A Comprehensive Analysis of Indian Stock Market IPOs

I. Introduction

Understanding the evaluation of companies before investing in them is one of the abilities an investor can have to close a good profitable deal. This can affect the returns of an investor and the ability to hire shares of valuable companies thereby raising funds for future. For our analysis on profitable investing, we will be studying about IPO (Initial Public Offering). An IPO or a stock market launch is a public offering where shares of a company are sold to public or institutional investors [1]. The IPO occurs only when a company is offering people to buy its shares for the very first time for public ownerships. This act makes the company public. In a primary market, any investor is capable of buying shares of companies which are issued for the first time by processing information on IPO. The offerings of the company are listed on a stock exchange. The funds required for buying the shares are directly transferred to the companies in order to raise the equity capital of that particular company. In return, the company provides assurity to the investor by claiming to have profitable closing. When we as investors buy the shares of a particular company, we become the shareholder of the company. The prices of shares are not fixed over time and may rise and fall due to various company factors.

Problem Definition: Why an analysis on IPO required? To answer this question, we must understand the working of IPOs. When a company displays its finances/securities to general public, the money paid to company as a result of buying shares goes directly to private investors as well as the company for the purpose of increasing working capital [1]. Hence, the IPO allows any company to attract a wide pool of potential investors for its future growth. Moreover, the sells once bought by investors cannot be resold to the company. The investors thus have to step forward into unpredictable nature of the IPO market. For this reason, an analysis on companies' growth and future prospect needs to be taken into consideration before buying their shares.

Along with this, as mentioned in [2], in regard to the pricing process for Indian IPOs in particular, some public issues managed during the initial period may be overpriced. Thus, it becomes important for investors to have a detailed information before buying shares. In addition to this, the Indian primary markets, funds are locked in for at least three months. It should be interesting to see whether this factor induces additional underpricing [2].

Also, the Indian primary markets have witnessed a boom during the last few years. According to the statistics in [2], both the number of new issues coming to the market and the total amount sought to be raised have increased in leaps and bounds. In 1992-93, 528 companies made public issues in the Indian capital markets and raised the equivalent of 1.955 billion [2]. Of the 528 issues, 231 issues were made by new companies, which raised over Rs.32120 million from the investing public. In 1993-94, there were 713 public issues of equity [2]. At least 300 of these were in the form of initial public offerings (IPOs). In comparison, the annual number of IPOs in US averaged 516 during the 80s and is currently reckoned to be about 322 in the first half of 1990s. Furthermore, there is consistent increase in the number of potential investors in India with passage of time [2]. Estimates show that there were at least 15 million individual investors in India at the end of 1993 [2]. For all the reasons stated above, it becomes necessary to analyze the company IPO data before buying shares of the company.

Importance of solving the problem: Due to volatile tendency of markets, the rise and fall of prices can take place in short periods of time. All the companies in IPO function differently and it may happen that market prices for one company provides better fortune than another. For predicting which IPO shares can give rise to early profits, the investors need to make a detailed analysis before investing. For this analysis, various aspects of an IPO need to be addressed and studied. These include a brief study on Listing day prices, Bidding price, Profit margins etc. If thorough analysis is performed on the data consisting of companies with IPO processing, the investor can extract knowledge that investing/buying in Which IPOs will lead to monetary growth in future. For this analysis we will use various Clustering algorithms to make our observations.

II. Dataset

For our analysis of clustering algorithm on IPO, we have selected Indian stock IPO dataset available on Kaggle [3]. By doing analysis on this dataset, it will help an investor to find out if he should start bidding for whichever companies based on various parameters. This dataset consists of companies listed in the Indian Stock Market from August 2006 till the present [3]. Some of the features include Company Name, Listing Date, Listing Day stock opening price, closing price, high, low, number of times subscribed by retail investors, etc. To find out the profit margins for each of the company and along with that, how beneficial they are to investors, a comparison is made between initial prices on the day of bidding and final price on the listing day open.

The dataset consists of Listing day entries for two types of stock exchanges of India namely NSE and BSE. The two main stock exchanges of India are NSE (National Stock Exchange) and BSE (Bombay Stock exchange). Discussing about difference between them, the number of companies listed in BSE is much higher than those in NSE [1]. But when it comes to trading of volumes, NSE gets the better of BSE. Due to this, discovery of prices becomes much more easier in NSE.

The raw dataset consists of 49 features and 1026 data instances [3]. For our analysis, we need to find out that investing in which set of IPOs give more profit when the stocks are sold. For this purpose, we will not require all the features available in the dataset but only few important ones. The important features taken into consideration are Bid Price From, Bid Price To, NSE Listing day open, NSE SME Listing day open, BSE Listing day open, BSE SME Listing day open. These features will be useful for estimation of profit procured by comparing bidding prices and Listing day gains. Note that we are using listing day prices of both NSE and BSE stock exchanges. This is because price of stocks may vary in both of them and some investors may wish to do trading or buy stocks from both the stock exchanges. However, we must keep in mind that only few companies permit trading of shares in BSE. We do have NaN values in some important features in our dataset. For simplicity purpose, we convert these values to 0.

