Fin. Manag. Persp. 2025, 15(2), 34-53; DOI: 10.48308/jfmp.2025.239954.1498

Financial Management Perspective
Journal hompage: https://jfmp.sbu.ac.ir/

Original Article

Prediction of Tehran Stock Exchange Total Index Using Natural Gradient Boosting and SHAP Values

Meysam Amiry*
Moslem Peymani Forooshani**
Mohammadali Dehghan Dehnavi ***
Madjid Ghods****

Abstract

Introduction: The stock market represents one of the most pivotal components of developing economies. Consequently, extensive research employing both technical and fundamental analyses has sought to predict financial time series in order to assist investors with their trading decisions. In this regard, machine learning models have emerged as effective tools for addressing a variety of challenges. Nevertheless, despite the notable performance of machine learning models in this area, two significant criticisms persist. The first concerns the lack of interpretability of the results; in such models, the process by which inputs are transformed into outputs, as well as the contribution of each input to the model's output, is not clearly defined. The second issue pertains to the reliability of the predictions generated by these models, as this reliability cannot be directly inferred from the model itself. Accordingly, this study utilizes the latest methods developed in the field of machine learning to address these two issues.

Method: Considering that the selection of input features plays a crucial role in shaping the output of these models, this study employs a systematic approach to extract features used in related research over the past five years through a systematic review using the Scopus scientific database. Ultimately, 34 features with daily available data were selected as inputs for the model. In the next step, the Natural Gradient Boosting

Received; 14 May 2025 Accepted; 17 August 2025 .
* Assistant Prof., Department of Finance and Banking, Faculty of Management and Accounting, Allameh Tabataba'i University, Tehran, Iran.
** Associate Prof., Department of Finance and Banking, Faculty of Management and Accounting, Allameh Tabataba'i University, Tehran, Iran.
*** Assistant Prof., Department of Finance and Banking, Faculty of Management and Accounting, Allameh Tabataba'i University, Tehran, Iran.
**** Ph.D. Candidate in Financial Engineering, Department of Finance and Banking, Faculty of Management and Accounting, Allameh Tabataba'i University, Tehran, Iran. (**Corresponding Author**), **Email**: madjid_ghods@atu.ac.ir

model was utilized to predict the data of the Tehran Stock Exchange Total Index from March 2010 to January 2025. The performance of this model was evaluated using the RMSE, MAE, and MAPE metrics and compared with the latest machine learning methods for time series prediction. Subsequently, SHAP values were employed to interpret the results of the Natural Gradient Boosting model. This approach allowed for the assessment of the contribution of each feature to the estimation of the model's output. SHAP values provide a powerful tool for evaluating the impact of each input feature on the output estimation, offering valuable insights to users of machine learning models.

Results and Discussion: A comparison of the error values of the proposed model with those of other machine learning models indicates superior predictive performance for the proposed approach. Unlike conventional machine learning models, which provide a single prediction as the best estimate, the proposed model outputs a probability distribution that can be described by its parameters. In this study, the assumed parametric form of the distribution is the normal distribution, which is characterized by its mean and standard deviation. In fact, the predicted value corresponds to the mean of the estimated distribution. For forecasting the Tehran Stock Exchange index, the most influential features are the closing price, the EMA indicator, and the SMA indicator. Interpretation of the predicted standard deviation parameter reveals that the ATR indicator, closing price, and TEMA indicator have the greatest impact on this parameter. As the relative values of these variables increase, the standard deviation of the estimated distribution also increases, indicating that the corresponding prediction is less reliable.

Conclusion: The findings of this study demonstrate that the Natural Gradient Boosting model can serve as an effective tool for predicting the Tehran Stock Exchange Total Index. The interpretation of results using SHAP values enables the identification of the most important input features and the manner in which the output is formed from these features, thereby aiding in model optimization. This approach not only enhances prediction accuracy but also assists market participants and policymakers in making more informed decisions regarding risk management and resource allocation. Ultimately, comparisons with other models indicate that this method can be employed as a practical and reliable solution for financial market analysis.

Keywords: Stock Price Prediction, Natural Gradient Boosting, SHAP Values, Tehran Stock Exchange.

How to Cite: *Mohammadiaghdam, S. , Peymany Foroushany, M. , Amiry, M. and bahrani, M. (2025). Predicting Iran's Yield Curve: Combining Factor Model with Machine Learning Approach. Financial Management Perspective, 15(1), 34-53. doi: 10.48308/jfmp.2025.239954.1498 (In Persian).*

چشم انداز مدیریت مالی ،۱۴۰۴ دوره ۱۵، شماره ۲ صص ۳۴ –۵۳، DOI: 10.48308/jfmp.2025.239954.1498

چشم انداز مدیریت مالی
Journal hompage: https://jfmp.sbu.ac.ir/

نوع مقاله: پژوهشی

پیش‌بینی شاخص کل بورس اوراق بهادار تهران با استفاده از تقویت گرادیان طبیعی و مقادیر SHAP

میثم امیری*
مسلم پیمانی فروشانی**
محمدعلی دهقان دهنوی***
مجید قدس****

چکیده

هدف: بازار سهام یکی از کلیدی‌ترین عناصر اقتصادهای درحال‌توسعه است. به همین جهت مطالعات گسترده‌ای با استفاده از تحلیل‌های تکنیکال و بنیادی، به پیش‌بینی سری‌های زمانی مالی پرداخته‌اند تا بتوانند به سرمایه‌گذاران در معاملاتشان یاری رسانند. در همین راستا مدل‌های یادگیری ماشین توانسته‌اند به‌عنوان ابزاری کارآمد برای مسائل گوناگون، نقش‌آفرینی کنند. اما علی‌رغم عملکرد قابل‌توجه مدل‌های یادگیری ماشین در این حوزه، دو ایراد مهم به آنها وارد است. مسئله اول تفسیرناپذیری نتایج است که در این مدل‌ها نحوه تبدیل ورودی‌ها به خروجی‌ها و یا سهم هر یک از ورودی‌ها در شکل‌دادن به خروجی مدل مشخص نمی‌باشد. دوم، قابل‌اتکا بودن نتایج حاصل از پیش‌بینی این مدل‌هاست که به طور مستقیم از مدل قابل‌استخراج نمی‌باشد. به همین جهت در این پژوهش از جدیدترین روش‌های ارائه شده در حوزه یادگیری ماشین برای پاسخ‌گویی به این دو مسئله استفاده شده است.

روش: باتوجه‌به اینکه در این مدل‌ها انتخاب ویژگی‌های ورودی از اهمیت بسیار زیادی در شکل‌دهی خروجی مدل برخوردار است، در این پژوهش با روشی نظام‌مند، از طریق مرور سیستماتیک ویژگی‌های به کار گرفته شده در پژوهش‌های مرتبط در پنج سال اخیر استخراج شده و نهایتاً ۳۴ ویژگی که داده‌های آنها به شکل روزانه موجود هستند به‌عنوان ورودی مدل انتخاب شده‌اند. در گام بعد، از مدل تقویت گرادیان طبیعی برای پیش‌بینی داده‌های شاخص کل بورس اوراق بهادار تهران از ابتدای سال ۱۳۸۹ تا سال ۱۴۰۳ استفاده شده است. عملکرد این مدل با استفاده از معیارهای RMSE، MAE و MAPE سنجیده

تاریخ دریافت: ۱۴۰۴/۰۲/۲۴ تاریخ پذیرش: ۱۴۰۴/۰۵/۲۶

* استادیار گروه مالی و بانکداری، دانشکده مدیریت و حسابداری، دانشگاه علامه طباطبایی، تهران، ایران.
** دانشیار گروه مالی و بانکداری، دانشکده مدیریت و حسابداری، دانشگاه علامه طباطبایی، تهران، ایران.
*** استادیار گروه مالی و بانکداری، دانشکده مدیریت و حسابداری، دانشگاه علامه طباطبایی، تهران، ایران.
**** دانشجوی دکترای مهندسی مالی، گروه مالی و بانکداری، دانشکده مدیریت و حسابداری، دانشگاه علامه طباطبایی، تهران، ایران. (**نویسنده مسئول**)
Email: madjid_ghods@atu.ac.ir

شده و با جدیدترین روش‌های یادگیری ماشین برای پیش‌بینی سری‌های زمانی مقایسه شده است. در ادامه از مقادیر SHAP برای تفسیر نتایج مدل تقویت گرادیان طبیعی استفاده شده است و سهم هر یک از ویژگی‌ها در تخمین خروجی مدل ارزیابی شده است. مقادیر SHAP ابزاری قدرتمند برای سنجش اثرگذاری هر یک از ویژگی‌های ورودی بر تخمین خروجی فراهم می‌آورد که اطلاعات ارزشمندی را در اختیار کاربران مدل‌های یادگیری ماشین قرار می‌دهد.

یافته‌ها: مقایسه مقادیر خطای مدل ارائه شده با سایر مدل‌های یادگیری ماشین، نشان از عملکرد بهتر این مدل در پیش‌بینی دارد. این مدل، برخلاف سایر مدل‌های یادگیری ماشین که یک پیش‌بینی به‌عنوان بهترین حدس را به‌عنوان خروجی ارائه می‌دهند، یک توزیع احتمال که بر اساس پارامترهای آن قابل‌توصیف است ارائه می‌دهد. شکل توزیع پارامتریک در نظر گرفته شده در این پژوهش تابع توزیع نرمال است که با پارامترهای میانگین و انحراف معیار قابل‌توصیف می‌باشد. در واقع مقدار پیش‌بینی شده، همان میانگین توزیع تخمین زده شده است. برای پیش‌بینی شاخص بورس اوراق بهادار تهران، اثرگذارترین ویژگی‌ها قیمت پایانی، اندیکاتور EMA و اندیکاتور SMA هستند. تفسیر پارامتر انحراف معیار پیش‌بینی انجام شده، نشان می‌دهد که بیشترین اثرگذاری در این پارامتر را اندیکاتور ATR، قیمت پایانی و TEMA دارند که هر چقدر مقدار نسبی این متغیرها بیشتر باشد، انحراف معیار توزیع تخمینی بیشتر بوده و بنابراین پیش‌بینی انجام شده قابلیت اتکای کمتری خواهد داشت.

