IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 4, April 2017

Financial Markets- Investment Strategy through Predictive and Statistical Analysis using Big Data and MATLAB

Vishal Kannan[1]

Student, Mechanical Engineering, RV College of Engineering, Bengaluru, India [1]

Abstract: Predicting public company stock upswing and downswing is a very complex and mathematical problem. Accurate prediction of public stock performances can be beneficial to hedge funds, private investors as well as the common public interested in investment in stock.The biotechnology industry has grown rapidly in recent years, doubling in size between 1993 and 1999. 150,800 of the US jobs were generated directly by biotechnology companies, while the remaining 286,600 jobs were generated by companies supplying inputs to the industry, or by companies providing goods and services to biotechnology employees. Biotechnology is a field of science which has enormous potential in growing into a global crisis savior with recent developments ranging from easy surgical techniques and cheap pharmaceuticals to solutions in carcinogenic problems and nanotechnology. Due to the presence of solutions to such immensely concerning humanitarian problems, Biotechnology would soon serve as a backbone to human sustenance and evolution. All this would only mean that the stock prices of such a sector would boom to great heights in the near future. Due to this reason, in depth analysis of stocks of public Biotechnology companies is very important. In this paper, an elegant real-time public stock performance prediction methodology based on an aggregation of historical stock performances and organizational stock properties using a set of top US Biotechnology public companies has been introduced. In this paper, a striking co relation between the stock price and market capitalization swings of the companies with the number of employees has been reported. With these temporal stock swings, repeater operator analysis has been used to establish a statistical veracity of the novel metric of these top biotech companies. This corresponds to a result of profit, 85% of the time. A MATLAB model has also been developed to aid automate the stock performance prediction methodology reported in this. This method lends itself to broaden a large scale statistical validation as well as development of more advanced and complex models for stock performance prediction. For this analysis, exactly 9,234 data points were considered by manual data mining.

Keywords: Earnings per share, P/E Ratio, Shares Outstanding, Market Capitalization, Innovation Potential (IP).

I. INTRODUCTION

Stock markets were always a very intriguing study for me because of its lack of predictability and its ability to baffle an entire country and its global economy. In 2014, China's economy crashed and left an entire nation awestruck. China's total bank debt has grown from $14 trillion in 2008 to $25 trillion today – more than double the total size of the US commercial banking sector. This was a major concern for the big-time investors and hedge funds as they lost massive amounts of money. This was the motivating factor in this paper. In this paper, a new methodology is introduced to improve the predictive power of stock market prediction. On 19th September 2014, Hilary Clinton released a statement of a social medial website about a fight against the price of pharmaceutical prescription products. This led to a sudden drop in stock price which shook the confidence of even fearless investors. Such debates that arise in the biotechnology stock market causes a lot of liquidity in the stock which attracted me to study more on this and start research on Big data analysis of Biotechnology stock.

Big data analytics is the process of collecting, organizing and analysing large sets of data to discover patterns and other useful information. Big data analytics can help organizations to better understand the information contained within the data and will also help identify the data that is most important to the business and future business decisions. Analysts working with big data basically want the knowledge that comes from analysing the data. As the technology that helps an organization to break down data silos and analyse data improves, business can be transformed in all sorts of ways. The biggest challenge of big data analysis is the sheer volume of the data itself. To maximize the capability of this and reduce cumbersomeness, learning of software such as MATLAB which as tremendous compiling power is essential.
Enterprises are increasingly looking to find actionable insights into their data. Many big data projects originate from the need to answer specific business questions. With the right big data analytics platforms in place, an enterprise can boost sales, increase efficiency, and improve operations, customer service and risk management.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 4, April 2017

There are two prices that are critical for any investor to know: the current price of the investment he or she owns, or plans to own, and its future selling price. Most investors have a certain behaviour of staying of stocks which have risen sharply recently with the assumption have the "shoot" is over. Investors also avoid on stocks that have been dropping consistently assuming a technical problem inside the company leading to the poor performance of the company. There are generally three types of investors, namely full time investors (aim for small frequent profits), weekly and monthly investors (view stock fluctuation every week) and long term investors (aim to get large profits over a course of a few years generally also partner with the company to provide funds).

