# A survival analysis method for stock market prediction

**Conference Paper** · October 2015

DOI: 10.1109/BESC.2015.7365968

CITATIONS
4

READS
3,168

**5 authors**, including:

Guangliang Gao
Nanjing University of Science and Technology
**10** PUBLICATIONS **204** CITATIONS
SEE PROFILE

Zhan Bu
Nanjing University of Finance and Economics
**51** PUBLICATIONS **1,071** CITATIONS
SEE PROFILE

Zhiang Wu
Nanjing Audit University
**122** PUBLICATIONS **1,214** CITATIONS
SEE PROFILE

# A Survival Analysis Method for Stock Market Prediction

Guangliang Gao[‡], Zhan Bu[*†], Lingbo Liu[†] Jie Cao[†], and Zhiang Wu[†]
[‡]*School of Computer Science and Engineering,*
*Nanjing University of Science and Technology, Nanjing, China*
[†]*Jiangsu Provincial Key Lab. of E-Business, Nanjing University of Finance and Economics, China*
[*]*Corresponding author: buzhan@nuaa.edu.cn*

***Abstract*—Stock market prediction focus on developing approaches to determine the future price of a stock or other financial product. The key task of stock market prediction is to determine the timing for the buying or selling of stock, undoubtedly, it is very difficult due to the high volatility and nonlinear relationships driven by short-term fluctuations in investment demand. In this work, we address this problem by adopting the Cox's hazard model to predict a stock's future rising or dropping probabilities. Specifically, we define the problem of Buy-and-Sell-Point Prediction from the survival analysis perspective. The Cox's hazard model is proposed as the model of choice for this prediction problem due to several reasons including the ability to model the dynamics in the stock movement, and to easily incorporate different types of technical indexes as covariates. In the experiment, we apply the trained model for the stock market forecasting on six stocks in Shanghai Stock Exchange. The results show that the proposed model is superior to several baseline models in terms of accuracy, and the stock return evaluations have revealed that the profits produced by the proposed model are higher.**

***Keywords*-Stock Market Prediction; Buy-and-Sell-Point Prediction; Survival Analysis; Hazard Model; Technical Index**

## I. INTRODUCTION

Sock market prediction is regarded as a challenging task due to the high volatility and nonlinear relationships, driven short-term fluctuations in investment demand. Some researchers even found that many standard econometric models are unable to produce better predictions than the random walk model [1], which has also encouraged researchers to develop more predictable models.

In the filed of stock market forecasting, most early models were depended on conventional statistical methods such as the time-series models and multivariate analysis [2], [3], [4], [5]. In these methods, the stock movement was modeled as a function of time series and was solved as a regression problem. However, stock prices are difficult to predict due to their chaotic nature. Furthermore, there are some assumptions about the variables used in statistical methods, which may not be suitable for those datasets that do not follow the statistical distributions.

Recent studies reveal that the computational intelligence methods are able to simulate the volatile stock markets well and produce better predictive results than traditional statistical methods [6], [7], [8]. Of various computational intelligence methods, the artificial neural networks (ANNs) are considered as a class of strong alternatives to predicting stock price movement, because ANNs are particularly well suited for finding accurate solutions in an environment characterized by complex, noisy, or partial information [9]. Furthermore, ANNs can map any nonlinear function without any a priori assumption about the data [10]. However, most computational intelligence methods are black-box methods, and the rules mined from it are not easily understandable.

Besides the statistical methods and the computational intelligence methods, technical analysis is one of popular approaches used by investors to make investment decisions, and many researchers have been focusing on technical analysis to increase their investment returns [11], [12], [13], [14]. The technical analysis method assumes that stock price and volume are the two most relevant factors in determining the future direction and behavior of a particular stock or market, and that the technical indicators, coming from a mathematical formula, based on stock price and volume, can be applied to predict price fluctuations and also provide data for investors, enabling them to determine the timing for the buying or selling of stock [11]. Although various technical indicators are proposed to forecast stock market trends, those indicators are mainly based on personal experience, which could result in erroneous judgments of market signals.

To provide a good solution to the above problems, this study proposed a novel survival analysis framework which tracks the stock's future rising or dropping probabilities. More generally, we defined four states for a stock according to its one-day return: a stock with at least $\alpha$ one-day rise or at least $\beta$ one-day drop are considered as two "events" in the survival analysis literature; on the contrary, a stock has no more than $\alpha$ one-day rise or no more than $\beta$ one-day drop are considered as the two survival states during its movement. Then we applied the Cox's hazard model [15] to predict the stock's future rising or dropping probabilities, which can be used as the indicators to determine the timing for the buying or selling of stock. The hazard based models are preferred over the standard regression based methods for this problem due to their ability to model particular aspects of duration data such as censoring. More importantly, the Cox's hazard regression model is used as it can incorporate the effects of covariates. Those covariates are set as some classical technical indexes which are based on the histori-

cally movement of a stock. In the experiment, we applied the model for the stock movement forecasting on six famous stocks in Shanghai Stock Exchange. Two validation indexes, hit rate and return, are considered together to evaluate the performance of the trained model. The results showed that the proposed model is superior to several baseline models in both the accuracy and return rate.

