THE QUEST TO PREDICT LABOR: A DATA-DRIVEN LOOK AT WEEKLY WORK HOURS

Group Members:

Alwyn Munatsi

Lucia Shumba

Chidochashe Makanga

Bekithemba Nkomo

THE QUEST TO PREDICT LABOR: A DATA-DRIVEN LOOK AT WEEKLY WORK HOURS

Key Takeaway: Understanding *how much* people work is critical in an economy marked by rising inequality and persistent wage gaps.

This project explores which demographic, socioeconomic, and occupational variables best predict the number of hours an individual works per week.

To establish the real-world context, consider two key statistics:

- The U.S. Gini index, a measure of income inequality, rose to 0.494 in 2023 one of the highest values recorded since 1967.
- In 2023, women's median weekly earnings were approximately 83% of men's, highlighting ongoing disparities in the labor market.

This project uses machine learning to identify the key drivers of labor supply and build a model to predict it.

OUR GUIDING RESEARCH QUESTIONS

Key Takeaway: *Our investigation is structured to answer three fundamental questions about labor supply.*

RQ 1: The Drivers

Which demographic, educational, and occupational attributes most strongly influence weekly labor supply (hours worked per week)?

RQ 2: The Prediction

Can machine learning regression models accurately predict weekly hours worked using demographic and socioeconomic features?

RQ 3: The Synergy

Does an ensemble model outperform individual regression models when predicting weekly hours worked?

THE FOUNDATION: THE UCI ADULT (CENSUS) DATASET

Key Takeaway: *Our analysis is based on a well-established public dataset containing over 32,000 records from the 1994 U.S. Census.*

Source: UCI Machine Learning Repository, "Adult (Census Income)" dataset.

Scope

32,561 Rows **15** Columns

Features

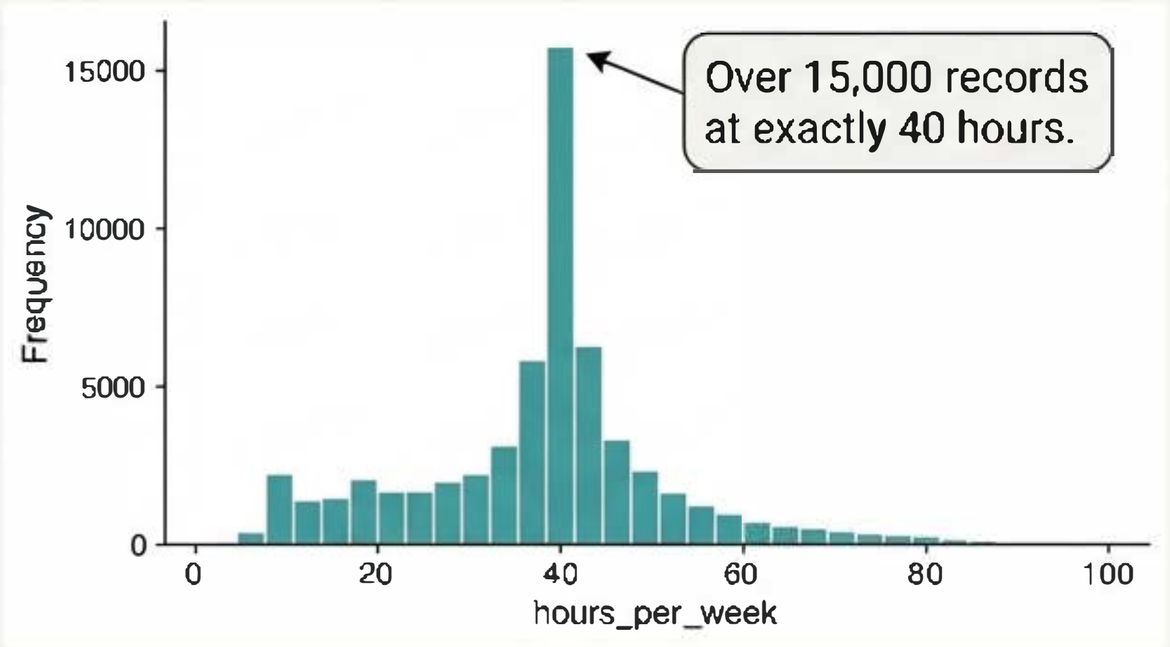
- **Numerical:** age, capital_gain, capital_loss
- **Categorical:** workclass, education, marital_status, occupation, race, sex, etc.

