SCHOOL OF COMPUTER SCIENCE AND ENGINEERING (SCOPE)
October 2020

CSE3020 - Data Visualization

FINAL PROJECT REPORT

Anomaly Detection in Time Series Data using LSTM in Keras

FALL SEMESTER 2020-21

SUBMITTED BY

17BCE2383 TEK RAJ AWASTHI

17BCE2380 ALOK SINHA

17BCE2393 BIBEK SINGH

UNDER THE GUIDANCE OF

PROF. RAMANI S

ASSISTANT PROFESSOR (SR) SCOPE

VIT UNIVERSITY, VELLORE

VIT®
UNIVERSITY
(Estd. u/s 3 of UGC Act 1956)

TABLE OF CONTENTS

ABSTRACT

The problem of anomaly detection on time series is to predict whether a newly observed time series is normal or not, to a set of training time series. It is extremely helpful in many observing applications for example video surveillance, signal recognition and finance, IT, medical, and energy. However, identifying abnormalities in streaming data is a difficult task, requiring detectors to process data in real-time, not batches, and learn while simultaneously
making predictions. The perfect anomaly detector would detect all anomalies as soon as possible, trigger no false alarms, work with real-world time-series data across a variety of domains, and automatically adapt to changing statistics Based on some existing outlier detection algorithms, we propose an instance-based anomaly detection algorithm. We also propose a local instance summarization approach to reduce the number of distance computation of time series, so that abnormal time series can be efficiently detected. Experiments show that the proposed algorithm achieves much better accuracy than the basic outlier detection algorithms. It is also very efficient for anomaly detection of time series.

CHAPTER 1

1.1 INTRODUCTION

Anomaly Detection in the data mining field is the identification of the data of a variable or events that do not follow a certain pattern. Anomaly detection helps to identify the unexpected behavior of the data with time so that businesses, companies can make strategies to overcome the situation. It also helps the firms to detect the error and frauds that are going to happen at particular time, or it helps to learn from past histories of data that showed unusual behavior.

Time Series is a sequence of numerical data collected at different points in time in successive order. This is not a cross-sectional data. This is an observation on the value of a variable at different times. Detecting anomalies in Stock index dataset means analyzing past data of stock price, recognize its pattern and finding the breakouts in the closing price of stock index.

1.2 IMPORTANCE OF STUDY PROPOSED

In modern competitive financial market, participants use machine learning, data analytics to gain insights on the historical dataset and then take effective approaches. So, anomaly detection helps to increase the speed of execution and to win competitive financial market. Machine learning algorithm's implementation helps the companies to find simple and effective approaches for detecting the anomalies. Since machine learning algorithms are able to learn from data and make predictions so applying these algorithms in anomaly detection of time series data carries huge impact on its performance.

Time Series data are very important for prediction. These data are used for understanding past outcomes, predicting future outcomes, making progress strategies, and more. There are various application of anomaly detection in time series data in different domain topics.

CHAPTER 2

2.1 BACKGROUND OF PROBLEM

In modern competitive financial market, participants use machine learning, data analytics to gain insights on the historical dataset and then take effective approaches. So, anomaly detection helps to increase the speed of execution and to win competitive financial market. Machine learning algorithm's implementation helps the companies to find simple and effective approaches for detecting the anomalies. Since machine learning algorithms are able to learn from data and make predictions so applying these algorithms in anomaly detection of time series data carries huge impact on its performance. There are various application of anomaly detection in time series data in different domain topics.

2.2 EARLIER STUDY & MOTIVATION BEHIND TAKING THE PROJECT

From the big data perspective, anomaly detection in financial data has widely been ignored despite many organizations store, process and disseminate financial market data for interested customers to assist them to make informed decision and create competitive advantages. Since machine learning algorithms are able to learn from data and make predictions so applying these algorithms in anomaly detection of time series data carries huge impact on its performance. LSTM Auto-encoders are an unsupervised learning technique, although they are trained using supervised learning methods.