## II. Buy-and-Sell-Point Prediction (BSPP)

The successful prediction of a stock's future price could yield significant profit, since the first stock market opened, numerous forecasting methods have been employed. However, economic environments and political situations both affect stock price variations which make it even more difficult for researchers and investors to forecast stock market. In this work, we adopt a unique methodology for analyzing the dynamic stock market by directly modeling the future rising or dropping probabilities.

### A. Problem Statement

Let $\gamma_t = \frac{P_t - P_{t-1}}{P_{t-1}} * 100\%$ to be one-day return, here $P_t$ is the closing price at time $t$. Given two marks of $\gamma_t$: $\alpha$ and $\beta$, which hold that $0 \leq \alpha \leq 10\%$, $-10\% \leq \beta \leq 0$. We define four possible states $\{\mathcal{R}_\alpha, \check{\mathcal{R}}_\alpha, \mathcal{D}_\beta, \check{\mathcal{D}}_\beta\}$ for each stock as follows:

- $\mathcal{R}_\alpha$ state: if a stock has at least $\alpha$ one-day rise;
- $\check{\mathcal{R}}_\alpha$ state: if a stock has no more than $\alpha$ one-day rise;
- $\mathcal{D}_\beta$ state: if a stock has at least $\beta$ one-day drop;
- $\check{\mathcal{D}}_\beta$ state: if a stock has no more than $\beta$ one-day drop.

Given the current time point $T_c$, the problem of buy-and-sell-point prediction can be converted to estimate the future rising or dropping probabilities at time point $T_c$, which can be formally defined as follows:

**Definition 1** (Buy-and-Sell Point Prediction)**.** *Let $T_\alpha$ to be the last time the stock was in $\mathcal{R}_\alpha$ state, and $T_\beta$ to be the last time the stock was in $\mathcal{D}_\beta$ state. The problem of buy-and-sell-point prediction is to predict the future $\alpha$-rising or $\beta$-dropping probabilities, $p_\alpha(T_c - T_\alpha)$ and $p_\beta(T_c - T_\beta)$ as shown in Fig. 1, where $T_c - T_\alpha$ is the time the stock has already been in $\check{\mathcal{R}}_\alpha$ state, and it has already been in $\check{\mathcal{D}}_\beta$ state for $T_c - T_\beta$ time. Given a divergence threshold $\theta$, the trading strategy is as follows:*
***Buy** stock with open price if $p_\alpha(T_c - T_\alpha) - p_\beta(T_c - T_\beta) \geq \theta$;*
***Sell** stock with close price if $p_\beta(T_c - T_\beta) - p_\alpha(T_c - T_\alpha) \leq \theta$.*

### B. A Stock's Survival Analysis

Survival analysis [16], [15] is a branch of statistics that deals with analysis of time duration until one or more events happen, such as death in biological organisms and failure in mechanical systems. In this work, we define the problem of Buy-and-Sell-Point Prediction from a survival analysis perspective. More generally, survival analysis involves the modelling of time to event data; in the context of stock

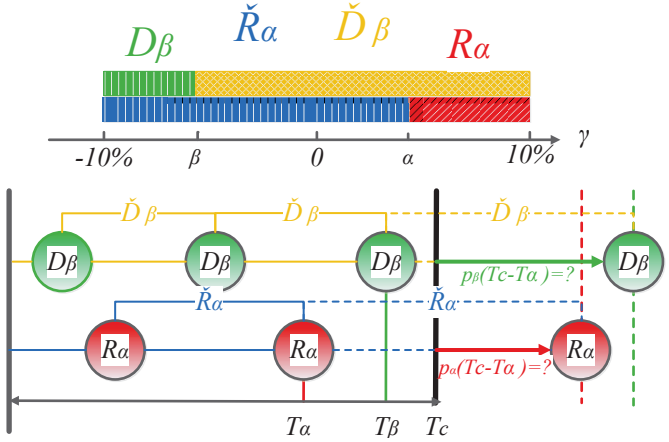


Figure 1: Problem Statement of BSPP.

market forecast, a stock with at least $\alpha$ one-day rise ($\mathcal{R}_\alpha$) or at least $\beta$ one-day drop ($\mathcal{D}_\beta$) are considered as two "events" in the survival analysis literature. Therefore, we attempt to answer questions such as: what is the proportion of a stock which will stay in states $\check{\mathcal{R}}_\alpha$ or $\check{\mathcal{D}}_\beta$ past a certain time? and at what rate will the stock price rise or fall?

