Whether cumulative stock return can reach target value after IPO —Based on China A-share market

Ai Jianxuan, Fu Jintao, Huang Chenxi, Gong Chao, Zhao Enping

June 29, 2021

1 INTRODUCTION

1.1 Business Concern

MARKET CURRENT SITUATION Since 2014, China's A-share market began to show the phenomenon that the new stocks keep at limit up price for several days after IPO. On the one hand, these new shares have a wealth effect, which can stimulate investors to participate in the trading these stocks. But on the other hand, these new shares will lead to scarce trading volume and extremely low turnover rate, thus creating the inflated stock price and valuation bubble. Finally it will invisibly amplifies the market risk and distorts the market's asset pricing function. In reality, there are also a lot of cases that show that after the open trading day, the number of new shares that show upward trend is much lower than the number of new shares that show downward trend , and some of the new shares after open trading day will fall by more than 50%, even some will directly fall to lower than the initial IPO issuing price. Therefore, there are certain risks in the purchase of new shares.

INVESTMENT RESTRICTION For individual investors, it is very difficult to buy new stocks during the keeping limit up price period or before IPO days. Only after open trading day, individual investors are able to buy or sell the new shares. As for open trading day, this concept means the day that stock can be traded after IPO while new stock cannot be traded if it keeps limit-up price. Then there is a problem, if individual investors only randomly invest in the new shares after open trading day without considering the price trend, it may cause some losses. In other words, if individual investors can effectively screen out certain stocks with higher or cumulatively higher return than a certain target value after open trading day, then investors will have more chance to gain profit than just random picking.

MACHINE LEARNING PURPOSE For the above research purposes, this paper, from the perspective of individual investors, focuses on finding a trading and investment strategies after new shares end keeping limit up price period. In this paper, we use six algorithm models of machine

learning to predict whether the stock return of the first **1, 3, 5, 7, 10** days after open trading day can reach target value **10%**. Then we use this result as the criteria to select new stock investment. Through such a forecast, we would be able to select new shares for investment to earn a profit for certain holding periods before open trading day.

1.2 Data Selection And Data Descriptions

As for the object of study,we have extracted 10 years of data, ranging from 2011 to 2021, for all IPO stocks in China. In total, we have around 2200 data points and 40 over variables. Within these variables, we select 18 variables (shown in Figure1.1) as explanatory variables to predict whether the stock return can achieve **10%** with the first **1, 3, 5, 7, 10** days holding periods after open trading day. The descriptions of selected variables as listed in Figure 1.1:

Variable Names	Variable Description
Stock Code	nominal variable
Company Name	nominal variable
IPO Date	The date the stock goes public
Purchase Open Day	The date on which a stock can be traded after IPO
Board Type	Four boards: 1. Shanghai Stock Exchange (SSE) 2. Growth Enterprise Market (GEM) 3. Sci-Tech innovation board (STAR) 4. Shenzhen Stock Exchange (SZSE)
Industry Type	Three main industries: 1.Agriculture ; 2.Industrial ; 3.Services
Market Trend	Market index growth during IPO Day and IPO Open Day: 1.if growth rate>0, market trend=1 2.if growth rate=0, market trend=0 3.if growth rate<0, market trend=-1
IPO Month	The month of the IPO date
Keep Limit Up Days	Cannot trade during these days
Issue Price	The offering price listed in the prospectus
Before Open Day Price	The price just the day before stock trading day
Industry PE	The PE of the industry that stock belongs to
Before Open Day PE	The PE of stock just the day before stock trading day
IPO Market Cap	The market cap listed in the prospectus
Before Open Day Market Cap	The market cap just the day before stock trading day
Claw back Ratio	Given-out stocks are required to be returned
Total Issuance cost (%)	The cost of issuing stock
Face Value	The price of stock per share
Y1	Y1 is a binary variable: 1.if return≥target value, Y=1 2.if return<target value, Y=0

Figure 1.1: Data Descriptions For All X And Y Variables

1.3 Data Processing

1.3.1 Check For Correlations

CORRELATION In statistics, multicollinearity is a phenomenon in which one predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy. In order to avoid this situation, we need to select variables carefully. When there is a problem of multicollinearity between two variables, we need to make a tradeoff to discard one of them .After selecting, we have obtained 20 variables as our machine learning data, just shown as below in Figure 1.2:

Month	Keep Limit Up Days	Issue_Price	Industry_PE	Before Open Day Price
Clawback_Ratio	Total Issuing Cost	Face Value	IPO_MktCap	Before Open Day PE
Mkt Trend	industry_=_1.0	industry_=_2.0	industry_=_3.0	Before Open Day Market Cap
board_=_1.0	board_=_2.0	board_=_3.0	board_=_4.0	Y1

Figure 1.2: Data Descriptions For All X And Y Variables

1.3.2 Deal With Imbalanced Sample

IMBALANCED SAMPLE This problem refers to the extremely uneven samples from different categories in the dataset. Taking binary classification problem as an example, if the number of positive category is much larger than that of negative category, usually the data with the sample category ratio exceeding 4:1 (or 3:1) can be called as imbalanced data. Before we deal with imbalanced sample problem, we can see the ratio of $Y = 0 : Y = 1 \approx 4 : 1$, which means imbalanced data structure is our concern. In practice, if our sample size is relatively small, we are more likely to use oversampling beacause it makes better use of data.

SMOTE OVERSAMPLING In this project, we used SMOTE method for oversampling, namely the synthesis of minority type of oversampling technology, which is an improved scheme to solve the problem that random oversampling is prone to lead to overfitting. After we complete smote oversampling, we can get balanced data and now $Y = 0 : Y = 1 = 1 : 1$.

1.3.3 Search For Best Parameter

GRIDSEARCHCV When we use these six machine learning models, we actually have some problems about parameters. In this situation, we can use GridSearchCV to help searching best parameters, the purpose of this method is to automatically adjust the parameters, as long as you input the range of parameters,this method will provide optimal results and parameters.

2 MODEL DESCRIPTION & MODEL BUILDING

In this project, we actually use a total of 6 models respectively to analyze our problem as follows: **(1)**Logistic Regression(LR), **(2)**Random Forest(RF), **(3)**Support Vector Machine(SVM), **(4)**K-Nearest Neighbors(KNN), **(5)**Multilayer Perceptron(MLP), **(6)**Gaussian Naive Bayes(Gaussian NB).

2.1 Model 1 : Logistic Regression

CONCEPT As for logistic regression, it's a linear Regression, but ultimately it is used as a classifier, which means it learns fitting parameters from the sample set, fits the target values to [0,1],

and then discretizes the target values to achieve classification. Sigmoid function is used to correlate the real markup of the classification task with the predicted value of the linear regression model. The specific calculation formula of Sigmoid function is as follows:

$$\sigma_{(z)} = \frac{1}{1+e^{-z}}$$

Then we can extend this formular into our logistic regression equation:

$$P(y=1|x;\theta) = \frac{1}{1+e^{-\theta Tx}}$$

Where$P(y=1|x;\theta)$means the probability of predicting the dependent variable to be 1,namely is positive.

2.2 Model 2 : Random Forest

CONCEPT Random forest algorithm has the advantages of fewer parameters to be adjusted, anti-overfitting, fast classification, strong ability to process large sample data, ranking the importance degree of each feature in classification, and strong anti-noise ability. In practical application, random forest model is mainly used for classification prediction and regression prediction.

2.3 Model 3 : Support Vector Machine

CONCEPT One of the most basic ideas of support vector machine is that the principle of structured risk minimization is superior to the traditional principle of empirical risk minimization,. which shows many unique advantages in solving small sample, nonlinear and high-dimensional pattern recognition problems. SVM is derived from classification problem, which is a kind of forward neural network in essence.

2.4 Model 4 : K-Nearest Neighbors

CONCEPT KNN is a machine learning algorithm that can be used for both classification and regression. A smaller K value means that the model becomes more complex and prone to overfitting. If the K value is large, the generalization error can be reduced, but the training error will increase. A large K value means that the model becomes simple and underfitting is easy to occur. When $K = 1$, the k-nearest neighbor algorithm is the nearest neighbor algorithm. K value is generally selected by cross validation method.

2.5 Model 5 : Multilayer Perceptron

CONCEPT A multilayer perceptron (MLP) can be thought of as a logistic regression classifier whose input data is first transformed using the learned nonlinear transformation. Map the input data into a linearly separable space. This middle layer is called the hidden layer. A multilayer perceptron with a single hidden layer is sufficient to be a universal approximator.

