Income Classification Report

John Enright

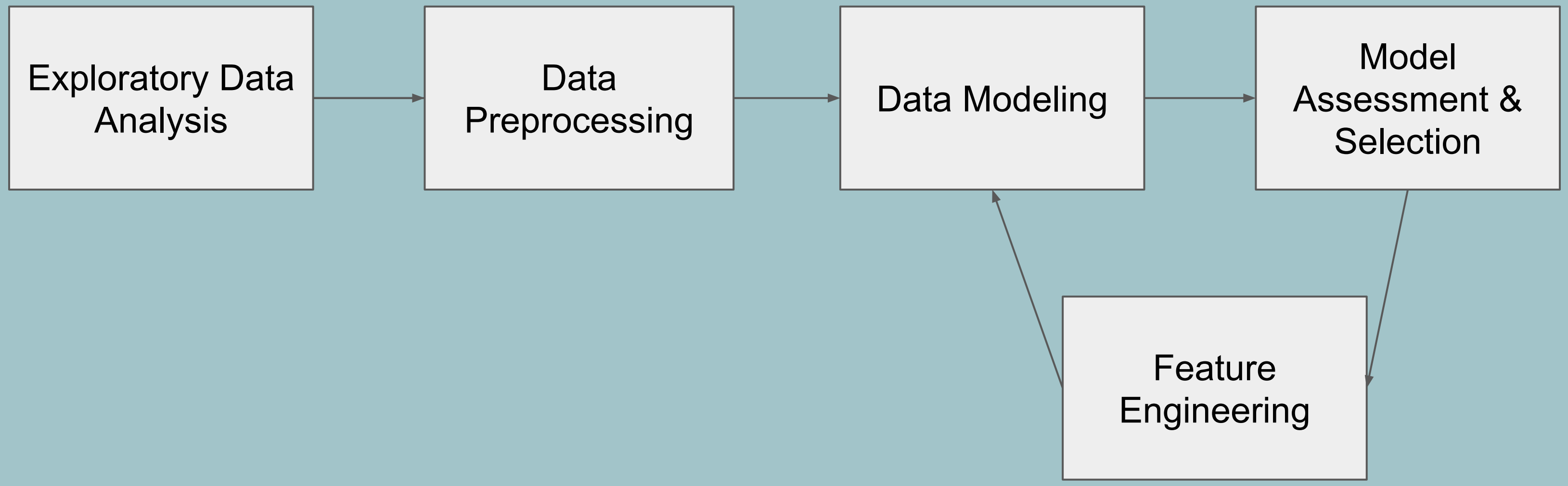
Outline

- Introduction
- Approach
- Exploratory data analysis
- Data cleaning & preprocessing
- Data modeling
- Model assessment and selection
- Conclusions & final thoughts

Introduction

- How can we identify an individual's income level based on a set of characteristics?
    - More or less than $50,000
- US Census Bureau Data
    - Collected every 10 years
    - Impacts the allocation of government funding
    - Examines demographic characteristics of subpopulations
    - Survey of ~ 300,000 individuals

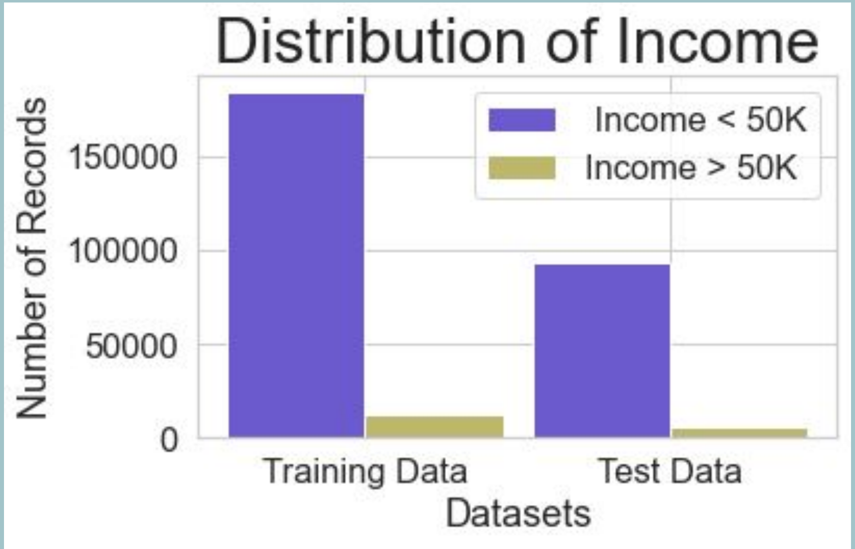
Approach



Exploratory Data Analysis
Data Preprocessing
Data Modeling
Model Assessment & Selection
Feature Engineering

