Kharazmi University

Mathematical Research
Year 2025, Volume 11, Issue 3, pp. 73–103

Print ISSN: 2588-2546
Online ISSN: 2588-2554
DOI: xxxx

A Data-Driven Approach for Portfolio Optimization Using Machine Learning and Deep Learning Algorithms

A. Najaf Najafi[(1)], **M. Najaf Najafi**[(2)1]

[(1)] Department of Industrial Engineering, Faculty of Industrial Engineering, University of Science and Technology, Tehran, Iran

[(2)] Khorasan Razavi Agricultural and Natural Resources Research and Education Center, AREEO, Mashhad, Iran

Received: 19 August 2024 Accepted: 19 November 2025 Published online: 17 December 2025

Abstract: In today's complex and dynamic financial markets, portfolio optimization presents a significant challenge for investors. As such, capital market investors grapple with fundamental questions regarding which stocks to buy, at what time, and in what quantities. This research aims to provide a novel approach to portfolio optimization using a mean-variance model based on predictions from traditional machine learning and deep learning algorithms, offering solutions to these crucial questions. Drawing on the emergence of data-driven methods, this study compares the performance of various machine learning and deep learning algorithms in forecasting stock prices on the Tehran Stock Exchange. The dataset comprises the closing prices of nine major symbols from the Tehran Stock Exchange over a 1000-day period. The findings suggest that traditional machine learning models, particularly linear regression, outperform deep learning models in predicting prices. Furthermore, the mean-variance portfolio optimization approach leverages optimal stock selection and allocation to maximize returns while minimizing risk. This research serves as a practical tool for portfolio managers and risk analysts, facilitating improved risk management and investment portfolio performance.

Keywords: Portfolio optimization, Stock price prediction, Machine learning, Deep learning.

[1]Corresponding author

E-mail addresses: (A. Najaf Najafi) ali_najaf80@ind.iust.ir, (M. Najaf Najafi) mnajafi.mhd@gmail.com

پژوهش‌های ریاضی
سال ۱۴۰۴، دوره ۱۱، شماره ۳، ص۷۳ـ۱۰۳
شاپا چاپی: ۲۵۴۶ـ۲۵۸۸
شاپا الکترونیکی: ۲۵۵۴- ۲۲۸۸
DOI: xxxx

ارایه روشی مبتنی بر داده جهت بهینه سازی سبد سهام با استفاده از الگوریتم های یادگیری ماشین و یادگیری عمیق

علی نجف نجفی[(۱)]، مسعود نجف نجفی[(۲)][۱]

[(۱)] گروه مهندسی صنایع، دانشکده مهدسی صنایع، دانشگاه علم وصنعت ایران، تهران، ایران.

[(۲)] مرکز تحقیقات و آموزش کشاورزی و منابع طبیعی خراسان رضوی، سازمان تحقیقات، آموزش و ترویج کشاورزی، مشهد، ایران.

تاریخ دریافت: ۱۴۰۳/۵/۲۸ تاریخ پذیرش: ۱۴۰۴/۸/۲۸ تاریخ انتشار۱۴۰۴/۹/۲۶

چکیده: در دنیای پیچیده و متغیر بازارهای مالی، بهینه‌سازی سبد سهام به عنوان یک چالش اساسی برای معامله‌گران مطرح است. از این رو معامله گران بازار سرمایه همواره با سوالات اساسی همچون اینکه از چه سهامی، در چه بازه زمانی و به چه مقدار باید خریداری کنند رو به رو هستند. هدف اصلی این پژوهش نیز با ارایه یک رویکرد نوین برای بهینه سازی سبد سرمایه با استفاده از مدل میانگین-واریانس مبتنی بر پیش‌بینی از طریق الگوریتم‌های یادگیری ماشین سنتی و یادگیری عمیق، به منظور پاسخگویی از این دست از سوالات می باشد. با توجه به ظهور روش‌های داده‌محور، این مطالعه به مقایسه عملکرد الگوریتم‌های مختلف یادگیری ماشین و یادگیری عمیق در پیش‌بینی قیمت سهام بازار بورس ایران نیز می‌پردازد. داده‌های مورد استفاده شامل قیمت بسته شدن ۹ نماد بزرگ از بازار بورس تهران در بازه زمانی ۱۰۰۰ روزه است. نتایج این مطالعه نشان می‌دهد که مدل‌های یادگیری ماشین سنتی، به ویژه رگرسیون خطی، در مقایسه با مدل‌های یادگیری عمیق، عملکرد بهتری در پیش‌بینی قیمت‌ها دارند. همچنین بهینه‌سازی سبد سهام با استفاده از روش میانگین-واریانس به دنبال انتخاب و تخصیص بهینه سرمایه به سهم های موجود برای رسیدن به بهترین بازده و کمترین ریسک را ارائه می‌دهد. این تحقیق به عنوان ابزاری عملی برای مدیران پرتفوی و تحلیلگران ریسک، بهبود مدیریت ریسک و عملکرد سبد سرمایه‌گذاری را تسهیل می‌کند.

واژه‌های کلیدی: بهینه سازی سبد سرمایه گذاری، پیش بینی قیمت سهام، یادگیری ماشین، یادگیری عمیق.

[۱] نویسنده مسئول

E-mail addresses: (A. Najaf Najafi) ali_najaf80@ind.iust.ir, (M. Najaf Najafi) mnajafi.mhd@gmail.com

مقدمه

در دنیای امروز، بازارهای مالی به عنوان یکی از مهم‌ترین و پیچیده‌ترین حوزه‌های اقتصادی شناخته می‌شوند. معامله‌گران بازار سرمایه همواره با چالش‌های متعددی مواجه هستند، از جمله اینکه از چه سهامی، در چه بازه زمانی و به چه مقدار باید خریداری کنند. این سوالات اساسی نیازمند رویکردهای نوین و کارآمد برای بهینه‌سازی سبد سرمایه‌گذاری هستند. هدف اصلی این پژوهش، ارائه یک مدل مبتنی بر داده جهت بهینه‌سازی سبد سهام با استفاده از الگوریتم‌های یادگیری ماشین و یادگیری عمیق است. به‌ویژه، این تحقیق به بررسی مدل میانگین-واریانس-در-خطر (MVaR) می‌پردازد که به پیش‌بینی بازدهی دارایی‌ها و محاسبه ریسک کمک می‌کند. به ویژه با ظهور روش های داده محور و تکنیک های یادگیری ماشین، بهینه سازی سبدهای سرمایه گذاری در سال های اخیر به طور قابل توجهی مورد توجه قرار گرفته است. روش های سنتی بهینه سازی پرتفوی، مانند رویکرد میانگین-واریانس (MV) که توسط آقای مارکوویتز ارائه شد، اساس تحلیل مالی مدرن را بنا نهاده است (Chen et al., 2006). با این حال، این روش ها غالباً به مفروضاتی متکی هستند که ممکن است در شرایط پویای بازار برقرار نباشند، و این امر ضرورت کاوش روش های قوی تر را آشکار می کند (Sheng et al., 2012).

مطالعات اخیر پتانسیل الگوریتم های یادگیری ماشینی در پیش بینی بازدهی سهام را نشان داده اند که این یک جزء حیاتی در بهینه سازی پرتفوی است. به عنوان مثال، تحقیقات نشان داده اند که مدل های یادگیری ماشینی مانند درخت تصمیم و جنگل تصادفی می توانند به دلیل توانایی خود در ذخیره روابط پیچیده و غیرخطی در داده ها، در پیش بینی قیمت سهام از روش های آماری سنتی پیشی بگیرند (Pawaskar, 2022). علاوه بر این، تکنیک های یادگیری عمیق، از جمله شبکه های حافظه طولانی کوتاه مدت (LSTM) و شبکه‌های عصبی کانولوشنی (CNN)، عملکرد مناسب در وظایف پیش بینی سری زمانی را نشان داده اند و آنها را به گزینه های مناسبی برای پیش بینی بازدهی سهام تبدیل می کنند (Fischer and Krauss, 2018).

یکپارچه سازی روش های یادگیری ماشینی و یادگیری عمیق در فرآیندهای انتخاب پرتفوی در مطالعات مختلف مورد بررسی قرار گرفته است. به عنوان مثال، مدل های ترکیبی که تکنیک های رگرسیون سنتی را با الگوریتم های یادگیری ماشینی ترکیب می کنند، برای افزایش دقت پیش بینی و بهینه سازی عملکرد پرتفوی پیشنهاد شده اند (Srivinay et al., 2022). با این حال، در ادبیات مربوط به مقایسه تحلیل روش های سنتی یادگیری ماشینی و رویکردهای یادگیری عمیق در زمینه بهینه سازی پرتفوی، به ویژه با استفاده از چارچوب میانگین-واریانس، شکاف وجود دارد.

این مطالعه با ارائه رویکردی از ترکیب مدل تحقیق در عملیات میانگین-واریانس و پیش بینی سهام ، قصد دارد این شکاف را پر کند. پیش‌بینی نمادها با استفاده از الگوریتم‌های متنوع یادگیری ماشین و یا یادگیری عمیق انجام می‌شود. در ادامه با مقایسه قابلیت پیش بینی روش‌های سنتی یادگیری ماشین با روش‌های یادگیری عمیق در بازار بورس ایران، این تحقیق به دنبال شناسایی روش های مؤثر برای بهبود مدیریت ریسک و عملکرد پرتفوی در بازارهای مالی ناپایدار است. استفاده از داده های مالی تاریخی از بازار بورس تهران امکان کاربرد عملی این تکنیک ها را فراهم می کند و بینش هایی در مورد اثربخشی آنها در سناریوهای دنیای واقعی ارائه می‌دهد.

علاوه بر این، بررسی الگوریتم های مختلف یادگیری ماشینی، از جملهCatBoost ، RNN، DNN، LSTM، Elastic Net، رگرسیون افزایش گرادیان، KNNوSVR ، امکان مقایسه جامع بین روش های سنتی و یادگیری عمیق را فراهم می کند. این تحلیل مقایسه ای برای درک نقاط قوت و ضعف هر رویکرد ضروری است و در نهایت مدیران پرتفوی و تحلیلگران ریسک را در فرآیندهای تصمیم گیری خود راهنمایی می کند. باتوجه به پیچیدگی فزاینده بازارهای مالی و نیاز به ابزارهای پیشرفته تر برای پیمایش عدم قطعیت ها برجسته می شود. با فراوان تر و پیچیده تر شدن داده های مالی، بهره گیری از تکنیک های پیشرفته یادگیری ماشینی برای پیش بینی سهام و بهینه سازی پرتفوی نه تنها مفید است، بلکه برای دستیابی به نتایج سرمایه گذاری برتر ضروری است (Kumbure et al., 2022). این مطالعه بینش های ارزشمندی را به حوزه تحلیل مالی ارائه می دهد و یک ابزار عملی و کارآمد برای مدیران پرتفوی و تحلیلگران ریسک ارائه می دهد.

