CENSUS DATA'S RELATIONSHIP TO INCOME

Capstone

Anthony Stachowitz

August 9, 2018

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 job choices? Are there basic life decisions that influence how our future will unfold? Is there advice we can give our children that will, if followed, give them an advantage?

I have analyzed a data set 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 of 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 categories, four of the categories were used to predict the fifth variable.

The outcome I am predicting is whether you are going to make more than $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.

PRESCIENCE

The census data used in my capstone is from 1994. There are some considerations that must be takin into account.

Factors such as Marital Status and Occupation have had major changes over this period. For example, gay marriage was not legal or officially counted back in 1994, and 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 a spilt point is a little low in 2018. I believe a split of $75,000 or $100,000 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.

THE DATA

The data I used comes from the 1994 census bureau by way of Data Mining and Visualization Silicon Graphics. It includes 48,842 instances split into a training set of 32,561 instances and a testing set of 16,281 instances.

Variable	Description
Age	Age of the person
WorkClass	Type S
FnlWgt	Weight based on socio-economic situation
Education	Highest education achieved
Education number	Number associated with education achieved
MaritalStatus	Marital Status or Divorce information
Occupation	Overall class of employment
Relationship	Current status in the family
Race	Race
Sex	Sex
CapitalGain	Capital Gains
CapitalLoss	Capital Losses
Hours per week	Hour worked per week
NativeCountry	Birth country
Salary	Salary amount

DATA WRANGLING

Headers were added to the data.

```
names(income_data) <- c("Age", "WorkClass", "FnlWgt", "Education", "Education
-num", "MaritalStatus", "Occupation", "Relationship", "Race", "Sex", "Capital
Gain", "CapitalLoss", "HoursPerWeek", "NativeCountry", "Salary")
```

The salary range under $50,000 was changed to be represented by the binary variable 1. The salary range over $50,000 was changed to the binary variable 0. This allowed us to use our regression analysis.

```
income_data <- mutate(income_data, LessThen_50 = ifelse(grepl(">50K", Salary)
, 0, 1))
```

The data set was cleaned by first removing categories and variables that I was not using.

The factors I used in my final analysis are WorkingClass, Education, MaritalStatus, and Occupation.

```
income_data <- income_data[, -c(1, 3, 5, 8:15)]

income_data

## # A tibble: 32,561 x 5

##    WorkClass        Education MaritalStatus         Occupation LessThen_50

##    <fct>            <fct>     <fct>                 <fct>            <dbl>

##  1 State-gov        Bachelors Never-married         Adm-cleri~           1

##  2 Self-emp-not-inc Bachelors Married-civ-spouse    Exec-mana~           1

##  3 Private          HS-grad   Divorced              Handlers-~           1

##  4 Private          11th      Married-civ-spouse    Handlers-~           1

##  5 Private          Bachelors Married-civ-spouse    Prof-spec~           1

##  6 Private          Masters   Married-civ-spouse    Exec-mana~           1

##  7 Private          9th       Married-spouse-absent Other-ser~           1

##  8 Self-emp-not-inc HS-grad   Married-civ-spouse    Exec-mana~           0
##  9 Private          Masters    Never-married         Prof-spec~           0
## 10 Private          Bachelors Married-civ-spouse    Exec-mana~           0
## # ... with 32,551 more rows
```

All the data was converted to factors.

```
income_data$Occupation <- as.factor(income_data$Occupation)
income_data$WorkClass <- as.factor(income_data$WorkClass)
income_data$Education <- as.factor(income_data$Education)
income_data$MaritalStatus <- as.factor(income_data$MaritalStatus)
```

Missing values are turned into NA’s.

