Penalized Logistic Regressions with Technical Indicators Predict Up and Down Trends

Huifeng Jiang (jianghuifeng0221@163.com)
Chongqing Technology and Business University https://orcid.org/0000-0002-5842-9876
Xuemei Hu
Chongqing Technology and Business University
Hong Jia
Chongqing Technology and Business University

Research Article

Keywords: Penalized logistic regressions, Up and down trends, Coordinate descent algorithm, Support vector machine, Artificial neural network

Posted Date: December 22nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1098354/v1

Penalized logistic regressions with technical indicators predict up and down trends

Huifeng Jiang[1], Xuemei Hu[2*] and Hong Jia[1]

[1]Research Center for Economy of Upper Reaches of the Yangtse River, Chongqing Technology and Business University, 19 Xuefu Avenue, Chongqing, 400067, China.

[2*]School of Mathematics and Statistics, Chongqing Key Laboratory of Social Economy and Applied Statistics, Chongqing Technology and Business University, 19 Xuefu Avenue, Chongqing, 400067, China.

*Corresponding author(s). E-mail(s): huxuem@163.com;
Contributing authors: jianghuifeng0221@163.com;
1055744669@qq.com;

Abstract

Predicting up and down trends for stock prices is an important puzzle in the financial field. [5] proposed logistic regression with 6 technical indicators to predict up and down trends for Google's stock prices. In this paper we further propose the five penalized logistic regressions with 19 technical indicators: ridge ($\boldsymbol{L_2}$), lasso ($\boldsymbol{L_1}$), elastic net(EN), smoothly clipped absolute deviation (SCAD) and minimax concave penalty (MCP) to improve the prediction accuracy. Firstly, we combine the iterative weighted least square algorithm with the coordinate descent algorithm, and apply a training set to obtain parameter estimators and probability estimators. Then we adopt a test set to construct confusion matrices and receiver operating characteristic (ROC) curves, and apply them to assess their prediction performances. Finally we compare the proposed five prediction methods with logistic regression, support vector machine (SVM) and artificial neural network (ANN), and found that the MCP penalized logistic regression performs the best. Therefore, we develop a new efficient prediction method to predict up and down trends for stock prices.

Keywords: Penalized logistic regressions, Up and down trends, Coordinate descent algorithm, Support vector machine, Artificial neural network

JEL Classification: C13 , C53 , G11

1 Introduction

Stock market exists some inherent characteristics such as model uncertainty, parameter instability and noise accumulation. These characteristics make the stock market prediction more complex. Different viewpoints spring up in economic and finance. For example, both efficient market hypothesis and random walk theory assumed that the stock market was unpredictable, whereas Dow theory and [7] assumed that financial market was predictable, proposed many technical indicators, and developed the technical analysis methods for finance market, and [3] systematically summarized the economic forecasting problems, emphasized the challenges from stock price forecasting, and provided the strategies to improve the forecasting performances. In recent years, machine learning methods were proposed to predict stock market. For example, [11] developed support vector regression and a two-step kernel learning method for financial time series prediction. [9] proposed adaptive artificial neural network to predict the second day closing price of stock market index. [2] systematically reviewed the progress on artificial intelligence, neural network and support vector machine in predicting the change of stock price or direction. [13] proposed a novel stock price trend prediction system that can predict both stock price movement and its interval of growth (or decline) rate within the predefined prediction durations. [12] introduced a new method to simplify noisy-filled financial temporal series via sequence reconstruction by leveraging motifs (frequent patterns), and then utilized a convolutional neural network to predict up and down trends for stock prices. [8] used machine learning and deep learning algorithms to significantly reduce the risk of trend prediction. [10] proposed a comprehensive customization of feature engineering and deep learning-based model to predict the price trends for stock markets in China. [6] applied machine learning method with emotional and situational features to predict the future trends for stocks, etc..

It is an important issue in the financial world to predict up and down trends for stock prices. Even small improvements in predictive performance can be very profitable. To improve the predictive performance, [5] proposed logistic regression with 6 technical indicators to predict the up and down trends for Google's stock prices. Here we introduce the five penalties: L_2, L_1, EN, SCAD and MCP for logistic regressions with 19 technical indicators, and propose the five penalized logistic regressions to further improve prediction accuracy. Firstly, we combine the iterative weighted least square algorithm with the coordinate descent algorithm, and apply a training set to obtain parameter estimators and probability estimators. Secondly, we adopt a test set to construct two-class confusion matrixes and ROC curves, and apply them

to assess prediction performances. Finally we compare the proposed five prediction methods with logistic regression, SVM and ANN, and found that the MCP penalized logistic regression performs the best. Therefore, we recommend the MCP penalized logistic regression to predict up and down trends for stock prices and bring richer economic benefit for investors.

