Predicting if income exceeds $50K per year based on US Census Data with Logistic Regression

April 29th, 2024

Yoon, Hae Young (Angela)

Abstract – This project evaluates the predictive performance of a machine learning model on the UCI Census Income dataset, which categorizes individuals into income brackets based on demographic and employment data. We aim to predict whether an individual's income will be greater than $50K annually.

Introduction

The Census Income dataset available at the *UC Irvine Machine Learning Repository*, is a collection of data from the 1994 US Census database. The dataset includes 48,842 instances and 14 attributes.

The primary challenge posed by the dataset is to predict whether an individual earns more than $50,000 a year based on the census attributes. We first explore the data at face value in order to grasp a better understanding of the trends and representations of certain attributes. We then apply this to a model to predict whether an individual made more or less than $50,000 in 1994. Then in the next section, we evaluate its performance. We rigorously evaluate and compare the performance of model(s) by conducting repeated experiments. Finally, we find out what features are of significance, what methods are most effective and also identify the best model for predicting income based on the provided attributes.

Explanatory Analysis

The Dataset – Data Description

The data utilized in this study is the UCI Adult dataset, commonly known as the "Census Income" dataset. The UCI Adult Dataset is composed of 48,842 entries, each with 14 attributes. The dataset is split into a training set of size 32,561 entries and a test set of size 16,281 entries.

The dataset includes a mix of continuous and categorical variables described in the following:

- `age`: the age of an individual
  - Continuous variable, Int greater than 0
- `workclass`: the employment status of an individual
  - Categorical variable; Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
- `fnlwgt`: final weight. The number of people the census takers believe that entry represents
  - Continuous variable, Int greater than 0
- `education`: the highest level of education achieved by the individual
  - Categorical variable; Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
- `education-num`: the highest level of education in numerical form
  - Continuous variable, Int greater than 0
- `marital status`: marital status of the individual

    - Categorical variable; Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
- `occupation`: the individual's occupation
    - Categorical variable; Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
- `relationship`: represents what this individual is relative to others
    - Categorical variable; Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
- `race`: the race of the individual
    - Categorical variable; White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
- `sex`: the biological gender of the individual
    - Categorical variable; Male, Female
- `capital-gain`: capital gains for an individual
    - Continuous variable, Int greater than 0
- `capital-loss`: capital losses for an individual
    - Continuous variable, Int greater than 0
- `hours-per-week`: The number of hours the individual works per week
    - Continuous variable, Int greater than 0
- `native country`: country of origin of the individual
    - Categorical variable; United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands
- `target`: whether or not the individual makes more than $50,000 annually
    - Categorical variable; <=50K, >50K

Missing Data Handling

Basic analysis revealed missing values in key categorical columns – workclass, occupation, and native-country. Given the nature of the missing data, different imputation strategies were employed:

- Workclass - used the mode of the column
- Occupation - used a weighted random sampling based on the existing distribution to maintain the original data structure
- Native Country - used the mode stratified by race to acknowledge potential cultural and demographic correlations

Target Variable

I wanted to start by looking at our target variable first. This is an expected and obvious observation.

- The number of people earning more than $50K annually is one third of the people earning less than it.
- We should also keep in mind that this data was collected back in 1994, so $50K back then valued more than the same amount today.

The target variable in the UCI Adult dataset is significantly imbalanced, with 24,720 individuals (approximately 75.92%) earning "<=50K" annually, representing the majority class, and 7,841

individuals (approximately 24.08%) earning ">50K," constituting the minority class. This distribution suggests a skew towards lower income brackets and indicates the need for models that can handle imbalanced classification tasks effectively.

