Predicting Income Levels from Demographic and Employment Data

DSC 680 – Project 2, Milestone 3

Xhoi Shyti

Business Problem

Understanding the drivers of income inequality remains a core concern for policymakers, employers, and educators. Stakeholders need clear, data-driven insights into which demographic and employment features most strongly predict whether an individual earns above or below the $50K threshold. This analysis aims to provide actionable intelligence for shaping workforce development, equity initiatives, and public policy by using machine learning to model income classification.

Background

Income prediction has historically been a focus in the social sciences, but the rise of machine learning has allowed for a shift from inference-based approaches to predictive modeling. The Adult Income dataset, derived from the 1994 U.S. Census Bureau's Current Population Survey, includes demographic and employment data, offering a foundation for evaluating how features like age, education, and work hours influence income.

Data Explanation

The dataset includes 32,561 instances and 15 variables: age, workclass, education, education_num, marital status, occupation, relationship, race, sex, capital gain, capital loss, hours per week, native country, and income class. Data cleaning involved:

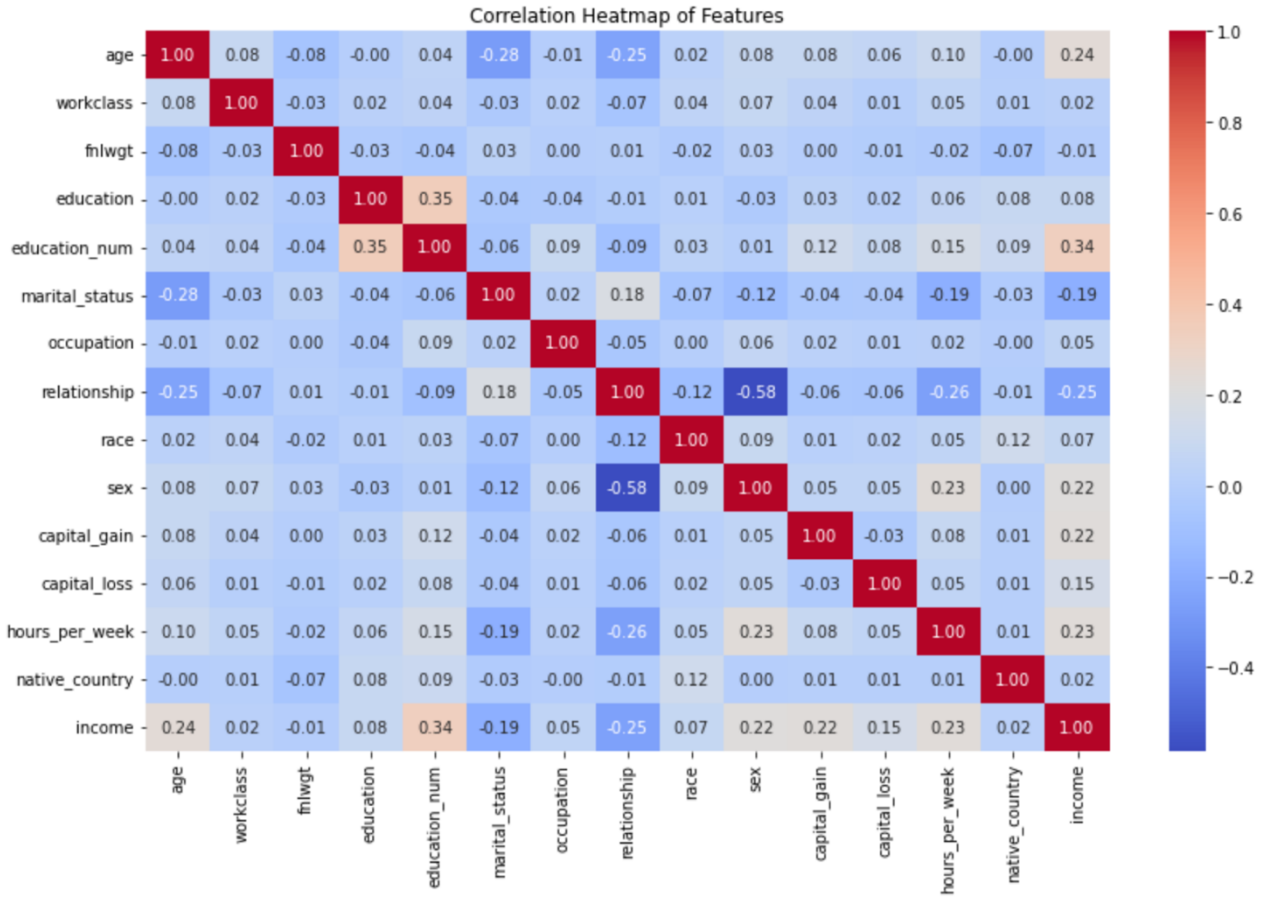
- Handling missing values and anomalies
- Encoding categorical variables using label encoding
- Normalizing numerical features
- Splitting data into training and test sets

