

# Predicting Stock Change Basing on Candlestick Patterns

## 1 Problem Define

Stock Prices are considered to be very dynamic and susceptible to quick changes because of the underlying nature of the financial domain and in part because of the mix of known parameters, features (like daily highest price, closing price, P/E Ratio etc.) and unknown factors (like election results, rumors, news etc.).

Candlestick Patterns Analysis is one part of Technical Analysis, and has been applied to real stock trading market during past decades mostly in Japan, China, as well as United States. Candlestick patterns are indicative because the prices changes in several days partly show the public confidence in the stock so that they are able to tell the trend in the near future. Through practice, it has been proved that candlestick patterns can be generally to any stock market, regardless of time or company. Actually, some of the patterns are recognized through huge amount of historic data, which makes them more reliable.

### 1.1 What We have

We downloaded the historic daily stock price for 500 companies from 2003 to 2013, and stored them into MySQL database. The table looks like the following:

Field	Type	Null	Key	Default	Extra
symbol	varchar(10)	NO	PRI	NULL	
sector	varchar(100)	YES		NULL	
industry	varchar(100)	YES		NULL	
date	date	NO	PRI	NULL	
open	float(14,2)	YES		NULL	
high	float(14,2)	YES		NULL	
low	float(14,2)	YES		NULL	
close	float(14,2)	YES		NULL	
volumn	int(11)	YES		NULL	
adjClose	float(14,2)	YES		NULL	

- symbol: company's stock symbol in NASDAQ market;
- sector: describing the specific sector of an industry for the company;
- industry: describing the industry of the company;
- date: the stock exchange date;
- open: the open price of the day;

- high: the highest price of the day;
- low: the lowest price of the day;
- close: the close price of the day;
- volumn: the stock exchange volumn of the day;
- adjClose: the adjusted close price of the day (adjusted by Yahoo, never used).

## 1.2 What We Want To Predict

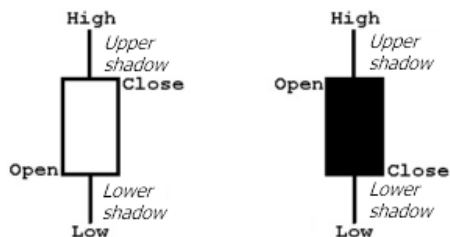
We want to predict the next day's stock trend according to its previous stock prices. And we use the trend to label our data to enable supervised learning. The trend(label) is defined as:

If (today's close price > previous 10 days' average close price):  
 Label = UP;  
 Else if (today's close price < previous 10 days' average close price):  
 Label = Down;  
 Else:  
 Label = Keep;

In our system, we want that when we input latest several weeks historic stock price of one company, we could be able to tell the next day's trend.



## 1.3 What Is A Candlestick

We know the daily open, close, high, low price of one company, so we are able to draw one candlestick for per company per day in the following way:



There are several candlestick types:

Candlestick	Type		Type
	Long Candlesticks		Spinning Tops
	Short Candlesticks		Hanging man; Hammer

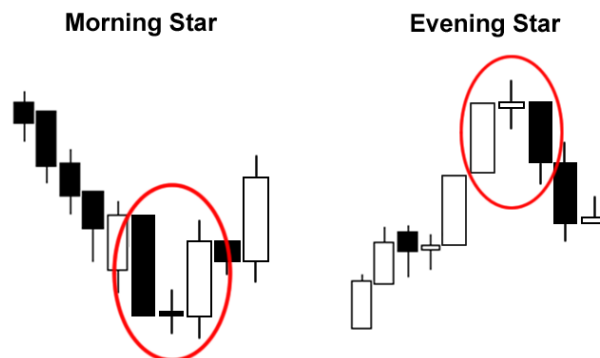
	Marubozu		Doji
---	----------	---	------

If we draw the candlestick for one company over a longer period of time, we can get a candlestick chart:



## 1.4 What is Candlestick Pattern

A candlestick pattern can be formed by one or more single candlestick. These candlesticks together tell the support of the market and potential trend of the stock. Like the following patterns:



Through decades of observation as well as the understanding of the patterns, people have summarized over a hundred patterns that mean something. In our system, we picked out 54 most indicative patterns(according to several candlestick analysis forums and websites).