2.3 OBJECTIVE OF PROPOSED STUDY

Anomaly Detection in the data mining field is the identification of the data of a variable or events that do not follow a certain pattern. Anomaly detection helps to identify the unexpected behavior of the data with time so that businesses, companies can make strategies to overcome the situation and can make informed decision and create competitive advantages.

In this project, we'll build a model for Anomaly Detection in Time Series data of S&P 500 stock Index dataset from 1988 to 201, to detect and predict anomalies. By anomalies I mean detecting sudden price change in S&P index, using Deep Learning in Keras with Python code. Specifically, we'll be designing and training an LSTM Auto-encoder using Keras API, and Tensorflow2 as back-end. Along with this we have also created interactive charts and plots with plotly python and Seaborn for data visualization and displaying results within Jupyter Notebook.

2.4 IDENTIFICATION OF PROBLEM

In this project we will be working with S&P 500 Index to detect and predict anomalies. We'll work with this data, but captured from 1986 and 2018. By anomalies means sudden stock price change in S&P 500 index dataset stock market index that tracks the stock performances of top 500 large-cap US companies listed in stock exchanges. This index represents the performances of stock market by reporting the risks and reporting of the biggest companies so that stock shareholders gain insights on the historical dataset and then take effective approaches

CHAPTER III

.

3.1 METHODOLOGY ADOPTED

We have designed deep learning model in keras API. Specifically, we have designed and trained an LSTM Auto-encoder unsupervised model with supervised data

LSTM stands for Long Short-term Memory, which is also an artificial neural network similar to Recurrent Neural Network (RNN). It processes the data's passing on the information as it propagates. It has a cell, allows the neural network to keep or forget the information.
The LSTM model in this project is employed as below:

1. Train an LSTM auto-encoder on the Johnson & Johnson's stock price data from 1986 to 2018. We assume that there were no anomalies and they were normal.

2. Using the LSTM auto-encoder to reconstruct the error on the test data.

3. If the reconstruction error for the test data is above the threshold, we label the data point as an anomaly.

We have used plotly, and we'll use a sub-module graph_objects from plotly and populated the figure using add_trace() method which helps to plot different types of charts in the same figure. And Scatter mode is set to 'line' plot. Legend value is set to 'close' which is closing stock value and then update the figure layout.

The dataset is splitter into training and testing set. We have taken 80% of data frame for training and remaining 20% for testing. Then we have standardized our target vector by removing the mean and scaling it to unit variance. We have created the subsequences before using the data to train our model.

As required by LSTM network, we reshaped our input data into shape and sample by n time steps by n features. In our case, the n is equal to 1 i.e. one feature. We have built an LSTM Auto-encoder network and visualize the architecture and data flow.
We have used sequential model from keras, and created one LSTM layer with the number of cells to be 128. Input shape is equal to no. of time steps divided by no. of features. Then we have added the Dropout regularization to 0.2.
Now we have mirrored the encoder in reverse fashion i.e. decoder. Time Distributed function creates a dense layer with number of nodes equal to the number of features. And the model is compiled finally using adam optimizer function which is gradient descent optimizer.
Then Auto-encoder model is trained using 100 epochs.

- Then we have plotted the training and validation loss matrix.
- Calculated MAE loss on the training data.
- Made the max MAE loss value in the training data as the reconstruction error model.
- If the reconstruction loss for a data point in the test set came greater than this reconstruction error model value then we labelled this data point as an anomaly.

3.2 DATASET COLLECTION

In this project, we have taken time series data of S&P 500 index dataset, but captured from 1986 and 2018.