For a multivariate analysis at a later stage, we would also use some of the remaining 16 features along with the ones mentioned above, namely NSE Listing day low, NSE Listing day high, NSE Listing day last trade, NSE Listing day volume, NSE SME Listing day low, NSE SME Listing day high, NSE SME Listing day last trade, NSE SME Listing day volume, and for BSE, BSE Listing day low, BSE Listing day high, BSE Listing day last trade, BSE Listing day volume, BSE SME Listing day low, BSE SME Listing day high, BSE SME Listing day last trade, BSE SME Listing day volume. Note that SME (Small and Medium Enterprises) is a platform of NSE where SMEs whose post issue paid up capital is less than or equal to Rs. 25 crores. As explained in [4], the platform is expected to offer a new and alternate asset asset class to informed investors having longer investment horizon. The platform shall allow new, early stage ventures and small quality companies to raise much needed growth capital as they grow, mature and transit to NSE's main board [4]. This platform is being founded on the following 4 cornerstone pillars of Credibility, Transparency, Liquidity and Growth.

III. Methods and Tools

A. Preprocessing Steps

Here we list out some preprocessing steps implemented in our analysis. A detail explanation for the same is discussed in Results and Explanations section.

- Grouping of related attributes in different lists for analysis purpose.
- Missing value imputation by replacing NaN values with 0.
- Type casting of important attributes for converting some string values into float.
- Finding Average of different lists with respect to feature values for analysis purpose.
- Minmax Scaling to transform values between 0 and 1 which is also the default range of MinMaxScaler.

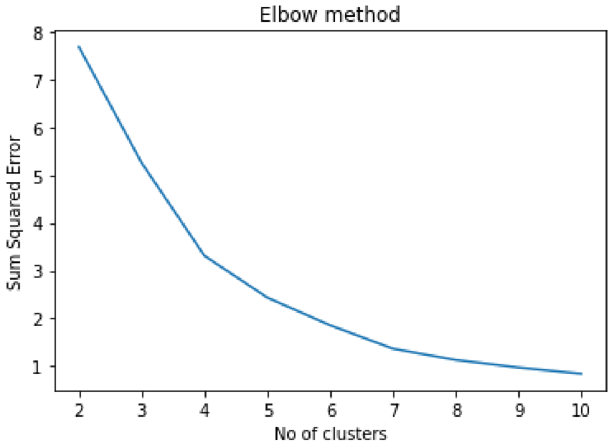

Fig. 1. Elbow method for Optimum K

B. Learning Methods

In this section we will briefly explain the methods which we implemented in our project. The practical analysis of the dataset with respect to our project is discussed in Results and Explanations section at a later stage. There we try to extract meaningful knowledge about Indian IPOs by using clustering analysis and graph plots for visualization.

1) Elbow method for Optimum value of K

The most important task we need to accomplish first is finding right number of clusters. For K-means algorithm, it means finding the right values of K or optimum values. For this purpose, firstly we will look at elbow method for finding number of clusters. Elbow method does the work for computing a metric which tells us how good the clustering algorithm. From these assumptions, we plot different values of K. Initially we observe that there is a rapid change in the number of clusters. However, after certain iterations we observe that the value for number of clusters is not changing and it remains constant [5]. The point at which this value starts remaining constant and not changing anymore is called the Elbow point. For this reason, method for finding optimum value of K is called Elbow method referencing an elbow of human being [5].

There are many metrics which can be used for calculating change in K. For our analysis, we are using Sum of Squares metric. This metric can be termed as average distance of every point in dataset from its corresponding cluster center after clustering process. Thus, it is the distance of each datapoint from its corresponding cluster centroid. A smaller value of

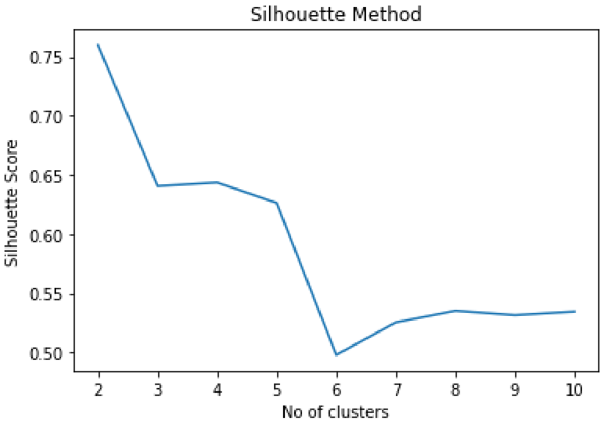

Fig. 2. Silhoutte score for optimum value of K

Sum of Squares metric is better because at that point, the clusters tend to be cohesive or better formation of cluster takes place [5]. This is because points are closer to their centroids. In our analysis, we found that when the number of clusters or Elbow point is 5, there was a decrease in change of value with respect to Sum of Squares metric. Thus, optimum value of K according to elbow method is 5. Hence, we have one value K = 5.

2) Silhouette score for optimum value of K

In this method the metric used for calculating optimum value of K is called Silhouette Score. This metric is calculated using 2 quantities A and B [5]:

- A = Mean of Intra cluster distance Suppose we want to calculate score for any particular point in a cluster. Then A can be calculated as mean of distance between the particular point and all other points in the cluster.
- B = Mean nearest cluster distance For mean nearest cluster distance, we find another cluster nearest to the data point of different cluster for which we want to predict silhouette score. Then we find the average distance between the points and all other points from the nearest cluster [5]. Mean of this average distance is value for point B

Now the Silhouette score = (B-A)/MAX(A,B)

Here we want B to be larger and A to be small for optimum value of K. We can find the silhouette score for each point in the dataset by taking its mean value. For our dataset, after plotting silhouette score for finding number of clusters, we observed that silhouette score is highest for K value of 2.

