Data Scientist Technical Assessment and Presentation

What makes a $50k+ a year income earner?

About me - David Rickheim

Computer Science & Engineering

CARLSON SCHOOL OF MANAGEMENT

Medtronic

Abbott

Medtronic

Agenda

1. Problem Statement
2. Exploratory Data Analysis
3. Data Preparation
4. Data Modeling
5. Model Assessment
6. Results

1: The problem

Problem statement

Can we predict who makes over $50k annually?

How may we differentiate this group (what predicts a >$50k earner)?

Importance

Determines qualifications for many things, such as:

- SNAP
- School lunches
- Medicaid

Census Data

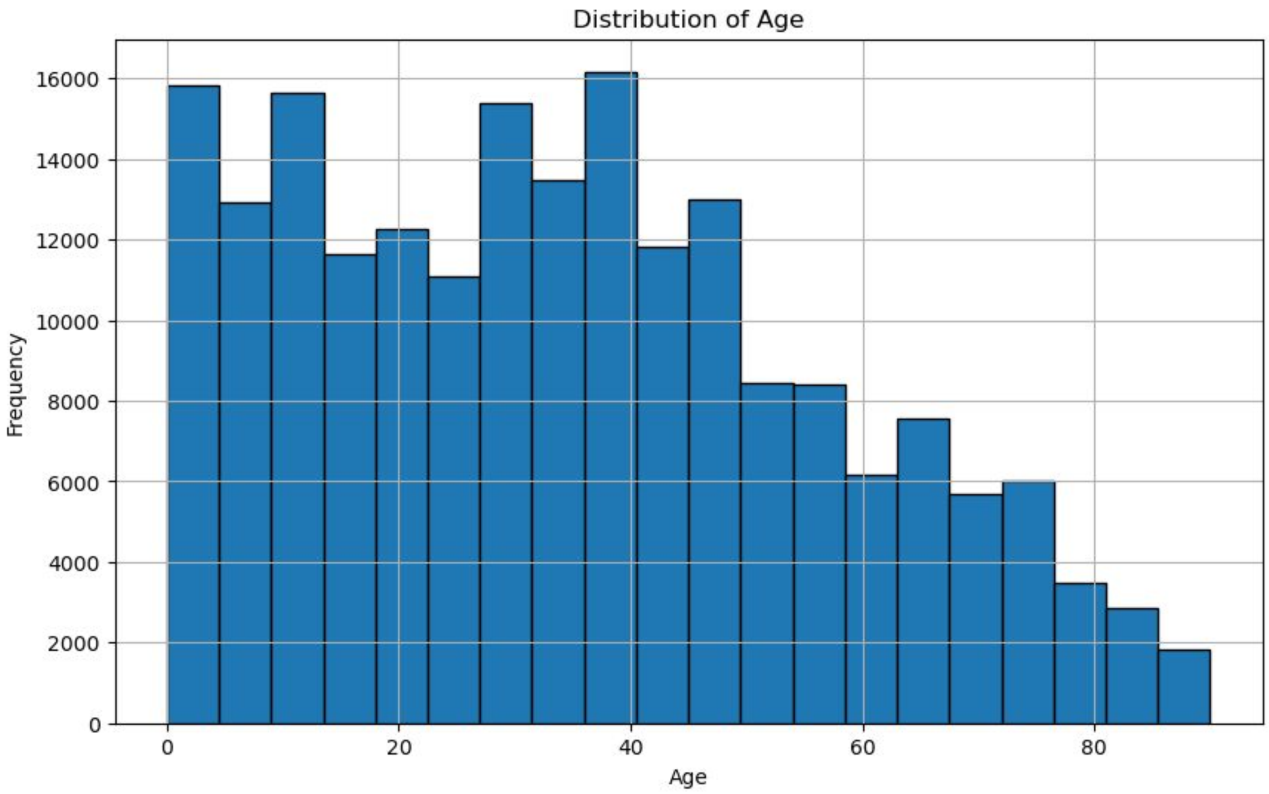
- Annual Social and Economic (ASEC) Supplement
  - 1993 / 1994
- Hourly wage, hours worked, demographic info, occupation, education

