Marriage Success Predictor

Supervised Machine Learning Algorithm

Overview

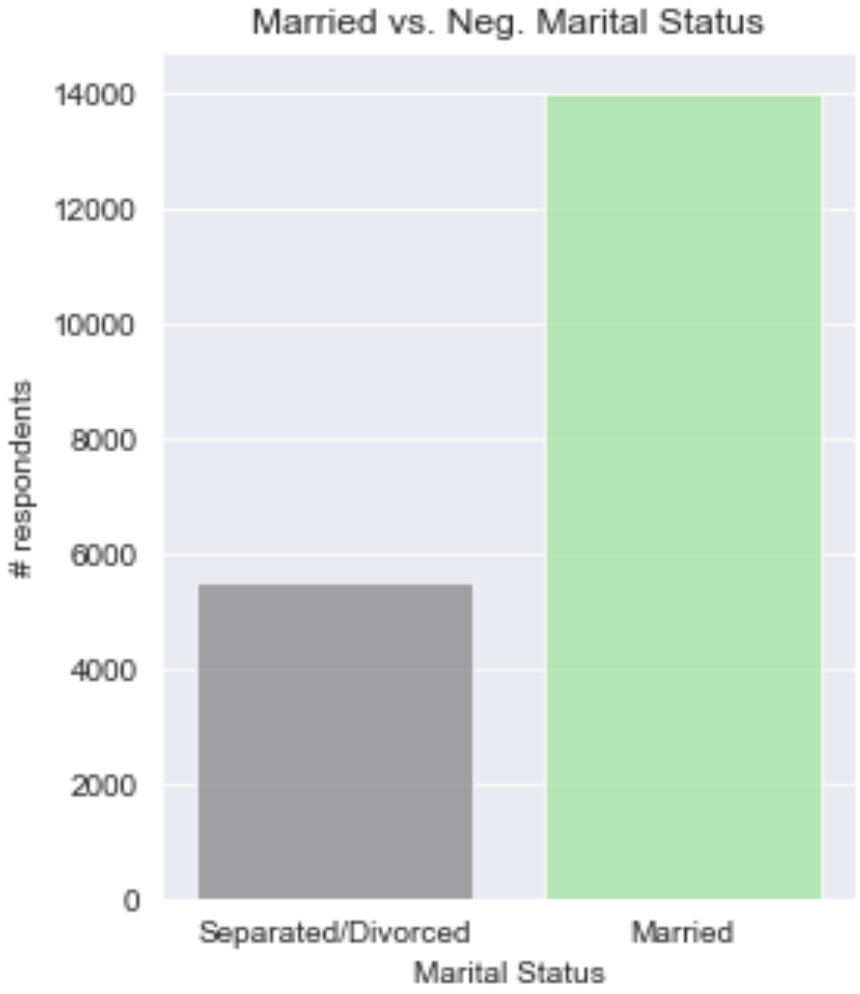
- To predict if a marriage will end in divorce
  - Binary classification
- Examine individual respondents from census survey
- Included 32,000 records
  - Removed those who were never married or widowed
  - Left with about 20k record
- Several models with similar accuracy and F-1 scores
  - Settled on an XGBoosted Tree model for F-1 score

Null Accuracy

Assuming every marriage ends in divorce:

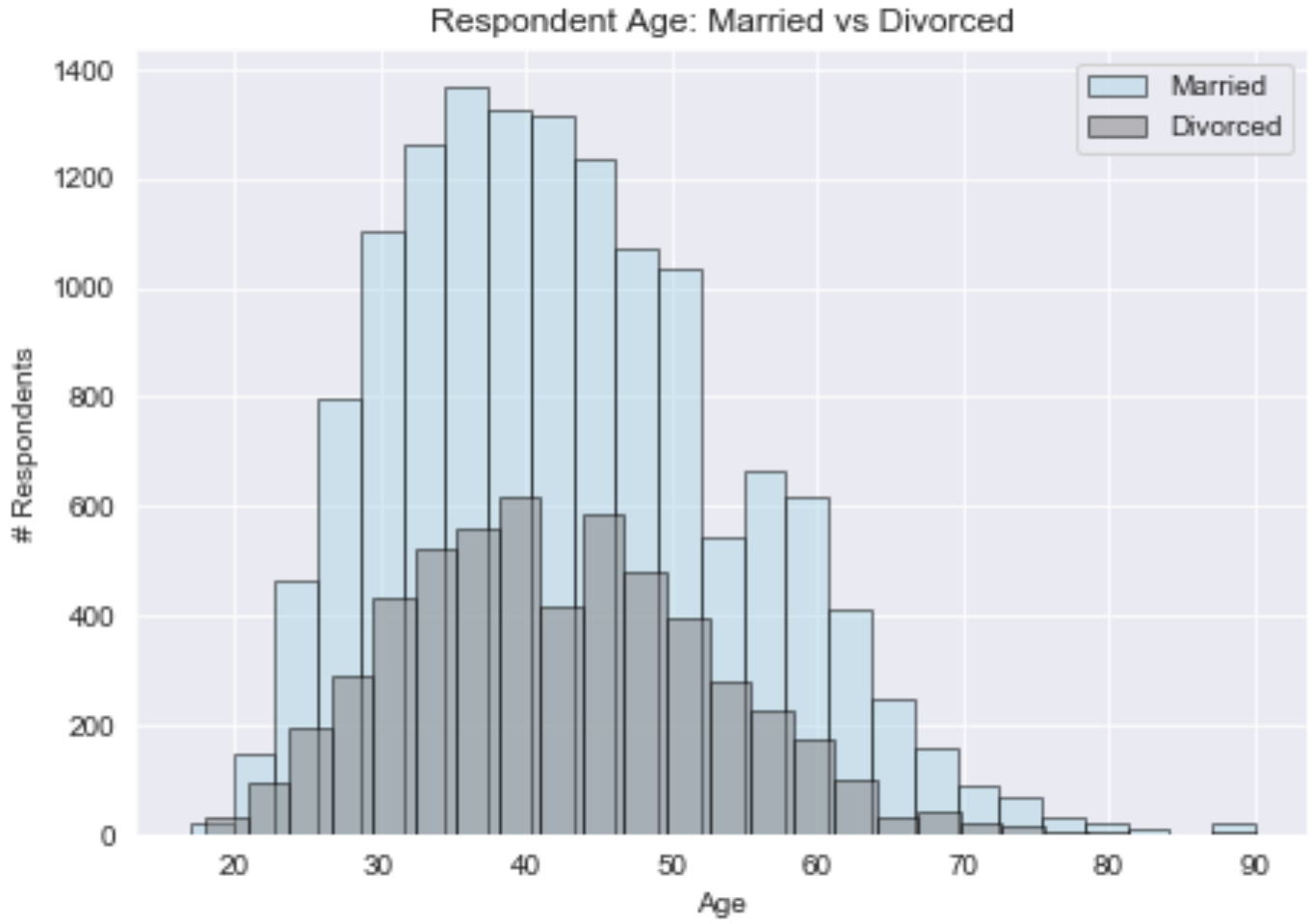
Null Accuracy = 28.2%

US Divorce Rate is closer to 42-45%
(survey limitations)



