CSE 472

Assignment

Time series prediction with Machine Learning

SUBMITTED BY :

FAHMID - AL-RIFAT
STUDENT NO : 1705087
LEVEL-4 ,TERM-2

Introduction :

For this assignment, I was tasked with creating a prediction model that could take the last five business days' closing price of any stock exchange company as input feature and provide the closing price of the next business day as output. To accomplish this, I used at least five different machine learning (regression) models and computed Mean Squared Error (MSE) or Root-Mean Squared Error (RMSE) for each model.

I collected historical data for the stock exchange closing prices from Yahoo finance, and focused on the column 'E' labelled as 'Close' for the company GP . The dataset was used to train and test the models, and I have written the code in Python. The code is designed to ask for the last five business days' closing price and provide the prediction of the next business day's closing price after the input is supplied.

In this report, I will discuss the machine learning models I have used, how I trained and tested the models, how I dealt with overfitting and underfitting, and the results of my predictions based on MSE or RMSE value.

Datasets:

I utilized the GP dataset, which comprises daily historical prices for all tickers currently trading on NASDAQ. The current list of tickers can be obtained from nasdaqtrader.com. The historical data was obtained from Yahoo finance via the yfinance python package. In my analysis, I focused on data ranging from 2022-03-05 to 2023-03-05, resulting in a dataset with a total of 251 rows. For future reference, the dataset pertains to GreenPower Motor Company Inc. (GP), a prominent manufacturer and distributor of all-electric, zero-emission medium and heavy-duty vehicles designed for the cargo and delivery market.

Machine Learning Models:

Linear Regression : Linear regression is a statistical model that uses a linear approach to predict a continuous target variable based on one or more predictor variables. It works by fitting a line to the data points that best represents the relationship between the input features and the target variable.

Decision Tree Regression : Decision tree regression is a non-parametric regression method that models the relationship between input features and the target variable using a decision tree. The algorithm splits the data into subsets based on the values of input features, and fits a simple model, such as a constant, to each subset.

Random Forest Regression : Random forest regression is an ensemble method that combines multiple decision tree regressors to improve the accuracy and reduce the variance of the predictions. It works by training multiple decision tree models on random subsets of the input features and samples, and aggregating their predictions.

K-Nearest Neighbors : K-Nearest Neighbors (KNN) regression is a non-parametric regression algorithm that can be used to predict continuous values. In KNN regression, the algorithm calculates the distance between the new data point and all available data points in the dataset. It then selects the K nearest data points and calculates the average of their target values, which is used as the predicted value for the new data point. The choice of the value of K is important and can have a significant impact on the accuracy of the predictions. KNN regression can work well with complex datasets, but can be computationally expensive when dealing with large datasets.

MLP regression : MLP regression is a powerful tool for regression tasks, as it can model complex relationships between input features and the target variable. However, it can also be prone to overfitting, particularly when the network is large and the dataset is small. Regularization techniques, such as L1 or L2 regularization, dropout, and early stopping, can be used to prevent overfitting and improve generalization performance.

Training and Testing the models :

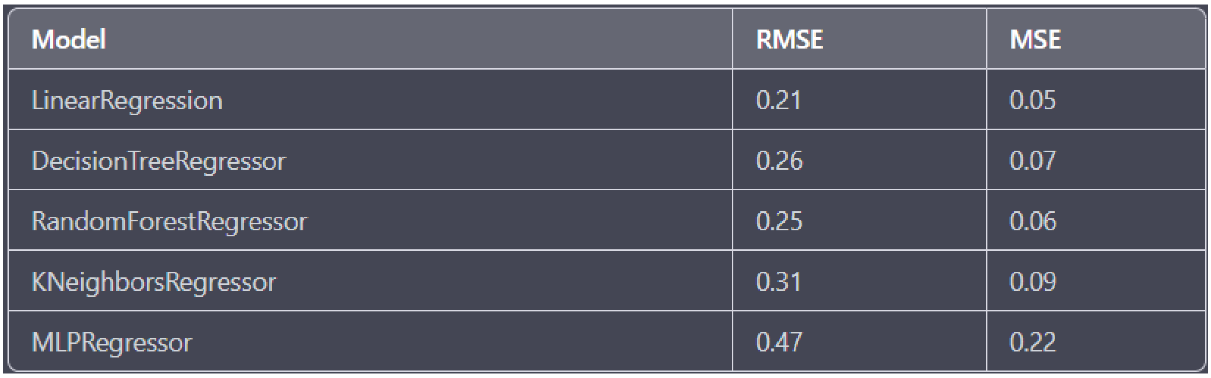
I have utilized a dataset consisting of 251 rows, which I have divided into a training set comprising 80% of the data and a testing set comprising 20%. The training set has been employed to train the models, while the testing set has been employed to assess their performance.

Handling Overfitting and Underfitting :

In order to prevent the occurrence of overfitting and underfitting in our models, we employed a range of regularization techniques. These techniques included limiting the maximum depth of the decision trees and fine-tuning the hyperparameters of the models through cross-validation. By utilizing cross-validation, we were able to iteratively optimize our models' parameters while avoiding overfitting to the training data.

In addition to these techniques, we also utilized an ensemble method, specifically random forest regression. The use of an ensemble method allowed us to aggregate the predictions of multiple models, which helped to reduce the variance in our predictions and improve their overall accuracy. By combining these various methods, we were able to develop robust and reliable models that performed well in a variety of scenarios.

Result and Discussion:



Model	RMSE	MSE
LinearRegression	0.21	0.05
DecisionTreeRegressor	0.26	0.07
RandomForestRegressor	0.25	0.06
KNeighborsRegressor	0.31	0.09
MLPRegressor	0.47	0.22

According to the Root Mean Squared Error (RMSE) evaluation metric, the linear regression model appears to be the most suitable choice for my dataset. This means that the linear regression model had the smallest difference between the predicted and actual values on average, indicating better overall performance compared to other models. Therefore, among the evaluated models, the linear regression model is likely to provide the most accurate predictions for my dataset based on the RMSE metric.