2. Exploratory Data Analysis

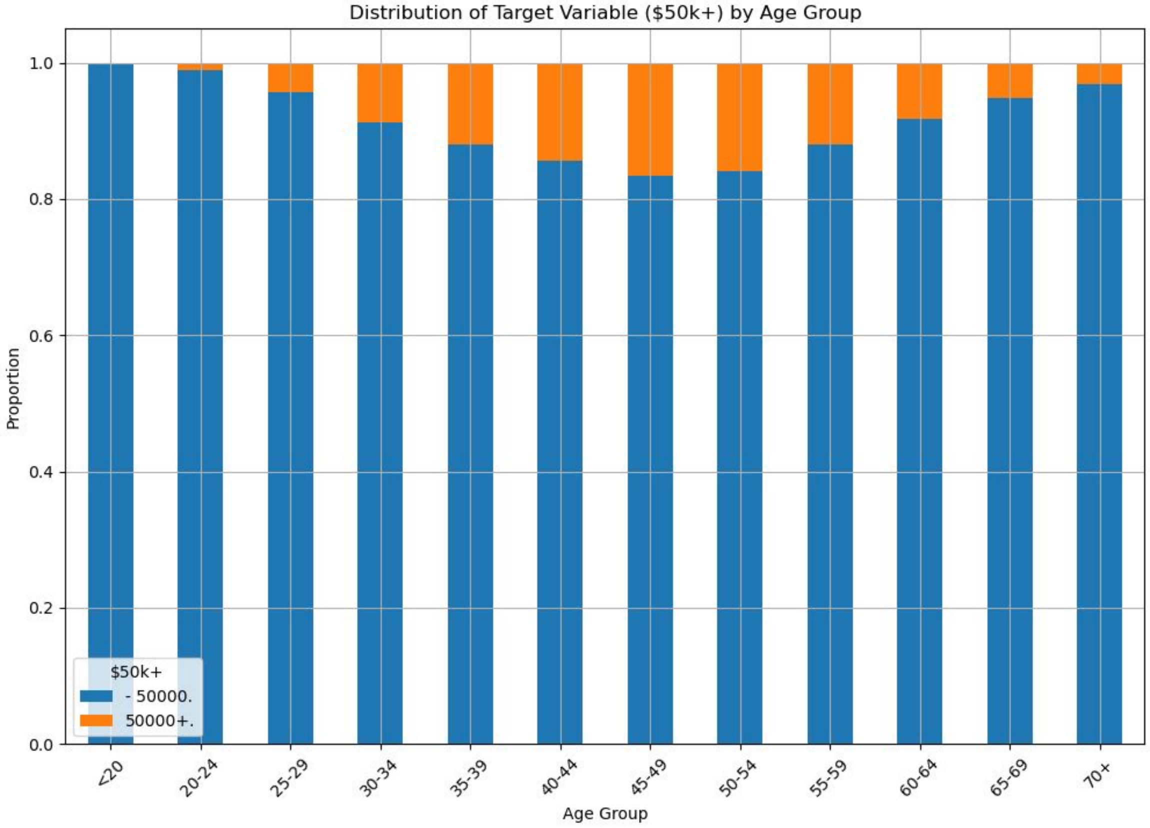
2. The data - key statistics

	Training Data	**Testing Data**
Inputs / features / columns	41	41
Target variable	Income bracket ($50k+)	Income bracket ($50k+)
Rows / Records	199,522	99,761
# of $50k+	12,382	6,186
% of $50k+	6.2%	6.2%

