Financial Research Journal

Print ISSN: 1024-8153
Online ISSN: 2423-5377

Inspecting the Predictive Power of Artificial Intelligence Models in Predicting the Stock Price Trend in Tehran Stock Exchange

Mahdi Heidari*
*Corresponding Author, Assistant Prof., Department of Economics, Faculty of Economics and Finance, Khatam University, Tehran, Iran. E-mail: m.heidari@teias.institute

Hamidreza Amiri
MSc. Student, Department of Economics, Faculty of Humanities, Khatam University, Tehran, Iran. E-mail: h.amiri@teias.institute

Abstract

Objective: Time series prediction methods based on artificial intelligence have been widely developed in recent years. Given that these data have large dimensions in the field of investment and stock price forecasting, traditional data analysis methods have low predictive power. This study examines the predictive power of a variety of models based on machine learning in the Tehran Stock Exchange.

Methods: After collecting data from 150 large companies listed on the Tehran Stock Exchange from 2012 to 2021, we want to predict the stock price trend- the movement direction of the price- and then validate each method and compare their accuracy. In these methods, we allocate part of the data to the learning section and the rest to the test section. We take these periods as training and trading sets. These methods include linear models, autocorrelation models, trees, and neural networks.

Results: Deep learning models show better performance than other models and have an accuracy of about 70 percent. Also, we show the time series of the best-performance model accuracy of portfolios of some large industries. The best-performance model of DL in this study is Recurrent Neural Networks. In addition, we show that shallow learning models have higher accuracy and most models perform better in predicting descending stock trends.

Conclusion: In this study, after trying to use the models very carefully, the result is that these models do not provide stunning results to investors.

Keywords: Stock price prediction, Machine learning, Investment, Tehran stock exchange.

Citation: Heidari, Mahdi & Amiri, Hamidreza (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. https://doi.org/10.22059/FRJ.2022.320064.1007149 (*in Persian*)

Financial Research Journal, 2022, Vol. 24, No.4, pp. 602-623
Published by University of Tehran, Faculty of Management
https://doi.org/10.22059/FRJ.2022.320064.1007149
Article Type: Research Paper

Received: May 07, 2021
Received in revised form: January 05, 2022
Accepted: July 25, 2022
Published online: January 19, 2023

شاپا چاپی: ۸۱۵۳-۱۰۲۴
شاپا الکترونیکی: ۵۳۷۷-۲۴۲۳
تحقیقات مالی

بررسی قدرت مدل‌های مبتنی بر هوش مصنوعی در پیش‌بینی روند قیمت سهام بورس اوراق بهادار تهران

مهدی حیدری*
* نویسنده مسئول، استادیار، گروه اقتصاد، دانشکده علوم انسانی، دانشگاه خاتم، تهران، ایران. رایانامه: m.heidari@teias.institute

حمیدرضا امیری
کارشناس ارشد، گروه اقتصاد ، دانشکده علوم انسانی، دانشگاه خاتم، تهران، ایران. رایانامه: h.amiri@teias.institute

چکیده

هدف: در سال‌های اخیر، روش‌های پیش‌بینی داده‌های سری زمانی مبتنی بر هوش مصنوعی و یادگیری عمیق گسترش بسیاری یافته است. با توجه به اینکه این داده‌ها در حوزه سرمایه‌گذاری و پیش‌بینی قیمت سهام ابعاد بزرگی دارند، روش‌های سنتی تحلیل داده، به‌سختی می‌توانند به یادگیری آن‌ها بپردازند. در این پژوهش، قدرت مدل‌های مختلف مبتنی بر یادگیری ماشین، در پیش‌بینی روند قیمت سهام در بورس اوراق بهادار تهران بررسی شده است.

روش: پس از جمع‌آوری داده‌های ۱۵۰ شرکت بزرگ پذیرفته‌شده در بورس اوراق بهادار تهران، از سال ۱۳۹۰ تا ۱۳۹۹، با تنظیم دقیق روش‌های یادگیری ماشین برای هر یک از سهام، به پیش‌بینی روند قیمت سهام و صحت‌سنجی هر یک از روش‌ها پرداختیم و آن‌ها را با هم مقایسه کردیم. در این روش‌ها، در هر مرحله یادگیری، بخشی از داده‌ها را به بخش یادگیری و ارزیابی و بقیه را به بخش آزمون اختصاص دادیم. این روش‌ها عبارت بودند از: مدل‌های خطی، مدل‌های خودهم‌بسته، جنگل تصادفی و شبکه‌های عصبی.

یافته‌ها: مدل‌های مبتنی بر یادگیری عمیق نسبت به سایر مدل‌ها عملکرد بهتری از خود نشان می‌دهند و در پیش‌بینی روند کوتاه‌مدت قیمت سهام، از دقتی حدود ۷۰ تا ۸۰ درصد برخوردارند. همچنین، مدل‌های یادگیری کم‌عمق دقت بالاتری داشتند. به‌طور کلی، بیشتر مدل‌ها در پیش‌بینی روندهای منفی سهام، عملکرد بهتری نشان می‌دهند.

نتیجه‌گیری: در این پژوهش، تلاش شد تا مدل‌ها با دقت بسیار به‌کار گرفته شوند. نتایج پژوهش نشان داد که برخلاف یافته‌های پژوهش‌های گذشته، این مدل‌ها نتایج خیره‌کننده‌ای در اختیار سرمایه‌گذاران قرار نمی‌دهند.

کلیدواژه‌ها: پیش‌بینی قیمت سهام، یادگیری ماشین، سرمایه‌گذاری، بورس اوراق بهادار تهران.

استناد: حیدری، مهدی و امیری، حمیدرضا (۱۴۰۱). بررسی قدرت مدل‌های مبتنی بر هوش مصنوعی در پیش‌بینی روند قیمت سهام بورس اوراق بهادار تهران. *تحقیقات مالی*، ۲۴(۴)، ۶۰۲- ۶۲۳.

تحقیقات مالی، ۱۴۰۱، دوره ۲۴، شماره ۴، صص. ۶۰۲- ۶۲۳
ناشر: دانشکدهٔ مدیریت دانشگاه تهران
نوع مقاله: علمی پژوهشی

تاریخ دریافت: ۱۴۰۰/۰۲/۱۷
تاریخ ویرایش: ۱۴۰۰/۱۰/۱۵
تاریخ پذیرش: ۱۴۰۱/۰۵/۰۳
تاریخ انتشار: ۱۴۰۱/۱۰/۲۹
doi: https://doi.org/10.22059/FRJ.2022.320064.1007149

مقدمه

یکی از گزاره‌های مهم تئوری قیمت‌گذاری[1] این است که بازده‌های انتظاری مقطعی[2]، باید با میزان در معرض قرار گرفتن آن‌ها در برابر عوامل ریسکی[3] توضیح داده شوند (فاما و فرنچ[4]، ۱۹۹۲)؛ بنابراین برای رسیدن به این نتیجه، بایستی به‌دنبال عواملی با بیشترین اثرگذاری بود. در سال‌های اخیر، جست‌وجو برای یافتن مهم‌ترین عوامل تأثیرگذار بر بازده انتظاری شدت گرفته است و بخش بزرگی از ادبیات اقتصاد مالی به این حوزه اختصاص دارد. پس از آنکه نظریهٔ پورتفولیوی مدرن (مارکوویتز[5]، ۱۹۵۲) ساخته شد، مدل‌های بیشتری برای توضیح ارتباط بازده اضافی[6] پورتفولیو با بازده اضافی بازار توسعه یافتند. در یکی از آخرین مدل‌ها، هاروی، لیو و ژو[7] (۲۰۱۶) قدرت پیش‌بینی‌کنندگی ۳۰۰ عامل را بررسی کردند. با توجه به ابعاد بسیار زیاد این تعداد متغیر، همان طور که کاکرن[8] (۲۰۱۱) توضیح می‌دهد، با «باغ وحشی از عامل‌ها»[9] مواجهیم و در این میان، باید مهم‌ترین آن‌ها را انتخاب کنیم. توجه به این نکته مهم است که پیش‌بینی بازده قیمت دارایی و سهام، به‌دلیل وجود محیط دارای اعوجاج[10] و همین طور فرض وجود بازار کارا، بسیار دشوار است (مالکیل و فاما[11]، ۱۹۷۰). به هرحال مدل‌های اقتصاد مالی، در حالی که تلاش خود را کرده‌اند تا به ارتباطی منسجم بین عوامل پیش‌بینی‌کننده و بازده‌های آینده دست یابند، با توجه به حضور عوامل غیرخطی پیچیده، نتوانسته‌اند به توضیحات و دقت خوبی دست پیدا کنند.

بعد از کاربردهای بسیار موفق روش‌های یادگیری ماشین، در بسیاری از حوزه‌هایی مانند پردازش زبان طبیعی[12] که شامل یادگیری توالی[13] اعداد می‌شوند، حدس‌های اولیه نشان دادند که شاید بتوان از این روش‌ها در فهمیدن ساختارهای غیرخطی بازارهای مالی استفاده کرد (هوک[14]، ۲۰۰۹). داده‌های مالی مانند قیمت و بازده سهام، به‌صورت سری زمانی هستند و از دیرباز مورد توجه محققان و سرمایه‌گذاران بوده‌اند. با توجه به اینکه اعوجاج فراوان، ابعاد بالا و همین طور شکست‌های ساختاری[15]، جزئی از طبیعت بازارهای مالی است، مدل‌های سادهٔ مالی توان کافی برای پردازش و پیش‌بینی آن‌ها را ندارند.

در این پژوهش ما علاقه‌مند بودیم تا قدرت پیش‌بینی روش‌های مختلف یادگیری ماشین را در بازار سهام تهران بررسی کنیم. اخیراً این دسته از پژوهش‌ها روی بورس اوراق بهادار آمریکا انجام شده و نشان داده است که مدل‌های

1. Asset Pricing Theory
2. Cross-Section of Expected Returns
3. Risk Factors
4. Fama & French
5. Markowitz
6. Excess Return
7. Harvey, Liu & Zhu
8. Cochrane
9. Factor Zoo
10. Noisy Environment
11. Malkiel & Fama
12. Natural Language Processing
13. Sequence
14. Huck
15. Structural Breaks

یادگیری عمیق، در مقایسه با سایر روش‌ها نتایج بسیار خوبی را کسب می‌کنند (گو، کلی و خیو[1]، ۲۰۲۰). نکتهٔ بسیار حائز اهمیت این است که شرایط شبیه‌سازی‌شده باید به‌دقت پایش شود تا مدل‌ها خطاهای زیاد و نتایج اشتباه نداشته نباشند.