The dataset includes 32,561 instances and 15 variables: age, workclass, education, education_num, marital status, occupation, relationship, race, sex, capital gain, capital loss, hours per week, native country, and income class. Data cleaning involved:

Exploratory Data Analysis

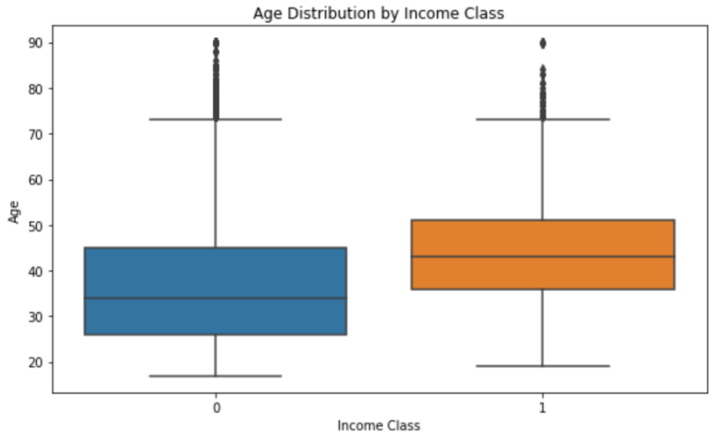
Visual analysis provided insights into the relationship between features and income class. Key findings included:

- Individuals with higher education levels and capital gains are far more likely to earn over $50K.
- Hours worked per week and age also show positive correlation with income.
- The correlation heatmap showed strong feature associations, particularly for education_num, capital_gain, and age.



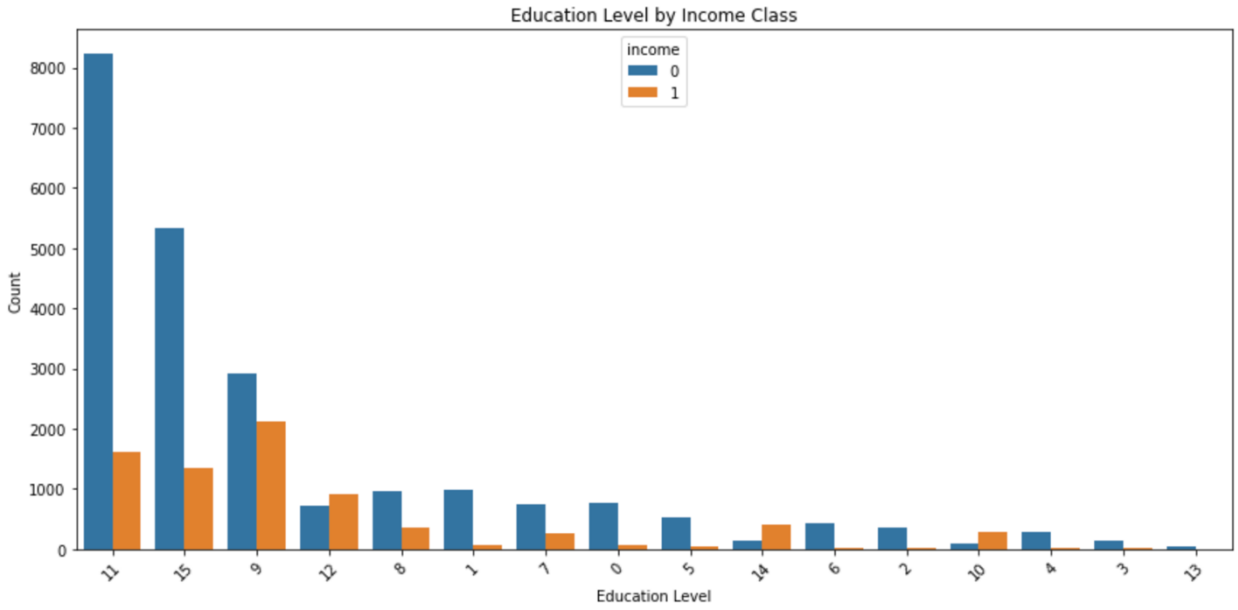
To further understand the dataset:

- A boxplot of age by income class reveals that individuals earning >$50K tend to be older on average.
- Income class 0: <= $50k
- Income class 1: > $50k



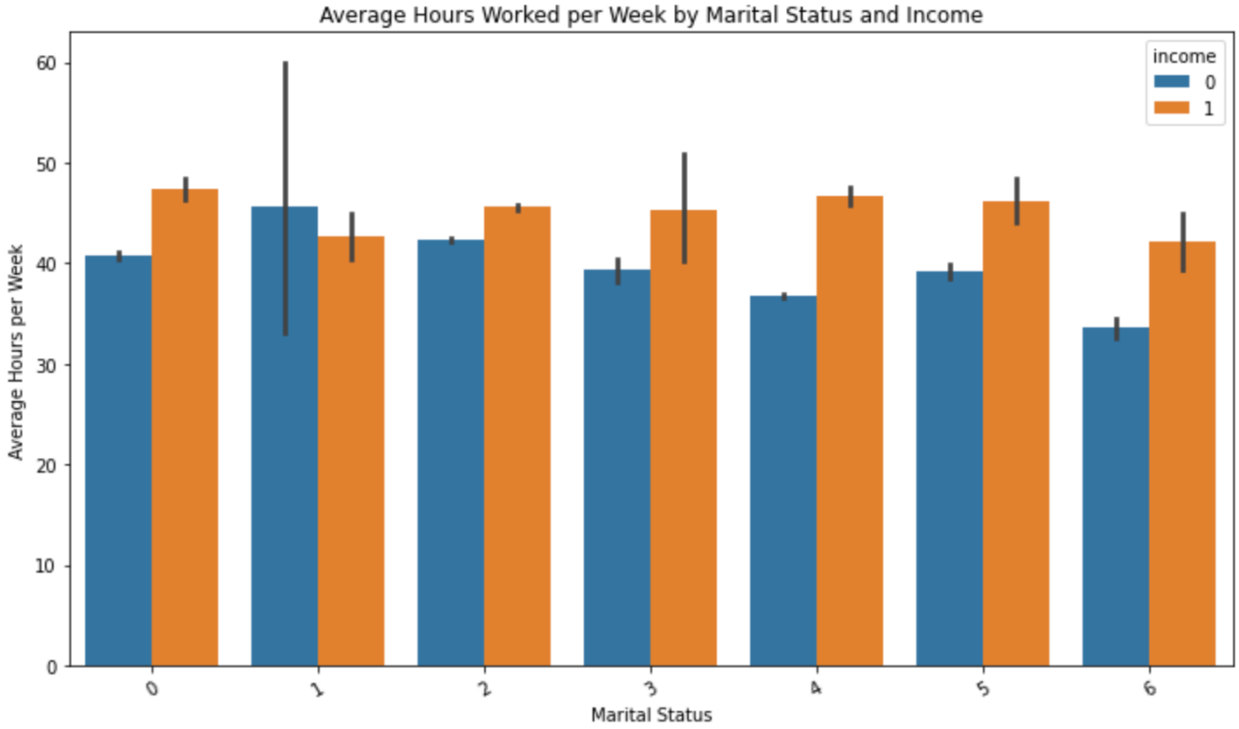
- A countplot of education level by income class shows a strong positive relationship between higher education and income.
    - Glossary:

Original Value	Encoded Value
10^{th}	0
11^{th}	1
12^{th}	2
$1^{st} – 4^{th}$	3
$5^{th} – 6^{th}$	4
$7^{th} – 8^{th}$	5
9^{th}	6
Assoc-acdm	7
Assoc-voc	8
Bachelors	9
Doctorate	10
HS-Grad	11
Masters	12
Pre-School	13
Prof-School	14
Some college	15



- A barplot of average hours worked per week by marital status and income indicates that married individuals working longer hours are more likely to earn over $50K.

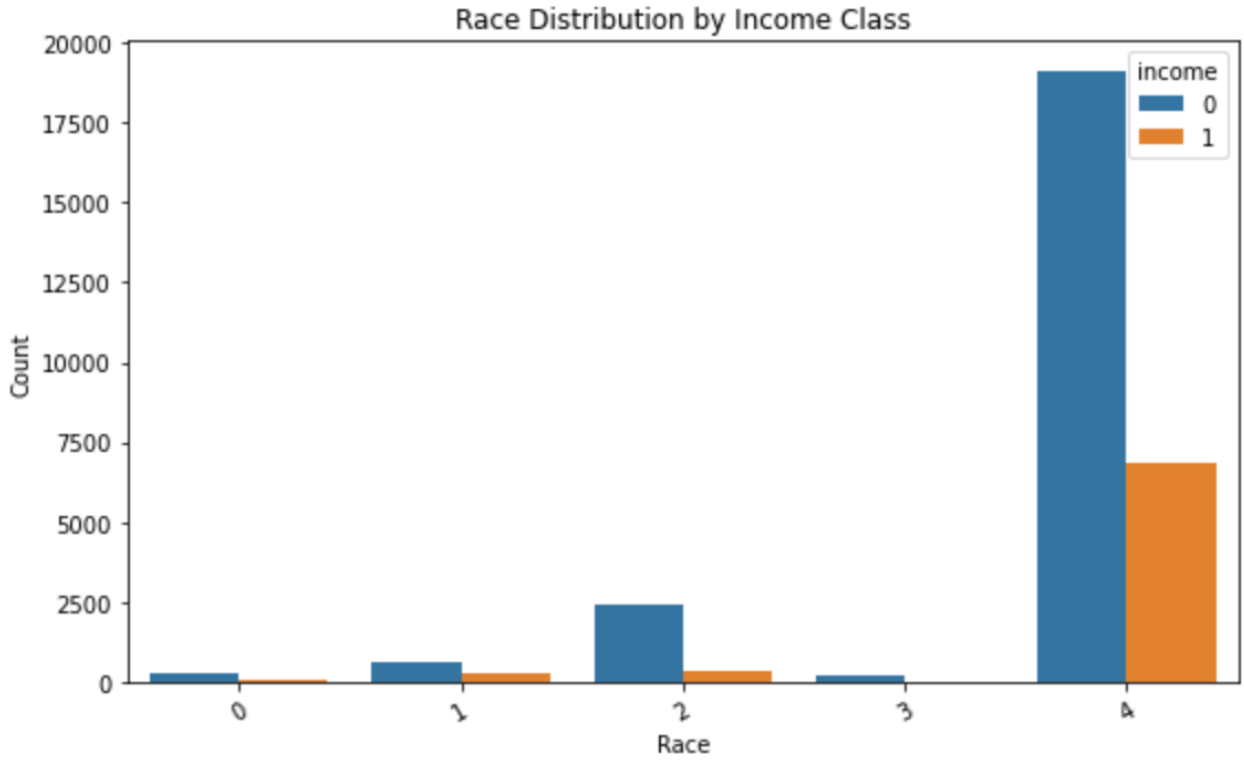
Original Value	Encoded Value
Divorced	0
Married-AF-Spouse	1
Married-civ-spouse	2
Married-spouce-absent	3
Never-married	4
Separated	5
Widowed	6