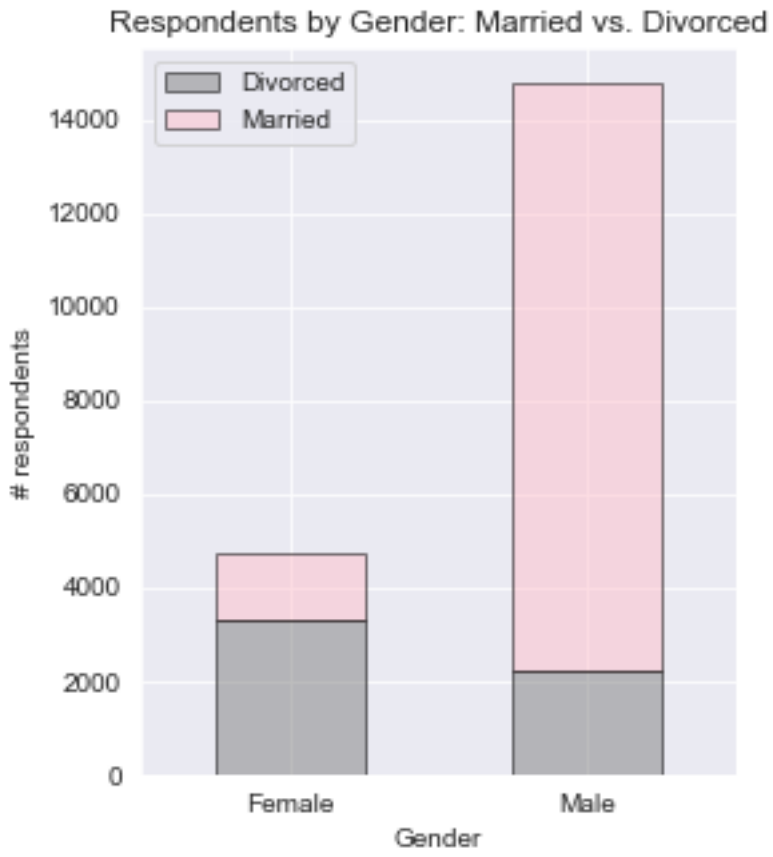
Data Features

- **Age**
- Gender
- Native Country
- Race
- Education
- Occupation Type
- Hours Worked
- Income
- Capital Gain/Loss



Data Features

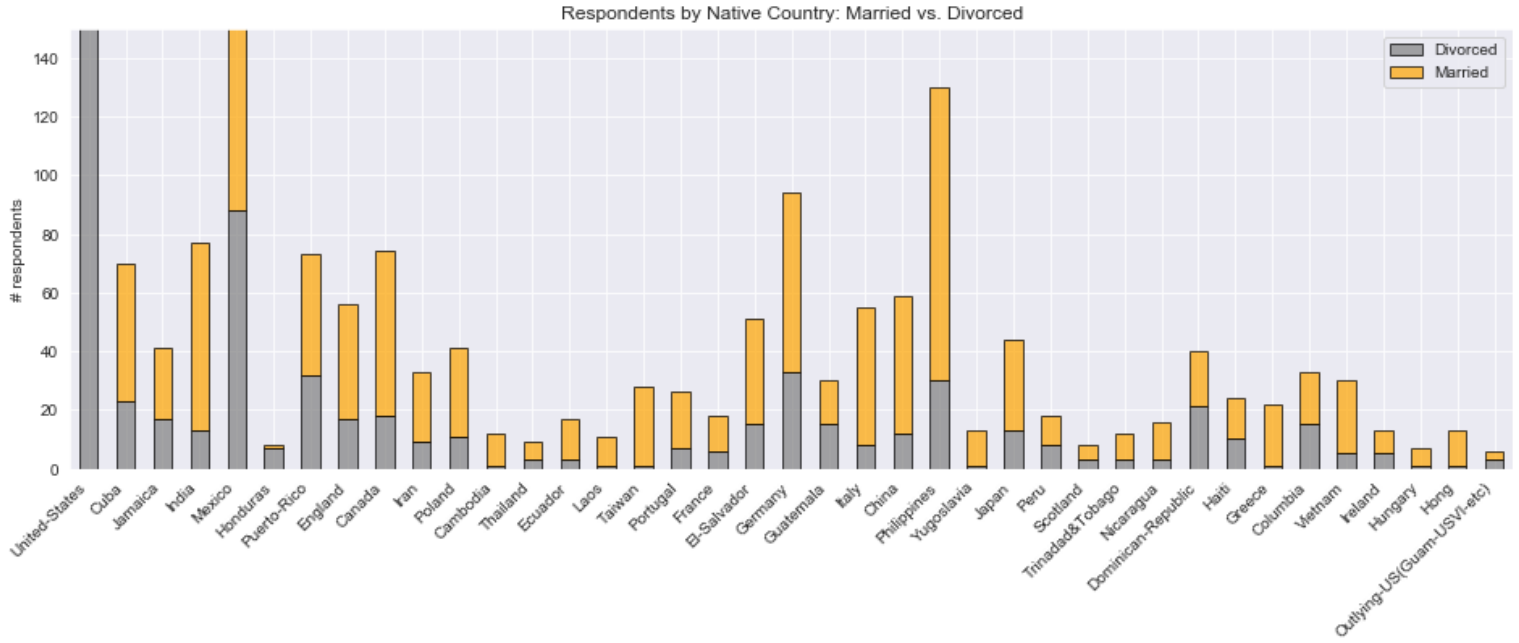
- Age
- **Gender**
- Native Country
- Race
- Education
- Occupation Type
- Hours Worked
- Income
- Capital Gain/Loss

