David Lam Capstone 3 Report

Introduction

The purpose of this capstone project is to develop a model to predict stock movements, either predict stock price level, or identify positive or negative movements. More specifically we'd like to identify when the market might turn.

To accomplish this we will leverage twitter data to perform sentiment analysis on the SP500 index and the JD.com stock. We would then like to combine this with time series data for these two investment vehicles to see if we can make any meaningful predictions.

One constraint we have is that the Twitter data does not look back so I need to download data every day. We may need to download data for at least 30 trading days before we have enough data to establish meaningful relationships with stock price. For the purposes of this capstone, we will use what data we can obtain, knowing that we will need to iterate on this model once we get more data in the future.

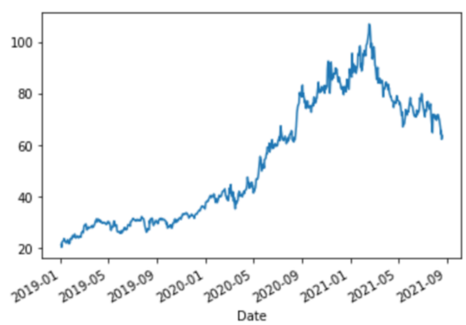
Given this constraint, the focus of this capstone shifted towards time series analysis and forecasting for JD.com and the S&P500.

Data

For sentiment analysis we leveraged Twitter data for JD.com and the S&P 500. Although we didn’t have enough data to perform meaningful analysis we are in the process of gathering daily tweet data to be able to do sentiment analysis at a later date. Below is a view of the twitter data with sentiment scores. We just need more daily data to obtain an average sentiment per day time series.

	created_at	text	retweet_count	favourite_count	date	month	week	day	sentiment
0	2021-08-24 21:52:06	The Nasdaq surpassed the 15,000 level for the ...	34	163	2021-08-24 21:52:06	8	34	24	0.7269
1	2021-08-25 09:31:14	US stock futures tread water after S&P 500...	9	14	2021-08-25 09:31:14	8	34	25	0.0000
2	2021-08-25 15:30:06	Join me & my @cfraresearch colleagues for ...	0	0	2021-08-25 15:30:06	8	34	25	0.2960
3	2021-08-25 15:24:51	@MacroAlf @CyberSpaceGal Based on Pe ratios of...	0	0	2021-08-25 15:24:51	8	34	25	0.4215
4	2021-08-25 15:22:56	Excess fiscal and #FederalReserve pumped liqui...	0	1	2021-08-25 15:22:56	8	34	25	-0.1513

Afterwards, we used the yahoo finance api wrapper yfinance to download JD.com and SPY daily data. Below is a plot of the JD.com price data:



Here is the S&P500 ETF (SPY):

From a feature engineering standpoint we created multiple data points as listed here for both stock series data.

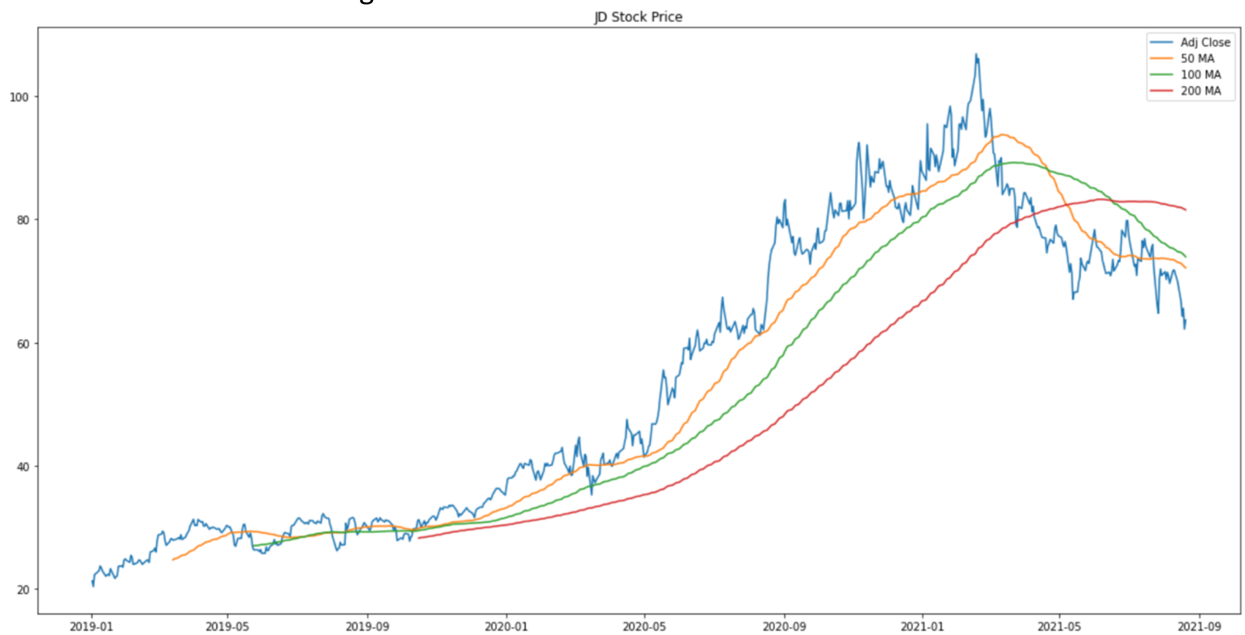
0	Open	Open Price
1	High	Daily High Price
2	Low	Daily Low Price
3	Close	Closing Price
4	Adj Close	Adjusted Close after share splits etc…
5	Volume	Volume of shares traded
6	high_minus_low	difference between high and low price
7	high_minus_low_pct_adjclose	% difference between high and low price
8	lagged_1	1 day lagged price
9	lagged_5	5 day lagged price
10	lagged_10	10 day lagged price
11	lagged_20	20 day lagged price
12	shifted_1	Price shifted by 1 day