## 1.5 Convert Price to Candlesticks

One candlestick is describes as the following structure:

```
public class SingleCandle {  
    public double open, high, low, close;  
    public int volumn;  
    public CandleType type;  
    public boolean bull;  
    public double bodysize;  
    public TrendType trend;  
}
```

Remember that to get the trend of the candlestick, we need previous 10 days' prices.

## 2 Data Preparation

### 2.1 Instance

As described above, we have the historic daily stock price for 500 companies from 2003 to 2013. We use six month's stock prices to form one instance for machine learning, and the last day's trend becomes the label of the instance. One instance is defined using the following structure:

```
public class Instance {  
    private ArrayList<SingleCandle> candles;  
    private ArrayList<SingleCandle> patternCandles;  
    private SingleCandle.TrendType label;  
    private String symbol;  
    private String time;  
    private PatternRecognizer pattern;  
}
```

- candles: contain all candles from the selecting 6 months;
- patternCandles: latest 11 candlesticks for extracting candlestick candles.
- label: the latest(last day of the six months) candle's trend;
- pattern: extracted features.

Each year, we select 6 months as (Jan...Jun), (Feb...Jul), (Mar...Aug)...(Jul...Dec). We


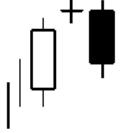









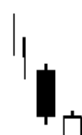

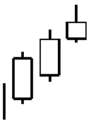



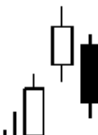
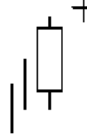




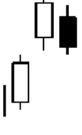
have all together about 30,000 instances. (500 companies X 11 years X 7 instances/year).





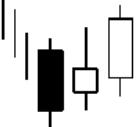

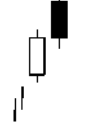

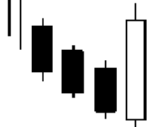



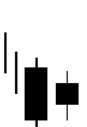


Among the 30,000 instances, there are 15145 labeled as “UP”, 13964 labeled as “DOWN”, and 92 labeled as “KEEP”.

## 2.2 Features:

We select 54 candlestick patterns, and see if they exist in the latest 11 but 10 candlesticks (the last one is used for labeling):

Pattern	Feature	Pattern	Feature	Pattern	Feature
	F1: Hammer (bullish)		F19: Abandoned Baby (bullish)		F37: Identical Three Crows (bearish)
	F2: Shooting Star (bearish)		F20: Abandoned Baby (bearish)		F38: Unique Three River Bottom (bullish)
	F3: Belt Hold (bullish)		F21: Morning Star (bullish)		F39: Concealing Baby Swallow (bullish)
	F4: Belt Hold (bearish)		F22: Evening Star (bearish)		F40: Breakaway (bullish)
	F5: Engulfing (bullish)		F23: Morning Doji Star (bullish)		F41: Breakaway (bearish)

	F6: Engulfing (bearish)		F24: Evening Doji Star (bearish)		F42: Kicking (bullish)
	F7: Harami Cross (bullish)		F25: Bearish Upside Gap Two Crows		F43: Kicking (bearish)
	F8: Harami Cross (bearish)		F26: Two Crows (bearish)		F44: Bearish On Neck Line
	F9: Harami (bullish)		F27: Three Star in the South (bullish)		F45: In Neck Line (bearish)
	F10: Harami (bearish)		F28: Deliberation (bearish)		F46: Thrusting Line (bearish)
	F11: Doji Star (bullish)		F29: Three White Soldiers (bullish)		F47: Upside Gap Three Methods (bullish)
	F12: Doji Star (bearish)		F30: Three Black Crows (bearish)		F48: Upside Gap Three Methods (bearish)
	F13: Piercing Pattern (bullish)		F31: Three Outside Up (bullish)		F49: Upside Tasuki Gap (bullish)

	F14: Dark Cloud Cover (bearish)		F32: Three Outside Down (bearish)		F50: Downside Tasuki Gap (bearish)
	F15: Meeting Lines (bullish)		F33: Three Inside Up (bullish)		F51: Three-Line Strike (bullish)
	F16: Meeting Lines (bearish)		F34: Three Inside Down (bearish)		F52: Three-Line Strike (bearish)
	F17: Matching Low (bullish)		F35: Three Stars (bullish)		F53: Inverted hammer (bullish)
	F18: Homing Pigeon (bullish)		F36: Three Stars (bearish)		F54: Hanging Man (bearish)

### 3 Baseline Experiment

We used 54 Boolean features to do the experiment, each Boolean value represent the existence the certain candlestick pattern.