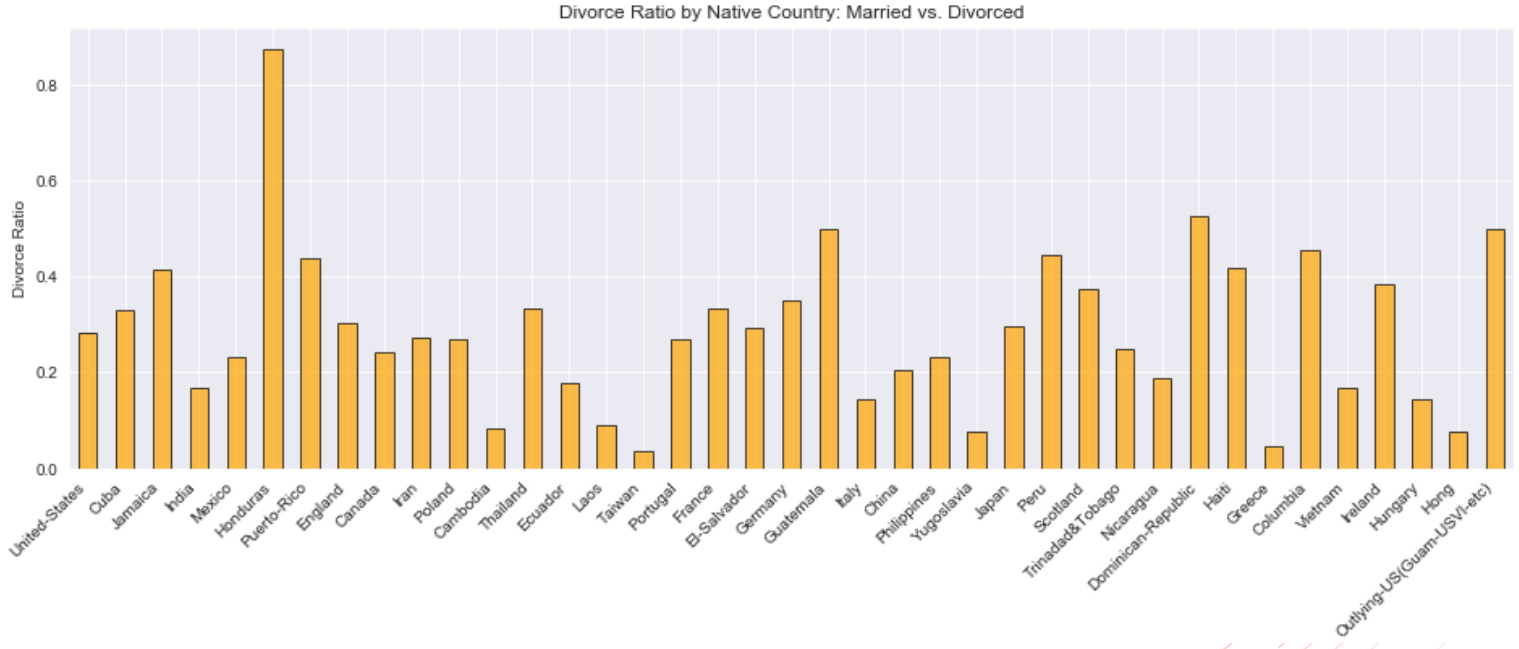


F: 69.04% M: 15.09%

Data Features

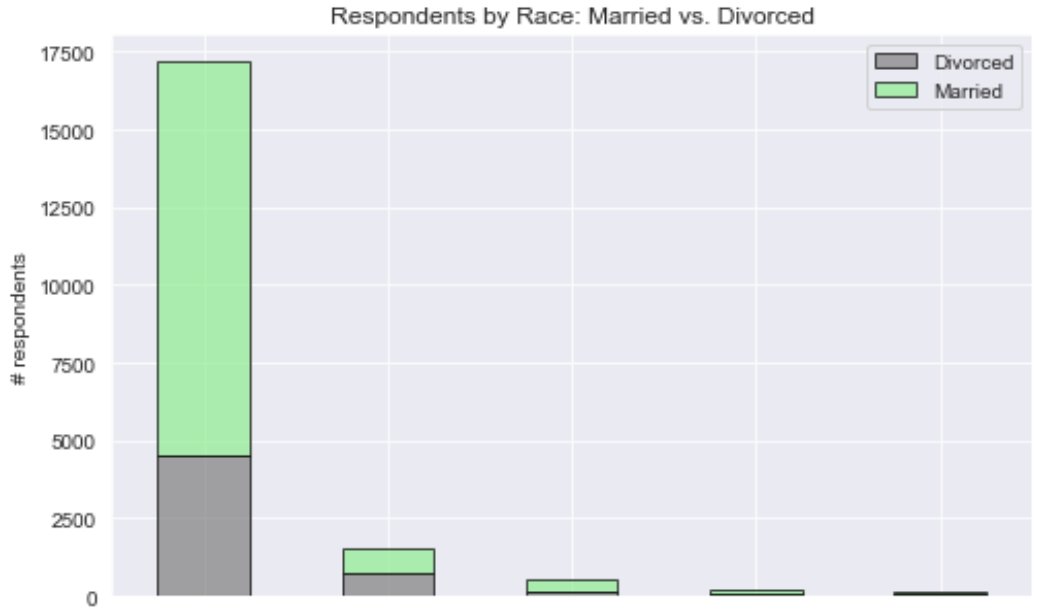
- Age
- Gender
- **Native Country**
- Race
- Education
- Occupation Type
- Hours Worked
- Income
- Capital Gain/Loss