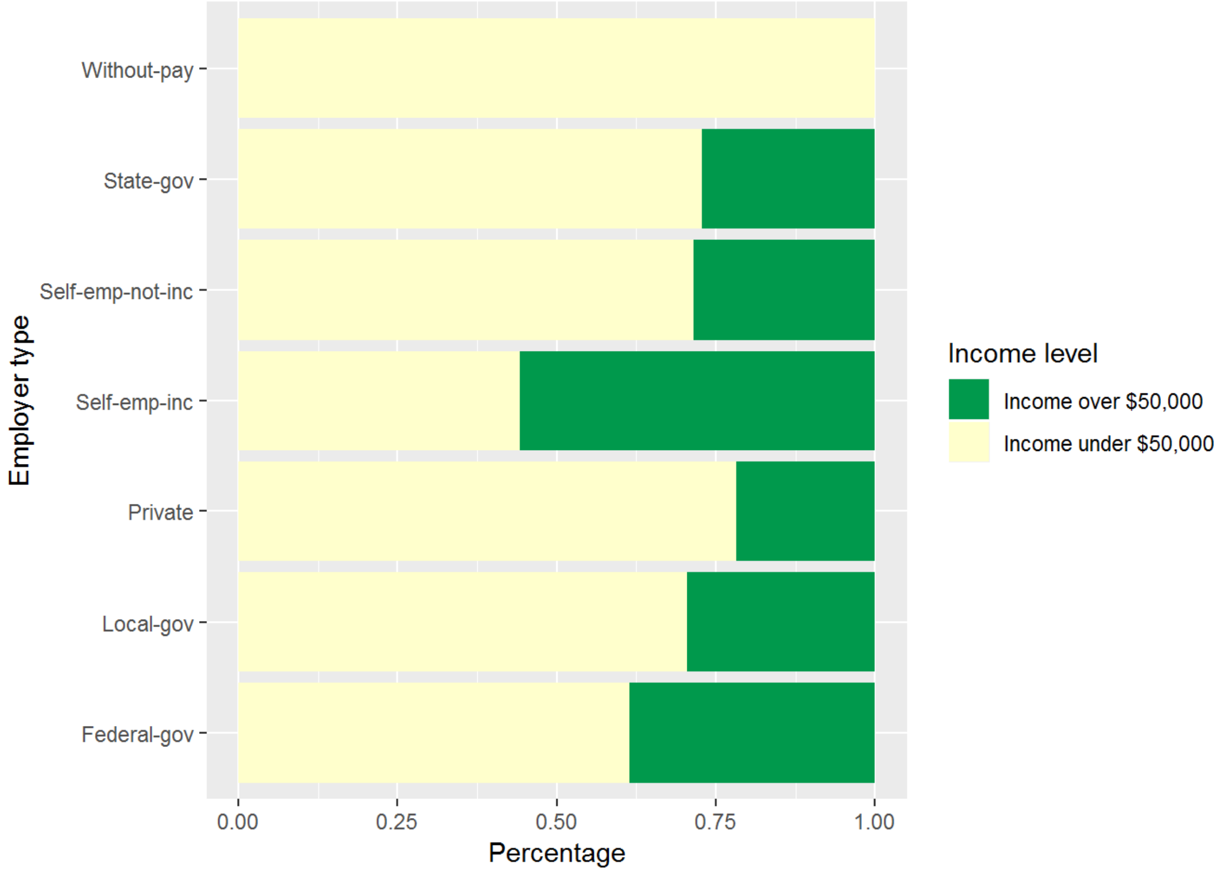
The data was checked for NA’s and the data with NA’s was removed and rechecked.

```
income_data[ income_data == "?" ] <- NA
sapply(income_data,function(x) sum(is.na(x)))
##      WorkClass      Education MaritalStatus    Occupation   LessThen_50
##           1836              0             0          1843             0
income_data <- income_data[complete.cases(income_data), ]
sapply(income_data,function(x) sum(is.na(x)))
##      WorkClass      Education MaritalStatus    Occupation   LessThen_50
##              0              0             0             0             0
str(income_data)
## Classes 'tbl_df', 'tbl' and 'data.frame':    30718 obs. of  5 variables:
##  $ WorkClass    : Factor w/ 9 levels "?","Federal-gov",..: 8 7 5 5 5 5 5 7
5 5 ...
##  $ Education    : Factor w/ 16 levels "10th","11th",..: 10 10 12 2 10 13 7
12 13 10 ...
##  $ MaritalStatus: Factor w/ 7 levels "Divorced","Married-AF-spouse",..: 5
3 1 3 3 3 4 3 5 3 ...
##  $ Occupation   : Factor w/ 15 levels "?","Adm-clerical",..: 2 5 7 7 11 5
9 5 11 5 ...
```

```
##  $ LessThen_50  : num  1 1 1 1 1 1 1 0 0 0 ...
```

