IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, July 2022
Impact Factor 6.252

Risk Reduction by Stock Portfolio Selection using LSTM and K-means Clustering

Rishabh Sharma and Dr Shipra Arora
Dronacharya College of Engineering, Gurgaon, India

Abstract: *The stock market is very complex to understand. In the stock market, the stock prices keep fluctuating and it is tough to track the price of stocks. Most users find it difficult to choose the right stocks for their investment and there is a lot of confusion when selecting the right stocks in fear of losing your hard-earned money. In this paper, we discuss how to create a stock portfolio on the basis of stock prediction and then select multiple stocks using K-means clustering to reduce risk as per the profit. Many people aren't aware of price fluctuation and lose their money in specific stocks. This paper helps to understand the benefit of diversification of stock portfolios to reduce the risk of stock market fluctuations.*

Keywords: K-means, Clustering, Stock Market, Price Prediction, Pattern Matching

I. INTRODUCTION

A stock market is a place where shares of publically listed companies are traded. The stock market is an aggressive market where it is difficult to predict the future stock price. These prices fluctuate on the basis of supply and demand. As we know, it is very difficult to predict the price of the stock, but if we can predict the stock price, it will definitely help to analyse the market and make a profit in the long term. But in real life, it is not trustworthy as prices are not only dependent on demand and supply trends. There are many external factors involved in deciding the stock prices, for example, market shutdown due to covid pandemic.

In this paper, we've used the LSTM model to predict the price of stocks in future and we conclude that LSTM has the potential to predict the price very accurately apart from some special external factors like the pandemic. So if the Customer buys the stocks for which LSTM doesn't do well, he/she/they may lose their invested money. So to overcome this risk we've used the K-means clustering algorithm to cluster all stocks which have predicted prices and make a group of stocks in such a manner that we diversify our portfolio by considering their predicted price and risk factor involved in such stocks. Hence it helps to reduce the risk of investment as a diversified portfolio protects us from sudden crashes in a particular sector.

II. RELATED WORK

Bini B. S. & Mathew T. [2] made a price prediction using clustering and regression techniques. The regression technique is used to predict the price and clustering to find the pattern between data. First of all, Clustering is done on the available stocks validation index and the K-means algorithm appears to be more efficient among all clustering algorithms. Then cluster is passed to multiple regression to find a better-predicted price.

Li, Y., & Wu, H [3] discussed the ways to improve the K-means clustering by not randomly selecting centroid for the first run. It first finds the minimum and maximum distance between data points, to understand the centroid value more accurately. Centroid value not to be adjacent and should be with minimum distance to nearest data points.

Nayak, A., Pai, M. M. M., & Pai, R. M [4] suggested a theoretical model predict the stock price by collecting historical data, and current news and media reports. This model appears to be more accurate as it considers external environment data. But it does not guarantee the operator's effect on the price prediction.

Alqaryouti, O., Farouk, T., & Siyam [6] used clustering techniques to reduce the risk of the stock market, by creating portfolios using clustering techniques. By normalizing the prices, clusters of stocks are created to analyse different stocks of different segments/ sectors.

Moghar, A., & Hamiche, M. [9] using LSTM model to predict the price of US stocks using epoch. Some stocks predict the price is very similar but it fails to predict for some stocks like NKE stocks. LSTM model uses time-series data to predict the data for upcoming series.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, July 2022
Impact Factor 6.252

III. METHODOLOGY

The goal of the study is to find out the diversified portfolio of stocks as per similarities of profit expected and risk involved in that particular stock. To predict the price of a stock in future, the ***'Close'*** attribute of stocks is used while on other hand to identify the risk involved in a stock ***'Beta'*** attribute is used. These values are defined as:

- **Close**: It is the closing price of a stock for a particular day. It does not include any variation in the market for that particular day.
- **Beta**: It is used to define the volatility of the stocks i.e. how much a stock's price moves according to the overall market.