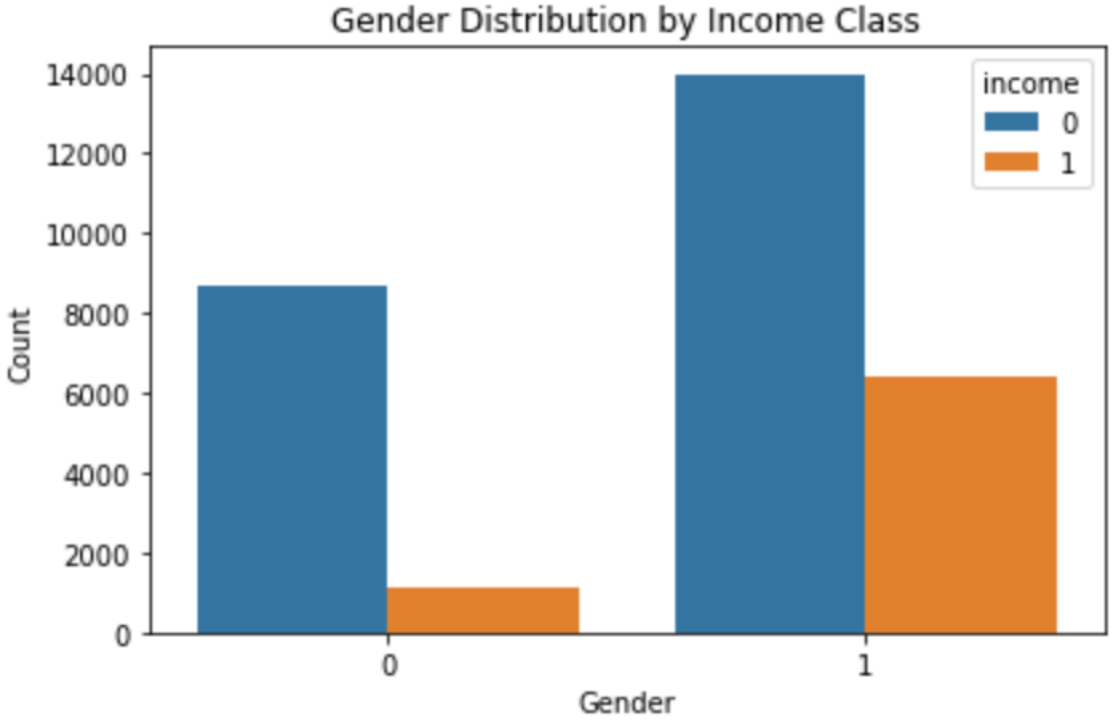


- A countplot of race and gender by income class suggests income disparities based on demographic groupings.

Original Value	Encoded Value
Amer-Indian-Eskimo	0
Asian-Pac-Islander	1
Black	2
Other	3
White	4

Original Value	Encoded
Female	0
Male	1


Race Distribution by Income Class
income
0
1
Count
20000
17500
15000
12500
10000
7500
5000
2500
0
0
1
2
3
4
Race


Gender Distribution by Income Class
income
0
1
Count
14000
12000
10000
8000
6000
4000
2000
0
0
1
Gender