نتیجه‌گیری: مدل تقویت گرادیان طبیعی می‌تواند به‌عنوان ابزاری مؤثر در پیش‌بینی شاخص کل بورس اوراق بهادار تهران مورد استفاده قرار گیرد. تفسیر نتایج با استفاده از مقادیر SHAP، امکان شناسایی مهم‌ترین ویژگی‌های ورودی و همچنین نحوه تشکیل خروجی از ویژگی‌های ورودی را فراهم کرده و به بهینه‌سازی مدل کمک می‌نماید. این رویکرد نه‌تنها دقت پیش‌بینی را بهبود می‌بخشد، بلکه به فعالان بازار سرمایه و سیاست‌گذاران کمک می‌کند تا تصمیم‌گیری‌های آگاهانه‌تری در مدیریت ریسک و تخصیص منابع داشته باشند. در نهایت، مقایسه با سایر مدل‌ها نشان می‌دهد که این روش می‌تواند به‌عنوان یک راهکار عملی و قابل‌اعتماد در تحلیل بازارهای مالی به کار گرفته شود.

کلیدواژه‌ها: پیش‌بینی قیمت، تقویت گرادیان طبیعی، مقادیر SHAP، بورس اوراق بهادار تهران

استناد دهی: محمدی اقدم, سعید , پیمانی فروشانی, مسلم , امیری, میثم و بحرانی, محمد . (۱۴۰۴). پیش‌بینی منحنی بازده ایران: ترکیب مدل عاملی با رویکرد یادگیری ماشین. *چشم انداز مدیریت مالی*, ۱۵(۲), ۳۴-۵۳ .

۱. مقدمه

پیش‌بینی شاخص‌های بازار سهام از جمله موضوعات کلیدی در حوزه مالی به شمار می‌رود که توجه پژوهشگران و سرمایه‌گذاران زیادی را به خود جلب کرده است. اما این موضوع با چالش‌های قابل‌توجهی همراه است؛ زیرا که بازارهای مالی تحت‌تأثیر عوامل مختلفی چون تغییرات اقتصادی، سیاست‌های مدیران شرکت‌ها، وضعیت روانی سرمایه‌گذاران و رویدادهای سیاسی قرار دارند[۷]. اگرچه مطالعات گسترده‌ای با استفاده از روش‌های آماری و یا یادگیری ماشین برای پیش‌بینی و اثرگذاری این ویژگی‌ها صورت پذیرفته است اما داده‌های مالی به دلیل نویزی بودن، نوسانات بالا و پیچیدگی‌های غیرخطی چالش‌های بزرگی برای تحلیل و پیش‌بینی ایجاد می‌کنند. روش‌های سنتی آماری، همچون رگرسیون خطی و میانگین متحرک، به دلیل محدودیت‌هایشان در مدل‌سازی ویژگی‌های غیرایستا و پیچیده بازار، کارایی کمتری دارند. ازطرف دیگر، مدل‌های یادگیری ماشین توانایی بالایی در استخراج ویژگی‌های مؤثر و تحلیل داده‌های پیچیده دارند[۲۱]. پژوهشگران استفاده از ورودی‌های گوناگونی مانند شاخص‌های تکنیکال و یا حتی اخبار را مورد مطالعه قرار می‌دهند[۲۳]. استفاده از ورودی‌های متعدد نیز می‌تواند منجر به پیچیدگی زیاد و عدم توانایی مدل‌های یادگیری ماشین در شناسایی الگوها گردد. به‌علاوه نحوه اثرگذاری هر یک از ورودی‌ها، در تشکیل خروجی مدل، مورد سؤال بوده و میزان اثرگذاری هر ورودی برای تشکیل خروجی محل ابهام بوده است و تفسیر نتایج مدل‌ها نیز از اهمیت بالایی برخوردار است. ابزارهایی مانند SHAP [۱] قابلیت تحلیل و تفسیر تأثیر هر متغیر بر نتایج پیش‌بینی را فراهم می‌کنند. این مدل‌های توضیحی[۲] امکان شناسایی متغیرهای کلیدی و ارزیابی تأثیر آن‌ها بر پیش‌بینی شاخص‌های مالی را فراهم کرده و به تصمیم‌گیری‌های دقیق‌تر کمک می‌کنند. استفاده از مدل‌های توضیحی، علاوه بر افزایش شفافیت و قابلیت اعتماد به مدل‌های یادگیری ماشین، امکان ایجاد استراتژی‌های معاملاتی کارآمدتر را نیز فراهم می‌آورد[۱۳].

اگرچه به طور کلی بخش قابل توجهی از مسائل یادگیری ماشین نظارت شده[۳]، مسائل رگرسیون هستند اما به طور کلی روش‌های یادگیری ماشین به کار گرفته شده برای مسائل مالی، روش‌هایی هستند که یک خروجی مشخص به‌عنوان بهترین حدس[۴] را ارائه می‌نمایند. اما در بسیاری از مسائل مهم است که بتوان عدم قطعیت حول جواب ارائه شده را کمی‌سازی کرد. به منظور اینکه بتوان به سوالات مختلف در خصوص احتمال وقوع خروجی‌های مختلف با دانستن ورودی‌های مشخص x پاسخ داد، باید به جای تخمین یک مقدار نقطه‌ای $E[y \mid x]$، توزیع احتمال شرطی $P(y \mid x)$ را به ازای ورودی x تخمین زد. به این کار رگرسیون احتمالی[۵] می‌گویند[۹]. البته تخمین‌های احتمالی در مسائل طبقه‌بندی رایج است اما روش‌های موجود غیرمنعطف و کند بوده و استفاده از آنها نیاز به دانش بسیار تخصصی دارد. هر روش رگرسیون تخمین میانگینی[۶] با فرض همسانی واریانس و فرض یک مدل غیر شرطی برای نویز، قابلیت تبدیل به رگرسیون احتمالی را دارد، اما همسانی واریانس یک فرض بسیار محدود کننده است و به علاوه به شناخت دقیق اطلاعات آماری فرآیند نیاز دارد. روش‌های بیزین[۷] به طور طبیعی تخمین‌هایی با عدم قطعیت قابل پیش‌بینی ارائه می‌دهند اما راه حل‌های قطعی برای مدل‌های بیزین محدود به مسائل ساده بوده و محاسبات مربوط به شبکه‌های عصبی بیزین و مدل رگرسیون درختی جمعی بیزین[۸] بسیار دشوار است[۵]. به علاوه مدل‌های بیزین در داده‌های حجیم عملکرد ضعیفی از خود نشان می‌دهند[۳۰]. مدل‌های بیزین یادگیری عمیق نیز که عملکرد بسیار

[1] Shapley Additive Explanations
[2] Explainable AI (XAI)
[3] Supervised Machine Learning
[4] Best guess
[5] Probabilistic Regression
[6] Mean-estimating regression model
[7] Bayesian methods
[8] Bayesian Additive Regression Trees (BART)

خوبی در وظایف ادراکی (پردازش صوت و تصویر) دارند، در حوزه رگرسیون عملکردی مشابه سایر روش‌های معمول دارند و این در حالی است که نیازمند تنظیم بسیاری از پارامترها و دانستن اطلاعاتی در مورد فرآیند (پیش از پردازش)، استفاده از آنها را بسیار دشوار می‌سازد.

ماشین‌های تقویت گرادیان[1] روش‌های ماژولار هستند که عملکرد بسیار مناسبی در داده‌های ساختاریافته و حتی با اندازه‌های کوچک از خود نشـان می‌دهند[۴]. در مسـائل طبقه‌بندی خروجی این روش‌ها به طور پیش‌فرض، احتمالی اسـت اما در مسـائل رگرسـیون تنها یک عدد اسـکالر را به‌عنوان خروجی نتیجه می‌دهند. در شـرایطی که مربع خطاها به‌عنوان تابع هزینه اسـت این عدد اسـکالر را می‌توان به‌عنوان میانگین شـرطی توزیع گائوسـی با یک مقدار واریانس ثابت (نامشـخص) تفسـیر کرد. به همین جهت تفسـیرهای احتمالی زمانی که واریانس ثابت فرض شـده اسـت کاربرد چندانی ندارد. توزیع پیش‌بینی شـده خروجی باید از دو درجه آزادی برخوردار باشـد تا هم مقدار خروجی و هم دقت آن خروجی را به شکل موثر نشان دهد. دقیقاً همین مسئله، چالش اصلی GBM و سایر روش‌های دیگر است که روش تقویت گرادیان طبیعی این مسئله را با تقرب چند پارامتری حل کرده است[۲].

در این پژوهش ابتدا با مروری سـیسـتماتیک بر پژوهش‌های گذشـته، ویژگی‌های مورد اسـتفاده در این پژوهش‌ها اسـتخراج شـده و پس از پایش ویژگی‌ها، از آنها به‌عنوان ورودی‌های مدل اسـتفاده شـده اسـت. در ادامه از تقویت گرادیان طبیعی برای پیش‌بینی شـاخص اسـتفاده شـده اسـت که به‌جای ارائه یک پیش‌بینی نقطه‌ای، یک توزیع را به‌عنوان خروجی تخمین می‌زند و امکان بررسـی عدم قطعیت و قابل‌اتکا بودن تخمین انجام شـده را به شـکل غیرمستقیم فراهم می‌آورد. در گام بعد عملکرد مدل با چند روش یادگیری ماشین دیگر مورد مقایسه قرار گرفته است. در نهایت مقادیر SHAP برای هر یک از ورودی‌ها مورد محاسبه قرار گرفته و سهم ویژگی‌های ورودی در تشکیل پارامترهای توزیع تخمین زده شده، نشان داده شده است.

