Contents

1. Introduction

In order driven markets prices are determined by the publication of orders to buy or sell shares. Examples of order driven markets around the world include Nasdaq OMX, London Stock Exchange, Tokyo Stock Exchange and Australian Securities Exchange. In these markets, participants may submit limit orders or market orders. The aim of the project is to predict stock market's short response following a large trade or a series of small trades.[1]

1.1 Terminology[2]

Limit orders

Limit orders specify the price at which a trader is willing to transact. Orders that do not execute immediately may be stored for later execution in a limit order book.

Market orders

Market orders execute immediately against orders in the limit order book.

An understanding of the dynamic interplay between market and limit orders, and the state of the limit order book is important to traders, exchanges and regulators.

Limit order book

The limit order book represents a pool of trading interest over a range of prices.

Bid and ask

The **bid** and **ask** are *the best potential* prices that buyers and sellers are willing to transact at: the bid for the buying side, and the ask for the selling side.

Bid-ask spread

A bid-ask spread is the amount by which the ask price exceeds the bid. This is essentially the difference in price between the highest price that a buyer is willing to pay for an asset and the lowest price for which a seller is willing to sell it.

Standing buy (sell) orders with the highest bid (lowest ask) price have the highest probability of execution. The difference between the two best prices is the bid-ask spread. Competition between market participants ensures that in equilibrium the size of the spread is small.

Changes to the state of the order book occur in the form of trades and quotes. A quote event occurs whenever the best bid or the ask price is updated. A trade event takes place when shares are bought or sold.

Liquidity

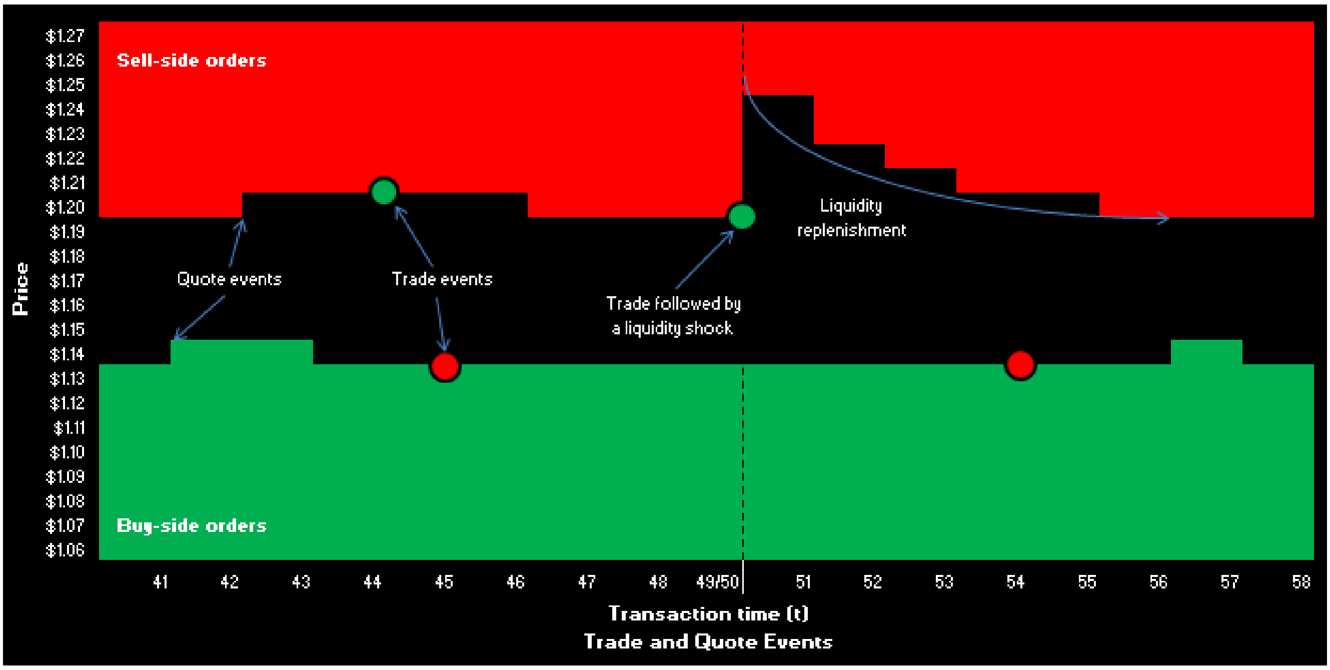
Liquidity is the ability of market participants to trade large amounts of shares at low cost and quickly. Market resiliency (also known as liquidity replenishment) is the time component of liquidity and refers to how a market recovers after liquidity has been consumed. Resiliency is very important to market participants, particularly traders wishing to reduce their market impact costs by splitting large orders across time.