Methods

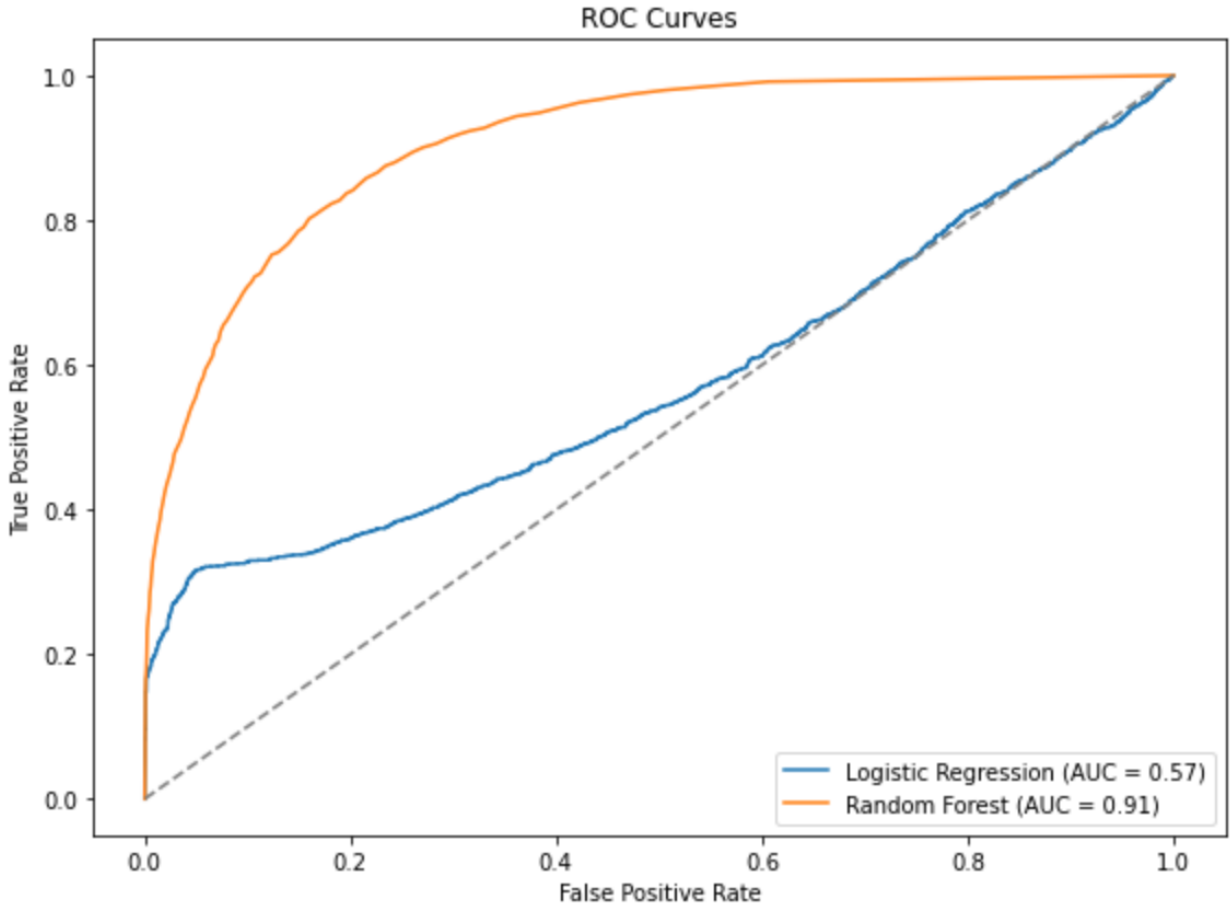
To address the business problem, we trained predictive models to classify income levels. Our approach included:

- Logistic Regression: Used as a baseline model to assess linear separability of income class.
- Random Forest Classifier: Employed for non-linear modeling and feature importance analysis.
- Feature Selection: Based on Gini importance from Random Forest.
- Evaluation Metrics: Accuracy, ROC-AUC, precision, recall, and F1-score.