FORMULAR EXPRESSION Specifically, given a small batch sample $X \in R^{n\times d}$. Its batch size is n, the number of inputs is d.Suppose that the multilayer perceptron has only one hidden layer, where the number of hidden units is h. Note that the output of the hidden layer (also known as the hidden layer variable or the hidden variable) is H, where $H \in R^{n\times h}$. Since both the hidden layer and the output layer are fully connected layers, the weight parameters and deviation parameters of the hidden layer can be set as $W_h \in R^{d\times h}$ and $b_h = R^{1\times h}$. The weight and deviation parameters of the output layer are respectively $W_0 \in R^{h\times q}$ and $b_h = R^{1\times q}$.

Let's start with the design of a multilayer perceptron with a single hidden layer. Its output $O \in R^{n \times q}$ is shown as below and if we combine them, then we can get:

$$H = XW_h + b_h$$
$$O = HW_0 + b_0$$
$$O = (XW_h + b_h W_0) + b_0 = XW_h W_0 + b_h W_0 + b_h W_0 + b_0$$

2.6 Model 6 : Gaussian Naive Bayes

CONCEPT Different from other classifiers, Naive Bayes is a classification algorithm based on probability theory. It assumes that the likelihood of the feature is a Gaussian distribution:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2})$$

3 PERFORMANCE METRICS ANALYSIS

3.1 Selected Peformance Metrics

In our project, we select six performance metrics to measure how good or bad a model is.

PRECISION SCORE The question that this metric answer is of all passengers that labeled as survived, how many actually survived ? High precision relates to the low false positive rate.

$$Precision = TP/(TP + FP)$$

RECALL SCORE As for recall, the question recall answers is: Of all the passengers that truly survived, how many did we label ?

$$Recall = TP/(TP + FN)$$

F1 SOCRE F1 Score is the weighted average of Precision and Recall.

$$F1 = 2 \times (Recall + Precision)/(Recall + Precision)$$

FBETA SCORE Fbeta-measure provides a configurable version of the F1 measure to give more or less attention to the precision and recall measure when calculating a single score.

AREA UNDER CURVE The ROC curve is fully known as receiver operating characteristic curve. It is based on a series of different dichotomies (cut-off value or determination threshold) and takes the true positive rate (sensitivity) as the ordinate. The false positive rate (specificity) is plotted on the abscissa. AUC is defined as the area under curve. The larger AUC value, the better classifier performance.

LOG LOSS This is the loss function used in (multinomial) logistic regression and extensions of it such as neural networks, defined as the negative log-likelihood of a logistic model that returns y-pred probabilities for its training data y- true. For a single sample with true label $y \in 0, 1$ and a probability estimate $p = Pr(y = 1)$,the log loss is:

$$L_{log}(y, p) = -(ylog(p) + (1 - y)log(1 - p))$$

3.2 Peformance Metrics Result

MODELS AND HOLDING-PERIODS COMPARISION We use six models to predict whether the stock return of the first **1,3,5,7,10** days after open trading day can reach target value **10%**. The performance metrics of these 6 models are shown as below in Figure 3.1:

Holding Period	Performance Metrics	LR	RF	SVM	MLP	KNN	NB
1-Day	Precision	0.5697	0.9598	0.8640	0.7084	0.5359	0.5359
	Recall	0.5906	0.7749	0.8649	0.7403	0.8984	0.8984
	F1	0.5799	0.8575	0.8645	0.7240	0.6714	0.6714
	F0.5	0.5738	0.9161	0.8642	0.7146	0.5830	0.5830
	AUC	0.5980	0.9317	0.9309	0.8123	0.8944	0.6961
	Log Loss	14.8753	4.4789	4.7157	9.8137	15.2942	15.2942
3-Days	Precision	0.5786	0.9206	0.8895	0.7340	0.8526	0.5515
	Recall	0.6002	0.7225	0.7746	0.6999	0.6750	0.8188
	F1	0.5892	0.8096	0.8281	0.7165	0.7535	0.6591
	F0.5	0.5828	0.8728	0.8639	0.7269	0.8100	0.5900
	AUC	0.6124	0.8964	0.8995	0.7967	0.8629	0.6415
	Log Loss	14.2834	5.7984	5.4891	9.4514	7.5379	14.4575
5-Days	Precision	0.5747	0.9177	0.8913	0.7258	0.7867	0.5362
	Recall	0.6016	0.7260	0.7580	0.6861	0.7409	0.9292
	F1	0.5878	0.8107	0.8192	0.7054	0.7631	0.6800
	F0.5	0.5799	0.8717	0.8610	0.7175	0.7771	0.5858
	AUC	0.6015	0.8947	0.8981	0.7997	0.8464	0.6485
	Log Loss	14.4206	5.7955	5.7175	9.7958	7.8640	14.9476
7-Days	Precision	0.5532	0.9569	0.8622	0.7242	0.7798	0.5087
	Recall	0.6221	0.7300	0.8075	0.6995	0.7688	0.9261
	F1	0.5856	0.8282	0.8339	0.7116	0.7742	0.6567
	F0.5	0.5658	0.9009	0.8506	0.7191	0.7775	0.5591
	AUC	0.6019	0.9043	0.8969	0.7910	0.8419	0.6498
	Log Loss	14.8705	5.1154	5.4327	9.5766	7.5740	16.3577
10-Days	Precision	0.5704	0.9478	0.8732	0.7481	0.8040	0.5200
	Recall	0.5677	0.7055	0.7951	0.6820	0.7585	0.9187
	F1	0.5691	0.8089	0.8323	0.7135	0.7806	0.6567
	F0.5	0.5699	0.8869	0.8564	0.7338	0.7945	0.5694
	AUC	0.5959	0.8920	0.8946	0.7942	0.8545	0.6484
	Log Loss	14.8054	5.7396	5.5165	9.4308	7.3418	16.0022

Figure 3.1: Performance Metrics Of Six Models

According to this listed table, we can do some analysis from two perspectives. One perspective is that we can see for each specific model,which holding period can be most accurately predicted by this model . Another perspective is that we can focus on each holding period to find which model can be best used to do prediction.

1. As for the first perspective, we want to find out which holding period can be best predicted by one specific model. For **Logistic Regression Model**, we can see it can be best used to predict 3-Days holding period by looking at AUC score (61.24%). For **Random Forest Model**, **Support Vector Machine Model**, **Multilayer Perceptron Model**, **K-Nearest**

Neighbors Model and **Guassian Naive Bayes Model**, we can see they can be best used to predict 1-Day holding period by looking at AUC score is (93.17%) ,(93.09%) ,(81.23%) ,(89.44%) and (69.62%) respectively.

2. As for the second perspective, if holding-period is 1-Day, we can see that the Random Forest is the best model to do prediction. The precision score of RF is 0.9598, which means RF model can have 95.98% probability to correctly predict whether stock return can achieve 10% for a given holding period. The recall of RF is 0.7749, which means RF model can have 77.49% probability to correctly identify all stock that have potential to get 10% yield. Actually,because precision and recall rates tend to trade off, we can use F1 score to combine these two ratios. These two F scores are pretty good, indicating RF model is very good at prediction. AUC is 0.9317 that means the model has pretty good performance at distinguishing between the positive and negative classes. As for log loss, it's useful to compare models not only on their output but on their probabilistic outcome. We can see from the table that log loss of Random Forest model is relatively smaller than others.
3. Follow up on the above point, if holding-period is 3-Days, 5-Days,7-Days and 10-Days, we can also derive the similar results since the Random Forest model and Support Vector Machine model are the best two models that perform very well. Overall, we can see that the precision score, F0.5 score and AUC are all near to 90%, which means the prediction of our model is satisfactory. From log loss perspective, we can also conclude that RF model and SVM model are better because of comparatively lower log loss value than others.

AUC OR ROC COMPARISION The reason why we concentrate more on ROC is that ROC is an evaluation that is independent of thresholds, which means it can easily detect the influence of arbitrary threshold on the generalization performance of the learner. We choose **1-day** as an example to show our AUC comparision and we can clearly to derive the conclusion that Random forest is the best model and Support Vector Machine is next shown in Figure 3.2.