For our dataset the optimum values of K = [2,3,4,5]. However, we will not use the value of K = 2. This is because only 2 clusters will not be able to justify the clustering of our dataset. Hence, for accurate clustering of our dataset, we will take optimum values of k = [3,4,5]. By observing graph for silhouette method, we can say that there is not much change is values of silhouette score for K values of 3 and 4. Also from the Elbow method, we have found K=5 as optimum number of clusters.

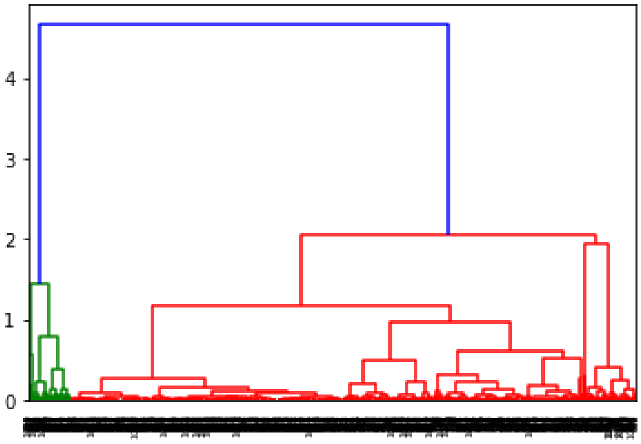

Fig. 3. Dendrogram for Visualizing relation between Clusters

3) Dendrogram for Hierarchical Representation of Data

Dendrogram is a visual representation of the hierarchy developed among the points in the dataset. It is used with hierarchical clustering for visualizing how different clusters are related to each other having similar or different features. Initially each datapoint is taken as a cluster. Then at each level two or more clusters are combined with each other to form a different cluster comprising of other different clusters. These grouping can be done in relation to the features of data points and similarity among them. If the clusters are close to eachother, then they can be grouped as one cluster as a whole. An important use of cluster is to allocate datapoints to clusters by observing the feature difference among them.

Another way to visualize or interpret a dendrogram is to observe the height at which two clusters are joined together. This height can indicate the order in which two clusters are combined with one another. A clearer dendrogram can be created if the heights of layers can differentiate the distance between two clusters. In this way we can deduce how to clusters are different from eachother.

A Dendrogram can also be used to find optimum value of K for a particular dataset. This can be done by drawing a horizontal which can intersect vertical lines of a cluster. The number of times horizontal line pierces through the vertical line, is the number of K values. However, it is important to note that this process can only be performed when there is visible distance among the clusters. If the vertical lines are too close to each other, then there is no need to pass a horizontal line through them. Also, we cannot determine the total number of clusters using dendrogram, but we can find optimum value of K for Kmeans and hierarchical clustering algorithms. However, Dendrograms can give a brief idea about number of clusters if we do not have true evidence to support the number of clusters.

4) Agglomerative Hierarchical Clustering

As we have seen how dendrogram is used to visualize the clusters that are placed in a hierarchy, agglomerative hierarchical clustering method is used to create clusters of that type. Agglomerative clustering is a type of Bottom-Up approach for creating clusters. In this algorithm, suppose we have a total of X objects/datapoints. Initially we assume that each object is a cluster in itself so we have a total of X clusters. Then we find the distance between each pair of clusters. If the distance between the two or more clusters is very less or the clusters are closest to eachother, then these clusters are merged into a single cluster. After completion of this step we have X-1 clusters as one cluster is reduced due to merging process. With this reduction, there is a need to update distance again as new cluster is formed by merging of 2 clusters. In this way we keep on repeating these steps until all the clusters are merged with each-other. Hence the repletion continues until only a single distance measure remains. This method gives rise to a diagrammatic structure called dendrogram.

Note that here, the distance between two clusters is found by observing the similarity of features among the objects. If two objects are only similar to each other if they are having similar features. Hence, using agglomerative clustering technique, we can analyze whether an object falls into which cluster by observing similarity distance between the clusters. In this way we can also differentiate among the type of clusters and which features are responsible for their characteristic.

An advantage of hierarchical clustering is that is easier to understand and implement. If the dataset is comparatively smaller, then this is a very useful method as smaller clusters are generated which can be very easy to visualize by discovering similarities and differences. With the visualization of dendrogram, we can understand the bigger picture of the dataset. However, there are disadvantages if there is noisy data and so the merging once completed cannot be reversed.

5) Birch Clustering

The BIRCH clustering algorithm stands for Balanced Iterative Reducing and Clustering using Hierarchies. This method is designed for clustering a large amount of numerical data. It is a kind of mixture of hierarchical clustering and other clustering methods such as iterative partitioning [6]. BIRCH method helps to overcome the limitations of hierarchical clustering such as scalability of data and the inability to re-merge the data in case of noisy data or outliers. Note that BIRCH algorithm can only be implemented on continuous variables whose values can be represented on coordinates. The 2 concepts introduced in BIRCH algorithm are [6]:

- Clustering Feature (CF)
- Clustering Feature Tree (CF Tree): This tree is used to summarize the representation of clusters. It is a height balanced tree which stores the Clustering Feature (CF) for hierarchical clustering.

