NetFlix Market Cap Time Series Analysis

Instructor: Arthur Small

Author: Qimin Luo

Abstract

Netflix is one of the most successful video websites in the United States. Futher, it is also the most profitable video website in the world. It is interesting to explore its future development. In this experiment, we apply Prophet procedure to build a time series model and make preditions about its future market cap. Additionally, we also use GARCH model to fit data which is popular in the finance field. Further, we make preditions based on the best model.

Introduction

Motivations

Netflix is already one of the biggest media companyies in the world. Depsite this, there are some fast-growing media companies like Hulu, Youtube is running after it. We are interested to explore its future development. We plan to apply Prophet procedure to build a time series model to predict its future market cap. Its future development is impossible to be determined only by market cap. However, it is expected to find out some clues about future car development trend.

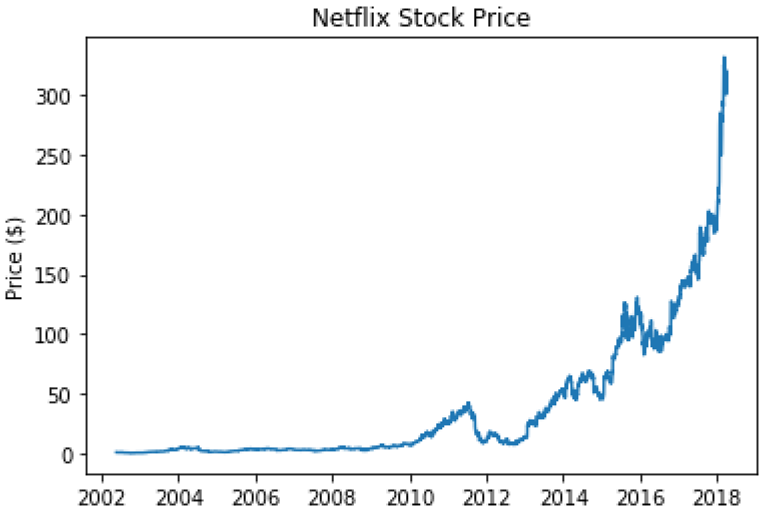
Literature Review

The autoregressive conditional heteroscedasticity (ARCH) model is created by Robert F. Engle. In financial markets, analysts observe something called volatility clustering in which periods of low volatility are followed by periods of high volatility and vice versa. ARCH models are able to correct for the statistical problems that arise from this type of pattern in the data. As a result, they have become mainstays in modeling financial markets that exhibit volatility. [1]

To better fit seasonal data, prophet procedure is explored by Facebook. Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. [2]

Data and Data Generating Process

It is a pretty complicated and time-consuming process to retrive and well-structured data from Internet manully. In this case, I use a convenient python financial data package -- Quandl. It contains almost unlimited financial data. Further, it can provide us with well-formated data.



The market cap of a company can be affected by lots of variables. It is reason take variables like daily stock prices and stock vlomue into account. However, some social and political event can also have crucial effecs on market cap. We are not able to add these things into model so we have to skip these unpredictable factors.

In this case, it is clear that variables like open price, daily highest price, daily lowest price and close price have directive influence on its market cap. Further, stock volume is also important because it is not fair to determine market cap by its stock price. For example, some companies have very high stock prices but its stock volume is small. In this case, its market cap is limited by volume. Addtionally, we conisder some adjusted factors like Adj. High since adjusted variables sometimes can demonstrate information clearly.

$$Cap\ value \ = \ Adj.Close\ Price\ *\ Stock\ volume$$

	Date	Open	High	Low	Close	Volume	Ex-Dividend	Split Ratio	Adj. Open	Adj. High	Adj. Low	Adj. Close	Vol
0	2005-01-03	12.48	12.60	11.52	11.92	1619200.0	0.0	1.0	1.782857	1.800000	1.645714	1.702857	113344
1	2005-01-04	11.92	11.95	11.25	11.66	2478900.0	0.0	1.0	1.702857	1.707143	1.607143	1.665714	173523
2	2005-01-05	11.74	11.74	11.09	11.20	1818900.0	0.0	1.0	1.677143	1.677143	1.584286	1.600000	127323
3	2005-01-06	11.20	11.37	11.01	11.05	1181900.0	0.0	1.0	1.600000	1.624286	1.572857	1.578571	82733
4	2005-01-07	11.11	11.55	11.00	11.12	1070100.0	0.0	1.0	1.587143	1.650000	1.571429	1.588571	74907

