Adult Income Prediction

Nena Esaw

Project Description

- **This project aims to predict whether an adult's income will exceed $50,000, leveraging attributes such as occupation, education, age and other demographics.**

- **The Adult Income dataset was extracted from the 1994 US Census database. The dataset is a repository of 48,842 entries.**

Adult Income Dataset

The dataset was sourced from Kaggle - Adult Income Dataset.

For this dataset, there were 48,852 rows and 15 columns.

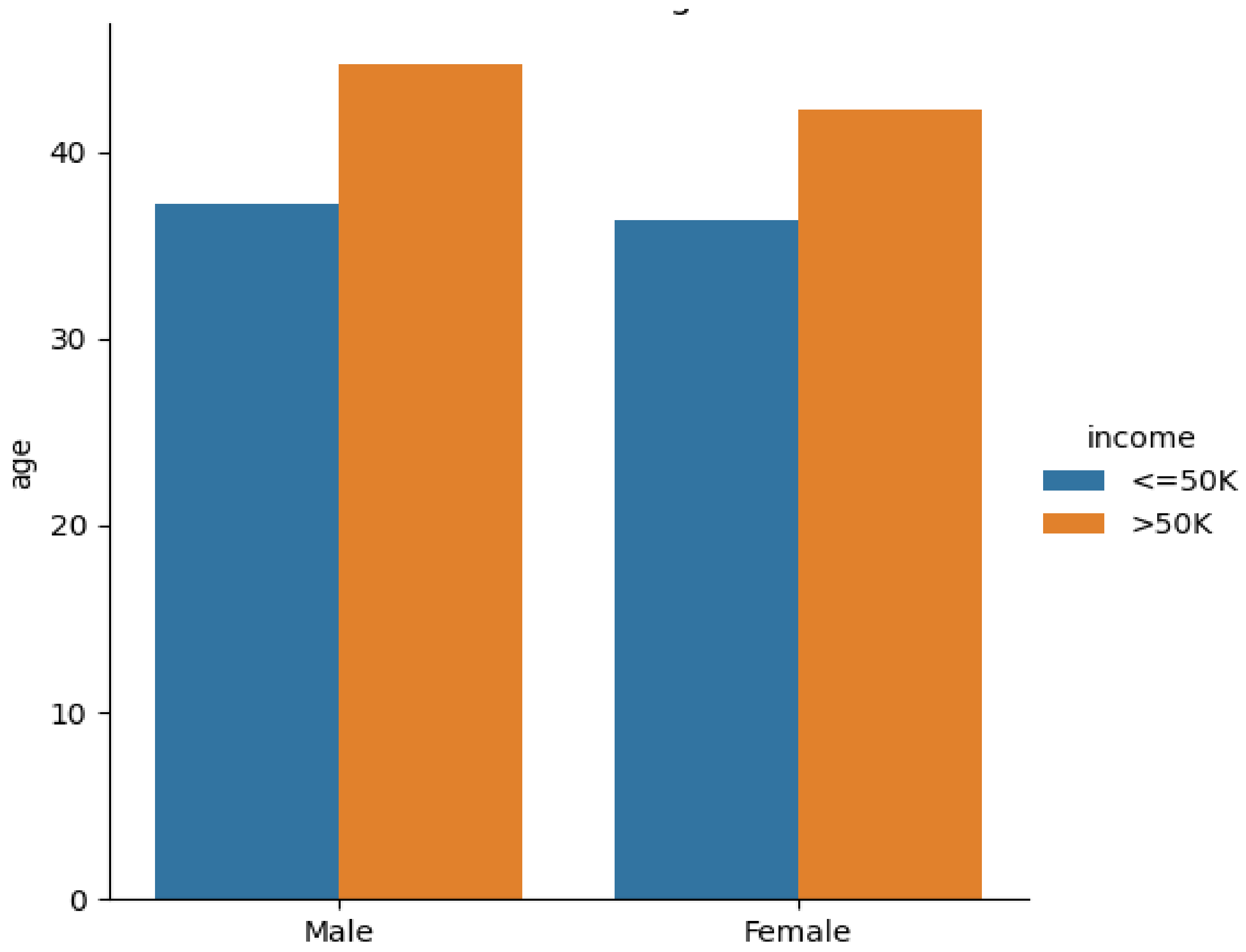
Age: The age of the individual	Workclass: Private, Self-emp-not-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
FNLWGT: Final Weight	Education: Preschool - 12th, HS-Grad, Prof-school, Assoc-acdm, Some-college, Bachelors, Masters, Doctorate
Education-Num: Education in numerical form	Marital Status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
Occupation: Tech Support, Craft-Repair, Other-Service, Sales, Exec-Managerial, Prof-Specialty, Handlers-Cleaners, Machine-op-inspect, Adm-Clerical, Farming-Fishing, Transport-moving, Priv-house-serv, Protective-Serv, Armed-Forces	Relationship: Wife, Own-Child, Husband, Not-in-family, Other-Relative, Unmarried
Race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black	Sex: Male, Female
Capital-Gain: Capital gain for an individual	Capital-Loss: Capital loss for an individual
Hours-per-week: The hours an individual has reported to work per week	Native-Country: Country of origin for an individual

Stakeholders

This predictive analysis delves into the intricate dynamics of adult income. By dissecting several demographics and employing advanced methodologies, this study aims to unravel the underlying patterns that govern the attainment of incomes in the excess of $50,000. This analysis will serve individuals seeking to understand the contributing factors behind earning a salary suppressing $50,000.

