No Gain,
      capitalloss=No Loss}           => {Wining50K= <=50K} 0.1123771  0.9752132 1.284556
```

Figure 20 Most important rules from the top 10 for woman winning <50K

winning less than 50K. The variable white seems to affect a lot as well as race, making the support lower just to 0.23 so 70%. Still, the biggest impact is the variable woman. It has been found that largest rules for woman earning more than 50K are being white and married being wife.

```
> inspect(head(rules_subsetBlack, n = 10, by= "confidence"))
    lhs                                rhs                  support confidence     lift
[2] {marital-status= Never-married,
     race= Black,
     capitalgain=No Gain,
     capitalloss=No Loss,
     nativeCountry= United-States}  => {Wining50K= <=50K} 0.03418305  0.9858282 1.298538
[3] {marital-status= Never-married,
     race= Black,
     capitalgain=No Gain}           => {Wining50K= <=50K} 0.03906634  0.9852827 1.297820
[7] {marital-status= Never-married,
     race= Black}                   => {Wining50K= <=50K} 0.04029484  0.9747400 1.283933
[8] {marital-status= Never-married,
     race= Black,
     nativeCountry= United-States}  => {Wining50K= <=50K} 0.03630221  0.9744435 1.283542
[9] {workclass= Private,
     race= Black,
     sex= Female,
     capitalgain=No Gain}           => {Wining50K= <=50K} 0.03049754  0.9650146 1.271122
```

```
> inspect(head(rules_subsetBlack, n = 10, by= "support"))
    lhs                                rhs                  support confidence     lift
[1] {race= Black}                   => {Wining50K= <=50K} 0.08406020  0.8761204 1.154030
[4] {race= Black,
     capitalgain=No Gain,
     capitalloss=No Loss}           => {Wining50K= <=50K} 0.07920762  0.9084185 1.196574
[5] {race= Black,
     nativeCountry= United-States}  => {Wining50K= <=50K} 0.07619779  0.8760593 1.153950
[8] {race= Black,
     capitalgain=No Gain,
     capitalloss=No Loss,
     nativeCountry= United-States}  => {Wining50K= <=50K} 0.07171376  0.9075010 1.195365
[9] {workclass= Private,
     race= Black}                   => {Wining50K= <=50K} 0.05992015  0.8965993 1.181005
```

Figure 21 Most important rules from the top 10 for race = Black winning <50K

Given that race = Black in the data set accounts for just 0.095 of the cases, having 0.084 earning less than $50K with 0.87 of confidence is quite impacting. One can say that this variable totally affects the probability of winnning more than $50K. No Loss, No Gain or native Country do not affect much (neither support or confidence change much). Also, if you are black and female, even if you work in the private sector, the probability of you winning less than $50K is 0.965. Again, if you are black and you have never been married, with support 0.363 and confidence 0.974 you will not be winning more than $50K. Also, being married= hubsand and race = black are the variables that affect more (with higher confidence and support) when winning more than $50K.

Discussion and comparison

The main results got from the analysis are that variables such as being white, Husband or married with spouse with high education (at least bachelor) affect highly your chances of winning more than $50K, whereas the variables of being a woman, race = Black, not being married or being lower than 30 years affect almost void the chances of winning more than $50K.These results were expected given the high level of discrimination known in the US 1990s as well as the possibility for the major part of the population to study, biasing them towards having to work from little and leaving education. Also, as I have previously said, it seems to exist some concern for woman not winning more than $50K given that it is not easy to find rules with interesting variables winning more than $50K rather than being married, probably to someone who does win much more than 50K who has contacts for her to get a decent job. These

results should make think to the US society about potential problems in the population and try to look for good explanations on them.

Criticism

The apriori algorithm does generate too many rules a lot of times, some of them not being relevant neither useful or causal even if they look like so. Nevertheless, variables like being black and/or being woman as by itself seem to have a huge impact on population. It would be necessary though trying to find more concrete rules that could help classify and explain the issues on population better as well as getting the help of an expert in the topic to do more accurate searches by knowing what are we looking for.

Forward steps

My work on analyzing this data set has been little, using just one algorithm to extract conclusions to see potential rules.It would be appropiate and necessary using other algorithms like C4.5, Naive-Bayes or Knn (like it has been done in orhter articles) in order to confirm better my evaluations and insights about the data set and so the US society in 1994.

7. CODE

```
#install.packages("stringr")
#install.packages("plyr")
#install.packages("stats")
#install.packages("arules")
#install.packages("arulesViz")
#install.packages("stringr")
library(plyr)
library(stats)
library(arulesViz)
library(arules)
library(stringr)

###Reading data set

data<- read.csv("C:/Users/Carles/Desktop/Data mining/dataset.1.txt",sep =",", stringsAsFactors = FALSE)
class(data)
table(data[,1])

##############################################################################################
#####################################CLEANINGDATA#############################################
# ##########################################################
##############################################################################################
#########################
```