در این تحقیق، با هدف برآورده کردن نیازهای معامله‌گران بازار سرمایه، به پیش‌بینی قیمت سهام با استفاده از داده‌های تاریخی و به کارگیری مدل‌های مختلف یادگیری ماشین و یادگیری عمیق پرداخته شده است. به این منظور، داده‌های مربوط به سهام بورس جمع‌آوری و پس از پیش‌پردازش، به ۱۳ مدل داده‌کاوی مختلف که هر یک جز الگوریتم‌های پر استفاده بودند، منتخب و اعمال شدند. این الگوریتم‌ها شاملLightGBM ، CatBoost، RNN، DNN، CNN، LSTM، Elastic Net، Gradient Boosting Regression، KNN، SVR، XGBoost، Random Forestو رگرسیون خطی هستند. هدف از تنوع در نظر گرفته شده در بهره اول، یافتن الگوریتمی متناسب با هر سهم به نحوی است که کمترین میزان خطا را داشته باشد.این پژوهش همچنین تأثیر عوامل مختلف مدل بر نتایج بهینه‌سازی را بررسی می‌کند و به تحلیل‌های مالی ارزشمندی می‌پردازد. به‌طور کلی، بینش‌های حاصل از این مقایسه، نقش حیاتی مدل‌سازی پیش‌بینی در انتخاب پرتفوی مبتنی بر داده‌ها را برجسته می‌کند. اثربخشی این مدل‌ها می‌تواند به طور قابل‌توجهی بر استراتژی‌های سرمایه‌گذاری و

فرآیندهای تصمیم‌گیری تأثیر بگذارد و انتخاب روش مناسب برای پیش‌بینی دقیق مالی را ضروری می‌سازد. در نهایت، این تحقیق به عنوان ابزاری عملی و کارآمد برای مدیران پرتفوی و تحلیلگران ریسک در جهت پیمایش عدم اطمینان‌های بازار عمل می‌کند و می‌تواند به بهبود مدیریت ریسک و عملکرد پرتفوی در بازارهای مالی پویا کمک کند.

۱. مواد و روش‌ها

در این بخش، به ارائه برخی از اطلاعات و تعاریف اساسی مورد نیاز در تبیین کار انجام شده پرداخته می شود. این پژوهش با بهرگیری از Python 3.12 بر روی یک سیستم CPU Core i7 (نسل ۱۳) پیاده‌سازی شده و کتابخانه‌های کلیدی مانند pandas، xgboost، sklearn، tensorflow، catboost، lightgbm، cvxpy و matplotlib مورد استفاده قرار گرفته است. داده‌ها از یک فایل Excel به یک DataFrame pandas بارگذاری می‌شوند و با تعریف ستون‌ها و ویژگی تأخیری 'Price_Lagged' برای پیش‌بینی سری زمانی آماده می‌شوند. داده‌ها به مجموعه‌های آموزشی و آزمایشی (نسبت ۸۰-۲۰) تقسیم می‌شوند تا مدل ها آموزش داده شوند. RMSE برای ارزیابی مدل بر روی داده‌های آزمایشی محاسبه می‌شود و یک نمودار پراکندگی به‌صورت بصری مقادیر واقعی و پیش‌بینی‌شده را برای بینش‌های بیشتر مقایسه می‌کند.

۱.۱ الگوریتم‌های سنتی یادگیری ماشین

در این مطالعه، از دو دسته اصلی الگوریتم‌های داده کاوی برای پیش‌بینی قیمت‌ها استفاده شده است: الگوریتم‌های سنتی یادگیری ماشین و الگوریتم‌های یادگیری عمیق. هر یک از این دسته‌ها شامل چندین الگوریتم مختلف است که به تفصیل در زیر توضیح داده می‌شوند.

۱.۱.۱ رگرسیون خطی

رگرسیون خطی یک روش یادگیری نظارت‌شده است که خروجی‌های پیوسته را با ایجاد روابط بین متغیرهای مستقل و وابسته پیش‌بینی می‌کند. هدف این روش ایجاد یک معادله خطی برای انجام پیش‌بینی‌ها با یافتن "خط بهترین برازش" است که تفاوت بین مقادیر مشاهده‌شده و پیش‌بینی‌شده را به حداقل می‌رساند. خط بهینه با فرمول خطی زیر مشخص می‌شود:

$$Y = b_0 + b_1X_1 + b_2X_2 + ... + b_nX_n$$

در این فرمول:

۱. Y نمایانگر متغیر هدفی است که ما قصد داریم پیش‌بینی کنیم.

۲. $X_1, X_2, ..., X_n$ به متغیرهای توضیحی یا ویژگی‌ها اشاره دارند.

۳. b_0 نمایانگر عرض از مبدأ مقدار Y زمانی که تمام مقادیر X صفر هستند.

۴. $b_1, b_2, ..., b_n$ پارامترهایی هستند که ارتباط بین هر متغیر توضیحی و متغیر هدف را تعریف می‌کنند.

رگرسیون خطی فرض می‌کند که بین پیش‌بین‌ها و متغیر هدف یک رابطه خطی وجود دارد. هدف مدل محاسبه ضرایب $b_0, b_1, ..., b_n$ است که واریانس‌های مربعی کل بین مقادیر پیش‌بینی‌شده و مقادیر مشاهده‌شده در مجموعه داده‌های آموزشی را کاهش می‌دهد. این فرآیند معمولاً به عنوان "تناسب مدل" شناخته می‌شود.

رگرسیون خطی به دلیل سادگی و قابلیت تفسیر بالا، به طور گسترده‌ای در تحلیل داده‌ها و پیش‌بینی‌ها مورد استفاده قرار می‌گیرد. این روش به ویژه در زمینه‌های مالی و اقتصادی، مانند پیش‌بینی قیمت سهام و تحلیل ریسک، کاربرد دارد. با توجه به اینکه رگرسیون خطی می‌تواند به راحتی روابط بین متغیرها را شناسایی کند، به عنوان ابزاری مؤثر برای تصمیم‌گیری در زمینه‌های مختلف، از جمله بهینه‌سازی پورتفولیو، شناخته می‌شود (Su et al., 2012).

۲.۱.۱ جنگل تصادفی (Random Forest)

جنگل تصادفی یک الگوریتم مبتنی بر درخت تصمیم است که از چندین درخت تصمیم برای بهبود دقت پیش‌بینی استفاده می‌کند. این الگوریتم با ایجاد چندین درخت تصمیم و ترکیب پیش‌بینی‌های آن‌ها، به کاهش واریانس و بهبود دقت کمک می‌کند. در جنگل تصادفی، هر درخت به‌طور مستقل از یک زیرمجموعه تصادفی از داده‌ها و ویژگی‌ها آموزش می‌بینند. این روش به‌ویژه در برابر اورفیتینگ مقاوم است و به‌خوبی می‌تواند با داده‌های بزرگ و پیچیده کار کند. جنگل تصادفی به‌عنوان یک روش پیشرفته‌تر نسبت به رگرسیون درختی شناخته می‌شود و به‌دلیل دقت بالای آن در مسائل مختلف، به‌طور گسترده‌ای مورد استفاده قرار می‌گیرد (Breiman, 2001).

XGBoost ۳.۱.۱

XGBoostیک الگوریتم تقویت شده است که از تکنیک‌های درختی برای بهبود دقت پیش‌بینی استفاده می‌کند. این الگوریتم با استفاده از روش‌های بهینه‌سازی و منظم‌سازی، به‌طور مؤثری می‌تواند دقت مدل را افزایش دهد. XGBoost

به‌عنوان یکی از بهترین الگوریتم‌ها در مسابقات یادگیری ماشین شناخته می‌شود و به دلیل سرعت و کارایی بالای آن، به‌طور گسترده‌ای در مسائل پیش‌بینی استفاده می‌شود. این الگوریتم از تکنیک‌های مانند Gradient Boosting و Regularizationبرای کاهش خطا و جلوگیری از اورفیتینگ استفاده می‌کند (Chen and Guestrin, 2016).

۴.۱.۱ رگرسیون بردار پشتیبان (Support Vector Regression - SVR)

رگرسیون پشتیبانی یک الگوریتم مبتنی بر نظریه پشتیبانی وکتور است که برای پیش‌بینی مقادیر پیوسته استفاده می‌شود. این الگوریتم سعی می‌کند یک تابع را پیدا کند که حداکثر حاشیه را بین نقاط داده و تابع پیش‌بینی ایجاد کند. SVR به ویژه در داده‌های با ابعاد بالا و غیرخطی عملکرد خوبی دارد و می‌تواند با استفاده از هسته‌های مختلف، روابط پیچیده را مدل‌سازی کند. این الگوریتم به‌عنوان یک روش پیشرفته‌تر نسبت به رگرسیون خطی شناخته می‌شود و به‌خوبی می‌تواند با داده‌های نویزی که دارای توزیع غیرنرمال هستند، کار کند (Vapnik, 2013).

۵.۱.۱ K-نزدیکترین همسایه(K-Nearest Neighbors - KNN)

KNN یک الگوریتم غیرپارامتریک است که برای پیش‌بینی مقادیر پیوسته و طبقه‌بندی استفاده می‌شود. این الگوریتم با محاسبه فاصله بین نقاط داده و انتخاب K نزدیک‌ترین همسایه، پیش‌بینی می‌کند. KNN به‌دلیل سادگی و عدم نیاز به آموزش مدل، به‌طور گسترده‌ای در مسائل پیش‌بینی استفاده می‌شود. با این حال، این الگوریتم به‌ویژه در داده‌های بزرگ و با ابعاد بالا ممکن است به‌طور قابل توجهی کند باشد و به‌راحتی تحت تأثیر نقاط دورافتاده قرار گیرد (Cover and Hart, 1967).

۶.۱.۱ رگرسیون گرادیان تقویتی (Gradient Boosting Regression)

رگرسیون گرادیان تقویتی یک الگوریتم مبتنی بر درخت تصمیم است که به‌طور تدریجی درختان جدیدی را به مدل اضافه می‌کند تا خطاهای مدل قبلی را کاهش دهد. این الگوریتم با استفاده از روش‌های بهینه‌سازی، به‌طور مؤثری می‌تواند دقت پیش‌بینی را افزایش دهد. رگرسیون گرادیان تقویتی به‌عنوان یک روش پیشرفته‌تر نسبت به جنگل تصادفی شناخته می‌شود و به‌خوبی می‌تواند با داده‌های پیچیده و غیرخطی کار کند (Friedman, 2001).

۷.۱.۱ شبکه الاستیک (Elastic Net)

یک الگوریتم رگرسیون است که ترکیبی از رگرسیون Lasso و Ridge را ارائه می‌دهد. این الگوریتم به‌دلیل توانایی‌اش در انتخاب ویژگی‌های مهم و جلوگیری از اورفیتینگ، به‌طور گسترده‌ای در مسائل پیش‌بینی استفاده می‌شود. شبکه الاستیک به ویژه در داده‌های با ابعاد بالا که دارای همبستگی بالایی بین ویژگی‌های داده هستند، عملکرد خوبی دارد و می‌تواند به‌عنوان یک روش پیشرفته‌تر نسبت به رگرسیون خطی در نظر گرفته شود (Zou and Hastie, 2005).

۲.۱ الگوریتم‌های یادگیری عمیق

یادگیری عمیق، از ساختارهای پیچیده‌ای تشکیل شده‌اند که می‌توانند الگوهای پیچیده را شناسایی کنند. این الگوریتم‌ها به خصوص الگوریتم های یادگیری عمیق، با استفاده از لایه‌های متعددی از نورون‌ها، قادر به یادگیری ویژگی‌های پیچیده از داده‌ها هستند. شبکه‌های عصبی به‌ویژه در مسائل پیش‌بینی با داده‌های بزرگ و پیچیده، عملکرد بسیار خوبی دارند. این الگوریتم‌ها می‌توانند به‌طور خودکار ویژگی‌های مهم را از داده‌ها استخراج کنند و به‌دلیل توانایی‌شان در یادگیری غیرخطی، به‌طور گسترده‌ای در مسائل مختلف مورد استفاده قرار می‌گیرند (Goodfellow, 2016).

