The B(eat)T(he)W(allstreet) Project

Erdős Institite Data Science Boot Camp 2021

The BTW team

May 14, 2021

"I am quite comfortable holding Berkshire, I think our business is better than the average in the market"

— Charlie Munger, *vice chairman of Berkshire Hathaway, Inc.*

"I recommend S&P 500 index fund and that's for a long long time and I've never recommended Berkshire to anybody."

— Warren Buffett, *chairman and CEO of Berkshire Hathaway, Inc.*

Problem Description and Data Sets

We would like to make predictions on the market indexes like S&P 500, Nasdaq etc, which are important indicators of U.S. stock market.

For the dataset, we will use the historical data including daily/monthly high, low, open, close prices and volume to make predictions.

As a more advanced approach if time permits, we will try to investigate other data on macroeconomics like 10-year U.S. treasury bond yield, p/e ratio, etc. to see if they will provide us with a more accurate estimate for market indexes

Stakeholders of the Project

- Policy makers in Federal Reserve System
- Investors (including the investment bankers) of U.S. Stock market
- Financial analysts

Planned modeling approach 1: Random Forests

One model we are considering is Random Forests (which is also to be covered in our Boot Camp), which is a type of ensemble learning method constructed by multiple Decision Trees during training process and output the mode or mean of individual Decision Trees.

The procedure for constructing Decision Trees begins with the dataset which have multiple features. The next step is find the best features in the dataset and split the data into sub-data, which contains the possible values with the best features. Then we repeat the procedure above by recursively splitting the sub-data. Decision Trees algorithm has the drawback of being sensitive to noises, which would cause the over-fitting problem.

Random Forests model could potentially remove or ease the issue of overfitting.

Random Forests

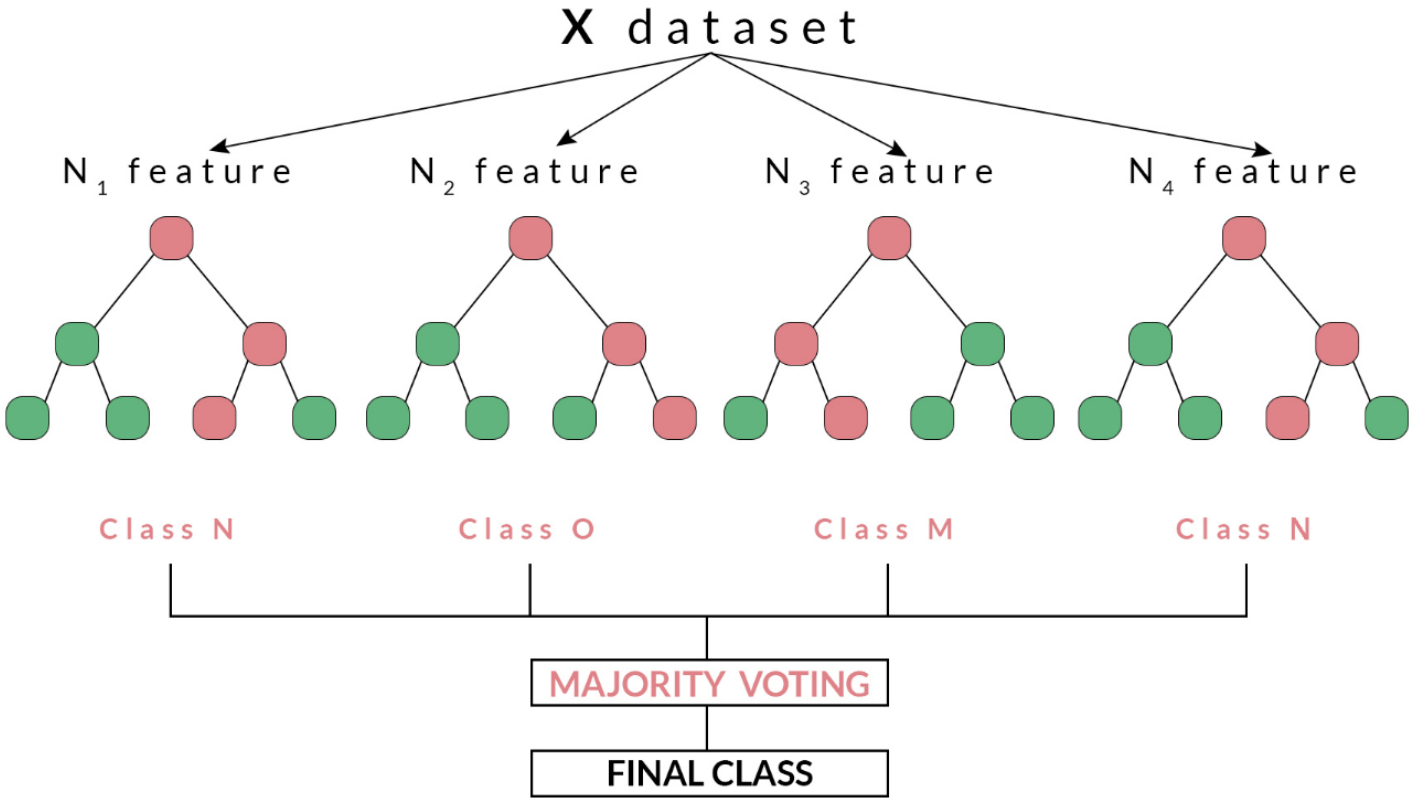

Figure:
https://blog.quantinsti.com/random-forest-algorithm-in-python/

Planned modeling approach 2: Facebook Prophet

Another model we are considering is Facebook's Prophet:

`https://facebook.github.io/prophet/`

Facebook's Prophet, based on Stan, 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. Prophet uses a linear model but it also takes into account the carrying capacity for growth, meaning that as the data approaches the top, its growth rate could be reduced significantly, which will normally happen in the real world.

Prophet is robust to outliers, missing data, and dramatic changes in the time series.

Facebook Prophet

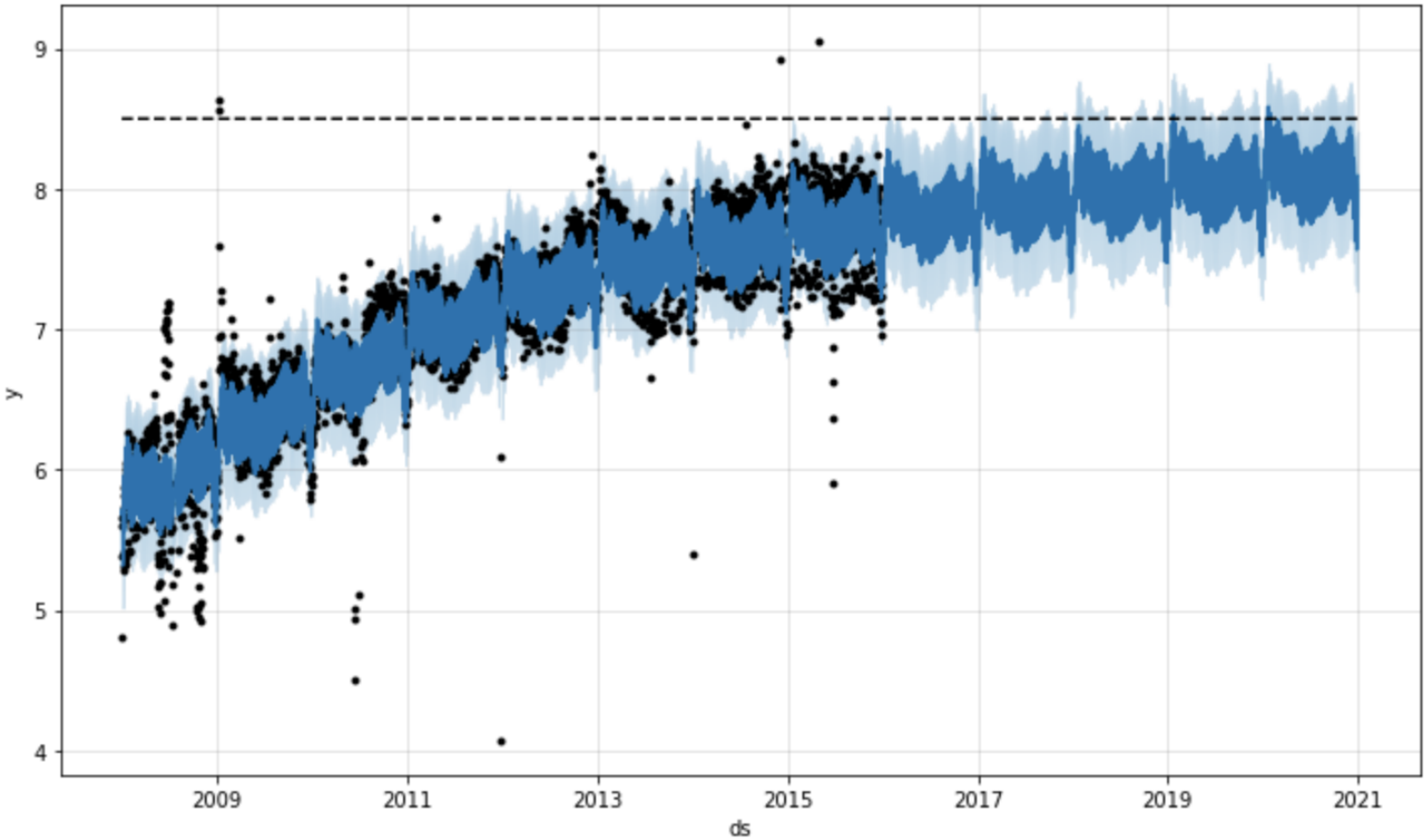

Figure: https://facebook.github.io/prophet/docs/saturating_forecasts.html

Thank you!