برخی از مدل‌های یادگیری ماشین، بر روش رگرسیون[2] و برخی دیگر، بر روش دسته‌بندی[3] استوارند. نمونهٔ بسیار ساده از مدل‌های یادگیری ماشین مبتنی بر روش رگرسیون، همان رگرسیون خطی[4] است. در این پژوهش تأکید بر ترکیب این دو بوده است که جزئیات آن در ادامه بررسی خواهد شد. در ادامهٔ این مقاله، ابتدا پیشینهٔ پژوهش‌های انجام‌شده معرفی و مختصری از روش‌ها مرور شده است؛ سپس در قسمت روش‌شناسی پژوهش، پس از پرداختن به معرفی اطلاعات داده‌های استفاده شده، مدل‌های آزمون‌شده، معرفی و هریک از آن‌ها با جزئیات فنی تحلیل و بررسی می‌شود. در انتها نیز، نتیجهٔ این بررسی‌ها بیان و پیشنهادهایی برای پژوهش‌های بعدی ارائه خواهد شد.

پیشینهٔ پژوهش

با توجه به اهمیت زیاد پیش‌بینی قیمت‌ها در بازار سهام، سرمایه‌گذاران از روش‌های متنوعی برای این کار استفاده می‌کنند. عمدهٔ این روش‌ها را می‌توان به دو صورت طبقه‌بندی کرد: برخی از آن‌ها گذشته‌نگرند و با پیروی از الگوهای قیمتی تلاش می‌کنند که بازارها را بهتر پیش‌بینی کنند. برخی از روش‌های دیگر، به‌صورت آینده‌نگر سعی می‌کنند تا با تحلیل سناریوهای مختلف، این مهم را انجام دهند.

روش‌های یادگیری ماشین بر منطق اول استوارند و تلاش می‌کنند که با تخمین بهتر روابط، به کمینه‌کردن خطای پیش‌بینی بپردازند. با توجه به اینکه اصطلاح «یادگیری ماشین» در خصوص زمینهٔ بحث، می‌تواند معناهای مختلفی داشته باشد، لازم است منظور خود را از این اصطلاح توضیح دهیم. ما از این اصطلاح به‌عنوان روشی برای تخمین مدل‌های با ابعاد بالا و استفاده از آن برای پیش‌بینی بهتر و همچنین، به‌عنوان روشی برای انتخاب بهینهٔ مجموعهٔ متغیرها، برای جلوگیری از مشکل پوشش بیش از حد[5] مدل‌ها استفاده می‌کنیم. با توجه به اینکه در بسیاری از مسائل اقتصاد مالی، با مجموعهٔ بسیار بزرگی از متغیرها روبه‌رو هستیم، طبیعت این مدل‌ها توانایی آن‌ها را برای توضیح و پیش‌بینی بهتر داده‌ها افزایش می‌دهد.

در سال‌های اخیر، محققان بسیاری در حوزه‌های مختلف مالی، مانند پیش‌پرداخت وام، قصور در پرداخت‌ها و امتیازدهی اعتبار[6] (سیریگنانو، سادوانی و گیسیک[7]، ۲۰۱۶) و همچنین انتخاب سبد دارایی (هیتون و پولسون[8]، ۲۰۱۷)، به استفاده از این روش‌ها روی آورده‌اند. موضوع پیش‌بینی بازارها و تخمین صرف ریسک نیز در کانون توجه محققان و

1. Gu, Kelly & Xiu
2. Regression
3. Classification
4. Linear regression
5. Overfitting Problem
6. Credit Scoring
7. Sirignano, Sadhwani & Gieseche
8. Heaton & Polson

سرمایه‌گذاران قرار گرفته است. مرتون[1] (۱۹۸۱) توضیح داد که پیش‌بینی‌های بازده بازارها، ارزش اقتصادی دارد و همچنین کمپبل و تامپسون[2] (۲۰۰۸) نشان دادند که بسیاری از مدل‌های رگرسیون پیش‌بینی‌کننده، بهتر از میانگین تاریخی بازارها عمل می‌کنند. در بسیاری از پژوهش‌های اقتصادی محققان به‌دنبال توضیح تغییرات داده‌ها هستند و سعی می‌کنند با تخمین مدل‌های قدرتمند[3]، به توضیحات دقیق‌تری از روابط علّی بین داده‌ها دست پیدا کنند. در این مدل‌ها، تخمین‌ها عموماً به‌صورت قدرت توضیح‌دهندگی بالای داخل نمونه[4] سنجیده می‌شوند؛ اما با توجه به ماهیت پُر از اعوجاج بازارهای مالی، پژوهشگران هنوز به روابط علّی دقیقی دست پیدا نکرده‌اند. در پژوهش حاضر تلاش کردیم که با تخمین بهتر مدل‌ها، به قدرت توضیح‌دهندگی بالای آن‌ها در پیش‌بینی خارج نمونه[5] دست پیدا کنیم و به آزمون هریک از مدل‌ها بپردازیم.

گو و همکاران (۲۰۲۰) نشان دادند که روش‌های یادگیری عمیق مبتنی بر شبکه‌های عصبی چندلایه‌ای[6]، بهتر از سایر روش‌های یادگیری ماشین می‌توانند بازار را پیش‌بینی کنند. تخمین آن‌ها از صرف ریسک انتظاری[7]، به مدل قیمت‌گذاری مقطعی[8] نگاشت می‌شود. ذکر این نکته حائز اهمیت است که عمدهٔ روش‌های قیمت‌گذاری در بازارهای مالی، بر مبنای مدل‌های خطی چندعاملی بنا شده و این مدل‌ها توسط فاما و فرنچ در سال‌های ۱۹۹۳ و ۲۰۱۵ توسعه یافته است.

بسیاری از کاربردهای روش‌های یادگیری ماشین، به یافتن بهترین متغیرهای پیش‌بینی‌کننده معطوف شده است. گیگلیو و خیو[9] (۲۰۱۷) با استفاده از یکی از روش‌های کاهش ابعاد، به‌نام تحلیل مؤلفه‌های اصلی[10] نشان دادند که صرف ریسک یک عامل در یک مدل عاملی خطی، با کاهش ابعاد و کنترل سایر عوامل، می‌تواند فضای عوامل را به‌خوبی توضیح بدهد.

در پژوهش اخیری که بروگارد و زارعی[11] (۲۰۱۹) انجام دادند، تلاش کردند که با استفاده از روش‌های یادگیری ماشین، قیمت‌گذاری‌های نادرست[12] را پیدا کنند. مدل آن‌ها توانست قیمت‌گذاری‌های نادرست را با پیداکردن سری‌های زمانی ناسازگار[13]، در قالب یک مدل قیمت‌گذاری کشف کند. محققان بر این نکته تأکید کردند که قیمت‌گذاری‌های نادرست، همچنان در بازارهای دارایی وجود دارد؛ اما گذر زمان نشان می‌دهد که این قیمت‌گذاری‌ها به‌سمت بازارهای کارا[14] در حال حرکت است.

1. Merton
2. Campbell & Thompson
3. Robust
4. In-sample
5. Out-of-sample
6. Multilayer Neural Networks
7. Expected Risk Premium
8. Cross-Sectional Asset Pricing Model
9. Giglio & Xiu
10. Principal component analysis (PCA)
11. Brogaard & Zarei
12. Mispricing
13. Inconsistent
14. Efficient markets

در سال‌های اخیر، کاربردهای روش‌های یادگیری ماشین در علوم مالی، به حوزه‌های بسیار گسترده‌ای راه پیدا کرده است. بیانچی، بوچنر و تامونی[1] (۲۰۲۰) به اندازه‌گیری صرف ریسک اوراق قرضه[2]، در یک مدل رگرسیون بر مبنای مدل‌های مختلف یادگیری ماشین پرداختند. نتایج آن‌ها نشان داد که مدل‌های مبتنی بر یادگیری عمیق و شبکه‌های عصبی، می‌توانند تخمین صرف ریسک اوراق قرضه را به‌خوبی بهبود دهند.

در پژوهشی که روی شاخص کل بورس اوراق بهادار تهران با استفاده از رأی اکثریت انجام شد، خروجی مدل‌های یادگیری ماشین، دقت بسیار بالایی را در پیش‌بینی روند نشان داد. ورودی مدل‌های یادگیری ماشین در این پژوهش، از خروجی چندین مدل تحلیل تکنیکال تغذیه می‌شده است (افشاری راد، علوی و سینایی، ۱۳۹۷). در پژوهشی دیگر در زمینهٔ پیش‌بینی روند شاخص کل بورس اوراق بهادار تهران، از طریق جداسازی مؤلفه‌های غیرخطی و متلاطم سری زمانی شاخص کل قیمت و ترکیب مدل ماشین بردار پشتیبان[3] و بهینه‌سازی ازدحام ذرات[4] با بهره‌گیری از مدل GJR، خطاهای کمتری نسبت به سایر مدل‌ها تا آن زمان گزارش شده است (درودی و ابراهیمی، ۱۳۹۵).

برخی از پژوهشگران نیز به پیش‌بینی‌کنندگی متغیرهای تکنیکال در بازار بورس اوراق بهادار تهران پرداخته‌اند. سیف، جمشیدی نوید، قنبری و اسماعیل‌پور (۱۴۰۰) با استفاده از متغیرهای امواج الیوت[5] و شاخص قدرت نسبی[6] و با کمک ماشین بردار پشتیبان و درخت تصمیم، روند شاخص کل بورس اوراق بهادار تهران را پیش‌بینی کردند. آن‌ها برای پیش‌بینی از روش‌های یادگیری ماشین با نظارت استفاده کرده‌اند.

یکی از استراتژی‌های معاملاتی که در بازارهای مختلف دارایی از آن استفاده می‌شود، استراتژی معاملات زوجی[7] است. فلاح‌پور و حکیمیان (۱۳۹۸) با استفاده از روش یادگیری تقویتی[8] با هدف ماکزیمم‌سازی بازده و مینیمم‌سازی ریسک‌های منفی در معاملات زوجی، به طراحی یک سیستم معاملاتی پرداختند.

در مطالعه‌ای (فخاری، ولی‌پور خطیر و موسوی، ۱۳۹۶) تلاش شده است تا قدرت مدل‌های شبکهٔ عصبی و آریما ارزیابی شود. در این مطالعه، از قیمت سهام چند شرکت سرمایه‌گذاری به‌عنوان ورودی و خروجی مدل‌ها استفاده شده است.

