Stock Market Prediction Using Ensemble of Graph Theory, Machine Learning and Deep Learning Models

Pratik Patil
Computer Science Department,
San Jose state University
San Jose, USA
pratik.patil@sjsu.edu

Ching-Seh (Mike) Wu
Computer Science Department
San Jose state University
San Jose, USA
ching-seh.wu@sjsu.edu

Katerina Potika
Computer Science Department
San Jose State University
San Jose, USA
katerina.potika@sjsu.edu

Marjan Orang
Economics Department
San Jose State University
San Jose, USA
marjan.orang@sjsu.edu

ABSTRACT

Efficient Market Hypothesis (EMH) is the cornerstone of the modern financial theory and it states that it is impossible to predict the price of any stock using any trend, fundamental or technical analysis. Stock trading is one of the most important activities in the world of finance. Stock price prediction has been an age-old problem and many researchers from academia and business have tried to solve it using many techniques ranging from basic statistics to machine learning using relevant information such as news sentiment and historical prices. Even though some studies claim to get prediction accuracy higher than a random guess, they consider nothing but a proper selection of stocks and time interval in the experiments. In this paper, a novel approach is proposed using graph theory. This approach leverages Spatio-temporal relationship information between different stocks by modeling the stock market as a complex network. This graph-based approach is used along with two techniques to create two hybrid models. Two different types of graphs are constructed, one from the correlation of the historical stock prices and the other is a causation-based graph constructed from the financial news mention of that stock over a period. The first hybrid model leverages deep learning convolutional neural networks and the second model leverages a traditional machine learning approach. These models are compared along with other statistical models and the advantages and disadvantages of graph-based models are discussed. Our experiments conclude that both graph-based approaches perform better than the traditional approaches since they leverage structural information while building the prediction model.

CCS Concepts

- **Information systems → Spatial-temporal systems**

ICSIM '20, January 12–15, 2020, Sydney, NSW, Australia
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7690-7/20/01…$15.00
DOI: https://doi.org/10.1145/3378936.3378972

Keywords

Big data analytics; Stock market; machine learning; deep learning; graph theory; financial networks; time series forecasting; spatio-temporal

1. INTRODUCTION

A stock is a share or ownership of a part of a publicly listed company. These shares are issued by the company to exchange for traders to trade. These stocks are sold by the owners of the company to raise money/funding for further development of the company. When the company is first listed on an exchange, it is called an Initial Public Offering where the initial selling price of that stock is set by the owners. The price of the stock after the initial public offering is decided by the equilibrium of buy and sell orders, which can also be thought of as demand and supply equilibrium.

It is easy to understand the demand and supply is the root cause of price change, however, demand and supply are based on several variables and factors like inflation, positive or negative news, market sentiment, socio-economic factors, trends, and many more.

Since the beginning of the stock market, the goal of the speculators/investors has been to predict the price of the stock as to buy low and sell high, thus earning a profit. For the purpose of this paper, stock prices have been assumed to be a time series of equal intervals and different models have been proposed to forecast time series.

A time series has huge significance in Econometrics, Social Science, Healthcare, Cyber Security and can be defined as a sequence of observations collected over regular time intervals. A time series describes the behavior and change in characteristics of a process over time. If we can understand the process with the help of statistics, its description or graphical representation, then we can model it and use it to forecast the future behavior of the process.

For this paper, we will limit our scope of study to time series in econometrics, especially company stock prices.

Since the price of the stock is dependent on a huge number of factors like socio-ecological and sentimental factors, there is a huge amount of noise in the time series which makes it extremely difficult to model using any of the statistical methods.

According to the hypothesis in random walk theory, the prices of a stock market are defined randomly and therefore are impossible to forecast. However, there have been extensive advances in the statistical modeling, deep learning methods, and the availability of huge amount of news and sentimental data. These factors have increased the probability of predicting the stock prices than that of a random process.

There are 3 basic approaches which has been used traditionally to predict stock price: technical analysis, fundamentals of a company, sentiments of the market.

The objective of this research is to explore different useful features which can be extracted from graph structural behavior to use in machine learning prediction engine. The paper aims to combine graph theory and machine learning by creating a hybrid model for prediction. The aim is to create a generic framework for any time series prediction and not just related to stock prediction.

2. RELATED WORK

2.1 Machine Learning

Machine learning has been used widely for forecasting stock price. The focus of majority of the studies was to predict change in price for near-term (less than a minute), short-term (1 day ahead or more) and long term (a few months ahead).

Most studies have considered a subset of stocks limited to less than 10 for their study. The set of predictor variables used in the study ranges from simple time series data of stock, google trend, news sentiment data, to characteristics of the company. Based on the existing work, most of the researchers focus on the near-term predication [1] and long-term [2].