Exploratory Data Analysis

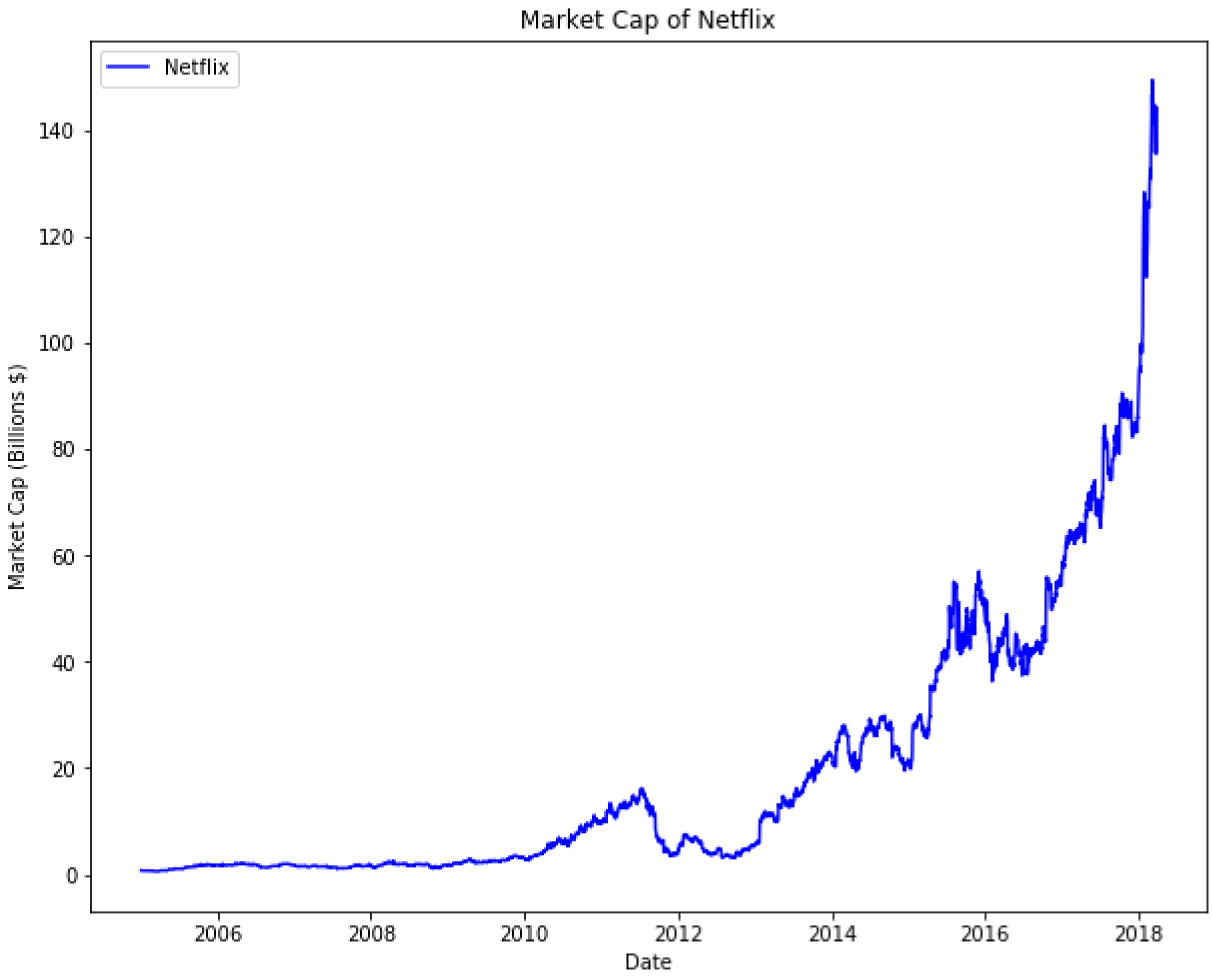
Data Format

After the calculations and data cleaning, the data format is shown below. It is a time series data framework. There are only two attributes: 'Date' and 'Cap'. The 'Cap' indicates the Netflix market cap which we are interested in.

	Date	cap
0	2005-01-03	0.781611
1	2005-01-04	0.764563
2	2005-01-05	0.734400
3	2005-01-06	0.724564
4	2005-01-07	0.729154

Visualization of Cap Value

We plot the data below. The blue curve notes the trend of Netflix market cap with time. The market cap grows slow before 2016 but it increases rapidly later. And its trend seems not linear and it might need log transformation to make it stationary.



Statistical Model

FB Prophet

Data Generating Process

Prophet Model can be roughly represented by these formulas below.

$$y(t) = g(t) + h(t) + s(t) + \epsilon(t)$$

$$g(t) = (k + \alpha(t)\delta) \cdot t + (m + \alpha(t)^T\gamma) \quad (1)$$

$$s(t) = \sum_{n=1}^{N}(a_n cos(\frac{2\pi nt}{p}) + b_n sin(\frac{2\pi nt}{p})) \quad (2)$$