Liquidity Shock

A liquidity shock is defined as any trade that changes the best bid or ask price. Shocks to liquidity may occur when a large trade (or series of smaller trades) consumes all available volume at the best price. Following a liquidity shock the spread may be temporarily widened, and/or result in permanent price shifts (see Figure 1). In time, investors replenish the order book with new orders to buy and sell. Market resiliency is that part of mean reversion where the spread and depth partially or completely revert to former levels following a liquidity shock.

The figure depicts a limit order book in event time. The event types are either 'quote', whereby the best bid or the ask price is updated, or 'trade', whereby shares are transacted. The red (green) coloured section represents standing orders to sell (buy). A liquidity shock is defined as a trade that changes the best bid or ask. Liquidity replenishment is the stock price mean reversion following a shock.

Figure 1 Liquidity Replenishment after a Shock[1]



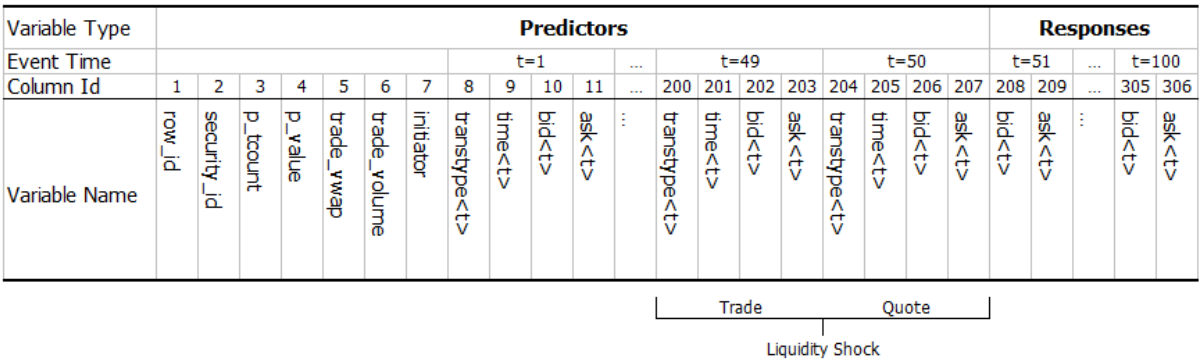
1.2 Objective

To determine the bid ask spread with a quantifiable amount of accuracy, measured in terms of reduction in mean square error of the same.

1.3 Data Description

Recent trade and quote data from the London Stock Exchange (LSE) is provided. The pre-processed dataset comprises observations of the limit order book before and after a liquidity shock (a trade that results in widening of the bid-ask spread).[1]

Table 1 Data Schema



Variable Type	Predictors																				Responses				
Event Time								t=1				...	t=49				t=50				t=51		...	t=100	
Column Id	1	2	3	4	5	6	7	8	9	10	11	...	200	201	202	203	204	205	206	207	208	209	...	305	306
Variable Name	row_id	security_id	p_tcount	p_value	trade_vwap	trade_volume	initiator	transtype<t>	time<t>	bid<t>	ask<t>	⋮	transtype<t>	time<t>	bid<t>	ask<t>	transtype<t>	time<t>	bid<t>	ask<t>	bid<t>	ask<t>	⋮	bid<t>	ask<t>

Trade
Quote
Liquidity Shock

Table 2 Sample data

Variable Name	Description	Type	Example
Row_id	Unique row identifier	Integer	6
Security_id	Unique security identifier	Integer	24
P_tcount	Count of previous day's on market trades in current security	Integer	670
P_value	Sum of previous day's on market trade values in current security	Integer	65000
Trade_vwap	Volume weighted average price of the trade causing the liquidity shock	Double	4250
Initiator	Whether trade is buyer or seller initiated	String	B
Trans_type	Whether the time series event is a trade or a quote at event time t	String	T
Time	Event time at event time t	String	10:05:36:488
Bid	Best buy price at event time t	Double	213.85
Ask	Best Sell price at event time t	Double	214.10

Total number of independent securities: 102

Assumption

Different events of liquidity shocks of same security have been considered as independent of each other.

Splitting the data into train and test sets

Subsets of 50000 data points have been taken, split by 70-30 % rule into train and test data sets. The same rule has been implemented in all the models that have been developed.

