Key Factors Influencing Income

Insights from US Census Data

Problem Overview

- Income disparity is influenced by various demographic, educational, and occupational factors.
- By analysing US Census data, this study aims to identify key characteristics that distinguish individuals earning above or below $50,000 per year.
- These insights can support informed decision-making in workforce strategy, policy formulation, and economic planning.

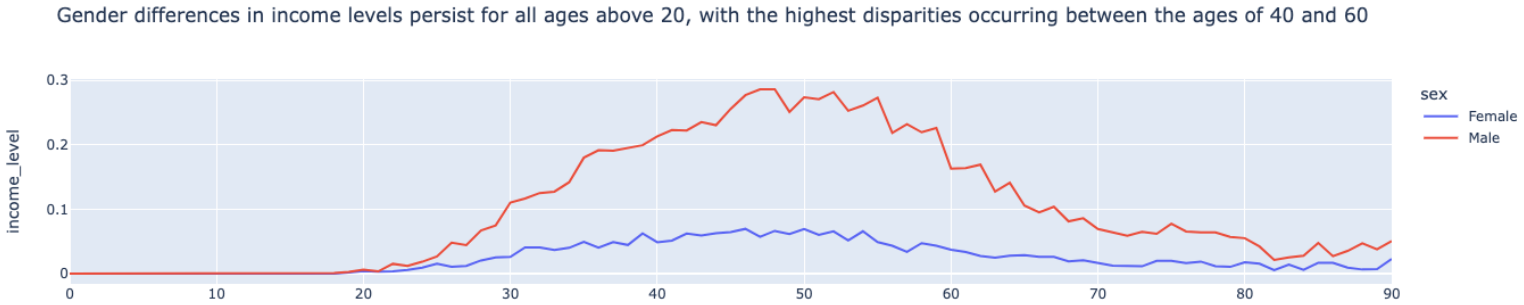
Data Overview	○ **Dataset:** US Census data (~300 individuals) ○ **Target Variable:** Income category (>50K or ≤50K per year) ○ **Key Features:** ▪ Age, Education, Gender, Race ▪ Occupation, Industry, Employment status ▪ Capital gains/losses, Work hours, Marital status

Data Processing Pipeline

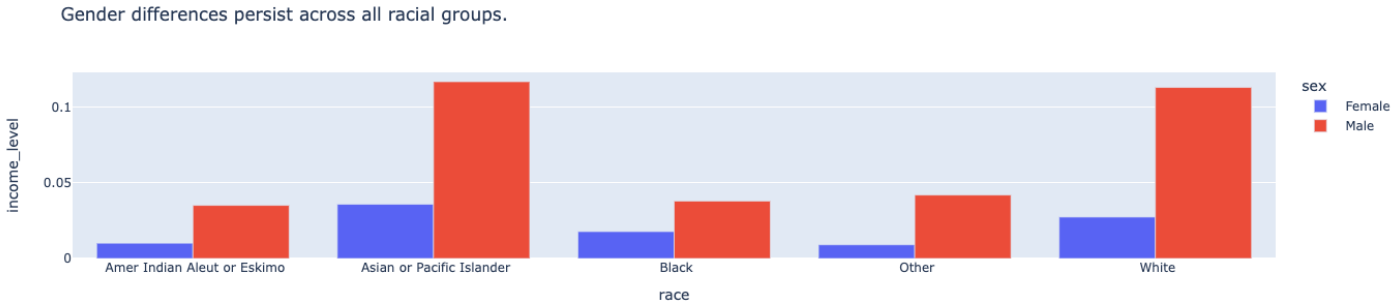
1. Added column names for readability
2. Checked and handled missing values
    - No null value identified from the dataset
    - Some categorical values ad ‘Not in the universe’, ‘?’ are treated as separate category
3. Removed duplicates
4. Addressed dataset imbalance via subsampling
    - Sampled 10% negative class to achieve relatively balanced dataset for modelling
5. Converted target variable to binary (>50K = 1, ≤50K = 0)
6. Feature engineering: Grouped age into bins