13	shifted_5	Price Shifted by 5 days
14	shifted_10	Price Shifted by 10 days
15	shifted_20	Price Shifted by 20 days
16	pct_change_1	1 day return
17	pct_change_5	5 day return
18	pct_change_10	10 day return
19	pct_change_20	20 day return
20	50_ma	50 day moving average
21	100_ma	100 day moving average
22	200_ma	200 day moving average
23	50_std	50 day standard deviation
24	100_std	100 day standard deviation
25	200_std	200 day standard deviation
26	cumulative_return	Cumulative return
27	up_or_down	up or down return day
28	consecutive_count	consecutive up or down days

<u>Technical Analysis</u>

We performed some technical analysis by looking at price and various moving averages. Here we see a lot of resistance at the 50 day moving average level and we can see that when the 50 day moving average crossed below the 100 day moving average, the market moved lower by a large margin. By the time the 100 day moving average crossed below the 200 day moving average, we can see that the recovery had already taken place.

For JD we have the following:

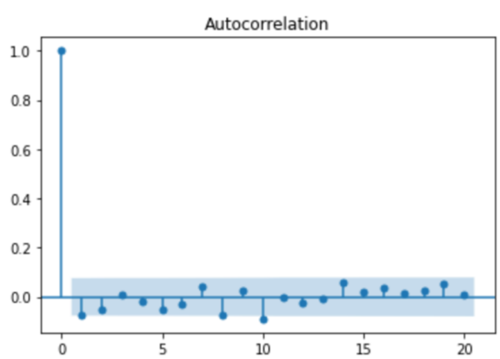


Here we see that when the 50 day moving average crosses below the 100 day we can see a dip in price. In fact the 100 day moving average crossed below the 200 day moving average in June of 2021. This is known as the death cross and is thought to signal further down performance. WHen we look further down at the S&P we can see this isn't always the case as when the 100 day moving average crossed below the 200 day moving average, the recovery was already on it's way. So the speed of recovery will impact this signal.

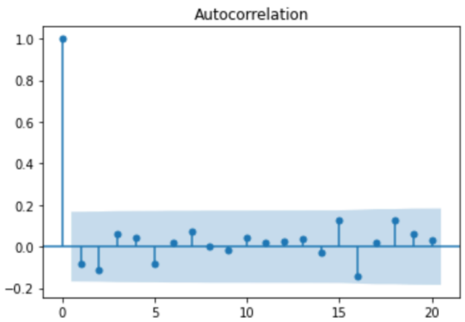
For JD it may appear that this is not on the way to recovery though.

<u>Modeling Journey</u>

We first started by trying an ARMA model for JD and SP500.