Ages of records

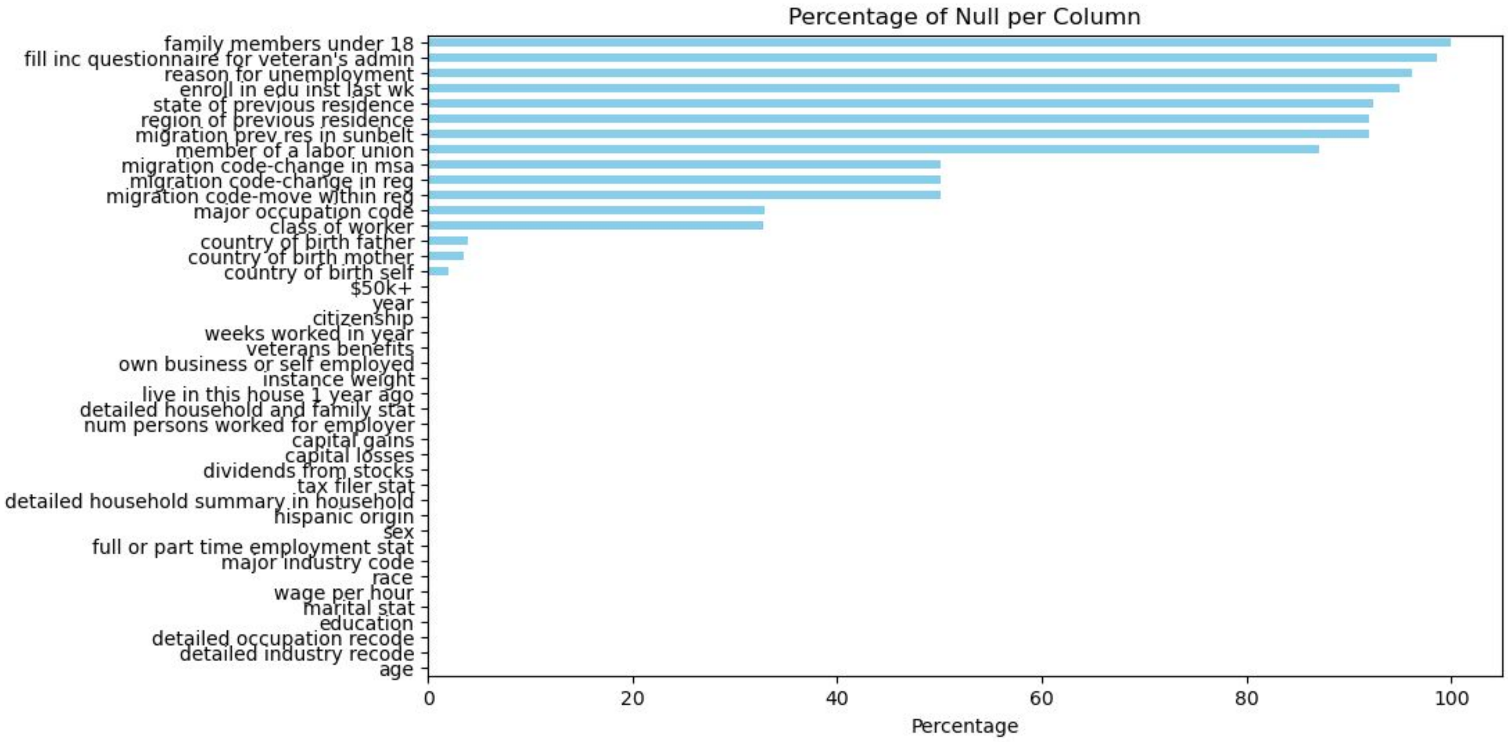


Children very rarely make $50k



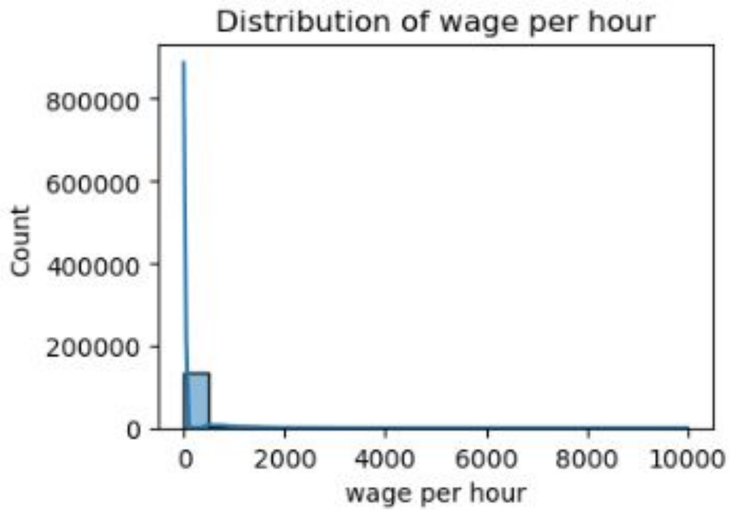
- Vast majority don’t have a job code or industry code
- Not to be included in analysis

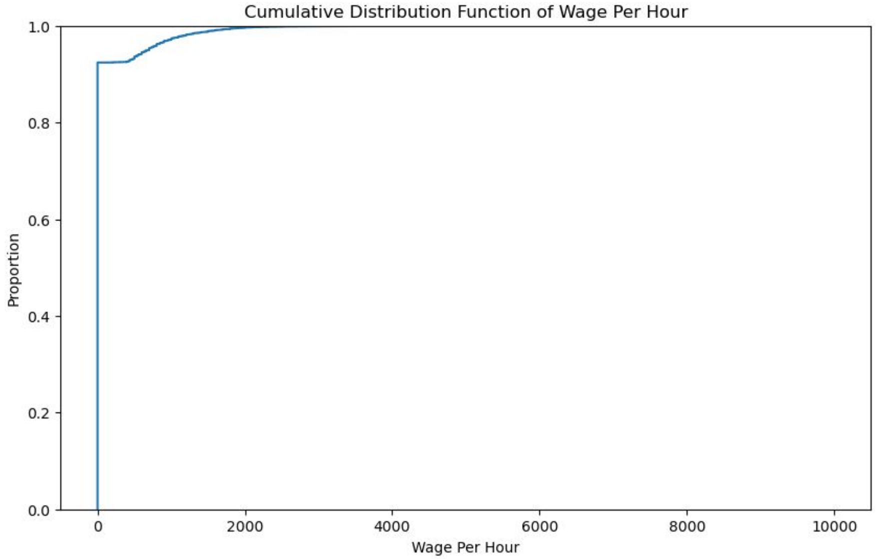
Not in universe - a common code for “nulls”



Question mark (“?”) a common null value for country fields

Skewness of wages

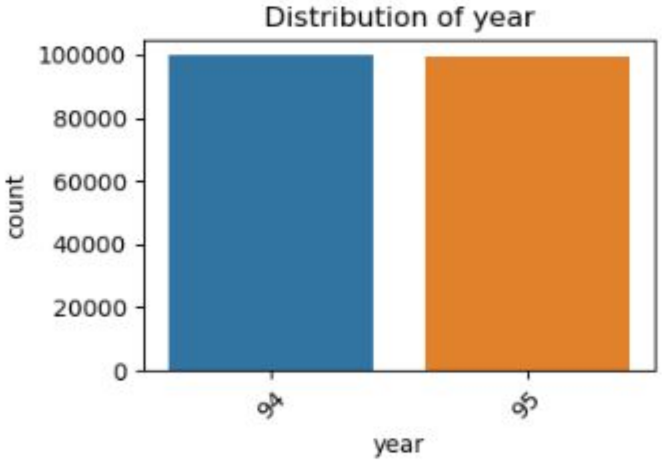


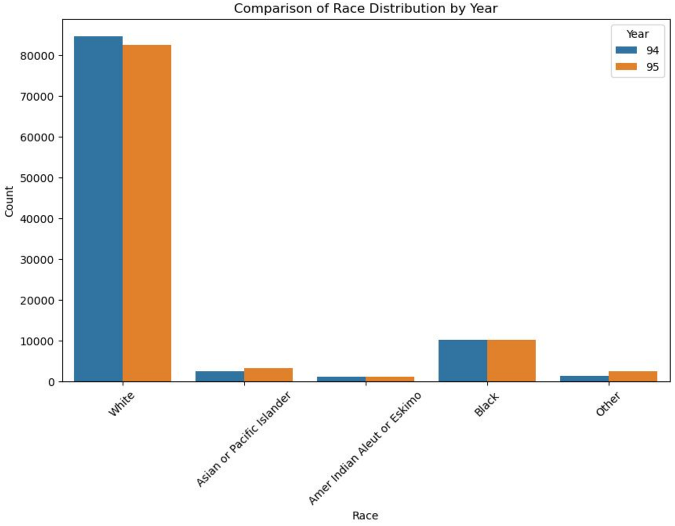


- Other income sources (capital gains, dividends) follow a similar pattern

We have two years of data

- Similarity between the data represented by years ‘94’ and ‘95’ are analyzed using statistical tests
  - Chi-squared & t-tests
- No significant differences
- Extra attention given to protected classes
  - Race, sex, hispanic origin, citizenship, veterans benefits