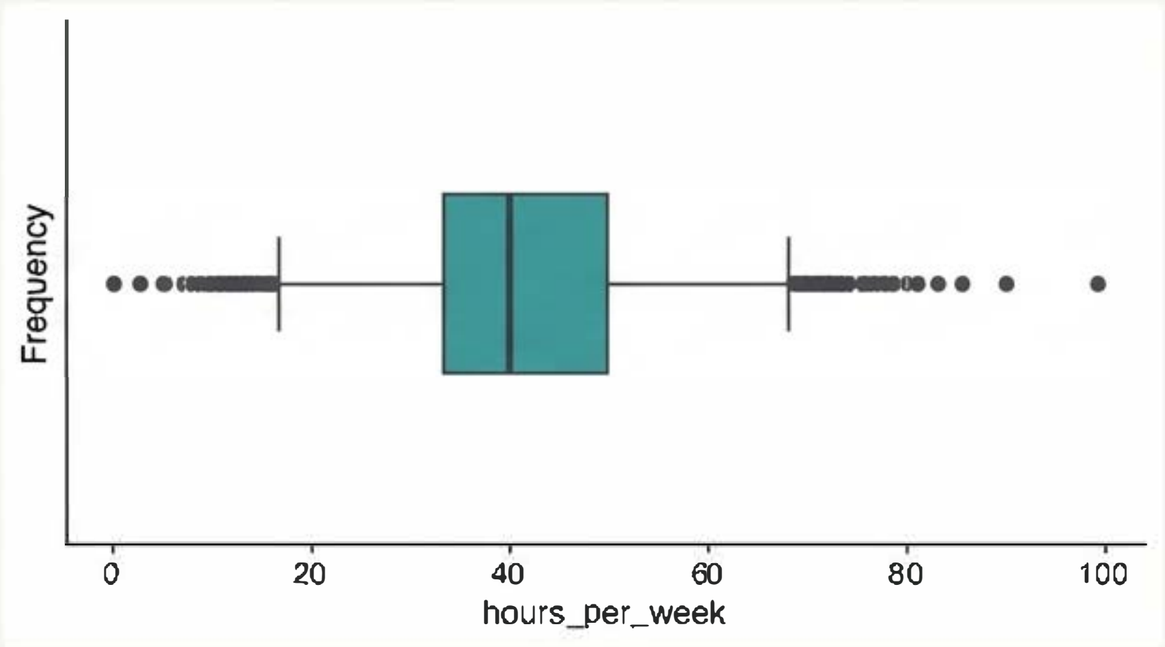
Target Variable:
`hours_per_week` (A continuous measure of weekly labor supply)

	`age`	`sex`	`hours_per_week`
[illegible]	40	male	40
[illegible]	50	male	50
[illegible]	50	male	35
[illegible]	40	male	60
[illegible]	39	male	45
[illegible]	50	male	50
[illegible]	35	male	60
[illegible]	40	male	45
[illegible]	50	male	40
[illegible]	45	male	35
[illegible]	40	male	50
[illegible]	35	male	35
[illegible]	45	male	60
[illegible]	50	male	45

THE TARGET: A STRONG PEAK AT THE 40-HOUR WORK WEEK

Key Takeaway: *The distribution of weekly hours is not normal; it is dominated by a sharp peak at 40 hours, with significant outliers on both ends.*