۱.۲.۱ حافظه طولانی کوتاه مدت (LSTM)

LSTM یک نوع خاص از شبکه‌های عصبی بازگشتی (RNN) است که برای پردازش داده‌های توالی‌دار طراحی شده است. این الگوریتم به‌خوبی می‌تواند وابستگی‌های طولانی‌مدت در داده‌ها را شناسایی کند و به‌ویژه در مسائل پیش‌بینی زمانی کاربرد دارد .LSTM به‌عنوان یک پیشرفت نسبت بهRNN های سنتی شناخته می‌شود، زیرا می‌تواند مشکلاتی مانند ناپایداری گرادیان را حل کند. این الگوریتم با استفاده از واحدهای حافظه، قادر به حفظ اطلاعات مهم در طول زمان است و به‌خوبی می‌تواند با داده‌های توالی‌دار کار کند (Hochreiter and Schmidhuber, 1997).

۲.۲.۱ شبکه‌های عصبی کانولوشنی (CNN)

شبکه‌های عصبی کانولوشنی (CNN) به‌طور خاص برای پردازش داده‌های تصویری طراحی شده‌اند، اما می‌توانند در

مسائل پیش‌بینی زمانی نیز استفاده شوند. این الگوریتم با استفاده از لایه‌های کانولوشن، قادر به استخراج ویژگی‌های مهم از داده‌ها است CNN .به‌عنوان یک روش پیشرفته‌تر نسبت به شبکه‌های عصبی سنتی شناخته می‌شود و می‌تواند به‌طور مؤثری در مسائل پیچیده‌تر عمل کند. این الگوریتم به‌ویژه در شناسایی الگوها و ویژگی‌های محلی در داده‌ها بسیار مؤثر است (LeCun et al., 1998).

۳.۲.۱ شبکه‌های عصبی عمیق (DNN)

شبکه‌های عصبی عمیق به‌عنوان یک نوع پیشرفته از شبکه‌های عصبی، شامل چندین لایه پنهان هستند که به آن‌ها اجازه می‌دهد تا ویژگی‌های پیچیده‌تری را یاد بگیرند. این الگوریتم‌ها به‌دلیل توانایی‌شان در یادگیری از داده‌های بزرگ و پیچیده، به‌طور گسترده‌ای در مسائل مختلف مورد استفاده قرار می‌گیرند DNN .می‌تواند به‌عنوان یک پیشرفت نسبت به شبکه‌های عصبی ساده در نظر گرفته شود، زیرا می‌تواند الگوهای پیچیده‌تری را شناسایی کند و به‌دلیل عمق بیشتر، دقت پیش‌بینی را افزایش دهد(Bengio, 2009).

۴.۲.۱ شبکه‌های عصبی بازگشتی (RNN)

شبکه‌های عصبی بازگشتی به‌طور خاص برای پردازش داده‌های توالی‌دار طراحی شده‌اند. این الگوریتم‌ها با استفاده از اتصالات بازگشتی، قادر به حفظ اطلاعات از ورودی‌های قبلی هستند و به‌خوبی می‌توانند وابستگی‌های زمانی را شناسایی کنند RNN .به‌عنوان یک روش پیشرفته‌تر نسبت به شبکه‌های عصبی سنتی شناخته می‌شود و به‌ویژه در مسائل پیش‌بینی زمانی و پردازش زبان طبیعی کاربرد دارد (Elman, 1990).

CatBoost ۵.۲.۱

CatBoost یک الگوریتم یادگیری تقویتی است که به‌طور خاص برای کار با داده‌های دسته‌ای طراحی شده است. این الگوریتم به‌دلیل توانایی‌اش در پردازش داده‌های با ویژگی‌های دسته‌ای و جلوگیری از اورفیتینگ، به‌طور گسترده‌ای در مسائل پیش‌بینی استفاده می‌شود. CatBoost به‌عنوان یک روش پیشرفته‌تر نسبت به XGBoost شناخته می‌شود و به‌خوبی می‌تواند با داده‌های پیچیده و غیرخطی کار کند (Dorogush et al., 2018).

۶.۲.۱ LightGBM

LightGBMیک الگوریتم یادگیری تقویتی است که به‌دلیل سرعت و کارایی بالای آن در پردازش داده‌های بزرگ شناخته می‌شود. این الگوریتم با استفاده از تکنیک‌های بهینه‌سازی، می‌تواند به‌طور مؤثری دقت پیش‌بینی را افزایش دهد. LightGBMبه‌عنوان یک روش پیشرفته‌تر نسبت به XGBoost و CatBoost شناخته می‌شود و به‌خوبی می‌تواند با داده‌های پیچیده و غیرخطی کار کند(Ke et al., 2017).

در این مقاله، به بررسی بهینه‌سازی سبد سهام با استفاده از روش میانگین_ واریانس با پیش‌بینی قیمت نمادها ازطریق الگوریتم‌های سنتی یادگیری ماشین و یادگیری عمیق می‌پردازد. هدف از استفاده این دو رویکرد (بهینه‌سازی سبد و پیش‌بینی قیمت نمادها)، شناسایی و پیش‌بینی بازده‌های بالقوه سهام و بهینه‌سازی ترکیب نماد ها سبد سهام برای کاهش ریسک و افزایش بازده می باشد. این مطالعه به بررسی چالش‌ها و فرصت‌های موجود در استفاده از این دو روش داده کاوی در تحلیل مالی می‌پردازد و به دنبال ارائه بینش‌های جدید در زمینه بهینه‌سازی سبر سرمایه‌گذاری است.

۲. بهینه‌سازی سبد سهام با استفاده از روش میانگین_ واریانس

فرض کنید $n \geq 2$ تعداد دارایی های پرخطری که سرمایه گذار تصمیم می گیرد برای مدت زمان ثابت T سرمایه گذاری کند. همچنین فرض کنید r_i نشان دهنده نرخ بازده دارایی i است. در این صورت نرخ بازده مورد انتظار دارایی i با فرمول $\mu i = \frac{1}{T}\sum_{t=1}^{T} r_{it}$ محاسبه می شود. فرض کنید x_i نسبت ثروتی است که در دارایی i سرمایه گذاری شده است، به طوری که $\sum_{i=1}^{n} x_i = 1$. در این صورت، نرخ بازدهی سبد سرمایه گذاری با $r_p = \sum_{i=1}^{n} r_i x_i$ و بازدهی مورد انتظار سبد سرمایه گذاری با $\mu_p = \sum_{i=1}^{n} \mu_i x_i$ محاسبه می شود. واریانس پرتفوی به این روش $\sigma_p^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} \sigma_{ij}\, x_i x_j$ محاسبه می گردد.

این مقاله به مدل بهینه‌سازی پرتفوی میانگین-واریانس پرداخته است. مسئله بهینه‌سازی پرتفوی میانگین-VaR به صورت زیر بیان می‌شود:

$$\min VaR = a^* \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} \sigma_{ij} x_i x_j} - \sum_{i=1}^{n} \mu_i x_i$$

$$s.t: \sum_{i=1}^{n} \mu_i x_i = \mu_{fix},$$

$$\sum_{i=1}^{n} x_i = 1, 0 \le x_i \le 1, \forall i = 1, 2, ..., n$$

مدل میانگین-واریانس، واریانس را به عنوان یک معیار ریسک با مدل بهینه سازی پرتفوی میانگین-واریانس ترکیب می کند. هدف این مدل، به حداقل رساندن واریانس پرتفوی در حالی که اطمینان حاصل می شود بازده مورد انتظار به یک مقدار حداقل از پیش تعیین شده برسد و درصد سهم های سرمایه گذاری دارای مجموع یک باشند (Sheng et al., 2012).

۳. خطای جذر میانگین مربعات (RMSE)

خطای میانگین مربعات (RMSE) یک معیار پذیرفته شده برای ارزیابی عملکرد مدل در هواشناسی، کیفیت هوا و تحقیقات اقلیمی است. RMSE ریشه دوم میانگین مربعات خطاها را محاسبه می کند و فرض می کند که نمونه های گرفته شده نااریب بوده و همچنین از خطاهای مربوط مشاهده یا روش مقایسه نیز چشم پوشی می کند. در مقابل نحوه محاسبه خطا با این روش آورده شده است (Chai and Draxler, 2014).

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} e_i^2}$$

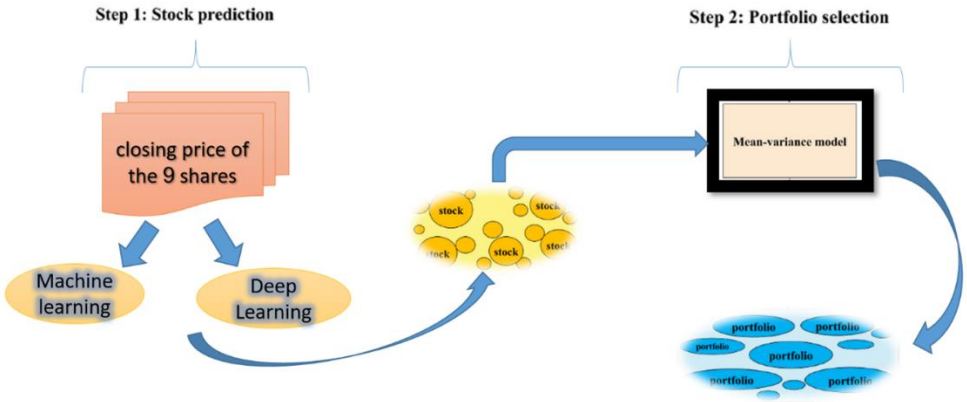
۴. داده های مورد استفاده

داده‌های مورد استفاده در این پژوهش شامل ۹ نماد بزرگ از صنایع مختلف در بازار بورس تهران است. این نماد ها شامل بپاس، فملی، فولاد، غنوش، وتجارت، خودرو، کحافظ، شپنا، وساخت می باشد. مجموعه داده به صورت یک فایل اکسل با ۹ ستون شامل قیمت بسته شدن این سهم ها در بازه‌ای تقریبی به طول هزار روز از فروردین ۱۳۹۹ تا اردیبهشت ۱۴۰۳ سازماندهی شده است. به منظور جلوگیری از تأثیر همبستگی نمادهای صنعتی خاص بر روی نتایج، سعی شد تا از گروه‌های صنعتی متفاوتی نمادها انتخاب شوند. ویژگی در نظر گرفته شده برای هر سهم، قیمت پایانی آن سهم در بازار معاملاتی آن روز بوده است. داده‌ها تماماً از سایت رسمی اطلاع رسانی بازار سرمایه کدال گرفته شدند. همچنین سعی شد تا بازه طولانی

چهار ساله اتخاذ شود تا حوادث مقطعی تأثیر کمتری بر روی نتایج داشته باشند. در انتها یک جدول تک ستونی از قیمت پایانی هر نماد در هر روز برای تحلیل‌های صورت گرفته در این تحقیق استفاده شد.

