Report

Dataset Overview and Target Understanding

The Adult dataset, also known as the "Census Income" dataset, was extracted from the 1994 Census database. It consists of 48,842 instances and 14 features, categorized as a multivariate classification task within the social science subject area. The primary objective is to predict whether an individual's annual income exceeds a threshold of $50,000. The target variable is binary, consisting of the classes `>50K` and `<=50K`.

https://archive.ics.uci.edu/dataset/2/adult

EDA Findings

- **Feature Diversity:** The dataset contains a mix of categorical traits (such as workclass, occupation, and race) and continuous integers (including age, capital-gain, and hours-per-week).
- **Missing Values:** Categorical features like `workclass` and `occupation` contain missing values, often represented by a **'?'** placeholder in the raw data.
- **Class Imbalance:** The data is skewed, with approximately 76% of individuals earning `<=50K` and 24% earning `>50K`.
- **High-Impact Outliers:** Exploratory analysis shows an arbitrary cap of **99,999** in the `capital-gain` column, which is almost exclusively associated with high earners.
- **Redundancy:** There is a direct relationship between `education` (categorical) and `education-num` (numeric), making one of them redundant for mathematical modeling.

Preprocessing Decisions and Reasoning

- **String Cleaning:** We must strip leading whitespace from categorical strings because the raw data includes spaces (e.g., `' Private'`) that prevent accurate matching.
- **Dropping Noise:** The `fnlwgt` (final weight) feature was removed because it is a population estimate rather than an individual trait, often acting as noise that leads to overfitting.
- **Handling Missing Data:** Missing values (marked as **'?'**) were replaced and filled using the mode of the column to maintain dataset size.
- **Encoding:** Categorical features were transformed using One-Hot Encoding to convert non-numeric labels like `occupation` into binary columns that a model can process mathematically.

ML Modeling Foundations

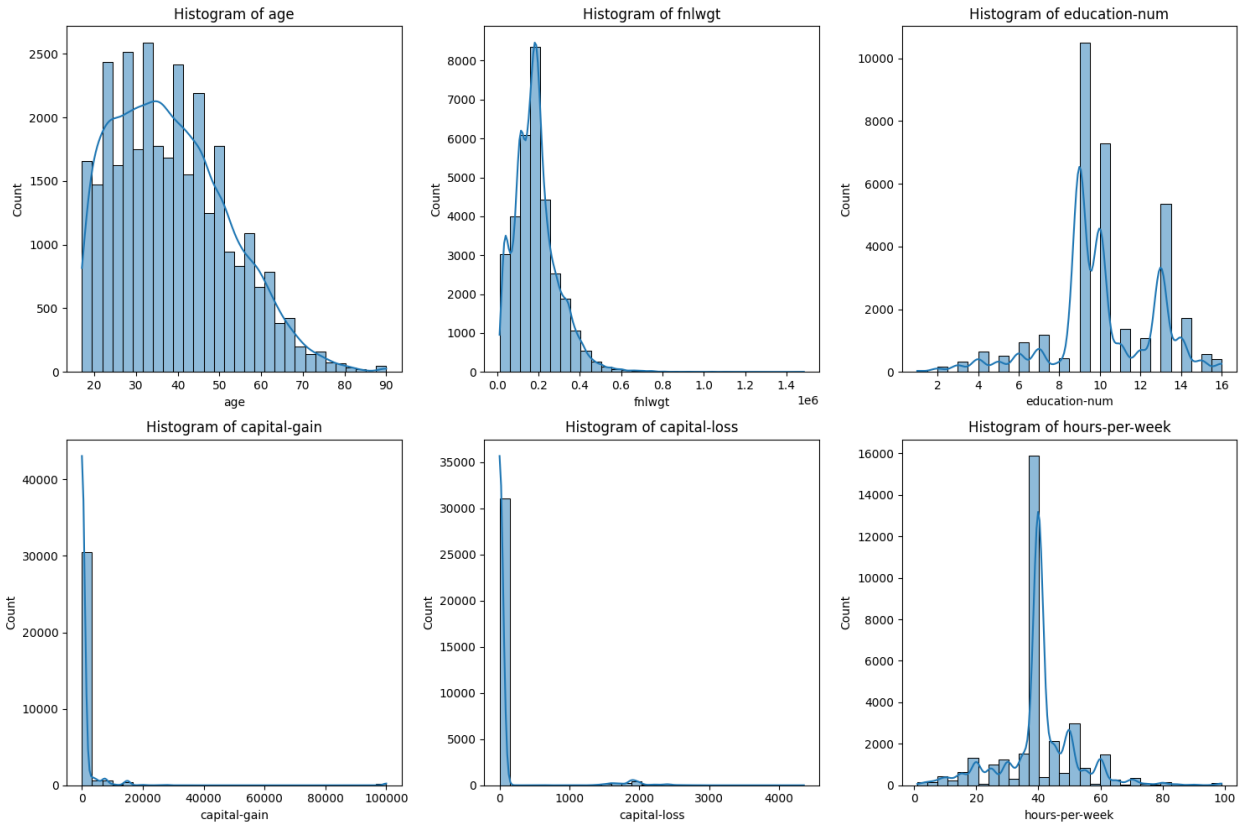
- **Baseline Importance:** A simple baseline, such as a **Majority Class Classifier** (predicting `<=50K` for everyone), establishes a performance floor of ~76% accuracy; any useful model must beat this while maintaining a high F1-score.