۲. مبانی نظری و پیشینه پژوهش

افشـاری‌راد و همکاران (۱۳۹۷) با اسـتفاده از داده‌های شـاخص کل به بررسـی روش‌های تکنیکال پیش‌بینی قیمت سهام و مدل‌های هوشمند یادگیری ماشین پرداختند. در این پژوهش ابتدا ۲۵ روش مختلف انتخاب شده و با استفاده از تکنیک‌های کاهش ابعاد ۵ روش از میان این روش‌ها برگزیده شده و برای تجمیع نهایی از روش رای اکثریت برای تصمیم‌گیری نهایی استفاده شده است. آنها بیان کردند که با استفاده از این روش علاوه بر دقت پیش‌بینی بالا، هیچ محدودیتی برای انتخاب روش‌های تکنیکال وجود ندارد و به نوعی انعطاف پذیری بالای این روش در به کارگیری سـایر ابزارها به‌عنوان مزیت آن معرفی شـده اسـت[۱]. صـالحی و گرشـاسـبی (۱۳۹۸) در پژوهش خود به پیش‌بینی شـاخص کل بورس اوراق بهادار تهران در بازه زمانی ۱۳۸۹ تا ۱۳۹۵ با اسـتفاده از سـیسـتم اسـتنتاج عصـبی- فازی انطباق‌پذیر پرداختند. آنها در ابتدا تعیین کردند که از ۶۸ ویژگی ورودی، ۱۰ ویژگی نهایی انتخاب شود و با استفاده از الگوریتم رقابت اسـتعماری ۱۰ ویژگی که منجر به کمترین خطا در پیش‌بینی مدل می‌شـوند را انتخاب کردند[۲۲]. محبی و همکاران (۱۴۰۰) با اسـتفاده از ۶۹ ویژگی و با اسـتفاده از روش انتخاب ویژگی پیشـنهادی خود اقدام به پیش‌بینی شـاخص کل بورس اوراق بهادار تهران نموده‌اند. آنها از روش‌های RBF، MLP، SVR و DNN برای این موضـوع اسـتفاده کرده و بیان داشـتند که اسـتفاده از روش پیشـنهادی آنها منجر به انتخاب ۷ ویژگی می‌شـود که اسـتفاده از این ویژگی‌ها کمترین میزان خطا در پیش‌بینی را به همراه خواهد داشت[۱۸]. سـهرابی و همکاران (۱۴۰۱) به بررسـی دقت پیش‌بینی جهش‌های شـاخص سـهام بر اسـاس روش‌های مختلف یادگیری ماشـین در بورس اوراق بهادار تهران پرداختند و با مقایسه روش‌های جنگل تصادفی، ماشین بردار پشتیبان، شبکه عصبی مصنوعی و شبکه

[1] Gradient Boosting Machines (GBMs)

عصبی بازگشتی استوار بر یادگیری عمیق به این نتیجه دست یافتند که روش یادگیری ماشین استوار بر شبکه عصبی بازگشتی حافظه طولانی کوتاه مدت نتیجه بهتری نسبت به سایر مدل‌ها در افق‌های زمانی ۱، ۳ و ۶ روزه دارد[۲۷]. حیدری و امیری (۱۴۰۱) به بررسی قدرت مدل‌های مبتنی بر هوش مصنوعی در پیش‌بینی روند قیمت سهام پرداختند. برای این منظور آنها داده‌های مربوط به ۱۵۰ شرکت بزرگ پذیرفته شده در بورس اوراق بهادار تهران در بازه زمانی سال‌های ۱۳۹۰ تا ۱۳۹۹ را جمع‌آوری نموده و با به کارگیری روش‌های یادگیری ماشین برای هر یک از این شرکت‌ها، به پیش‌بینی روند تغییرات قیمت سهام و آزمون هر یک از روش‌ها پرداخته و با یکدیگر مقایسه کردند. روش‌های به کارگرفته شده در این پژوهش شامل جنگل تصادفی، شبکه عصبی، مدل‌های خطی و مدل‌های خود همبسته بوده و نتایج آنها حاکی از آن بود که مدل‌های مبتنی بر یادگیری عمیق نسبت به سایر مدل‌ها عملکرد بهتری از خود نشان می‌دهند و در پیش‌بینی روند کوتاه مدت قیمت سهام، از دقتی حدود ۷۰ تا ۸۰ درصد برخوردارند. همچنین، مدل‌های یادگیری کم عمق دقت بالاتری داشته و به طور کلی، بیشتر مدل‌ها در پیش‌بینی روندهای منفی سهام، عملکرد بهتری نشان می‌دهند. آنها بیان کردند که طبق نتایج پژوهش آنها برخلاف یافته‌های پژوهش‌های گذشته، این مدل‌ها نتایج خیره کننده‌ای در اختیار سرمایه‌گذاران قرار نمی‌دهند[۱۱]. وزیری کردستانی و همکاران (۱۴۰۱) با استفاده از داده‌های قیمت شرکت‌های بورسی و فرابورسی، ۶۰ شرکت را به‌عنوان سهام ارزشی انتخاب نموده و با استفاده از یک رویکرد ترکیبی جدید با عنوان PSO-BiLSTM به پیش‌بینی قیمت سهام این شرکت‌ها پرداختند و در نهایت نتیجه‌گیری کردند که خطای این روش نسبت به سایر روش‌ها کمتر است[۲۹]. محبی و همکاران (۱۴۰۱) با به کارگیری ۳۴ ویژگی و با استفاده از شبکه‌های عصبی RBF[1] و یادگیری عمیق به پیش‌بینی شاخص کل بورس اوراق بهادار تهران پرداختند. آنها از الگوریتم‌های کمینه افزونگی، بیشینه وابستگی[2] و تحلیل مولفه اساسی[3] برای انتخاب ویژگی و کاهش ابعاد استفاده کردند و بیان داشتند که شبکه‌های عصبی RBF تنها با شش ویژگی که از روش الگوریتم کمینه افزونگی و بیشینه وابستگی به دست آمده‌اند، بهترین عملکرد را دارد[۱۹]. حاج سیدجوادی و همکاران (۱۴۰۱) با استفاده از یک الگوی ترکیبی از تبدیل موجک و تقویت گرادیان درختی به پیش‌بینی قیمت پسته در بورس کالای کشور پرداختند. آنها بیان کردند که با به کارگیری تبدیل موجک، خطای داده‌های قیمت کاهش یافته و داده‌ها از یک روند باثبات برخوردار شدند و در نتیجه این روش عملکرد بهتری در مقایسه با سایر روش‌ها از خود نشان می‌دهد[۱۰]. کیانی زاده و همکاران (۱۴۰۲) به مقایسه دقت مدل‌های منتخب یادگیری ماشین شامل، شبکه عصبی، رگرسیون لجستیک، نزدیک‌ترین همسایه K، ماشین بردار پشتیبان و اعتبار سنجی ضربدری برای پیش‌بینی قیمت سهام برای ۱۲ شرکت منتخب بورس اوراق بهادار تهران که از طریق روش حذف سیستماتیک انتخاب شده‌اند پرداخته و بیان کردند که از میان روش‌های یادگیری ماشین، الگوریتم ماشین بردار پشتیبان بیشترین قدرت پیش‌بینی کنندگی در قیمت سهام را به خود اختصاص داده است[۱۴]. شیخ زاده و رحمانی (۱۴۰۱) با به کارگیری ۳۰ ویژگی و اندیکاتور و استفاده از سه روش یادگیری ماشین به پیش‌بینی شاخص کل بورس اوراق بهادار تهران پرداختند. آنها از ۷ روش مختلف انتخاب ویژگی به منظور کاهش تعداد ویژگی‌های ورودی‌های این مدل‌ها استفاده کردند. آنها نتیجه‌گیری کردند که اندیکاتورهای RSI، Jaw، Senkou-Span A، CCI، William %R، Senkou Span B، Teeth و OBV نسبت به سایر اندیکاتورها و ویژگی‌ها بیشتر توسط

[1] Radial Basis Function
[2] Minimum Redundancy Maximum Relevance
[3] Principal Component Analysis

الگوریتم‌های انتخاب ویژگی انتخاب شده‌اند، بنابراین احتمالاً می‌توانند در تشخیص روند نسبت به سایر ویژگی‌ها و اندیکاتورها موثرتر باشند[۲۶].

گاندمال و کومار[۱] (۲۰۱۹) در پژوهش خود چندین مدل پیش‌بینی یادگیری ماشین را بر اساس مدل مورد استفاده، داده‌ها، معیار ارزیابی و نرم‌افزار مورد استفاده مورد بررسی قرار دادند[۸]. سیزر[۲] و همکاران (۲۰۲۰) مدل‌های یادگیری عمیق، شبکه‌های عصبی بازگشتی، شبکه‌های عصبی باور عمیق، شبکه‌های یادگیری تقویتی و شبکه‌های LSTM را مورد بررسی قرار دادند و بیان کردند که شبکه‌های LSTM پرکاربردترین روش برای پیش‌بینی شاخص بازار سهام هستند زیرا که پیاده‌سازی آنها آسان بوده و عملکردی مناسب در داده‌های سری زمانی از خود نشان می‌دهند[۲۴]. تاکار و چادهاری[۳] (۲۰۲۱) به بررسی ۹ مدل مختلف یادگیری عمیق برای پیش‌بینی شاخص سهام پرداختند. آنها آزمون‌های مقایسه‌ای بر اساس ویژگی‌های ورودی این مدل‌ها انجام داده و بیان کردند که شبکه عمیق کیو[۴] بهترین عملکرد را بین انواع مختلف شبکه‌ها دارا است[۲۸]. کو[۵] و همکاران (۲۰۲۱) از ۴ روش مختلف انتخاب ویژگی برای انتخاب بهترین ویژگی‌ها برای پیش‌بینی ورشکستگی شرکت‌های کوچک و متوسط استفاده کردند. آنها اهمیت قابل‌توجه انتخاب ویژگی در برای بهبود عملکرد مدل‌های پیش‌بینی‌کننده را نشان دادند[۱۵]. سلیک[۶] و همکاران (۲۰۲۳) مدلی ترکیبی از الگوریتم‌های یادگیری ماشین و مدل‌های تفسیرکننده ارائه دادند. آنها از LIME[۷] که یک روش تفسیر محلی است استفاده کرده و بیان کردند که ترکیب پیشنهاد شده در پیش‌بینی جهت تغییرات شاخص‌های سهام از دقت بسیار زیادی برخوردار است[۳]. تان و همکاران (۲۰۲۳) به بررسی ۳۲ پژوهش در حوزه پیش‌بینی سری‌های زمانی مالی با استفاده از روش‌های یادگیری ماشین و مدل‌های استخراج ویژگی یا انتخاب ویژگی پرداختند. آنها با تشریح انواع روش‌های ترکیب مدل‌های یادگیری ماشین و انتخاب ویژگی/استخراج ویژگی بیان کردند که پرکاربردترین مدل‌های یادگیری ماشین برای پیش‌بینی قیمت، ماشین بردار پشتیبان و الگوریتم جنگل تصادفی بوده‌اند. همچنین تحلیل همبستگی، تحلیل مؤلفه اساسی و خود رمزگذار[۸] پرکاربردترین روش‌های انتخاب و استخراج ویژگی بوده‌اند[۱۲].