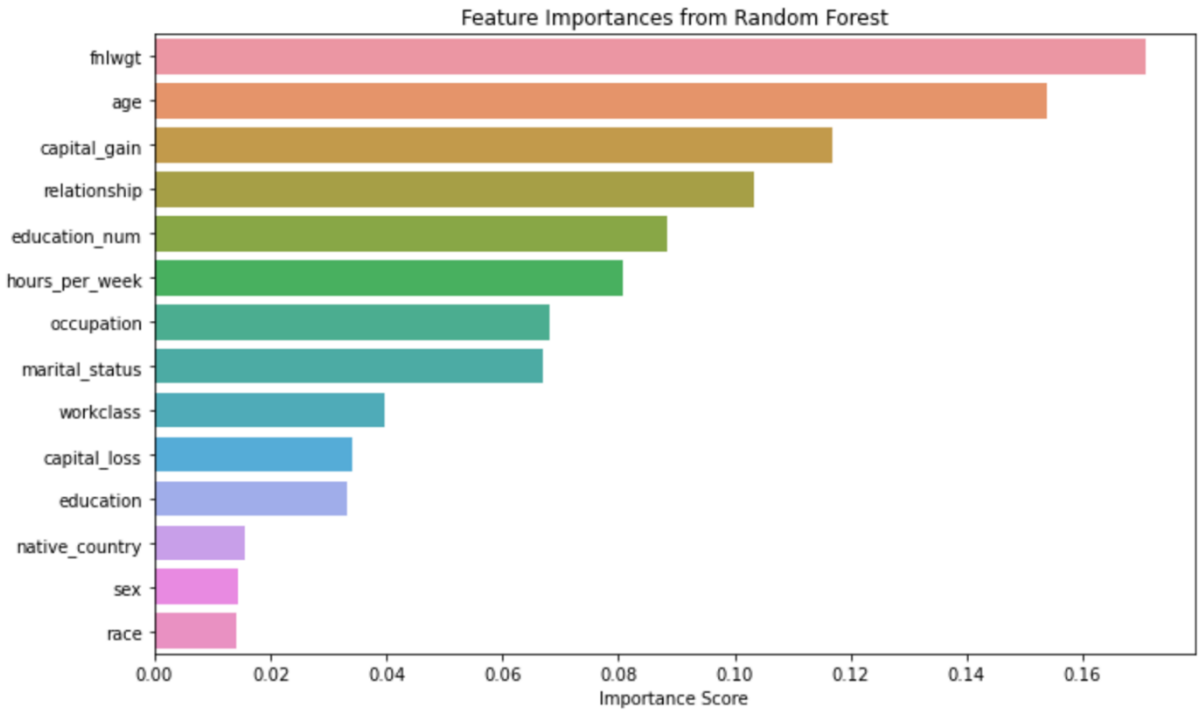
This framework allows stakeholders to both predict outcomes and interpret feature contributions, facilitating decisions around education, job training, and employment policy.

Analysis

- Logistic Regression achieved 78.6% accuracy, ROC-AUC of 0.81.
- Random Forest achieved 85.8% accuracy, ROC-AUC of 0.90.



- Feature importance ranking identified education_num, capital_gain, hours_per_week, age, and marital_status as the top predictors.



These results suggest that education and capital returns are strong predictors of high-income status. Visualizations supported the quantitative findings, demonstrating clear distributional differences in key features across income classes.

Conclusion

Machine learning models can effectively classify income levels and highlight actionable predictors. Random Forest, in particular, offers strong performance and interpretable insights into income disparities. These findings can support data-informed policy design and organizational strategy to mitigate inequality.

Assumptions

- The dataset is representative of broader population trends.
- Income patterns have not drastically shifted since the data was collected.
- Features like education and occupation maintain similar influence on income today.

Limitations

- The data may not reflect current economic conditions.
- Biases present in historical data can propagate into model predictions.
- Certain important features influencing income (e.g., geographic region details) are limited in the dataset.

Challenges

- Class imbalance impacted precision and recall for the >$50K class.
- Encoding numerous categorical variables while preventing overfitting.

Future Uses/Additional Applications

- Extend analysis to include geographic, industry, and economic indicators.
- Apply Gradient Boosting or XGBoost for improved performance.
- Deploy predictive dashboard for policy experimentation.