در بسیاری از پژوهش‌ها محققان تلاش کرده‌اند که به‌جای قیمت دارایی، به پیش‌بینی بازده قیمت‌ها بپردازند. دلیل این امر آن است که بازده قیمت‌ها، ساختار نرمال و مانایی[9] در میانگین و واریانس دارند. با توجه به اینکه در بازار سهام ایران با پدیدهٔ دامنهٔ نوسان مواجهیم، بسیار مهم است که روند قیمت‌های رو به بالا یا پایین را حدس بزنیم. از این رو متغیر هدف[10] مدل‌های پژوهش حاضر، علامت بازده فردا خواهند بود که می‌تواند ۱، به‌معنای صعود قیمت یا منفی ۱، به‌معنای نزول قیمت باشد. در پژوهش حاضر تلاش می‌کنیم به بررسی انواع مختلفی از روش‌های یادگیری ماشین در

1. Bianchi, Buchner &Tamoni
2. Bonds
3. Support vector machine
4. Particle swarm optimization
5. Elliott wave
6. Relative strength index (RSI)
7. Paired trading
8. Reinforcement learning method
9. Stationary
10. Target

پیش‌بینی روند قیمت سهام در بورس اوراق بهادار تهران بپردازیم تا از میان روش‌های مختلف، بهترینِ آن‌ها را برای پژوهش‌های آتی شناسایی کنیم.

روش‌شناسی پژوهش

در این پژوهش به بررسی نتایج حاصل از تخمین ۱۷ مدل یادگیری ماشین پرداخته شده است که ابتدا هر یک از آن‌ها به‌طور دقیق تشریح می‌شود. شایان ذکر است که برای پیشبرد این پژوهش، ترتیب خاصی را در نظر گرفتیم. در گام نخست، داده‌های موجود را به دو دسته تقسیم کردیم. از دستهٔ اول که آن را دستهٔ یادگیری می‌نامیم، به‌عنوان دادهٔ آموزشی[1] و از دستهٔ دوم، به‌عنوان دادهٔ آزمون[2] داخل بازه برای ارزیابی و تنظیم مدل به‌منظور پیش‌بینی خارج از بازه استفاده کردیم. این فرایند به‌صورت چرخشی از اولین روز آزمون آغاز و تا آخرین روز داده تکرار می‌شود. در گام بعدی، ابعاد و ویژگی‌های متغیرهای استفاده شده را که به هر یک از آن‌ها متغیر خصیصه[3] می‌گوییم، به بحث می‌گذاریم و در نهایت با توجه به نتایج، میزان موفقیت هر یک از مدل‌ها را بررسی می‌کنیم.

مجموعه متغیرهای خصیصه و هدف

با توجه به روشی که کراووس[4] (۲۰۱۷) استفاده کرده است، در این پژوهش نیز ۷۰ درصد دادهٔ هر شرکت را به‌عنوان دادهٔ آموزشی و بقیه را به‌عنوان دادهٔ آزمون جهت ارزیابی[5] مدل و استفاده از آن برای پیش‌بینی خارج از بازه در نظر می‌گیریم. این نحوه از تقسیم داده‌ها، در هر بار تکرار آموزش برای پیش‌بینی خارج از بازه تکرار می‌شود. به عبارت دیگر، قسمت داده‌های آموزش، پس از هر بار پیش‌بینی بزرگ‌تر می‌شود. بسیاری از مدل‌های یادگیری ماشین برای اینکه واگرا نشوند، به ورودی‌های نرمال نیاز دارند. برخی از مجموعه متغیرهایی که به‌عنوان متغیر خصیصه استفاده می‌کنیم، مانند بازده روزانه، به‌شکل استاندارد هستند؛ اما برخی دیگر باید به‌صورت استاندارد تغییر کنند. این کار برای هر متغیر خصیصه در خصوص هر سهم با استفاده از رابطهٔ ۱ انجام می‌شود.

$$F_t = \frac{f_t - \bar{f}}{\sigma_f} \qquad \text{رابطهٔ ۱)}$$

رابطهٔ ۱، نحوهٔ محاسبهٔ مقدار استاندارد خصیصهٔ f در زمان t را نشان می‌دهد که در آن، $\bar{f}$ مقدار میانگین خصیصه و σ_f انحراف معیار آن است.

با توجه به‌وجود پدیدهٔ صف در بورس اوراق بهادار تهران و اهمیت زیاد آن در روند یک سهم، به جای اینکه مقدار دقیق بازده را با عنوان متغیر هدف تعیین کنیم، از علامت بازده روزانه برای دسته‌بندی خروجی مدل‌ها استفاده می‌کنیم؛ از این رو، چنانچه قیمت سهم افزایش یابد، از مقدار ۱ و اگر قیمت کاهش پیدا کند، از مقدار ۱- استفاده می‌کنیم. برای

1. Training
2. Test
3. Feature
4. Krauss
5. Validation

مثال، فرض کنید که بازهٔ اطمینان بازده (R_{low}, R_{high}) و نقطهٔ تخمین را R_{point} در نظر بگیریم. اگر مقدار $R_{high} - R_{low}$ بزرگ باشد، نقطهٔ تخمینی بی‌استفاده می‌شود؛ زیرا نشان می‌دهد که خطای تخمین زیاد و نقطهٔ واقعی بسیار متفاوت است. همچنین در حالتی که این مقدار کوچک و نزدیک به صفر باشد، می‌توان گفت که روند تخمینی، همان مقدار واقعی است و با خطای کوچکی مواجهیم.

داده‌های پژوهش

در این پژوهش، از داده‌های بورس اوراق بهادار تهران از سال ۱۳۹۰ تا ۱۳۹۹ به‌صورت روزانه استفاده شده است. اطلاعات مربوط به حدود ۵۰۰ شرکت که از ابتدا تا انتهای این بازه در بازار حضور داشته‌اند، جمع‌آوری شده و در نهایت با توجه به نقدشوندگی بالاتر سهام بزرگ در مقایسه با سهام کوچک، از ۱۵۰ سهم بزرگ بازار برای ارزیابی‌های مربوطه در این پژوهش استفاده می‌شود. با توجه به قوانین بازار، هر سهمی ممکن است در یک بازهٔ بسته باشد؛ از این رو، در این مدت تغییرات قیمتی نخواهد داشت و آن ردیف‌ها از داده خالی می‌شوند. برای حل این مشکل، قیمت سهام در این مدت با استفاده از شاخص صنعت آن سهم پُر می‌شود. برخی از داده‌هایی که در این پژوهش استفاده شده است، داده‌های مربوط به شرکت بوده و برخی دیگر داده‌های اقتصاد کلان است. در مواقعی که نیاز به پُرکردن روزهای خالی به‌شکلی نبود که در مورد قیمت توضیح داده شد، داده‌های آخرین روز کاری آن سهم مبنا قرار داده شده است.

شایان ذکر است که برخی از متغیرها برای آنکه معنادار شوند، با اعمال تغییراتی در آن‌ها وارد داده‌ها شده‌اند. برای مثال، می‌دانیم که حجم معاملات یک سهم طی ده سال اخیر، روند صعودی شدیدی داشته است، در این مواقع از نسبت حجم معاملات به میانگین روز یا ماه‌های قبل استفاده کرده‌ایم. همچنین از داده‌های ۱ درصد بالا و پایین توزیع هر یک از خصیصه‌ها هم، به‌دلیل نویزی بودن آن‌ها، چشم‌پوشی شده است.

جدول ۱. میانگین آماره‌های مختلف روزانه در صنایع بزرگ مختلف

صنعت	تعداد سهم	میانگین بازده	انحراف معیار	چولگی[1]	کشیدگی[2]
فلزات اساسی	۱۶	۰/۲۰	۱/۲۰	۰/۵۹	۳/۱۶
محصولات شیمیایی	۱۸	۰/۱۸	۱/۲۲	۰/۴۸	۳/۷۶
خودرو	۱۲	۰/۲۵	۲/۰۱	۰/۳۶	۱/۲۵
داروسازی	۱۴	۰/۲۴	۱/۲۷	۱/۰۶	۴/۶۶
بانک‌ها و مؤسسه‌های اعتباری	۱۰	۰/۱۵	۱/۴۶	۰/۰۶	۲/۸۹
سرمایه‌گذاری	۱۲	۰/۱۹	۱/۲۴	۰/۳۳	۱/۳۲
انبوه‌سازی، املاک و مستغلات	۸	۰/۱۵	۱/۷۳	۰/۲۰	−۰/۰۲
همه شرکت‌ها	۱۵۰	۰/۱۸	۱/۹۹	۰/۲۰	۰/۰۰۷

1. Skewness
2. Kurtosis

در جدول ۱ مشخصات آماری شرکت‌های حاضر در صنایع منتخب بزرگ آورده شده است. مقادیر این جدول با استفاده از تشکیل سبدهای سهام با وزن یکسان برای هر صنعت به‌دست آمده است. همچنین، مقادیر میانگین بازده و انحراف معیار به‌صورت درصد بیان شده است.

همان‌طور که گفته شد، داده‌های مربوط به هر یک از یک سهم‌ها دارای ابعاد بالا و چندین خصیصه‌اند. در جدول ۲ می‌توان هر یک از آن‌ها را مشاهده کرد. با توجه به اینکه صنایع و سهام مختلف می‌توانند با یکدیگر هم‌بستگی یا هم‌بستگی با تأخیر داشته باشند، در داده‌های مربوط به هر یک از سهام، از شاخص صنایع مختلف دیگر این مجموعه داده استفاده شده است.

