**Combining Time Series Models to compute closing stock price**

Anirudh Balakrishnan

PSG College of Technology, Coimbatore

Email: anirudhbala13@gmail.com

Contents

[Understanding the Data 4](#_Toc163942727)

[Analysis of Data using Correlation coefficient and Distance Metric 4](#_Toc163942728)

[Distance metric to infer relationship between companies 6](#_Toc163942729)

[Principal Component Analysis and Clustering 7](#_Toc163942730)

[Hierarchical Clustering 7](#_Toc163942731)

[Clustering using Minimum Spanning Tree 9](#_Toc163942732)

[Time Series Analysis 10](#_Toc163942733)

[Autoregressive Moving Average (ARMA) 10](#_Toc163942734)

[Autoregressive Integrated Moving Average (ARIMA) 10](#_Toc163942735)

[Akaike Information Criterion 11](#_Toc163942736)

[Combining time series models for prediction 13](#_Toc163942737)

[Convex Combination 13](#_Toc163942738)

[Results for Banking Index 15](#_Toc163942739)

[Results for Auto Index 16](#_Toc163942740)

[Conclusion 18](#_Toc163942741)

[Keywords 19](#_Toc163942742)

[References 20](#_Toc163942743)

**Abstract**

The behaviour of the stock market has always piqued the interest of people, given its highly volatile and unpredictable nature. Being able to predict the trend of the stock market serves as an interesting task, since the stock market is highly influenced by a multitude of economic and non-economic factors. Some of the key factors include the GDP growth, employment rates, consumer spending. Technological advancements can have a massive impact on the industry leading to a shift in the market dynamics. Quite surprisingly, natural disasters such as earthquakes, hurricanes too can disrupt supply chains thereby impacting the stock prices.

In this research paper, we aim to apply the techniques of machine learning and time series analysis to develop a single model which can be used to predict the closing stock prices of BSE Banking Index and BSE Auto Index. Leaving the uncontrollable elements to the hands of nature, what we have in hand is rich data of stock prices and look to utilize them in an efficient manner to predict future stock prices. We work on closing stock price data collected from January 1 2022 to November 30 2023. The task in hand is to be able to accurately predict the closing stock price of all the Banks in the BSE Banking Index and for all the companies in the BSE Auto index.

Initial stages of the research involved understanding the relationships within the banks and within the auto index companies. The strength of relationships is quantified using Pearson’s correlation coefficient. A more intuitive metric known as the distance metric is then introduced in order to infer relationships. We then apply a dimensionality reduction technique, popularly known as Principal Component Analysis. We utilize the reduced data and perform clustering, which will group similar performing banks in one cluster. Similarly, the auto index companies that are similar in performance are combined into one cluster. We utilize two clustering techniques, namely Hierarchical Clustering and Minimum Spanning Tree and compare their performances.

The most vital part post clustering involves finding the best time series model for each cluster. This is an iterative procedure and we finalize the best model as the one having the lowest AIC score (Akaike Information Criterion score). The intuition behind this is explained in the later stages of the extended abstract. We use the individual time series models to predict the closing stock price values for December 2023 and compare it with the actual values. The model is validated using MSE (Mean Square Error).

The final stage of the research involved combining the individual time series models into a single model, that can be used to predict the closing stock price of all the banks in the BSE Banking Index. Similarly, the individual models developed for the BSE Auto Index companies are also combined into a single model to predict the closing stock price for all the Auto Index companies. The models are then used to predict the closing stock price of December 2023 and validated using Mean Square Error. The subsequent sections of the extended abstract describe the flow of the project and the results obtained at each stage.

# Understanding the Data

Data has been collected from the official website of the Bombay Stock Exchange. We are primarily interested in predicting the closing stock price of banks in BSE Banking Index and the companies in the BSE Auto Index. Given below is a snippet of the data collected.