در پژوهش‌های گذشته در حوزه پیش‌بینی سری‌های زمانی، از روش‌هایی استفاده شده است که هدف آنها صرفاً به حداقل رساندن خطای پیش‌بینی بوده و تنها یک خروجی به‌عنوان تخمین نهایی ارائه می‌نمایند و هیچ توضیحی در خصوص عدم قطعیت پیش‌بینی و یا اینکه احتمال وقوع پیش‌بینی‌های انجام شده به چه شکلی است قابل استنتاج نمی‌باشد. در سال‌های اخیر تفسیر نتایج مدل‌های یادگیری ماشین از اهمیت روزافزونی برخوردار شده است و مقادیر SHAP ابزاری بسیار کارآمد در این حوزه است، اما در هیچ یک از پژوهش‌های داخلی از این ابزار برای تفسیر نتایج مدل‌های یادگیری ماشین استفاده نشده است.

[1] Gandhmal & Kumar
[2] Sezer
[3] Thakkar & Chaudhari
[4] Deep Q-network model
[5] Kou
[6] Celik
[7] Local Interpretable model-agnostic explanations
[8] Autoencoder

۳. روش‌شناسی پژوهش

در پژوهش‌های حوزه پیش‌بینی، ثابت شـده که انتخاب ورودی‌های مؤثر از اهمیت بسـیار زیادی برخوردار اسـت. به‌طورکلی این داده‌ها می‌تواند شـامل داده‌های کمی، گزارش‌های شـرکت‌ها و شـاخص‌های مربوط به احسـاسـات سـرمایه‌گذاران باشـند[۲۰]. به همین جهت در گام اول، به‌منظور انتخاب ورودی‌های مناسـب برای مدل پیشـنهادی، مروری سیستماتیک بر روی پژوهش‌های انجام‌گرفته در این حوزه صورت پذیرفته است.

در بخش دوم با استفاده از روش‌های کمی به پیش‌بینی شاخص کل بورس اوراق بهادار تهران پرداخته شده است. برای بررسـی عملکرد مدل پیشـنهاد شـده از داده‌های شـاخص کل بورس اوراق بهادار تهران از سـال ۱۳۸۹ تا ۱۴۰۳ استفاده شده است. در نهایت به‌منظور ارزیابی عملکرد مدل پیشنهادی، از معیارهای MAE، RMSE و MAPE استفاده شده و نتایج آن با مدل‌های MLP، AdaBoost و XGBoost مقایسه شده است.

به‌منظور تعیین ورودی‌های مدل از مرور سـیسـتماتیک پژوهش‌های انجام شـده در این حوزه، از طریق پایگاه علمی اسکوپوس[1] با قیود زیر استفاده شده است:

جدول ۱ *. قیدهای استفاده شده برای مرور سیستماتیک*

عنوان	Stock market prediction Or Stock market forecasting
عنوان، کلمات کلیدی، چکیده	And not Carbon, copper, news, social media, oil, text, stock price
حوزه	Limited to: Computer science Economics, econometrics, finance Business, management, accounting
دوره زمانی	2020- present time
زبان	English

نتیجه این جست‌وجو ۱۲۳ مقاله بوده است که از این میان، یک مقاله به دلیل در دسترس نبودن از طریق پایگاه‌های علمی، ۵ مقاله به دلیل اسـتفاده از مدل‌های تحلیل متن و پیام، ۹ مقاله به دلیل بی ارتباط بودن و یا کیفیت نامناسـب پژوهش و ۴۵ مقاله به دلیل اسـتفاده از داده‌های با تواتر دقیقه‌ای کنار گذشـته شـده‌اند. در نهایت ۶۳ مقاله که از داده‌های روزانه بهره برده بودند انتخاب شدند.

با استخراج تمام ویژگی‌های مورد استفاده در این پژوهش‌ها که عمدتاً اندیکاتورهای تکنیکال هستند، ویژگی‌هایی را که اطلاعات مربوط به آنها به شـکل روزانه در دسـترس یا قابل محاسـبه هسـتند به‌عنوان ورودی انتخاب شـده‌اند. در این پژوهش از قیمت دلار بازار آزاد به‌عنوان شـاخص قیمت برابری ارز کشـور میزبان در برابر ارزهای خارجی، از میانگین موزون نرخ اوراق خزانه اسـلامی به‌عنوان شـاخص نرخ بهره و از قیمت سـکه رفاه به‌عنوان شـاخص قیمت کامودیتی‌ها استفاده شده است. با توجه به اینکه ضـریب همبسـتگی پیرسون[2] برای متغیرهای کمترین[3]، بازگشـایی[4]، بیشـترین[5] و پایانی[6] معادل ۰.۹۹۹ اسـت، مطابق یانز[7] و همکاران (۲۰۲۴) [۳۱] فقط از ویژگی قیمت پایانی به‌جای همه آنها استفاده شده است. ویژگی‌های مورد استفاده در این پژوهش در جدول ۲ ارائه شده است:

[1] Scopus
[2] Pearson Correlation Coefficient
[3] Low
[4] Open
[5] High
[6] Close
[7] Yanez

جدول ۲. ویژگی‌های استخراج شده از پژوهش‌های قبلی برای به کارگیری در این پژوهش

ردیف	ویژگی	ردیف	ویژگی
۱	previous closing prices (prev_close)	۱۸	On Balance Volume (OBV)
۲	trading volume	۱۹	Money Flow Index (MFI)
۳	Simple Moving Average (SMA)	۲۰	Stop And Reverse (SAR)
۴	Moving Average Convergence Divergence (MACD)	۲۱	Standard Deviation (SD)
۵	Relative Strength Index (RSI)	۲۲	Accumulation/Distribution Oscillator (ADO)
۶	Commodity Channel Index (CCI)	۲۳	DirectionalI ndex (DX)
۷	Exponential Moving Average (EMA)	۲۴	Kaufman's Adaptive Moving Average (KAMA)
۸	Momentum (MOM)	۲۵	Triple Exponential Moving Average Oscillator (TRIX)
۹	price change	۲۶	Percentage Price Oscilator (PPO)
۱۰	Stochastic %K (STCK)	۲۷	Time Series Forecast (TSF)
۱۱	Average Directional Index (ADX)	۲۸	Triple Exponential Moving Average (TEMA)
۱۲	william%R	۲۹	Force Index (FI)
۱۳	Rate of Change (ROC)	۳۰	Year High
۱۴	Weighted Moving Average (WMA)	۳۱	Year Low
۱۵	Stochastic %D (STCD)	۳۲	Interest Rate (IR)
۱۶	Average True Range (ATR)	۳۳	USD/IRR (Dollar)
۱۷	Bolinger Band (BB)	۳۴	Gold

تقویت گرادیان طبیعی[1] برای پیش‌بینی: هدف پیش‌بینی در ساختار استاندارد روش‌های یادگیری ماشین، تخمین یک مقدار اسکالر $E[y \mid x]$ است، جایی که x بردار ویژگی‌های مشاهده شده و y هدف پیش‌بینی است. در این روش محققان به دنبال دستیابی به یک توزیع احتمال به شکل $P_\theta(y \mid x)$ (با تابع توزیع تجمعی F_θ) هستند. در واقع در این رویکرد فرض می‌شود که $P_\theta(y \mid x)$ یک شکل پارامتریک دارد و باید پارامترهای P ($\theta \in R^p$) از توزیع را به‌عنوان تابعی از x تخمین زد.

قاعده امتیازدهی: برای شروع کار نیاز به یک هدف برای فرایند یادگیری است. در مدل‌های پیش‌بینی نقطه‌ای، پیش‌بینی‌ها با داده هدف در قالب یک تابع که به آن تابع هزینه[2] می‌گویند مقایسه می‌شوند. مشابه این موضوع در ادبیات رگرسیون احتمالی قاعده امتیازدهی[3] نامیده می‌شود که توزیع احتمال تخمینی را با مشاهدات مقایسه می‌کند.

قاعده امتیازدهی صحیح S، یک توزیع احتمال پیش‌بینی شده P و یک مشاهده y (خروجی) را به‌عنوان ورودی گرفته و یک امتیاز $S(P, y)$ به پیش‌بینی انجام شده می‌دهد به‌طوری‌که توزیع واقعی مشاهدات بالاترین امتیاز را در این روش دریافت می‌کند[۹]. به بیان ریاضی، قاعده امتیازدهی S یک قاعده صحیح امتیازدهی است اگر و فقط اگر

$$\forall P, Q \; E_{y \sim Q}[S(Q, y)] \le E_{y \sim Q}[S(P, y)] \qquad (۱)$$

به‌طوری‌که Q توزیع واقعی، خروجی‌ها y و P هر توزیع دیگر است (مثلاً خروجی یک مدل پیش‌بینی احتمالی). از آنجایی که در اینجا یک توزیع پارامتریک وجود دارد، بنابراین می‌توان هر توزیع را بر حسب پارامترهای آن بیان کرد و امتیاز را به شکل $S(\theta, y)$ نشان داد. پرکاربردترین روش امتیازدهی، امتیازدهی لگاریتمی $\mathcal{L}$ است که وقتی بهینه‌یابی می‌شود همان تخمین حداکثر درست‌نمایی را به دست می‌دهد.

$$\mathrm{L}(\theta, y) = -\log P_\theta(y) \qquad (۲)$$

[1] Natural Gradient Boosting
[2] Loss Function
[3] Scoring Rule

گرادیان طبیعی تعمیم‌یافته[1]: در زمانی که از گرادیان نزولی استاندارد برای یافتن پارامترهایی که قاعده امتیازدهی را کمینه می‌کنند استفاده شود، در جهت منفی گرادیان تابع نسبت به پارامترها در هر نقطه x حرکت می‌شود. گرادیان معمولی قاعده امتیازدهی S در زمانی که تابع توزیع پارامتری P_θ که برحسب پارامترهای آن θ توصیف می‌شود و y خروجی است به شکل $\nabla S(\theta, y)$ نمایش داده می‌شود. در واقع جهت این گرادیان، جهت بیشترین افزایش (با حرکت بی نهایت کوچک در هر پارامتر در این جهت) قاعده امتیازدهی است که به شکل زیر بیان می‌شود:

$$\nabla S(\theta, y) \propto \lim_{\epsilon \to 0} \arg\max_{d:\|d\|=\epsilon} S(\theta + d, y) \qquad (۳)$$

این گرادیان نسبت به پارامتری‌سازی مجدد ثابت نیست. اگر فرض شود که P_θ به شکل $P_{z(\theta)}(y)$ پارامتری‌سازی شود به شکلی که $P_\theta(y \in A) = P_\psi(y \in A)$ برای تمام حالات A که $\psi = z(\theta)$، در این حالت اگر گرادیان نسبت به θ محاسبه شود، یعنی گام‌های بی نهایت کوچک در جهت θ به $\theta + d\theta$، توزیع به دست آمده نسبت به حالتی که گرادیان نسبت به ψ محاسبه می‌شد متفاوت خواهد بود. به عبارت بهتر:

$$P_{\theta+d\theta}(y \in A) \neq P_{\psi+d\psi}(y \in A) \qquad (۴)$$

یعنی انتخاب نحوه پارامتری‌سازی به شکل قابل‌ملاحظه‌ای جهت حرکت برای کمینه‌سازی تابع هدف را تحت‌تأثیر قرار می‌دهد و این در حالی است که حداقل تابع موردنظر تغییری نکرده است. مسئله این است که فاصله بین دو متغیر دقیقاً معادل فاصله دو توزیع که این دو متغیر آن را توصیف می‌کنند نیست. در واقع این موضوع منجر به پیدایش گرادیان طبیعی (که با $\tilde{\nabla}$ نشان داده می‌شود) در هندسه اطلاعات شد[۲].

