Data Story

Anthony Stachowitz

July 13, 2018

Data Story

Introduction

There is a lot of conversation about the future of jobs in this country. People are worried about innovation and if it’s going to take away jobs. Will self-driving trucks take away good blue-collar jobs? Are algorithms going to take over the financial sector? Should everyone major in robotics, so they can fix all the robots that are surely going to be doing all the factory jobs?

These are conversations that we are going to have to have as technology changes the employment landscape, But, are there things we can do to prepare for the future that has nothing to do with advances in tech? Are the generalities we can make about out type of employment and industry without worrying about specific jobs? Are there basic life decisions that influence how our future will unfolds? Is there advice we can give our children that will, if followed, give them an advantage?

I have analyzed a data set from extracted from the census bureau by Data Mining and Visualization Silicon Valley (via the UCI machine learning repository) to help answer some of these questions. The data includes two sets. The training set 32,561 instances of data and the testing set includes 16,281 instances of date. I removed some variables to condense it to five categorical variables, four of the variables were used to predict the fifth variable.

The variable I am predicting is whether you are going to make more $50,000 a year. The categorical variables that I am using are type of employment, industry, marital status, and education level completed. Although you would assume going to school will help your chances, what is the best education level? Does your marital status have any effect on your income level, or is it more important then what industry you pick? Do you have a better chance to earn a living by owning your own business or should you try to get a government job… and what government, federal, state, local? These are some questions I will answer in capstone project.

Data set

Fields

```
str(income_data_f)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    32561 obs. of  5 variables:
##  $ WorkClass    : chr  "State-gov" "Self-emp-not-inc" "Private" "Private" ...
##  $ Education    : chr  "Bachelors" "Bachelors" "HS-grad" "11th" ...
##  $ MaritalStatus: chr  "Never-married" "Married-civ-spouse" "Divorced" "Married-civ-spouse" ...
##  $ Occupation    : chr  "Adm-clerical" "Exec-managerial" "Handlers-cleaners" "Handlers-cleaners"
...
##  $ LessThen_50  : num  1 1 1 1 1 1 1 0 0 0 ...
```

Limitations

The limitations of this data set, in my opinion, include the year the data was taken from and the amount that was used for the monetary split point of $50,000.

The data was from 1994 which is over 20 years ago. In my opinion factors such as Marital Status and Occupation have had major changes over the period. For example, gay marriage was not legal or officially counted back in 1994, and just social attitudes towards relationships have changed. With the massive growth of the tech industry and the internet, there have obviously also been major changes in the type of occupations there are in the tech sector. For instance, there was no such thing as a Data Scientist in 1994… at least not officially.

The amount of $50,000 as spilt point is a little low in 2018. I believe a split of $75,000 would have been better as a point that represents the separation of the middle class in America. This data is also being used to represent a possible future outcome. A $50,000 per year salary in 10 to 20 years will probably not represent an amount much over the poverty rate in some parts of the United States.

Cleaning and wrangling

The data I used comes from the 1994 census bureau by way of Data Mining and Visualization Silicon Graphics. It was 48,842 instances split into a training set of 32,561 instances and a testing set of 16,281 instances. I further decreased the number by removing variables that I was not using and removing rows that was missing data in one or more of the variables. The factors are WorkingClass, Education, MaritalStatus, and Occupation. The variables I am using for working class are "Federal-gov", "Local-gov", "Never-worked", "Private", "Self-emp-inc", "Self-emp-not-ing", "State-gov", "Without-pay".

The variables I am using for Education are "10th", "11th", "12th", "1st-4th", "5th-6th", "7th-8th", "9th", "Assoc-acdm", "Assoc-voc", "Bachelors", "Doctorate", "HS-grad", "Masters", "Preschool", "Prof-school", "Some-college".

The variables I am using for Marital Status are "Divorced", "Married-AF-spouse", "Married-civ-spouse", "Married-spouse-absent", "Never-married", "Separated", "Widowed".

The variables I am using for Occupation are "Adm-clerical", "Armed-Forces", "Craft-repair", "Exec-managerial", "Farming-fishing", "Handlers-cleaners", "Machine-op-inspct", "Other-service", "Priv-house-serv", "Prof-specialty", "Protective-serv", "Sales", "Tech-support", "Transport-moving".

I added an extra column to represent a person with an income over $50,000 with a 0, and an income under $50,000 with a 1.