The object of primary interest is the survival function, conventionally denoted $\mathcal{S}$, which is defined as:

$$\mathcal{S}(t) = Pr(T > t), \tag{1}$$

where $t$ is elapsed time, $T$ is a random variable denoting the time of $\mathcal{R}_\alpha$ or $\mathcal{D}_\beta$, and $Pr$ stands for probability. In this work we use days as the unit of time for our analysis. That is, the survival function is the probability that the time of an event is later than some specified time $t$. The hazard function, conventionally denoted $\lambda$, is defined as the event rate at time $t$ conditional on survival until time $t$ or later (that is, $T \geq t$):

$$\lambda(t) = \lim_{dt \to 0} \frac{Pr(t \leq T \leq t + dt)}{dt \cdot \mathcal{S}(t)} = \frac{f(t)}{\mathcal{S}(t)} = -\frac{\mathcal{S}'(t)}{\mathcal{S}(t)}. \tag{2}$$

Given a set of covariates $X(t) = (x_1(t), \cdots, x_p(t))$, note that the covariates may be static or may vary with time, the hazard function can be written as a function of an underlying hazard function and a covariant function:

$$\lambda(t|X(t)) = \lambda_0(t)g(X(t)) = \lambda_0(t)g(x_1(t), \cdots, x_p(t)). \tag{3}$$

The underlying hazard function, $\lambda_0(t)$, represents how the risk changes with time, and $g(X(t))$ represents the effect of variables/covariates. $\lambda_0(t)$ can be interpreted as the hazard function when all covariates are ignored, and is also called the baseline hazard function. The Cox's proportional hazard model [15] is just one of several approaches that attempts to evaluate survival curves taking into account other covariates that may effect the survival, which assumes that $g(X(t))$ is an exponential function of the covariates, that is,

$$g(X(t)) = \exp \sum_{j=1}^{p} b_j x_j(t), \tag{4}$$

and the hazard function is

$$\lambda(t|X(t)) = \lambda_0(t) \exp \sum_{j=1}^{p} b_j x_j(t) = \lambda_0(t) \exp BX(t)', \tag{5}$$

where $B = (b_1, \cdots, b_p)$ denotes the coefficients of covariates, which can be estimated from the data observed (e.g., the historical stock trend), and indicate the magnitude of the effects of their corresponding covariates.

With Eq. 2, the hazard function can alternatively be represented in terms of the cumulative hazard function:

$$\Lambda(t|X(t)) = \sum_{u=0}^{t} \lambda(u|X(u))du, \tag{6}$$

$$\mathcal{S}(t|X(t)) = \exp(-\Lambda(t|X(t))). \tag{7}$$

For the two "events" of a stock ($\mathcal{R}_\alpha$ and $\mathcal{D}_\beta$), we denote the corresponding survival functions as $\mathcal{S}_\alpha(t)$ and $\mathcal{S}_\beta(t)$, the corresponding cumulative hazard functions as $\Lambda_\alpha(t|X(t))$ and $\Lambda_\beta(t|X(t))$, and the corresponding hazard functions as $\lambda_\alpha(t|X(t))$ and $\lambda_\beta(t|X(t))$. Furthermore, given the current time point $T_c$, the last time the stock was in $\mathcal{D}_\beta$ state $T_\alpha$, and the last time the stock was in $\mathcal{D}_\beta$ state $T_\beta$, the future $\alpha$-rising or $\beta$-dropping probabilities can be computed as follows:

$$\begin{aligned} p_\alpha(T_c - T_\alpha) &= 1 - \mathrm{S}_\alpha(T_c - T_\alpha) \\ &= 1 - \exp(- \sum_{u=1}^{T_c - T_\alpha} \lambda_\alpha(u|X(u))du) \end{aligned} \tag{8}$$

$$\begin{aligned} p_\beta(T_c - T_\beta) &= 1 - \mathrm{S}_\beta(T_c - T_\beta) \\ &= 1 - \exp(- \sum_{u=1}^{T_c - T_\beta} \lambda_\beta(u|X(u))du) \end{aligned} \tag{9}$$

## III. Experimental Validation

In this section, we illustrate the use of the Cox's proportional hazard model for predicting financial market movements based on capturing the potential buying and selling points. A set of candidate stocks was selected based on the following criteria: (1)capital size; (2)monthly sales; (3) earnings per shares(EPS); (4)transaction volume per day; and (5)marginal accounts. Based on these factors, this study selected six famous stocks in Shanghai Stock Exchange (SHSE) for testing, including China Petrochemical Corporation (SINOPEC), China National Petroleum Corporation (CNPC), Agricultural Bank of China (ABC), Ping An Insurance (PINGAN), CRRC Corporation Limited (CRRC), and China State Construction Engineering Corporation Limited (SCSEC). The data set used includes daily closing prices and turnover rate from January 2007 to June 2015, and the prices are decoded into returns $\gamma_t$.

