STATS 112 FINAL PROJECT

Predicting Nasdaq (QQQ) intraday volatility

Luke Shuman

Introduction

- The **QQQ** is an exchange-traded-fund (ETF) that tracks the **Nasdaq**, the largest stock exchange in the world. This consists of America's 100 largest tech stocks (FAANG, TSLA, NVIDIA, etc.)
- QQQ is known to be one of the most **volatile** stock tickers in the market (aggressive growth stocks)
- Tech multiples are highly sensitive to **macro factors** such as interest & inflation rates, GDP growth, & unemployment levels

Question:

Can one use macro factors to predict how much in pts (in stock pts where 1 pt = $1) how much the market will move **over the course of a day**?

Because it would be a bit unrealistic to use **macro factors** to predict the movement of a stock every single day, we predict its average **intraday** move over **monthly** periods

Terminology

- QQQ → stock ticker tracking Nasdaq
- “Intraday volatility” → Combined Vol score = (close price - open price) + (high price - low price)
- Macroeconomic factors: inflation rates, interest rates, GDP growth, unemployment rates

Step 1: Data Collection

- **2 DATASETS** (QQQ historical quotes & US macroeconomic data)
- 1st data source: **MarketStack API** (https://marketstack.com/) to retrieve daily QQQ historical quotes from 6/2005 - 16/2/23
- 2nd data source: **Kaggle csv** (https://www.kaggle.com/datasets/federalreserve/interest-rates) downloaded csv file


```
"data": {
        "name": "Microsoft Corporation",
        "symbol": "MSFT",
        "has_intraday": false,
        "has_eod": true,
        "country": null,
        "stock_exchange": {
            "name": "NASDAQ Stock Exchange",
            "acronym": "NASDAQ",
            "mic": "XNAS",
            "country": "USA",
            "country_code": "US",
            "city": "New York",
            "website": "www.nasdaq.com"
        },
        "eod": [
            {
                "open": 235.9,
                "high": 237.47,
                "low": 233.17,
                "close": 236.94,
                "volume": 24307569.0,
                "adj_high": 237.47,
                "adj_low": 233.15,
                "adj_close": 236.94,
                "adj_open": 235.9,
                "adj_volume": 25332837.0,
                "split_factor": 1.0,
                "dividend": 0.0,
                "symbol": "MSFT",
                "exchange": "XNAS",
                "date": "2021-03-01T00:00:00+0000"
            },
           [...]
        ]
}
```

Data Cleaning

	Year	Month	Federal Funds Target Rate	Federal Funds Upper Target	Federal Funds Lower Target	Effective Federal Funds Rate	Real GDP (Percent Change)	Unemployment Rate	Inflation Rate	Rate Difference	year-month
0	2006	1	4.25	0.00	0.00	4.29	4.9	4.7	2.1	0.04	2006-1
1	2006	2	4.50	0.00	0.00	4.49	0.0	4.8	2.1	-0.01	2006-2
2	2006	3	4.50	0.00	0.00	4.59	0.0	4.7	2.1	0.09	2006-3
3	2006	4	4.75	0.00	0.00	4.79	1.2	4.7	2.3	0.04	2006-4
4	2006	5	4.75	0.00	0.00	4.94	0.0	4.6	2.4	0.19	2006-5
...	...	...	...	...	...	...	...	...	...	...	...
129	2016	10	0.00	0.50	0.25	0.40	1.9	4.8	2.1	0.40	2016-10
130	2016	11	0.00	0.50	0.25	0.41	0.0	4.6	2.1	0.41	2016-11
131	2016	12	0.00	0.50	0.25	0.54	0.0	4.7	2.2	0.54	2016-12
132	2017	1	0.00	0.75	0.50	0.65	0.0	4.8	2.3	0.65	2017-1
133	2017	2	0.00	0.75	0.50	0.66	0.0	4.7	2.2	0.66	2017-2

FILTERING

- QQQ index dataset: create daily volatility metric called "Combined Vol" by subtracting close price column - open price column and taking absolute value and adding to the absolute value of daily high - daily low. This is the formula I use to gauge volatility.

