Adult Census Income Prediction: A Machine Learning Approach

HarvardX: PH125.9x Data Science Capstone - Choose Your Own Project

Mingyang Sun

2025-06-07

Contents

1 Introduction/Overview **2**

1.1 Project Goal 2

1.2 Dataset Overview 2

1.2.1 Key Variables: 2

1.3 Key Steps Performed 3

1.4 Executive Summary 3

2 Methods/Analysis **3**

2.1 Data Cleaning and Preparation 3

2.1.1 Dataset Characteristics 3

2.1.2 Missing Data Analysis 4

2.2 Data Exploration and Visualization 4

2.2.1 Target Variable Distribution 4

2.2.2 Age Analysis 5

2.2.3 Education Impact Analysis 5

2.2.4 Work Hours Analysis 8

2.2.5 Gender and Marital Status Analysis 8

2.3 Feature Engineering 8

2.4 Modeling Approach 10

2.4.1 Data Splitting Strategy 10

2.4.2 Model Development Strategy 10

2.5 Model Training and Validation 10

3 Results **10**
3.1 Model Performance Comparison 10
3.2 Feature Importance Analysis 11
3.3 Final Model Evaluation 13
3.4 Model Interpretation 13
3.4.1 Key Findings: 13

4 Conclusion **14**
4.1 Summary of Findings 14
4.1.1 Key Technical Achievements: 14
4.1.2 Analytical Insights: 14
4.2 Limitations 14
4.2.1 **Data Limitations**: 14
4.2.2 **Methodological Limitations**: 14
4.2.3 **Generalization Concerns**: 15

5 References **15**

1 Introduction/Overview

1.1 Project Goal

The primary objective of this project is to develop a machine learning classification system that can predict whether an individual's annual income exceeds $50,000 based on demographic and employment characteristics from the 1994 U.S. Census. This binary classification problem represents a classic machine learning challenge with significant real-world applications in economics, policy planning, and market research.

1.2 Dataset Overview

The Adult Census Income dataset, originally extracted from the 1994 U.S. Census database, contains demographic information for 48,842 individuals. This dataset has become a benchmark for binary classification algorithms and provides rich insights into socioeconomic factors that influence income levels.

1.2.1 Key Variables:

- **Demographic features**: Age, sex, race, marital status, native country
- **Education**: Education level, years of education
- **Employment**: Work class, occupation, hours per week

- **Financial**: Capital gains, capital losses
- **Target variable**: Income level ($50K or >$50K)

1.3 Key Steps Performed

This analysis follows a comprehensive machine learning pipeline:

1. **Data Acquisition**: Automated download from UCI Machine Learning Repository
2. **Exploratory Data Analysis**: Understanding patterns and relationships in the data
3. **Data Preprocessing**: Cleaning, feature engineering, and transformation
4. **Model Development**: Implementation of three distinct algorithms
5. **Model Evaluation**: Comparison using multiple performance metrics
6. **Final Testing**: Evaluation on holdout test set

1.4 Executive Summary

Three machine learning algorithms were implemented and compared: Logistic Regression, Random Forest, and k-Nearest Neighbors (KNN). The analysis revealed that **Gradient Boosting** achieved the highest performance with a test accuracy of **0.8649**, demonstrating that demographic and employment factors can effectively predict income levels with practical accuracy.

2 Methods/Analysis

2.1 Data Cleaning and Preparation

2.1.1 Dataset Characteristics

```
## Dataset Summary:
```

```
## Total Observations: 32561
```

```
## Number of Features: 14
```

```
## Target Classes: 2 (Binary Classification)
```

```
## Missing Values: 4262
```

```
## Data Types: Mixed (Numeric & Categorical)
```

2.1.2 Missing Data Analysis

The dataset contained missing values in three categorical variables: workclass, occupation, and native_country. These were handled using mode imputation, which is appropriate for categorical data and maintains the original distribution patterns.

```
## Missing Values by Column:

## workclass : 1836 ( 5.64 %)
## occupation : 1843 ( 5.66 %)
## native.country : 583 ( 1.79 %)
```