### A. Data Processing

During the time window from January 2007 to June 2015, the duration time a stock in $\check{\mathcal{R}}_\alpha$ or $\check{\mathcal{D}}_\beta$ states are collected. Each duration time observation can be associated with a set of covariates influencing its magnitude. Hence, the data can be represented as a set of tuples: $< T, X(T), tag >$, where $T$ is the duration time observation, $X(T)$ is the vector of covariates associated with that observation and $tag$ is the status variable. To predict the future $\alpha$-rising probability, $tag$ is set to 0 when the time variable represents the actual observation of duration time (the "event" $\mathcal{R}_\alpha$ has happened) whereas it is set to 1 when the time variable represents a censored observation.

In this work, we selected several technical indexes as covariates which are related to the typical visitation patterns of a stock for the Buy-and-Sell-Point prediction problem. Such covariates seek to predict the future stock price movement based on how their visitation behavior has been historically. For example, a stock will become vulnerable to a sell off after the market has moved for a substantial period of time. Therefore, we constructed the following covariates based on stock activities observed from the last "event" time to current observation time.

- Accumulated Rate of Change (AcROC). This covariate is defined as the cumulative gains from the time point when the latest "event" happened to the current time point: $x_1(T) = \sum_{t=T_c-T}^{T_c-1} \gamma_t$.
- Average Rate of Change (AvROC). This covariate is defined as the average gains from the time point when the latest "event" happened to the current time point: $x_2(T) = \frac{\sum_{t=T_c-T}^{T_c-1} \gamma_t}{T}$.
- Accumulated Turnover (AcT). This covariate is used as the cumulative turnover rate from the time point when the latest "event" happened to the current time point: $x_3(T) = \sum_{t=T_c-T}^{T_c-1} \mu_t$, where $\mu_t$ is the turnover rate at time point $t$.
- Average Turnover (AvT). This covariate is used as the average turnover rate from the time point when the latest "event" happened to the current time point: $x_4(T) = \frac{\sum_{t=T_c-T}^{T_c-1} \mu_t}{T}$.
- Stochastic $K\%$ ($K\%$). This covariate refers to the point of a current price in relation to its price range over a period of time: $x_5(T) = \frac{P_{T_c-1} - LL_T}{HH_T - LL_T} * 100\%$, where $HH_T$ and $LL_T$ mean lowest low and highest high in the last $T$ days, respectively.
- Stochastic $D\%$ ($D\%$). This covariate measures the average $K\%$ over the last $n$ days: $x_6(T) = \frac{\sum_{t=T_c-T}^{T_c-1} K_t\%}{T}$.
- Stochastic $J\%$ ($J\%$). This covariate is actually a derived form of the Stochastic with the only difference being an extra line: $x_7(T) = 3K_{T_c-1}\% - 2D_{T_c-1}\%$.

- Relative strength index (RSI). This covariate is intended to chart the current and historical strength or weakness of a stock on the closing prices of a recent trading period: $x_8(T) = 100 - \frac{100}{1+RS}$, where $RS = \frac{\sum_{t=T_c-T,\gamma_t\geq 0}^{T_c-1} \gamma_t}{\sum_{t=T_c-T}^{T_c-1} \gamma_t}$.
- Psychological Line (PSY). The covariate measures the ratio of the number of rising periods over the total number of periods. It reflects the buying power in relation to the selling power: $x_9(T) = \frac{\sum_{t=T_c-T}^{T_c-1} I(\gamma_t)}{T}$, where $I(\gamma_t) = 1$ if $\gamma_t \geq 0$ and 0 otherwise.

### B. Experimental Setup

To determine which covariates are the key driver of stock price movement, we use the standard survival package in SPSS for estimating the Cox's model. Note that the model can be broken down into two factors. The first factor represents the effect of the covariates on the hazard rate, which is independent of the baseline hazard function and can be learnt by maximizing the partial likelihood. Once the regression coefficients have been learnt, the non-parametric form of the baseline hazard function is estimated using KaplanCMeier estimator [17]. In the experiment, to maintain a reasonable size of training data, we set $\alpha$ and $\beta$ to be $1\%$, and $-1\%$ respectively. Furthermore, to show the robustness of the the Cox's hazard model, we use a $k$-fold CV estimation procedure to perform independent experiments for each stock. In the $k$-fold CV estimation procedure, each training data is divided into $k$ non-overlapping groups. We train two Cox's models (one for the buy point and the other for the sell point) using the first $k-1$ groups of training data and test the trained models on the $k$-th group. We repeat this procedure until each of the groups is used as a test set once. We then take the average of the performance measurements over the folds. In the experiment, $k$ is set to 5. This is a reasonable compromise considering the computational complexity and modeling robustness. Therefore, an estimate from fivefold CV is likely to be more reliable than an estimate from a common practice only using a single testing set.