3. Data Preparation

Data Preparation Challenges

Nulls & Skewness

Drop where >50% null

Impute other nulls

Scale / transform skewed variables

Year column

Not well explained in data dictionary

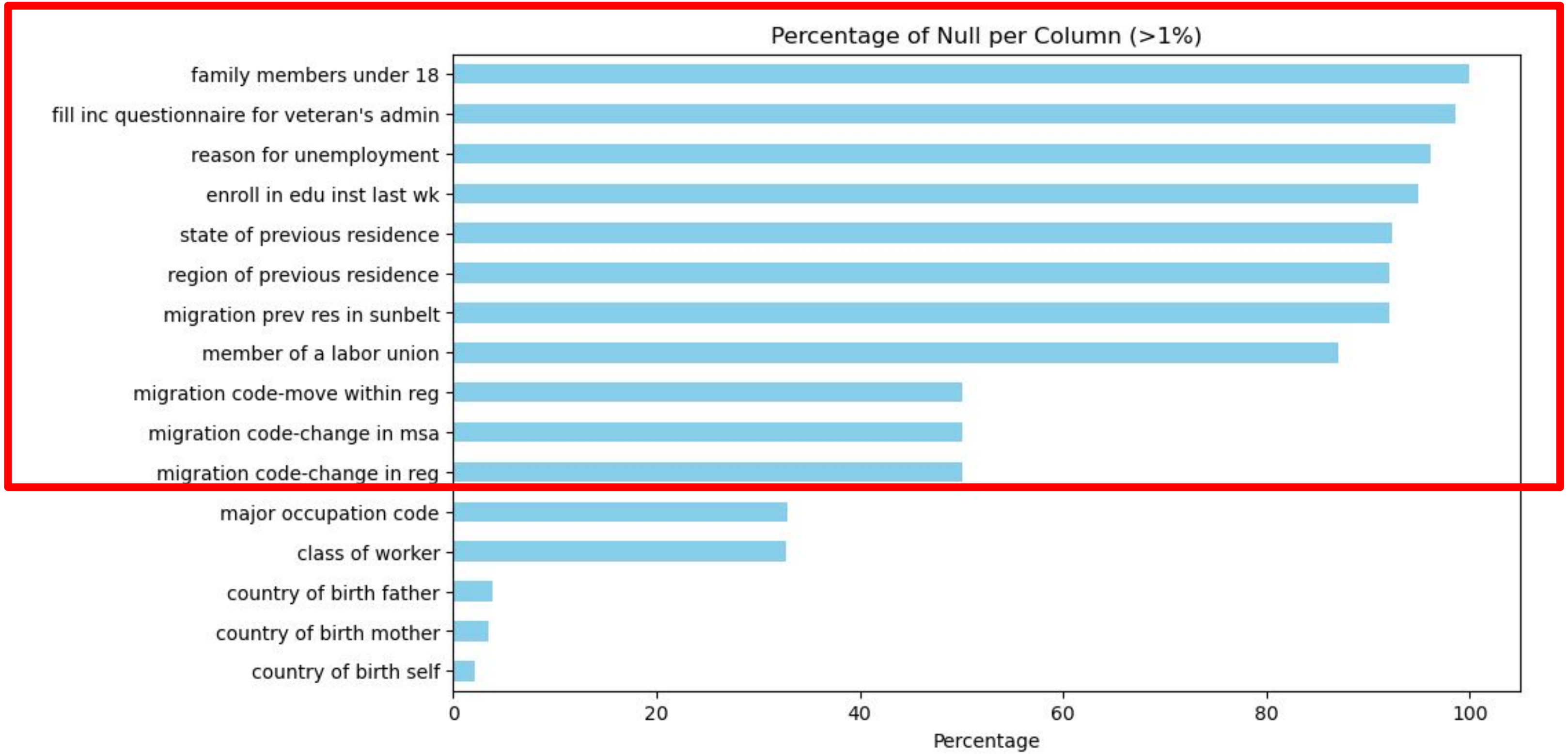
We elect to choose one year (‘94’) as the data seems to be very similar

Sample selection

Who do we want to predict?

We remove children <18

Dropping inputs with >50% null



Feature engineering: annual salary

- Noticeably missing from the dataset is annual salary.

- Assuming 40 hour work weeks, we will use "wage per hour" and "number of weeks worked in year" to estimate annual salary.

