FAIR INCOME PREDICTION: ADDRESSING GENDER BIAS IN AUTOMATED DECISION-MAKING

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

STORY: SALARY RECOMMENDATION SYSTEM

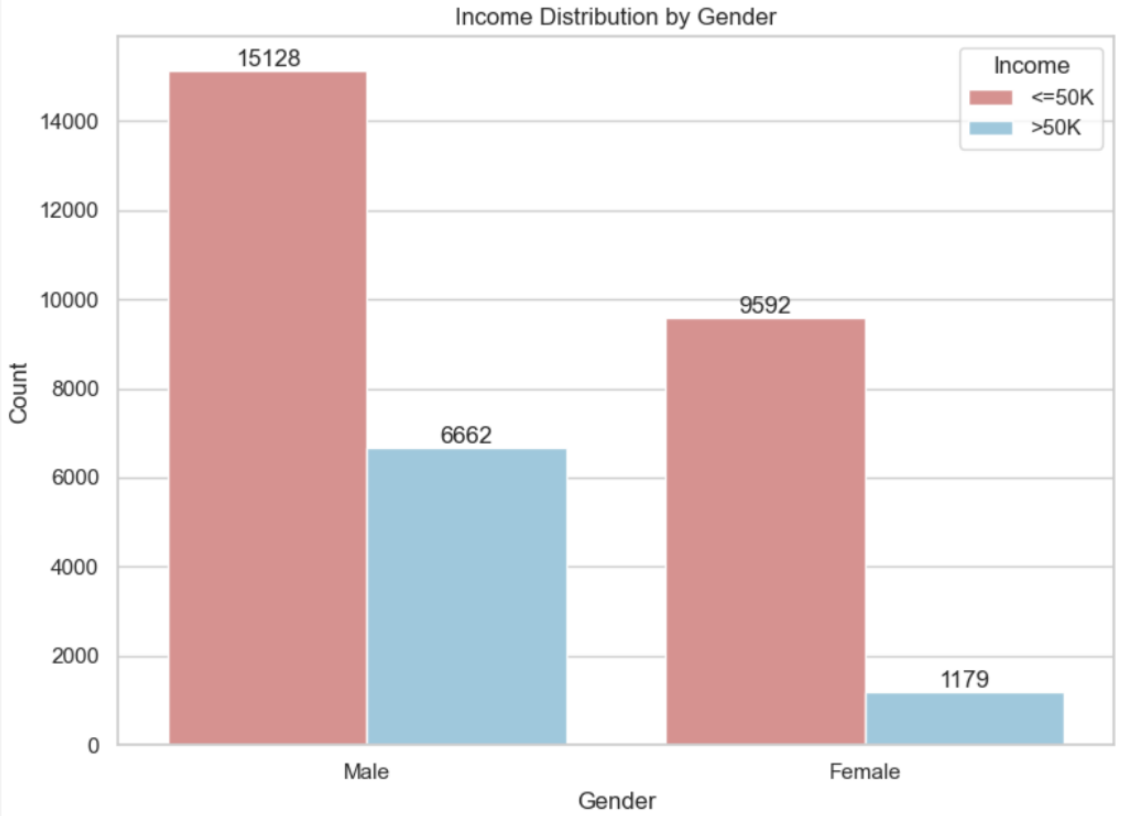
- A recruiting platform uses ML to predict whether job candidates are likely to earn >$50K annually. This prediction influences:
- Initial salary offers
- Job level recommendations
- Benefits eligibility

We have been tasked to create a model to predict high-income earners based on demographic and employment-related features, helping the platform automate income classification and better understand candidate salary patterns.

