Predicting Stock Price Movements post Quarterly Financial Report Release Using Linear Regression

Raunak Kumar
Mathematics Department
IIT Bombay
Mumbai, India
kraunak1402@gmail.com

***Abstract*—Use Linear Regression and distributed representation to predict stock price movements on the day of quarterly financial report release based on the reports data. The researchers notes that while textual data, such as news and financial reports, has been used to predict stock prices, exact words may convey different sentiments across different sectors, which can negatively impact performance. The study focused on the immediate temporary effects of quarterly financial reports from Pfizer Inc. (PFE) and employed distributed representation for predicting stock price movements on that particular day. The study results provide insight into the effectiveness of these techniques for predicting stock price changes relative to the opening price of the stock for that single day.**

I. INTRODUCTION

Stock market prediction has been an active area of research for many years. In recent years, there has been a growing interest in the use of machine learning and natural language processing techniques for stock market prediction. The main objective of this study is to investigate the effectiveness of Linear Regression and distributed representation for predicting temporary stock price movements for a single day based on quarterly financial reports.

II. DATA DESCRIPTION

The dataset used in this project was created using web scraping. The following sites were referred:

- https://investors.pfizer.com/Investors/Financials/Quarterly-Results/
- https://in.tradingview.com/chart/?symbol=NYSE%3APFE
- https://www.barchart.com/stocks/quotes/PFE/balance-sheet/quarterly?reportPage=1

Web Scrapping was done using python script and some manual changes were done to organize data as formats of data were different due to coming for different sources.

CSV file, named "pfizer_data.csv" was created to store all data.

"pfizer_data.csv" consists of the data for 152 row entries and 29 feature columns.

A few must-known columns are given below:

- Release Date: Impact of Quarterly Report on Investor on that day
- Cash and cash equivalents: Liquid current assets found on a business's balance sheet.
- PPE Net: Value of all buildings, land, furniture, and other physical capital that a business has purchased to run its business.

III. DATA CLEANING

The original data consist of 152 rows × 29 columns.

	column_name	Missing Number	percent_missing
Common Shares	Common Shares	17	11.184211
Retained earnings	Retained earnings	42	27.631579
Other shareholders' equity	Other shareholders' equity	72	47.368421
TOTAL.4	TOTAL.4	137	90.131579
Total Liabilities And Equity	Total Liabilities And Equity	147	96.710526

Fig. 1. Missing values table in the data

Release Date	Cash & Cash Equivalents	Marketable Securities	Receivables	Inventories	Income taxes - deferred	Other current assets	TOTAL	PPE Net	Investments And Advances	Intangibles
31-Jan-23	416000	22316000	10952000	8981000	3577000	5017000	51259000	16274000	15069000	94745000
1-Nov-22	1298000	34825000	16076000	9513000	2544000	6147000	70403000	15441000	13888000	77592000
28-Jul-22	1780000	31524000	15155000	10454000	2583000	5970000	67466000	15244000	18962000	78956000
3-May-22	2470000	21427000	13225000	9979000	3117000	4202000	54420000	15109000	20737000	80027000
8-Feb-22	1944000	29125000	11479000	9059000	4266000	3820000	59693000	14882000	21526000	74354000
...	...	...	...	...	...	...	...	...	...	...
15-Apr-99	1166000	3039000	3470000	1704000	1370000	10749000	1548000	2197000	10769000	19025000
19-Jan-99	1552000	2377000	2914000	1828000	1260000	9931000	1756000	2200000	3956000	18302000
13-Oct-98	2723000	930000	2860000	1681000	1680000	9874000	1717000	2471000	10574000	18177000
11-Aug-98	1089000	958000	3110000	1842000	1260000	8259000	1315000	2768000	10631000	16508000
14-Apr-98	1053000	795000	2913000	1837000	810000	7408000	1360000	3048000	4408000	15955000

100 rows × 23 columns

Fig. 2. Quarterly Data of Pfizer

- Dropped rows containing missing values, specifically the value ”NaN”.
- Removed all columns having all 0s.
- Removed the features that have more than 40 % missing data.
- Missing values in a dataset were imputed using the KNN imputation method.
- Remove highly correlated features
- After cleaning, ”pfizer_data.csv” consists of the data for 100 row entries and 23 feature columns.

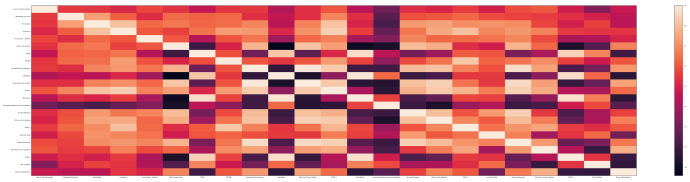

Fig. 3. Correlation Matrix

IV. Methodology

- 1) Financial entries which are effectively feature set for out dataset were visualize using heatmap figures. Visualization revealed some of the features were highly corelated with each other.
- 2) As some of entries financial data are just arithmetically related to other entries, so to reduce inefficiency in training of the machine learning model, highly corelated features were dropped.
- 3) The quarterly financial reports obtained from online sources did not have stock price included. So exact date of the report release were obtained from other online source and stock price were retrieved for that particular date.
- 4) Change in the stock price (Closing Price - Opening Price) of the company was considered as Y for our training and testing dataset.
- 5) Since stock price of any company is also affected by market mood on that particular date, hence change in NASDAQ index value was added as another feature for our dataset. Stock price and financial date for the company was taken from American market so NASDAQ was considered as market benchmark for the purpose.

V. Results

R-squared score for testing set predictions is 0.10, and Mean Squared Error (MSE) is 0.97

The R-squared score may indicate that my model is not appropriate for my data or data is skewed or has outliers.

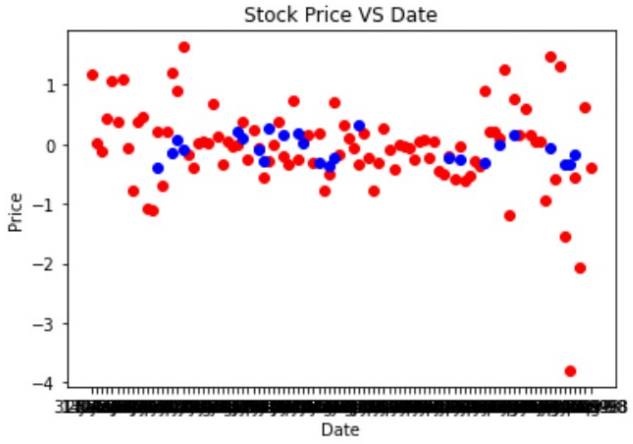

Fig. 4. Result

References

https://www.investopedia.com/what-quarterly-report.asp

https://investors.pfizer.com/Investors/Financials/Quarterly-Results/

https://scikit-learn.org/stable/modules/generated/sklearn.impute.KNNIm

https://towardsdatascience.com/are-you-dropping-too-many-correlated-features-d1c96654abe6

https://seaborn.pydata.org/generated/seaborn.heatmap.html

https://scikit-learn.org/LinearRegression.html

https://www.w3schools.com/python/matplotlib_scatter.asp