۲. نتایج و بحث

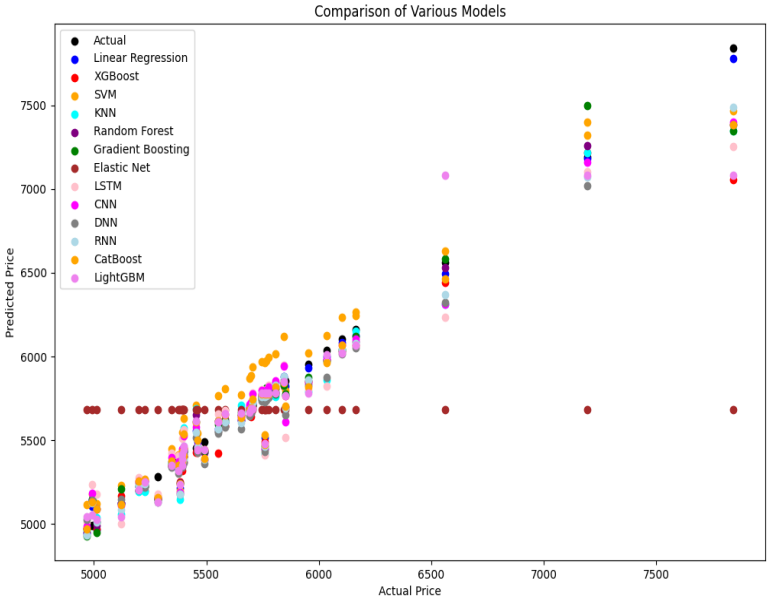
در این تحقیق با هدف اصلی برآورده کردن نیاز های معاملگران بازار سرمایه، در گام نخست به پیش‌بینی قیمت سهام با استفاده از داده‌های تاریخی و به کارگیری مدل‌های مختلف یادگیری ماشین و یادگیری عمیق پرداخته شد. به این منظور داده‌های مربوط به سهام بورس جمع‌آوری و پس از پیش‌پردازش، به ۱۳ مدل داده کاوی مختلف که هر یک جز الگوریتم های پر استفاده بودند، منتخب و اعمال شدند. این الگوریتم ها شامل LightGBM ، CatBoost، RNN، DNN، CNN، LSTM، Elastic Net، Gradient Boosting Regression، KNN، SVR، XGBoost، Random Forest و رگرسیون خطی اعمال بودند. هدف از تنوع در نظر گرفته شده در بهره اول یافتن الگوریتمی متناسب با هر سهم، به نحوی که کمترین میزان خزا را داشته باشد بود. هدف بعدی مقایسه این بین روش های سنتی یادگیری ماشین و روش های ایدگیری عمیق می باشد. از این رو مدل‌ها به منظور پیش‌بینی قیمت‌ها در یک بازه ۴۰ روزه آموزش داده شدند و نتایج به دست آمده با استفاده از معیار جذر میانگین مربعات ارزیابی گردیدند. این معیار کمک می‌کند تا میزان خطای پیش‌بینی هر مدل به طور دقیق اندازه‌گیری شود.



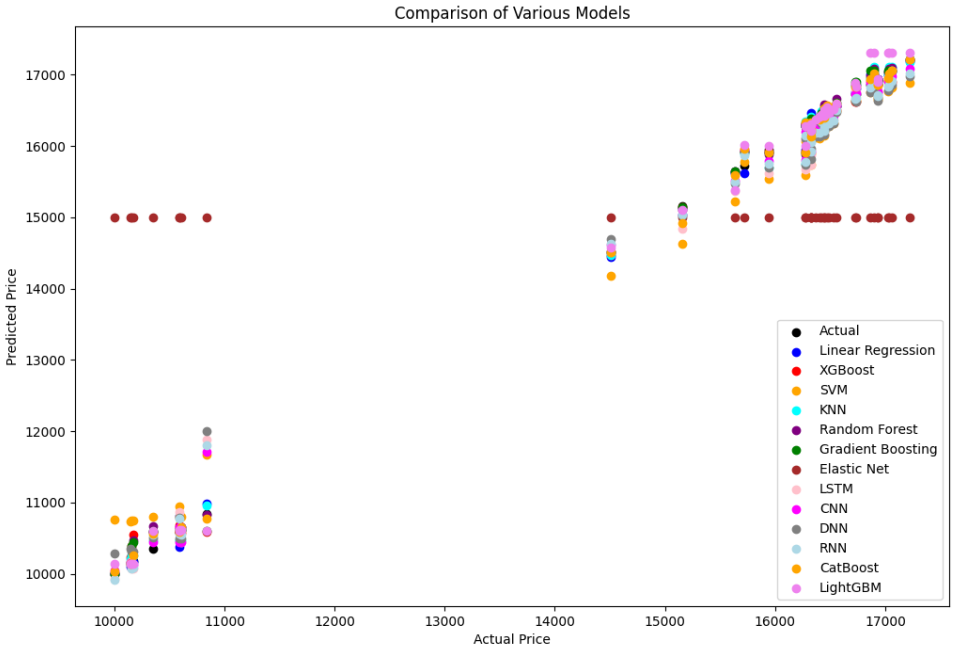
شکل۱: خلاصه رویکرد ارایه شده

در ادامه حاصل پیش بینی هر یک از الگوریتم های داده کاوی برای هر نه نماد در شکل۲ آورده شده است.

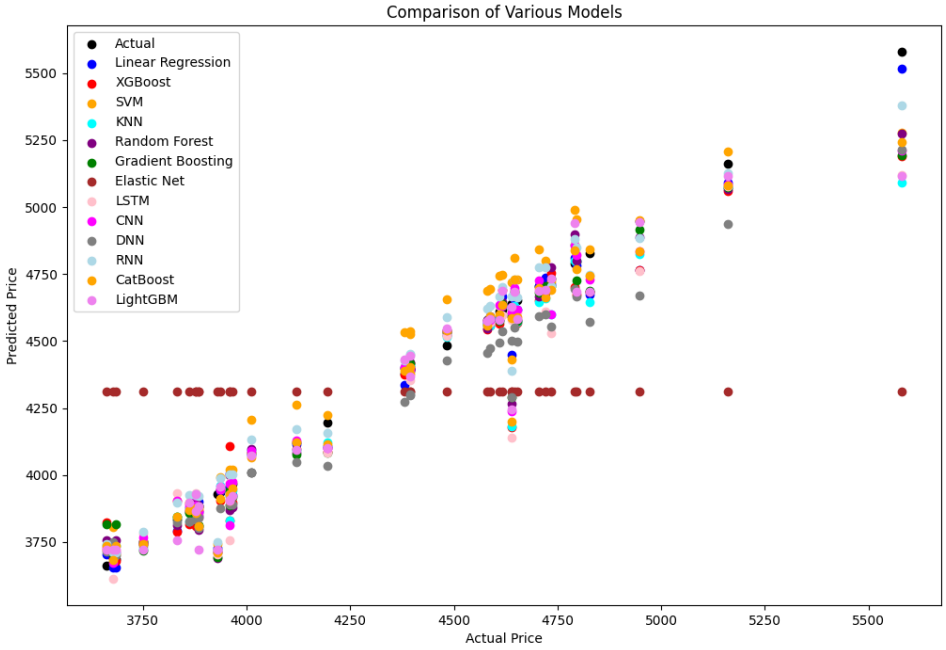
فمل



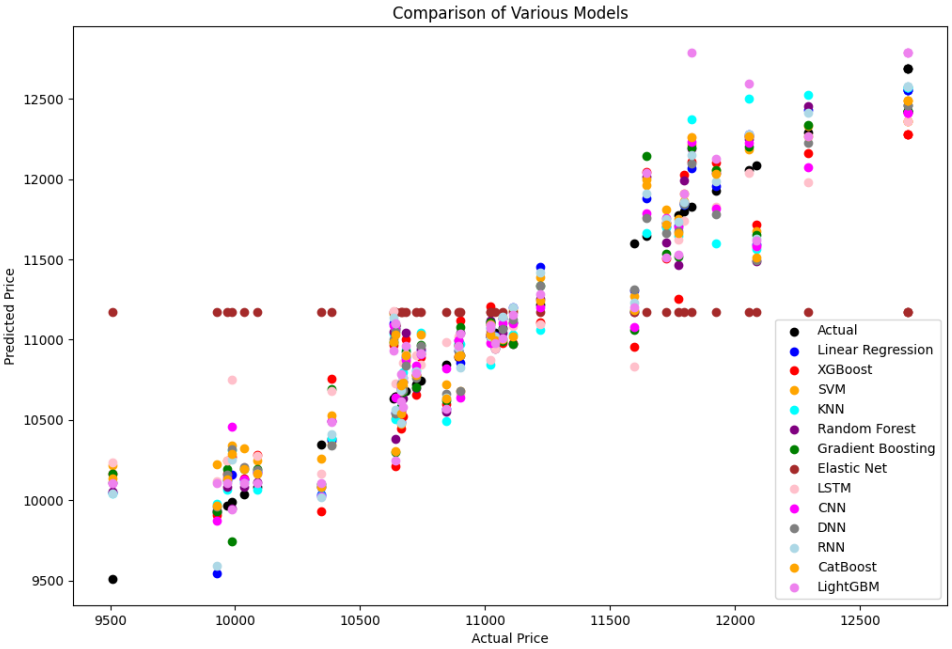
پپاس:



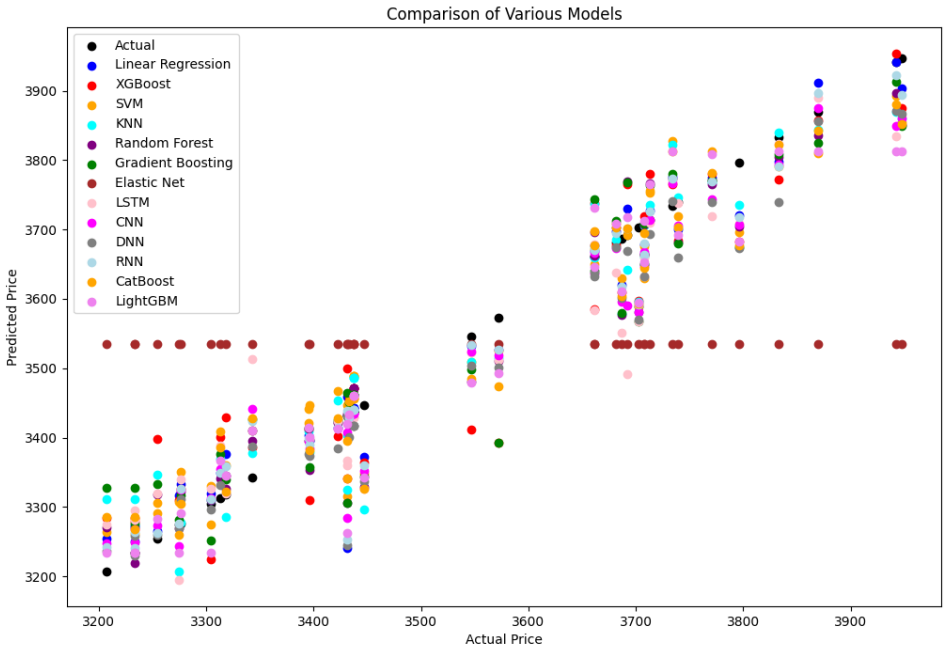
فولاد:



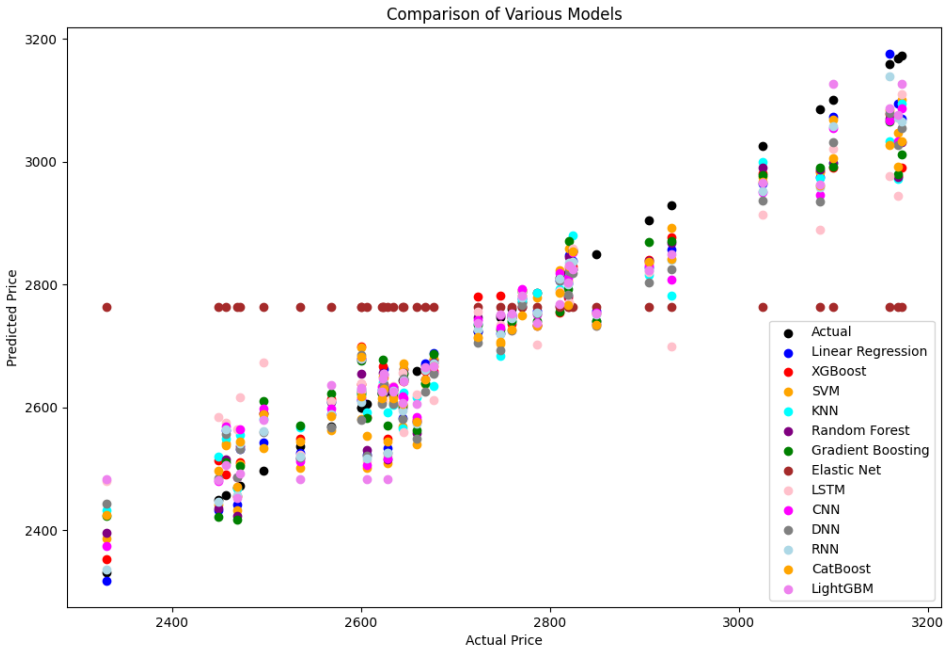
غنوش:



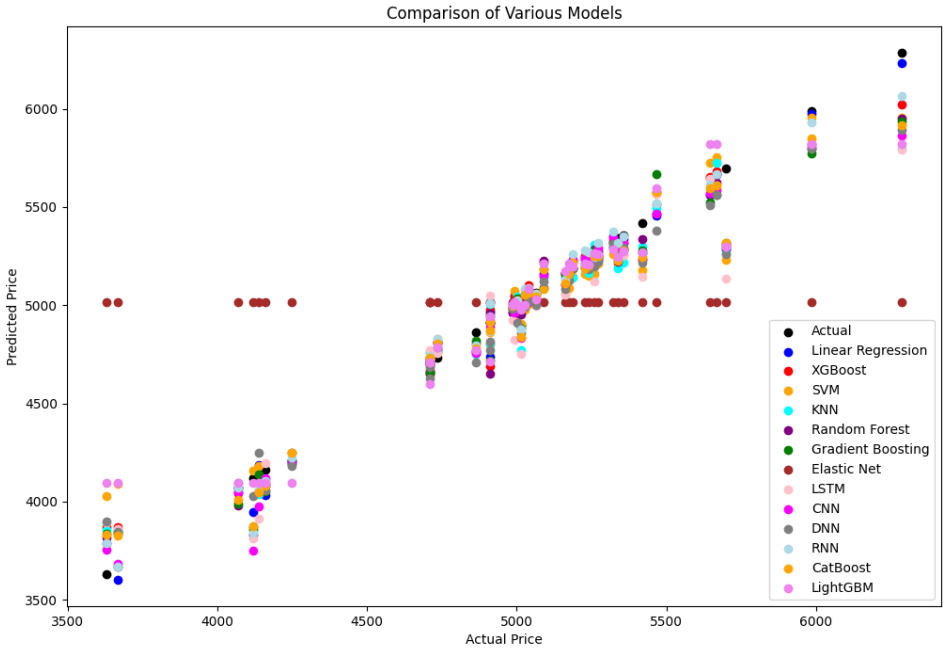
کحافظ:



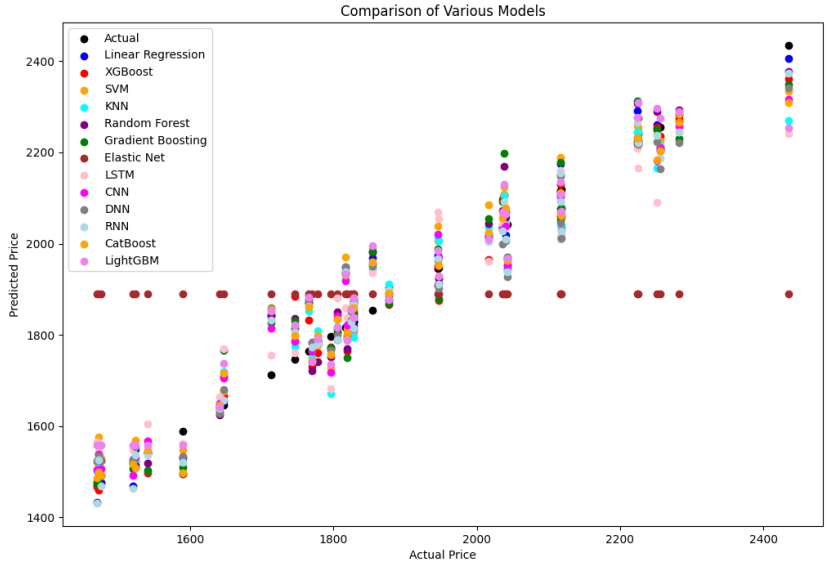
خودرو:



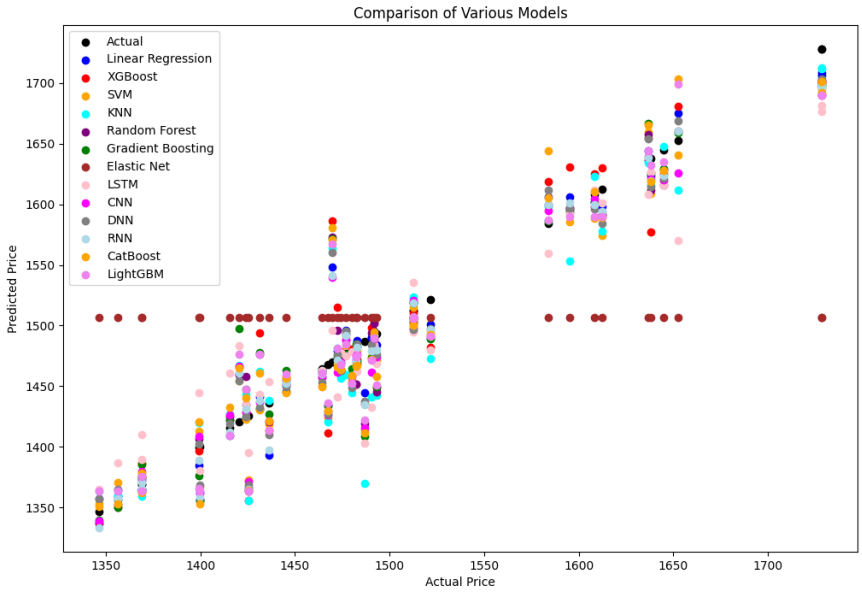
شپنا:



وساخت

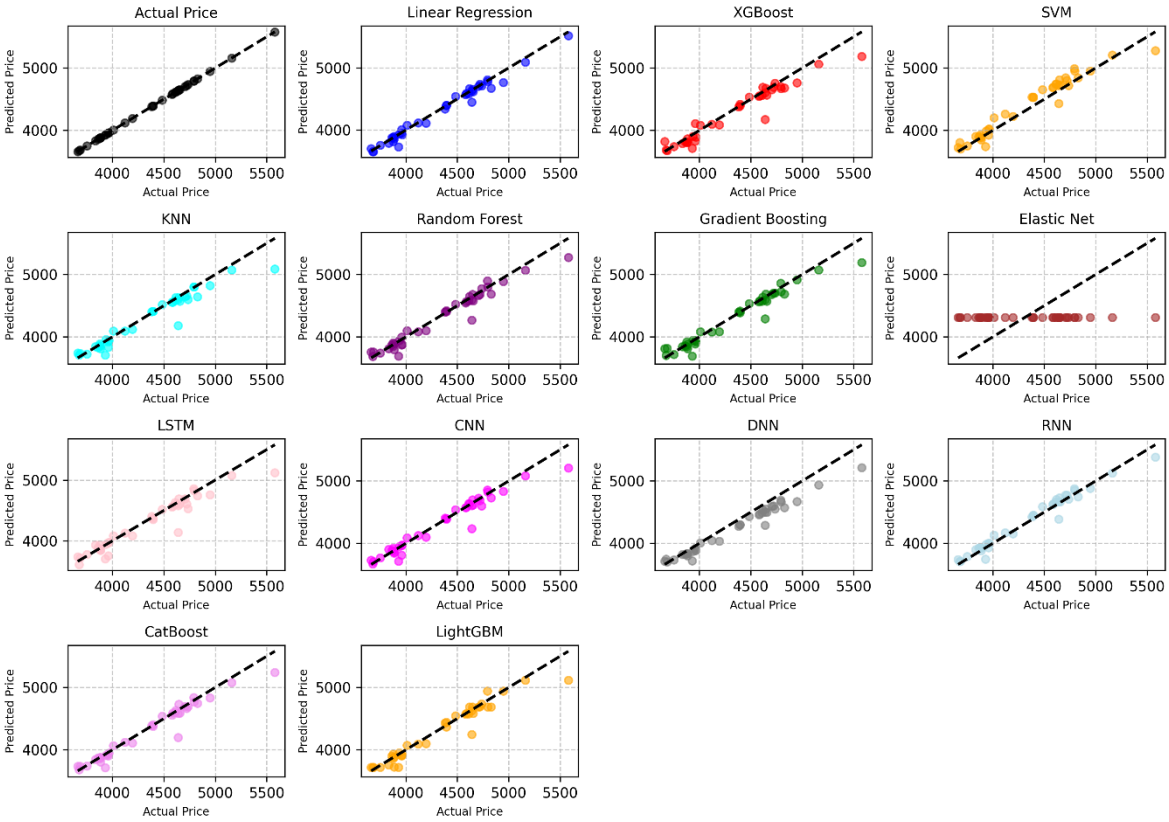


وتجارت:



شکل۲: حاصل پیش بینی حاصل از هر نماد

این نتایج نشان‌دهنده اهمیت انتخاب مدل مناسب در پیش‌بینی قیمت سهام و تأثیر آن بر تصمیم‌گیری‌های سرمایه‌گذاری است. همچنین با توجه به روند قیمتی که هر سهم در بر می گیرد، الگوریتم های مختلف نتایج متفاوتی را در بر داشتند که ضرورت امتحان روش های مختلف بر روی سهم خای متفاوت را تصدیق می کند. در ادامه جهت نمایش بهتر در روند های حاصل شده از مقادیر پیش بینی شده با مقدار واقعی نماد فولاد، شکل۳ آورده شده است.



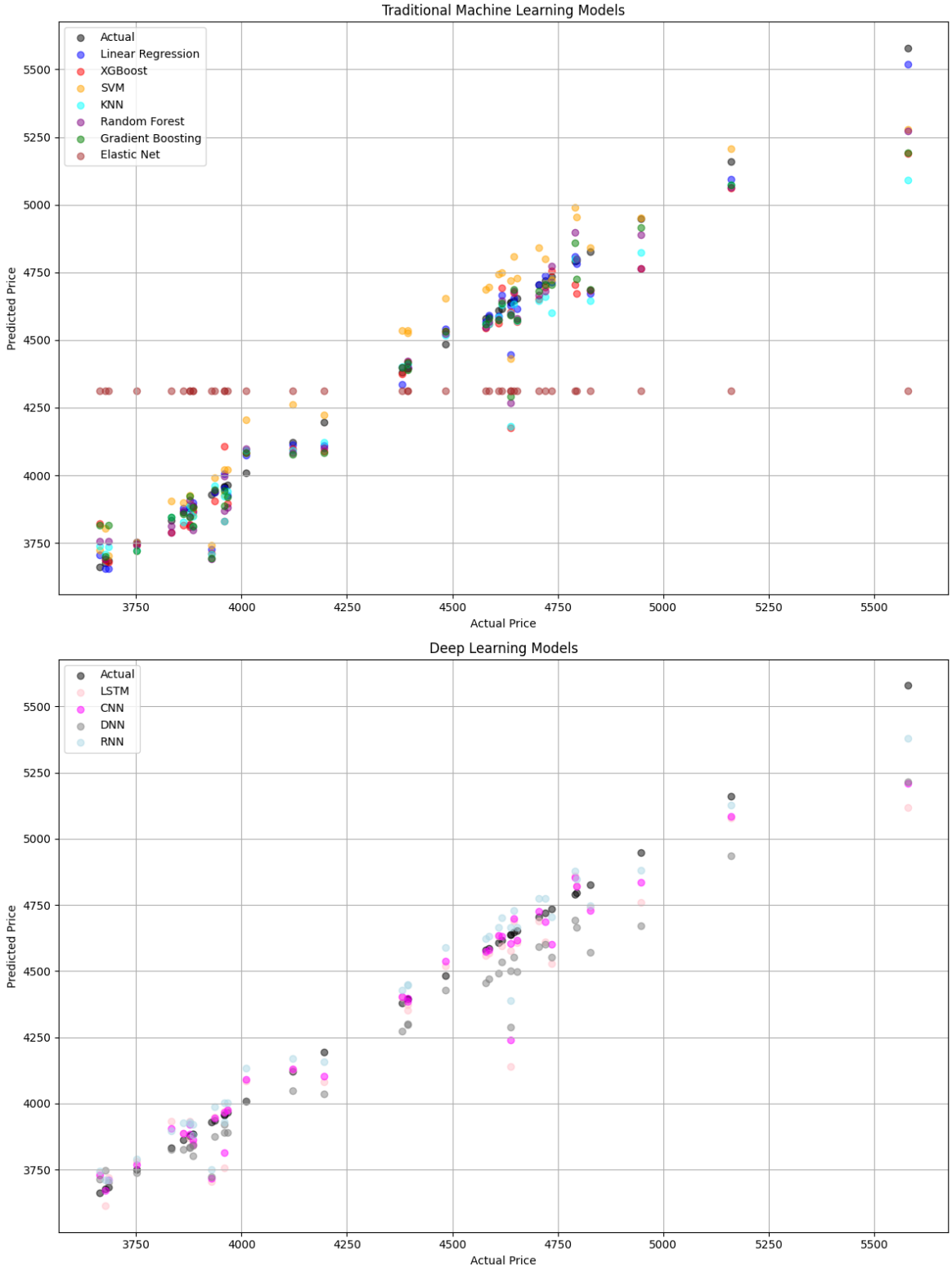
شکل۳: تفکیک نتایج حاصل از هر الگوریتم در نماد فولاد

