Predicting income from the 1994 US Census

Nicole McBride

Median household income in the United States

Goal

The What: Predict whether an individual's income will be greater than $50,000 based on several attributes from the 1994 US Census data.

The Why: The local state government wants to see how well a model from 30 years ago would correctly classify against today's median household income. Insight into how demographic characteristics have changed. First step is to build a model using the 1994 data.

Method

Context

- UCI Machine Learning Repository / Kaggle
- 1994 US Census Data
- **32,561** people, 14 features (8 categorical)
- Predictive variable:
  - Greater than $50k
  - Less than or equal to $50k

Analysis Steps

1. Exploratory Analysis
2. Data cleaning
3. Feature Engineering
4. Extensive validation

Features

Capital Loss
Capital Gain
Age

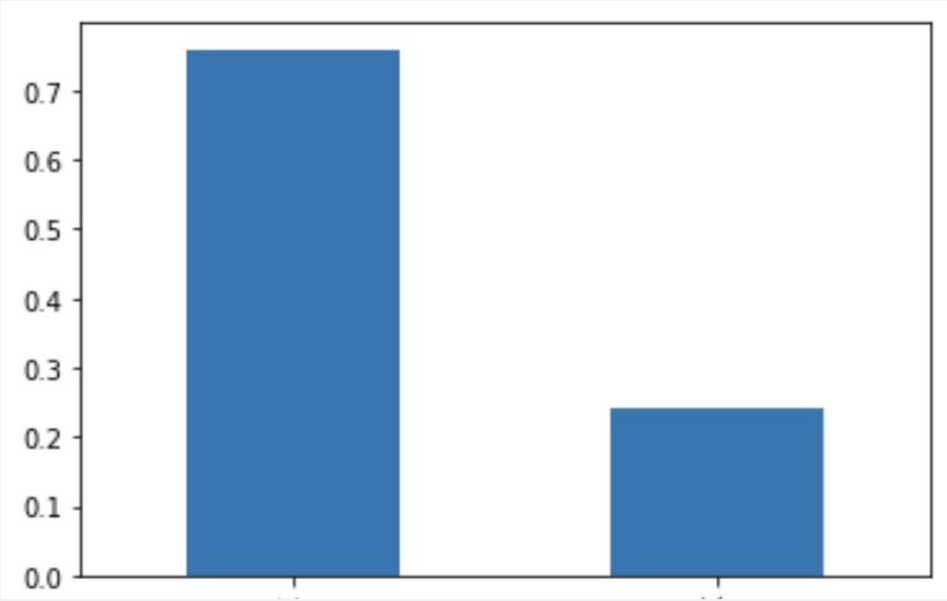
Weekly working hours

Education

Occupation
Employment Status
Entry Representation

Native Country
Sex
Marital Status

Race
Relationship Status



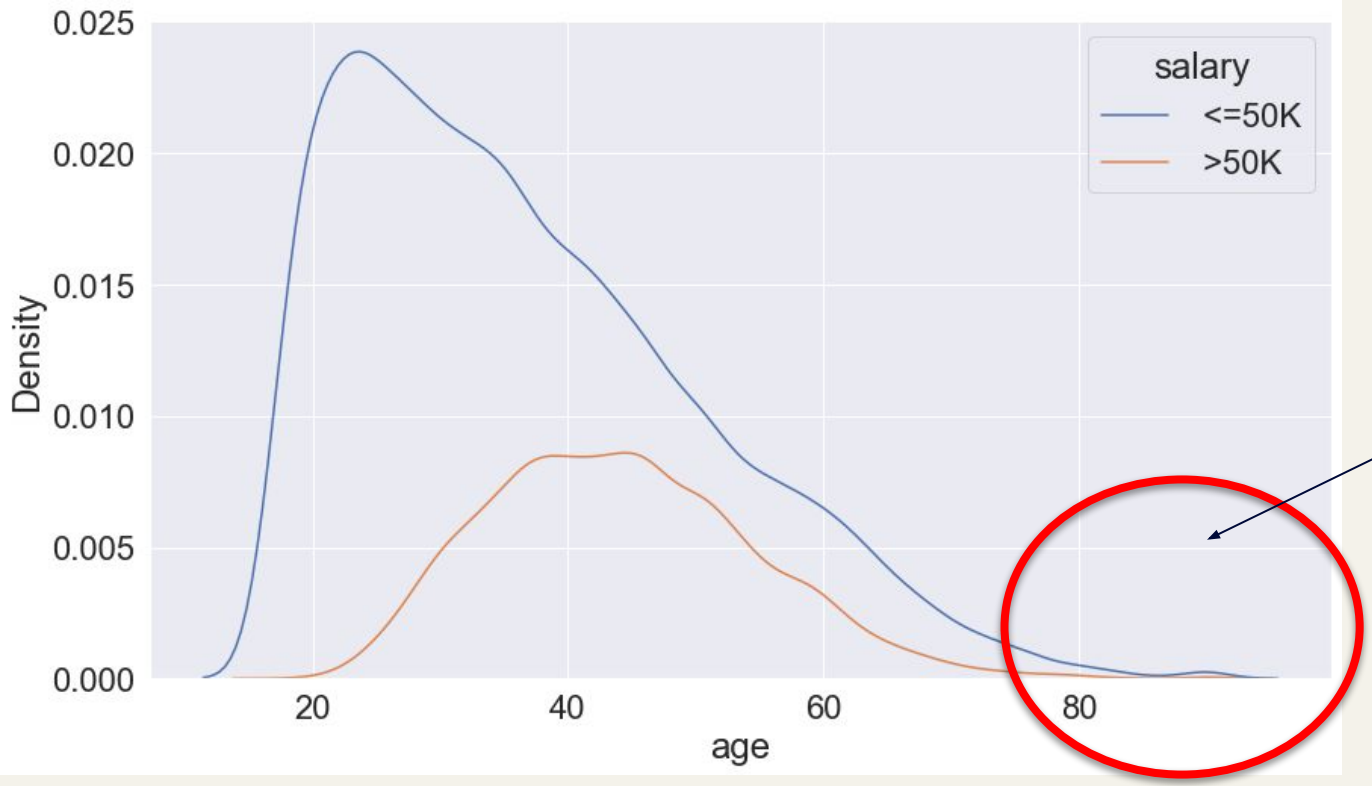
<=$50k >$50k

24% >$50k

76% <=$50k

Exploration

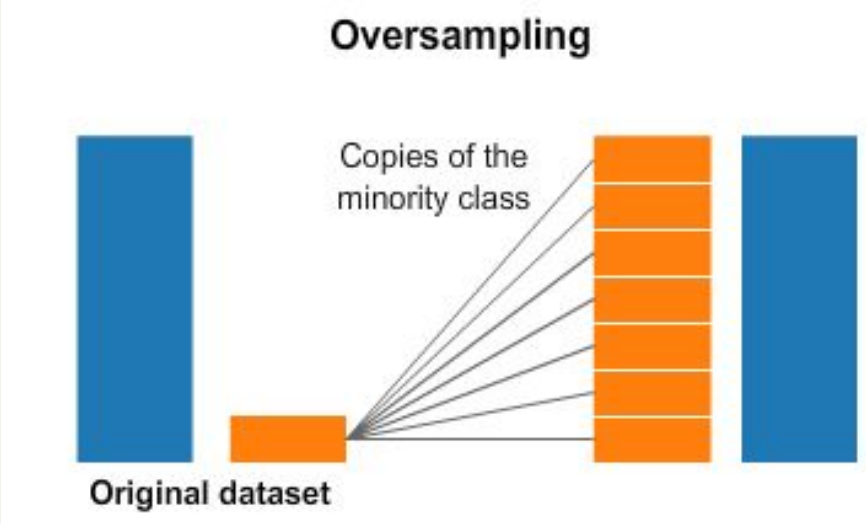
What's the context of these salaries in the elderly?