II. DISADVANTAGES OF CURRENT PREDICTION METRICS

A. Earnings per share

Earnings per share is the ratio of profit in a particular time period and current share price. If earnings per share is high, the profit is high and share price is low. But this is based on past performance and does not tell us about future potential. It merely tells us whether a company did well in chosen time frame or not.

B. P/E Ratio

P/E ratio is the ratio of market price of share and the earnings per share. If P/E ratio is high, it implies that the market price is high. It could also mean that the market price is high compared to the earnings per share. Hence it put the investor in a dilemma whether to invest or not and puts the pressure back on the investor.

III.PROPOSED METHODOLOGY

A. Definitions and terms

- Shares Outstanding

Shares outstanding refers to all shares currently owned by stockholders, company officials, and investors in the public domain, but does not include shares repurchased by a company.

- Market Capitalization

Market Capitalization is the product of number of shares outstanding and the current share price. This gives the investor a general idea of the worth of the company.

- Initial Public Offering

An initial public offering (IPO) is the first sale of stock by a private company to the public. IPOs are often issued by smaller, younger companies seeking the capital to expand, but can also be done by large privately owned companies looking to become publicly traded.

- Share Price

B. The Method

- The proposed method of stock prediction is illustrated using the following steps-

1. The historical shares outstanding and the stock prices of the companies were recorded.
2. The total number of employees at the time of investment were recorded.
3. Using historical shares outstanding and share price, the historical market capital was calculated.
4. A self-generated metric know as innovation potential was calculated (Innovation potential is the ratio of market capital in hundred thousand and the current number of employees.
5. A hypothesis was made that the companies with a higher innovation potential are most likely to grow in the next three months. (It is expected that the investor holds the invested stock for the time period of three months)
6. Receiver Operator analysis was used to test the correctness is the proposed idea.
7. Thresholding of the companies were done as "innovative", "non-innovative" and others.
8. The proposed innovative companies were indeed innovative with a chance of profit 85% of the times.

- What an investor must do

1. Input a spreadsheet with historical shares outstanding, share price and number of employees.
2. Feed this input into the MATLAB code.
3. Observe the Innovation Potential (IP) values.
4. Invest in the stocks that have a high IP values according to the provided thresholds.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 4, April 2017

	A	B
1		Neurocrine Biosciences Inc
2	Stock Symbol	NBIX
3	Founded	1996
4	Went Public on (CrunchBase)	1996
5	Minimum Private Investment in Million (CrunchBase)	10
6	Headquarters	CA
7	Medical keywords	S disorders, Endocrine disord
8	Number of Employees (MorningStar)	94
9	Date - Number of Employees (MorningStar)	31-12-2014
10	Number of Full Time Emloyees (MorningStar)	
11	NUMBER OF EMPLOYEES (Bloomberg)	94
12	Num of Employees (Linkedin)	148
13	VALUATION PER EMPLOYEE(in 100K USD per Employee)	500.00
14	VALUATION PER EMPLOYEE (in 100K - MorningStar)	
15	Valuation/employee according to Linkedin	317.57
16	Shares Outstanding JUNE 14(END,MIL)	75.92
17	Shares Outstanding JULY 14	75.92
18	Shares Outstanding AUGUST 14	75.96
19	Shares Outstanding SEPTEMBER 14	76
20	Shares Outstanding OCTOBER 14	76.01
21	Shares Outstanding NOVEMBER 14	76.27
22	Shares Outstanding DECEMBER 14	76.47
23	Shares Outstanding JANUARY 15	77.16
24	Shares Outstanding FEBRUARY 15	81.26
25	Shares Outstanding MARCH 15	85.37
26	Shares Outstanding APRIL 15	85.42
27	Shares Outstanding MAY 15	86
28	Shares Outstanding JUNE 15	85.75
29	Shares Outstanding JULY 15	85.87
30	Share Price JUNE 14	14.84
31	Share Price JULY 14	13.58
32	Share Price AUGUST 14	16.31
33	Share Price SEPTEMBER 14	15.67
34	Share Price OCTOBER 14	17.64
35	Share Price NOVEMBER 14	19.93
36	Share Price DECEMBER 14	22.34
37	Share Price JANUARY 15	33.55
38	Share Price FEBRUARY 15	39.05
39	Share Price MARCH 15	39.71
40	Share Price APRIL 15	43.37
41	Share Price MAY 15	43.86
42	Share Price JUNE 15	47.76
43	Share Price JULY 15	51.5
44	Market Capital JUNE 14(MIL)	1126.6528
45	Market Capital JULY 14	1030.9936
46	Market Capital AUGUST 14	1238.9076
47	Market Capital SEPTEMBER 14	1190.92	0.269298477
48	Market Capital OCTOBER 14	1340.8164	0.303194013
49	Market Capital NOVEMBER 14	1520.0611	0.343725975
50	Market Capital DECEMBER 14	1708.3398	0.386300764
51	Market Capital JANUARY 15	2588.718	0.585377535
52	Market Capital FEBRUARY 15	3173.203	0.717545036
53	Market Capital MARCH 15	3390.0427	0.766578221
54	Market Capital APRIL 15	3704.6654	0.837722726
55	Market Capital MAY 15	3753.1002	0.848675114
56	Market Capital JUNE 15	4095.42	0.926082665
57	Market Capital JULY 15	4422.305	1
58			
59	Mean valuation/employee in 100K	83	
60	STD valuation/employee in 100K	96	
61	Mean + STD	179	
62	Mean - STD	0	
63	Mean + 2*STD	275	

