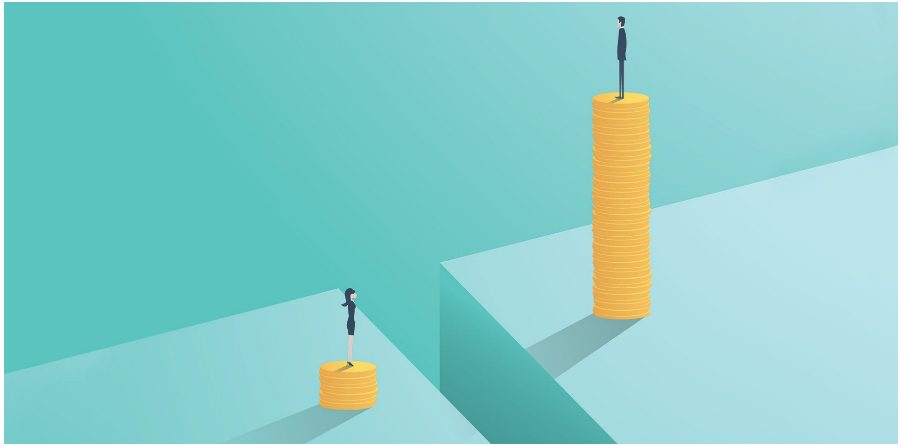
TEAM 8 - Adult Income Prediction

Fernanda Lin, Kyle Blackburn, Mansi Tolla, Lyufan Pan, Honyang Liu

Problem

- The inequality gap in the US has steadily risen over the past 40 years[1]
- We are interested in predicting income levels based on 1994 US census data
  - Predict >= $50k/year OR < $50k
- $50K per year is defined as the bottom threshold for the middle class



Dataset

- **Adult Income Dataset** from Kaggle[1]
- Data extracted from 1994 Census database by Ronny Kohavi and Barry Becker
  - Unclean dimensions: 48.8K x 15
  - Clean dimensions: 41.1K x 11
    - Removed 4 unnecessary fields
    - Removed all incomplete rows ~7K
- Target field - income
    - <=$50K and >$50K
- Predictors - 24 socioeconomic factors/demographics

