Stock Price Prediction with K-nearest Neighbors (KNN) and Random Forest

Wan Chi Fung,Victor

2020/9/13

1. Introduction

Predicting stock price is one of the popular fields where machine learning is applicable. Market trends, news, price action, and performance of the industry are some of the key factors affecting the stock price. This project will focus on the price action and aims to predict stock price on the next day whether it is up or down, utilizing KNN, random forest, and financial indicators to maximize the accuracy of the model.

2. Methods

2.1 Dataset Preperation

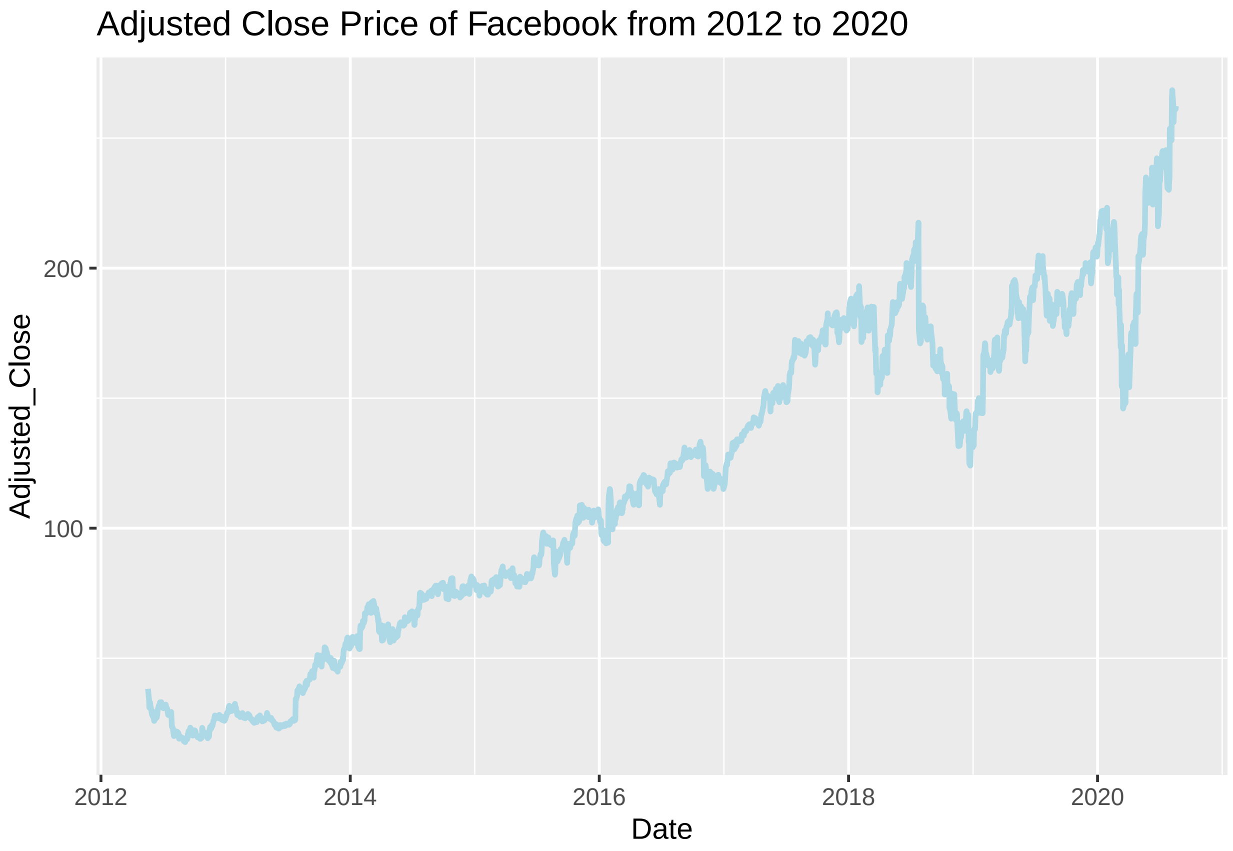
2.2 Data Wrangling

The dataset used in the Facebook stock price dataset was downloaded from Kaggle. The dataset consists of 7 columns of information on the Facebook stock from 2012 to 2020:

- Date : Date with stock price recorded
- Open : Opening price of the stock
- High : Maximum price of the stock for the day
- Low : Minimum price of the stock for the day
- Close : Closing price of stock for the day
- Adjusted Close : Amending a stock's closing price to reflect that stock's value after accounting for any corporate actions
- Volume : The physical number of shares traded of that stock on a particular day

Date	Open	High	Low	Close	Adjusted_Close	Volume
2012-05-18	42.05	45.00	38.00	38.23	38.23	573576400
2012-05-21	36.53	36.66	33.00	34.03	34.03	168192700
2012-05-22	32.61	33.59	30.94	31.00	31.00	101786600
2012-05-23	31.37	32.50	31.36	32.00	32.00	73600000
2012-05-24	32.95	33.21	31.77	33.03	33.03	50237200
2012-05-25	32.90	32.95	31.11	31.91	31.91	37149800

2.2.1 The facebook stock price



2.3 Adding Indicators

As momentum changes before the price changes, some price momentum indicators are added to the dataset and used as the model training feature to increase the accuracy of the model.

2.3.1 Change in Price and Signal Flag

Signal Flag indicates that the stock price on that day is increased, decreased, or unchanged. Change in price indicates how much do the prices are different between two Consecutive days. More precisely, the change in price is the difference between the adjusted closing price of the consecutive day. Signal Flag is the value that the model is going to predict. It is composed of three values. When the change in price is larger than 0, Signal Flag is 1. When the change in price is equal to 0, Signal Flag is 0. When the change in price is less 0, Signal Flag is -1.

$$Changeinprice = C_t - C_{t-1}$$

where C__t is the adjusted closing price at time t.

2.3.2 Relative Strength Index (RSI)

RSI is popular momentum indicator determines the stock is oversold or overbrought. RSI values are ranged from 0 to 100. Normally, RSI above 70 is overbrought and below 30 is oversold.

$$RSI = 100 - \frac{100}{1 + RS}$$

where RS is the relative strength factor and defined as the ratio between two smoothed moving average.

2.3.3 Moving Average Convergence Divergence (MACD)

MACD consist of two part, the MACD value and the signal value. When the MACD value is above the signal value, it indicates a buy signal. When the MACD value is below the signal value, it indicates a sell signal. The MACD and Signal line are defined as:

$$MACD = EMA_{12}(C) - EMA_{26}(C)$$

$$SignalLine = EMA_9(MACD)$$

Where C is the adjusted closing price and EMA_9 is the 9 day Exponential Moving Average

2.3.4 Stochastic Oscillator

Stochastic Oscillator measures the level of the closing price relative to the low-high range over a period of time. The fask K in stochastic oscillator will be used in this project.

$$K = 100 * \frac{C - L_{14}}{H_{14} - L_{14}}$$

where C is the adjusted closing price, L_{14} is the lowest Low over the past 14 days and H_{14} is the highest high over the past 14 days.

2.3.5 Williams %R

Williams %R ranges from -100 to 0. When the value is above -20, it indicates a sell signal. When its value is below -80, it indicates a buy signal.

$$R = \frac{H_{14} - C}{H_{14} - L_{14}} * (-100)$$

where C is the adjusted closing price, L_{14} is the lowest Low over the past 14 days and H_{14} is the highest high over the past 14 days.

2.3.6 Price Rate Of Change

It measures the most recent change in price with respect to the price in n days ago.

$$PROC_t = \frac{C_t - C_{t-n}}{C_{t-n}}$$

where $PROC_t$ is the Price Rate of Change at time t and C_t is the adjusted closing price at time t.