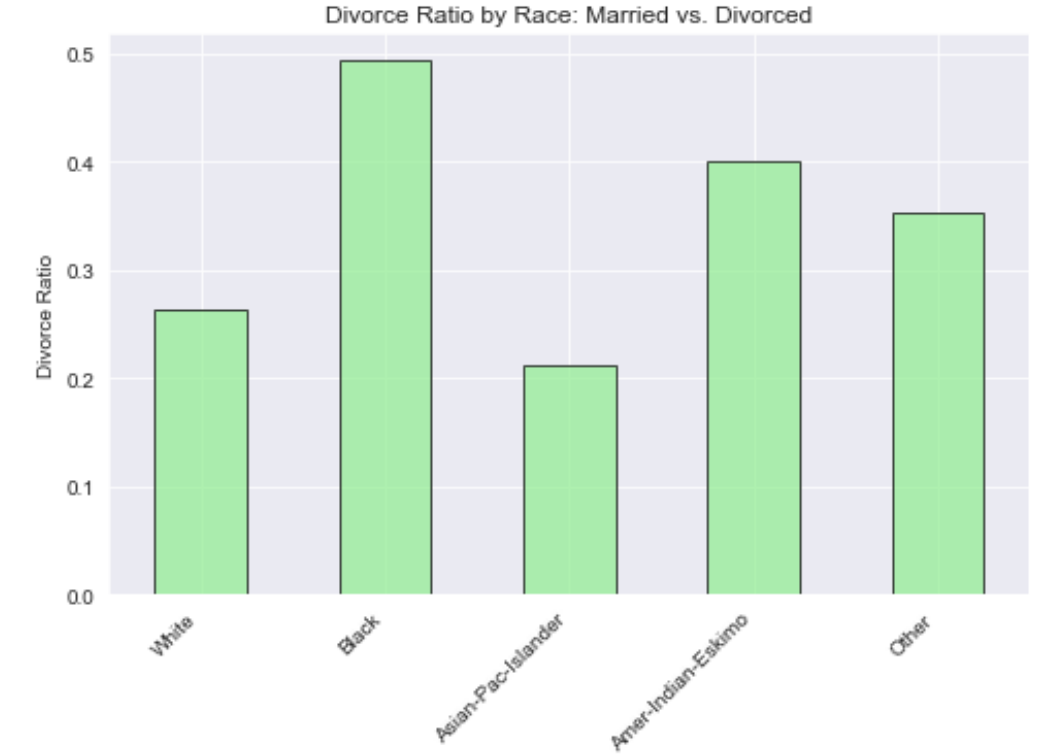




Data Features

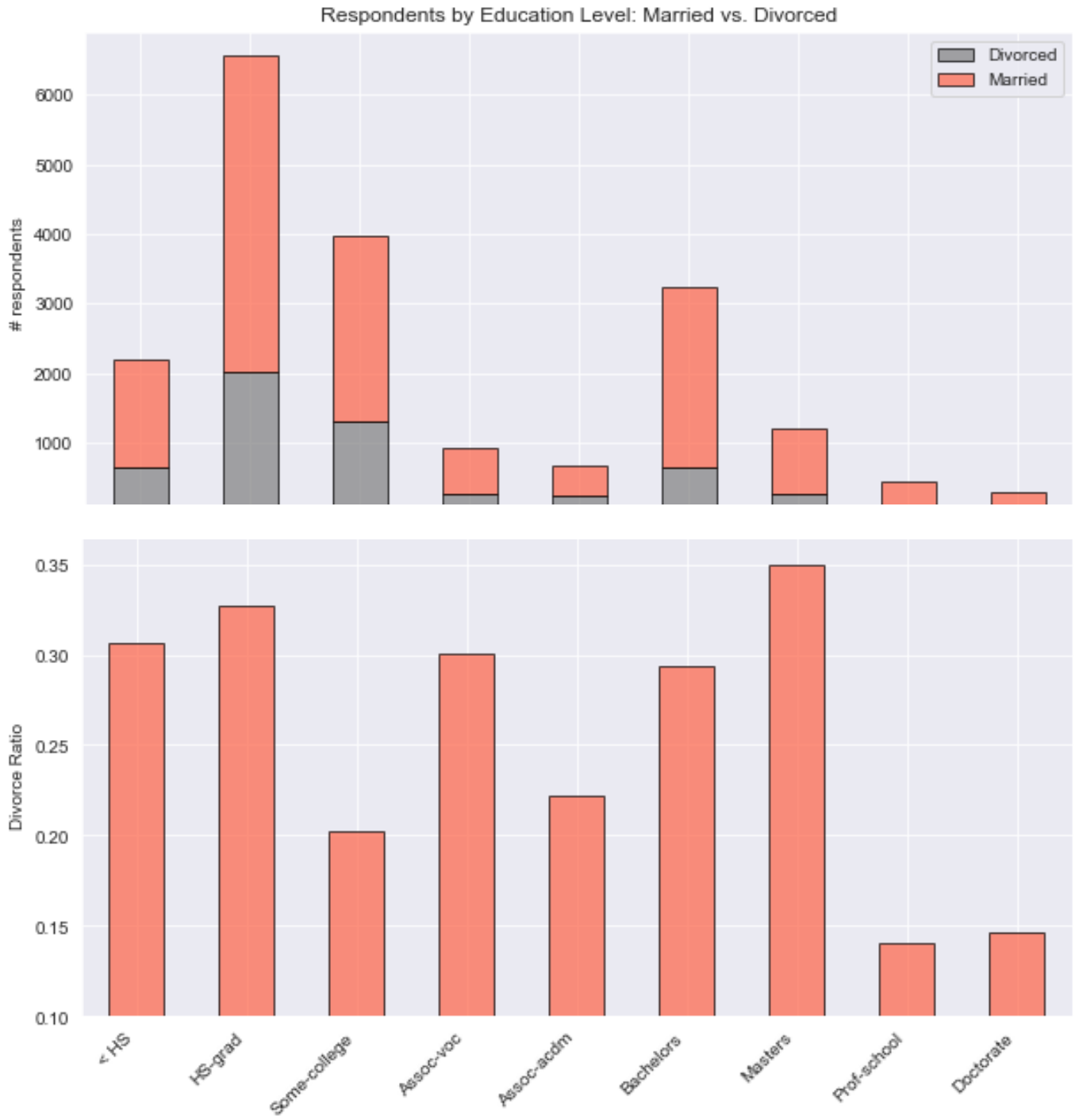
- Age
- Gender
- Native Country
- **Race**
- Education
- Occupation Type
- Hours Worked
- Income
- Capital Gain/Loss