-filter for dates from 1/1/2006 to 2/1/2017 to match econ dataset available data

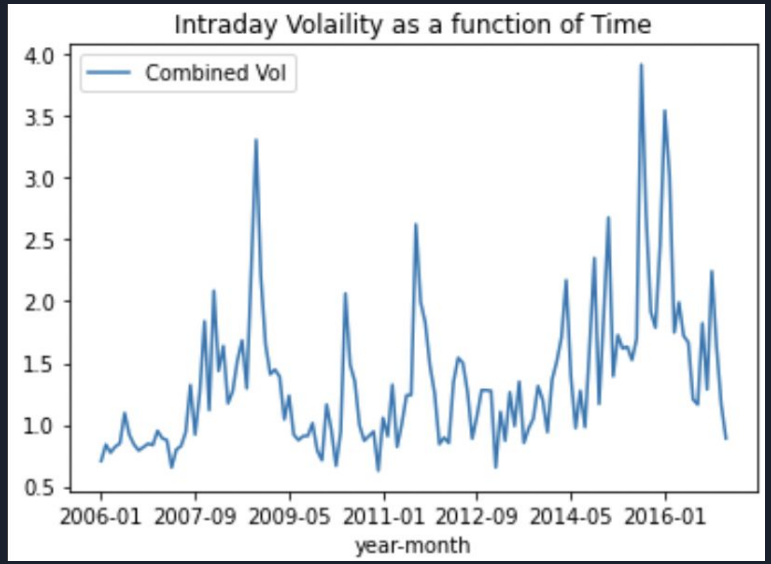
-clean date format into a standardized format YYYY-MM that can be used to calculate monthly intraday movement mean values (USED TO MERGE DATASETS AS SHARED KEY)

- create monthly intraday average by taking the mean "Combined Vol" value for the month(i.e. the average intraday move over a one-month interval)

	high	date	open	close	low	Daily Move	Combined Vol	year-month
0	127.24	2017-02-09	126.63	126.96	126.56	0.33	1.01	2017-02
1	126.68	2017-02-08	126.12	126.50	125.88	0.38	1.18	2017-02
2	126.55	2017-02-07	126.06	126.29	125.97	0.23	0.81	2017-02
3	125.85	2017-02-06	125.42	125.83	125.35	0.41	0.91	2017-02
4	125.81	2017-02-03	125.49	125.68	125.33	0.19	0.67	2017-02
...	...	...	...	...	...	...	...	...
2784	42.82	2006-01-19	42.40	42.52	42.29	0.12	0.65	2006-01
2785	42.46	2006-01-18	42.04	42.21	42.02	0.17	0.61	2006-01
2786	42.79	2006-01-17	42.66	42.70	42.50	0.04	0.33	2006-01
2787	43.05	2006-01-13	42.94	42.98	42.73	0.04	0.36	2006-01
2788	43.29	2006-01-12	43.16	43.00	42.85	0.16	0.60	2006-01

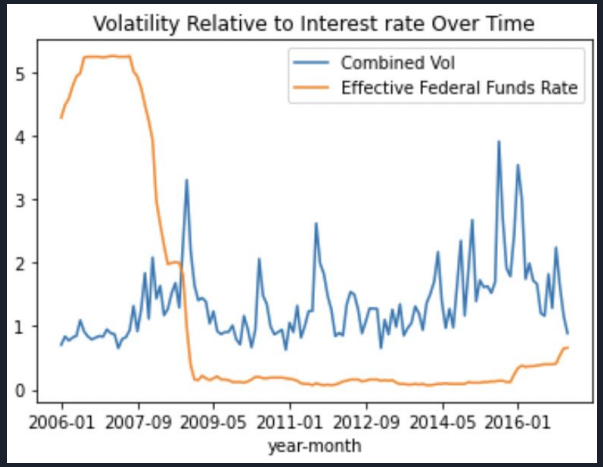
Data Exploration

What economic factors contribute to QQQ intraday volatility?