```
colnames(data)<-c("Age", "workclass", "fnlwgt", "education","education-num", "marital-status", "occupation", "relationship", "race", "sex","capitalgain", "capitalloss", "hoursperweek", "nativeCountry", "Wining50K")

##Droping education num because it stands for the same as education but continuous
drops<- c("education-num")

data<- data[ , !(names(data) %in% drops)]

table(data[["Wining50K"]])
### <=50K   >50K
### 24719   7841


##wining +50K == to 1, less than 50k equals to 0

levels(data[["Wining50K"]])<-c(0,1)

#Data does not contain NA
apply(data, 2, anyNA)


###################################APRIORI
#####################################################
#install.packages("arules")



#colnames(data)<-c("Age", "workclass", "fnlwgt", "education","education-num", "marital-status", "occupation", "relationship", "race", "sex","capitalgain", "capitalloss", "hoursperweek", "nativeCountry", "Wining50K")

min(data[,1]); max(data[,1]);
hist(data[,1])
thirties<-which(data[,1]<30)
fourties<-which((data[,1]>=30)&(data[,1]<40))
fifties<-which((data[,1]>=40)&(data[,1]<50))
sixties<-which((data[,1]>=50)&(data[,1]<65))
pensioners<- which(data[,1]>=65)

data[thirties,1]<-"<30s"
data[fourties,1]<-"30-39s"
data[fifties,1]<-"40-49s"
data[sixties,1]<-"50-64s"
data[pensioners,1]<-">65"

table(data[,1])
    ### <30s    >65 30-39s 40-49s 50-64s
    ### 9711   1336   8612   7175   5726

colnames(data)[c(3,5,11,13)]##[1] "fnlwgt" "education-num" "capitalgain"   "hoursperweek"

#### Final weight: The \# of people the census takers believe that observation represents. We will be ignoring this variable
hist(data[,3])
min(data[,3]); max(data[,3]);
```

```
belowhundred<- which(data[,3]<100000)
hundred<- which((data[,3]>=100000)&(data[,3]<200000))
twohundred<- which((data[,3]>=200000)&(data[,3]<300000))
threehundred<-which((data[,3]>=300000)&(data[,3]<400000))
fourhundred<- which(data[,3]>=400000)

data[belowhundred,3]<-"<100M"
data[hundred,3]<-"100-200M"
data[twohundred,3]<-"200-300M"
data[threehundred,3]<-"300-400M"
data[fourhundred,3]<-">400M"

table(data[,3])
class(data[,3])


#### "capital gain"

hist(data[,10])
min(data[,10]); max(data[,10]);
NoGain<-which(data[,10] == 0)
Gain<-which(data[,10]>1)
data[NoGain,10]<-"No Gain"
data[Gain,10]<-" Gain"
table(data[,10])

#### "capital loss"

hist(data[,11])
min(data[,11]); max(data[,11]);
NoLoss<- which(data[,11]==0)
Loss<- which(data[,11]>0)
data[NoLoss,11]<-"No Loss"
data[Loss,11]<-" Loss"

table(data[,11])
#### "hoursperweek"
min(data[,12]); max(data[,12]);
hist(data[,12])
lesstwenty<- which(data[,12]<20)
twenty<- which((data[,12]>=20)&(data[,12]<35))
thirtyfive<-which((data[,12]>=35)&(data[,12]<45))
fortyfive<-which((data[,12]>=45)&(data[,12]<60))
PlusSixty<-which(data[,12]>=60)

data[lesstwenty,12]<-"<20 hours"
data[twenty,12]<-"20-34 hours"
data[thirtyfive,12]<-"35-44 hours"
data[fortyfive,12]<-"45-60 hours"
data[PlusSixty,12]<-">60 hours"

table(data[,12])

Myframe<- as.data.frame(data)


DataFactor<-Myframe
```

```
for(i in 1:ncol(data)){
  #levels(data[,i])<-names(table(data[,i]))
  DataFactor[,i]<- as.factor(data[,i])
}
##########################################################################################
#########################
####################################A                                                PRIORI
ALGORITHM#########################################################################
##########################################################################################
#########################

####MAIN FUNCTIONS TO CHECK FROM THE RULES STATED, which ones are the repeated variables
most variables
MainWordsList<-function(data= resultswin50_global , start=2, stop=22, numbercommas=6){
  require(stringr)
  ####just taking into account those which has more than 5 variables, counted as comma separated
  data<-data[which(str_count(data[,1], ',')>=numbercommas),]
  data[,1]<- as.character(data[,1])
  for(i in 1:length(data[,1])){
    ##Getting rid of last variable and first variable for counting most important features
    data[i,1]<- substr(data[i,1],start,nchar(data[i,1])-stop)
  }

  featWin50<-table(unlist(strsplit(as.character(data[,1]), split = c(","))))
  return(featWin50)
}