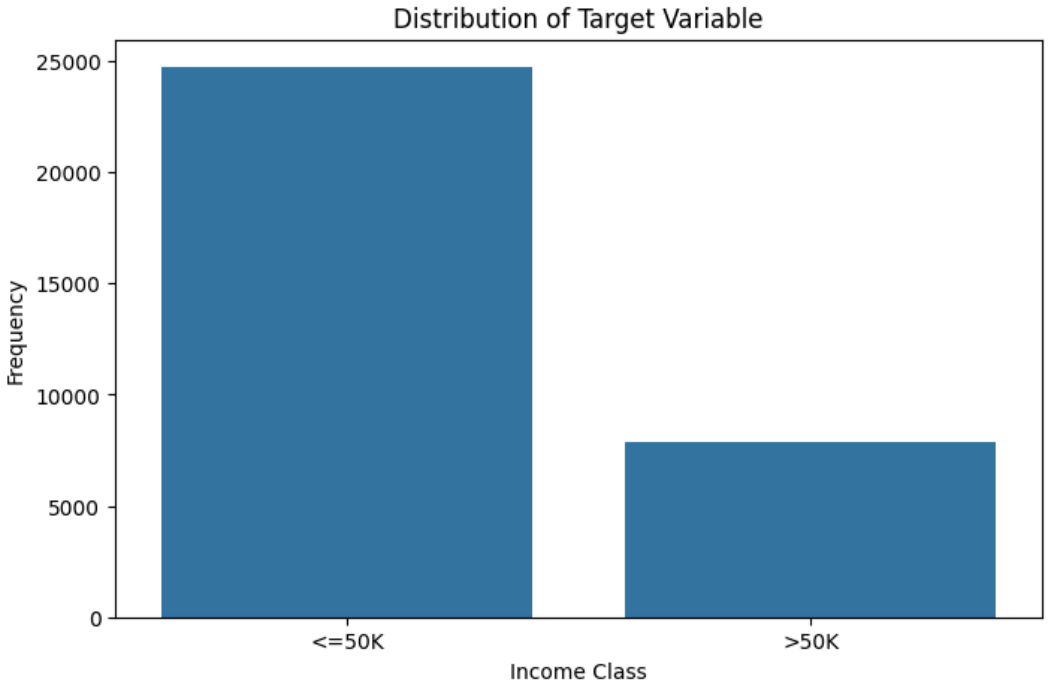

Figure 1. Target Variable Distribution

Categorical Data Analysis

We visualized relationships between various categorical features and the target variable ('income >50K or <=50K'):

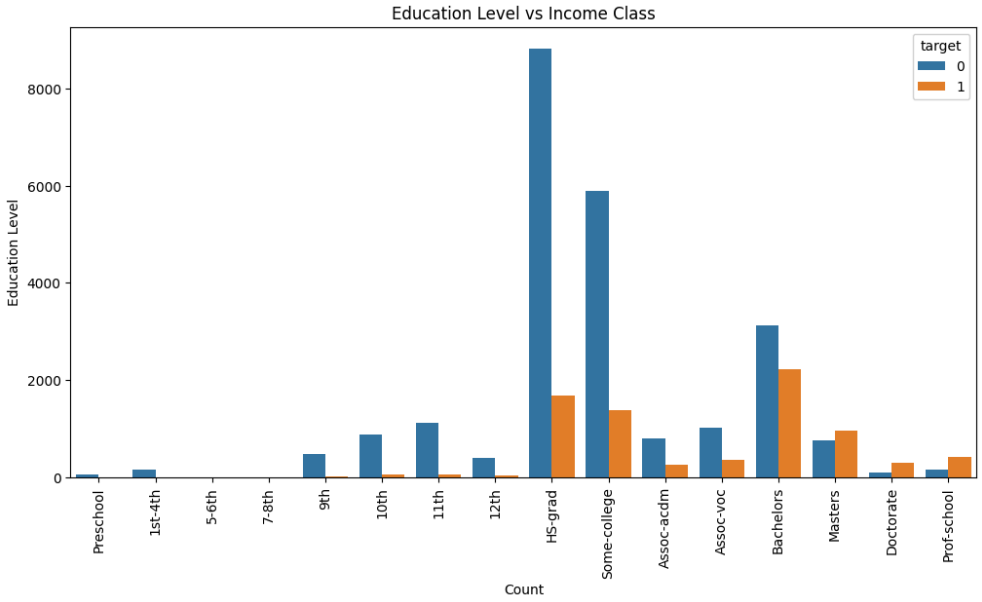

Figure 2: Education vs. Target

- Higher education correlates with higher income: Individuals with advanced degrees (Bachelor's and above) are more frequently represented in the >50K income class.

- The majority of individuals who did not pursue education beyond high school are in the <=50K income class.
- A significant spike in frequency is observed at the 'HS-grad' level, predominantly in the <=50K income class.
- Professional and doctoral-level education show a relatively higher proportion of individuals earning >50K compared to other education levels.

