Predicting the Income Level of Working Adults in 1994 US

DTSA 5509 Final Project

June 2024
Joseph Bae

Data

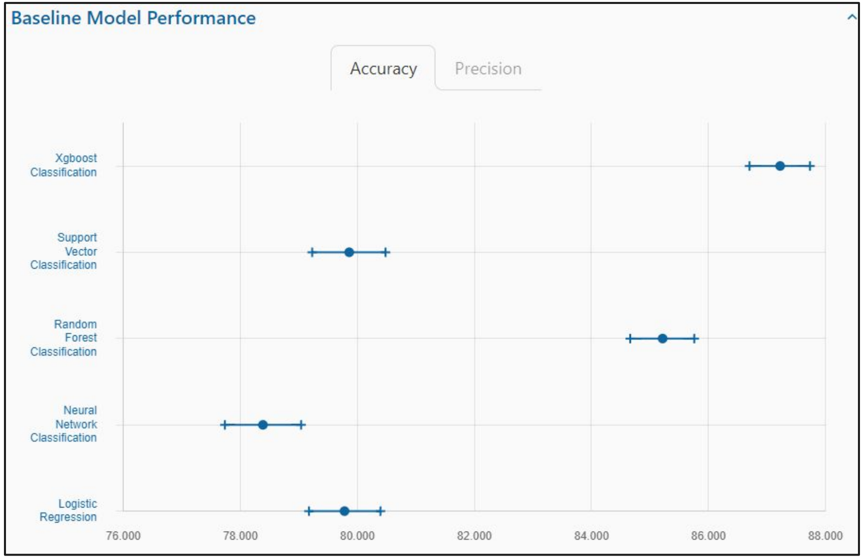
Source file: Adult.DATA[1] from UCI Machine Learning Repository
https://archive.ics.uci.edu/dataset/2/adult

- Data is collected from the 1994 US Census results
- 32,561 samples with 15 data points each
- Small data size: 3.8MB
- Data points revolve around demographics, work, and finances for US working adults.
- The 15th data point is whether each person makes above or below $50K USD annually.
    - **This is what we'll be predicting using a classification model**

[1]Adult. UCI Machine Learning Repository. (n.d.). https://archive.ics.uci.edu/dataset/2/adult

Goals & Motivation

- Use and compare the classification models we learned about throughout the course
- Score an accuracy in line with baselines on UCI (80% - 85% for accuracy metric)



Columns/Features

Variable	Data Type	Variable Type	Description
age	integer	nominal	The age in years of this person.
workclass	string	categorical	Occupation type as far as being self-employed, government worker, unemployed, etc.
fnlwgt	integer	continuous	"final weight", a set of weights given to each observation that represent how many people each observation represents.
education	string	categorical	The highest level of schooling completed.
education-num	integer	ordinal	A numerical representation of the educational column.
marital-status	string	categorical	Different categorizations ranging from single, to married/divorced, or widowed.
occupation	string	categorical	The industry that this person works in.
relationship	string	categorical	The relationship of this person in their family (wife, husband, unmarried, not-in-family, etc.)
race	string	categorical	The ethnicity of this person.
sex	string	categorical	The gender of this person.
capital-gain	integer	continuous	Total capital gain for this person.
capital-loss	integer	continuous	Total capital loss for this person.
hours-per-week	integer	discrete	The number of hours this person works in a week.
native-country	string	categorical	The birth country for this person before coming to the US.
income	string	categorical	Indication of whether this person makes below or above $50K.

Data Cleaning and Munging

- Extra whitespace was removed from all entry values
- About 3.7K samples were removed for having missing values '?' in a few data points
- 'age' and 'hours-per-week' numerical data points were normalized so they add up to 1
- Categorical data points were converted to 0/1 columns (aka dummy variables, one-hot encoding)
    - i.e., a data point with N category values is converted into N separate binary columns, 1 for each category
- Target variable 'income' was encoded as (-1 = below $50k annual) (1 = above $50k annual).
- Data points removed for being vague/undefined, or had many missing values.
    - fnlwgt
    - education-num
    - capital-gain
    - capital-loss
    - native-country

Exploratory Data Analysis (EDA)