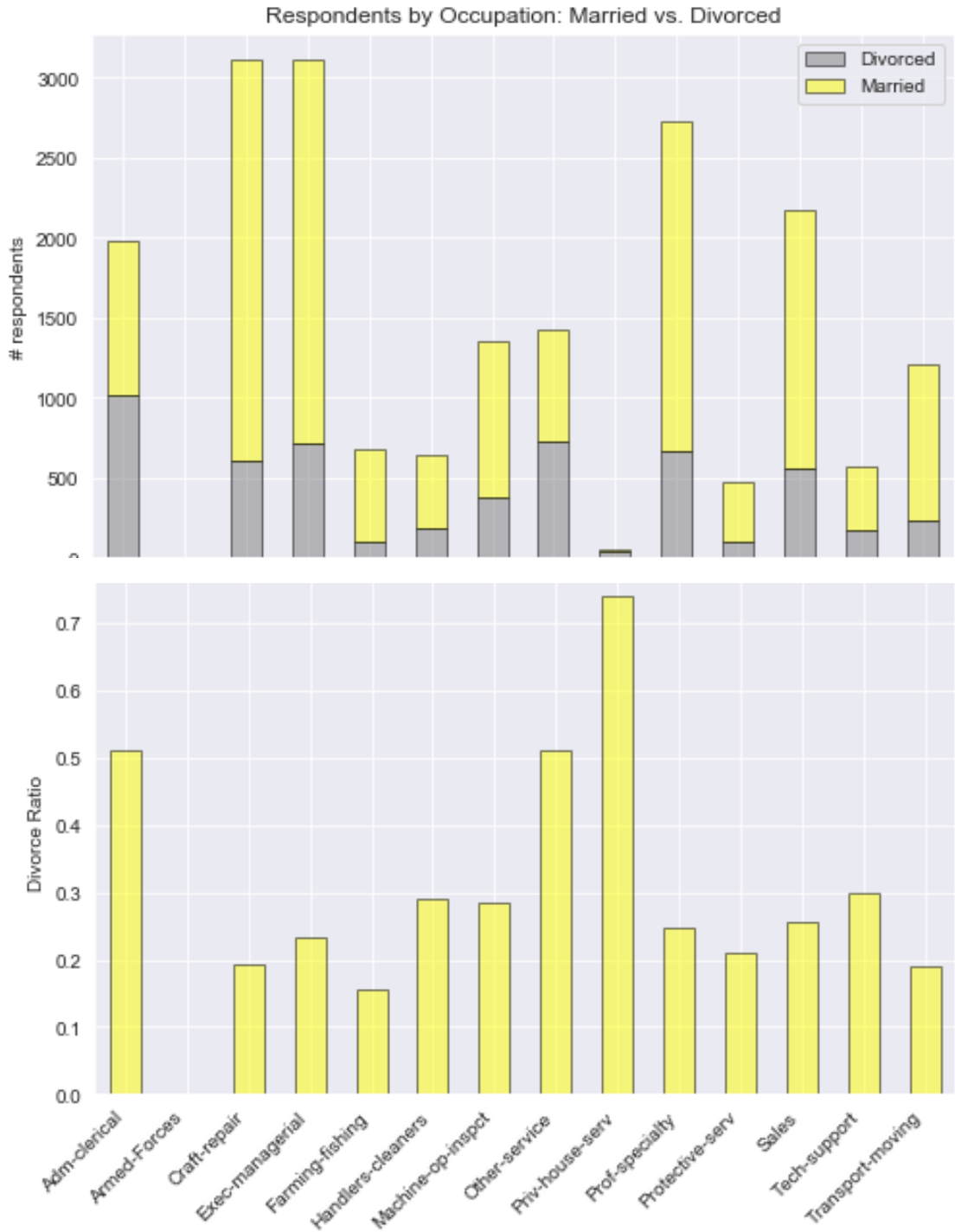
Data Features

- Age
- Gender
- Native Country
- Race
- **Education**
- Occupation Type
- Hours Worked
- Income
- Capital Gain/Loss



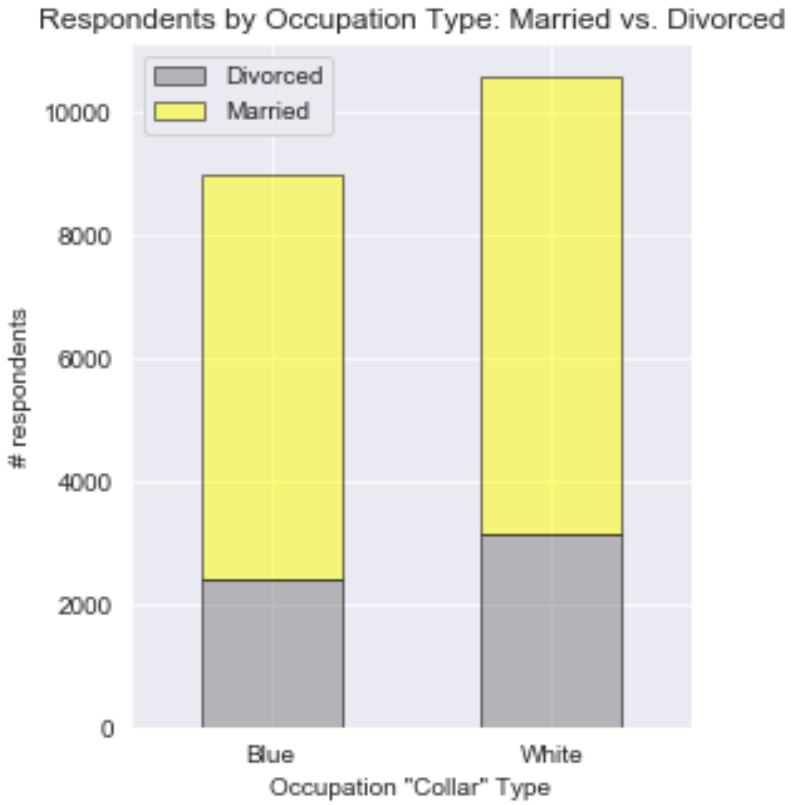
Data Features

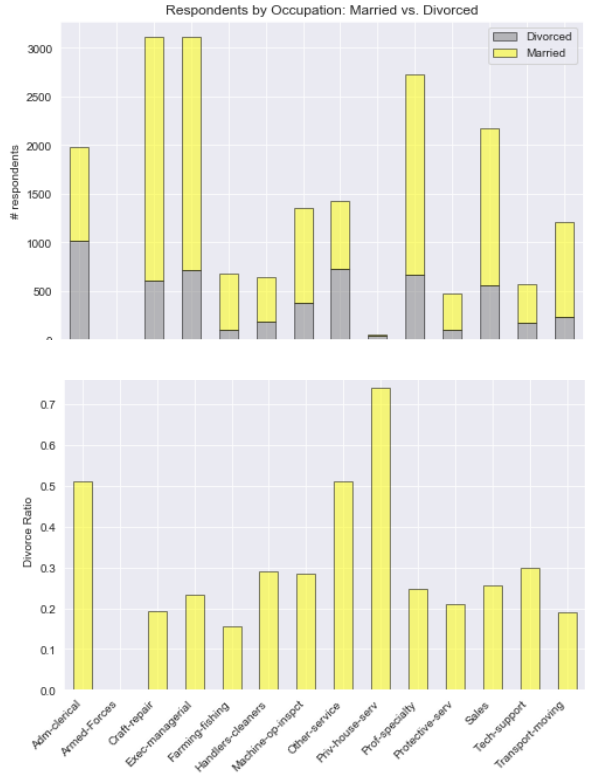
- Age
- Gender
- Native Country
- Race
- Education
- **Occupation Type**
- Hours Worked
- Income
- Capital Gain/Loss



