Income Categorization

Irene Anibogwu, Eric Heidbreder
Josh Mizraji, Juhee Sung-Schenck

Table of Contents

Problem Statement

Can we predict if a person's income is in excess of $50,000 given certain profile information?

Why it matters?

By predicting if a person’s income is in excess of $50,000, we can further utilize the model to improve business decisions such as if someone gets approved for a loan and identify areas of improvement for a more financially stable life.

Process

01 Clean the data

02 Exploratory data analysis

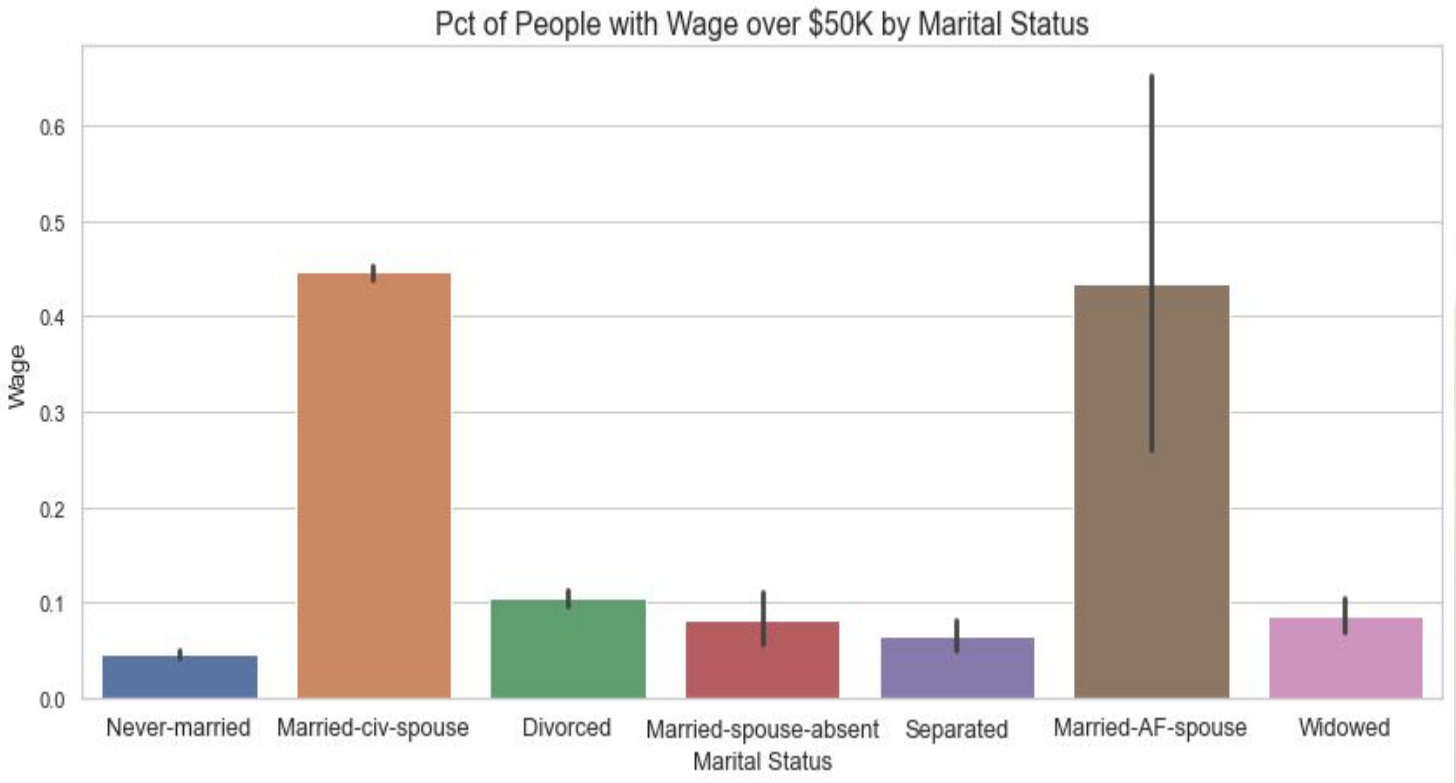
03 Engineer the features

04 Preliminary modeling

05 Modeling with hyperparameters

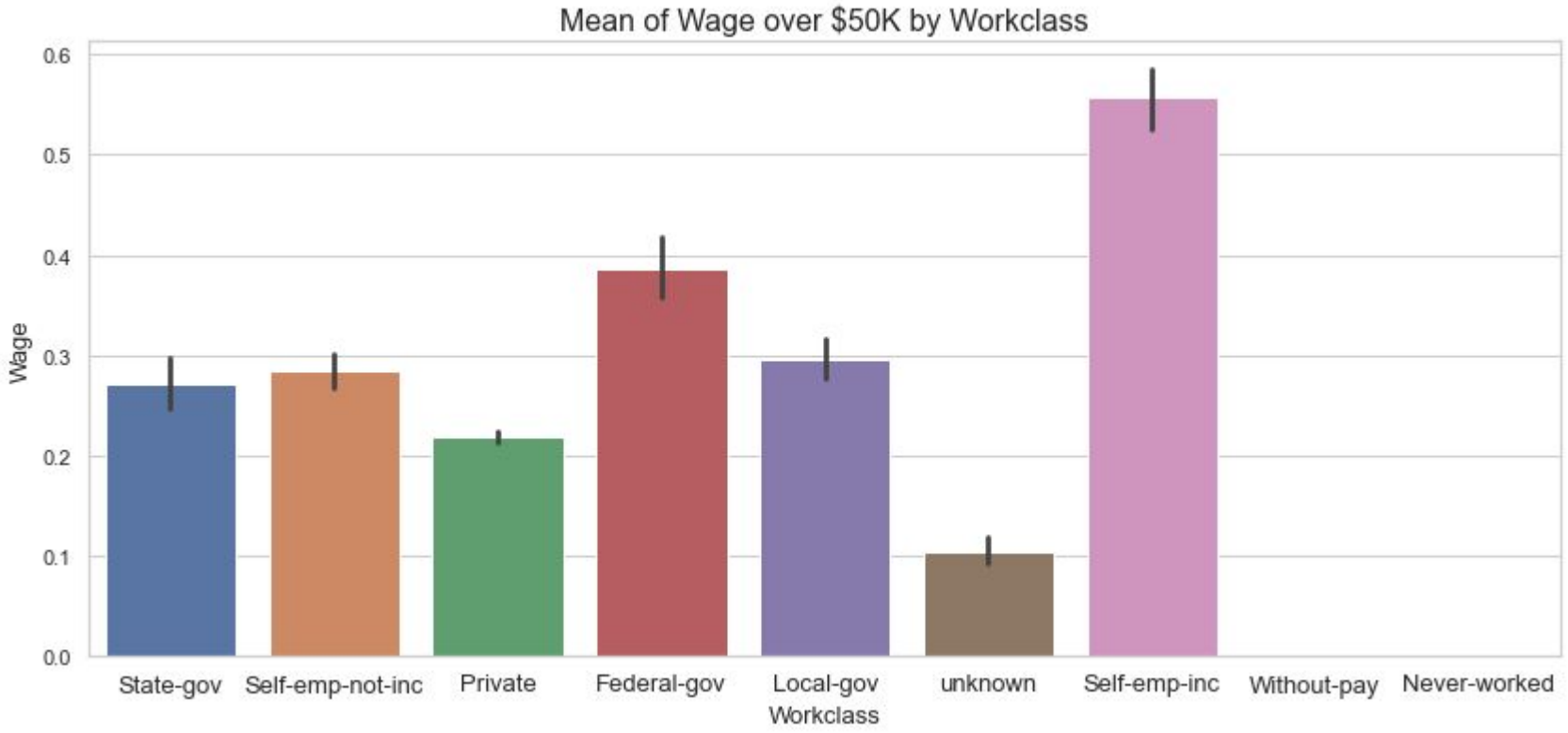
06 Analyze and compare performances

Exploratory Data Analysis



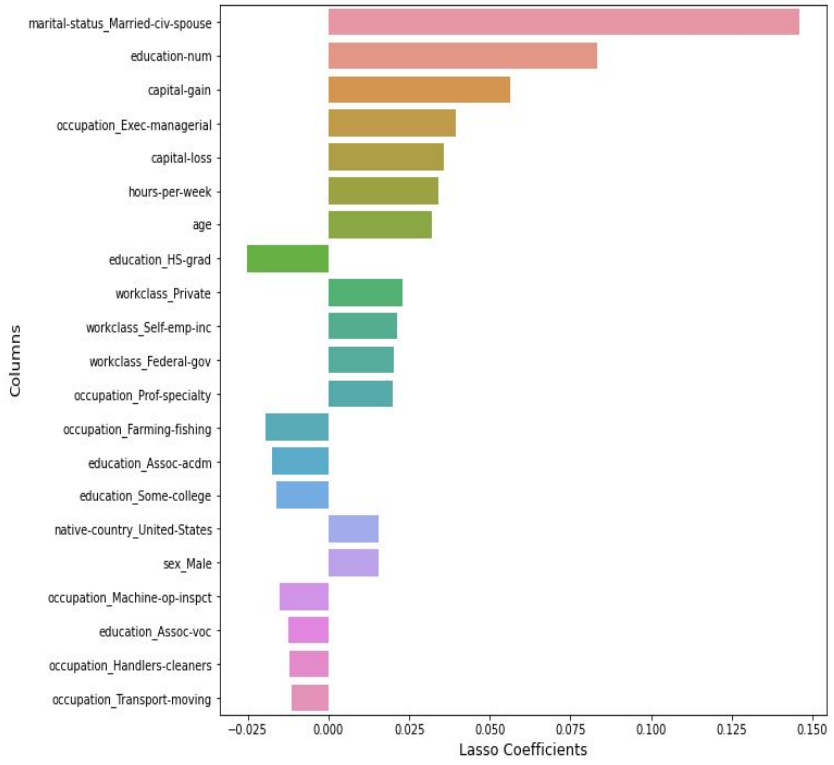
Exploratory Data Analysis