***Source**: https://www.kaggle.com/pdquant/sp500-daily-19862018*

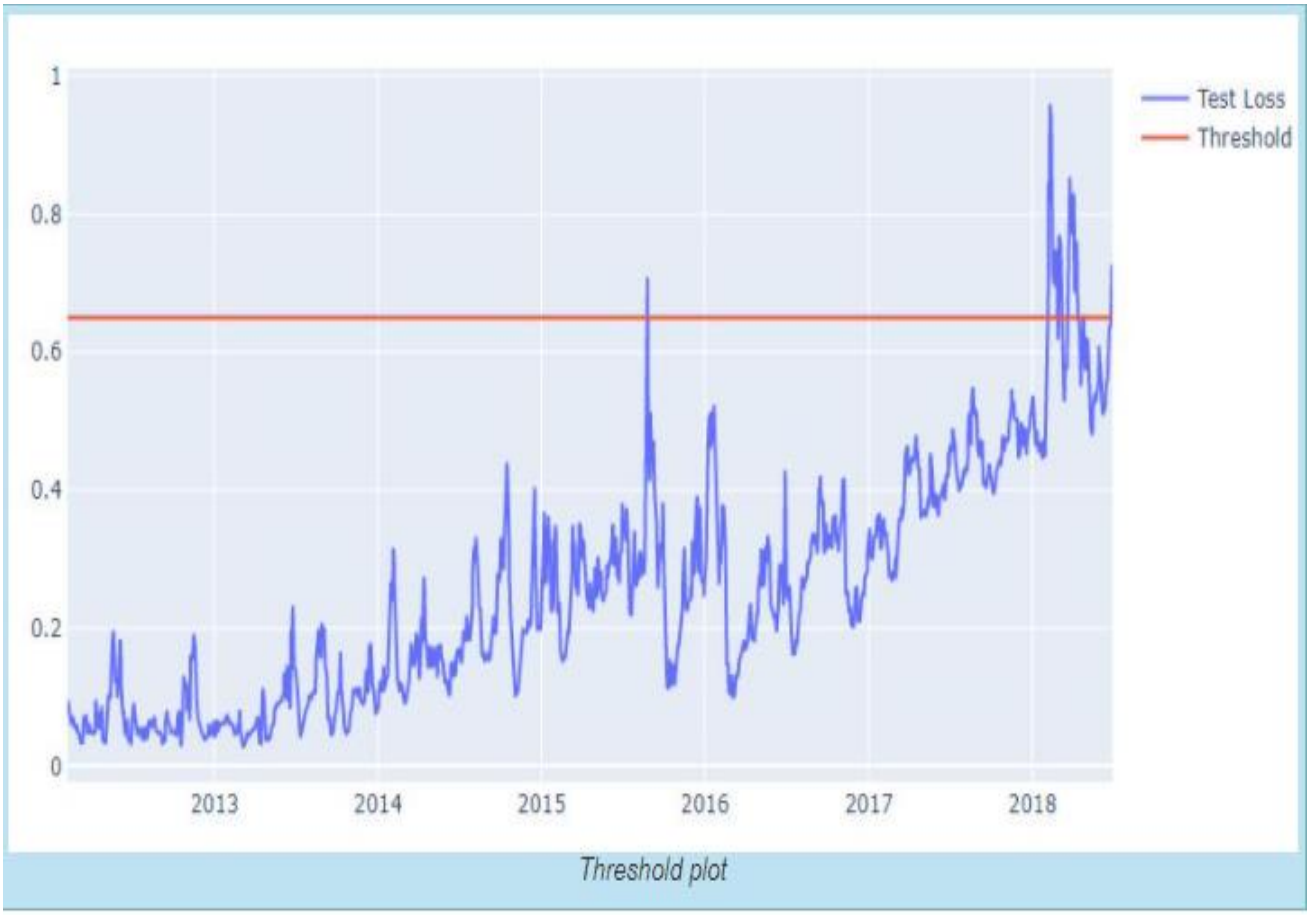
3.3 COMPUTATION TOOLS USED

We have designed deep learning model in keras API. we have designed and trained an LSTM Auto-encoder using Keras API, and Tensorflow2 as back-end. Along with this we will also create interactive charts and plots with plotly in python and seaborn for data visualization and displaying results.

CHAPTER IV

4.1 RESULT

The anomalies in S&P500 index data is detected using LSM Auto-encoder model, using sequential model from keras API. The plot below looks like we are thresholding extreme values quite well. All the values above the horizontal orange line are classified as Anomalies.