Fig 1.1 Snippet of Stock data collected from BSE Website

# Analysis of Data using Correlation coefficient and Distance Metric

We combine the closing stock prices into a single data-frame and find the correlation coefficient in order to understand the strength of association between them. The correlation coefficient provides a numerical measure of the degree of association between the entities. It also provides the direction of relationship between the variables. A high positive value of correlation coefficient indicates both the entities have a strong relationship in the same direction, while a high negative value indicates that the entities move in the opposite direction. The correlation values obtained are visualized as a heatmap as shown in Figures 2 and 3.

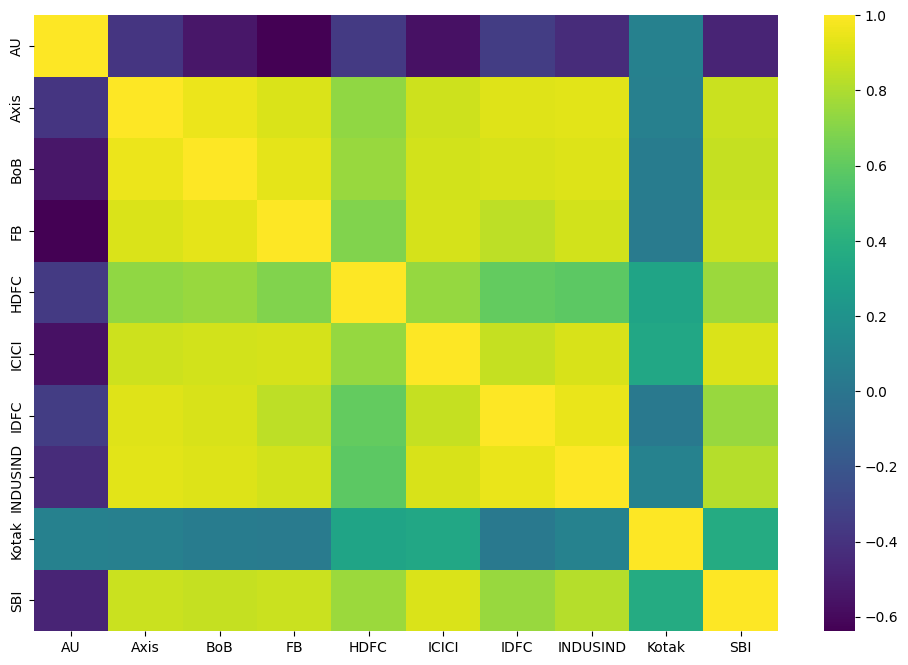


Fig 2. Correlation Heatmap for BSE Banking Index

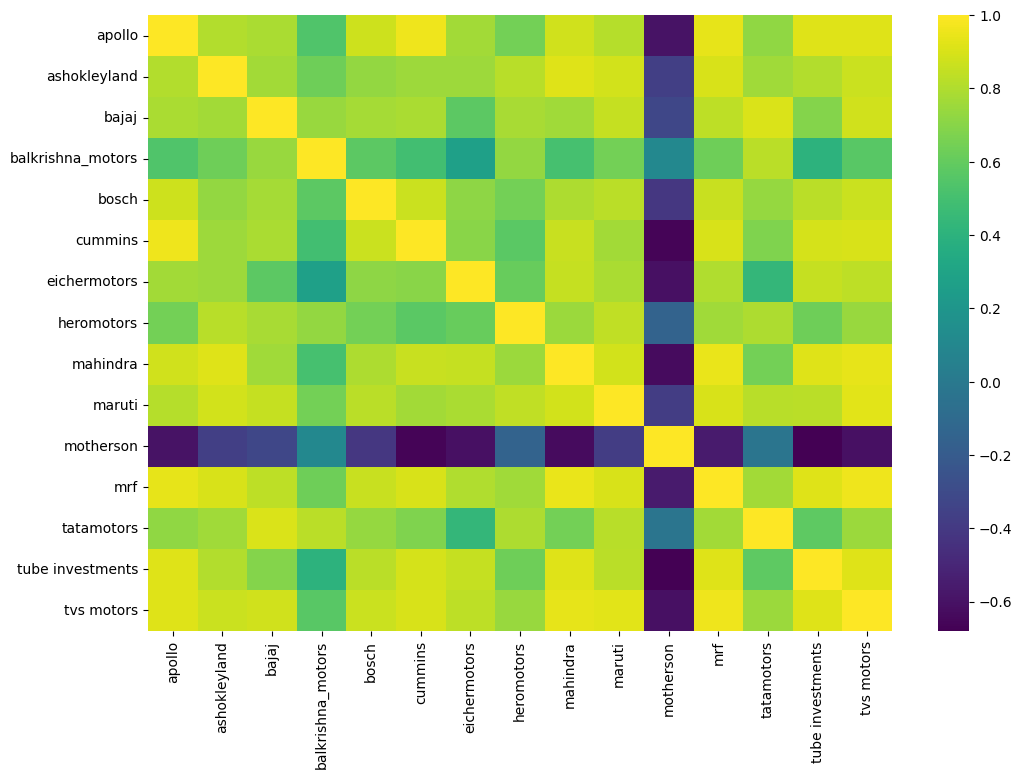


Fig 3. Correlation Heatmap for BSE Auto Index

From **Fig 2** it is evident that Axis bank (Axis) and Bank of Baroda (BoB) have a high positive correlation of 0.95, whereas Axis and AU have a correlation coefficient of -0.38. Similarly, the relationships between the other banks can be inferred.

From **Fig 3** we can infer that TVS and MRF have a strong positive relationship as indicated by the correlation value of 0.95, while Apollo and Motherson have a correlation value of -0.6 which indicates that the two companies are not similar to each other. Similarly, the relationships between the other companies can be inferred.

# Distance metric to infer relationship between companies

The distance metric **d=1−*ρ2*** (where *ρ* is the correlation coefficient) is another way to quantify the relationship between variables. It is inversely related to correlation coefficient (as *ρ increases the distance metric decreases and vice versa).* The distance metric is intuitive and easier to interpret. A value closer to 0 indicates that the entities are extremely similar in their behaviour whereas a higher value of the metric indicates stronger levels of dissimilarity. The distance metric is also robust to nonlinear relationships, thus making it a better metric for comparison than the correlation coefficient. While the results obtained from the correlation coefficient and distance metric are largely same, the distance metric is easier to interpret.

