Automated stock recommendations using Financial Indicators and Machine Learning

Utkarsh Sharma  
B. Tech CSE  
ASET, Amity University, Noida  
skhiearth@gmail.com

Simran Gogia  
B. Tech CSE  
ASET, Amity University, Noida  
simrangogia20@gmail.com

Abstract

Stock market is suggested and regarded as one of the high-yielding long-term investments, yet a majority of people don't capitalize on the same. Dubious advice and attempts to ‘beat the market’ usually give rise to skepticism and distrust among first-time investors. This paper proposes a subjective, low-risk stock market advising platform that leverages Machine Learning clustering (K-Means) on basic Financial Indicators that are used to track the performance of stocks in the exchange to serve as an aid in investment decision, particularly for first-time investors. The results suggest that clustering-powered subjective recommendations can prove to be a low-risk advising tool.

Keywords: Stocks, Recommendations, Finance, Analysis, Machine Learning, Clustering Algorithms

Introduction

The fact that stock investments return well is well established. Stock investments are highly advisable (Blanchard et al. 1993) for all – due to low barrier of entry, trust and profitability. However, the lack of reliable information and advice regarding investments for beginners is a major hurdle in absolute adoption of stocks as a mainstream investment channel among those with limited capital.

Inspirations for this research include the financial problems concerning SMEs and individuals (Beauregard et al. 2020) generated due to the COVID-19 pandemic and review reports (Ramnath et al. 2008) (Morgan et al. 2003) addressing the forecasts made by financial analysts about stock recommendations. Provided the economic limitations of first-time investors, subjective stock recommendations pose themselves as a more viable option than risky advice – in that Machine Learning clustering aims to group stocks that have performed well within an exchange together.

Using key financial metrics, we can observe the trends and patterns in the financials of such companies, and with the help of Machine Learning – these metrics can be used for recommendations. Listed stocks from the most prominent exchanges from Asia-Pacific and Europe are used for this analysis. The following key financial indicators are considered while clustering the stocks in the different exchanges:

**Gross Profit Ratio**

Gross profit ratio is defined as the profitability ratio between gross profit and net sales.

\[
\text{Gross Profit Ratio} = \frac{\text{Gross profit}}{\text{Net sales}}
\]

The GP Ratio is a good indicator of the financial performance and effective resource allocation (Nariswari et al. 2020) of a company and helps in analyzing its profitability.
Automated stock recommendations using Financial Indicators and Machine Learning

**P/E Ratio**

This indicator is included to determine the possible overvaluation or undervaluation of the ticker in question – reflecting the market’s belief on the company’s grossing capacity. A high P/E Ratio typically indicates over-valuation (Shen 2000).

\[
P/E \text{ Ratio} = \frac{\text{Market value per share}}{\text{Earnings per share}}
\]

**Price cash flow ratio**

Used for equating the company’s market worth to the cash flow, PCF ratio provides a reliable sign of long-term returns.

\[
PCF \text{ ratio} = \frac{\text{Share price}}{\text{Operational cash flow per share}}
\]

**EPS diluted**

It is an indication of the company’s potential earnings per share if all eligible securities were converted.

\[
\text{EPS diluted} = \frac{\text{Net income–Preferred Stock Dividends}}{\text{Average Outstanding Shares + Dilutive Shares}}
\]

It is a constructive metric because it takes into account all convertible securities.

**Price to book value ratio**

Used for indication of a share’s worth according to the book of accounts.

\[
PBV \text{ ratio} = \frac{\text{Market price per share}}{\text{Book value per share}}
\]

**Asset turnover ratio**

A measurement of the ability of company management to effectively use the net assets available for sales revenue generation.

\[
AT \text{ ratio} = \frac{\text{Net Sales}}{\text{Average Total Assets}}
\]

**Dividend yield percentage**

Indication of the expected return percentage in the means of dividends on the investments made at the prevailing market price.

**Price earning to growth ratio**

It draws the relationship between a stock’s P/E ratio and projected earnings’ growth rate over a period of time – giving an estimation about the actual value taking into account the predictable growth.

\[
\text{PEG ratio} = \frac{\text{P/E Ratio}}{\text{EPS Growth}}
\]

This paper is divided into the following sections: Related Work, Proposed Approach, Methodology, Results and Discussion, Conclusion and Future Scope, and References.
Related Work

In paper (Sanboon et al. 2019), the authors propose a model to capture long-term dependencies in stock price data using LSTM to develop a Deep Learning model for predicting buy and sell recommendations in the Thai Stock Exchange.