Before we proceed with CF and CF tree, we first need to find some measurements for clustering process. Given a d-dimensional datapoints/objects in a cluster, we first need to find [6]:

- Centroid i.e., datapoint located in the middle of the cluster.
- Radius of the cluster: Calculated as average distance from the member object to the centroid.
- Diameter of the cluster: Average pairwise distance within a cluster.

Both Radius and Diameter represent the tightness of the cluster around the centroid which is the center point of the cluster

Furthermore, a cluster feature (CF) has following parameters:

- Number of points in the cluster N
- Linear Sum of the N points denoted by LS
- Square sum of the datapoints denoted as SS

A CF tree is a tree where new points are inserted in an incremental way. It consists of all the CFs. All the non-leafy nodes in this tree has descendants or children [6]. The leaf nodes does the work of storing the sum of the CFs of their children. This is helpful in summarizing the clustering information about the children. The CF tree has 2 parameters:

- Branching Factor: It specifies the maximum number of children per non-leaf node
- Threshold: It specifies the maximum number of diameter of subclusters stored at the leaf node of the tree.

BIRCH applies a multiphase clustering technique these phases are as below [6]:

- Phase 1: BIRCH scans the dataset to build an in memory CF tree, which can be viewed as multilevel compression of the data which tries to preserve the inherent clustering structure of the data. So an initial CF tree is built in phase 1 of BIRCH.
- Phase2: In phase 2, BIRCH applies a selecting clustering algorithm to cluster the leaf nodes of the CF tree. Any clustering algorithm such as K-means can also be applied to categorize the data while minimizing the input output operations using BIRCH.

In BIRCH algorithm, for each input of datapoint, we first find the closest leaf entry. Then we add that particular point to leaf entry and update the CF. If the entry diameter for each input of datapoint is greater than the maximum diameter, then we split the leaf node as well as the parent node.

Advantages of BIRCH clustering algorithm include scalability of dataset i.e., clustering on very large dataset can be performed easily using BIRCH. Also, better quality of clusters can be created than other clustering techniques. A disadvantage is that the algorithm does not perform well when the clusters are not spherical in shape.

6) K-Means Clustering

K-Means Clustering technique is a method using which we can reduce the distance of points within a cluster by taking an object as a centroid of cluster. By doing so, we can make the cluster as compact as possible. In this algorithm, we calculate the distances from a point or a centroid to another datapoints to assign it to a cluster. Hence it can be called a distance based or a centroid based algorithm.

K-means algorithm works in steps to form clusters which are associated to a centroid. The first step of K-means clustering is to choose a value for the total number of clusters or value of K. For our dataset we have already found optimum values for K which is [3,4,5]. Let us assume K = [3] for understanding. In the next step we randomly choose a centroid equal to the number of K. For K=3, we will randomly select 3 different centroids or 3 points which are farthest to each other. For all the remaining points, we assign them to each of these 3 centroids to form 3 different clusters. Now after creation of these 3 clusters, we will find a new centroid for each of these clusters. This centroid can only be a datapoint which exists in that cluster. The new centroid can be formed by calculating the distance of each point to all the other points. The point which is closest to other points in a cluster is the new centroid. We can repeat the step for finding new centroid until the new centroid does not change and that point becomes the final centroid of the cluster. After performing multiple iterations, if we are obtaining the same centroid then we can finally say a cluster is formed. When 3 clusters are finally formed, all the different points from these clusters remain in their same respective clusters.

Advantage of implementing K-means is that if we have a greater number of features, then K-means algorithm performs faster computationally than hierarchical clustering. Additionally, K-means clustering produce more compact clusters where objects in the clusters are tightly related to eachother. This is an advantage of distance-based algorithm. A disadvantage of K-means clustering is that the data is not structured in proper way as compared to hierarchical clustering. This may result in a difficulty while finding clusters as the data is unstructured.

7) Multivariate Analysis using K-Means

To finally visualize the data with respect to each feature and coming to conclusions, it is necessary to find out how the clusters are grouped and their relation with all the features. The clustering techniques we implemented before are useful for the process of dimensionality reduction but, it is just one step towards practical analysis of dataset [7]. It becomes important to find out characteristics of each clusters/groups with respect to multiple features. Hence, the name Multivariate analysis is given to this technique.

In this method, more than two features are chosen and analysis is done by visualizing the behavior of groups with respect to those features [7]. As we do not have any target vector for our dataset, we will try to understand patterns for various groups with respect to different attributes. The algorithm works by assigning an identity or cluster to features. When the features which are important for our analysis are assigned with this identity, then we can extract valuable information from the data provided [7]. For clustering, we are using K-means clustering and for a meaningful visualization of data, we will be plotting a Line Polar graph for multiple variables. A more detailed explanation with respect to our project is given in Explanation part in Results.

