Jonathan Cosme

Project Luther

Contents

1 **Background**
Purpose, brief time series background

2 **Models**
Data overview, and model walk-through

3 **Conclusion**
Model evaluation, and further consideration

Background

- Purpose of project
- Time series background

Purpose

A:
Compare Simple & Complex

We will **build two models**:
- “**Simple**” model: 1 variable.
- “**Complex**” model: 13 variables.

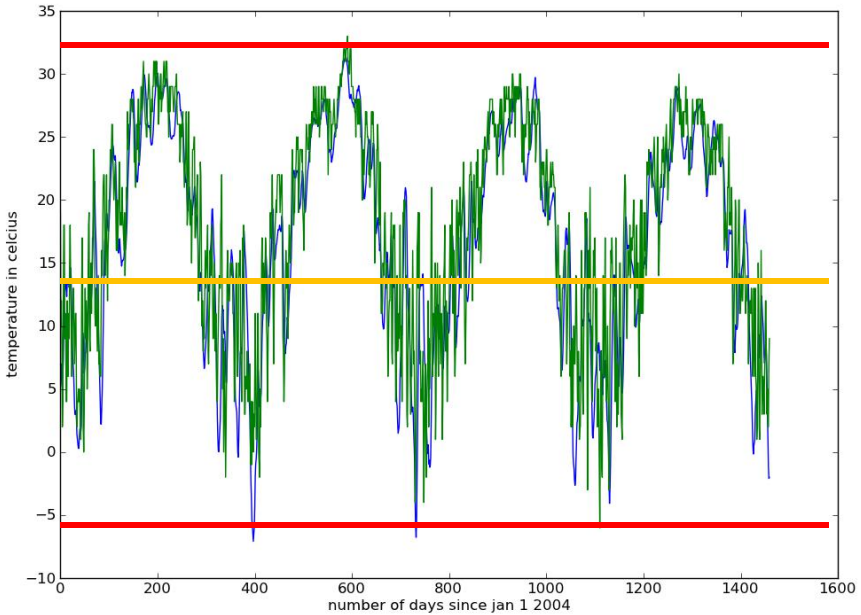
and **compare** the **results**.

B:
Compare NN to rest

We will **build** a **Neural Network** model, and **compare** it’s performance **to** the “**Simple**” **and** “**Complex**” models.

Auto-Regression

- Need **Stationarity**
- E.g. **can** build an AR model **on temperature** data.

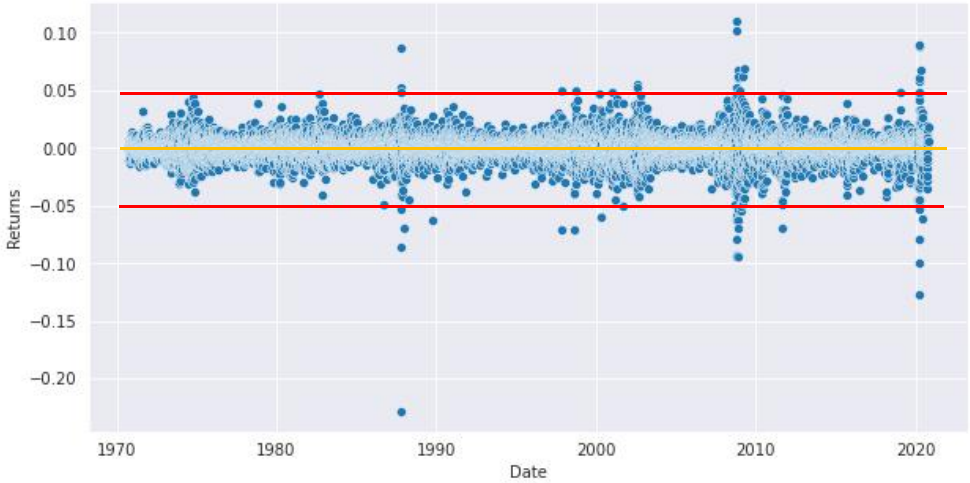


Auto-Regression

- Need Stationarity
- E.g. **cannot** build an AR model **on** stock **prices**!