DTSA 5509 Final Project

June 2024
Joseph Bae

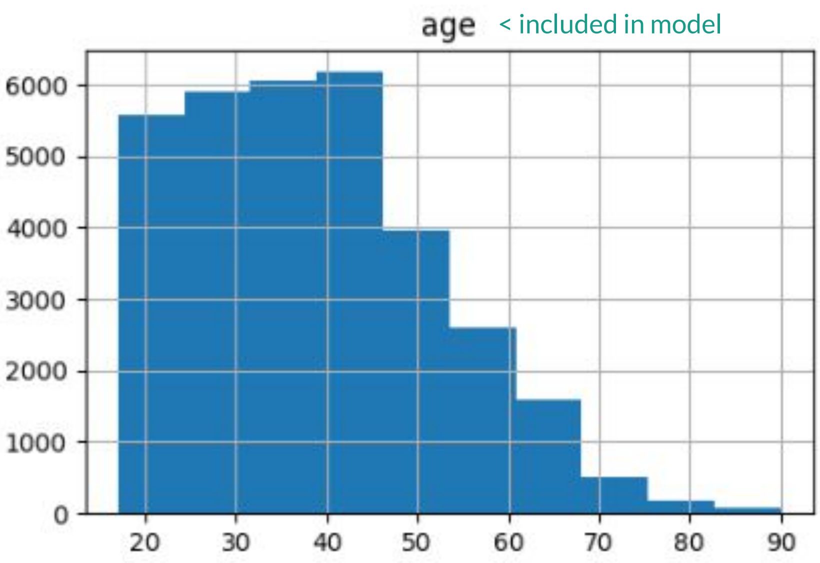
EDA Process

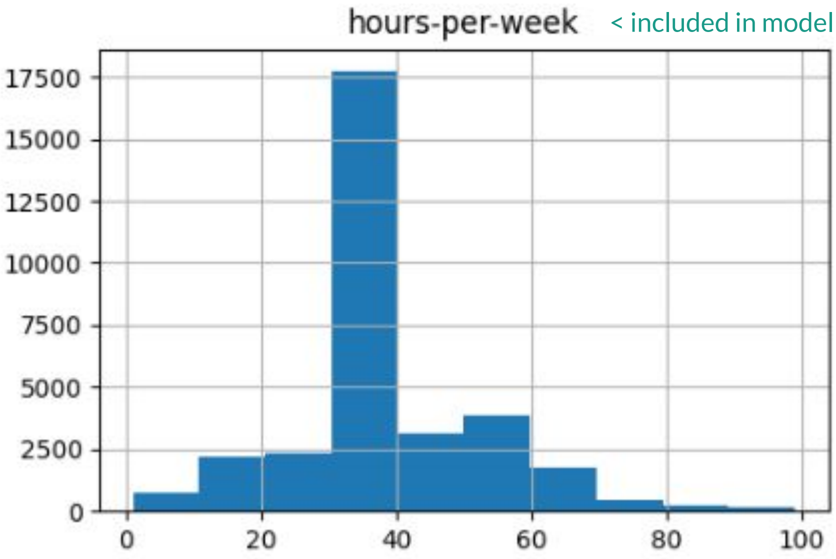
- Plot histogram for numerical variables
- Plot distribution for categorical variables using bar chart
- Assess how realistic the values look, compare it to distributions[2] from Bureau of Labor Statistics
- Keep data bias in mind and determine how much it exists in the dataset
- Justification will be given for removed data points:
  - fnlwgt
  - education-num
  - capital-gain
  - capital-loss
  - native-country

[2](N.d.). *Labor Force, Employment, and Earnings*. https://www2.census.gov/library/publications/1996/compendia/statab/116ed/tables/labor.pdf

Data Point: age and hours-per-week

Data below seems to align with reality of working adults in 1994 US, comparing to BLS publication[2]

