Decoding Income: Data-Driven Solutions for Economic Equity

Adel Hassen
January 30th, 2025

Executive Summary

- Successfully identified influential factors determining whether an individual earns an income above or below $50k per year
- Key takeaways
    - Education level, capital gains, and weeks worked during a year are strong predictors of income
    - Demographic features like sex and age show disparities
    - Predictive modeling achieved **94.76%** accuracy in classifying income groups

These insights can drive data-driven policies and economic mobility programs.

Why does this Matter?

Understanding characteristics associated with earning above or below $50K per year to enable key stakeholders to take mitigative actions

Policymakers → Design policies to remove barriers towards upward economic mobility

Nonprofits & Education → Guide career pathways for low-income individuals

Scope and Constraints

- Scope
  - Exploratory Data Analysis
  - Data Preparation
  - Modeling
  - Results
- Constraints
  - Time
  - Resources
  - Data
  - Human Capital

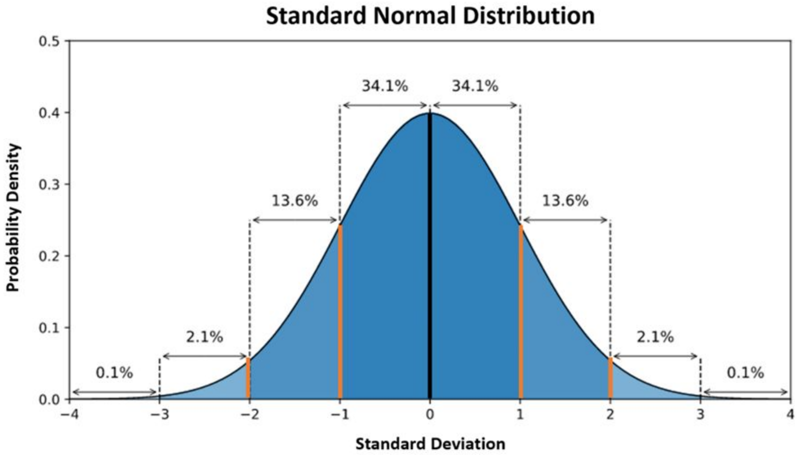
Data Overview

- US Census data from 1994-1995
- Roughly 300,000 records
- 40 features
  - 7 numerical
  - 33 categorical
- Bias and privacy considerations
  - Potential historical biases in census data
  - Not PII (Personally Identifiable Information) but must be careful
  - Demographic data requires special attention when used in modeling

