Expert Systems With Applications 173 (2021) 114651

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A data mining framework for financial prediction

Misuk Kim

Department of Data Science, Sejong University, 207 Neungdong-ro, Gwangjin-gu, Seoul 05006, Republic of Korea

ARTICLE INFO

Keywords:
Data mining framework
Financial prediction
Prediction metrics
Feature selection
Prediction modeling

ABSTRACT

In the financial markets, because real-time transactions directly relate to profit, it is important to process and analyze data on a real-time basis. In practice, decisions influenced by experts' experiences from fundamental and technical analysis occur frequently compared to decisions using prediction algorithms. A domain-specific data mining framework was proposed recently to reduce related cost. Therefore, this study proposes a novel data mining framework suitable for financial markets according to expert knowledge. The proposed framework predominantly considers the following three perspectives as the standards for the effectiveness of research: interpretability, proper prediction metrics, and reporting methods. We applied our framework to the real-world financial prediction problems, such as the 3–10 year treasury spread forecasts. Consequently, we achieved an 84% prediction performance on the spread prediction and used hierarchical information to provide additional insight. In addition, we obtained practical knowledge and synergies through extraction of critical variables that can be used as a quick and accurate data-driven decision making support tool by active agents in the real world.

1. Introduction

Through several different channels via diverse means, financial markets play important roles in the well-being of enterprises and the overall economy. A financial market transfers real economic resources, offers dividends or interest to the market participants, creates liquidity, and enables trade among the investors in the market. Therefore, it is important to analyze factors that are linked organically to the financial market and to derive latest information from the huge data available. Many investors use several sources of information to predict target values and develop strategies to gain an edge in competitions. Previous studies have been based on fundamental and technical analysis, such as evaluating the value of a company using the company's performance and reliability, or predicting future target values according to historical data rather than a company's fundamentals. Recently, because of advancements in computing, several data mining models are proposed, and analysis using complex financial data have been attempted. Because financial markets generate big data in real-time, an advantage in competition is the ability to process various modes of data in real-time and characterize the financial market in the form of data-driven guidelines that are deliverable to business decision makers. Therefore, it is important to establish a standardized process capable of deriving data-driven insights that are directly related to performance in the financial markets.

Recently, several studies conducted on data mining frameworks have proposed standardized processes that depict the characteristics of each categorized domain. In particular, data mining frameworks have been proposed for manufacturing processes, such as semiconductor, copper-clad laminate manufacturing, and industrial products with significant manufacturing cost savings (Kim et al., 2017; Kang et al., 2017). In addition, in the construction field, a framework has been developed to extract useful features for smart environment technology (Yu, Fung, & Haghighat, 2013). Therefore, as domain-specific data mining frameworks are being researched, a data mining framework suitable for the financial market is required.

In the financial market, there are three main perspectives for a framework. First, model selection and dimensionality reduction should be performed with caution, because it is important to determine variables that are the drivers of the analysis result; hence, wrapper methods should be used among dimensional reduction algorithms. Second, it is appropriate to use suitable metrics that represent numeric estimates of financial market performance. In some cases, it can be important to approximate the range of a target measure of interest instead of computing exact values. For example, selling a stock at a given time can be based on a range containing the target price rather than the exact amount because of the transaction cost. Thus, the financial market requires a framework that utilizes a measure that indicates the prediction accuracy of the target value, along with an evaluation metric considering the tolerances of the target value. Finally, real-time transactions in the financial market are based on intuition and rapidly produced

E-mail address: misuk.kim@sejong.ac.kr.

https://doi.org/10.1016/j.eswa.2021.114651
Received 21 May 2020; Received in revised form 5 December 2020; Accepted 21 January 2021
Available online 4 February 2021

analysis results. Hence, the presentation of analysis results should include modes of result descriptions such as tables, charts, and diagrams, with detailed interpretations, such as a list of variables that contributed significantly in deriving such results. There may exist a hierarchical relationship among variables, that investors and active traders must be informed prior to making selling/buying decisions.

In this study, we define a novel data mining framework that depicts the above- described characteristics of a financial market. This data mining framework can be used as a guideline for designing prediction models for the financial market, and it can support data-driven, decision-making processes through quantitative and qualitative reporting. Therefore, a data mining framework according to the characteristics of the financial market is defined through a variable selection and prediction model, proper evaluation metrics, and reporting. Thereafter, a proposed data mining framework is applied to predict the financial indicators related to the Korean bond market.

The remainder of this paper is structured as follows: in Section 2, we will present a brief review of literature on domain-specific data mining framework and methodologies for predicting target values in the financial market. Then, in Section 3, we examine the data mining framework for financial markets used in this research. Thereafter, experimental results that apply to real financial data are discussed in Section 4. Finally, we propose future research directions and present the conclusions of this study in Section 5.

2. Literature review

2.1. Domain-specific data mining framework

Recently, data mining frameworks are proposed to solve domain-specific tasks. Kim et al. (2017) proposed a data mining framework of virtual metrology to provide high predictive performance and business-relevant information using data collected through sensors in the manufacturing domain. They identified optimal models by defining accuracy metrics for copper-clad laminate particularly designed for the manufacturing domain using several prediction models and dimensional reductions. From their experimental results, they proposed practical guidelines for monitoring the inputs in lieu of reducing production costs. Furthermore, Kang et al. (2009) detected faulty wafers by applying a similar data mining framework to semiconductor manufacturing. Kang et al. (2017) proposed a data mining process consisting of problem definition, preprocessing, modeling, and visualization, to efficiently analyze faulty industrial products. Fan, Xiao, and Yan (2015) proposed a generic framework for knowledge discovery, that consists of data exploration, data partitioning, knowledge discovery, and post-mining stages. In the process, they extracted hidden knowledge from building operational data stored in building automation systems, such as methods to change operation strategy that detect and diagnose non-typical and abnormal operations and sensor fault occurring. De Silva, Yu, Alahakoon, and Holmes (2011) proposed a data mining framework using incremental summarization and pattern characterization, and actionable knowledge was extracted by applying the framework to actual data measured by electricity meters. A data mining framework using k-modes clustering methodology and association rule mining has been proposed to address difficulties in analyzing accident data that depict the heterogeneous nature of road accident data (Kumar & Toshniwal, 2015). In this framework, incidents were clustered into six groups, and trend analysis for each cluster showed correlations between incidents within the cluster.