این نمودار، مقایسه بصری کاملی از مدل‌های مختلف یادگیری ماشین و یادگیری عمیق در پیش‌بینی قیمت ارائه می‌دهد. هر زیرنمودار، رابطه بین قیمت‌های واقعی و پیش‌بینی‌های مربوط به آن توسط مدل‌های مختلف را نشان می‌دهد. خط مورب در هر زیرنمودار به عنوان معیاری برای پیش‌بینی‌های کامل عمل می‌کند، جایی که مقادیر پیش‌بینی شده با مقادیر واقعی

مطابقت دارند، مدل مربوطه عملکرد بهتری در پیش‌بینی ارائه داده است. همانطور که قابل مشاهده است، برخی مدل‌ها هم‌راستایی نزدیک‌تری با خط مورب نشان می‌دهند، که نشان‌دهنده عملکرد پیش‌بینی بهتری است. تراکم نقاط در اطراف این خطوط نشان می‌دهد که این مدل‌ها دچار بیش برازش نشده و منجر به پیش‌بینی‌های دقیق‌تر می‌شود. برعکس، مدل‌هایی که پراکندگی وسیع‌تری از پیش‌بینی‌ها را نشان می‌دهند، نشان‌دهنده چالش‌هایی در پیش‌بینی دقیق قیمت‌ها هستند، که نشان‌دهنده زمینه‌های احتمالی برای بهبود است. این تحلیل نشان می‌دهد که در حالی که مدل‌های سنتی عملکرد بهتری را نشان می‌دهند، برخی مدل‌های پیشرفته در حفظ دقت در سراسر طیف قیمت‌ها با مشکل مواجه هستند. این نوسانات، اهمیت انتخاب مدل در دستیابی به پیش‌بینی‌های قابل اعتماد را برجسته می‌کند. نتایج، ضرورت توجه دقیق به ویژگی‌های مدل و مناسب بودن آن‌ها برای مجموعه داده‌های خاص را نشان می‌دهند.

به طور کلی، بینش‌های حاصل از این مقایسه، نقش حیاتی مدل‌سازی پیش‌بینی در انتخاب پرتفوی مبتنی بر داده‌ها را برجسته می‌کند. اثربخشی این مدل‌ها می‌تواند به طور قابل‌توجهی بر استراتژی‌های سرمایه‌گذاری و فرآیندهای تصمیم‌گیری تأثیر بگذارد، و انتخاب روش مناسب برای پیش‌بینی دقیق مالی را ضروری می‌سازد. این مقایسه بصری به عنوان ابزاری ارزشمند برای درک نقاط قوت و ضعف تکنیک‌های مختلف مدل‌سازی، راهنمایی تحقیقات آینده و کاربردهای عملی در این زمینه عمل می‌کند.

برای مقایسه نتایج حاصل از روش های سنتی یادگیری ماشین و روش های یادگیری عمیق شکل۴ در نظر گرفته شده است. لازم به ذکر است این شکل حاصل از نتایج نماد فولاد می باشد.



شکل۴: مقایسه روش های یادگیری ماشین سنتی با روش های یادگیری عمیق.

همانطور که گفته شد در این مقاله، یک مقایسه از عملکرد مدل‌های مختلف یادگیری ماشین سنتی و یادگیری عمیق در پیش‌بینی قیمت ارائه می شود. نمودارهای پراکندگی، رابطه بین قیمت‌های واقعی و قیمت‌های پیش‌بینی شده توسط هر مدل را نشان می‌دهند و امکان ارزیابی بصری دقت آن‌ها را فراهم می‌کنند.

در زیر نمودار اول با عنوان "مدل‌های یادگیری ماشین سنتی"، پیش‌بینی‌های چندین الگوریتم، از جمله رگرسیون خطی، XGBoost، ماشین بردار پشتیبان (SVM)، کا امین نزدیک ترین همسایگی (KNN) ، جنگل تصادفی، تقویت گرادیان و شبکه الاستیک قابل مشاهده است. خط مورب سیاه نشان دهنده قیمت های واقعی در داده های تست هستند که در آن مقادیر پیش‌بینی‌شده به طور کامل با مقادیر واقعی مطابقت دارند. هر چه نقاط به این خط نزدیک‌تر باشند، پیش‌بینی‌های مدل بهتر خواهد بود. قابل توجه است که رگرسیون خطی عملکرد قوی‌ای را نشان می‌دهد و نرخ خطای کمتری را نشان می دهد. در مقابل، مدل‌هایی مانند KNN و XGBoost پراکندگی وسیع‌تری از مقادیر تست را نشان می‌دهند، که بیانگر نوسانات بالاتر و احتمالاً قابلیت اطمینان کمتر در پیش‌بینی‌های آن‌ها است.

نمودار دوم با عنوان "مدل‌های یادگیری عمیق"، عملکرد معماری‌های پیشرفته شبکه عصبی، از جمله LSTM، CNN، DNN و RNN را به نمایش می‌گذارد. در اینجا، پیش‌بینی‌های مدل‌های LSTM و RNN به نظر می‌رسد که بیشتر به دور خط مورب جمع شده‌اند، که نشان می‌دهد این مدل‌ها قادر به درک الگوهای نهفته در داده‌ها به طور موثر هستند. با این حال، مدل‌های DNN و CNN توزیع گسترده‌تری از پیش‌بینی‌ها را نشان می‌دهند، که نشان می‌دهد آن‌ها ممکن است در مقایسه با LSTM و RNN به خوبی به داده‌های ندیده تعمیم نیابند.

تحلیل نمودارهای پراکندگی، نقاط قوت و ضعف هر مدل را برجسته می‌کند. مدل‌های یادگیری ماشین سنتی، به ویژه رگرسیون خطی، عملکرد قوی‌ای را نشان می‌دهند، در حالی که مدل‌های یادگیری عمیق مانند LSTM، نویدبخش درک روابط پیچیده هستند، اما ممکن است برای دستیابی به نتایج مطلوب به تنظیم بیشتر نیاز داشته باشند. این تحلیل مقایسه‌ای، اهمیت انتخاب مدل در مدیریت پرتفوی مبتنی بر داده را برجسته می‌کند، جایی که پیش‌بینی دقیق قیمت برای تصمیم‌گیری آگاهانه ضروری است.

نکته جالب توجه این است که روش‌های یادگیری ماشینی سنتی در مقایسه با روش‌های یادگیری عمیق، عملکرد پایاپایی را در زمینه پیش‌بینی قیمت نشان دادند. در میان مدل‌های سنتی، رگرسیون خطی به عنوان موثرترین مدل ظاهر شد و

پیش‌بینی‌های دقیقی ارائه کرد. سادگی و قابلیت تفسیر این مدل آن را به یک انتخاب قوی تبدیل کرد، که نشان می‌دهد روابط اساسی در داده‌ها ممکن است نسبتا ساده باشند.

در حالی که XGBoost و دیگر مدل‌های سنتی مانند SVM، KNN، جنگل تصادفی و افزایش گرادیان، درجات مختلفی از اثربخشی را نشان دادند، اما عملکرد رگرسیون خطی را پشت سر نگذاشتند. قابل توجه است که Elastic Net به طور قابل توجهی با مشکل مواجه شد، که نشان می‌دهد این مدل برای این مجموعه داده‌ی خاص کمتر مناسب است.

از طرف دیگر، مدل‌های یادگیری عمیق، به ویژه LSTM، توانایی درک الگوهای پیچیده در داده‌ها را نشان دادند. با این حال، عملکرد آن‌ها از مدل‌های سنتی ساده‌تر فراتر نرفت. این نشان می‌دهد که در حالی که یادگیری عمیق می‌تواند قدرتمند باشد، اما همیشه لازم یا مفید نیست، به خصوص زمانی که مجموعه داده‌ها به اندازه کافی بزرگ یا پیچیده نیستند تا از مزایای این تکنیک‌های پیشرفته بهره‌مند شوند.

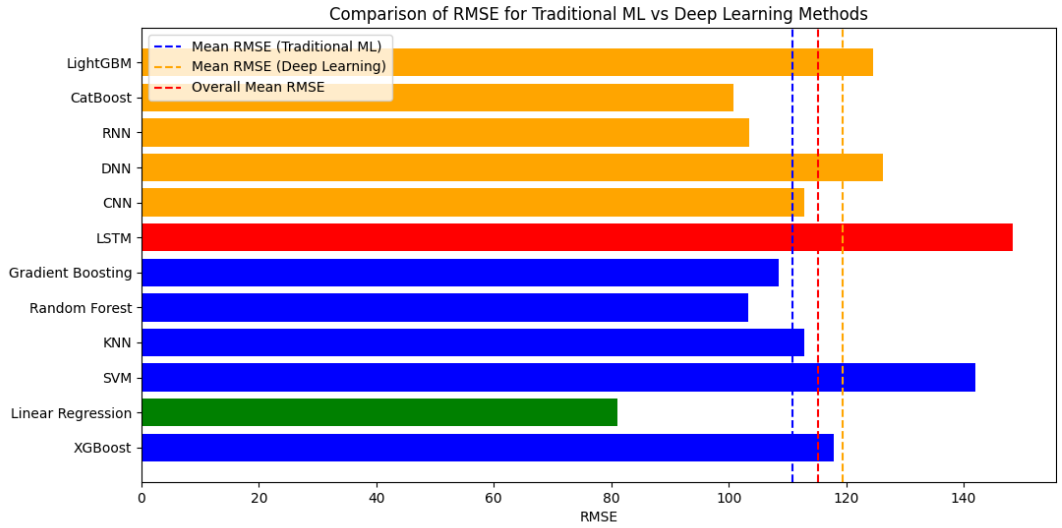
پراکندگی گسترده‌تر پیش‌بینی‌ها از مدل‌های یادگیری عمیق مانند CNN، DNN و RNN نشان می‌دهد که این مدل‌ها ممکن است به طور کلی به داده‌های آزمون تعمیم نیافته باشند، که ممکن است به دلیل فرا یادگیری یا داده‌های آموزشی ناکافی باشد. این امر بر اهمیت انتخاب مدل بر اساس ویژگی‌های مجموعه داده‌ها تاکید می‌کند.

