2020

Predicting Stock Prices Using Gaussian Process Regression

PREPARED BY:
FUAT CEM ÖZYAZICI & ALP GÖZAYDIN

KOÇ UNIVERSITY | QMBU 450 – Advanced Data Analysis Using Python | Instructor: Asst. Prof. David Carlson

Preface

In financial markets, traders make binary decisions and make or lose money based on their decisions. According to the famous trader and trading coach Mark Douglas, there are five fundamental truths about trading, which are very important to understand the uncertainty concept of trading. They are:

1. Anything can happen.
2. You don’t need to know what is going to happen next in order to make money.
3. There is a random distribution between wins and losses for any given set of variables that define an edge.
4. An edge is nothing more than an indication of a higher probability of one thing happening over another.
5. Every moment in the market is unique.

“When traders understand that trading is simply a probability game, concepts like right or wrong or win or lose no longer have the same significance. Putting on a winning trade or even a series of winning trades require absolutely no skill. On the other hand, creating consistent results and being able to keep what you have created require skill. Making money consistently is a by-product of acquiring and mastering mental skills. When you genuinely accept the risks, you will be at peace with any outcome. It is the ability to believe in the unpredictability of the game at the micro level and simultaneously believe in the predictability of the game at the macro level that makes the casino and the professional blackjack card counter effective and successful at what they do. **The hard, cold reality of trading is that every trade has an uncertain outcome. Trading is not about being right or wrong. It’s a probability game.”**

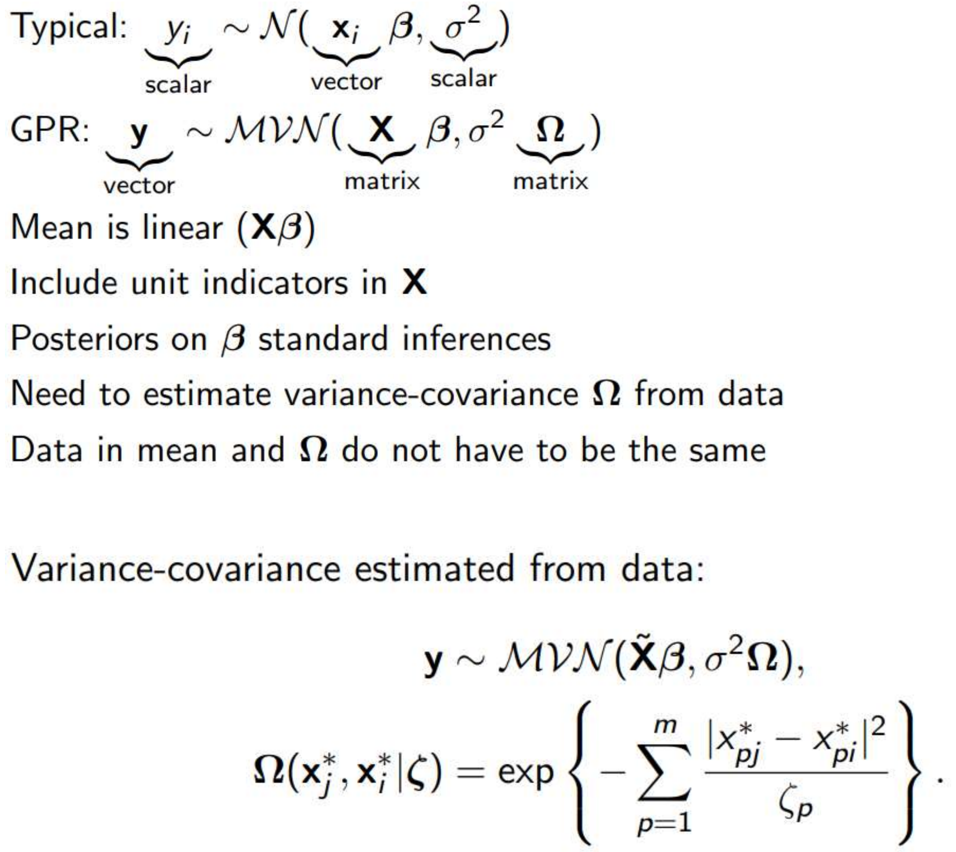
Mark Douglas – Trading in the Zone

In this project, we will use Gaussian Process Regression, to predict IBM's and Microsoft's stock prices for year 2018, and Q3-Q4 of 2019. Our hypothesis is that, if we create a gaussian process regression for the available historical data starting from 2008 of IBM and MSFT stocks, all of the close prices of the trading days in 2018, and 2019 Q3-Q4 will lay in the 95% confidence interval of our prediction model.