The rest of this paper is organized as follows. In Section 2, we establish the five penalized logistic regressions with 19 technical indicators. In Section 3, we apply the training set to learn the five penalized logistic regressions, and obtain parameter estimators and probability estimators. In Section 4, we adopt the testing set to obtain confusion matrices and ROC curves to evaluate their prediction performances. In Section 5, we compare the proposed five prediction methods with logistic regression, SVM and ANN.

2 Penalized Logistic Regressions

Let C_t be the closing price of a given stock at the end of the t-th trading day, $Z_t = C_{t+1} - C_t$ be the stock excess return,

$$Y_t = \begin{cases} 1, & \text{if} \quad Z_t > 0, \\ 0, & \text{if} \quad Z_t \leq 0, \end{cases} \tag{1}$$

represents the direction indicator function, where $Y_t = 1$ represents up trend, and $Y_t = 0$ represents down trend. Our main goal is to predict stock return movement directions using the current and past data. In the following we apply a training sample $D = \{x_t, y_t\}_{t=1}^n$ to learn up and down trends for stock prices and construct a two-category classification rule that may be hidden deeply in the raw data set, where x_t is the sample from the predictor vector X_t whose distribution is usually unknown. It is well-known that logistic regression is a powerful two-category classification method. Note that [7] proposed many technical indicators and developed the technical analysis methods for finance market. Therefore, we combine logistic regression with the technical analysis method and proposed the following logistic regression with 19 technical indicators

$$P\left(X_t;\ \beta_0, \beta\right) = P\left(Y_t = 1 \mid X_t;\ \beta_0, \beta\right) = \frac{\exp\left(\beta_0 + X_t^\top \beta\right)}{1 + \exp\left(\beta_0 + X_t^\top \beta\right)}, \tag{2}$$

$$1 - P\left(X_t;\ \beta_0, \beta\right) = P\left(Y_t = 0 \mid X_t;\ \beta_0, \beta\right) = \frac{1}{1 + \exp\left(\beta_0 + X_t^\top \beta\right)}, \tag{3}$$

where the parameter vector $\beta = (\beta_1, \beta_2, \ldots, \beta_{19})^\top$ are unknown, and the predictor vector $X_t = (X_{t,1}, X_{t,2}, \ldots, X_{t,19})^\top$ are composed of 19 technical indicators listed in Table 1. When X_t exists multi-collinearity, the probabilities perform poor. To improve them, one can introduce the penalized functions for logistic regression, construct the penalized logistic regressions for modelling the two-class classification problem and removing technical indicators that are irrelevant to the future stock price direction.

Let $x_t = (x_{t,1}, x_{t,2}, \ldots, x_{t,19})^\top$ and y_t be the observation samples for X_t and Y_t, respectively. Given the training set $\{x_t, y_t\}_{t=1}^n$, we obtain the log-likelihood function

$$L(\beta) = \sum_{t=1}^{n} \left\{ y_t \left(\beta_0 + x_t^\top \beta\right) - \log\left[1 + \exp\left(\beta_0 + x_t^\top \beta\right)\right] \right\}, \tag{4}$$

the negative log-likelihood

$$l(\beta) = -L(\beta) = -\sum_{t=1}^{n} \left\{ y_t \left(\beta_0 + x_t^\top \beta\right) - \log\left[1 + \exp\left(\beta_0 + x_t^\top \beta\right)\right] \right\}, \tag{5}$$

and the penalized negative log-likelihood function

$$Q(\beta) \equiv l(\beta) + p_{\lambda,\gamma}(\beta), \tag{6}$$

where the penalized functions $p_{\lambda,\gamma}(\beta)$ defined in Table 2 with the tuning parameter λ and the regularization parameter γ. By minimizing the penalized negative log-likelihood function (6), one can obtain the parameter vector estimator

$$\beta^{\text{new}} = \arg\min_{\beta} \left\{ -\sum_{t=1}^{n} \left\{ y_t(\beta_0 + x_t^\top \beta) - \log\left[1 + \exp\left(\beta_0 + x_t^\top \beta\right)\right] \right\} + p_{\lambda,\gamma}(\beta) \right\}, \tag{7}$$

where the intercept term β_0 does not be penalized.

3 Parameter Estimators and Probability Estimators

The penalized negative log-likelihood (6) is not differentiable. In order to obtain its minimum, the penalized logistic regression need be transformed into a convergent quadratic problem and apply the iteration procedure to obtain the corresponding estimator. [1] introduced the coordinate descent algorithm for the penalized logistic regressions, and obtained the iterative parameter estimators, see Table 3.