Data points example of Innovative Companies

DOI10.17148/IJARCCE.2017.6428

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 4, April 2017

	A	P
1		Chimerix Inc
2	Stock Symbol	CMRX
3	Founded	2002
4	Went Public on (CrunchBase)	2013
5	Minimum Private Investment in Million (CrunchBase)	256.4
6	Headquarters	NC
7	Medical keywords	Antivirals
8	Number of Employees (MorningStar)	
9	Date - Number of Employees (MorningStar)	
10	Number of Full Time Emloyees (MorningStar)	
11	NUMBER OF EMPLOYEES (Bloomberg)	104
12	Num of Employees (Linkedin)	111
13	VALUATION PER EMPLOYEE(in 100K USD per Employee)	221.15
14	VALUATION PER EMPLOYEE (in 100K - MorningStar)	
15	Valuation/employee according to Linkedin	207.21
16	Shares Outstanding JUNE 14(END,MIL)	35.4
17	Shares Outstanding JULY 14	35.53
18	Shares Outstanding AUGUST 14	35.67
19	Shares Outstanding SEPTEMBER 14	36.41
20	Shares Outstanding OCTOBER 14	36.44
21	Shares Outstanding NOVEMBER 14	36.48
22	Shares Outstanding DECEMBER 14	41.03
23	Shares Outstanding JANUARY 15	42.06
24	Shares Outstanding FEBRUARY 15	41.09
25	Shares Outstanding MARCH 15	41.31
26	Shares Outstanding APRIL 15	41.32
27	Shares Outstanding MAY 15	43.58
28	Shares Outstanding JUNE 15	45.85
29	Shares Outstanding JULY 15	46.07
30	Share Price JUNE 14	21.94
31	Share Price JULY 14	22.72
32	Share Price AUGUST 14	22.71
33	Share Price SEPTEMBER 14	27.62
34	Share Price OCTOBER 14	31.04
35	Share Price NOVEMBER 14	31.41
36	Share Price DECEMBER 14	40.26
37	Share Price JANUARY 15	41.78
38	Share Price FEBRUARY 15	41.41
39	Share Price MARCH 15	37.69
40	Share Price APRIL 15	37.53
41	Share Price MAY 15	41.85
42	Share Price JUNE 15	46.2
43	Share Price JULY 15	53.74
44	Market Capital JUNE 14(MIL)	776.676
45	Market Capital JULY 14	807.2416
46	Market Capital AUGUST 14	810.0657
47	Market Capital SEPTEMBER 14	1005.6442	0.4061893
48	Market Capital OCTOBER 14	1131.0976	0.456861127
49	Market Capital NOVEMBER 14	1145.8368	0.46281443
50	Market Capital DECEMBER 14	1651.8678	0.667205186
51	Market Capital JANUARY 15	1757.2668	0.709776849
52	Market Capital FEBRUARY 15	1701.5369	0.68726701
53	Market Capital MARCH 15	1556.9739	0.628876633
54	Market Capital APRIL 15	1550.7396	0.62635854
55	Market Capital MAY 15	1823.823	0.736659534
56	Market Capital JUNE 15	2118.27	0.85558949
57	Market Capital JULY 15	2475.8018	1