2. Approaches for predicting Bid-Ask spread

Before starting with modelling, a benchmark has been set to consider a model as adequate if its RMSE is far less than this.

2.1 Benchmark Model

To set a lower bound for the error of the models that have been used later, average of first 50 bid prices and ask prices have been used to estimate bid and ask prices. The corresponding RMSEs are:

28.6 and 29.4.

2.2 Empirical Model

This is the first model that has been used is an empirical model. It is based on the primary assumption that bid ask spread takes an exponential decay after liquidity shock, reaching the market state before liquidity shock. The equation for the same is:

$$\frac{St - \mu}{S0 - \mu} = e^{-\alpha(t-49)}$$

where t is the event occurring after the liquidity shock. St is the spread when an event t occurs. S0 is the spread of the event that occurs immediately after the shock.µ is the mean spread across events before shock.

Following are the RMSE values for few of the events:

Table 3: Event with time stamp and corresponding RMSE

Event	**RMSE**
S 52	3.40
S 57	500+
S 62	Very high

The increase in RMSE as time increases can be attributed to decrease in prediction power as the effect of variables further away is very less.

The high RMSE even at S52 has raised a doubt of if the data is actually following the assumption or not.

To verify the same, spread vs time graphs have been plotted for different securities and the results are as follows:

Figure2: Spread Vs time plots.

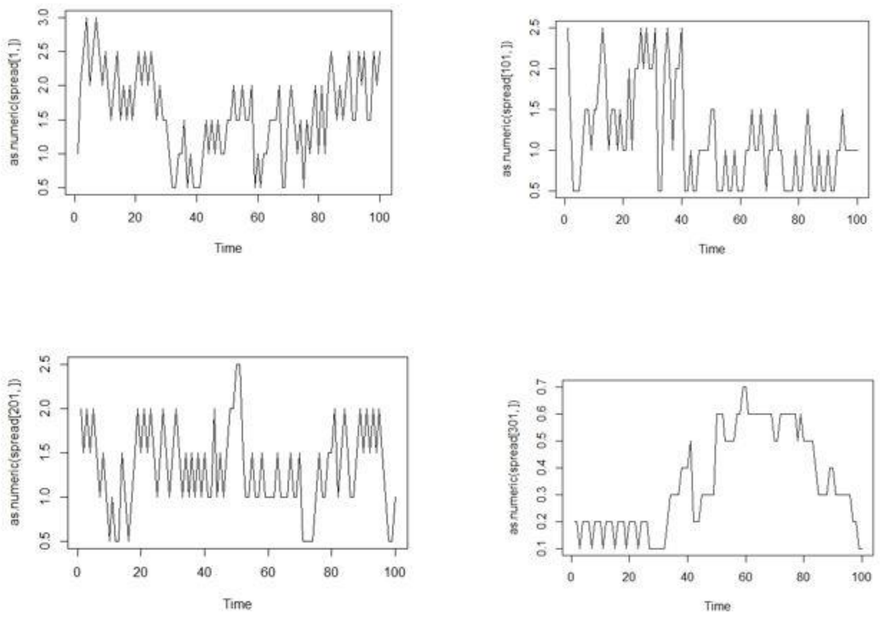
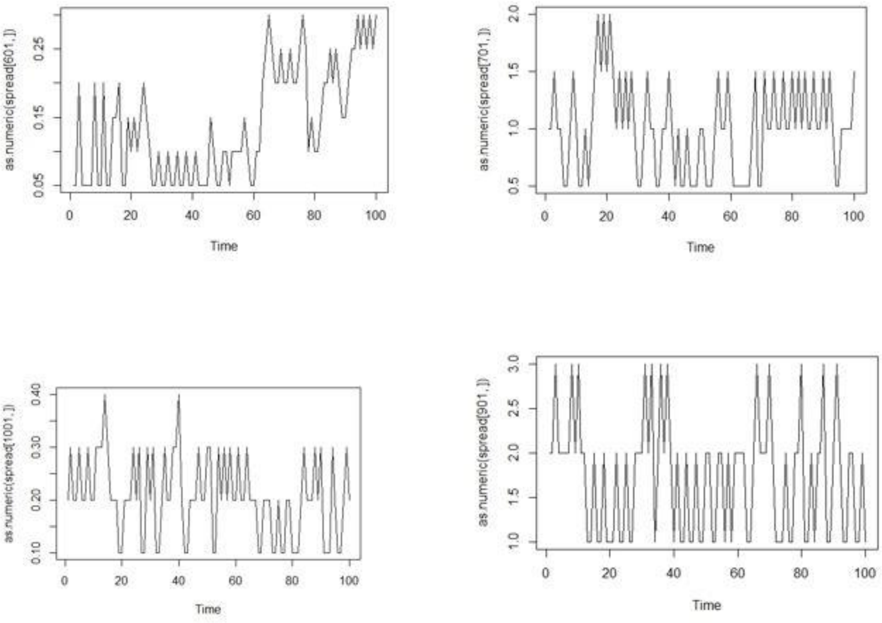