جدول ۲. مشخصات و توضیحات متغیرهای خصیصهٔ سهم‌ها

متغیر	روش محاسبه
انحراف معیار بازده ۷ و ۳۰ روز اخیر	$std(r_{t-i:t}), i = 7, 30$
انحراف معیار ارزش معاملات ۷ و ۳۰ روز اخیر	$std(value_{t-i:t}), i = 7, 30$
نسبت قیمت به شاخص صنعت	$price_t / industry_t$
نسبت قیمت به شاخص کل بورس	$price_t / tepix_t$
نسبت شاخص کل به نرخ دلار	$tepix_t / dollar_t$
نسبت قیمت به نرخ دلار	$price_t / dollar_t$
بازده نسبت به قیمت ۱، ۳، ۲۰، ۵۰ و ۱۲۰ روز قبل	$r_{t-i:t}, i = 1, 3, 20, 50, 120$
بازده صنعت ۱، ۳، ۲۰، ۵۰ و ۱۲۰ روز قبل	$r^{industry}{}_{t-i:t}, i = 1, 3, 20, 50, 120$
بازده شاخص کل ۱، ۳، ۲۰، ۵۰ و ۱۲۰ روز قبل	$r^{tepix}{}_{t-i:t}, i = 1, 3, 20, 50, 120$
بازده دلار ۱، ۳، ۲۰، ۵۰ و ۱۲۰ روز قبل	$r^{dollar}{}_{t-i:t}, i = 1, 3, 20, 50, 120$
بازده صنایع مختلف ۱، ۳، ۲۰، ۵۰ و ۱۲۰ روز قبل	$r^{industries}{}_{t-i:t}, i = 1, 3, 20, 50, 120$
نسبت ارزش معاملات امروز سهم به میانگین ۵۰ روز اخیر	$value_t / avg(value_{t-50:t})$
نسبت تعداد افراد معاملات امروز سهم به میانگین ۵۰ روز اخیر	$count_t / avg(count_{t-50:t})$
نسبت ارزش معاملات خرید امروز افراد حقیقی سهم به میانگین ۵۰ روز اخیر	$value^{ind.buy}{}_t / avg(value^{ind.buy}{}_{t-50:t})$
نسبت ارزش معاملات فروش امروز افراد حقیقی سهم به میانگین ۵۰ روز اخیر	$value^{ind.sell}{}_t / avg(value^{ind.sell}{}_{t-50:t})$
نسبت امروز قیمت به شاخص صنعت به میانگین ۵۰ روز اخیر	$\frac{(price_t / industry_t)}{avg(price_{t-50:t} / industry_{t-50:t})}$
نسبت امروز قیمت به نرخ دلار به میانگین ۵۰ روز اخیر	$\frac{(price_t / dollar_t)}{avg(price_{t-50:t} / dollar_{t-50:t})}$
نسبت امروز قیمت به شاخص صنعت به میانگین ۵۰ روز اخیر	$\frac{(price_t / tepix_t)}{avg(price_{t-50:t} / tepix_{t-50:t})}$
نسبت امروز شاخص کل به قیمت دلار به میانگین ۵۰ روز اخیر	$\frac{(tepix_t / dollar_t)}{avg(tepix_{t-50:t} / dollar_{t-50:t})}$

مدل

در حالت عمومی می‌توانیم مقدار بازده را با یک مدل افزودنی[1] پیش‌بینی تخمین، به‌صورت رابطهٔ ۲ بنویسیم (گو و همکاران، ۲۰۲۰).

$$r_{i,t+1} = \mathbb{E}_t[r_{i,t+1}] + \varepsilon_{i,t+1}$$ رابطهٔ ۲)

با توجه به رابطهٔ بالا، بازده سهم i در زمان $t+1$ یعنی $r_{i,t+1}$ از دو قسمت تشکیل شده است. هدف ما این است که مقدار خطای تخمینی بخش انتظاری بازده، یعنی $\mathbb{E}_t[r_{i,t+1}]$، را که به‌صورت تابعی از متغیرهای خصیصه است، به‌گونه‌ای تخمین بزنیم که خطای قسمت دوم، یعنی $\varepsilon_{i,t+1}$، کمینه شود. به بیان دیگر، قدرت توضیح‌دهندگی خارج از نمونه[2] برای بازده محقق شدهٔ $r_{i,t+1}$ بیشینه شود. بخش انتظاری رابطهٔ ۲ را به‌صورت زیر نیز می‌توانیم بازنویسی کنیم.

$$\mathbb{E}_t[r_{i,t+1}] = g^*[z_{i,t}]$$ رابطهٔ ۳)

در رابطه‌های ۲ و ۳، سهام به‌صورت $i = 1, \dots, N$ است و N تعداد کل سهام را نشان می‌دهد. زمان نیز به‌صورت $t = 1, \dots, T$ است که تمام روزهای موجود در داده را دربرمی‌گیرد. با توجه به رابطهٔ ۳ می‌توانیم بگوییم که در این مدل می‌خواهیم به‌وسیلهٔ همه متغیرها تا زمان t، بازده‌های زمان $t+1$ را تخمین بزنیم.

همچنین با توجه به فرم تابع $g^*(.)$ باید گفت که پیش‌بینی‌ها با استفاده از دادهٔ شرکت i است و فقط از داده‌های تا زمان t استفاده می‌شود.

مدل خطی

ما تخمین‌های خود را ابتدا از یک مدل خطی آغاز می‌کنیم که ساده‌ترین مدل یادگیری ماشین محسوب می‌شود. یکی از مدل‌های خطی معروف، روش حداقل مربعات معمولی[3] است که در مدل‌های اقتصادسنجی نیز بسیار کاربرد دارد. همان‌طور که انتظار می‌رود، این مدل احتمالاً بسیار ضعیف است و ما از آن برای مقایسهٔ مدل‌های پیچیده‌تر استفاده خواهیم کرد. می‌توان با استفاده از روابط ۲ و ۳ یک مدل خطی بر مبنای حداقل مربعات معمولی نوشت.

$$g[z_{i,t}, \beta_i] = {z'}_{i,t}\beta_i$$ رابطهٔ ۴)

در رابطهٔ ۴، β نشان‌دهندهٔ ضرایب مدل خطی و ${z'}_{i,t}$ مجموعه متغیرهای خصیصه مدل خطی است. به‌کمک این رابطه می‌توانیم به این سؤال پاسخ دهیم که حتی اگر مدل واقعی ما غیرخطی باشد، آیا با یک مدل جبری خطی می‌توان آن را با دقت تخمین زد. طبق روش حداقل مربعات معمولی، تابع هدف[4] ما برای کمینه‌کردن آن، مجذور مجموع مربعات انحراف از خطاست.

1. Additive
2. Out of sample
3. Ordinary least squares
4. Objective Function

$$obj_i(\beta_i) = \sum_{t=1}^{T} (r_{i,t+1} - g[z_{i,t}, \beta_i])^2 \quad \text{رابطهٔ ۵)}$$

کمینه‌کردن این تابع هدف، همان مدل حداقل مربعات معمولی را به ما خواهد داد. ما در روش‌های خطی، نه از همهٔ متغیرها، بلکه از متغیرهای بازده اخیر، بازده دلار بازار آزاد و اندازهٔ شرکت استفاده می‌کنیم.

مدل‌های خطی با پارامترهای تصحیح شده

یکی از مشکلات اساسی مدل‌های تخمین و پیش‌بینی، بیش‌برازش[1] آن‌هاست که به افزایش خطای خارج از نمونه منجر می‌شود. زمانی گفته می‌شود که مدل دچار خطای بیش‌برازش شده است که در زمان تخمینِ دادهٔ داخلِ بازه[2]، خطای بسیار کمی داشته باشد؛ اما در صورتی که هنگام استفاده از آن مدل برای تخمین دادهٔ خارج از بازه با خطای بزرگی روبه‌رو شویم، باید به‌شکلی مشکل را رفع کنیم. برای حل این مشکل در مدل‌های خطی، در تابع هدف یک جریمه[3] یا شرط در نظر گرفته می‌شود. یکی از این شرط‌ها که در رابطهٔ ۵ نشان داده شده است، باعث محدود شدن فضای متغیرهای خصیصه شده و بسیاری از آن‌ها صفر می‌شوند و از این طریق، مسئله بیش‌برازش تا حدی حل می‌شود.

در این پژوهش ما از ۴ تابع هدف جریمه استفاده خواهیم کرد که عبارت‌اند از: لَسو[4]، خط الرأس[5]، الاستیک نت[6] و لَسوی گروهی[7]. در جدول ۳ هریک از این توابع جریمه آورده شده است. همان‌طور که در جدول مشاهده می‌شود، در تابع جریمهٔ الاستیک‌نت اَبَرمتغیر[8] ρ وجود دارد که با قرار دادن مقادیر مختلف، می‌توان به توابع جریمه دیگر رسید. برای نمونه، اگر $\rho = 0$ باشد، به روش لَسو می‌رسیم. همچنین λ ابرمتغیر روش لاگرانژ است.

جدول ۳. توابع جریمه برای مدل خطی

تابع جریمه	معادله
Lasso	$\lambda \sum_j \lvert\beta_j\rvert$
Ridge	$\frac{1}{2}\lambda \sum_j {\beta_j}^2$
Elastic net	$\lambda(1-\rho)\sum_j \lvert\beta_j\rvert + \frac{1}{2}\lambda\rho \sum_j {\beta_j}^2$
Group Lasso	$\lambda \sum_j \lVert\beta_j\rVert$

برای نمونه اگر بخواهیم به‌روش لَسو دست پیدا کنیم، باید شرط تابع جریمهٔ لَسو در جدول ۳ را به‌کمک روش حل لاگرانژ به تابع هدف معرفی شده در رابطهٔ ۵ اضافه کنیم که در این صورت به رابطهٔ ۶ خواهیم رسید.

1. Overfitting
2. In-sample
3. Penalty
4. Lasso
5. Ridge
6. Elastic net
7. Group lasso
8. Hyperparameter

$$objLASSO_i(\beta_i) = \sum_{t=1}^{T}\left(r_{i,t+1} - g[z_{i,t}, \beta_i]\right)^2 - \lambda\left(\sum_i |\beta_i|\right)$$ رابطهٔ ۶)

مدل آریما[1]

مدل آریما یا میانگین متحرک خودهم‌بستهٔ یکپارچهٔ[2] یک مدل برای داده‌های سری زمانی است که از ترکیب یک فرایند خودهم‌بسته[3] با یک فرایند میانگین متحرک[4] به‌دست می‌آید. برای آنکه نتایج مدل آریما قابل اطمینان باشد، باید داده‌های ورودی مانا باشند (ژنگ[5]، ۲۰۰۳). یک مدل سادهٔ خودهم‌بسته برای متغیر سری زمانی α_t می‌تواند به شکل خطی باشد، همان‌طور که در رابطهٔ ۷ آمده است. این مدل را به‌صورت $AR(p)$ نشان می‌دهند که p نشان‌دهندهٔ مقادیر تأخیر[6] است.

$$\alpha_t = c + \sum_{i=1}^{p} \phi_i \alpha_{t-i} + \varepsilon_t$$ رابطهٔ ۷)

در رابطهٔ ۷، ϕ_i مقادیر خودهم‌بستگی[7] است و مقادیر باقی‌مانده یا ε_t به‌صورت نویز سفید[8] با میانگین صفر و واریانس ${\sigma_\varepsilon}^2$ است. مقدار c نیز ثابت است. همچنین یک مدل سادهٔ میانگین متحرک که به‌صورت $MA(q)$ نشان داده می‌شود، در رابطهٔ ۸ بیان شده است.

$$\alpha_t = \mu + \sum_{i=0}^{q} \theta_i \varepsilon_{t-i}$$ رابطهٔ ۸)

در رابطهٔ ۸، θ_i وزن‌های مقادیر تصادفی[9] متغیر سری زمانی و غیرصفر است. همچنین $\theta_0 = 1$ است. مانند رابطهٔ ۹ فرض می‌کنیم که مقادیر باقی‌ماندهٔ نویز سفید با میانگین صفر و واریانس ${\sigma_\varepsilon}^2$ هستند.