Data points example of Mediocre Companies

DOI10.17148/IJARCCE.2017.6428

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 4, April 2017

	A	E2
1		AEterna Zentaris inc
2	Stock Symbol	AEZS
3	Founded	1991
4	Went Public on (CrunchBase)	2000
5	Minimum Private Investment in Million (CrunchBase)	
6	Headquarters	QC
7	Medical keywords	endrocrine therapy and oncology,drug discovery
8	Number of Employees (MorningStar)	
9	Date - Number of Employees (MorningStar)	
10	Number of Full Time Emloyees (MorningStar)	
11	NUMBER OF EMPLOYEES (Bloomberg)	56
12	Num of Employees (Linkedin)	49
13	VALUATION PER EMPLOYEE(in 100K USD per Employee)	1.96
14	VALUATION PER EMPLOYEE (in 100K - MorningStar)	
15	Valuation/employee according to Linkedin	2.24
16	Shares Outstanding JUNE 14(END,MIL)	56.51
17	Shares Outstanding JULY 14	57.2
18	Shares Outstanding AUGUST 14	57.89
19	Shares Outstanding SEPTEMBER 14	60
20	Shares Outstanding OCTOBER 14	62.5
21	Shares Outstanding NOVEMBER 14	65.51
22	Shares Outstanding DECEMBER 14	65.51
23	Shares Outstanding JANUARY 15	73
24	Shares Outstanding FEBRUARY 15	81
25	Shares Outstanding MARCH 15	90.56
26	Shares Outstanding APRIL 15	92.5
27	Shares Outstanding MAY 15	95.89
28	Shares Outstanding JUNE 15	138
29	Shares Outstanding JULY 15	182.3
30	Share Price JUNE 14	1.16
31	Share Price JULY 14	1.18
32	Share Price AUGUST 14	1.16
33	Share Price SEPTEMBER 14	1.32
34	Share Price OCTOBER 14	1.12
35	Share Price NOVEMBER 14	0.52
36	Share Price DECEMBER 14	0.6
37	Share Price JANUARY 15	0.53
38	Share Price FEBRUARY 15	0.65
39	Share Price MARCH 15	0.54
40	Share Price APRIL 15	0.56
41	Share Price MAY 15	0.29
42	Share Price JUNE 15	0.28
43	Share Price JULY 15	0.18
44	Market Capital JUNE 14(MIL)	65.5516
45	Market Capital JULY 14	67.496
46	Market Capital AUGUST 14	67.1524
47	Market Capital SEPTEMBER 14	79.2
48	Market Capital OCTOBER 14	70
49	Market Capital NOVEMBER 14	34.0652
50	Market Capital DECEMBER 14	39.306
51	Market Capital JANUARY 15	38.69
52	Market Capital FEBRUARY 15	52.65
53	Market Capital MARCH 15	48.9024
54	Market Capital APRIL 15	51.8
55	Market Capital MAY 15	27.8081
56	Market Capital JUNE 15	38.64
57	Market Capital JULY 15	32.814

1	
0.883838384	
0.430116162	
0.496287879	
0.488510101	
0.664772727	
0.617454545	
0.654040404	
0.351112374	
0.487878788	
0.414318182	

Data points example of Non-Innovative Companies

IV. RESULTS

After generating the hypothesis, many exciting results were seen. Firstly, the thresholding of these stocks were done and the success rate of profits in investment was very high irrespective of the threshold. As expected, the companies with a very high IP value has an increase in stock price almost always, from the time the IP value was calculated.