2.2. Analysis methods for predicting target value in financial markets

Studies on the analysis methods for predicting target values such as commodity prices, and returns in the financial market are classified into three categories: fundamental analysis, technical analysis, and data mining technologies.

Fundamental analysis evaluates a company's value using the performance and reliability of the company. The desired value is estimated through numerical indicators of the fundamentals of the company. The analysis is performed under the assumption that if a company operates effectively, additional capital is created, and the desired value increases accordingly (Abarbanell & Bushee, 1997). Fundamental analysis is a widely used method because fund managers use the most rational, objective, and publicly available information such as financial statement analysis (Rodríguez-González, 2012). Furthermore, technical analysis, also called chart analysis, predicts future target values using past trends under the assumption that some patterns appear consistently and produce the same results (Leigh, Modani, Purvis, & Roberts, 2002; Gavrilovic & Zimonjic, 2017). Chart patterns such as candle stick pattern, the head and shoulders reversal pattern, and the cup and handle continuation pattern, consider the major concept and aids identifying trading signals of future target value movements. Additionally, there are prediction methods that use several data mining techniques. Recently, as the performance of computer processing has improved and various models have been proposed, research is being conducted to assist investors to find hidden patterns by applying data mining techniques to financial data (Kannan, Sekar, Sathik, & Arumugam, 2010). Models widely used in financial data analysis typically use algorithms such as artificial neural networks (ANNs), support vector machines (SVMs), and decision trees to discover meaningful information in large amounts of financial data (Qiu, Liu, & Wang, 2012; Pan, Xiao, Wang, & Yang, 2017). For example, Chatzis, Siakoulis, Petropoulos, Stavroulakis, and Vlachogiannakis (2018) forecasted the stock market crisis using logistic regression and extreme gradient boosting. In particular, several models that widely use linear models in time-series forecasting, such as the autoregressive moving average (ARMA), autoregressive integrated moving average (ARIMA), and generalized autoregressive conditional heteroscedasticity (GARCH), are integrated with data mining methodologies to forecast the stock market (Herui, Xu, & Yupei, 2015; Rather, Agarwal, & Sastry, 2015). Pai and Lin (2005) used SVMs that perform well on nonlinear regression problems with ARIMA in the stock price forecasting domain. This model improves prediction performance by capturing the complexity of linear and nonlinear structures. Similarly to the work of Pai and Lin (2005) and Wang, Zou, Su, Li, and Chaudhry (2013) predicted three well-known time-series datasets using an ANN and ARIMA hybrid model that extracts nonlinear patterns from time-series data. The financial market is also being actively researched across all markets, but other fine-tuning studies reflect the market and economic characteristics of emerging markets. Mostafa (2010) used MLP and a generalized regression neural network (GRNN) that does not require an estimate of the number of hidden units to forecast the closed price movement of the Kuwait stock exchange and made perfect forecasts and better understood the inner dynamics of the stock market. Xiong and Lu (2017) proposed a hybrid model using ARIMA and a back-propagating neural network and applied this model to the Main Board market and the Growth Enterprise market of the Chinese stock market. They demonstrated the robustness and generalization of the proposed model. In addition to these studies, Preis, Moat, and Stanley (2013) identified online precursors of stock prices from search volume data provided by Google Trends to identify optimal trading strategies, and Moat et al. (2013) studied the relationship 120 between Wikipedia's views change and total stock market movements.

3. Data mining framework for the financial market

In this section, we propose a novel data mining framework that focuses on three perspectives of financial data analysis: interpretation, proper accuracy metrics, and suitable reporting. The overview of the proposed framework is shown in Fig. 1 and is described in 3.1–3.5 for each process.

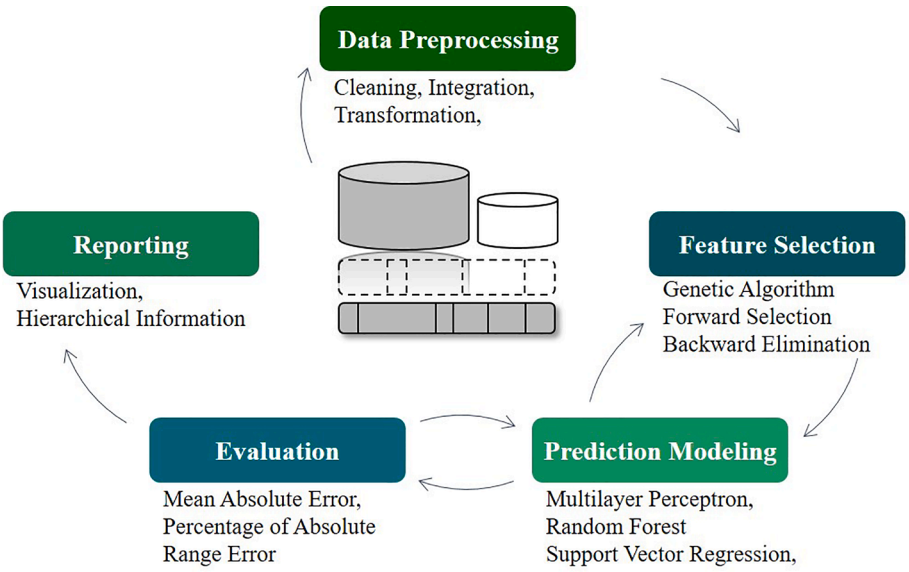

Fig. 1. Data mining framework.

