Predicting Wages From Census Data

Jurgen Arias, Djuwita Carney,
Larry Curran, Eileen Palmer

Objective

Build a classification model to predict whether a person's income is **more** or **less** than $50,000.

Constraints

- Number of features (20 features)
- Time (7 hours)

Original Features

- Age
- Work class
- Final weight
- Education
- Education-num
- Marital Status
- Occupation
- Relationship
- Sex
- Capital gain
- Capital loss
- Hours per week
- Native country

Original Features

- Age
- Work class
- Final weight
- ~~Education~~
- Education-num
- Marital Status
- Occupation
- Relationship
- Sex
- Capital gain
- Capital loss
- Hours per week
- Native country

Original Features

- Age
- Work class
- Final weight
- ~~Education~~
- Education-num
- Marital Status
- Occupation
- Relationship
- Sex
- Capital gain
- Capital loss
- Hours per week
- Native country

Original Features

★= Dummies

- Age
- ★ Work class
- Final weight
- ~~Education~~
- Education-num
- ★ Marital Status
- Occupation

- ★ Relationship
- ★ Sex
- Capital gain
- Capital loss
- Hours per week
- ★ Native country

Logistic Regression

Gaining Insights Into Our Features

With all features...
Accuracy = 80%

Determined
Top 20 features

More Feature Engineering

Polynomial Features

Strongest correlations from original features:

Age & Education

Strongest coefficients from logistic regression:

Marital Status: Civic Spouse & Marital Status: Never Married

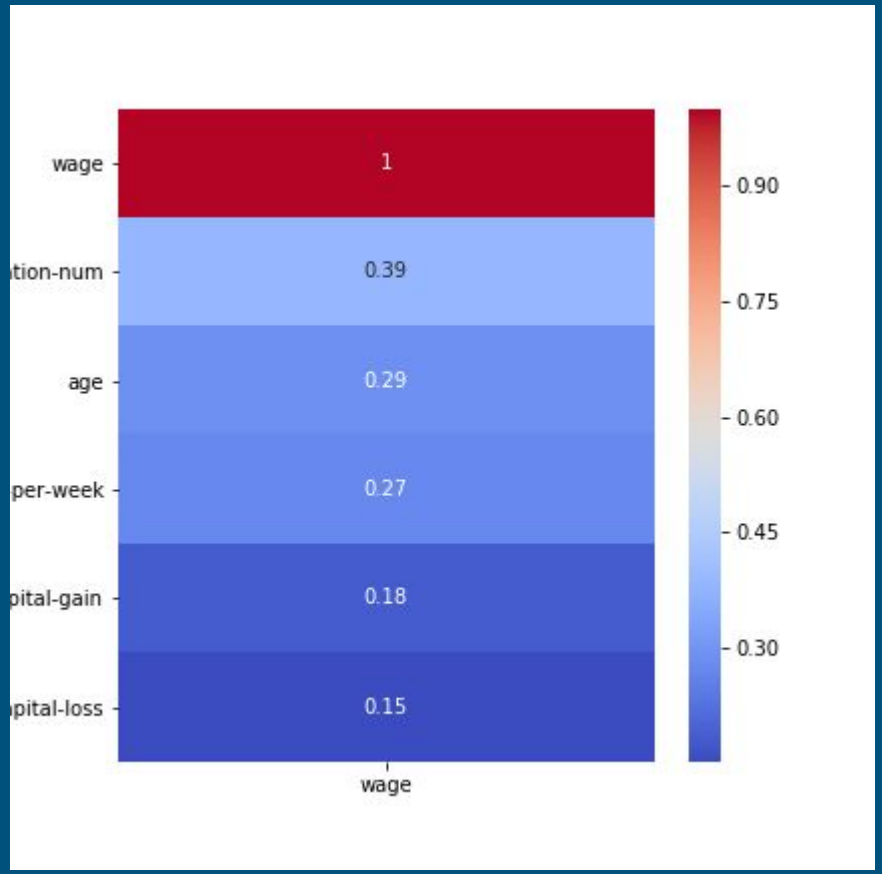
Top 20 Features

```
['relationship_ Own-child',
 'relationship_ Wife',
 'relationship_ Other-relative',
 'marital-status_ Married-civ-spouse',
 'marital-status_ Married-civ-spouse^2',
 'sex_ Male',
 'relationship_ Unmarried',
 'marital-status_ Separated',
 'marital-status_ Never-married',
 'marital-status_ Never-married^2',
 'workclass_ Self-emp-not-inc',
 'native-country_ United-States',
 'native-country_ Mexico',
 'workclass_ Local-gov',
 'workclass_ Private',
 'workclass_ State-gov',
 'education-num',
 'relationship_ Not-in-family',
 'marital-status_ Married-spouse-absent',
 'workclass_ Self-emp-inc']
```