Threshold plot

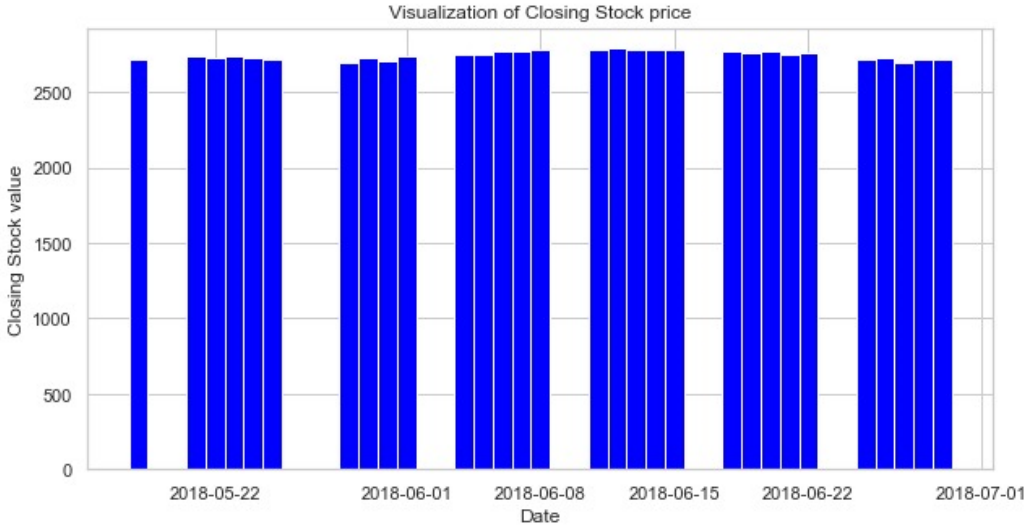
The plots used in these projects are

- Bar Graph
- Line Graph
- Distribution Plot
- Scatter plot

BAR GRAPH

A bar graph (also known as a bar chart or bar diagram) is a visual tool that uses bars to compare data among categories. A bar graph may run horizontally or vertically. The important thing to know is that the longer the bar, the greater its value.
Bar graphs consist of two axes. On a vertical bar graph, the horizontal axis (or x-axis) shows the data categories. In this example, they are days. The vertical axis (or y-axis) is the scale. The colored bars are the data series.
Bar graphs have three key attributes:

• A bar diagram makes it easy to compare sets of data between different groups at a glance.
• The graph represents categories on one axis and a discrete value in the other. The goal is to show the relationship between the two axes.
• Bar charts can also show big changes in data over time.