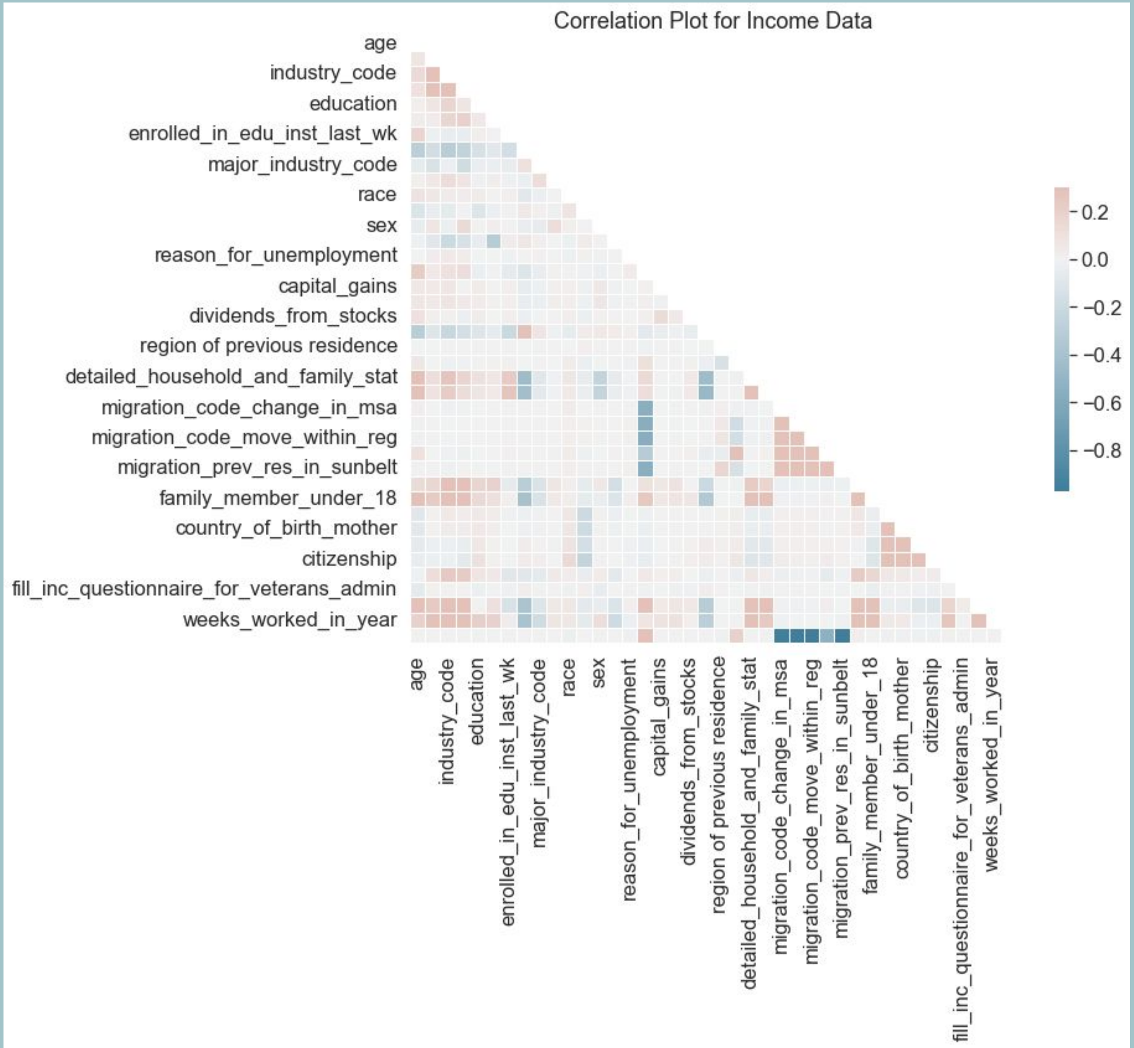
Exploratory Data Analysis

- Imbalanced data
  - **93%** individuals make over $50,000
  - Can introduce classification problems
- Duplicate or conflicting instances
  - Training → 46716
  - Test → 99762
- 40 characteristics (features) for an individual
  - 33 are continuous values
  - 7 are categorical
- Example features
  - Age
  - Sex
  - Race
  - Education level
  - Class of work