Figure3: Spread Vs time plots



As observed from the above plots, except for one of these securities, nothing else is following the expected phenomenon of exponential decay.

Main reason for this would be that the shock might not have occurred exactly at time stamp 50 but somewhere near it or there are multiple shocks in the given time stamp because of which the response is deviating from expectation. This has called for the motivation of using machine learning models to predict the market response.

2.3 Feature Extraction

Basic statistics have been extracted and used as features. The following are the extracted features:

Mean values of ask and bid prices, median values of ask and bid price, minimum values of ask and bid prices, maximum values of ask and bid prices and standard deviation of ask and bid prices.

2.4 Prediction Models using Machine Learning applications

2.4.1 Model using Random Forest Regression

Average RMSE per stock: 38.45

Figure4: Error Vs Number of trees

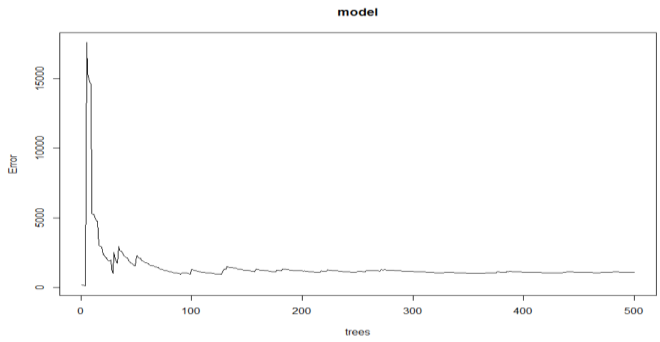
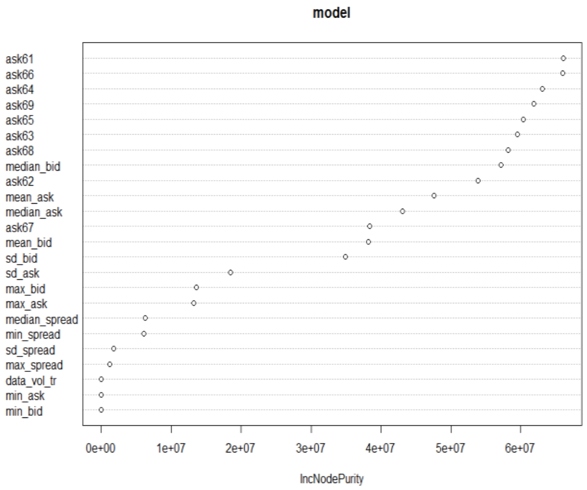


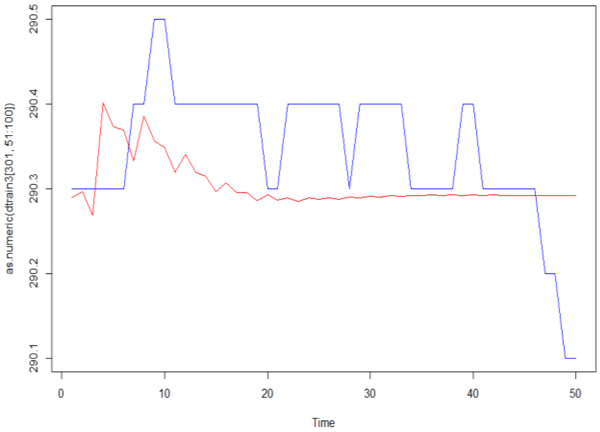
Figure5: Dependencies of ask72 on other variables



The above graph clearly indicated that ask price at a particular point of time is dependent on its values at previous time instances. The same has been observed for bid price also. This indicates that the data is time series and hence there might be dependencies of current values of prices on just previous values. Hence AR model has been applied.

2.4.2 AR Model

Figure6: Results of AR model



As observed in the above graph, the small trend has been predicted but not able to predict the fall perfectly for all the securities for a less number of data.RMSE is low, but the model might not be able to predict the fall if it has high price dip.