$$h(t) = Z(t)\mathbf{k} \quad (3)$$

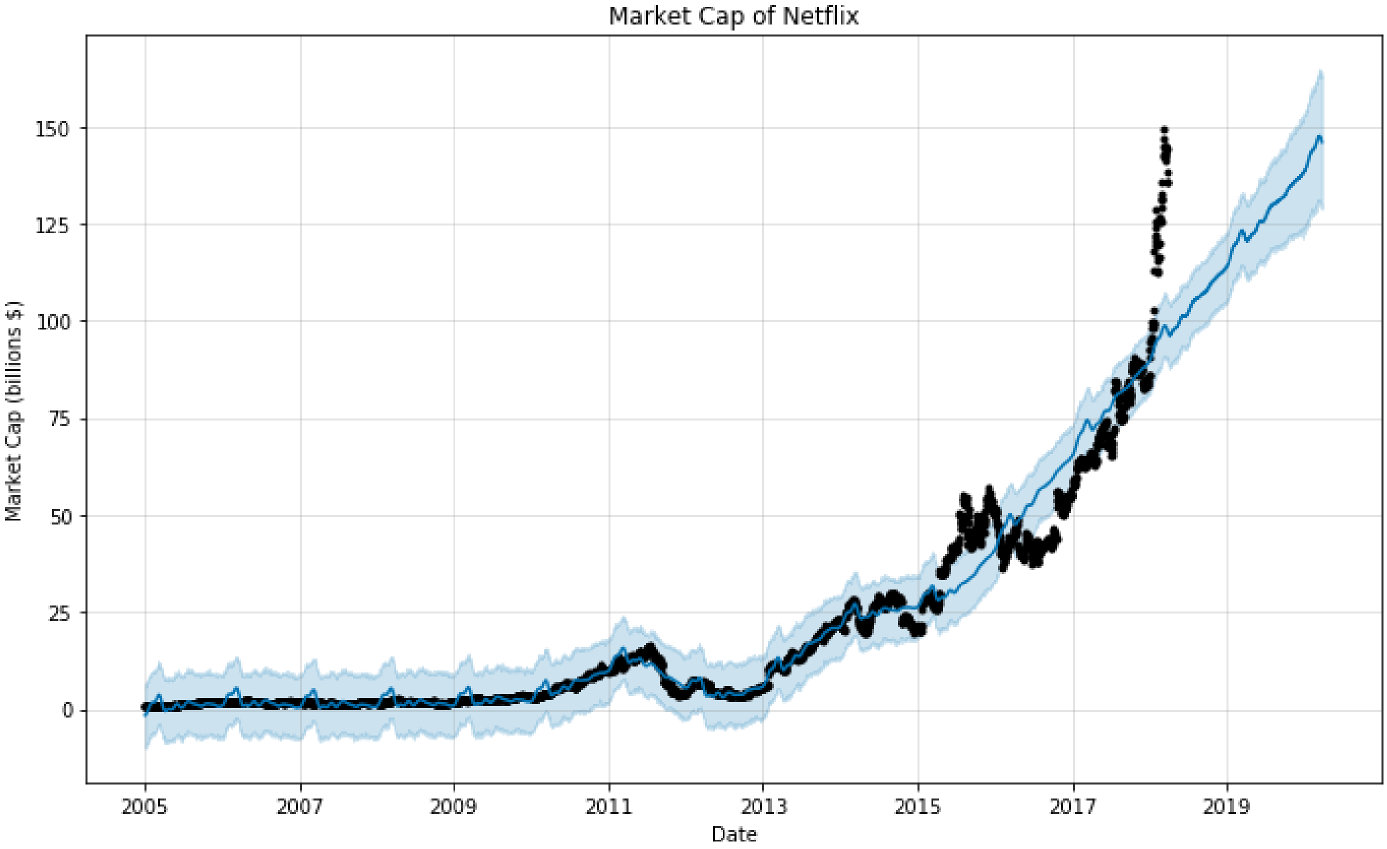
$$Z(t) = [1(t \in D_1), \ldots, 1(t \in D_L)], \mathbf{k} = (k_1, \ldots, k_L)^T$$

Model Discussion

I decide to apply Prophet Model which is explored by Facebook. Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. The financial data usually have seasonal variations and it also related with holidays. Therefore, I think this model might fit the data well.

Model Fitting Process

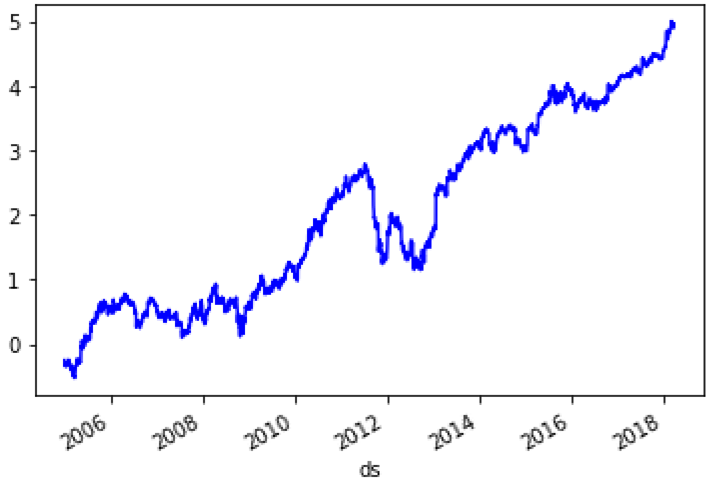
The fitting result is shown in the figure below. The black points indicate the actual data point and the blue line represents the model we fit and our predictions. Additionallly, the light blue region area states the upper bound and the lower bound of our model. However, the model didn't fit the data well based on the plot. The blue line is not close to our data and the variation area seems wide. In this case, we decided to use log transformation on the data.