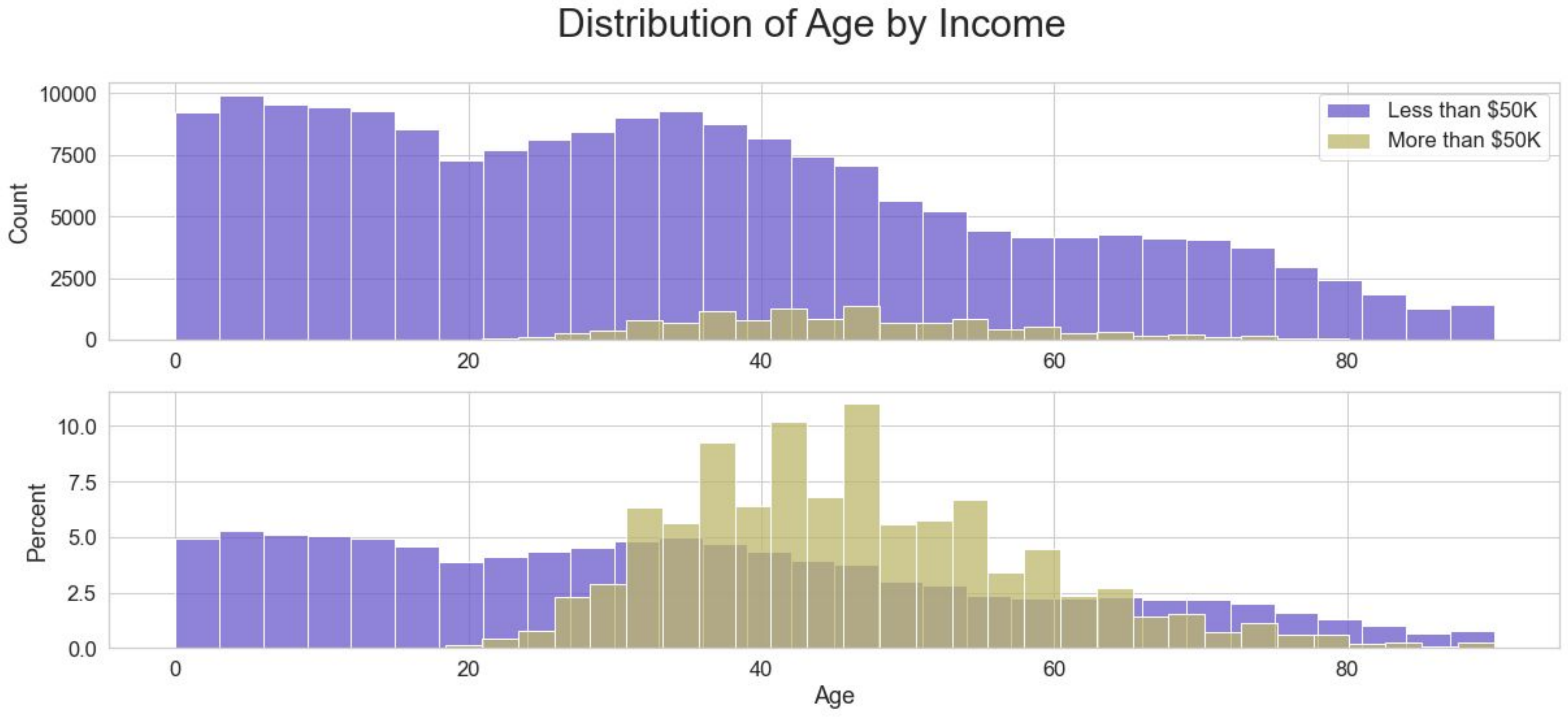


Correlation between features

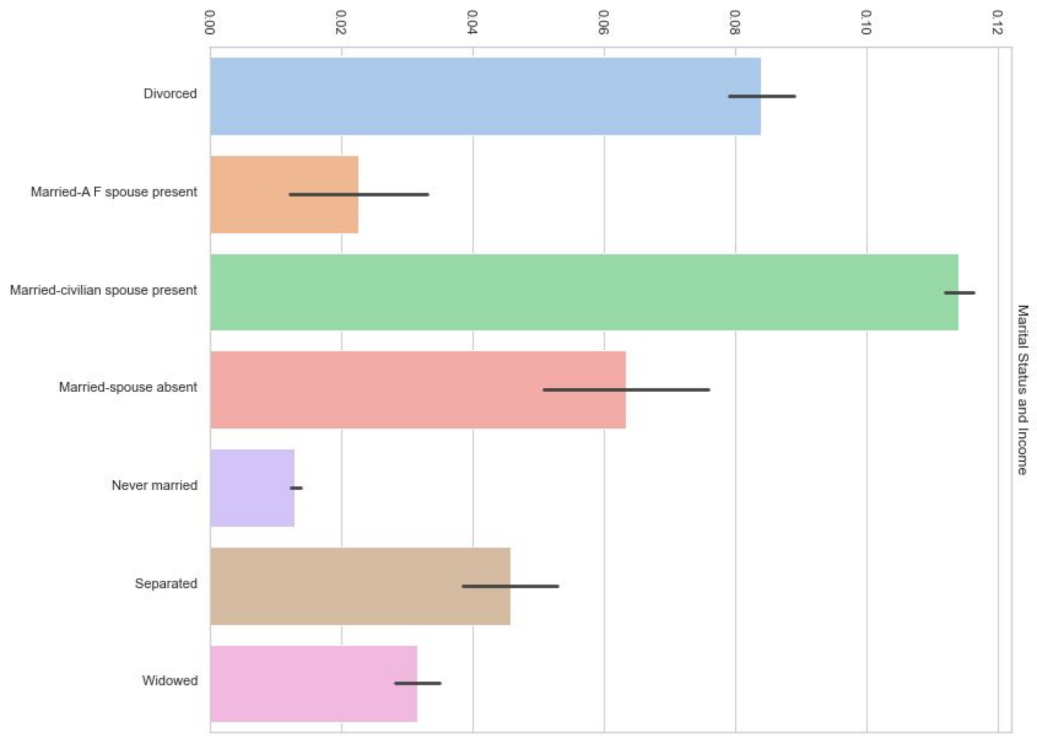