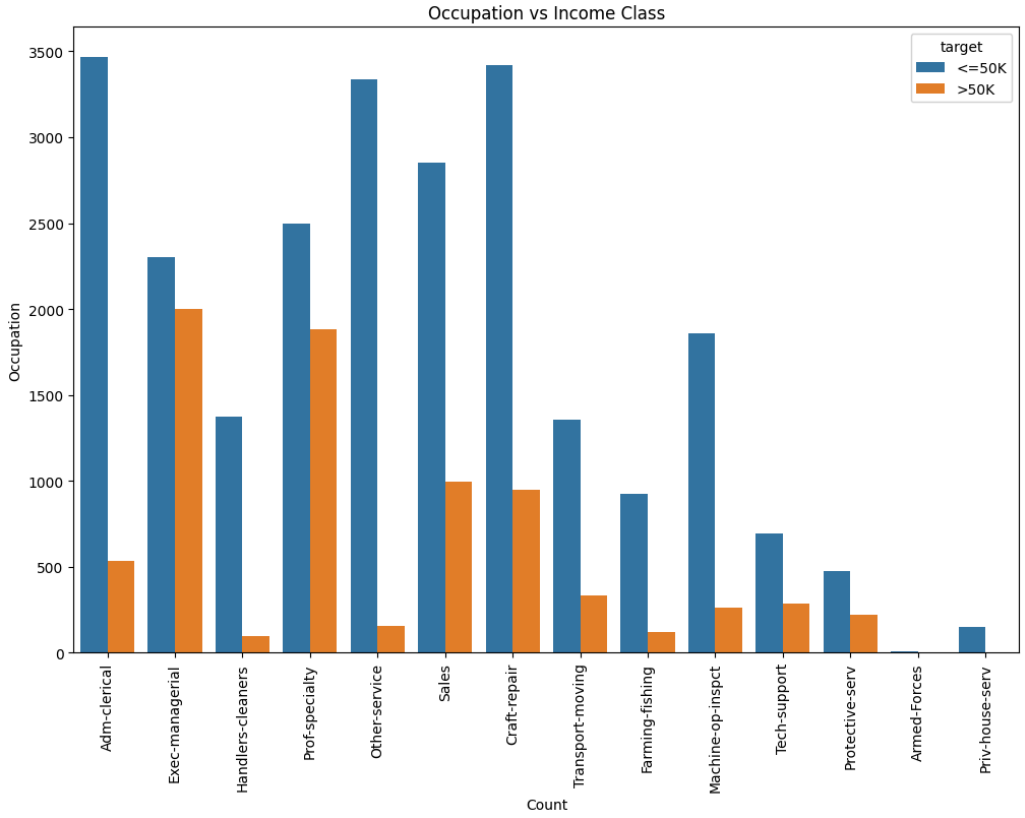

Figure 3: Occupation vs. Target

- Certain occupations such as 'Exec-managerial' and 'Prof-specialty' have a higher proportion of individuals earning >50K compared to others.
- Occupations like 'Handlers-cleaners', 'Farming-fishing', and 'Machine-op-inspct' show a higher count of individuals earning <=50K.
- Technical, protective service, and armed forces occupations have a more balanced distribution between the two income classes, but still lean towards <=50K.
- Overall, there is a visible trend that more specialized and managerial roles tend to be associated with higher income brackets.

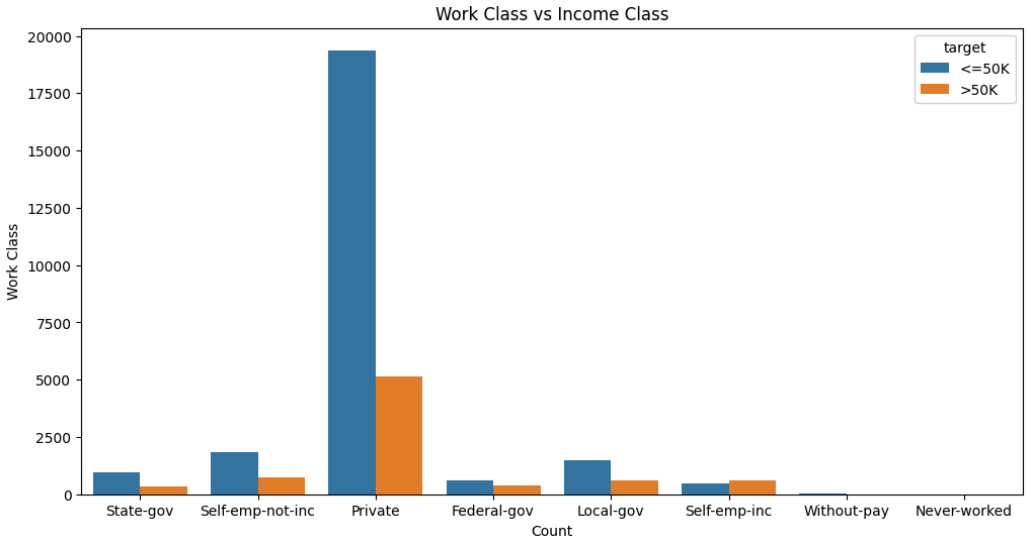

Figure 4: Work class vs. Target

- The 'Private' work class has the highest overall count, with a significantly larger proportion earning <=50K.
- Individuals in the 'Self-emp-inc' work class show a relatively higher likelihood of earning >50K compared to other work classes.
- 'Federal-gov' employees also exhibit a substantial count in the >50K income class, indicating higher earnings in this sector.
- Work classes such as 'State-gov', 'Local-gov', and 'Self-emp-not-inc' predominantly fall within the <=50K income category.
- There are very few or no individuals in the 'Without-pay' and 'Never-worked' categories who earn >50K, as expected.

