DS 3000 Final Project: Stock Price Predictor

Team 17:
Forrest Meng (meng.fo@northeastern.edu)
Michael Kim (kim.mich@northeastern.edu)

Introduction

- Investing in the stock market provides opportunities to grow wealth and achieve financial freedom.
- Successful investing requires extensive knowledge and understanding of a company's performance.
- By understanding a company’s performance, you could confidently decide to buy, sell, or hold the company’s stock

Project Goals

1. Can we successfully predict the price of a company given information about its recent performance?
2. Does the company's industry affect our ability to predict its stock price? If so, what might cause the unpredictability?

Data Description

1. **Historical Stock Prices** – yfinance library
2. **Company performance indicators** scraped from macrotrends.net
   - Quarterly earnings reports
     - EPS
     - P/E Ratio
     - Price-to-Book Ratio
     - Price-to-Sales Ratio
     - Debt-Equity Ratio
     - Current Ratio

macrotrends

Search over 200,000 charts.

Stock Screener | Stock Research | Market Indexes | Precious Metals | Energy | Commodities | Exchange Rates

Procter & Gamble Current Ratio 2006-2020 | PG

Prices | Financials | Revenue & Profit | Assets & Liabilities | Margins | Price Ratios | Other Ratios | Other Metrics

Current Ratio | Quick Ratio | Debt/Equity Ratio | ROE | ROA | ROI | Return On Tangible Equity

Procter & Gamble Current Ratio Historical Data

Date	Current Assets	Current Liabilities	Current Ratio
2020-12-31	$24.66B	$31.74B	0.78
2020-09-30	$26.03B	$30.01B	0.87
2020-06-30	$27.99B	$32.98B	0.85
2020-03-31	$27.14B	$32.90B	0.83
2019-12-31	$18.92B	$30.16B	0.63
2019-09-30	$21.93B	$30.25B	0.73
2019-06-30	$22.47B	$30.01B	0.75
2019-03-31	$22.31B	$28.37B	0.79
2018-12-31	$24.43B	$31.25B	0.78
2018-09-30	$23.35B	$29.22B	0.80
2018-06-30	$23.32B	$28.24B	0.83

Data Description

Historical Prices

In [38]:
```
df_prices = get_prices('pg')
df_prices.head()
```

Out[38]:

	date	quarterly avg close
0	2005Q2	35.227041
1	2005Q3	36.539654
2	2005Q4	38.401567
3	2006Q1	36.216768
4	2006Q2	38.805659

Company Performance Metrics

In [6]:
```
company_name = 'procter-gamble'
ticker = 'pg'

# compile a df of data of given metrics
x_feat_list = ['eps-earnings-per-share-diluted', 'pe-ratio', 'price-book', 'price-sales',
               'debt-equity-ratio', 'current-ratio']
df_metrics = get_metrics(company_name, ticker, x_feat_list)
df_metrics.head()
```

Out[6]:

	date	eps-earnings-per-share-diluted	pe-ratio	price-book	price-sales	debt-equity-ratio	current-ratio
0	2020Q4	1.47	26.14	7.02	4.90	1.47	0.78
1	2020Q3	1.63	26.27	7.01	4.97	1.47	0.87
2	2020Q2	1.07	23.68	6.21	4.35	1.58	0.85
3	2020Q1	1.12	60.65	5.78	3.98	1.58	0.83
4	2019Q4	1.41	71.69	6.52	4.55	1.43	0.63

Joined Together: →

In [8]:
```
df_joined = merge_df(df_prices, df_metrics)
df_joined.head()
```

Out[8]:

	date	quarterly avg close	eps-earnings-per-share-diluted	pe-ratio	price-book	price-sales	debt-equity-ratio	current-ratio
0	2006Q4	42.162891	0.84	15.00	2.03	1.95	1.10	0.79
1	2007Q1	41.574533	0.74	14.25	1.97	1.87	1.04	0.74
2	2007Q2	43.255557	0.67	13.34	1.90	1.84	1.07	0.78
3	2007Q3	48.297804	0.92	14.78	2.15	2.08	1.10	0.82
4	2007Q4	45.539909	0.98	14.85	2.22	2.14	1.12	0.86

Method: Multiple Linear Regression

- Input: joined dataframe
- Output: (equation, r2 score)
- Some companies had less data points, n_splits was lower
- Cross Validation (n_splits=10)