EDA Insights | Gender Disparities in Income

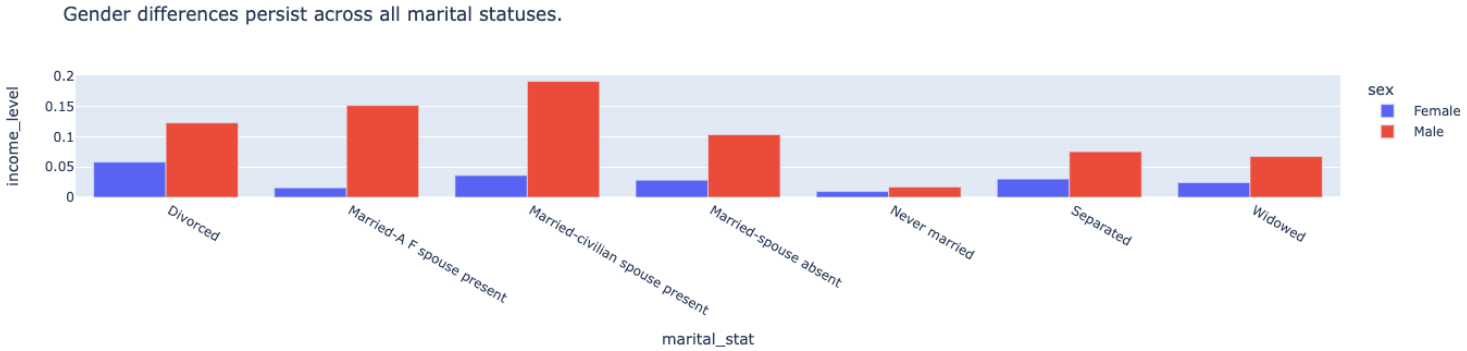
1. Gender differences in income levels persist for all ages above 20
2. The highest disparities occur between the ages of 40 and 60



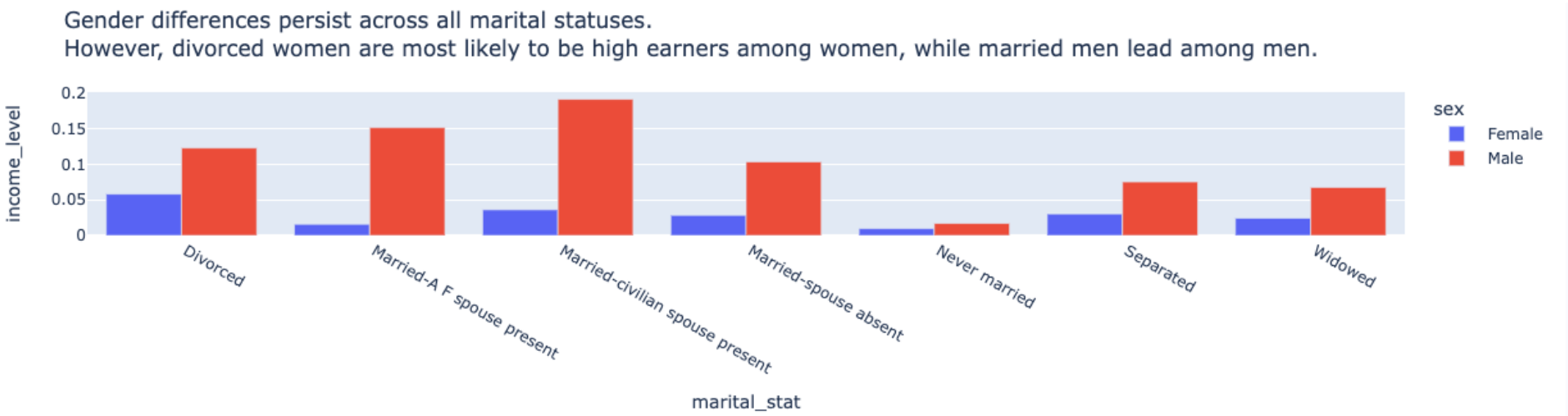
3. Gender differences persist across all racial groups
4. Gender differences persist across all marital statuses



So What? Indicates systemic income inequality that may require policy intervention or targeted career advancement programs