|  |  |
| --- | --- |
| Range of distance values | Type of relationship |
| 0.75<=d<=1 | Weak |
| 0.36<=d<0.75 | Moderate |
| 0<=d<0.36 | strong |

# Principal Component Analysis and Clustering

With a thorough analysis of the relationships within the data, we now aimed to find as many clusters as possible from within the banks and auto index companies. The idea is to club similar companies into a single cluster, which makes it easier to generalize the data and also ease the process of further analysis as will be described in the subsequent sections of the paper. The precursor to the clustering process was to perform Principal Component Analysis on the distance matrices (distance metric values organized as a matrix) of the banks and auto indices respectively. This step was done in order to reduce the dimensions of the date we worked with, with minimum loss of data. The core principle of Principal Component Analysis, is to find a set of “k” linearly independent vectors onto which the original data points can be projected such that maximum variance is captured. What this means is that we should be able to identify the original data points on this new dimension and we must also be able to reconstruct the original data by minimizing the reconstruction error. The distance matrix for the Bankex data is a 10\*10 matrix and for the Auto Index the same is a 15\*15 matrix. Thus, we apply PCA to both these matrices and obtain data of reduced dimensions.

## Hierarchical Clustering

Hierarchical Clustering is a popular method in data mining and machine learning used for grouping similar data points into clusters. It is a recursive approach that looks to partition the data into a hierarchy of nested clusters with each level representing. There are two approaches to hierarchical clustering, namely agglomerative clustering (bottom up) and divisive clustering (top down). In agglomerative clustering, each data point is considered a cluster and points are merged iteratively based on their similarity or proximity until all data points belong to a single cluster. This is the clustering approach that we have used in this research project. We perform agglomerative clustering using the complete linkage method on the reduced data. The steps of this method are as given below.