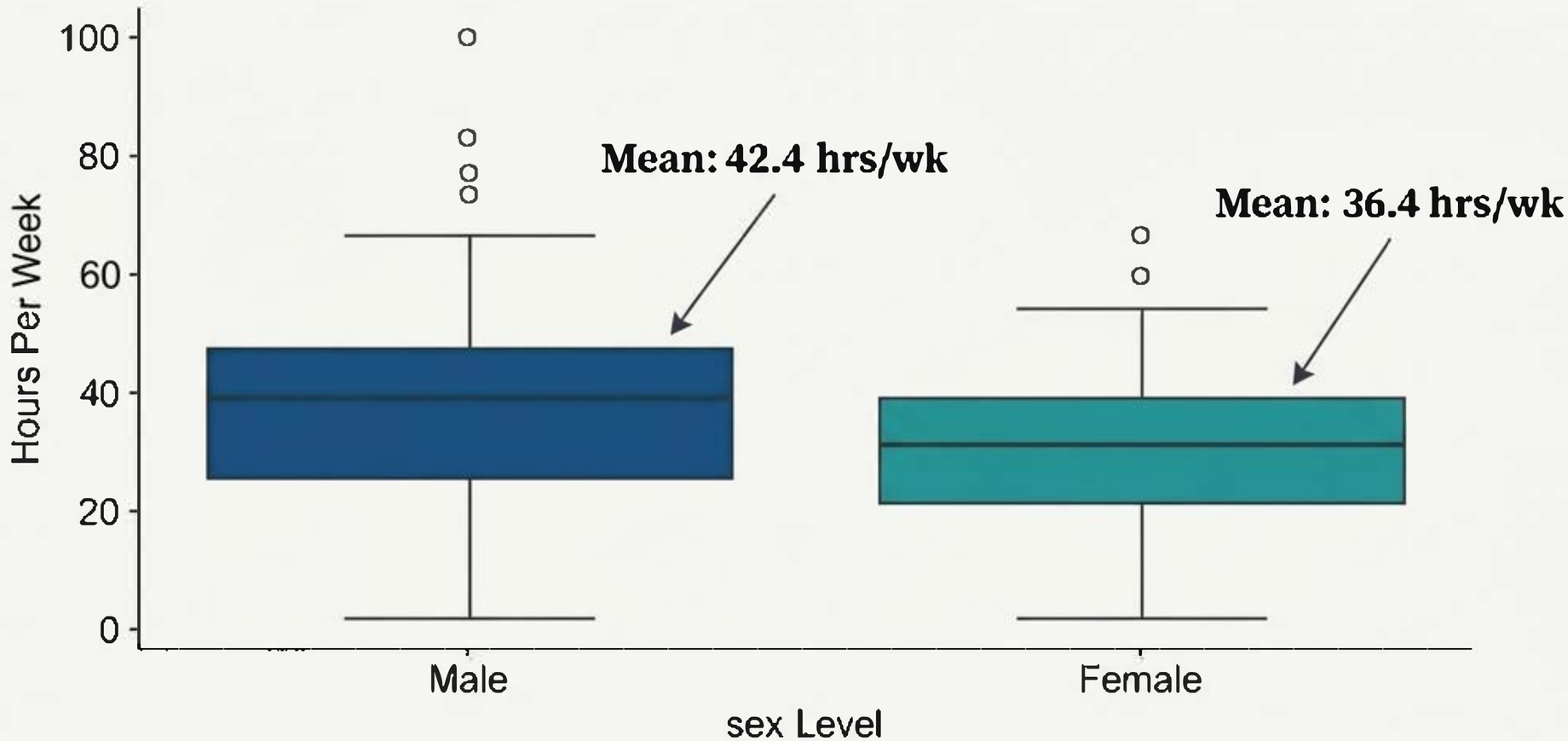
Key Statistics & Insight

Mean: 40.4 hours	**Median:** 40 hours	**Std. Dev:** 12.3 hours	**Range:** 1 to 99 hours

Insight: The data confirms the cultural standard of a 40-hour work week. This concentration presents a unique modeling challenge compared to a normally distributed target.

