Exploring the Relationship Between Education, Marital Status, and Salary

5/4/22

Victor Chen, Tristan Duerk

Project Description:

For our final project, we decided to explore personal salary data. We were curious to see how education level and marital status affected whether or not someone made over $50k annually, as we predicted these to be the 2 most important factors in determining one's career and salary. We hypothesized that those with a higher education level and those who are married to a present and living partner are more likely to make over $50k annually than those with a lower education level and either have absent or deceased partners or have never been married. We decided to explore this through logistic regression and a decision tree classifier.

Resources:

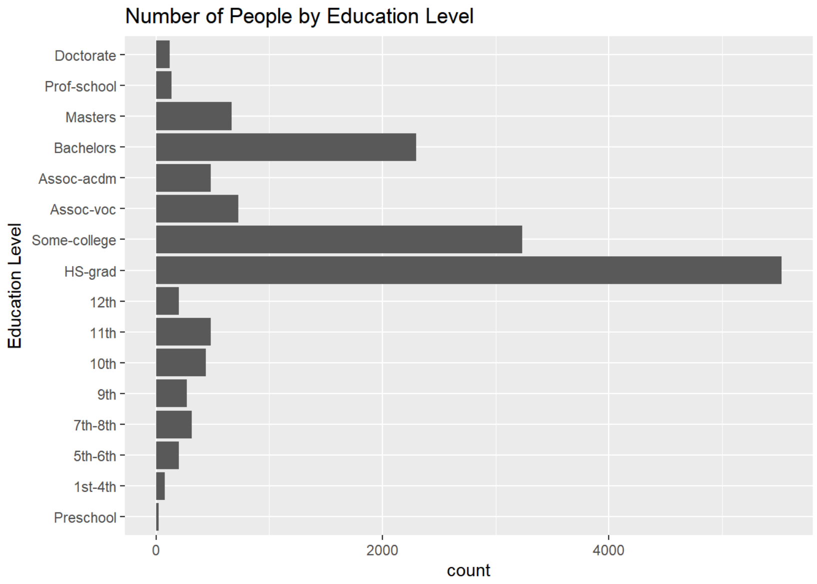
Datatset Link: https://www.kaggle.com/datasets/ayessa/salary-prediction-classification

Our dataset contains 15 columns with over 32k observations. The variables of interest are "education" (categorical, education level) and "marital.status" (categorical) as predictor variables and "salary" (categorical, <=50k or >50k) as the outcome variable.

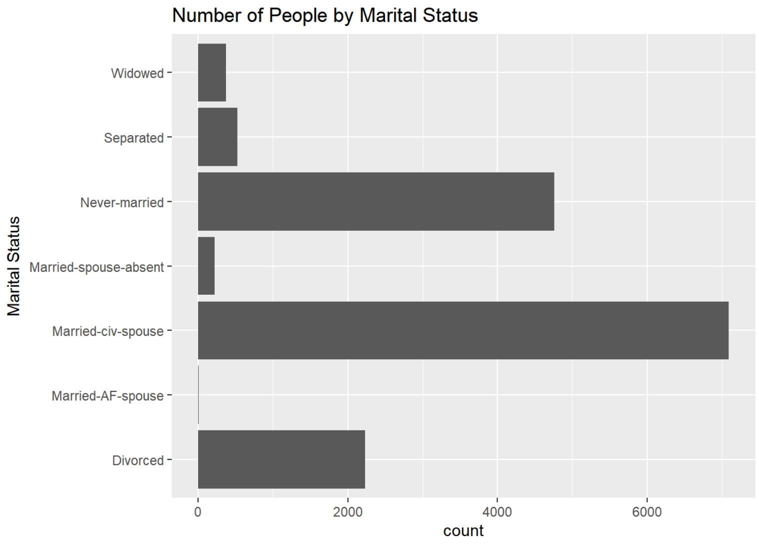
For the logistic regression the libraries ggplot2, dplyr, tidyverse, ROCR, sjPlot, sjmisc, and sjlabelled were used. The "age" variable was also filtered to only contain observations with age = 40 to make sure the comparison was between people at a similar "place" in life where there would be a significant amount of people making greater than $50k annually.

For the decision tree model, the libraries rpart, tree, randomForest, and gbm were used. The data was not subset or filtered and the model directly worked with the original observations.