More beta value means the stock is highly volatile as compared to the market. Overall market's beta value is 1.0.

For the purpose of predicting stock price LSTM network is used due to its greater accuracy for demand forecasting. It can predict future values based on previous sequential data. And the stock price available with us is time-series sequential data. On another hand, K-means performs clustering on a given dataset with accuracy irrespective of the shape and size of the cluster. And it provides freedom to decide how many possible portfolios (K-value) to choose from.

3.1 Long Short Term Memory Network (LSTM)

Long Short Term Memory (LSTM) is used for building a model to predict the price of a stock in the future. LSTM is a `Recurrent Neural Network` for learning long term dependency. LSTM consist of four interacting layers creating a chain-like structure.

LSTM works in a three-step process:

1. It uses a sigmoid function to decide which information is omitted from the given time step.
2. Uses a combination of two functions in the second layer, i.e sigmoid and tan(h). The main motive of the layer is to provide an importance level to the given values.
3. The final step is to generate output. The final output is decided on the basis of the sigmoid layer and then the tan(h) function by generating a value between -1 to 1

LSTM is generally used for processing and predicting time-series data. Here it is used to calculate the future price by generating the 'Close' attribute value of stock in time-series format to predict the price at a given time.

3.2 K-means Clustering

"Clustering is the process of dividing a population or set of data points into various groups so that data points in the same group are more similar and data points in other groups are more dissimilar. K-Means Clustering is an unsupervised learning algorithm used in machine learning to handle clustering problems. The K-means algorithm finds k centroids and then assigns each data point to the cluster with centroids as near as possible. It uses Euclidean distance to find clusters Centroid" [12].

$$\text{Euclidean Distance} = \sqrt{(Xo - Xc)2 + (Yo - Yc)2} \quad (1)$$

where,
Xo represent observered Value of X,
Xc represent centroid value of X,
Yo represent observered Value of Y,
Yc represent centroid value of Y

K-means is an iterative algorithm that splits the unlabeled dataset into k distinct clusters, with each dataset belonging to only one of them. Hence it performs the Euclidean distance formula iteratively to compute the final result.

$$J = \sum_{k=1}^{K} \blacksquare \sum_{i=1}^{j} \blacksquare (Euclidean\ Distance)^2 \quad (2)$$

where: j: the number of data items in a data set,
K: number of clusters

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, July 2022
Impact Factor 6.252

3.3 Risk Calculation

Generally it is considered that small-cap stocks are riskier as compared to mid-cap and large-cap stocks. But based on the past results small-cap stocks can give more profit as compared to other stocks. So it is mandatory to maintain a balance between risk and profit while investing. Hence, Risk calculation is the main aspect of investing in the stock market. In this paper, to calculate the risk we have used the profit claimed and volatility factor of that particular stock.

$$Risk\ Factor = (\alpha - \gamma) \times \beta \quad (3)$$

Where α : *expected profit*
γ : *risk free profit*
β : *beta value of stock*

Risk-Free Profit represents the profit that is available in the market without any risk as per state guidance. For example in the current scenario, we have taken '7' as a value. Because it is common FD rates available in the market in India.

IV. DATA SETS

Data consists of various stocks of all market cap i.e. small-cap, mid-cap and large-cap. The data used was reported by the Bombay Stocks Exchange on a daily basis from 1 January 2015 to 31 December 2020 (open market days only). The 'Close' and 'Beta' attribute value is extracted from the data set and used for price prediction for 31 December 2020.

V. RESULTS

5.1. Price Prediction

To predict the price of Stocks, the 'Close' value is extracted from the dataset for each dataset and LSTM Model is applied to train the machine for the date from 1 Jan 2015 to 31 October 2017. And then LSTM Model is used to predict the future price for 31 December 2020. During these days, the Indian market does not have a major impact, the price predicted by LSTM is very similar to the actual market.