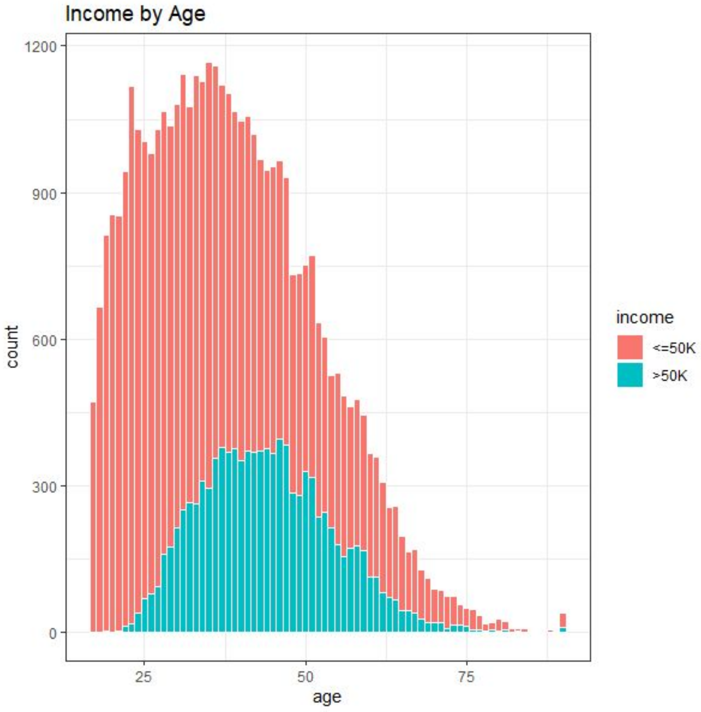
Dataset

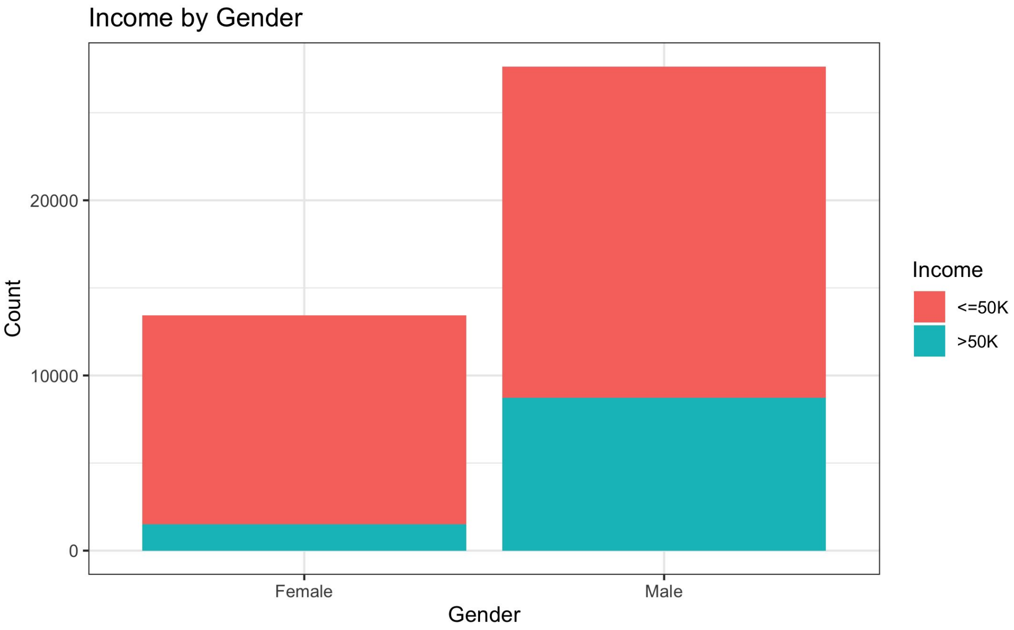
	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relationship	race	gender	capital-gain	capital-loss
1	25	Private	226802	11th	7	Never-married	Machine-op-inspct	Own-child	Black	Male	0	0
2	38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	Husband	White	Male	0	0
3	28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	Husband	White	Male	0	0
4	44	Private	160323	Some-college	10	Married-civ-spouse	Machine-op-inspct	Husband	Black	Male	7688	0
5	18	?	103497	Some-college	10	Never-married	?	Own-child	White	Female	0	0

	ID	age	income	capChange	gender	hoursPerWeek	skilled	workClass_Federal.gov	workClass_Local.gov	workClass_Private	workClass_Self_employe
1	1	25	0	0	0	40	1	0	0	1	0
2	2	38	0	0	0	50	0	0	0	1	0
3	3	28	1	0	0	40	0	0	1	0	0
4	4	44	1	7688	0	40	1	0	0	1	0
5	6	34	0	0	0	30	0	0	0	1	0