A CLEAR GENDER GAP IN WEEKLY HOURS WORKED

Key Takeaway: *Men in the dataset work, on average, 6 more hours per week than women, a statistically significant difference.*

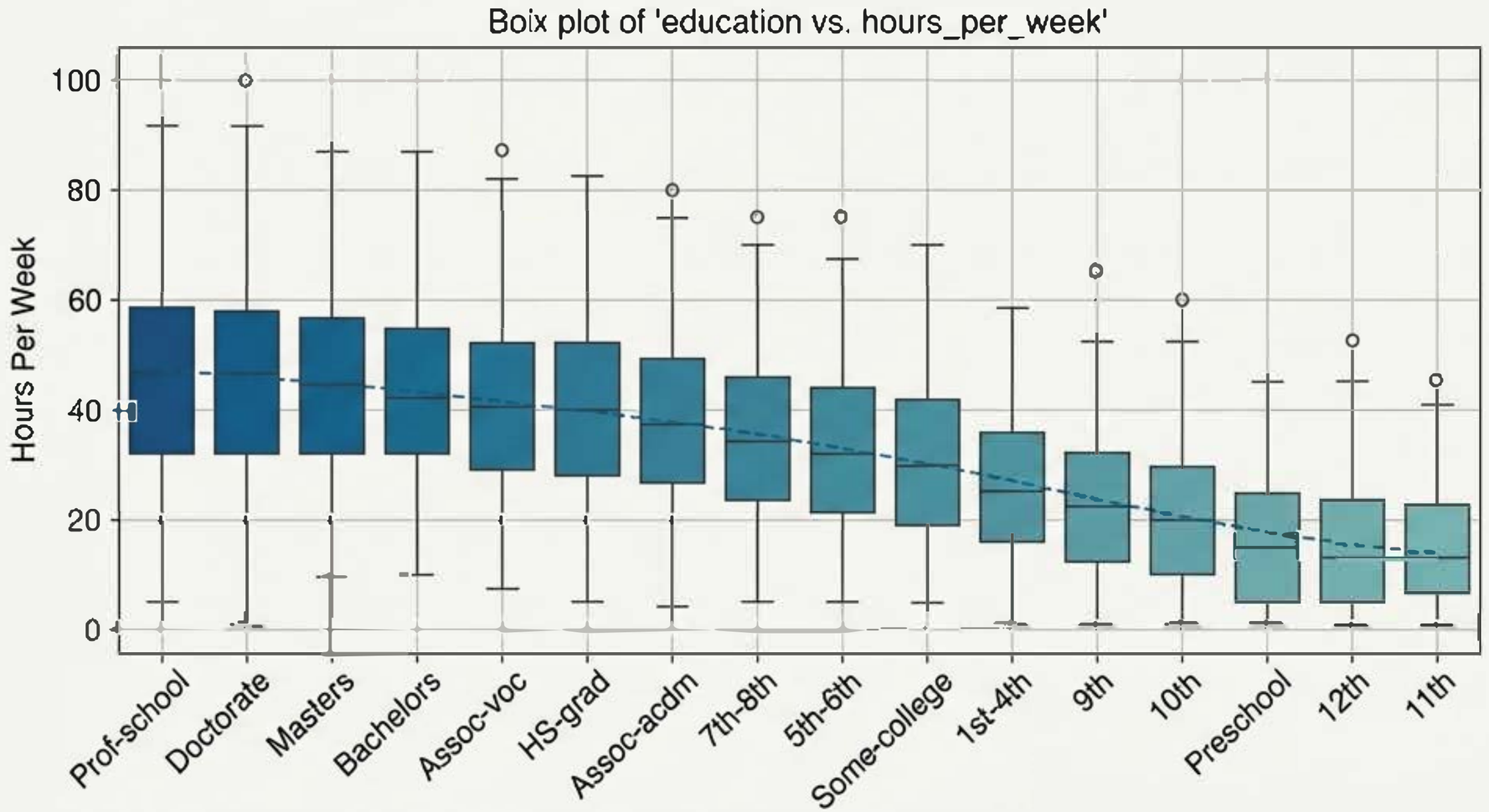


Statistical Significance

A one-way ANOVA test confirms a significant difference between the groups (F-statistic: 1807.06, p-value: < 0.001). This indicates `sex` is a strong candidate for a predictive feature.

