**Stock Price Correlation Coefficient Prediction**

**with ARIMA-LSTM Hybrid Model**

**Introduction**

When constructing and selecting a portfolio for investment, evaluation of its expected returns and risks is considered the bottom line. Investors then select a portfolio on the efficient frontier, depending on their risk tolerance. It’s seen correlation of any two

assets may as well deviate from mean historical correlation coefficients subject to financial conditions. Thus, the correlation s not stable.

An alternative is the Constant Correlation model, which sets all pairs of assets’ correlations equal to the mean of all the correlation coefficients of the assets in the portfolio. The stock correlation data can also be represented as time series data -

deriving the correlation coefficient dataset with a rolling time window – application of neural networks in forecasting future correlation coefficients can be expected to have successful results as well. Rather than circumventing by predicting individual asset returns to compute the correlation coefficient, they cast predictions directly on the correlation coefficient value itself.

Looking at these ways this paper suggests a hybrid model of the ARIMA and the neural network to predict future correlation coefficients of stock pairs. The model adopts the Recurrent Neural Network with Long Short-Term Memory cells(LSTM). In the first phase, the ARIMA model catches the linear tendencies in the time series data. Then, the LSTM

model attempts to capture non-linearity in the residual values, which is the output of the former phase.

**Various Financial Models for Correlation prediction**

The Full Historical model is the simplest method to implement the portfolio correlation estimation. This model adopts the past correlation value to forecast future correlation coefficient. The Constant Correlation model assumes that the full historical model

encompasses only the information of the mean correlation coefficient. Applying the Constant Correlation model, all assets in a single portfolio have the same correlation coefficient. The Single-Index model presumes that asset returns move in a systematic way with the‘single-index’, that is, the market return. The Multi-Group model is a model that applies the Constant Correlation model to each pair of business sectors.

**The ARIMA-LSTM Hybrid Model**

Time series data is assumed to be composed of the linear portion and

the nonlinear portion.

xt = Lt + Nt + εt Lt represents the linearity of data at time t, while Nt signifies nonlinearity. The εvalue is the error term. The Autoregressive Integrated Moving Average (ARIMA) model is one of the traditional statistical models for time series prediction. The model is known to perform decently on linear problems. The Long Short-Term Memory (LSTM) model can capture nonlinear trends in the dataset. The two models are consecutively combined to encompass both linear and nonlinear tendencies in the model. The ARIMA model is fundamentally a linear regression model accommodated to track linear tendencies in stationary time series data.

To build an ARIMA model the methodology consists of three iterative steps.

1. Model identification and model selection the type of model,

2. Parameter estimation the model parameters are adjusted to optimize the model.

3. Model Checking residual analysis is performed to better the model.

After stationarity conditions are satisfied, the auto-correlation function (ACF) plot and the partial auto-correlation function (PACF) plot are examined to select the model type. An optimization process utilizing mathematical error metrics such as the Akaike Information Criterion (AIC) is used to estimate the parameters. A model that has a small AIC value is generally considered a better model. They have used the maximum likelihood estimator for the computation. This method tends to be slow, but produces accurate results. If residual analysis concludes that the residual value does not suffice standards, the three steps are iterated until an optimal ARIMA model is attained.

The accuracy of an LSTM neural network for stock price prediction generally tops other non-neural-network models. The RNN is a type of sequential model that performs effectively on time series data. The LSTM cell adopted in this paper comprises four interactive neural networks, each representing the forget gate, input gate, input candidate gate, and the output gate. Depending on the task, further activation functions such as the Softmax or Hyperbolic tangent can be applied.

**Research Methodology ARIMA**

Over here they have used the price data of S&P500 firms. On that they randomly select 150 stocks from the fully imputed price dataset. Using the fully imputed 150 set of price data, they compute the correlation coefficient of each pair of assets with a 100-day time window. Before fitting an ARIMA model, the order of the model must be specified. The ACF plot and the PACF plot aids the decision process. The method they have used to compute the log likelihood function for the AIC metric is the maximum likelihood estimator. After fitting the

ARIMA model, they generate predictions for each 21 time steps to compute the residual value.

**LSTM**

They have used the residual values, derived from the ARIMA model, of the 150 randomly selected S&P500 stocks as input for the LSTM model.The architecture of the model for the task is an RNN neural network that employs 25 LSTM units. The last outputs of 25 LSTM units is merged into a single value with a fully connected layer. Then,the value will be passed through a doubled-hyperbolic tangent activation function to output a single final prediction. The predictive performance on the train dataset will be high, but will be poor on other newly introduced data. To monitor this problem, a separate set of development dataset is used. They train the LSTM model with the train dataset until the predictive performances on the train dataset and development dataset become similar to each other. The dropout method is one of the widely used methods to prevent overfitting.

There are mainly two types of regularization methods, Lasso regularization (L1) and Ridge regularization (L2). These regularizers keep the weight values of each network in the LSTM model from becoming too large. It turns out that not applying any regularization performs better. Another problem is the vanishing/exploding gradient. The solution for that problem is the LSTM cell itself, it is capable of connecting large time intervals without loss of information.

The walk-forward optimization method is used as the evaluation method but this process is computationally expensive. Rather than training a new model for each rolling train-set window, they train a single model with the first window and apply it to three time intervals the development set and the test1/test2 set. They selected an optimal model with the Mean Squared Error (MSE) metric as the cost function of the model.

**Results And Evaluation**

After around 200 epochs, the train dataset’s MSE value and development dataset’s MSE value started to converge. Among the models, they selected the 247th epoch’s model. The epoch was decided based on both the overfitting metric and the performance metric. With

the selected ARIMA-LSTM hybrid model, the MSE, RMSE and MAE values of the prediction were calculated. Then, the metric values were compared with that of other financial models. Among the financial models, the Constant Correlation model performed the best on our 150 S&P500 stocks’ dataset. The ARIMALSTM’s MSE value was nearly two thirds of that of other equivalent models. The MAE metric also showed clear out-performance. They tested the final model on different assets in the S&P500 firms. Though there are some variations in the results compared to the Test1 & 2, which may be due to a relatively small sample size, and the outstanding performance of the model makes it negligible which suggest that ARIMA-LSTM model is robust.

**Conclusion**

The purpose of empirical study was to propose a model that performs superior to extant financial correlation coefficient predictive models. They adopted the ARIMA-LSTM hybrid model in an attempt to first filter out linearity in the ARIMA modeling step, then predict

non-linear tendencies in the LSTM recurrent neural network. The testing results showed that the ARIMA-LSTM hybrid model performs far superior to other equivalent financial models. Thus, the ARIMA-LSTM model is considered as a correlation coefficient predictor for

portfolio optimization.