Let us start with an introduction to gaussian processes.

Gaussian Processes

The gaussian process is an alternative Bayesian approach to regression problems. "GP defines a priori over functions that can be converted into a posteriori once we have observed a few data points." (Jain)



Typical: $\underbrace{y_i}_{\text{scalar}} \sim \mathcal{N}(\underbrace{\mathbf{x}_i}_{\text{vector}}\beta, \underbrace{\sigma^2}_{\text{scalar}})$

GPR: $\underbrace{\mathbf{y}}_{\text{vector}} \sim \mathcal{MVN}(\underbrace{\mathbf{X}}_{\text{matrix}}\beta, \sigma^2 \underbrace{\mathbf{\Omega}}_{\text{matrix}})$

Mean is linear ($\mathbf{X}\beta$)

Include unit indicators in $\mathbf{X}$

Posteriors on β standard inferences

Need to estimate variance-covariance $\mathbf{\Omega}$ from data

Data in mean and $\mathbf{\Omega}$ do not have to be the same

Variance-covariance estimated from data:

$$\mathbf{y} \sim \mathcal{MVN}(\tilde{\mathbf{X}}\beta, \sigma^2\mathbf{\Omega}),$$

$$\mathbf{\Omega}(\mathbf{x}_j^*, \mathbf{x}_i^* | \zeta) = \exp\left\{ -\sum_{p=1}^{m} \frac{|x_{pj}^* - x_{pi}^*|^2}{\zeta_p} \right\}.$$

(Carlson)

$$p(f_*|X, X_*, f) \sim N(\mu_*, \Sigma_*)$$ (Jain)

$$\begin{pmatrix} f \\ f_* \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu \\ \mu_* \end{pmatrix}, \begin{pmatrix} K & K_* \\ K_*^T & K_{**} \end{pmatrix}\right)$$ (Jain)

White noise kernel, which adds a white noise of variance to the covariance matrix:

$$k(x_i, x_j) = \sigma^2 \delta(i, j)$$ (Jain)

Squared exponential kernel:

$$k_{SE}(x_i, x_j) = \sigma^2 \exp\left[- \left(\frac{x_i - x_j}{l}\right)^2 \right]$$ (Jain)

More mathematical information about the gausian processes can be found at http://www.gaussianprocess.org/gpml/chapters/RW.pdf

Applying GPs to MSFT and IBM Stock Price Prediction

In this project, we downloaded all available historical price data for MSFT and IBM from Yahoo Finance to train on. The CSV files downloaded from Yahoo Finance has 6 columns: Date, Open, High, Low, Close, Adjusted Close. We will use Date and Adjusted Close columns. Adjusted Close is the adjusted close price for stock splits and dividends, so that there is no gap in the time series. It will be the target variable in our data.