C. Tools/Libraries

For the analysis we use the following tools and Libraries:

- Google Colab for coding: It is a free Jupyter notebook environment that stores all the data on cloud. It allows us to combine and execute the executable code and text in a single document.
- Numpy library for mathematical calculations and working with arrays
- Pandas library for python programming required for data analysis and manipulation.
- sklearn.metrics library for calculating Silhouette score
- sklearn.cluster library for applying clustering algorithms
- matplotlib for plotting graphs for each methods
- scipy.cluster.hierarchy library for visualizing dendrogram
- plotly.express library for implementing line polar graphs

IV. Results and Explanation

The Results of our Clustering Analysis on Indian Stock Market IPOs are as follows:

A. Explanation for Pre-processing of data

- Dataset is selected by filtering out the irrelevant features and only keeping the relevant features.
- Dataset contains features which varies and are not related to other IPOs. Hence, all NaN values are imputed with 0 instead of mean/median.
- All the highly relevant features such as Bid Price, Open Price, High Price, Low Price, Last Trade (Close) Price are averaged and Volume is summed up so that all the computation can be performed on single values for each relevant feature.
- The percent change between the Bid Price Average and the Open Price Average is used to measure Profit Margin. To analyse and cluster the data, the Profit Margin value and Bid Price Average value will be used in all subsequent clustering methods.
- Min-Max Scaling is used to scale all of the values from 0 to 1.

B. Comparison of Clustering Methods

By further comparison between the methods implemented on our dataset, we can say that there is not much difference in clusters for Agglomerative hierarchical and K-means Clustering. However, a minute noticeable difference visible by analyzing the graphs, is that K-means clustering tend to make more compact clusters than Agglomerative clustering. This can be because of the distance-based method used by K-means clustering. But BIRCH clustering method seems to do much better clustering than other two. Group of IPOs in BIRCH tend to include a greater number of datapoints/IPO in their cluster. So, clusters created by BIRCH are larger than those created by Agglomerative and K-means. Thus, making visualization of the dataset much easier.

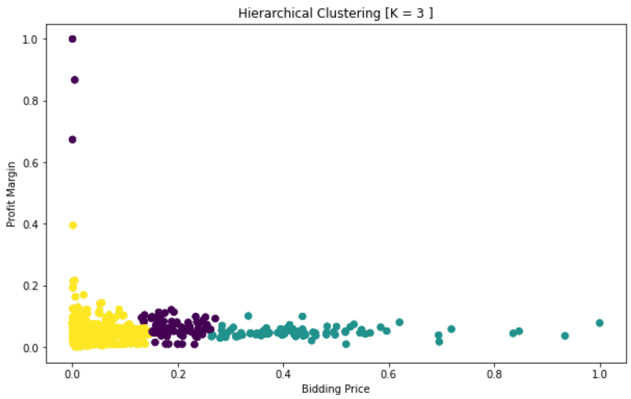

Fig. 4. Agglomerative Heirarchical for K = 3

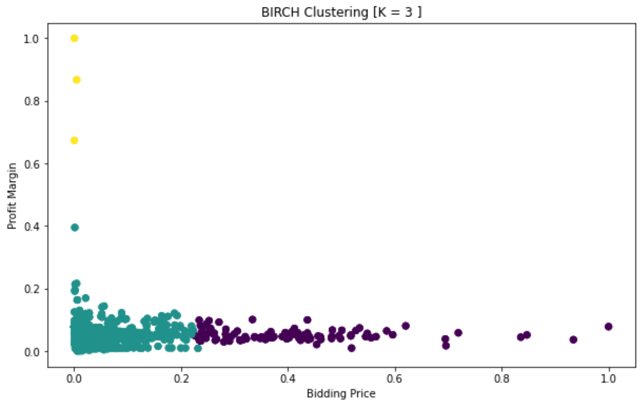

Fig. 5. BIRCH Algorithm for K = 3

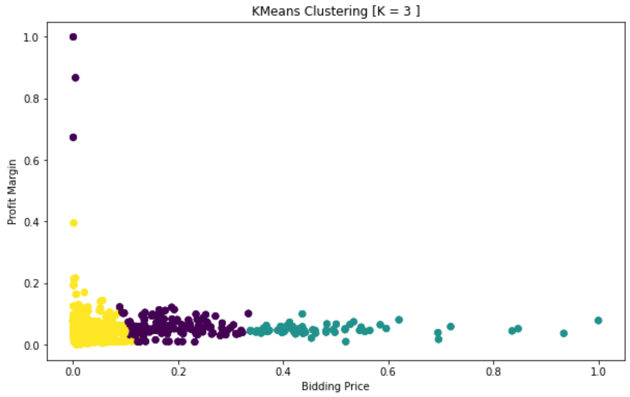

Fig. 6. KMeans for K = 3

C. *Analysis of IPOs on Profit Margins and Bidding Price*

We plotted graphs for Profit Margin Vs Bidding Price to visualize analytical observation on group of IPOs. The different colors in the graph indicate sets/groups of IPOs which are converted to clusters using the Clustering methods. Each data point in the cluster represents a single IPO. Hence, for K = [3,4,5], we have 3, 4 and 5 groups of IPO clusters respectively.