Data Features

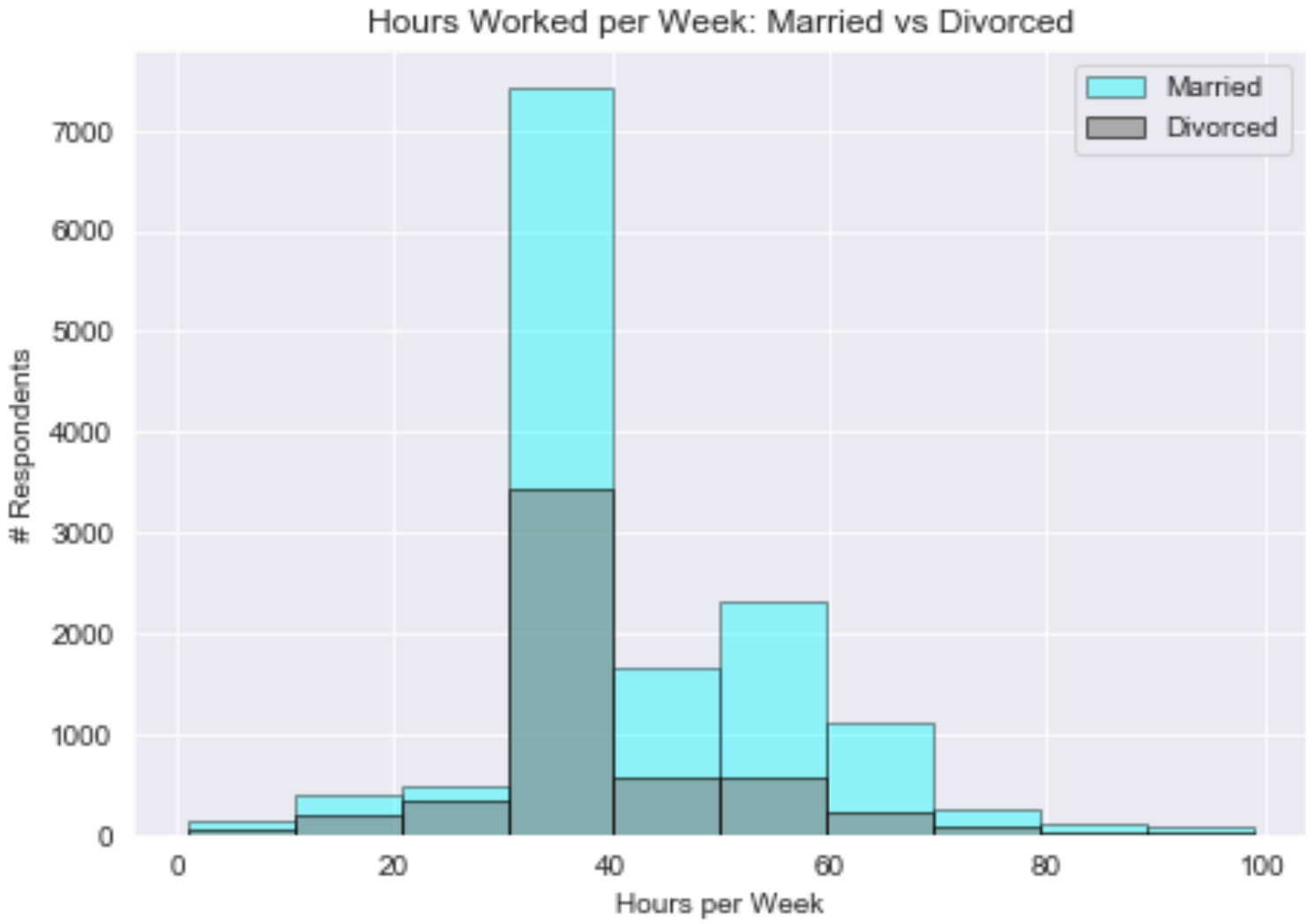
- Age
- Gender
- Native Country
- Race
- Education
- **Occupation Type**
- Hours Worked
- Income
- Capital Gain/Loss





Data Features

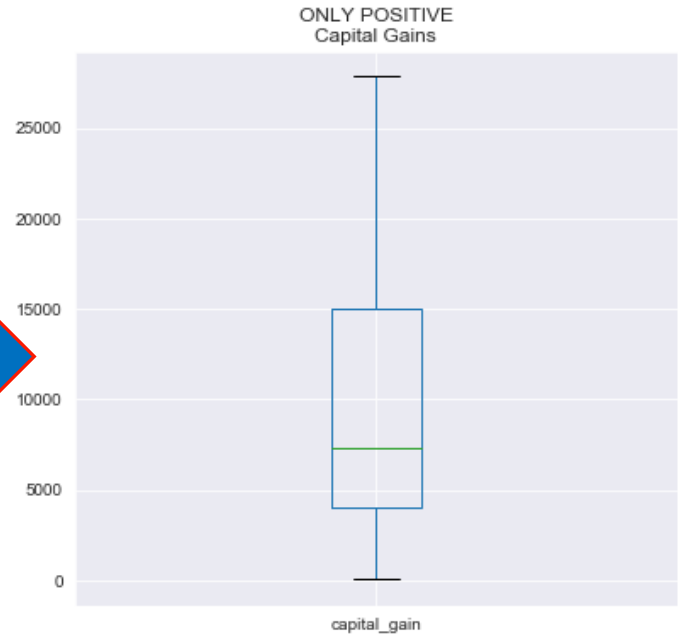
- Age
- Gender
- Native Country
- Race
- Education
- Occupation Type
- **Hours Worked**
- Income
- Capital Gain/Loss