When the ground-truth of stock movement is known, we utilize two factors, i.e., `Accuracy` and `Return`, together to evaluate the performance of the trained model. Accuracy measures the percentage of correct predictions of the model which can be computed as follows:

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

where TP, TN, FP and FN represent true positive, true negative, false positive and false negative, respectively. We treat the upward trend cases as the positive class here. As for our method, accuracy is calculated according to the suggested buy or sell points.

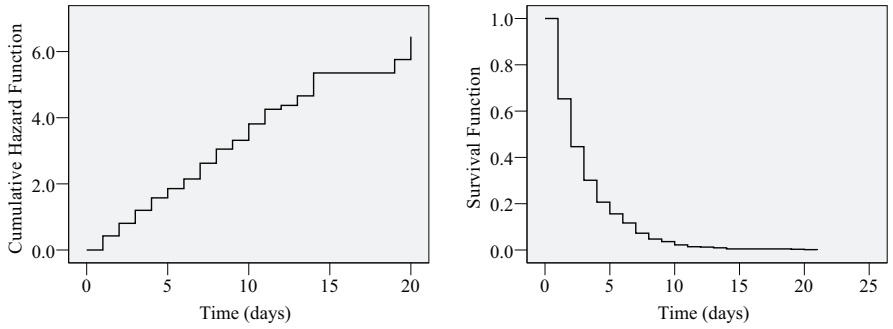
We analyze the return gained by an investor who uses the predictive outcomes of each approach to trade the stocks. The trading strategy adopted by an investor is as follows: if an approach forecasts an upward trend, the investor takes a buy action; if there is a downward trend from the forecasting, a sell action is taken; otherwise, s/he holds the stock till next sell point or waits for the next buy point. With this trading strategy, the stock holding period can be expressed as a periodic sequence: $(T_{B1}, T_{S1}), (T_{B2}, T_{S2}), \cdots, (T_{Bl}, T_{Sl})$, thus the return can be computed as:

$$Return = \prod_{i=1}^{l} \prod_{T_{Bi}}^{T_{Si}} (1 + \gamma_t) - 1. \tag{11}$$

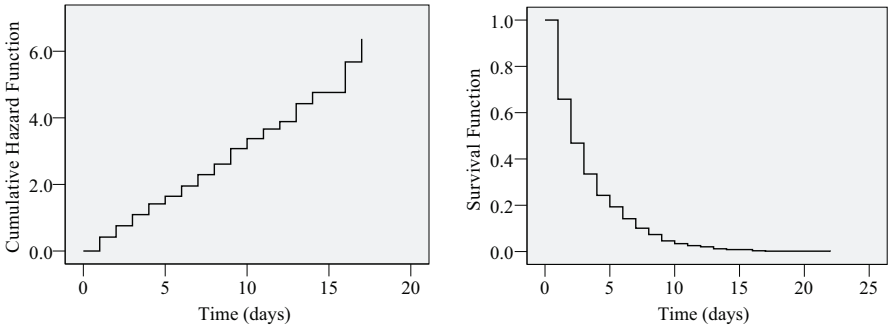
### C. Model Parameters

In this subsection, we analyze the results of the experimental evaluation of the Cox's proportional hazard model. We only discuss the parameters of model trained on China Petrochemical Corporation (SINOPEC) stock data from January 2007 to June 2015. By setting $\alpha = 1\%$ and $\beta = -1\%$, two models are respectively trained, one is for the future rising prediction (the rise model) and the other is for the future dropping estimation (the drop model). The importance of the covariates for the BSPP problem can be assessed using different importance indicates as shown in Tables I and II. Note that the regression coefficients and the significance score for the covariates are directly obtained from the output of SPSS for fitting the Cox's model. The regression coefficient tells us how much a unit change in the value of the covariate impact a stock's future movement. In the default setting, the value of the coefficient was statistically significant at a significance level of $0.05$. With reference to Tables I and II, we find that the key determinants for the rising and dropping predictions are slightly different expect AcROC, AvROC, AcT, AvT, RSI and PSY. For the rise model, most of the covariates associated with the typical patterns of visitation (AcROC, AvROC, AcT, AvT, $K\%$, RSI and PSY) are found to be highly significant for predicting the future rising probability variable. While for the drop model, AcROC, AvROC, AcT, AvT, $D\%$, $J\%$, RSI and PSY can be considered to be some key drivers for the downward trend forecasting. For each significant covariate, we also computed its average influence on the baseline hazard function ($\exp(b\bar{x})$). This provided an average score for how much the covariate impacted the magnitude of the baseline hazard function. As shown in the last columns in Tables I and II, RSI and PSY impacted the stock future movement the most on an average for SINOPEC dataset.

