Adult Census Income Project

HarvardX PH125.9x - Data Science: Capstone - Choose Your Own Project

Joe Li Man Hon

December 7, 2025

1. Introduction

Predictive analysis plays a central role in machine learning and data science by turning historical data into actionable forecasts about future events and behaviors. It combines statistical modeling and machine learning algorithms to uncover patterns that support better decision-making in areas such as risk scoring, customer targeting, and demand forecasting. In practice, predictive models help organizations move from reactive reporting to proactive strategy, allowing them to anticipate outcomes and optimize interventions rather than simply describing what has already happened. This focus on foresight makes predictive analysis a key driver of value in data-driven projects across industries.

This project exemplifies binary classification in machine learning by using the Adult Census Income dataset (hereinafter referred to as "the adult dataset") to predict whether an individual's annual income exceeds $50,000 beased on demographic like age, education, occupation and weekly work hours. The data was extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics) [1].

This report begins with exploratory data analysis utilizing common visualization methods, followed by descriptions on the development, training, and testing phases of the predictive algorithm. It then presents results from each model iteration and concludes with an evaluation of the overview of models' accuracy, its limitations, and potential areas for further improvement.

In this project, the code was run, and the report was compiled using R Markdown in RStudio Version 2025.09.2+418.

1.1 Evaluation of Algorithm

This project divides the dataset into separate training and testing sets to ensure a robust evaluation of algorithm. The models will be trained on the training set and then evaluated by comparing its predictions to the actual outcomes in the testing set. Accuracy, defined as the proportion of correctly predicted instances, is used as the primary metric to assess the performance of the algorithm. This approach helps to gauge how well the model generalizes to unseen data and identifies potential areas for optimization.

1.2 Preprocessing and Partitioning

1.2.1 Processing Question Marks

The adult dataset consists of 4,262 question marks ("?") representing missing values across various features, for example `occupation`. To streamline downstream processing and enable standard R functions, these question marks are systematically replaced by `NA`. This preprocessing step ensures data integrity for exploratory analysis while facilitating the subsequent partitioning of the cleaned dataset into training and testing sets.

1.2.2 Near Zero Variance and Dimension Reduction

During the preprocessing stage, near zero variance (NZV) predictors, that is, those with minimal unique values or skewed frequencies, are identified and removed to prevent model instability. The $nearZeroVar$ function identifies three variables: `capital.gain`, `captial.loss` and `native.country`. Among these variables, `capital.gain` has **91.7%** zeros, `capital.loss` **95.3%** zeros and `native.county` **89.6%** the United-States. Figure 1 below plots the disparity of these variables.

1.2.3 Partitioning

The adult dataset is partitioned into training and testing set. The training set (train_set) contains 80% (26,048 observations) of the adult data, while the testing set (test_set) 20% (6,513 observations).

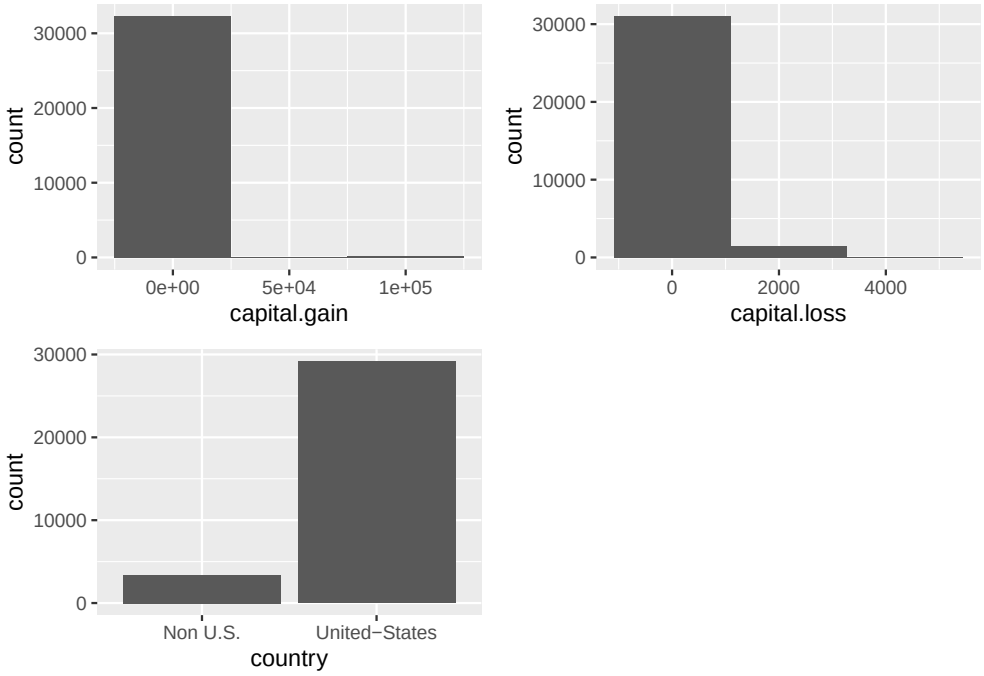


Figure 1: Disparity in the Three Near Zero Variables

2. Exploratory Data Analysis

2.1 Overview on the dataset

The `train_set` data frame comprises of 26,048 rows and 12 columns, derived from an 80/20 partition of the adult dataset. The target variable shows class imbalance, with **24.1%** of observations having annual income exceeding $50,000 and the remaining **75.9%** below that threshold.