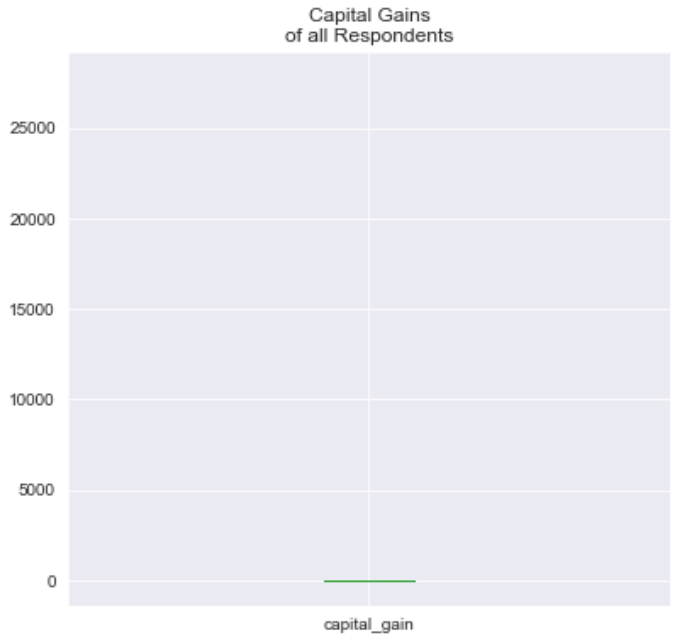


Data Features

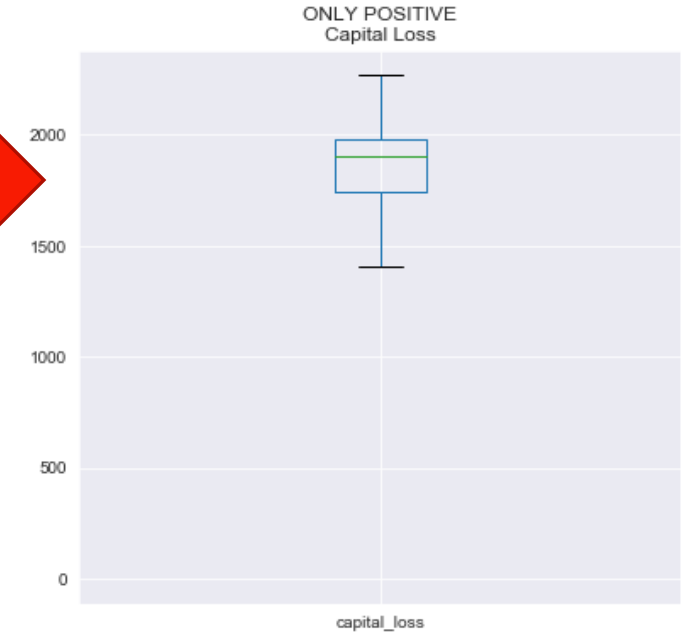
- Age
- Gender
- Native Country
- Race
- Education
- Occupation Type
- Hours Worked
- **Income**
- **Capital Gain/Loss**