2.3.7 On Balance Volume

On balance volume (OBV) utilizes changes in volume to estimate changes in stock price. It cumulatively adds the volumes on days when the price group, and subtracts the volume on the days when price go down, compared to the prices of the previous day.

$$OBV(t) = OBV(t-1) + Vol(t) \qquad \text{if } C(t) > C(t-1)$$

$$OBV(t) = OBV(t-1) - Vol(t) \qquad \text{if } C(t) < C(t-1)$$

$$OBV(t) = OBV(t-1) \qquad \text{if } C(t) = C(t-1)$$

where $C(t)$ is the adjusted closing price at time t.

2.3.8 NAs

Some indicators are using past 14 days or 9 days data, for the first 14 dates of the indicator would become NA. Therefore, the first 14 dates of the data cannot be used.

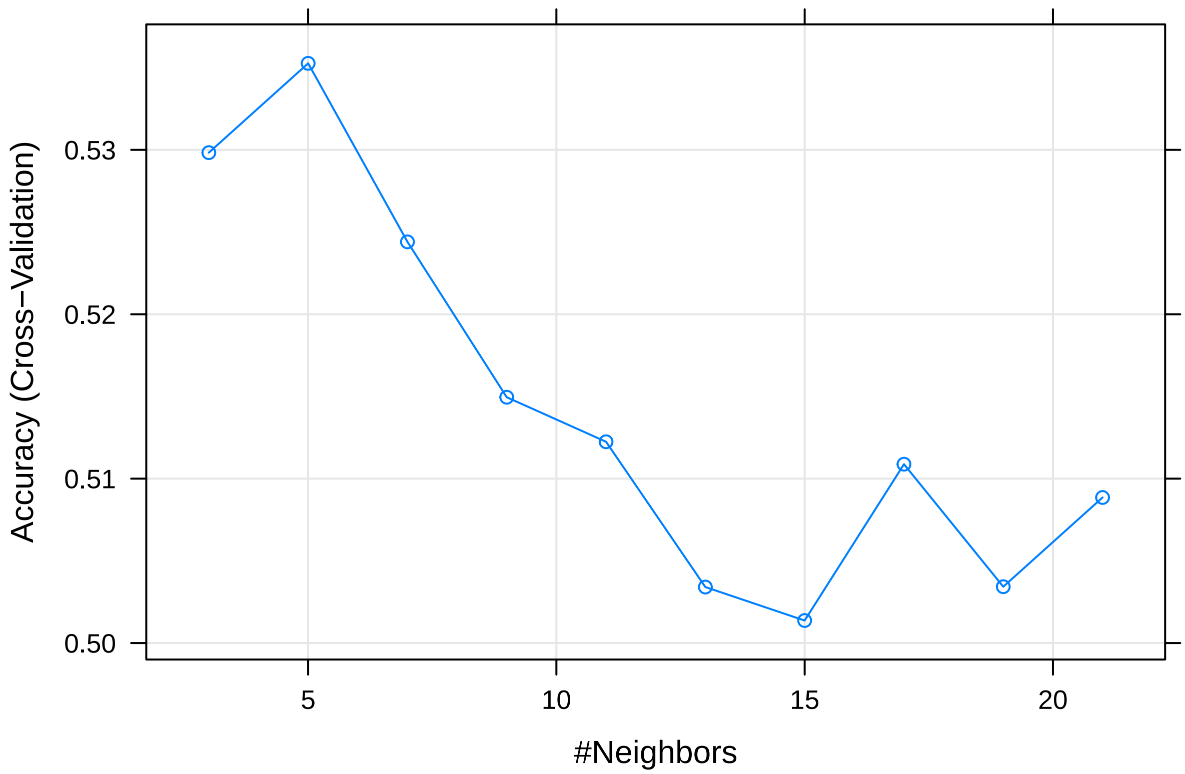
2.4 Create train set, test set and validation set

The train set, test set and validation are split from the facebook stock price dataset with a approximate ratio of 7:2:1.

2.5 Model Building

2.5.1 k-nearest neighbors (KNN)

To solve this classification problem , KNN method with cross validation is used. Apart from the indicators mentioned, the open of the stock is also considered to improve the accuracy of the model.



k = 5 will be used for training.

method	Accuracy
KNN	0.5229111

The Accuracy of the KNN model is 0.5229111, which is slightly better than guessing.