2.2 Plots about the dataset

Further exploration reveals the distribution of `income >50K` versus `<=50K` across key categorical features, highlighting their predictive importance. Figures below show stacked bar plots created with `ggplot2::geom_bar(position = "fill")` for some of the features such as `sex`, `race`, `education`, `occupation`, `workclass` and `relationship`. These visualizations demonstrate stark disparities—for instance, males and certain education levels like "Bachelors" or higher show significantly higher proportions above the $50K threshold—informing feature selection and model interpretation.

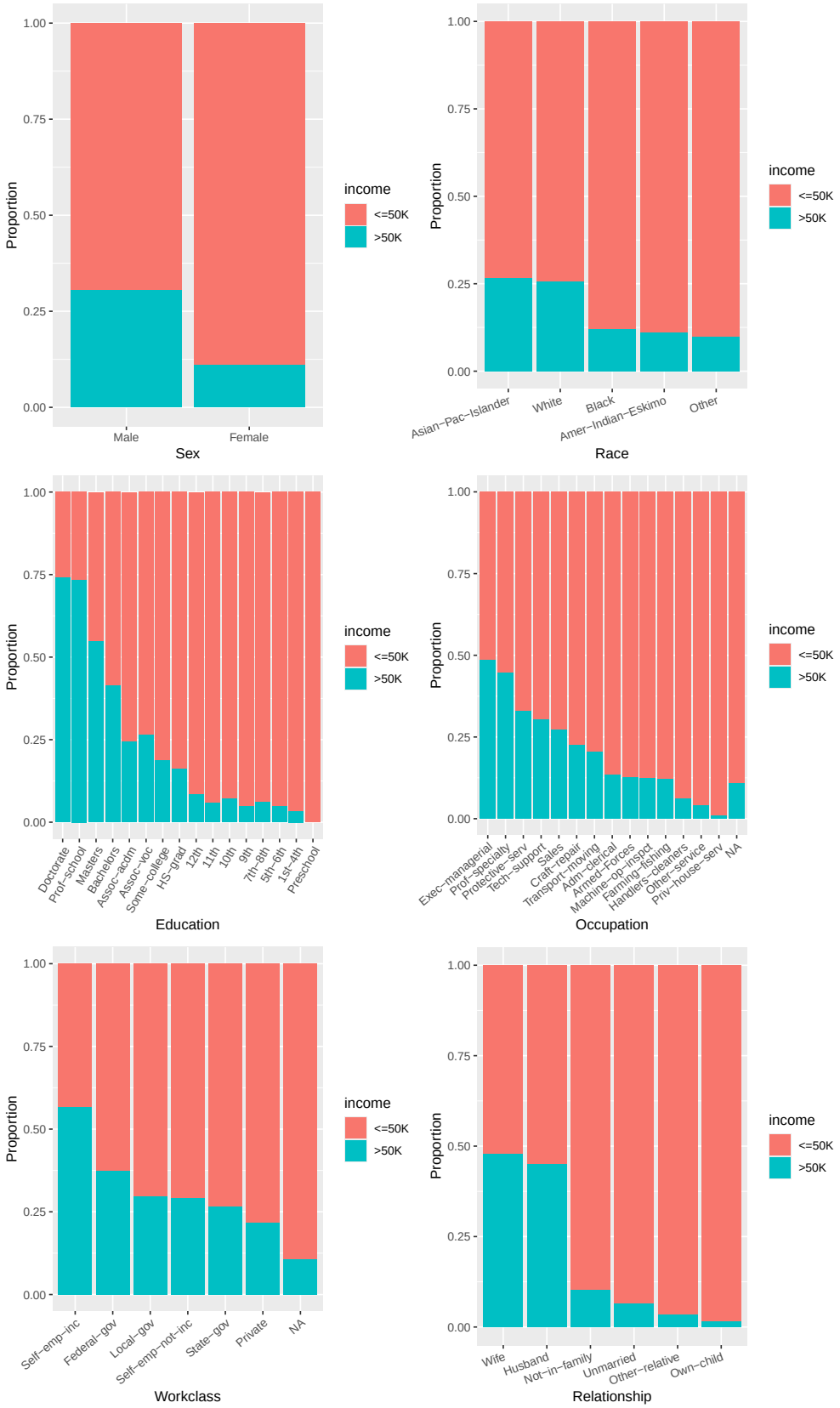


Figure 2: Distribution of Income

3. Methods

3.1 Factor Encoding Categorical Variables

The dataset contains many categorical data that the data frame itself considers them as characters. Categorical characters in the dataset are converted to factors using `as.factor()` to establish meaningful levels, then encoded to numeric via `as.numeric()` for machine learning compatibility. This label encoding preserves category order while creating integer representations suitable for models like kNN.

3.2 Random Guess

The simplest prediction method randomly assigns predicted values to each observation without using any features. This random guess establishes a minimum performance threshold that all models must exceed to demonstrate predictive value beyond chance.

Method	Accuracy
Random Guess	0.4957777

3.3 Generalized Linear (GLM) Model

Before training mathematical models like GLM, missing values must be addressed. Below reveals `NA`s in the columns `workclass` and `occupation` after replacing "?" values:

```
##            age      workclass         fnlwgt      education  education.num
##              0           1488              0              0              0
## marital.status     occupation   relationship           race            sex
##              0           1494              0              0              0
## hours.per.week         income
##              0              0
```

To enable immediate model training, these columns are temporarily excluded from the feature set.

Method	Accuracy
GLM	0.7664671

3.4 LOESS Model

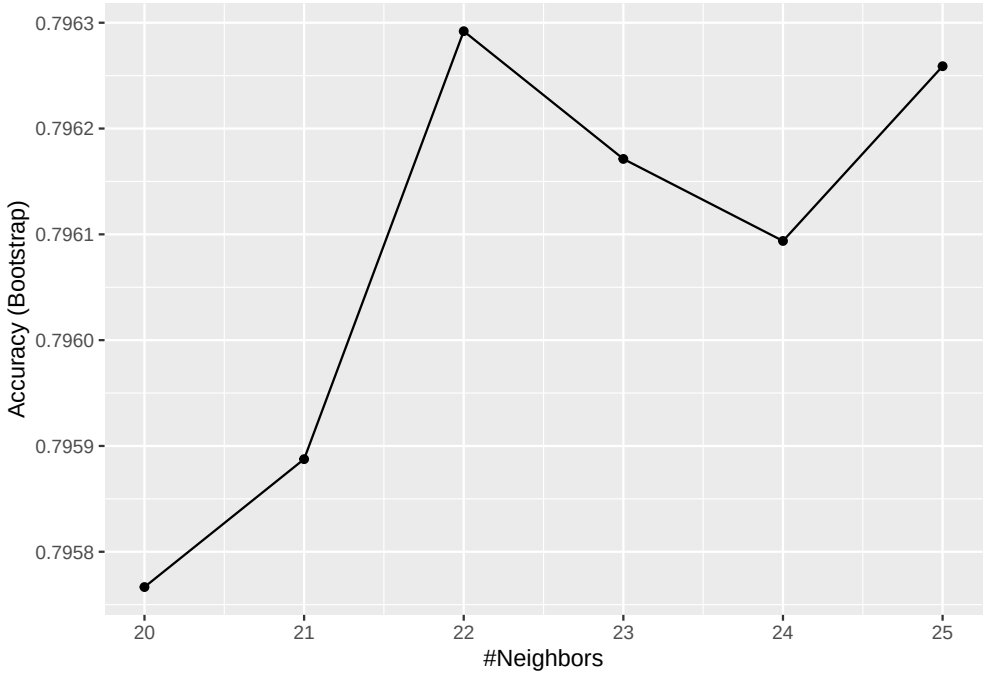
Next, a LOESS (Local weighted regression) model is trained on the training set.

Method	Accuracy
LOESS	0.7899585

3.5 k-Nearest Neighbors (kNN) Model

During tuning of the optimal k value, the accuracy curve exhibited a progressively increasing convex shape with k values rising up to 180 without reaching a clear peak (the lengthy computation omitted for brevity). This behavior stems from the extreme numeric range of the `fnlwgt` feature (10,000 to >1,000,000), which dwarfs other variables like `age` (16–90), dominating Euclidean distance calculations in kNN. Removing `fnlwgt` enabled efficient convergence on an optimal k. Notably, accuracy improved to around 0.796, versus 0.74–0.75 when including it, confirming that excluding `fnlwgt` enhances kNN performance on this dataset.

```
##    k
## 3 22
```



Method	Accuracy
kNN	0.7947183

3.6 Random Forest

Next, a Random Forest model is trained with `mtry = 2` (number of randomly selected predictors per tree split).