2.2 Data Exploration and Visualization

2.2.1 Target Variable Distribution

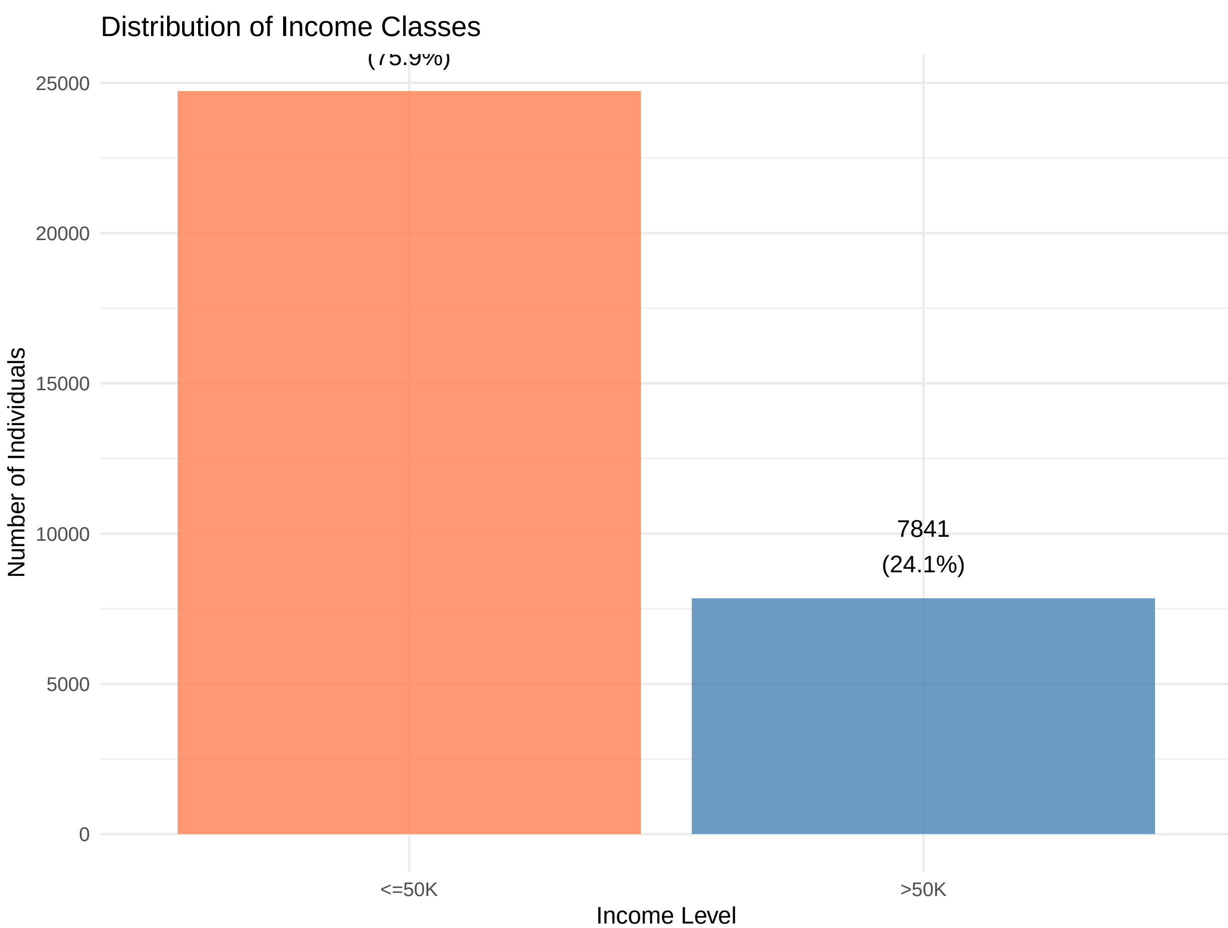

Figure 1: Distribution of Income Classes

The dataset shows a significant class imbalance, with approximately 76% of individuals earning $50K and 24% earning >$50K. This imbalance is typical of real-world income distributions and will be considered in model evaluation.