در فضای توزیع‌ها، دیورژانس می‌تواند به‌عنوان یک معیار فاصله در نظر گرفته شود.

همان‌طور که بیان شد، هر قاعده امتیازدهی مناسب در رابطه (۱) صدق می‌کند. بنابراین:

$$D_S(Q \| P) = E_{y \sim Q}[S(P, y)] - E_{y \sim Q}[S(Q, y)] \qquad (۵)$$

همیشه بزرگ‌تر مساوی صفر بوده و می‌تواند به‌عنوان معیار فاصله توزیع P و توزیع Q از یکدیگر در نظر گرفته شود. بدین ترتیب دیورژانس، مستقل از شکل پارامتری‌سازی توزیع‌ها، می‌تواند به‌عنوان فاصله دو توزیع به کار برود. اگر چه به‌طورکلی دیورژانس‌ها متقارن نیستند، اما برای تغییرات با مقادیر بسیار کوچک می‌توان آنها را متقارن در نظر گرفت و از آنها به‌عنوان معیار فاصله استفاده کرد. در این کاربرد دیورژانس یک خمینه آماری[2] نتیجه می‌دهد که هر نقطه در این خمینه معادل یک توزیع احتمال است[۶].

گرادیان طبیعی تعمیم‌یافته در فضای ریمانی، جهت حداکثر افزایش را نشان می‌دهد که نسبت به شیوه پارامتری‌سازی مستقل است:

$$\tilde{\nabla} S(\theta, y) \propto \lim_{\epsilon \to 0} \arg\max_{d:D_S(P_\theta \| P_{\theta+d})=\epsilon} S(\theta + d, y) \qquad (۶)$$

اگر این مسئله بهینه‌سازی حل شود گرادیان طبیعی به شکل زیر قابل بیان است:

$$\tilde{\nabla} S(\theta, y) \propto \mathrm{I}_S(\theta)^{-1} \nabla S(\theta, y) \qquad (۷)$$

که $\mathrm{I}_S(\theta)$ معیار ریمانی از خمینه آماری در نقطه θ است که توسط قاعده امتیازدهی نتیجه شده است. با انتخاب $S = \mathrm{L}$ (تابع حداکثر درست نمایی) و حل مسئله بهینه‌سازی بالا:

$$\tilde{\nabla} \mathrm{L}(\theta, y) \propto \mathrm{I}_\mathrm{L}(\theta)^{-1} \nabla \mathrm{L}(\theta, y) \qquad (۸)$$

$$\mathrm{I}_\mathrm{L}(\theta) = E_{y \sim P_\theta}[\nabla_\theta \mathrm{L}(\theta, y) \nabla_\theta \mathrm{L}(\theta, y)^T] \qquad (۹)$$

[1] The Generalized Natural Gradient
[2] Statistical Manifold

استفاده از گرادیان طبیعی در فرایند یادگیری منجر می‌شود که یادگیری باثبات بیشتر و بهره‌وری بیشتری صورت پذیرد[۲].

تقویت گرادیان طبیعی: تقویت گرادیان[1] یک روش یادگیری نظارت شده است که در آن تعداد زیادی یادگیرنده ضعیف[2] در یک فرایند تجمعی ترکیب می‌شوند. در این روش، مدل طبق یک فرایند متوالی آموزش می‌بیند به شکلی که هر یادگیرنده به‌وسیله مانده گروه یادگیرنده‌های قبلی آموزش می‌بیند. خروجی یادگیرنده آموزش‌دیده با استفاده از یک پارامتر به نام نرخ یادگیری[3] مقیاس شده و به گروه یادگیرنده‌های قبلی اضافه می‌شود. این فرایند توسط هر یادگیرنده ضعیفی قابل پیاده‌سازی است اما معمولاً از درخت تصمیم برای این موضوع استفاده می‌شود زیرا که در عمل، نتیجه مناسبی به همراه دارد[۴].

در فرایند آموزش درخت تصمیم، داده‌ها به بخش‌های مختلف تقسیم می‌شوند که هر بخش در یک برگ درخت تحلیل شده (همه داده‌ها نهایتاً در یک برگ قرار می‌گیرند) و مقدار پیش‌بینی شده توسط هر برگ به پیش‌بینی کلی اضافه شده و این فرایند به شکلی صورت می‌پذیرد که تابع خطا را کمینه نماید.

الگوریتم تقویت گرادیان طبیعی (NGBoost) یک روش یادگیری نظارت شده برای پیش‌بینی است که از تقویت گرادیان برای تخمین پارامترهای توزیع احتمال شرطی $P(y \mid x)$ به‌عنوان تابعی از x استفاده می‌نماید. به‌طورکلی الگوریتم از سه جز تشکیل شده است که پیش از اجرای آن باید آنها را تعیین کرد:

۱- یادگیرنده پایه

۲- توزیع احتمال پارامتری P_θ

۳- قاعده امتیازدهی مناسب

برای یک ورودی جدید x، پیش‌بینی $y \mid x$ از یک تابع توزیع شرطی P_θ حاصل می‌شود که پارامترهای θ آن از ترکیب جمعی خروجی M یادگیرنده پایه به دست آمده است؛ به طور مثال اگر توزیع نرمال استفاده شود: $\theta = (\mu, \log\sigma)$

برای بدست آوردن پارامترهای θ برای یک x هر کدام از یادگیرنده‌های پایه f^m، مقدار x را به‌عنوان ورودی دریافت می‌کند. منظور از f^m مجموعه یادگیرنده‌های ضعیف در مرحله m برای هر پارامتر است. به طور مثال برای توزیع نرمال با پارامترهای μ و $\log\sigma$ دو یادگیرنده پایه به شکل f^m_μ و $f^m_{\log\sigma}$ وجود خواهد داشت که به شکل $f^m = (f^m_\mu, f^m_{\log\sigma})$ بیان می‌شود. پیش‌بینی‌های انجام شده با فاکتور مقیاس هر مرحله $\rho^{(m)}$ و نرخ یادگیری η مقیاس می‌شوند.

$$y \mid x \sim P_\theta(x), \quad \theta = \theta^0 - \eta\sum_{m=1}^{M}\rho^{(m)}.f^{(m)}(x) \qquad (۱۰)$$

مدل در هر مرحله با استفاده از یادگیرنده‌های پایه و فاکتور مقیاس هر مرحله به شکل متوالی آموزش می‌بیند. الگوریتم یادگیری در مرحله اول با تخمین‌های اولیه‌ای از پارامترها آغاز می‌شود به طوری که نتیجه قاعده امتیازدهی در این مرحله برای تمام داده‌های آموزش کمینه شود. این مقادیر اولیه برای تمام پارامترها خواهد بود. سپس در هر مرحله m، الگوریتم برای هر نمونه i مقدار گرادیان طبیعی $g_i^{(m)}$ قاعده امتیازدهی را نسبت به پارامترهای تخمین زده شده در مرحله قبل $\theta_i^{(m-1)}$ محاسبه می‌کند. باید توجه داشت که ابعاد $g_i^{(m)}$ با ابعاد θ برابر است. یک گروه از یادگیرنده‌های پایه برای آن مرحله f^m برای پیش‌بینی اجزای $g_i^{(m)}$ برای هر x_i به کار می‌روند. خروجی یادگیرنده

[1] Gradient Boosting
[2] Weak Learner or Base Learner
[3] Learning Rate

پایه در هر طبقه برای محاسبه گرادیان طبیعی در هر مرحله مورد استفاده قرار می‌گیرد. سپس گرادیان طبیعی محاسبه شده با استفاده از فاکتور مقیاس $\rho^{(m)}$، تعدیل می‌شود زیرا که تخمین‌های زده شده محلی بوده و در فضای بسیار دورتر از پارامترهای فعلی ممکن است کاملاً درست نباشند. فاکتور مقیاس به گونه‌ای انتخاب می‌شود که با حرکت در جهت گرادیان طبیعی به حداقل‌سازی قاعده امتیازدهی بپردازد. پس از مشخص شدن فاکتور مقیاس $\rho^{(m)}$، پارامترهای پیش‌بینی شده برای هر نمونه $\theta_i^{(m)}$ با کم کردن گرادیان طبیعی مقیاس شده از پارامترهای مرحله قبل $\theta_i^{(m-1)}$ پیش‌بینی می‌گردد که البته این مقدار با نرخ یادگیری کم η (معمولاً ۰.۱ یا ۰.۰۱) تعدیل می‌گردد.

SHAP: همان‌طور که بیان شد روش‌های یادگیری ماشین پتانسیل بالایی برای پژوهش‌ها در حوزه مالی دارند؛ اما عدم توانایی در تفسیر نتایج مانع از استفاده جامع از ظرفیت‌های این روش‌ها شده است. برای حل این موضوع لاندبرگ و لی[1] (۲۰۱۷) رویکرد SHAP را برای توضیح و تفسیر خروجی‌های مدل‌های یادگیری ماشین معرفی کردند[۱۷]. SHAP به کاربران کمک می‌کند تا نتایج مدل‌های پیچیده را تفسیر نمایند. این روش در ابتدا در سال ۱۹۵۳ توسط شپلی بر اساس تئوری بازی‌ها معرفی شد[۲۵].