##### RULES FOR WIN +50K###################
Mytransmatrix <- as(DataFactor, "transactions")

Win50<- apriori(Mytransmatrix, parameter = list(supp = 0.02, conf = 0.5, target = "rules"),
       appearance = list(rhs = c("Wining50K= >50K"),
                         default="lhs"))
plot(Win50, method = NULL, measure = "support", shading = "lift",
     interactive = FALSE, data = NULL, control = NULL)

rules_subsetWin50 <- subset(Win50, subset=(lift>1.05& confidence>0.7))
plot(rules_subsetWin50, method = NULL, measure = "support", shading = "lift",
     interactive = FALSE, data = NULL, control = NULL)
##larger by support
inspect(head(rules_subsetWin50, n = 10, by= "support"))

##larger by confidence
inspect(head(rules_subsetWin50, n = 10, by= "confidence"))

resultswin50_global<- as(Win50, "data.frame");
resultswin50_subset<- as(rules_subsetWin50, "data.frame");

featWin50_subset<- sort(MainWordsList(resultswin50_subset,2,22,1))
featWin50_global <- sort(MainWordsList(resultswin50_global,2,22,1))
##### RULES FOR WIN -50K###################

NoWin50<- apriori(Mytransmatrix, parameter = list(supp = 0.10, conf = 0.7, target = "rules"),
              appearance = list(rhs = c("Wining50K= <=50K"),
                                default="lhs"))

plot(NoWin50, method = NULL, measure = "support", shading = "lift",
```

```
    interactive = FALSE, data = NULL, control = NULL)

rules_subsetNoWin50 <- subset(NoWin50, subset=(lift>1.05& confidence>0.7))
plot(rules_subsetNoWin50, method = NULL, measure = "support", shading = "lift",
    interactive = FALSE, data = NULL, control = NULL)
##larger by support
inspect(head(rules_subsetNoWin50, n = 10, by= "support"))

##larger by confidence
inspect(head(rules_subsetNoWin50, n = 10, by= "confidence"))

resultsNowin50_global<- as(NoWin50, "data.frame");
resultsNowin50_subset<- as(rules_subsetNoWin50, "data.frame");
dim(resultsNowin50_subset)
featNoWin50_subset<- sort(MainWordsList(resultsNowin50_subset,2,22,1))
featNoWin50_global<- sort(MainWordsList(resultsNowin50_global,2,22,1))


table(data[,8])
##### RULES FOR WIN -50K###################
###Rules for female
NoWin50Female<- apriori(Mytransmatrix, parameter = list(supp = 0.10, conf = 0.7, target = "rules"),
            appearance = list(rhs = c("Wining50K= <=50K"),
                        items = c("sex= Female"),
                        default="lhs"))
rules_subsetWoman <- subset(NoWin50Female, (lhs %in% "sex= Female"))
inspect(head(rules_subsetWoman, n = 10, by= "confidence"))
inspect(head(rules_subsetWoman, n = 10, by= "support"))

table(data[,9])[1]/sum(table(data[,9]))
###   Female
###  0.3308047


###Rules for female to win 50K
Win50Female<- apriori(Mytransmatrix, parameter = list(supp = 0.02, conf = 0.2, target = "rules"),
                appearance = list(rhs = c("Wining50K= >50K"),
                            items = c("sex= Female"),
                            default="lhs"))
rules_subsetWomanwin50 <- subset(Win50Female, (lhs %in% "sex= Female"))
inspect(head(rules_subsetWomanwin50, n = 10, by= "confidence"))
inspect(head(rules_subsetWomanwin50, n = 10, by= "support"))

table(data[,9])[1]/sum(table(data[,9]))
###   Female
###  0.3308047


###Rules for race black
NoWin50Black <-apriori(Mytransmatrix, parameter = list(supp = 0.03, conf = 0.5, target = "rules"),
              appearance = list(rhs = c("Wining50K= <=50K"),
                          items = c("race= Black"),
                          default="lhs"))
rules_subsetBlack <- subset(NoWin50Black, (lhs %in% "race= Black"))
inspect(head(rules_subsetBlack, n = 10, by= "confidence"))
inspect(head(rules_subsetBlack, n = 10, by= "support"))
```

```
table(data[,8])[3]/sum(table(data[,8]))
###    Black
####0.09594595


###Rules for race black to win 50K
Win50Black <-apriori(Mytransmatrix, parameter = list(supp = 0.005, conf = 0.2, target = "rules"),
                appearance = list(rhs = c("Wining50K= >50K"),
                                  items = c("race= Black"),
                                  default="lhs"))
rules_subsetBlackwin <- subset(Win50Black, (lhs %in% "race= Black"))
inspect(head(rules_subsetBlackwin, n = 10, by= "confidence"))
inspect(head(rules_subsetBlackwin, n = 10, by= "support"))

table(data[,8])[3]/sum(table(data[,8]))
###    Black
####0.09594595
```