1. Initialization: Start with each data point as a singleton cluster. Each data point is considered a cluster of size one.
2. Calculate Distance matrix: Compute the pairwise distance between all pairs of clusters. We have used Euclidean distance for the same. In complete linkage, the distance between two clusters is defined as the maximum distance between any two points in the clusters.
3. Merge Closest Clusters: Find the pair of clusters with the smallest distance and merge the two clusters into one.
4. Update Distance Matrix: Recalculate the distance matrix to reflect the newly formed cluster and all other clusters based on the complete linkage criterion.
5. Repear steps 3 and 4 until all points are in a single cluster.

The results of Agglomerative clustering can be visualized using a Dendrogram. A dendrogram is a tree-like visualization that illustrates the order and manner in which clusters are merged, as well as the distance or similarity between clusters.

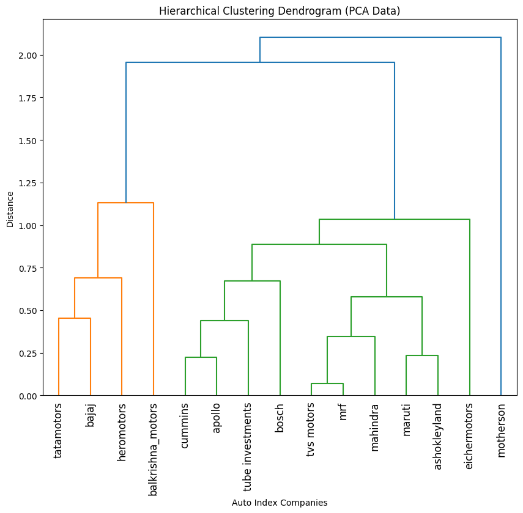
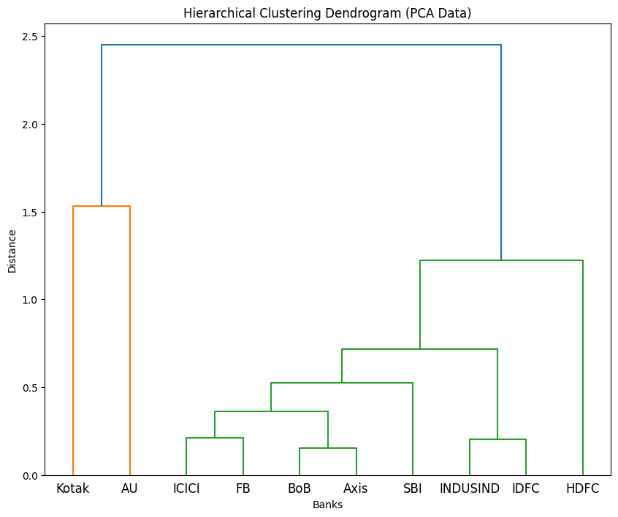
 

Fig 4. Dendrogram for Auto Index Fig 5. Dendrogram for Bankex

From Fig 4 we infer that TVS motors and MRF have the smallest distance between them and hence are a part of the 1st cluster that was identified. Also, the cluster containing Tata Motors, Bajaj, Hero Motors and Balkrishna Motors seem to have a significant level of dissimilarity from the cluster containing all the other companies.

From Fig 5 we infer that Bank of Baroda and Axis Bank have the smallest distance between them and hence are a part of the 1st cluster that was identified. The cluster containing Kotak and AU seems to have a significant level of dissimilarity from the cluster containing all the other banks.

## Clustering using Minimum Spanning Tree

Before delving into time series analysis, an alternate approach to cluster data points by constructing a Minimum Spanning Tree from the reduced data was explored. The minimum spanning tree is constructed and we place a threshold value and try to identify the clusters obtained.