در این رویکرد خروجی مدل به شکل رابطه‌ای خطی از متغیرهای ورودی توضیح داده می‌شود. به بیان دیگر مقدار عددی به دست آمده این روش $g(z')$ برای مدل $f(x)$ به وسیله تجزیه خروجی مدل به یک تابع خطی باینری $z' \in \{0,1\}^M$ از مقادیر SHAP $\phi_i \in R$ محاسبه می‌شود.

$$f(x) = g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \qquad (۱۱)$$

که در آن M تعداد متغیرهای ورودی، z' شکل ساده شده متغیرهای ورودی و ϕ_0 مقدار ثابتی است که از حذف اثر تمام متغیرهای دیگر محاسبه می‌شود. به طور کلی با استفاده از SHAP جهت اثر هر یک از ویژگی‌های ورودی قابل محاسبه است.

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!(M-|z'|-1)!}{M!}[f_x(z') - f_x(z' \setminus i)] \qquad (۱۲)$$

که در اینجا x' بردار متغیرهای ورودی، f مدل آموزش دیده و $[f_x(z') - f_x(z' \setminus i)]$ مقدار اثر ویژگی‌ها بدون اثر ویژگی iام و z' اثر ورودی‌های حاضر در مدل است[۱۷].

با استفاده از SHAP اثر هر ویژگی را بر روی خروجی به طور محلی[2] (برای هر خروجی مشخص) و یا کلی[3] می‌توان محاسبه کرد. بدین ترتیب می‌توان با درنظرگرفتن ترکیب همبستگی‌های ورودی و برهم‌کنش بین ویژگی‌ها، عوامل اثرگذار و شدت اثرگذاری هر یک از عوامل را بر روی خروجی و جهت آن را تشخیص داد. با توجه به قابلیت تعاملی SHAP می‌توان مسئله رابطه بین ویژگی‌ها را نیز مورد بررسی و تحلیل قرار داد. با استفاده از مقدار SHAP برای ویژگی‌های ورودی می‌توان دست به انتخاب ویژگی زد. هر ویژگی که اثرگذاری کمی در تخمین خروجی دارد، از بار اطلاعاتی مفید کمتری برخوردار بوده و قابل حذف است.

٤. تحلیل داده‌ها و یافته‌ها

معیارهای ارزیابی استفاده شده برای بررسی خطاهای مدل‌ها و مقایسه آنها شامل MAE، RMSE و MAPE می‌باشد که جدول ۳ مقادیر خطاها را برای داده‌های آموزش و داده‌های تست نشان می‌دهد. در اینجا

[1] Lundberg & Lee
[2] Local explanation
[3] Global explanation

فرض شـده خروجی مدل دارای توزیع نرمال بوده که این توزیع با پارامترهای میانگین و انحراف معیار قابل توصـیف است. در واقع مقدار پیش‌بینی مدل ارائه شده همان میانگین توزیع خروجی است.

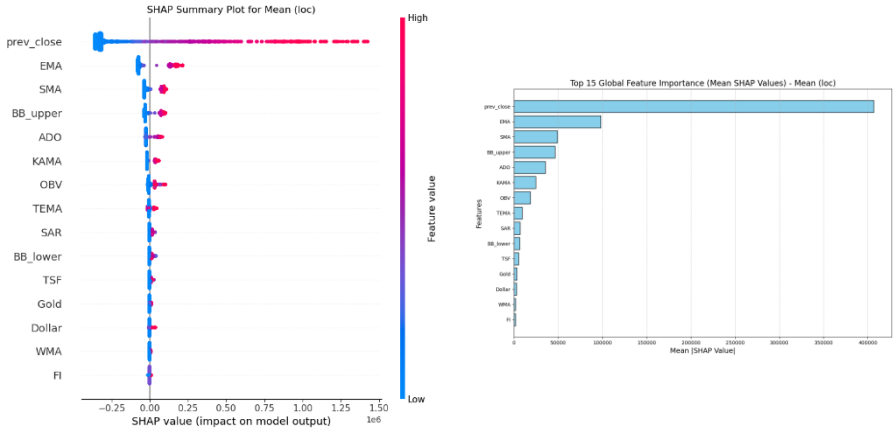
جدول ۳ *. خطای پیش‌بینی مدل ارائه شده و مقایسه آن با سایر مدل‌ها*

	داده‌های آموزش				
مدل	**MAE**	**MAE (%)**	**RMSE**	**RMSE (%)**	**MAPE (%)**
NGB	۳،۹۰۱	۰/۶۴	۷،۷۰۶	۱/۲۷	۱/۲۹
MLP	۱۳،۵۹۷	۲/۲۴	۲۰،۱۸۱	۳/۳۳	۱۰/۹۵
XGB	۸۴۵	۰/۱۴	۱،۳۱۷	۰/۲۲	۰/۶۵
AdaBoost	۸،۹۶۸	۱/۴۸	۱۳،۱۸۰	۲/۱۸	۷/۵۲
	داده‌های تست				
مدل	**MAE**	**MAE (%)**	**RMSE**	**RMSE (%)**	**MAPE (%)**
NGB	۶،۲۲۵	۱/۰۰	۱۴،۰۲۰	۲/۲۶	۱/۳۱
MLP	۱۳،۱۷۲	۲/۱۲	۱۹،۴۴۲	۳/۱۳	۱۰/۰۰
XGB	۶،۷۴۹	۱/۰۹	۱۵،۲۷۰	۲/۴۶	۱/۳۴
AdaBoost	۹،۶۴۳	۱/۵۵	۱۵،۶۳۲	۲/۵۲	۶/۴۶

همان‌طور که در جدول ۳ نشان داده شده است، مدل NGBoost در پیش‌بینی داده‌های شاخص نسبت به بقیه روش‌ها کمترین میزان خطا را نشـان داده اسـت. این موضـوع از آن جهت اهمیت دارد که در مدل‌های معمول، تابع هدف به شـکلی تعریف می‌شـود که فاصـله مقدار پیش‌بینی از مقدار هدف کمینه شـود. در حالی که در NGBoost هدف کمینه کردن فاصـله توزیع احتمال خروجی مدل از مقدار هدف اسـت. در واقع این مدل نه فقط امکان بررسـی توزیع خروجی را فراهم می‌آورد بلکه حتی خطای پیش‌بینی آن در اسـتفاده از داده‌های شـاخص کل بورس تهران از روش‌های دیگر کمتر است.

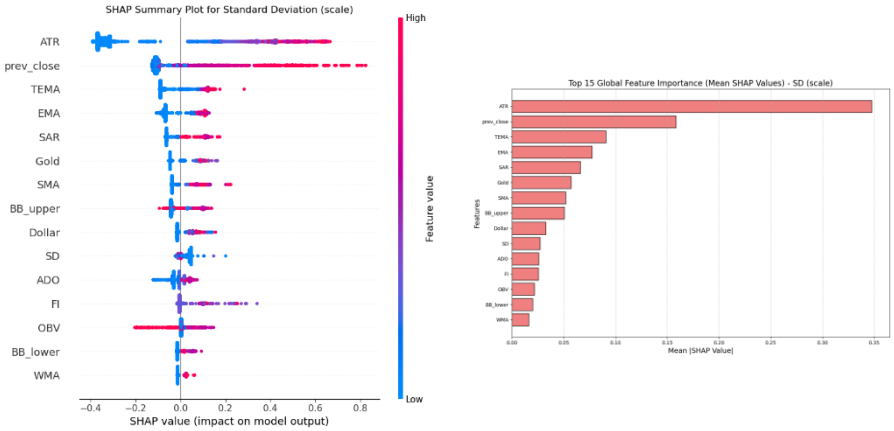
تفسـیر ویژگی‌ها: مقادیر SHAP برای تفسـیر اثرگذاری ورودی‌ها در تشـکیل مقدار خروجی به کار می‌روند که این مقادیر به کاربران امکان می‌دهد اهمیت هر یک از ویژگی‌ها در تشکیل خروجی را ارزیابی نمایند.

در نمودار ۱ مقادیر SHAP برای پیش‌بینی مدل با استفاده از NGBoost نشان داده شده است.



الف) متوسط اثرگذاری هر ویژگی بر مقدار میانگین توزیع احتمال تخمینی

ب) مقادیر SHAP (برای مقدار میانگین توزیع احتمال تخمینی)



د) مقادیر SHAP (برای مقدار انحراف معیار توزیع احتمال تخمینی)

ج) متوسط اثرگذاری هر ویژگی بر مقدار انحراف معیار توزیع احتمال تخمینی

نمودار ۱ *.مقادیر SHAP برای ویژگی‌های ورودی در پیش‌بینی با استفاده از NGBoost*

در نمودار ۱–الف، ۱۵ ویژگی ورودی که بر اساس متوسط مقادیر SHAP آنها، بیشترین تأثیر بر تشکیل مقدار میانگین توزیع احتمال تخمین زده شده را داشتند نشان داده شده است. همانطور که مشاهده می‌شود قیمت پایانی، بالاترین اهمیت را در تخمین شاخص برای روز بعد برخوردار است. همچنین اندیکاتورهای EMA، SMA، باند بالایی بولینگر و ADO در رده‌های بعدی به لحاظ اهمیت در پیش‌بینی خروجی قرار دارند. نمودار ۱–ب مقدار اثرگذاری هر یک از این ویژگی‌ها را در هر یک از نقاط نشان می‌دهد. در این نمودار هرچقدر مقدار نسبی همان متغیر بالاتر باشد رنگ آن قرمزتر و هر چقدر مقدار نسبی آن متغیر کمتر باشد رنگ آن نقطه آبی‌تر نشان داده شده است.