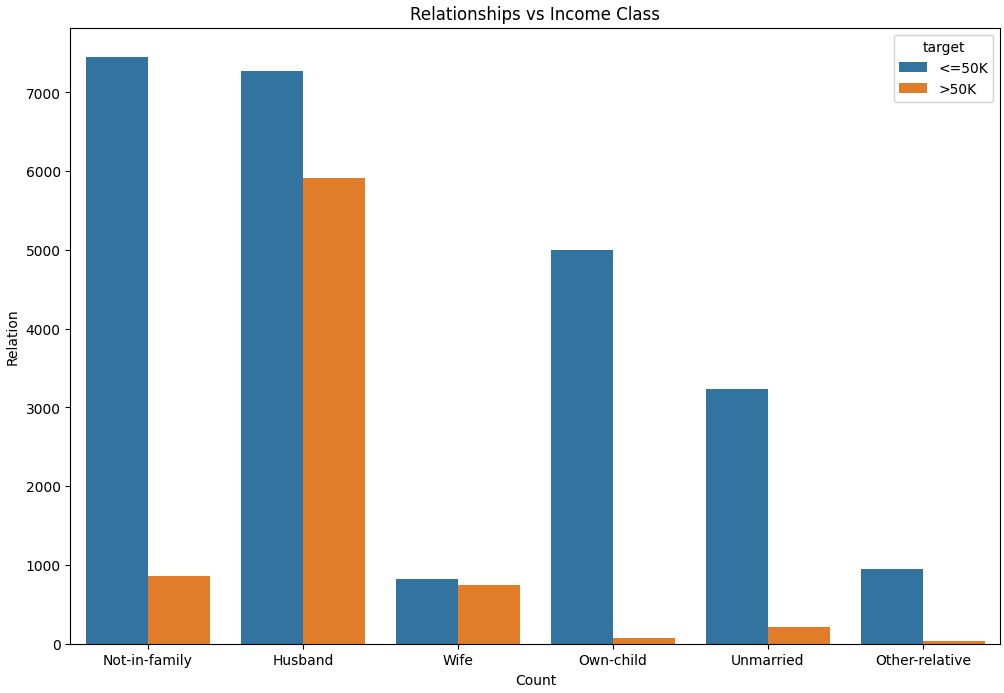

Figure 5: Relationships vs. Target

- "Husbands" represent a substantial portion of individuals earning >50K, indicating a higher income potential for this demographic.
- The "Wife" category also shows a significant number of individuals earning >50K, though less than "Husbands".
- Those categorized as "Not-in-family" and "Unmarried" have a larger presence in the <=50K income class.
- "Own-child" and "Other-relative" categories have the lowest counts in the >50K income class, suggesting these groups are less likely to have higher earnings.
- Across all relationship statuses, there is a noticeable trend where there are more individuals in the <=50K category than in the >50K, with the exception of "Wife" where the distribution is more balanced.