Most of the studies use multiple predictor variables derived from the time series data. These variables are called as indicator in financial terms. Supervised Machine Learning algorithms have been used along with the above indicators as features. In addition to indicators some studies use news data as input features to the model. Most of the studies have compared different machine learning algorithm using the same feature set [3]. The study by Alice et. al. compares logistic regression, support vector machine (SVM), and bayesian network and the conclude that SVM are better for smaller time series data [3]. Kim Kj et. al. have compared SVM and artificial neural network (ANN) and they have concluded that SVM are superior for stock price prediction [4]. Study by Huang W. et. al. treated prediction as classification problem and compared linear discriminant analysis, elman back propagation, SVM and quadratic discriminant [5]. Studies have been done on the usage of google trend and number of google searches for a specific stock as input variables to the model. [6-10]. They conclude that the increase in number of searches implies that the prices will increase within two weeks and will go down within the year. The study by H.J. Kim et. al. used SVM and ANN using google trends, and indicators using in [11] and tested three hypothesis that proved that google trends, and the state machine learning algorithms does not perform well in plausible framework for market investment [12].

2.2 Graph Theory

It has been proposed by researchers that there exists an underlying structure in the stock market that can be used to understand the behavior of stock markets. In [1], Susan George and Manoj Changat asserted that such underlying structure can be used by governments to prevent a financial crisis such as the one experienced in 2008. They constructed a graph based on the stock market trend correlations and simply calculated various statistics such as the degrees of the nodes. They argue that by simply looking at the resulting graph, various insights can be derived. For example, they discovered that many banks have high degrees, and argued that nodes with a high number of edges are crucial in preventing failures in the stock market. Additionally, these authors argued that the topology of the resulting graphs also describes the market. For example, based on their analysis, a chain-like topology means that there are no powerful companies with high market capitalization, whereas a star-like topology indicates the existence of a few very powerful corporations that have a strong influence over the market. The future work proposed by [1] was to apply community detection algorithms to the graphs constructed from the stock market data.

A very few researchers have tried to model the stock market problem as a graph problem. However, no researchers have significant results so far. Since the stock market does not have inherent or obvious graph structure, it becomes difficult to create a graph representation of stock market.

Researchers have used correlation analysis to construct a graph of stock market, where nodes in the graph are stocks and edges represent the price fluctuation relationship between those stocks [10]. Ghazale et. al. have developed a trend detection algorithm for stock price using graphs [13, 14]. They are the first to create a graph where every node in a graph is a state transition and relationship is defined based on the change in trend of stocks. They also explore community detection on stock market graphs. Communities in a graph exhibit an interesting characteristic that groups nodes with similar properties and behavior together. Louvain community detection algorithm was used to create this trend detection algorithm in graph [15, 16].

2.3 Deep Learning

Recently a lot of work has been done using deep learning techniques such as recurrent neural networks. Especially long short-term memory (LSTM) are excellent at modelling sequential time series data. De Mello Assis et. al. has trained 30 models and studied the performance of ensemble model for predicting stock price on Brazilian index [17]. Using hypothesis testing they proved that top 5 ANN had an accuracy above 50% [17]. Zhang et. al. have used multiple source as input features, especially financial news and articles [18]. They show that investor opinions through news and social media has significant effect on the market volatility. Sentiment analysis was done on financial news and fed as an input to the model. They concluded that accurate consistent news information significantly increases the accuracy of the model. Jiahong li et al have used LSTM with news sentiment analysis and achieved an accuracy of 88%.

Recently an emerging branch of deep learning known as reinforcement learning (RL) have showed promising opportunities in the field of stock market prediction. Reinforcement learning agent is based on a certain set of actions and a goal which in this case is to maximize the profit [19]. A deep recurrent Q-learning was employed in this study and the results of the experiment were positive profit. This is the first positive results by a pure deep reinforcement learning algorithm under transactional costs and therefore RL provides promising opportunities for researchers in this field.

3. DATASET

The raw data used in this paper is the time series data, that is the stock price data collected at an equal interval of time. This stock prices data was collected for 30 stocks from the top 30 fortune 500 companies listed below in Table 1. This dataset was collected for two different intervals, 1-day interval, and 1 min interval. Another dataset that is used in this paper is news dataset. This data contains 1 million financial news collected from a financial news website.