Using JRip doing 10-fold cross-validation, the performance is 57.0943% for correctness, and Kappa statistic is 0.2.

The data is sparser than our expectation, and go through each feature, we find that some feature (f5, f6, f21, f22, f25, f26, f37, f38, f41, f43, f44, f45) never occurs. We also expect that for bullish features, there would be more “UP” labels than “DOWN” labels, and for bearish features, there would be more “DOWN” labels than “UP” labels, however, this is also not the case.

Rules:

(f12 <= 0) and (f11 >= 1) and (f35 <= 0) and (f33 <= 0) => label=DOWN (3447.0/1152.0)
---

```
(f12 <= 0) and (f7 >= 1) and (f33 <= 0) => label=DOWN (3402.0/1319.0)
(f36 >= 1) and (f7 >= 1) => label=DOWN (734.0/305.0)
(f34 >= 1) => label=DOWN (2199.0/1035.0)
(f8 <= 0) and (f12 <= 0) and (f19 >= 1) => label=DOWN (1030.0/425.0)
(f12 <= 0) and (f20 >= 1) and (f7 <= 0) => label=DOWN (481.0/173.0)
(f36 >= 1) and (f11 >= 1) and (f35 <= 0) and (f8 <= 0) => label=DOWN (420.0/150.0)
(f36 >= 1) and (f20 >= 1) => label=DOWN (134.0/54.0)
(f36 >= 1) and (f8 <= 0) and (f19 >= 1) and (f16 <= 0) => label=DOWN (129.0/51.0)
(f8 <= 0) and (f12 <= 0) and (f7 <= 0) and (f30 <= 0) and (f48 >= 1) => label=DOWN (165.0/71.0)
=> label=UP (17326.0/6915.0)
```

## 4 Improvements

### 4.1 Changing Boolean features to numeric features

We think that Boolean features might not best represent the occurrences, so we decide to change Boolean features to numeric features, representing how many times the patterns occur.

Using JRip doing 10-fold cross-validation, the performance is 60.2572% for correctness, and Kappa statistic is 0.2055.

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.643	0.439	0.607	0.643	0.625	0.61	UP
	0.564	0.356	0.597	0.564	0.58	0.612	DOWN
	0	0	0	0	0	0.627	KEEP
Weighted Avg.	0.603	0.397	0.6	0.603	0.601	0.611	

Rules:

```
(f12 <= 0) and (f11 >= 1) and (f35 <= 0) and (f33 <= 0) => label=DOWN (3447.0/1152.0)
(f12 <= 0) and (f7 >= 1) and (f33 <= 0) => label=DOWN (3402.0/1319.0)
(f36 >= 1) and (f12 <= 1) and (f8 <= 0) => label=DOWN (2754.0/1216.0)
(f12 <= 0) and (f34 >= 1) => label=DOWN (1485.0/672.0)
(f12 <= 0) and (f8 <= 0) and (f19 >= 1) and (f7 <= 0) => label=DOWN (905.0/358.0)
(f12 <= 0) and (f8 <= 0) and (f20 >= 1) and (f7 <= 0) => label=DOWN (445.0/156.0)
(f36 >= 1) and (f12 <= 1) and (f34 >= 1) => label=DOWN (165.0/63.0)
(f12 <= 0) and (f8 <= 0) and (f11 >= 1) and (f11 >= 2) => label=DOWN (338.0/157.0)
=> label=UP (16526.0/6473.0)
```

To solve the problem that some features do not occur, we need to loosen the condition for detecting certain pattern. We can use soft equals instead of strong equals while we determine some patterns in the code, because sometimes we



visually see that the candlesticks are have identical price, however, they might be different on the lower digits.

## 4.2 Loosening Conditions

We use soft equals instead of hard equals in the code, and hope that we can extract those patterns who never occur before. However, those features remain unseen. Meanwhile, the less strict conditions bring more other patterns' occurrence.

