ORIE 5255 Project Report
Predicting the Direction of Baijiu Stock Price

Rui Dai (rd576), Ying Guan (yg532), Yang Shang (ys2222)

1 Introduction

Stock market prices influenced the decision of investors and the overall market. Investors always want to have the correct prediction towards the price of the stock market. Due to the uncertainties and variables that influence the market value, predicting the specific stock price after a range of time might be too challenging. However, we try to predict the trends of stock market prices using the Machine Learning method. The stock market prices have chaotic and volatile nature and therefore come with high risk. In reality, traders love to own the stock whose values are expected to rise and refrain from buying the stock whose values are expected to fall in the future. The goal for our project is to maximize capital gain and to minimize loss involving investment in the stock market.

We would like to predict the up and down direction according to the historical data. If the stock price rises, the label will be +1. Otherwise, the label will be -1 with decreasing price. Predicting the direction of future stock market prices should be treated as a classification problem instead of a regression problem. We use a class of powerful machine learning algorithms known as ensemble learning and several technical indicators to build a model to forecast the future trends of the stock market prices. Our model tends to support or change decision making of some investors and let them maximize their profits.

2 Literature Review

The paper "Predicting the direction of stock market prices using random forest[1]" used random forest to forecast the trend of stock for Apple, Samsung, and GE. It used approximately 7,000 daily data and forecasted the outcome for 30, 60 and 90 days ahead.

For data processing, it applied the exponential smoothing method to add a higher weight to recent data and decrease the impact of dated data. It used six technical indicators that are widely used in the financial industry to indicate whether the market is performing bearish or bullish, including Relative Strength Index, Stochastic Oscillator, Williams %R, Moving Average Convergence Divergence, Price Rate of Change and On Balance Volume.

When fitting the data, it shows that the model actually performs best when predicting 90 days ahead. The accuracy decreases as time length decreases. In addition, the accuracy decreases as

[1] https://arxiv.org/pdf/1605.00003.pdf

the number of trees increases and stabilizes at around 45 trees. As a result, the model was able to predict all three stocks with prediction accuracy above 85%.

3 Data

3.1 Stock Prices and Volumes

Chinese Baijiu has attracted a lot of attention among investors in recent years, especially for Baijiu stocks. The Baijiu industry stock index rose 92% in 2019 and 120% in 2020, and investors are interested in whether this trend will continue. However, little research has been done trying to do forecasts or develop trading ideas in this area. Inspired by the research in US stocks mentioned in the literature review session, our team would like to focus on two top companies in the Baijiu industry and try to predict the direction of their stock price movement. The two stocks are Kweichow Moutai and Wuliangye, traded in Shanghai Stock Exchange and Shenzhen Stock Exchange, respectively.

Stock data will be pulled from Yahoo Finance, which has six variables: open, high, low, close, adjusted close and volume. To forecast the future, recent data matters more than data long ago, so exponential smoothing is applied on the time series which applies more weights to recent data and exponentially decreasing weights to past observations. Denote Y as the original time series and S as the adjusted time series, we have

$S_0 = Y_0\,;\ for\ t > 0,\ S_t = \alpha Y_t + (1 - \alpha) S_{t-1}$

where α is the optimized weight chosen by Python statsmodels package.

After exponential smoothing, we calculate technical indicators from the data as features to be fed into the model, since the technical indicators are widely used parameters among investors to predict the direction of stock price movement.

3.2 Feature Extraction - Technical Indicators

Based on previous research mentioned in the literature review session, our team decide to use the following 6 technical indicators which are widely used by investors to make buy/sell decisions: Relative Strength Index, Stochastic Oscillator, Williams %R, Moving Average Convergence Divergence, Price Rate of Change and On Balance Volume.

