Systematic Investing Final Project
Stock Price Prediction Using the Naïve Bayes: A Case Study of Oil Industry in China

Ruijie (Cherry) Cai, Zhiyu Chen

May 17, 2019

This project examines the validity and efficiency of using Naïve Bayes algorithm to predict the Chinese stock market. In particular, we posit the oil industry as the focus of our analysis and set out to explore trading strategies on historical data. The result suggests high potentials of applying Naïve Bayes algorithm to forecast stock movements in China when long-period training data is available.

1 Introduction

The use of machine learning algorithms has been accentuated over recent years, driving various innovations in the financial industry. Naïve Bayes, with its simple but powerful algorithm, has showcased strong practical significance to stock price prediction in the previous study [1]. Interested in applying machine learning on making trading decisions, we implement Naïve Bayes algorithm by using the scikit-learn library in python and processing raw stock data into appropriate feature vectors. Naïve Bayesian Bernoulli classifier is deployed in the project.

Here, we choose the oil industry as the unit of our analysis to undertake corresponding modeling and analysis work. The oil industry draws our attention for its importance as the essential energy source for the world's industrial development. China, being the fourth-greatest oil producer in the world, controls the monopolized oil industry with three major state-owned companies—PetroChina, Sinopec, and China National Offshore Oil Corporation (CNOOC). Given the nature of the oil industry, the global macroeconomic environment has a great impact on stock performance within this sector and prices move rapidly in responding to external events. Despite the weak performance of crude oil through recent years, the embedded systematic behavior across oil stocks gives rise to algorithm-driven trading strategies in this space.

In this project, we first run Naïve Bayes algorithm on a generalized case, using historical data of listed A-share stocks in the oil industry from April 19, 2013, to April 19, 2019. Next, we delve into investigating a special pair: PetroChina and Sinopec. Historical trading data of these two stocks covers the period from March 22, 2010, to April 19, 2019. Using the performance of the traditional pair trading strategy, as the baseline, we are particularly interested in exploiting the co-integrated relationship within the two and draw on a Naïve Bayes predictive model to make profitable bets on stock price movements.

2 Background

2.1 Pair Trading Strategy

Our baseline model is the traditional pair trading strategy. The traditional pairs trading strategy is a ”market-neutral” strategy that matches a long position with a short position in a pair of highly correlated instruments such as two stocks, exchange-traded funds (ETFs), currencies, commodities or options” [2]. Profit of this strategy originates from betting on the mean-reversion behavior of price difference between the two chosen instruments. It has its comparative advantage since pair traders are able to benefit from a variety of market conditions if proper entry thresholds are set.

In this project, we employ the Dickey-Fuller test on our pair companies: PetroChina and Sinopec. The t-stats from the hypothesis test suggests a promising co-integrated relationship between stocks of these two leading Chinese oil companies. However, the result of the pair trading strategy is not satisfying. Since the short selling in the Chinese market (A-share) is limited, the gain from the short-selling side is not available.

2.2 Naïve Bayes

Naïve Bayes is a method, describing a set of supervised learning algorithms based on applying Bayes' theorem with the "Naïve" assumption of conditional independence between every pair of features given the value of the class variable.

We use the Bayesian method to infer the model parameters and learn from the available data. As a general Bayesian theorem, we state the following mathematical equation:

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)} \tag{1}$$

where
• A and B are events and P(B)≠0
• P(A| B) is the likelihood of event A occurring given that B is true, and conversely for P(B | A)
• P(A) and P(B) are the probabilities of observing A and B independently of each other [3].

2.2.1 Naïve Bayes Bernoulli Classifier

Since Naïve Bayes has been efficiency in inductive learning algorithms for machine learning and data mining, we take advantage of its competitive performance in classification [4].

Naïve Bayes Bernoulli Classifier implements the Bayes training and classification algorithms for data that is distributed according to multivariate Bernoulli distributions. Multiple features, considered as the input, will be a set of binary-valued variable, the Bernoulli/Boolean variable [5].