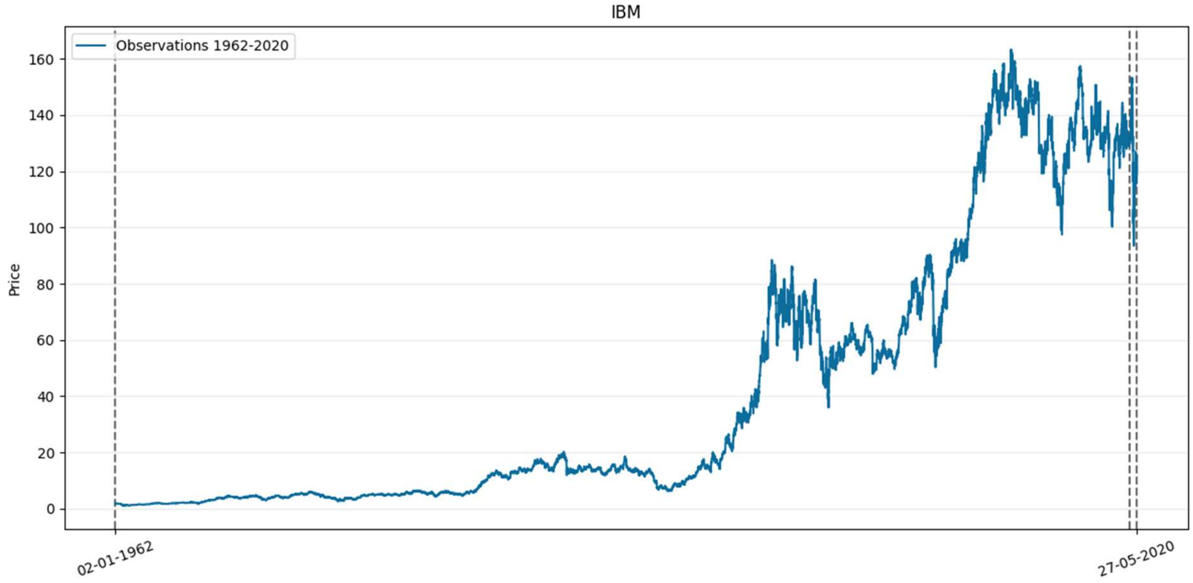
In other regression models, you usually model the entire time series as one, but for Gaussian Process Regression, we will divide the data into different time series: we will take each year separately, as a separate time series of stock data to make each time series an independent input variable for the regression model. So our model will predict future adjusted close prices given multiple yearly (252 trading days) time series as the input. Because Gaussian Process models are distributions over functions, our model will predict mean and uncertainty for each future date.

Because we assume the prior on the prediction distribution to be zero mean, we need to normalize our data. This will also help us to set kernel covariance matrix scale parameters. To get the posterior distribution, we also have to invert the covariance matrix.

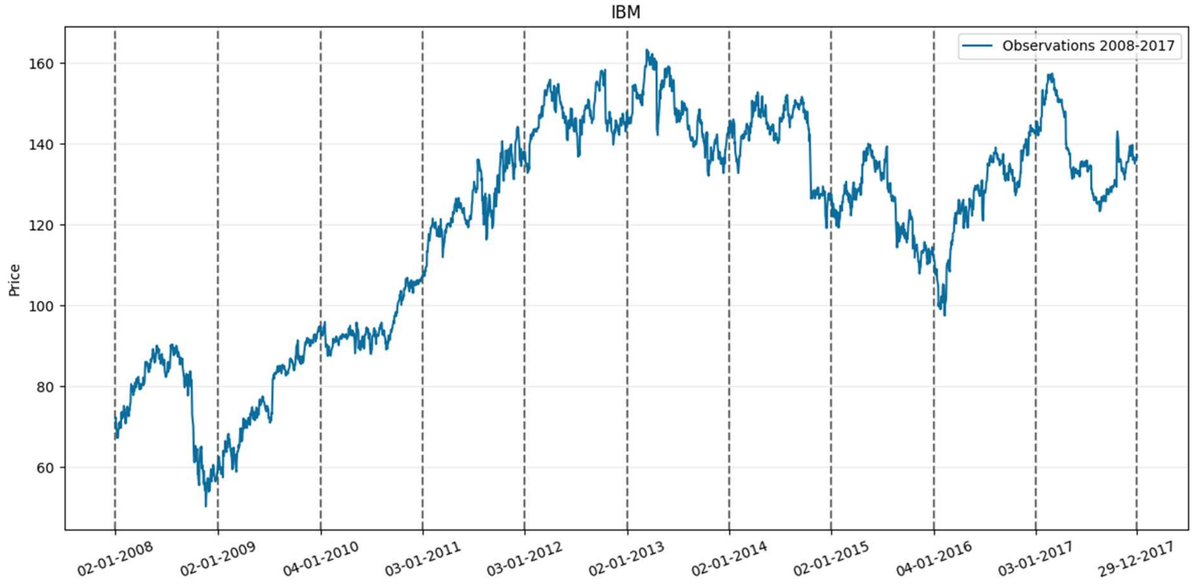
Our biggest challenge was to correctly arrange the CSV files and make the plotting. We tried following the steps provided in the book "Machine Learning Projects with Tensorflow" by Ankit Jain, Armando Fandango and Amita Kapoor, to implement the theory in code. Our independent variable X is the year and the day in that year, dependent variable Y is the adjusted close prices. We used squared exponential kernel, RBF with lengthscale = 1 and 63 variance, white noise 1e-10. Because RBF is infinitely differentiable and easy to understand, we chose it as our kernel.