3.2.1 Relative Strength Index (RSI)

RSI indicates whether the stock is overbought or oversold. The formula is

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

$$RS = \frac{Average\ Gain\ over\ the\ past\ 14\ days}{Average\ Loss\ over\ the\ past\ 14\ days}$$

3.2.2 Stochastic Oscillator (%K)

%K keeps track of the stock price momentum which changes before the price changes. The formula is

$$\%K = 100 * \frac{Current\ Price - Lowest\ Low\ over\ the\ past\ 14\ days}{Highest\ High\ over\ the\ past\ 14\ days - Lowest\ Low\ over\ the\ past\ 14\ days}$$

3.2.3 Williams %R

Williams %R is a buy/sell signal which follows the momentum. The formula is

$$\%R = (-100) * \frac{Highest\ High\ over\ the\ past\ 14\ days - Current\ Price}{Highest\ High\ over\ the\ past\ 14\ days - Lowest\ Low\ over\ the\ past\ 14\ days}$$

3.2.4 Moving Average Convergence Divergence (MACD)

If MACD is below the Signal Line, it indicates a sell signal.
If MACD is above the Signal Line, it indicates a buy signal.
The formula is

$$MACD = EMA_{12}(Close) - EMA_{26}(Close)$$
$$Signal\ Line = EMA_{9}(MACD)$$

where $EMA_n(time\ series)$ denotes the n-day exponential moving average on the time series.

3.2.5 Price Rate of Change (PROC)

PROC is the most recent change in price with respect to the price in n days ago. The formula is

$$PROC_t = \frac{Close_t - Close_{t-n}}{Close_{t-n}}$$

We picked n to be 14 in this project.

3.2.6 On Balance Volume (OBV)

OBV reflects the buying and selling trends of a stock. The formula is

$$OBV_t = \begin{cases} OBV_{t-1} + Volume_t, & \text{if } Close_t > Close_{t-1} \\ OBV_{t-1} - Volume_t, & \text{if } Close_t < Close_{t-1} \\ OBV_{t-1}, & \text{if } Close_t = Close_{t-1} \end{cases}$$

3.3 Label

For the label, our team focuses on predicting the direction of stock price movement after 30 days, and the label reflects the direction of the stock price movements using -1, 0 and 1. That is,

$$Label_t = Sign(Close_{i+30} - Close_i)$$

4 Model

4.1 Logistic Regression

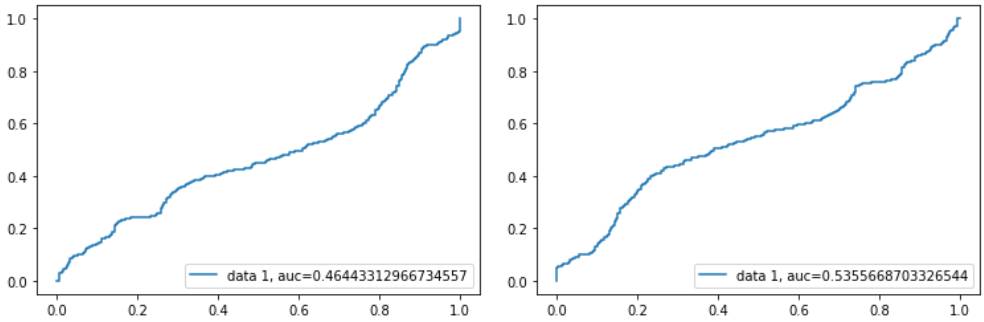
We first tried logistic regression, the simplest linear classifier. As an extension of linear regression, the model can be represented by the following formula

$$ln\frac{P(Y=1)}{1-P(Y=1)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \epsilon$$

The model is implemented with Python sklearn, and the results are as follows

Stock	Accuracy	AUC
Kweichow Moutai	68.59%	0.4644
Wuliangye	34.48%	0.5356

Figure 4.1.1 ROC Curves for Moutai (left) and Wuliangye (right)



After a first trial, the results are not satisfactory, so we decided to switch to other models instead of digging further into logistic regression.

4.2 Support Vector Machines(SVM)

4.2.1 Introduction to SVM

Support Vector Machines algorithm is a supervised machine learning model that works well for classification problems. It classifies y variables using the following optimization problem:

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i \\ \text{subject to} \quad & y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0. \end{aligned}$$

Different from logistic regression, SVM maps features into a higher dimension and tries to find a linear separating hyperplane in this higher dimensional space. This allows SVM to separate data that's not linear separable in nature.

