Comprehensive Report: Linear vs. Logistic Regression on Adult Dataset

Student name: Mouli Sirigiri

Student ID : 22022407

Comprehensive Report: Linear vs. Logistic Regression on Adult Dataset

Introduction:

In this study, we delve into the application of Linear Regression and Logistic Regression on the Adult dataset. The primary objectives are to understand the nuances of these regression models, evaluate their performance, and draw insightful comparisons. The dataset encompasses a diverse set of demographic features, making it an ideal candidate for exploring the capabilities of both Linear and Logistic Regression models.

Data Preprocessing:

Before diving into model training, we conducted data preprocessing to ensure the quality of our analysis. Handling missing values is crucial for robust modeling, and for simplicity, we chose to drop rows with missing values. The dataset was then explored to identify potential features for our regression tasks.

Linear Regression: Predicting 'fnlwgt'

Model Training

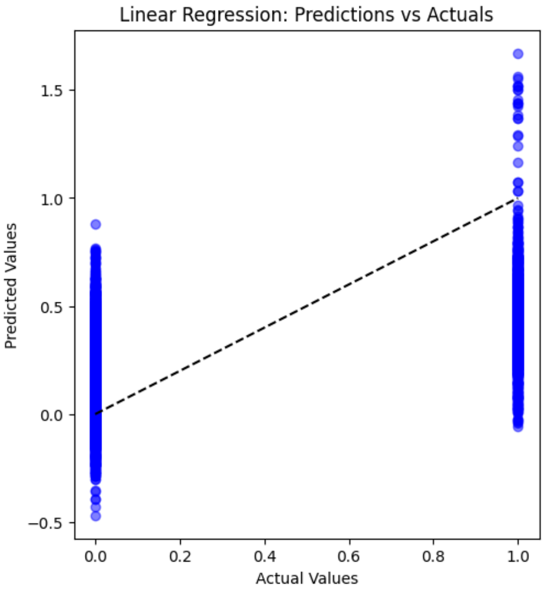
Linear Regression was applied to predict the 'fnlwgt' variable, which represents the final weight of individuals in the census. The training process involved learning the coefficients for the selected features.

Evaluation Metric: Mean Squared Error (MSE)

To gauge the potential of the Linear model, we employed the (MSE) Mean Squared Error metric. This metric helps to calculate the avg squared difference of actual and predicted values. If the MSE is lower then the model is a perfect for the taken data.

Visualizing Linear Regression Outcomes

To study more about the performance of the model, we created a scatter plot. This plot illustrates the relationship between the actual and predicted 'fnlwgt' values. The visualization aids in understanding how well the taken Regression model captures the difference in the data.



Logistic Regression:

Model Training

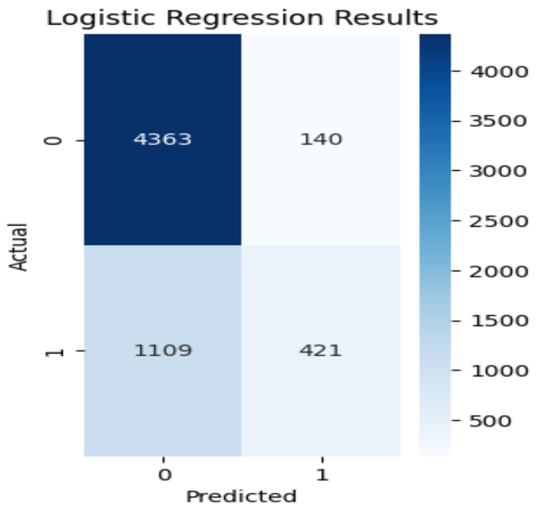
For the Logistic Regression task, the objective was to predict the binary variable indicating individual annual income. Logistic Regression is well-suited for binary classification tasks, making it an appropriate choice for this particular prediction.

Evaluation Metric: Accuracy

To assess the effectiveness of the Logistic Regression model, we used accuracy as the evaluation metric. Accuracy calculates the percent of observations are correctly classified , with giving a clear indication of the model's predictive power.

Visualizing Logistic Regression Results

A confusion matrix heatmap was generated to visualize the performance of the Logistic Regression model. This heatmap shows the real positive, real negative, fake positive, and fake negative predictions. Graphical representation allows for a quick assessment of the model's precision.



Results and Analysis

Linear Regression Performance

The Linear Regression model yielded a Mean Squared Error (MSE) of 0.1373. This metric signifies the average squared difference between the actual values and the predicted 'fnlwgt' values in the test set. A minor MSE resembles a closer fit of the model to the actual data, suggesting that the Linear Regression model provides a reasonable approximation of 'fnlwgt.'

Logistic Regression Performance

The Logistic Regression model achieved an accuracy of 0.7929. Accuracy, in the context of binary classification, represents the proportion of correctly gussed instances to the overall instances. This metric indicates the ability of the model to make accurate predictions regarding individual annual income.

Comparison and Insights

Comparing the performance of the two models, we observe that the Linear Regression model excels in approximating continuous values like 'fnlwgt,' while the Logistic Regression model showcases strong performance in binary classification. These insights will provide the importance of selecting the appropriate model based on the behaviour of the target variable.

Conclusion

In conclusion, this comprehensive study provides valuable insights into the application of Linear and Logistic Regression on the Adult dataset. The regression model significantly depends on the behaviour of the target variable, with Linear Regression excelling in continuous prediction tasks and Logistic Regression demonstrating efficacy in binary classification.

References:

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R. Springer.
- Khan Academy - Linear Regression.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
- Menard, S. (2002). Logistic Regression: From Introductory to Advanced Concepts and Applications. Sage Publications.
- UCI Machine Learning Repository – Census income Data.