The next Step Performed is Data Cleaning and therefore all those stocks are dropped which are having negative returns. Because no one wants to add this type of stock to their portfolio. Some of them are *Coal India, Sun Pharma, ITC, ONGC, FEL, Suzlon,* and *ZEE Learn.* After that Error of Prediction is calculated using actual and predicted values. And Profit Column is added using Predicted Price (*Refer to Figure 1*)

	Stock Name	Initial Price	Predicted Price	Actual Price	Error In Prediction	Profit Percentage
0	Reliance Industries Limited	439.84	1998.070700	1984.65	0.676225	354.272167
1	TCS	1274.10	2664.544700	2870.20	7.165191	109.131520
2	Infosysis	493.95	1222.232700	1255.85	2.676856	147.440571
3	Bajaj Finance	348.47	4544.156200	5296.00	14.196446	1204.030820
4	Hdfc Bank	475.65	1330.781900	1436.75	7.375542	179.781751
5	Hdfc	1123.45	2423.118000	2558.60	5.295161	115.685433
6	Wipro	207.11	359.989960	386.25	6.798716	73.815827
7	APL Apollo Tubes	33.73	365.182620	442.33	17.441137	982.664157
8	Indian Hotels Company	113.01	119.595290	116.43	2.718621	5.827175
9	Asian Paints	750.35	2406.834200	2764.45	12.936237	220.761538
10	Relaxo Footwear	144.45	872.830600	811.20	7.597461	504.244098
11	Voltas	240.90	803.921630	826.00	2.672926	233.715911
12	MRF	37976.95	75546.440000	75712.40	0.219198	98.927086
13	Godrej Properties	256.15	1376.679700	1431.55	3.832929	437.450595
14	BATA	649.53	1848.079300	1578.85	17.052241	184.525626
15	TVS Motors	276.85	482.303200	485.55	0.668685	74.211017
16	Bajaj Auto	2454.10	3487.157200	3447.20	1.159120	42.095155
17	MPS Infoterence	0.11	0.191205	0.19	0.634253	73.822800
18	Orchid Pharma	65.75	109.873540	126.13	12.888655	67.108046
19	Tata Elxsi	302.13	1689.672600	1832.85	7.811736	459.253500
20	Finotex Chemicals	19.36	58.572163	66.45	11.855285	202.542164
21	Century Plywood	160.45	238.985080	233.05	2.546698	48.946762
22	TV18 Broadcast	31.15	31.539450	30.95	1.904523	1.250241
23	Deepak Nitrite	82.35	911.204200	942.40	3.310250	1006.501761
24	Laurus Lab	96.10	341.984830	353.15	3.161594	255.863507
25	Globus Spirits	67.80	253.494920	322.55	21.409109	273.886313

Figure 1. Predicted Price and Profit Calculated Table

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, July 2022
Impact Factor 6.252

As we can see the maximum Error in Prediction is found in *APL Appolo tubes* and the least in *MPS Infotechnics*. So we can conclude that this model work in an efficient way since the error in prediction average is less than 7% (refer to Figure 2).

	Initial Price	Predicted Price	Actual Price	Error In Prediction	Profit Percentage
count	26.000000	26.000000	26.000000	26.000000	26.000000
mean	1849.376538	4057.209073	4127.383846	6.769415	282.990598

Figure 2. Mean Value For Each Table Entity

5.2. Portfolio Construction

To construct the portfolio we used the benefits of clustering to find similarities between given datasets. So that similar stocks can be constructed into a portfolio as per profit or risk constraints. To perform clustering we calculated Risk as per the Risk Formula (refer to Formulae 3) by extracting the Beta value for each stock. Then we used the scaler property to compress the profit percentage value and risk value from 0 to 1 (refer to Figure 4) as this helps to plot a smooth graph. By using the K value as 3, K-means clustering is applied to the Compressed value of risk and profit and they are distributed in different clusters (refer to Figure 4) Here K-means clustering is used to iterate 10 times with different centroid values for the most accurate result. At last K-means plot is drawn to graphically visualise the shape of the cluster (refer to Figure 3).