On observing the data, it has been deduced that the post shock events follow 5 different trends namely:

1. Monotonically increasing
2. Monotonically decreasing
3. Multi modal increase
4. Multi modal decrease
5. No noticeable fluctuations

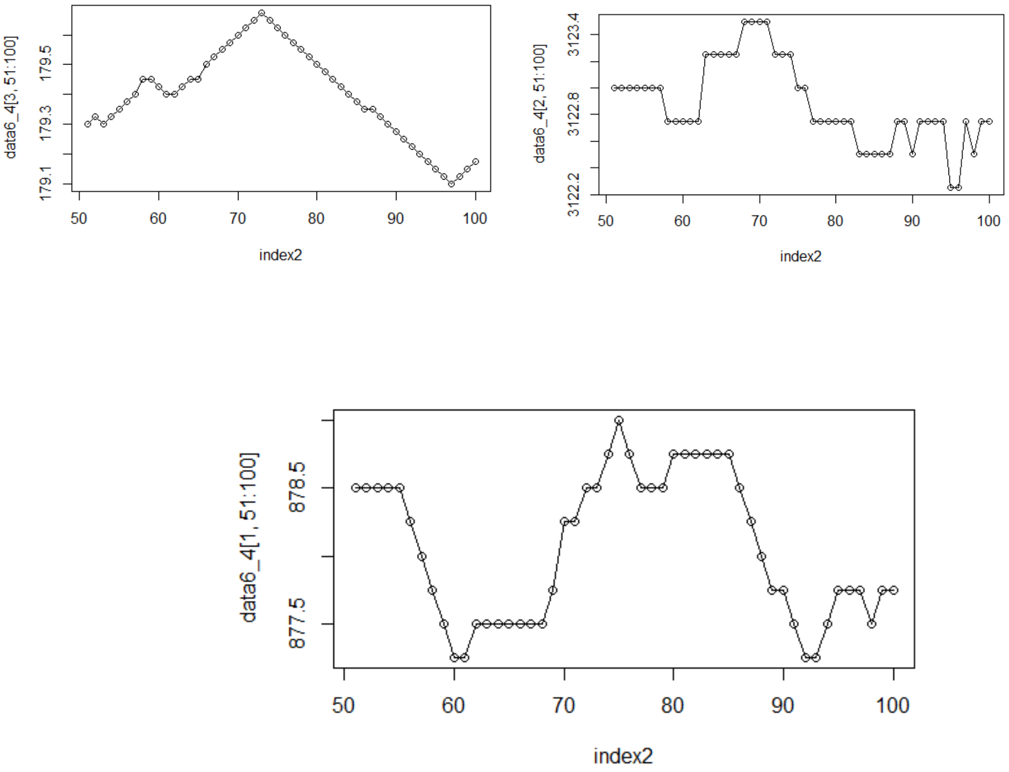
2.4.3 Clustering and Classification

K-means clustering was applied on post shock events of the training data and 5 clusters were identified, each of which represents a particular trend of post shock events.

Features have been used in random forest and cluster number as dependent variable on training data set. This was used to train the classifier, predictions of the clusters were made on test data.

Prediction models were applied on each of the clusters.

Figure7: Spread Vs Time for Random securities belonging to same cluster



From the above plots, it can be observed that all the three securities have decreasing pattern of bid ask spread and have been classified in the same group which justifies the correctness of the clustering.

2.4.3.2 Classifier model:

The purpose of Classifier model is to label the given data point into one of the five groups mentioned above. From given data, around 50000 data labels are taken, and using K-means clustering technique, they were classified into one of the groups. Considering this labelled data as input, following classified models are trained to predict the label of the given data point.

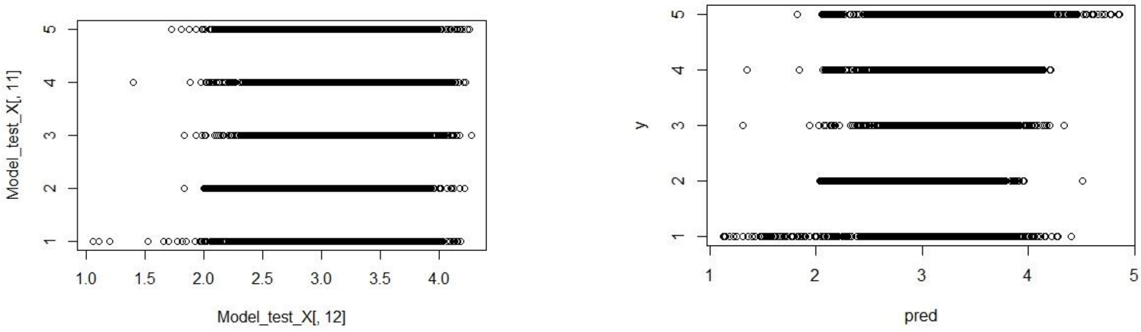
Input data for classifier model: 50000 data points are labelled into five clusters using clustering model explained above. Out of 37500 points taken as training data, to train classifier model and rest of the data will be used to test the accuracy of the classifier model.