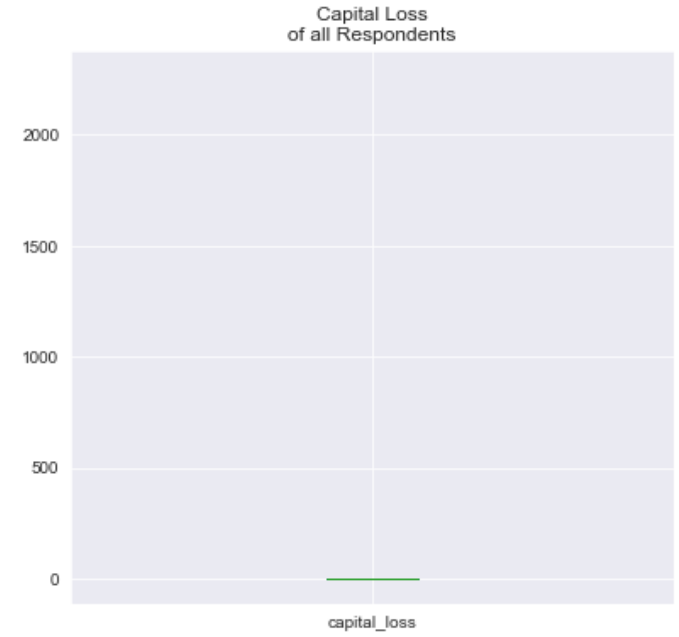
10.5%



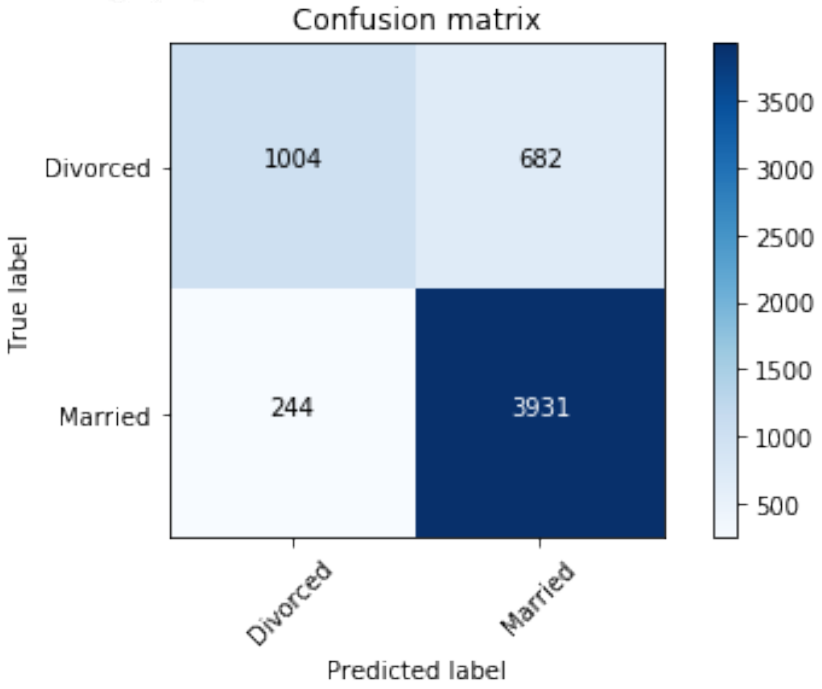


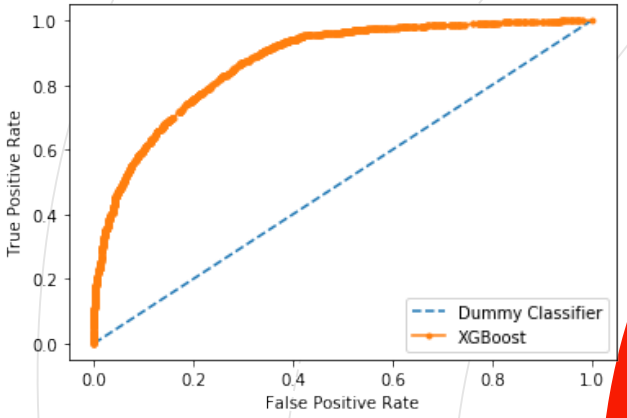
5.7%





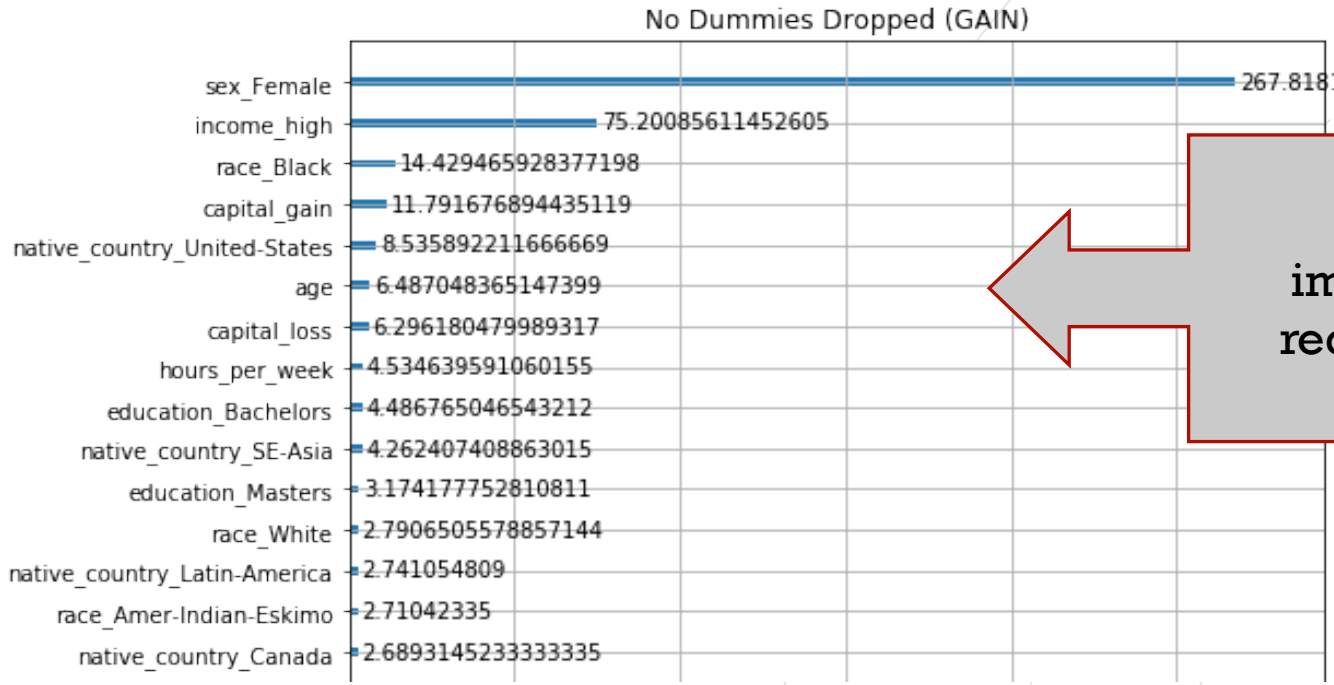
GBoosted Decision Tree Ensemble



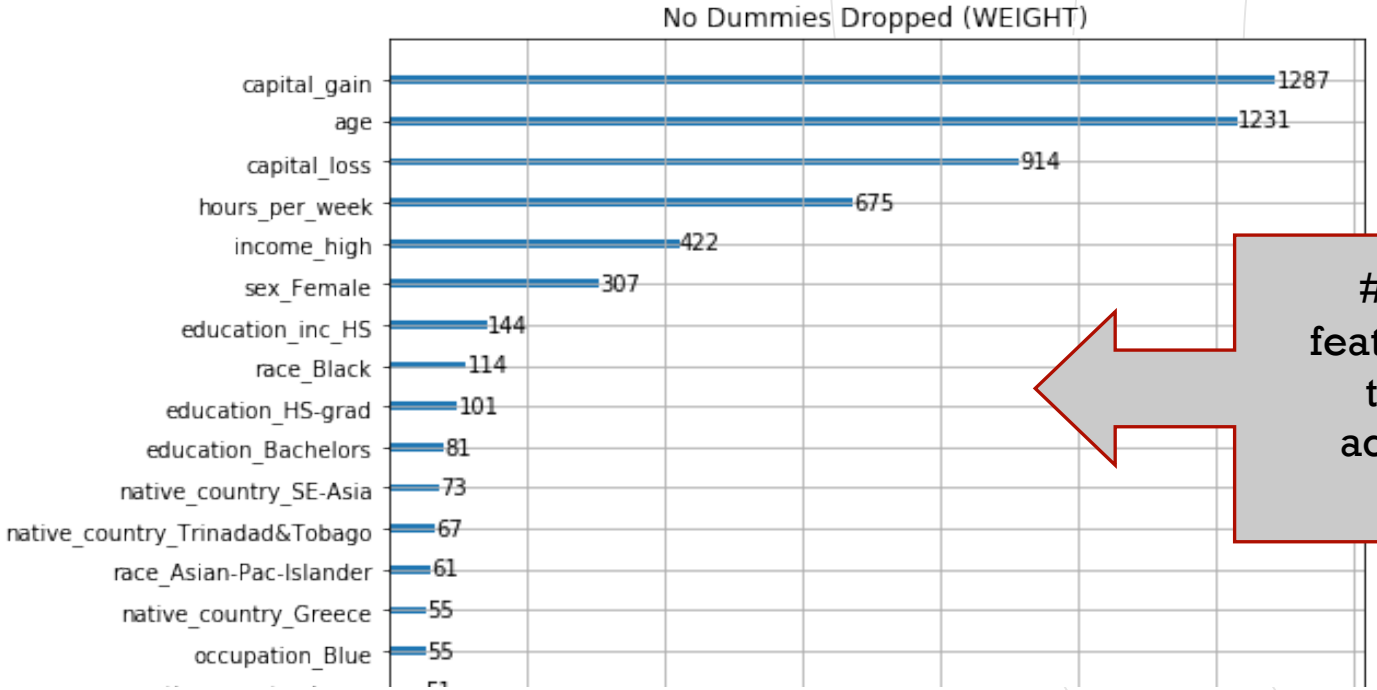


	precision	recall	f1-score	support
0	0.80	0.60	0.68	1686
1	0.85	0.94	0.89	4175
accuracy			0.84	5861
[illegible]				5861
weighted avg	0.84	0.84	0.83	5861

Model



Gini impurity reduction



Feature Importance

times feature used to split across all trees

Conclusions

- What should I focus on to avoid a divorce?
  - Work > 40hrs/week
    - prove value to your spouse?
  - Invest your money
    - for better or worse?
  - Make sure you're above the $85k threshold

What's Next?

- Clustering
  - To test hypotheses from EDA (occupations with higher divorce rates)
- SHAP – individualized interpretation of feature importance
- Source current data
- Source length of marriage
- Source continuous income data (what is the cutoff), geographic location (state by state), # kids in household, marriage #