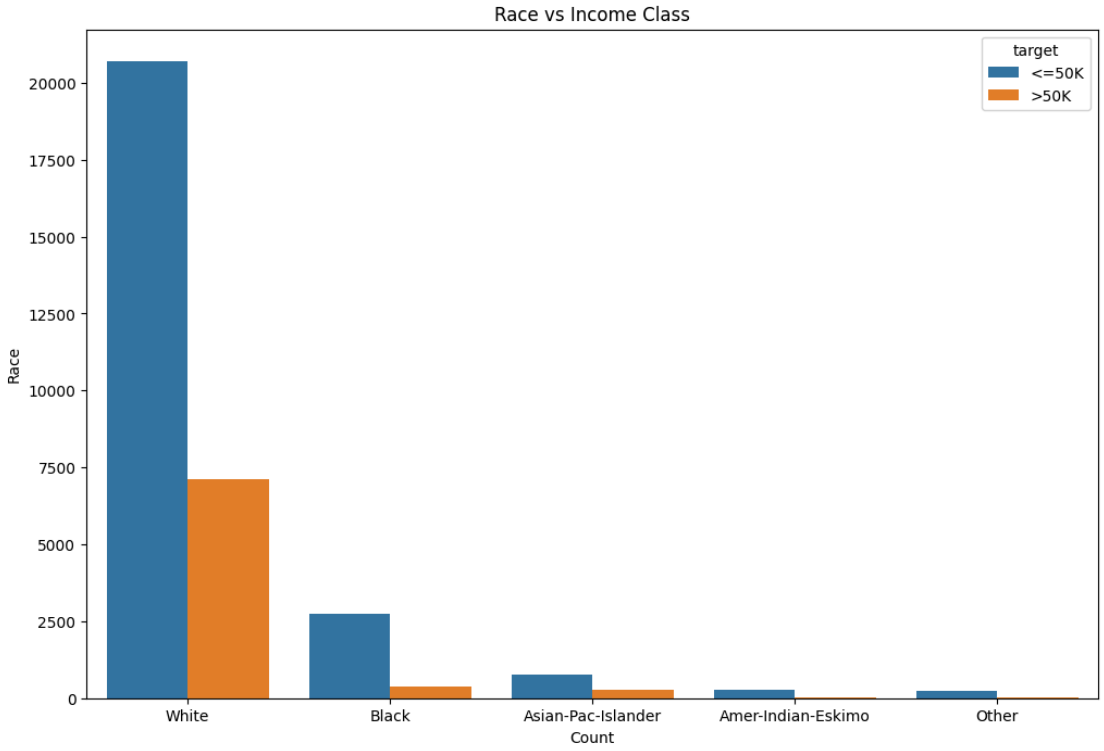

Figure 6: Race vs. Target

- The "White" race category has the highest counts in both income classes, but particularly dominates the <=50K income class.
- The "Black" and "Asian-Pac-Islander" categories show more individuals in the <=50K income class, with a relatively small proportion in the >50K income class.
- "Amer-Indian-Eskimo" and "Other" race categories have the least representation in the dataset, with the majority falling into the <=50K class.
- Overall, there is a visible disparity in income distribution across different races, with the "White" race category having a larger representation in the >50K income class compared to other races.

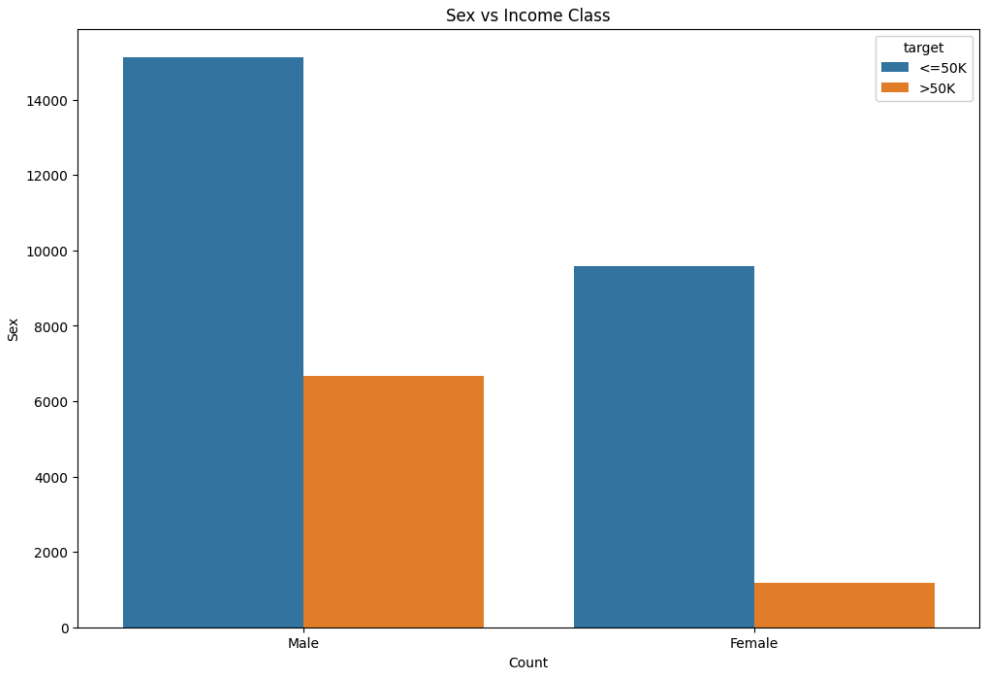

Figure 7: Sex vs. Target

- The "White" race category has the highest counts in both income classes, but particularly dominates the <=50K income class.
- The "Black" and "Asian-Pac-Islander" categories show more individuals in the <=50K income class, with a relatively small proportion in the >50K income class.
- "Amer-Indian-Eskimo" and "Other" race categories have the least representation in the dataset, with the majority falling into the <=50K class.
- Overall, there is a visible disparity in income distribution across different races, with the "White" race category having a larger representation in the >50K income class compared to other races.