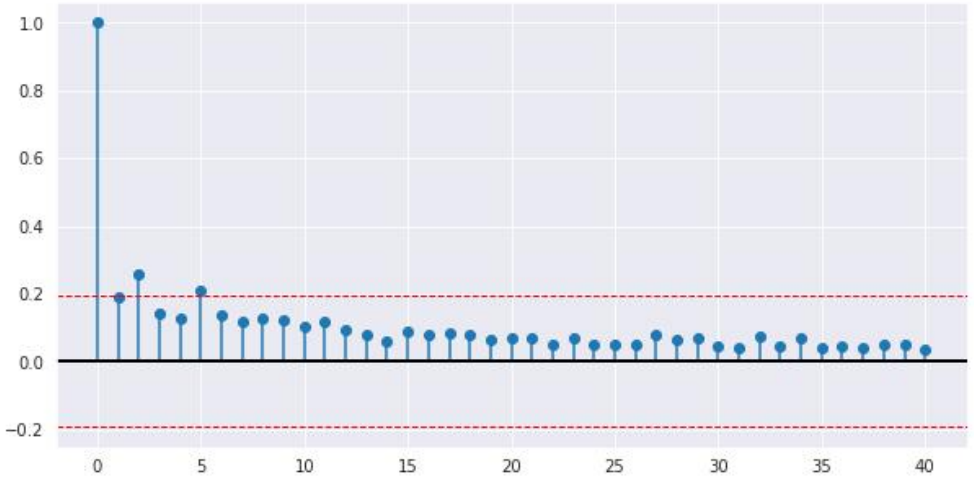
Auto-Regression

- Need Stationarity
- We **use** the **return** (continous % change) **of** stock **prices**.



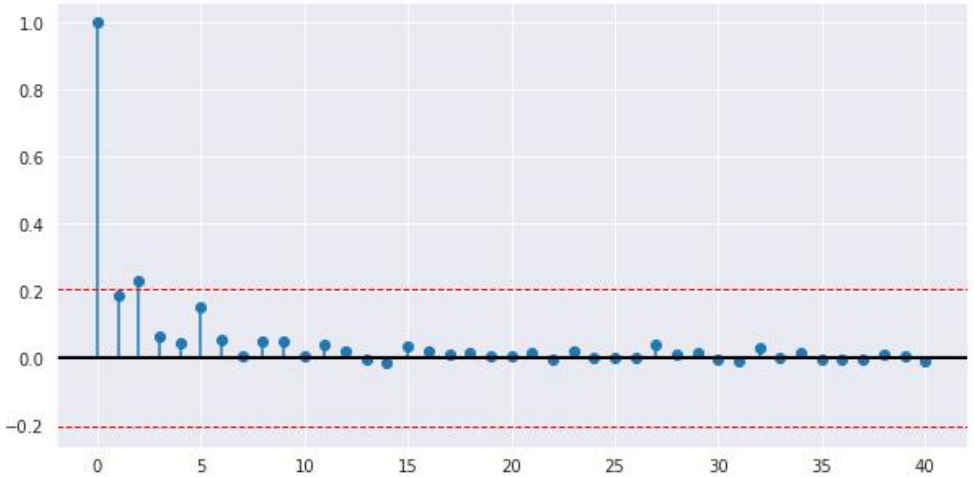
Auto-Regression

- Step 1: Plot the **ACF**



Auto-Regression

- Step 1: Plot the ACF
- Step 2: Plot the **PACF**



Models

- Data used in models
- Simple Model
- Complex Model
- Neural Network Model

The Data

- Scrape **S&P 500 components** from Wikipedia
- Used **Yahoo Finance API** to obtain **price** information **from Oct 1, 2019 to Sep 29, 2020**
- **Scraped** components **stats from Yahoo Finance** Page