حال با توجه به روابط ۷ و ۸ می‌توان به رابطهٔ اصلی مدل آریما دست یافت که در رابطهٔ ۹ بیان شده است. در مدل آریما، منظور از یکپارچی میزان (d) اعمال تفاضلات است که به‌واسطهٔ آن، دادهٔ غیرمانا[10] به دادهٔ مانا تبدیل می‌شود.

$$\alpha_t = c + \sum_{i=1}^{p} \phi_i \alpha_{t-i} + \varepsilon_t + \sum_{i=0}^{q} \theta_i \varepsilon_{t-i}$$ رابطهٔ ۹)

1. ARIMA
2. Autoregressive integrated moving average
3. Autoregressive process
4. Moving average process
5. Zhang
6. Lag
7. Autocorrelation
8. White noise
9. Stochastic
10. Non-stationary

در این رابطه، مقادیر خودهم‌بستگی و وزن‌های مقادیر تصادفی غیرصفر هستند. یک رابطهٔ آریما به‌صورت $ARIMA(p,d,q)$ بیان می‌شود. در این پژوهش برای یافتن مقادیر مناسب متغیرهای آریما، به‌صورت متوالی برای هر سهم نتیجهٔ این مقادیر محاسبه شده و در نهایت بهترین آن‌ها به عنوان نتیجه در نظر گرفته شده است.

مدل جنگل تصادفی[1]

در مدل‌های قبلی که به‌صورت خطی به تخمین داده می‌پرداختند، بسیاری از پیچیدگی‌های موجود در داده از جمله روابط غیرخطی و همچنین روابط بین متغیرها می‌توانست نادیده گرفته شود. با توجه به پُراعوجاج بودن قیمت در بازارها، در این پژوهش قصد داریم تا قدرت پیش‌بینی مدل‌های غیرخطی را نیز بسنجیم. یکی از این مدل‌ها که در سال‌های اخیر بسیار مورد توجه پژوهشگران حوزه یادگیری ماشین بوده است، مدل‌های درخت تصمیم و به‌خصوص مدل جنگل تصادفی است (بریمن[2]، ۲۰۰۱).

ایدهٔ اصلی این روش در دو قسمت بیان می‌شود. قسمت اول آن مربوط به دسته‌بندی[3] است. در این تکنیک، زیرمجموعه‌ای از داده‌ها به هرکدام از دسته‌های مدل داده می‌شود؛ به این معنا که هر یک از دسته‌ها باید مدل خود را با توجه به دادهٔ دریافت‌شده بسازد. فایدهٔ این تکنیک در آن است که واریانس درخت‌های تصمیم[4] را کاهش داده و باعث بهبود معنادار در عملکرد پیش‌بینی خارج از بازه می‌شود. نکتهٔ دیگر مدل جنگل تصادفی این است که محدود کردن مجموعه متغیرهای هر دسته، باعث کاهش هم‌بستگی[5] درخت‌های مجزا می‌شود. امروزه روش جنگل تصادفی، در پیش‌بینی‌های عمومی کاربردهای موفقی دارد و توانسته است عملکرد نسبتاً خوبی از خود در پیش‌بینی خارج از بازه نشان دهد.

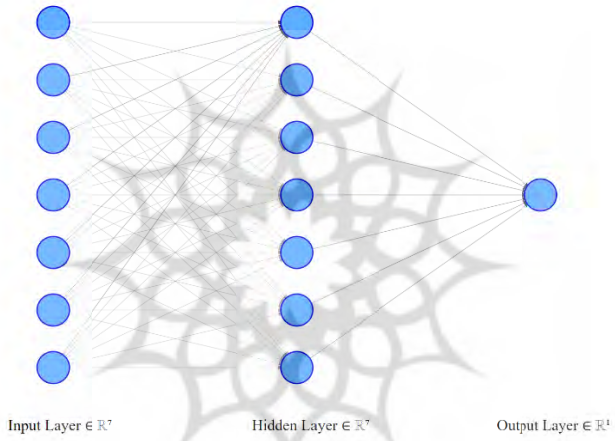
مدل شبکهٔ عصبی پیش‌خور[6]

یکی از مدل‌های غیرخطی که به بررسی آن می‌پردازیم، مدل شبکهٔ عصبی پیش‌خور است که در اغلب حوزه‌های پیش‌بینی به‌صورت محسوسی از بقیهٔ مدل‌ها بهتر عمل کرده است. این مدل‌ها یک لایهٔ ورودی[7] دارند که به هر یک از ورودی‌های آن، به‌اصطلاح لایهٔ نورون گفته می‌شود. این شبکه یک لایهٔ خروجی[8] هم دارد که می‌تواند یک یا چند متغیر باشد. همچنین در میان این دو لایه، یک یا چندین لایهٔ نهان[9] وجود دارد که وظیفهٔ تبدیلات غیرخطی را برعهده دارند. در هر یک از نورون‌های لایه‌های پنهان، یک تابع فعال‌سازی وجود دارد که آن را روی هر یک از سیگنال‌ها قبل از فرستادن به لایهٔ بعدی اِعمال می‌کند. تعداد نورون‌های لایهٔ ورودی، به همان تعداد متغیرهای خصیصه و تعداد

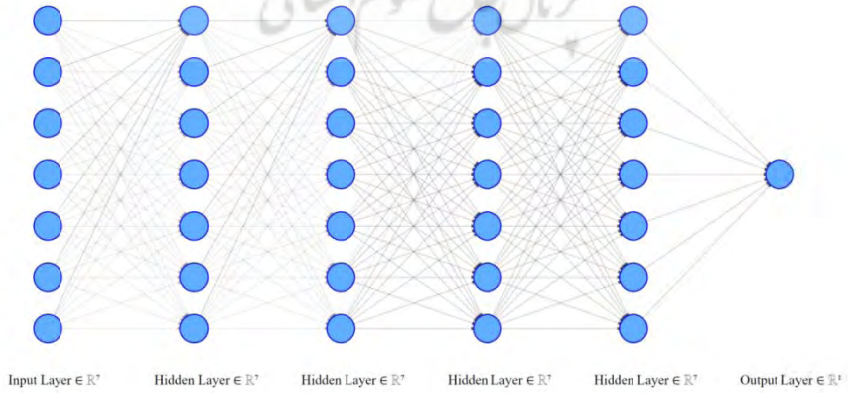
1. Random forest
2. Breiman
3. Bagging
4. Decision trees
5. Correlation
6. Feed forward neural network
7. Input layer
8. Output layer
9. Hidden layer

نورون‌های لایهٔ خروجی، به همان تعداد متغیر هدف است. در این بین، هریک از سیگنال‌های ورودی شبکه با توجه به ابعاد بالای ورودی می‌توانند جهت پیش‌بینی بهتر، ضعیف یا قوی شوند تا وزن‌های پارامترهای مدل تنظیم شود. وظیفهٔ لایهٔ خروجی این است که آخرین سیگنال‌های رسید از لایه‌های نهان را به‌عنوان تخمینی از متغیر هدف تجمیع کند. نمونه‌ای از مدل شبکهٔ عصبی پیش‌خور در شکل‌های ۱ و ۲ دیده می‌شود.

یکی از سؤال‌های مهم پژوهش حاضر این است که کدام‌یک از انواع بسیار عمیق یا کم‌عمق شبکه‌های عصبی دقت بیشتری دارند. منظور از شبکه‌های کم‌عمق، شبکه‌ای است کمتر از ۴ لایه و منظور از شبکه‌های بسیار عمیق، شبکه‌ای است که بیش از ۵ لایه دارد. در شکل ۱ نمونه‌ای از شبکهٔ کم‌عمق و در شکل ۲ نمونه‌ای از شبکهٔ عمیق را می‌توان دید.



شکل ۱. نمونه‌ای از شبکهٔ عصبی پیش‌خور کم عمق با ۷ ورودی، یک لایهٔ نهان و یک خروجی



شکل ۲. نمونه شبکه عصبی پیش‌خور عمیق با ۷ ورودی، ٤ لایه نهان و یک خروجی

به‌صورت شهودی می‌دانیم که شبکه‌های بسیار عمیق برای تخمین داده‌های بزرگ مناسب‌اند و برعکس؛ از این رو، ما این فرضیه را در داده‌های بورس اوراق بهادار تهران آزمون می‌کنیم و برای این آزمون، از شبکه‌های عصبی یک تا پنج لایه‌ای و همچنین از یکسوساز خطی[1] در این شبکه استفاده می‌کنیم.

$$ReLU(x) = \max(0, x)$$ رابطهٔ ۱۰)

رابطهٔ ۱۰ یک تابع یکسوساز خطی است که ما از آن به‌عنوان تابع فعال‌سازی هر نورون استفاده می‌کنیم.

مدل حافظهٔ طولانی کوتاه‌مدت[2]

توسعهٔ شبکه‌های حافظهٔ طولانی کوتاه‌مدت در سال‌های اخیر، باعث رشد چشمگیر دقت مدل‌های مبتنی بر یادگیری عمیق شده است. گریوز[3] (۲۰۱۳) و اُلاه[4] (۲۰۱۵) اخیراً تحت عنوان مدل‌های ریاضی، به توسعهٔ این روش یادگیری عمیق پرداخته‌اند. این مدل در ادبیات اقتصاد مالی نیز در سال‌های اخیر بسیار مورد توجه بوده است. در پژوهشی که چن، پلجر و ژو[5] (۲۰۱۹) انجام داده‌اند، توانسته‌اند با تخمین تابع تنزیل فاکتور تصادفی[6] توسط شبکه‌های حافظهٔ طولانی کوتاه‌مدت، به دقت بسیار خوبی دست پیدا کنند.

