Census Income Prediction: A Machine Learning Approach

Sankit (2022446), Sanyam Barwar (2022447),
Sanyam Garg (2022448), Vivan Rangra (2022583)

Supervisor: Dr. Jainendra Shukla
Indraprastha Institute of Information Technology, Delhi
jainendra@iiitd.ac.in

Abstract

Predicting income levels based on demographic and so cioeconomic factors is essential for various applications like policy-making, targeted marketing, and financial plan ning. This project focuses on comparing multiple machine learning models—such as Logistic Regression, Decision Trees, and Random Forest—to identify the most effective approach for predicting whether an individual's income is above $50,000. The Census Income dataset is used to train and evaluate the models. Initial results indicate that Ran dom Forest provides the highest accuracy, though further fine-tuning and testing of models like Support Vector Ma chines and Neural Networks are required.

1. Introduction

Income prediction is a crucial problem with practical implications for governments and businesses. Predicting whether an individual's income exceeds $50,000 based on features like age, education, occupation, and marital status can aid in better decision-making for policy development, targeted financial services, and marketing strategies. In this project, we tackle this problem using machine learning al gorithms to assess which model offers the best trade-off be tween accuracy and computational efficiency.

Binary classification is employed to predict two classes: income greater than $50,000 or income less than or equal to $50,000. Our motivation stems from the growing demand for data-driven solutions in the socio-economic sector. The project also serves as a comparative study of several ma chine learning algorithms, aiming to find the most reliable model for income classification.

1.1. Problem Statement

Income prediction based on demographic data is a com plex task due to the multiple factors influencing income

levels. Our objective is to classify individuals into one of two categories: those earning more than $50,000 and those earning less. We aim to determine which machine learn ing model is best suited for this binary classification task in terms of accuracy, interpretability, and computational cost.

2. Literature Survey

A comprehensive literature review was conducted to un derstand existing methods in income classification using machine learning algorithms. The two most influential stud ies that guided our approach are outlined below:

- Paper 1: Census Income Prediction using Machine Learning Techniques. This paper applies multiple machine learning models—such as Logistic Regres sion, Decision Trees, and Random Forest—on the UCI Adult Income dataset. It highlights that Random For est outperforms other models in terms of accuracy and robustness, particularly by reducing overfitting through ensemble methods. This study strongly influ enced our decision to prioritize Random Forest in our experimentation, considering its balance between ac curacy and generalization.
- Paper 2: Comparative Analysis of Machine Learn ing Algorithms in Socioeconomic Data. This study compares various machine learning models, including Neural Networks, Decision Trees, and SVMs, across several socioeconomic datasets, focusing on the trade offs between accuracy and interpretability. Complex models like Neural Networks showed higher accuracy, while simpler models like Decision Trees were

easier to interpret. The paper influenced our decision to test both complex models for accuracy and simpler models for interpretability, ensuring that our model selection is well-rounded.

These findings underscore the importance of balancing model accuracy with interpretability. Based on the insights
from these papers, we decided to focus on Random For est and Logistic Regression as part of our experimental ap proach.

3. Dataset and Preprocessing

3.1. Dataset Overview

The dataset used in this study is the Census Income Dataset, often referred to as the Adult dataset. It contains demographic and employment-related attributes of individ uals, such as age, education, marital status, occupation, and work class. The target variable is binary: whether an indi vidual earns more than $50K per year or less than or equal to $50K. The dataset contains 48,842 instances and 14 fea tures, including both categorical and numerical data types.

Figure 1. Dataset Description

3.2. Data Preprocessing

1. Data Cleaning

- Fixed income column by merging (”≤50K.”, ”>50K.”) with (”≤50K”, ”>50K”)
- Replaced ”?” with ”Unknown” in features - ’workclass’, ’occupation’, and ’native-country’
- Removed duplicate ”education” column, which is a duplicate feature of ”education-num”
- Dropped approximately 1200 rows with missing values

2. Feature Engineering

- Applied label encoding to categorical features
- Used standard scaling for continuous numerical features

4. EDA

EDA is crucial as it helps understand data characteristics, identify potential problems (like class imbalance, missing values, or outliers). It provides insights that inform model selection, feature engineering strategies, and helps antici pate potential challenges in the machine learning pipeline,
ultimately leading to more effective model development and better results.