	Current	6/30/2020	3/31/2020	12/31/2019	9/30/2019
Market Cap (intraday) [5]	1.95T	1.56T	1.10T	1.29T	995.15B
Enterprise Value [3]	1.97T	1.58T	1.10T	1.30T	1.01T
Trailing P/E	34.60	28.52	20.02	24.70	19.01
Forward P/E [1]	29.67	24.33	19.65	22.17	17.27
PEG Ratio (5 yr expected) [1]	2.82	2.02	1.58	2.03	2.04
Price/Sales (ttm)	7.39	6.12	4.34	5.25	4.09
Price/Book (mrq)	26.99	19.93	12.28	14.23	10.32
Enterprise Value/Revenue [3]	7.20	26.44	18.88	14.11	15.76
Enterprise Value/EBITDA [6]	23.78	95.15	66.00	43.87	50.16

The Data

Volume	DateDelta	Return	HighLow	Market Cap (intraday) 5	Enterprise Value 3	Trailing P/E	Forward P/E 1	PEG Ratio (5 yr expected) 1	Price/Sales (ttm)	Price/Book (mrq)	Enterprise Value/Revenue 3	Enterprise Value/EBITDA 6
3271500.0	1.0	-0.037305	1.044509	9.454000e+10	1.082200e+11	19.83	15.87	4.12	3.03	9.38	13.54	44.10
2972300.0	1.0	-0.017387	1.016750	9.454000e+10	1.082200e+11	19.83	15.87	4.12	3.03	9.38	13.54	44.10
3025700.0	1.0	-0.002702	1.030834	9.454000e+10	1.082200e+11	19.83	15.87	4.12	3.03	9.38	13.54	44.10
2646400.0	1.0	0.003794	1.010687	9.454000e+10	1.082200e+11	19.83	15.87	4.12	3.03	9.38	13.54	44.10
2471200.0	3.0	-0.014871	1.011920	9.454000e+10	1.082200e+11	19.83	15.87	4.12	3.03	9.38	13.54	44.10
...	...	...	...	...	...	...	...	...	...	...	...	...
2017300.0	1.0	0.004057	1.040024	3.060000e+09	5.670000e+09	15.31	24.27	NaN	0.44	1.32	2.73	48.09
3239000.0	1.0	-0.043440	1.089782	3.060000e+09	5.670000e+09	15.31	24.27	NaN	0.44	1.32	2.73	48.09
2027400.0	1.0	-0.007275	1.036829	3.060000e+09	5.670000e+09	15.31	24.27	NaN	0.44	1.32	2.73	48.09
1841600.0	1.0	-0.002132	1.029639	3.060000e+09	5.670000e+09	15.31	24.27	NaN	0.44	1.32	2.73	48.09
1633700.0	3.0	-0.012579	1.044836	3.060000e+09	5.670000e+09	15.31	24.27	NaN	0.44	1.32	2.73	48.09

