Adult Income Data Analysis

Objective

The Adult Income dataset comprises information on individuals in the United States, such as their age, job class, degree of education, marital status, occupation, and income.
The purpose of this study is to forecast whether a person earns more or less than $50,000 per year depending on demographic and socioeconomic data. The "Adult" dataset from the UCI Machine Learning Repository [1] was utilized in this experiment.

Data Exploration

Data was collected from
https://archive.ics.uci.edu/ml/machine-learning-databases/adult/

The dataset contains 32,561 instances and 14 attributes. The attributes are as follows:

- age: continuous numerical variable representing the age of the individual
- workclass: categorical variable representing the type of work the individual does, such as Private, Self-emp-not-inc, etc
- fnlwgt: continuous numerical variable representing the number of people the census taker believes that observation represents
- education: categorical variable representing the highest level of education achieved, such as Bachelors, Some-college, 11th, etc
- education-num: continuous numerical variable representing the highest level of education achieved in numerical form
- marital-status: categorical variable representing the marital status of the individual, such as Married-civ-spouse, Divorced, Never-married, etc
- occupation: categorical variable representing the type of job the individual has, such as Tech-support, Craft-repair, Other-service, Sales, etc
- relationship: categorical variable representing the relationship status of the individual, such as Wife, etc
- race: categorical variable representing the race of the individual, such as White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, or Black
- sex: categorical variable representing the gender of the individual, either Male or Female
- capital-gain: continuous numerical variable representing the capital gains of the individual
- capital-loss: continuous numerical variable representing the capital losses of the individual
- hours-per-week: continuous numerical variable representing the number of hours the individual works per week

- native-country: categorical variable representing the country of origin of the individual, such as United-States, Cambodia, etc

  These features provide information about an individual's demographics, socioeconomic status, and employment.

We need to perform some data pre-processing before we can use this dataset for machine learning algorithms.

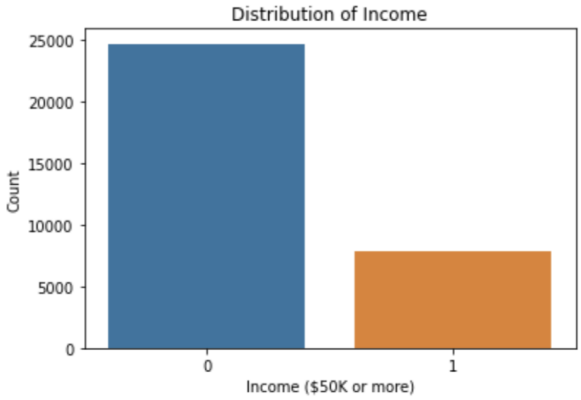
Data Pre-processing

We first removed the rows with missing values in the dataset. Next, we converted the categorical features into numerical features using one-hot encoding and numerical variables were standardized to have zero mean and unit variance using StandardScaler from scikit-learn.

The target variable income was transformed to a binary variable, with values of <=50K being labelled as 0 and values of >50K being labelled as 1.

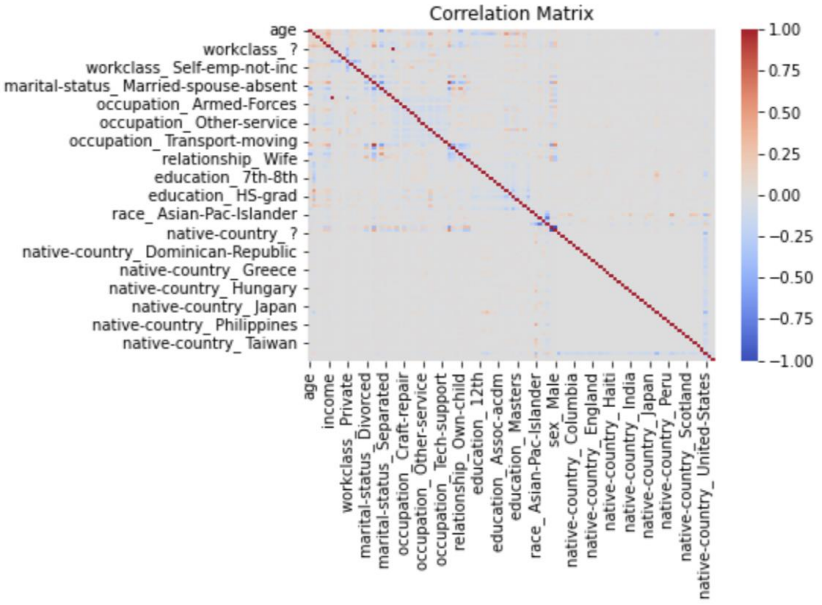
Exploratory Data Analysis

We first examined the distribution of the target variable income in the following plot:



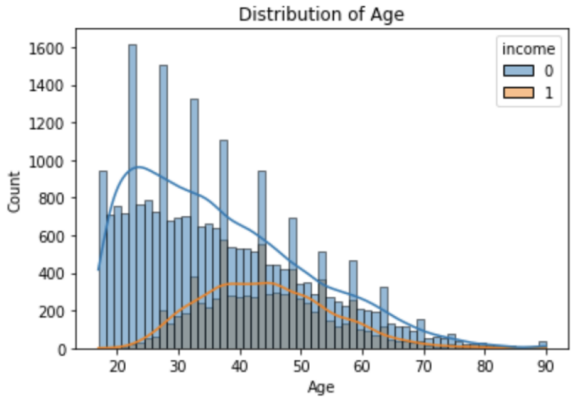
The dataset is imbalanced, with a larger number of individuals earning less than or equal to $50K per year.

We then examined the correlation matrix of the input features to see how they are related to each other. The following heatmap shows the correlation coefficients between the variables:



The heatmap demonstrates that there are no significant correlations between the input variables, indicating that multicollinearity is unlikely.

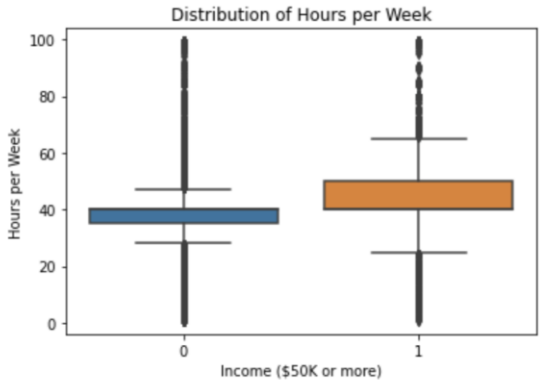
We also looked at the distribution of the age variable with income as the colour. The graphic below depicts the age distribution for both income groups:



The figure illustrates that the age distribution is comparable between the two income categories, with both groups peaking around the age of 35-40.

The age distribution is comparable between the two income categories, but those earning more than $50,000 per year work longer hours on average than those earning less than or equal to $50,000 per year.

Finally, we looked at how the hours-per-week variable was distributed across both income categories. The boxplot below depicts the distribution of weekly hours worked for both groups:



The boxplot shows that individuals earning more than $50K per year tend to work longer hours on average than those earning less than or equal to $50K per year.

These insights can be useful for building predictive models for income classification.

Modelling

The dataset was randomly split into training and testing sets with a 80:20 ratio, and a random state of 42 was used for reproducibility.

We train the following four classification algorithms on the dataset model with default hyperparameters and evaluate their performance using the test set:

- Logistic Regression: We used the Logistic Regression model from scikit-learn with default hyperparameters.
- Naive Bayes: We used the GaussianNB model from scikit-learn for Naive Bayes classification.
- Linear Discriminant Analysis (LDA): We used the Linear Discriminant Analysis model from scikit-learn.
- Quadratic Discriminant Analysis (QDA): We used the Quadratic Discriminant Analysis model from scikit-learn.

Model Evaluation

Evaluation of the models was performed on the test data using classification report, confusion matrix, and ROC curve. The performance of each model was compared using accuracy, precision, recall, and F1-score for the positive class.