Table 1: List of 30 stocks used in the paper

Ticker	Company Name	Sector
WMT	Walmart	Consumer
XOM	Exxon Mobil Corporation	Energy
AAPL	Apple Inc.	Technology
UNH	UnitedHealth Group Incorporated (DE)	Healthcare
MCK	McKesson Corporation	Healthcare
CVS	CVS Health Corporation	Healthcare
AMZN	Amazon.com Inc	Technology
T	AT&T Inc	Communication Services
GM	General Motors Company	Auto Mobile
F	Ford Motor Company	Auto Mobile
ABC	AmerisourceBergen Corporation	Healthcare
CVX	Chevron Corporation	Energy
CAH	Cardinal Health Inc	Healthcare
COST	Costco Wholesale Corporation	Consumer Defensive
VZ	Verizon Communications Inc	Communication Services
KR	Kroger Company	Consumer
GE	General Electric Company	Industrials
WBA	Walgreens Boots Alliance Inc	Healthcare
JPM	JP Morgan Chase & Co	Financial Services
GOOGL	Alphabet Inc	Technology
HD	Home Depot Inc	Consumer
BAC	Bank of America Corporation	Financial Services
WFC	Wells Fargo & Company	Financial Services
BA	The Boeing Company	Industrials
PSX	Phillips 66	Energy
ANTM	Anthem Inc	Healthcare
MSFT	Microsoft Corporation	Technology
UNP	Union Pacific Corporation	Industrials
PCAR	PACCAR Inc	Industrials
DWDP	DowDuPont Inc	Basic Materials

3.1 Data Collection

- *1-Day interval dataset:*

Daily closing prices of these 30 stocks were collected by scrapping the website https://finance.yahoo.com. Two years of data was collected starting from 2017-05-01 to 2019-04-01. Fix_yahoo_finance was used in scrapping the daily closing price of each stock. This data was only used in creating the graph of companies and calculating the correlation matrix because it has more variance than minute level data and thus is able to model richer information. It was not used for training and predicting the GCN model.

- *1-minute interval dataset.*

Minute level data of stock prices are not publicly available since it requires substantial storage. However, there are specific financial data companies which store this data. This data is exposed to data scientists using a restful API for a small amount of fee.

For the purpose of this paper, minute-level data for above 30 stocks were collected form the website https://api.tiingo.com. This data is later used in graph convolution networks (GCN) and linear machine learning (ML) models for training and forecasting.

The stock market is open for trading for 6.5 hours daily from Monday to Friday, which is 390 minutes daily. Therefore, everyday contained 390 data points. Data was collected for 44 days for all 30 stocks. Finally, every time series (stock) had 17,204 data points.

- *News dataset*

1 million financial news were collected from a financial news website. The data was collected since the year 2007. The dataset lists the news and the companies mentioned in that particular piece of news. This data was later used to generate the causality-based graph.

4. MODELLING STOCKS INTO GRAPHS

The goal of this paper is to model the stock price prediction problem into a graph problem. Here, to create a stock graph, we can assume that the stocks are nodes of a graph. However, the most important question is how to define the relationship between two nodes/stocks. In this paper, the relationship between any two nodes is defined in two ways:

- *Correlation based relationship.*
- *Causation based relationship*

4.1 Correlation Based relationship

The correlation coefficient is a numerical measure of the statistical relationship between two variables. This number is between -1 and 1 representing linear dependence between two variables. The correlation coefficient is calculated between two columns of data. If both variables rise and fall together then the correlation coefficient is positive with the higher magnitude and vice versa.

To define the relationship between two nodes. We calculate a correlation coefficient between two stocks and if it is greater than a certain threshold then we create an edge between those two nodes.

To make this easier, correlation matrix is calculated and if the absolute value of correlation coefficient of stock I, j >threshold (0.5), then we create an edge between stock I and j by putting 1 at location matrix[I,j] other we put 0. Eventually, we get an adjacency matrix which can be interpreted as a graph. The graph structure changes as the threshold vary from 0-1.

In this paper, we have used the following 3 correlation coefficient to generate the graph.

- *Pearson correlation coefficient*
- *Spearman correlation coefficient*
- *Kendall correlation coefficient*

All the 3 correlation coefficients are calculated on the top 30 stocks from fortune 500 mentioned above. Every stock price is a daily closing price with 1-day interval.