3.1. Data preprocessing

In general, the process of collecting and preprocessing data in data analysis is important and time-consuming. For numerical data, relatively simple preprocessing, such as null value removal, and categorical variable processing is required. However, for textual data, there are somewhat complicated preprocessing such as parsing, stop-words elimination, and tagging. In this section, we discuss in the detail preprocessing of numerical real-word data. Most collected real-world data contains null values. If some variable has several null values, the variable can be removed to enable the analysis, otherwise recorded with the null values deleted or replaced with average values. Furthermore, we regard variables having the same values of all records as no information, we delete from raw data. Finally, we normalize a data to obtain variables in the same range, between 0 and 1.

3.2. Feature selections

Before using a large amount of data in prediction models, we apply dimensionality reduction methodologies to reduce the time complexity of models by deleting some variables. There are many different feature selection methodologies, mainly the filter, wrapper, and embedded methods. The filter approach informs users of the impact of each feature by providing feature scores instead of giving the best feature subset, so it reveals relevance among features. The wrapper approach provides the best subset according to its usefulness by training a prediction model with a subset of features. The embedded methods that combine the filter and wrapper methods, can be built in various forms depending on the use of various linear models. In this study, we use methodologies that are representative of the wrapper approaches, such as genetic algorithm (GA), forward selection (FS), and backward elimination (BE), to find the best subsets, concise models, and promote ease of interpretation. GA, a meta-heuristic optimization algorithm, improves fitness functions by generating a set of candidate solutions for an optimal set of solutions through the evolutionary process of natural selection and genetics. GA simulates the survival of the fittest over successive generations in each iteration and each individual in the population is encoded as a string representing a set of candidate solutions to the problem. The candidate solutions with good fitness values have the opportunity to reproduce from the current generation to the next generation through genetic operations (Shukla & Tiwari, 2011). When using GA for feature selection, each gene in each candidate solution corresponds to a variable and is assigned a value of 0 or 1. A value of 1 means that the variable corresponding to the gene is selected (Yu & Cho, 2006). FS is an iterative method that commences with no variables in the model and provides a subset that produces the best performing results. During the FS iterations, it uses the model-selection statistical criterion to test the appending of each variable, adds the variable that significantly improves the model after appending, and repeats this process until no further improvement is possible. When the procedure stops, there are no other variables that meet the entry criterion. The disadvantage of FS is that the variable cannot be deleted once added. BE begins with all input variables and sequentially removes them. It tests the deletion of each variable using the variable removal criterion statistically and deletes the variable that significantly improves the model after deletion. The iteration stops when the stop criterion is reached; in other words, there are no variables to eliminate as the removal criterion in the candidate subset. The downside of BE is that variables cannot be inserted once removed. The procedure of feature selection algorithms can be summarized as shown in Fig. 2. The optimal subset of variables is obtained by providing feedback on predictive performance until some condition is reached using predicting models.

3.3. Prediction modeling

With the recent advances in various deep learning technologies, rather complex models are demonstrating good performance. However, many financial market participants are non-professionals who do not specialize in deep learning-related skills and, due to both training and inference time aspects, complex models are often not suitable for solving financial market problems that require fast real-time responses. Thus, the prediction models for the domain-specific framework of the financial market were built using three representative simple and fast regression algorithms, namely, multilayer perceptron (MLP), random forest (RF), and support vector regression (SVR). The descriptions of prediction models are as follows:

MLP, a feed-forward artificial neural network, consists of an input layer, several hidden layers, and an output layer. Each node of the hidden layers and output layer commonly uses nonlinear activation functions, such as sigmoids, hyperbolic tangent, rectifier linear unit (ReLU), and softmax. We introduce MLP architecture for the prediction problems (Rumelhart, Hinton, & Williams, 1986). The regression function of the MLP is based on a linear combination of nonlinear basis functions $h_i(x)$in the form

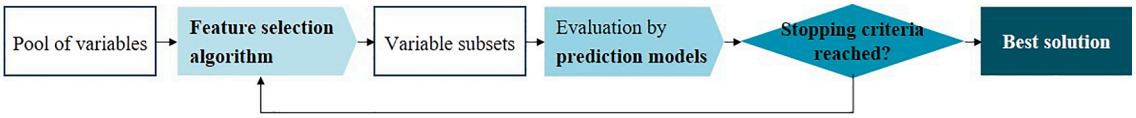

Fig. 2. The prodedure of feature selection algorithms.

$$y_i = f\left(w_0 + \sum_{i=1}^{N} w_i h_i(x)\right), \tag{1}$$

where w_i and $h_i(x)$ are weights and activation functions. Generally, the logistic sigmoid function is selected as the activation functions as follows:

$$h_i(x) = \frac{1}{1 + exp(-w_{io} - \sum_{j=1}^{p} w_{ij} x_j)}, \tag{2}$$

where x_j, w_{ij}, and $h_i(x)$ are j^{th} input variable, weights, and i^{th} hidden node, that consists of the hidden layer, respectively. Given a training set $\{x_i, y_i\}$, $i = 1, ..., N, y_i \in R, x_i$ is an input feature vector and y_i is its target value to be estimated by the regression function, we minimize the error function,

$$E(w) = \frac{1}{2} \sum_{i=1}^{N} ||f(x_i) - y_i||^2. \tag{3}$$

A back propagation algorithm is commonly used to derive optimal parameters (weights and biases) in the training phase. Back propagation algorithms iteratively update parameters to minimize the error function until convergence. In the inference phase, the output values of test data can be estimated using these parameters. Although the training time of an MLP model is considerable owing to model complexity, the model has the advantage of determining a nonlinear relationship between a target variable and predictor variables.