Here we also apply coordinate descent algorithm to obtain the parameter estimators $\widehat{\beta_0}$ and $\widehat{\beta}$ for the five penalized logistic regressions. Based on the parameter estimators $\widehat{\beta_0}$ and $\widehat{\beta}$, we compute the probability estimators

$$\widehat{P}\left(Y_t = 1 \mid X_t;\, \widehat{\beta_0}, \widehat{\beta}\right) = \frac{\exp\left(\widehat{\beta_0} + X_t^\top \widehat{\beta}\right)}{1 + \exp\left(\widehat{\beta_0} + X_t^\top \widehat{\beta}\right)}, \tag{8}$$

Table 1 19 technical indicators and their formulae

Indicators	**Descriptions**	**Formulae**
$X_{t,1}$(WMA)	Weighted Moving Average.	$WMA_t = [nP_t + (n-1)P_{t-1} + \ldots + P_1]/n!$.
$X_{t,2}$(DEMA)	Double Exponential Moving Average.	$DEMA_t(n) = 2EMA_t(n) - EMA_t(EMA_t(n))$, $EMA_t(n) = [2P_t + (n-1)EMA_{t-1}(n)]/(n+1)$.
$X_{t,3}$(ADX)	Average Directional Movement Index measures the strength of a trend.	$ADX_t = [(n-1)ADX_{t-1} + DX_t]/n$, $DX_t = [(+DI_t) - (-DI_t)]/[(+DI_t) + (-DI_t)]$, $+DI_t = H_t - H_{t-1}, -DI_t = L_{t-1} - L_t$.
$X_{t,4}$(MACD)	Moving Average Convergence Divergence compares a fast exponential moving average with a slow exponential moving average.	$MACD_t = EMA_t(s) - EMA_t(t), s < t$.
$X_{t,5}$(CCI)	Commodity Channel Index measures the current price relative to an average price.	$CCI_t = (M_t - SM_t)/0.015D_t$, $M_t = (H_t + L_t + C_t)/3$, $SM_t = \sum_{i=1}^{n} M_{t-i+1}/n$, $D_t = \sum_{i=1}^{n} \mid M_{t-i+1} - SM_t \mid /n$.
$X_{t,6}$(MO)	Momentum provides the difference of a series over two observations.	$MO_t(k) = P_t - P_{t-k}$.
$X_{t,7}$(RSI)	Relative Strength Index measures velocity magnitude of directional price movements.	$RSI_t(n) = 100 - 100/[1 + RS_t(n)]$, $RS_t(n) = UP_{avg}(n)/DOWN_{avg}(n)$.
$X_{t,8}$(ATR)	Average True Range.	$TR_t = \mathrm{Max}\left[(H_t - L_t), (H_t - C_t), (L_t - C_t)\right]$, $ATR_t(n) = \frac{1}{n}\sum_{t=1}^{n} TR_t$.
$X_{t,9}$(CLV)	Close Location Value is a metric utilized in technical analysis to assess where the closing price of a security falls relative to its day's high and low prices.	$CLV_t = \frac{C_t - L_t - (H_t - C_t)}{H_t - L_t}$.
$X_{t,10}$(CMF)	Chaiken Money Flow compares the whole volume with regard to the Close, High and Low prices.	$CLV_t = [(C_t - L_t) - (H_t - C_t)]/(H_t - C_t)$, $CMF_t = \sum(CLV_t \times VO_t)/\sum VO_t$.
$X_{t,11}$(CMO)	Chande Momentum Oscillator.	$CMO_t = \frac{SU_t - SD_t}{SU_t + SD_t} \times 100$.
$X_{t,12}$(EMV)	Ease of Movement Value.	$BR_t = \frac{V_t}{H_t - L_t}, EMV_t = \frac{MPM_t}{BR_t}$, $MPM_t = \left(\frac{H_t + L_t}{2}\right) - \left(\frac{H_{t-1} + L_{t-1}}{2}\right)$.
$X_{t,13}$(MFI)	Money Flow Index uses price and volume data for identifying overbought or oversold signals in an asset.	$TP_t = \frac{H_t + L_t + C_t}{3}, RMF_t = TP_t \times V_t$, $MFR_t = \frac{14PPMF_t}{14PNMF_t}$, $MFI_t = 100 - \frac{100}{1 + MFR_t}$.
$X_{t,14}$(ROC)	Rate Of Change.	$ROC_t = C_t/C_{t-n} \times 100$.
$X_{t,15}$(VHF)	Vertical Horizontal Filter can distinguish the types of market.	$VHF_t = \frac{HCP_t - LCP_t}{\sum\mid C_{t-i+1} - C_{t-i} \mid}$.
$X_{t,16}$(SAR)	Parabolic Stop-And-Revers is used to determine the direction of a trend and the potential reversal of a price.	$SAR_t = SAR_{t-1} + AF\,(H_{t-1} - SAR_{t-1})$, $SAR_t = SAR_{t-1} + AF\,(L_{t-1} - SAR_{t-1})$.
$X_{t,17}$(TRIX)	Triple Smoothed Exponential Oscillator is to filter price noise and insignificant price movements.	$TR_t(n) = EMA\left(EMA\left(EMA\left(C_t, n\right), n\right), n\right)$, $TRIX_t(n) = 100 \times (TR_t(n)/TR_{t-1}(n) - 1)$.