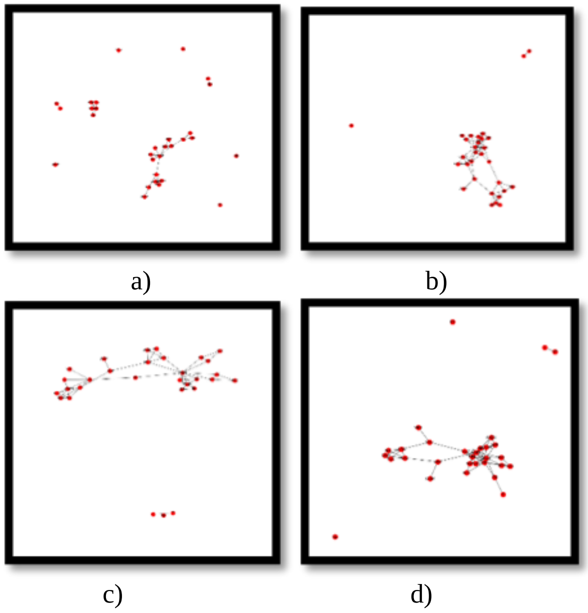


Figure 1: a) Graph from pearson coef. with threshold 0.5
b) Graph constructed from spearman coef. with threshold 0.4
c) Graph from Kendall tau coef. with threshold 0.3
d) Graph from news co-mentions (Causation)

4.2 Causation based relationship

Financial news was used in generating another type of stock graph. In this graph, the relationship between two nodes/stock is based on co-mention of those two stocks in the same news article. The assumption that two stocks mentioned in the same news must be somehow related to each other was used here.

Number Therefore, if a piece of news mentions Amazon, Microsoft, and Apple in the same article, then we create an edge between each of these 3 nodes/stocks with an edge weight of 1. If another news mentions Apple and Microsoft in the article, and if there is already an edge between these nodes then we increase the edge weight by 1 to represent a stronger correlation. This, when done over a dataset of 1 million news, would remove any outliers or wrong mention of any two companies in the same news article.

5. APPROACHES AND IMPLEMENTATION

Time series forecasting has traditionally been done using statistical algorithms like auto regressive integrated moving average (ARIMA), holtwinters and even using LSTMs and recurrent neural unites recently. However, each model needs to be built for every stock/time series and thus all these models are independent of each other. They do not leverage the Spatio-temporal relationship between different companies.

In this paper two novel models are proposed based on the graph which leverage the Spatio-temporal relationship between different stocks/companies. The first model is based on deep learning convolutional neural network and the other model is based on a traditional machine learning algorithm.

5.1 Graph Based Deep Learning Models

In this model, we create a graph convolution neural network (GCN), in which we construct convolution layers based on graph and its structure instead of using regular recurrent units and convolutions. In this model, we create multiple blocks of spatio-temporal convolution which are comprised of graph convolution layers. We extract spatio-temporal features from the graph of many time series (stock prices) using convolution structures.

Figure 2 describes a general architecture for a Spatio-temporal graph convolutional neural network. The adjacency matrix A is passed to the architecture along with the feature matrix X. In this case the feature matrix is a vector containing time series of the corresponding node. Then there is a GCN layer which calculates the spatial dependency from input A and the next 1-D CNN layers [20] calculates the temporal dependency from the feature vector X. There is a Multi layered perceptron (MLP) layer after this which does a linear transformation to predict at each node/stock [20].

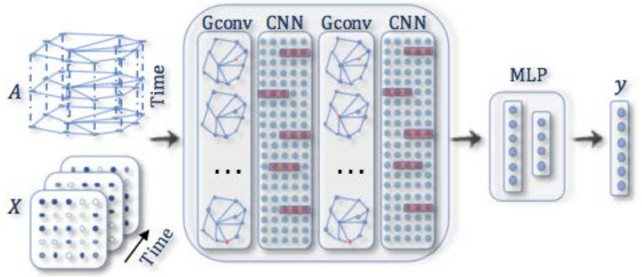


Figure 2: Graph Convolution Neural Network Architecture

There are two inputs to this model:

1. Adjacency matrix of the graph (Weight matrix)
2. Time series (stock prices) of every node in the graph.

The adjacency matrix is calculated using two ways, correlation coefficient and using news co-mentions. For the time series, the interval is set to 1 minute. In the US, the stock market is open for 6.5 hours, which makes the time series to have 390 data points per day. The linear interpolation is used to fill NA or missing values. Normalization is done to scale all the time series within the range 0 and 1. Prediction is done for 3 ,6 and 9-time steps ahead [20]. Since the data is normalized, we calculate the merged Root mean squared error (RMSE), Mean absolute percentage error (MAPE), and mean absolute error (MAE) for model evaluation.

5.2 Graph Based Traditional Machine Learning Models

In this model, we create a linear regression model for predicting stock prices. A separate model is needed for predicting every stock. To build a model for a stock we extract the features from that stock's time series. We also derive features from time series of other stocks. The additional stocks which are used in the feature set are determined from the communities in the graph structure.

5.2.1 Community Detection:

One of the most relevant and useful features of a graph structure is its clustering of nodes or community. A community structure within a graph is defined as a set of nodes which can be grouped together and whose nodes are densely connected internally. In real world graphs, nodes within a community can be considered to be related in similar manner.