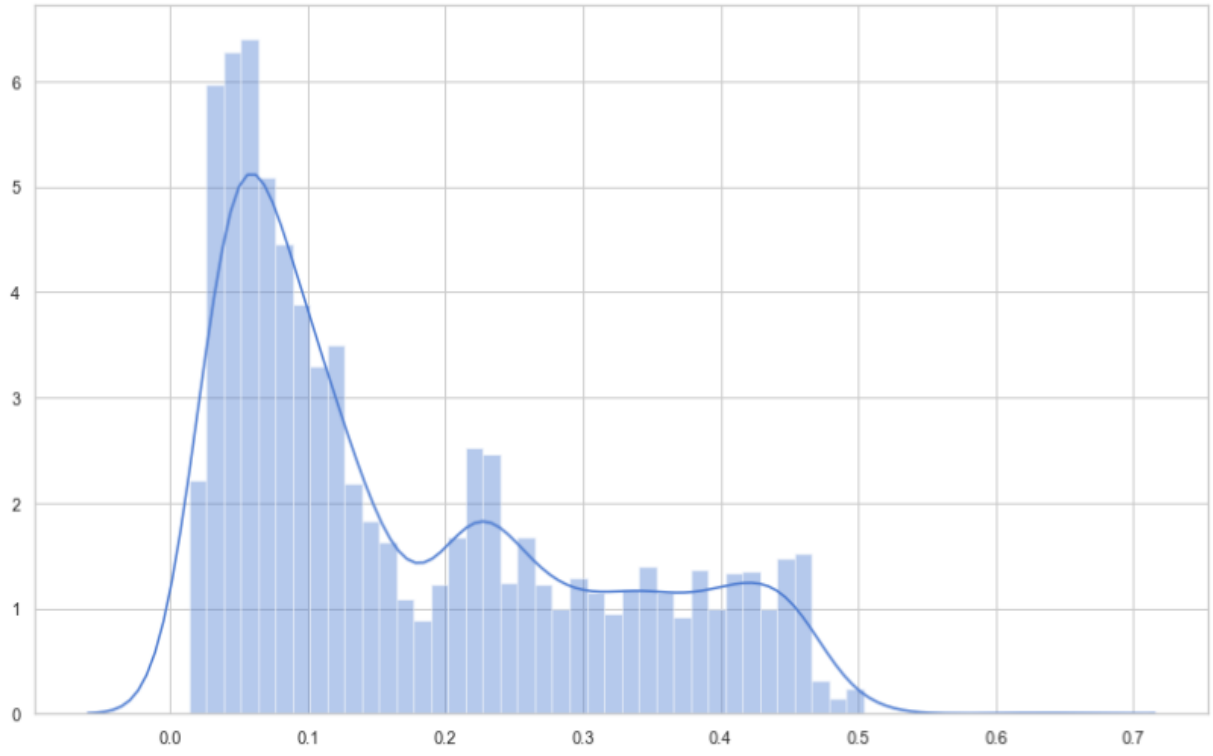


LINE GRAPH

Line graph, also known as a line chart, is a type of chart used to visualize the value of something over time. The line graph consists of a horizontal x-axis and a vertical y-axis. Most line graphs only deal with positive number values, so these axes typically intersect near the bottom of the y-axis and the left end of the x-axis. The point at which the axes intersect is always (0, 0). Each axis is labelled with a data type. For example, the x-axis shows days while the y-axis shows revenue. Here, we have plotted the Date vs stock closing price.