EXPLORATORY DATA ANALYSIS

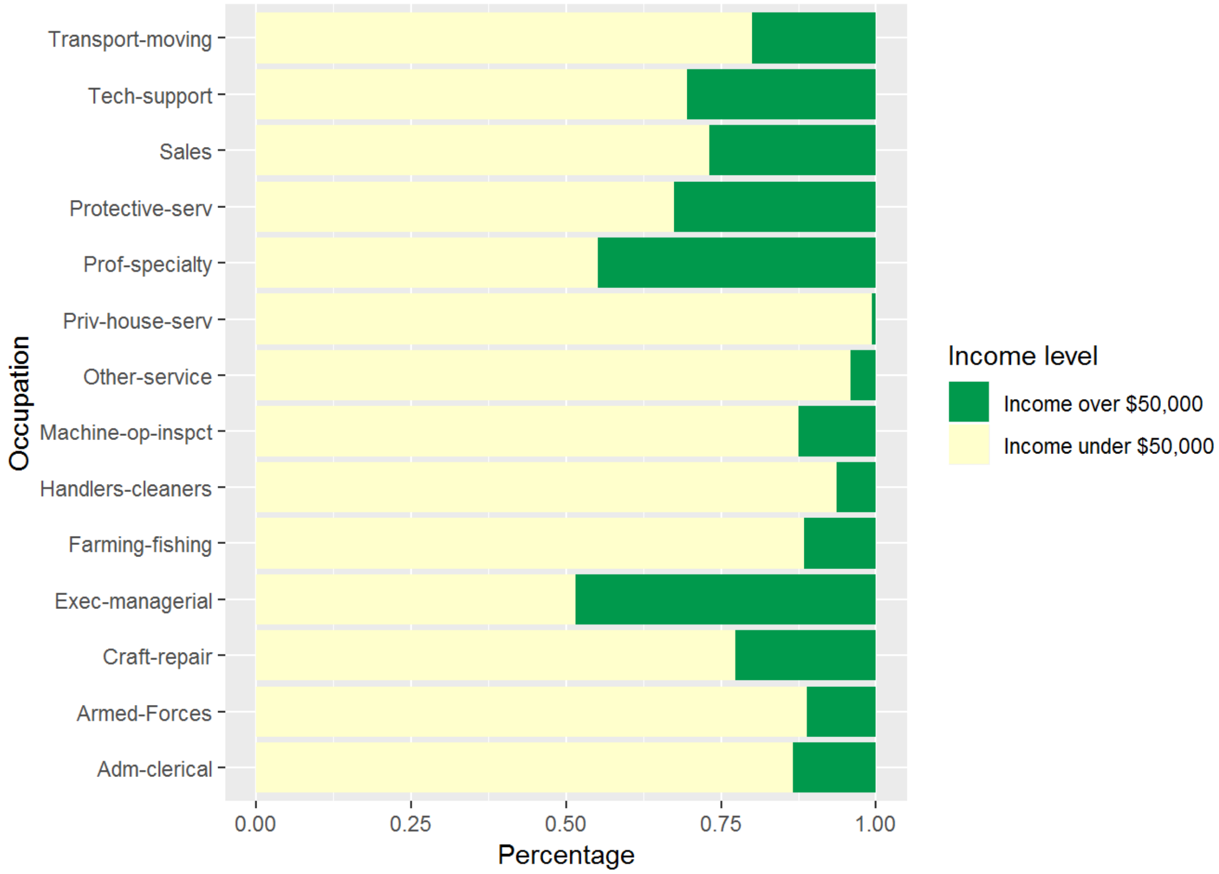
Working Class



The variable that has the strongest relationship to having an income over $50,000 is self-employed at a corporation.