16 people who were >= 80 years old & worked MORE THAN 50 hours/week

....

DROPPED



UNBALANCED

<=$50k: 14,802

>$50k: 4,725

BALANCED

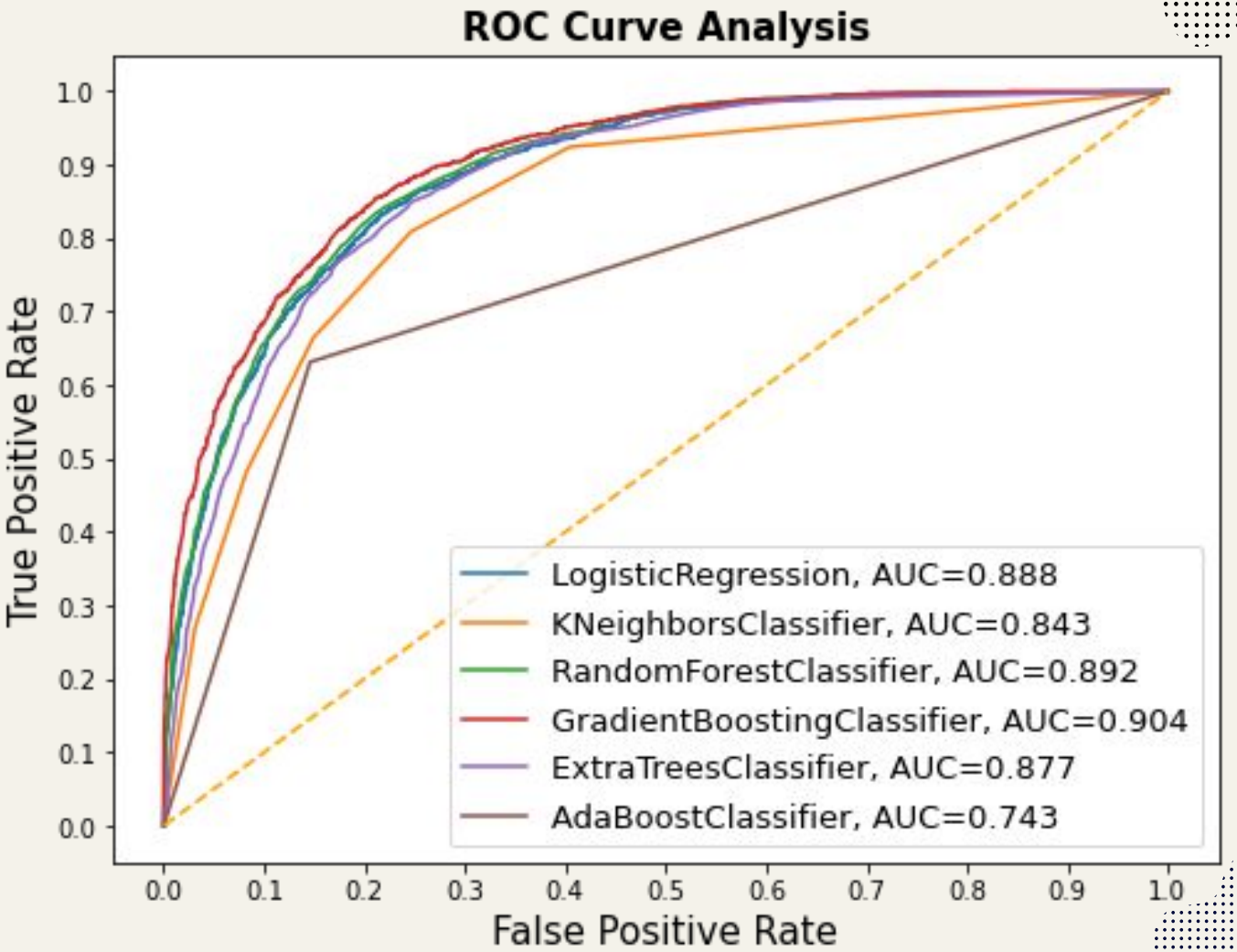
<=$50k: 14,802

>$50k: 14,802

The AUC helps us distinguish between those making more than/equal to $50k vs less than $50k.