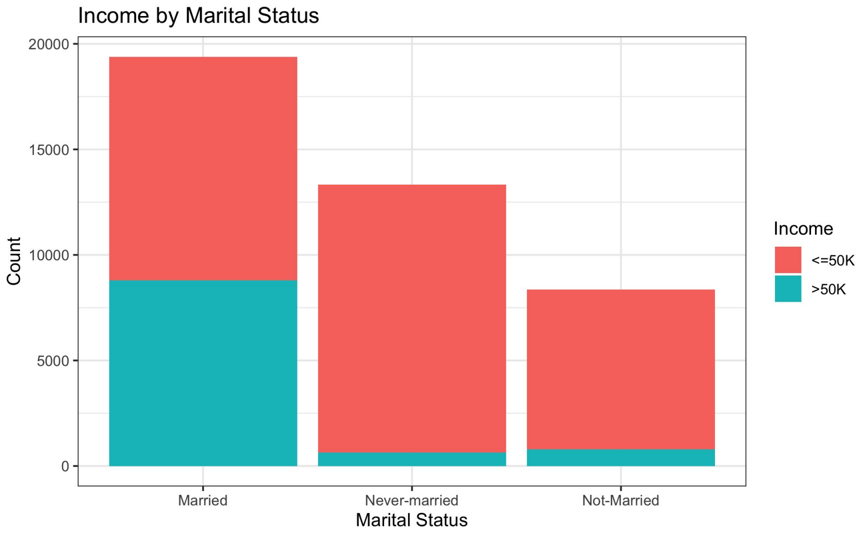
*Not all columns visible

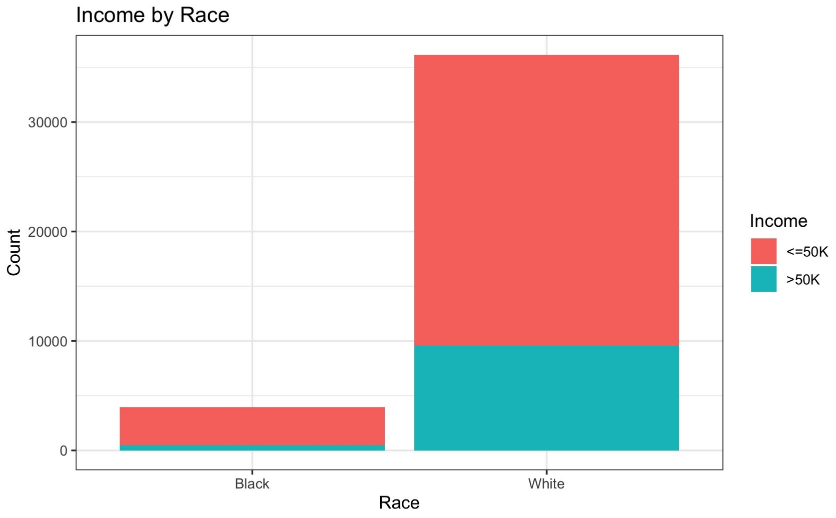
Exploration





Exploration





Main Results

Model	Test MSE
Random Forest	0.1049572
Forward Selection	0.12361100
Backward Selection	0.12368289
Lasso	0.12369716
Elastic Net	0.12369737
Ridge	0.12376343
Boosting Trees	0.13491678

Lasso/Ridge/Elastic Net

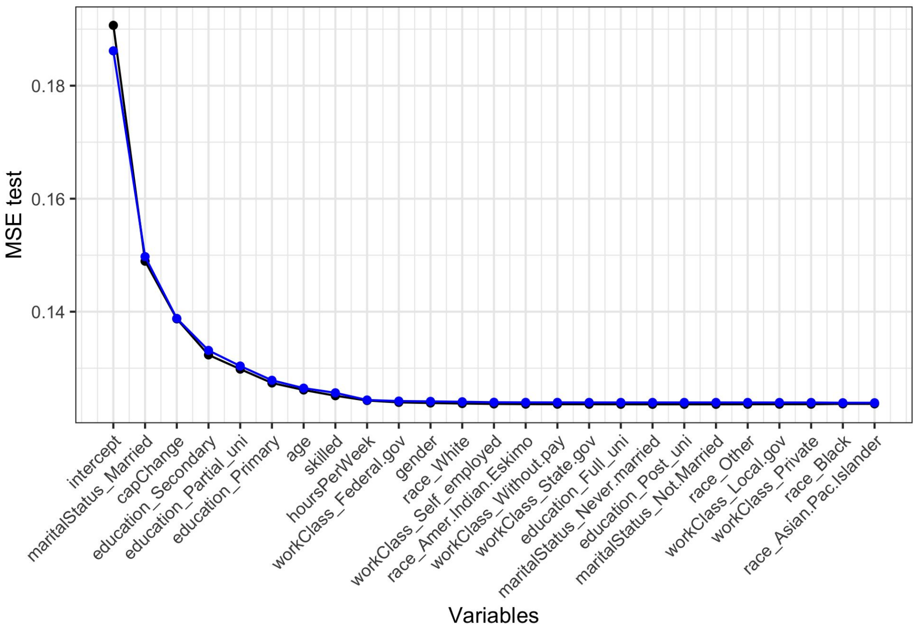
- Lasso reduced the number of predictors down to 13

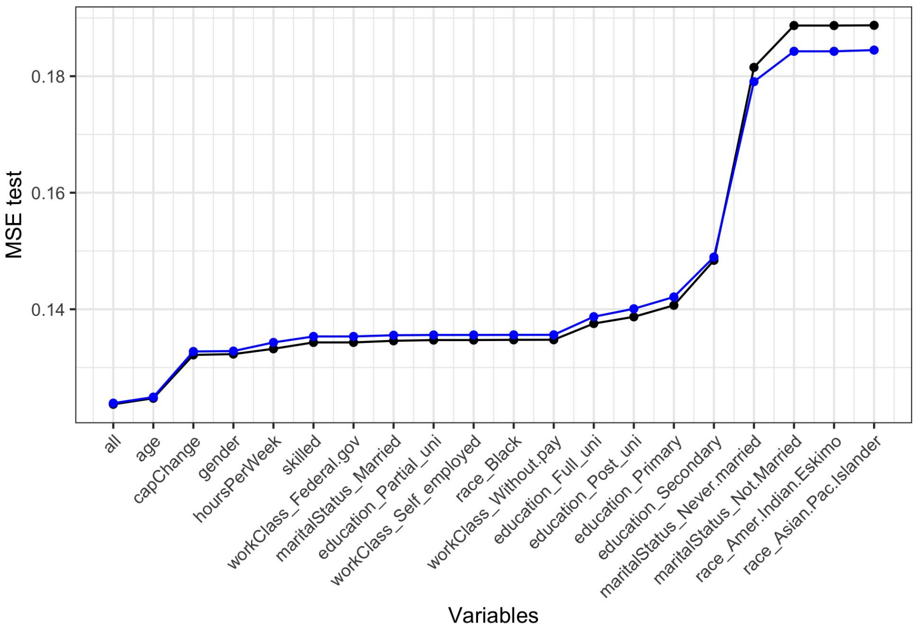
maritalStatus_Married1	2.97E-01	0.2970958
education_Primary1	-1.63E-01	0.1626783
education_Full_uni1	1.19E-01	0.1192258
education_Post_uni1	1.15E-01	0.1151875
education_Secondary1	-6.44E-02	0.06435764