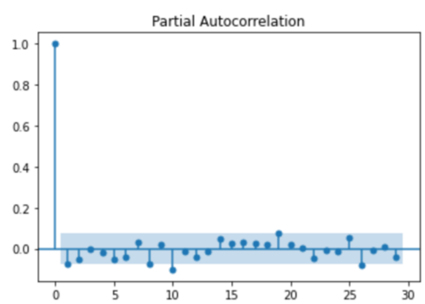


Daily returns do not appear to have any autocorrelation for JD.com



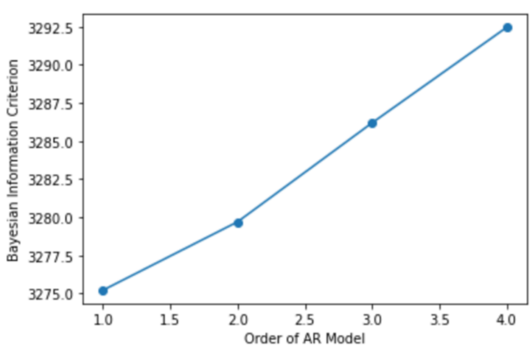
JD.com weekly returns appears to have zero autocorrelation at all lags

But after conducting the Dickey Fuller test, we find that both JD daily and weekly returns were stationary.

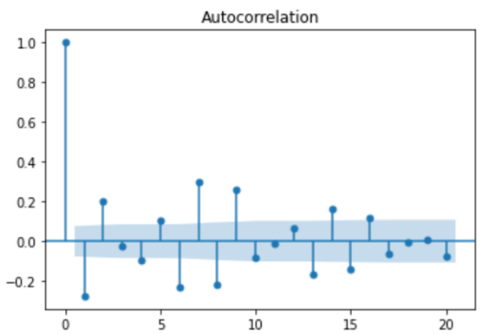


JD 1 day return does not have a significant PACF

Based on this I decided that an AR model for JD was not worth the effort since there are no significant PACF's. This is confirmed by trying to run various AR models for JD daily data and we find that the BIC increases:

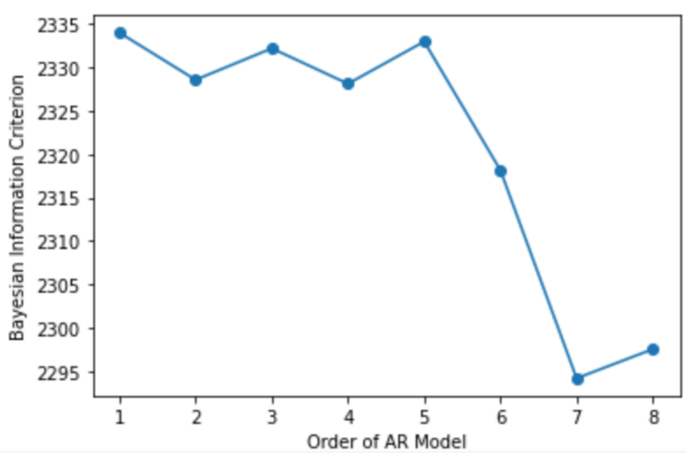


We have better luck with the S&P500 daily returns:



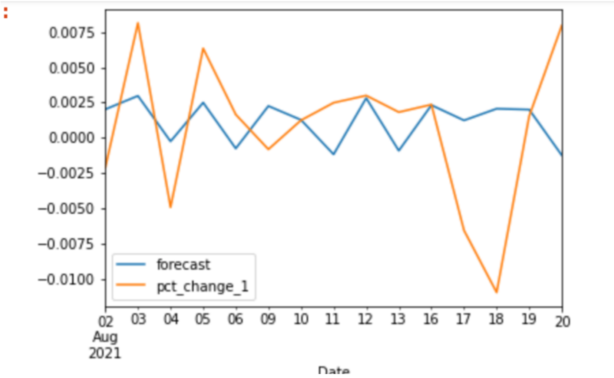
Daily SP500 returns appear to be autocorrelated and mean reverts.

The Dickey Fuller test also indicates that the daily return series for the SP500 is stationary. The Autocorrelation indicates that the returns are autocorrelated and mean reverts.



We also see that the AR(7) model has the lowest BIC.