Indicators	Descriptions	Formulae
$X_{t,18}$(WPR)	William's indicator is a dynamic technical indicator that determines whether the market is overbought or bought.	$WPR_t = (H_{t-n} - C_t) / (H_{t-n} - L_{t-n}) \times 100.$
$X_{t,19}$(SNR)	Signal to Noise Ratio can see the trend direction of the stock.	$SNR_t = \mid C_t - C_{t-n} \mid / ATR_n.$

Table 2 Penalized functions

Penalized functions	Formulae
EN	$p_{\lambda,\gamma}(\beta) = \lambda\left[(1-\gamma)\|\beta\|_2 + \gamma\|\beta\|_1\right]$, $\lambda \in (0,\infty), \gamma \in (0,1)$. $\gamma = 0$, EN becomes L_2 penalty $p_{\lambda(\beta)} = \lambda\|\beta\|_2$. $\gamma = 1$, EN becomes L_1 penalty $p_\lambda(\beta) = \lambda\|\beta\|_1$.
MCP	$p_{\lambda,\gamma}(\beta) = \begin{cases} \lambda\beta - \frac{\beta^2}{2\gamma}, & \text{if } \beta \le \gamma\lambda, \\ \frac{1}{2}\gamma\lambda^2, & \text{if } \beta > \gamma\lambda. \end{cases}$ $\lambda \ge 0$, $\gamma > 1$.
SCAD	$p_{\lambda,\gamma}(\beta) = \begin{cases} \lambda\beta, & \text{if } \beta \le \lambda, \\ \frac{\lambda\gamma\beta - 0.5(\beta^2+\lambda^2)}{(\gamma-1)}, & \text{if } \lambda < \beta \le \lambda\gamma, \\ \frac{\lambda^2(\gamma+1)}{2}, & \text{if } \beta > \lambda\gamma. \end{cases}$ $\lambda \ge 0, \gamma > 2$.

$$\widehat{P}\left(Y_t = 0 \mid X_t;\, \widehat{\beta_0}, \widehat{\beta}\right) = \frac{1}{1 + \exp\left(\widehat{\beta_0} + X_t^\top \widehat{\beta}\right)}. \tag{9}$$

4 Two-class Prediction Performance

The two-class confusion matrix is a cross table of the true class and the predicted class. It accurately describes the two-class classification results, see Table 4.

Accuracy is the proportion of population samples that are correctly predicted

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{10}$$

and is the simplest index to evaluate prediction performance. However, it cannot reflect the losses from two types of errors. Therefore, a ROC curve is introduced to evaluate prediction performance. Suppose that $TPR(c) = P(X_1 < c)$ represents the true positive rate at the threshold c, and $FPR(c) = P(X_2 < c)$ represents the false positive rate at the threshold c. By setting the different threshold c, we calculate $\{(TPR(c), FPR(c))\}$ or (Sensitivity, 1-Specificity) to draw a ROC curve, where

$$\text{Sensitivity(True positive rate,TPR)} = TP/(TP + FN), \tag{11}$$

$$\text{Specificity(1-False positive rate,1-FPR)} = TN/(TN + FP). \tag{12}$$

Table 3 Penalized functions and parameter estimators for penalized logistic regressions

