Department of Computer Science and Electronic Engineering

CF969-7-SU-CO –Big Data for Computational Finance

Assignment

Report on:

"Deep Learning for Stock Market Prediction" by M. Nabipour, P. Nayyeri, H. Jabani, A. Mosavi, E. Salwana, and Shahab S.

By:

NATARAJA PILLAI RENJITH

Id : 2111151

Supervised by:

Dr. Panagiotis Kanellopoulos

Date: (29/07/2022)

What is the paper about (i.e., what is the topic)?

Stock market price prediction is one of the challenging and classical problems in machine learning. For the last couple of years several machine learning tools and techniques including deep learning approaches were used to build an effective prediction model. The study aims to build a machine learning model that can predict the future stock market price based on the previous data collected from the opening, close, low high, and prices of different groups. The data used here is based on the Iranian stock market and is obtained from the repository of the Tehran Securities Exchange Technology Management Co (TSETMC). The data collected is based on the last 10 years of the historical records and for their convenience four generous groups which are Diversified Financials, Petroleum, Non-metallic minerals, and Basic metals were chosen for the study.

The dataset used had 10 features which are given below:

Simple n-day moving average = $\frac{C_t+C_{t-1}+\ldots+C_{t-n+1}}{n}$
Weighted 14-day moving average = $\frac{n*C_t+(n-1)*C_{t-1}+\ldots+C_{t-n+1}}{n+(n-1)+\ldots+1}$
Momentum = $C_t - C_{t-n+1}$
Stochastic K% = $\frac{C_t-LL_{t\_t-n+1}}{HH_{t\_t-n+1}-LL_{t\_t-n+1}} \times 100$
Stochastic D% = $\frac{K_t+K_{t-1}+\ldots+K_{t-n+1}}{n} \times 100$
Relative strength index (RSI) = $100 - \frac{100}{1+\frac{\sum_{i=1}^{n-1} UP_{t-i}}{\sum_{i=1}^{n-1} DW_{t-i}}}$
Signal(n)$_t$ = MACD$_t$ × $\frac{2}{n+1}$ + Signal(n)$_{t-1}$ × (1 − $\frac{2}{n+1}$)
Larry William's R% = $\frac{HH_{t\_t-n+1}-C_t}{HH_{t\_t-n+1}-LL_{t\_t-n+1}} \times 100$
Accumulation/Distribution (A/D) oscillator: $\frac{H_t-C_t}{H_t-L_t}$
CCI (Commodity channel index) = $\frac{M_t-SM_t}{0.015D_t}$

Table:1 .Selected features

Different machine learning algorithms including the ensemble models were used for prediction. The ensemble models used here are random forest, Adaboost, Gradient boost and XGBoost. Also, the deep learning techniques ANN, RNN and LSTM were implemented for prediction. The predictions are made for 1, 2 ,5, 10, 15, 20 and 30 days in future. Various evaluation methods like Mean Absolute Percentage Error, Mean Absolute Error, Relative Root Mean Square Error and Mean Squared Error were used to measure the performance of the model. Finally, a comparative study is done for the performance of different machine learning approaches based on the evaluation results.

In general the paper can be divided into five parts. The aim and the relevance of the study is described in the first section, followed by the details of all the ML models used for the study along with the reason for selecting them in the second part. The third part is the research data section, were the 10 selected features are described in detail and the evaluation methods and results were shown in the fourth and fifth sections respectively.

• How do the authors approach the problem? I.e., what is the method they use?

Dataset and Pre-processing:

The dataset is collected from the Iranian Stock Market for the period of ten years from November 2009 to November 2019. It has 10 features (independent variables) which is calculated from the opening, close, low high, and prices of the groups and one target variable which is the price prediction. The features were selected by domain experts and based on previous studies. The details of the features are mentioned in table 2.

SL No	Features		
1	Simple n-day moving average	SMA	Average of prices in a selected range
2	Weighted 14-day moving average	WMA	weighted average of the last n values
3	Momentum	MOM	calculates the speed of the rise or falls in stock prices
4	Stochastic K%	STCK	momentum indicator over a particular period of time
5	Stochastic D%	STCD	measures the relative position of the closing prices in comparison with the amplitude of price oscillations in a certain period
6	Larry William's R%	LWR	evaluates oversold and overbought levels
7	Signal(n)t	MACD	Indicates the relationship between two moving averages of a share's price
8	Accumulation/Distribution (A/D) oscillator	ADO	used to find out the flow of money into or out of stock.
9	Relative strength index (RSI)	RSI	an oscillator moves between 0 to 100
10	Commodity channel index	CCI	measures the difference between the historical average price and the current price

Table:2

After obtaining the dataset the next step was data pre-processing. This is the process of identifying irrelevant information's and replacing or deleting it, checking for any missing values/ null values, converting categorical data to numerical data etc. The IQR (Interquartile Range) is used to eliminate the outliers which represents the errors in the measurement or variable which doesn't add any weightage in prediction. The dataset is clean and all the unwanted data were removed. The features that choose has different value range, it is better to standardize it before feeding to the ML model. The normalisation method is used here to scale the all the feature values into a range of 0 to 1. This normalisation is done independently for all the features and for all the groups.

Model Creation and Evaluation:

The dataset is divided into training and testing data in the ratio 70:30, with a 20% cross validation data from the training dataset. The values were normalised prior to the split independently for all the groups. Since it is a supervised learning regression problem the ML learning models used here are : Decision Tree, Random Forest, Adaboost, Gradient Boosting, and XGBoost. The three deep learning models

used were Artificial neural networks (ANN), Recurrent neural network (RNN), and Long short-term memory (LSTM). The same dataset is used for all the models except RNN and LSTM.