HIGHER EDUCATION CORRELATES WITH MORE HOURS WORKED

Key Takeaway: *There is a clear, statistically significant trend where individuals with higher levels of education tend to work more hours per week.*



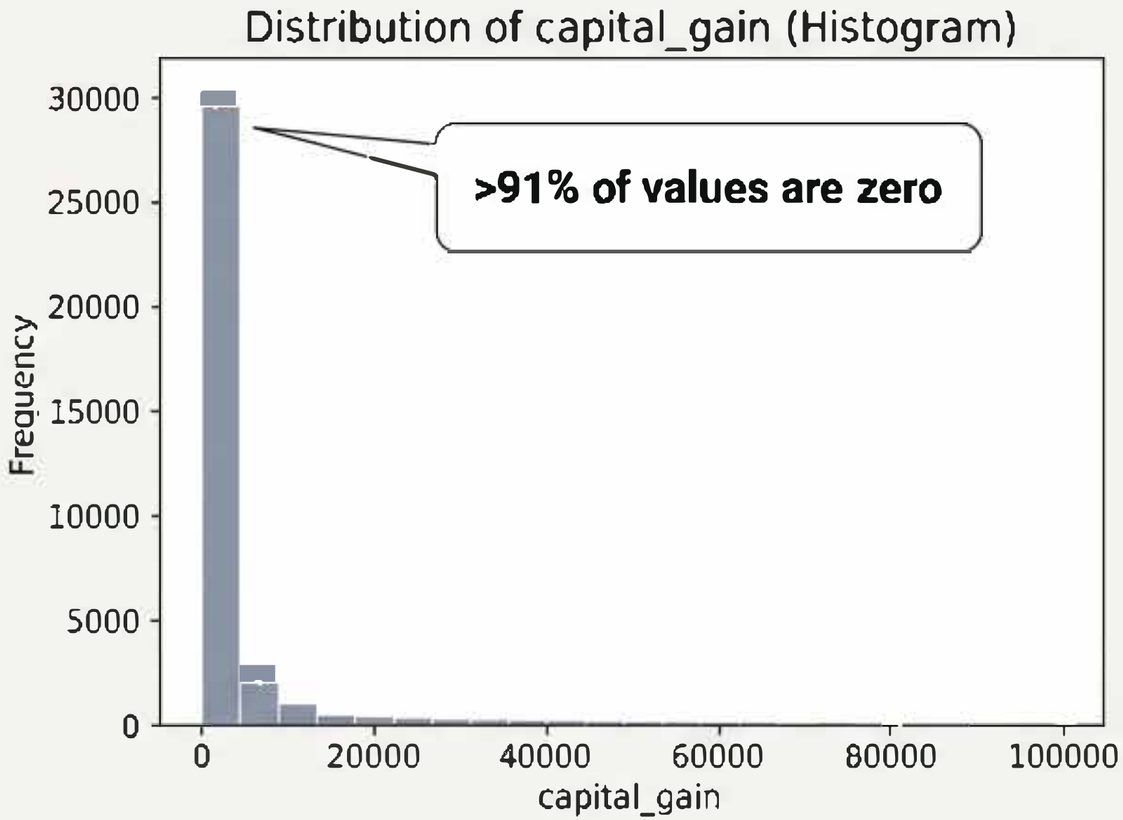
Key Trend: Mean weekly hours increase with educational attainment.