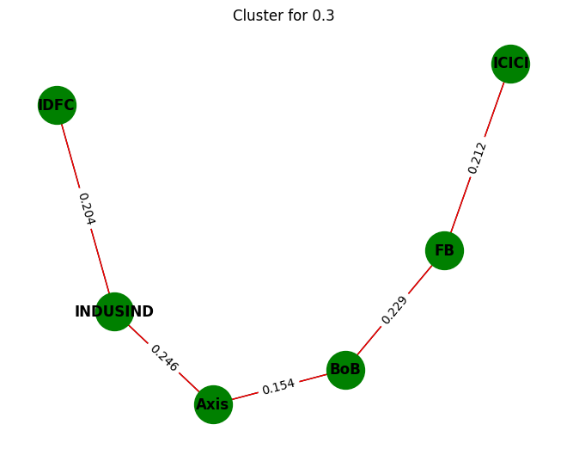


Fig 6. Clustering using MST for Bankex data

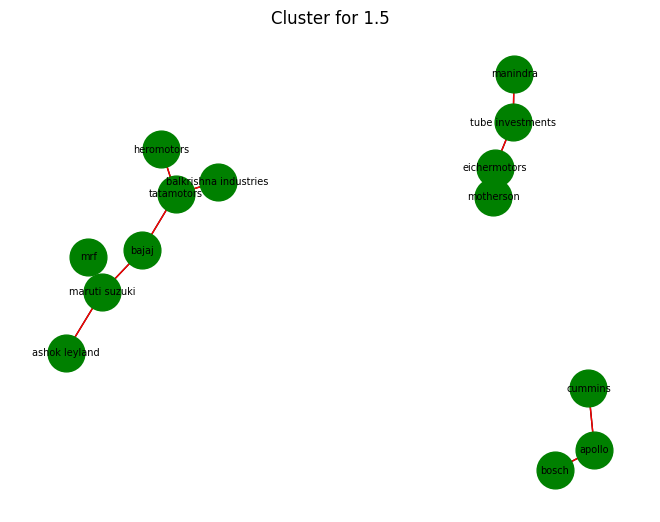


Fig 7. Clustering using MST for Auto Index data

The drawback of the minimum spanning tree is that it only computes the distances once and returns clusters based on the threshold whereas a dendrogram uses a linkage matrix that is updated iteratively and thus is able to infer a larger number of clusters. Thus,for the data we had, hierarchical clustering proved to be the better method as it provided richer insights.

# Time Series Analysis

This section describes the most vital part of the project, which is the development of Time Series Models (ARMA/ARIMA/AR/MA) for the individual clusters. Before describing the methodology let us briefly look at what ARMA, ARIMA, AR and MA models mean.

## Autoregressive Moving Average (ARMA)

* **Autoregressive component (AR)**
  + The ARMA model includes an autoregressive component, denoted by AR(p), where "p" represents the order of the autoregressive process. This component captures the linear relationship between the current value of the time series and its past values. Mathematically, an AR(p) process can be expressed as a linear combination of the previous "p" observations, where the coefficients represent the weights assigned to each lagged observation.
* **Moving Average component (MA)**
  + The ARMA model also includes a moving average component, denoted by MA(q), where "q" represents the order of the moving average process. This component captures the linear relationship between the current value of the time series and its past error terms (or residuals). Mathematically, an MA(q) process can be expressed as a linear combination of the previous "q" error terms, where the coefficients represent the weights assigned to each error term.

## Autoregressive Integrated Moving Average (ARIMA)

* **Autoregressive Component**
  + Similar to the ARMA model, ARIMA includes an autoregressive component (AR(p)), which captures the linear relationship between the current value of the time series and its past values.
* **Moving Average Component**
  + ARIMA also includes a moving average component (MA(q)), which captures the linear relationship between the current value of the time series and its past error terms.
* **Integrated Component**
  + The "I" in ARIMA stands for integrated, indicating that ARIMA models incorporate differencing to achieve stationarity. Differencing involves computing the differences between consecutive observations to remove trends or seasonal patterns from the time series data. The order of differencing (denoted by "d") represents the number of times differencing is applied to achieve stationarity.

Before developing models for each cluster we need to analyse the time series components (trend, seasonality) for each cluster. This analysis showed that the data followed a non-stationary distribution. This means that the statistical properties (mean,variance) do not remain constant with time. So we had to make the data stationary before developing ARMA(p,q) model. The data is made stationary by differencing successive points. In this case, differencing was done once, and the data was found to be stationary. It is easier to analyse stationary data since the statistical properties remain constant.