The Data

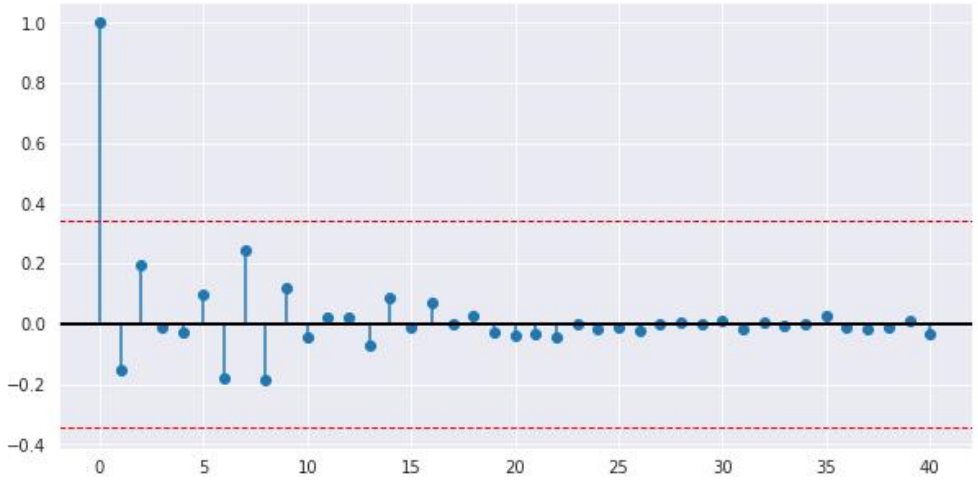
Volume	DateDelta	Return	HighLow	Market Cap (intraday) 5	Enterprise Value 3	Trailing P/E	Forward P/E 1	PEG Ratio (5 yr expected) 1	Price/Sales (ttm)	Price/Book (mrq)	Enterprise Value/Revenue 3	Enterprise Value/EBITDA 6
3271500.0	1.0	-0.037305	1.044509	9.454000e+10	1.082200e+11	19.83	15.87	4.12	3.03	9.38	13.54	44.10
2972300.0	1.0	-0.017387	1.016750	9.454000e+10	1.082200e+11	19.83	15.87	4.12	3.03	9.38	13.54	44.10
3025700.0	1.0	-0.002702	1.030834	9.454000e+10	1.082200e+11	19.83	15.87	4.12	3.03	9.38	13.54	44.10
2646400.0	1.0	0.003794	1.010687	9.454000e+10	1.082200e+11	19.83	15.87	4.12	3.03	9.38	13.54	44.10
2471200.0	3.0	-0.014871	1.011920	9.454000e+10	1.082200e+11	19.83	15.87	4.12	3.03	9.38	13.54	44.10
...	...	...	...	...	...	...	...	...	...	...	...	...
2017300.0	1.0	0.004057	1.040024	3.060000e+09	5.670000e+09	15.31	24.27	NaN	0.44	1.32	2.73	48.09
3239000.0	1.0	-0.043440	1.089782	3.060000e+09	5.670000e+09	15.31	24.27	NaN	0.44	1.32	2.73	48.09
2027400.0	1.0	-0.007275	1.036829	3.060000e+09	5.670000e+09	15.31	24.27	NaN	0.44	1.32	2.73	48.09
1841600.0	1.0	-0.002132	1.029639	3.060000e+09	5.670000e+09	15.31	24.27	NaN	0.44	1.32	2.73	48.09
1633700.0	3.0	-0.012579	1.044836	3.060000e+09	5.670000e+09	15.31	24.27	NaN	0.44	1.32	2.73	48.09

We will have multi-colinearity

Simple Model

Simple Model

- Step 1: Plot the **ACF**
- Note: **No** significant **auto-correlation** in series. Thus, an **AR model is not appropriate** for this data.



Simple AR(1) Model

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.024
Model:	OLS	**Adj. R-squared:**	0.024
Method:	Least Squares	**F-statistic:**	1527.
Date:	Tue, 06 Oct 2020	**Prob (F-statistic):**	0.00
Time:	16:11:30	**Log-Likelihood:**	1.1495e+05
No. Observations:	61507	**AIC:**	-2.299e+05
Df Residuals:	61505	**BIC:**	-2.299e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0026	0.000	-17.030	0.000	-0.003	-0.002
x1	-0.1559	0.004	-39.083	0.000	-0.164	-0.148

Omnibus:	34108.526	**Durbin-Watson:**	1.946
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1747436.161
Skew:	-1.951	**Prob(JB):**	0.00
Kurtosis:	28.819	**Cond. No.**	26.5

Complex Model

Complex Model

- We regress all 13 variables, (plus the 13 “notMissing” encodings), on T+1 returns.

1

Trailing P/E	Forward P/E 1	PEG Ratio (5 yr expected) 1	Price/Sales (ttm)	Price/Book (mrq)	Enterprise Value/Revenue 3	Enterprise Value/EBITDA 6
19.83	15.87	4.12	3.03	9.38	13.54	44.10
19.83	15.87	4.12	3.03	9.38	13.54	44.10
19.83	15.87	4.12	3.03	9.38	13.54	44.10
19.83	15.87	4.12	3.03	9.38	13.54	44.10
19.83	15.87	4.12	3.03	9.38	13.54	44.10
...	...	...	...	...	...	...
15.31	24.27	NaN	0.44	1.32	2.73	48.09
15.31	24.27	NaN	0.44	1.32	2.73	48.09
15.31	24.27	NaN	0.44	1.32	2.73	48.09
15.31	24.27	NaN	0.44	1.32	2.73	48.09
15.31	24.27	NaN	0.44	1.32	2.73	48.09