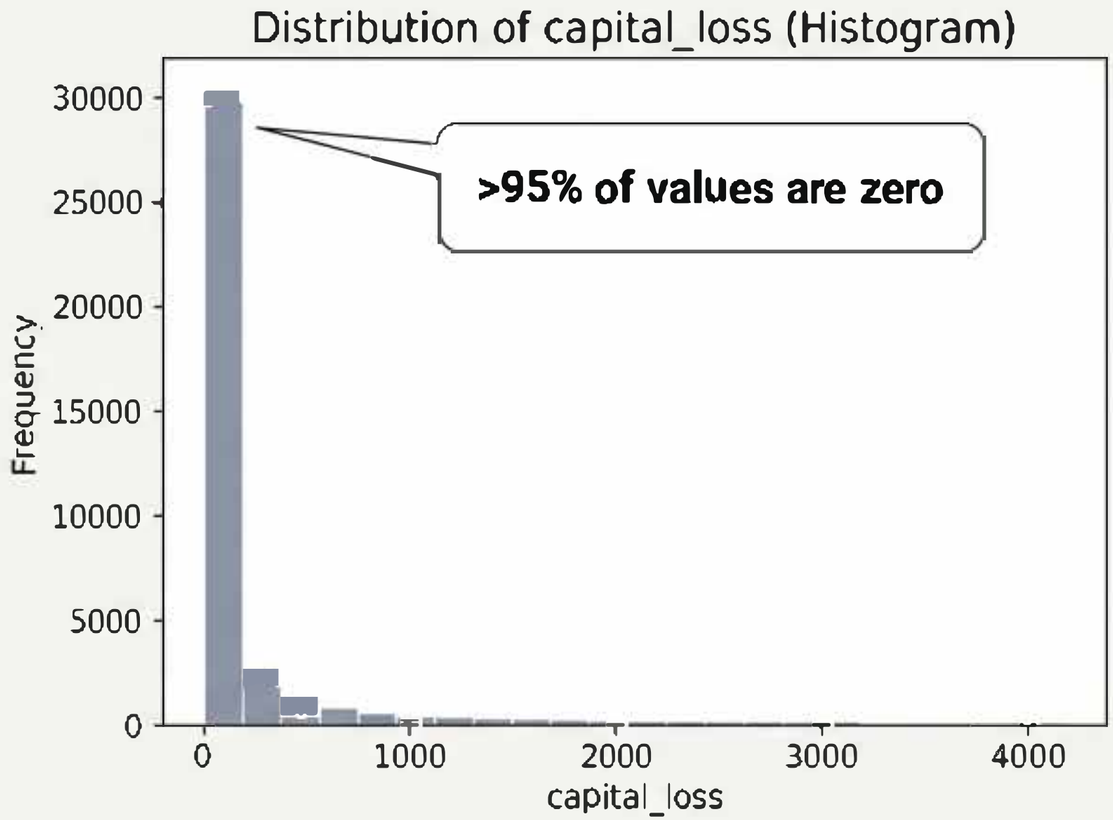
- Profschool: **47.4 hours**
- Doctorate: **47.0 hours**
- Bachelors: **42.6 hours**
- HS-grad: **40.6 hours**

Statistical Significance: ANOVA results confirm this is not due to chance (F-statistic: 85.11, p-value: < 0.001).

THE DATA CHALLENGE: EXTREME SPARSITY IN FINANCIAL FEATURES

Key Takeaway: *The `capital_gain` and `capital_loss` features are dominated by zeros, making them difficult for many models to use effectively without transformation.*





- **Observation:** For both features, the 25th, 50th, and 75th percentiles are all 0. This extreme right-skew and sparsity can cause traditional regression models to perform poorly.
- **Modeling Consideration:** This requires careful preprocessing, such as creating binary `has_gain` flags or using models robust to skewed, sparse data like tree-based algorithms.

DATA PREPARATION AND SELECTION

Key Takeaway: *A systematic pipeline was used to clean the data, handle missing values, and transform features for optimal model performance.*

1. Data Cleaning

Duplicates: The 24 duplicate rows were considered not true duplicates and hence were not removed from the dataframe.

Missing Values: Handled `?` entries found in `workclass`, `occupation`, and `native country`

2. Feature Selection & Engineering

Dropped Columns: `fnlwgt` (statistical weight) and `education` (redundant with `education_num`).

Categorical Encoding: Converted features like `workclass`, `occupation`, and `sex` to a numerical format using One-Hot Encoding.

3. Numerical Scaling

Standardization: Applied `StandardScaler` to numerical features like `age` to prevent features with larger values from disproportionately influencing models.

THE MODELS : FIVE CONTENDERS FOR PREDICTION

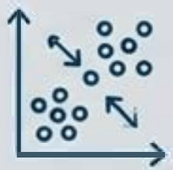
Key Takeaway: *We developed five distinct regression models, each chosen for its unique ability to handle the characteristics of our dataset.*

NEGATIVE BINOMIAL REGRESSION

A count model chosen to address the discrete, overdispersed nature of the `hours_per_week` variable.