- Relationship



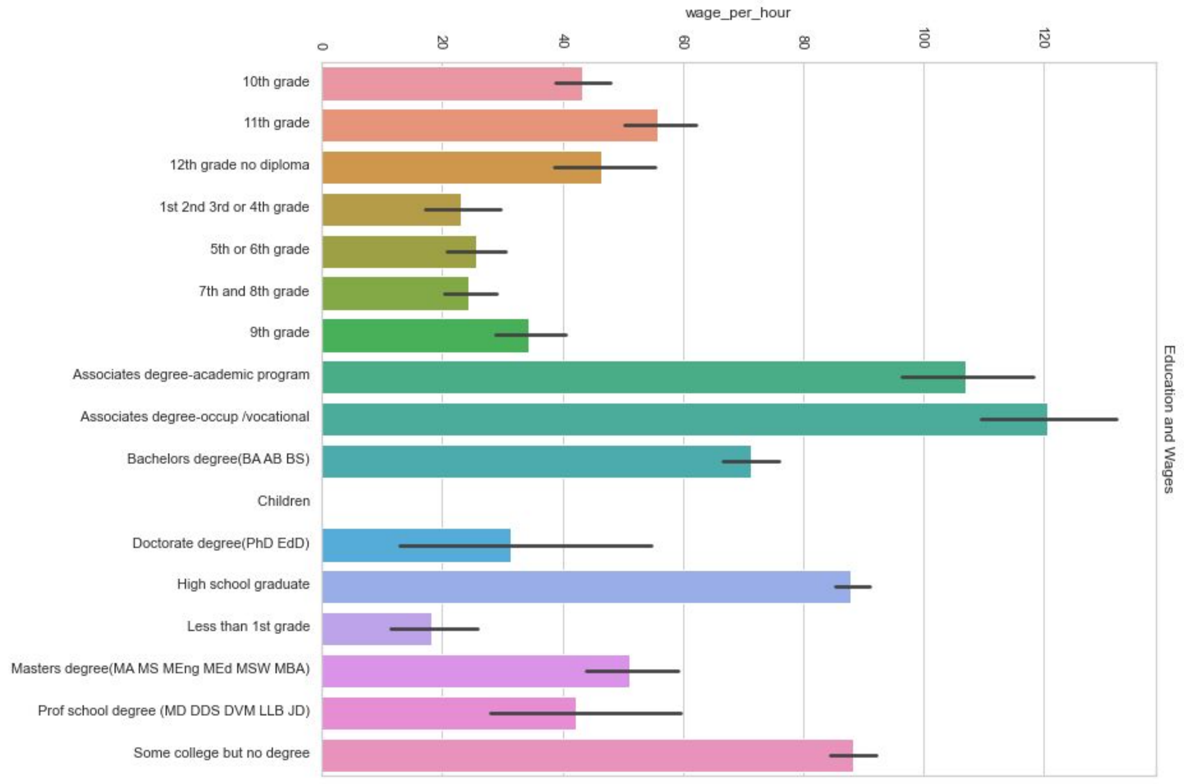
Exploratory Data Analysis