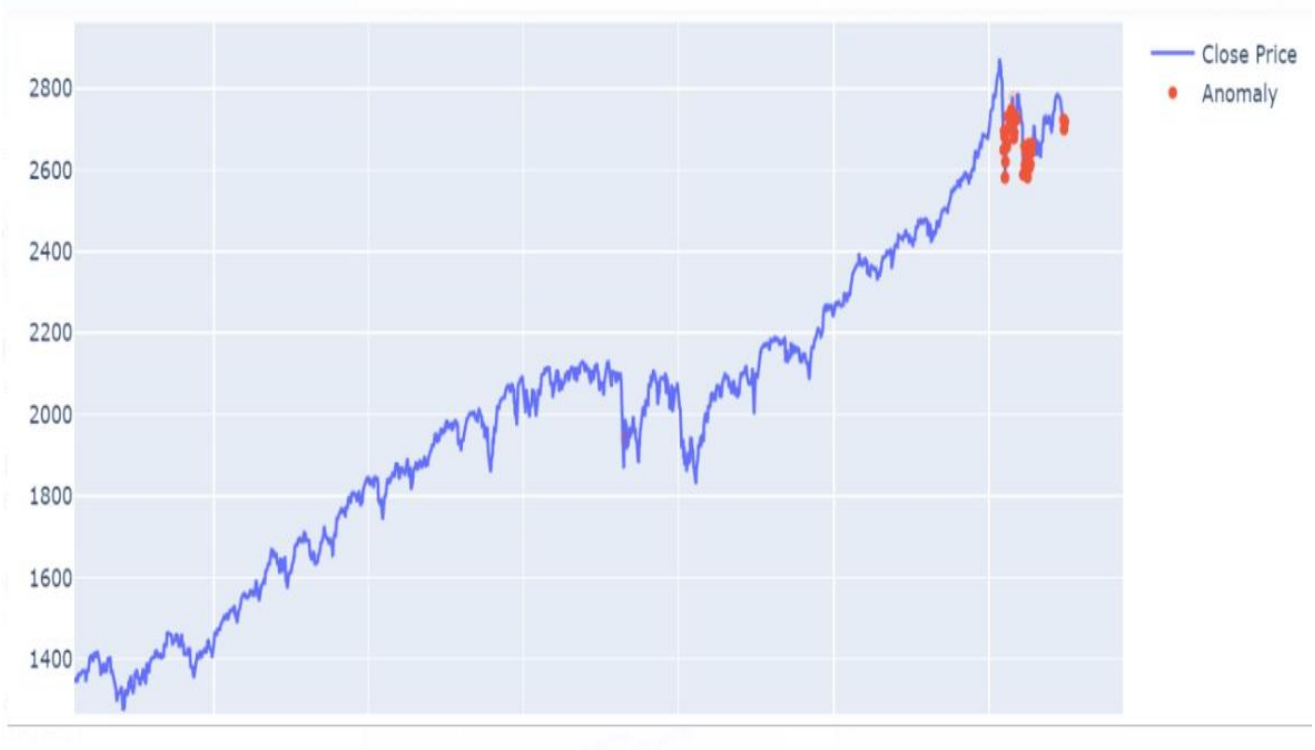
DISTRIBUTION PLOT

A distribution plot displays a distribution and range of a set of numeric values plotted against a dimension. You can display this chart in three different ways, you can just have the value points displayed showing the distribution, or you can display the bounding box which shows the range or use a combination of both. In the distribution plot shown below, you can see there a range and distribution of the average sales values displayed for each product group. The outer bounding box shows the range whereas inner individual dots show the distribution. Each range and distribution box show how data values for a product group is distributed over the average sales range.



SCATTER PLOT

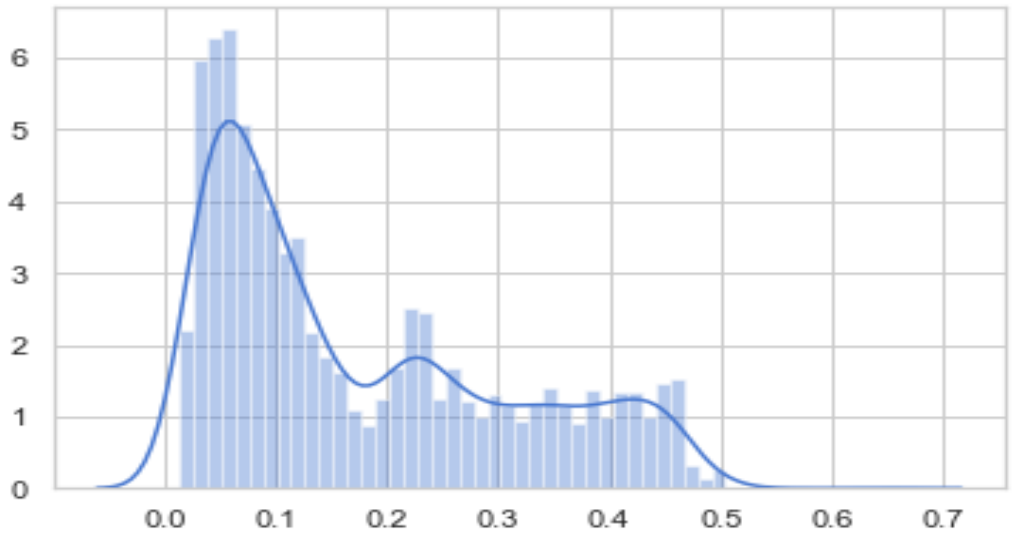
A scatter plot (aka scatter chart, scatter graph) uses dots to represent values for two different numeric variables. The position of each dot on the horizontal and vertical axis indicates values for an individual data point. Scatter plots are used to observe relationships between variables. Scatter plots' primary uses are to observe and show relationships between two numeric variables. The dots in a scatter plot not only report the values of individual data points, but also patterns when the data are taken as a whole.