Exploratory Data Analysis

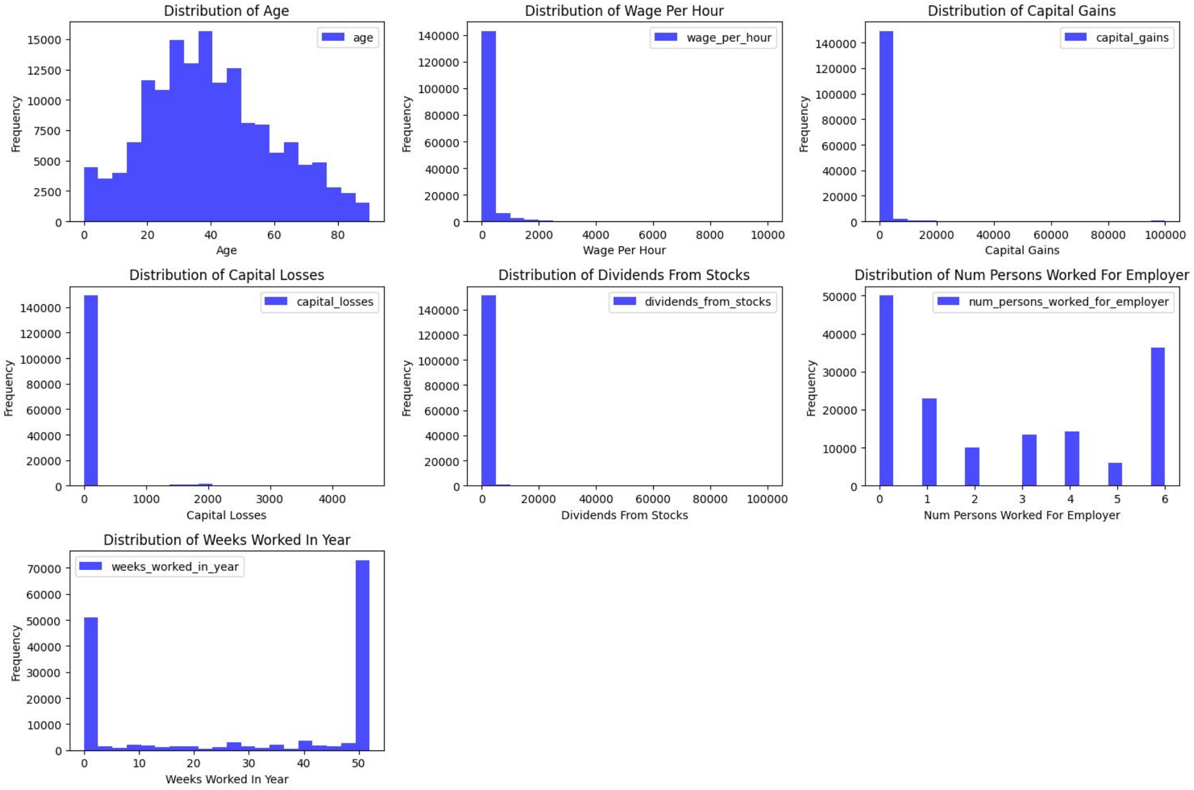
Exploratory Data Analysis Key Terms

- **Outlier:** Data point far from others
- **Skewed Data:** Data asymmetry in distribution
- **Distribution:** Spread of data values
- **Normally Distributed:** Symmetric, bell-shaped data distribution



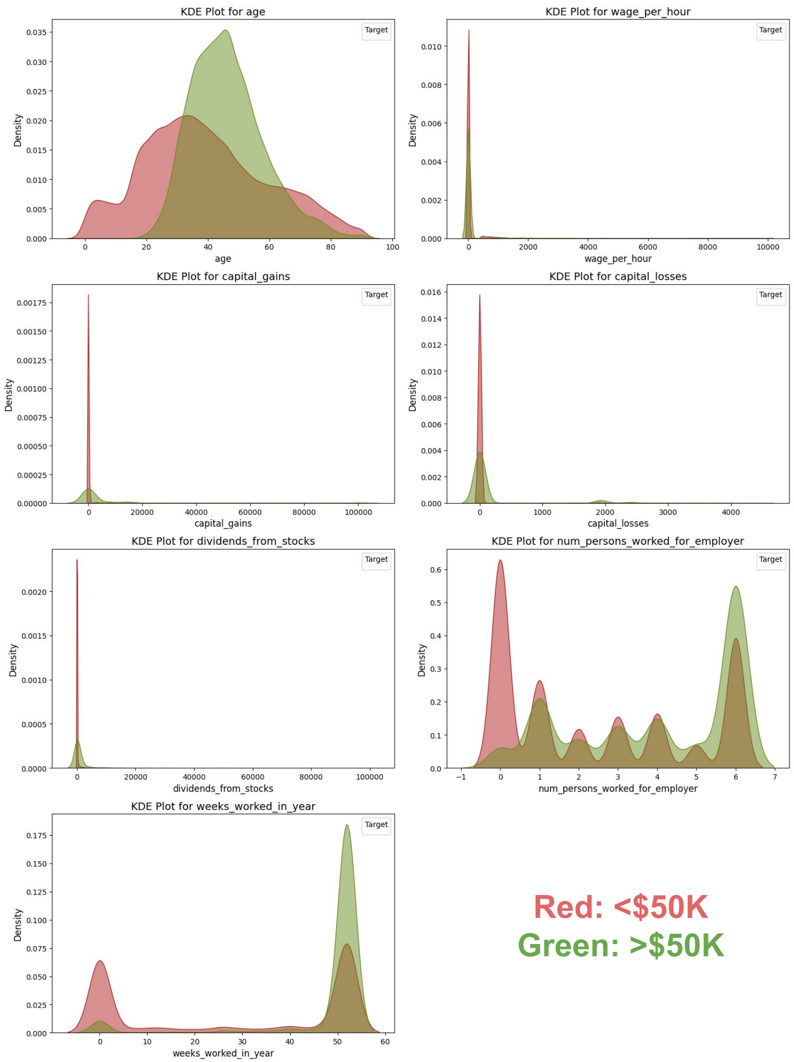
Distribution of Numerical Data

- Multiple features contain outliers and are skewed right
- Weeks worked in a year and number of persons worked for employer have bimodal (two peaks) distributions
- Age is close to normally distributed but looks a bit skewed