Fig. 2 shows the baseline cumulative hazard functions and the survival functions for both rise and drop models, from which we noticed that the survival functions have a sharply declining shape typical of processes exhibiting inertia. Hence, the longer a stock stays in $\check{\mathcal{R}}_\alpha$ or $\check{\mathcal{D}}_\alpha$ state the more likely $\mathcal{R}_\alpha$ or $\mathcal{D}_\alpha$ will happen in the future. For example, in the rise model, the survival function has a value of 0.001 at 20 days. This suggest that a stock will have $\alpha$-




(a) Baseline cumulative hazard function for the rise model (b) Survival function for the rise model




(c) Baseline cumulative hazard function for the drop model (d) Survival function for the drop model

Figure 2: The baseline cumulative hazard function and the survival function computed on SINOPEC training dataset.

Table I: Covariate importance indicators for the rise model with $\alpha = 1\%$.

| Covariates | Coefficient | Significance | Average $\bar{x}$ | $\exp(b\bar{x})$ |
|---|---|---|---|---|
| AcROC | 10.80 | **0.000** | $2.24e^{-3}$ | 1.025 |
| AvROC | 11.90 | **0.031** | $1.10e^{-3}$ | 1.013 |
| AcT | -2.60 | **0.011** | $6.95e^{-3}$ | 0.982 |
| AvT | 7.90 | **0.001** | $2.77e^{-3}$ | 1.022 |
| $K\%$ | 1.25 | **0.000** | $4.92e^{-1}$ | 1.848 |
| $D\%$ | - | 0.925 | $5.95e^{-1}$ | - |
| $J\%$ | - | 0.932 | $2.84e^{-1}$ | - |
| RSI | -2.21 | **0.000** | $2.63e^{-1}$ | 0.559 |
| PSY | -10.13 | **0.000** | $1.32e^{-1}$ | 0.263 |

Table II: Covariate importance indicators for the drop model with $\beta = -1\%$.

| Covariates | Coefficient | Significance | Average $\bar{x}$ | $\exp(b\bar{x})$ |
|---|---|---|---|---|
| AcROC | 45.20 | **0.000** | $1.92e^{-3}$ | 1.091 |
| AvROC | -45.20 | **0.000** | $-1.10e^{-2}$ | 1.051 |
| AcT | -2.60 | **0.003** | $7.56e^{-2}$ | 0.822 |
| AvT | 10.60 | **0.000** | $2.28e^{-2}$ | 1.273 |
| $K\%$ | - | 0.100 | $5.21e^{-1}$ | - |
| $D\%$ | -0.88 | **0.002** | $4.15e^{-1}$ | 0.694 |
| $J\%$ | 0.45 | **0.000** | $7.32e^{-1}$ | 1.390 |
| RSI | -3.86 | **0.000** | $2.83^{e}-1$ | 0.335 |
| PSY | -9.09 | **0.000** | $1.43e^{-1}$ | 0.273 |

rising with $1 - 0.1\% = 99.9\%$ probability if it has already stayed in $\check{\mathcal{R}}_\alpha$ for 20 days.

### *D. Prediction results*

To evaluate our approach, we take the following methods which are either typically used in financial markets:

- ARIMA [3]: This is a statistical method for analyzing and building a forecasting model which best represents a time series by modeling the correlations in the data. we use it as a baseline method.
- Logistic [5]: We use this approach with indicators for the stock movement prediction.
- ANN [18]: We use the back-propagation algorithm with indicators for different stocks to train the model.

The results of `Accuracy` and `Return` are reported in Table III. Note that the values in Table III are the average prediction performance over the fivefold CV experiments. From both technique and business perspectives, the baseline methods of ARIMA and Logistic do not achieve a good performance except on the dataset CRRC, this is because both ARIMA and Logistic are built on stationary data, and pays no attention on the underlying complex hidden interactions between the different technical indexes. As dataset CRRC follows the linear distribution in the observation time window, both the two models perform well. However, most financial markets, especially the stock markets, are not linear, the statistical models can not produce significant instructions for stock exchanging. ANN outperforms ARIMA and Logistic, this is because it constructs predictions on

Table III: Performance of comparative methods in Chinese stock market.

