INFERENCE FROM RELATED WORK

The inference from the related works in the literature review section highlighted the need for which this research work has been proposed.

From table 1, the Linear regression model was used for stock prediction and R squared metric was the chosen metric. However, one major improvement this proposed work seeks to cover is the extension of choice metric to the other three metrics, which are MAE, RMSE, and MSE.

In addition, a larger dataset will be used against the short-term dataset and the period will be over thirty years. The purpose of the proposed improvement is to provide a more reliable and tested perspective on the use of the Linear regression model to predict stock prices both for short-term and long-term investors.

For ARIMA, data used in the reviewed literature were raw and not clean. For the reduction of errors, data cleaning will be carried out in this work. Also, as an update to the prior related work, more data will be employed. The purpose of a larger data set is perhaps a lesser error value can be achieved. Four error metrics will be used against two in the prior works captured in table 1 of the literature review.

A common deep learning model which was also used in table 1 is LSTM. As an update to the work reviewed in the literature review section, the same error metric used for linear regression, AR, and ARIMA models will be used in the LSTM model. This work attempts to improve upon the short-sightedness of using scholastic gradient descent by using the Adam optimizer. We also aim is to present a perspective on the strength of all models proposed in stock price prediction based on choice datasets. In addition, provide each model character in handling time series data.

3.0 Research Methods

The proposed approach of this work is segregated into multiple steps, with each step detailed in this section. The first section encapsulates necessary information relating to datasets used in this research work. These are data sources and feature scaling. The following section illustrates the models for stock prediction. For this work, the Auto Regression (AR) model, Auto Regressive Integrated Moving Average (ARIMA), Linear Regression (LR), and LSTM with each representing a statistical model, and deep learning models. The final section explains the data description.

3.1 Dataset and Feature Scaling

Data preprocessing captures cleaning data of noise, removal of irrelevance, and finding missing values. Splitting data into train and test datasets without the initial process of preprocessing poses a risk that can be significant to the performance of a model. Another drawback of not doing data preprocessing is that there might be biases. To avoid errors and biases, critical steps of data preprocessing should be done prior to data being fed into the relevant testing models. Since we obtained stock prices over several years, there is the likelihood of datasets distributed over a wide range of values and outliers.

In this paper, we will use datasets for Apple Inc. from 1980 till 19/1/2023. We will be checking and treating missing values. For Netflix, datasets from 1/1/2020 to 19/1/2023 will be used. Each data is composed of date, open, close, high, low, and Adjusted close prices, and will be predicting close price.

The LSTM model performs better with normalized and scaled datasets. The learning process becomes faster for the sequence prediction with appropriate scaling for input and output variables. For this work, we will use the Min-Max-Scaler.

3.2 Models (AR, ARIMA, LINEAR REGRESSION, and LSTM)

The AR model predicts future values when it is builds-in previous values as input datasets. The past datasets are applied in modeling future behavior with randomness built in. The methodology includes the estimation of model parameters. This entails the identification of the behavior of partial autocorrelation (PACF). After parameter estimation, AR also allows the use of new data for prediction. With fitting, a lower Akaike information criterion (AIC) is better as this means a better-fit model.

For the ARIMA model, we will plot diagrams to present the datasets for more visual representation for both scaled, variations, show trends, seasonality, and prediction results. P and q values represent autocorrelation and moving averages. We will be computing rolling mean and standard deviation, checking stationarity, and selecting the best ARIMA model. ARIMA models dataset will be split into two, with training data and test data in a ratio of 80:20. The choice of 80:20 signifies sufficient training size to afford a good prediction.

The focus of our proposed linear regression model will be to detect the best fit. Since linear regression can be easy to understand and compute, the difference between the prediction and actual prices is expected to give a clearer perspective of the strength of the model. Like other models, the four-error metrics will be calculated in addition to the mean, standard deviation, and other values which are outputs of the model.

For LSTM, we will use Adam Optimizer. Adam Optimization algorithm is an upgrade from SGD due to the ability of the algorithm in terms of speed of execution. Deep learning models require adaptation of their neural networks such as weights and learning rates. The purpose of the optimization includes accuracy enhancement. Another benefit of using Adam is to limit actual loss. LSTM does not require p, q, d as we have in ARIMA however, needs hyperparameter tuning. For hyperparameter tuning, we suggest a total epoch of 100 however is subjective. Epoch refers to the period a complete pass occurs via the training dataset. Batch size is also required of LSTM, we, therefore, propose a batch size of between 32 and 64 inclusive.

The training and testing will be carried out using Python and we will install a couple of libraries. Sklearn, Karas, TensorFlow, NumPy, and pandas, are additional apart from the error metrics, matplotlib, and model libraries. We will import the Min-Max Scaler for the training dataset and reshape relevant data to data shape.

Due to the nature of MAE, MSE, and MAPE time series forecasting error metrics, we have chosen them in addition to RMSE as forecast metrics. The MAE affords some numeral stability and is easy to interpret as the average absolute difference between two variables x and y. The R-squared error metric has not been included because it will not show if a model is suitable for future forecasts or not. The MAPE was also included because it gives a percentage return demonstrating an easy error value.

3.3 Data Description

The use of statistical models, and machine and deep learning models is to provide a fairly safe prediction with higher accuracy. This is proposed as an improvement on existing work by other researchers. Data employed will be fine-tuned to ensure correct adaptation and interpretation for training and metrics. We will test for data stationarity using the Augmented Dickey-Fuller test. Stationarity in the AR model denotes without trend. We will use lagged values as prediction variables. We will be comparing predicted results against real prices.

Research Questions

How can a dataset provide improved stock price prediction accuracy?

What chance of correlation is possible between models and data volumes?

What does past price trend impute towards metric performance in stock price prediction?

How best can new models predict stock market price movements?