For the decision tree regressor all the parameters are set to default values. The hyperparameter Max depth is set to default so that the nodes will expand until all leaves are pure. The mean absolute error criterion is used as squared error function which measures the quality of the split. All the rest ML models are ensembled models, were randomforest regressor is based on bagging and the rest are boosting algorithms. The randomforest method consist of a no of decision trees and the prediction is carried out for all the individual models and majority voting is done for final output. The no of trees used were in the range of 50 to 500 and the maximum depth is set to 10, with a learning rate of 0.1. The rest parameters were set to default values.

The boosting Algorithms minimize the training error by combining a set of week learners into a strong learner and the mistakes from the past performance is corrected by sequentially training the week learners. The boosting algorithm used here are: AdaBoost, Gradient Boosting, and XGBoost. For all the 3 algorithms the number of trees is in the range of 50 t0 500 and the learning rate and depth is 0.1 and 10 respectively. Since the large no of tree depth may favour the chances for over fitting and pruning should be used to avoid that, which is unfortunately not used in this study. For the deep learning models LSTM and RNN the dataset is arranged in such a way that it will contain the features of more than a day. The parameters that varied for these models are no of epochs and the no of days, which they changed proportionally for training the network. For all the models the optimisation function used is adam with the value of beta1 and beta 2 as 0.9 and 0.999. The value of epochs with respect to the no of days is set in the range of 50 to 1000 and the learning rate of ANN, RNN, LSTM are 0.01, 0.0001 and 0.0005 respectively. The details of the parameter values are described in the table [3]

	Number of Neurons	**Activation Function**	**Optimizer**	**Learning Rate**	**Epochs**
ANN	500	Relu	Adam ($\beta1 = 0.9$, $\beta2 = 0.999$)	0.01	100, 200, 500, 1000
RNN	500	tanh	Adam ($\beta1 = 0.9$, $\beta2 = 0.999$)	0.0001	100, 200, 300, 500, 800, 1000
LSTM	200	tanh	Adam ($\beta1 = 0.9$, $\beta2 = 0.999$)	0.0005	50, 50, 70, 100, 200, 300

Table [3]

After training the model on the training dataset, the performance of the models were measured using the evaluation metrics like Mean Absolute Percentage Error (MAPE), Mean Absolute Error, Relative Root Mean Square Error, and Mean Squared Error. The values of these metrics were calculated separately for n number days, since the parameters values are changed with no of days.

• What are the results?

In this study both the tree based models and deep learning methods were able to do the predictions efficiently. It is noticeable that a no of experiments were conducted for all the groups and models using different parameters which helped to get a more generalised results. When the value for no of days increases the error value seems to be increasing correspondingly. That is, it is more accurate to predict the prices for a short period of time rather than a longer one. For example in random the MAE values is in the order of 15.51, 18.39, 24.46, 31.36, 37.28 and 42.66 for the days 1,2,5,10, 15, 20 and 30 respectively. Even though the training process and choosing the right parameter values are important, the role of the dataset in prediction is greater in this case. The data that fed into the model must be clean and the input features must be selected based on a thorough study of past experience.

The study was successful in creating a good prediction model. While comparing the model LSTM outperformed all the other models with a lower MAE values of 4.46, 5.21, 6.02, 6.84, 7.06, 7.25 and 10.03. But still all of them shows a noticeable performance in regression problems. The main disadvantage of the LSTM was the high run time of 80.902 ms per sample. Some of the other noticeable results from the study is listed below.

- The ensemble and deep learning models performed well compared to the DT
- The lowest average value of MAE is for diversified financial group with 6.7 and highest is 1653.79 for Petroleum.
- The three boosting Algorithms shows a similar performance for all groups, except for diversified financials were Adaboost outperformed other two.
- In deep learning models were able to do the prediction more effectively with lower error values.
- ANN showed the least performance compared to LSTM and RNN
- While analysing the RRMSE and MSE values DL models had the best fitting curve.

• What are strong points in the paper, in your view?

- A very well written introduction, with a strong description of the relevance of the study in the current market.
- Instead of taking a single stock group, the authors choose four different groups which gave us a generalised overview of price prediction in different market conditions.
- The approaches used in the study is remarkable, Normal ML algorithms to Latest Deep learning models, they authors tried multiple methods in the study.
- Almost all the popular evaluation methods were used to measure the model performance, that helped to standardise the overall results.
- A good detailing of the dataset features and the relevance of each feature in prediction.
- The independent scaling of the features helped to improve the model performance and thereby better prediction.
- The prediction is done for 1, 2 5,10, 15, 20 and 30 days in advance which helped understand the trend of model performance with future times.
- The results are presented very well in tables, which is very easy to interpret.

• What are weak points in the paper, in your view?

- More data analysis tools like EDA, PCA etc can be used to understand the feature importance and its trends.
- It should be better to present the graphical representation of the results, which is more is to compare the model performance.
- Even though there is a decent detailing of the results, the authors came up with only a few suggestions. It can be more elaborated.
- There was discussion about previous studies done based on the topic but should add more details as generic introduction.
- Instead of using many models they could have use a couple of models with different parameters and hyper parameter tuning is not clearly stated in the paper.
- There was only a short description about the data pre-processing. They can be elaborated based on the different approaches they have used to clean the data.
- The issues they faced during the studies and the methods employed to overcome is not mentioned in the paper. It will really help in future studies if they included it in the paper.