RF is an ensemble algorithm based on bootstrap aggregating, an integrated regression tree (Breiman, 2001). RF randomly extracted several variables, which improved the correlation between the regression trees. Given a training set $\{x_i, y_i\}$, multiple B trees are constructed by selecting multiple random samples with replacement of the training data set. After training, test data are predicted by averaging the predictions from B trees on the test data as follows:

$$\widehat{T} = \frac{1}{B} \sum_{b=1}^{B} \widehat{T_b}(x^{test}), \tag{4}$$

where $\widehat{T_b}$ is the prediction value from the b-th tree. Despite depending on the computational power, the RF creates a highly accurate prediction model and estimates the variables that are important to the prediction for several data sets, and, therefore, it is a popular model widely used in various fields.

SVM is widely known as classification problems and can also be used as a regression problem, called SVR (Nocedal & Wright, 2006). SVR uses the same principle as SVM to find the regression equation by minimizing errors and maximizing the margins. SVR can be formulated as follows:

$$\begin{aligned} \min_{w,b,\xi_i,\xi_i^*} \quad & \frac{1}{2} w^T w + C \sum_{i=1}^{N} (\xi_i + \xi_i^*) \\ \text{subject to} \quad & y_i - (w^T x + b) \leqslant \epsilon + \xi_i, \\ & (w^T x + b) - y_i \leqslant \epsilon + \xi_i^*, \\ & \xi_i, \xi_i^* \geqslant 0, \ i = 1, \ldots, N \end{aligned} \tag{5}$$

where C is a hyperparameter for tuning the tolerance of occurring outside the acceptable error rate, ϵ is the margin of error, and ξ_i, ξ_i^* are the slack variables, that are zero if the training data is in the epsilon-insensitive tube. This quadratic optimization problem with constraints as in Eq. (6) is termed the primal problem and can be transformed into a dual problem for computational convenience (Wolfe, 1961). The dual problem can be summarized as follows:

$$\max_{\alpha} \mathscr{L}(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{i=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j, \tag{6}$$

where α_i are Lagrange multipliers. In this dual problem, the kernel trick can be used to transform the data into a higher-dimensional feature space to create a linear separable space. Using these transformation methods to solve the dual problem, we can obtain the regression function. SVR, a deterministic algorithm, although provides accurate predictions in several applications, require significant times to converge.

3.4. Evaluation

We used two accuracy metrics for financial markets to evaluate the prediction performance: mean absolute error (MAE) and percentage of absolute range error (PARE) that are appropriate indicators for the target value with error tolerance. The commonly used MAE is a measure of the absolute errors between the predicted and actual values. Since MAE is the most intuitive measure, we selected the MAE over the MSE, and it is calculated as follows:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |\widehat{y_i} - y_i|, \tag{7}$$

where $\widehat{y_i}$ and y_i, are the predicted and the actual value, respectively. In addition, evaluating whether a predicted value is within a certain range, rather than predicting an exact value, can be a meaningful interpretation in the financial market. In other words, due to the nature of the financial target values, tolerances exist, PARE is calculated for the ratio of accuracy in the tolerance as follows:

$$PARE = \frac{1}{n} \sum_{i=1}^{n} I(|\widehat{y_i} - y_i| \leqslant \theta), \tag{8}$$

where θ and $I(x)$ are the tolerance and indicator functions, respectively, having the values 1 and 0 for true and false x, respectively. We can use the tolerances determined by experts in the field. The 10-fold cross validation was used to calculate MAE and PARE. The PARE is 1 if all the predicted values are in the error tolerance. We can use error tolerances determined by experts in the domain. The MAE and PARE were calculated using 10-fold cross validation.

3.5. Reporting

There are several methods to report analysis results based on the objective of the analysis and the modeling employed. Financial markets require rapid decision making and intuitive reporting. Visualization allows intuitive problem identification and assists decision making processes in practical sense. For example, objects can be clustered with similar information by compressing information from various variables, such as documents, and economic indicators, and projecting it in a two-dimensional plane. In addition, real-time information of significant variables extracted by the model can be visualized and used for monitoring as shown in Fig. 3. Fig. 3 shows an example of some variables collected in real-world. From the graph, the average value of The variable in Fig. 3(a) increases over time, and the values at some point in

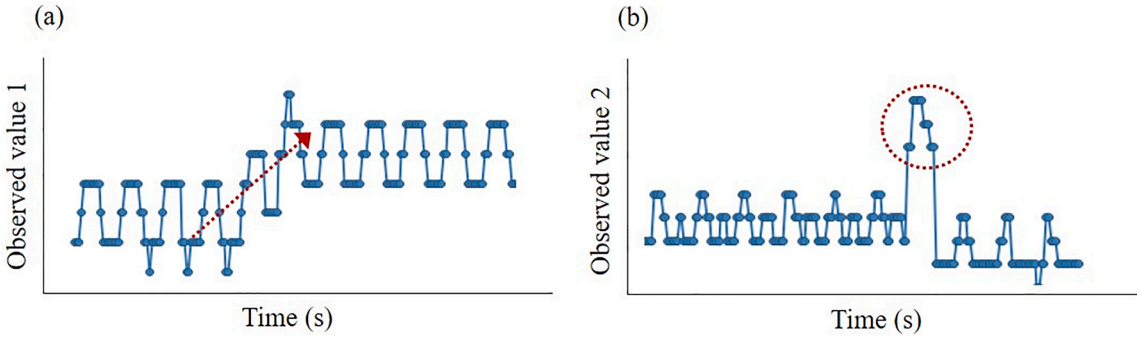

Fig. 3. The observed values as a function of time.