After being self-employed within a corporate structure, every other employment category (besides being unemployed) is within about 25% of each other. In other words, there is a fair chance that, if you are working, your chance of making over $50,00 is similar. Based on this bar chart, being self-employed within a corporate structure would be the best indicator among classification of work to predict a salary of above $50,000.

Occupation



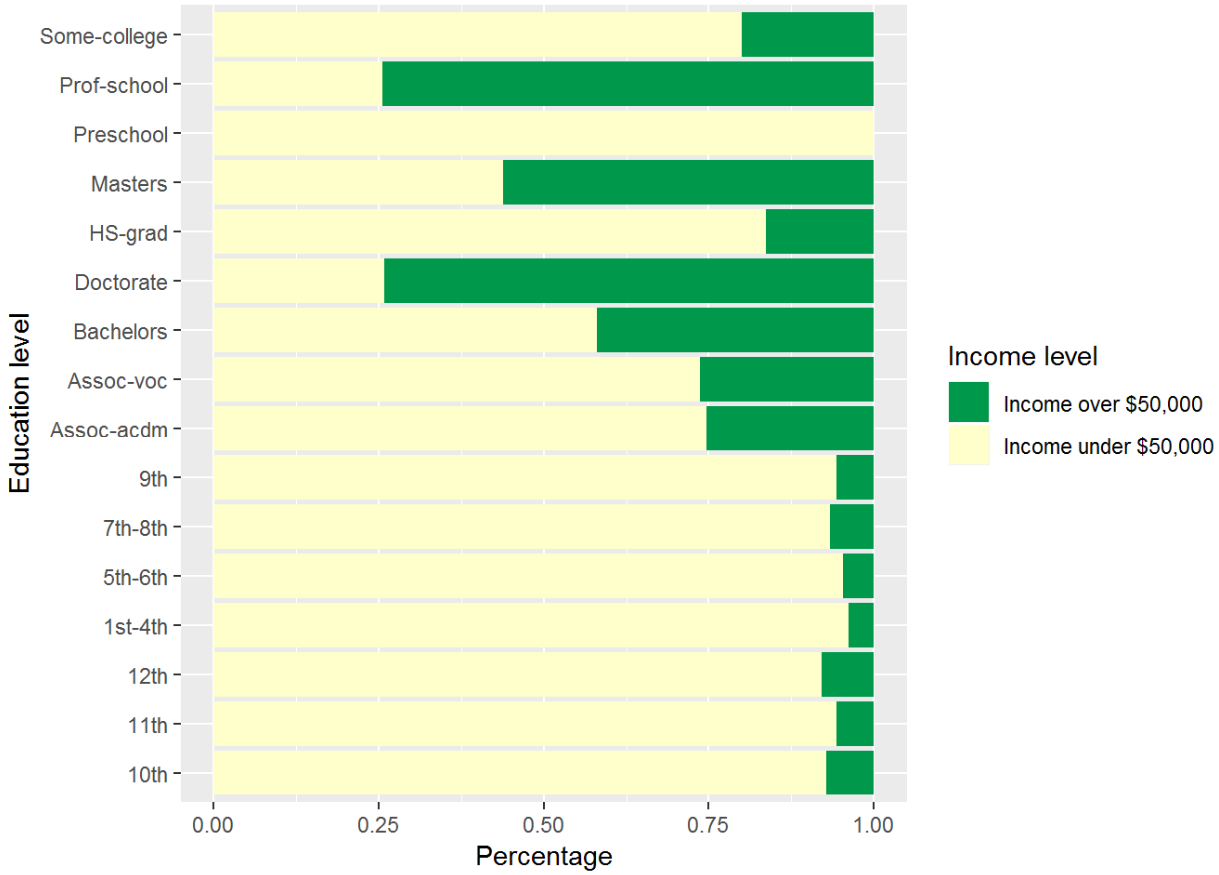
The occupation category has fourteen variables. They are set up, so most people can fit their specific job into a variable. This is the most subjective category of the four that I chose. It is also probably the most undependable category because of that reason. I struggled a little with the decision to include this category but it the end I figured that in this case more information would not cloud the results and I am individually measuring each of the variables in each category against income, so the occupation category will not influence the relevance of the other variables.