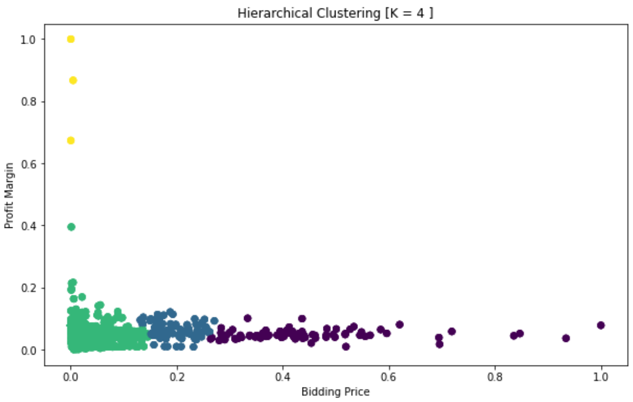

Fig. 7. Agglomerative Heirarchical clustering for k = 4

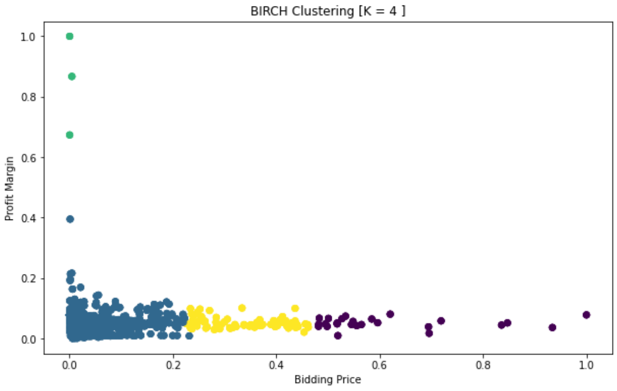

Fig. 8. BIRCH Algorithm for K = 4

:

By analyzing graphs, we can observe that there are many IPOs having most of the profit margin in the range of 0 to 0.2 for their bidding price. Hence only a little bit of profit is obtained by investing in some group of IPOs. In most of the cases, IPOs investing in lower number of bids (in the range of 0 to 0.2) made a good profit with respect to their individual bids. Number of IPOs bidding higher were much less than IPOs bidding lower. As the bidding price range goes from 0.2 to 0.8, we can observe that number of IPOs keeps getting lesser. So, we can say that majority of IPOs do not prefer to invest in stock of higher bidding price. However, profit margins also increased a little bit when higher bids were raised.

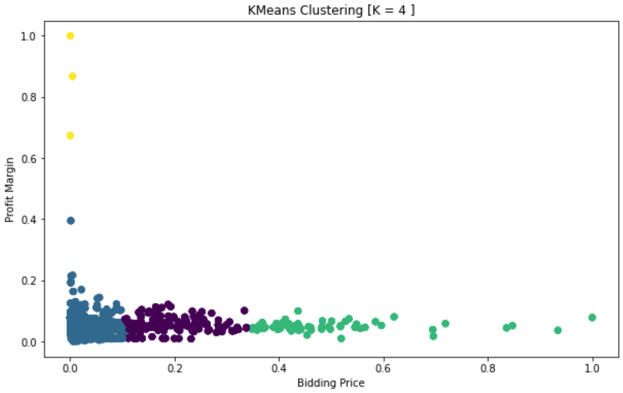

Fig. 9. KMeans for K = 4

D. *Multivariate analysis for analyzing behavior of IPOs on different attributes*

For practically analyzing how IPOs behaved with respect to multiple important variables, we performed a multivariate analysis on the dataset. We used K-means as clustering algorithm along with line Polar graph for meaningful visualization.

The features included in the analysis are Profit margin, Bid Price Average, Open Price Average, High Price Average, Low Price Average, Close Price Average, Volume Average. For K = 3, the clusters of IPOs are represented by Blue, Red and Green labels

By observing green set of IPOs, we can say that their Bid Price Average, Open Price Average, High Price Average, Low Price Average, Close Price Average have almost equal values. However, the Profit Margin and Volume Average are relatively very low for this group of IPOs. The red group of IPOs bidding take place in a medium range (Not too high and not too low) and the values of their Bid Price Average, Open Price Average, High Price Average, Low Price Average, Close Price Average were also similar. But this group of IPOs made slightly better profit than green IPOs. Hence, we can say that higher bids do not guarantee better profits in the long run. Furthermore, the blue IPOs group has low values for all the attributes. So, there can be a possibility that this group belongs to Small and Medium Enterprises. (SMEs)