Using natural language processing techniques, Robert et al. (Schumaker et al. 2009) present a quantitative stock prediction system based on daily financial news. Decision support systems are proposed (Gottschlich et al. 2014) to help incorporate crowd recommendations into investment decision.

In paper (Barber et al. 2003), the author’s primary concern is to assess and analyze the returns obtained on investments based on analysts’ stock recommendations and (Bradshaw 2004) deals with addressing how analysts use their forecasts to generate such recommendations. (Nair et al. 2017) use Machine Learning based clustering for generating stock trading recommendations based on price time-series data.

Proposed Approach

In this paper, the K-Means clustering algorithm is used to group stocks from various exchanges around the globe into an appropriate number of disjoint clusters – and to extract information from such clusters. All exchanges are kept mutually exclusive to prevent cross-exchange clusters.

An automated platform is set-up to aid on-the-flow cluster generation as the stocks move over time, accounting for new changes in the financial indicators of the companies.

![Diagram](Figure 1. Generalized Framework)

Methodology, Results and Discussions

Tools used

For implementation of the clustering algorithm and cluster analysis, Python (Sklearn) and R (dplyr, caTools, e1071, factoextra and ggplot2) are used, whereas the Pandas, requests and NumPy libraries of the Python language are used for scraping and data pre-processing. Python and R were selected due to the availability of well-kept popular libraries.
**Dataset, Cleaning and Pre-processing**

Due to unavailability of an exhaustive dataset that comprised of all the financial indicators, and crucially, targeted the time-frame concerning the pandemic, Web Scraping was used to collect the data.

The scrapped dataset comprised of more than 20,000 stock tickers from over ten stock exchanges from Asia Pacific and Europe – covering the different continents and eliminating location bias (Pirinsky et al. 2006). All stock data, across exchanges, was scaled (Standard Scaling) to have a mean value of zero and a standard distribution of unity. The standard score of a sample stock ticker, \( x \) is computed as:

\[
z = \frac{x - \mu}{s}
\]

where, \( \mu \) is the mean of the samples, and \( s \) is the standard deviation of the samples.

**Clustering**

We use K-Means Clustering to assemble similar tickers into a pre-defined set of clusters – to partition stocks based on financial patterns. All these clusters are disjoint and limited to the scope of one stock exchange. K-Means is the widely used iterative clustering algorithm. The algorithm requires the number of clusters \( K \) to be provided as input (Lloyd. 1982). Financial indicators are used as independent variables.

To initialize the number of clusters, \( k \), two techniques were used: Elbow Method and Silhouette Method. In the former, the number of clusters corresponding to a severe cliff in the sum of squares of the distances of each data point in all clusters to their respective centroids are considered as the optimal number of groupings. The silhouette method computes silhouette coefficients - a degree of how comparable a stock is to its specific group (cohesion) compared to other clusters - of each stock that measure how much a stock is similar to its own cluster paralleled to other groups.

For several exchanges, to boost the number of clusters, Principal Component Analysis was employed. It is vital to ensure a health clustering of data as it is known that the statistical power of results increase with an increase in number of groups – to better analyze the inter-cluster disparity. The ideal number of clusters varied across exchanges. The following table illustrates the number of clusters and the corresponding silhouette scores for the optimal number.

<table>
<thead>
<tr>
<th>Stock Exchange</th>
<th>Number of Clusters</th>
<th>Silhouette Score</th>
<th>PCA Components</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSE</td>
<td>3</td>
<td>0.66</td>
<td>1</td>
</tr>
<tr>
<td>ASX</td>
<td>2</td>
<td>0.68</td>
<td>1</td>
</tr>
<tr>
<td>MCX</td>
<td>4</td>
<td>0.68</td>
<td>1</td>
</tr>
<tr>
<td>HKSE</td>
<td>2</td>
<td>0.70</td>
<td>0</td>
</tr>
<tr>
<td>Osaka</td>
<td>2</td>
<td>0.74</td>
<td>0</td>
</tr>
<tr>
<td>Lisbon</td>
<td>6</td>
<td>0.72</td>
<td>1</td>
</tr>
<tr>
<td>Brussels</td>
<td>7</td>
<td>0.74</td>
<td>0</td>
</tr>
<tr>
<td>Amsterdam</td>
<td>7</td>
<td>0.73</td>
<td>0</td>
</tr>
<tr>
<td>XETRA</td>
<td>2</td>
<td>0.68</td>
<td>1</td>
</tr>
<tr>
<td>LSE</td>
<td>2</td>
<td>0.76</td>
<td>0</td>
</tr>
</tbody>
</table>

**Table 1. Ideal number of clusters across exchanges**

A silhouette score of positive unity indicates perfect distribution of data points into disjoint clusters. For most of the exchanges, the silhouette score can be termed as satisfactory provided the high level of variations among stocks of different industries and maturity.
Cluster Analysis

Financial Analysts can utilize patterns in the clusters formed by K-Means clustering to better study the correlation of the financial indicators in questions across different stocks.

