Gender Pay Gap: Predicting Income Based on Demographic Data

Katie Peterson

Supervised Learning Capstone

May 2018

Photo: Getty Images

Research Questions

What demographic data is the best determinant for a person to have a higher income?

- Level of education?
- Occupation?
- Race?

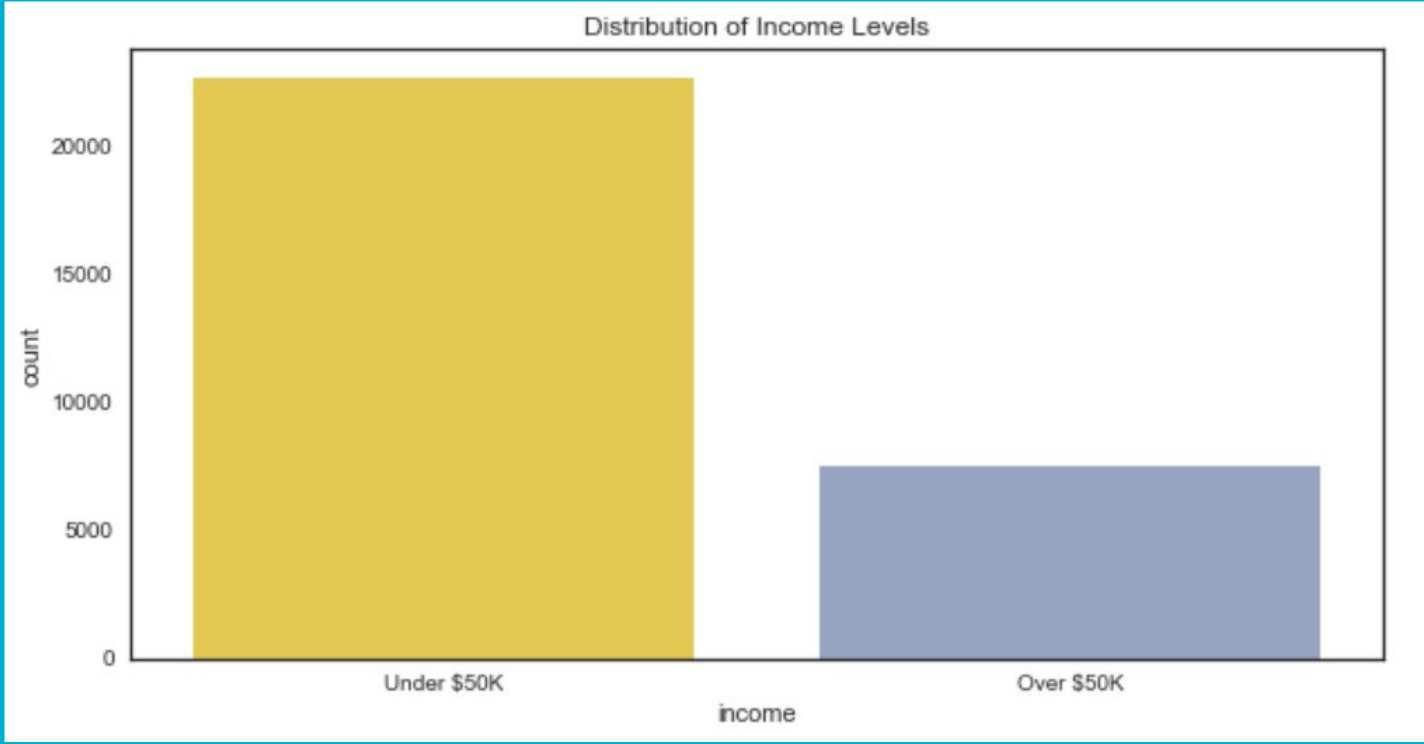
Do these features differ between men and women?

Data Set - 1994 Census Bureau Database

- ~32,000 working people over the age of 16, who made over $100 that year and who are representative of the larger population
- Tracked if income was over or under $50,000
    - Note: After accounting for inflation and cost of living increases, $50,000 in 1994 would be worth approximately $84,500 in 2018.