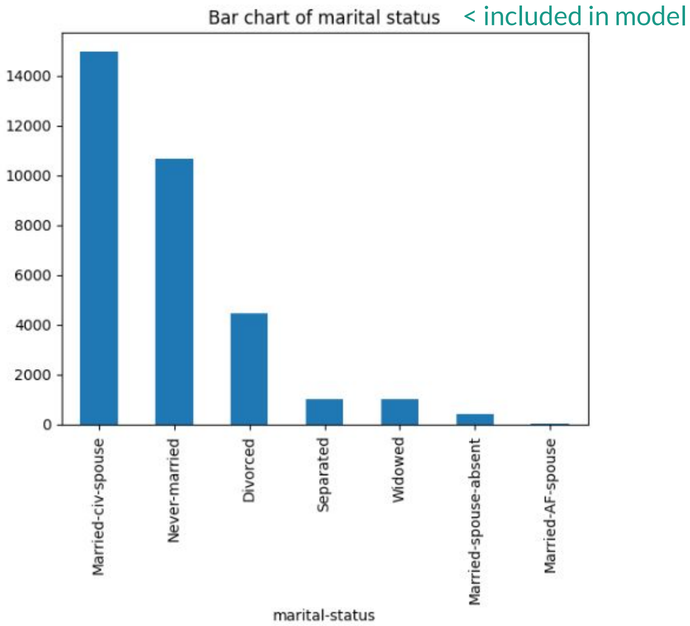




[2](N.d.). *Labor Force, Employment, and Earnings*. https://www2.census.gov/library/publications/1996/compendia/statab/116ed/tables/labor.pdf

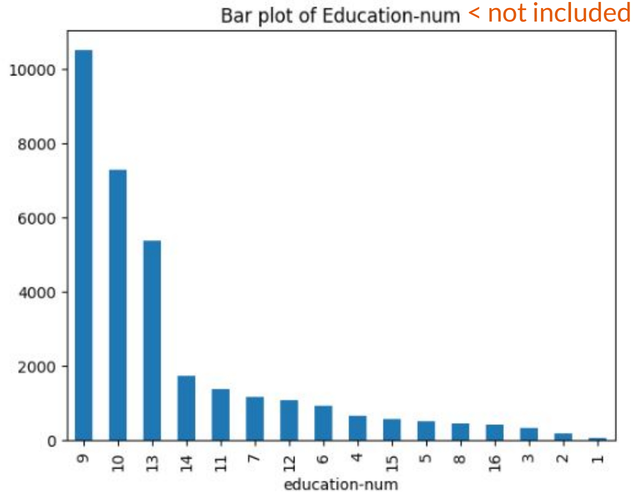
Data Point: marital status

Data below seems to align with reality of working adults in the US

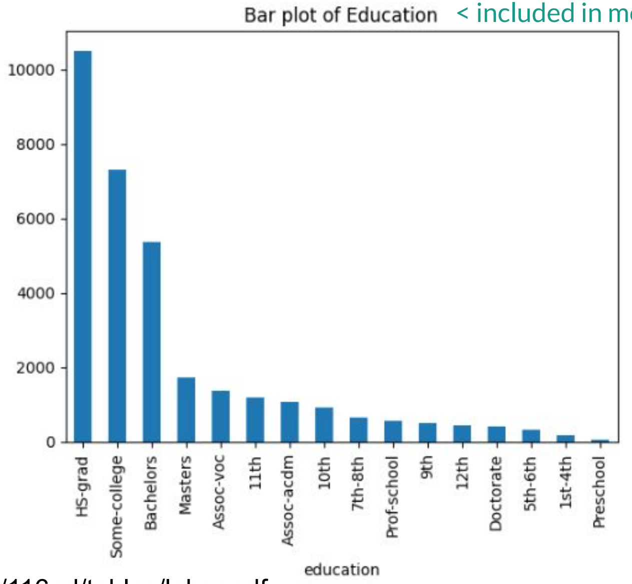


Data Point: education-num and education

- Education-num is a 1-to-1 encoding of Education, and can be excluded from the model.
- Category definitions are not that clear
- Breakout seems realistic for a workforce in the 1994 timeframe and align with stats from BLS[2]