2

notMissingTrailing P/E	notMissingForward P/E 1	notMissingPEG Ratio (5 yr expected) 1	notMissingPrice/Sales (ttm)	notMissingPrice/Book (mrq)	notMissingEnterprise Value/Revenue 3	notMissingEnterprise Value/EBITDA 6
1.0	1.0	1.0	1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0	1.0	1.0	1.0
...	...	...	...	...	...	...
1.0	1.0	0.0	1.0	1.0	1.0	1.0
1.0	1.0	0.0	1.0	1.0	1.0	1.0
1.0	1.0	0.0	1.0	1.0	1.0	1.0
1.0	1.0	0.0	1.0	1.0	1.0	1.0
1.0	1.0	0.0	1.0	1.0	1.0	1.0

3

Trailing P/E	Forward P/E 1	PEG Ratio (5 yr expected) 1	Price/Sales (ttm)	Price/Book (mrq)	Enterprise Value/Revenue 3	Enterprise Value/EBITDA 6
19.83	15.87	4.12	3.03	9.38	13.54	44.10
19.83	15.87	4.12	3.03	9.38	13.54	44.10
19.83	15.87	4.12	3.03	9.38	13.54	44.10
19.83	15.87	4.12	3.03	9.38	13.54	44.10
19.83	15.87	4.12	3.03	9.38	13.54	44.10
...	...	...	...	...	...	...
15.31	24.27	0.00	0.44	1.32	2.73	48.09
15.31	24.27	0.00	0.44	1.32	2.73	48.09
15.31	24.27	0.00	0.44	1.32	2.73	48.09
15.31	24.27	0.00	0.44	1.32	2.73	48.09
15.31	24.27	0.00	0.44	1.32	2.73	48.09

4

PEG Ratio (5 yr expected) 1	Price/Sales (ttm)	...	notMissingHighLow	notMissingMarket Cap (intraday) 5	notMissingEnterprise Value 3	notMissingTrailing P/E	notMissingForward P/E 1	notMissingPEG Ratio (5 yr expected) 1
4.12	3.03	...	1.0	1.0	1.0	1.0	1.0	1.0
4.12	3.03	...	1.0	1.0	1.0	1.0	1.0	1.0
4.12	3.03	...	1.0	1.0	1.0	1.0	1.0	1.0
4.12	3.03	...	1.0	1.0	1.0	1.0	1.0	1.0
4.12	3.03	...	1.0	1.0	1.0	1.0	1.0	1.0
...	...	...	...	...	...	...	...	...
0.00	0.44	...	1.0	1.0	1.0	1.0	1.0	0.0
0.00	0.44	...	1.0	1.0	1.0	1.0	1.0	0.0
0.00	0.44	...	1.0	1.0	1.0	1.0	1.0	0.0
0.00	0.44	...	1.0	1.0	1.0	1.0	1.0	0.0
0.00	0.44	...	1.0	1.0	1.0	1.0	1.0	0.0

Complex Model

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.037
Model:	OLS	**Adj. R-squared:**	0.037
Method:	Least Squares	**F-statistic:**	125.2
Date:	Tue, 06 Oct 2020	**Prob (F-statistic):**	0.00
Time:	16:16:50	**Log-Likelihood:**	1.1536e+05
No. Observations:	61507	**AIC:**	-2.307e+05
Df Residuals:	61487	**BIC:**	-2.305e+05
Df Model:	19		
Covariance Type:	nonrobust		