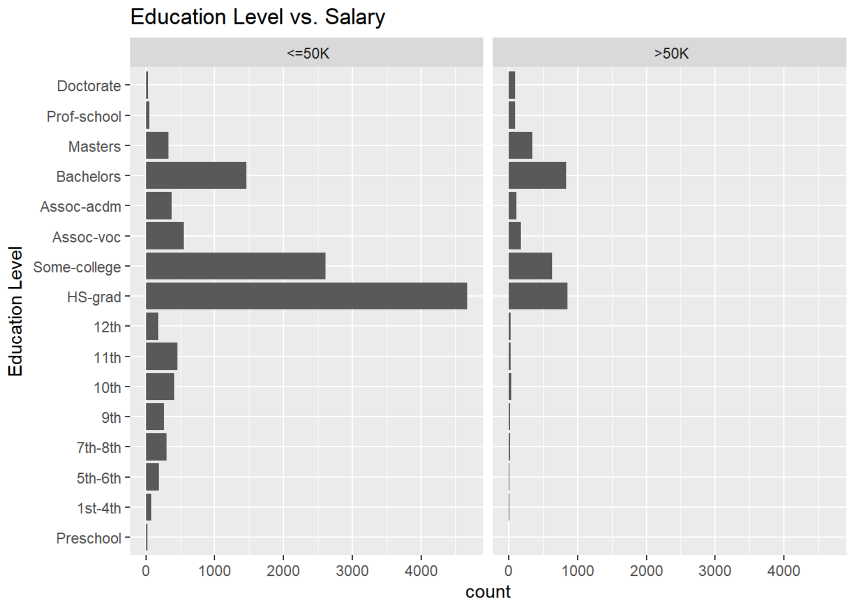
Plots and Models:



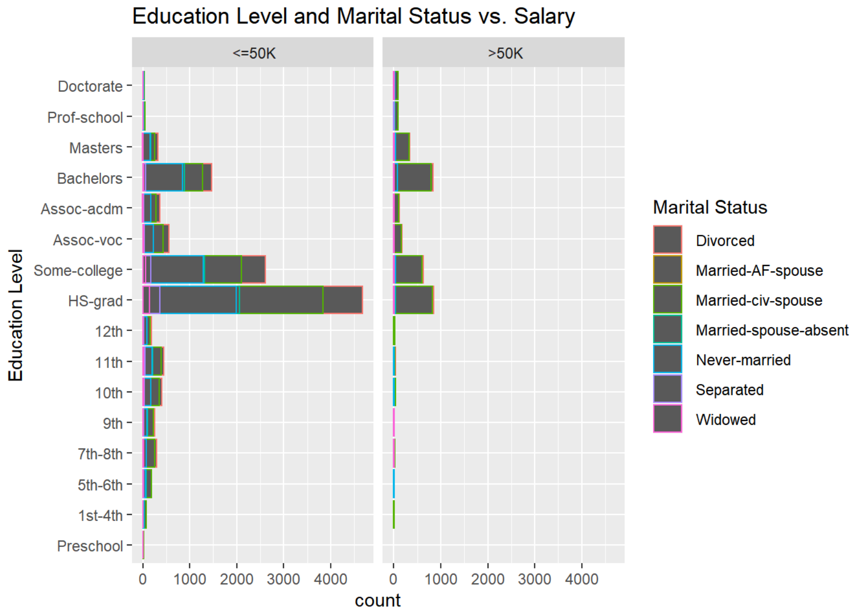
This plot shows the general distribution of the variable representing education level for all observations with an age of 40. Among these observations, it appears that most have graduated high school, done some college, or have a bachelors degree.



This plot shows the general distribution of the variable representing marital status for all observations with an age of 40. Among these observations, it appears that most are civilly married, never married, or divorced.