- **Overfitting vs. Underfitting:** Underfitting occurs if we use too few features (like only `age`), while overfitting happens if the model "memorizes" training quirks, such as specific values of `fnlwgt`.
- **Model Assumptions:** Logistic Regression assumes a linear relationship between features and income, which is often unrealistic as earning potential typically plateaus with age.
- **Real-World Risks:** Since the data is from **1994**, it suffers from temporal obsolescence; $50,000 is no longer a modern benchmark for high income. Furthermore, using features like `race` or `sex` risks building models that perpetuate historical societal biases.

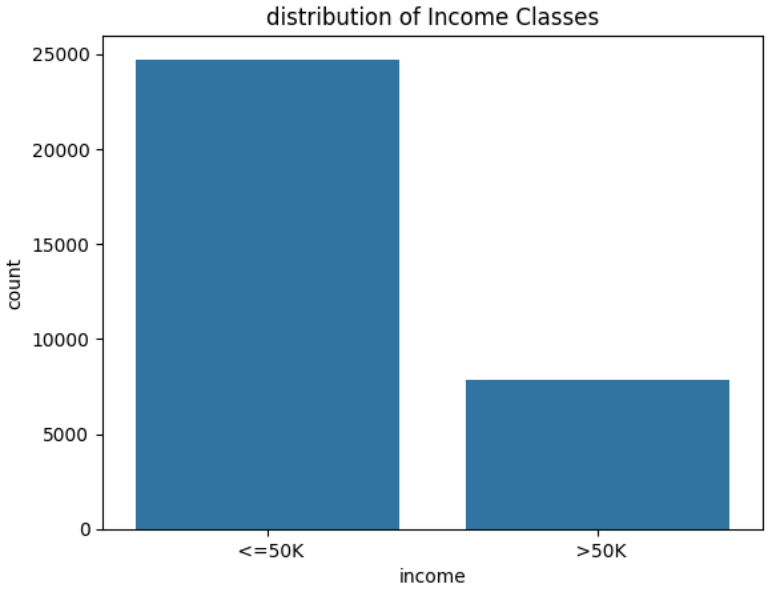
We've successfully improved our Gradient Boosting model's F1-score through threshold adjustment.