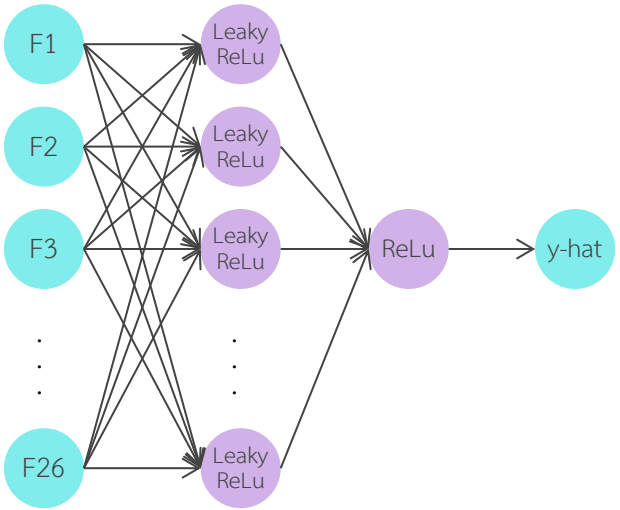
	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0077	0.001	11.747	0.000	0.006	0.009
x1	-3.736e-11	1.49e-11	-2.499	0.012	-6.67e-11	-8.06e-12
x2	0.0037	0.000	20.847	0.000	0.003	0.004
x3	-0.1600	0.004	-39.591	0.000	-0.168	-0.152
x4	-0.0623	0.004	-15.422	0.000	-0.070	-0.054
x5	-5.352e-16	5.37e-15	-0.100	0.921	-1.11e-14	9.99e-15
x6	3.975e-15	5.41e-15	0.735	0.462	-6.63e-15	1.46e-14
x7	-2.475e-06	1.34e-06	-1.844	0.065	-5.11e-06	1.56e-07
x8	-2.414e-05	1.02e-05	-2.363	0.018	-4.42e-05	-4.12e-06
x9	1.79e-05	7.92e-06	2.260	0.024	2.38e-06	3.34e-05
x10	0.0004	0.000	3.397	0.001	0.000	0.001
x11	1.31e-06	7.21e-06	0.182	0.856	-1.28e-05	1.54e-05
x12	-5.369e-05	3e-05	-1.790	0.073	-0.000	5.09e-06
x13	3.613e-06	1.16e-06	3.124	0.002	1.35e-06	5.88e-06
x14	0.0077	0.001	11.747	0.000	0.006	0.009
x15	0.0077	0.001	11.747	0.000	0.006	0.009
x16	0.0077	0.001	11.747	0.000	0.006	0.009
x17	0.0077	0.001	11.747	0.000	0.006	0.009
x18	0.0077	0.001	11.747	0.000	0.006	0.009
x19	0.0007	0.001	1.263	0.207	-0.000	0.002
x20	0.0009	0.001	1.215	0.224	-0.001	0.002
x21	-0.0004	0.001	-0.255	0.799	-0.003	0.002
x22	0.0006	0.000	1.399	0.162	-0.000	0.002
x23	0.0077	0.001	11.747	0.000	0.006	0.009
x24	-0.0004	0.001	-0.573	0.567	-0.002	0.001
x25	0.0007	0.001	1.263	0.207	-0.000	0.002
x26	9.651e-05	0.001	0.127	0.899	-0.001	0.002

Omnibus:	29744.239	**Durbin-Watson:**	1.963
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1585265.867
Skew:	-1.566	**Prob(JB):**	0.00
Kurtosis:	27.673	**Cond. No.**	5.88e+17

Neural Network Model

Neural Network Model

- We use 2 layers.
- L1: 26 nodes; leaky relu activation
- L2: 1 node; relu activation