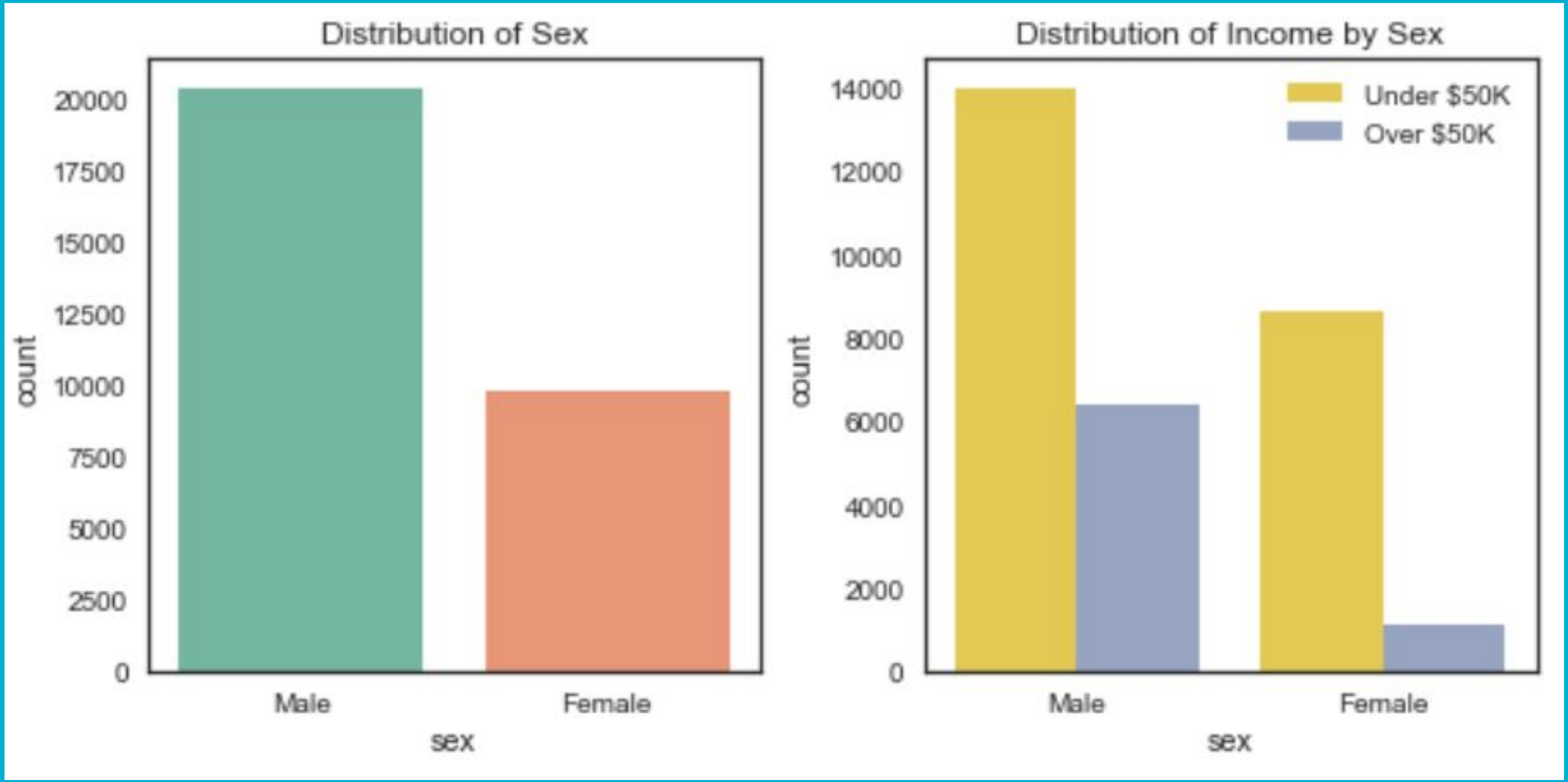
Data Source: UCI Machine Learning Repository
Source: Bureau of Labor and Statistics

Counts of Income Level

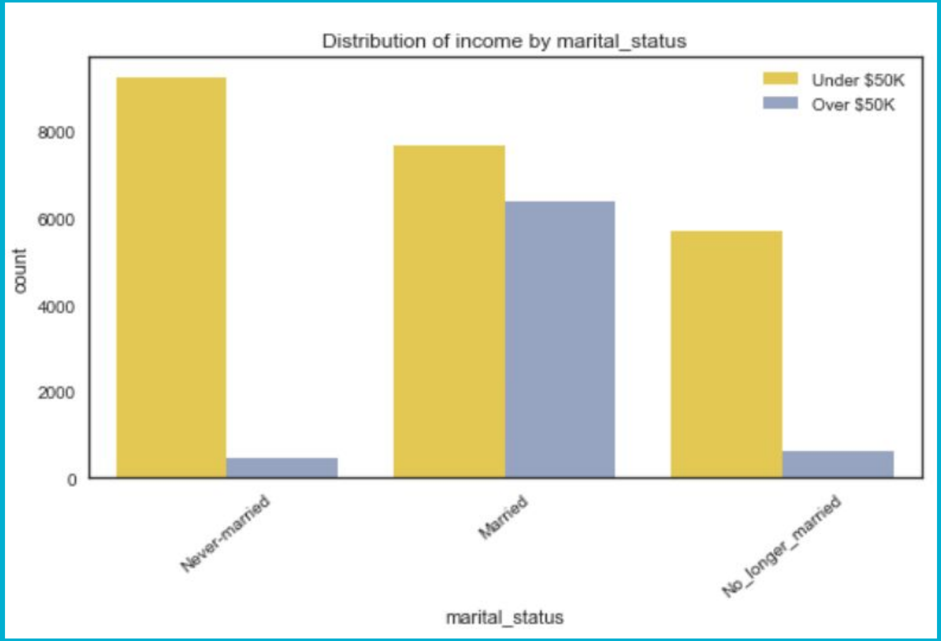


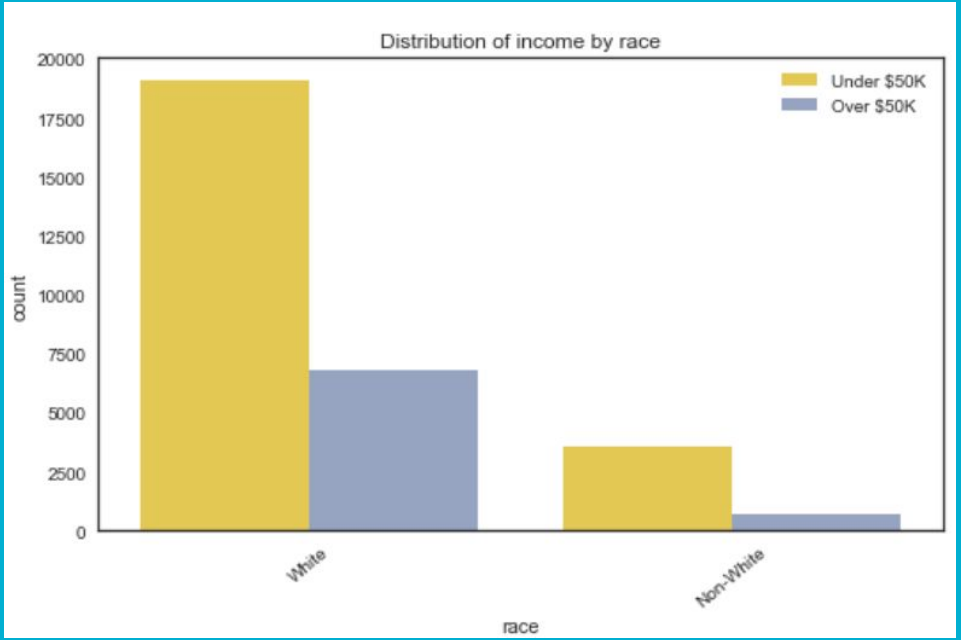
Income Level, by Sex

```
income         0      1
sex
Female      8670   1112
Male       13984   6396
5.86241470132775e-310
```