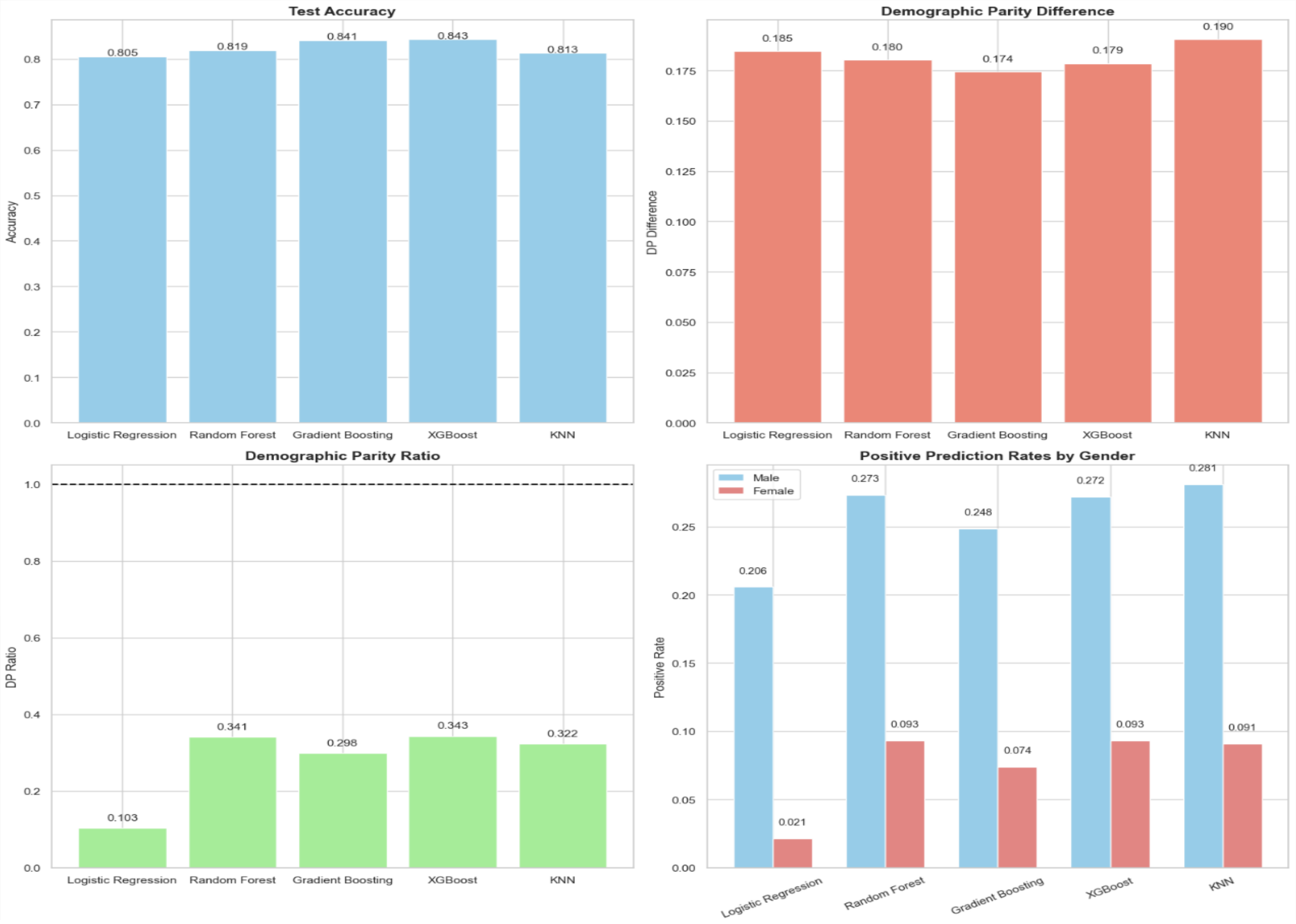
ADULT /CENSUS INCOME DATASET

- Task: Predict whether an individual earns > $50K per year
- Number of instances: 48,842 records
- Number of features: 14 attributes (numerical and categorical)
- Target variable: income (binary: ≤50K or >50K)
- Sensitive attribute: Gender (Male / Female)
- Key features:
  - Age, Workclass, Education level,
  - Marital status, Occupation, Relationship,
  - Race, Hours per week, Native country
- Mostly clean data but contains some missing values (represented as ?)

