Adult Income Prediction with Logistic Regression

A minimal logistic regression example in R using the UCI Adult Census Income dataset to predict whether an individual earns more than **50K** per year.

Dataset

- **Source**: UCI Adult Census Income dataset (adult.data)
- **Rows**: 32,561 observations
- **Key variables used in the model**:
  - AGE – Age in years
  - EDUCATIONNUM – Years of education
  - HOURSPERWEEK – Weekly working hours
  - CAPITALGAIN – Capital gains
  - CAPITALLOSS – Capital losses
  - ABOVE50K – Target; 1 if income >50K, 0 if <=50K (derived from the last column)

The data are read from the raw text file, column names are assigned, and ABOVE50K is created from the income label.

Objective

Predict whether an individual's income is **above 50K** (binary classification) using demographic and financial features, and evaluate the model with accuracy and ROC/AUC metrics.

Workflow

1. **Load and prepare data**
   - Read adult.data with read.csv() using header = FALSE and strip.white = TRUE.
   - Assign 15 descriptive column names and derive ABOVE50K as a binary target, then drop the original INCOME column.
   - Inspect data with head(), str(), and check class balance using table(ABOVE50K) (≈ 24,720 zeros vs. 7,841 ones).
2. **Balanced train/test split**
   - Split into:
     - input_ones: ABOVE50K == 1
     - input_zeros: ABOVE50K == 0
   - Sample 70% of the positive class and the same number of negatives to create a **balanced training set** trainingData.
   - Use the remaining positives and negatives as testData.
3. **Fit logistic regression model**
   - Model formula: using glm(..., family = binomial("logit")) on trainingData.
   - All predictors are highly significant ($p < 2e-16$) and have **positive** coefficients, meaning higher age, education, working hours, capital gains, and capital losses increase the odds of earning above 50K.

- Odds ratios from exp(coef()), for example:
    - AGE OR ≈ 1.052
    - EDUCATIONNUM OR ≈ 1.39
    - HOURSPERWEEK OR ≈ 1.047.

4. **Prediction and evaluation (cutoff = 0.5)**
    - Predict probabilities on testData with type = "response" and classify as 1 when p > 0.5.
    - Confusion matrix example:
        - True 0: 14,732 correctly predicted, 4,500 misclassified as 1.
        - True 1: 1,694 correctly predicted, 659 misclassified as 0.
    - Overall accuracy ≈ **76.1%**.
5. **ROC curve and AUC**
    - Use the ROCR package to compute ROC and AUC on predicted probabilities vs. actual ABOVE50K.
    - AUC ≈ **0.83**, which is typically interpreted as **good** discriminatory performance (0.8–0.9 range).
6. **Optimal cutoff and trade-offs**
    - Compute accuracy across thresholds and select the cutoff that maximizes accuracy (≈ **0.876**).
    - Best accuracy at this cutoff ≈ **90.3%** on the test set.
    - At this optimized cutoff:
        - Sensitivity ≈ **0.249** (true positive rate).
        - Specificity ≈ **0.983** (true negative rate).
    - This shows the trade-off: very high specificity but low sensitivity when optimizing purely for accuracy.

Files

- **R script**: Contains data loading, preprocessing, logistic regression model, predictions, confusion matrix, ROC/AUC, and optimal cutoff analysis (your glm and ROCR workflow).
- **Dataset**: adult.data from UCI, containing raw comma-separated records used to build the model.

You can save this text as README.md in your project and render or download it as a Markdown file from your development environment.

https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data