Report on Time Series Prediction With Machine Learning

Submitted by: Arif Shariar Rahman, ID: 1705095

Dataset

For the dataset, I have used **AAPL** dataset. Dataset was downloaded using **yfinance** library of python. The AAPL dataset refers to the stock prices of Apple Inc., a leading multinational technology company. The dataset includes historical stock prices for Apple's common stock traded on NASDAQ, covering a period of several years. The dataset provides information on the daily opening and closing prices, highest and lowest prices of the day, trading volume, and other related financial metrics. This information is valuable for investors and financial analysts who use it to track the performance of the company and make informed decisions about buying or selling Apple's stock. The AAPL dataset is widely used in financial research and is available from various sources, including financial data providers and online databases such as Yahoo Finance. For my model, I have used data from 2005-01-01 to 2022-01-01.

How to Run Code

Create a virtual environment in the project directory following this command:

pip install virtualenv

python3 -m venv env

Activate virtual environment by following this command:

source env/bin/activate

After the virtual environment is activated, run this command:

pip install -r requirements.txt

After all necessary libraries have installed, you can run '**machine_learning.py**' by following this command:

python3 machine_learning.py

Machine Learning Models Used

1. **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.

2. **Decision Tree Regression:** Decision tree regression is a non-parametric regression method that uses a decision tree to model the relationship between the input features and the target variable. It works by recursively splitting the data into subsets based on the values of the input features, and fitting a simple model (e.g., a constant) to each subset.
3. **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.
4. **Support Vector Regression:** Support vector regression is a regression method that uses support vector machines (SVMs) to find the best linear or nonlinear function that separates the input features and the target variable. It works by finding a hyperplane in the input feature space that maximally separates the data points, and using it to predict the target variable.
5. **Multi Layer Perception Regressor:** The Multi-Layer Perceptron (MLP) Regressor is a type of artificial neural network that is commonly used for regression tasks. It consists of an input layer, one or more hidden layers, and an output layer. The MLP uses non-linear activation functions and backpropagation algorithm to learn complex relationships between inputs and outputs. It is capable of modeling non-linear relationships and can handle large and complex datasets. The MLP Regressor is a popular choice for solving regression problems in various fields, including finance, engineering, and healthcare.

Training and Testing the Models

The dataset I have used in the models has 4280 rows. I have split the dataset into 80/20 ratio, where 80 is the ratio for training and 20 is the ratio for testing. I have used the training set to fit the models and the testing set to evaluate their performance.

Dealing With Overfitting and Underfitting

To avoid overfitting and underfitting, we used a combination of regularization techniques, such as setting the maximum depth of the decision trees, and tuning the hyperparameters of the models using cross-validation. We also used an ensemble method (random forest regression) to reduce the variance of the predictions and improve their accuracy.

Results based on MSE, RMSE

	Linear Regression	Decision Tree Regressor	Random Forest Regressor	Support Vector Regressor	Multi Layer Perception Regressor
MSE	1.187	1.9091	1.45629	27.3094	1.51272
RMSE	1.0897	1.38173	1.20677	5.22584	1.2299
MAE	0.48184	0.6434	0.54103	1.22538	0.5380

Future Works

There are several potential future works that can be done to improve the prediction model and make it more accurate and efficient. Some of them are:

1. **Feature engineering:** Adding more features such as technical indicators, economic indicators, or social media sentiment can improve the model's performance.

2. **Hyperparameter tuning:** Tuning the hyperparameters of the machine learning models can improve the model's performance. Techniques like grid search or random search can be used to find the optimal hyperparameters.

3. **Ensemble methods:** Ensemble methods such as bagging, boosting, or stacking can be used to combine the predictions of multiple models and improve the overall performance.

4. **Deep learning:** Using deep learning models such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) can be explored to capture more complex patterns in the data.

5. **Online learning:** Implementing an online learning algorithm can allow the model to learn from new data and adapt to changing market conditions.

6. **Transfer learning**: Transfer learning can be used to leverage pre-trained models on other related datasets to improve the performance of the model.

7. **Evaluation metrics:** Using additional evaluation metrics such as Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), or R-squared can provide a better understanding of the model's performance.