Log Transformation

We applied log transformation on the data and the new data is shown below. Further, we also plot the trend of the new data. Though it is still not stationary enough, the trend seems much better and more close to linear relationship.

	ds	y
0	2005-01-03	-0.246398
1	2005-01-04	-0.268451
2	2005-01-05	-0.308701
3	2005-01-06	-0.322185
4	2005-01-07	-0.315870



GARCH Model

Data Generating Process

The autoregressive conditional heteroscedasticity (ARCH) model is a statistical model for time series data that describes the variance of the current error term or innovation as a function of the actual sizes of the previous time periods' error terms[1]; often the variance is related to the squares of the previous innovations. The ARCH model is appropriate when the error variance in a time series follows an autoregressive (AR) model; if an autoregressive moving average (ARMA) model is assumed for the error variance, the model is a generalized autoregressive conditional heteroskedasticity (GARCH) model[3].

$r_t = c_1 + \sum_{i=1}^{R} \phi_i r_{t-i} + \sum_{j=1}^{M} \phi_j \epsilon_{t-j} + \epsilon_t$

$\epsilon_t = u_t \sqrt{h_t}$

$h_t = k + \sum_{i=1}^{q} G_i h_{t-i} + \sum_{i=1}^{p} A_i \epsilon_{t-i}^2$

Model Discussion

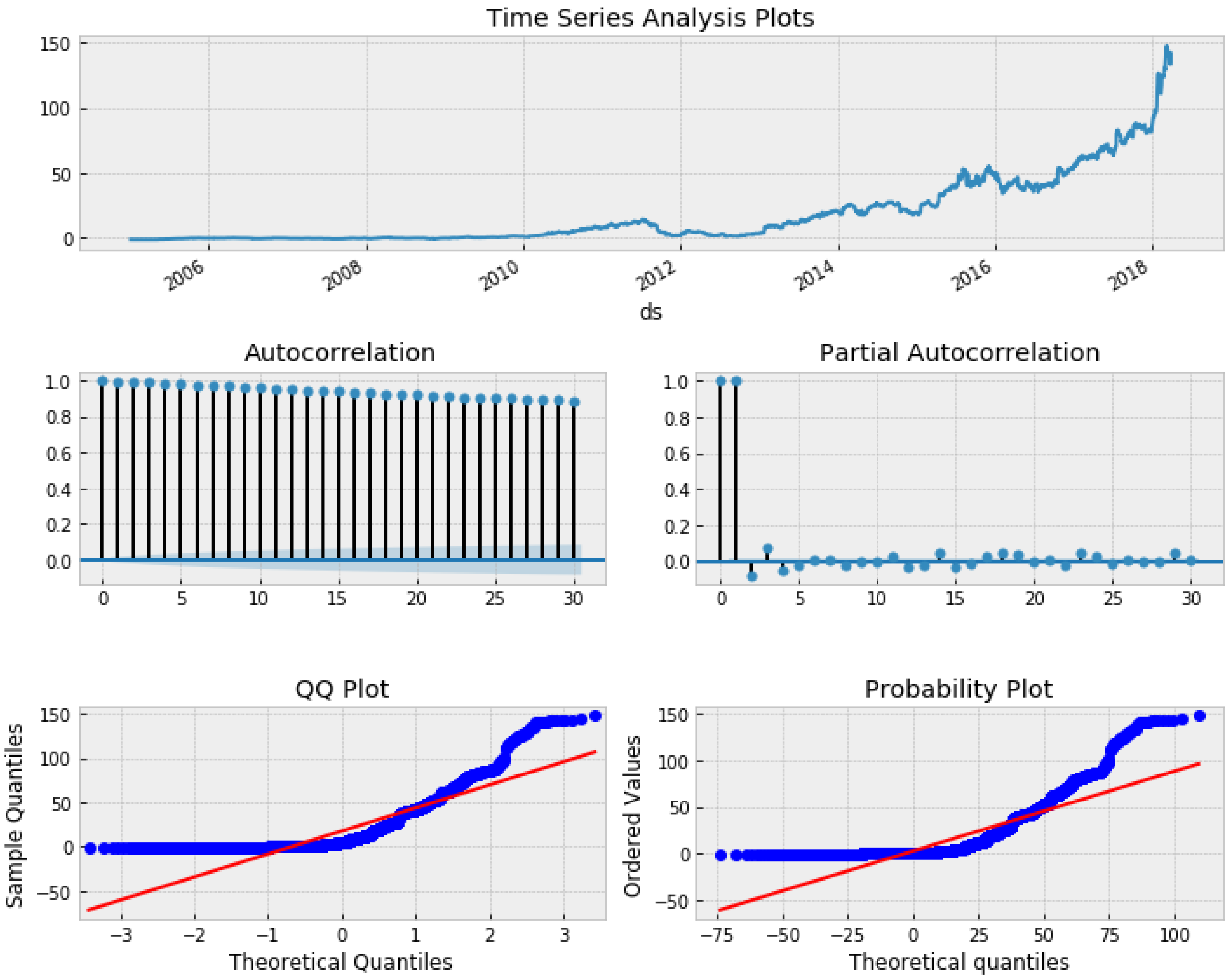
ARCH models are commonly employed in modeling financial time series that exhibit time-varying volatility and volatility clustering, i.e. periods of swings interspersed with periods of relative calm. ARCH-type models are sometimes considered to be in the family of stochastic volatility models, although this is strictly incorrect since at time t the volatility is completely pre-determined (deterministic) given previous values[4]. In this case, the variance of stock prices is not constant so the traditional regression model is not feasible here. The variance of stock price is likely to vary with time. Therefore, GARCH model maybe be suitable in this case.