Preliminary Logistical Regression formula

```
model <- glm(LessThen_50 ~ WorkClass + Education + MaritalStatus + Occupation,family=binomial(link='l
ogit'),data=income_data_f)
summary(model)
```

```
## 
## Call:
## glm(formula = LessThen_50 ~ WorkClass + Education + MaritalStatus + 
##     Occupation, family = binomial(link = "logit"), data = income_data_f)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -3.4495   0.0793   0.2575   0.5845   2.4744  
## 
## Coefficients: (1 not defined because of singularities)
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                           4.27902    0.17282  24.760  < 2e-16
## WorkClassFederal-gov                 -1.18550    0.13953  -8.496  < 2e-16
## WorkClassLocal-gov                   -0.46962    0.12701  -3.697 0.000218
## WorkClassNever-worked                11.11977  291.15371   0.038 0.969535
## WorkClassPrivate                     -0.63769    0.11185  -5.701 1.19e-08
## WorkClassSelf-emp-inc                -1.17868    0.13480  -8.744  < 2e-16
## WorkClassSelf-emp-not-inc            -0.37428    0.12368  -3.026 0.002477
## WorkClassState-gov                   -0.20785    0.13854  -1.500 0.133535
## WorkClassWithout-pay                 12.10596  204.23048   0.059 0.952732
## Education11th                         0.06519    0.19738   0.330 0.741206
## Education12th                        -0.37086    0.24104  -1.539 0.123912
## Education1st-4th                      0.77263    0.44883   1.721 0.085175
## Education5th-6th                      0.41805    0.29798   1.403 0.160637
## Education7th-8th                      0.42443    0.21885   1.939 0.052455
## Education9th                          0.36928    0.24855   1.486 0.137350
## EducationAssoc-acdm                  -1.24185    0.16338  -7.601 2.94e-14
## EducationAssoc-voc                   -1.24731    0.15676  -7.957 1.76e-15
## EducationBachelors                   -1.91459    0.14578 -13.134  < 2e-16
## EducationDoctorate                   -3.30344    0.19957 -16.553  < 2e-16
## EducationHS-grad                     -0.73860    0.14247  -5.184 2.17e-07
## EducationMasters                     -2.42527    0.15565 -15.582  < 2e-16
## EducationPreschool                   11.43366  106.24244   0.108 0.914298
## EducationProf-school                 -3.13069    0.18512 -16.912  < 2e-16
## EducationSome-college                -1.04403    0.14429  -7.236 4.63e-13
## MaritalStatusMarried-AF-spouse       -2.39984    0.46428  -5.169 2.35e-07
## MaritalStatusMarried-civ-spouse      -2.08355    0.05675 -36.717  < 2e-16
## MaritalStatusMarried-spouse-absent    0.11999    0.19953   0.601 0.547608
## MaritalStatusNever-married            0.85325    0.07153  11.928  < 2e-16
## MaritalStatusSeparated                0.27591    0.14502   1.903 0.057100
## MaritalStatusWidowed                 -0.12532    0.13260  -0.945 0.344593
## OccupationAdm-clerical                0.09969    0.08932   1.116 0.264401
## OccupationArmed-Forces                1.17453    1.26734   0.927 0.354048
## OccupationCraft-repair               -0.06954    0.07951  -0.875 0.381822
## OccupationExec-managerial            -0.89943    0.08113 -11.086  < 2e-16
## OccupationFarming-fishing             0.67258    0.12987   5.179 2.23e-07
## OccupationHandlers-cleaners           0.84740    0.13713   6.180 6.42e-10
## OccupationMachine-op-inspct           0.37896    0.09988   3.794 0.000148
## OccupationOther-service               1.02791    0.11530   8.915  < 2e-16
## OccupationPriv-house-serv             2.45531    1.01952   2.408 0.016027
## OccupationProf-specialty             -0.51055    0.08648  -5.904 3.55e-09
## OccupationProtective-serv            -0.59115    0.12289  -4.811 1.51e-06
## OccupationSales                      -0.33671    0.08352  -4.031 5.54e-05
## OccupationTech-support               -0.54547    0.11129  -4.901 9.51e-07
## OccupationTransport-moving                 NA         NA      NA       NA
## 
```

```
## (Intercept)                        ***
## WorkClassFederal-gov               ***
## WorkClassLocal-gov                 ***
## WorkClassNever-worked
## WorkClassPrivate                   ***
## WorkClassSelf-emp-inc              ***
## WorkClassSelf-emp-not-inc          **
## WorkClassState-gov
## WorkClassWithout-pay
## Education11th
## Education12th
## Education1st-4th                   .
## Education5th-6th
## Education7th-8th                   .
## Education9th
## EducationAssoc-acdm                ***
## EducationAssoc-voc                 ***
## EducationBachelors                 ***
## EducationDoctorate                 ***
## EducationHS-grad                   ***
## EducationMasters                   ***
## EducationPreschool
## EducationProf-school               ***
## EducationSome-college              ***
## MaritalStatusMarried-AF-spouse     ***
## MaritalStatusMarried-civ-spouse    ***
## MaritalStatusMarried-spouse-absent
## MaritalStatusNever-married         ***
## MaritalStatusSeparated             .
## MaritalStatusWidowed
## OccupationAdm-clerical
## OccupationArmed-Forces
## OccupationCraft-repair
## OccupationExec-managerial          ***
## OccupationFarming-fishing          ***
## OccupationHandlers-cleaners        ***
## OccupationMachine-op-inspct        ***
## OccupationOther-service            ***
## OccupationPriv-house-serv          *
## OccupationProf-specialty           ***
## OccupationProtective-serv          ***
## OccupationSales                    ***
## OccupationTech-support             ***
## OccupationTransport-moving
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 35948  on 32560  degrees of freedom
## Residual deviance: 23861  on 32518  degrees of freedom
## AIC: 23947
##
## Number of Fisher Scoring iterations: 13
```