نکته بعدی در خصوص اثرگذاری هر یک از ورودی‌ها در تشکیل انحراف معیار توزیع تخمین زده شده است. نمودار ۱–ج، متوسط مقادیر SHAP هر یک ویژگی‌ها را بر اساس اثرگذاری در تشکیل انحراف معیار توزیع تخمین زده شده نشان می‌دهد. هر چقدر که انحراف معیار توزیع تخمین زده شده کمتر باشد، احتمال انحراف از مقادیر تخمین زده شده کمتر و پیش‌بینی قابل اتکاتر خواهد بود. در طرف مقابل هر چقدر مقدار انحراف معیار توزیع تخمین زده شده بیشتر باشد، احتمال انحراف از مقادیر پیش‌بینی شده بیشتر بوده و قابلیت اتکای کمتری دارد. مطابق نمودار ۱–د، لندیکاتور ATR، قیمت پایانی، لندیکاتور TEMA، EMA و SAR بیشترین اثرگذاری بر مقدار انحراف معیار خروجی مدل را دارند. با توجه به اینکه نقاط آبی به معنی مقادیر نسبی کم متغیر و نقاط قرمز نشان دهنده مقادیر نسبی بالای متغیر است هر چقدر که مقدار نسبی اندیکاتور ATR بیشتر باشد مقدار انحراف معیار توزیع تخمین زده شده بالاتر خواهد بود. در طرف مقابل برای اندیکاتور OBV، نقاط قرمز در سمت چپ نمودار قرار دارد، یعنی مقادیر نسبی بالا برای این ویژگی، مقادیر SHAP منفی داشته و در جهت کاهش انحراف معیار و افزایش قابلیت اتکای مدل حرکت می‌نماید. هرچقدر که مقادیر نسبی OBV کاهش می‌یابد به واسطه افزایش انحراف معیار، قابلیت اتکای تخمین انجام شده نیز کاهش می‌یابد.

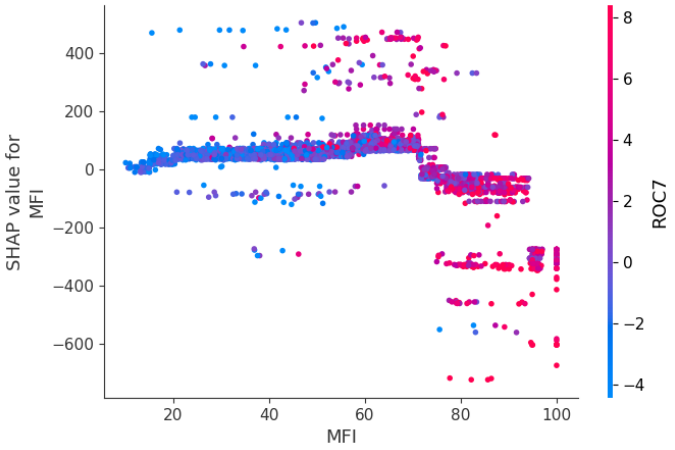
نمودار وابستگی: برای تحلیل بیشتر ارتباط هر یک از ویژگی‌های ورودی و خروجی مدل‌های یادگیری ماشین، می‌توان از نمودارهای وابستگی SHAP استفاده کرد. این نمودارها اثرگذاری هر یک از متغیرها را در تشکیل خروجی مدل نشان می‌دهند.

در نمودار ۲ برای نمونه، نمودار وابستگی اندیکاتور MFI نشان داده شده است. علت انتخاب این اندیکاتور این است که مقادیر آن همواره بین ۰ تا ۱۰۰ متغیر است و اطلاعات ارزشمندی را در اختیار تحلیلگران قرار می‌دهد اما هیچ محدودیتی در رسم نمودارهای وابستگی برای سایر ویژگی‌ها وجود ندارد. در این نمودار رنگ هر یک از نقاط بر

اساس مقادیر بازدهی ۷ روز گذشته آن مشخص شده است. نقاط آبی‌تر نشان‌دهنده بازدهی‌های به نسبت کمتر در ۷ روز و نقاط قرمزتر نشان‌دهنده بازدهی‌های بیشتر ۷ روزه می‌باشد.

در نمودار ۲، یک نوع تفکیک برای مقادیر کمتر از ۷۵ و بیشتر از ۷۵ اندیکاتور MFI مشاهده می‌شود به شکلی که برای مقادیر کمتر از ۷۵ آن، مقادیر SHAP مثبت بوده و برای مقادیر بالاتر از آن مقادیر SHAP منفی و در جهت کاهش خروجی مدل می‌باشد. همچنین در مقادیر بالای MFI تمرکز نقاط قرمز بسیار بیشتر بوده و هر چه MFI افزایش می‌یابد، مقادیر SHAP آن منفی‌تر می‌شود. به بیان دیگر الگوی یادگیری ماشین در زمان‌هایی که مقادیر MFI بالاتر از ۷۵ قرار گرفته است، پیش‌بینی خود را از روز بعد کاهش می‌دهد و هر چقدر مقدار این لندیکاتور به ۱۰۰ نزدیک‌تر باشد، کاهش مقدار تخمینی برای روز بعد نیز شدت می‌یابد. در سمت مقابل هر چه به سمت مقادیر کم MFI حرکت کرده می‌شود تمرکز نقاط آبی بیشتر می‌شود. این موضوع نشان دهنده آن است که مقادیر بالای MFI متناظر بازدهی‌های بالای روزهای گذشته بوده و الگوی یافت شده توسط مدل، مقادیر خروجی را در مقادیر بالای این ویژگی کاهش می‌دهد و بالعکس. این موضوع بدین شکل قابل تفسیر است که مدل ارائه شده بر مبنای داده‌های مورد استفاده، برای مقادیر بالاتر از ۷۵ این اندیکاتور، منطقه اشباع خرید را شناسایی کرده و تخمین خود را از مقدار شاخص روز بعد کاهش می‌دهد. اما برای مقادیر کم این اندیکاتور، تخمین خود را از مقدار شاخص روز بعد افزایش می‌دهد اما الگویی قدرتمند برای شناسایی مناطق اشباع فروش را شناسایی نکرده است.

تفکیک‌های مشابهی را در نمودار وابستگی اندیکاتور RSI برای مقادیر کمتر از ۴۵ این اندیکاتور و همچنین برای مقادیر کمتر از ۳۰ و بیشتر از ۷۰ اندیکاتور ADX نیز می‌توان مشاهده نمود. بررسی جامع همه نمودارهای وابستگی ویژگی‌های ورودی، اطلاعات ارزشمندی را در خصوص به کارگیری هر یک از آنها در اختیار کاربران قرار می‌دهد.



نمودار ۲ *. وابستگی اندیکاتور MFI در تخمین خروجی مدل (رنگ نقاط بر حسب بازدهی ۷ روزه)*

جدول ۴ خطای پیش‌بینی را در حالتی که، از ۱۵ ویژگی که دارای بیشترین اهمیت هستند، استفاده شود (به جای همه ویژگی‌ها) نشان می‌دهد. در این حالت خطای همه مدل‌ها (به جز MLP) با کاهش همراه بوده است. همانطور که در پژوهش‌های متعدد از جمله پژوهش تان[1] و همکاران (۲۰۲۳) اشاره شده است تعداد بسیار زیاد ویژگی‌های ورودی برای مدل‌های یادگیری ماشین چالش برانگیز است و استفاده از ابزارهای انتخاب ویژگی با کاهش مشکلات مربوط به بیش‌برازش مدل، حذف متغیرهای غیرمرتبط و کاهش بار محاسباتی منجر به کاهش خطای پیش‌بینی می‌گردد[۱۲].

[1] Htun

جدول ٤ . خطای پیش‌بینی مدل ارائه و مقایسه با سایر مدل‌ها با استفاده از ۱۵ ویژگی با بیشترین مقادیر *SHAP*

	داده‌های آموزش				
مدل	MAE	MAE (%)	RMSE	RMSE (%)	MAPE (%)
NGB	۳،۹۸۵	۰/۶۶	۷،۹۷۳	۱/۳۲	۱/۰۵
MLP	۱۹،۰۱۴	۳/۱۴	۳۴،۸۱۲	۵/۷۴	۹/۵۴
XGB	۱،۴۴۲	۰/۲۴	۲،۶۴۸	۰/۴۴	۰/۶۵
AdaBoost	۸،۹۷۰	۱/۴۸	۱۳،۱۶۵	۲/۱۷	۷/۴۹
	داده‌های تست				
مدل	MAE	MAE (%)	RMSE	RMSE (%)	MAPE (%)
NGB	۵،۹۷۶	۰/۹۶	۱۳،۵۳۶	۲/۱۸	۱/۱۱
MLP	۱۷،۶۲۲	۲/۸۴	۳۱،۵۵۰	۵/۰۸	۸/۲۷
XGB	۷،۰۰۶	۱/۱۳	۱۶،۱۶۱	۲/۶۰	۱/۱۸
AdaBoost	۹،۵۳۹	۱/۵۴	۱۵،۳۶۲	۲/۴۸	۶/۴۵

۵. بحث و نتیجه‌گیری

در این پژوهش با استفاده از رویکردی جدید در حوزه یادگیری ماشین به پیش‌بینی شاخص کل بورس اوراق بهادار تهران پرداخته شده است. همچنین با استفاده از مقادیر SHAP اثرگذاری ورودی‌های مدل در تشکیل خروجی نهایی مورد بررسی قرار گرفته است که این موضوع به سرمایه‌گذاران امکان می‌دهد تا علاوه بر داشتن تخمین بهتری از روند شاخص کل، با تحلیل و بینش بهتری در انتخاب ورودی‌ها و نحوه تبدیل ورودی‌های مدل به خروجی‌ها تصمیم‌گیری نمایند. در ابتدا برای انتخاب ورودی‌های مدل، پژوهش‌های مرتبط با این موضوع در ۵ سال گذشته مورد بررسی قرار گرفته و همه ورودی‌هایی که داده‌های آنها به شکل روزانه در بازار سرمایه کشور در دسترس هستند مورد استفاده قرار گرفتند. سپس با به‌کارگیری تقویت گرادیان طبیعی، به پیش‌بینی شاخص کل بورس اوراق بهادار تهران پرداخته و نشان داده شده است که خطای این مدل نسبت به مدل‌های یادگیری ماشین پیشتاز در این حوزه شامل پرسپترون چندلایه، AdaBoost و XGBoost کمتر می‌باشد. این در حالی است که در این مدل تابع هدف، فقط کمینه‌سازی خطای پیش‌بینی مدل نبوده و هدف کمینه‌سازی فاصله توزیع احتمال تخمینی از مقادیر واقعی شاخص می‌باشد. در ادامه با استفاده از مقادیر SHAP، اثرگذاری هر یک از ویژگی‌های ورودی در تشکیل خروجی مورد بررسی قرار گرفته است. تفسیر خروجی‌های مدل پیشنهادی در این پژوهش، با استفاده از مقادیر SHAP نشان می‌دهد که به ترتیب قیمت پایانی، اندیکاتورهای EMA، SMA، بلند بالایی بولینگر و ADO بیشترین اثرگذاری را در تشکیل مقدار پیش‌بینی شده دارند. این نتایج هم‌راستا با نتایج پژوهش کومار[1] و همکاران (۲۰۲۴) است که با استفاده از LSTM به پیش‌بینی چندین شاخص از جمله شاخص S&P500 پرداختند و با استفاده از LIME اثرگذاری هر یک از ویژگی‌ها را مورد بررسی قرار دادند و بیان داشتند که اثرگذارترین ویژگی در تشکیل خروجی و تخمین شاخص، قیمت پایانی است[۱۶]. از طرف دیگر شیخ‌زاده و همکاران (۱۴۰۱) در پژوهش خود بدون رتبه‌بندی اهمیت ویژگی‌های ورودی، بیان داشتند که به‌طورکلی اندیکاتورهایی مانند RSI، CCI، OBV و Williams%R احتمالاً اثرگذاری بیشتری در پیش‌بینی شاخص کل بورس اوراق بهادار تهران دارند که با نتایج این پژوهش متفاوت است[۲۶].