Model Fitting Process

Firstly, we use garch model fit the original data. Based on the result table below, R-squared is low and AIC and BIC are a bit large. Therefore, the model didn't fit the original data well. And we also made serval plots and they validated the conclusion. The PACF is fine but ACF plots demonstartes there exist high correlation between data points. What's more, the QQ plot indicates the model fitting is bad. As a result, we tried log transformation and difference transformation.

```
                     Constant Mean - GARCH Model Results
==============================================================================
Dep. Variable:                      y   R-squared:                      -0.464
Mean Model:             Constant Mean   Adj. R-squared:                 -0.464
Vol Model:                      GARCH   Log-Likelihood:               -8919.93
Distribution:                  Normal   AIC:                           17847.9
Method:            Maximum Likelihood   BIC:                           17872.3
                                        No. Observations:                 3330
Date:                Fri, Dec 11 2020   Df Residuals:                     3326
Time:                        14:00:09   Df Model:                            4
                               Mean Model
========================================================================
                 coef    std err          t      P>|t|  95.0% Conf. Int.
------------------------------------------------------------------------
mu             1.7959  2.807e-02     63.985      0.000 [  1.741,  1.851]
                               Volatility Model
=============================================================================
                 coef    std err          t      P>|t|       95.0% Conf. Int.
-----------------------------------------------------------------------------
omega      3.2474e-03  7.138e-04      4.550  5.376e-06  [1.848e-03,4.646e-03]
alpha[1]       0.9965  8.078e-03    123.364      0.000      [  0.981,  1.012]
beta[1]    6.0694e-04  1.345e-03      0.451      0.652 [-2.029e-03,3.242e-03]
=============================================================================

Covariance estimator: robust
```