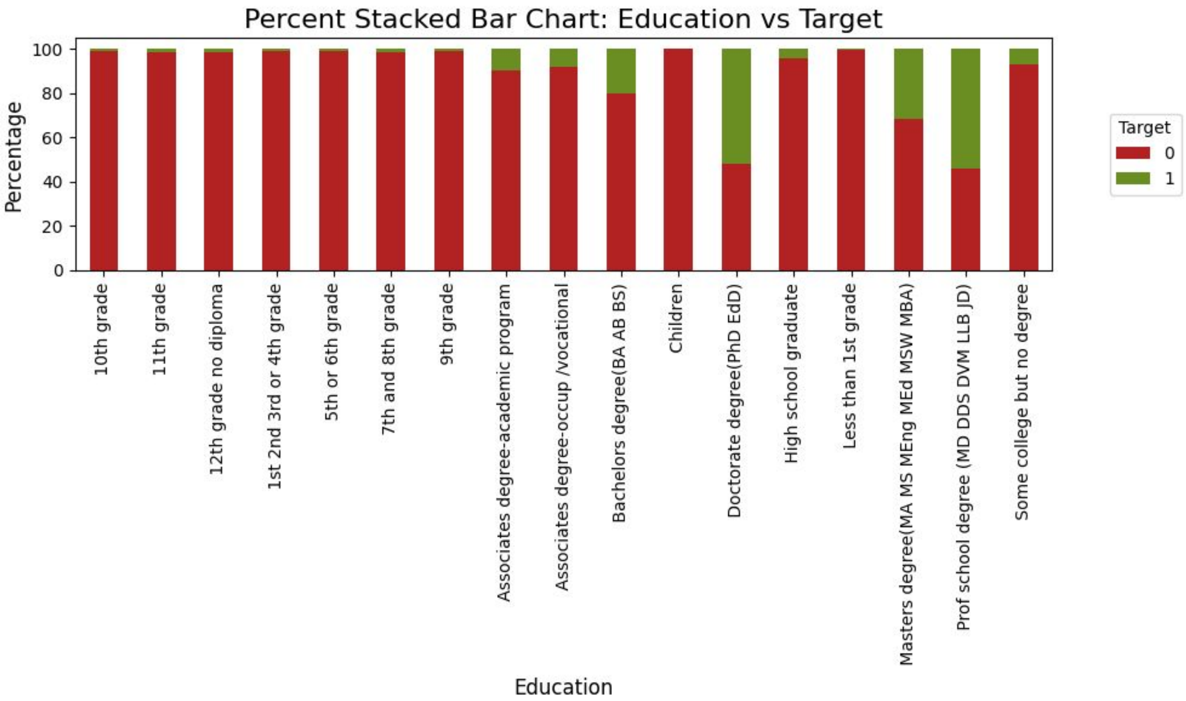


Distribution of Income Class

- Older individuals are more likely to earn >$50K than younger ones
- When number of persons worked for employer is 6, it increases chances of >$50K. When 0, increases chances of <$50K.
- 50-52 weeks worked increases >$50K likelihood while 0 weeks strongly correlates with <$50K