Regression Input:

	date	quarterly avg close	eps-earnings-per-share-diluted	pe-ratio	price-book	price-sales	debt-equity-ratio	current-ratio
0	2006Q4	42.162891	0.84	15.00	2.03	1.95	1.10	0.79
1	2007Q1	41.574533	0.74	14.25	1.97	1.87	1.04	0.74
2	2007Q2	43.255557	0.67	13.34	1.90	1.84	1.07	0.78
3	2007Q3	48.297804	0.92	14.78	2.15	2.08	1.10	0.82
4	2007Q4	45.539909	0.98	14.85	2.22	2.14	1.12	0.86

Regression Output:

In [31]:

```
x_feat_list = ['eps-earnings-per-share-diluted', 'pe-ratio', 'price-book',
               'price-sales', 'debt-equity-ratio', 'current-ratio']

model_str = regress_stock(df_joined, x_feat_list)
print(f'Equation: {model_str[0]}')
print(f'r2 score: {model_str[1]:.3f}')
```

```
Equation: quarterly avg close = 55.9420 + -2.3947 eps-earnings-per-share-diluted + -0.3486 pe-ratio + 31.6534 price-b
ook + -6.7322 price-sales + -53.8676 debt-equity-ratio + 6.3698 current-ratio

r2 score: 0.924
```

100 NASDAQ Companies

- Grouped by sector
- Cross-validated r2 values

In [44]:

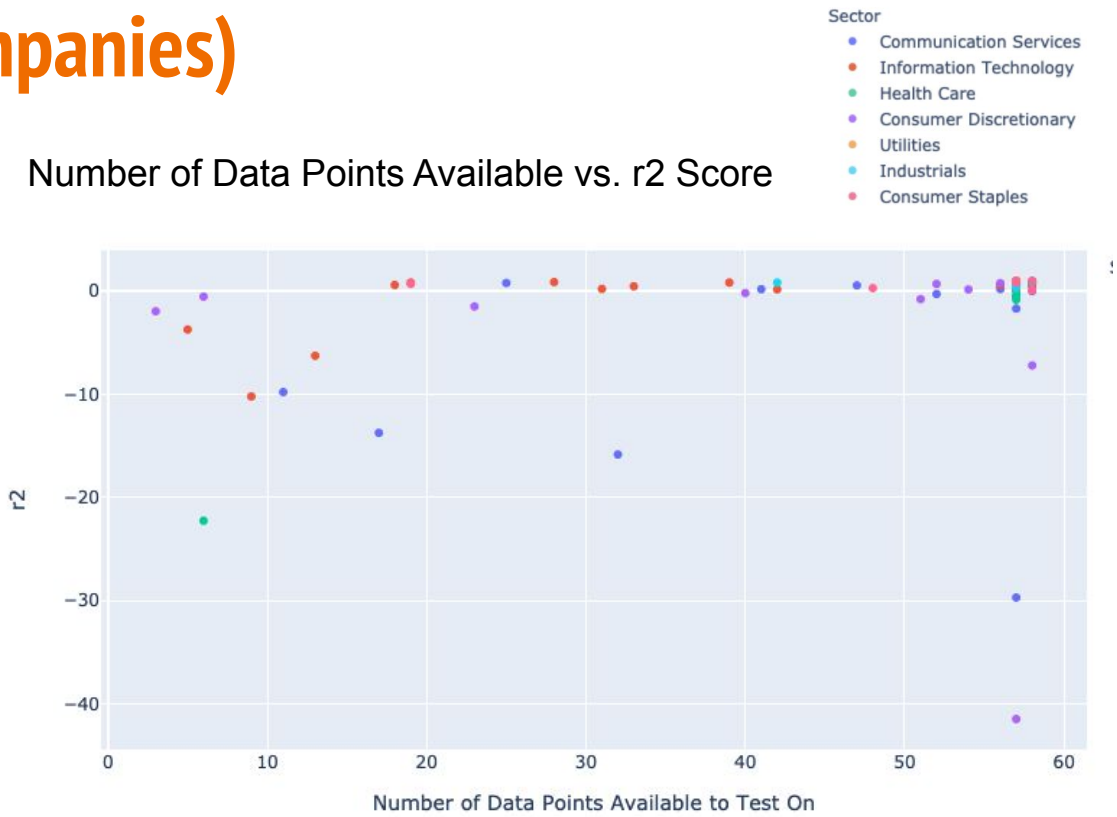
```
x_feat_list = ['eps-earnings-per-share-diluted', 'pe-ratio', 'price-book', 'price-sales',
               'debt-equity-ratio', 'current-ratio']

# import top 100 NASDAQ traded companies from external csv
df_nasdaq = pd.read_csv('NASDAQ-100-Companies.csv', usecols=['Company', 'Ticker', 'GICS Sector'])

# run each copany through our regression model
df_results = run_regression(df_nasdaq, x_feat_list)

# add each company's sector to our results
df_results['Sector'] = df_nasdaq.iloc[:, 2].values
df_results.head()
```