Fig. 3(b). lie outside the range commonly shared by the remaining. The field worker can easily detect this jumping data point from the visuals provided and suspect that there is a problem with the variable in the data recorded. Another effective reporting method is to provide a hierarchical listing of the set of input variables reported significant from the analysis. For example, in the financial market, the convention is that interest rates, exchange rates, stock prices, and other macroeconomic and microeconomic indicators are defined as the higher level categories, and 3-year treasury rate, LIBOR, USDKRW, KOSPI, unemployment rate, and consumer price index is classified into the lower level categories. Consequently, as shown in Fig. 4, it is possible to provide several information to the business through a reporting form that analyzes the category among the input variables with the characteristic of the hierarchical variables. Fig. 4(a) shows the distribution of variables for each higher level category after categorizing each variable and Fig. 4(b) shows the distribution of the relevant input variables categorized into a set of category 1, category 2, and category 3 variable for each target variable. Category 1 variables among input variables were selected for target variable 1, and category 3 variables were chosen for target variable 2. From Fig. 4, it can be seen that important category variables can appear differently depending on the target variables. Therefore, it is possible to provide various types of information to the business through the reporting form that analyzes the category among the input variables having the characteristic of the hierarchical variables and the visualization.

4. Application to financial market

4.1. Background

Bonds issued in other countries vary relative to characteristics and regulations involved, and several market participants actively use these bonds as means of asset managements. Bonds are issued by the government, national banks, high-endowed companies, and other institutions. The issue interest rates are calculated according to the credit rating of the issuing entity. Among several different bonds, those issued by the government mainly have maturities of 3, 5, 10, 20, and 30 years, and an efficient bond portfolio should be constructed because risks are calculated differently at maturity. Generally, each company allocates risk limits for each maturity and total risk limits to the traders. Risk is directly related to the quantity of the invested asset, that has a direct impact on profitability. Therefore, it is necessary to manage risks effectively to maximize operating profits in the financial market.

An example of risk management strategies is spread trading, that although affects risk limits by maturity, does not affect the total risk limits. It uses the difference in interest rate between short-term and long-term bonds to eliminate interest rate directional risk through the opposite direction trading, leading to its delta neutrality. Particularly, the difference between 3-year treasury bond rate and 10-year treasury bond rate, called the 3–10 year treasury spread, has functioned as an important market indicator, and traders use a strategy that generates additional profits by predicting the 3–10 year treasury spread. Fig. 5 shows the interest rates of the 3-year and 10-year treasury bonds and the 3–10 year treasury spread trends. The graph shows that the 3–10 year treasury spread are volatile, oscillating between 20 and 120 basis points (bp). From Fig. 5, although long-term and short-term spreads have predominantly positive values, interest rates often reverse owing to changes in market conditions, such as reflecting the investor's pessimistic outlook for the long-term economic growth and inflation latently expected. Because the 3–10 year treasury spread is unclear in terms of cap, floor, reference point, and many more, and because the volatility of the spread is huge owing to the complex influence of several economic indicators, it is difficult to predict the 3–10 year treasury spread. Notwithstanding these difficulties, if active traders can rapidly predict the treasury spreads using a several economic indicators, it can be an important leading indicator that aids rapid decision-making and depicts the long-term economic perspective and the short-term bond market direction. Hence, a novel data mining framework that analyzes the 3–10 year treasury spread and detect distinct signals can serve as an effective assistant tool for the financial agent in the market with their decision-

Fig. 4. (a) Distribution of category variables (b) Distribution of critical variables in each category for target variables.

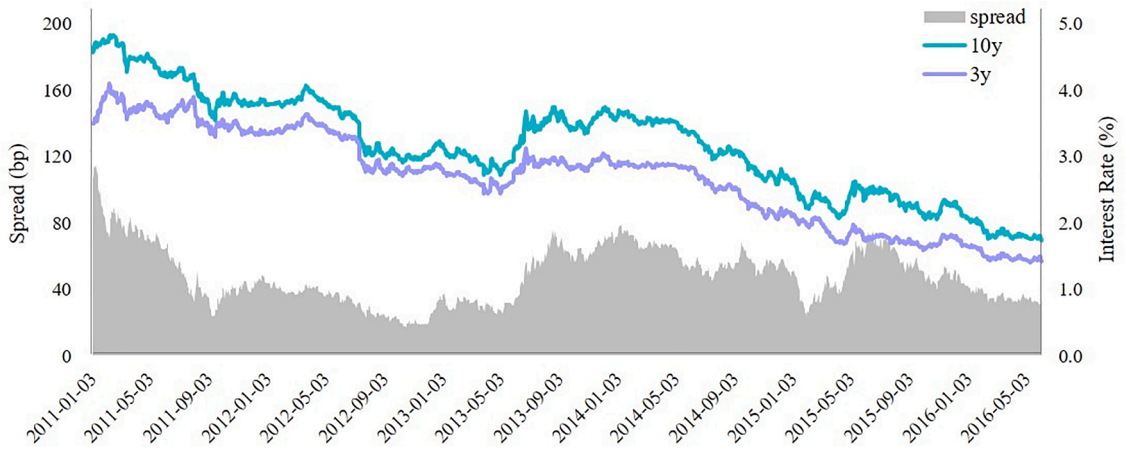

Fig. 5. The 3–10 year treasury spread trend.

making process.