In case of stock graph, communities within this graph should represent a sector or specific industry like healthcare, finance, technology etc. This is based on the assumption that stock prices of companies within same sector would rise and fall together because of similar opportunities and conditions for them. Therefore, community detection algorithm was used to select appropriate threshold value in this paper.

In this paper, Louvain community detection algorithm was used to create communities. Louvain is a modularity-based algorithm. Empirically a modularity value greater than 0.4 represents good community structure in the graph.

To predict a specific stock, a model is built by training on features extracted from other stocks in the same community.

22 features are extracted from the time series using statistical formulas. These features are called indicator in technical analysis of stock market. Below is the list of all the technical indicators extracted in this paper.

Table 2: List of all the statistical indicator(features) extracted from time series

Sr. No.	Indicator Name
1	Bollinger High Band Indicator
2	Relative Strength Index (RSI)
3	True strength index (TSI)
4	Bollinger Bands (BB)
5	Bollinger Low Band Indicator
6	Bollinger Bands (BB)
7	Donchian channel (DC)
8	Donchian High Band Indicator
9	Donchian channel (DC)
10	Donchian Low Band Indicator
11	Aroon Indicator (AI)
12	Aroon Indicator (AI)
13	Detrended Price Oscillator (DPO)
14	EMA
15	Moving Average Convergence Divergence (MACD)
16	Moving Average Convergence Divergence (MACD Diff)
17	Moving Average Convergence Divergence (MACD Signal)
18	Trix (TRIX)
19	Cumulative Return (CR)
20	Daily Log Return (DLR)
21	Daily Return (DR)
22	Bollinger Bands (BB)

To use the linear regression model as forecasting machine, and to formulate it as a supervised ML problem, we shifted the label, that is vector Y by an offset of h, where h is the number of time steps we need to forecast in future.

- Building Linear Models

In this part, two different linear models are built for every stock, each with different input feature matrix. We call it single model and a composite model. The single model derives its features from its own time series. The other composite model derives its features from all the neighboring stocks in the same community of the graph. The communities derived using the Louvain algorithm is used in this part. Below in figure 3, we have a list of stocks in each community. For example, we need to build both single and composite model for the stock AAPL (Apple. Inc.).

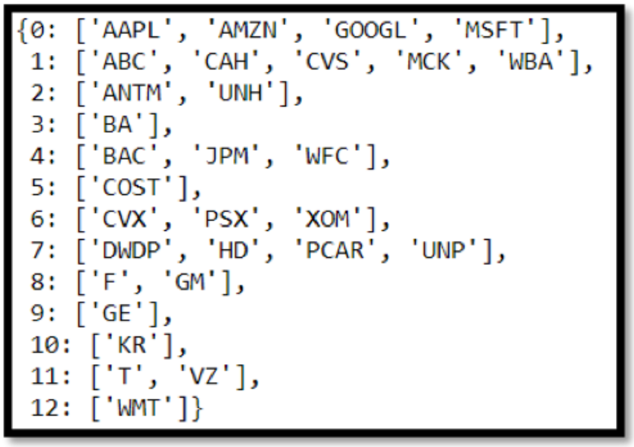
a) Single model:

22 features mentioned in table 2 are extracted from time series of AAPL for training and testing. The label will be offset according to the number of time step of forward prediction.

b) Composite model:

22 features mentioned in the table are extracted from all of the stocks in the community 0. Since there are 4 stocks in community 0, the feature matrix will have 22x4, that is 88 features for the time series AAPL.

We build these two models to show that the community-based graph approach has more information and would eventually perform better in the long run. The figure 3 shows the 12 communities derived from the stock graph



```
{0: ['AAPL', 'AMZN', 'GOOGL', 'MSFT'],
 1: ['ABC', 'CAH', 'CVS', 'MCK', 'WBA'],
 2: ['ANTM', 'UNH'],
 3: ['BA'],
 4: ['BAC', 'JPM', 'WFC'],
 5: ['COST'],
 6: ['CVX', 'PSX', 'XOM'],
 7: ['DWDP', 'HD', 'PCAR', 'UNP'],
 8: ['F', 'GM'],
 9: ['GE'],
 10: ['KR'],
 11: ['T', 'VZ'],
 12: ['WMT']}
```

Figure 3: Stocks aggregated based on communities

5.3 Statistical Model

In this study we have also implemented a traditional statistical model, auto regressive integrated moving average (ARIMA) which is very robust and generic for any type of time series forecasting. ARIMA is an acronym which breaks down as:

AR: Auto-regression. This models the relationship between the lagged parameters of the time series as variables.
I: Integrated. Differencing is integration is done to remove the stationarity in the time series. A time series is non-stationary if it has unit root. A stationary time series has a mean of 0 and std. deviation of 1.
MA: Moving average. This models the residual error terms as lagged parameters from moving average model.
Thus, ARIMA has 3 parameters, p, d and q.