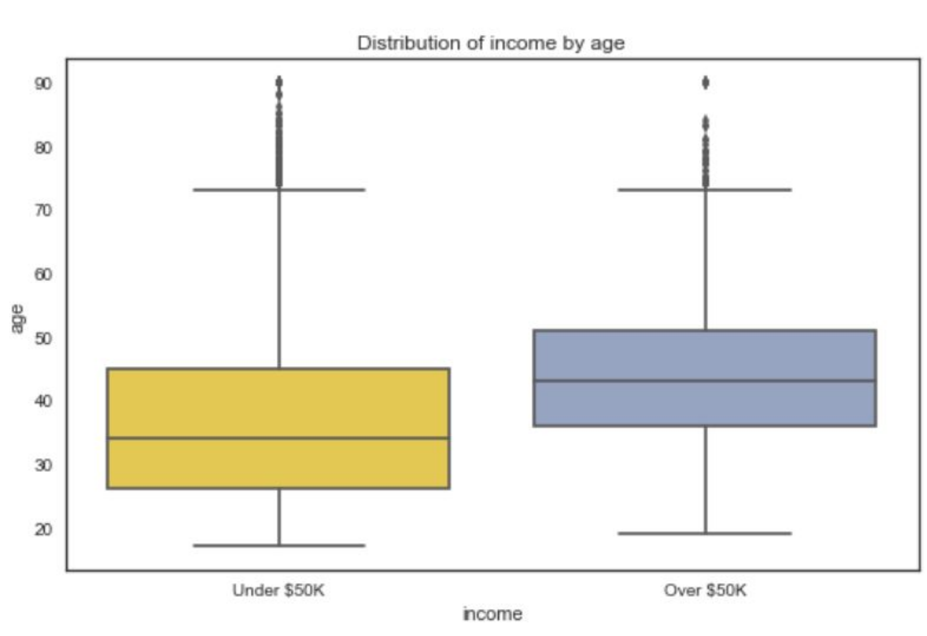


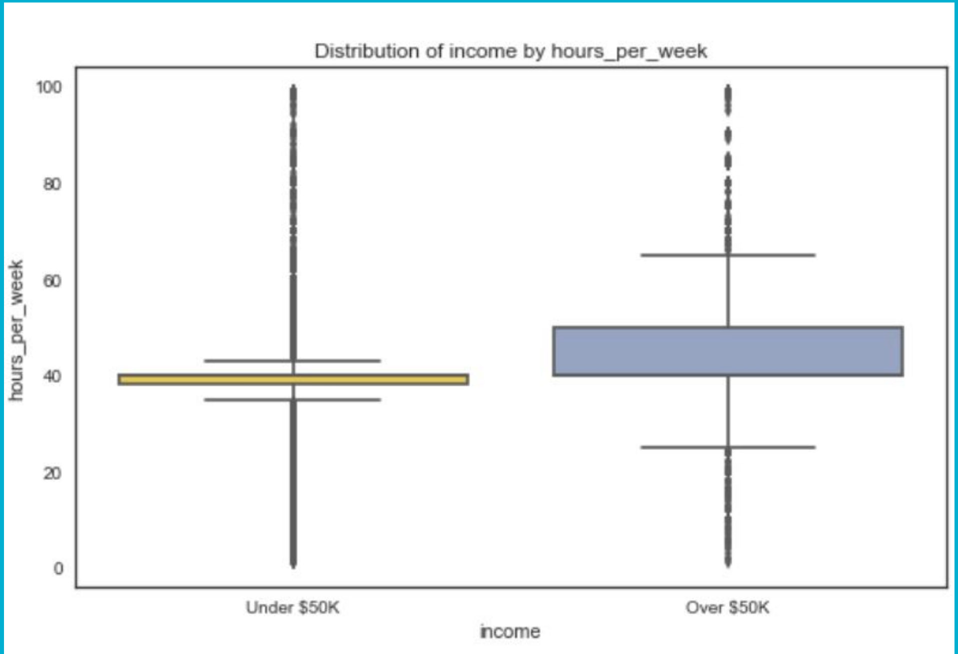
Other Interesting Insights - Marital Status and Race





Other Interesting Insights - Age and Hours per Week





Feature Engineering

Feature Engineering - Working Class

```python
# Creating new data frame with updated working class categories
inc = inc[inc['workclass'] != '?']
inc.workclass = inc.workclass.map({'Private':'Private',
                                   'Self-emp-not-inc':'Self_employed','Self-emp-inc':'Self_employed',
                                    'Local-gov':'Government', 'State-gov':'Government', 'Federal-gov':'Government',
                                    'Without-pay':'Not_working', 'Never-worked':'Not_working'})
inc.workclass.value_counts()
```

```
Private          22696
Government        4351
Self_employed     3657
Not_working         21
Name: workclass, dtype: int64
```