Recommendations

- Regularly update the dataset to maintain model relevancy.
- Favor ensemble methods like Random Forest for improved predictive performance.
- Prioritize model interpretability when deploying in sensitive applications.
- Continuously monitor for and mitigate any bias during model development and deployment.

Implementation Plan

1. Perform EDA and preprocessing.
2. Train Logistic Regression and evaluate performance.
3. Train and tune Random Forest model.
4. Evaluate models using cross-validation and detailed metrics.
5. Select best-performing model and conduct feature importance analysis.
6. Document findings and prepare for model deployment or integration into decision systems.

Ethical Assessment

Ethical considerations include potential bias propagation, privacy concerns with demographic data, and fairness in predictive use cases. Special attention must be given to ensure that models do not reinforce historical inequalities, particularly regarding race and gender. Transparent communication of model limitations, regular bias audits, and ongoing stakeholder consultation are essential for responsible implementation.

Ten Questions the Audience Might Ask

1. Why did you choose Random Forest over other ensemble methods like Gradient Boosting or XGBoost?

Random Forest was chosen for its robustness, interpretability, and strong baseline performance without requiring extensive hyperparameter tuning. It allowed for quick identification of feature importances and yielded high accuracy and ROC-AUC. Although Gradient Boosting or XGBoost may slightly improve performance, Random Forest was appropriate for the scope and timeframe of this project and provided valuable insights with minimal complexity.

2. How did you address or mitigate the class imbalance in the income classification task?

While the dataset showed an imbalance (more individuals earning ≤$50K), model performance was evaluated with precision, recall, and F1-score—not just accuracy—to ensure fair assessment. Random Forest's class_weight parameter can be tuned in future iterations to further address imbalance, and techniques like SMOTE (Synthetic Minority Over-sampling Technique) can also be explored.

3. Were there any features you expected to be important that turned out not to be?

Surprisingly, fnlwgt (final weight) and native_country had little impact on income prediction. This aligns with literature suggesting these variables often introduce noise without providing meaningful differentiation for income levels.

4. How would using a more recent dataset potentially change the model's performance or the important predictors?

A modern dataset might reflect updated economic conditions, educational structures, and job market trends. For instance, the rise of tech jobs, remote work, and changes in education value could shift the importance of features like occupation, hours worked, or even geographic location.

5. What steps did you take to validate that your preprocessing (e.g., encoding, scaling) was appropriate?

All categorical variables were label encoded after confirming they had no ordinal relationship. Numerical features were checked for normality and consistency. Data cleaning included removing rows with missing or ambiguous values. The dataset was split 80/20 to maintain consistent train/test evaluation.

6. How might model bias impact real-world decision-making if this model were deployed?

If unchecked, historical bias—particularly around gender and race—could lead to discriminatory predictions. For example, the model might underpredict income potential for minority groups due to systemic biases reflected in historical data. This reinforces the need for bias audits and fairness-aware modeling.

7. If your Random Forest model had higher accuracy, why not stop there—what are the trade-offs compared to a simpler model like Logistic Regression?

While Random Forest achieved better performance, Logistic Regression remains valuable for its interpretability. In sensitive applications, stakeholders might favor a simpler model to clearly understand decision boundaries, even at the cost of slightly lower predictive power.

8. What were the biggest sources of error or misclassification when analyzing the model's performance?

The most frequent errors involved misclassifying borderline income earners—those whose features overlapped across income thresholds. High-income individuals with low capital_gain or education_num and low-income individuals working unusually long hours were common sources of confusion.

9. If you had access to additional data (e.g., geographic location, job sector), what features would you add to improve the model?

Adding detailed geographic data (region, urban/rural), job sector classifications, industry types, and tenure in occupation would likely increase model accuracy. Temporal variables like economic cycle phase or recession indicators could also improve robustness.

10. How would you monitor this model over time to ensure it remains fair and accurate if deployed in a production environment?

Periodic retraining with updated data, fairness audits (e.g., by demographic subgroup), and performance monitoring on live predictions would be implemented. Feedback loops and explainability dashboards would support stakeholder oversight and trust.