The dataset contains 48,842 records with 15 features, exhibiting a significant class imbalance (76.07% making ≤$50k vs 23.93% making >$50k). This moderate-sized dataset provides sufficient data for model training with a manageable feature space, though balancing techniques may be needed to address the 3:1 class ratio.

4.1. Histograms

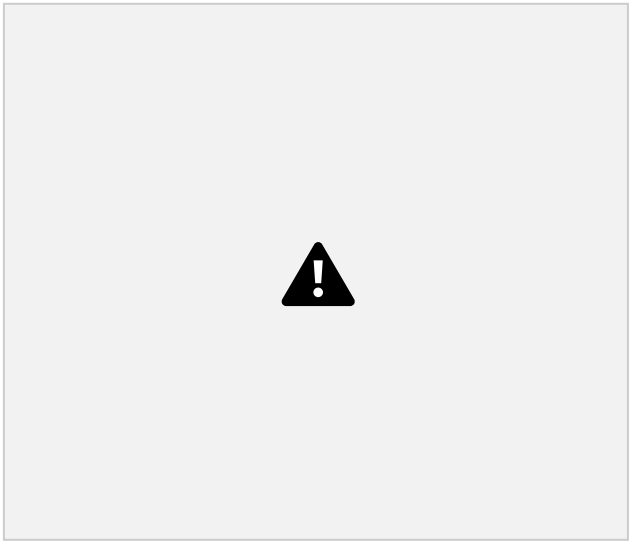

Figure 2. Histogram of numerical features

In Figure 2., the numerical features show distinct dis tributions: age follows a right-skewed normal distribution (20-90 years); education-num shows multiple peaks at 9, 10, and 13 years; hours-per-week has a strong peak at 40 with secondary peaks around 20 and 60. Financial features (fnlwgt, capital-gain, capital-loss) display highly skewed distributions with many zero values, suggesting potential need for logarithmic transformation. The working popula tion

predominantly clusters around 35-45 years of age, with standard 40-hour work weeks being most common.

4.2. Box and Violin Plots

Key Feature Relationships with Income

Strong Predictors

- Education-num: Higher education strongly correlates with income >$50K
- Capital-gain: Significantly higher values in >$50K group
- Age: Higher income group tends to be older

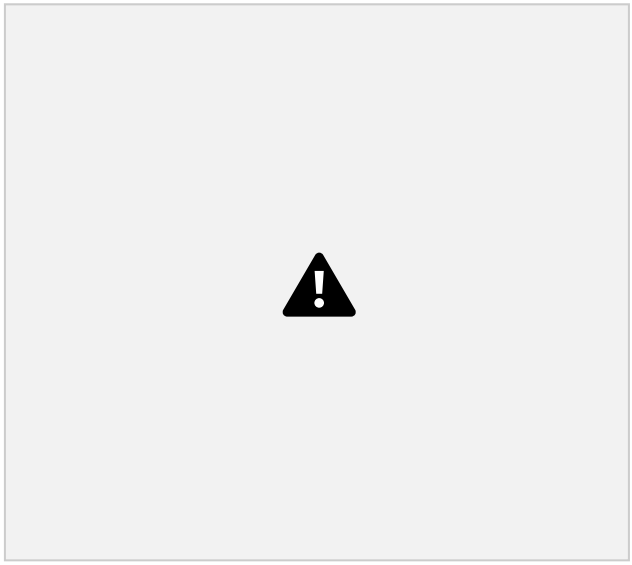

Figure 3. Box and Violin Plots w.r.t Income

- Hours-per-week: More hours worked correlates with higher income
- Marital-status & Relationship: Certain categories (e.g., married) more prevalent in higher income

Weak/Neutral Predictors

- Fnlwgt: No clear relationship with income •

Race: Minimal disparity between income groups •

Native-country: No strong distinguishing patterns •

Workclass: Slight variation but not significant

4.3. Pairplot

- Strong correlations are visible between several numer ical features:
  - Age shows positive correlation with capital-gain and hours-per-week
  - Education-num positively correlates with both income level and hours-per-week
  - Capital-gain has strong relationship with income, with higher gains clustering in the ¿50K group
- Categorical features show distinct clustering patterns:
  - Income classes show clear separation across education-num and hours-per-week
  - Workclass and occupation categories exhibit dis tinct patterns with income levels