2.5.2 Random Forest

Another popular classification model is random forest. It is a supervised machine learning algorithim that uses multiple decision trees in aggregate to help make more stable and accurate predictions.

method	Accuracy
KNN	0.5229111
Random Forest	0.6873315

The accuracy in random forest is much higher than that in KNN. It is close to 0.7. Random forest seems to be a better method for stock price prediction.

3. Result

3.1 Model evaluation on validation set

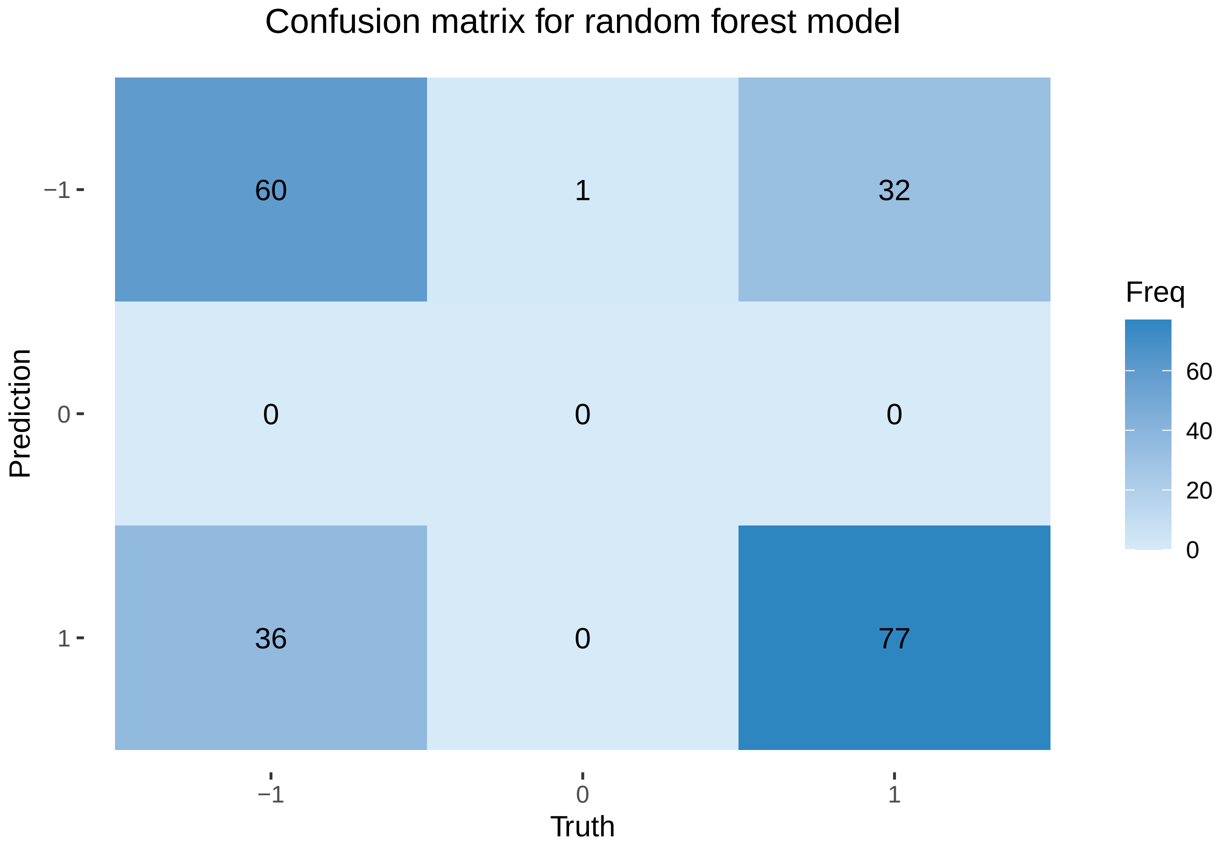
The random forest model is chosen for the final evaluation of our dataset.