شبکه‌های حافظهٔ طولانی کوتاه‌مدت، نمونه‌ای از شبکه‌های عصبی بازگشت‌کننده[7] هستند. این شبکه‌ها حاوی روابط داخل نورونی[8] هستند (مدسکر و جین[9]، ۱۹۹۹) که می‌توانند روابط غیرخطی موجود در رشته اعداد را بهتر توضیح دهند. شبکه‌های حافظهٔ طولانی کوتاه‌مدت یک لایهٔ ورودی، یک یا چندین لایهٔ نهان و یک لایهٔ خروجی دارند. تعداد نورون‌های لایهٔ ورودی، برابر با فضای متغیرهای ورودی و تعداد نورون‌های لایهٔ خروجی، برابر با تعداد متغیرهای هدف است که در اینجا با توجه متغیر هدف ما، یک نورون خواهد بود. ویژگی اصلی شبکه‌های حافظهٔ طولانی کوتاه‌مدت در قسمت لایه‌های نهان آن است که سلول‌های حافظه را شامل می‌شود. هر سلول حافظه سه درگاه[10] دارد تا بتواند حالت‌ها[11] را در خود ذخیره کند. این درگاه‌ها ورودی (i_t)، خروجی (o_t) و فراموشی (h_t) را شامل می‌شوند. در زیر به وظیفهٔ هر یک از درگاه‌ها اشاره شده است:

- درگاه فراموشی تعیین می‌کند که چه اطلاعاتی باید از حالت سلول حذف شود.
- درگاه ورودی تعیین می‌کند که چه اطلاعاتی وارد حالت سلول شود.
- درگاه خروجی تعیین می‌کند که چه اطلاعاتی به‌عنوان خروجی سلول تشخیص داده می‌شود.

1. Linear rectifier
2. Long short-term memory
3. Graves
4. Olah
5. Chen, Pelger & Zhu
6. Stochastic discount factor (SDF)
7. Recurrent neural networks (RNN)
8. Inter-neural
9. Medsker & Jain
10. Gate
11. States

اگر حالت سلول را s_t در نظر بگیریم، در هر زمان این گام‌ها طی خواهد شد: در گام نخست، لایهٔ شبکه تشخیص می‌دهد که کدام دسته از اطلاعات باید از حالت قبلی سلول، یعنی s_{t-1} حذف شود. در این لحظه، خروجی تابع فعال‌سازی با توجه به ورودی و خروجی سلول حالت محاسبه می‌شود. در این پژوهش ما از تابع فعال سازی سیگموید[1] استفاده خواهیم کرد. به‌طور خلاصه ما به‌روش زیر از شبکه‌های حافظهٔ طولانی کوتاه‌مدت استفاده می‌کنیم:

- داده‌های ورودی آموزش که شامل ۷۰ درصد داده‌های هر سهم است.
- شبکه‌های حافظهٔ طولانی کوتاه‌مدت با ۱۰، ۲۰ و ۳۰ لایهٔ نورون در لایهٔ نهان و نرخ حذف تصادفی[2] ۰/۱.
- لایه خروجی که نتیجه تخمین مدل از بازده فردای سهام خواهد بود.

اندازه‌گیری عملکرد

برای اینکه عملکرد هر یک از مدل‌ها را در این پژوهش بسنجیم، از نتایج ماتریس درهم‌ریختگی[3] استفاده می‌کنیم. با توجه به این نکته که در این پژوهش مدل‌ها به‌گونه‌ای تنظیم شده‌اند که بتوانند روند بازده سهام را تشخیص دهند، به‌عنوان عملکرد هر یک از مدل‌ها، موارد تشخیصی درست ارزیابی می‌شود. یادآوری می‌شود که با توجه به ماهیت پنل این پژوهش، دقت این مدل‌ها هم به‌صورت تجمیع شده در زمان ارائه شده، هم برای بهترین مدل، به‌صورت سری زمانی برای گروه‌های بزرگ صنایع نشان داده خواهد شد. دقت تخمین کلی برای هر مدل، به‌صورت دقت نهایی آن‌ها در پیش‌بینی روند رو به بالا یا پایین (بازده مثبت یا منفی) اندازه‌گیری می‌شود. روش اندازه‌گیری این دقت نیز، به‌صورت میانگین دقت مدل در زمان‌های مختلف در یک پورتفولیو با سایز یکسان (میانگین حسابی) است و برای نشان‌دادن دقت بهترین مدل در زمان‌های مختلف بازار، از دقت آن‌ها به‌صورت سری زمانی در پورتفولیو با اندازهٔ یکسان در زمان‌های یکسان استفاده می‌کنیم.

ذکر این نکته ضروری است که با توجه به اینکه مجموعهٔ بازده‌های فردا در این روش به دو صورت منفی یا مثبت است، دقت هر یک از مدل‌ها که به ۵۰ درصد نزدیک باشد، به معنی این است که مدل نمی‌تواند به‌درستی بازده فردا را تخمین بزند. همین‌طور دقت هر مدلی که فاصلهٔ بیشتری با ۵۰ درصد دارد، نشان‌دهندهٔ این موضوع است که مدل در تخمین درست‌تر بازده فردا، موفق‌تر عمل کرده است.

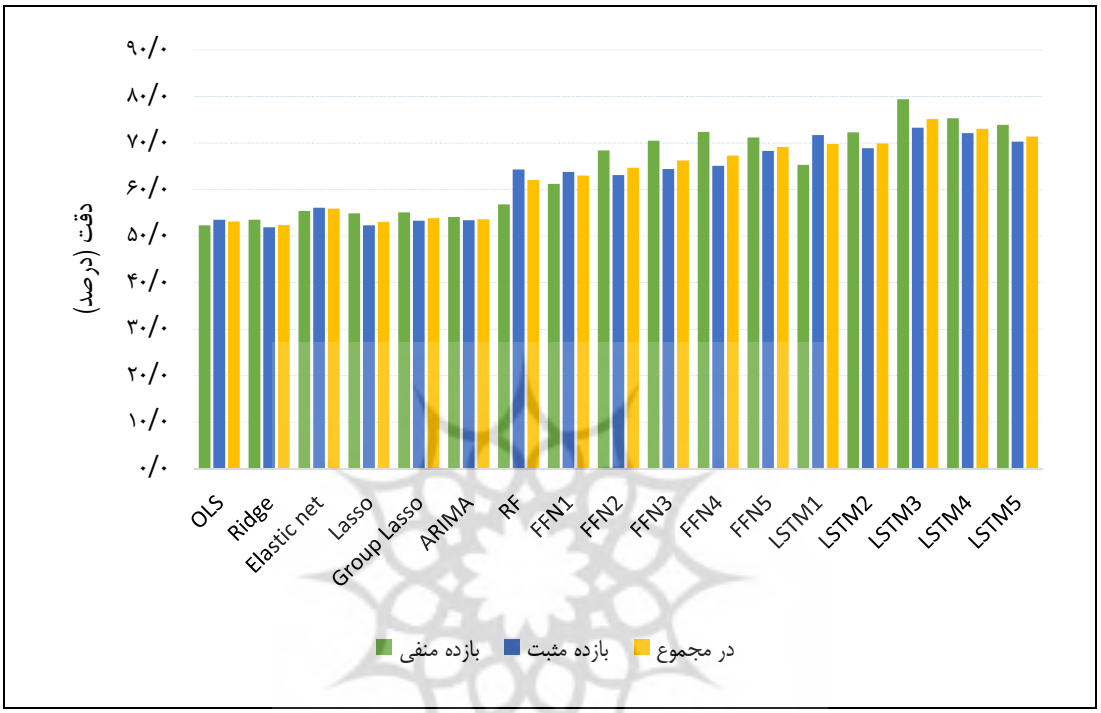
یافته‌های پژوهش

عملکرد کلی مدل‌ها

شکل ۳ مقایسه روش‌های مختلف مبتنی بر یادگیری ماشین را نشان می‌دهد. در این نمودار، دقت هریک از مدل‌ها در پیش‌بینی خارج از بازه را می‌توان مشاهده کرد. ما در این پژوهش از ۱۷ مدل یادگیری ماشین، شامل مدل‌های خطیِ

1. Sigmoid
2. Dropout rate
3. Confusion matrix

حداقل مربعات معمولی، مدل‌های با پارامتر تصحیح شده (ریج، الاستیک‌نت، لسو، لسوی گروهی) و آریما و همچنین از مدل‌های پیچیده‌تر جنگل تصادفی، شبکهٔ عصبی پیش‌خور (شامل ۵ مدل) و مدل حافظهٔ طولانی کوتاه‌مدت (شامل ۵ مدل) استفاده کرده‌ایم.



شکل ۳. نمودار قدرت پیش‌بینی مدل‌های مختلف یادگیری ماشین

همچنین در جدول ۴ می‌توان نتایج عددی دقت هر یک از مدل‌ها را دید. دقت مدل‌ها در هر کدام از روزهای با بازده منفی یا مثبت به تفکیک آورده شده است. همهٔ اعداد این جدول به‌صورت درصد هستند. همان‌طور که دیده می‌شود، معمولاً دقت تخمین مدل‌های یادگیری ماشین در روزهای منفی بازار بهتر از روزهای مثبت آن است. این جدول همچنین نشان می‌دهد که دقت مدل‌های غیرخطی به‌صورت محسوسی از مدل‌های خطی بهتر است. با تصحیح پارامترهای مدل‌های خطی و استفاده از مدل‌های ریج و الاستیک‌نت و همچنین لسو، شاهد بهبود عملکرد مدل‌های خطی هستیم؛ اما همچنان نمی‌توان به آن‌ها اتکا کرد.