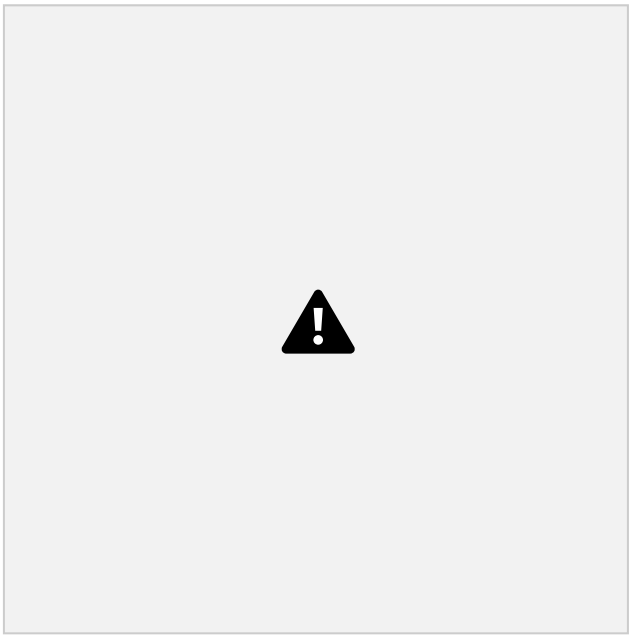

Figure 4. Pairplot of Features colored by Income

  - Marital status shows clustered patterns with age and income
- Distribution characteristics:
  - Capital-gain and capital-loss show highly skewed distributions with many zero values
  - Hours-per-week shows a normal-like distribution with some outliers
  - Age distribution is relatively uniform across working years (20-70)

5. Methodology

5.1. Train-Test Split

We split the data into 80-20 ratio, with 80% for train ing set and 20% for testing. Splitting the data is essential to evaluate a machine learning model's performance. The training set is used to train the model, while the testing set acts as unseen data to measure its predictive capability and generalization. This split helps avoid overfitting and en sures the model is robust when facing new data.

5.2. Model Selection

We implemented several models to assess their perfor mance:

- Logistic Regression: Chosen for its simplicity and ability to provide probabilistic outputs.
- Decision Trees: Selected for their ability to model non-linear relationships and for easy interpretability.
- Random Forest: A powerful ensemble learning method that reduces overfitting by combining multiple decision trees.

5.3. Model Comparison

- Logistic Regression: This model shows stable train ing and validation scores, suggesting a consistent per formance with no significant overfitting. It indicates that the model is underfitting slightly, with similar scores between the training and validation sets, imply ing it may not capture complex patterns in the data.
- Decision Trees: The Decision Tree model displays significant overfitting. The training accuracy is nearly perfect, but the validation accuracy lags behind, indi cating the model is highly tuned to the training data and struggles to generalize.
- Random Forest: This ensemble method, while also showing signs of overfitting, maintains a better valida tion score compared to the Decision Tree. The valida tion accuracy and F1-score suggest it performs better on unseen data, balancing complexity and generaliz ability.

Based on the analysis, the dataset exhibits class imbal ance, with individuals earning more than $50k represent ing 75% of the data. To address this imbalance and im prove model performance, oversampling techniques were employed.

5.4. Model Performance after Random Oversam pling

Random Over Sampling ensures a diverse and represen tative selection from the dataset, enhancing the Random Forest's ability to generalize. This combination helps re duce overfitting and allows the model to capture patterns from both dominant and minority classes, improving pre diction reliability on unseen data.

The following table summarizes the performance metrics for each model:

- Logistic Regression:
  - *Accuracy:* Train = 0.767 — Test = 0.767
  - *F1-score:* Train = 0.766 — Test = 0.766

  Logistic Regression maintains similar training and testing scores, indicating consistent performance.
- Decision Tree:
  - *Accuracy:* Train = 0.999 — Test = 0.922
  - *F1-score:* Train = 0.999 — Test = 0.927

  Despite oversampling, the Decision Tree continues to show overfitting, with a perfect training score but a lower test score.
- Random Forest:
  - *Accuracy:* Train = 0.999 — Test = 0.937
  - *F1-score:* Train = 0.999 — Test = 0.940

  The Random Forest model performs the best after oversampling, with high training and improved testing accuracy.