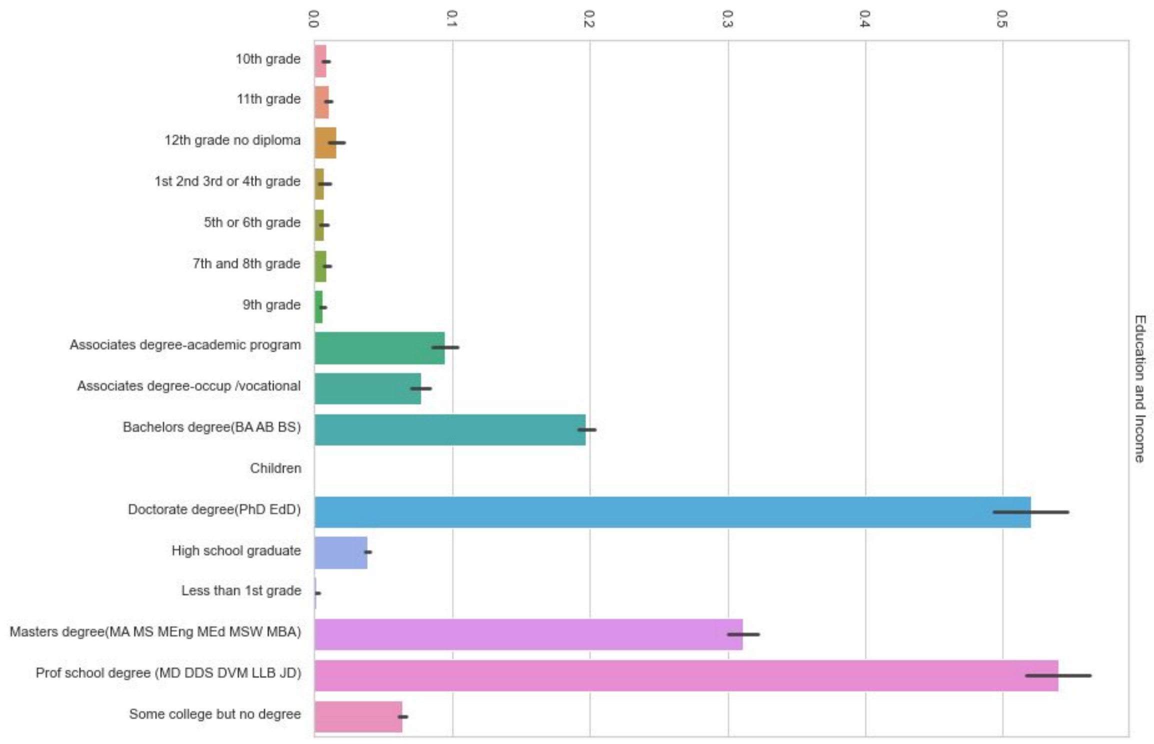
Exploratory Data Analysis - Marital Status