Name: workclass, dtype: int64

Feature Engineering - Education

```python
# Re-naming entries to generalize some of the smaller categories
inc.education = inc.education.map({'Preschool':'Dropout',
                                   '1st-4th':'Dropout',
                                   '5th-6th':'Dropout',
                                   '7th-8th':'Dropout',
                                   '9th':'Dropout',
                                   '10th':'Dropout',
                                   '11th':'Dropout',
                                   'HS-grad':'HS-grad',
                                   'Some-college':'Some-college',
                                   'Assoc-voc':'Some-college',
                                   'Assoc-acdm':'Some-college',
                                   'Bachelors':'Bachelors',
                                   'Masters':'Advanced-degree',
                                   'Prof-school':'Advanced-degree',
                                   'Doctorate':'Advanced-degree'})
inc.education.value_counts()
```

```
HS-grad            9969
Some-college       9118
Bachelors          5182
Dropout            3432
Advanced-degree    2631
Name: education, dtype: int64
```

Feature Engineering - Marital Status

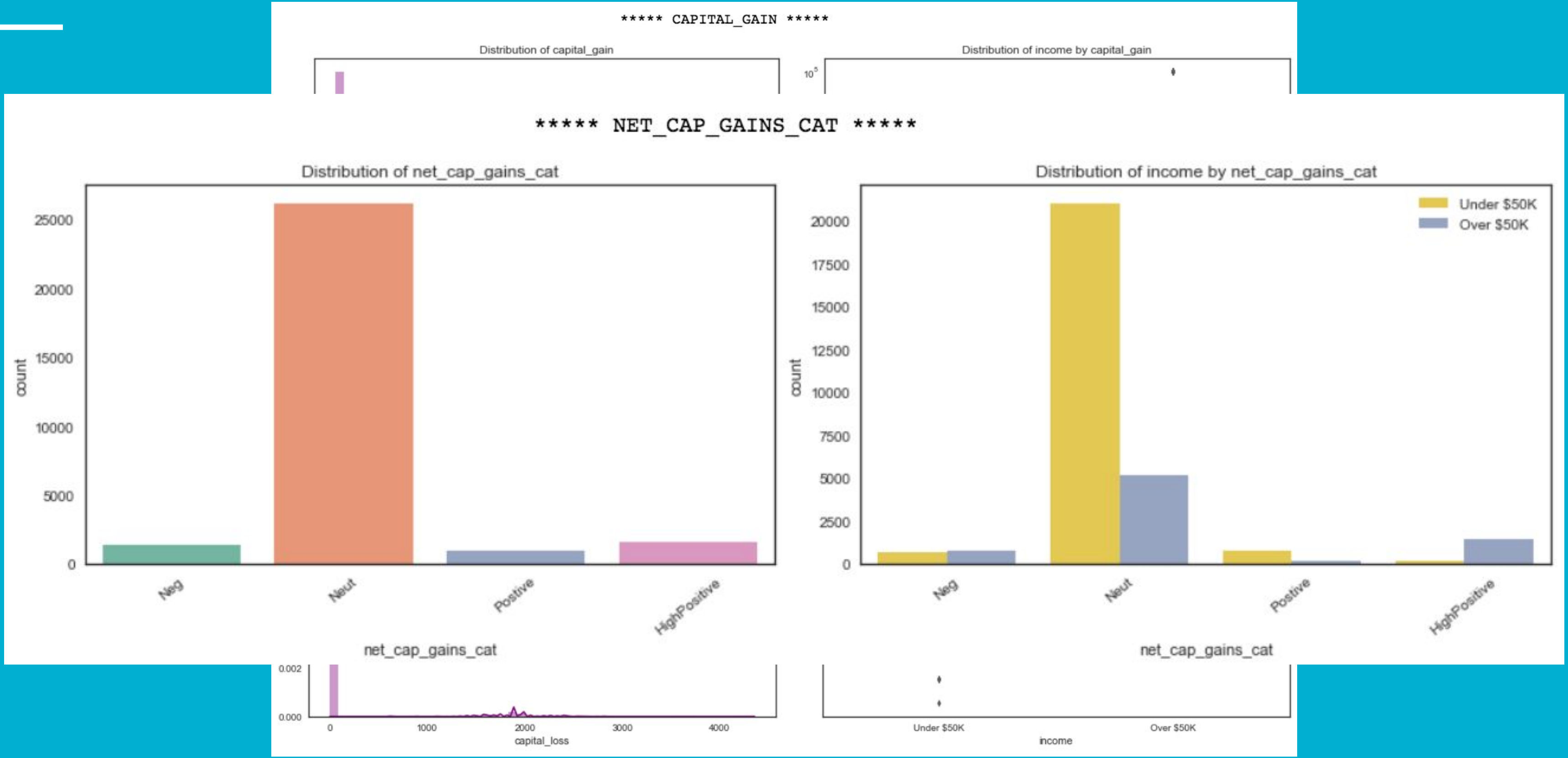
```
inc.marital_status.value_counts()
```

```
inc.marital_status = inc.marital_status.map({'Married-civ-spouse':'Married', 'Married-AF-spouse':'Married',
                                              'Divorced':'No_longer_married', 'Separated':'No_longer_married',
                                              'Married-spouse-absent':'No_longer_married', 'Widowed':'No_longer_married'
                                             'Never-married':'Never-married'})
inc.marital_status.value_counts()
```

```
Married              14361
Never-married         9917
No_longer_married     6447
Name: marital_status, dtype: int64
```