Penalized functions	**Parameter estimators**
EN	$\widehat{\beta}_j^{EN}(Z_j;\ \lambda) = \frac{S(Z_j,\lambda)}{\nu_j}.$
MCP	$\widehat{\beta}_j^{MCP}(Z_j;\ \lambda,\gamma) = \begin{cases} \frac{S(Z_j,\lambda)}{\nu_j - 1/\gamma}, & \mid Z_j \mid \leq \nu_j\lambda\gamma, \\ \frac{Z_j}{\nu_j}, & \mid Z_j \mid > \nu_j\lambda\gamma, \end{cases} \quad \gamma > 1/\nu_j.$
SCAD	$\widehat{\beta}_j^{SCAD}(Z_j;\ \lambda,\gamma) = \begin{cases} \frac{S(Z_j,\lambda)}{\nu_j}, & \mid Z_j \mid \leq \lambda(\nu_j+1), \\ \frac{S(Z_j,\gamma\lambda/(\gamma-1))}{\nu_j - 1/(\gamma-1)}, & \lambda(\nu_j+1) < \mid Z_j \mid \leq \nu_j\lambda\gamma, \\ \frac{Z_j}{\nu_j}, & \mid Z_j \mid > \nu_j\lambda\gamma, \end{cases} \quad \gamma > 1 + 1/\nu_j.$
Notation	$\nu_j = n^{-1}x_j^\top W x_j, j = 1,\cdots,p, \tilde{Y} = x^\top\beta^{(m)} + W^{-1}(Y-P),$ $Z_j = n^{-1}x_j^\top W\left(\tilde{Y} - x_{-j}\beta_{-j}\right) = n^{-1}X_j^\top W r + \nu_j\beta_j^{(m)}, \gamma = W^{-1}(y - P(x_t;\ \beta)),$ $x_{-j} = (x_1,\cdots,x_{j-1},0,x_{j+1},\cdots,x_p), \beta_{-j} = (\beta_1,\cdots,\beta_{j-1},0,\beta_{j+1},\cdots,\beta_p).$

Table 4 Two-class confusion matrix

	True class 1$(Y_t = 1)$	**True class 2** $(Y_t = 0)$
Predicted class 1$(\widehat{Y}_t = 1)$	TP	FP
Predicted class 2$(\widehat{Y}_t = 0)$	FN	TN

TP:True positive, FP:False positive, TN:True negative, FN:False negative.

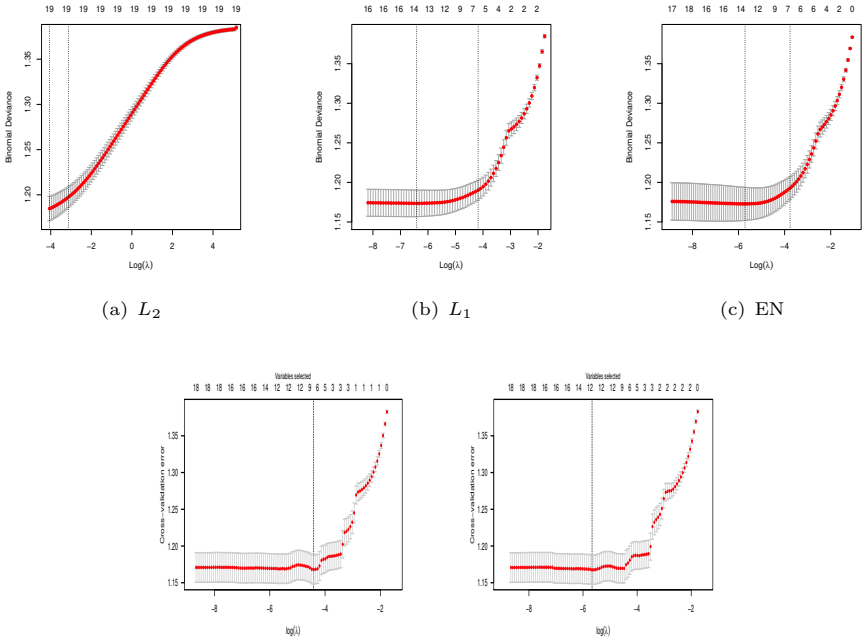
In Section 5 we adopt the R program package pROC to draw a ROC curve, compute AUC(the area under the ROC curve, a summary indicator of classification performance) and obtain the relevant statistics, the details can refer to Chapter 7 from [4].

5 Real Data Analysis

The stock market fluctuates greatly during December 2019 because of the novel coronavirus pandemic. Therefore we select Google's stock prices from January 2010 to November 2019 as the observation data, choose the 80% observation data as the training set to learn the up and down trends, and choose the remaining 20% observation data as the test set to predict the up and down trends for stock prices, where opening price (O_t), highest price (H_t), lowest price (L_t), closing price (C_t), volume (V_t) and adjusted price (A_t) are obtained from the Yahoo Finance port using R's getSymbols function. We consider 19 technical indicators: WMA, DEMA, ADX, MACD, CCI, Mo, RSI, ATR, CLV, CMF, CMO, EMV, MFI, ROC, VHF, SAR, TRIX, WPR and SNR as predictors, select Y_t defined (1) as response variables, and apply the aforementioned five penalized logistic regressions to predict up and down trends for Google's stock prices.