4.2. Data description

We used real-world data collected from January 1, 2011 to June 3, 2016, and the variables were selected according to the opinions of experts in the domain. First, we used INFOMAX terminal to collect 20 variables such as interest rates, exchange rates, and stock market index. In addition, we collected seven other economic indicators, that are announced monthly, on the website of the national indicator system. A total of 27 input variables were used as listed in Table 1, and output variables used were from the difference between the 3-year and 10-year treasury bond rates.

4.3. Experimental settings

The data were sorted according to Korean time because of the time difference between countries and because there is no value for each country's national holidays, in the case of public holidays, the value was replaced by the value of the previous business day. Monthly data were also converted to daily data according to the date of the announcement. All data values were preprocessed from −1 to 1 through min–max normalization.

The parameters required for each prediction model mentioned in Section 3.3 were optimized before the experiment. The experiment was repeated 10 times for each algorithm to determine the optimal parameters using the MAE criteria. We applied the algorithm-specific greedy search methodology to the search space for the parameters shown in Table 2 only and the default values provided by MATLAB were used for the other parameters.

h of MLP is the number of hidden nodes in the hidden layer, the *minimum leaf* of RF is the minimum number of observations per tree leaf for each regression tree, and C and $\in$ of SVR are the tradeoff and epsilon, respectively. Consequently, the hidden node of MLP is 18, the minimum leaf of RF is 1, and C and $\in$ of the SVR used are 100 and 2^{-5}, respectively. To determine the tolerance required when evaluating performance with the PARE, we set the PARE tolerance between 2 and 3 *bps* based on opinions from experts in the bond market sector.

Table 1
Lists of input variables.

Higher level category	Lower level category	Frequency
H1. Interest rate	*Policy rate*: South Korea, U.S., Japan, China, Australia, UK *Variable rate*: LIBOR, EURIBOR	Daily
H2. Exchange rate	USDKRW, USDJPY, EURUSD	Daily
H3. Stock market index	KOSPI, KOSDAQ, DOW, HSCEI, NIKKEI, FTSE, S&P, Shanghai	Daily
H4. Commodity	WTI	Daily
H5. Other economic indicator in South Korea	Unemployment rate, Consumer price index, Index of mining and manufacturing industrial product, Total production of small manufacturing firm, Trade balance	Monthly
H6. Other economic indicator in the U.S.	Unemployment rate of USA, Nonfarm payroll	Monthly

Table 2
Search space for each parameter.

Algorithm	Parameter	Search space
Multilayer perceptron	h	{2, 4, …, 20}
Random forest	*minimum leaf*	{1, 2, …, 5}
Support vector regression	C	$\{10^{-1}, 10^{0}, \ldots, 10^{2}\}$
	$\in$	$\{2^{-5}, 2^{-4}, \ldots, 2^{-3}\}$

4.4. Prediction results

Based on those optimal parameters mentioned in Section 4.3, the experiment was designed as shown in Table 3 using feature selection algorithms and prediction models discussed in Sections 3.2 and 3.3.

We examined nine different methods by applying feature selection algorithms to three prediction models, and the results of MAE are shown in Table 4. The μ values shown in Table 4 are calculated by averaging across 10 repetitions for each methods, and the σ values are the standard deviation of each method for the 10 repetitions. The MAE of spread is at minimum values if we chose the RF prediction model and GA variable selection. Moreover, because the standard deviation of all experiments is significantly small, the prediction models are stable with no significant effected from initial conditions, and the prediction models converge stably. In addition, independent 2 samples t-tests were conducted to determine if there is a performance difference between RF-GA model and other models. Consequently, the performance of the RF-GA model was statistically significantly different from that of other models at a significant level 0.05.

Table 3
Prediction models using feature selections.

	Genetic algorithm (GA)	Forward selection (FS)	Backward elimination (BE)
Multilayer perceptron (MLP)	MLP-GA	MLP-FS	MLP-BE
Random forest (RF)	RF-GA	RF-FS	RF-BE
Support vector regression (SVR)	SVR-GA	SVR-FS	SVR-BE

Table 4
MAE results of each algorithm for spread prediction.

MAE	MLP-GA	MLP-FS	MLP-BE	RF-GA	RF-FS	RF-BE	SVR-GA	SVR-FS	SVR-BE
μ	2.017	2.222	2.126	1.633	1.706	1.671	1.985	2.002	1.982
σ	0.011	0.047	0.094	0.010	0.040	0.013	0.009	0.011	0.011

We also experimented 10 times with the PARE metric for each prediction model using feature selections, and the results of PARE are shown in Table 5.

In the results above, the larger the tolerance the larger the PARE value is. More specifically, a 1 bp difference in the tolerance results in a performance difference of close to 20% for MLP and SVR, whereas for RF, there is a performance difference of approximately 14%. In the PARE metric, RF-GA has the best performance and, compared to other prediction models, the difference in values against the tolerance value is small, indicating that the spread prediction value is forecasted within 2 *bp* of the actual value. These results may support the interpretation of MAE values of RF-GA within 2 *bp*.

We compared the performances of prediction models with the MAE and PARE criteria. The best model was RF-GA based on PARE criteria for predicting the spread, and we select relevant input variables that were selected more than 50% for the ten repetitions and categorized input variables into high level categories as shown in Fig. 6. From Fig. 6, most economic indicators in South Korea and the United States are selected as important variables, and is consistent with indicators that are considered to affect bond yields in the field. Some component of the price index of stock variables are selected as important variables. Moreover, most exchange rates were selected as important variables, implying that the exchange rates are directly related to the influx of foreign capital to the domestic financial market. In the practice, although traders anticipate the announcement of economic indicators and monitor interest rates, they tend to consider that exchange rate fluctuations have insignificant impact on the 3–10 year treasury spread. Therefore, from prediction models, traders can identify variables, that were considered as having insignificant influence and use for monitoring.