Using JRip doing 10-fold cross-validation, the performance is 60.9631% for correctness, and Kappa statistic is 0.22.

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.642	0.424	0.615	0.642	0.629	0.616	UP
	0.579	0.356	0.603	0.579	0.591	0.618	DOWN
	0	0	0	0	0	0.636	KEEP
Weighted Avg.	0.61	0.39	0.607	0.61	0.608	0.617	

Rules:

(f12 <= 0) and (f19 >= 1) and (f33 <= 0) => label=DOWN (4258.0/1562.0)

(f11 >= 1) and (f35 <= 0) and (f33 <= 0) => label=DOWN (3809.0/1408.0)

(f12 <= 0) and (f7 >= 1) and (f33 <= 0) => label=DOWN (2170.0/877.0)

(f36 >= 1) and (f12 <= 1) and (f7 >= 1) => label=DOWN (599.0/251.0)

(f12 <= 0) and (f34 >= 1) => label=DOWN (1319.0/621.0)

(f36 >= 1) and (f12 <= 1) and (f8 <= 0) and (f30 <= 0) and (f35 <= 0) and (f27 <= 0) and (f7 <= 0) and (f16 <= 0) => label=DOWN (1379.0/621.0)

(f12 <= 0) and (f8 <= 0) and (f11 >= 2) => label=DOWN (325.0/151.0)

=> label=UP (15608.0/5953.0)

We think this is the best candlestick patterns can do, and we need to find other features that can support candlestick patterns to make it more indicative, like volumn.

## 4.3 Adding features

We want to make use of other features in addition to candlestick patterns, such as company's industry and sector, sudden volumn change.

Using JRip doing 10-fold cross-validation, the performance is 60.5355% for correctness, and Kappa statistic is 0.2124.

### === Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.631	0.421	0.613	0.631	0.622	0.613	UP
	0.581	0.367	0.597	0.581	0.589	0.615	DOWN
	0.12	0	0.5	0.12	0.193	0.65	KEEP
Weighted Avg.	0.605	0.393	0.605	0.605	0.605	0.614	

### Rules:

(f55 <= 0) and (f57 = Coal Mining) and (f7 <= 0) => label=KEEP (19.0/8.0)
(f12 <= 0) and (f19 >= 1) and (f33 <= 0) => label=DOWN (4258.0/1562.0)
(f11 >= 1) and (f35 <= 0) and (f33 <= 0) => label=DOWN (3809.0/1408.0)
(f12 <= 0) and (f7 >= 1) and (f33 <= 0) and (f34 >= 1) => label=DOWN (160.0/39.0)
(f12 <= 0) and (f7 >= 1) and (f33 <= 0) => label=DOWN (2010.0/838.0)
(f36 >= 1) and (f12 <= 1) => label=DOWN (2834.0/1331.0)
(f12 <= 0) and (f34 >= 1) and (f56 = Finance) => label=DOWN (252.0/101.0)
(f12 <= 0) and (f34 >= 1) and (f7 >= 1) => label=DOWN (127.0/50.0)
(f12 <= 0) and (f34 >= 1) and (f30 <= 0) and (f8 <= 1) => label=DOWN (716.0/340.0)
=> label=UP (15282.0/5812.0)

Sector and industry do not help a lot in the final rules. This makes sense, because candlestick analysis theory itself claims that it does not depend on company.

We also found that the volumn feature we add do not help, but hurt the performance. This might resulted from the wrong way we extract it. We currently count the times of its sudden change, however, according to the technical analysis, sudden volumn change is indicative when it sudden changes to a large volumn and meanwhile certain candlestick pattern occurs. Thus where it occur is more important than how many times it occur.

## 4.4 Continuing adding features

We want to use sudden volumn change position during this experiment. We also want to add RSI (relative strength index), a very helpful technical indicator that people use with candlestick patterns. This means we can make use of other candles in the instance.

RSI - A technical momentum indicator that compares the magnitude of recent gains to recent losses in an attempt to determine overbought and oversold conditions of an asset. It is calculated using the following formula:

$$RSI = 100 - 100 / (1 + RS^*)$$

\*RS = Average of x days' up closes / Average of x days' down closes.