به طور کلی، روش‌های یادگیری ماشینی سنتی در این تحلیل، قابل اعتمادتر و موثرتر برای پیش‌بینی قیمت ثابت شدند. نتایج بر ضرورت در نظر گرفتن دقیق ماهیت داده‌ها و اهداف تحلیل در هنگام انتخاب بین روش‌های سنتی و یادگیری عمیق تاکید می‌کند. برای انتخاب پرتفوی مبتنی بر داده‌ها، جایی که پیش‌بینی دقیق قیمت بسیار مهم است، روش‌های سنتی ممکن است راه حل مطمئن‌تری ارائه دهند، به خصوص زمانی که قابلیت تفسیر و سادگی ارزش زیادی داشته باشد. در جدول۱ مقادیر جذر مربعات خطای مربوط به هر روش و هر سهم آورده شده است.

جدول۱: مقادیر جذر میانگین مربعات خطا

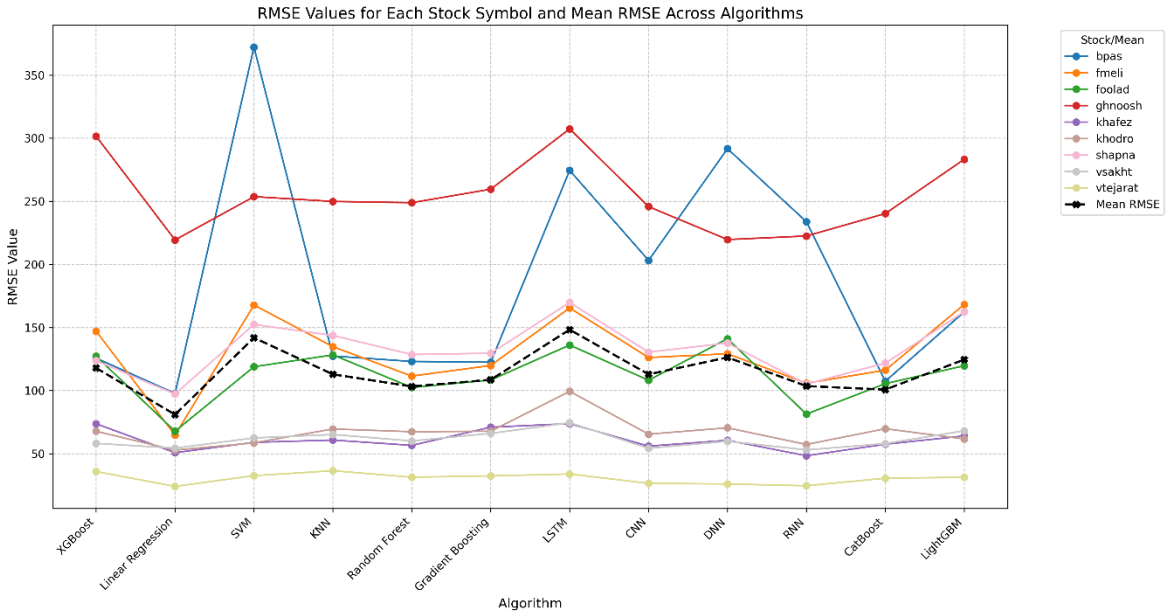
Model	bpas	fmeli	foolad	ghnoosh	khafez	khodro	shapna	vsakht	vtejarat	mean
XGBoost	125.7	147.1	127.2	301.6	73.7	67.8	123.7	58.2	35.9	117.9
Linear Regression	97.8	64.7	67.8	219.4	50.8	52.8	97.3	54.4	24.1	81.0
SVM	372.2	167.8	118.9	253.7	59.0	58.6	152.4	62.5	32.6	141.9
KNN	127.4	134.8	128.3	249.9	60.8	69.6	143.7	65.2	36.5	112.9
Random Forest	123.1	111.5	102.5	248.9	56.6	67.4	128.6	60.1	31.4	103.3
Gradient Boosting	122.5	119.9	108.4	259.6	71.0	67.8	129.6	66.1	32.4	108.6
Elastic Net	2591.3	546.9	460.8	813.0	213.2	213.8	571.9	251.5	97.6	640.0
LSTM	274.5	165.5	136.1	307.4	73.8	99.4	169.8	74.5	33.9	148.3
CNN	203.2	126.2	108.2	245.8	56.0	65.5	130.5	54.3	26.5	112.9
DNN	291.7	129.2	141.1	219.7	60.6	70.5	137.7	59.9	26.0	126.3
RNN	233.8	105.9	81.4	222.5	48.4	57.3	105.2	53.0	24.6	103.6
CatBoost	107.4	116.2	105.5	240.3	57.5	69.8	121.8	58.0	30.6	100.8
LightGBM	162.4	168.3	119.7	283.3	64.1	61.7	162.5	68.2	31.3	124.6

شکل۵ برای درک مناسب تر از جدول بالا ترسیم گردیده است. در حقیقت این نمودار از ستون اخر جدول۱ استفاده کرده و به میانگین خطای حاصل از سیزده سهم برای هر یک از روش هلی پیش بینی آورده شده است. لازم به ذکر است که با توجه به عدم پاسخ گویی مناسب روش شبکه الاستیک، این روش به عنوان داده پرت از نمودار حذف گردیده است و در تحلیل های بعد از این محاسبه نمی گردد.



شکل۵: نمودار میانگین خطا

نمودار میله‌ای ارائه شده، تحلیل مقایسه‌ای خطای میانگین مربعات (RMSE) را برای مدل‌های مختلف یادگیری ماشین و یادگیری عمیق کهبرای پیش‌بینی قیمت بسته شدن سهام استفاده شده‌اند، نشان می‌دهد.



شکل۶: میزان خطای پیش‌بینی به تفکیک هر روش و نماد

شکل ۶ به تفاوت خطا در هر روش را به تفکیک هر نماد نشان می‌دهد. هر خط روی نمودار، یک نماد را نشان می‌دهد و بیان می‌کند که چگونه میانگین مربعات خطا برای آن سهام در مدل‌های یادگیری ماشین و یادگیری عمیق مورد استفاده است. خط میانگین ارائه شده، میانگین خطا را در تمام نمادها برای هر الگوریتم نشان می‌دهد. این شکل با تفکیک میانگین مربعات خطا در الگوریتم‌ها و نمادها برای شناسایی مؤثرترین مدل‌های پیش‌بینی متناسب با ویژگی‌های منحصر به فرد ابزارهای مالی منفرد بسیار ارزشمند است.مقدار میانگین مربعات خطا پایین‌تر، دقت پیش‌بینی برتر را نشان می‌دهد. از شکل، مشخص است که عملکرد مدل در بین نمادهای سهام بسیار ناهمگن است. به عنوان مثال، سهام وتجارت به طور مداوم کمترین مقادیر میانگین مربعات خطا را در تقریباً تمام الگوریتم‌ها نشان می‌دهد، که بیان می‌کند پیش‌بینی آن نسبتاً آسان‌تر است. در مقابل نماد بپاس نسبت به اکثر مدل‌ها میانگین مربعات خطا بسیار بالاتری دارد، به ویژه با الگوریتم SVM که خطای پرت بالایی را برای این نماد خاص نشان می‌دهد.مدل‌های سنتی یادگیری ماشین مانند رگرسیون خطی، جنگل تصادفی، تقویت گرادیان، CatBoost و LightGBM عموماً عملکرد قوی نشان می‌دهند، که اغلب در اطراف یا زیر خط میانگین RMSE قرار دارند. روش یادگیری عمیق RNN به عنوان یک رویکرد شاخص، میانگین مربعات خطا بسیار پایینی را برای سهام‌هایی مانند

کحافظ و وتجارت ارائه می‌دهد. مدل‌های LSTM، CNN و DNN عملکرد متفاوتی را نشان می‌دهند، در برخی موارد عالی هستند در حالی که در برخی دیگر عملکرد ضعیفی دارند، که مستلزم بررسی دقیق کاربرد آنها بر اساس سهام به سهام است.

مدل‌ها به دو دسته روش‌های یادگیری ماشین سنتی و روش‌های یادگیری عمیق تقسیم‌بندی شده‌اند، که برای شفافیت بیشتر، با رنگ‌های متمایز نشان داده شده‌اند:

- مدل‌های یادگیری ماشین سنتی: با رنگ آبی نشان داده شده‌اند، شامل XGBoost، رگرسیون خطی، SVM، KNN، جنگل تصادفی و گرادیان تقویتی هستند. این مدل‌ها معمولاً به دلیل اتکای به تکنیک‌های آماری و ساختارهای ساده‌تر، برای تفسیر و آموزش سریعتر، شناخته شده‌اند.

- مدل‌های یادگیری عمیق: با رنگ نارنجی نشان داده شده‌اند، شامل LSTM، CNN، DNN، RNN، CatBoost و LightGBM هستند. مدل‌های یادگیری عمیق اغلب شامل ساختارهای پیچیده‌تر هستند و برای آموزش به مجموعه داده‌های بزرگتری نیاز دارند، که می‌تواند در برخی زمینه‌ها به بهبود عملکرد منجر شود، اما در صورت عدم مدیریت مناسب، احتمال بیش‌برازش را نیز افزایش می‌دهد.

مدل با کمترین RMSE با رنگ سبز برجسته شده است، که نشان می‌دهد عملکرد برتر آن در زمینه دقت پیش‌بینی است. در این مورد، رگرسیون خطی با کمترین مقدار میانگین RMSE تقریباً ۸۱.۰، نشان می‌دهد که در میان مدل‌های مورد آزمایش برای این وظیفه خاص، موثرترین مدل است. این یافته به ویژه قابل توجه است، زیرا رگرسیون خطی مدلی نسبتاً ساده است که می‌تواند از روابط خطی موجود در داده‌های قیمت سهام بهره‌مند شود. در مقابل، مدل با بیشترین RMSE با رنگ قرمز برجسته شده است، که نشان می‌دهد عملکرد نسبتاً ضعیف آن است. در اینجا، LSTM با بیشترین مقدار RMSE تقریباً ۱۴۸.۰، نشان می‌دهد که نسبت به سایر مدل‌ها برای پیش‌بینی قیمت سهام کمتر موثر است. این می‌تواند نشان دهد که فرضیات یا تکنیک‌های منظم‌سازی مدل ممکن است با ویژگی‌های داده‌های قیمت سهام همخوانی نداشته باشند.

