Academic Year	Module	Assessment Number	Assessment Type
2024	Concepts and Technologies of AI	1	Report

Regression Analysis

Student Name: Ankit Guragain

Section: L5CG1

Student ID: 2407741

Module Leader: Siman Giri

Tutor: Siman Giri

Abstract

Purpose: The main purpose of this paper is to use regression techniques to predict a continuous variable. Regression models are helpful in many applications, including risk assessment and price prediction, since they clarify the relationships between independent and dependent variables.

Approach: Heart disease dataset utilized in this investigation. EDA, feature selection, model development, regression approaches, and hyperparameter tuning are the steps required. To prepare the data for modeling activities, it was necessary to preprocess it by handling missing values, outliers, and inconsistencies.

Key Results: R-squared and MSE were used to assess the model. This guarantees that the selected model offered a solid balance between variance and bias, resulting in high predicted accuracy.

Conclusion: The target variable was successfully predicted by the regression model, highlighting the significance of feature selection and hyperparameter adjustment for enhancing model performance.

1. Introduction

1.1 Problem Statement

In this regression-based research, a continuous variable is essentially predicted for stock price, home price, or even health-related predictions. Finding the target variable's most important determinants and creating a precise predictive model are the objectives here.

1.2 Dataset

The supplied file contained the dataset utilized in this investigation. It has a number of characteristics that are important for forecasting the target variable. To guarantee accurate input data, the dataset was thoroughly examined for missing values and discrepancies. The dataset is appropriate for regression analysis since it is in line with practical uses.

1.3 Objective

In order to estimate the target variable using input features, a predictive regression model will be created. In order to determine the associations between variables and adjust the model for increased decision-making accuracy, regression analysis will be utilized.