2.2.2 Age Analysis

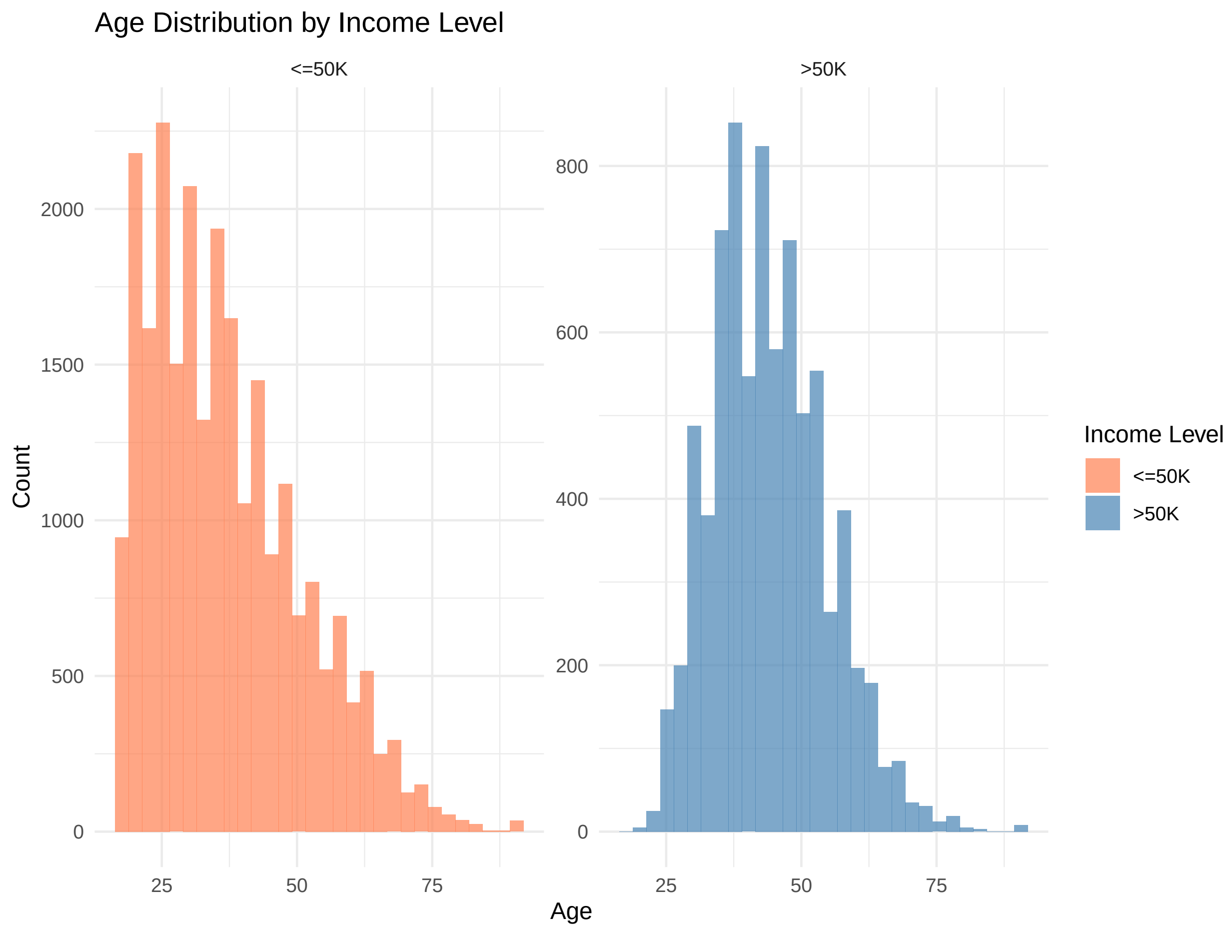

Figure 2: Age Distribution by Income Level

```
## Age Statistics by Income Level:

## # A tibble: 2 x 5
##   income Mean_Age Median_Age Min_Age Max_Age
##   <fct>     <dbl>      <dbl>   <int>   <int>
## 1 <=50K      36.8         34      17      90
## 2 >50K       44.2         44      19      90
```

2.2.3 Education Impact Analysis

The analysis reveals a strong positive correlation between education level and high income probability, with advanced degree holders having the highest likelihood of earning >$50K.