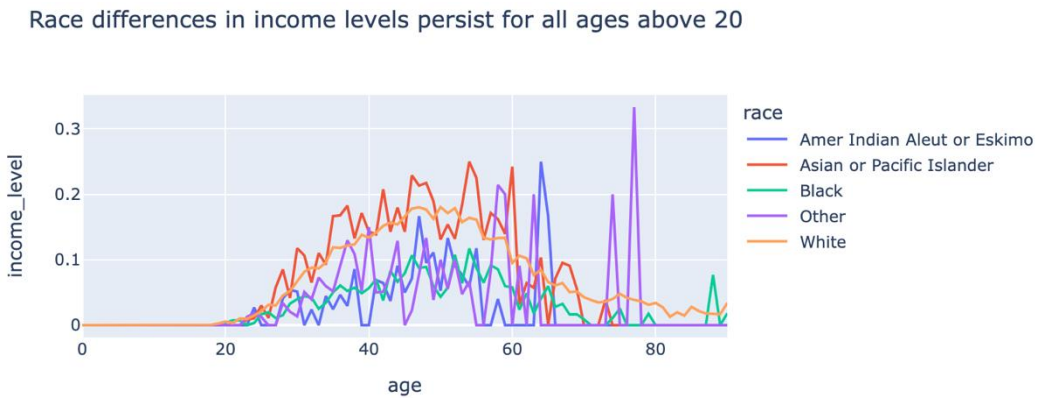
EDA Insights | Marital Status and Income

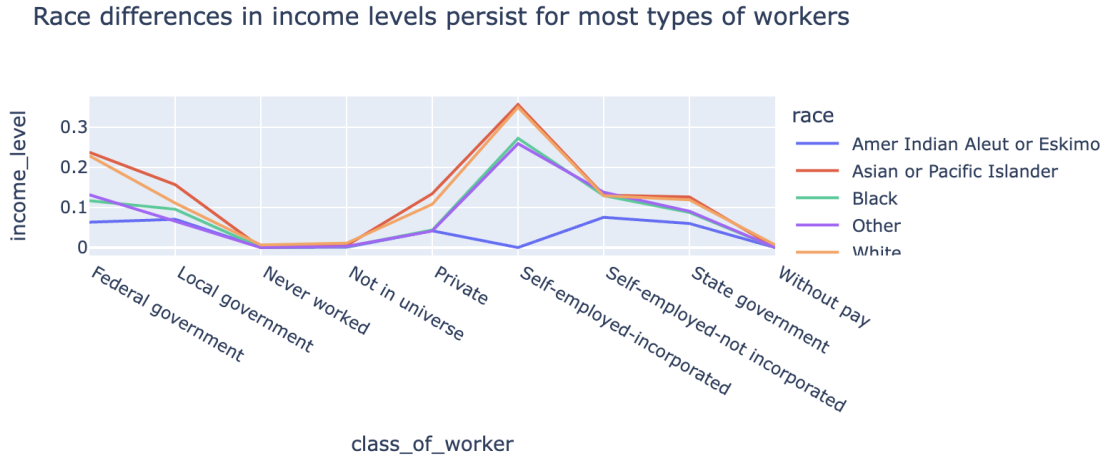


1. Among women, divorced individuals are most likely to be high earners
2. Among men, married individuals are most likely to be high earners

So What? Marital status appears to influence financial stability and career progression, potentially due to differences in household responsibilities, social expectations, and support structures. Divorced women who achieve high earnings may have a stronger career focus post-divorce, while married men benefit from traditional support systems that facilitate career advancement.