If x_i is a Boolean expressing the occurrence or absence of the i'th term from the vocabulary, then the likelihood of a document given a class y is given by [6]

$$P(x_i \mid y) = P(i \mid y)x_i + (1 + P(i \mid y))(1 - x_i) \tag{2}$$

3 Algorithm

3.1 Data Cleaning

To initiate this project, we acquired the data from “Tushare”, a Chinese financial data platforms for stocks, funds, futures, bonds, foreign exchange, industry big data, and block-chain data. The historical daily stock price list provides access to all the data we need. The key variable we selected for the model is the Historical Daily (i)Close Price, and (ii) Trading Volume. Close price measures the final price at which a stock is traded on a given trading day which represents the most up-to-date valuation of a security until trading commences again on the next trading day. As for the Trading Volume, under the context of a single stock trading on a stock exchange, it is commonly reported as the number of shares that changed hands during a given day. We cleaned the data using Python and then downloaded as .xlsx extension files for further analysis. Detailed procedure is documented in attached Data_Cleaning.ipynb file.

3.2 Naïve Bayes Modeling

3.2.1 Setting Features and Labels

Proper choice of features is essential to build an efficient Naïve Bayes predictive model. Due to the limited historical stock data, we endeavor to extract reasonable features from daily close prices: (i) 20-day rolling volatility and (ii) 5-day market momentum (iii) daily trading volume of each stock. All features are stored in data frames indexed by trading date in chronicle order. For days with no trading volume, corresponding feature values on those days are set to NAN because it does not make sense to include them in our training set if trades were suspended. For each feature, new value-labeled data frames are generated and different labels are assigned on the basis that all values are divided into five groups of equivalent data points. Regarding the two models of our interest, detailed treatments are slightly different given different settings, which will be explained in the ensuing part.

3.2.2 Stocks in oil industry

Our first layer of testing Naïve Bayes in Chinese stock prediction focused on a generalized case. After excluding stocks with several suspended trading days, we arrive at a modified list of 25 active stocks on Chinese exchange over the period from April 19, 2013, to April 19, 2019. To begin with the implementation of Naïve Bayes algorithm, we set up a trade calendar consisting of all covered trade dates to control the loop in the back test. During each iteration i, labeled feature information is reorganized into dummy variables to be read in by Bernoulli classifier. We initiated with training the classifier by the features of Day i and labeled return of Day i+1, and then using the features of Day i+1 to predict the label of the return of Day i+2. The trading basis is intuitively that buy stocks with predicted next-day return greater than 0.

3.2.3 PetroChina and Sinopec

We take a different approach of using Naïve Bayes to predict the pair (PetroChina and Sinopec). Bounded by the bounded amount of data for the two stocks, we train the classifier with features of Day i and labeled returns of Day i+1 from the first company, and then using the classifier to predict another company's Day i+1 return based on features of Day i. Specifically, to demonstrate this idea, we apply the information of Sinopec to predict PetroChina's stock. The implementation of the pair model is easier than the generalized case since the timing of the training set matches the predicted set. As a matter of fact that the data sets are prepared to reflect the corresponding relation between Day i features and Day i+1 labeled return, the classifier can fit and predict all at once.

All source codes for the above modeling are documented and appropriately commented in attached .ipynb files.

4 Simulation Results and Conclusion

In the generalized case, Naïve Bayes strategy produces a humble result with a Sharpe ratio of 0.3883 and a win/loss ratio of 0.5358. Judging from the equity graph(on next page), we can see that the strategy plays out well only in a particular period.

However, in the pair model for PetroChina and Sinopec, different tested periods yield divergent Sharpe ratios.

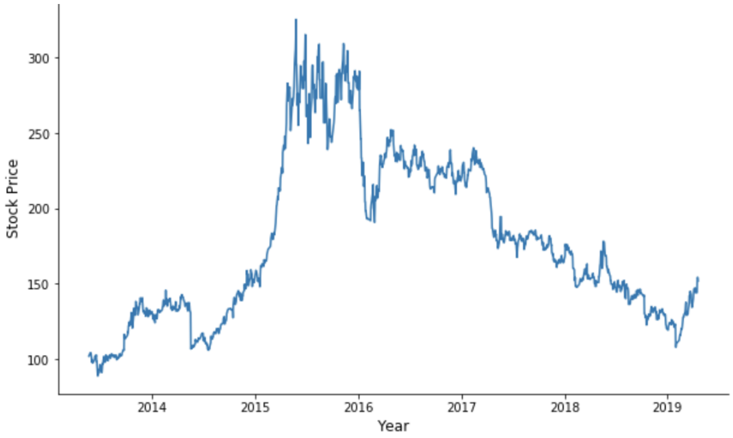

Figure 1: Industry Model Equity Graph

Below exhibits the visual presentation of sensitivity analysis for the pair model. Apart from the baseline pair trading strategy, three variations of Naïve Bayes model is examined.