method	Accuracy
Random Forest on validation set	0.6650485

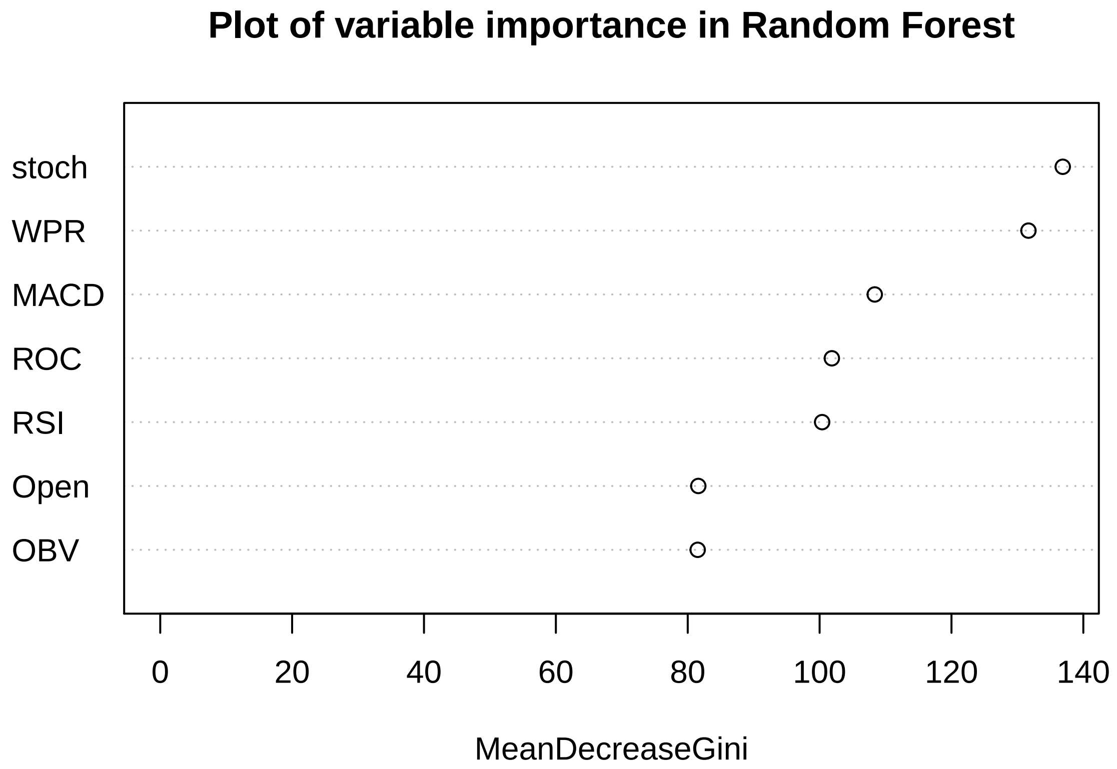
The Validation set accuracy is 0.6650485.

Coufusion Matrix

```
## Scale for 'fill' is already present. Adding another scale for 'fill', which
## will replace the existing scale.
```



The number of True positive (1) is larger than false positive and the number of True negative (-1) is larger than the false negative. A unchanged signal flag cannot be predicted correctlt and it is classfied into -1.



The Williams %R and Stochastic Oscillator had the largest mean decrease gini, which mean they can more significantly affect the stock price change.

4. Conclusion

To conclude, the model had a satisfactory result that could predict the daily uptrend and downtrend of the stock with an accuracy of 0.6650485.

4.1 Limitation

As there are numerous factors affecting stock price change in the market, such as company policy and news, the overall market trend, and the overall industry performance, the accuracy of the prediction is hard to improve without additional information on the factors mentioned.

4.2 Future work

A more precise result may be able to obtain through predicting the actual daily stock price instead of the daily uptrend and downtrend of the stock. Time series analysis, such as Long Short Term Memory and ARIMA could be applied to develop a more precise model on the stock price prediction.