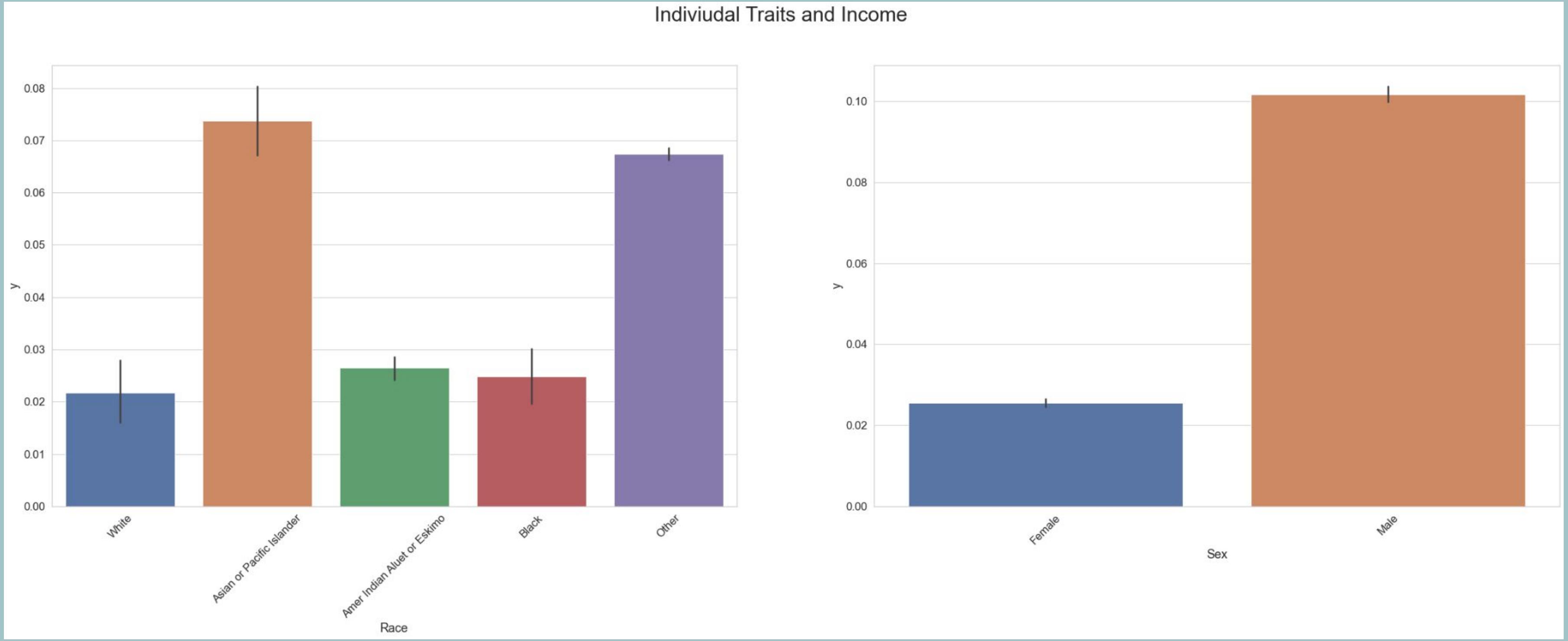
Exploratory Data Analysis - Education



Exploratory Data Analysis - Education



Exploratory Data Analysis - Race & Sex



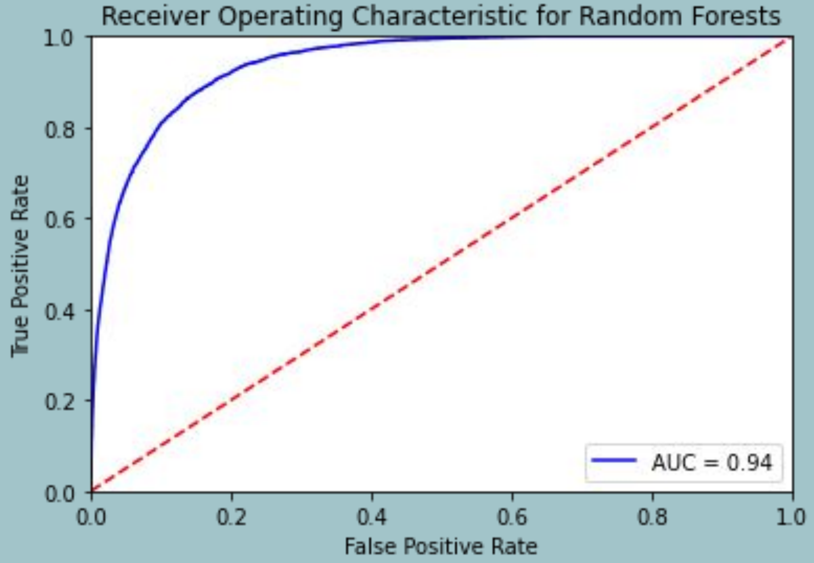
Data cleaning & preprocessing

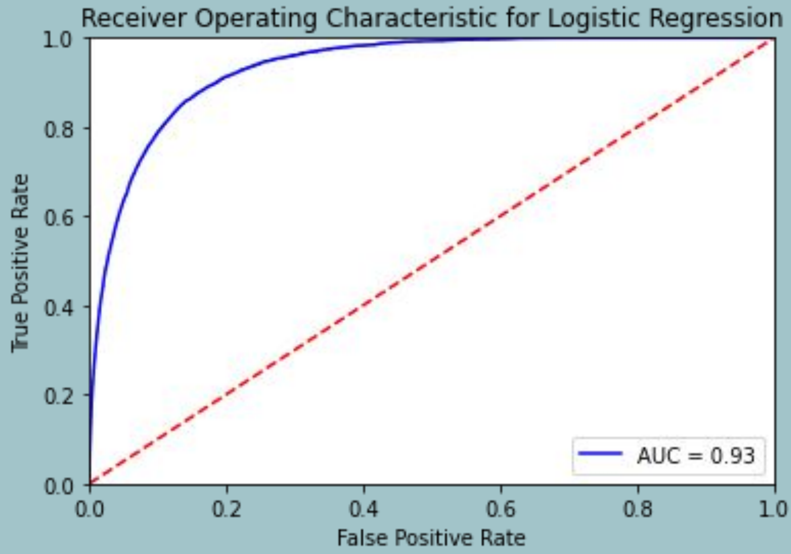
- Data
  - Labeled
  - Duplicates dropped
  - Categorical features are split into appropriate categories
  - Highly variant data is normalized
    - Age
    - Wages
  - Features dropped & transformed
    - Instance weight
    - Capital losses → ‘Has losses’
    - Capital gains → ‘Has gains’
    - Dividends → ‘Has stock’