- p: Number of lagged terms considered in building model.
- d: Number of times the consecutive terms are differenced to make time series stationary.
- q: This is the length of the moving average window.

If the value of either of p,d,or q is 0, then that component is not considered in building the model. Therefore, depending on the parameter value, we can build ARMA, AR, MA, I, or ARIMA model to fit the given time series.

6. EXPERIMENTS

6.1 Metric

In this paper we have built multiple models for each stock, and there are total 30 stocks in the graph approach. Therefore, in the linear models, 30 models need to be built as well, 30 single models and another 30 for composite models. Evaluating, visualizing and comparing the results of these many models in difficult. Therefore, all the metrics used in this paper have been merged to a single value for every time step forward prediction.

Since, all the stock prices are at different scale, the absolute error will be at different scales as well. To solve this problem, all the time series have been normalized, making all the stock prices in the range of 0 and 1. This makes the merged results from different models comparable.

- *Root mean squared error (RMSE)*
- *Mean absolute percentage error (MAPE)*
- *Mean absolute error. (MAE)*

6.2 Graph Convolutional Network

For GCN, the first input is the graph, G, represented as an adjacency matrix, and another input is the temporal time series at every node. Since there are 30 nodes in the graph, 30-time series were input to the model. Every time series had 17204 data points in total. Time series data cannot be split with percentage for training and testing as it is tightly bound to the days/time. Therefore, 34 days of data was used for training and 10 days of data was used for testing. This is roughly 75% - 25% train-test split. We train the model for 10 epochs with learning rate of 0.001. We use the optimizer function RMSProp for the gradient descent.

We build 4 GCN models, 1 for each of the input graph. There are 3 graphs generated from Pearson correlation, Spearman correlation, and Kendall correlation [21], each with an absolute threshold of 0.5. The 0.5 threshold is empirically used in other graph research and we have verified and supported this hypothesis by doing community detection using various threshold and studying the community features like sector and industry of every stock. For the graph generated using spearman rank coefficient, the threshold used is 0.4 and for Kendall Tau it is 0.3. This is done because Spearman rank and Kendall Tau model non-linear relationships and therefore the scale of the coefficient reduces as compared to the Pearson correlation, which models linear relationship. Figures in Figure 1 are the figures of the graph given as input. Another graph is generated from the new dataset. This graph is derived from causal relationship rather than correlation relationship. Therefore, we believe this graph holds huge amount of unknown hidden information than the correlation-based graphs. And this hidden information can be interpreted by GCNs. The Figures 1 represent the graphs used in the 4 GCN models.

The inputs to the model are 2 csv files. 1 file for weighted adjacency matrix of the graph and other for time series of every node. The model outputs the merged metrics RMSE, MAE, MAPE by combining the results of all the normalized time series in the graph. For RMSE the values were added, For MAPE the values were averaged, and for MAE the values were added.

6.3 Graph Based Linear Models

In this experiment, we build linear regression model from sklearn machine learning library. We have two different types of model for every stock:

1) Single model
2) Composite model

Therefore, instead of training 60 different models, we have selected 1 representative stock from each community, and built model for that stock. The representative stock is the one with the highest degree in the graph community.

The input time series is normalized to bring all the series to similar scale of 0-1. The input to the single model is a 22-column feature matrix for every data point collected per min for 30 days. The output vector is a 1-dimensional array which stores the forecasted stock price for the nth future time step. Here, n is the offset in the label vector y.

Similarly, for the composite model, we have 22* n number of features in the input matrix where n is the number of stocks in the community. The output vector is a similar 1-d vector with the future predicted values, n time steps ahead.

In the end we merge all the metrics RMSE, MAPE and MAE to get a single representative metric of the linear model.

6.4 ARIMA

In this part we created an ARIMA model for each of the 30 stocks. The parameter selection is the most difficult and time-consuming part of the model building. To compare multiple ARIMA models with different input parameters (p,d,q), Akaike information criterion (AIC) and Bayesian information criterion (BIC) is used. AIC and BIC explain how well the model with given parameters fits the time series.

In our paper we used a grid search for finding the optimal value of p, d, and q parameters. The values of p and q ranged from 0 to 5 and d ranged from 0-2. Therefore, for every stock, 6*3*6 = 108 models were built. The model with the lowest value of AIC and BIC was finalized for that stock. This process was repeated for each of the 30 stocks.

