Vikram Thirumaran

January 21, 2026

Prof Searcy/CS 422

Winter 2026

Project 1: Census Income Data Analysis and Classification Report

1. Correlation Heatmap Analysis

The correlation heatmap reveals the linear relationships between numerical features and the target income label.

Observations:
The attribute educational-num has the strongest positive correlation with the income, which reveals that the level of education plays a very important role in the prediction of the income being over 50K. The attributes age and capital-gain also tend to display positive correlation, which reveals that older persons along with members of the population who report higher capital gains are more likely to belong to the group of persons having a higher income. On the contrary, the final weight fnlwgt displays a near-zero correlation with the income category.

2. Education and Hours-Per-Week Histograms

The histograms provide a visual distribution of the educational attainment and work intensity of the individuals in the dataset.

Observations:
The histogram for education indicates strong points at "HS-grad," "Some-college," and "Bachelors," which correspond to the most common educational attainments in the labor market. The hours.per.week histogram has an enormous peak at 40 hours, marking the standard full-time employment in the United States. The lesser points along 50-60 hours reveal that some people in this data set work a substantial amount of overtime.

3. Performance Analysis (Learning Rate 0.05 vs. 0.75)

Model performance was evaluated using average log-likelihood and prediction accuracy across 500 iterations for two different learning rates.

Observations:
With a learning rate of 0.05, the log-likelihood curve is smooth and shows stable convergence, which reflects in steady and eventually plateauing accuracy increases. By contrast, the curve of

learning rate 0.75 exhibits a much faster initial convergence, though the log-likelihood curve may have slight oscillations when it approaches its optimum. While both learning rates reach similar final accuracy, the 0.05 rate is more robust against overshooting the global minimum, while 0.75 prioritizes speed at the expense of minor instability.

4. ROC Curves

The Receiver Operating Characteristic (ROC) curves describe how the True Positive Rate is related to the False Positive Rate for the logistic regression model.

ROC Figure:Learning Rate = 0.05

Summary: There is a strong tendency towards the top-left corner, which implies a good Area Under the Curve (AUC) value. This shows a proper separation of classes. A constant learning rate guarantees a well-generalized decision boundary.

ROC Figure: learning rate = 0.75

Conclusion:

The ROC Curve for the 0.75 learning rate is almost indistinguishable from the 0.05 learning rate in terms of their ultimate AUC value. This implies that although their routes to convergence were different, their destinations were the same – that of reaching an equally reliable set of weights that can classify income levels successfully.