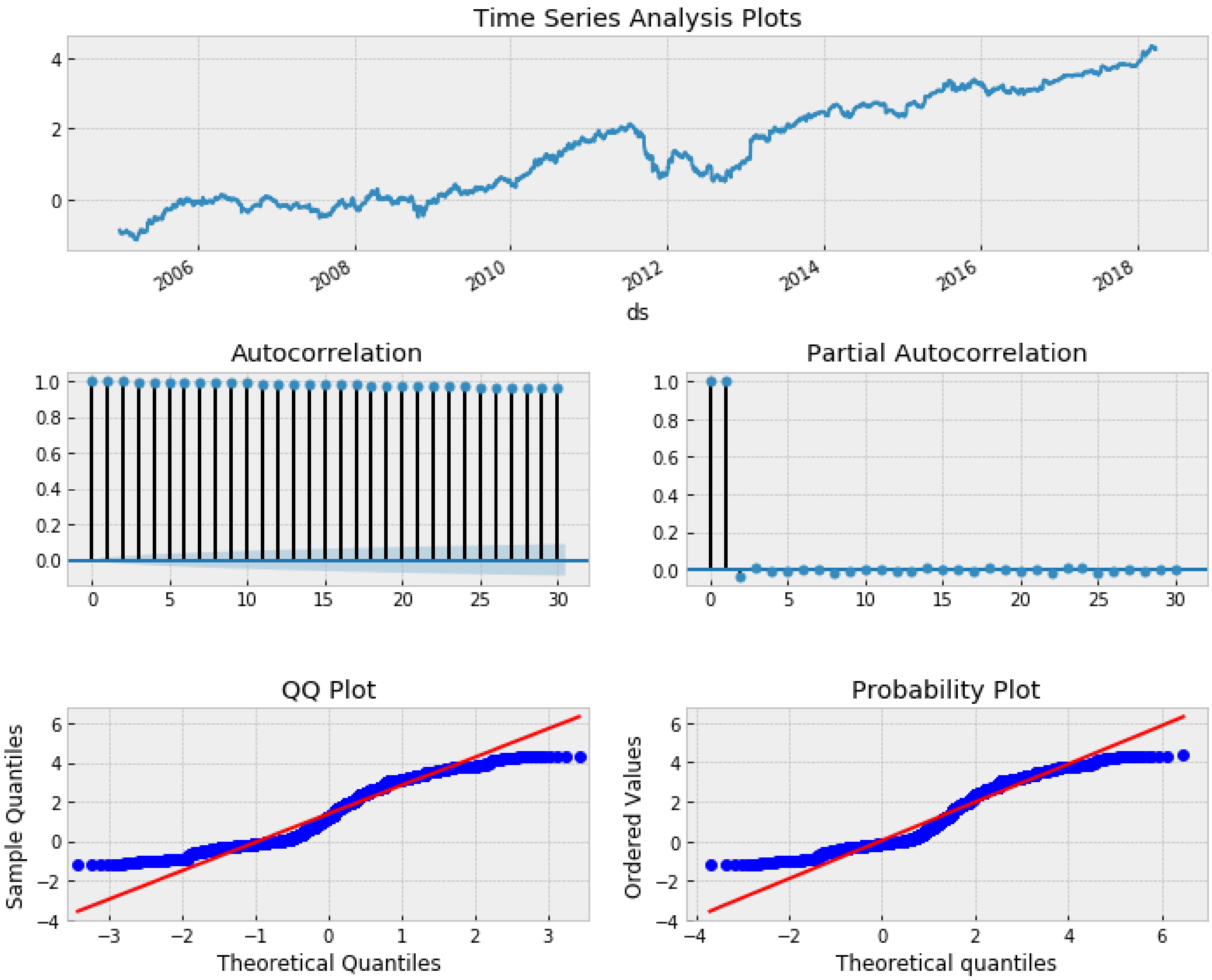


Model with Log Transformation

After log transformation, the statistics is much better. In the table below, the R-squared is over 0.90 and AIC and BIC are much smalller. The trend in the plot is also more stationary and QQ plot shows the model fitting is good. However, the problem in ACF plot still exist. Correlation sometimes is not a problem for the model. We don't need to care about the correlation problem if we only want good preditions. And in this case, we want a good forcasting of Netflix market cap so the correlation problem is not crucial here. Despite this, we also tried difference transformation to fix the correlation problem.

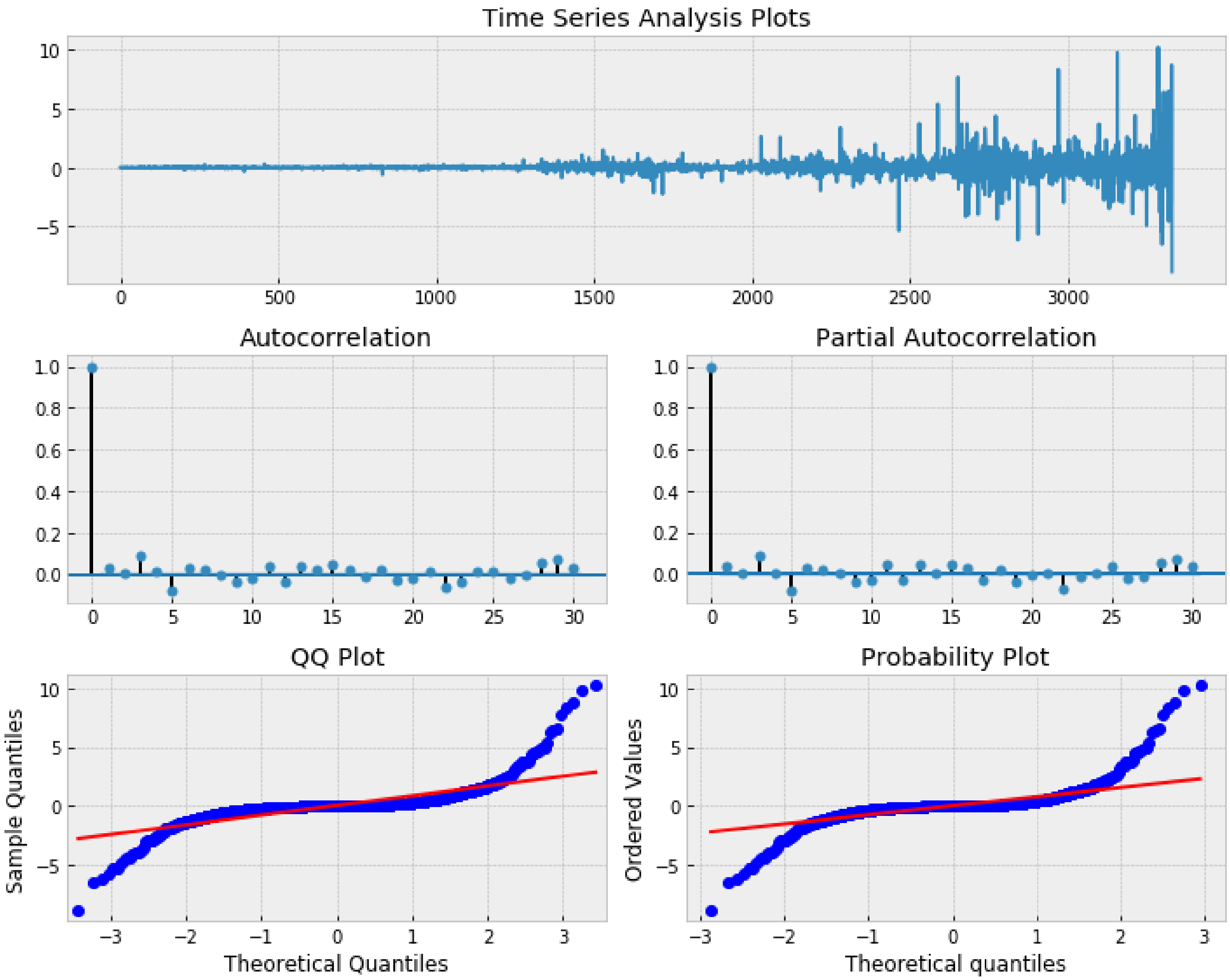
```
                     Constant Mean - GARCH Model Results
==============================================================================
Dep. Variable:                      y   R-squared:                      -0.925
Mean Model:             Constant Mean   Adj. R-squared:                 -0.925
Vol Model:                      GARCH   Log-Likelihood:               -3968.96
Distribution:                  Normal   AIC:                           7945.92
Method:            Maximum Likelihood   BIC:                           7970.36
                                        No. Observations:                 3330
Date:                Fri, Dec 11 2020   Df Residuals:                     3326
Time:                        14:01:26   Df Model:                            4
                               Mean Model
========================================================================
                 coef    std err          t      P>|t|  95.0% Conf. Int.
------------------------------------------------------------------------
mu             0.6481  6.483e-03     99.964      0.000 [  0.635,  0.661]
                               Volatility Model
=============================================================================
                 coef    std err          t      P>|t|       95.0% Conf. Int.
-----------------------------------------------------------------------------
omega      6.1575e-04  1.188e-04      5.184  2.174e-07  [3.829e-04,8.486e-04]
alpha[1]       1.0000  2.174e-02     45.992      0.000      [  0.957,  1.043]
beta[1]        0.0000  1.812e-02      0.000      1.000 [-3.551e-02,3.551e-02]
=============================================================================

Covariance estimator: robust
```