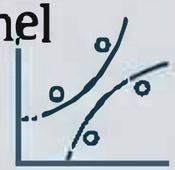
K-NEAREST NEIGHBORS (KNN) REGRESSOR

A non-parametric model that captures complex, non-linear patterns by looking at local data points.



SUPPORT VECTOR REGRESSOR (SVR)

A powerful model that can capture non-linear relationships using kernel transformations.

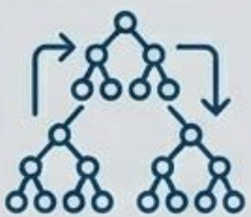


DECISION TREE REGRESSOR

An interpretable model that creates rule-based splits, adapting well to feature interactions.

XGBOOST REGRESSOR

A state-of-the-art gradient boosting algorithm known for its high accuracy and robustness.



EVALUATING PERFORMANCE: XGBOOST EMERGES AS THE CHAMPION

Key Takeaway: *After 5-fold cross-validation, the XGBoost Regressor demonstrated the best overall performance, particularly in minimizing prediction error (RMSE) and explaining variance (R^2).*

MODEL	RMSE	R^2
Negative Binomial	11.4226	0.1417
K-Nearest Neighbors	11.9291	0.0637
SVR	11.7476	0.0922
Decision Tree	11.3895	0.1467
XGBoost Regressor	**11.3870**	**0.1471**

Justification: XGBoost's superior performance, especially its ~15% R-squared value, makes it the preferred standalone model for predicting weekly hours worked.

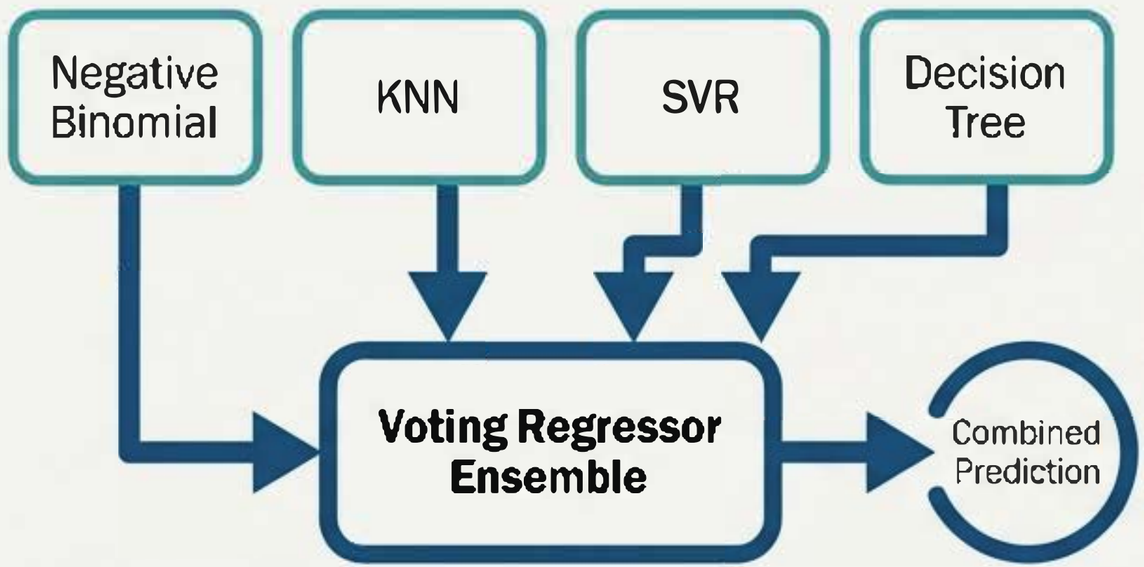
TESTING : APPLY XGBOOST TO TEST DATA

Model	RMSE	R^2
XGBoost (TRAINING)	11.3870	0.1471
XGBoost (TESTING)	**11.4715**	**0.1456**

Conclusion: The ensemble did not reduce RMSE or improve R^2 compared to the champion. The XGBoost model remains superior. It did however perform SVR(RMSE 11.7476, R^2 0.0922) and KNN(RMSE 11.9392, R^2 0.0637)

THE TWIST: CAN A TEAM OUTPERFORM THE CHAMPION?