4. Data Modeling

Model Selection

- Three models:
  - logistic regression - interpretable, relatively simple
  - random forest - interpretable, robust
  - XGBoost - superior performance, handles missing data better
- Main scoring metric will be AUC (area under curve)
  - Robust when dealing with an imbalanced dataset
- Secondary scoring metric will be recall of >$50k
  - We would like to avoid missing too many high earners

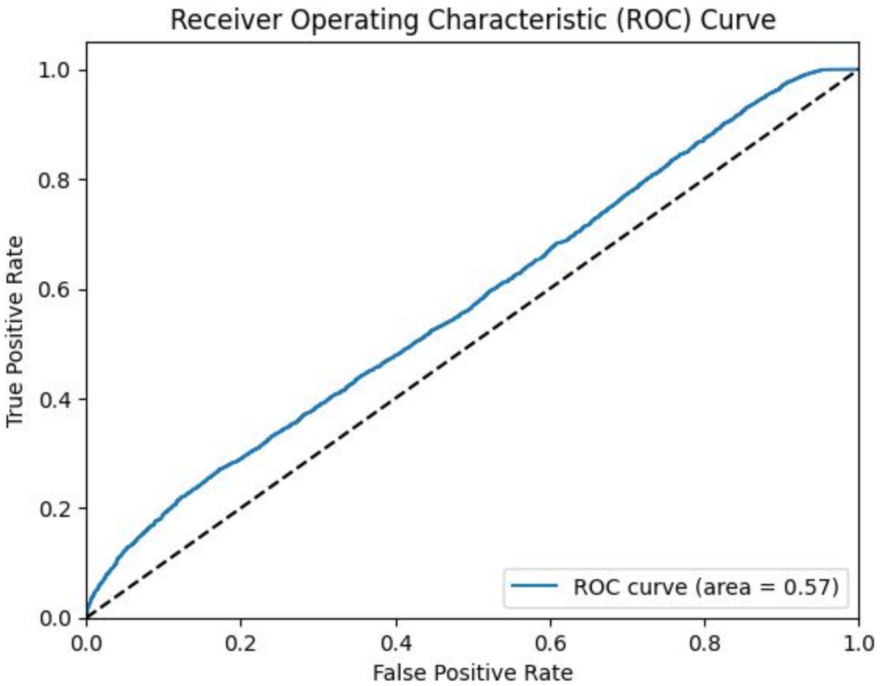
5. Model Assessment

Logistic Regression

	precision	recall	f1-score	support
Over50k	0.19	0.11	0.14	2859
Under50k	0.93	0.96	0.94	33088
accuracy			0.89	35947
macro avg	0.56	0.53	0.54	35947
weighted avg	0.87	0.89	0.88	35947

AUC ROC Score: 0.572879900791758

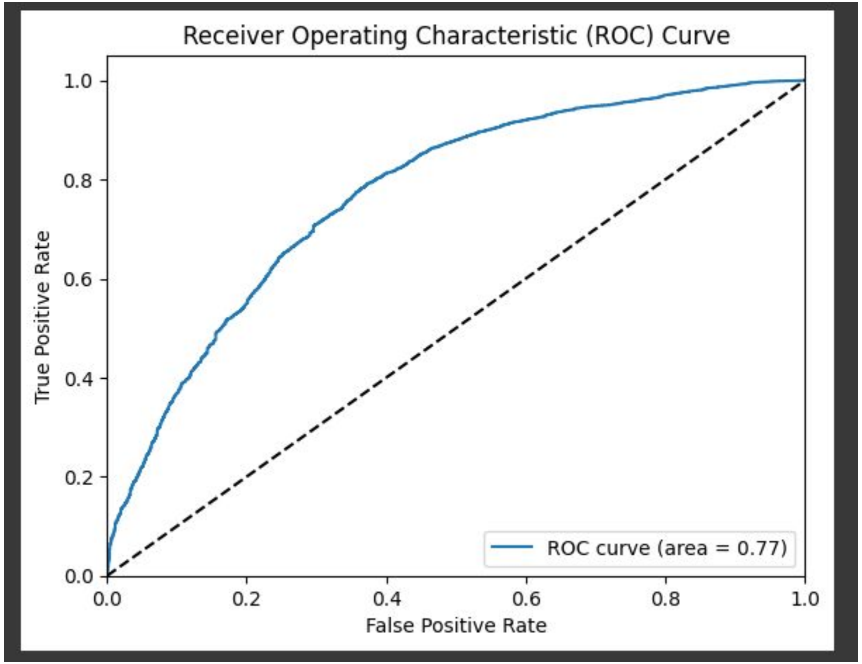
- AUC ROC: 0.57
- Model performs poorly
- 11% of >$50k captured



Random Forest

	precision	recall	f1-score	support
Over50k	0.40	0.15	0.22	2859
Under50k	0.93	0.98	0.95	33088
accuracy			0.91	35947
macro avg	0.66	0.57	0.59	35947
weighted avg	0.89	0.91	0.90	35947

AUC ROC Score: 0.6956459140533509



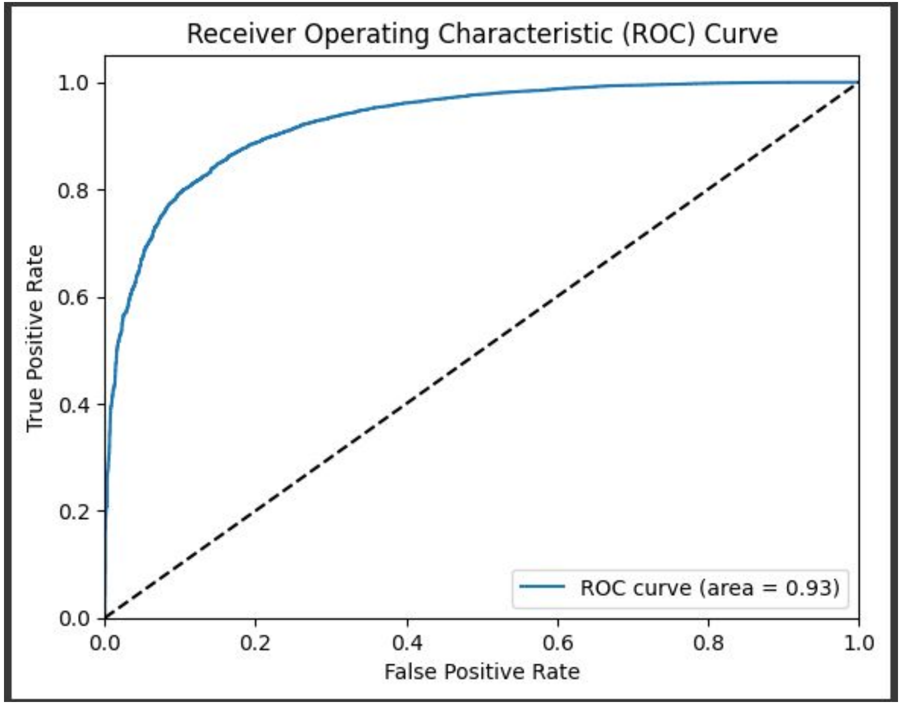
- AUC ROC: 0.77
- Model performance is fair
- 15% of >$50k captured