4.2.2 Data Featuring

One important thing to note for SVM is that the scale of features can have a great impact on the model. To account for this problem, data in training sets are scaled using min-max scale to the range of 1 to -1 and data in test sets are scaled according to the scale of training sets.

There are two hyperparameters that need to be tuned: gamma and C. Gamma controls the fitness of the model while C adds a penalty on the error term. Five-fold cross validation and grid search are performed on the training set to find the hyperparameters with the highest average accuracy. The graph below shows how model accuracy changes as the value of hyperparameters change for Moutai and Wuliangye.

Figure 4.2.1 Change of CV accuracy with C (left: Moutai, right: Wuliangye)

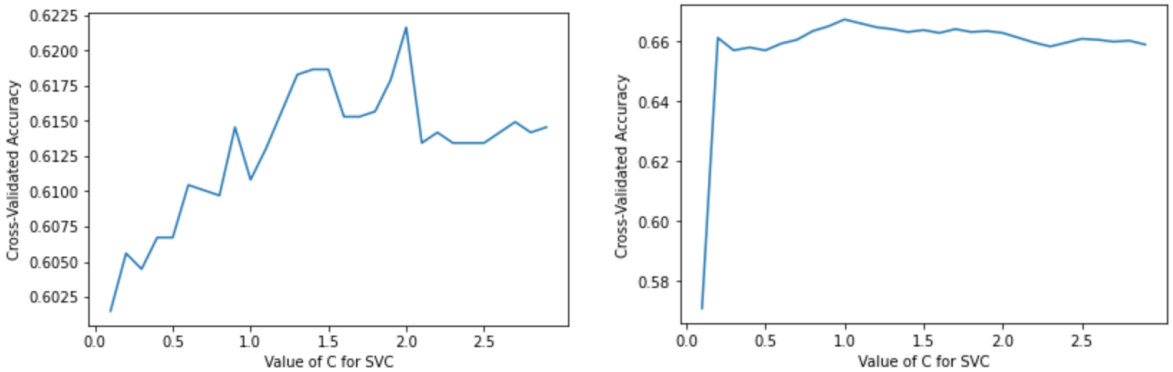
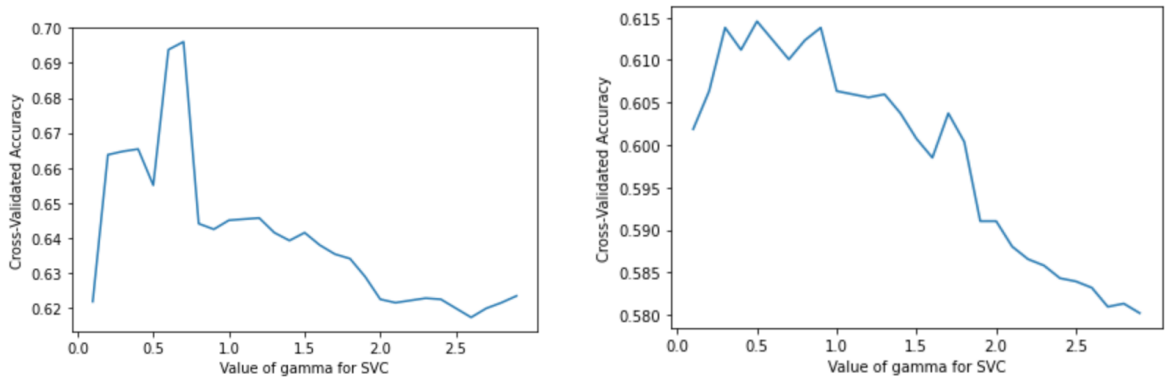


Figure 4.2.2 Change of CV accuracy with gamma (left: Moutai, right: Wuliangye)



4.2.3 Result Analysis

The table below shows the accuracy as well as the AUC for models fitted for Moutai and Wuliangye. We can see that compared to logistic regression, the prediction accuracy for Moutai is about the same yet it makes a more balanced prediction which is reflected by AUC score. The performance for predicting Wuliangye highly increased as well as the AUC score. However, we think this prediction is still not satisfying, mainly due to the limitation of predefined model

structure. In the next part, we use unsupervised learning methods to better capture the underlying trend of data.