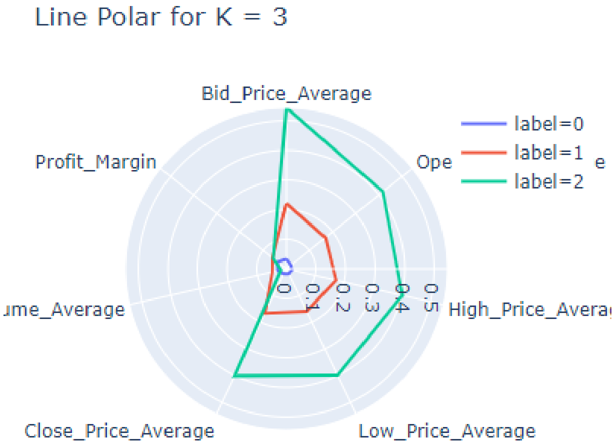

Fig. 10. Multivariate analysis for K = 3

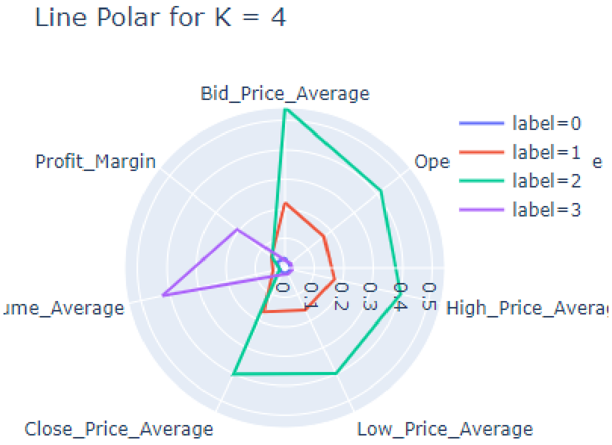

Fig. 11. Multivariate analysis for K = 4

Notice when the value of K increases to 4, a new purple cluster of IPOs is created. This group of IPOs have higher value of Volume Average and along with this, the profit margin is also increased while neglecting any increase in the values of Bid Price Average, Open Price Average, High Price Average, Low Price Average, Close Price Average. This is because the Volume Average value measures the number of shares traded in the stock. A high value of this feature indicates strong market value. Rising markets are directly dependent on the Volume of shares and a rise in market is viewed as healthy profits. So, cluster of IPOs having high Volume Average will also have better Profit Margin which can also be observed in our line polar graph for K = 4.

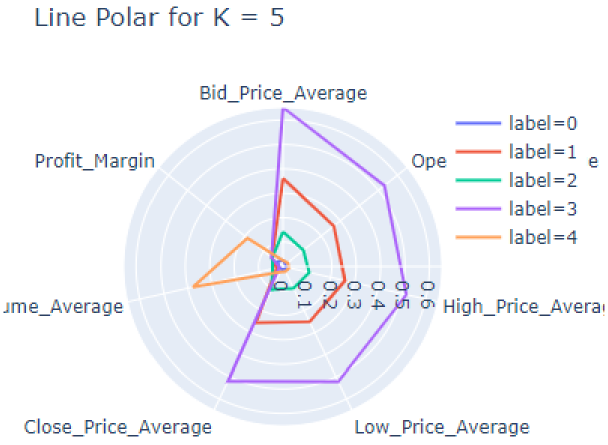

Fig. 12. Multivariate analysis for K = 5

Furthermore, by observing the pie charts, we can say that maximum number of IPOs are present in the blue cluster of IPOs which is around 80 percent of total number of IPOs. Next set of IPO cluster is the red colored one which comprises of 13 to 15 percent of IPOs. At last, the smallest cluster of IPOs have around 3 to 4 percent of total IPOs.

