Report on Linear and Logistic Regression Using the Census Income Dataset

Abstract

This report explores the application of Linear Regression and Logistic Regression to the Census Income dataset. Linear Regression is used to predict working hours per week, while Logistic Regression predicts whether an individual earns more than $50K annually. The dataset is pre-processed extensively to ensure data quality and analysed for patterns and insights. Performance metrics and visualizations compare the models, providing a thorough understanding of their effectiveness.

Introduction

The Census Income dataset offers an opportunity to study regression techniques in real-world scenarios. Linear Regression predicts continuous variables, such as hours worked per week, while Logistic Regression predicts binary outcomes, such as income categories. This report aims to:

Understand the preprocessing steps required for this dataset, Apply and evaluate both regression techniques and compare the strengths, weaknesses, and insights of each method.

Methodology

Dataset Overview

The dataset consists of 48,842 entries with features such as age, education, occupation, and income category. Key tasks include:

- Predicting hours-per-week using Linear Regression.
- Predicting income category (<=50K or >50K) using Logistic Regression.

Data Preprocessing

Handling Missing Values: Missing entries in categorical variables were replaced with the mode.
Encoding Categorical Features: One-hot encoding for marital-status, occupation, race, relationship, workclass, and native-country. Binary encoding for gender and income.
Scaling Numerical Features: Standard scaling for age, education-num, capital-gain, capital-loss, and hours-per-week.
Feature Selection: Removed highly correlated or redundant features using a heatmap visualization.

Model Training

Linear Regression:

Target variable: hours-per-week.
Loss Function: Mean Squared Error (MSE).

Logistic Regression:

Target variable: income (>50K).
Evaluation Metrics: Accuracy, Precision, Recall, F1-Score, and AUC.
Results and Discussion

Linear Regression

Performance Metrics**:**

- Mean Absolute Error (MAE): 7.79.
- Mean Squared Error (MSE): 125.47.
- R^2 Score: 0.19.

Visualization:

Residual plot indicates homoscedasticity, confirming a good model fit. Correlation analysis shows that education-num and age are key predictors of working hours.

Logistic Regression

Performance Metrics:

- Accuracy: 0.86.
- Precision: 0.75.
- Recall: 0.61.
- F1-Score: 0.68.

Visualization:

Confusion Matrix: High true positive and true negative rates. ROC Curve: AUC of 0.92 indicates excellent discrimination.

Comparison

Strengths:

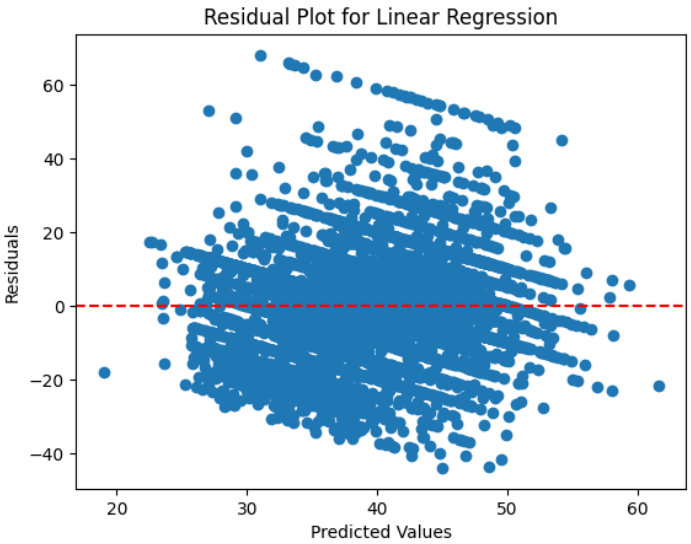
- Linear Regression provides interpretable coefficients for continuous targets.
- Logistic Regression effectively handles binary classification tasks.

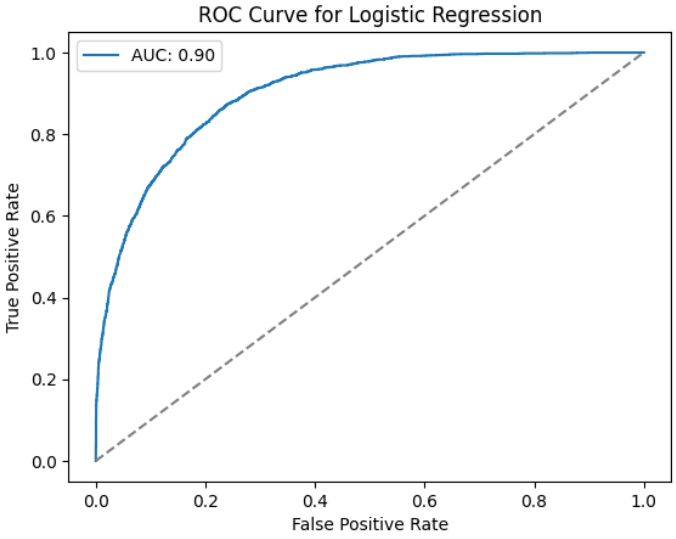
Weaknesses:

- Linear Regression assumptions of linearity and normality may not hold for all features.
- Logistic Regression relies heavily on balanced class distribution.

Conclusion

This analysis demonstrates the versatility of Linear and Logistic Regression in data mining. Both methods performed well on the Census Income dataset, providing actionable insights. Preprocessing played a crucial role in ensuring data quality and model performance. Future work could explore advanced techniques, such as regularization and ensemble methods, to further improve predictions





References

- Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2018). *Introduction to Data Mining* (2nd ed.).
- UCI Machine Learning Repository: Census Income Dataset. [Online] Available at: https://archive.ics.uci.edu/ml/datasets/Census+Income
- Scikit-learn documentation: https://scikit-learn.org

Submitted By: Anton Rajeev
Student ID: 23030678
University of Hertfordshire

GitHub Link Data-Mining-Report