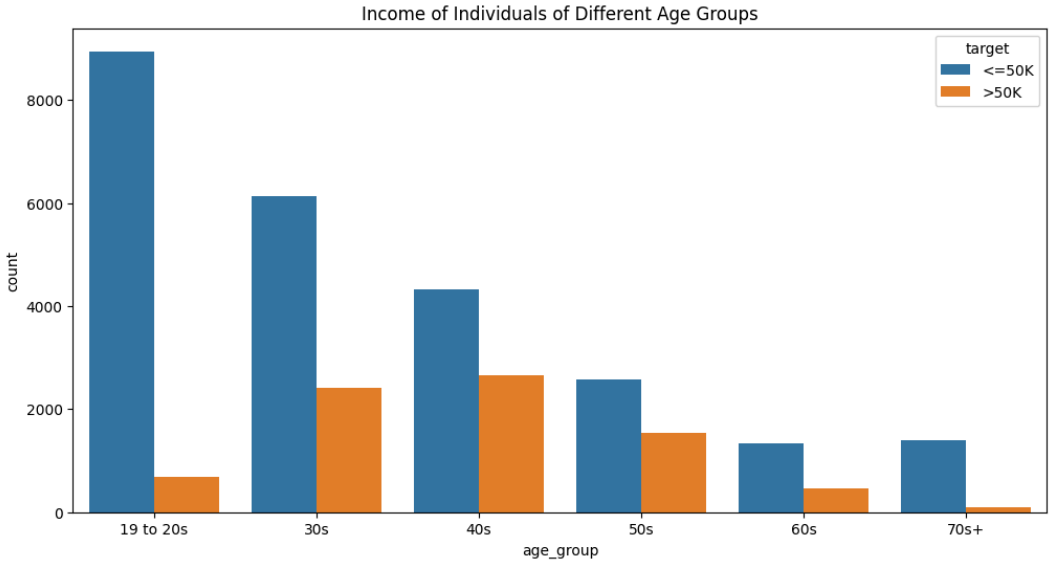

Figure 8. Age vs. Target

- A higher count of younger individuals (19 to 20s) fall into the <=50K income class, with very few exceeding the 50K mark, reflecting lower earnings among younger people.
- As age increases, the proportion of individuals earning >50K also increases, with those in their 40s showing a notable presence in the higher income class.
- The 50s age group still maintains a significant presence in the >50K class but starts to show a decline in count compared to the 40s.
- Individuals in the 60s and 70s+ age groups are predominantly in the <=50K class, which may reflect reduced earnings associated with retirement or decreased participation in the workforce.
- Overall, there is a trend indicating that income tends to increase with age until a certain point, after which it decreases, likely due to retirement.