Boosting Trees

- Used 100 trees
- Increasing the shrinkage & the interaction depth helped reduce MSE of the model

Forward/Backward Selection



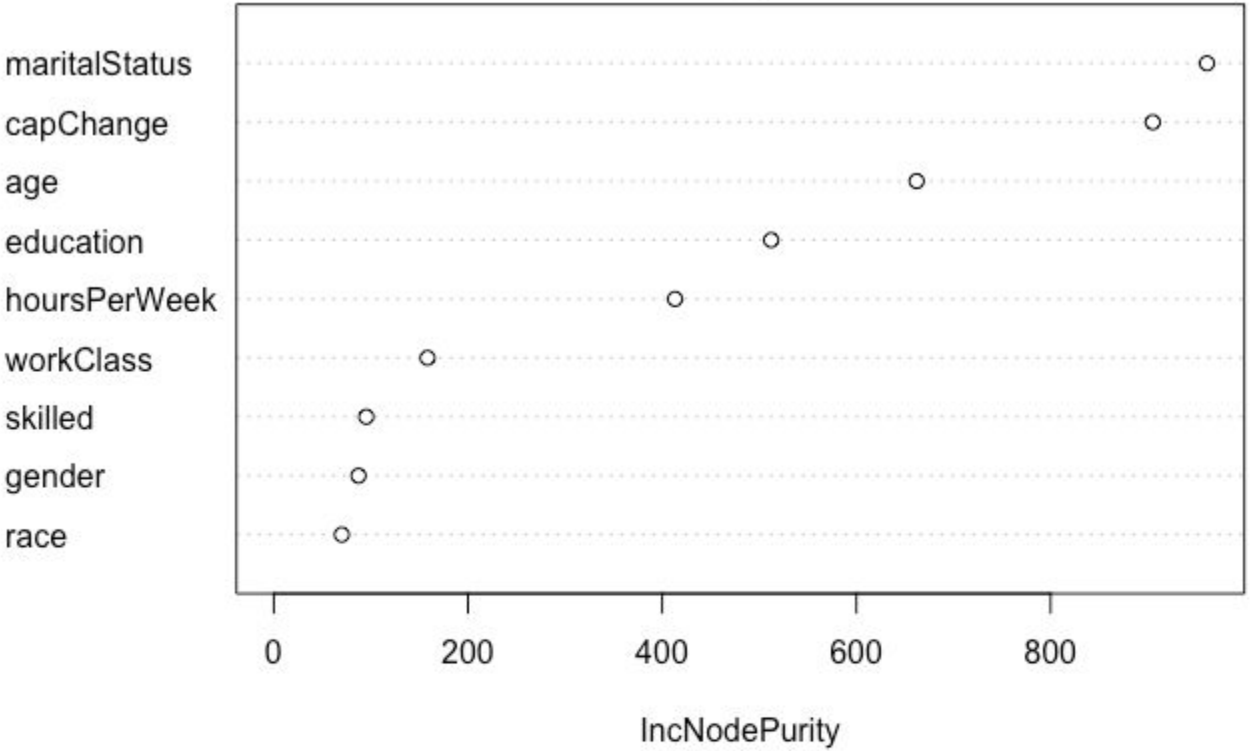


Best Model - Random Forest

- The random forest algorithm returned the best MSE_{test}
- Rest of models were very close in MSE_{test}
- No. of trees used: 100

- Recurring attributes across models (shared significance):
  - Marital Status
  - Capital Change
  - Race
  - Education

Best Model - Random Forest



Challenges

- Data Cleaning
  - Choosing relevant variables to use
    - Census data is specific
    - Grouping - years of school, marital status, employment status, occupation
  - Choosing unnecessary variables to omit (arbitrary figures, came with data, collinearity)
    - Fnlwgt
    - Educational-num
    - Relationship
    - Native country - data was > 95% USA
  - What to do with N/As (removed)
- Applying theoretical knowledge (classroom) to R environment (programming)

Learning Points

- Comparing ML algorithms to find make prediction
- Nuances of different ML packages
  - Differ across classification, prediction, etc.
- Familiarity with spreading categorical variables into dummy variables with significance
- Tuning hyperparameters and their impact on overall MSE

Thank You!

Questions?