The categories of prof-specialty (lawyer, accountant, etc.) and Exec-managerial (pres., vise-pres., etc.) have the percentage of people that earn over $50,000. The only real surprise to me was that the sales category percentage is not higher.

Unfortunately, while looking at this data there seems to be an issue with how subjective this data is You can easily fit a certain job in more than on variable. For instance, you can work for the armed forces and have almost any job fit into any one of the other variables. You can work in the transportation business answering phones to get clients who need to be move across country. And, in that case, you would work in transportation, sales, and Adm-clerical. This data set needs to be further parsed to be able to depend on the results of my final analysis.

Although, the final analysis will provide a new starting point for a more in-depth analysis in the future.

Education Level

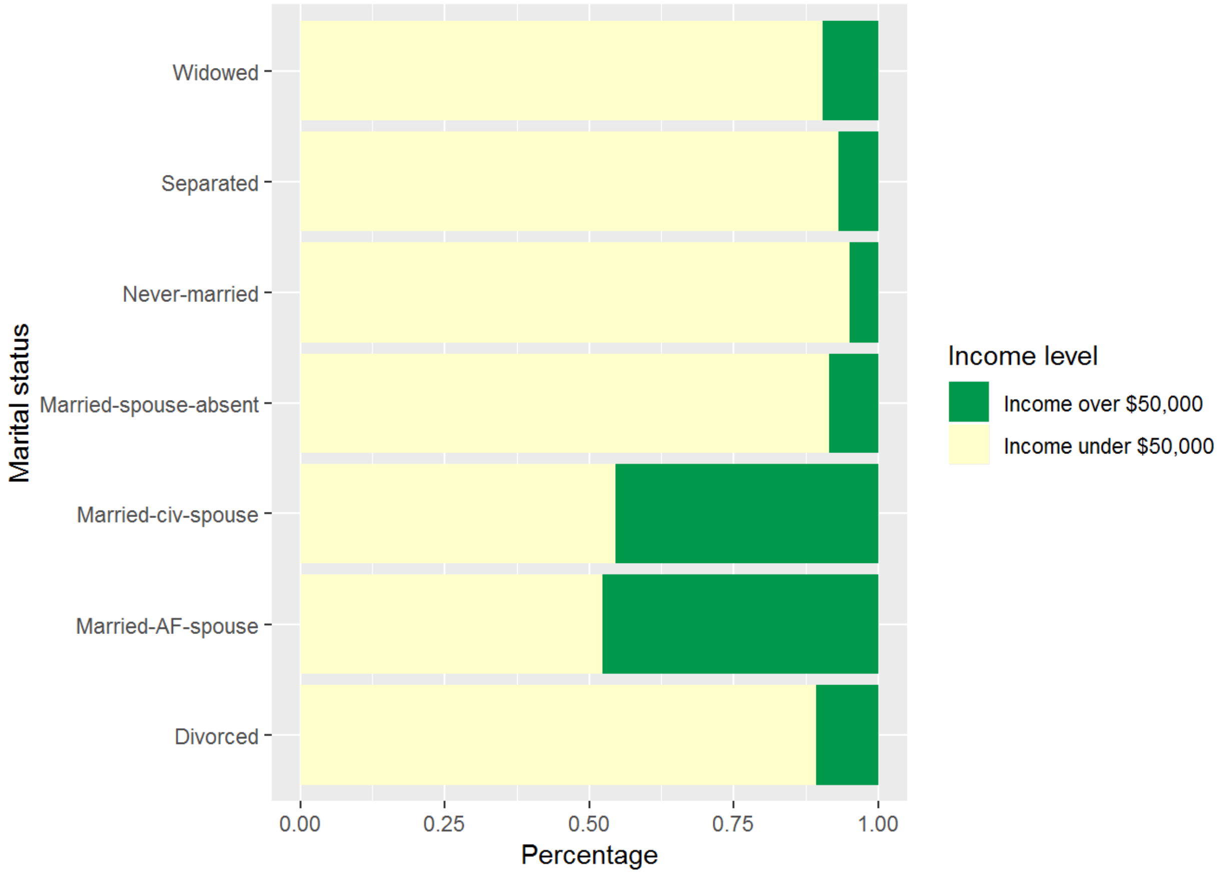


The education category has 16 variables. They range from pre-school to doctorate. The lower grades are grouped by two or four until you get to the 9th grade and each level represents one individual year after that.

This category has the least surprising results. The highest percentages are Prof-school and Doctorate with the rest of the variable falling in order with more education giving you a better chance of making over $50,000 per year.

This is also the category that I think will have the most obvious and predictable relationship to salary. Generally, most people believe the more education you obtain the higher the probability of achieving financial success over your lifetime. It will be interesting to see if the results go up proportionately with the higher level of education.

Marital Status



Marital status has six variables. They range from being married, divorced, and never married, with different classifications within those. This is the most interesting category for the question I am answering with this set of data.

This date suggests that there is a considerably better chance of making $50,000 if you are married than any other variable listed. It seems that getting divorced has a measurable effect on your income.

Most people wouldn't look to a person's marriage status as an indicator to what type of salary you would earn. But marriage is such an emotionally charged union, as most types of serious relationships are. Emotion, mental health and general wellbeing influences your day to day life, so you would probably expect that to spill over into your work life.

There are two problems with this data that jump out at me. The first is that according to this census from 1994, the divorce rate is low. The second is that this data doesn't include civil unions or gay marriage.

The divorce rate now is generally understood to be about 50% and the rate give visually by this data looks to be around 23%. At the least, this data may not relate correctly to the data for 2018 and would have to be updated in another study.

Gay marriage wasn't legal until recently so there was no official information on those unions. They were not included in this data. Therefore, information involving gay marriage cannot in any way be estimated. It has a negative effect on the other results in this set because it will change the percentages of the results by changing the number of every variable except the never-married variable.