6. Results and Analysis

6.1. Confusion Matrix and ROC Curve

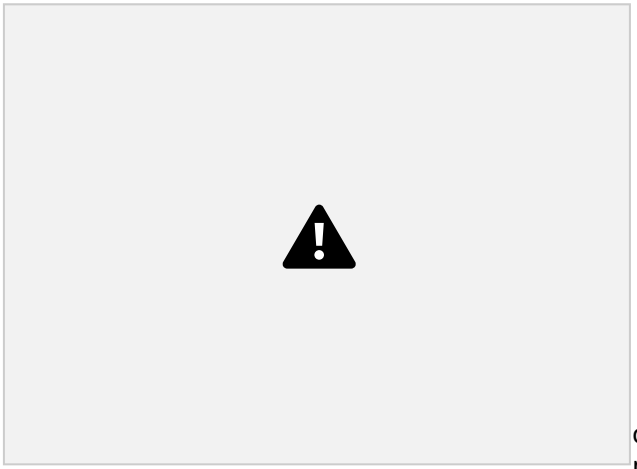

Figure 5. AUC-ROC Curve of all models used

We evaluated model performance based on several key metrics:

- Accuracy: This represents the overall correctness of the model. Random Forest achieved the highest accu racy at 93%, outperforming both Logistic Regression and Decision Trees.
- Precision and Recall: These metrics focus on the model's ability to correctly identify high-income in dividuals. Random Forest again performed best, strik ing a good balance between precision (identifying true positives) and recall (minimizing false negatives).
- F1-Score: A combination of precision and recall, providing a more balanced measure of model perfor mance. Random Forest had the highest F1-Score.
- AUC-ROC Curve: This metric assessed how well the model distinguishes between income levels. Random Forest achieved the highest AUC, indicating strong predictive capabilities across varying thresholds.

6.2. Bias and Variance

Figure 6. Bias and Variance Graphs

Comparative Analysis (of Figure 6.):

- Random Forest shows the best overall performance
- Decision Tree and Random Forest both suffer from overfitting but show promising validation improve ments
- Logistic Regression shows underfitting, suggesting it might be too simple for this problem
- More training data appears beneficial for tree-based models but doesn't help Logistic Regression much
- The learning curves suggest tree-based models (espe cially Random Forest) are more suitable for this par ticular problem

In summary, Random Forest outperformed the other models in almost every metric, making it the most promis ing candidate for further fine-tuning.

7. Conclusion

The primary outcome of our project so far is that Ran dom Forest provides the best performance for predicting in come levels. This model shows the highest accuracy, pre cision, recall, and F1-Score among the models tested. Ad ditionally, its robustness to overfitting makes it suitable for further exploration.

7.1. Future Work

The next steps involve:

- Feature Selection: Analyzing the importance of each feature and dropping those that do not contribute sig nificantly to the model's performance.
- Hyperparameter Tuning: Fine-tuning the parameters of Random Forest and other models to improve accu racy.
- Neural Networks and SVMs: Exploring more com plex models that may offer better accuracy at the cost of interpretability.
- Cross-Validation: Implementing k-fold cross validation to ensure that our results generalize well across different subsets of the data.

7.2. Individual Contributions

- Sameer Singh Godara: Visualization, report writing, and presentation preparation.
- Sanyam Barwar: Preprocessing, Exploratory Data Analysis (EDA), project management and finalization.
- Sanyam Garg: Data collection, feature engineering, and model comparison.

- Vivan Rangra: Model selection, initial training, and evaluation of traditional ML models.

8. References

References

[1] Ronny Kohavi and Barry Becker (1996). *Adult Cen sus Income Dataset*. UCI Machine Learning Repos itory. https : / / archive . ics . uci . edu / dataset/20/census+income

[2] S. Shukla, A. Kumar and H. Kumar (2023). *Bi nary Classification of Adult Census Income Dataset: Analysis and Comparison of Machine Learning and Deep Learning Techniques*. IEEE Interna tional Conference on Computing, Communica tion, and Intelligent Systems (ICCCIS), pp. 535- 540. https : / / ieeexplore . ieee . org / document/10181907

[3] Alghazwi, M., Alqahtani, S., Almutlaq, N. and Al shathri, S. (2022). *Adult Income Prediction using Ma chine Learning Classification Techniques*. Procedia Computer Science, Vol. 207, pp. 2192-2201. https: / / www . sciencedirect . com / science / article/pii/S1877050922021159

[4] AlDhuwayhi, F. (2020). *Census Income Prediction*. GitHub Repository. https : / / github . com / Faisal - AlDhuwayhi / Census - Income - Prediction