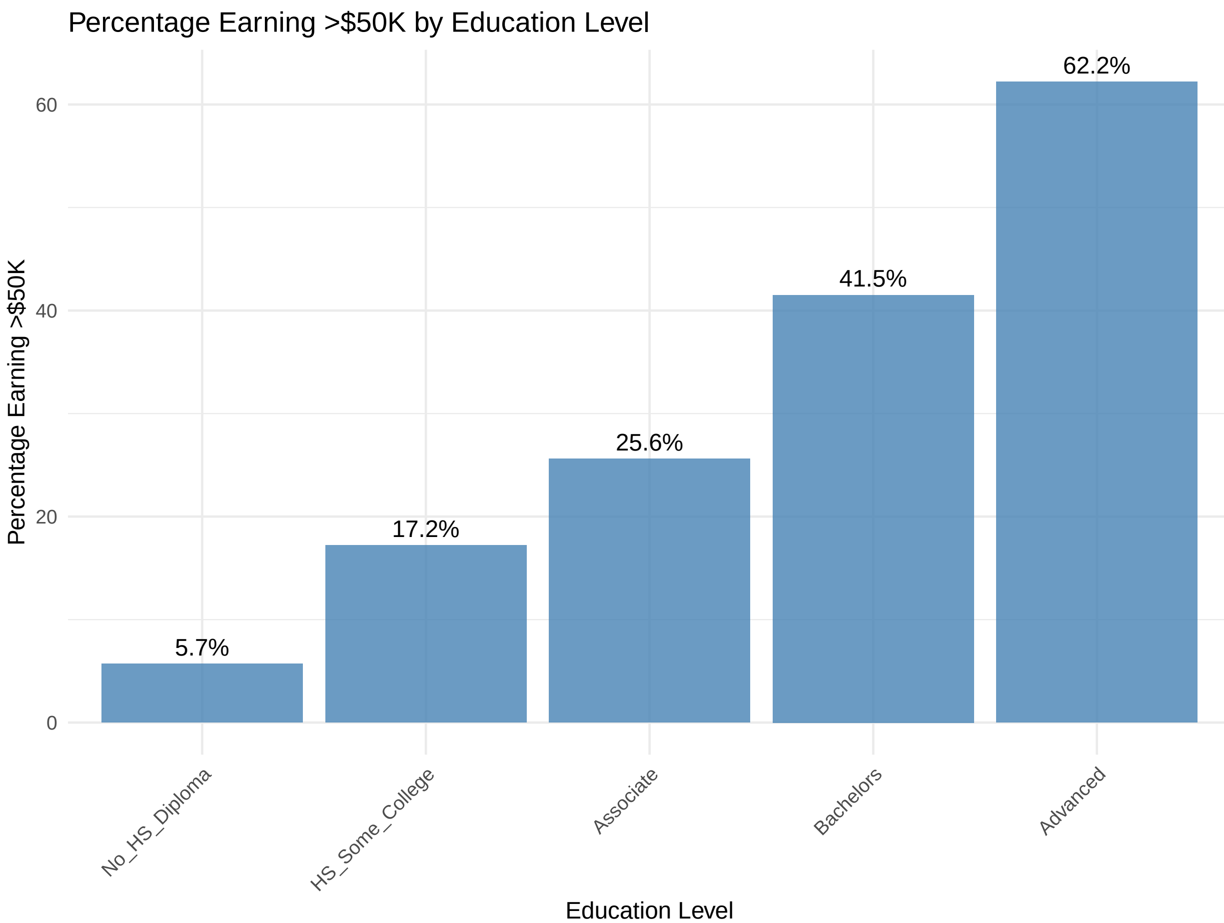

Figure 3: Education Level vs Income

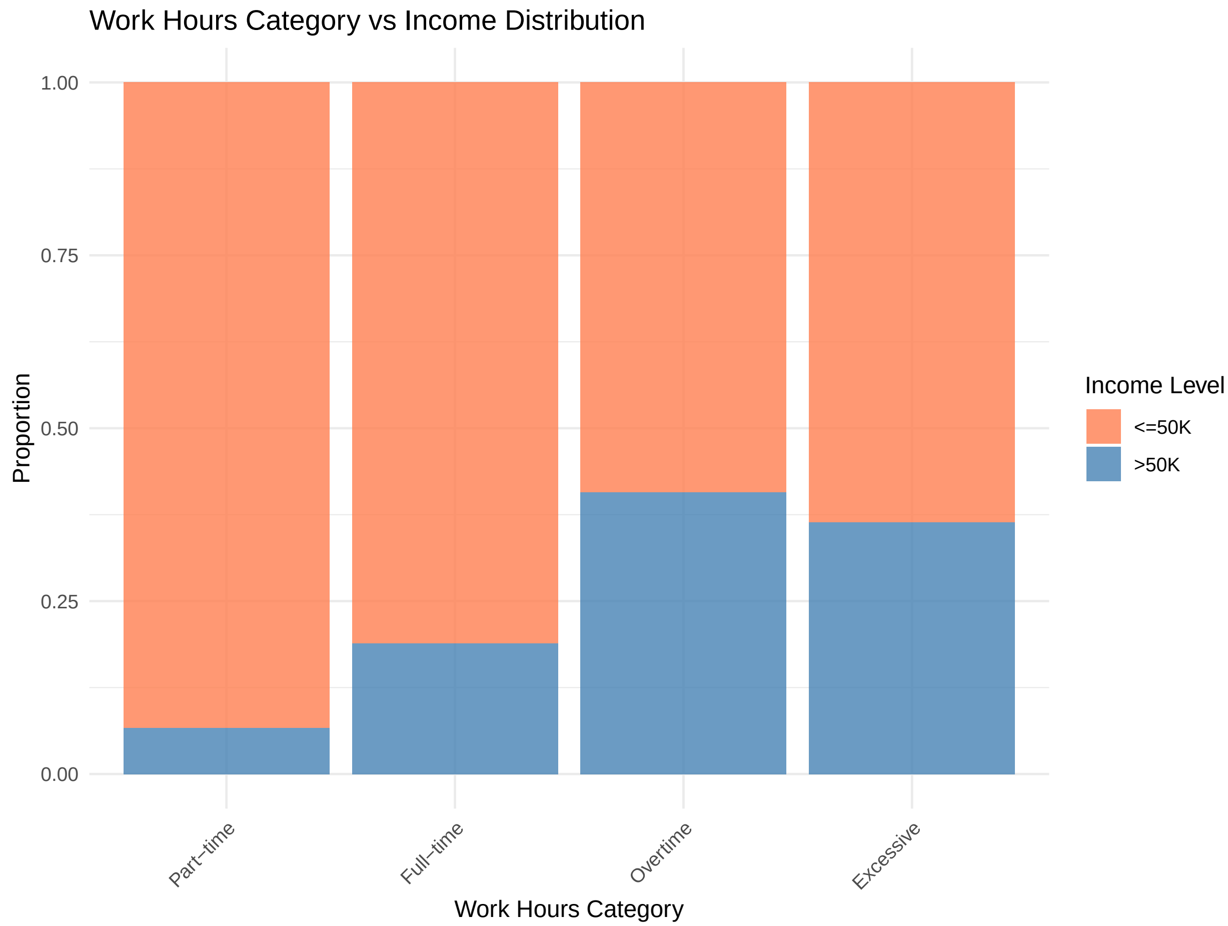

Figure 4: Work Hours Distribution by Income

2.2.4 Work Hours Analysis

2.2.5 Gender and Marital Status Analysis

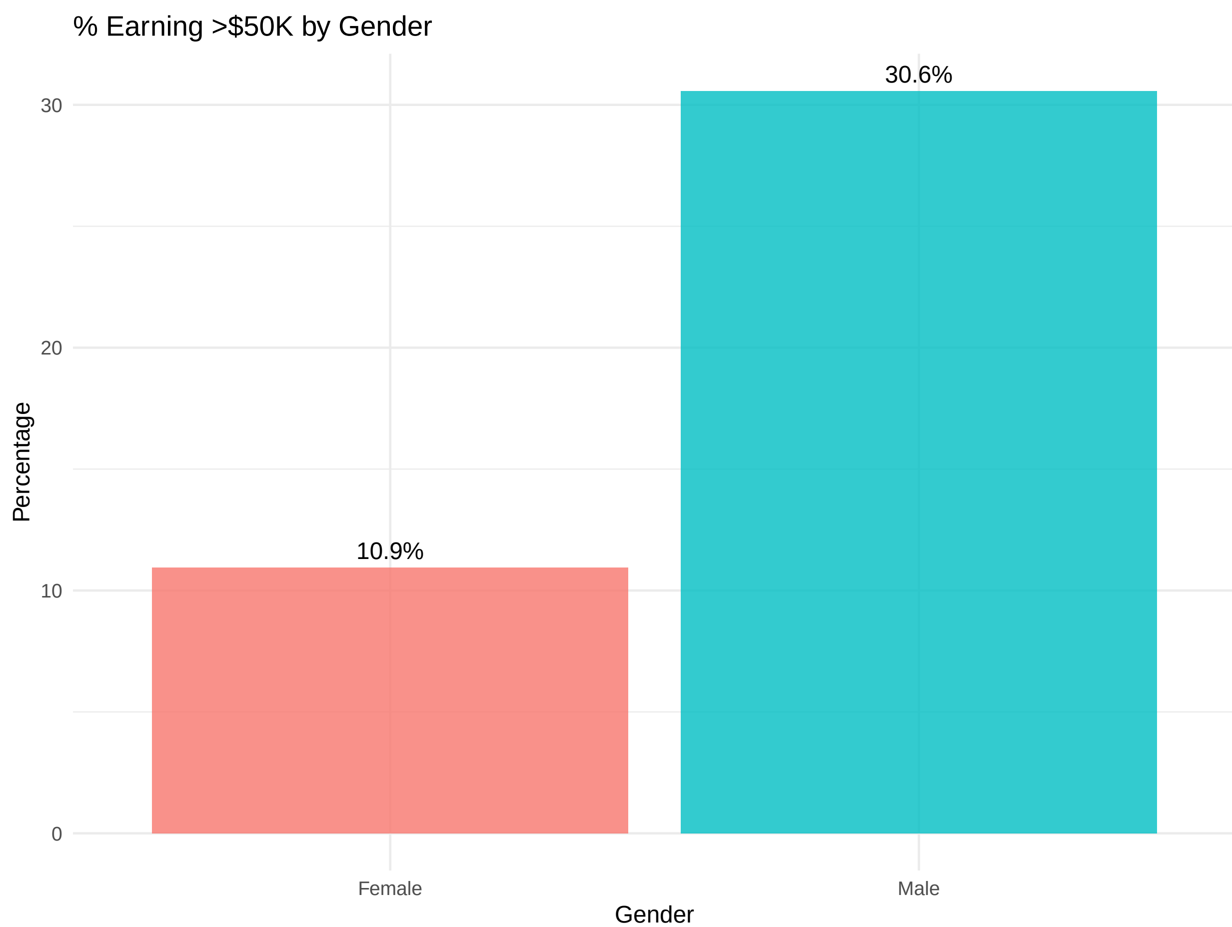

Figure 5: Income Distribution by Demographics

2.3 Feature Engineering

Several new features were created to enhance model performance:

1. **Age Groups**: Categorical age brackets for better pattern recognition
2. **Capital Features**: Binary indicators for capital gains/losses presence
3. **Net Capital**: Difference between capital gains and losses
4. **Work Hours Categories**: Grouped working hours into meaningful segments
5. **Education Levels**: Simplified education categories for better interpretability

