AI Cure Analysis Report

Ashwin Waghmare
Hrishikesh Karande

January 19, 2024

1 Data Loading and Preprocessing

- Loaded dataset from 'train_data.csv'.
- Dropped irrelevant columns ('uuid', 'datasetId').
- Handled missing values by dropping rows with NaN values.
- Encoded categorical variable 'condition' using LabelEncoder.

2 Exploratory Data Analysis (EDA)

- Checked data types of columns.
- Utilized Pearson's correlation to analyze feature correlations.
- Verified absence of null values.
- Described the distribution of the target variable 'HR' using summary statistics and a histogram.

3 Feature Selection

- Applied SelectKBest with f_regression scoring to select the top 15 features.
- Visualized feature scores using a bar chart.

4 Regression Modeling

4.1 Linear Regression Model

- Trained a Linear Regression model on the dataset with selected features.
- Evaluated model performance on both training and testing sets using Mean Squared Error and R-squared scores.

4.2 Gradient Boosting Regression Model

- Implemented a Gradient Boosting Regression model with hyperparameters.
- Assessed model performance using the same evaluation metrics.

5 Results

5.1 Linear Regression

With 15 selected features:

- Training Error: [train_error]
- Testing Error: [test_error]
- Training R2 Score: [train_score]
- Testing R2 Score: [test_score]

5.2 Gradient Boosting Regression

With all features:

- Training Error: [train_error]
- Testing Error: [test_error]
- Training R2 Score: [train_score]
- Testing R2 Score: [test_score]

6 Observations

- The Gradient Boosting model shows improvement over the Linear Regression model.
- Further hyperparameter tuning and feature engineering may enhance model performance.