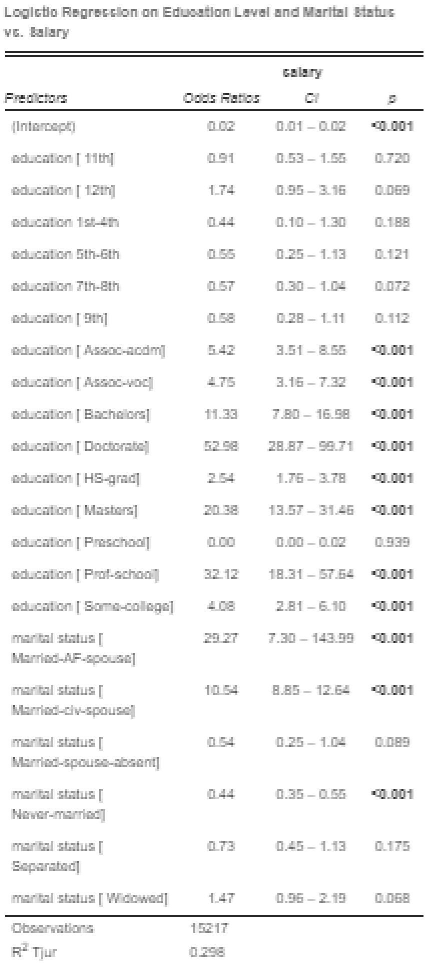


This plot shows the distribution of years educated when faceted by salary group. Among the group which makes less than or equal to $50k annually, high school level education appears to be much more common than college level education. Among the group which makes more than $50k annually, high school level education appears to be equally as common as college bachelors level education, supporting our hypothesis about education level.



This visualization adds the marital status variable to the above visualization. Among the group which makes less than or equal to $50k annually, the distribution of marital statuses is fairly diversified, while the majority of observations in the group which makes more than $50k annually appears to be composed mostly of those civilly married. This supports our hypothesis about marital status.

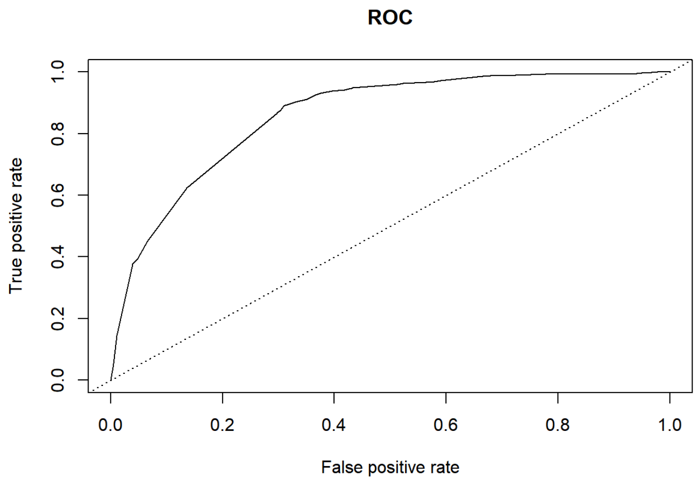
```
##
## Call:
## glm(formula = salary ~ education + marital.status, family = binomial(logit),
##     data = salary.df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1433  -0.5585  -0.2729  -0.1181   3.3335
##
## Coefficients:
##                                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)                           -4.13458    0.20908 -19.775  < 2e-16 ***
## education 11th                        -0.09831    0.27446  -0.358   0.7202
## education 12th                         0.55536    0.30581   1.816   0.0694 .
## education 1st-4th                     -0.82577    0.62661  -1.318   0.1876
## education 5th-6th                     -0.59033    0.38075  -1.550   0.1210
## education 7th-8th                     -0.56049    0.31105  -1.802   0.0716 .
## education 9th                         -0.55100    0.34642  -1.591   0.1117
## education Assoc-acdm                   1.68978    0.22659   7.457 8.82e-14 ***
## education Assoc-voc                    1.55718    0.21362   7.289 3.11e-13 ***
## education Bachelors                    2.42741    0.19782  12.271  < 2e-16 ***
## education Doctorate                    3.96991    0.31602  12.562  < 2e-16 ***
## education HS-grad                      0.93263    0.19421   4.802 1.57e-06 ***
## education Masters                      3.01437    0.21402  14.084  < 2e-16 ***
## education Preschool                  -12.09260  158.03911  -0.077   0.9390
## education Prof-school                  3.46956    0.29230  11.870  < 2e-16 ***
## education Some-college                 1.40540    0.19657   7.150 8.70e-13 ***
## marital.status Married-AF-spouse       3.37640    0.73264   4.609 4.06e-06 ***
## marital.status Married-civ-spouse      2.35555    0.09086  25.924  < 2e-16 ***
## marital.status Married-spouse-absent  -0.61334    0.36112  -1.698   0.0894 .
## marital.status Never-married          -0.82732    0.11749  -7.041 1.90e-12 ***
## marital.status Separated              -0.31979    0.23597  -1.355   0.1754
## marital.status Widowed                 0.38323    0.21018   1.823   0.0683 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 15777  on 15216  degrees of freedom
## Residual deviance: 11176  on 15195  degrees of freedom
## AIC: 11220
##
## Number of Fisher Scoring iterations: 13
```