< not included in model

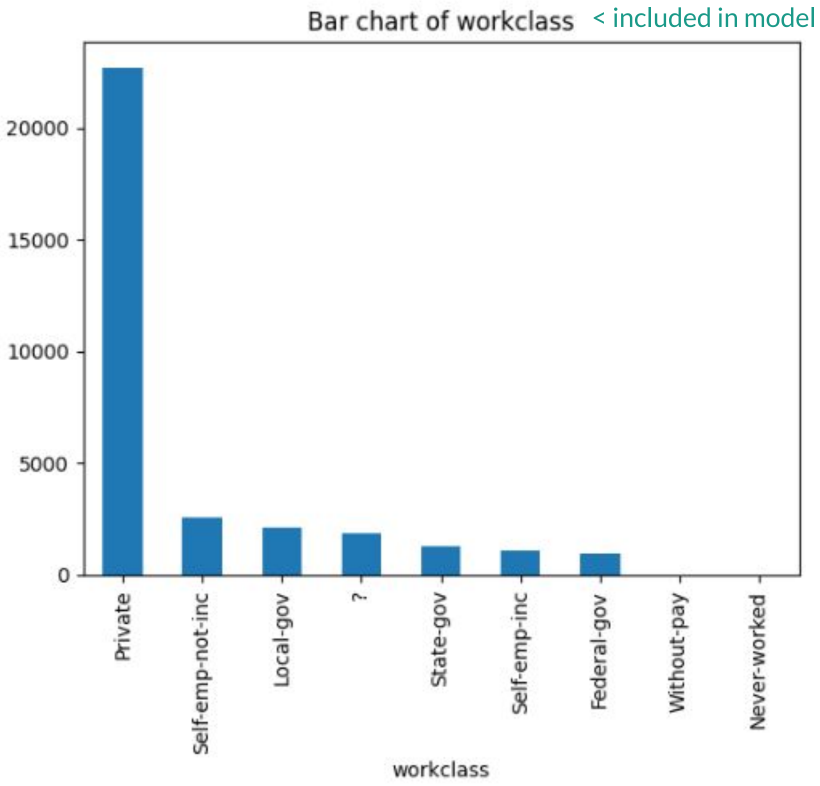


< included in model

[2](N.d.). *Labor Force, Employment, and Earnings*. https://www2.census.gov/library/publications/1996/compendia/statab/116ed/tables/labor.pdf

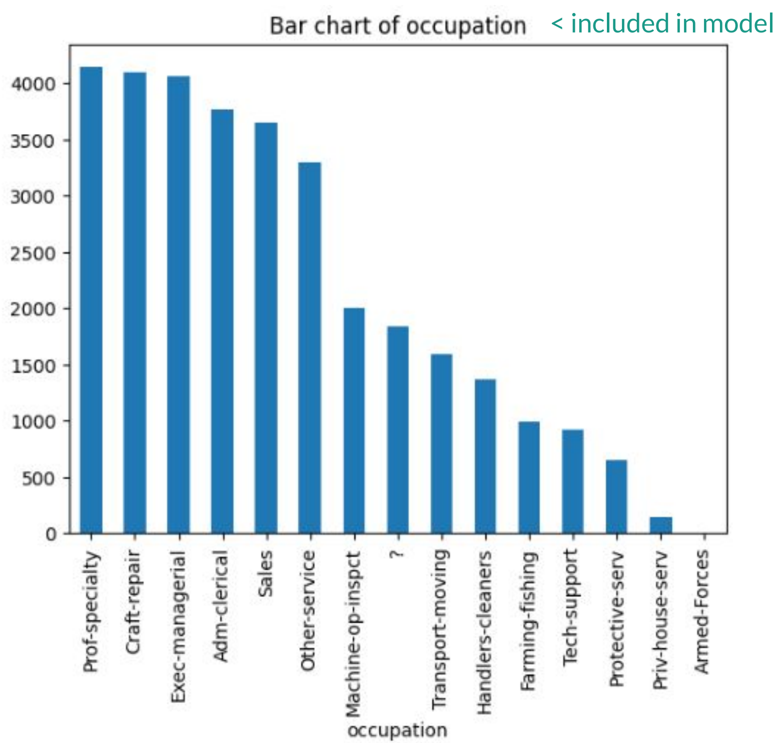
Data Point: workclass

- The vast majority of employees are Private employed which is realistic
- Other groups are similar in size and seem accurate enough for purposes of this project
- 1,836 missing values '?'