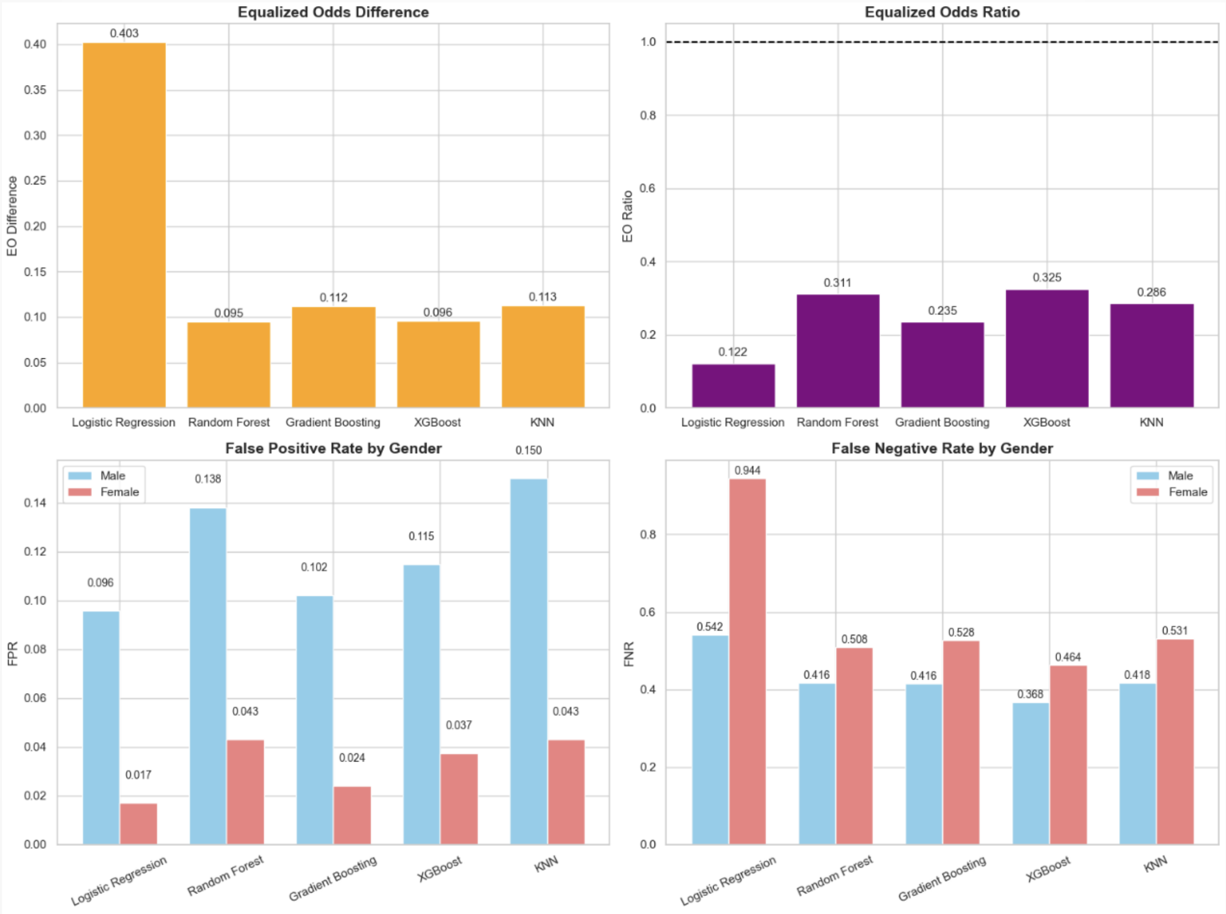
DISTRIBUTION OF INCOME AND GENDER IN THE DATASET



COMPARISON OF MODEL ACCURACY AND FAIRNESS METRICS



FAIRNESS METRICS – EQUALIZED ODDS & ERROR RATES



KEY FINDINGS

- High Accuracy != Fairness
- There is a gender gap in income, with males more likely to earn >$50K.
- Models inherit biases from the dataset, reflecting existing disparities.
- Positive prediction rates and error patterns differ by gender, highlighting systemic inequalities.
- Even accurate models can produce unfair outcomes if sensitive attributes are ignored.
- Understanding data distributions and disparities is critical before deploying ML systems.

THANK YOU

- Any Questions?