In the end, like other approaches in this paper, the results using metrics RMSE, MAPE and MAE were merged to get a single representative value for ARIMA model.

7. RESULTS

From the above experiments, we build total 7 models for stock price forecasting. Of these, 4 models are GCN models and two models are graph based linear models. The other model is statistical ARIMA model. All the experiments were performed on top 30 stocks from fortune 500. The results for every type of model were merged to get a representative value for that model. This is done for the sake of simplicity in comparing different models. Below are the results for all the models with 3, 6, and 9 steps forecasting along with the graphical comparison with the bar graphs in Figure 4, 5, and 6.

The best performing model is GCN based on graph constructed from new co-mentions. This is probably because the graph is causation based rather than correlation based. The next GCN model with Kendall tau correlation performs well, probably because of non-linear modelling by Kendall coefficient, the same applies for spearman rank as it models non-linear relationship as well. Clearly, GCN models perform better than the traditional statistical method used for time series forecasting. The composite community based linear model performs better than the single linear model. This supports our initial assumption that graph has rich structural information which can be leveraged not only in time series forecasting but in other problems like classification as well.

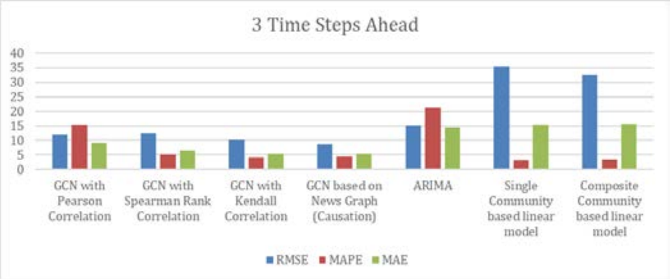


Figure 4: Comparison of all the models using metrics for 3-time step ahead forecasting

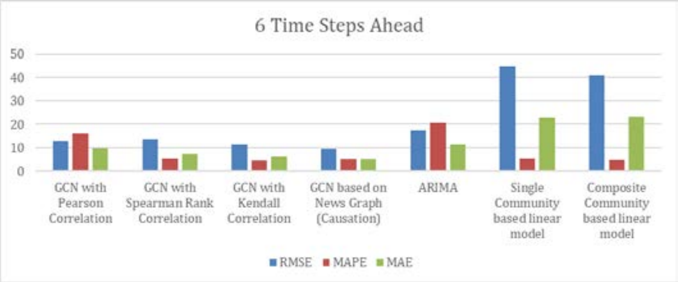


Figure 5: Comparison of all the models using metrics for 6-time step ahead forecasting

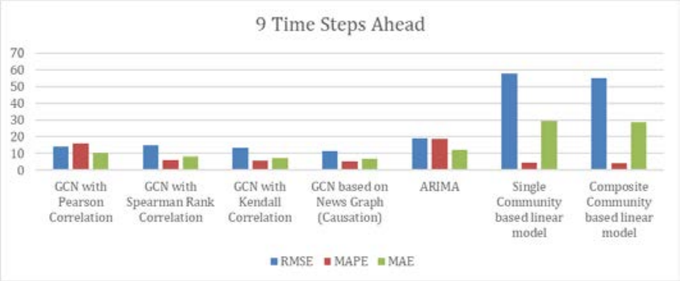


Figure 6: Comparison of all the models using metrics for 9-time step ahead forecasting

8. CONCLUSION AND FUTURE WORK

In this paper we have proposed a novel approach for forecasting stock prices using graphs and leveraging the Spatio-temporal relationship among the companies and its stock prices. In our experiments, overall, the graph-based models perform better than single linear and traditional statistical ARIMA model. We observe that modelling and leveraging graph's spatial structure to temporal features increases the accuracy of forecasting drastically. The experiments prove that graph-based models outperform other traditional and statistical model for stock prediction. In this paper we have shown two ways of generating graph, correlation-based and causation-based graphs. We observe that causal relationship holds more hidden information and can be extremely useful. Owing to the resources limitations we used only 30 nodes in the graph, however, the causal based GCN can give extremely accurate forecasts at larger scale with a bigger graph. Through our experiments, we also support our claim that the graph structure holds richer information which gives better results when leveraged. The graph structure can be leveraged in other machine learning problems like supervised and unsupervised classification as well.

Moreover, this Spatio-temporal GCN model for forecasting is not tied to stocks and can be used in any generic time series prediction if underlying graph representation is available.