4.5. Summary of data mining framework for financial markets

We propose a data mining framework that uses several economic data and quantitative indicators as the input and the 3–10 year treasury spread as the output. Regardless of this, if traders could exploit different quantitative indicators that can help effectively predict the spreads of interest rates, then this prediction can be used as an important indicator in the bond market. The framework used wrapper approaches which select variables that significantly affect the performance of prediction models to enhance the direct interpretation of the effect of financial variables. PARE integrated with MAE, was used to quantify the model performance, considering the tolerance of the target value. Moreover, we provided visualization and hierarchical information of significant variables to present the analysis results in an intuitive and steady-fast technique. From applying this framework to the bond spread prediction, the PARE of the RF-GA model recorded 84%, that can be considered as the benchmark prediction accuracy for business decision making in the field. Furthermore, tests have shown that differences in performance across different models employed were statistically significant. Finally, it was concluded that domestic and foreign economic indicators are important among existing higher level categories, and exchange rate, that was considered as having negligible influence on the spread, is indeed required for explaining the fluctuations of the spread.

Although only 27 input variables were used in this study, we predict an improvement in the performance of the prediction model using several economic indicators in the future. Although financial time series data is important, economic indicators are related to data on future economies, and hence, using adequate economic indicators can yield significant results. From prediction models, traders need to collect and monitor adequate exchange rate variables in the future. This framework is envisaged to be a powerful tool to support traders in making decisions.

5. Conclusion

In this study, we proposed a data mining framework designed specifically for the financial market with the following three characteristics: interpretation, prediction evaluation metrics, and reporting methods. Firstly, a prediction model should provide predictive power and interpretability. We proposed that such an objective can be achieved by employing wrapper approaches if developing a prediction model, wherein the extent of contribution of each variable in the analysis can be observed. Secondly, the financial prediction model requires evaluation using appropriate metrics. For example, it may be important to predict the range of a target outcome, rather than the exact value, to incorporate tolerance. In this case, PARE is an appropriate measure of evaluation. Finally, presenting the results from the analysis in several reporting modes is required, particularly to aid active traders in the market to instantly decide with detailed explanations.

Considering these characteristics, we applied the proposed data mining framework to predict the 3–10 year treasury spread in a given financial market. Some pre-defined data mining framework provided a predictive power, and traders identified that predictor variables such as exchange rate, that were treated as having insignificant influence, have a significant effect on the spread. The results can lead to an effective monitoring by focusing on these variables intensively, instead of the conventional overall monitoring. We propose that the results from our pre-defined data mining framework can assist traders to make data-driven decisions with appropriate reporting methods. Finally, these frameworks can be used as a study to analyze several factors organically linked through several data mining algorithms, such as, association analysis and cluster and to characterize the financial markets both for individual countries and worldwide. In the future, comparative studies that replace simple regression models with deep learning models will be possible.

CRediT authorship contribution statement

Misuk Kim: Conceptualization, Methodology, Software, Validation,

Table 5
PARE results of each algorithm for spread prediction.

PARE	θ	MLP-GA	MLP-FS	MLP-BE	RF-GA	RF-FS	RF-BE	SVR-GA	SVR-FS	SVR-BE
μ	2 *bp*	0.618	0.563	0.590	0.706	0.688	0.693	0.573	0.568	0.574
σ		0.008	0.013	0.019	0.010	0.008	0.009	0.002	0.007	0.006
μ	3 *bp*	0.780	0.738	0.760	0.839	0.832	0.833	0.778	0.778	0.780
σ		0.011	0.009	0.016	0.003	0.008	0.004	0.005	0.005	0.004

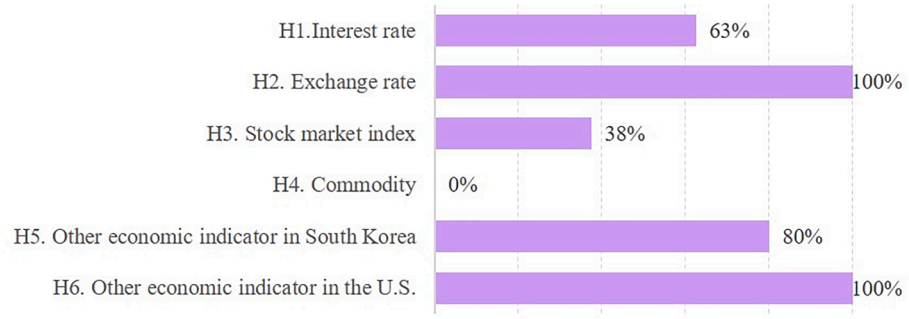

Fig. 6. Distribution of critical variables in each category.

Formal analysis, Investigation, Resources, Data curation, Visualization, Supervision, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1G1A1101195).

References

Abarbanell, J. S., & Bushee, B. J. (1997). Fundamental analysis, future earnings, and stock prices. *Journal of Accounting Research, 35*, 1–24. https://doi.org/10.2307/2491464

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324

Chatzis, S. P., Siakoulis, V., Petropoulos, A., Stavroulakis, E., & Vlachogiannakis, N. (2018). Forecasting stock market crisis events using deep and statistical machine learning techniques. *Expert Systems with Applications, 112*, 353–371. https://doi.org/10.1016/j.eswa.2018.06.032

De Silva, D., Yu, X., Alahakoon, D., & Holmes, G. (2011). A data mining framework for electricity consumption analysis from meter data. *IEEE Transactions on Industrial Informatics, 7*, 399–407. https://doi.org/10.1109/TII.2011.2158844

Fan, C., Xiao, F., & Yan, C. (2015). A framework for knowledge discovery in massive building automation data and its application in building diagnostics. *Automation in Construction, 50*, 81–90. https://doi.org/10.1016/j.autcon.2014.12.006

Gavrilovic, M., & Zimonjic, S. (2017). Technical analysis as a method of predicting price movements and future market trends. *Finance, Banking and Insurance*, 26.