خطوط عمودی نقطه‌چین، میانگین مقادیر RMSE را برای هر دو روش یادگیری ماشین سنتی (خط نقطه‌چین آبی) و روش‌های یادگیری عمیق (خط نقطه‌چین نارنجی) نشان می‌دهند. میانگین کلی RMSE با یک خط نقطه‌چین قرمز نشان داده شده است. این خطوط نقطه مرجعی برای ارزیابی عملکرد مدل‌های فردی در مقابل عملکرد متوسط دسته‌های مربوطه

ارائه می‌کنند. میانگین RMSE برای روش‌های یادگیری ماشین سنتی به طور قابل توجهی کمتر از روش‌های یادگیری عمیق است، که نشان می‌دهد به طور متوسط ، مدل‌های سنتی برای این مجموعه داده‌ها موثرتر هستند. نمودار به وضوح تفاوت‌های عملکرد بین روش‌های یادگیری ماشین سنتی و یادگیری عمیق را نشان می‌دهد. مدل‌های سنتی به طور کلی مقادیر RMSE کمتری را نشان می‌دهند، که نشان می‌دهد دقت پیش‌بینی بهتر نسبت به مدل‌های یادگیری عمیق در این تحلیل خاص دارند. این یافته نشان می‌دهد که برای مجموعه داده‌ای که استفاده شده است، روش‌های یادگیری ماشین سنتی ممکن است برای وظایف پیش‌بینی قیمت سهام مناسب‌تر باشند. نتایج همچنین بر اهمیت انتخاب مدل در تحلیل‌های پیش‌بینی تاکید می‌کنند. در حالی که مدل‌های یادگیری عمیق به دلیل توانایی آنها در گرفتن الگوهای پیچیده در مجموعه داده‌های بزرگ محبوبیت پیدا کرده‌اند، این تحلیل نشان می‌دهد که مدل‌های ساده‌تر هنوز هم می‌توانند در سناریوهای خاص، به ویژه زمانی که مجموعه داده به اندازه کافی بزرگ یا پیچیده نیست، عملکرد بهتری داشته باشند. علاوه بر این، وجود مقادیر پرت، همانطور که توسط RMSE بالای Elastic Net نشان داده شده است، نشان می‌دهد که پیش‌پردازش و انتخاب ویژگی دقیق، برای بهبود عملکرد مدل بسیار مهم است. ممکن است برای شناسایی ناهنجاری‌های بالقوه یا برای اصلاح مجموعه ویژگی مورد استفاده در مدل‌ها، بررسی بیشتر داده‌ها مفید باشد.

همانطور که بیان شد تحقیق حاضر یک الگوی جامع را برای سرمایه گذاران بازار سرمایه ارئه می‌دهد. سرمایه گذاران بازار سرمایه عموماً با این سؤال مواجه هستند که از چه سهمی و به چه مقداری باید سرمایه گذاری کنند. تا به حال در تحقیق به سعی بر این بود که به سؤال اول پاسخ دهد. پاسخ به سؤال دوم بر می‌گردد به موضوع بهینه سازی سبد سرمایه که با استفاده از یک رویکرد تحقیق در عملیاتی، بهینه‌ترین حالت ممکن را ارئه می‌دهد. بهینه سازی سبد سهام با استفاده از مدل آقای مارکوییتز و به طور خاص با بکارگیری روش بهینه سازی میانگین-واریانس (MVO) انجام شد. با توجه به خطای کمتر روش رگرسیون خطی، این تحلیل بر اساس داده‌های این روش برای نه سهم در طول ۴۰ روز انجام شده است. نتایج حاصل از مدل در جدول۲ آمده است.

جدول۲: نتایج حاصل از مدل میانگین_ واریانس

نماد	درصد
بپاس	۳۷.۳۳٪
فملی	۴۴.۲۷٪
فولاد	۰٪
غنوش	۰٪
کحافظ	۰٪
خودرو	۰٪
شپنا	۱۸.۴۰٪
وساخت	۰٪
وتجارت	۰٪

این درصد ها نشان می‌دهند که چه مقدار از کل سبد سرمایه گذاری باید به چه سهامی اختصاص یابد. بنابر نتایج بدست آمده از ۱۰۰ درصد کل مبلغ در نظر گرفته شده برای سرمایه گذاری باید ۳۷.۳۳ درصد از نماد بپاس، ۴۴.۲۷ از نماد فملی، ۱۸.۴ درصد از نماد شپنا به هر سهم اختصاص داده شود تا به بهترین ریسک و بازده در سبد سرمایه گذاری حاصل شود. قابل توجه است که این تحلیل نشان می‌دهد که باید سهم قابل توجهی به فملی(۴۴.۲۷٪) و سپس به بپاس (۳۷.۳۳٪) و در انتها به شپنا (۱۸.۴۰٪) اختصاص داده شود، در حالی که سایر نمادها در پرتفوی بهینه هیچ سهمی دریافت نکرده‌اند. این امر نشان می‌دهد که انتظار می‌رود این دو سهم بر اساس داده‌های تاریخی مورد تحلیل، بهترین بازده و کمترین ریسک را ارائه دهند. بازده پیش‌بینی شده پرتفوی ۲.۶۹٪ و ریسک پیش‌بینی شده پرتفوی (انحراف استاندارد) ۹.۱۸٪ محاسبه شده است. بازده پیش‌بینی شده، میانگین بازده پیش‌بینی شده از پرتفوی است که بر اساس عملکرد تاریخی سهام انتخاب شده محاسبه می‌شود. ریسک پیش‌بینی شده، نشان دهنده نوسان بازده پرتفوی است، و مقدار بالاتر آن نشان‌دهنده عدم اطمینان بیشتر

در دستیابی به بازده پیش‌بینی شده است. نتایج این بهینه‌سازی، نشان‌دهنده یک استراتژی سرمایه‌گذاری متمرکز است که عمدتا بر روی فملی و شپنا تمرکز دارد. این رویکرد با اصول نظریه مارکوییتز هم‌راستا است که به انتخاب دارایی‌هایی که حداکثر بازده را برای سطح مشخصی از ریسک فراهم می‌کنند، توصیه می‌کند. تحقیقات آینده می‌توانند به بررسی اثرات این یافته‌ها در شرایط مختلف بازار یا با بازه‌های زمانی مختلف بپردازند.

۳. نتیجه‌گیری

در این پژوهش، یک روش نوین مبتنی بر داده برای بهینه‌سازی سبد سهام با استفاده از الگوریتم‌های یادگیری ماشین و یادگیری عمیق ارائه شد. با توجه به پیچیدگی و نوسانات بازارهای مالی، این تحقیق به بررسی و مقایسه عملکرد مدل‌های مختلف در پیش‌بینی قیمت سهام و بهینه‌سازی سبد سرمایه‌گذاری پرداخت. نتایج به‌دست‌آمده نشان داد که مدل‌های یادگیری ماشین سنتی، به ویژه رگرسیون خطی، در پیش‌بینی قیمت‌های بازار بورس ایران عملکرد بهتری نسبت به مدل‌های یادگیری عمیق دارند. این یافته‌ها بر اهمیت انتخاب مدل مناسب در مدیریت پرتفوی تأکید می‌کند، به‌ویژه زمانی که دقت پیش‌بینی و قابلیت تفسیر داده‌ها از اهمیت بالایی برخوردار است.

با توجه به اینکه اکثر پژوهش‌های انجام شده بر روی بازارهای جهانی انجام شده‌اند، محدودیت و تحقیق نبود اطلاعات کافی جهت مقایسه خروجی‌ها با دیگر پژوهش‌ها می‌باشد. مسأله دیگر بالا بودن تنوع نمادها و روش‌های در نظر گرفته شده برای پیش‌بینی بود که به خصوص مصورسازی داده‌ها و نتایج را دشوار می‌کرد. در این تحقیق سعی شد تا الگوریتم‌های مورد توجه‌تر و معروف‌تر را برای پیش‌بینی استفاده کند، از این رو توصیه می‌شود تا در ادامه ارائه تصمیمی جامع‌تر به فعالین بازار سرمایه، از الگوریتم‌های دیگر پیش‌بینی استفاده شود. همچنین به مدل روش تحقیق در عملیاتی میانگین-واریانس، می‌توان محدودیت‌هایی مانند محدودت در حداکثر میزان تخصیص به یک سهم را اعمال کرد تا تنوع نمادها در سبد سرمایه بیش‌تر شود تا ریسک احتمالی کاهش یابد. همچنین برای نزدیک‌تر شدن به واقعیت، می‌توان از تکنیک‌های مواجهه با عدم قطعیت در تحقیق‌های آتی استفاده نمود تا ریسک‌های احتمالی در مدلسازی ریاضی لحاظ گردند.

تحلیل‌های انجام‌شده نشان داد که رگرسیون خطی به عنوان مؤثرترین مدل در این مطالعه ظاهر شد و توانست پیش‌بینی‌های دقیقی ارائه دهد. در مقابل، مدل‌های یادگیری عمیق مانند LSTM وCNN ، در حالی که توانایی درک الگوهای پیچیده را

دارند، نتوانستند به نتایج بهتری نسبت به مدل‌های سنتی دست یابند. این امر نشان‌دهنده این است که در شرایط خاص بورس ایران، به‌ویژه با داده‌های محدود، مدل‌های ساده‌تر می‌توانند عملکرد بهتری داشته باشند. از سوی دیگر وجود تورم دایمی در بازار و رشد قیمت ها در نگاه بلند مدت می تواند دلیلی دیگر بر این موضوع باشد. بهینه‌سازی سبد سهام با استفاده از روش میانگین-واریانس نشان داد که تخصیص سرمایه به سهام‌های خاص، به‌ویژه فملی و بپاس، می‌تواند به بهترین بازده و کمترین ریسک منجر شود. این نتایج می‌تواند به مدیران پرتفوی و تحلیلگران ریسک کمک کند تا تصمیمات بهتری در زمینه سرمایه‌گذاری اتخاذ کنند. در نهایت، این تحقیق به عنوان یک ابزار عملی برای بهبود مدیریت ریسک و عملکرد سبد سرمایه‌گذاری در بازار بورس ایران مطرح می‌شود و می‌تواند زمینه‌ساز تحقیقات آینده در این حوزه باشد. بررسی اثرات این یافته‌ها در شرایط مختلف بازار و با استفاده از داده‌های جدید می‌تواند به غنای بیشتر این حوزه کمک کند.

مراجع

[1] Bengio, Y. 2009. Learning deep architectures for AI. Foundations and trends® in Machine Learning, 2, 1-127.

[2] Breiman, L. 2001. Random forests. Machine learning, 45, 5-32.

[3] Chai, T. & Draxler, R. R. 2014. Root mean square error (RMSE) or mean absolute error (MAE). Geoscientific model development discussions, 7, 1525-1534.

[4] Chen, G., Chen, S., Fang, Y. & Wang, S. 2006. A possibilistic mean VaR model for portfolio selection. Advanced modeling and optimization, 8, 99-107.

[5] 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, 785-794.

[6] Cover, T. & Hart, P. 1967. Nearest neighbor pattern classification. IEEE transactions on information theory, 13, 21-27.

[7] V., Ershov, V. & Gulin, A. 2018. CatBoost: gradient boosting with categorical features support. arXiv preprint arXiv:1810.11363.

[8] Elman, J. L. 1990. Finding structure in time. Cognitive science, 14, 179-211.

[9] Fischer, T. & Krauss, C. 2018. Deep learning with long short-term memory networks for financial market predictions. European journal of operational research, 270, 654-669.

[10] Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

[11] Goodfellow, I. 2016. Deep Learning, MIT Press.

[12] Hochreiter, S. & Schmidhuber, J. 1997. Long short-term memory. Neural computation, 9, 1735-1780.

[13] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. & Liu, T.-Y. 2017. LightGBM: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.

[14] Kumbure, M. M., Lohrmann, C., Luukka, P. & Porras, J. 2022. Machine learning techniques and data for stock market forecasting: A literature review. Expert Systems with Applications, 197, 116659.

[15] Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86, 2278-2324.

[16] Pawaskar, S. 2022. Stock price prediction using machine learning algorithms. International Journal for Research in Applied Science & Engineering Technology (IJRASET), 10.

[17] Sheng, Z., Benshan, S. & Zhongping, W. 2012. Analysis of mean-VaR model for financial risk control. Systems Engineering Procedia, 4, 40-45.

[18] Srivinay, Manujakshi, B. C., Kabadi, M. G. & Naik, N. 2022. A hybrid stock price prediction model based on PRE and deep neural network. Data, 7, 51.

[19] Su, X., Yan, X. & Tsai, C. L. 2012. Linear regression. Wiley Interdisciplinary Reviews: Computational Statistics, 4, 275-294.

[20] Vapnik, V. 2013. The nature of statistical learning theory. Springer Science & Business Media.

[21] Zou, H. & Hastie, T. 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B: Statistical Methodology, 67, 301-320.