Stock	Accuracy	AUC
Kweichow Moutai	65.58%	0.6540
Wuliangye	77.12%	0.7060

4.3 Random forest

4.3.1 Introduction to Random Forest

Random forest model follows the steps as data collection, exponential smoothing, feature extraction, ensemble learning, stock market prediction. We use the exponentially smoothed time series data and technical indicators to predict the Wuliangye and Moutai's stock market price. The step of exponential smoothing can put more weight on recent data and removes random variation from the historical data. Since decision trees have the characteristics of low bias and high variance, random forest can overcome the problem by training multiple decision trees on different subspaces and the splitting method is decided on Shannon Entropy or Gini impurity. The best split is chosen by reducing the impurity and increasing the gain in information.

Gini impurity is given by $g(N) = \sum_{i \neq j} P(\omega_i)P(\omega_j)$ where $P(\omega_i)$ is the proportion of the population with class label i.

Shannon Entropy measures the disorder for the information and it is given by

$H(N) = -\sum_{i=1}^{i=d} P(\omega_i) log_2(P(\omega_i))$ where d is the number of classes considered.

4.3.2 Measurements of Models

By using the prediction result from models, we can decide on whether to buy or sell the stock. We predict the result as +1 for buying the stock since the price will increase after n days. Otherwise, if the result is -1, we sell the stock since the price will decrease after n days. Our model will be evaluated for its robustness and those indicators are Accuracy, Precision, Recall, Specificity.

$Accuracy = \frac{tp+tn}{tp+tn+fp+fn}$, $Precision = \frac{tp}{tp+fp}$, $Recall = \frac{tp}{tp+fn}$, $Specificity = \frac{tn}{tn+fp}$

$tp = true\ positive,\ tn = true\ negative,\ fp = false\ positive,\ fn = false\ negative$

We also use Receiver Operating Characteristic to measure the performance of the prediction. It is a curve of True Positive rate against False Positive Rate and it is a trade-off between specificity and sensitivity.

4.3.3 Random Forest-Moutai

We plot the adjusted price for Moutai to have a clear overview of the hugely increasing price from 2001 to 2021.

We apply two different splitting data methods to the Moutai dataset for better comparison. The first model splits the data using the sklearn package (sklearn.model_selection.train_test_split) which splits the data randomly. The second model splits the data as first 70% for the training set and the last 30% for the testing set.

Figure 4.3.1 Price of Moutai

Table 4.3.1 Measurements of Moutai stock

Model	Length of training set	Length of testing set	**Accuracy**	Recall	Precision	Specificity
Model 1	3536	1179	**0.88**	0.80	0.87	0.93
Model 2	3301	1414	**0.42**	0.82	0.37	0.18