| Model | Accuracy/Return | | | | | |
|---|---|---|---|---|---|---|
| | SINOPEC | CNPC | ABC | PINGAN | CRRC | SCSEC |
| ARIMA | 0.51/49.50% | 0.50/-40.30% | 0.57/-0.60% | 0.49/17.50% | 0.52/**590.50**% | 0.55/73.20% |
| Logistic | 0.54/46.40% | **0.54**/-33.20% | **0.61**/9.40% | 0.55/-2.30% | 0.54/589.70% | 0.57/74.50% |
| ANN | 0.53/56.26% | 0.52/20.22% | 0.52/31.47% | 0.53/36.89% | 0.52/333.12% | **0.57**/64.59% |
| Our model | **0.55**/**76.44**% | 0.53/**108.91**% | 0.52/**112.82**% | **0.57**/**223.00**% | **0.55**/278.93% | 0.56/**152.64**% |

the hidden coupled features. Our hazard model nearly outperforms all baselines regardless of technique or business perspective. This can be interpreted as follows: firstly, unlike those methods that predict market movements directly from the observations, the proposed model builds a survival analysis framework to learn the hidden features which removes the vulnerabilities of observations; secondly, it trains two models, one is rise model the other is drop model, which serve as the key factors driving market dynamics.

## IV. Related Work

The stock market is a complex system where decision-making can be very difficult, many researchers have used different methods to predict stock future movement. This section briefly reviews three main research methods used in stock market prediction.

Many researchers try to use statistical tools [2], [3], [4], [5] to predict stock price. The most popular methods include the regression model, such as the GARCH model [2], the ARIMA model [3], and the probabilistic model [4]. In ref. [20], GARCH model is used to test exchange rate currencies, the work shows that the output of GARCH model is similar to the traditional regression model and random walk strategy. ARIMA model combines the moving average stock price and also find similar patterns [3]. The probabilistic model offers an alternative to predict the stock movement. In the work of Bao and Yang [4], a new learning strategy combined with the probabilistic model was proposed to find stock data critical buy-and-sell point, and then a Markovian network is used to find the best trading signal probability. Although the statistical tools described above show some interesting results, they do not produce significant instructions for stock exchanging. A good stock market prediction method should consider not only stock price variations, but also other information which can help investors "buy low, sell high".

Computational intelligence methods are also popular in stock market forecasting [6], [7], [8], [21]. Recently, Artificial Neural Network (ANN) has gained wide attention, especially in problems whose solution spaces are so complex and large. The application of ANN for stock market prediction has its advantages and disadvantages. For example, Ref. [18] showed that ANN could be used successfully for short-term series prediction. The application of an ANN is especially effective when the link between the independent and dependent variables from stock price variation is nonlinear and very noisy, which is typical of a stock market. The main drawback of the computational intelligence methods is that they do not provide any insight into the underlying processes, and prevent us from obtaining a specific collection of rules. Therefore, decision-making that relies solely on ANN results is not advisable.

Technical index analysis is based on historically formed regularities in the stock exchange, and assumes that the same result will repeat in the future. It explores internal market information and assumes that all the necessary factors are in the stock exchange information [12]. Traditionally, Technical index analysis proposes a set of investment rules for the investor. However, the use of individual trading rules is not very effective from the investor's point of view [13], [14]. Technical analysis can help extract financial information from stock price pattern. However, this method must be combined with other methods to help investors make accurate decisions. Recently, many researchers have used computational intelligence tools in combination with technical analysis to better determine trading signals [22], [23]. To determine the exact moment for stock trading, analysts develop a group of technical rules based on technical indicators. The aim of each rule is to generate either a buy signal when a bull market is anticipated or a sell signal when a bear market is expected. The proposed method in this paper is still along this line, in which the sock movement is considered from the survival analysis perspective. Furthermore, different types of technical indexes are incorporated in the model.

## V. Conclusion

In this work, we proposed a survival analysis framework for the stock market forecasting. For each stock, we defined four possible states according to its one-day rise. We suggest that the future rising and dropping probabilities prediction can directly address the heart of the problem for buy-and-sell point detection. To facilitate such efforts, we formulated the problem of future rising and dropping probabilities prediction and applied several technical indexes as relevant covariates to the problem. The Cox's hazard model was proposed as the model of choice for this prediction problem due to several reasons including the ability to handle dynamics of stock price movement and to incorporate different types of covariates. The performance of the proposed model was found to surpass several selected baseline models in both the accuracy and return rate. In our future work, the proposed Cox's hazard model can be further accommodated

with several covariates, e.g., macroeconomic indexes, company fundamental information and public investment mood detected from online social network, which might enhance the prediction performance.

## ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China under Grant 71372188, Grant 61103229, Grant 61472079, and Grant 61502222, in part by the National Center for International Joint Research on E-Business Information Processing under Grant 2013B01035, in part by the National Key Technologies Research and Development Program of China under Grant 2013BAH16F01, in part by Industry Projects in Jiangsu ST Pillar Program under Grant BE2012185, in part by the Key/Surface Project of Natural Science Research in Jiangsu Provincial Colleges and Universities under Grant 12KJA520001, Grant 14KJA520001, and Grant 14KJB520015, in part by Natural Science Foundation of Jiangsu Province under Grant BK20150988, and in part by educational reformation funds of Nanjing University of Finance  Economics under Grant Project JGZ1504.