As shown in the left plot, three blue lines are representative of three different Sharpe ratios of (A) pair trading strategies over nine different tested periods when using different spreads as the trading basis (zdiff5, zdiff10, zdiff20). The red line shows the Sharpe ratios of (B) a Naïve Bayes strategy when we match the tested period of the training set with that of the predicted set. In general, Naïve Bayes fails to beat pair trading except for the last period of the year 2018-2019.

On the right, blue line displays the Sharpe ratios of the (C) Naïve Bayes strategy that uses training data of the year 2010-2019 to predict next-day returns over different tested periods, while the red line unveils some promising Sharpe ratios from the (D) strategy that predict returns of the year 2018-2019 relying on training data over different tested periods.

By visualizing series of Sharpe ratios from different strategies, we can easily tell that (D) strategy produces the most lucrative results across all tested periods. While the outcome casts a shadow on applying Naïve Bayes algorithm to predict stock movements over a long period of time, it confirms the

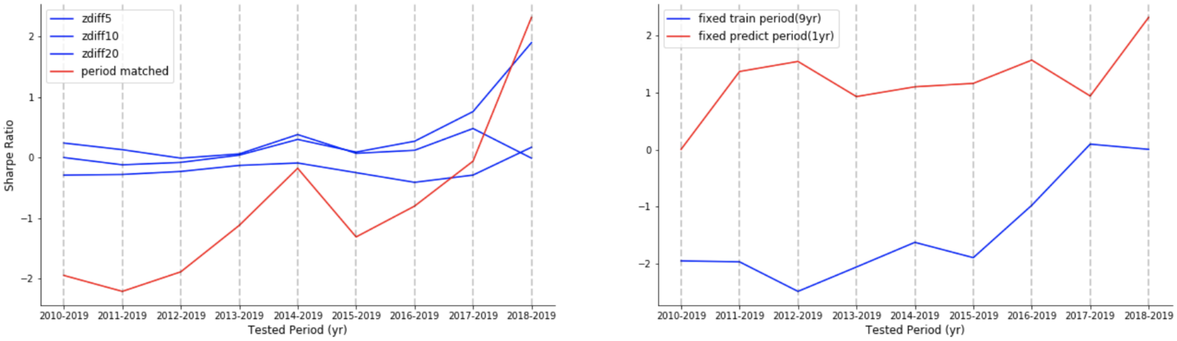

Figure 2: Pair Trading v.s. Naïve Bayes

legitimacy of implementing Naïve Bayes in stock prediction when the training set greatly outsizes the portion to be predicted. Moreover, the result indicates that 20-day rolling volatility, 5-day market momentum, and daily trading volumes are intuitive and efficient features for stock prediction under Naïve Bayes framework. Such observation also draws on a conclusion that some feasible and implementable real-world trading strategies should have generated the above features. Our study provides a solid basis for future trading strategy design.

In summary, this empirical study reveals the following aspects regarding Naïve Bayes prediction in Chinese stock market:

1) Naïve Bayes has its strength in modeling predictive decision-making system, by utilizing as much historical information as possible.
2) Naïve Bayes showcased its comparative advantage over the baseline pair trading strategy by leveraging the embedded similar behaviors in price movement across stocks.
3) Naïve Bayes can be generalized into many different cases for practical use, suggesting high potentials in stock prediction field. In this project, we illustrate two ways of modeling by testing Naïve Bayes on oil industry stocks or on a specific pair. The model can be further extended to test stocks from different industries to exploit the efficiency of Naïve Bayes algorithm.

References

[1] K. S. Kannan, P. S. Sekar, M. Sathik, and A. Assaf, "Stock market prediction and analysis using naïve bayes," 2016.

[2] T. Kanamura and et al, "The application of pairs trading to energy futures markets," 2008.

[3] B. Thomas and P. Richard, "An essay towards solving a problem in the doctrine of chance. by the late rev. mr. bayes, communicated by mr. price, in a letter to john canton, a. m. f. r. s.," pp. 370–418, 1763.

[4] Z. Harry, "The optimality of naïve bayes," 2004.

[5] V. Metsis and et al., "Spam filtering with naïve bayes – which naive bayes?," 2006.

[6] M. Andrew and N. Kamal, "A comparison of event models for naive bayes text classification," 1998.