Features for classifier model:

- Time 40-50 Mean and spread
- Min,MAx,Stdev of Mean and Spread
- Events/Quotes rate
- Trade volume and Trade amount before shock
- Buyer Initiated/Seller Intiatied

We used random forest and SVM model classifiers,and following are prediction accuracy of classifier model. Accuracy was poor, and probabilistic clustering and classifier model might be appropriate choice for the given problem.

Figure8: Prediction (Y-axis) vs. Labelled data by K –means (X-axis): (a) Random forest (b) SVM classifier

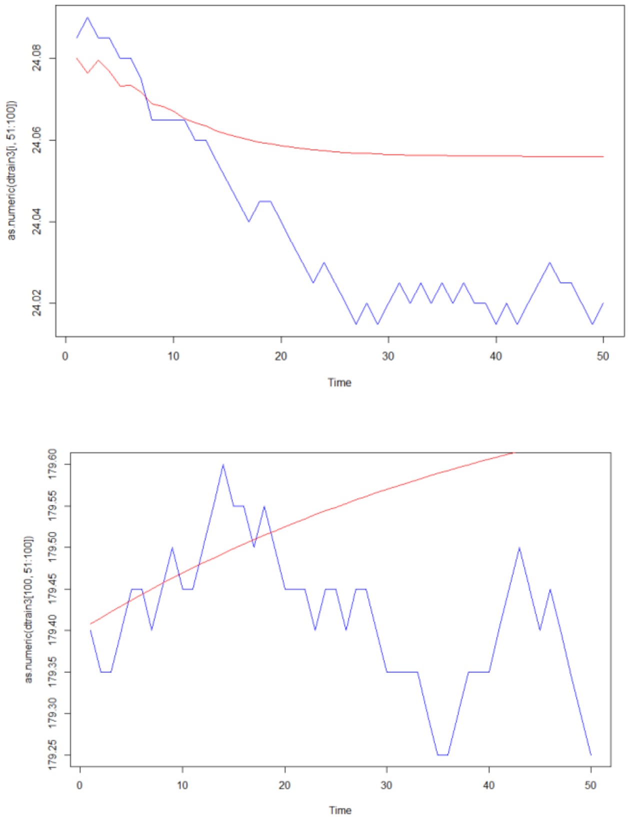


(a) **(b)**

2.4.3.2 AR model on clustered dataset

AR model has been implemented to capture linear dependencies. By using Auto Regressive model trend of the security has been captured. While the RMSE was found to be as low as 0.8, the prediction made was not satisfactory. The low RMSE is attributed to averaging the trend.

Figure9: Spread Vs Time accuracy for AR model:

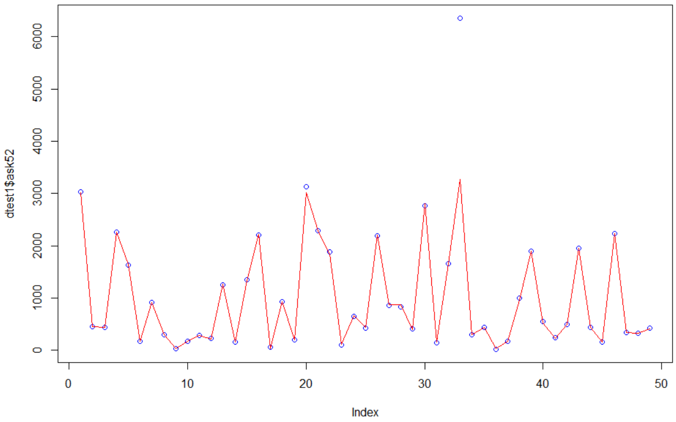


The above figure shows that the AR model accounts for the overall increase in the spread of the security, it does not account for the variations within, despite the clustering.