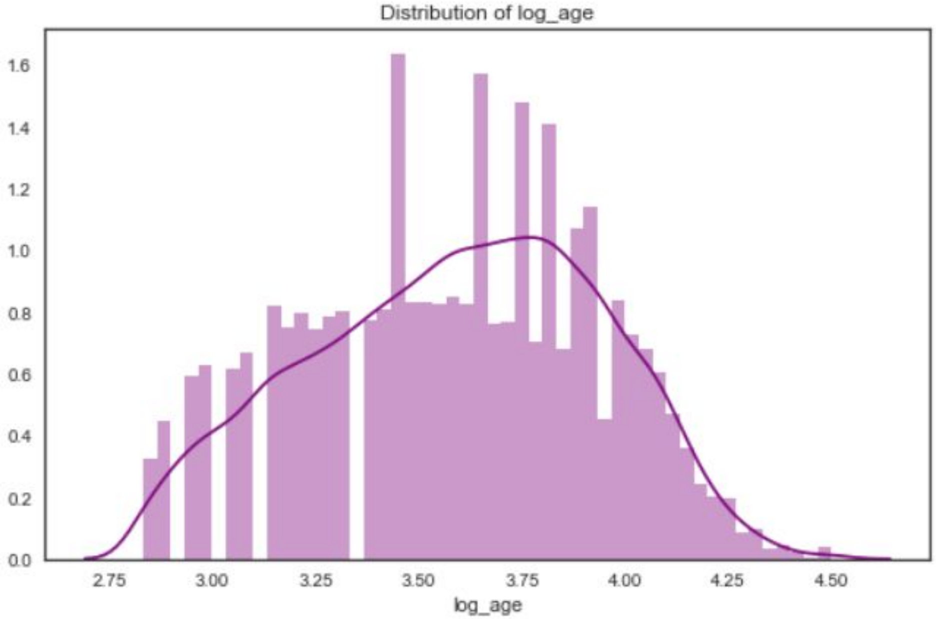
```
Married-AF-spouse            21
Name: marital_status, dtype: int64
```