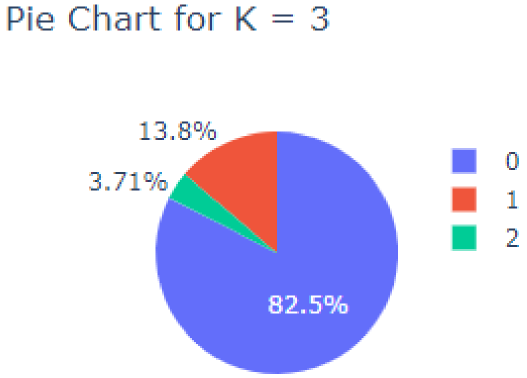

Fig. 13. Pie Chart for K = 3

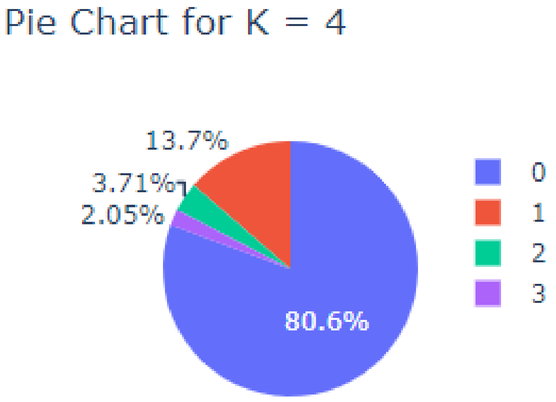

Fig. 14. Pie Chart for K = 4

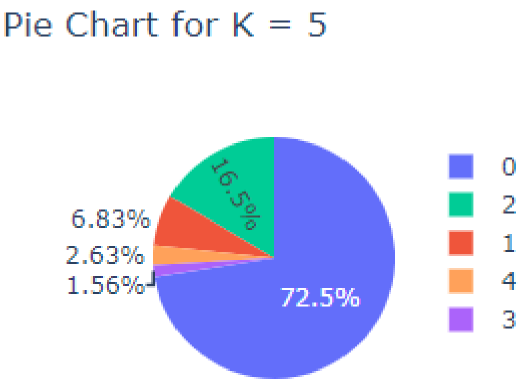

Fig. 15. Pie Chart for K = 5

E. Graphs for Clustering methods where K = 5

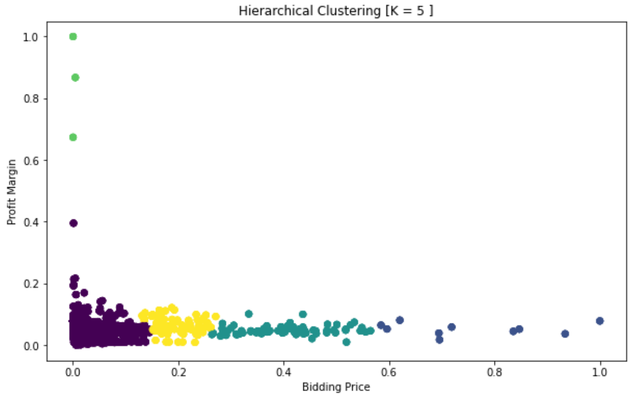

Fig. 16. Agglomerative Heirarchical Clustering for K = 5

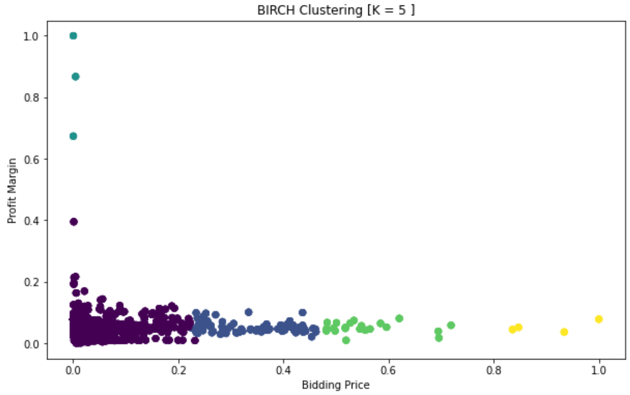

Fig. 17. BIRCH Algorithm for K = 5

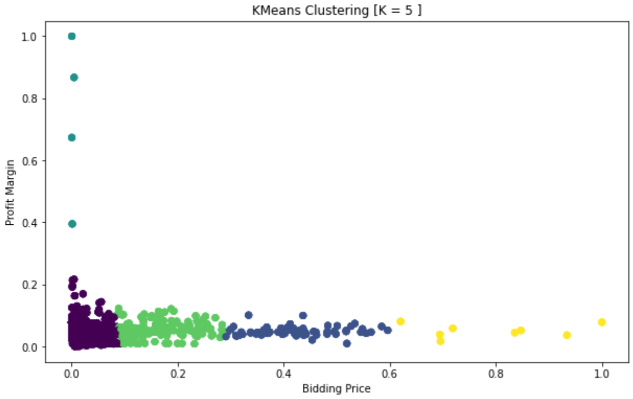

Fig. 18. KMeans for K = 5

V. Discussion and Conclusion

A. A summary of what has been done

We looked at data from Indian stock IPOs (Initial Public Offerings) on the BSE (Bombay Stock Exchange) and NSE (National Stock Exchange) for companies classified as SME (Small and Medium Enterprises) and Non-SME (Non-Small and Medium Enterprises). IPO's are distinguished based on investment by individual based on bidding price to profit margin as on first day of listing the stock in markets. There are multiple factors affecting stock price on listing day, hence we applied dimensionality reduction for features that are most prominent to determine the class for the stock.

B. A summary of the results

We had categorized IPO's into various classes based on their performance on listing day. It is similar to grading IPO's based on their performance in stock exchange market. Each IPO's has characteristics that can be used to identify them for their listing day performance.

C. The significance/implications of results

Results can be used to grade future IPO's , so as to calculate the risk factor and profit margin on Listing day. It can be proved very beneficial for individual investors who hold IPO's for short term and likely to perform intraday trading. This system can help classify new IPO's with very low processing and can be primarily useful for analyzing behavioral pattern for predictions.

D. Limitations

There are various factors affecting the listing day gain of IPO such as market behaviour, company news, economical structure and global trade. This analysis is mostly based on factors that are few of the most prominent. Also, current analysis is trained and limited to IPO's in Indian market only. IPO's are also very scarce comparatively over time and hence data is limited and can be vague on long run with changing trends of market. Currency volatility, Financial education, Political and Economic reforms can affect prominently on future assumptions too.

E. Future work

Current analysis is based on just listing day gains and classify them accordingly. However, performance of stocks are needed to be evaluated for bigger time frame. This project can be further developed to analyze gains for long term investment and can be filtered based on IPO sector, subscription, stake holders, etc.

References

[1] 1 IPO Wikipedia: URL https://en.wikipedia.org/wiki/Initial public offering
[2] Krishnamurti, C., and Kumar, P. (2002). The initial listing performance of Indian IPOs. Managerial Finance, 28(2), 39–51. doi:10.1108/03074350210767681
[3] Kaggle Dataset Indian stock IPO Results URL https://www.kaggle.com/adityap10/indian-stock-ipo-results
[4] NSE India: URL https://www1.nseindia.com/emerge/smeaboutus.htm
[5] Elbow method and Silhouette score:URL https://towardsdatascience.com/clustering-metrics-better-than-the-elbow-method-6926e1f723a6
[6] BIRCH Algorithm: Cory Maklin, URL: https://towardsdatascience.com/machine-learning-birch-clustering-algorithm-clearly-explained-fb9838cbeed9
[7] Multivariate Analysis: Mauricio Letelier, URL: https://towardsdatascience.com/clustering-with-more-than-two-features-try-this-to-explain-your-findings-b053007d680a