Method	Accuracy
Random Forest	0.8200522

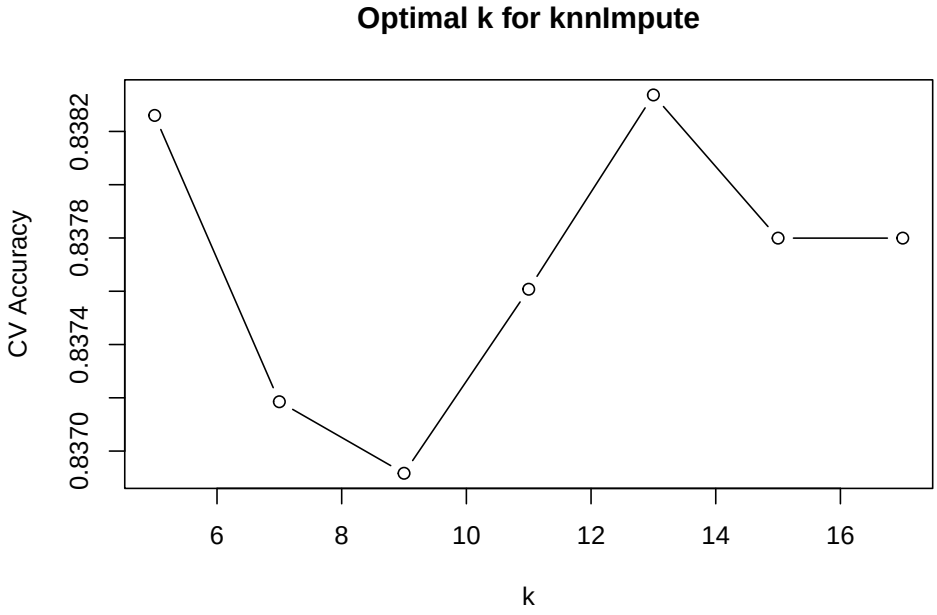
3.7 Imputation and Random Forest with Imputation

Recall that in section 3.3, `workclass` and `occupation` were omitted due to `NA` values. To handle missing data, here introduces data imputation. Data imputation is the clever process of filling in missing information in

datasets so that the analysis stays accurate [2]. For categorical data like the adult dataset, usual techniques are mode imputation and kNN imputation [3]. However, exploring `occupation` shows the distribution of outcomes are very different from the most frequent `Exec-managerial` to `NA`; mode imputation distorts outcome distributions.

Occupation	Income	Count	Percentage
Exec-managerial	<=50K	1658	51.38
Exec-managerial	>50K	1569	48.62
Missing (NA)	<=50K	1334	89.29
Missing (NA)	>50K	160	10.71

Therefore, kNN imputation is performed, using data from training set and then apply the result to both training and testing set, to correctly apply statistical missing data imputation and avoid data leakage [4]. Cross validation tunes the imputation k value.



Among the method used, the Random Forest has the highest accuracy, so the imputed training set will be trained with Random Forest.

Method	Accuracy
Random Forest with Imputation	0.8426224

4. Results

Below table summarizes the accuracies resulting from the methods described in section 3 above:

Method	Accuracy
Random Guess	0.4957777
GLM	0.7664671
LOESS	0.7899585
kNN	0.7947183
Random Forest	0.8200522
Random Forest with Imputation	0.8426224

The final algorithm achieved accuracy at **0.84262**.

5. Conclusion

This project applies multiple machine learning techniques to predict income levels (`>50K` vs. `<=50K`) using the Adult Census Income dataset. Preprocessing identifies and removes near-zero variance features and partitions data into 80/20 training and testing sets. Exploratory analysis reveals key disparities across demographic features, visualized through stacked bar plots. Various models are evaluated: random guess baseline, GLM, LOESS, kNN, and Random Forest achieving highest accuracy.

Regarding future improvement, the code in this project takes time; improvement on computing time should be sought. Also there is another method to handle relative large numerical feature (in this data, `fnlwgt`) called "scaling", which can be further explored in the future.

References

[1] https://www.kaggle.com/datasets/uciml/adult-census-income

[2] https://blog.mitsde.com/data-imputation-techniques-handling-missing-data-in-machine-learning/

[3] same as above

[4] https://machinelearningmastery.com/statistical-imputation-for-missing-values-in-machine-learning/

Irizarry, Rafael A., Introduction to Data Science: Data Analysis and Prediction Algorithms with R https://rafalab.dfci.harvard.edu/dsbook/

Irizarry, Rafael A., Introduction to Data Science: Statistics and Prediction Algorithms Through Case Studies https://rafalab.dfci.harvard.edu/dsbook-part-2/

https://www.kaggle.com/datasets/uciml/adult-census-income