The receiver operator analysis shows which threshold would be the best for an inventor for future models and investment plans. According to the receiver operator table given below, the IP threshold of (175,100) is chosen as the best. All companies above an IP value of 175 are considered innovative and all companies below IP values 100 is chosen as non-innovative. This threshold is chosen because of two reasons. Firstly, the percentage of success or the true positive percentage is highest at 85.71%. This means that, if were an investor to invest in a company with an IP value above 175, there is an 85.71% chance that he would make profits in the next 3 months. Secondly, if an investor were to invest in a company below IP value of 100, there is a 50% chance that he would get profits. This clearly highlights the liquidity of non-innovative companies which intern shows the risk of investment.

Hence an investor must always look to companies which have an IP value of above 175 before investment in order to ensure a higher chance of profit. It is important to note that this is most effective when there has been a recent crash in the markets. The investor must take full advantage of this time period by investing in companies having an IP value of above 175 and avoiding interest in companies with IP values below 100. This peculiar behaviour of innovative companies is studied even more in the next section.

IJARCCE

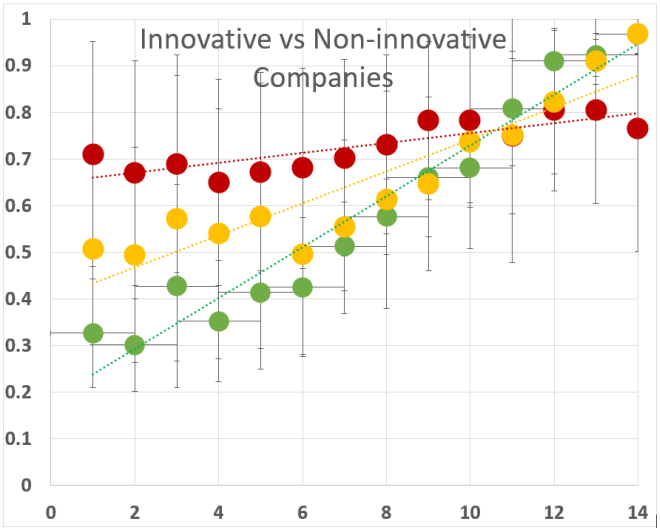
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 4, April 2017

CASE NUMBER	Threshold (IPI-1 >X - innovative; Y < IPI-1 < X - mediocre; IPI-1 < Y - noninnovative)	MARKET CAP						SHARE PRICE					
		TP	FP	T(TP+FP)	FN	TN	T(FN+TN)	TP	FP	T(TP+FP)	FN	PN	T(FN+TN)
1	IPI-1 - 200, 50	4	1	5	24	15	39	4	1	5	23	16	39
2	IPI-1 - 200, 150	4	1	5	34	32	66	4	1	5	35	31	66
3	IPI-1 - 200, 100	4	1	5	30	30	60	4	1	5	32	28	60
4	IPI-1 - 200, 75	4	1	5	29	17	46	4	1	5	28	28	46
5	IPI-I - 175, 125	6	1	7	34	22	56	6	1	7	32	24	56
6	IPI-I - 175, 100	6	1	7	30	30	60	6	1	7	32	28	60
7	IPI-1 - 150,100	9	2	11	30	30	60	9	2	11	32	28	60
8	IPI-1 - 125, 75	11	3	14	29	17	56	10	4	14	28	28	56
9	IPI-1 - 100, 50	14	4	18	24	15	39	13	5	18	23	16	39
10	IPI-1 - 75, 25	16	8	24	15	8	23	16	8	24	15	8	23

V. CONFIRMATION OF FEASIBILITY OF IP VALUE AS A STOCK MARKET PREDICTION METRIC

To make the analysis even more firm and to highlight the potential of such a statistical model, its conformation has been represented in the following way.