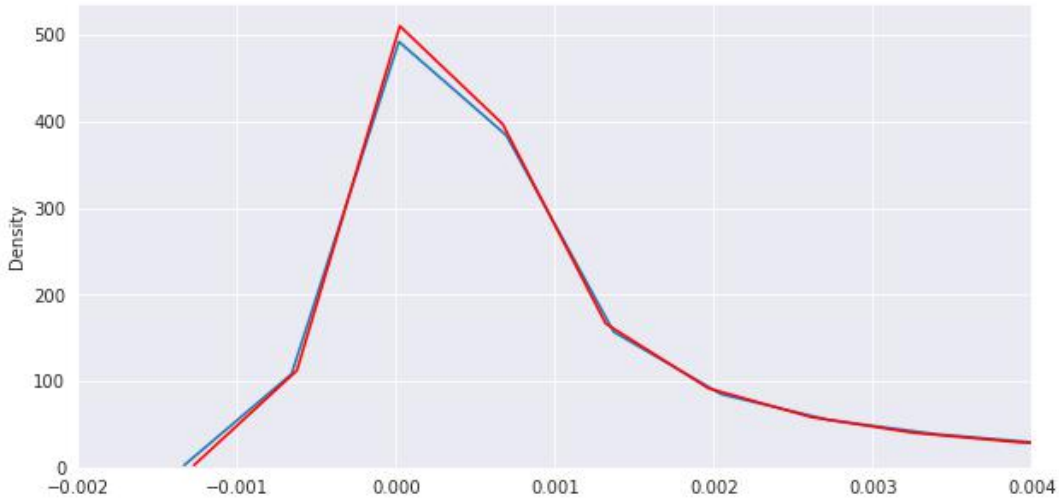


Conclusion

- Model Performance
- Further consideration

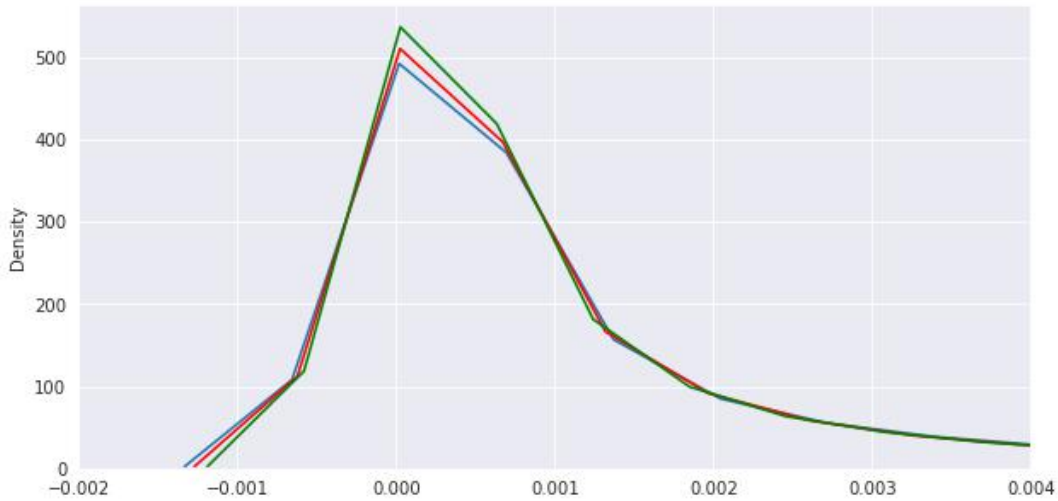
Simple vs Complex

- Dev set **MSE** for Simple model:
  0.0014718622471362882
- Dev set **MSE** for Complex model:
  0.0015187861917754437
- One-way ANOVA results:
  F-Stat = 2.8797 (pval = 0.0897)



NN vs Rest

- Dev set **MSE** for Neural Net model: 0.0014027121828562 31
- One-way ANOVA results: F-Stat = 7.003 (pval = 0.0081)
- Dev set **MSE** for Simple model: 0.0014718622471362882
- Dev set **MSE** for Complex model: 0.0015187861917754437



Further Consideration

- **Using levels** rather than returns makes **performance worse**, in addition to violating the stationarity condition of AR models.
- Using **more than 1 year of data** would also help build a better model.
- Note **S&P 500 components change over time**; it's important to make sure one gets correct components for the time period.
- Building a **categorical** predictive **model**, rather than a coninuous model (predict if returns will be positive or negative, rather than the value of the return) might be **more appropriate** for this data.
- Building a continuous **model for** predicting the **squared returns** (volatility) would also be **more appropriate** for this data. Serial correlation of volatility is a well known phenomenon.
- Combining the above two: build a categorical predictive model, then a model for predicting squared returns for each, would be an interesting thing to explore.

Thanks!

Does anyone have any questions?