<table>
<thead>
<tr>
<th>Financial Indicator</th>
<th>Cluster #1</th>
<th>Cluster #2</th>
<th>Cluster #3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gross Profit ratio</td>
<td>0.277</td>
<td>0.753</td>
<td>0.538</td>
</tr>
<tr>
<td>P/E ratio</td>
<td>33.242</td>
<td>72.948</td>
<td>30.499</td>
</tr>
<tr>
<td>Price Cash Flow ratio</td>
<td>68.632</td>
<td>98.680</td>
<td>9714.396</td>
</tr>
<tr>
<td>EPS diluted</td>
<td>5.049</td>
<td>19.547</td>
<td>24.351</td>
</tr>
<tr>
<td>Price to book-value ratio</td>
<td>7.416</td>
<td>7.873</td>
<td>15.340</td>
</tr>
<tr>
<td>Asset turnover ratio</td>
<td>1.080</td>
<td>0.717</td>
<td>1.010</td>
</tr>
<tr>
<td>Dividend yield percentage</td>
<td>1.299</td>
<td>0.815</td>
<td>1.239</td>
</tr>
<tr>
<td>Price earnings to growth ratio</td>
<td>1.813</td>
<td>8.866</td>
<td>-10.847</td>
</tr>
</tbody>
</table>

Table 2. Mean values of financial indicators in NSE clusters

<table>
<thead>
<tr>
<th>Financial Indicator</th>
<th>Cluster #1</th>
<th>Cluster #2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gross Profit ratio</td>
<td>0.208</td>
<td>0.793</td>
</tr>
<tr>
<td>P/E ratio</td>
<td>11.736</td>
<td>53.299</td>
</tr>
<tr>
<td>Price Cash Flow ratio</td>
<td>50.722</td>
<td>75.744</td>
</tr>
<tr>
<td>EPS diluted</td>
<td>0.210</td>
<td>0.111</td>
</tr>
<tr>
<td>Price to book-value ratio</td>
<td>5.165</td>
<td>9.381</td>
</tr>
<tr>
<td>Asset turnover ratio</td>
<td>1.832</td>
<td>0.875</td>
</tr>
<tr>
<td>Dividend yield percentage</td>
<td>2.738</td>
<td>2.639</td>
</tr>
<tr>
<td>Price earnings to growth ratio</td>
<td>209.035</td>
<td>340.632</td>
</tr>
</tbody>
</table>

Table 3. Mean values of financial indicators in Australian Securities Exchange clusters

Similar trends and patterns can be found in all exchanges across the three continents – providing a useful tool for financial analysts and individuals alike to evaluate stock recommendations.

Automated Recommendations

As time passes, the financial indicators of the companies in question are bound to change – affecting the relationship between stocks. To incorporate for such market movements, a platform should be set up that updates the considered data for every quarter (Kiger 1972) or daily movements – thereby updating the clusters of tickers and providing the most accurate disjoint groups of related stocks.

As part of a POC, a Python-based script was used for this continuous updating process and deployment of clusters.
Conclusion and Future Scope

Recommendations across a wide range of exchanges and the stock token therein are ought to be beneficial to first-time investors and those with superficial or no knowledge about the market. The simple input sets used in the model improve transparency and the subjective nature of the proposed approach helps reduce risk. The combined automated employed can be deployed in the form of a consumable REST API, leveraged by end-user application – a mobile app was developed as part of our POC. The source code for the clustering algorithms, the API, the end-user application and the original and clustered datasets are open-sourced and forthcoming researches are welcome to experiment on the same (Simran et al. 2020).

A similar recommendation and clustering engine can be extended beyond listed stocks to Mutual Funds, Bonds and Real estate as well – offering more diverse investment options. The lack of reliable data in these sectors and the lack of domain knowledge are the prime restraining factors for the authors as of writing.

Deeper domain knowledge can be used for refining the inputs to the Machine Learning models. We propose feature manipulation and selection of a different array of indicators. Although advances in objective stock recommendations have been largely unsatisfactory so far, advances in Neural Networks can be leveraged to build transparent recommendation engines based on similar input tuples.

References