Questionable Data Point: occupation

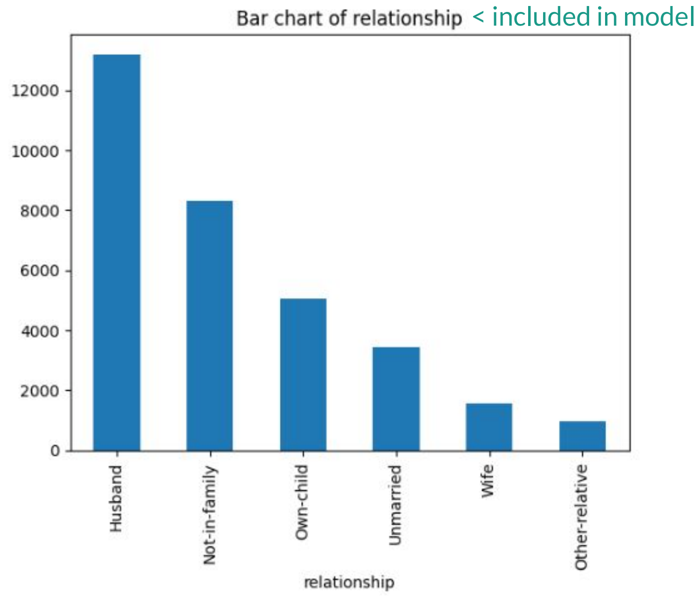
- Distribution roughly matches up with BLS publication[2] in order, but not proportions
- Catch-all groups that aren't broken out
  - Prof-specialty
  - Other service
  - Priv-house-serv(ice)
- Armed-Forces are underrepresented
  - Likely not surveyed
- Representation by occupation is questionable
  - Are there really more exec-managerial than sales?
  - More handlers-cleaners than farming-fishing?

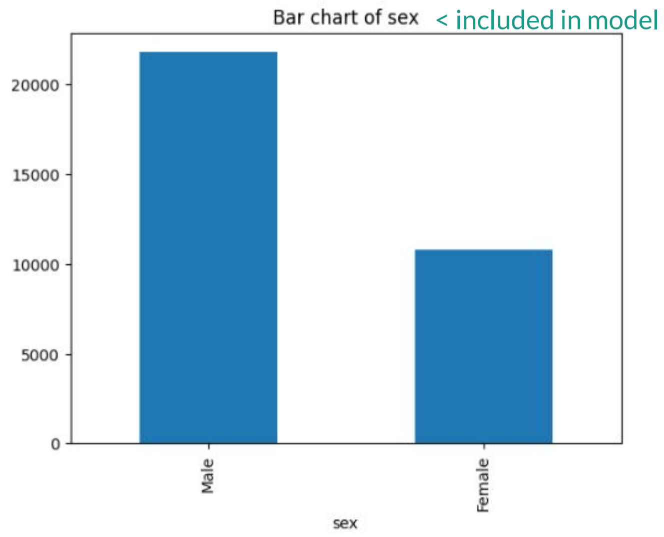


[2](N.d.). *Labor Force, Employment, and Earnings*. https://www2.census.gov/library/publications/1996/compendia/statab/116ed/tables/labor.pdf

Questionable Data Point: relationship and sex

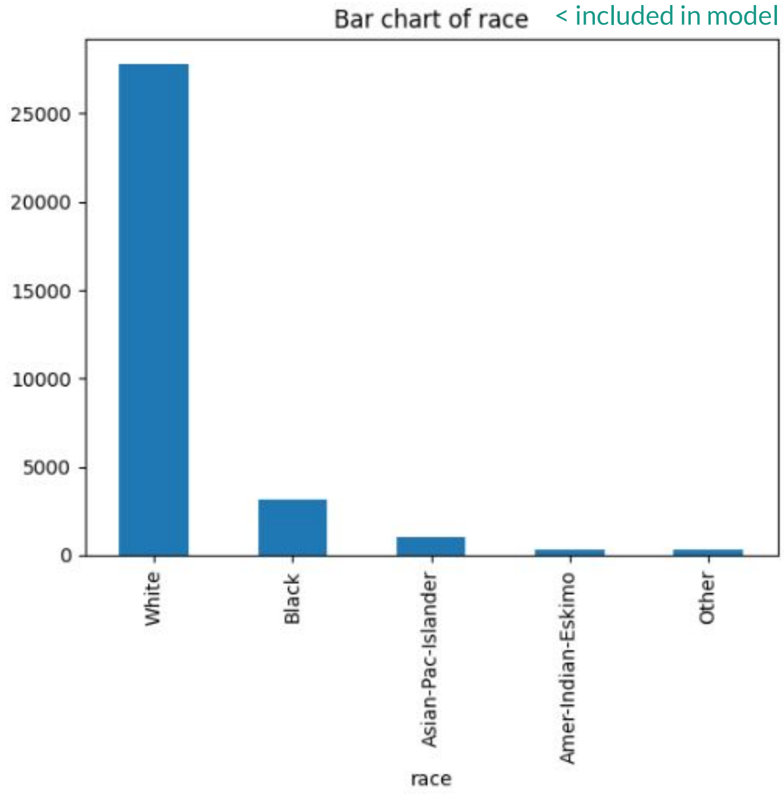
- Unclear definition of Not-in-family, Own-child
- Husband is overrepresented, and Wife are underrepresented
  - Model will be better trained for husbands than wives, and similarly for males than females
  - Doesn’t match with BLS stats which has males and females closer to a 55:45 split