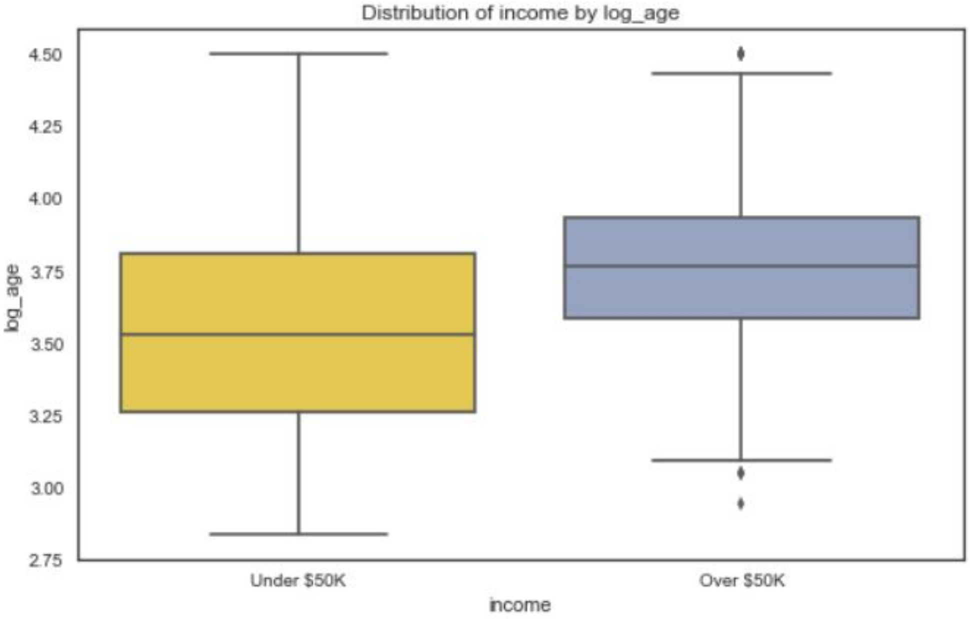
Feature Engineering - Capital Gains



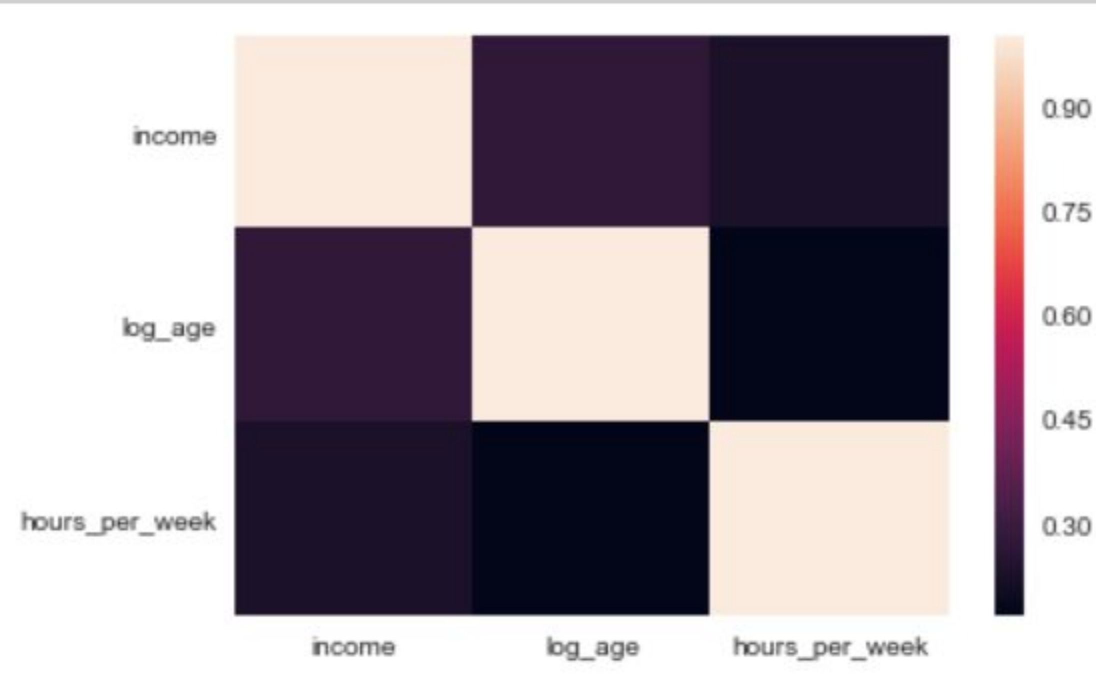
Feature Engineering - Age

***** LOG_AGE *****





Feature Engineering - Independence



Modeling

Logistic Regression

K Nearest Neighbors Classifier

Random Forest

Gradient Boosting Classifier

**All with under-sampling on training set

Logistic Regression

+ Provides probability scores
+ Robust to noise in data
+ Interpretability of odds ratios from coefficients
- Struggles with large number of categorical features

Default Settings

Accuracy: 85.28 (+/- 1)%

ROC Score: 0.9048 (+/- 0.01)

Optimized the regularization parameter, solver algorithm, and L1 (LASSO) vs. L2 (Ridge) regression penalties

Accuracy: 85.31 (+/- 1)%

ROC Score: 0.9049 (+/- 0.01)

K Nearest Neighbors Classifier

- \+ Classifies based on closeness of other known observations
- \+ Lazy learning responds to changes in inputs
- \- Longer computation time in test set
- \- High dimensionality reduces effectiveness

Default Settings

Accuracy: 82.4 (+/- 2)%

ROC Score: 0.8453 (+/- 0.03)

Optimized the number of neighbors used to compare and classify points

Accuracy: 82.9 (+/- 2)%

ROC Score: 0.8751 (+/- 0.03)

Random Forest

\+ Typically high performer
\+ Guards against overfitting
\+ Provides feature importance
\- Black box
\- Not able to predict outside sample
\- Optimization is computationally expensive

Default Settings

Accuracy: 83.0 (+/- 1)%

ROC Score: 0.8610 (+/- 0.03)

Optimized the number of estimators, minimum samples split, maximum depth

Accuracy: 85.12 (+/- 2)%

ROC Score: 0.9055 (+/- 0.01)

Gradient Boosting Classifier

+ Minimizes loss function
+ Subsampling and learning rate help prevent overfitting
+ Robust to outliers and missing data
- Can be prone to overfitting
- Optimization can be computationally expensive

Default Settings

Accuracy: 85.7 (+/- 2)%

ROC Score: 0.9093 (+/- 0.01)

Optimized the minimum samples split, minimum samples per leaf, maximum depth, number of features considered, fraction of observations used to subsample, and number of estimators

Accuracy : 85.4 (+/- 2)%

ROC Score: 0.9097 (+/- 0.01)

Overall Model Analysis

	Model	Mean_Accuracy_Train	Mean_Accuracy_Test	ROC_AUC_Score
3	Gradient_Boost	0.825457	0.856653	0.909662
2	Random_Forest	0.815599	0.851081	0.905456
0	Logistic_Regression	0.816308	0.853203	0.904994
1	KNN	0.793836	0.829469	0.845342

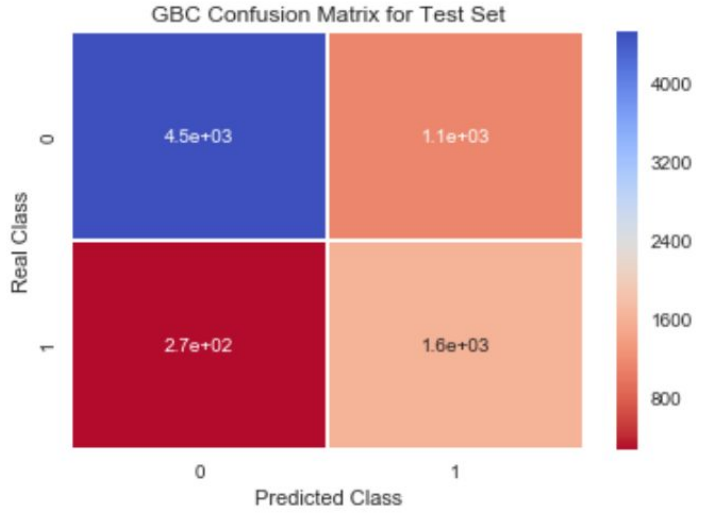
Error Analysis - Gradient Boosting

Precision (positive outcomes correctly predicted) was higher for predicting incomes under $50,000