Prediction Results

IBM adjusted close prices between 1962-2020:

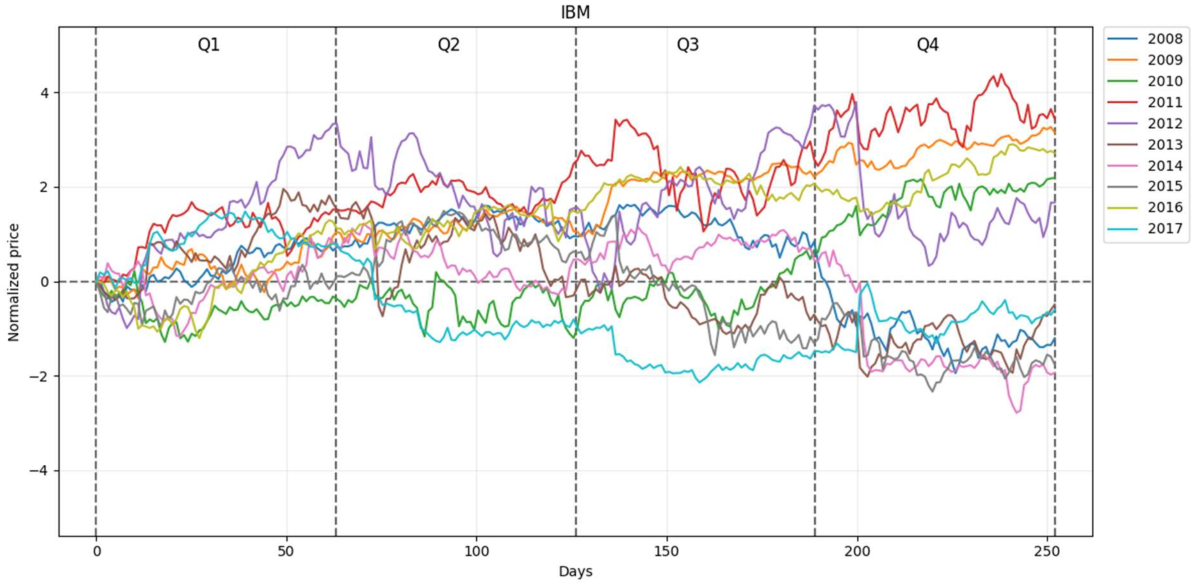


We used IBM historical data between 2008-2017 to train our model:

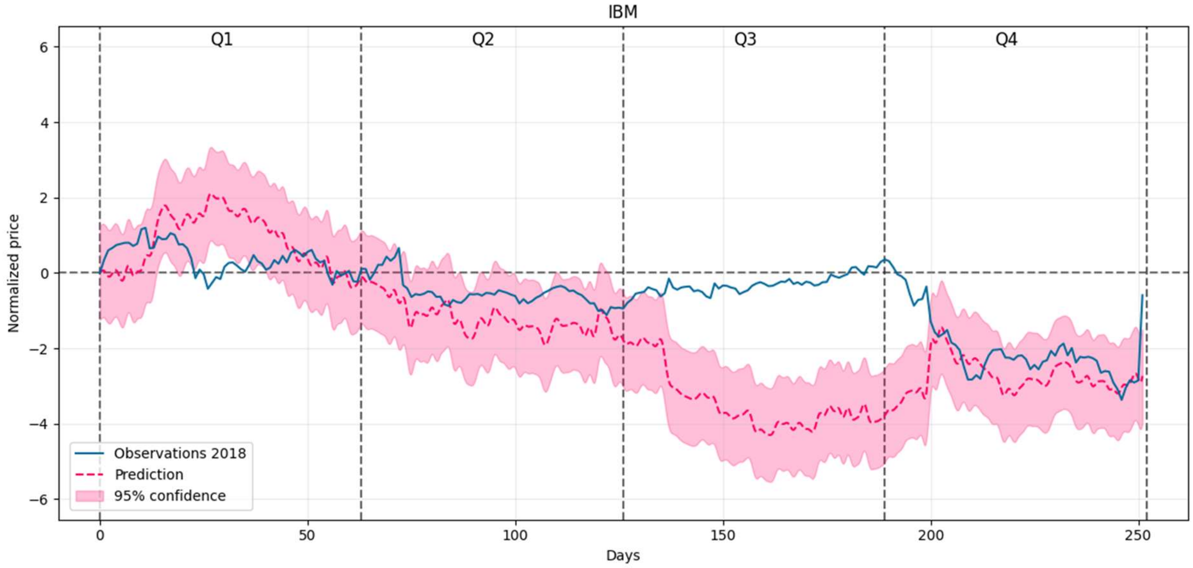


As seen on the plot, we separated each year as different time series.

Normalized adjusted close prices for these years:

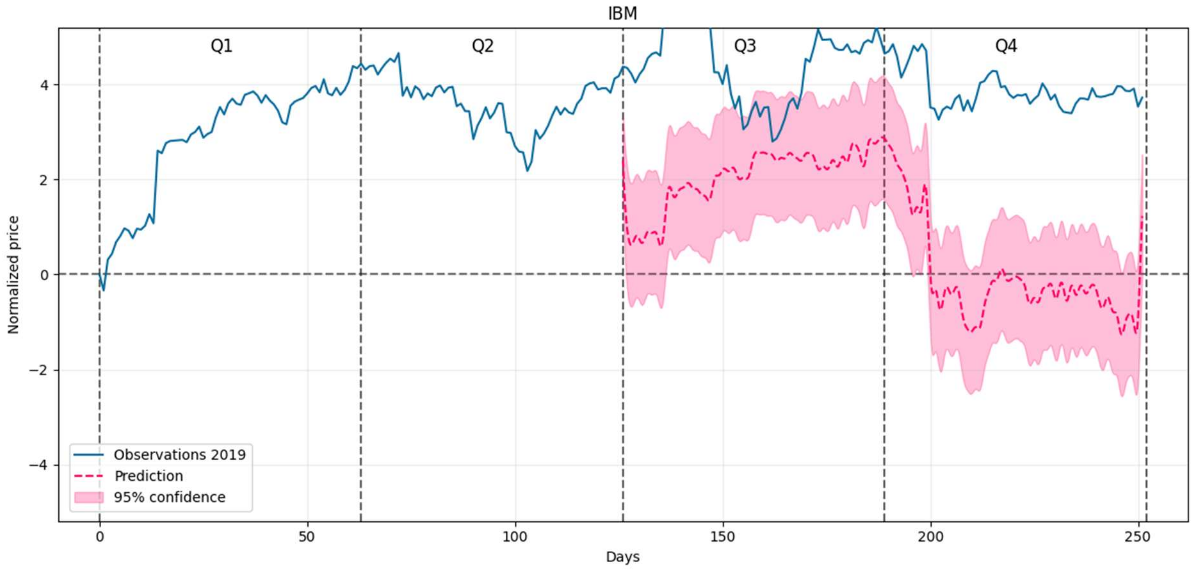


Prediction of 2018 with 95% confidence interval on plot:



The results we obtained amazed us. Even though our hypothesis was that the prediction’s 95% interval will include the actual prices, we were not expecting this much an accuracy. It seems like only the Q3 of 2018 is predicted falsely.