Here is our forecast of the SP500 with this AR model:



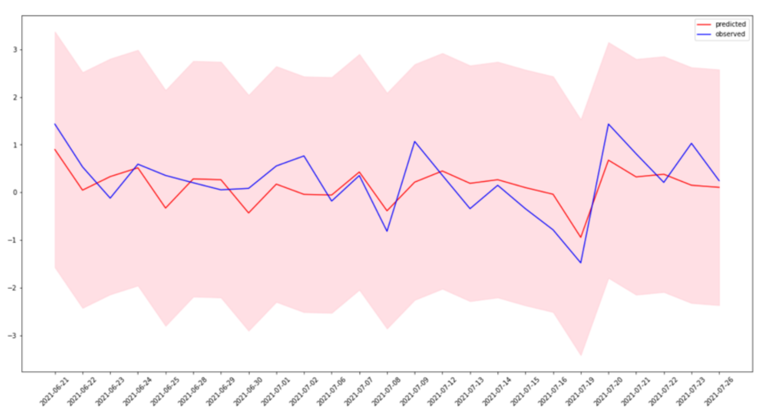
The Mean Absolute Error for this model is 0.85% which is fairly high from a practical standpoint.

We checked for Cointegration between the S&P 500 and JD but it was not the case. Next we tried to use an ARIMAX model for the S&P500 and we used the daily percentage change of volume as an exogenous variable. This produced the best ARIMAX result.

```
                              ARMA Model Results
==========================================================================
====
Dep. Variable:             pct_change_1   No. Observations:
645
Model:                       ARMA(7, 0)   Log Likelihood               -1064
.630
Method:                         css-mle   S.D. of innovations              1
.260
Date:                  Sat, 28 Aug 2021   AIC                           2149
.259
Time:                          14:39:33   BIC                           2193
.952
Sample:                      01-03-2019   HQIC                          2166
.600
                           - 07-26-2021
==========================================================================
=============
                          coef    std err          z      P>|z|      [0.02
5      0.975]
--------------------------------------------------------------------------
-------------
const                   0.1616      0.049      3.320      0.001       0.06
6       0.257
volume_pct_change_1    -1.0661      0.138     -7.733      0.000      -1.33
6      -0.796
ar.L1.pct_change_1     -0.2009      0.039     -5.205      0.000      -0.27
6      -0.125
ar.L2.pct_change_1      0.1267      0.039      3.211      0.001       0.04
9       0.204
ar.L3.pct_change_1      0.0682      0.039      1.733      0.083      -0.00
9       0.145
```

```
ar.L4.pct_change_1     -0.0862       0.039      -2.195       0.028      -0.16
3       -0.009
ar.L5.pct_change_1     -0.0193       0.039      -0.491       0.623      -0.09
6        0.058
ar.L6.pct_change_1     -0.1346       0.039      -3.455       0.001      -0.21
1       -0.058
ar.L7.pct_change_1      0.2160       0.039       5.600       0.000       0.14
0        0.292
                                    Roots
=============================================================================
===
                  Real          Imaginary           Modulus         Freque
ncy
-----------------------------------------------------------------------------
---
AR.1           -0.9967           -0.5017j            1.1158           -0.4
258
AR.2           -0.9967           +0.5017j            1.1158            0.4
258
AR.3           -0.2984           -1.2147j            1.2509           -0.2
883
AR.4           -0.2984           +1.2147j            1.2509            0.2
883
AR.5            0.9052           -0.9353j            1.3016           -0.1
276
AR.6            0.9052           +0.9353j            1.3016            0.1
276
AR.7            1.4027           -0.0000j            1.4027           -0.0
000
-----------------------------------------------------------------------------
---
```

Below are our forecast results for one day look ahead:

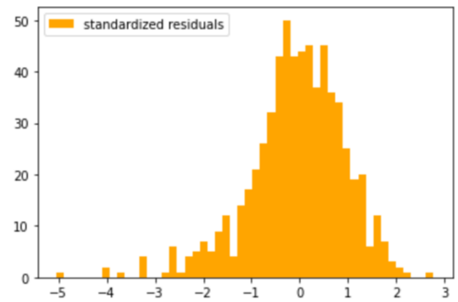


This is an in sample forecast, and I am still trying to figure out how to do an out of sample forecast.

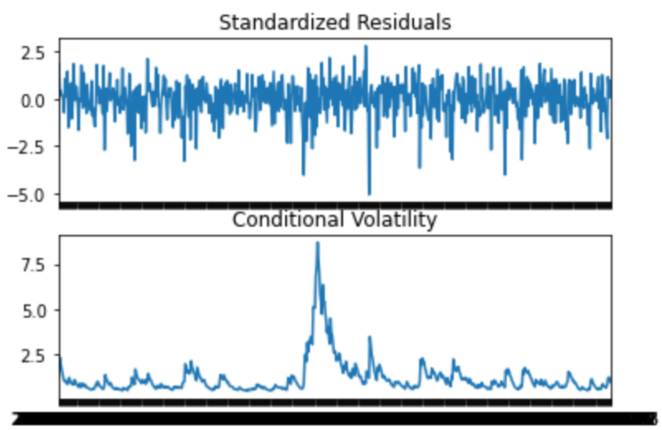
Finally we tried to use a GARCH model to predict to see if it could perform better than ARMA and ARIMAX. We used a GARCH(1,1) model to achieve the below results.