This paper considers 30 nodes graph owing to the complexity of the model and its training period. It will be interesting to study the results and performance of a more complex network (graph). The GCN model proposed in this paper is susceptible to exploding gradient problem as nodes with higher degree will have larger value in their convoluted feature representation, whereas nodes with smaller degree will have smaller value in feature representation. A solution to this problem can reduce the complexity of the model training. It will also be interesting to check the performance of GCN on a more generic time series forecasting problems.

9. REFERENCES

[1] Rechenthin, Michael David. "Machine-learning classification techniques for the analysis and prediction of high-frequency stock direction." PhD (Doctor of Philosophy) thesis, University of Iowa, 2014.

[2] Milosevic, Nikola. (2016). Equity forecast: Predicting long term stock price movement using machine learning.

[3] Alice Zheng: Using AI to make prediction on stock market, 2017

[4] Kim Kj. Financial time series forecasting using support vector machines. Neurocomputing. 2003; https://doi.org/10.1016/S0925-2312(03)00372-2

[5] Huang W, Nakamori Y, Wang SY. Forecasting stock market movement direction with support vector machine. Comput Oper Res. 2005; 32(10):2513-2522. https://doi.org/10.1016/j.cor.2004.03.016

[6] Ni LP, Ni ZW, Gao YZ. Stock trend prediction based on fractal feature selection and support vector machine. Expert Syst Appl. 2011; 38(5):5569-5576. https://doi.org/10.1016/j.eswa.2010.10.079

[7] Kumar D, Meghwani SS, Thakur M. Proximal support vector machine based hybrid prediction models for trend forecasting in financial markets. J Comput Sci. 2016; 17:1-13. https://doi.org/10.1016/j.jocs.

[8] Qiu M, Song Y, Akagi F. Application of artificial neural network for the prediction of stock market returns: The case of the Japanese stock market. Chaos Solitons Fractals. 2016; 85:1-7. https://doi.org/10.

[9] Preis, T., Moat, H. S., & Stanley, H. E. (2013). Quantifying Trading Behavior in Financial Markets Using Google Trends. Scientific Reports,3(1). doi:10.1038/srep01684

[10] R. N. Mantegna, "Hierarchical structure in financial markets," The European Physical Journal B—Condensed Matter and Complex Systems, vol. 11, no. 1, pp. 193–197, 1999. View at Google Scholar · View at Scopus.

[11] H.-J. Kim, Y. Lee, B. Kahng, and I.-M. Kim, "Weighted scale-free network in financial correlations," Journal of the Physical Society of Japan, vol. 71, no. 9, pp. 2133–2136, 2002.

[12] H.-J. Kim, I.-M. Kim, Y. Lee, and B. Kahng, "Scale-free network in stock markets," Journal of the Korean Physical Society, vol. 40, no. 6, pp. 1105–1108, 2002. View at Google Scholar · View at Scopus

[13] M. Gałązka, "Characteristics of the Polish stock market correlations," International Review of Financial Analysis, vol. 20, no. 1, pp. 1–5, 2011.

[14] Ghazale, P.P., Zhao, L., Zheng, Q., & Zhang, J. Time Series Trend Detection and Forecasting Using Complex Network Topology Analysis. 2018 International Joint Conference on Neural Networks (IJCNN), 1-7, 2018.

[15] https://en.wikipedia.org/wiki/Louvain_Modularity, 2019

[16] Blondel et. al., "Fast unfolding of communities in large networks," arXiv, 2008. [Available] https://arxiv.org/abs/0803.0476

[17] de Mello Assis, Julia & Pereira, Adriano & Couto e Silva, Rodrigo. Designing Financial Strategies based on Artificial Neural Networks Ensembles for Stock Markets. 1-8. 10.1109/IJCNN.2018.8489688, 2018.

[18] Zhang, Xi & Qu, Siyu & Huang, Jieyun & Fang, Binxing & Yu, Philip. Stock Market Prediction via Multi-Source Multiple Instance Learning. IEEE Access. PP. 1-1. 10.1109/ACCESS.2018.2869735, 2018.

[19] Yi Huang, Chien. (2018). Financial Trading as a Game: A Deep Reinforcement Learning Approach, 2018

[20] https://www.semanticscholar.org/paper/Spatio-Temporal-Graph-Convolutional-Networks%3A-A-for-Yu-Yin/4b1c78cde5ada664f689e38217b4190e53d5bee7/figure/0, 2019

[21] Xu Weichao, Yunhe Hou, Hung Y. S. & Yuexian Zou, 2010. Comparison of Spearman’s rho and Kendall’s tau in normal and contaminated normal models. Manuscript submitted to IEEE Transactions on Information Theory (http://arxiv.org/PS_cache/arxiv/pdf/1011/1011.2009v1.pdf)