Out[44]:

	name	ticker	data_points	r2	equation	Sector
0	activision-blizzard	ATVI	57	0.520633	quarterly avg close = -19.4851 + 9.1000 eps-ea...	Communication Services
1	adobe	ADBE	58	0.941850	quarterly avg close = -83.1323 + 64.3756 eps-e...	Information Technology
2	amd	AMD	57	0.871801	quarterly avg close = 1.5358 + -0.3527 eps-ear...	Information Technology
3	align-technology	ALGN	57	0.425821	quarterly avg close = -219.8135 + 35.6708 eps-...	Health Care
4	alphabet	GOOGL	57	0.653044	quarterly avg close = -289.9697 + 55.4176 eps-...	Communication Services

Visualization (All Companies)

- Higher r2 with more data to train and test on
- Difference between sectors

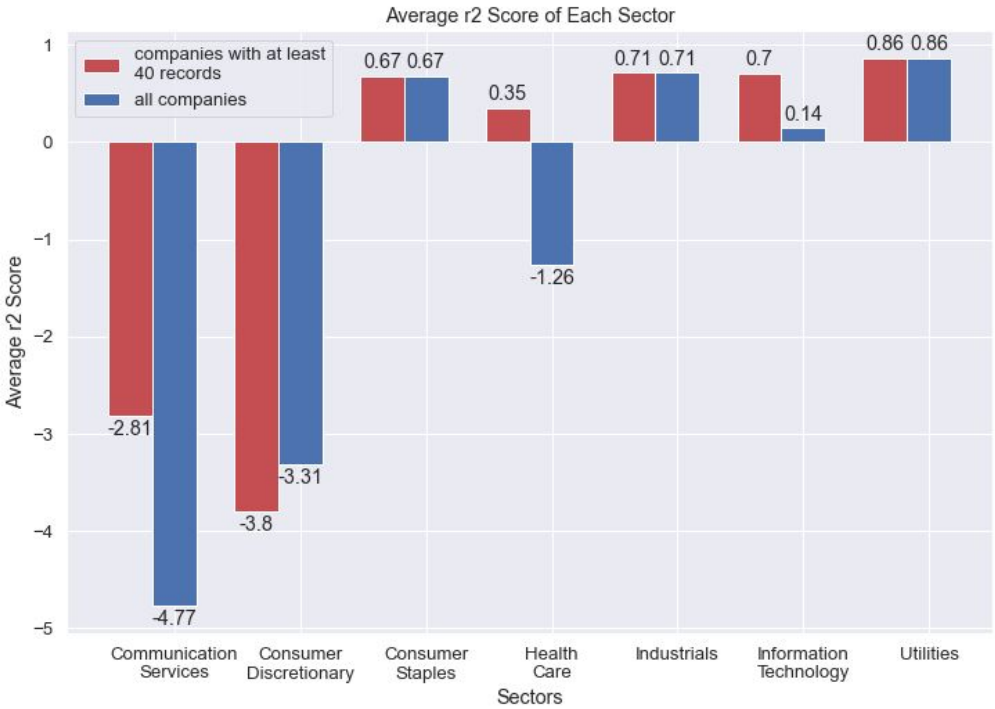


Visualization (By Sector)

- Split by number of data points available

    - Companies with more data had higher r2 values

- Speculative and hyped stocks



Discussion

Successes:

- Our model successfully predicted stock prices when given enough input data
- We’ve shown strong evidence that the company’s industry affects our ability to predict its stock price using just performance metrics

Some issues we faced:

- Some stocks are very volatile and change dramatically due to speculation
- We don't consider external factors about the U.S. and global economies or anything else that might have affected a company (or an entire sector)
- We had limited data

Conclusion

Overall, we think this work could be used as a reference for people to learn more about a company’s performance and determine good prices to buy, sell, or hold the company's stock. But it would probably be unwise to trust this model with your life savings.