در ادامه اثرگذاری ویژگی‌های ورودی مدل بر مقدار انحراف معیار توزیع احتمال خروجی مورد بررسی قرار گرفت که نشان داده شد اندیکاتور ATR، قیمت پایانی، اندیکاتورهای TEMA، EMA و SAR بیشترین اثرگذاری را بر

[1] Kumar

انحراف معیار توزیع تخمینی دارند. انحراف معیار بیشتر برای توزیع نرمال به این معنی است که احتمال وقوع مقادیر متفاوت از مقدار میانگین بیشتر خواهد بود و در نتیجه قابلیت اتکای کمتری به پیش‌بینی مدل (میانگین توزیع) وجود خواهد داشت. در ادامه با توجه به محدودیت‌های گزارش‌دهی، مقادیر SHAP برای سه ویژگی RSI، ADX و MFI نیز مورد بررسی قرار گرفت و اثرگذاری هر یک از این ویژگی‌ها در تشکیل خروجی تشریح شد، اما مقادیر SHAP برای همه ویژگی‌ها قابل بررسی بوده و می‌تواند اطلاعات بسیار ارزشمندی را به کاربران ارائه نماید. در نهایت نشان داده شد که با استفاده از ۱۵ ویژگی که بیشترین مقادیر SHAP را دارا بودند خطای پیش‌بینی مدل‌های NGBoost، XGBoost و AdaBoost کاهش می‌یابد.

پیشنهادها و محدودیت‌ها: برای پژوهش‌های بعدی می‌توان رویکردی متفاوت در انتخاب ویژگی‌های ورودی مدل انتخاب کرده و مقادیر SHAP را برای ورودی‌های دیگر محاسبه نمود که این موضوع می‌تولند اطلاعات ارزشمندی را در نحوه اثرگذاری ویژگی‌های مختلف بر تشکیل خروجی نشان دهد. این ابزار می‌تواند به مدیران سرمایه‌گذاری کمک نماید تا اثرگذاری عوامل مختلف را بهتر شناخته و تصمیمات بهتری اتخاذ نمایند. بررسی عملکرد مدل پیشنهادی در این پژوهش بر روی داده‌های قیمت سهام، شاخص‌های صنایع و یا قیمت صندوق‌های سرمایه‌گذاری، اطلاعات ارزشمندی از اثرگذاری و دینامیک موجود بین عوامل مختلف در بازار سرمایه را نمایان خواهد ساخت. همچنین بررسی ویژگی‌های توزیع احتمال تخمینی می‌تولند از جنبه‌های دیگری از جمله تخمین ریسک مورد ارزیابی قرار گیرد. به‌طورکلی این مدل در کنار مقادیر SHAP می‌تواند قابلیت‌های گسترده‌ای را در حوزه‌هایی مانند مدیریت ریسک، استراتژی‌های معاملاتی و بهینه‌سازی پرتفو ایجاد نماید. نهادهای نظارتی با استفاده از مدل‌های توضیح‌پذیر می‌توانند اثرگذاری هر یک از عوامل بر شاخص کل را بررسی نموده و با دیدی بهتر به تنظیم ریزساختارهای بازار سرمایه بپردازند.

تعارض منافع: برای ارائه مطلب و نگارش این مقاله هیچ‌گونه کمک مالی از هیچ فرد، نهاد و سازمانی دریافت نشده است و نتایج و دستاوردهای این مقاله به نفع یا ضرر سازمان یا فردی خاص نخواهد بود. حضور نویسندگان در این پژوهش به عنوان شاهدی بی‌طرف ولی متخصص بوده است و نویسندگان هیچ‌گونه تعارض منافعی ندارند.

Reference

1. Afsharirad, E., Alavi, S. E., & Sinaei, H. (2018). Developing an Intelligent Model to Predict Stock Trend Using the Technical Analysis. *Financial Research Journal*, *20*(2), 249-264.
2. Amari, S.-I. (1998). Natural gradient works efficiently in learning. *Neural computation*, *10*(2), 251-276 .
3. Çelik, T. B., İcan, Ö., & Bulut, E. (2023). Extending machine learning prediction capabilities by explainable AI in financial time series prediction. *Applied Soft Computing ,109876* ,132.
4. Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining ,
5. Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees .
6. Dawid, A. P., & Musio, M. (2014). Theory and applications of proper scoring rules. *Metron*, *72*(2), 169-183 .

7. Enke, D., & Thawornwong, S. (2005). The use of data mining and neural networks for forecasting stock market returns. *Expert Systems with applications*, *29*(4), 927-940 .
8. Gandhmal, D., & Kumar, K. (2019). Systematic analysis and review of stock market prediction techniques. Comput Sci Rev 34: 100190. In.
9. Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, *102*(477), 359-378 .
10. Haj Seyed Javady, S. M. R., heydari, r., & Abbasi, F. (2023). Forecasting the future price of pistachio in agricultural commodity exchange using of the hybrid model of Wavelet-XGBoost. *Agricultural Economics*, *17*(1), 79-108.
11. Heidari, M., & Amiri, H. (2022). Inspecting the Predictive Power of Artificial Intelligence Models in Predicting the Stock Price Trend in Tehran Stock Exchange. *Financial Research Journal*, *24*(4), 602-623.
12. Htun, H. H., Biehl, M., & Petkov, N. (2023). Survey of feature selection and extraction techniques for stock market prediction. *Financial Innovation*, *9*(1), 26 .
13. Jabeur, S. B., Mefteh-Wali, S., & Viviani, J.-L. (2024). Forecasting gold price with the XGBoost algorithm and SHAP interaction values. *Annals of Operations Research*, *334*(1), 679-699 .
14. Kianizadeh, H., Baghani, A., & hamidian, m .(2023) .Comparing the accuracy of selected Machin learning models for stock price prediction in stock exchange market. *Journal of Securities Exchange*, *16*(62), 75-102.
15. Kou, G., Xu, Y., Peng, Y., Shen, F., Chen, Y., Chang, K., & Kou, S. (2021). Bankruptcy prediction for SMEs using transactional data and two-stage multiobjective feature selection. *Decision Support Systems*, *140*, 113429 .
16. Kumar, P., Hota, L., Tikkiwal, V. A., & Kumar, A. (2024). Analysing Forecasting of Stock Prices: An Explainable AI Approach. *Procedia Computer Science*, *235*, 2009-2016 .
17. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30* .
18. Mohebbi, S ,.Fadaeinejad, M. E., & Hamidizadeh, M. r. (2021). The Proposed Algorithm to Select Appropriate Features for Predicting Tehran Stock Exchange Index. *Financial Management Perspective*, *11*(34), 35-67.
19. Mohebi, S., Fadaeinejad, M. E., Osoolian, M., & Hamidizadeh, M. R. (2022). Feature Selection for the Prediction Model of the Tehran Stock Exchange Index by Dimensionality Reduction Techniques. *Financial Research Journal*, *24*(4), 577-601.
20. Nti, I. K., Adekoya, A. F., & Weyori, B. A. (2020). A systematic review of fundamental and technical analysis of stock market predictions. *Artificial Intelligence Review*, *53*(4), 3007-3057 .
21. Park, H. J., Kim, Y., & Kim, H. Y. (2022). Stock market forecasting using a multi-task approach integrating long short-term memory and the random forest framework. *Applied Soft Computing*, *114*, 108106 .
22. Salehi, M., & Garshasbi, F. (2019). Tehran Stock Exchange Index Forecasting Using Approach Adaptive Neural-Fuzzy Inference System and Imperialist Competitive Algorithm. *Business Intelligence Management Studies*, *8*(29), 5-34.
23. Schumaker, R. P., & Chen, H. (2009). A quantitative stock prediction system based on financial news. *Information Processing & Management*, *45*(5), 571-583 .
24. Sezer, O. B., Gudelek, M. U., & Ozbayoglu, A. M. (2020). Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied Soft Computing*, *90*, 106181 .
25. Shapley, L. S. (1953). A value for n-person games .
26. Sheikhzadeh, M. J., & Rahmany, S. (2023). Identification of effective indicators on predicting trends of total index of Tehran Stock Exchange using feature selection and

classification algorithms. *Financial Engineering and Portfolio Management*, *56*(14), 142-159.
27. SOHRABI, M., Seyed Mozaffar, S. M., Chirani, E., & Kheradyar, S. (2022). Modeling the Prediction of Stock Market Jumps Based on the Recurrent Neural Network and Deep Learning. *Journal of Securities Exchange*, *15*(59), 245-268.
28. Thakkar, A., & Chaudhari, K. (2021). A comprehensive survey on deep neural networks for stock market: The need, challenges, and future directions. *Expert Systems with applications*, *177*, 114800 .
29. Vaziri Kordestani, J., Farid, D., Nazemi Ardakani, M., & Hosseini Bamakan, S. M. (2022). Evaluation of PSO-BiLSTM method for stock price forecasting using stock price time series data (Case study: Iran Stock Exchange and OTC stock). *Financial Management Strategy*, *10*(4), 125-150.
30. Williams, C. K., & Rasmussen, C. E. (2006). *Gaussian processes for machine learning* (Vol. 2). MIT press Cambridge, MA .
31. Yañez, C., Kristjanpoller, W., & Minutolo, M. C. (2024). Stock market index prediction using transformer neural network models and frequency decomposition. *Neural Computing and Applications*, *36*(25), 15777-15797 .