5.1 Tuning Parameter Selection

Variable selection is determined by the regularization parameter λ. In order to select the appropriate λ, we combine the coordinate descent algorithm with the ten fold cross-validation method to calculate the whole solution paths for model parameters, select a specific solution path from the whole solution paths, and apply the binomial deviance as the risk measurement. Then we get the mean cross validation error curve and the one-standard-deviation band shown in Figure 1. In particular, the parameter estimators for the MCP penalized logistic regression and the SCAD penalized logistic regression depend on the selection of regulating parameters γ and λ. They are usually chosen by cross-validation.



(a) L_2 (b) L_1 (c) EN

(d) MCP (e) SCAD

Fig. 1 The relationship between ten-fold cross-validation λ and model error

Figure 1(a)(b)(c) represent the 10 fold cross validation deviation curves for L_2, L_1 and EN, respectively. The number above each graph indicates the number of the selected variables. The left vertical line corresponds to $log(\lambda)$ when the minimum mean square error occurs, the right vertical line represents the corresponding $log(\lambda)$ when 1 times standard error occurs, and $log(\lambda)$ between the two vertical lines indicate that their errors are within a minimum standard error range(i.e.,the "one-standard-error" rule). We often use the rule

to select the best model. From Figure 1 we observe that the range of "one-standard-error" for L_2, L_1 and EN is $0.0173 \sim 0.0401$, $0.0020 \sim 0.0154$ and $0.0033 \sim 0.0213$, respectively. However, for MCD and SCAD, there is only one vertical line and corresponds to the $log(\lambda)$ when the average minimum error occurs, see Figure 1 (d)(e). We evaluate the prediction performance at each λ and γ value, select the best model corresponding to $\lambda = 0.0121$, $\gamma = 5$ for MCP and $\lambda = 0.0035$, $\gamma = 10$ for SCAD, and obtain the final five regressions. We compare the five regressions with logistic regression(LR), found that the parameters for LR and L_2 choose all 19 variables, whereas the other four penalized logistic regressions choose the different variables. This phenomenon indicates that there exists serious multi-collinearity between 19 variables, see Table 5.

Table 5 Selected variables for LR and the five penalized logistic regressions

Coefficient	LR	L_2	L_1	EN	MCP	SCAD
β_0	0.4918	-0.1008	-0.1049	-0.1050	-0.1137	-0.1520
β_1	0.2401	0.0159			0.0311	
β_2	-0.2343	0.0087				
β_3	0.0016	0.0031				
β_4	-0.0192	-0.0383				
β_5	-0.1959	-0.3103	-0.5788	-0.5451	-0.5068	-0.4751
β_6	0.0373	0.0869				
β_7	-0.0313	-0.1191		-0.1234	-1.0464	-1.0114
β_8	0.0214	0.0485	0.0260	0.050	0.0015	0.0956
β_9	-0.2859	-0.1568	-0.1622	-0.1761		-0.0992
β_{10}	0.3466	0.1977	0.0917	0.1443		0.1333
β_{11}	0.0208	0.4683	0.8422	0.8287	1.4888	1.6237
β_{12}	0.0507	0.0258				0.0488
β_{13}	0.0145	0.3582	0.3496	0.3981	0.4470	0.3876
β_{14}	-9.7227	0.1061				-0.2134
β_{15}	0.6910	0.0715	0.0101	0.0361	0.0122	0.0975
β_{16}	-0.0057	0.0033				
β_{17}	1.3665	0.1115		0.0368	0.4117	0.4375
β_{18}	-0.9247	0.1081				
β_{19}	-0.0983	-0.0250				-0.0918

5.2 Predicted Accuracy

We take advantage of the training set to study the Google stock price trends. In the following we apply the test set and the ROC curve to evaluate the prediction accuracy. According to the predicted class from the training set and the actual class from the test set, we establish the following second-class confusion matrix.

According to Table 6 we calculate Accuracy, Sensitivity and Specificity for LR as follows:

$$\text{Accuracy} = \frac{191 + 164}{191 + 164 + 84 + 51} \approx 0.724,$$

Table 6 Two-class confusion matrix

	Actual 1($Y_t = 1$)	**Actual 2** ($Y_t = 0$)
Predicted 1($\widehat{Y}_t = 1$)	191	84
Predicted 2($\widehat{Y}_t = 0$)	51	164

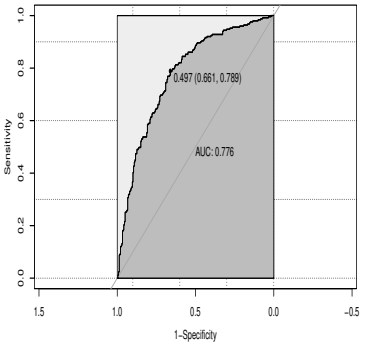
$$\text{Sensitivity} = \frac{191}{191+51} \approx 0.789, \qquad \text{Specificity} = \frac{164}{164+84} \approx 0.661.$$

Similarly, we calculate Accuracy, Sensitivity and Specificity for the five penalized logistic regressions listed in Table 7.