2.4.3.3 Random Forest Model on clustered dataset

From previous results, it has been concluded that random forest canpredict the trend of securities RMSE was high due to model inaccuracy, caused byabsence of liquidity shock in several securities. The same model when applied on these clusters has shown a significant decrease in the RMSE.

Figure10: Applying Random Forest in one of the clusters

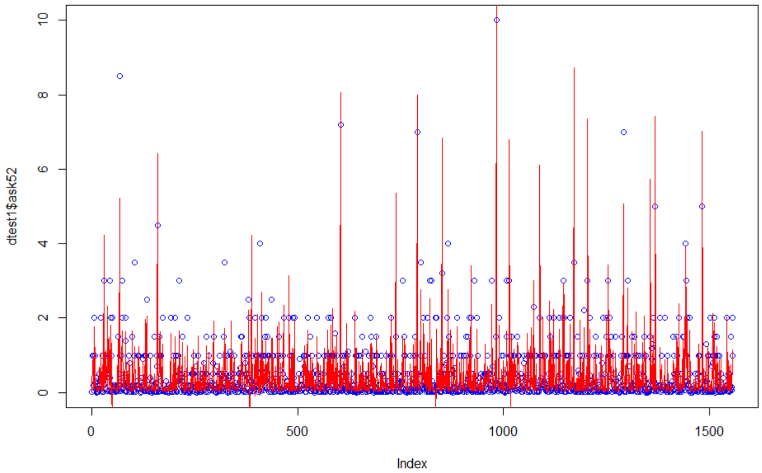


From the plot, it can be observed that random forest has been able to predict normal situation and trend but is not able to capture the outlier related data.

2.5 Neural Network Model on clustered dataset

Hence Neural network has been applied on the data which usually tries to fit the data depending on their structure.Neural Networks implementation on the whole dataset without clustering resulted in a RMSE value of 1.8. But when neural network was trained for each cluster separately. RMSE obtained from the given cluster is 0.6. Given below is figure showing the prediction using neural nets for 50000 dataset.

Figure11. Result of applying neural networks



In order to contribute for the non-linear dependencies in the data and also it's time series nature, a variant of ARNN model namely NARX has been tried to implement.

2.5.1 NARX Model

The Non-Linear Auto-Regressive model is one relates the current value of a time series where one would like to explain or predict to both past values of the same series and current and past values of the driving (exogenous) series — that is, of the externally determined series that influences the series of interest. General form a NARX model is:[2]

$$y_t = F(y_{t-1}, y_{t-2}, y_{t-3}, \ldots, u_t, u_{t-1}, u_{t-2}, u_{t-3}, \ldots) + \varepsilon_t$$

where, *Y* is the output.

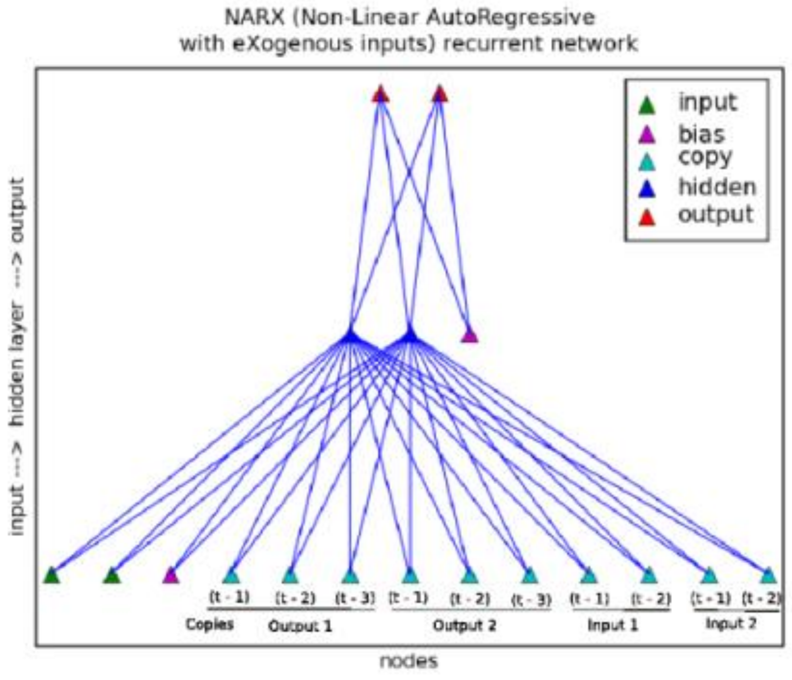
F is a non-linear function. It can be a polynomial or a neural network