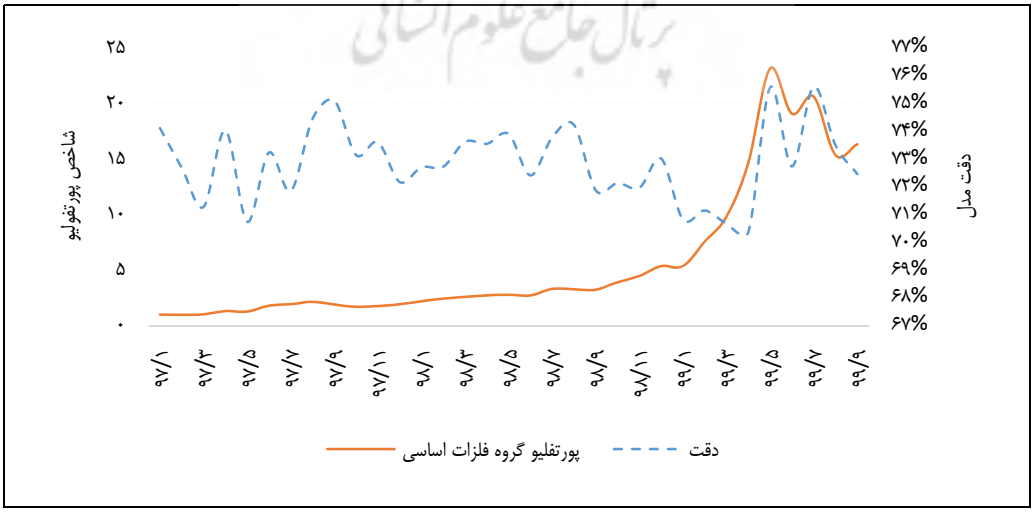
با توجه به نتایج می‌توان فهمید که مدل‌های یادگیری عمیق، در حالت غیرعمیق‌تر دقت بهتری دارند. در شبکه‌های عصبی پیش‌خور، مدلی که دارای ۴ لایه است و همچنین در مدل حافظهٔ طولانی کوتاه‌مدت، مدلی که دارای ۳ لایه است، دقت بهتری از خود نشان داده است.

با توجه به اینکه در پژوهش مدل حافظه طولانی کوتاه‌مدت بهتر از سایر مدل‌ها توانسته است به تخمین بازده فردا بپردازد، به‌صورت سری زمانی نیز آن را برای صنایع بزرگ بورس اوراق بهادار تهران بررسی می‌کنیم.

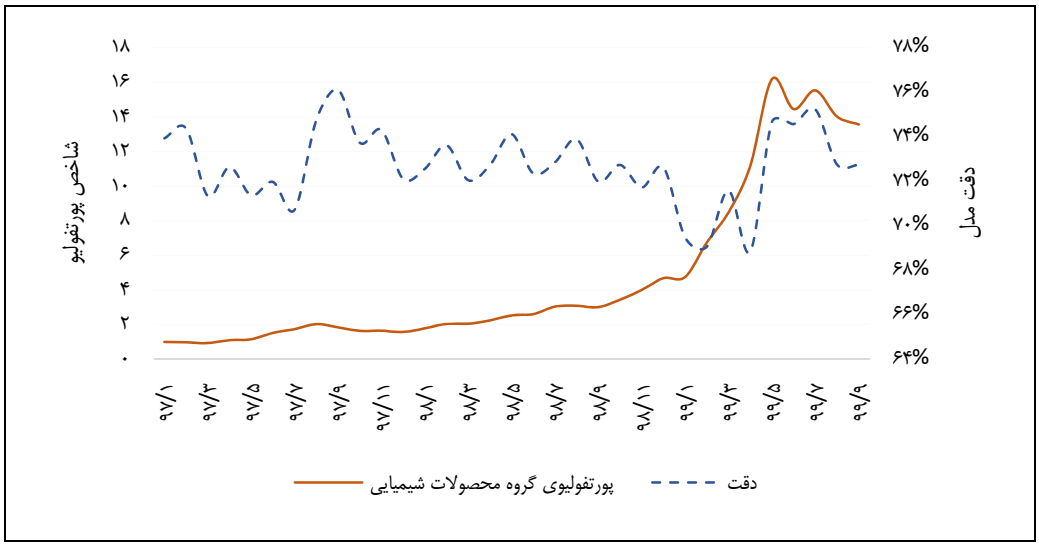
جدول ٤. دقت مدل‌های یادگیری ماشین

مدل	بازده منفی	بازده مثبت	در مجموع	F1 score
OLS	۵۲/۳	۵۳/۵	۵۳/۱	۵۲/۶
Ridge	۵۳/۵	۵۱/۹	۵۲/۴	۵۳/۰
Elastic Net	۵۵/۴	۵۶/۱	۵۵/۹	۵۵/۵
Lasso	۵۴/۹	۵۲/۳	۵۳/۱	۵۴/۱
Group Lasso	۵۵/۱	۵۳/۳	۵۳/۸	۵۴/۶
ARIMA	۵۴/۱	۵۳/۴	۵۳/۶	۵۳/۹
RF	۵۶/۸	۶۴/۳	۶۲/۱	۵۹/۰
FFN1	۶۱/۲	۶۳/۸	۶۳/۰	۶۲/۰
FFN2	۶۸/۴	۶۳/۱	۶۴/۷	۶۶/۶
FFN3	۷۰/۵	۶۴/۴	۶۶/۲	۶۸/۴
FFN4	۷۲/۴	۶۵/۱	۶۷/۳	۶۹/۸
FFN5	۷۱/۲	۶۸/۳	۶۹/۲	۷۰/۱
LSTM1	۶۵/۳	۷۱/۷	۶۹/۸	۶۷/۴
LSTM2	۷۲/۳	۶۸/۹	۶۹/۹	۷۱/۱
LSTM3	۷۹/۴	۷۳/۳	۷۵/۱	۷۷/۰
LSTM4	۷۵/۳	۷۲/۱	۷۳/۱	۷۴/۱
LSTM5	۷۳/۹	۷۰/۳	۷۱/۴	۷۲/۵

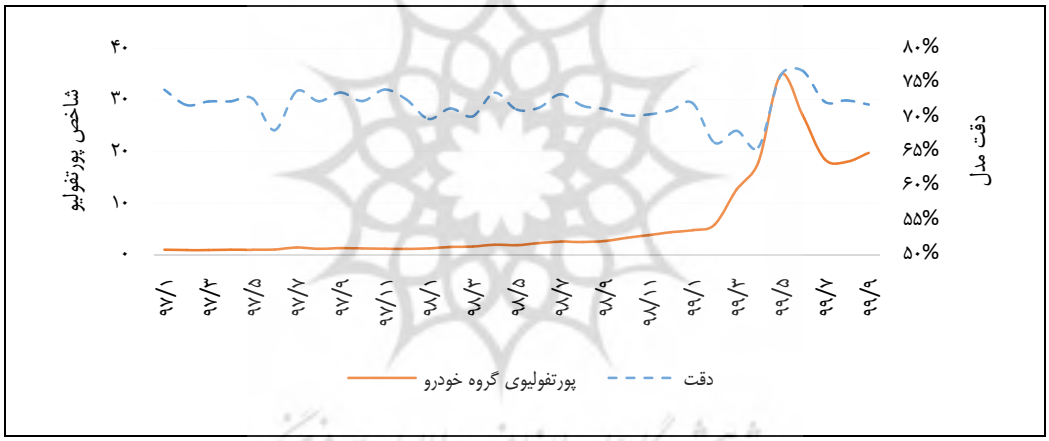
شکل‌های ۴ تا ۸ دقت مدل در طول زمان را برای صنایع مختلف نشان می‌دهند. یکی از نتایج این پژوهش این یافته است که مدل‌های مبتنی بر یادگیری عمیق در پیش‌بینی روندهای نزولی عملکرد بهتری از خود نشان می‌دهند. به طول مثال در بازه فروردین تا مرداد سال ۱۳۹۹، این مدل‌ها دارای دقت کمتری بوده‌اند که علت آن را می‌توان در شکست ساختاری قیمت‌های بورس اوراق بهادار تهران در آن زمان دانست که مورد مشابه یا نزدیک آن قبلا مشاهده نشده بود.



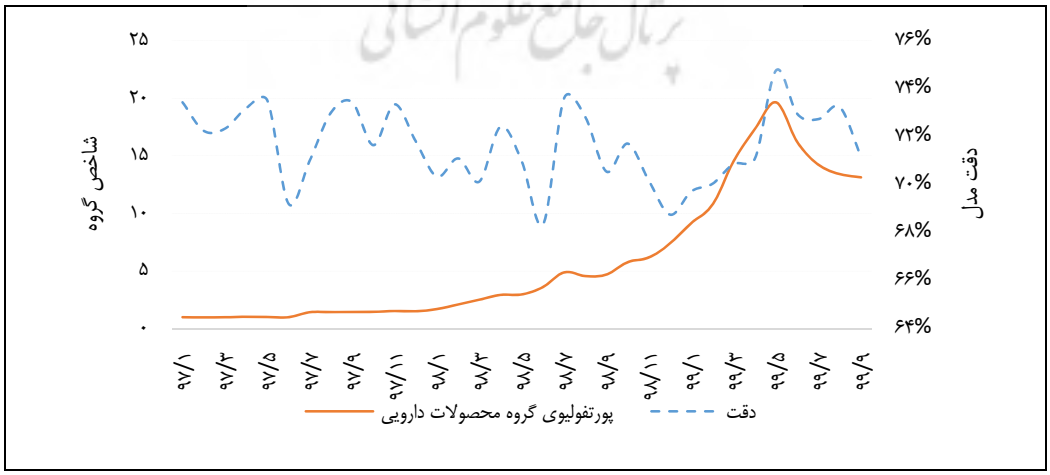
شکل ٤. نمودار عملکرد مدل حافظهٔ طولانی کوتاه‌مدت در طول زمان آزمون خارج بازه در گروه فلزات اساسی



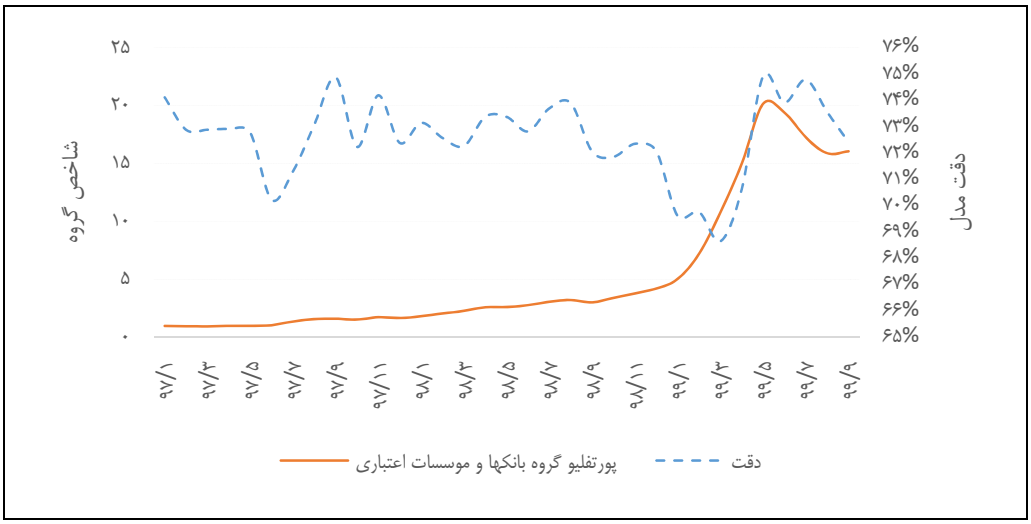
شکل ۵. نمودار عملکرد مدل حافظهٔ طولانی کوتاه‌مدت در طول زمان آزمون خارج بازه در گروه محصولات شیمیایی