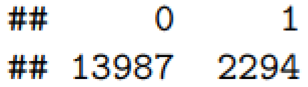
Logistic Regression on Education Level and Marital Status vs. Salary

	salary		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	0.02	0.01 – 0.02	**<0.001**
education [11th]	0.91	0.53 – 1.55	0.720
education [12th]	1.74	0.95 – 3.16	0.069
education 1st-4th	0.44	0.10 – 1.30	0.188
education 5th-6th	0.55	0.25 – 1.13	0.121
education 7th-8th	0.57	0.30 – 1.04	0.072
education [9th]	0.58	0.28 – 1.11	0.112
education [Assoc-acdm]	5.42	3.51 – 8.55	**<0.001**
education [Assoc-voc]	4.75	3.16 – 7.32	**<0.001**
education [Bachelors]	11.33	7.80 – 16.98	**<0.001**
education [Doctorate]	52.98	28.87 – 99.71	**<0.001**
education [HS-grad]	2.54	1.76 – 3.78	**<0.001**
education [Masters]	20.38	13.57 – 31.46	**<0.001**
education [Preschool]	0.00	0.00 – 0.02	0.939
education [Prof-school]	32.12	18.31 – 57.64	**<0.001**
education [Some-college]	4.08	2.81 – 6.10	**<0.001**
marital status [Married-AF-spouse]	29.27	7.30 – 143.99	**<0.001**
marital status [Married-civ-spouse]	10.54	8.85 – 12.64	**<0.001**
marital status [Married-spouse-absent]	0.54	0.25 – 1.04	0.089
marital status [Never-married]	0.44	0.35 – 0.55	**<0.001**
marital status [Separated]	0.73	0.45 – 1.13	0.175
marital status [Widowed]	1.47	0.96 – 2.19	0.068
Observations	15217		
R^2 Tjur	0.298		

These two figures show the results of the logistic regression performed on “salary~education+marital.status”. The results show that high school graduate level of education, any sort of college level of education, being civilly married, being married away from your spouse, and having never been married are all significant in determining whether or not one makes more than $50k annually. These findings support our hypotheses, as those who have a high school diploma level of education, those who have some form of college level of education, those who are civilly married, and those who are married away from their spouse are all more likely to make more than $50k annually, while those who have never been married are less likely to make more than $50k annually.



The above visual is the ROC curve of the regression. The area under the curve was 0.8585692, meaning that the regression was fairly accurate. This is further supported by the misclassification rate being as low as 0.1622865.



```
##     0     1
## 13987  2294
```

The above screenshot shows the results of our “bag.salary<-randomForest(salary ~ education + marital.status, data=train.data, mtry=2, importance=TRUE)” bagged random forest tree prediction model. The model predicted more people overall to make less than or equal to $50k (represented by 0 which was 13987 people) a year compared to more than $50k a year (represented by 1 which was 2294 people). This model proved to be fairly accurate with a misclassification rate of 0.1793502. Keep in mind that this was across all observations in the dataset with no filter or subset.

Summary:

Our hypothesis about education was generally affirmed. Our findings suggest that those with a higher level of education are more likely to make over $50k a year.

Our hypothesis about marital status was also generally affirmed. Our findings suggest that those who are married to a present and/or living partner are more likely to make over $50k a year.

Here are some other interesting findings:

- In general, there is a greater probability that someone makes less than or equal to $50k a year compared to more than $50k a year.
- Those who finished highschool are much more likely to make over $50k a year than people who have not finished high school even if they have completed most of high

school (drastic jump, almost no chance of making more than 50k a year unless you at least complete high school)

- It is not until you reach masters or higher degree level of education where you have more people making over $50k annually than less than or equal to $50k annually.
- Remember, this data is from 1994 and worldwide. If you were focusing on modern-day America, we hypothesize that there would be a lot more observations making more than $50k a year.