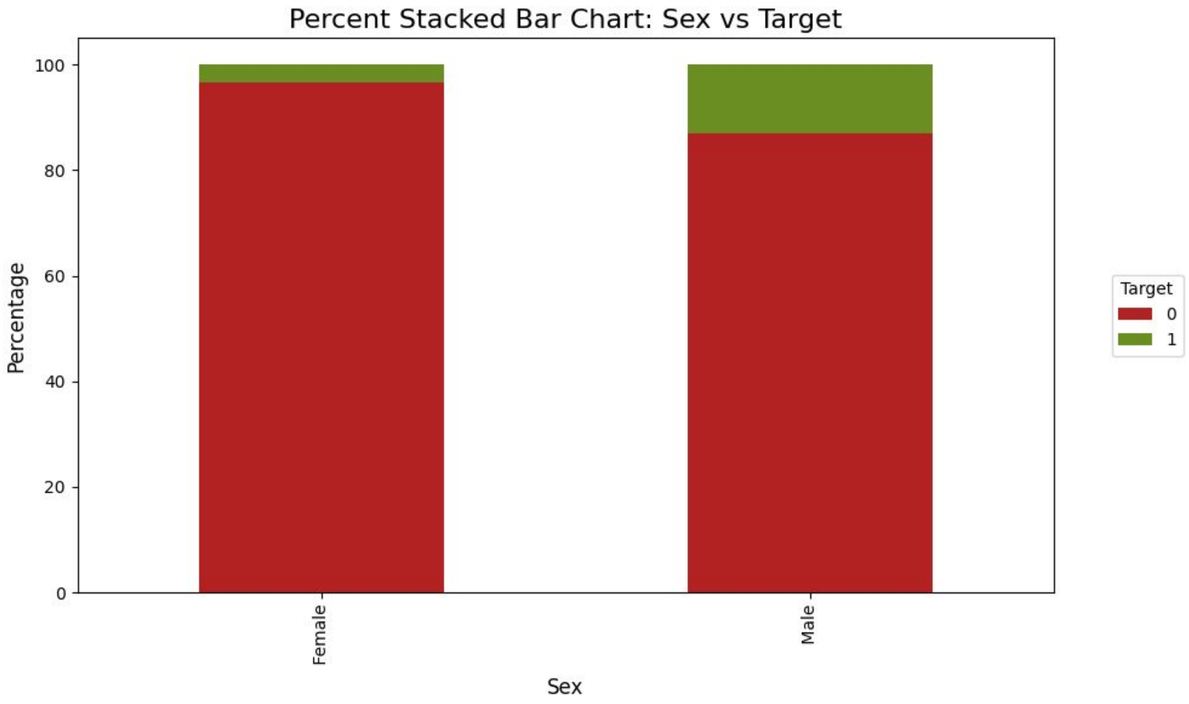
Education



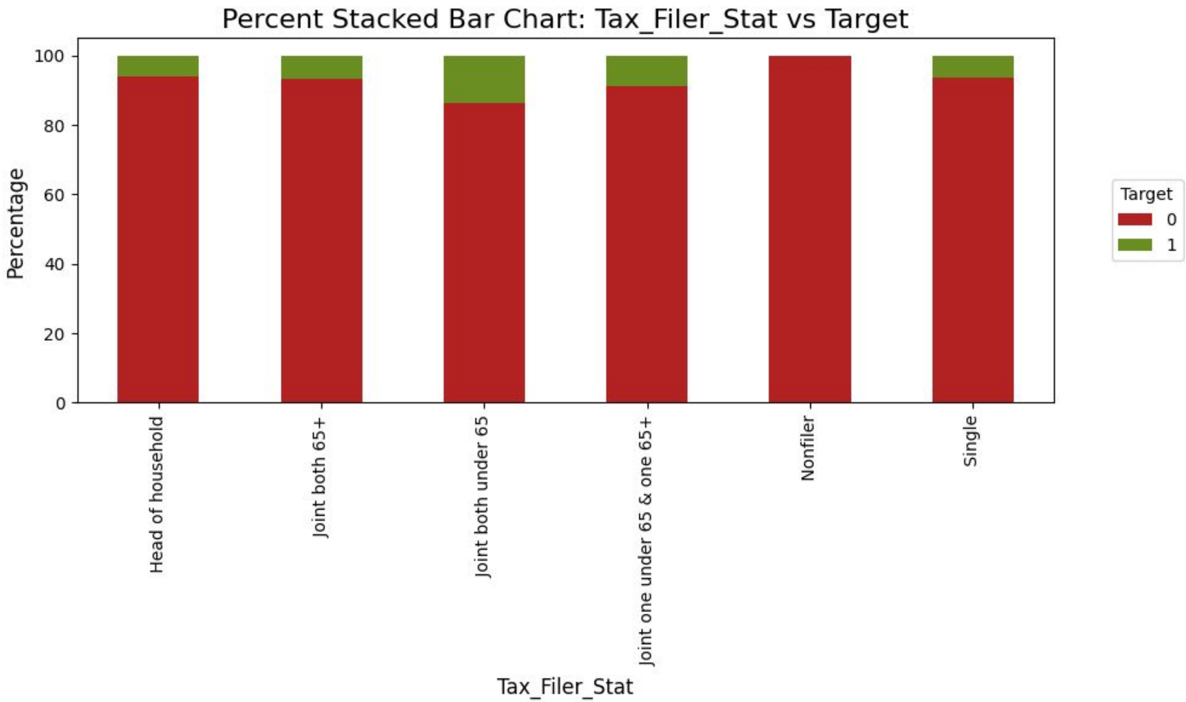
- Advanced degrees show the highest percentage of income >$50K
  - PhD
  - Prof school degree
  - Masters degree

Sex

- Males have higher percentage of earning >$50K
- Highlights gender pay gap



Tax Filer Status



- Joint both under 65 shows the highest percentage of income >$50K
- Nonfiler is nearly 100% <$50K