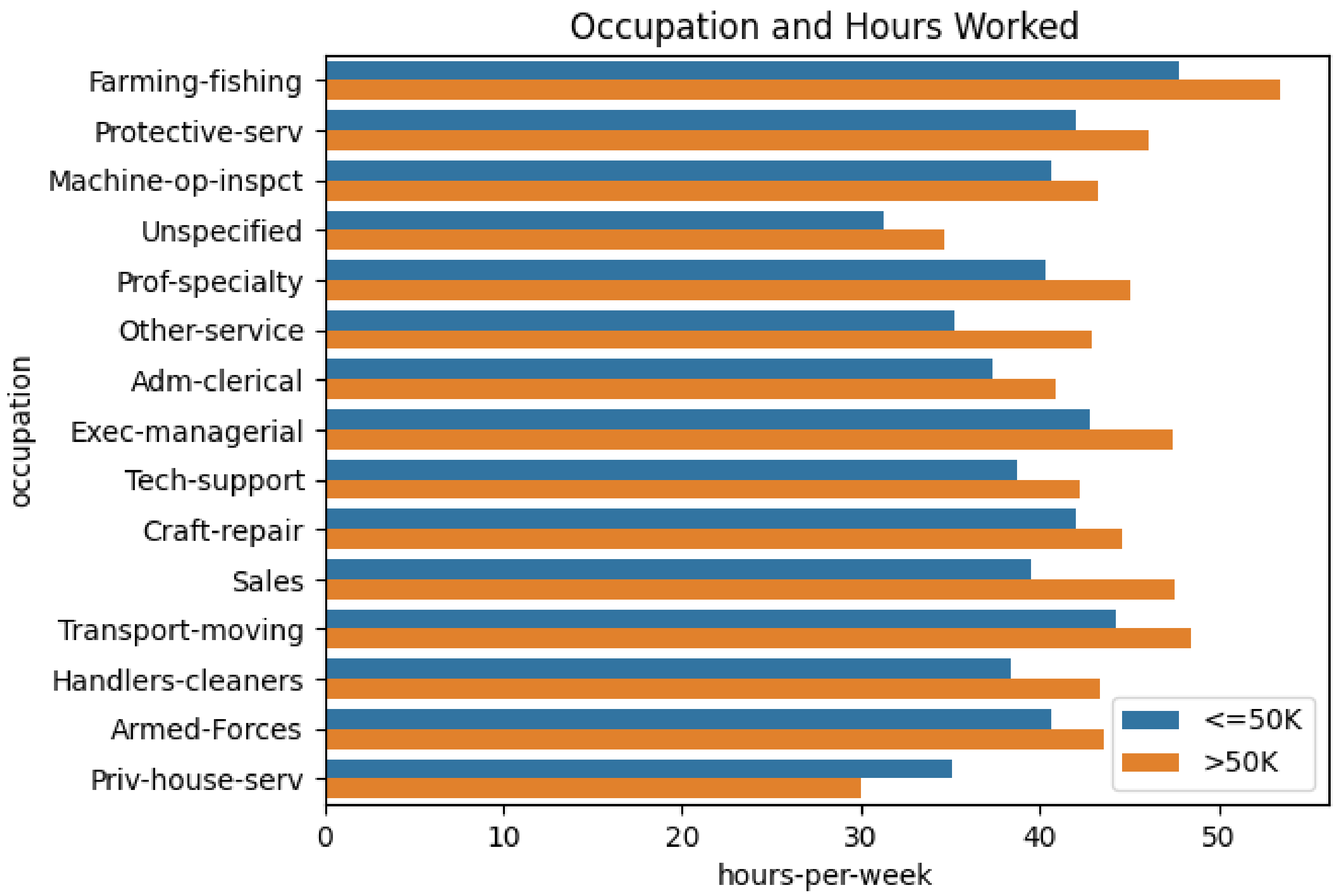
Key Findings

- We can see that more male over the age of 40 generate a higher income over 50k than compared to females.
- Females over the age 40 make over 50k as well but not as much as males.
- We can determine that over the over of 40 for both male and females they generate a higher income.



Key Findings

- This graph shows up that Farming-fishing occupation generates the highest income and work more than 50 hours per week.
- We can also see that most of the work class that work over 40 hours per week generate over 50k income.



Key Findings

- The KNN predictive model that used showed limitations in accurately predicting the income of individuals earning over $50,000.
- The model was more biased towards salaries that were below $50,000.
- The bias might have stemmed from the inherent imbalance in the dataset, where instances with salaries below 50k significantly outnumbered those above.
- Addressing this bias and striving for a more balanced model output becomes crucial to ensure equitable predictions across income categories.

Challenges

- Despite implementing class balancing techniques such as SMOTE, the data remained highly imbalanced, with no noticeable improvements.
- There were columns that had to be removed because of missing data, such as capital gains/losses. There isn't a lot of information explaining if the data is missing because individuals did or did not have this information.
- Workclass and occupation had data that was unknown or missing.
- Working with small dataset caused limitations.

Recommendations

- The KNN model might predict better if the data were more balanced. There were about 4330 values that were unknown or missing data. Having a more accurate dataset would improve the predictions for this model.
- Incorrectly estimating a person's income to be over $50,000 should not be used with this model due to the high false negatives.
- Using current data, the data used was from 2014. Might not have as much missing data.