As we can see from the cluster that the stock with higher return makes one cluster which includes stocks like *Bajaj Finance, APL Appolo tubes,* and *Deepak Nitrite*. So for the person who wants a higher return without considering the amount of risk, this cluster is best. So it helps to diversify the portfolio, if LSTM fails due to external factors for specific stocks, then the overall risk can be distributed. Similarly, stocks with less risk are clustered into a single group.

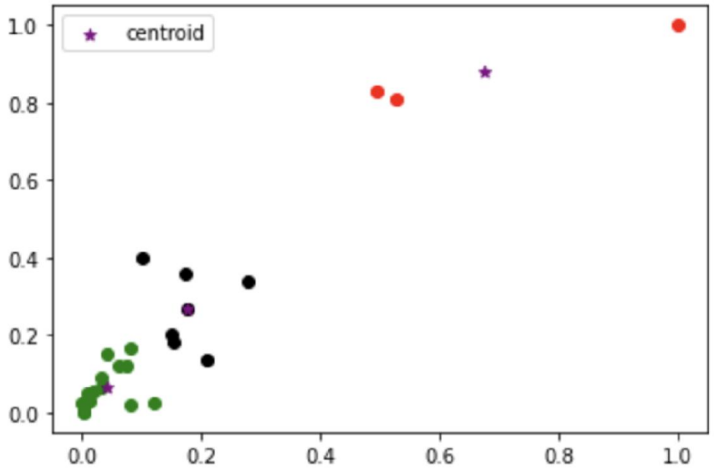

Figure 3. Cluster Graph Plotted with Centroid

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, July 2022
Impact Factor 6.252

	Stock Name	Expected Profit Percentage	Beta Value	Risk Value	Compressed Profit	Compressed Risk Value	cluster
0	Reliance Industries Limited	354.272167	1.16	402.835714	0.268670	0.176232	0
1	TCS	109.131520	0.64	65.364173	0.057694	0.017541	2
2	Infosysis	147.440571	0.67	94.095183	0.090664	0.031051	2
3	Bajaj Finance	1204.030820	1.80	2154.655477	1.000000	1.000000	1
4	Hdfc Bank	179.781751	1.09	188.332109	0.118498	0.075364	2
5	Hdfc	115.685433	0.86	93.469473	0.063334	0.030757	2
6	Wipro	73.815827	0.42	28.062647	0.027300	0.000000	2
7	APL Apollo Tubes	982.664157	1.18	1151.283705	0.809485	0.528179	1
8	Asian Paints	220.761538	0.54	115.431230	0.153766	0.041084	2
9	Relaxo Footwear	504.244098	0.48	238.677167	0.397741	0.099038	0
10	Voltas	233.715911	0.87	197.242843	0.164915	0.079555	2
11	MRF	98.927086	0.50	45.963543	0.048911	0.008418	2
12	Godrej Properties	437.450595	1.44	619.848857	0.340256	0.278279	0
13	BATA	184.525626	0.89	157.997807	0.122580	0.061100	2
14	TVS Motors	74.211017	0.75	50.408263	0.027640	0.010508	2
15	Bajaj Auto	42.095155	0.96	33.691349	0.000000	0.002647	2
16	MPS Infoterence	73.822800	4.26	284.665128	0.027306	0.120664	2
17	Orchid Pharma	67.108046	3.32	199.558711	0.021527	0.080644	2
18	Tata Elxsi	459.253500	0.88	397.983080	0.359020	0.173950	0
19	Finotex Chemicals	202.542164	2.41	471.256616	0.138086	0.208406	0
20	Century Plywood	48.946762	0.81	33.976877	0.005897	0.002781	2
21	Deepak Nitrite	1006.501761	1.08	1079.461902	0.830000	0.494406	1
22	Laurus Lab	255.863507	1.42	353.386180	0.183976	0.152979	0
23	Globus Spirits	273.886313	1.31	349.621070	0.199487	0.151208	0

Figure 4. Cluster Formed using K-means clustering

We have calculated the average risk and expected return for the cluster formed. So that we can predict or expect the amount of return in a given time period by keeping risk factors in knowledge (refer to Figure 5.). The same can be used to derive risk to profit analysis of the portfolio better.