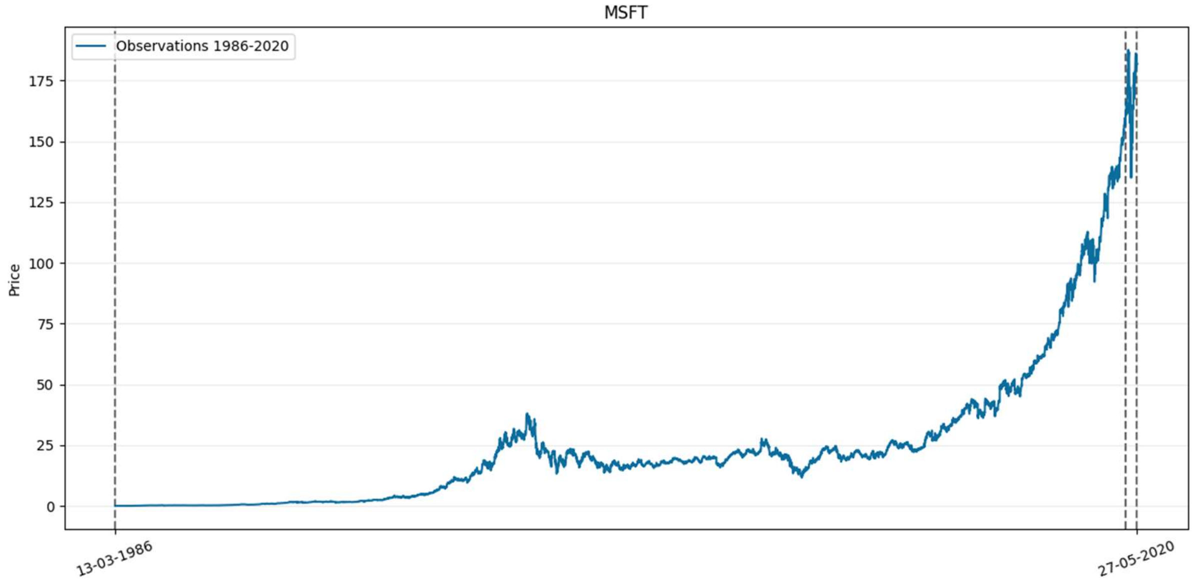
We then trained our model with data including 2019 Q1-Q2 to predict Q3 and Q4 of 2019:



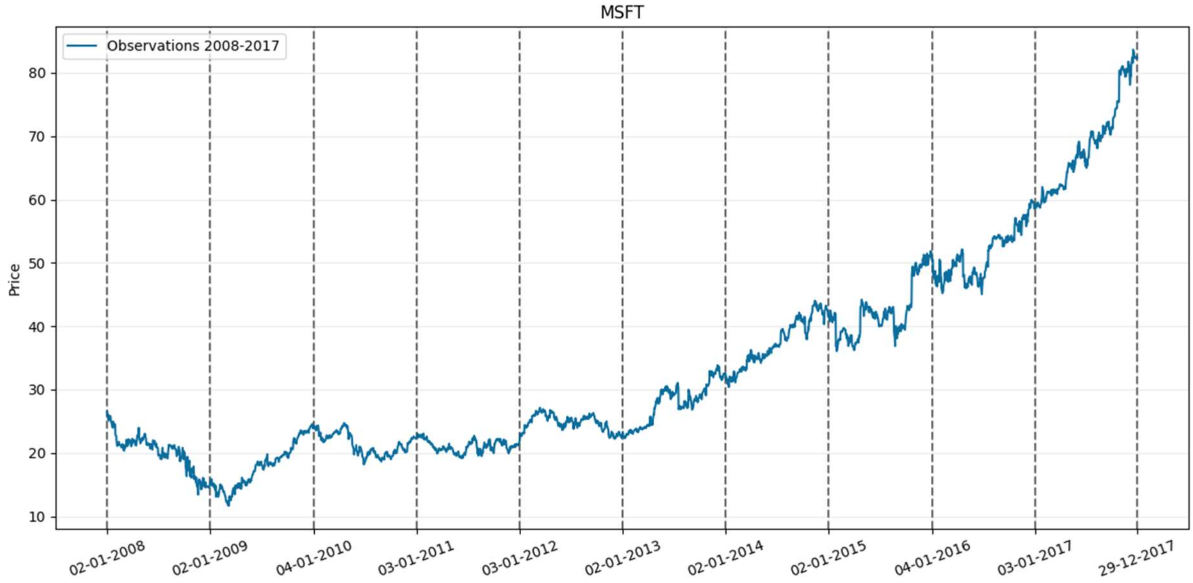
The model's prediction is way more inaccurate in predicting the other half of the year given the first half and past as input.

We then ran the same process with Microsoft Stock data.