## REFERENCES

[1] A.-S. Chen and M. T. Leung, "Regression neural network for error correction in foreign exchange forecasting and trading," *Computers & Operations Research*, vol. 31, no. 7, pp. 1049–1068, 2004.

[2] R. Gencay, "Linear, non-linear and essential foreign exchange rate prediction with simple technical trading rules," *Journal of International Economics*, vol. 47, no. 1, pp. 91–107, 1999.

[3] A. Timmermann and C. W. Granger, "Efficient market hypothesis and forecasting," *International Journal of Forecasting*, vol. 20, no. 1, pp. 15–27, 2004.

[4] D. Bao and Z. Yang, "Intelligent stock trading system by turning point confirming and probabilistic reasoning," *Expert Systems with Applications*, vol. 34, no. 1, pp. 620–627, 2008.

[5] J. R. Stock and D. M. Lambert, *Strategic logistics management*. McGraw-Hill/Irwin Boston, MA, 2001, vol. 4.

[6] P.-C. Chang, C.-H. Liu, and C.-Y. Fan, "Data clustering and fuzzy neural network for sales forecasting: A case study in printed circuit board industry," *Knowledge-Based Systems*, vol. 22, no. 5, pp. 344–355, 2009.

[7] C.-H. Cheng, T.-L. Chen, and L.-Y. Wei, "A hybrid model based on rough sets theory and genetic algorithms for stock price forecasting," *Information Sciences*, vol. 180, no. 9, pp. 1610–1629, 2010.

[8] L.-J. Cao and F. E. Tay, "Support vector machine with adaptive parameters in financial time series forecasting," *Neural Networks, IEEE Transactions on*, vol. 14, no. 6, pp. 1506–1518, 2003.

[9] G. Grudnitski and L. Osburn, "Forecasting s&p and gold futures prices: An application of neural networks," *Journal of Futures Markets*, vol. 13, no. 6, pp. 631–643, 1993.

[10] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.

[11] S.-C. Chi, W.-L. Peng, P.-T. Wu, and M.-W. Yu, "The study on the relationship among technical indicators and the development of stock index prediction system," in *Fuzzy Information Processing Society*. IEEE, 2003, pp. 291–296.

[12] G. S. Atsalakis and K. P. Valavanis, "Surveying stock market forecasting techniques–part ii: Soft computing methods," *Expert Systems with Applications*, vol. 36, no. 3, pp. 5932–5941, 2009.

[13] M. A. Dempster, T. W. Payne, Y. Romahi, and G. W. Thompson, "Computational learning techniques for intraday fx trading using popular technical indicators," *IEEE Transactions on neural networks*, vol. 12, no. 4, pp. 744–754, 2001.

[14] A. W. Lo, H. Mamaysky, and J. Wang, "Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation," National bureau of economic research, Tech. Rep., 2000.

[15] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 187–220, 1972.

[16] C. McGilchrist and C. Aisbett, "Regression with frailty in survival analysis," *Biometrics*, pp. 461–466, 1991.

[17] M. Hollander, D. A. Wolfe, and E. Chicken, *Nonparametric statistical methods*. John Wiley & Sons, 2013.

[18] A. F. De Souza, F. D. Freitas, and A. G. Coelho de Almeida, "Fast learning and predicting of stock returns with virtual generalized random access memory weightless neural networks," *Concurrency and Computation: Practice and Experience*, vol. 24, no. 8, pp. 921–933, 2012.

[19] L. Yu, H. Chen, S. Wang, and K. K. Lai, "Evolving least squares support vector machines for stock market trend mining," *Evolutionary Computation, IEEE Transactions on*, vol. 13, no. 1, pp. 87–102, 2009.

[20] F. X. Diebold and R. S. Mariano, "Comparing predictive accuracy," *Journal of Business & economic statistics*, 2012.

[21] W. Huang, K. K. Lai, Y. Nakamori, S. Wang, and L. Yu, "Neural networks in finance and economics forecasting," *International Journal of Information Technology & Decision Making*, vol. 6, no. 01, pp. 113–140, 2007.

[22] W. Dai, J.-Y. Wu, and C.-J. Lu, "Combining nonlinear independent component analysis and neural network for the prediction of asian stock market indexes," *Expert systems with applications*, vol. 39, no. 4, pp. 4444–4452, 2012.

[23] R. K. Nayak, D. Mishra, and A. K. Rath, "A naïve svm-knn based stock market trend reversal analysis for indian benchmark indices," *Applied Soft Computing*, 2015.