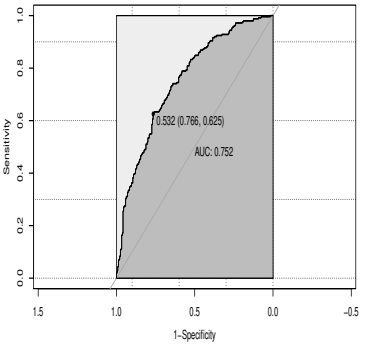
Table 7 Comparison of prediction accuracy for the six methods

	LR	$\mathbf{L_2}$	$\mathbf{L_1}$	**EN**	**MCP**	**SCAD**
Sensitivity	0.789	0.625	0.681	0.749	0.781	0.773
Specificity	0.661	0.766	0.720	0.678	0.678	0.686
Accuracy	0.724	0.694	0.705	0.712	0.732	0.731

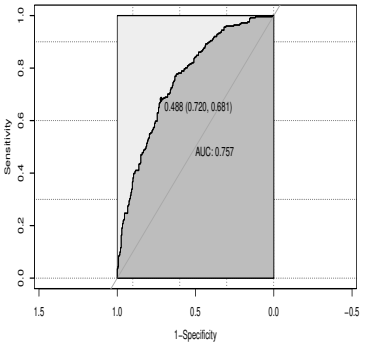
From Table 7 we observe the following facts: (1)For EN and L_1, Accuracy is higher than that of L_2, but is lower than that of LR; (2)Accuracy for MCP is higher than that of SCAD, whereas Accuracy for SCAD is higher than that of EN and LR. However, Accuracy is the simplest index to evaluate the prediction, and it cannot fully reflect the corresponding loss of two kinds of errors. Therefore, in the following we first compute Sensitivity and Specificity corresponding to different thresholds for the six methods, and then apply them to draw the ROC curve to evaluate Accuracy, see Figure 2.



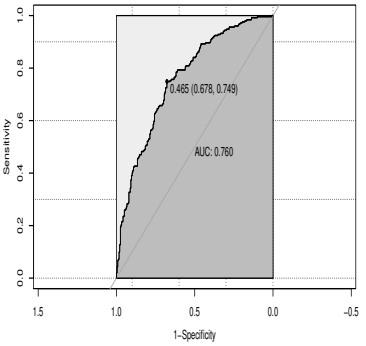
(a) ROC for LR



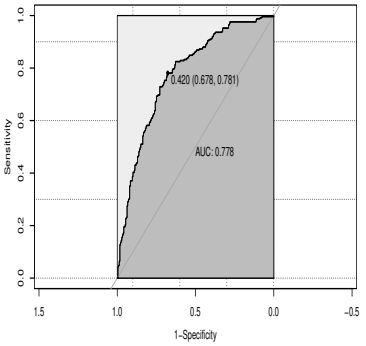
(b) ROC for L_2



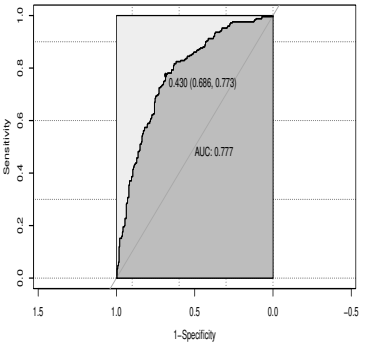
(c) ROC for L_1



(d) ROC for EN



(e) ROC for MCP



(f) ROC for SCAD

Fig. 2 ROC curves of six models for predicting the movement directions of stock price

In Figure 2, the AUC corresponding to LR, L_2, L_1, EN, MCP and SCAD are 0.776, 0.752, 0.757, 0.760,0.778 and 0.777, respectively. Combined with Accuracy listed in Table 7, it can be concluded that among the six methods, the MCP penalized logistic regression with technical indicators performs the best in terms of in terms of Accuracy. In order to further explain the superiority to the MCP penalized logistic regression in predicting stock prices trends movement, we compare the prediction results for the MCP penalized logistic regressions with those for SVM and ANN, see Table 8.