Epsilon is the error

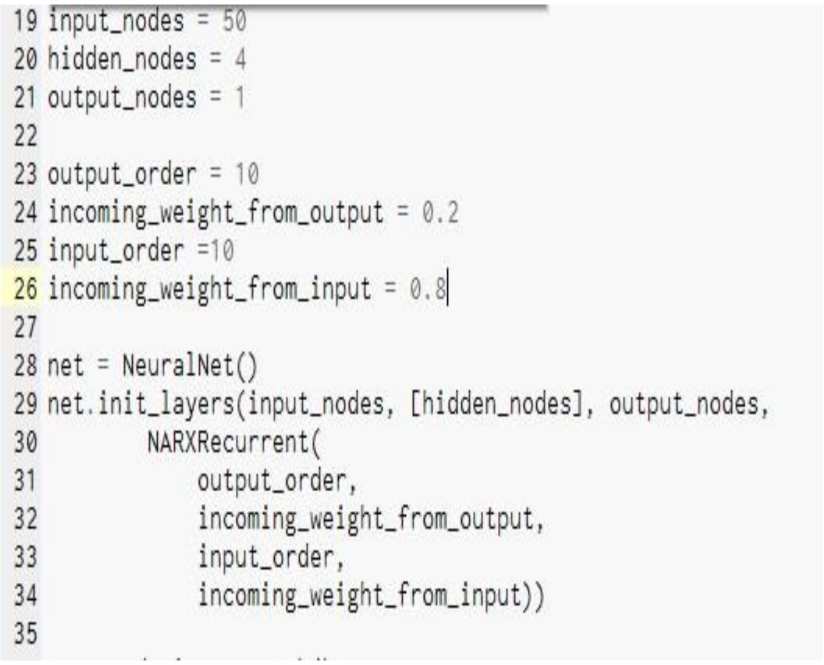
NARX neural network is a recurrent network which takes copies of both the input and the output. The transfer of the copy value in this case replaces the previous value.

This form with the mellifluous name can take copies from the output and input layers. Multiple levels of copies can be maintained, such as times **t-1**, **t-2**, **t-3**, and so on. In this network's nomenclature, the number of copies are referred to as the 'order'. The transfer of the copy value in this case replaces the previous value. In addition, a weighting factor is applied to a transfer so that a copy value is attenuated at each level. A sample NARX network is shown below

Figure12



In the project, a NARX Model with 50 input nodes (the ask/bid price during the first 50 timestamps), 4 hidden nodes, 10 copies and 1 output was used. The idea is that the model would output the tenth value, i.e., t=60. This value is dependent on the forecasted values of the previous nine timestamps along with a non-linear activation function. The model used in the project is shown below. However,it did not converge.



```
input_nodes = 50
hidden_nodes = 4
output_nodes = 1

output_order = 10
incoming_weight_from_output = 0.2
input_order =10
incoming_weight_from_input = 0.8

net = NeuralNet()
net.init_layers(input_nodes, [hidden_nodes], output_nodes,
        NARXRecurrent(
            output_order,
            incoming_weight_from_output,
            input_order,
            incoming_weight_from_input))

```

3. Conclusions

Prediction accuracy increased by a significant amount when the data is clustered based on the variability in price and spread of the securities

While the AR model resulted in the lowest RMSE, it did not model the fluctuations and trend in the data.

The Random Forest models the trend post liquidity shock very neatly but RMSE is higher due to existence of outliers.

Implementing Neural networks resulted in values, which were much closer to the actual values, than the other models.

ARNN model if converged, might have given a better accuracy as it does all the work an ANN does and also includes the time series nature of the data.

4. Possible Improvements

Rectifying NARX model and verify if there has been an improvement in the RMSE.

Clustering shocks based on security ids to identify interdependencies among them, use it as a feature to predict response of the market.

Extend this model to general identification of liquidity shocks and implement back testing on it.

5. References and Literature review

References:

1. Kaggle Algorithmic Trading Challenge
2. Wikipedia
3. Trading Challenge’s forum

Literature Review:

1. Measuring the Resiliency of an Electronic Limit Order Book-Jeremy Large*.
2. Winning the Kaggle Algorithmic Trading Challenge with the Composition of Many Models and Feature Engineering-Ildefons Magrans de Abril.
3. Market Response to Liquidity Shocks in the Limit Order Book -Kartikey Asthana*.
4. Information Shocks, Liquidity Shocks, Jumps, and Price Discovery —Evidence from the U.S.Treasury Market-George J. Jiang, Ingrid Lo, and Adrien Verdelhan.