Now that we have made the data stationary, we now aim to find the best time series model for each bank. This is an iterative procedure and we try to find the order of AR (p) and MA within a range of (0,3). We use the AIC measure (Akaike Information Criterion), as a metric for choosing the best model. The model with the lowest AIC value is chosen to be the best fit for each cluster.

### Akaike Information Criterion

AIC of Akaike Information Criterion, is a measure used for model selection, particularly in the context of statistical modelling. It is designed to balance the trade-off between the goodness of fit of the model and the complexity of the model. In other words, it penalizes models for being too complex while rewarding them for explaining the data well.

AIC= -2 \*ln(L)+2\*k

Where:

* L is the likelihood of the model given the data
* K is the number of parameters in the model

The first term, -2\*ln(L) penalizes models based on how well they fit the data. Models that fit the data well have a higher likelihood (L), resulting in a lower value for this term. The second term, 2\*k penalizes models based on their complexity. More complex models with more parameters will have a higher value for this term.

The goal is to minimize the AIC value. Lower AIC values indicate a better balance between model fit and complexity.

In the context of time series modelling the best model is the one that adequately captures the underlying patterns and dynamics of the data while being as simple as possible. Time series models with low AIC values are preferred because they achieve this balance (they fit the data well without being overly complex).

Once the time series models were obtained, the values for December 2023 were predicted and mean square error was used to evaluate the models.

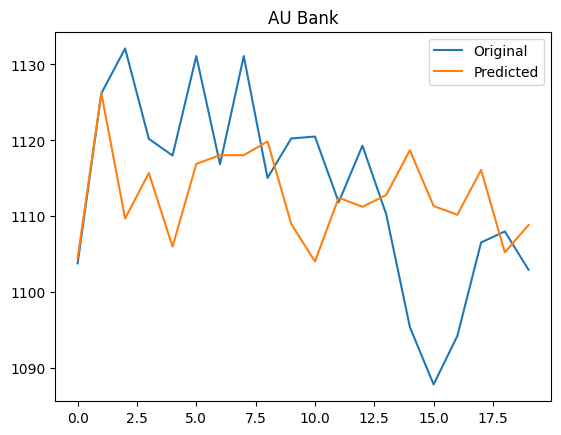


Fig 8. Actual vs Predicted values for AU Bank

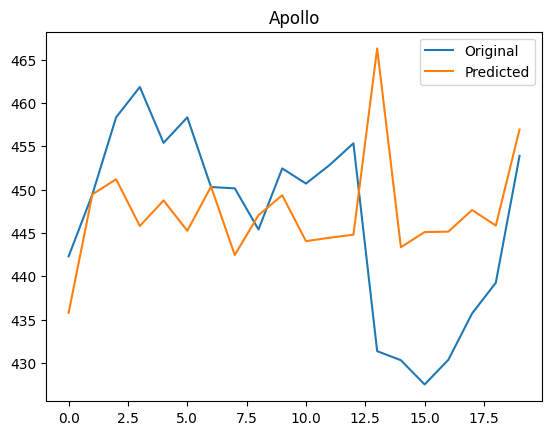


Fig 9. Actual vs Predicted values for Apollo

# Combining time series models for prediction

Now that we have obtained individual models for banks and auto index companies, we tried to extend the idea by trying to combine the individual models into one single model. So, all the models for banks are combined into one single model. Similarly, all the individual auto-index models are combined into one single model. The idea is to combine each model by using different weights for each with the constraint that all the weights should add up to 1.

Essentially, we are trying to find convex combinations of all the companies. We generate random numbers for the weights and normalize them, in order to adhere to the constraint. We multiply the weights with the corresponding values predicted by each model and add them. We do this for the month of December 2023. We obtain all the predicted values and find the mean square error by comparing it with the actual values for December 2023. The weights for which the MSE is the least, is considered to be the optimal weights for each model. However, for each bank, the optimal weights obtained are not the same and differ.