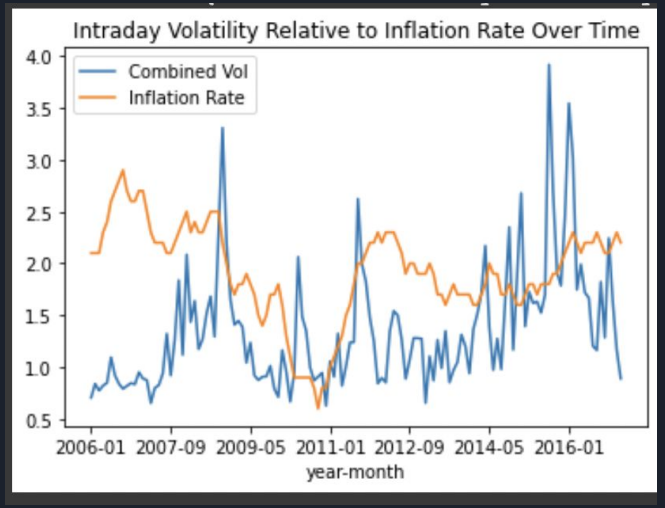


DATA EXPLORATION PART 1

Interest Rates vs Intraday Vol

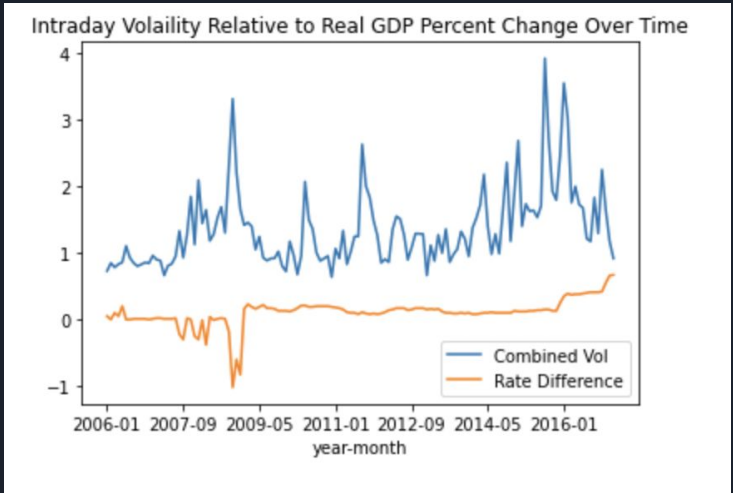


Inflation vs Intraday Vol

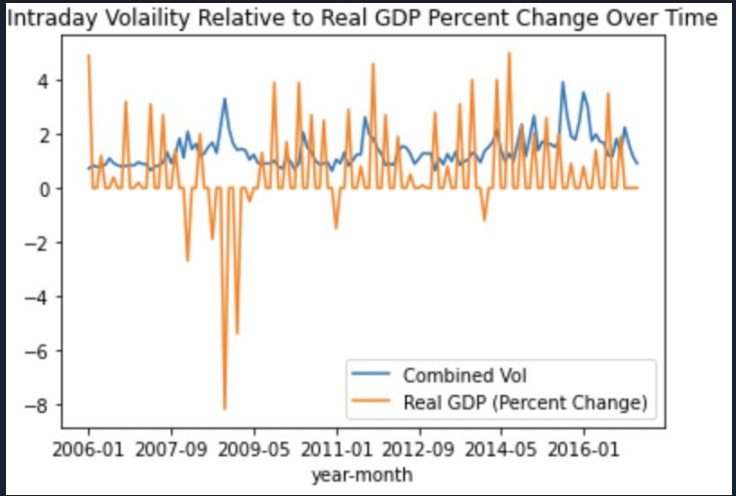


DATA EXPLORATION PART 2

Interest rate difference (target - actual) vs Intraday Vol



Real GDP change vs Intraday Vol



MACHINE LEARNING

How accurately can we predict average intraday movement of QQQ over monthly time periods given economic metrics?

3 Models:

1. **K-Nearest Neighbors**
2. **Gradient Boosting Regressor**
3. **Random Forest Regressor**

*y label is quantitative so using regression

MACHINE LEARNING: K-Nearest Neighbors

Step 1: Feature Selection → What features produce lowest RMSE?

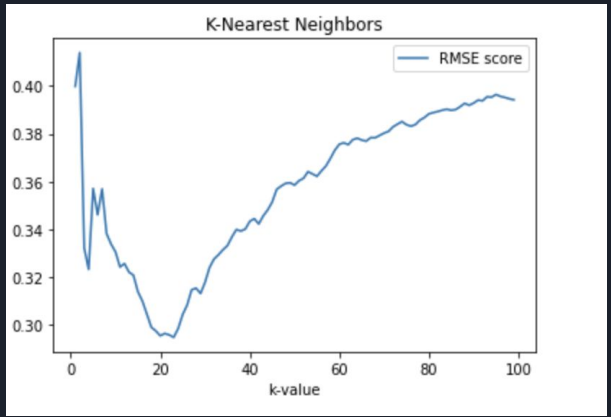
Result: all quantitative variables (i.e. Interest rates, Rate Difference, GDP growth, Inflation, unemployment rates)

Step 2: Finding Optimal DIstance metric → Cosine

Step 3: Finding Optimal K-value → **k = 22**

	features	RMSE score
0	[all quantitative variables]	0.315683
1	[Interest & Inflation Rates]	0.363914
2	[GDP Change, Rate Difference, Unemployment Rate]	0.394535
3	[Unemplyment & GDP Change]	0.348626

	distance metric	RMSE score
0	euclidean	0.315683
1	manhattan	0.323301
2	chebyshev	0.312177
3	minkowski	0.315683
4	cityblock	0.323301
5	cosine	0.295577