	Cluster Name	Profit Mean	Risk Mean
0	Cluster 1	79.984767	86.511184
1	Cluster 2	1059.340257	1452.657411
2	Cluster 3	324.612814	381.710059

Figure 5. Cluster Mean Value

As we can clearly see, cluster 1 has less profit close to 80 per cent and low risk because it contains stocks which are less volatile and have fewer growth chances. However, Cluster 3 contains stocks which are more volatile and hence have more growth chances. The variation in the mean value of profit and risk is due to the selection of stocks in each cluster. The more the cluster we chose, the lesser the variation will be seen.

VI. CONCLUSION AND FUTURE WORK

The fundamental contribution of this paper is to highlight the risk involved in the stock market. So in this paper, we have predicted the price for various stocks and concluded that price varies from actual and predicted values for some stocks. But the overall mean for the price variation is very less. So to avoid stock risk by investing in a single stock this paper helps to identify the similar stocks as per profit and related risk and suggests a portfolio to reduce the risk of monetary losses. However LSTM model works well with 'Close' value, but as we know the market is very volatile and aggressive, so we can add multiple related attributes like market Cap, Short term external factors, etc which are known beforehand to make the model more reliable. So it will definitely enhance the model and make it more efficient.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, July 2022
Impact Factor 6.252

REFERENCES

[1]. G. Kesavaraj and S. Sukumaran, "A study on classification techniques in data mining", 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), pp. 1-7, 2013.

[2]. Bini, B. S., & Mathew, T., " Clustering and Regression Techniques for Stock Prediction", Procedia Technology, pp. 1248–1255, 2016.

[3]. Li, Y., & Wu, H., "A Clustering Method Based on K-Means Algorithm", Physics Procedia, pp. 1104–1109. 2012

[4]. Nayak, A., Pai, M. M. M., & Pai, R. M., "Prediction Models for Indian Stock Market", Procedia Computer Science, pp. 441–449, 2016

[5]. Gosain, A., & Dahiya, S. " Performance Analysis of Various Fuzzy Clustering Algorithms: A Review ", Procedia Computer Science, pp. 100–111, 2016

[6]. Alqaryouti, O., Farouk, T., & Siyam, N, " Clustering Stock Markets for Balanced Portfolio Construction", Advances in Intelligent Systems and Computing, pp. 577–587, 2019

[7]. Idrees, S. M., Alam, M. A., & Agarwal, P " A Prediction Approach for Stock Market Volatility Based on Time Series Data", IEEE Access, pp. 17287–17298, 2019

[8]. Rahul, Sarangi, S., Kedia, P., & Monika, "Analysis of various approaches for stock market prediction", Journal of Statistics and Management Systems,,pp. 285–293, 2020

[9]. Moghar, A., & Hamiche, M., " Stock Market Prediction Using LSTM Recurrent Neural Network", Procedia Computer Science, pp. 1168–1173, 2020

[10]. Mohebi, A., Aghabozorgi, S., Ying Wah, T., Herawan, T., & Yahyapour, R., " Iterative big data clustering algorithms: A review" , Software - Practice and Experience, pp. 107–129, 2016

[11]. Hotta, K., Xie, H., & Zhang, C., "Component-based nearest neighbour subspace clustering", IET Image Processing, 2022

[12]. Sharma, R., " ANALYSIS OF K-MEANS CLUSTERING ALGORITHM", IRJMETS, 2022