Data Preparation

Data Preparation Key Terms

- **Ordinal feature:** categorical data with a clear order
- **Nominal feature:** categorical data with no clear order
- **Encoder:** convert categorical data into a numerical representation
  - **Ordinal encoder:** Preserves order of categories
  - **One hot encoder:** transform into binary representation
- **Normalization:** scales data to specific range, usually between 0 and 1

One Hot Encoding

id	color
1	red
2	blue
3	green
4	blue

One Hot Encoding

id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

Ordinal Encoding

Original Encoding	Ordinal Encoding
Poor	1
Good	2
Very Good	3
Excellent	4

Data Preprocessing

- Ordinal features (education feature)
  - Rank in order
  - Ordinal encoder to convert to integers
  - Normalize data between 0 and 1
- Nominal features
  - One hot encoder
- Numerical features
  - Normalize data between 0 and 1

Simple data pipeline that converts all features into numerical values with range 0 to 1

Modeling

Modeling Key Terms

- **Machine Learning (ML) Model:** mathematical system that learns patterns from data to make predictions
- **Classification:** ML task predicting categories
- **Training data:** Data used to train a machine learning model
- **Test data:** Data used to evaluate a trained model's performance
- **Feature Selection:** Choosing the most important variables for the model
- **Recursive Feature Elimination:** Iteratively removing less important features
- **Sampling:** Adjusting data distribution for fair model training
- **Random Oversampling:** Duplicating minority class examples to balance data
- **SMOTE:** Creating synthetic data points to balance classes
- **Tree-Based Model:** Decision-making model using if-then rules like a flowchart
- **Boosted Tree:** Ensemble of trees improving weak learners iteratively
- **LightGBM Classifier:** Fast, efficient boosted tree model optimized for large datasets
- **Hyperparameter:** Settings that control how a model learns
- **Hyperparameter Tuning:** Optimizing hyperparameters for better performance

Machine Learning Approach

- Problem type: Binary classification of income class
  - Below $50K per year (92%)
  - Above $50K per year (8%)
- Steps
  - Use lazypredict to test basic models
  - Select best model and perform hyperparameter tuning
  - Try improving model (feature selection, sampling, etc.)
  - Retrain on full training data
  - Return results

Best Model

- LightGBM Classifier was the best model
- Qualities of LightGBM
  - Based on decision trees
  - Boosted tree method
  - Handles missing values and categorical features
  - Optimized for speed
- Accuracy
  - Default hyperparameters: **94.28%**
  - Hyperparameter tuning: **94.50%**

Attempted Improvements

- Feature selection with recursive feature elimination
- Oversampling
  - Random oversampling
  - SMOTE

Model	Accuracy	F1 Score (Weighted)
Default hyperparameters	**94.28%**	**93.68%**
Feature selection	93.23%	92.20%
Random oversampling	85.99%	88.45%
SMOTE	93.25%	93.14%

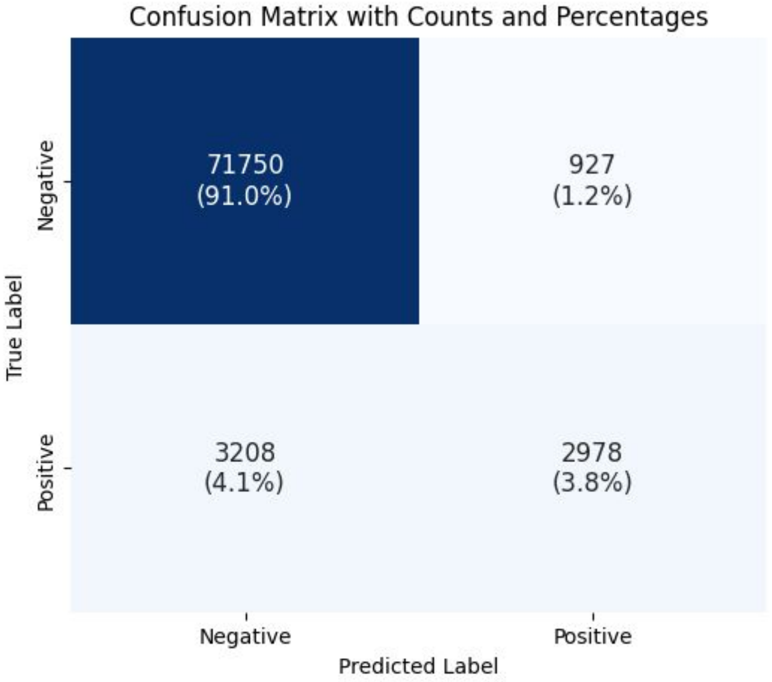
Results

Results Key Terms

- **Accuracy:** Correct predictions out of total predictions
- **Precision:** True positives out of predicted positives
- **Recall:** True positives out of actual positives
- **F1-Score:** Balance of precision and recall
- **Confusion Matrix:** Table showing actual vs. predicted values
- **Baseline Model:** Simple model for performance comparison
- **Feature Importance:** Measure of a feature’s impact on predictions