Exploratory Data Analysis



Feature Engineering

- Created Polynomial features on: age, hours-per-week, and educational-num
- Created dummy variables: workclass, education, marital-status, occupation, and native-country
- Binarized:
  sex and wage



Findings

Model	Logistic Regression
Training Accuracy	81.6%
Testing Accuracy	81.9%

Model	Naive Bayes
alpha	1
Train Accuracy	77.9%
Test Accuracy	77.6%

Findings

Model	Decision Trees
max_depth	7
min_samples_leaf	4
min_samples_split	20
Training Accuracy	86.1%
Testing Accuracy	85.9%

Model	Random Forest
max_depth	9
max_features	None
n_estimators	70
Training Accuracy	87.2%
Testing Accuracy	86.5%

Findings

Model	Bagging
n_estimators	2000
max_samples	300
max_features	18
Train score	85.6%
Test score	85.6%

Model	AdaBoost
n_estimators	2500
Train Accuracy	86.9%
Test Accuracy	87.2%

Findings

Model	SVC
C	10
Degree	2
Train Accuracy	86.6%
Test Accuracy	85.9%

Model	Voting Classifier
Train Accuracy	87.3%
Test Accuracy	87.3%

Voting Classifier contained the following models:

- AdaBoost
- Bagging
- Random Forest
- Decision Tree

Best Model

Model	AdaBoost
n_estimators	2500
Train Accuracy	86.9%
Test Accuracy	87.2%

*__Note__: This model requires a large amount of computing power to run.

Conclusion

- In conclusion, we found Adaboost to be the best model with a training accuracy of 86.9% and test accuracy of 87.2%
- Close 2nd goes to random forest with a 87.2 training accuracy and 86% testing accuracy score
- Model can be used to predict categorize whether someone makes 50,000 a year, and we can derive loan approval with 87% accuracy.
- Need better processing speed to run multiple jobs and find better models