Herui, C., Xu, P., & Yupei, M. (2015). Electric load forecast using combined models with hp filter-sarima and armax optimized by regression analysis algorithm. *Mathematical Problems in Engineering, 2015*. https://doi.org/10.1155/2015/386925

Kang, P., Lee, H.-J., Cho, S., Kim, D., Park, J., Park, C.-K., & Doh, S. (2009). A virtual metrology system for semiconductor manufacturing. *Expert Systems with Applications, 36*, 12554–12561. https://doi.org/10.1016/j.eswa.2009.05.053

Kang, S., Kim, E., Shim, J., Cho, S., Chang, W., & Kim, J. (2017). Mining the relationship between production and customer service data for failure analysis of industrial products. *Computers & Industrial Engineering, 106*, 137–146. https://doi.org/10.1016/j.cie.2017.01.028

Kannan, K. S., Sekar, P. S., Sathik, M. M., & Arumugam, P. (2010). Financial stock market forecast using data mining techniques. In *Proceedings of the international multiconference of engineers and computer scientists*. Vol. 1. 10.1.1.302.5160.

Kim, M., Kang, S., Lee, J., Cho, H., Cho, S., & Park, J. S. (2017). Virtual metrology for copper-clad laminate manufacturing. *Computers & Industrial Engineering, 109*, 280–287. https://doi.org/10.1016/j.cie.2017.04.016

Kumar, S., & Toshniwal, D. (2015). A data mining framework to analyze road accident data. *Journal of Big Data, 2*, 26. https://doi.org/10.1186/s40537-015-0035-y

Leigh, W., Modani, N., Purvis, R., & Roberts, T. (2002). Stock market trading rule discovery using technical charting heuristics. *Expert Systems with Applications, 23*, 155–159. https://doi.org/10.1016/S0957-4174(02)00034-9

Moat, H. S., Curme, C., Avakian, A., Kenett, D. Y., Stanley, H. E., & Preis, T. (2013). Quantifying wikipedia usage patterns before stock market moves. *Scientific Reports, 3*, 1801. https://doi.org/10.1038/srep01801

Mostafa, M. M. (2010). Forecasting stock exchange movements using neural networks: Empirical evidence from kuwait. *Expert Systems with Applications, 37*, 6302–6309. https://doi.org/10.1016/j.eswa.2010.02.091

Nocedal, J., & Wright, S. (2006). Numerical optimization. *Springer Science & Business Media*. https://doi.org/10.1038/srep01801

Pai, P.-F., & Lin, C.-S. (2005). A hybrid arima and support vector machines model in stock price forecasting. *Omega, 33*, 497–505. https://doi.org/10.1016/j.omega.2004.07.024

Pan, Y., Xiao, Z., Wang, X., & Yang, D. (2017). A multiple support vector machine approach to stock index forecasting with mixed frequency sampling. *Knowledge-Based Systems, 122*, 90–102. https://doi.org/10.1016/j.knosys.2017.01.033

Preis, T., Moat, H. S., & Stanley, H. E. (2013). Quantifying trading behavior in financial markets using google trends. *Scientific Reports, 3*, 1684. https://doi.org/10.1038/srep01684

Qiu, W., Liu, X., & Wang, L. (2012). Forecasting shanghai composite index based on fuzzy time series and improved c-fuzzy decision trees. *Expert Systems with Applications, 39*, 7680–7689. https://doi.org/10.1016/j.eswa.2012.01.051

Rather, A. M., Agarwal, A., & Sastry, V. (2015). Recurrent neural network and a hybrid model for prediction of stock returns. *Expert Systems with Applications, 42*, 3234–3241. https://doi.org/10.1016/j.eswa.2014.12.003

Rodríguez-González, A., Colomo-Palacios, R., Guldris-Iglesias, F., Gómez-Berbís, J. M., & García-Crespo, A. (2012). Fast: Fundamental analysis support for financial statements. Using semantics for trading recommendations. *Information Systems Frontiers, 14*, 999–1017. https://doi.org/10.1007/s10796-011-9321-1

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*, 533–536. https://doi.org/10.1038/323533a0

Shukla, S. K., & Tiwari, M. K. (2011). Ga guided cluster based fuzzy decision tree for reactive ion etching modeling: a data mining approach. *IEEE Transactions on Semiconductor Manufacturing, 25*, 45–56. https://doi.org/10.1109/TSM.2011.2173372

Wang, L., Zou, H., Su, J., Li, L., & Chaudhry, S. (2013). An arima-ann hybrid model for time series forecasting. *Systems Research and Behavioral Science, 30*, 244–259. https://doi.org/10.1002/sres.2179

Wolfe, P. (1961). A duality theorem for non-linear programming. *Quarterly of Applied Mathematics, 19*, 239–244. https://doi.org/10.1090/qam/135625

Xiong, L., & Lu, Y. (2017). Hybrid arima-bpnn model for time series prediction of the chinese stock market. In *2017 3rd International conference on information management (ICIM)* (pp. 93–97). IEEE. doi: 10.1109/INFOMAN.2017.7950353.

Yu, E., & Cho, S. (2006). Ensemble based on ga wrapper feature selection. *Computers & Industrial Engineering, 51*, 111–116. https://doi.org/10.1016/j.cie.2006.07.004

Yu, Z., Fung, B. C., & Haghighat, F. (2013). Extracting knowledge from building-related data–a data mining framework. In *Building simulation* (pp. 207–222). Springer. Vol. 6. doi: 10.1007/s12273-013-0117-8.