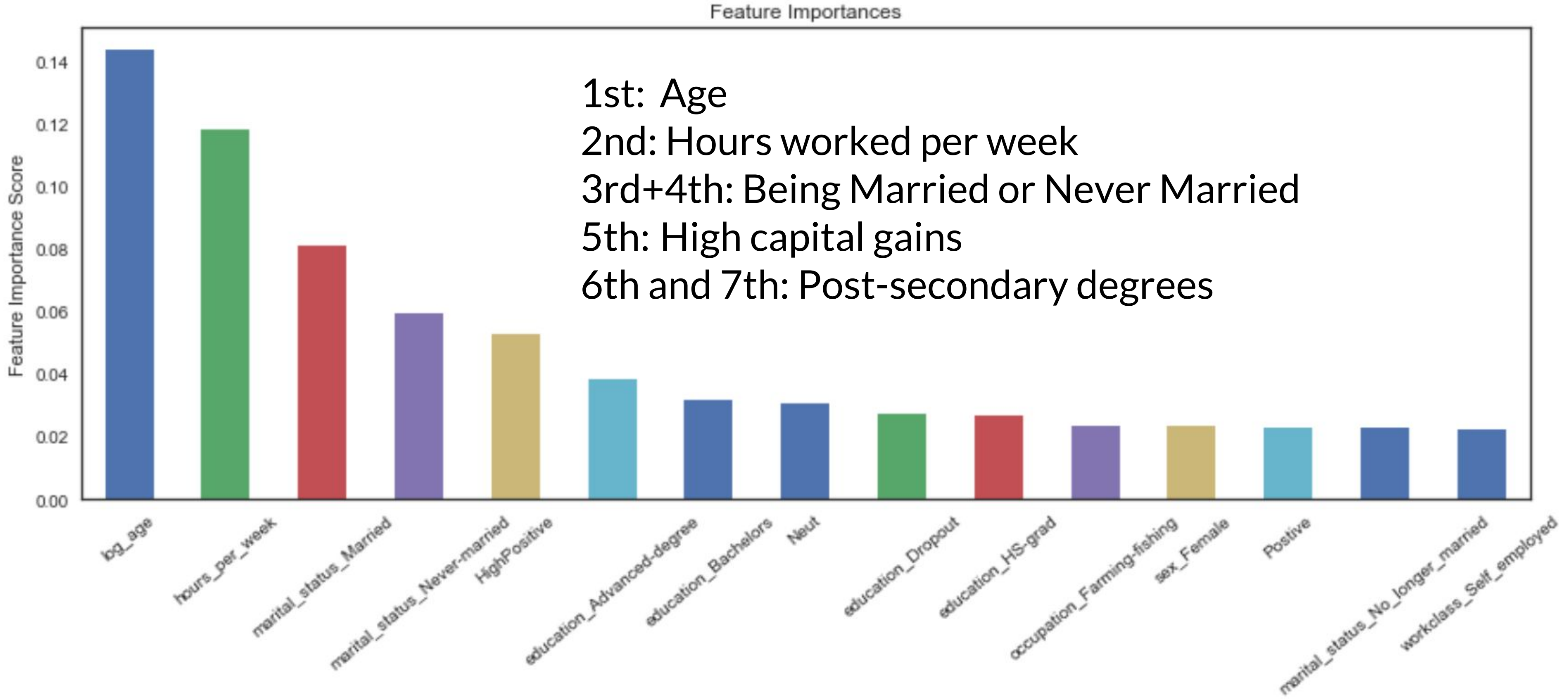
Recall (actual positives correctly identified) was higher for predicting incomes over $50,000

F1-score (weighted average of precision and recall) was higher for predicting incomes under $50,000

	precision	recall	f1-score	support
Under $50K	0.94	0.80	0.86	5662
Over $50K	0.58	0.86	0.70	1879
avg / total	0.85	0.81	0.82	7541