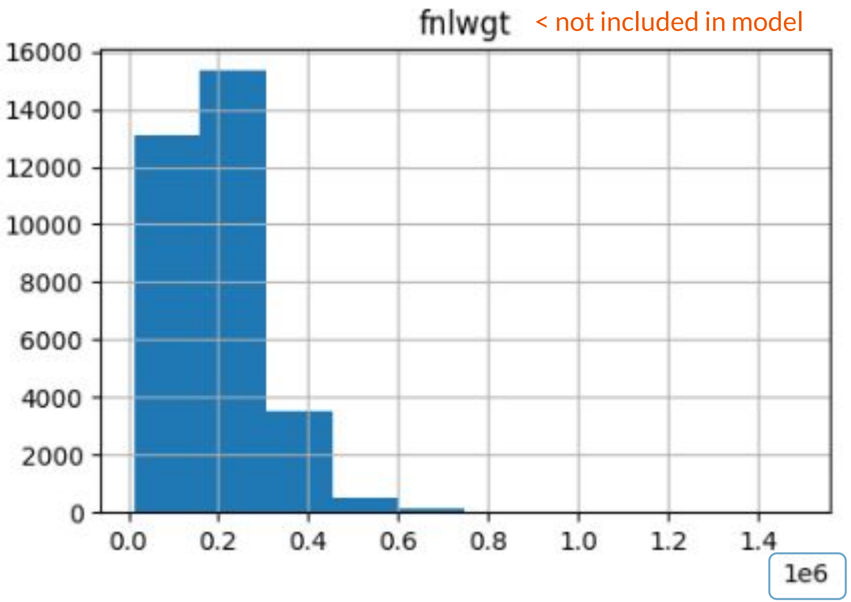
Questionable Data Point: race

- White is overwhelming majority
- Proportions seem aligned with BLS publication
- Data bias: other races are underrepresented
- Acknowledge the model will be best trained for White adult workers



< included in model

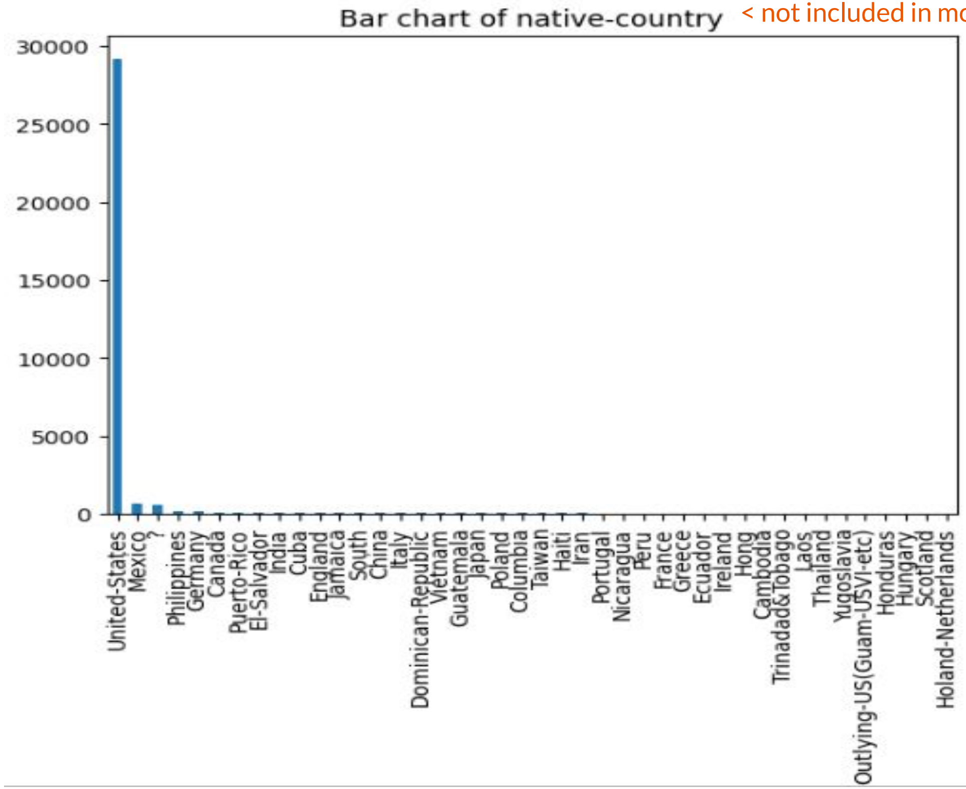
Bad Data Points - not included in the model



