untitled1

November 19, 2024

To tackle this problem, we will follow a structured approach. Here's a step-by-step solution framework:

0.0.1 Solution Approach

1. Problem Understanding The goal is to predict labor earnings for 1978 based on demographic, educational, and employment-related factors using data from 1974 and 1975.

2. Data Preparation

- **Data Exploration**: Understand the data distribution for Age, Race, Education, etc., and examine how these features relate to earnings.
- **Data Cleaning**: Handle missing values, outliers, or incorrect entries.
- **Feature Engineering**:
  - Convert categorical variables (like Race, Hispanic, Married) into numerical forms using encoding techniques.
  - Normalize or scale numeric variables like Age and Education levels.
- **Check for Bias**: Since the dataset is based on a copy of the U.S. version, it is important to analyze the racial and ethnic disparities in data and consider fairness in predictions.

3. Model Selection

- Use **Linear Regression** as the target variable (earnings) is continuous.

0.0.2 Linear Regression Assumptions and Verification

Linear Regression relies on several assumptions. Here's how to check and verify them:

1. Linearity

- The relationship between independent variables and the dependent variable must be linear.
- **How to check**:
  - Plot scatter plots of each independent variable vs. earnings.
  - Check for linear trends.

2. Independence

- Observations must be independent.
- **How to check**:
  - Ensure that the data collection process doesn't introduce dependencies.
  - Look for correlations in residuals (Durbin-Watson test).

3. Homoscedasticity

- The variance of residuals should remain constant across levels of predicted values.
- **How to check**:
  - Plot residuals vs. predicted values. The plot should not show a funnel shape.
  - Conduct a Breusch-Pagan or White test.

4. Normality of Residuals

- Residuals should be normally distributed.
- **How to check**:
  - Plot a histogram or Q-Q plot of residuals.
  - Perform a statistical test like the Shapiro-Wilk test.

5. Multicollinearity

- Independent variables should not be highly correlated with each other.
- **How to check**:
  - Calculate the Variance Inflation Factor (VIF). A VIF > 10 indicates multicollinearity.

0.0.3 4. Model Building

- Split the data into training and testing sets.
- Train the Linear Regression model using the training set.
- Evaluate using metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and (R^2).

0.0.4 5. Model Evaluation

- Compare the predicted vs. actual values for the 1975 data as a validation step.
- If the model performs well on 1975, proceed to predict 1978 earnings.

0.0.5 6. Ethical Considerations

- Be cautious of biases introduced by variables like Race or Hispanic origin.
- Fairness metrics should be considered to ensure no group is disadvantaged.

Would you like to dive into the coding implementation for this, including data preprocessing or checking the assumptions?