EDA Insights | Race Disparities in Income



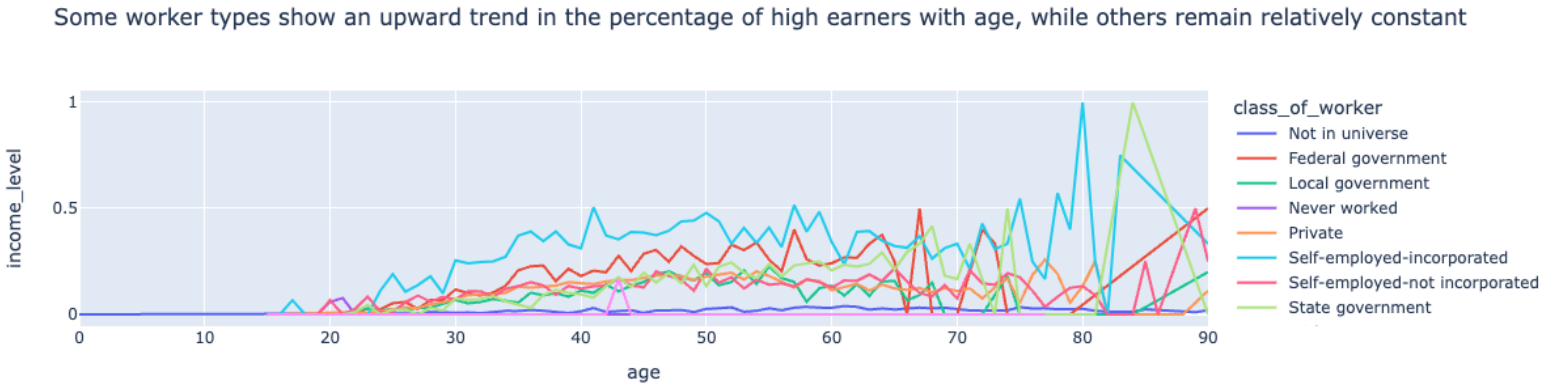


1. Race differences in income levels persist for all ages above 20
2. Race differences in income levels persist for most types of workers

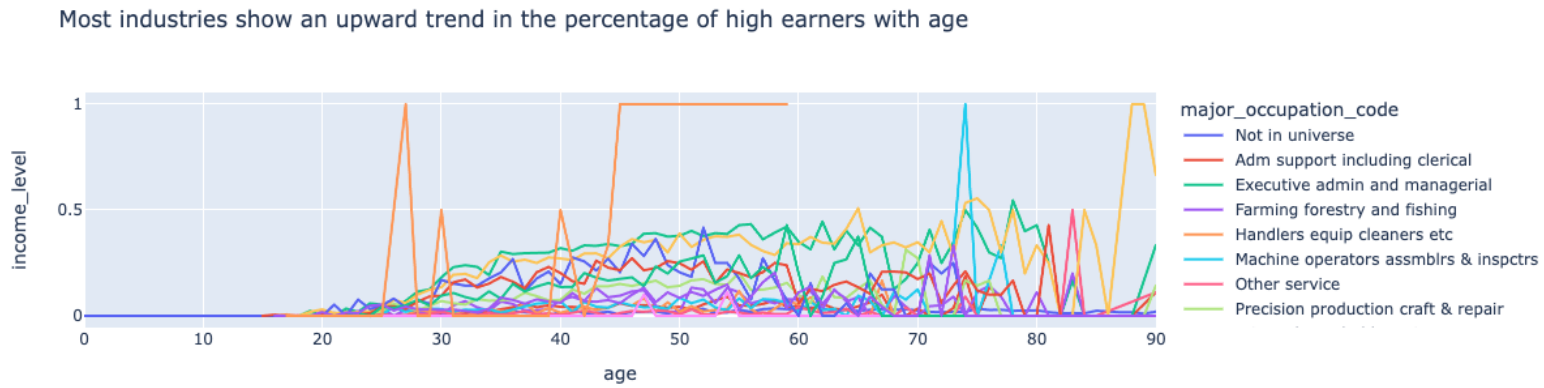
So What? Highlights the need for further investigation into the impact of structural barriers and career opportunities across racial groups

EDA Insights | Industry and Occupation Trends

1. Some worker types show an upward trend in the percentage of high earners with age, while others remain relatively constant
2. Most industries show an upward trend in the percentage of high earners with age, but they have different income levels at all ages



So What? Suggests that career growth potential varies significantly by industry and worker type. Additionally, industry disparities indicate that while some sectors provide consistent income growth, others may have structural limitations affecting earnings, influencing long-term career stability and financial planning.



Modelling Approach and Evaluation

1. **Data Preparation:** Converted categorical features into numerical format

2. **Individual Models Tried:**
    1. CatBoost (without categorical transformation)
    2. LightGBM*
    3. Random Forest*
    4. Logistic Regression*

3. **Stacking Model:**
    - Combined results from individual models
    - Used Logistic Regression as meta-model