Numerical Data Analysis

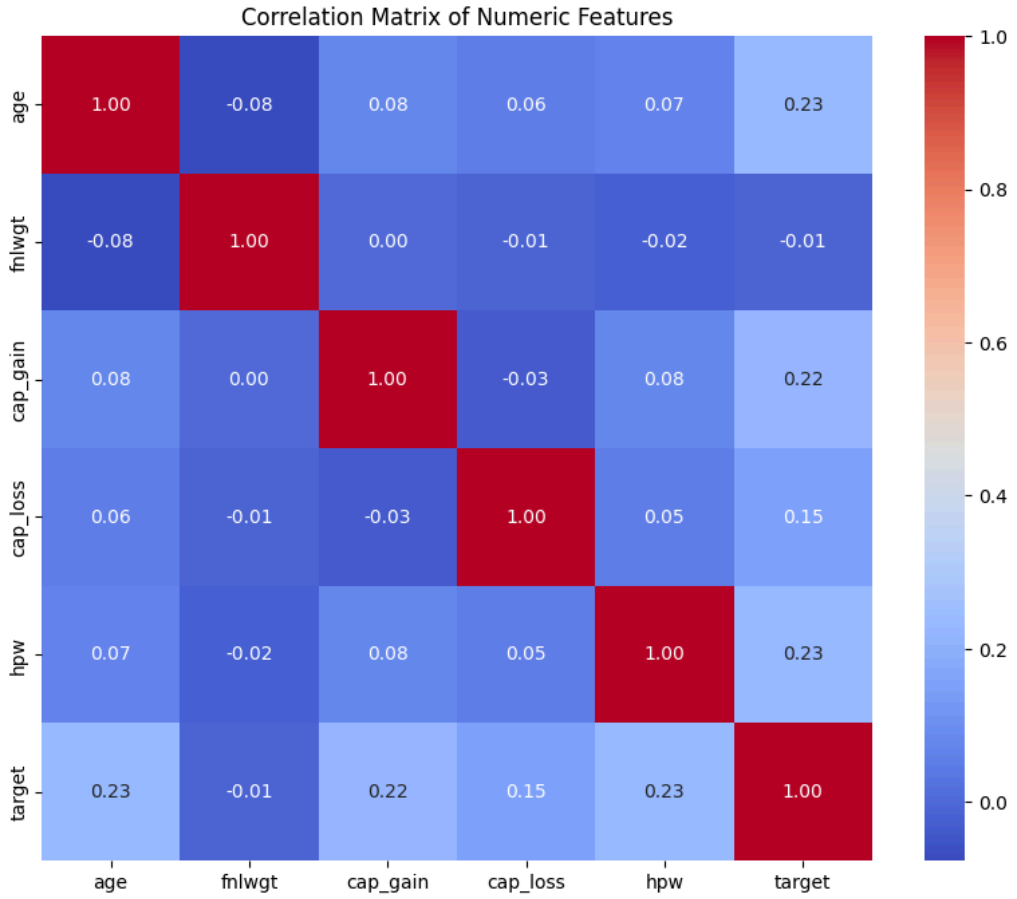

Figure 9: Correlation matrix of numerical data

- Age shows a positive correlation with the target variable, suggesting that as age increases, so does the likelihood of earning >50K.
- Capital gain and capital loss both have notable positive correlations with the target, indicating that higher capital gains and losses are associated with higher income brackets.
- Hours per week (hpw) worked also has a positive correlation with the target, which aligns with the expectation that working more hours could lead to higher income.
- The fnlwgt (final weight) feature shows a very low and slightly negative correlation with the target, implying it has little to no linear relationship with income level.

- There are no strong correlations between the numerical features themselves, suggesting limited multicollinearity concerns within these variables.

Predictive Task

Model Implementation \ Logistic Regression

Here, I'll be using a Logistic Regression Model. Instead of building the model from scratch, I will be utilizing the Scikit-learn package for an automated model. Since this prediction project itself is a binary classification problem, I was considering using one of Naïve Bayes or Logistic Regression models for its straightforwardness and computational efficiency. Logistic Regression was ultimately chosen due to its suitability for binary classification problems. This model estimates the probability that a given input point belongs to a certain category, which aligns perfectly with the dichotomous nature of our target variable.

Unlike how Naïve Bayes classifiers make a strong assumption about the independence of the features given the class label, which is rarely true in practice, Logistic Regression can handle situations where features may have some degree of correlation.

I set out to fine-tune the hyper parameters of the Logistic Regression model to predict whether individuals earn more than $50,000 per year, using the UCI Adult dataset. To accomplish this, I utlized a k-fold cross validation approach alongside GridSearchCV, which is a systematic way of browsing through multiple combinations of parameter tunes, cross-validating as it goes to determine which hyper-parameter combination gives the best performance.

In the hyperparameter tuning phase of my Logistic Regression model, I applied k-fold cross-validation using GridSearchCV to identify the most effective combination of hyperparameters. The search across the predefined grid resulted in the selection of **C=0.1** with an **L2 penalty** as the optimal parameters.

- C=0.1 indicates that a slight regularization if beneficial for the model we're using
- L2 Regularization is also effective because it shrinks the coefficients towards zero but does not force them to be exactly zero

Model Performance

In the Model Performance section of my report, I will summarize the effectiveness of the Logistic Regression model that was trained and evaluated on the UCI Adult dataset.

Metrics

	Accuracy	Precision	Recall	F1 Score
LogisticRegression with C=0.1	0.8043	0.6807	0.3258	0.4409

These results were consistent across an additional 3 separate runs, as evidenced by the negligible std values for each metric (close to 0). The lack of variability suggests that the model is **consistent** and **stable**, providing **reliable predictions** regardless of different subsets of data used in training due to random initialization.

Takeaways:

- **Accuracy** – while the model shows a strong ability to accurately predict the lower-income class, its performance in correctly identifying the higher-income class is moderate.
- **Lower recall value** – indicates that there is room for improvement, particularly in terms of minimizing false negatives.Potential strategies include trying alternative classification algorithms that might deal better with the class imbalance
- **Lower F1 Score** – score is closer to the precision, reflecting that the precision has a stronger influence in this case due to the class imbalance in the dataset

Roc Curve and AUC

I also thought the ROC was essential for evaluating the performance of this classification model, especially because our dataset was imbalanced as well.