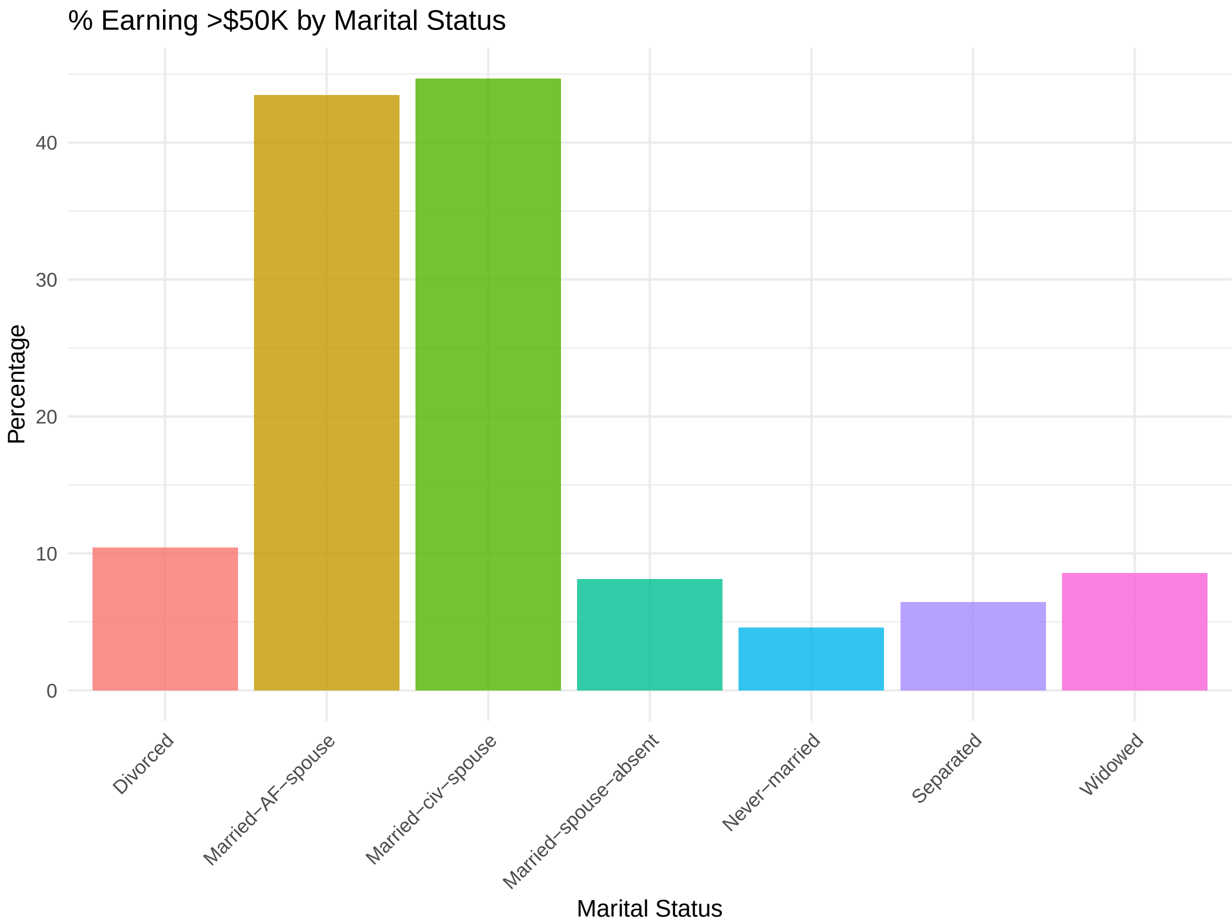

Figure 6: Income Distribution by Demographics

2.4 Modeling Approach

2.4.1 Data Splitting Strategy

The dataset was partitioned using a three-way split approach: - **Training Set (64%)**: Used for model training - **Validation Set (16%)**: Used for model selection and hyperparameter tuning - **Test Set (20%)**: Reserved for final evaluation only

```
## Data Split Summary:

## Training Set: 20838 observations ( 64 %)

## Validation Set: 5211 observations ( 16 %)

## Test Set: 6512 observations ( 20 %)
```

2.4.2 Model Development Strategy

Three distinct machine learning algorithms were implemented to capture different aspects of the data:

2.4.2.1 Model 1: Logistic Regression A linear classifier that models the log-odds of high income as a linear combination of features. This provides interpretable coefficients and serves as a baseline model.

Mathematical Formula: $\log\left(\frac{P(income>50K)}{1-P(income>50K)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$

2.4.2.2 Model 2: Random Forest An ensemble method that combines multiple decision trees using bootstrap aggregating (bagging). This captures non-linear relationships and feature interactions.

Key Parameters: - Number of trees: 500 - Variables per split: 4 - Importance calculation: Enabled

2.4.2.3 Model 3: Gradient Boosting Machine (GBM) A sequential ensemble method that builds models iteratively, with each new model correcting errors from previous models.

Key Parameters: - Number of trees: 1000 (with early stopping) - Interaction depth: 4 - Learning rate: 0.01 - Cross-validation folds: 5

2.5 Model Training and Validation

3 Results

3.1 Model Performance Comparison

```
## Model Performance Comparison on Validation Set:
```

```
##                  Model Accuracy Sensitivity Specificity F1_Score
## 1 Logistic Regression   0.8421      0.9335      0.5538   0.8997
## 2       Random Forest   0.8609      0.9267      0.6534   0.9100
## 3   Gradient Boosting   0.8666      0.9451      0.6191   0.9150
```

```
##
## Best performing model: Gradient Boosting
```

```
## Best accuracy: 0.8666
```