Key Takeaway: *We constructed an ensemble model to test if combining the predictions of multiple models could yield even greater accuracy.*



Ensemble Strategy

A `VotingRegressor` was created to average the predictions of multiple base models.

The "Team"

The ensemble combines the predictions of our four "weak learners": Negative Binomial, KNN, SVR, and Decision Tree.

Can this collective approach, which leverages the diverse strengths of each base model, achieve a lower error and higher R^2 than the standalone XGBoost champion?

FINAL VERDICT: ENSEMBLE FAILED TO OUTPERFORM THE CHAMPION

The Voting Regressor ensemble did not improve upon the champion model's performance. The ensemble's RMSE (11.5561) is higher than the champion's RMSE (11.3870), and its R^2 (0.1329) is lower than the champion's R^2 (0.1471), indicating that the ensemble performed worse across both key metrics.

Model	RMSE	R^2
Champion (XGBoost)	11.3870	0.1471
Ensemble (Voting Regressor)	**11.5561**	**0.1329**

Conclusion: The ensemble did not reduce RMSE or improve R^2 compared to the champion. The XGBoost model remains superior. It did however outperform SVR(RMSE 11.7476, R^2 0.0922) and KNN(RMSE 11.9392, R^2 0.0637)

ANSWERING OUR QUESTIONS, PART 1: THE STRONGEST PREDICTORS OF LABOR SUPPLY

Key Takeaway: *Our analysis confirms that an individual's education, sex, occupation, and marital status are the most influential factors in determining weekly work hours.*

RQ1: Which attributes most strongly influence weekly labor supply?

EDUCATION

- **Higher education** levels consistently correlate with more hours worked.

SEX

- **Males** work significantly more hours per week on average than females.

OCCUPATION & WORKCLASS

- **Self-employed** individuals and those in occupations like **'Exec-managerial'** and **'Farming-fishing'** work the most hours.

RELATIONSHIP

- Individuals who are listed as **'Husband'** work the most hours on average.

ANSWERING OUR QUESTIONS, PART 2: THE SUCCESS OF PREDICTIVE MODELING

***Key Takeaway**: Machine learning models can predict weekly hours with moderate success, and xgboost approach provided the best performance.*

RQ2: Can ML models accurately predict weekly hours worked?

Yes. Our best model, the xgboost regressor, was able to account for approximately

15% of the variance ($R^2 = 0.1456$) in weekly hours worked based on the available features.

RQ3: Does an ensemble model outperform individual models?

No. The 'VotingRegressor' ensemble achieved a slightly higher Root Mean Squared Error **(11.5561)** than the best individual model, XGBoost **(11.3870)**. However it outperformed SVR(RMSE 11.7476, R^2 0.0922) and KNN(RMSE 11.9392, R^2 0.0637)

THE STORY OF LABOR IS A STORY OF DEMOGRAPHICS

Key Takeaway: *While predicting the exact number of hours a person works is a complex challenge, this project successfully demonstrates that a significant portion of labor supply is explained by an interconnected web of personal and professional characteristics.*

We began by contextualizing labor supply within broad economic trends of inequality.

Through rigorous analysis of the UCI Adult dataset, we identified key predictors like education, sex, and occupation.

We built and validated a series of machine learning models, culminating in an xgboost model that could explain nearly 15% of the variance in work hours.

Concluding Thought: This data-driven approach provides a quantitative foundation for better understanding - and potentially shaping - workforce participation and equity.

REFERENCES

- Final Project Notebook
- Final Project Proposal
- DAV - 6150 Lecture Notes
- U.S. Census Bureau. (2024). Income inequality in the United States: Gini Index historical reports. https://www.census.gov/
- U.S. Bureau of Labor Statistics. (2024). Highlights of women’s earnings: 2023. https://www.bls.gov/