Visualization:

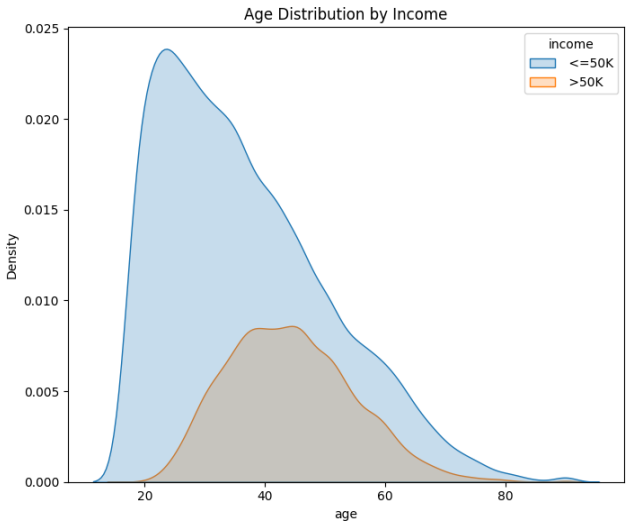
Features Distribution:

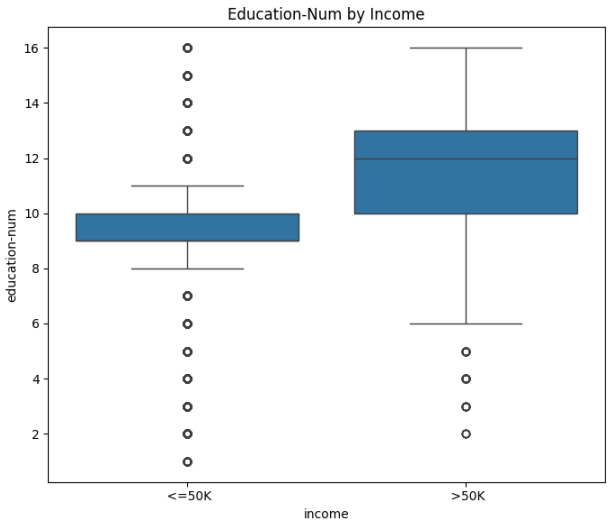


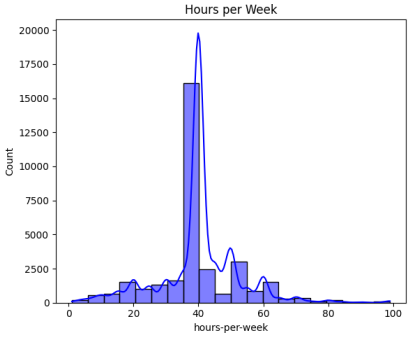
Target Distribution:

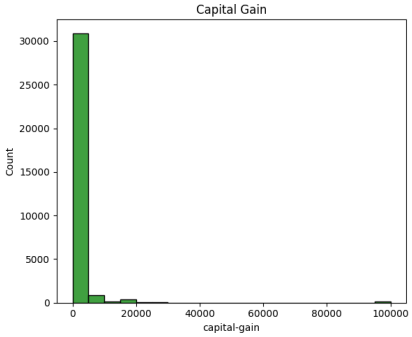


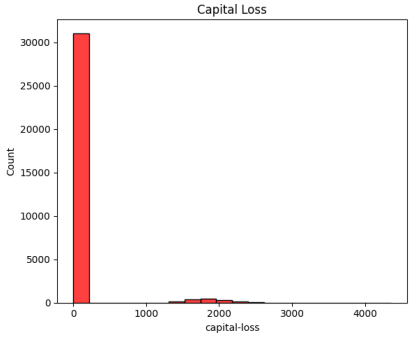
Feature vs. Target comparison:






Hours per Week
Count
hours-per-week
0
2500
5000
7500
10000
12500
15000
17500
20000
0
20
40
60
80
100


Capital Gain
Count
capital-gain
0
5000
10000
15000
20000
25000
30000
0
20000
40000
60000
80000
100000


Capital Loss
Count
capital-loss
0
5000
10000
15000
20000
25000
30000
0
1000
2000
3000
4000