Final Project Report

Predicting Income Levels Using the UCI Adult Census Dataset

# Abstract

This project aimed to predict whether an individual earns more than $50K annually using demographic and employment-related attributes from the UCI Adult Census dataset. The analysis included comprehensive data preprocessing, exploratory data analysis (EDA), and the development of two classification models: Logistic Regression and Random Forest. The models were evaluated using accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC). Logistic Regression achieved slightly better performance in terms of AUC, while Random Forest demonstrated competitive accuracy and recall. The results highlight key income predictors, including education level, capital gains, and hours worked per week, and provide insights for future model improvement.

Keywords: income prediction, logistic regression, random forest, machine learning, data preprocessing, exploratory data analysis

# 1. Introduction

The UCI Adult Census dataset contains demographic and employment data for individuals in the United States. The primary objective of this project was to develop predictive models to determine whether an individual’s annual income exceeds $50K. Predicting income levels is valuable for socio-economic analysis, policy-making, and targeted public programs, as it can reveal factors that influence earning potential.

# 2. Dataset Description

## 2.1 Overview

The dataset consists of over 48,000 records and 15 attributes, including demographic information, work-related variables, and the target variable income.

## 2.2 Variable Descriptions

|  |  |  |
| --- | --- | --- |
| Feature | Description | Type |
| age | Age of the individual | Numeric |
| workclass | Type of employment | Categorical |
| fnlwgt | Census final weight (sample size indicator) | Numeric |
| education | Highest education level attained | Categorical |
| education\_num | Years of education | Numeric |
| marital\_status | Marital status | Categorical |
| occupation | Type of occupation | Categorical |
| relationship | Relationship status | Categorical |
| race | Race of the individual | Categorical |
| sex | Gender of the individual | Categorical |
| capital\_gain | Monetary gains from capital investments | Numeric |
| capital\_loss | Monetary losses from capital investments | Numeric |
| hours\_per\_week | Average hours worked per week | Numeric |
| native\_country | Country of origin | Categorical |
| income | Target variable — 0 for <=50K, 1 for >50K | Target |

# 3. Methodology

## 3.1 Preprocessing

The preprocessing pipeline included handling missing values, trimming whitespace from categorical data, encoding the target variable as binary integers, removing duplicates, applying one-hot encoding to categorical features, and scaling numerical features using StandardScaler while excluding the target variable.

## 3.2 Exploratory Data Analysis (EDA)

Target Distribution: The dataset is imbalanced, with approximately 75% of individuals earning <=$50K and 25% earning >$50K.  
  
Correlation Analysis: education\_num, capital\_gain, and hours\_per\_week showed the strongest positive correlations with income.  
  
Key Insights: Higher education levels, professional occupations, and significant capital gains are associated with higher income. Longer working hours also correlate with higher income levels.

# 4. Modeling and Validation

## 4.1 Models Used

Logistic Regression: Selected for interpretability and suitability for linearly separable data.  
Random Forest Classifier: Selected for its ability to model complex, non-linear interactions.

## 4.2 Evaluation Strategy

The data was split into 80% training and 20% testing sets. Five-fold cross-validation was performed to assess generalization performance. Metrics included accuracy, precision, recall, F1-score, and AUC.

# 5. Results

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | CV Mean Accuracy | Test Accuracy | Precision | Recall | F1 Score | AUC |
| Logistic Regression | 0.8524 | 0.8490 | 0.7434 | 0.5938 | 0.6602 | 0.9046 |
| Random Forest | 0.8530 | 0.8478 | 0.7300 | 0.6091 | 0.6641 | 0.8995 |

Confusion Matrix Insights: Logistic Regression correctly classified 6,854 low-income and 1,431 high-income cases but misclassified 979 high-income cases. Random Forest correctly classified 6,895 low-income and 1,468 high-income cases with slightly fewer false negatives.  
  
ROC Curve Analysis: Both models showed strong discriminatory performance, with Logistic Regression achieving a slightly higher AUC.

# 6. Discussion

Both models achieved similar performance, with Logistic Regression slightly outperforming Random Forest in terms of AUC (0.9046 vs. 0.8995). Logistic Regression was selected as the preferred model due to its interpretability and marginally better ranking performance. Random Forest remains a strong alternative for capturing non-linear feature interactions. Addressing class imbalance through oversampling methods such as SMOTE may further improve recall.

# 7. Conclusion

This project successfully built predictive models for income classification using the UCI Adult Census dataset. Both Logistic Regression and Random Forest achieved approximately 85% accuracy and AUC values around 0.90. Education level, capital gains, and hours worked per week emerged as key predictors of income. Future work could include hyperparameter tuning, oversampling, and additional feature engineering.

# References

Dua, D., & Graff, C. (2019). UCI machine learning repository: Adult data set. University of California, Irvine, School of Information and Computer Sciences. https://archive.ics.uci.edu/ml/datasets/adult

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

Wickham, H., & Grolemund, G. (2017). R for data science: Import, tidy, transform, visualize, and model data. O’Reilly Media. https://r4ds.had.co.nz/