Predictive Model

Binomial Linier Regression

```
Call:
glm(formula = LessThen_50 ~ WorkClass + Education + MaritalStatus + 
    Occupation, family = binomial(link = "logit"), data = income_data)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-3.4527   0.0004   0.2627   0.5864   2.4850  

Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)    
(Intercept)                         3.183101   0.177233  17.960  < 2e-16 ***
WorkClassLocal-gov                  0.715125   0.104388   6.851 7.35e-12 ***
WorkClassPrivate                    0.547025   0.086547   6.321 2.61e-10 ***
WorkClassSelf-emp-inc               0.005631   0.113435   0.050  0.96041    
WorkClassSelf-emp-not-inc           0.810610   0.101350   7.998 1.26e-15 ***
WorkClassState-gov                  0.974774   0.116737   8.350  < 2e-16 ***
WorkClassWithout-pay               13.283264 203.702993   0.065  0.94801    
Education11th                       0.013560   0.199883   0.068  0.94591    
Education12th                      -0.340111   0.247827  -1.372  0.16995    
Education1st-4th                    0.727442   0.450960   1.613  0.10672    
Education5th-6th                    0.463630   0.315349   1.470  0.14150    
Education7th-8th                    0.418560   0.224325   1.866  0.06206 .  
Education9th                        0.362343   0.253766   1.428  0.15333    
EducationAssoc-acdm                -1.228764   0.166367  -7.386 1.51e-13 ***
EducationAssoc-voc                 -1.213716   0.159777  -7.596 3.05e-14 ***
EducationBachelors                 -1.905437   0.148699 -12.814  < 2e-16 ***
EducationDoctorate                 -3.266194   0.203979 -16.012  < 2e-16 ***
EducationHS-grad                   -0.732734   0.145212  -5.046 4.51e-07 ***
EducationMasters                   -2.420237   0.158815 -15.239  < 2e-16 ***
EducationPreschool                 11.399078 110.939745   0.103  0.91816    
EducationProf-school               -3.151988   0.189543 -16.629  < 2e-16 ***
EducationSome-college              -1.031666   0.147082  -7.014 2.31e-12 ***
MaritalStatusMarried-AF-spouse     -2.533957   0.482825  -5.248 1.54e-07 ***
MaritalStatusMarried-civ-spouse    -2.078186   0.057286 -36.277  < 2e-16 ***
MaritalStatusMarried-spouse-absent  0.108060   0.203135   0.532  0.59475    
MaritalStatusNever-married          0.848054   0.072233  11.740  < 2e-16 ***
MaritalStatusSeparated              0.255255   0.145548   1.754  0.07947 .  
```

```
MaritalStatusWidowed              -0.152596   0.136347  -1.119  0.26306    
OccupationArmed-Forces             1.070579   1.266781   0.845  0.39805    
OccupationCraft-repair            -0.170936   0.070666  -2.419  0.01557 *  
OccupationExec-managerial         -0.999383   0.068536 -14.582  < 2e-16 ***
OccupationFarming-fishing          0.570972   0.124839   4.574 4.79e-06 ***
OccupationHandlers-cleaners        0.747974   0.131965   5.668 1.44e-08 ***
OccupationMachine-op-inspct        0.279358   0.093405   2.991  0.00278 ** 
OccupationOther-service            0.930297   0.108368   8.585  < 2e-16 ***
OccupationPriv-house-serv          2.360394   1.018562   2.317  0.02048 *  
OccupationProf-specialty          -0.608484   0.073456  -8.284  < 2e-16 ***
OccupationProtective-serv         -0.693413   0.115233  -6.018 1.77e-09 ***
OccupationSales                   -0.436858   0.072891  -5.993 2.06e-09 ***
OccupationTech-support            -0.648725   0.101731  -6.377 1.81e-10 ***
OccupationTransport-moving        -0.099341   0.089326  -1.112  0.26609    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 34483  on 30717  degrees of freedom
Residual deviance: 23016  on 30677  degrees of freedom
AIC: 23098

Number of Fisher Scoring iterations: 13
```

Now we can analyze the fitting and interpret what the model is telling us.
First, we can see that the following variables are not statistically significant:

WorkClassSelf-emp-inc, WorkClassState-gov, WorkClassWithout-pay, Education11th, Education12th, Education1st-4th, Education5th-6th, Education9th, EducationPreschool, MaritalStatusMarried-spouse-absent, MaritalStatusWidowed, OccupationArmed-Forces, OccupationTransport-moving

As for the statistically significant variables, **MaritalStatusMarried-civ-spouse** has the lowest p-value suggesting a strong association of the being married to a civilian with the probability of having an income over $50,000. The negative coefficient for this predictor suggests that all other variables being equal, the marital status of being married to a civilian is less likely to have an income of less than $50,000.