### Convex Combination

Convex combination refers to a way of combining (or weighting) several vectors such that weights are non-negative and sum to one. Geometrically, this results in a point that lies within the convex hull of the vectors being combined.

Mathematically, given vectors v1,v2, v3…..vn and non-negative weights w1, w2, w3….wn , such that w1+w2+……wn=1, their convex combination is given by:

w1v1+w2v2+………wnvn

Here the weights determine how much each vector contributes to the combination. Since the weights are non-negative and sum to one, they represent the portions or percentages of each vector to include in the combination.

In the context of our research the weights imply the contribution of each model to the combined model which is used to predict the stock prices for all the individual banks and auto index companies.

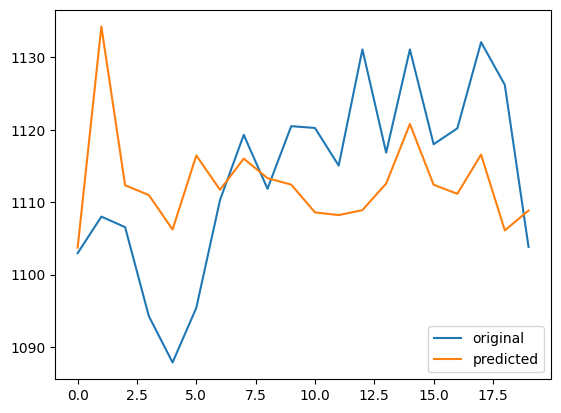


Fig 10. Result of combined model for AU Bank

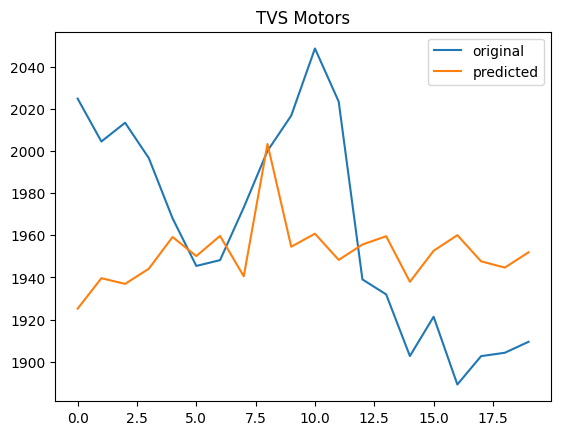


Fig 11. Result of the combined model for TVS Motors

### Results for Banking Index

|  |  |
| --- | --- |
| Cluster | Optimal Weights |
| Axis-BoB | [0.2900908185568677, 0.1524629561451581, 0.16409131743929273, 0.12183493565498126, 0.20030569843004845, 0.0712142737736518] |
| FB-ICICI | [0.0071210689715655836, 0.978916653618481, 0.0015543086689191403, 0.00835130339030742, 0.0004312569048826612, 0.003625408445844083] |
| Kotak | [0.020316246228239077, 0.00018691873677370853, 0.9732368889893044, 0.0001104447084499066, 0.0003186409448512167, 0.005830860392381636] |
| AU | [0.21754372427531643, 0.0013686949201799892, 0.020309733491459935, 0.3318849826022745, 0.2215067034353732, 0.20738616127539602] |
| HDFC | [0.02527495253781422, 0.012124815524673552, 0.1700736760180994, 0.09174695993993652, 0.6474999853155002, 0.05327961066397612] |
| SBI | [0.08265113648584528, 0.3138914650107552, 0.037191688180750214, 0.06779111737540507, 0.010028512925950287, 0.4884460800212939] |