MSFT Adjusted Close Prices between 1986-2020:

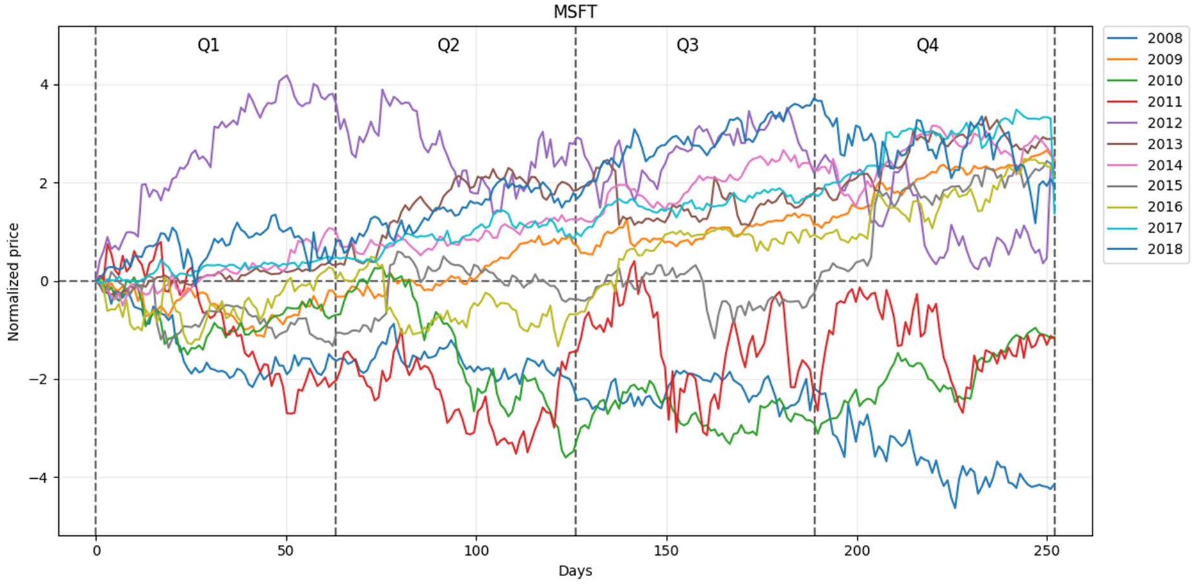


We used 2008-2017 MSFT stock data to train the model for phase 1:

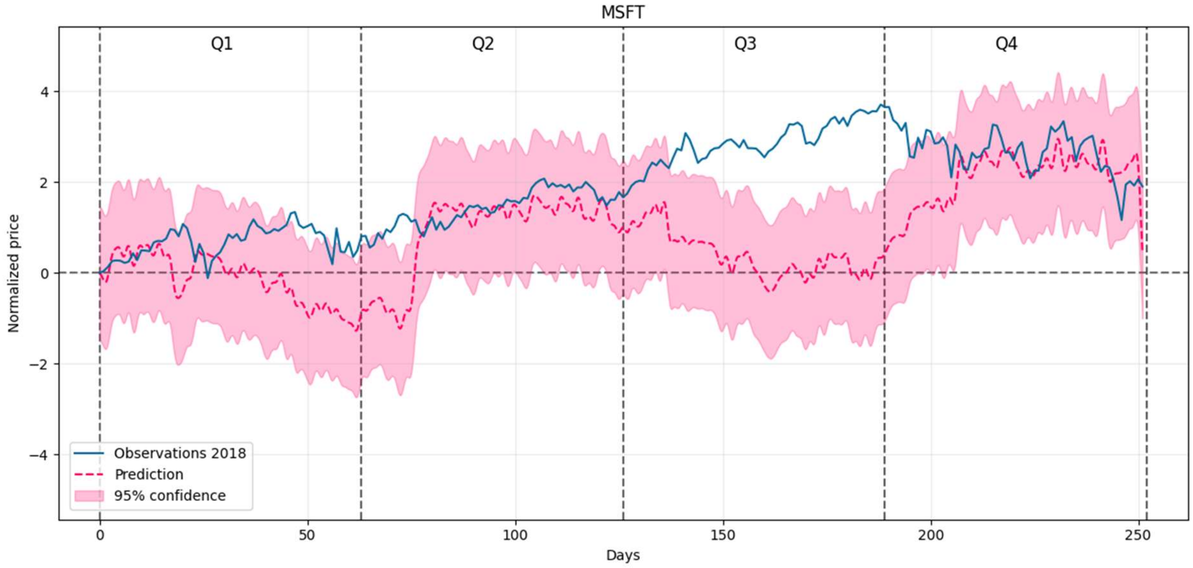


Again, we separated each year as different time series data.