The higher the AUC value, the better the model is at distinguishing between these two classes.

Our Gradient Boosting model gives us the best value!



Gradient Boosting Classifier

What GBC model yielded best results?

Max depth: 8

Max features: 6

N estimators: 150

All other hyperparameters were set to default.

Scores

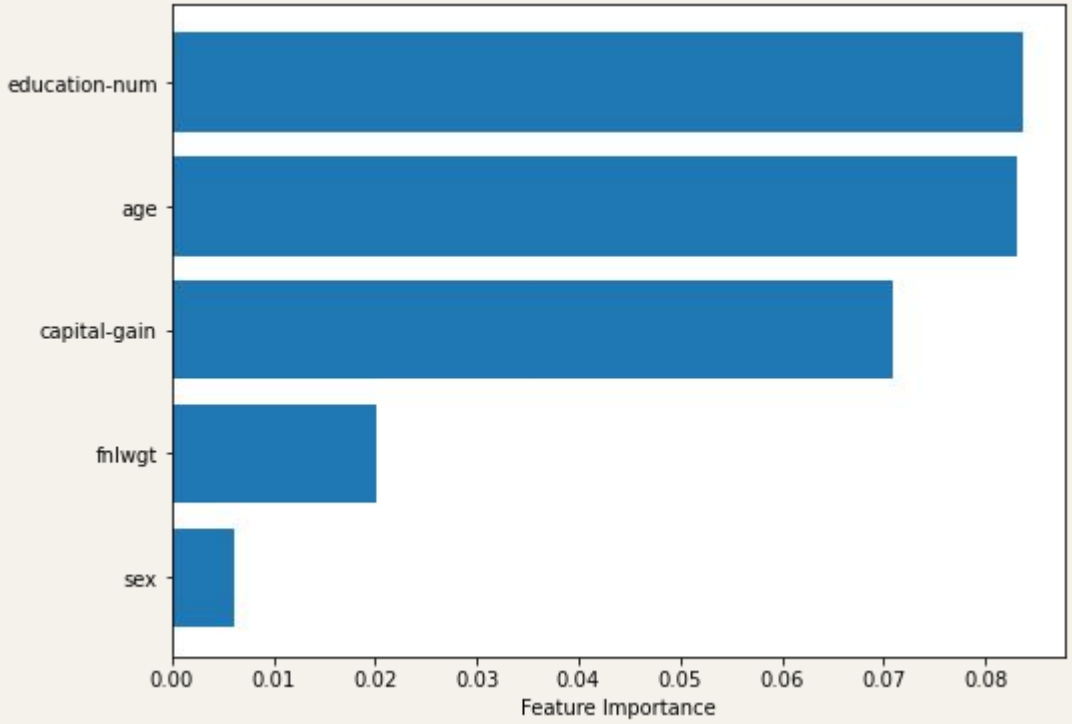
Accuracy	.86
Precision	.71
Recall	.7
F1 Score	.7

Confusion Matrix

		Predicted Class	
		<= $50k	> $50k
True Class	<= $50k	**4560** TRUE NEGATIVE	435
	> $50k	457	**1057** TRUE POSITIVE

Good false positive to false negative ratio!

Top 5 Most Important Features



The higher the score, the greater effect this variable on our model in predicting income classification (>= or <$50k).

Future Work

Deeper feature engineering

Hyperparameter tuning for other models

Deeper feature importance

Different dataset

THANKS!

Questions?