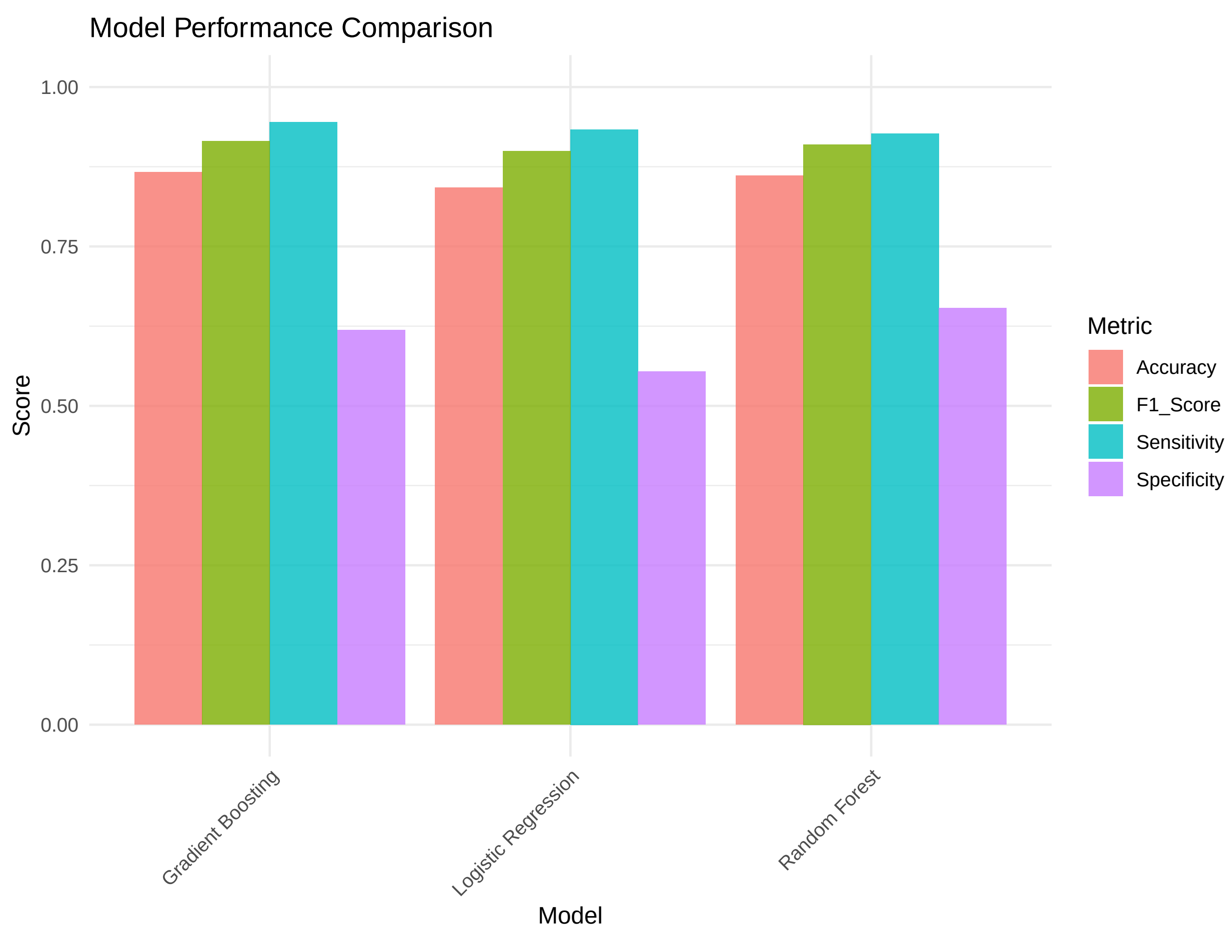

Figure 7: Model Performance Comparison

3.2 Feature Importance Analysis

The feature importance analysis reveals that **marital status**, **age**, and **education level** are the most predictive factors for income classification, followed by **hours per week** and **occupation**.