fnlwgt

- Represents "final weight", and how many people match each observation
- Not fully documented, and definition is unclear e.g., taking the sum of this column exceeds the population of the US by a great margin.
- Not interpretable

Bad Data Points - not included in the model



< not included in model

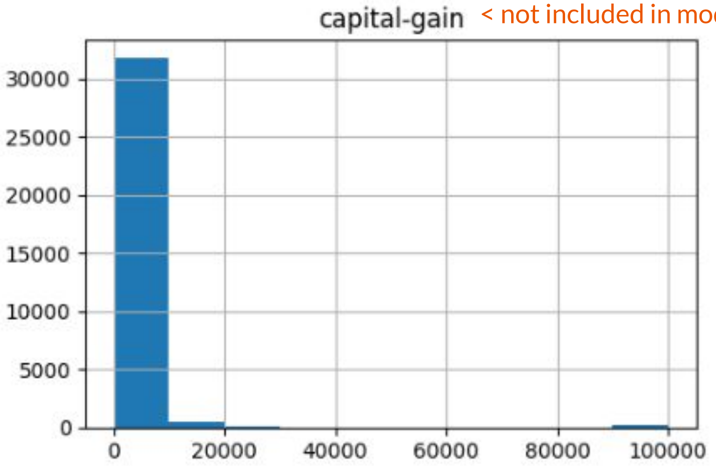
native-country

- Nearly all records are workers born in the United States
- Data bias: immigrant workers are underrepresented
- Regardless of bias, this is not a useful data point as nearly everything is a single value.

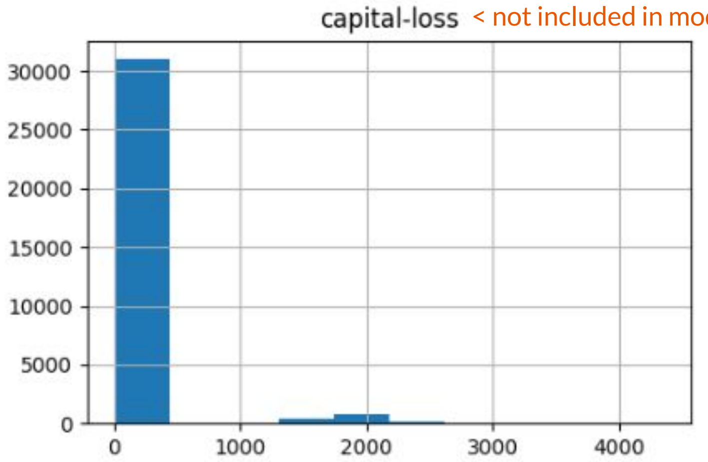
Bad Data Points - not included in the model

capital-gain and capital-loss

- The majority of these data points are zeros
- Doesn’t add any helpful information to the model