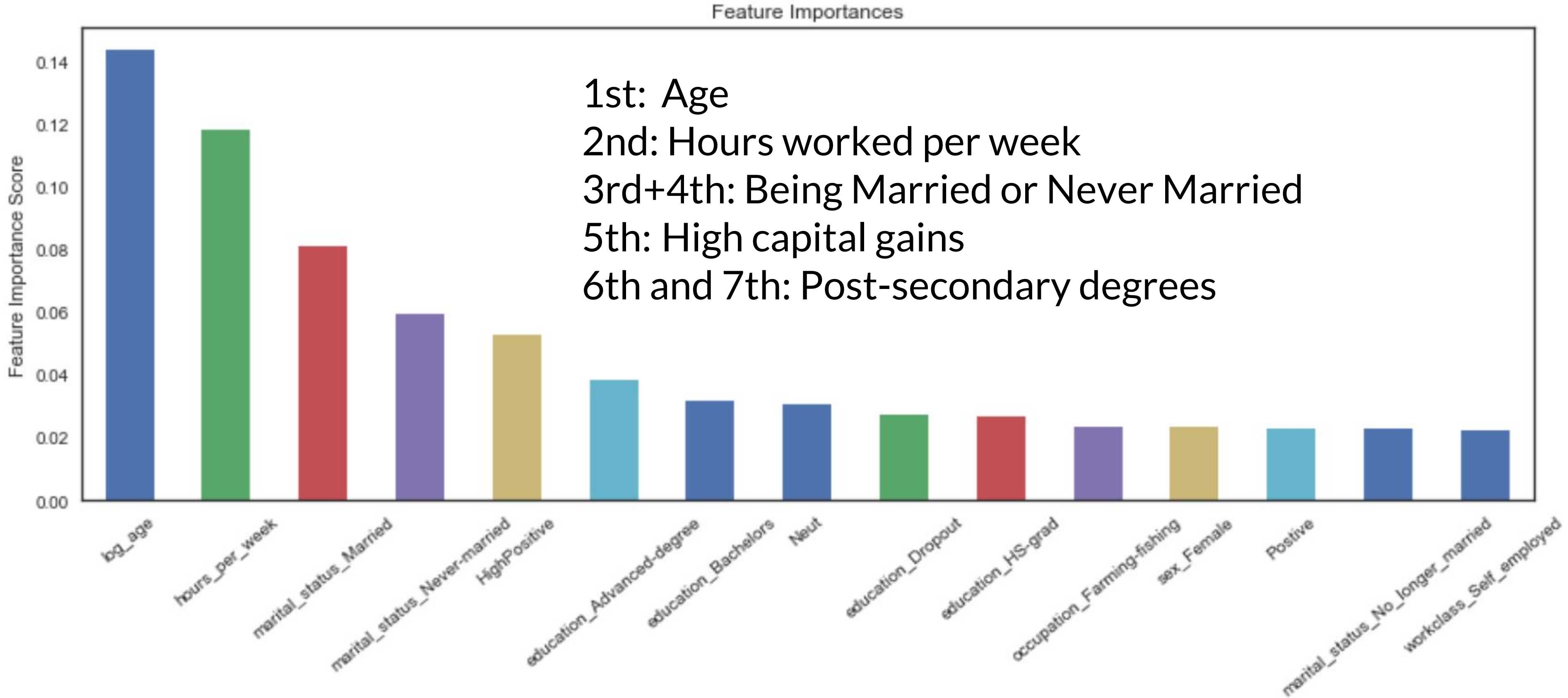


Model Interpretation – Feature Importances

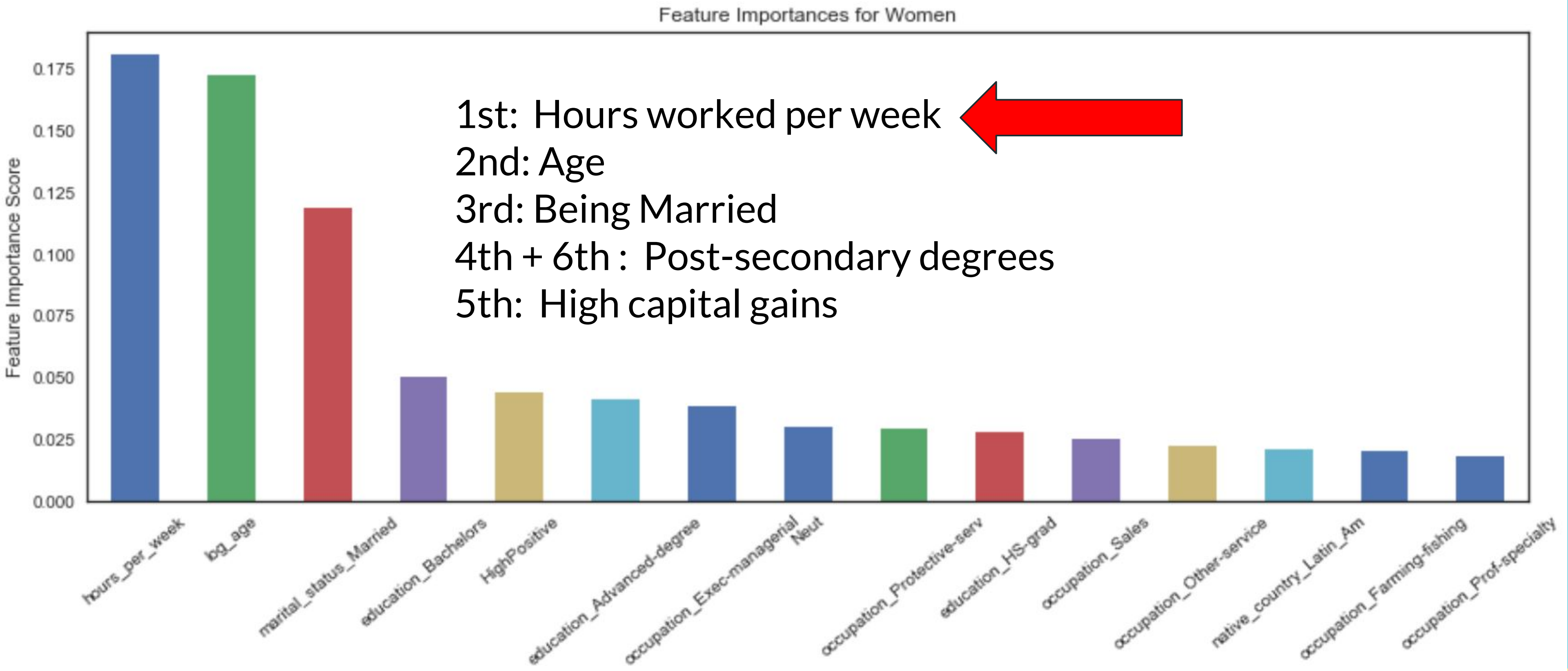


1st: Age
2nd: Hours worked per week
3rd+4th: Being Married or Never Married
5th: High capital gains
6th and 7th: Post-secondary degrees

Model Interpretation – Feature Importances



Gender Pay Gap - Model with only Females



1st: Hours worked per week
2nd: Age
3rd: Being Married
4th + 6th : Post-secondary degrees
5th: High capital gains

Conclusion

- While people can't change their age (without waiting), they can change all of the other demographic indicators that are indicative of earning more money

- Demographic Indicator / Characteristics of Individual

Demographic Indicator	Characteristics of Individual
Being married	Interpersonal skills, commitment
Number of hours worked per week	Grit, persistence, passion
High capital gains	Risk/reward
Bachelor's degree and other advanced degrees	Critical thinking skills, discipline

Conclusion

- While people can't change their age (without waiting), they can change all of the other demographic indicators that are indicative of earning more money

Demographic Indicator	Characteristics of Individual
○ Being married	Interpersonal skills, commitment
○ Number of hours worked per week	Grit, persistence, passion
○ High capital gains	Risk/reward
○ Bachelor's degree and other advanced degrees	Critical thinking skills, discipline

Final Thoughts

- Opportunities for further exploration
  - How have these indicators changed since 1994?
  - How do these indicators compare to the income levels of other developed countries?
  - What indicators are most important for predicting if minority races earn higher incomes?

Thanks!

[Link to codebook]