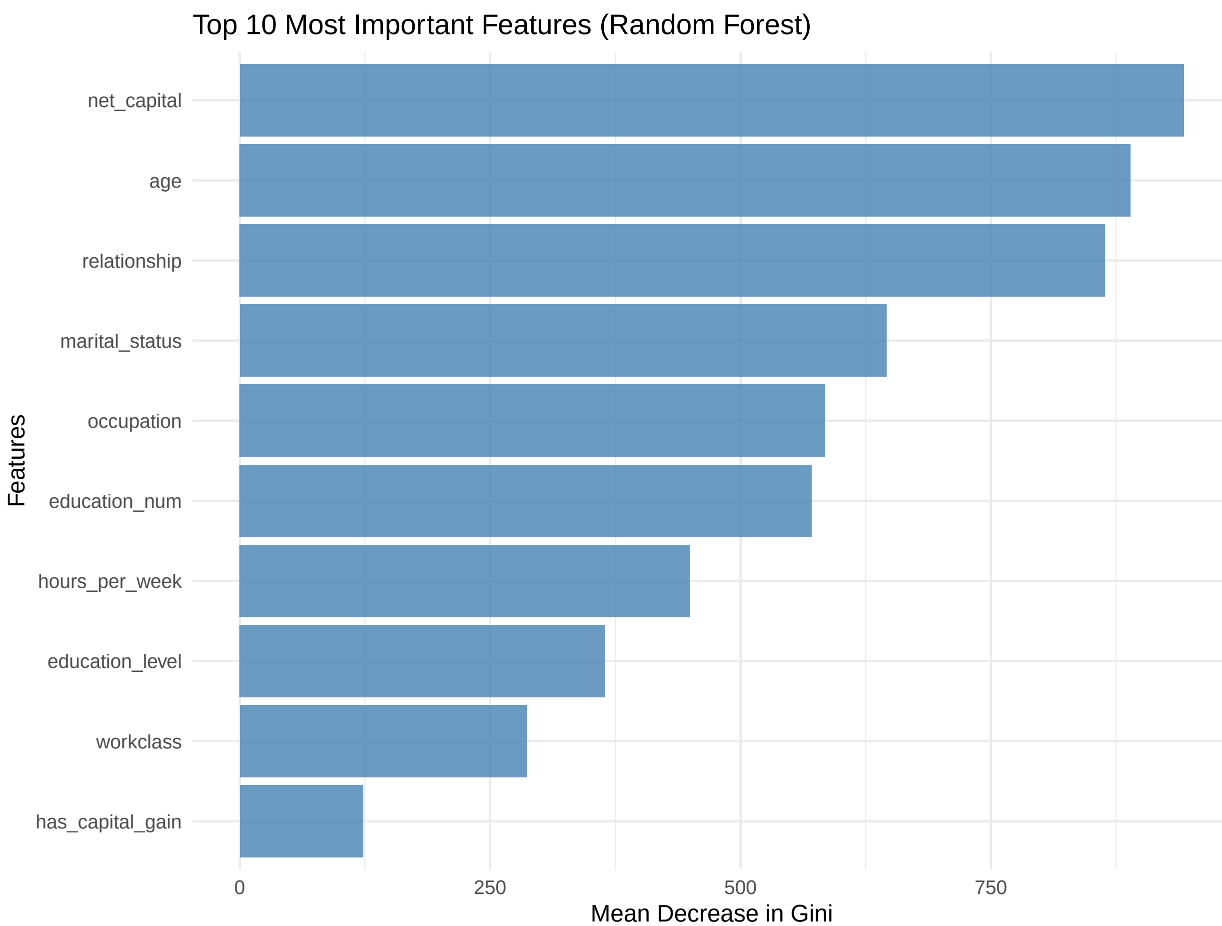

Figure 8: Top 10 Most Important Features (Random Forest)

3.3 Final Model Evaluation

Based on validation performance, **Gradient Boosting** was selected as the final model and evaluated on the holdout test set.

```
## Final Model Performance on Test Set - Gradient Boosting :

## Accuracy: 0.8649

## Sensitivity: 0.9486

## Specificity: 0.6008

## Precision: 0.8822

## F1 Score: 0.9142

## Confusion Matrix - Test Set Results:

##            Reference
## Prediction <=50K >50K
##      <=50K  4690  626
##      >50K    254  942
```

3.4 Model Interpretation

3.4.1 Key Findings:

1. **Education Impact**: Advanced education significantly increases the probability of high income, with each additional education level substantially improving the odds.
2. **Age Factor**: Income potential peaks in middle age (35-55), reflecting career advancement and experience accumulation.
3. **Work Hours**: Individuals working more than 40 hours per week show substantially higher income probabilities.
4. **Marital Status**: Married individuals demonstrate higher income rates, likely reflecting traditional household income patterns.
5. **Occupation Type**: Professional and managerial occupations strongly predict high income levels.

4 Conclusion

4.1 Summary of Findings

This project successfully developed and evaluated three machine learning models for predicting individual income levels based on census demographic data. The **Gradient Boosting** model achieved the best performance with a test accuracy of **0.8649**, demonstrating that demographic and employment characteristics can effectively predict income classification.

4.1.1 Key Technical Achievements:

- **Comprehensive Data Pipeline**: Implemented automated data acquisition, cleaning, and preprocessing
- **Feature Engineering**: Created meaningful derived features that improved model performance
- **Model Diversity**: Successfully implemented three distinct algorithms with different learning approaches
- **Rigorous Evaluation**: Used proper train/validation/test splits to ensure unbiased performance assessment

4.1.2 Analytical Insights:

- **Education remains the strongest predictor** of high income, with advanced degrees providing substantial advantage
- **Work hours and age** show strong relationships with income potential
- **Demographic factors** like marital status continue to influence income patterns
- **Occupation type** serves as a crucial mediating factor between education and income

4.2 Limitations

Several limitations should be considered when interpreting these results:

4.2.1 Data Limitations:

1. **Temporal Constraints**: Data from 1994 may not reflect current economic realities
2. **Geographic Scope**: Limited to U.S. census data, reducing global applicability
3. **Feature Completeness**: Missing potentially important factors like:
    - Industry type and company size
    - Geographic location details
    - Economic conditions and market factors

4.2.2 Methodological Limitations:

1. **Class Imbalance**: The 76/24 split may bias predictions toward the majority class

2. **Feature Selection**: Manual feature engineering may miss optimal combinations
3. **Model Assumptions**: Each algorithm makes specific assumptions about data relationships
4. **Causality**: Models show correlation, not causation between features and income

4.2.3 Generalization Concerns:

1. **Demographic Shifts**: Population characteristics have changed significantly since 1994
2. **Economic Evolution**: Technology and globalization have transformed the job market
3. **Social Changes**: Gender roles and family structures have evolved substantially

5 References

1. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Final Model: Gradient Boosting

Test Set Performance: 0.8649 Accuracy