شکل ۶. نمودار عملکرد مدل حافظهٔ طولانی کوتاه‌مدت در طول زمان آزمون خارج بازه در گروه خودرو



شکل ۷. نمودار عملکرد مدل حافظهٔ طولانی کوتاه‌مدت در طول زمان آزمون خارج بازه در گروه محصولات دارویی



شکل ۸. نمودار عملکرد مدل حافظهٔ طولانی کوتاه‌مدت در طول زمان آزمون خارج بازه در گروه بانک‌ها و مؤسسه‌های اعتباری

نتیجه‌گیری و پیشنهاد

با توجه به استفاده روزافزون از مدل‌های یادگیری ماشین در پیش‌بینی داده‌های مختلف از جمله داده‌های مالی، تلاش کرده‌ایم در این پژوهش به بررسی مدل‌های پرکاربردِ تخمین و پیش‌بینی در بورس اوراق بهادار تهران بپردازیم. ویژگی‌های منحصربه‌فرد داده‌های بورس اوراق بهادار تهران، مانند وجود دامنهٔ نوسان و اطلاعات مربوط به معاملات حقیقی و حقوقی، موجب علاقه‌مندی پژوهشگران برای ارزیابی مدل‌های یادگیری ماشین برای پیش‌بینی در این حوزه می‌شود. ما در این پژوهش، به بررسی و مقایسهٔ عملکرد مدل‌های مختلف یادگیری ماشین در پیش‌بینی روند قیمت سهام در بورس اوراق بهادار تهران پرداخته‌ایم. مطالعات مختلفی برای ارزیابی هر یک از مدل‌ها انجام گرفته است که برخی از آن‌ها با نتایج ما هم‌خوانی دارند. نکتهٔ بسیار مهمی که در استفاده از مدل‌های یادگیری ماشین در پیش‌بینی‌ها باید در نظر گرفته‌شود، تنظیم دقیق پارامترهای مدل و همچنین داده‌های ورودی برای جلوگیری از خطای آینده‌نگر[1] است؛ چرا که بسیاری از پژوهش‌های حاضر داخل کشور، برای روش‌های خود، دقت‌های بالای ۹۰ درصد گزارش کرده‌اند که این موضوع با پژوهش‌های معتبر خارجی هم‌خوانی ندارد.

یکی از یافته‌های ما در پژوهش حاضر این است که در میان مدل‌های مختلف یادگیری ماشین، برای داده‌های بورس اوراق بهادار تهران، مدل‌های شبکهٔ عصبی و حافظهٔ طولانی کوتاه‌مدت نسبت به مدل‌های دیگر دقت بیشتری دارند. به عبارت دیگر، این مدل‌ها بهتر توانسته‌اند به یادگیری روابط غیرخطی میان متغیرها دست پیدا کنند. یکی دیگر از یافته‌های ما این است که در مدل‌های شبکهٔ عصبی، آن‌هایی که عمق کمتری دارند، نسبت به شبکه‌های بسیار عمیق از دقت بهتری برخوردارند. همچنین یافتهٔ دیگر ما این است که مدل‌های مبتنی بر یادگیری ماشین، در روزهای منفی بازار

1. Look-ahead bias

سهام، در پیش‌بینی روند سهام عملکرد بهتری از خود نشان می‌دهند. ما در این پژوهش به‌دنبال استراتژی بهینهٔ معاملاتی در بورس اوراق بهادار تهران نبودیم؛ اما همان طور که به نظر می‌رسد، هیچ‌یک از این مدل‌ها تضمینی برای یافتن یک استراتژی معاملاتی سودآور نمی‌دهند. بهترین دقت مدل‌ها هنوز زیر ۸۰ درصدند که این خود می‌تواند انگیزه‌ای برای پژوهش‌های بعدی باشد تا استراتژی‌های مبتنی بر مدل‌های یادگیری عمیق را بیازمایند. برای پژوهش‌های آتی پیشنهاد می‌شود که داده‌های مربوط به سودآوری، فروش و عملکرد شرکت‌ها، به‌عنوان دادهٔ ورودی، برای تقویت دقت مدل‌ها استفاده شوند. در طول سالیان اخیر، متغیرهای مختلفی به‌صورت جداگانه بررسی شده‌اند. یکی از مزیت‌های مدل‌های پیشرفته یادگیری ماشین، این است که می‌توانند از حجم بالاتر داده‌ها برای پیش‌بینی دقیق‌تر استفاده کنند. همچنین در مدل‌هایی که تعداد متغیرهای ورودی‌شان بسیار زیاد است، می‌توان از روش‌های مختلف کاهش ابعاد برای جلوگیری از بیش‌برازش مدل‌ها استفاده کرد. در نهایت امیدواریم یافته‌های ما بتواند به کاربردهای بیشتر روش‌های یادگیری ماشین در علوم اقتصادی و مالی کمک کند.

منابع

افشاری راد، الهام؛ علوی، سید عنایت اله و سینایی، حسنعلی (۱۳۹۷). مدلی هوشمند برای پیش‌بینی روند سهام با استفاده از روش‌های تحلیل تکنیکال. *تحقیقات مالی*، ۲۰(۲)، ۲۴۹-۲۶۴.

سیف، سمیرا؛ جمشیدی نوید، بابک؛ قنبری، مهرداد؛ اسماعیلی‌پور، منصور (۱۴۰۰). پیش‌بینی روند بورس سهام ایران با استفاده از نوسان نمای موج الیوت و شاخص قدرت نسبی. *تحقیقات مالی*، ۲۳(۱)، ۱۳۴-۱۵۷.

درودی، دیاکو؛ ابراهیمی، سید بابک (۱۳۹۵). ارائه روش هیبریدی نوین برای پیش‌بینی شاخص کل قیمت بورس اوراق بهادار. *تحقیقات مالی*، ۱۸(۴)، ۶۱۳-۶۳۲.

فخاری، حسین؛ ولی پور خطیر، محمد؛ موسوی، سیده مائده (۱۳۹۶). بررسی عملکرد شبکه عصبی بیزین و لونبرگ مارکوات در مقایسه با مدل‌های کلاسیک در پیش‌بینی قیمت سهام شرکت‌های سرمایه‌گذاری. *تحقیقات مالی*، ۱۹(۲)، ۲۹۹-۳۱۸.

فلاح پور، سعید؛ حکیمیان، حسن (۱۳۹۸). بهینه‌سازی استراتژی معاملات زوجی با استفاده از روش یادگیری تقویتی، با به‌کارگیری دیتاهای درون‌روزی در بورس اوراق بهادار تهران. *تحقیقات مالی*، ۲۱(۱)، ۱۹-۳۴.

References

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. (*in Presian*)

Bianchi, D., Büchner, M., & Tamoni, A. (2020). Bond risk premia with machine learning. *WBS Finance Group Research Paper*, (252).

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Brogaard, J., & Zareei, A. (2019). Machine learning and the stock market. *Available at SSRN 3233119*.

Campbell, J. Y., & Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies*, *21*(4), 1509-1531.

Chen, L., Pelger, M., & Zhu, J. (2019). Deep learning in asset pricing. *Available at SSRN 3350138*.

Cochrane, J. H. (2011). Presidential address: Discount rates. *The Journal of finance*, *66*(4), 1047-1108.

Dorodi, D., & Abrahimi, S. B. (2017). Presenting a new hybrid method for predicting the Stock Exchange price index. *Financial Research Journal*, 18 (4), 612-632. (*in Persian*)

Fakhari, H., Valipour Khatir, M. & Mousavi, M. (2017). Investigating Performance of Bayesian and Levenberg-Marquardt Neural Network in Comparison Classical Models in Stock Price Forecasting. *Financial Research Journal*, 19 (2), 229-318. (*in Persian*)

Fallahpour, S., & Hakimian, H. (2019). Paired Trading Strategy Optimization Using the Reinforcement Learning Method: Intraday Data of Tehran Stock Exchange. *Financial Research Journal*, 21(1), 19- 34. (*in Persian*)

Fama, E. F., & French, K. R. (1992). The cross section of expected stock returns. *The Journal of Finance*, *47*(2), 427-465.

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33, 3-52.

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of financial economics*, *116*(1), 1-22.

Giglio, S., & Xiu, D. (2017). *Inference on risk Premia in the presence of omitted factors* (No. w23527). National Bureau of Economic Research.

Graves, A. (2013). Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223-2273.

Harvey, C. R., Liu, Y., & Zhu, H. (2016). … and the cross-section of expected returns. *The Review of Financial Studies*, *29*(1), 5-68.

Heaton, J. B., Polson, N. G., & Witte, J. H. (2017). Deep learning for finance: deep portfolios. *Applied Stochastic Models in Business and Industry*, *33*(1), 3-12.

Henriksson, R. D., & Merton, R. C. (1981). On market timing and investment performance. II. Statistical procedures for evaluating forecasting skills. *Journal of business*, 513-533.

Huck, N. (2009). Pairs selection and outranking: An application to the S&P 100 index. *European Journal of Operational Research*, *196*(2), 819-825.

Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research*, *259*(2), 689-702.

Malkiel, B. G., & Fama, E. F. (1970). Efficient capital markets. *A review of theory and empirical work Journal of Finance*, 25, 383, 417.

Markowitz, H. (1952). Portfolio selection. *The journal of finance*, 7(1), 77-91.

Medsker, L., & Jain, L. C. (Eds.). (1999). *Recurrent neural networks: design and applications*. CRC press.

Olah, C. (2015). Understanding lstm networks.

Seif, S., Jamshidinavid, B., Ghanbari, M. & Esmaeilpour, M. (2021). Predicting Stock Market Trends of Iran Using Elliott Wave Oscillation and Relative Strength Index. *Financial Research Journal*, 23(1), 134-157. (*in Persian*)

Sirignano, J., Sadhwani, A., & Giesecke, K. (2016). Deep learning for mortgage risk. *arXiv preprint arXiv:1607.02470*.

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, *50*, 159-175.