< not included in model



< not included in model

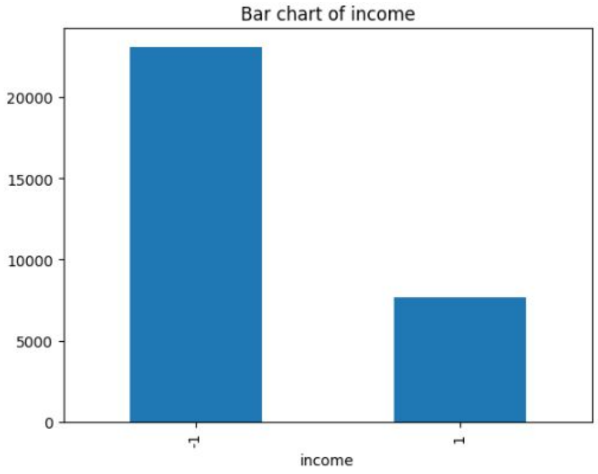
Class Balance and Correlation

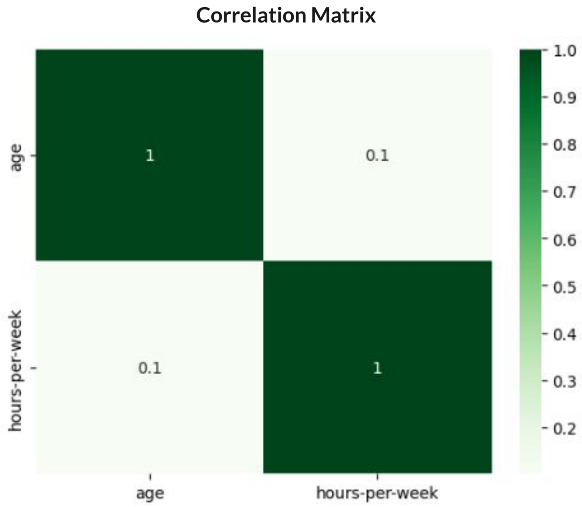
DTSA 5509 Final Project

June 2024
Joseph Bae

Class Balance and Correlation

- Majority of income labels are -1, so indicates a class imbalance though not severe.
    - Some algorithms sensitive to imbalance like decision tree will be affected.
    - Utilize class_weight = ‘balanced’
- There is little correlation (0.1) between age and hours-per-week √
    - We don’t need to be concerned with issues regarding collinearity affecting our models.





Model Building and Tuning

DTSA 5509 Final Project

June 2024
Joseph Bae

Models Trained and Compared

Logistic Regression	K-Nearest Neighbors - KNN
Single Decision Tree	Random Forest (bagging)
Ada Boost (boosting)	Gradient Boost (boosting)
Support Vector Machine - SVM	

Approach to building each model

- Training and test datasets were the same for each model
- Whenever possible, class_weight = ‘balanced’ was use to account for the class imbalance.
- Grid Search with 5-fold cross validation was done to find the correct hyperparameters.
- Once hyperparameters were found, model was trained and CPU time taken was recorded.
- Accuracy and F1 score were computed, and F1 score was used to assess and compare models.

Model	Hyperparameters found via Grid Search	Training Accuracy	Test Accuracy	Training F1 Score	Test F1 Score	Training Time
Logistic Regression	C = 15.56	82.8%	82.6%	61.3%	60.6%	63 ms
KNN	n_neighbors = 24	84.0%	83.2%	65.4%	63.6%	16 ms
Single Decision Tree	max_depth = 5, max_leaf_nodes = 10	74.3%	73.7%	60.9%	59.7%	47 ms
Random Forest	n_estimators = 11, max_depth = 8, max_leaf_nodes = 12	72.7%	72.3%	60.7%	59.6%	63 ms
Ada Boost	n_estimators = 100, learning_rate = 1, max_depth (base tree) = 2	82.8%	83.1%	60.1%	60.3%	1880 ms
Gradient Boost ☆	loss = exponential, learning_rate = 1, n_estimators = 100, max_depth = 2	84.4%	83.7%	66.8%	65.4%	1480 ms
SVM	kernel = rbf*, C = 10, gamma = 0.1	83.8%	82.7%	64.0%	61.2%	2080 ms

** Grid search for SVM was taking very long, so kernel was run through grid search separately from other parameters.*

Closing thoughts

- GradientBoost performed the best and had acceptable runtime.
- SVM took the longest to train, and didn't perform better than models that took shorter time.
- KNN was the quickest to train, and performed well (2nd best).
- GridSearchCV started to take a while (few minutes) for the boosting algorithms, and almost an hour for SVM.
- A stronger understanding of acceptable ranges for each parameter would likely allow me to be more efficient with GridSearch, and find better hyperparameters particularly for SVM and decision tree.

Github

https://github.com/jDyn90/dtsa5509

- Adult.DATA file
- Jupyter notebook with full code used for data cleaning, munging, EDA, and model building & tuning
- PDF of this slide deck