Income Classification Using Logic Regression

I. Analytical Framework

a) Dataset: UCI Adult Census Dataset
b) Source: https://archive.ics.uci.edu/dataset/2/adult
c) Limitations: Uses census data from 1994. This would not reflect current income distributions. Some features may have incomplete data.

1) Technology Platform
   a) Tools: Python, Google Colab, Scikit-learn, Pandas, Matplotlib, Seaborn.
   b) Ownership: Public dataset, no restrictions.
   c) Constraints: Colab isn't the quickest.
2) Business Alignment
   - Being able to classify income is useful in marketing, governments, insurance, and banking industries. The dataset lacks other financial information.
3) Methodology
   a) Approach: machine learning with binary logistic regression.
   b) Justification: computationally efficient.
4) Data Prep
   a) Missing Values: Dropped rows with missing values.
   b) Omitted Fields: Some fields like fnlwgt had no effect on what was needed.
   c) Encoding: Applied target and frequency coding.
   d) Data Split: 80/20 testing, training.

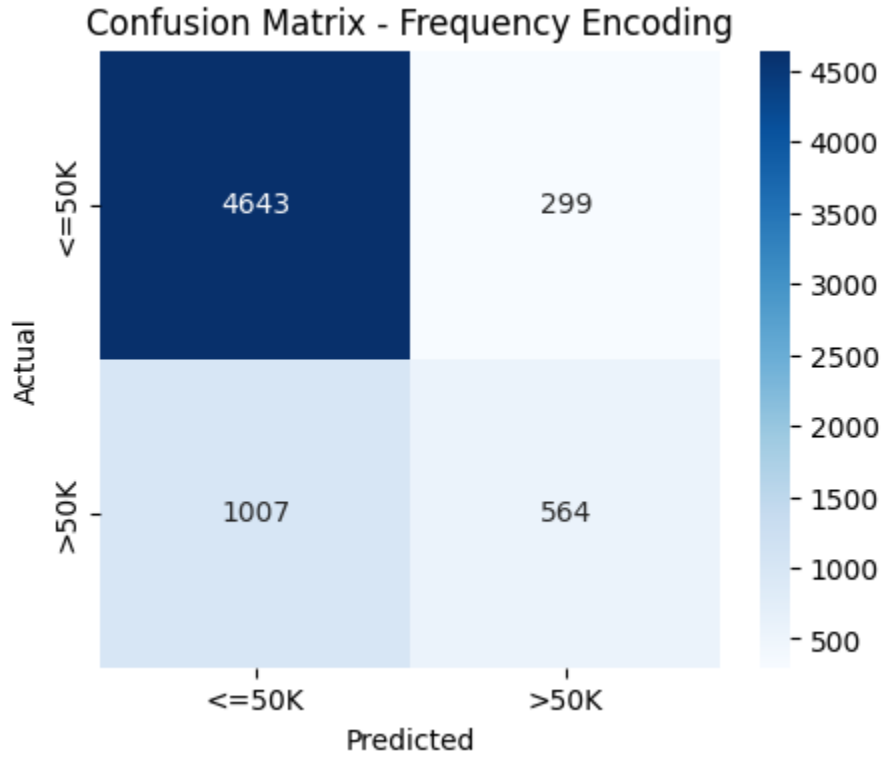
II. Machine Learning Techniques & Methodology

1) Modeling Approach
   a) Machine Learning Technique: Binary Logistic Regression.
   b) Encoding Techniques:
      i) Target Encoding: encodes categorical values based on the mean target value.
      ii) Frequency Encoding: Replaces the categorical values with the frequency counts.
2) Training & Validation Strategy
   a) Train-Test Split: 80/20
   b) Standardisation: StandardScaler
   c) Eval Metrics:
      i) Accuracy
      ii) Confusion Matrix
      iii) Precision, Recall, F1-Score
3) Limits
   a) Imbalance: Income >50k is not really in the dataset

III. Results

1) Model Performance Comparison:
   Target Encoding Model Accuracy: 0.8544449562413634
   Frequency Encoding Model Accuracy: 0.7994779671426377

2) Confusion Matrix Analysis



a)

b) Frequency Encoding:
   i) High False Negatives (1007): Model is bad at predicting earners >50k
   ii) Lower True Positives (564): Poor recall for high-income

c) Target Encoding:
   i) Higher Accuracy: Good improvement for high income.
   ii) Lower False Negatives: Comparable to frequency encoding.

3) Error Analysis
   a) Target Encoding: Did better because we included the target variable.
   b) Frequency Encoding: Lead to a feature bias making a bad classification.
   c) Scaling: improved on convergence leading to a stable result.

4) Suggestions
   a) Use advanced models.
   b) Tune parameters so that the logistic progression has a better generalization