Difference Transformation

After difference transformation, the correlation problem is fixed according to the ACF plot and PACF plot. Further, the QQ plot indicates the model fitting is good. However, the model with log transformation is closer to linear relationship compared with the model with difference transformation.



Results

GARCH Simulation Forecasts

Simulation-based forecasts use the model random number generator to simulate draws of the standardized residuals, e_{t+h}. These are used to generate a pre-specified number of paths of the variances which are then averaged to produce the forecasts. In models like GARCH which evolve in the squares of the residuals, there are few advantages to simulation-based forecasting.

Assume there are B simulated paths. A single simulated path is generated using

$$\sigma^2_{t+h,b} = \omega + \alpha\epsilon^2_{t+h-1,b} + \beta\sigma^2_{t+h-1,b}$$

$$\epsilon_{t+h,b} = e_{t+h,b}\sqrt{\sigma^2_{t+h,b}}$$

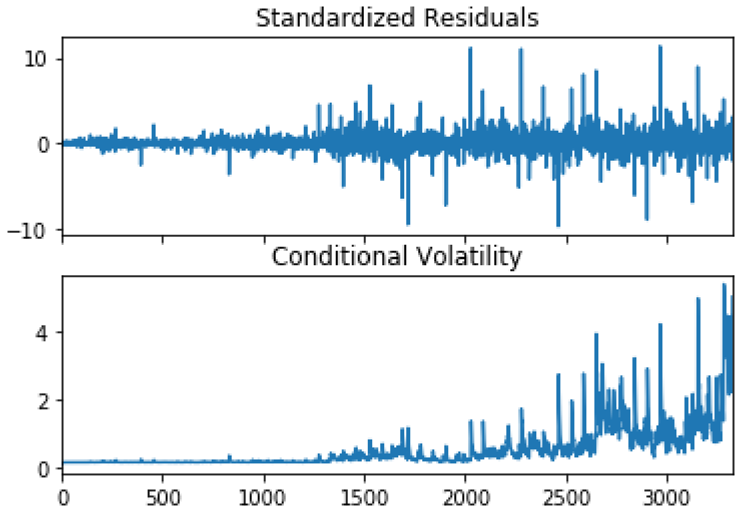
where the simulated shocks are $e_{t+1,b}, e_{t+2,b}, \ldots, e_{t+h,b}$ where b is included to indicate that the simulations are independent across paths. Note that the first residual, ϵ_t, is in-sample and so is not simulated.

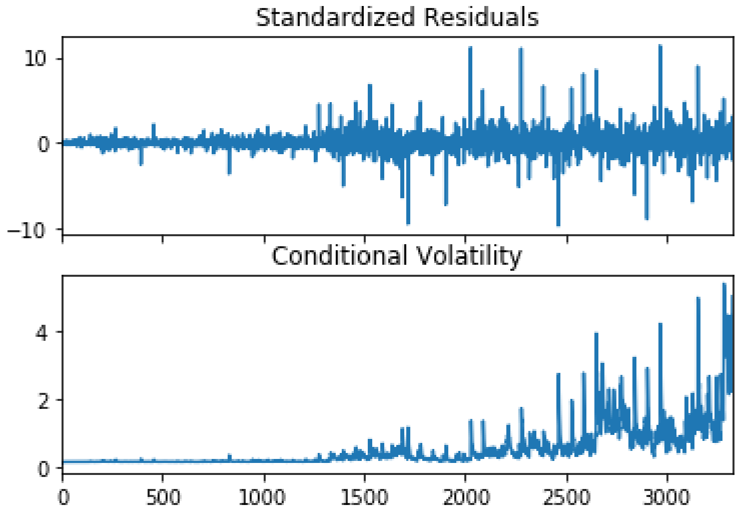
The final variance forecasts are then computed using the B simulations

$$E_t[\epsilon^2_{t+h}] = \sigma^2_{t+h} = B^{-1} \sum_{b=1}^{B} \sigma^2_{t+h,b}$$