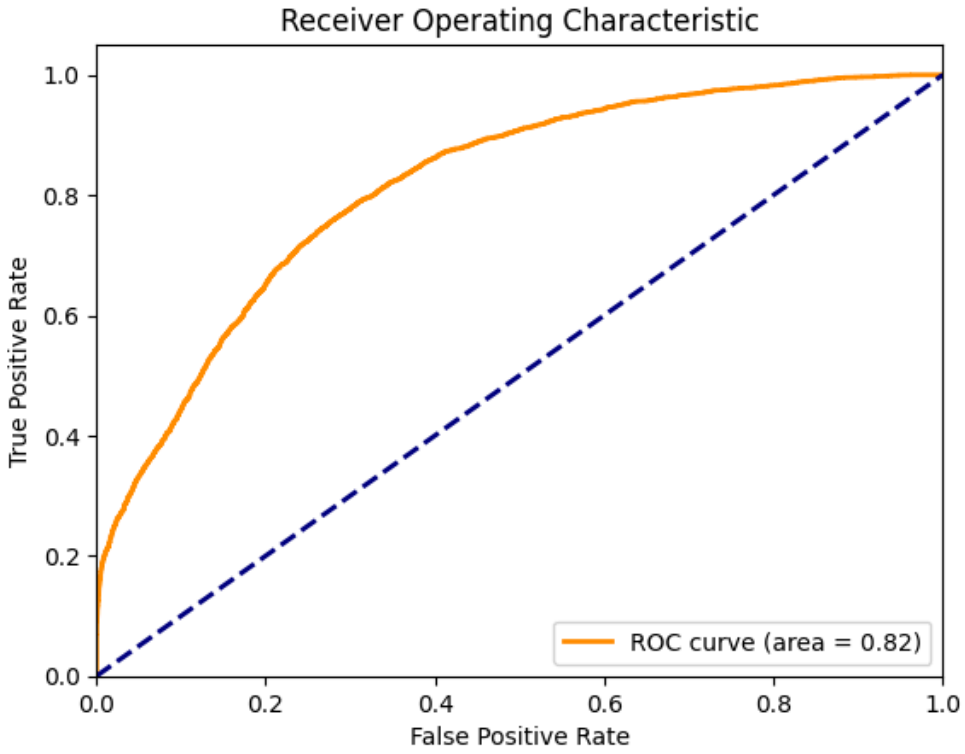

Figure 9: ROC Curve

In our model's case, the ROC curve shows a good balance between the true positive rate and the false positive rate, which is indicated by the curve being closer to the top-left corner of the plot. An AUC value of 0.82 means that there is an 82% chance that the model will be able to tell between a randomly chosen positive instance and a randomly chosen negative instance.

The fact that the ROC curve of our model lies significantly above the diagonal line (which represents random chance) indicates that the model has learned to classify salaries effectively. Though, there is still room for improvement, and might benefit from exploring further model tuning or using alternative classifiers to improve this metric further.

Results

I wanted to go a little beyond to see if other models were actually able to improve these metrics. I experimented with Naïve Bayes and k-Nearest Neighbors (kNN) in addition to Logistic Regression to establish a comparative analysis of their performance. Here's a summary of the outcomes for each model using key classification metrics:

Naïve Bayes

Accuracy	Precision	Recall	F1 Score
0.7993	0.6625	0.3063	0.4189

kNN

- After applying a grid search to determine the best parameters, the kNN model's best configuration was found to be with 10 neighbors, using the Minkowski distance metric and distance weighting.
- The best cross-validation accuracy achieved was 82.68%, and the test accuracy mirrored this at 82.68%, indicating the model's stability and generalization capability.

Model	Accuracy
kNN	0.822
Naïve Bayes	0.799
Logistic Regression	0.804
SVM	n/a (took too long to run)

The kNN model outperformed Naïve Bayes and our original Logistic Regression model, showing that it could be a more robust classifier for this particular imbalanced dataset. Though, I was surprised to see that the accuracy of the Naïve Bayes model performed worse than the Logistic Regression model.

It's also essential to consider the trade-off between model complexity and interpretability. While kNN had higher accuracy, it is a more computationally intensive model (took longer to run), especially as the size of the dataset grows. Conversely, Logistic Regression and Naïve Bayes models offer simplicity and speed, which can be advantageous for very large datasets or real-time predictions.