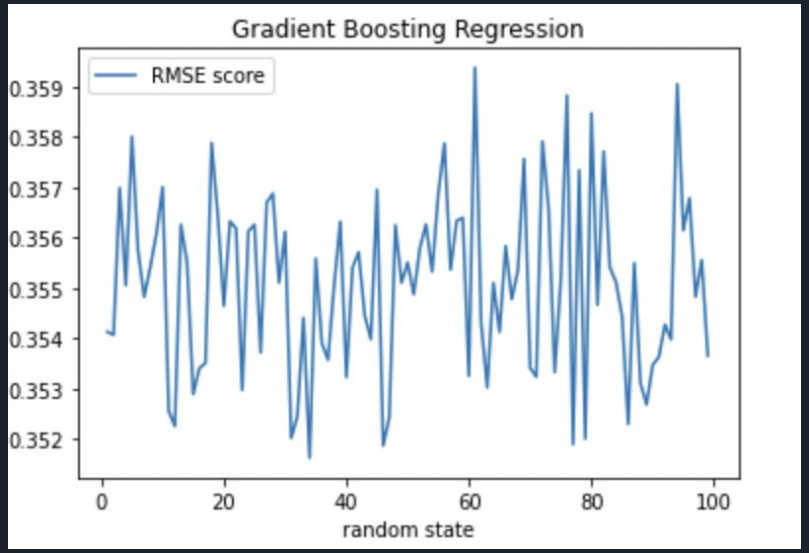


MACHINE LEARNING: GRADIENT BOOSTING REGRESSOR

Features = all quantitative variables

Distance = cosine

Optimal Random State Value = 34



	random state	RMSE score
33	34	0.351628
45	46	0.351871
76	77	0.351897
78	79	0.352011
30	31	0.352024
...	...	...
4	5	0.358011
79	80	0.358472
75	76	0.358828
93	94	0.359053
60	61	0.359379

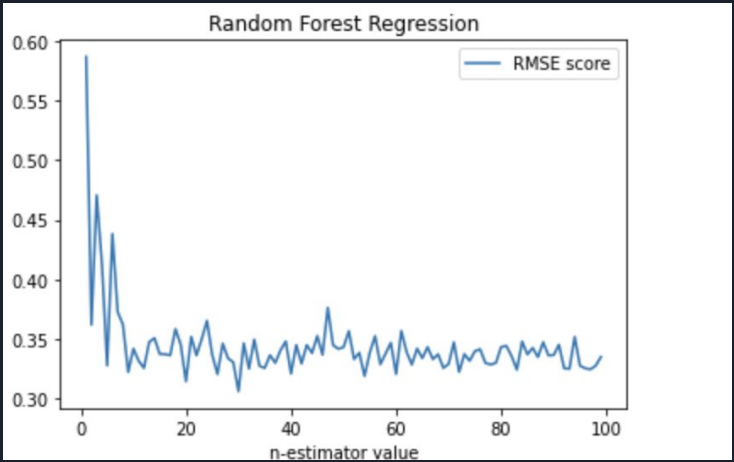
99 rows × 2 columns

MACHINE LEARNING: RANDOM FOREST REGRESSOR

Features = all quantitative variables

Optimal n-estimator value: n = 30

Distance = cosine



This is much better than Gradient Boosting Regression with a RMSE of .30...almost an improvement of .06!

	n-estimator value	RMSE score
29	30	0.305855
19	20	0.314282
53	54	0.318655
25	26	0.320252
59	60	0.320325
...	...	...
46	47	0.376160
3	4	0.413267
5	6	0.438234
2	3	0.470749
0	1	0.587141

OVERALL RESULTS

In the end, because the RMSE is so high relative to the Nasdaq's movement in pt values (almost 50% error), using macro economic factors **probably isn't the best approach to predict intraday volatility - even over monthly time frames**

MODEL	OPTIMAL PARAMETER VALUE	RMSE
K-NEAREST NEIGHBORS	22	.295
GRADIENT BOOSTING REGRESSOR:	30	.359
RANDOM FOREST REGRESSOR:	7	.30

.3 RMSE / (.5 < pt move < .8) ~ > 50% error → not ideal!

year-month	Daily Move	Combined Vol
2006-01	0.217692	0.720000
2006-02	0.297895	0.839474
2006-03	0.237826	0.773935
2006-04	0.267368	0.822084
2006-05	0.282727	0.849550
...	...	...
2016-10	0.377143	1.285581
2016-11	0.838571	2.243662
2016-12	0.536190	1.635476
2017-01	0.320000	1.166815
2017-02	0.237143	0.908571

Thank you!