Report on Polynomial Regression Model for Heart Rate Prediction

Introduction

This report details the development and evaluation of a Polynomial Regression model designed to predict heart rates based on a set of input features. The dataset utilized for training and testing purposes is sourced from a CSV file named 'train_data.csv.'

Data Loading and Initial Exploration

The initial step involves loading the dataset and conducting an initial exploration to comprehend its structure and characteristics. The 'reading_csv_file' function is employed to read the CSV file and store the data in a Pandas DataFrame. The dataset is then displayed to provide an overview of its content.

Data Wrangling

Data wrangling is a critical step in preparing the dataset for model training. This includes removing irrelevant features, encoding categorical variables, and addressing missing values. The 'drop_features' function is used to eliminate unnecessary columns, and label encoding is applied to the 'condition' feature using the 'label_encoding' function. We dropped ‘uuid’ and ‘datasetid’ as they were totally unrelated parameters.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is performed to visualize the relationships between input features and the target variable (heart rate). The 'data_plots' function generates scatter plots for each feature against the heart rate. Additionally, the correlation between each input feature and the target variable is calculated and displayed using the 'correlation_determinance' function.

We used polynomial regression based on the correlation matrix and various other numerical data analysis. Based on the numerical analysis of data we concluded that tree-based models are not suitable. And the data is too linear to use deep linear models. Since this is a medical data we cannot use any gaussian model. Also the data is not suitable for classical linear model and hence we use polynomial regression.

Model Training

The Polynomial Regression model is trained using the 'PolynomialFeatures' and 'LinearRegression' classes from Scikit-Learn. The data is split into training and testing sets, and the model is fitted using the 'fit' method of the 'Pipeline' class. Training set contains 80% and testing set contains 20% of the data.

Model Evaluation

The performance of the model is evaluated using the R-squared score on the testing set. This metric provides insights into how well the model captures the variability in the target variable.

Model Deployment

The trained model is saved for future use. In this instance, it is saved using the 'joblib' library. Saving the model allows for easy retrieval and application to new data for heart rate predictions.

Conclusion

The choice of Polynomial Regression in this context may be motivated by the desire to capture potential non-linear relationships between the input features and the target variable (heart rate). Polynomial Regression allows for the modeling of more complex relationships than simple linear regression.

Here are some reasons why Polynomial Regression might be chosen:

1. Non-Linearity in Data: If the exploratory data analysis (EDA) reveals that the relationship between the input features and the target variable is not strictly linear, Polynomial Regression can better capture non-linear patterns. The scatter plots generated in the EDA phase might indicate curvature or trends that cannot be effectively modeled by a linear function.

2. Feature Interaction: Polynomial Regression introduces terms that represent the interaction between different features. This is particularly useful when there are indications that the impact of one feature on the target variable is dependent on the value of another feature.

3. Flexibility Polynomial Regression is a flexible model that can fit a wide range of data patterns. By using higher-degree polynomial terms, the model can adapt to more intricate relationships present in the data.

4. Trade-off between Bias and Variance: While a higher-degree polynomial might lead to overfitting, careful regularization and validation techniques can be applied to find a balance between bias and variance. This trade-off aims to ensure the model generalizes well to new, unseen data.

It's important to note that the choice of a regression model should be guided by a combination of data characteristics, domain knowledge, and the specific goals of the analysis. Polynomial Regression is a valuable tool, but its effectiveness depends on the underlying patterns in the data and the potential risk of overfitting when using higher-degree polynomials.

The Polynomial Regression model demonstrates its efficacy in predicting heart rates based on the provided features. Further refinement, parameter tuning, and validation on new datasets can be conducted to enhance the model's accuracy and generalization capabilities.