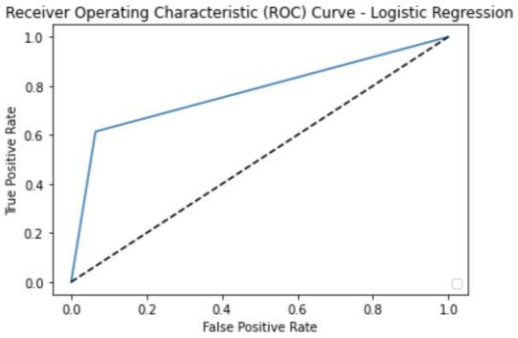
Performance of logistic regression

Confusion matrix
[[4622 320]
 [607 964]]

Classification report

	precision	recall	f1-score	support
0	0.88	0.94	0.91	4942
1	0.75	0.61	0.68	1571
accuracy			0.86	6513
macro avg	0.82	0.77	0.79	6513
weighted avg	0.85	0.86	0.85	6513

ROC curve



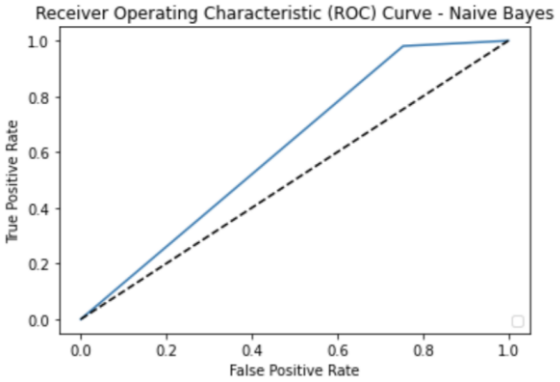
Performance of Naïve Bayes

Confusion matrix
[[1215 3727]
 [30 1541]]

Classification report

	precision	recall	f1-score	support
0	0.98	0.25	0.39	4942
1	0.29	0.98	0.45	1571
accuracy			0.42	6513
macro avg	0.63	0.61	0.42	6513
weighted avg	0.81	0.42	0.41	6513

ROC curve



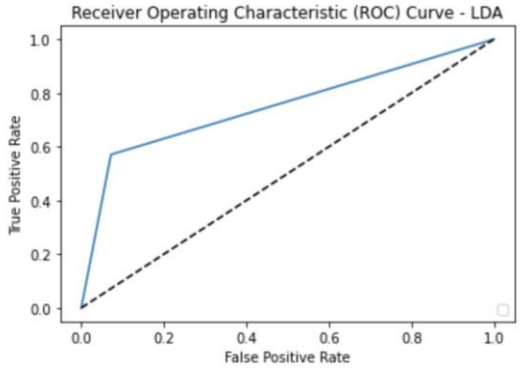
Performance of LDA

Confusion matrix
[[4586 356]
 [674 897]]

Classification report

	precision	recall	f1-score	support
0	0.87	0.93	0.90	4942
1	0.72	0.57	0.64	1571
accuracy			0.84	6513
macro avg	0.79	0.75	0.77	6513
weighted avg	0.83	0.84	0.84	6513

ROC curve



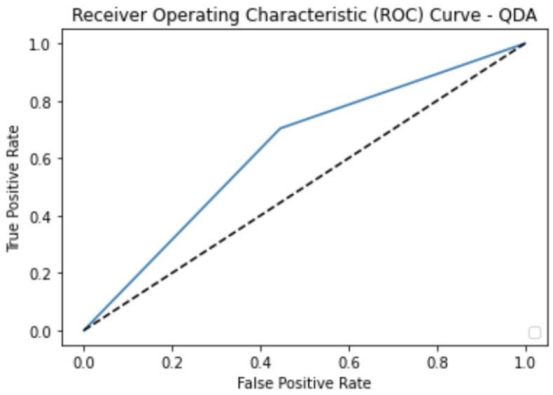
Performance of QDA

Confusion matrix
[[2741 2201]
 [465 1106]]

Classification report

	precision	recall	f1-score	support
0	0.85	0.55	0.67	4942
1	0.33	0.70	0.45	1571
accuracy			0.59	6513
macro avg	0.59	0.63	0.56	6513
weighted avg	0.73	0.59	0.62	6513

ROC curve



The logistic regression and LDA had similar performance, with accuracy scores ranging from 84% to 87%. The Naive Bayes and QDA models performed slightly worse.

Conclusion

The choice of model depends on the specific needs and goals of the application. However, based on the evaluation metrics, the logistic regression model may be a good choice for predicting high income individuals in this dataset. It is worth noting that further feature engineering and hyperparameter tuning may improve the performance of the models.

References

[1] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. Available at: http://archive.ics.uci.edu/ml/datasets/Adult.