Assessing the predictive ability of the model

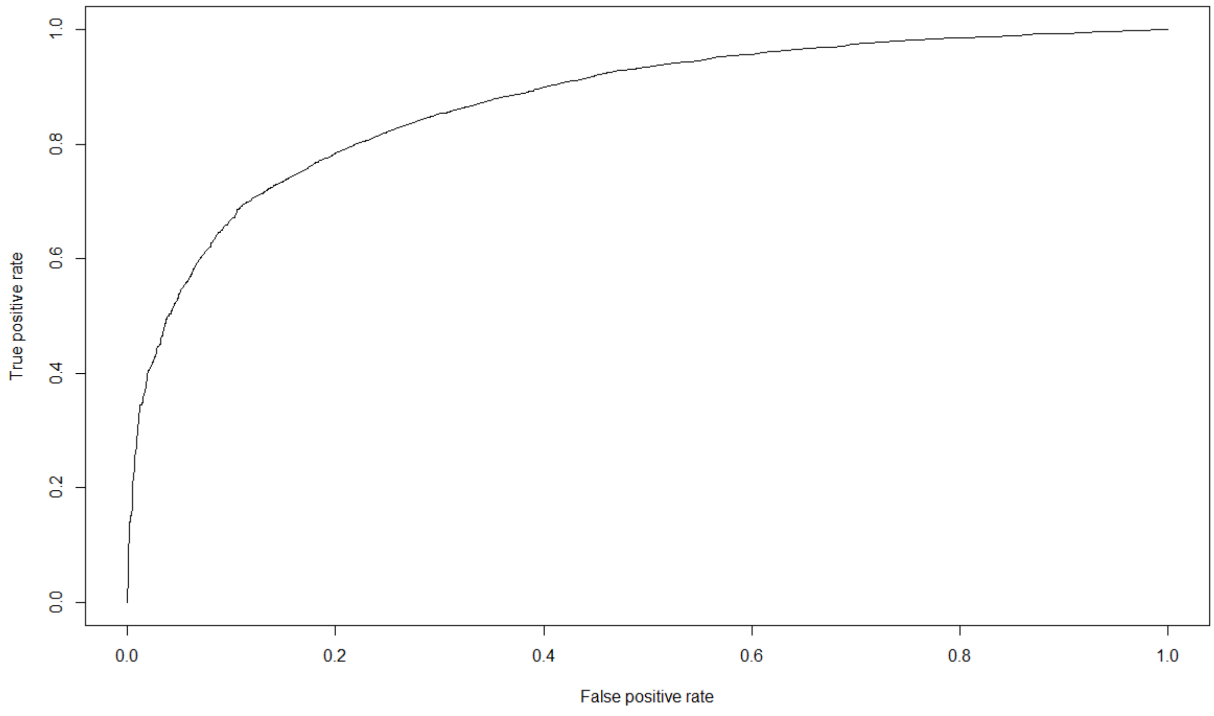
```
> fitted.results <- predict(model1,newdata=income_test_data,type='response')
> fitted.results <- ifelse(fitted.results > 0.5,1,0)
> misClasificError <- mean(fitted.results != income_test_data$LessThen_50)
> print(paste('Accuracy',1-misClasificError))
```

```
[1] "Accuracy 0.827424094025465"
```

By setting the parameter type=response, The output probabilities in the form of P(y=1|X). Our decision boundary will be 0.5. If P(y=1|X) > 0.5 then y = 1 otherwise y=0.

The accuracy rate is 0.83 tested against the test data set of 15,315 (after cleaning). This is a very good result.

As a last test I am going to plot a ROC curve and calculate the AUC (area under the curve. A model with a good predictive ability should have an AUC value closer to 1 then 0.5.



```
> library(ROCR)
> p <- predict(model, newdata=income_test_data, type="response")
> pr <- prediction(p, income_test_data$LessThen_50)
> prf <- performance(pr, measure = "tpr", x.measure = "fpr")
> plot(prf)
>
> auc <- performance(pr, measure = "auc")
> auc <- auc@y.values[[1]]
> auc
[1] 0.8730041
```

The AUC value is 0.87 which is closer to 1 then 0.5. This model has good predictablity.

Conclusion

The exploratory data analysis gave us a good idea where the machine learning model would take us. The exploratory data showed a somewhat surprising relationship between being married, civilian or armed forces, and earning over $50,000. The surprise was not that people who are married, civilian or armed forces, earn more than $50,000, it was the rate at which they earn over $50,000 compared to every other classification within the marital status field.

The machine leering model also confirmed that being married had a strong relationship to earning over $50,000. The regression model was further able to dig deeper and show the difference between being married in a civilian family vs being married in an armed forces family. This distinction is important because just using the exploratory data, you would believe that the being married in an armed forces family may be a slightly better predictor then being in civilian family. This is incorrect. The regression model shows that being married in a civilian family is more significant than being married in an armed forces family along with every other variable tested in this model. Being married in an armed forces family is still significant but not as significant being married in a civilian family or as other variables in the model.

The relevant finding in this model is that a social construct, marriage, has a stronger relationship with earing money then more typical factors like education and type of work. This analysis is a good starting point for further analysis using more current census information and updated laws and societal norms.