Normalized adjusted close prices for 2008-2018 MSFT data:

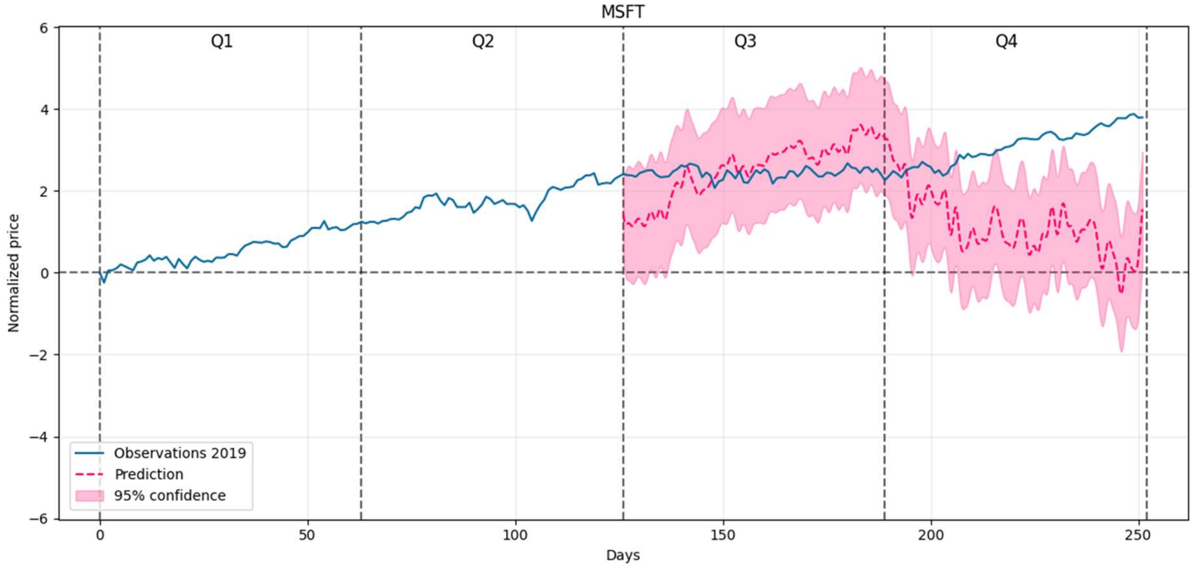


Prediction of 2018 with 95% confidence interval on plot:



The MSFT prediction results are as good as IBM results, it seems like in 2018 Q3, our model failed because of a common market impact, in other quarters, the prediction confidence interval includes almost every actual data point, which is fascinating.

MSFT 2019 Q3-Q4 Prediction:



Again, the model seems not good at predicting quarterly data. The yearly prediction performance is way better.

Discussion and Results

The Gaussian Process Regression model we used in this project seems surprisingly accurate in predicting yearly adjusted close prices of IBM and MSFT stocks. Our hypothesis was that all the actual prices would fall in the 95% interval of the GPR model's prediction. However they did not. We are seriously amazed by the accuracy in yearly prediction, especially the end of year prices.

As undergraduate students, we pushed our limits to learn and implement the Gaussian Process Regression to stock price prediction by scanning resources online, as well as books written on topic of machine learning. We are very grateful because our instructor, Mr. David Carlson has broadened our horizon on topic of advanced data analysis using machine learning, and we will keep learning on this topic and improving ourselves through rest of our lives.

Since trading is just a probability game as Mark Douglas defines, we think that our Bayesian approach which builds on unpredictability, just like financial markets in micro level, was a great choice to implement a prediction model.

References

Jain, Ankit, et al. *TensorFlow Machine Learning Projects: Build 13 Real-World Projects with Advanced Numerical Computations Using the Python Ecosystem*. Packt Publishing, 2018.

Carlson, David. *QMBU 450 – Koç University, Advanced Data Analysis Using Python Lecture Notes*. 2020.

Bailey, Katherine. Gaussian Processes for Dummies. https://katbailey.github.io/post/gaussian-processes-for-dummies/

Douglas, Mark. *Trading in the Zone: Master the Market with Confidence, Discipline and a Winning Attitude*. Prentice Hall, 2000.