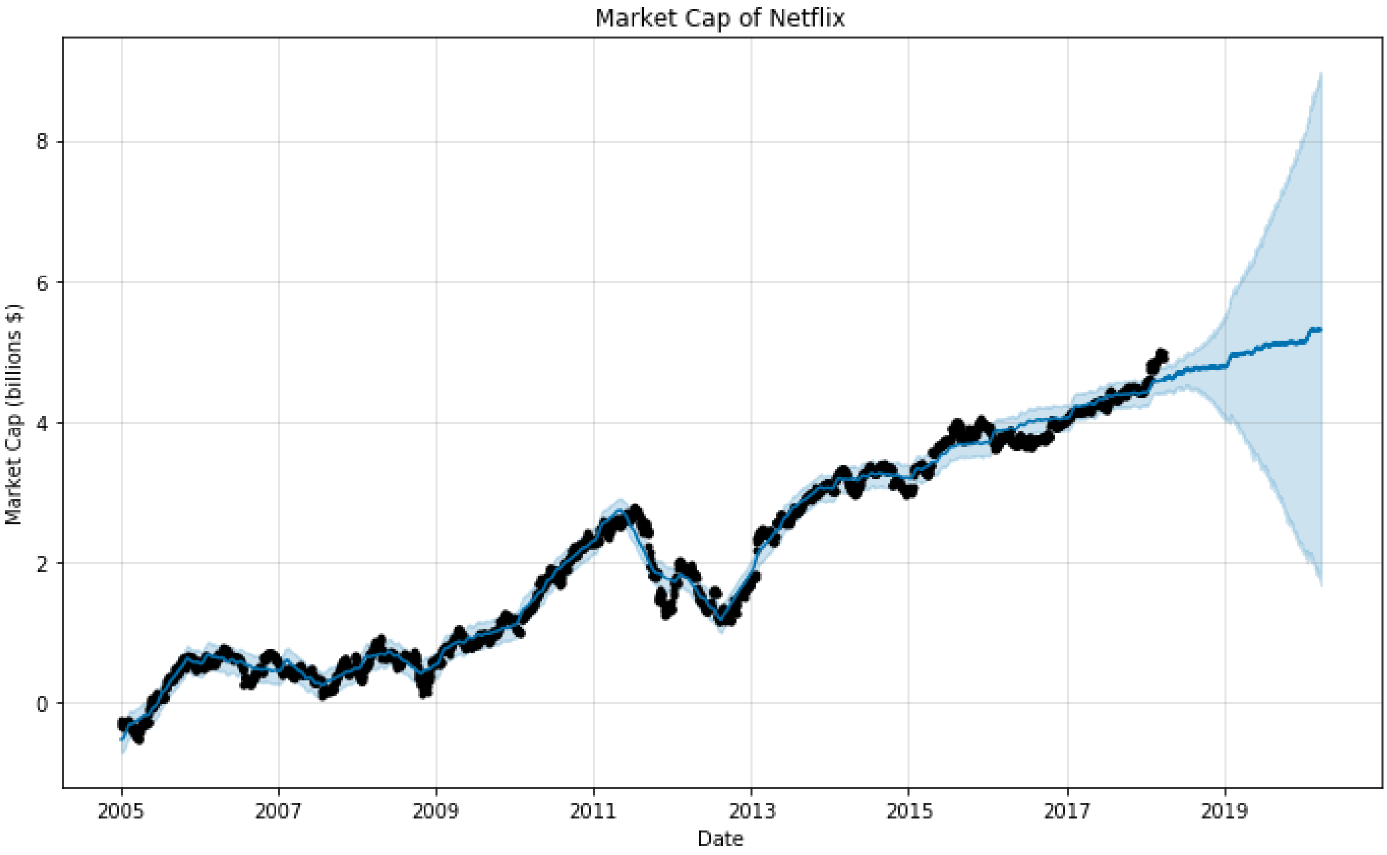
Discussion and Conclusion

We used GARCH model predict volatilities but there is a trade-off. If we are more interested in the relationship between data points, the model with difference transformation is more suitable. If we only care about the preditions, we could apply the model with log transformation though it has the correlation problem.





The prophet also has a good performance with log transformation. We can conclude that the market cap of Netflix will keep high increasing rate based on our model and the fact actually validates that. Despite this, the model might have overfitting problem. One simple way of mitigating ovefitting is to perform extensive back-testing. In practice, it means the process needs to split the input dataset over dozens - if not hundreds - of incremental date thresholds, and re-train all the forecasting models and re-assess them each time. Backtesting requires a lot of processing power. In the future work, this method can be used to solve overfitting problem. Moreover, the Netflix had a big jump during the quarantine. It is hard to consider the unexpected events in the model and it is always a big issue in time series analysis. Prophet model partly takes these events into considerations but it is not enough. I hope there will a more advanced model can solve this problem in the future for better preditions.



Reference

[1] Engle, Robert F. (1982). "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation". Econometrica. 50 (4): 987–1007. doi:10.2307/1912773. JSTOR 1912773.

[2] https://facebook.github.io/prophet/ (https://facebook.github.io/prophet/)

[3] Bollerslev, Tim (1986). "Generalized Autoregressive Conditional Heteroskedasticity". Journal of Econometrics. 31 (3): 307–327. CiteSeerX 10.1.1.468.2892. doi:10.1016/0304-4076(86)90063-1.

[4] Brooks, Chris (2014). Introductory Econometrics for Finance (3rd ed.). Cambridge: Cambridge University Press. p. 461. ISBN 9781107661455.