Model Evaluation and Metrics

After retraining on full training set:

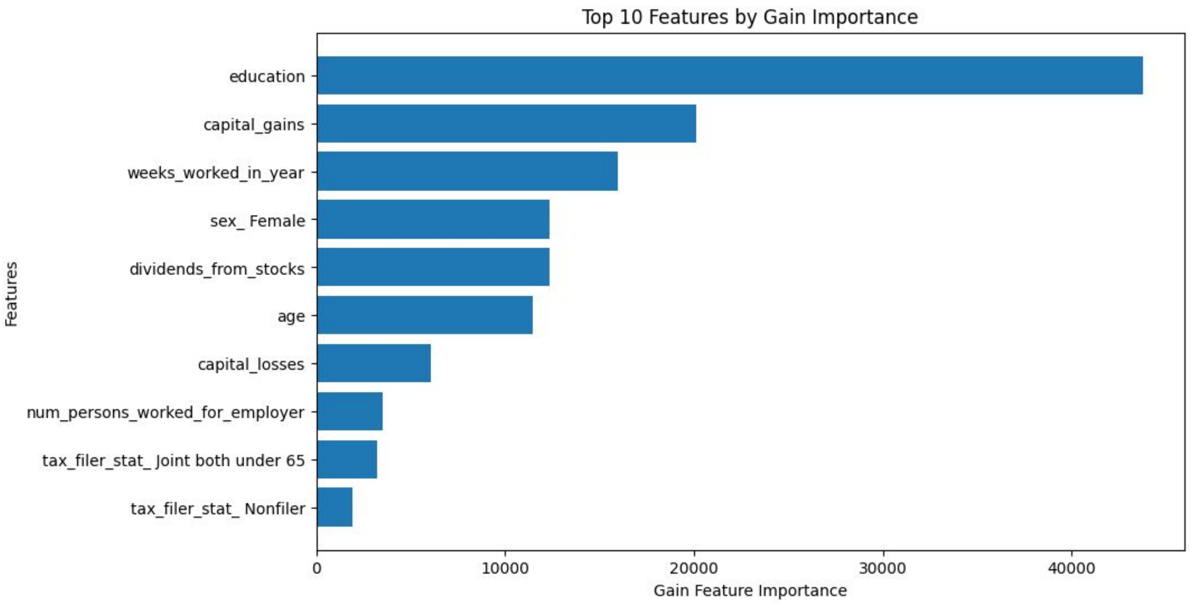


- Accuracy: 94.76%
- F1 Score (Weighted): 94.20%
- Precision (Weighted): 94.19%
- Recall (Weighted): 94.76%

Our model outperformed the 93.8% majority-class baseline

Feature Importance

- Education is most important feature
- Capital gains ranks second
- Demographic features like sex and age rank in the top 10
- Tax filer status is in the top 10



Conclusion

Recommendations

- Education is the strongest predictor of income class
- Demographic features rank highly as predictors of income class highlighting disparities
- Practical applications
  - Policymakers: design programs and policies around upskilling, workforce development, and wage fairness
  - Nonprofits & educators: Develop career guidance programs and align educational curricula with economic opportunities

Drive Adoption

- Dataiku Data Science Studio (DSS)
  - End-to-end machine learning workflow on one platform
  - Streamlined data preparation and feature engineering
  - Collaboration and reproducibility
  - Scalability and deployment
- Expert support
  - Data science advisors
  - Learning modules for Dataiku DSS
  - Office hours
  - Masterclasses

Next Steps

Short-Term

1. Deeper dive into each feature to build understanding
2. Explore missing values and steps to impute them
3. Impact of outliers and skewness
4. Rebinning categories
5. Trying more feature selection methods
6. Error analysis of ML models

Long-Term

7. Creating more features and incorporating new data
8. Responsible ML practices
9. Production-level ML pipeline with MLOps workflow (experiment tracking, orchestration, monitoring, CI/CD, etc.)

Questions and Discussion