4.2 INTERPRETATION OF RESULT

Below is the visualization of closing stock price data using time series forecasting. And it shows that stock index value is heavily increasing after 1995 and then fluctuating in between 2000 to 2010, and after that stock price value is increasing constantly.

Below picture is the plot of distribution loss of training mean absolute error is, using seaborn. So, we set the threshold as 0.65 as no value is larger than that. Then a data frame containing loss and anomalies values. Then we created a Boolean-valued column called an anomaly, to track whether the input in that corresponding row is an anomaly or not using the condition that the loss is greater than the threshold or not.



INFERENCE & ANALYSIS

The main method for the analysis part is to see the change in the price movement between the closing and opening time respectively. The unpredicted and irregularities between the prices will help to understand the on-going activities of the stock market.
RNNs like LSTMs, however, are used to understand the relationship between past values of a variable and some target (be it a class or a future state value of the variable). Where the patterns of future variable values deviate from the values predicted by the algorithm, we have a chance of finding anomalies and novel values of other kinds.
RNN generally has a better chance of learning whether the anomaly to be detected is data dependent, as in the previous data values influence the later, than CNN. Now that would be completely based upon whether your time series data actually gives you more information combined to some intervals like maybe it's about days of a month. Your RNN in this case can definitely help you since it might just learn that certain minute bad days let's say consecutive 4–5 days may mean in the 6th day you're going to get the anomaly you wish to detect.

CHAPTER V

5.1 CONCLUSION

The project introduces the anomaly behavior of the stock market price of SP index. The dataset contains the price of opening and closing price which is being interpreted how the change in stock market price has made significant changes to the real world. The digitization of the stock market dashboard shows how the data are fluctuated with the term of Buying and Selling prices at opening and closing time period. The various factor determines the rate of stock prices at different level. The condition is being predicted and generalized with the current scenario. Different illegal and unauthorized manipulation in the stock price has made drastic changes in the stock market which is being predicted in a different manner. The anomaly behavior is predicting when there is sudden change in the graph which might give some other output. These variations may lead to increase or decrease the price of stock at different levels. The main method for the analysis part is to see the change in the price movement between the closing and opening time respectively. The unpredicted and irregularities between the prices will help to understand the on-going activities of the stock market.

5.2 SCOPE FOR FURTHER STUDY

Despite the fact that the all calculation beats the other tried peculiarity identifiers, the outcomes show there is still opportunity to get better. it is ready for researchers to evaluate their algorithms and report their results. To move beyond, future work on it will involve adding more real-world data files to the corpus. We will also anticipate incorporating multivariate.
anomaly detection and categorical data. Over time we hope researchers can use it to test and develop a large number of anomaly detection algorithms for the specific purpose of applying them to real time streaming applications.

5.3 REFERENCES

[1]. Boniol, P., & Palpanas, T. (2020). Series2graph: Graph-based subsequence anomaly detection for time series. *Proceedings of the VLDB Endowment*, *13*(12), 1821-1834.

[2]. Su, Y., Zhao, Y., Niu, C., Liu, R., Sun, W., & Pei, D. (2019, July). Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25thACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2828-2837).

[3]. Wen, T., & Keyes, R. (2019). Time series anomaly detection using convolutional neural networks and transfer learning. *arXiv preprint arXiv:1905.13628*.

[4]. Tseng, K. C., Kwon, O., & Tjing, L. C. (2012). Time series and neural network forecasts of daily stock prices. *Investment Management and Financial Innovations*, (9, Iss. 1), 32-54.

[5]. Wang, H., Tang, M., Park, Y., & Priebe, C. E. (2013). Locality statistics for anomaly detection in time series of graphs. *IEEE Transactions on Signal Processing*, *62*(3), 703-717.

[6]. Teng, M. (2010, December). Anomaly detection on time series. In *2010 IEEE International Conference on Progress in Informatics and Computing* (Vol. 1, pp. 603-608). IEEE.