4. Model Evaluation:

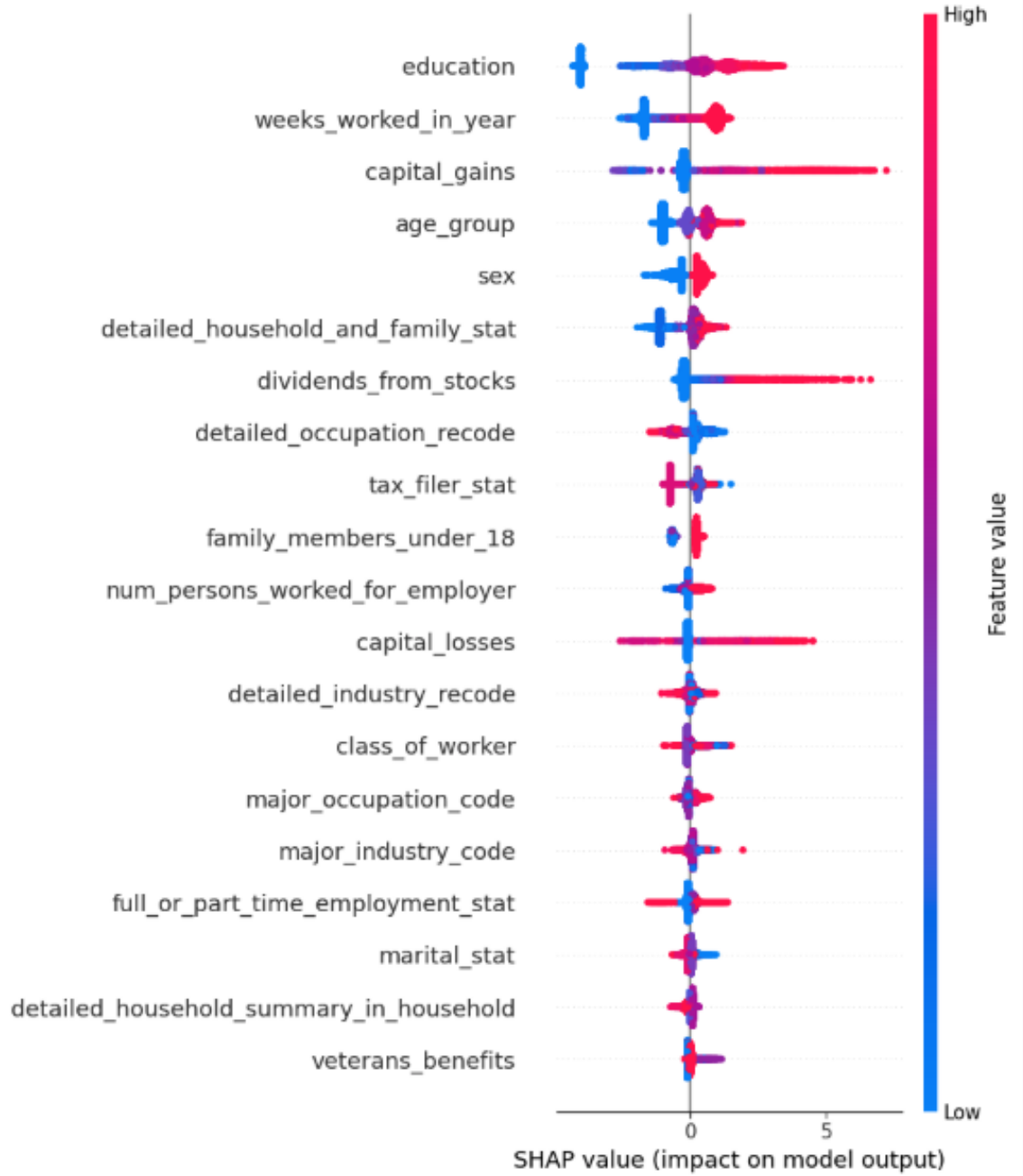
- Evaluated performance using: Accuracy, Precision, Recall, F1-score, Average Precision Score, ROC AUC Score
- Given dataset imbalance, **F1-score was the key metric**
- **Best Model:** Stacking (Meta-Model)
    - Superior out-of-sample performance
    - Better generalisation ability

Model	In-Sample F1 Score (on subsampled data)	Out-Of-Sample F1 Score
Catboost	0.853	0.411
LightGBM	0.851	0.466
Random Forest	0.840	0.477
Logistic Regression	0.798	0.418
Stacking Model	0.835	0.478

* with categorical transformation

Model Insights | Feature Importance (SHAP)

1. **Education level, Age, Sex, Occupation, and Industry** are strong predictors.
    - These features contribute significantly to the model's ability to predict income >$50K.
2. **Weeks worked in year** is a strong indicator as expected.
    - The more they worked, the higher their income will be.
3. **Income-related features** are also strong indicators, as expected.
    - Individuals with higher capital gains or stock dividends are more likely to earn >$50K.