Key of graph
X axis- Months in numeric form (June 2014 to July 2015)
Y axis- Mean Normalized values of stock price of the three different thresholds against its maximum within the time period of June 2014 to July 2015 (Green represents the predicted innovative companies, red indicates the predicted non-innovative companies and yellow indicate the predicted mediocre companies)

Explanation
In the above plot, the data points of the individual months in the 14-month period are represented as green, red and yellow dots for innovative, non-innovative and mediocre companies respectively. Further, to show the steepness of the curves, straight line curve fitting was done (shown as dotted lines).
It is important to note that these values are normalised against a company's own maximum values of stock price, so the slope of the curve must be looked at and not the y-axis values at each month.
Now, it can be clearly observed that the innovative companies have the steepest graph. This implies that these companies have shot up at the fastest rate in terms of stock price during the recovery of the biotech stocks. The non-innovative companies seem to have an almost flat curve indicating that there was no improvement of the stock price after the crash of the markets. Naturally, the mediocre companies have performed as suspected, a mediocre rise in stock price.
This analysis proves that the proposed big data analytics prediction model proposed for stock markets is indeed applicable.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 4, April 2017

VI. MATLAB MODELLING

For the ease of the investor a MATLAB code was created to increase the automation of the process. This code is disclosed below.

```
%step1-read the file
[allnum,lab,eve]=xlsread('Datainput.xlsx');

%step2-Multiply individual shares outstanding
%with the share price (multiply row 5 with row 19, row 6 with row 20 and so on)
so=eve(6:19,2:end);
sp=eve(21:34,2:end);
marcap=cell2mat(so).*cell2mat(sp);

%for each company (collumn in eve)
for i=2:size(eve,2)
%step4-Read date of announcement of number of employees
%creating two variables anyear and anmonth to represent dates for
%each company
[anyear(i-1),anmonth(i-1)]=datevec(eve{4,i},'dd/mm/yy');%indexin
into (i-1) to have values from first cell in array
%(prevent writing from second cell)

end

%step5.1-match andate with SO dates and identify the corresponding SO price
for i=1:length(anyear)
for j=6:19
      [y, m] = datevec(eve{j,1}, 'mm/dd/yyyy');
if (y==anyear(i)) && (m==anmonth(i))
         SOprice_andate(i) = eve{j,i+1};%1+1 so that companoes start from second collumn

end
end
end

%step5.2-match andate with SP dates and identify the corresponding SP price
for i=1:length(anyear)
for j=21:34
      [y, m] = datevec(eve{j,1}, 'mm/dd/yyyy');
if (y==anyear(i)) && (m==anmonth(i))
         SPprice_andate(i) = eve{j,i+1};%1+1 so that companoes start from second collumn

end
end
end

%step5.3 multiply SO and SP identified above and divide by No of employees
%to get IP
for i=2:size(eve,2)
anIP(i-1)=SOprice_andate(i-1)*SPprice_andate(i-1)*10/eve{3,i};%*10 as it is calculated in 100k
end

%Step6 write IP values back to file
filename='Datainput.xlsx';
sheet=1;
xlRange='B40';
xlswrite(filename,anIP,sheet,xlRange);
```

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 4, April 2017

VII. CONCLUSION

With this statistical model, one can confidently invest in the companies in the Biotechnology sector just after a fall in the global market and expect profits in the next three months, as the companies try to get out of a recent economic meltdown. A 85.71% success rate for profits has been shown in this illustrative example after the companies. Biotechnology stocks have a great potential in growth and the US have invested heavily in this sector. This sector has shown a lot of promise in recent years and will continue to help investors to make their share of profits. The model proposed will help these investors and hedge funds. This will also help as a component in further in-depth research in this subject

REFERENCES

[1] http://www.bloomberg.com/research/sectorandindustry/industries/industrydetail.asp?code=352010&firstrow=20
[2] http://finance.yahoo.com/;_ylt=AhKpu4y5NXAD3xbq0dUbaH5.FJF4
[3] https://ycharts.com/
[4] http://www.gurufocus.com/
[5] http://www.morningstar.com/
[6] https://www.macroaxis.com/
[7] http://in.mathworks.com/index.html?s_tid=gn_logo
[8] https://www.crunchbase.com/
[9] https://in.linkedin.com/

BIOGRAPHY

Vishal Kannan who hails from Bangalore (Karnataka) is pursing B.E in Mechanical Engineering at RV College of Engineering, Bangalore. His area of interest includes mechatronics, design, thermodynamics and physics. Other interests include finance, investment and emerging technologies. He was selected for and has successfully completed, a general management program for young leaders at Stanford University.