XGBoost

	precision	recall	f1-score	support
0	0.64	0.52	0.57	2859
1	0.96	0.97	0.97	33088
accuracy			0.94	35947
macro avg	0.80	0.75	0.77	35947
weighted avg	0.93	0.94	0.94	35947

AUC ROC Score: 0.9272389540427831

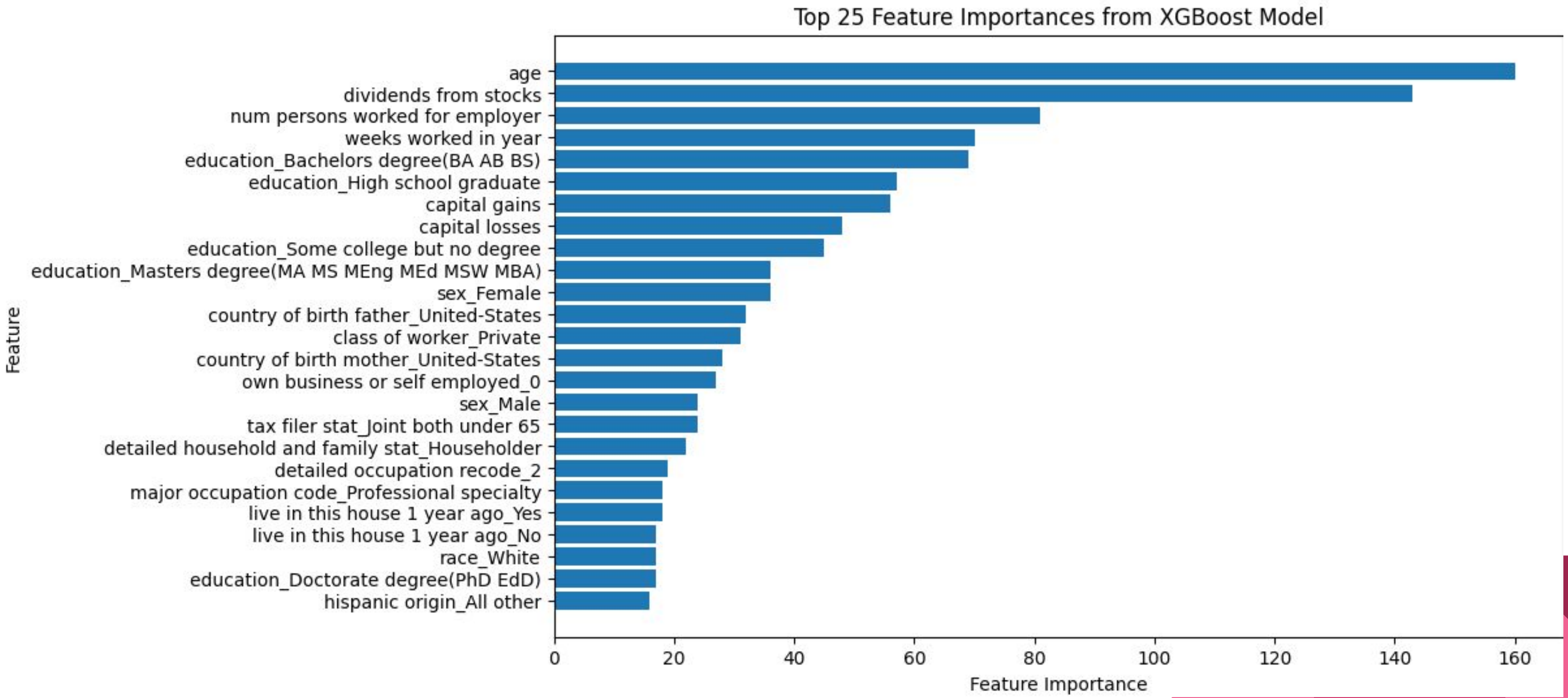
- AUC ROC: 0.93
- Model performance fairly well
- 52% of >$50k captured