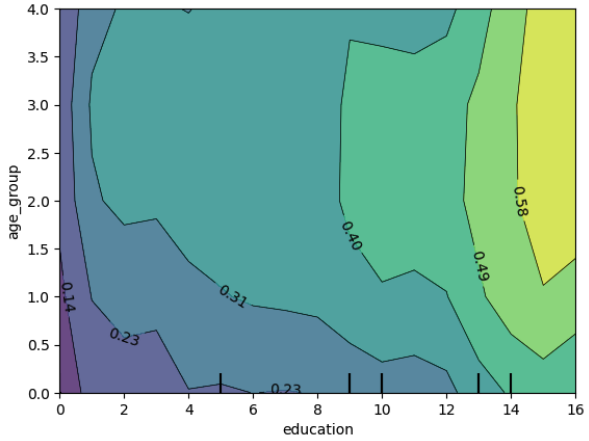


Model Insights | Partial Dependence Plots

Inspected Partial Dependence Plot for pairs of features to understand their joint impact on the predicted outcomes

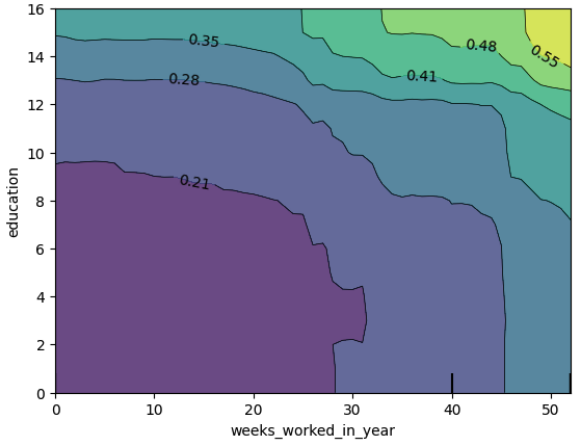
Age & Education:

- For highly educated individuals, age has a stronger impact on income when they are younger, with diminishing returns as they age.
- For individuals with limited education, age has little effect on income, meaning their earnings do not grow significantly with age.
- For individuals with moderate education, both age and education jointly influence income.



Weeks Worked & Education:

- Highly educated individuals see a significant increase in income with more weeks worked, with the highest income growth rate compared to others.
- Individuals with lower levels of education benefit from working more weeks, but the income difference becomes noticeable only after a significant increase in workweeks, not with minor changes.



Summary & Recommendations

Key Findings

- **Main Factors Driving Income Above $50K:**
  - **Top Features**: Education, Age, Sex, Occupation, Industry, Weeks Worked, Capital Gains.
  - **Education & Age**: Higher education amplifies early-career earnings growth, but impact plateaus over time.
  - **Work Effort Matters**: More weeks worked leads to higher earnings, especially for highly educated individuals.

- **Income Disparities Exist**
  - **Gender Gap**: Women earn less than men across all ages, races, and marital statuses.
  - **Race Gap**: Racial income differences persist beyond age 20.
  - **Marital Status Influence**: Divorced women and married men are more likely to be high earners.

- **Career Growth Varies by Industry**
  - Some industries offer consistent income growth, while others have structural limitations.

Recommendations

- **Policy & Corporate Actions**
  - Address systemic gender and racial wage gaps through targeted career advancement programs and policy interventions.
  - Support divorced women's career growth and married men's balanced work-life strategies to ensure equitable financial stability.

- **Industry & Career Guidance**
  - Encourage career transitions into high-growth industries with strong long-term earning potential.
  - Promote higher education and lifelong learning to maximise income growth opportunities.

Future Improvements

- **Expand Feature Set**
  - Collect additional socioeconomic factors such as job tenure, education quality, and social networks to refine predictions
  - Explore geographic variations in income disparities

- **Model Enhancement**
  - Explore deep learning models and additional feature engineering for better insights
  - Explore causal modelling to identify more complex interaction structures and distinguish between **correlation and causation** to provide more actionable policy recommendations

- **Refine Model Interpretability**
  - Leverage explainable AI techniques for deeper insights into decision-making
  - Further analyse interaction effects between race, gender, and education

- **Real-World Applications**
  - Use insights to inform HR policies, wage negotiations, and career development programs
  - Provide data-driven career guidance for individuals seeking higher earnings potential