Table 8 Sensitivity, Specificity, Accuracy and AUC for MCP, SVM and ANN

	MCP	**SVM**	**ANN**
Sensitivity	0.781	0.705	0.725
Specificity	0.678	0.653	0.732
Accuracy	0.732	0.686	0.729
AUC	0.778	0.679	0.759

From Table 8, we can observe that among the aforementioned three methods, MCP performs the best in terms of Sensitivity, Accuracy and AUC. The reason that SVM performs the worse may be that Gaussian kernel function is a typical local kernel function, and it only affects the data points in a small area near the test point, and has strong learning ability and weak generalization performance. In addition, ANN is unstable, so we choose the average of the 10 predicted results as the final values, and they are worse than MCP. Obviously, the MCP penalized logistic performs best in predicting the trend of stock price ups and downs. Therefore, we recommend the MCP penalized logistic regressions to predict the stock price trend movements.

6 Conclusion

Based on Murphy's technical analysis method, we combine technical indicators with penalized logistic regression and propose the five penalized logistic regressions to predict the up and down trends of Google's stock price. The prediction results show that the MCP penalized logistic regression with technical indicators is superior to the other prediction methods such as logistic regression, the other four penalized logistic regressions, SVM and ANN. For other stock price trends prediction problems, we can also apply statistical charts, data analysis, empirical knowledge and the penalized method to extract some important technical indicators that may affect stock price trends movement, establish some penalized logistic regressions with the different technical indicators to predict these stock price trends, and apply the corresponding confusion matrix and ROC curves to assess the prediction accuracy. Therefore, here we combine technical indicators with MCP penalized logistic regressions and provide the effective method to improve the prediction accuracy.

References

[1] Breheny P, Huang J (2011) Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. Annals of Applied Statistics 5(1):232-253.

[2] Cavalcante R C, Brasileiro R C, Souza V L F, Nobrega J P, Oliveira A L I (2016) Computational intelligence and financial markets: a survey and future directions. Expert Systems with Applications 55(15):194-211.

[3] Elliott G, Granger C, Timmermann A (2013) Handbook of economic forecasting. North HollandElsevier.

[4] Hu X M, Liu F (2020) Estimation theory and model recognition of high-dimensional statistical models. Beijing: Higher Education Press.

[5] Hu X M, Jiang H F (2021) Logistic regression model with technical indicators predicts ups and downs for google stock prices. System Science and Mathematics 41(3):1-22.

[6] Khan W, Malik U, Ghazanfar M A, Azam M A, Alyoubi K, Alfakeeh A (2020) Predicting stock market trends using machine learning algorithms via public sentiment and political situation analysis. Soft Computing 24(15):11019-11043.

[7] Murphy J J (1999) Technical analysis of the financial markets. New YorkPrentice Hall Press.

[8] Nabipour M, Nayyeri P, Jabani H, Shahab S, Mosavi A (2020) Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis on the Tehran stock exchange. IEEE Access 99(8):150199-150212.

[9] Nair B, Sai S G, Naveen A N, Lakshmi A, Venkatesh G S, Mohandas V (2011) A ga-artificial neural network hybrid system for financial time series forecasting. Information Technology and Mobile Communication 147(2):499-506.

[10] Shen J Y, Shafiq M O (2020) Short-term stock market price trend prediction using a comprehensive deep learning system. Journal Of Big Data 7(1):66-98.

[11] Wang L, Zhu J (2010) Financial market forecasting using a two-step kernel learning method for the support vector regression. Annals of Operations Research 174(2):103-120.

[12] Wen M, Li P, Zhang L F, Chen Y (2019) Stock market trend prediction using high-order information of time series. IEEE Access 7: 28299-28308.

[13] Zhang J, Cui S C, Xu Y (2018) A novel data-driven stock price trend prediction system. Expert Systems with Applications 97(1):60-69.

Statements & Declarations

Funding This research was supported by the Fifth Batch of Excellent Talent Support Program of Chongqing Colleges and University (68021900601), the Natural Science Foundation of CQ CSTC (2018jcyjA2073), Science and Technology Research Program of Chongqing Education Commission (KJZD-M202100801), the Program for the Chongqing Statistics Postgraduate Supervisor Team (yds183002), Chongqing Social Science Plan Project (2019WT592020YBTJ102), Open Project from Chongqing Key Laboratory of Social Economy and Applied Statistics (KFJJ2018066) and Mathematic and Statistics Team from Chongqing Technology and Business University (ZDPTTD201906).

Competing Interests The author declares that they have no relevant financial or non-financial interests to disclose.

Author Contributions Xuemei Hu provided the basic idea and improve the writing to the manuscript. Huifeng Jiang collected data, provided the figures and tables, and finished the basic writing. Hong Jia improved the program.

Data Availability The datasets analyzed during the current study are available in the Yahoo Finance, uk.finance.yahoo.com.

Ethical approval This article does not contain any studies with human participants or animals performed by the author.