Optimizing XGBoost

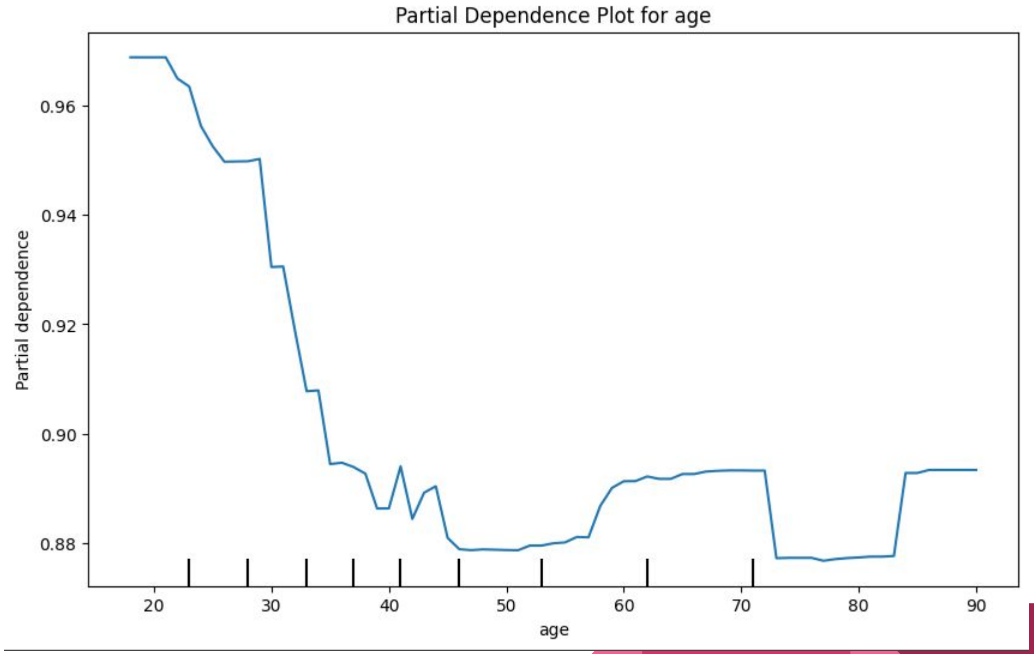
6. Results

Most important variables



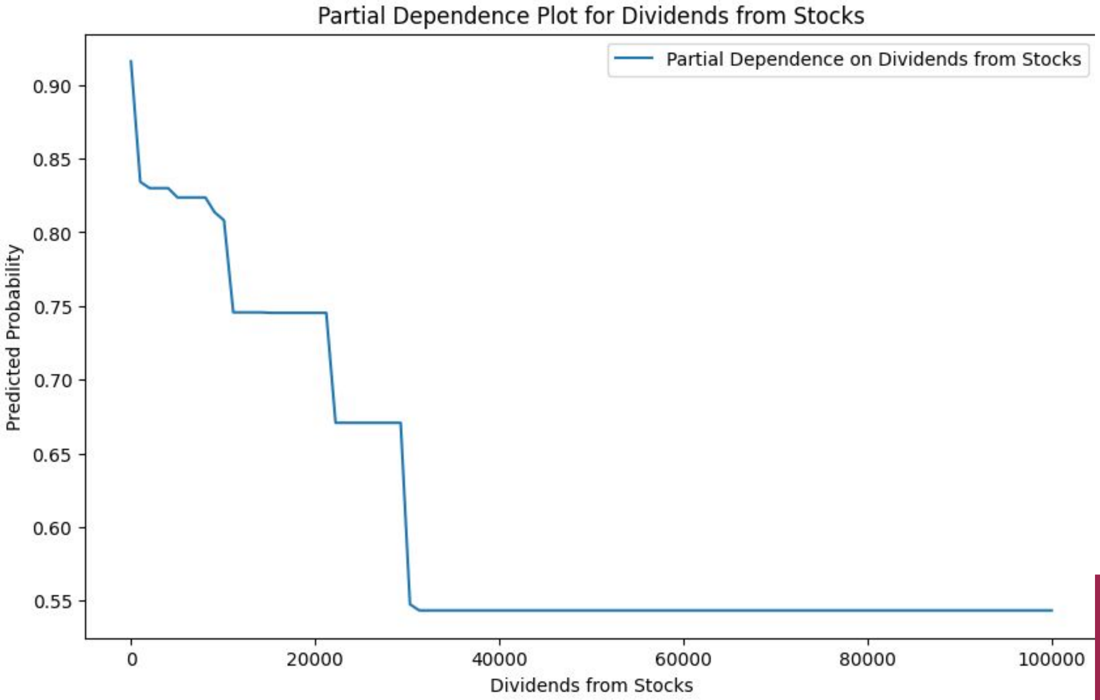
How age impacts income

- Higher plot line means more likely to make less than $50k
- Likelihood of >$50k **increases** as one approaches ~45 years old, then stagnates



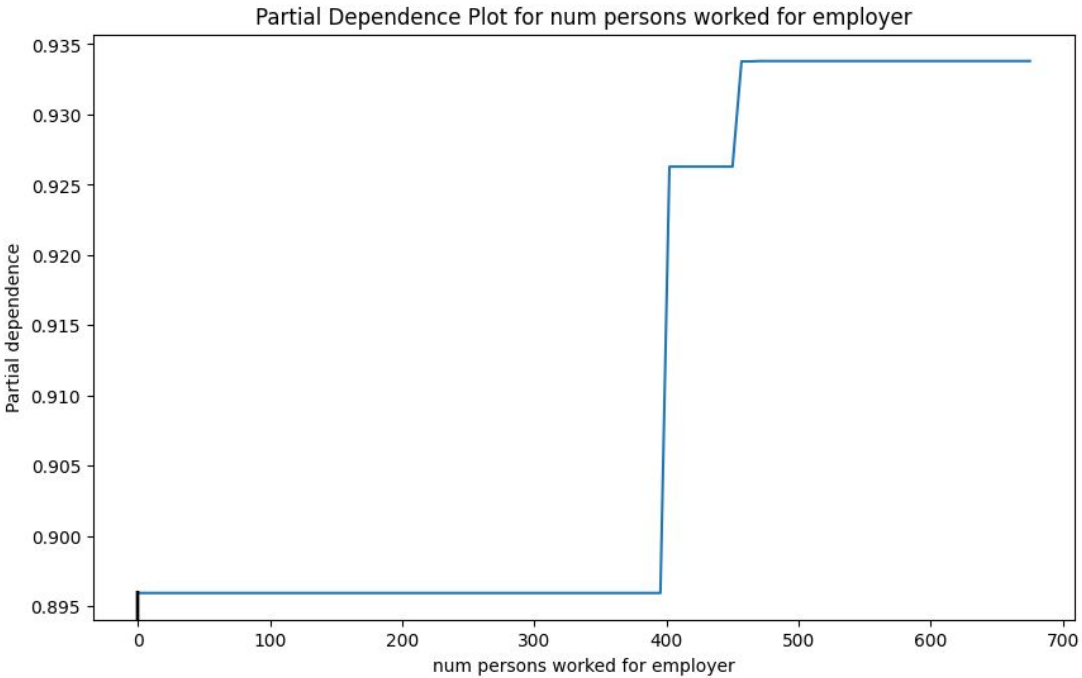
Dividend earners are likely to make >$50k