Figure 4.3.2 Model 1 Performance

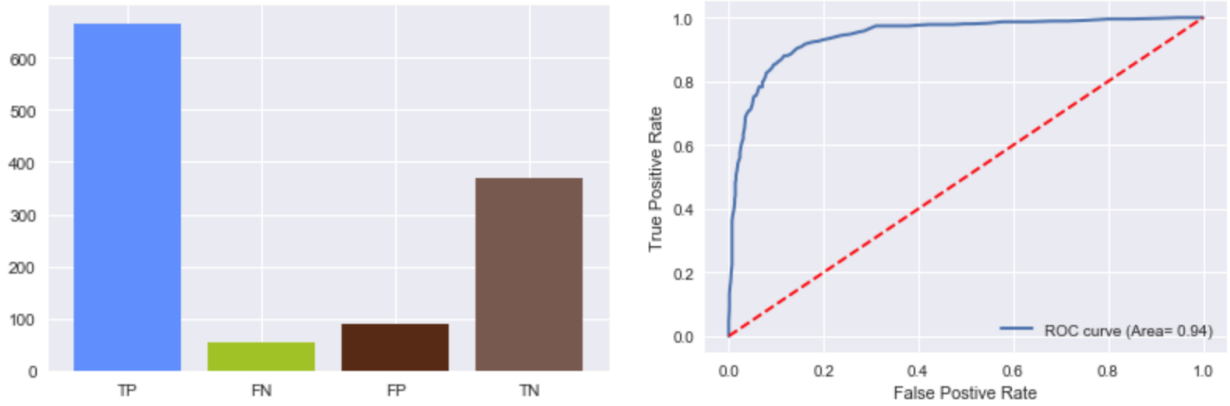
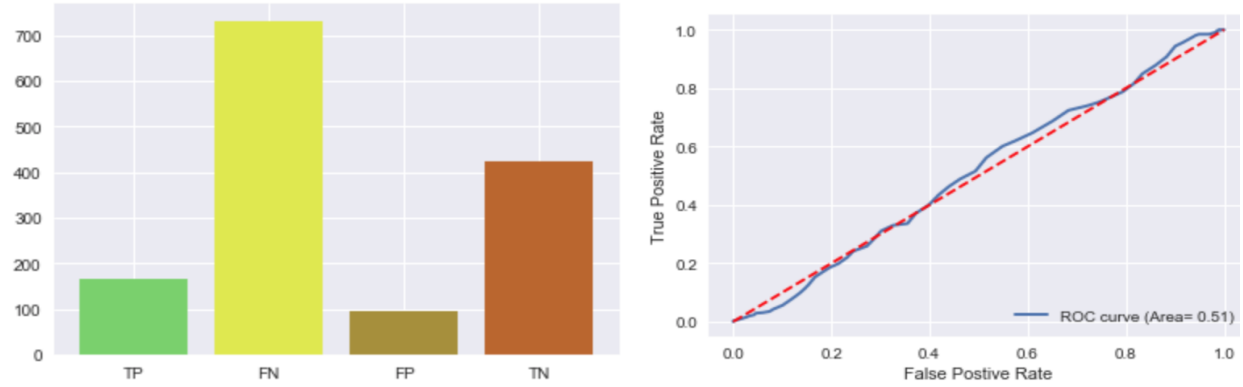


Figure 4.3.3 Model 2 Performance



Model 1 achieves higher accuracy, recall, precision and specificity than model 2. By analyzing and computing measurements, we have high true positive cases and the ROC curve is approaching the top and left-hand border which indicates that the result is accurate. For the model 2, the ROC is approaching 45 degrees diagonal of the space and indicates the model has low accuracy.

4.3.4 Random Forest-Wuliangye

Figure 4.3.4 Price of Wuliangye

Wuliangye stocks also have high stock prices and the prices have increased sharply these years.

Model 1 achieves high accuracy for Moutai stock and model 2 performs badly for the prediction. We apply another stock to check the accuracy of the models. For Wuliangye stock, the first model also splits the data using the sklearn package (sklearn.model_selection.train_test_split) which splits the data randomly. The second model splits the data as first 70% for the training set and the last 30% for the testing set.

Table 4.3.2 Measurements of Wuliangye stock

Model	Length of training set	Length of testing set	**Accuracy**	Recall	Precision	Specificity
Model 1	4104	1369	**0.85**	0.83	0.83	0.86
Model 2	3832	1641	**0.51**	0.99	0.50	0.04