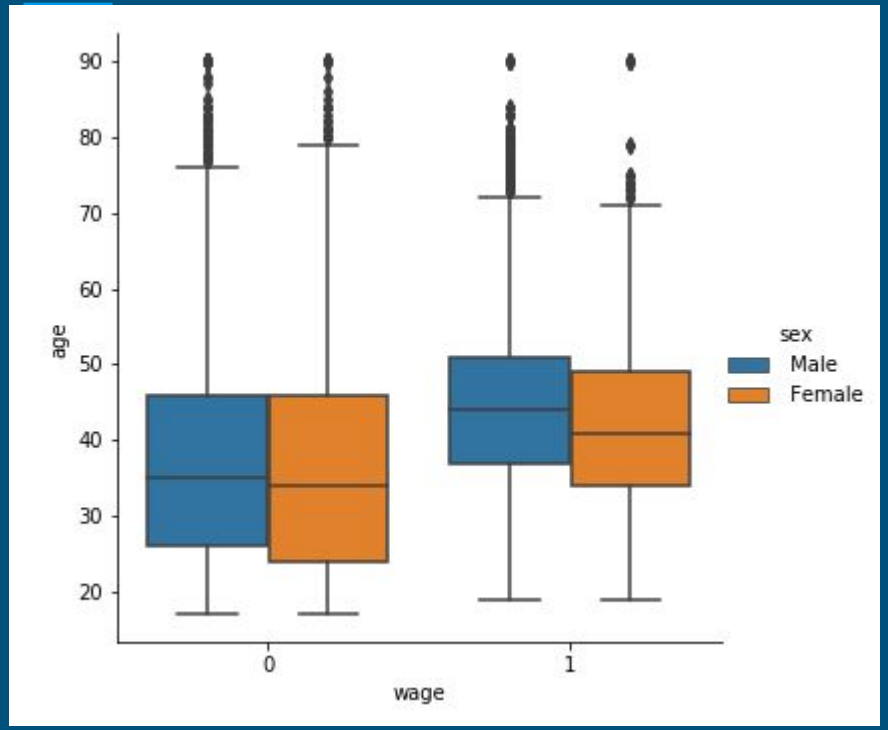
Correlations

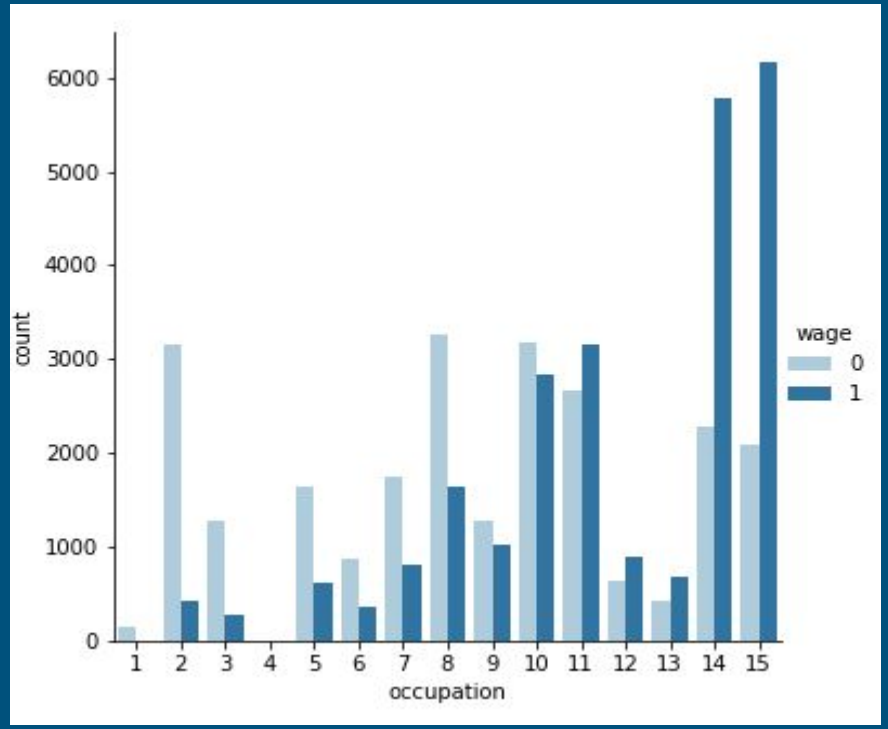
Highlights the top correlations

Useful for feature engineering

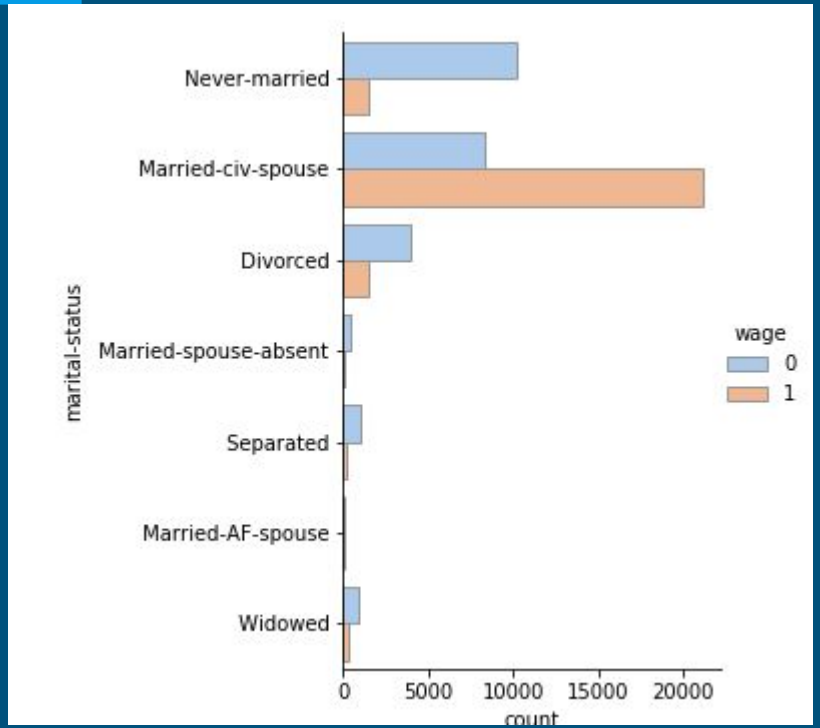


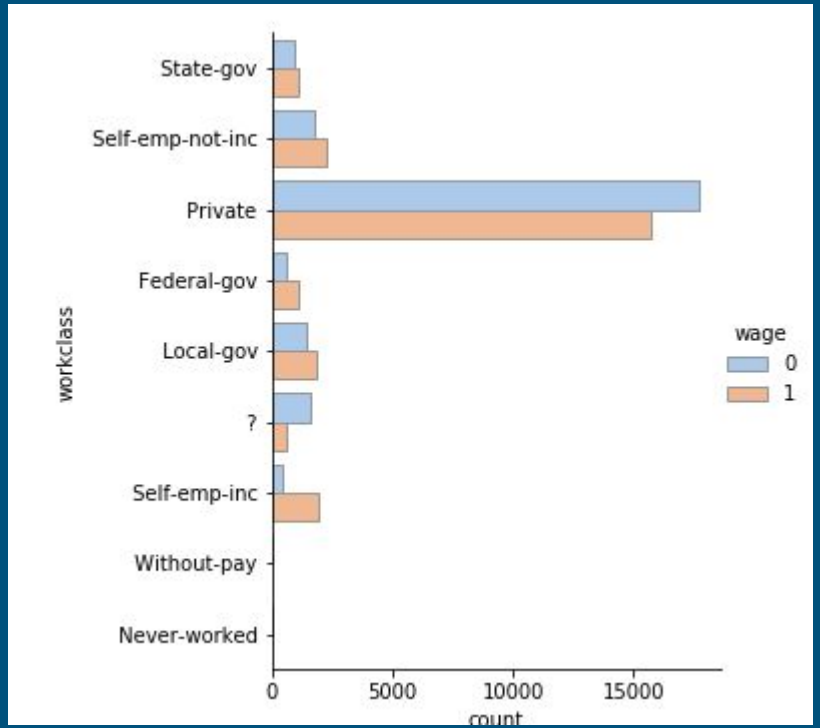
Visualization of Top Features





Visualization of Top Features, Cont'd





New logistic regression model scored

99% accuracy on test data

Other Models with Top 20 Features

Accuracy Scores on Test Data

Logistic Regression 98.92%	*Random Forest* xx.xx%
Gradient Boosting xx.xx%	*Support Vector Machine* xx.xx%

Conclusions

- Used logistic regression to identify the 20 most important predictors of income, including polynomial features.

- With logistic regression, built a model that was able to classify whether a person's income is more or less than $50,000 with 99% accuracy.

- Used top 20 features in other models, with varying degrees of success.