Data modeling

- Classification methods
  - Logistic regression classifier
  - Random forests classifier
- Goal
  - Predict income level with individual characteristics as inputs
- Evaluate predictability on the test set
  - F1 score
  - **ROC Area under curve**
- Metrics represent predictability of each respective class

Model results - first pass

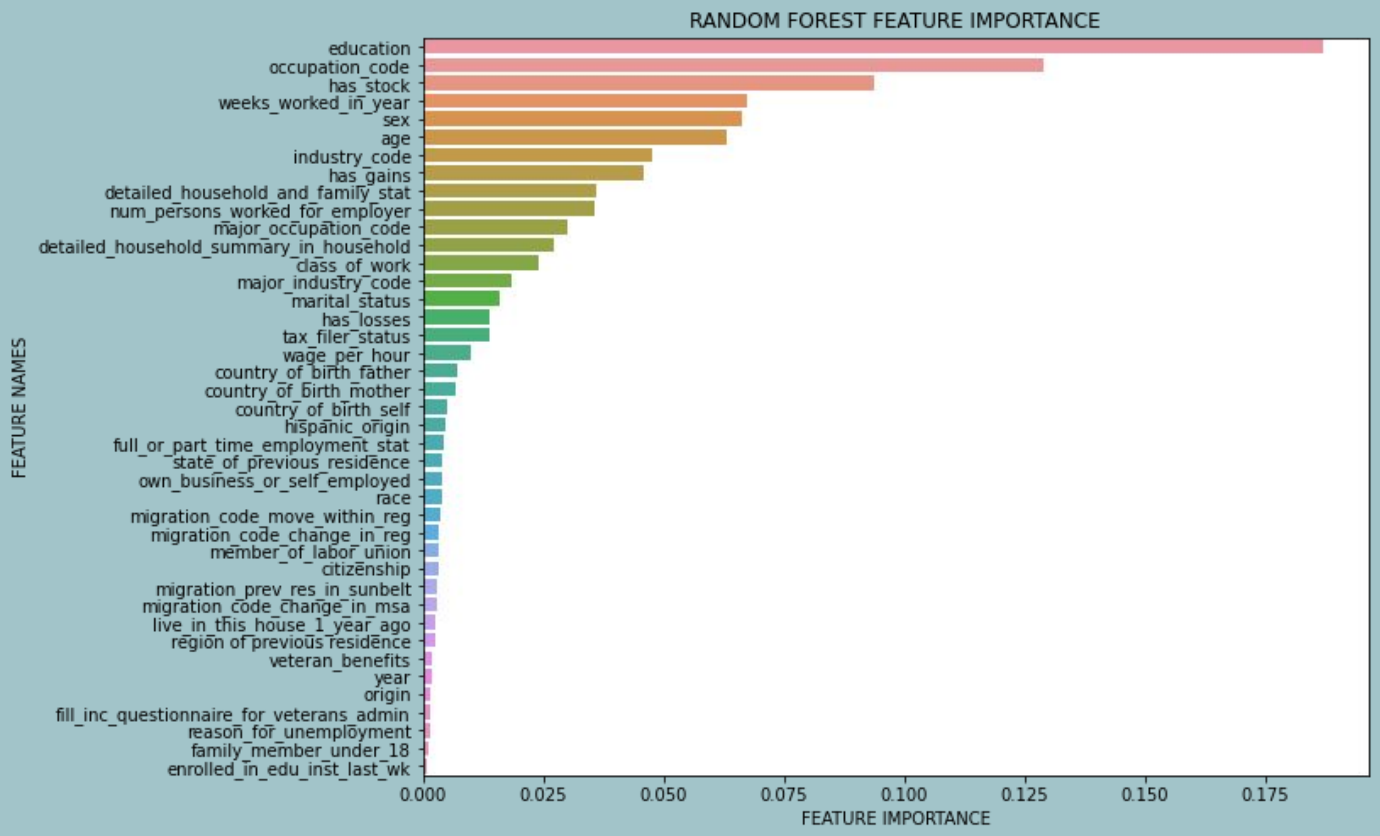




	F1 Score
Under $ 50,000	0.97
Over $ 50,000	0.41

	F1 Score
Under $ 50,000	0.97
Over $ 50,000	0.44

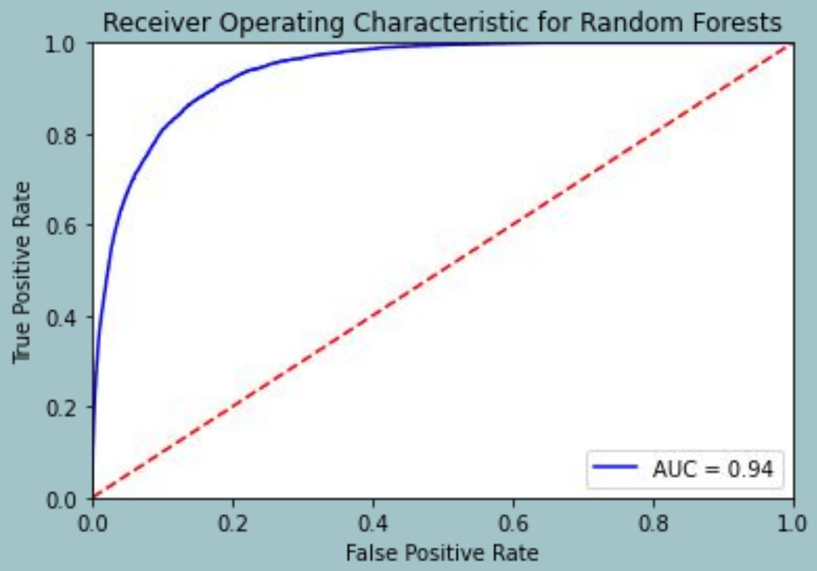
Model results - important features



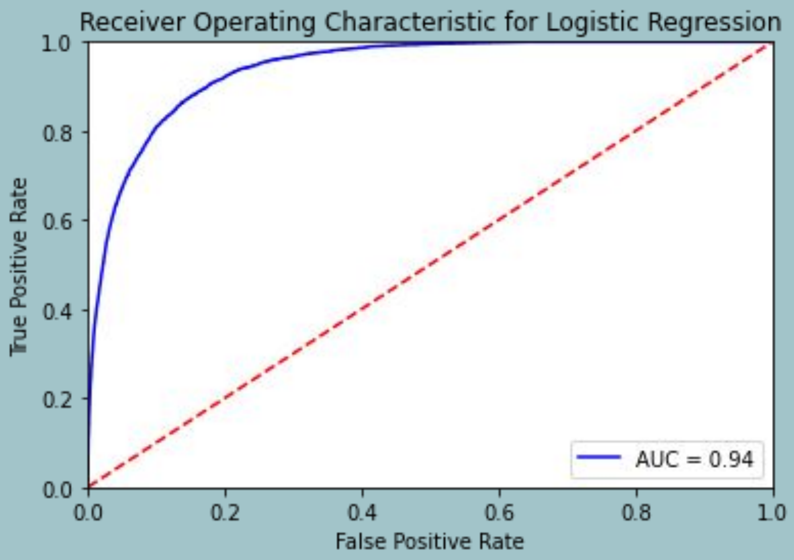
Feature engineering

- Take the top “important” features
  - Transform the data set
  - Number of features 41 → 12
- Retrain models
  - Compare to non-engineered models

Model results - reduced features



	F1 Score
Under $ 50,000	0.97
Over $ 50,000	0.46



	F1 Score
Under $ 50,000	0.97
Over $ 50,000	0.43

Modeling takeaways

- The logistic classifier is slightly stronger at predicting the minority class
- Random forests model performs stronger in other metrics
- Random forests model gives key insights into feature importance
- Feature engineering did not improve model performance
  - Lose key information
- Average accuracy of predicting income level ~ 86 %
  - 95 % predicting under $50k
  - 76% predicting over $50k
- Average f1 score ~ 70%

Other key takeaways

- Highest probability to make over $50k
    - Middle aged individuals
    - Asian & pacific islanders
    - Males
    - Married individuals
    - Higher levels of education (doctorate, professional)
- Occupation and where they work are related to income
- Higher e
- An individual having stock dividends is related to income
    - Disposal income?

Further considerations

- Improving model performance
  - Hyperparameter tuning
  - Ensemble methods
- Experimenting with other classifiers
  - Support vector machines
  - K nearest neighbors
  - Neural Networks
- Investigating occupation code's relationship to income
  - Deriving occupation titles to the codes
- Getting new data
  - 1994 & 1995
  - Balancing the classes

Thanks!

John Enright