### Results for Auto Index

|  |  |
| --- | --- |
| Cluster | Optimal Weights |
| TVS-MRF | [0.7696281026624708, 0.04501026450062281, 0.03428668599916671, 0.00227786772573437, 0.047679297434965504, 0.004872057381412266, 0.003986463789565384, 0.03108291048693746, 0.02199872114641309, 0.03199886500667489, 0.007178763866036623] |
| Maruti-Ashok Leyland | [4.277537231760183e-06, 0.9999534875623918, 1.0036425412286165e-05, 2.9748579803354468e-06, 2.97788195262637e-06, 3.776862751943092e-07, 5.588135722573045e-06, 8.274650968530693e-06, 2.494198172387867e-06, 5.507568310043062e-06, 4.00349558259615e-06] |
| Cummins-Apollo | [0.0019625253536085124, 0.0015903496699354108, 0.9750297840829868, 8.60333946447482e-05, 0.003833837031966539, 0.0015886253266311483, 0.000816780917533225, 0.0034376364993810634, 0.003050949209916545, 0.0046317862485598885, 0.003971692264836059] |
| Bajaj | [0.015796095474175053, 0.006142189162550027, 0.004511965541722955, 0.7774972451267883, 0.023826714562142706, 0.04379958664819511, 0.026389149812022863, 0.04411258617386931, 0.015390404640585127, 0.03877824971843759, 0.0037558131395110227] |
| Balkrishna Motors | [0.009958228750747483, 0.041439083937512305, 0.028643090569681843, 0.02867301694778767, 0.7432134707519004, 0.0024420686241396583, 0.036335781087989745, 0.00044537675675865844, 0.030787860980253084, 0.020103760193174888, 0.05795826140005418] |
| Bosch | [0.006344480246616851, 0.0006940586536170742, 0.004641655231982205, 0.0022321869508436514, 0.0030815665292795143, 0.975605280657629, 0.002861426515508218, 0.002032175236851427, 0.0009247694428035201, 0.0011813157892732358, 0.00040108474559530176] |
| Eicher | [0.03185588425659057, 0.0026940486476589383, 0.04466431538322062, 0.03507043495559993, 0.030710999302673302, 0.014655994433327463, 0.8087227263219208, 0.0011972152192940182, 0.022662781981189865, 0.004239194482475409, 0.00352640501604908] |
| Hero Motors | [0.017255178454562228, 0.030411241098109044, 0.00769201407280477, 0.04894787561600281, 0.05295514079611845, 0.009752374474660715, 0.05676521352301903, 0.7424733637392292, 0.005018535794838945, 0.015094013913522015, 0.013635048517132846] |
| Mahindra | [0.0033660453014245305, 0.0008915124817573053, 0.0012961618307881, 0.0034748410701140268, 0.0005092303223672294, 0.00030489438232023797, 0.004703381514704245, 0.0024453745413642617, 0.9778019406856379, 0.005114872237459966, 9.174563206220232e-05] |
| Motherson | [4.1533549359068385e-05, 0.0016805145687959736, 0.0011910357543782675, 0.0006938673021915845, 0.0004551690651349763, 6.11853293799498e-05, 0.0027393378474235173, 0.001107033218827176, 0.003960407007575141, 0.9868238893498396, 0.0012460270070946853] |
| Tube Investments | [0.00296454579869432, 0.010013966592252982, 0.006922958175361282, 0.006172189867765016, 0.00708113812902151, 0.0038534575011619186, 4.663617559096405e-06, 0.004062130949084663, 0.005991283121258516, 0.002939297854207022, 0.9499943683936336] |

# Conclusion

At the end of this research, we have two models (one for Banking Index and the other for Auto Index). For each model we have a different combination of weights for different companies. Each combination is used to predict the closing stock price of the respective companies and is evaluated using mean square error. Using a combined model makes the task of prediction easier since we only have to play with the weights. The weights give us a numerical measure of percentage of contribution to the prediction of stock prices.

# Keywords

* Time Series Analysis
* Hierarchical Clustering
* Convex Combination
* Principal Component Analysis
* ARMA, ARIMA

# References

1. Richard I. Levin, David S. Rubin, Sanjay Rastogi, Masood Hussain Siddiqui, “Statistics for Management Seventh Edition.”
2. Trevor Hastie, Robert Tibshirani and Jerome Friedman, “The elements of Statistical Learning”, Springer, 2013.