Figure 4.3.5 Model 1 Performance

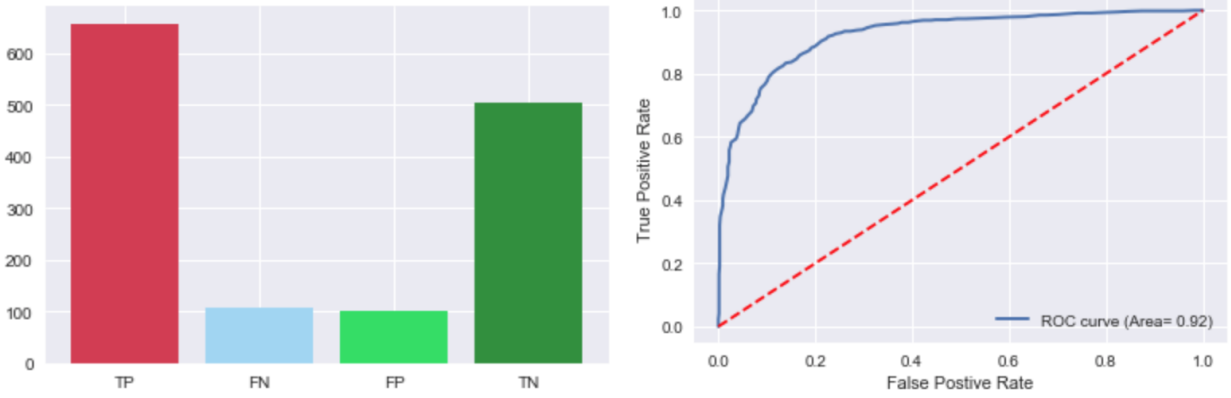
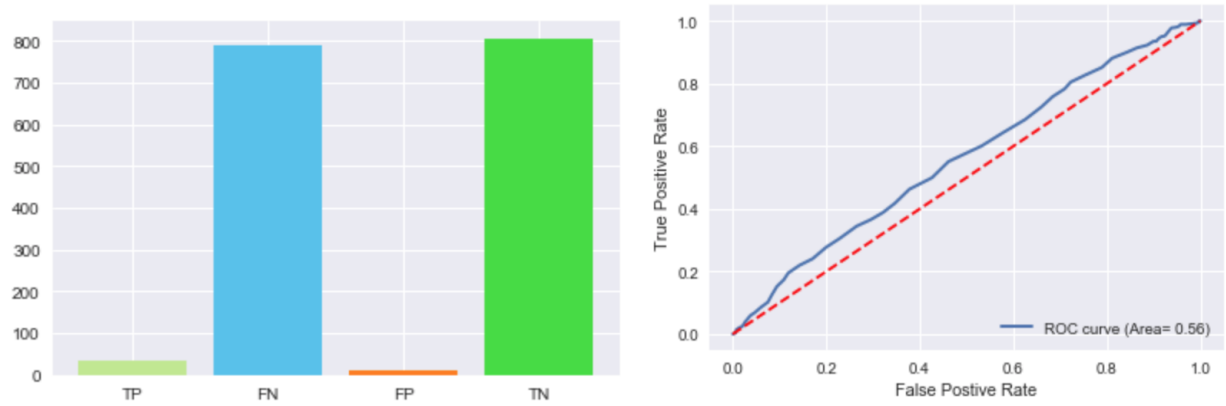


Figure 4.3.6 Model 2 Performance



For Wuliangye stock, model 1 has higher accuracy than model 2. Model 2 performs a little better than the model 2 of Moutai stock.

4.3.5 Analysis of Models

We achieve a high accuracy for both Moutai and Wuliangye stocks. However, the model 1 has problems regarding the high accuracy. The model might be overfitting and the main problem is data leakage. The model 1 splits the data randomly and the splitting method is not realistic in the real world. Due to the method of exponential smoothing and intrinsic autocorrelation of the data, the model has data leakage problems.

5 Future Work

The features we've selected are solely based on the performance of the stock itself which might be one reason that limits the predictive power of data. We believe the performance of the model can be improved by incorporating more features that reflect the performance of the Baijiu market, such as macro economical data that reflects the performance of the general market. In addition, inspired by the referenced paper, it might be a good idea to extend the prediction period, predicting the performance of stock 60 days as well as 90 days ahead, compare and contrast the performance of different models.

6 Conclusion

In this paper, we dug into one of the most heated markets in Mainland China, the Chinese Baijiu stock market and tried to forecast the top 2 stocks in the market, namely Moutai and Wuliangye using logistic regression, SVM and random forest. In addition to features that can be retrieved directly from stock itself, such as close price and volume, we also incorporated some of the most valued technical features in the industry, namely Strength Index, Stochastic Oscillator, Williams %R, Moving Average Convergence Divergence, Price Rate of Change and On Balance Volume. Data featuring were performed based on the different assumptions for each model. By comparing

the prediction accuracy on the testing set and AUC score, we found that the random forest method performs the best on both stocks, with prediction accuracy as high as 88% and 85% for Moutai and Wuliangye respectively.

References

Khaidem et al. (2016). Predicting the Direction of Stock Market Prices using Random Forest.