- Likelihood of >$50k **increases** as one approaches $30k worth of stock dividends annually

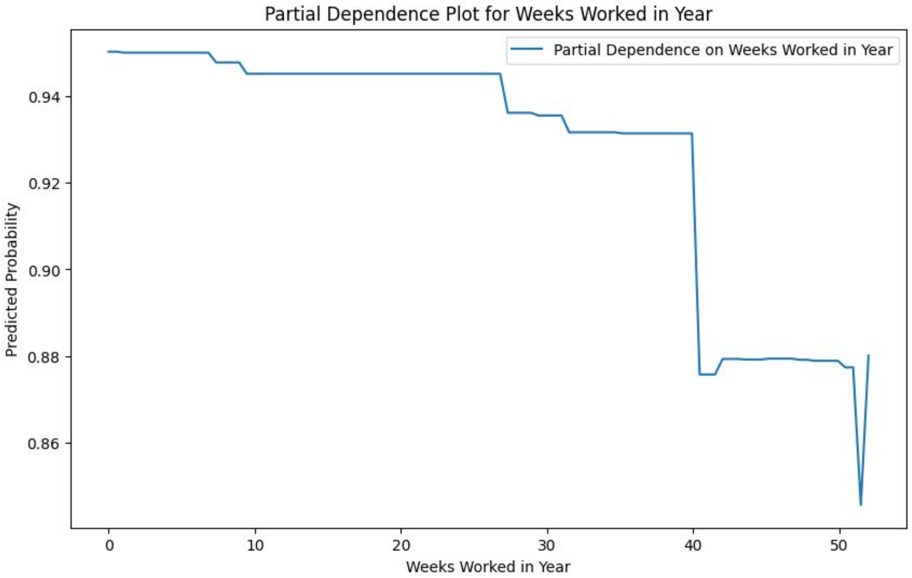


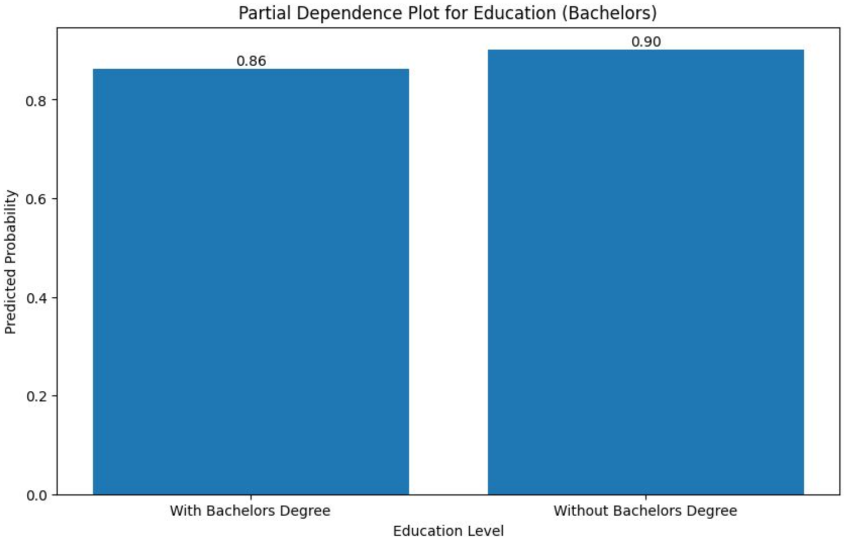
Employees at larger employers less likely >$50k

- Likelihood of >$50k **decreases** as the size of one's employer reaches 450 employees



Weeks Worked a Year and Bachelor’s Holders





Recommendations

- Enable / encourage employees to work 40 hours / week
  - Encourage full-year employment and maximized work week
  - Seasonal employers should creatively find employment for temps
- Promote higher education
  - At least at the bachelor's level, employees have a higher probability of earning >$50k
- Personalized Development Plans (PDPs)
  - PDPs should be based on employee's current skills and their personal as well as business gaps
- Leverage predictive analytics for major employment decisions
  - Promotions, salary changes, development / coaching opportunities
  - And hiring

Considerations for improvements

- Family structure opens many possibilities
- Feature engineering
  - Grouping highschool and middle school dropouts
  - Scaling variables differently
- More models
  - KNN - nearest neighbor analysis
  - Deep Learning

Q&A