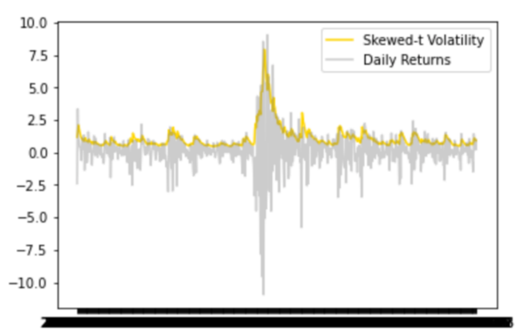
```
Iteration:      4,   Func. Count:     27,   Neg. LLF: 889.966977792967
Iteration:      8,   Func. Count:     50,   Neg. LLF: 889.58575315945
Optimization terminated successfully    (Exit mode 0)
            Current function value: 889.5857515420125
            Iterations: 10
            Function evaluations: 59
            Gradient evaluations: 10
                     Constant Mean - GARCH Model Results
==========================================================================
====
Dep. Variable:           pct_change_1   R-squared:                       0
.000
Mean Model:             Constant Mean   Adj. R-squared:                  0
.000
Vol Model:                      GARCH   Log-Likelihood:               -889
.586
Distribution:                  Normal   AIC:                           178
7.17
Method:            Maximum Likelihood   BIC:                           180
5.05
                                        No. Observations:
645
Date:                Sat, Aug 28 2021   Df Residuals:
644
Time:                        15:45:53   Df Model:
1
                                Mean Model
==========================================================================
                 coef    std err          t      P>|t|    95.0% Conf. Int.
--------------------------------------------------------------------------
mu             0.1402  2.835e-02      4.946  7.567e-07 [8.466e-02,  0.196]
                              Volatility Model
==========================================================================
==
                 coef    std err          t      P>|t|      95.0% Conf. In
t.
--------------------------------------------------------------------------
--
omega          0.0555  1.838e-02      3.018  2.541e-03 [1.945e-02,9.148e-0
2]
alpha[1]       0.3152  7.966e-02      3.956  7.609e-05     [  0.159,  0.47
1]
beta[1]        0.6806  5.468e-02     12.446  1.464e-35     [  0.573,  0.78
8]
==========================================================================
==

Covariance estimator: robust
```



The residuals appear slightly negatively skewed. Let's see if we can improve this.

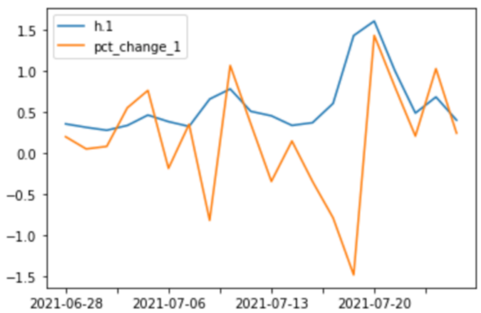




The skewed Student's t-distribution assumption gives us a GARCH model in line with actual observations

Most practitioners prefer to separate the mean and volatility models. We will not do that here but that is a future step. We also want to address the asymmetry issue with stock market returns since volatility in down markets is higher than in upmarkets.

So what we find is that GARCH had the better predictions, smaller Mean absolute error and smallest BIC score.



We can see based on this forecast that there is some asymmetry in volatility that we're not accounting for. It appears that using GARCH to model is very promising. Should explore this further.

Recommendations

Going forward, I would recommend that we explore the SP500 time series more with a GARCH model that separates the mean and volatility models. This means we set the mean parameter to zero when performing garch and using an SARIMAX model to estimate the mean.

I would also recommend that we continue to collect twitter data to see if this is an exogenous variable that we can use to improve the SARIMAX model.

Even though we may be able to obtain some promising forecasts, it's important to note that markets are highly competitive and that the professional institutional quant traders have far more resources to improve investment performance. Hence, we should keep this in mind and favour being conservative in our risk taking while using such models.