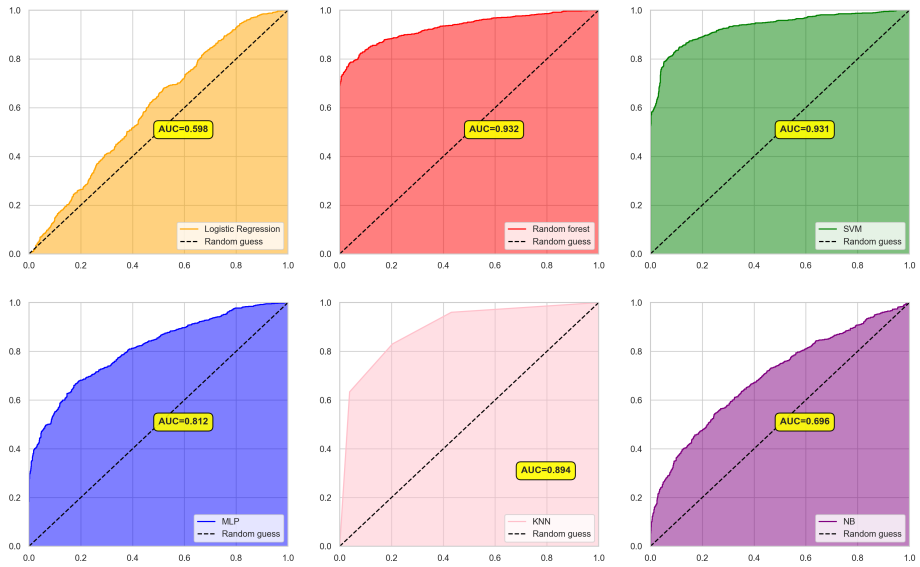

Figure 3.2: AUC (1-Day) Of Six Models

4 CONCLUSION

In this project, we have totally dealed with six models (LR, RF, SVM, MLP, KNN ,NB), six performance matrics (Precision, Recall, F1, F0.5, AUC, Log Loss) for five different holding periods (1, 3, 5, 7, 10 days). Besides, we set 10% stock return as our target value to find out the predicted performance of our models. When buliding prediction model, we choose nearly 20 input features to conduct supervisory training and optimize parameters for each classification model. In this part, we want to conclude from two views: **(1)**trading suggestion , **(2)**model prediction.

4.1 Trading Suggestion

As for trading strategy, we need to consider the choice underlying stocks and the holding period of strategy. Focusing on the choice of trading underlying stocks, if we just invest into the whole new stocks in holding periods without filtering, we would have large probability of losing money in the long-run. In other words, if we are able to select certain stocks that have potentail to achieve target value in selected holding periods, then we can conduct long chasing trading for these new stocks to obtain higher return. Looking at the choice of holding periods, our result shows that short-term holding periods are recommended. The reason why we prefer short-term investment rather than long-term investment is due to the existence of a continuous keeping at limit up price on our A stock market. As a result, the valuation premium before IPO often exceeds its resaonable valuation range. Therefore, there is high probability that new stocks would like to go down back to its reasonable valuation range in the short-term after open trading day. It's sensible for investors to sell the holding new stocks within short-term periods, like 1-3 days.

4.2 Model Prediction

If we choose the holding period after open-trading days to be 1-day,3-days,5-days,7-days and 10-days respectively, we can compare the result by looking at these six performance metrics. Generally speaking, for all the holding period(**1, 3, 5, 7, 10** days), random foreset**(RF)** and support vector machine**(SVM)** relatively perform very well,which means they show **higher** precision score, recall score, F1 score, F0.5 score, AUC value and **lower** log loss. In other words, Random Forest Model and Support Vector Machines Model can give pretty good prediction, which is close to 95%. Especially for Random Forest Model, it has strong adaptability and good resistance to overfitting in the prediction situation of classification based on features.

REFERENCES

[1] Kibbey Tohren C.G., Jabrzemski Rafal, O'Carroll Denis M. Source allocation of per- and polyfluoroalkyl substances (PFAS) with supervised machine learning: Classification performance and the role of feature selection in an expanded dataset[J] Chemosphere, 2021, 275.

[2] Lee You Won, Choi Jae Woo, Shin Eun Hee Machine learning model for predicting malaria using clinical information[J] Computers in Biology and Medicine, 2021, 129.

[3] M. Prashanthi Reddy;T. Uma Devi Prediction of Diagnosing Chronic Kidney Disease using Machine Learning: Classification Algorithms [J] International Journal of Innovative Technology and Exploring Engineering,2020.