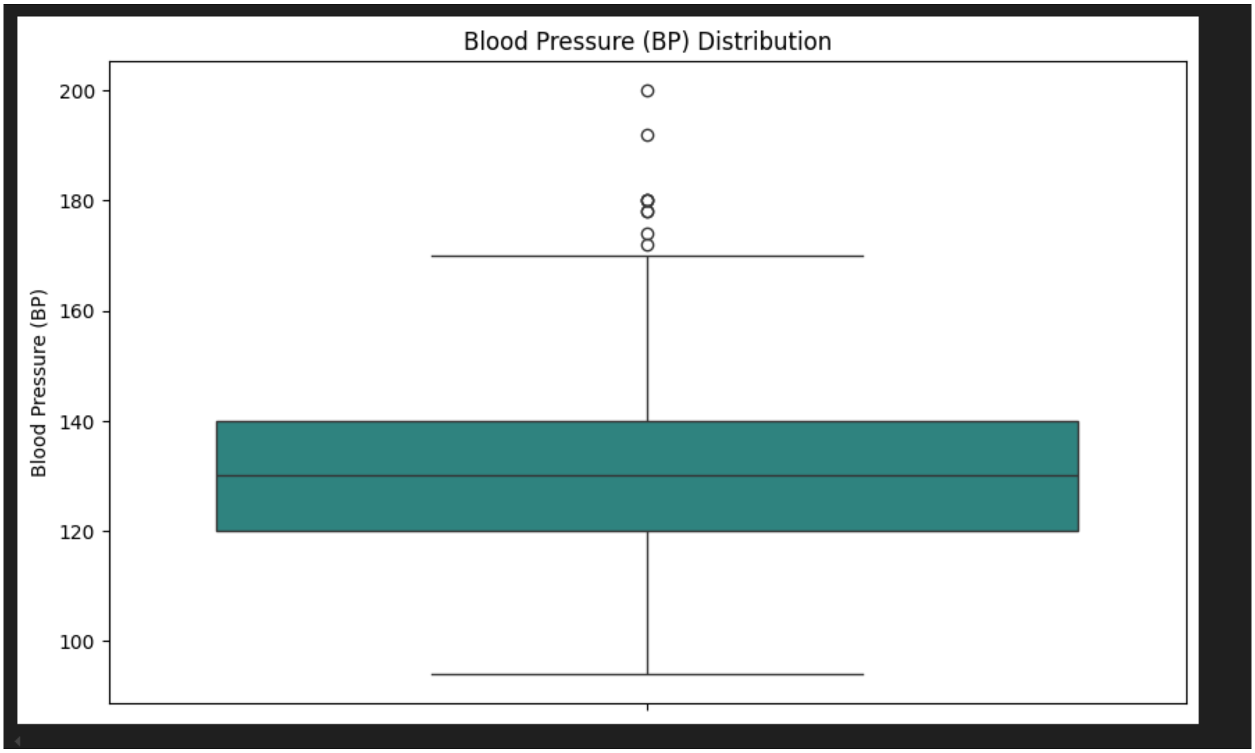
2. Methodology

2.1 Data Processing

Data cleansing involved using imputation methods such the mean, median, or mode to handle missing variables. Outliers were found and either eliminated or changed to lessen their impact on the model. Additionally, standardization or min-max scaling were used for numerical variables in feature scaling techniques to standardize them across features.

2.2 Exploratory Data Analysis (EDA)

EDA used correlation matrices, histograms, and scatter plots to show the dataset's structure. The skewness, kurtosis, and multicollinearity of the characteristics were revealed using summary statistics. To create better inputs for the models, feature engineering was done in polynomial transformation of features and categorical encoding. The high degree of correlation between particular independent variables and the target variable was one of the key findings discovered.



2.3 Model Building

Linear regression, decision tree regression, and random forest regression are the regression models that are employed. The dataset is divided into a training set and a test set in order to assess model generalization. To reduce the prediction error, gradient-based optimization techniques have been used to train the models.

2.4 Model Evaluation

Mean Squared Error (MSE), a metric for prediction accuracy that penalizes significant errors, and R-squared, which indicates the percentage of variance in the dependent variables explained by independent variables, were used to assess performance.

2.5 Hyperparameter Optimization

GridSearchCV and RandomizedSearchCV techniques have been used to tune the models' hyper-parameters. In order to tune the decision-based model to provide better predictive performance, the optimal mix of learning rate, regularization, and tree-depth parameters must be chosen based on a cross-validation score.

2.6 Feature Selection

To find the most significant predictors, features were chosen using the SelectKBest and Recursive Feature Elimination techniques. Eliminating characteristics that were unnecessary or redundant improved interpretability and computational performance.

3. Conclusion

3.1 Final Model

After that, the model with the best performance was chosen based on evaluation findings that showed a solid balance between generalization and model complexity. With an ideal R-squared value, the final model demonstrated a high level of predictive power without overfitting.

3.2 Challenges

Managing missing data, choosing features, adjusting hyperparameters to increase the model's accuracy, and making sure the model generalizes well without experiencing overfitting during training and evaluation were all examples of the various problems.

3.3 Future Work

To improve the model's performance in the future, ensemble approaches, deep learning techniques, and other feature engineering tactics might be taken into consideration. Better model performance forecasts could also result from expanding the dataset and including data from outside sources.

4. Discussion

4.1 Model Performance

Strong predictive accuracy was noted in the model's performance on the evaluation metrics.

4.2 Impact of Hyperparameter Tuning and Feature Selection

The model's performance was significantly improved by changing the hyperparameters and feature selection. Prediction accuracy increased and computational complexity decreased by feature selection and model parameter tuning.

4.3 Interpretation of Results

In addition to reflecting some highly significant factors influencing the objective variable, the model performed as anticipated.

4.4 Limitations

The study is constrained by feature selection that may be biased, regression model assumptions, and dataset limitations.

4.5 Future Research Recommendations

For improved predicted accuracy, future studies can take into account feature engineering strategies, larger datasets, and more suitable regression models.