فصلنامه مهندسی مالی و مدیریت اوراق بهادار
شماره چهلم / پائیز ۱۳۹۸

آزمون عملکرد الگوریتم جنگل‌های تصادفی و الگوریتم شبکه عصبی عمیق در استراتژی آربیتراژ آماری

علیرضا فضل زاده[۱]
جعفر حقیقت[۲]
فرانک پورکیوان[۳]
وحید احمدیان[۴]

تاریخ دریافت مقاله : ۹۷/۰۸/۱۵ تاریخ پذیرش مقاله : ۹۸/۰۲/۲۸

چکیده

در این تحقیق به آنالیز اثر بخشی الگوریتم جنگل‌های تصادفی در زمینه آربیتراژ آماری پرداخته شده است، همچنین برای سنجش عملکرد الگوریتم جنگل‌های تصادفی در زمینه آربیتراژ آماری نسبت به دیگر مدل‌های ارائه شده در پژوهش‌های پیشین، مقایسه نتایج بدست آمده از کاربرد این الگوریتم با الگوریتم شبکه‌های عصبی عمیق انجام شده است. مدل‌های مورد نظر با اطلاعات مربوط به قیمت سهام آموزش داده شده و خروجی بدست آمده از این تکنیک، سهام را بر اساس موقعیت خرید و فروش طبقه‌بندی کرده است. با استفاده از این استراتژی موقعیت‌های سودآوری در بازار سهام برای کسب سود شناسایی می‌شود. نتایج نشان داد مدل جنگل‌های تصادفی دارای خطای طبقه‌بندی کمتری نسبت به مدل شبکه عصبی عمیق می باشد، بنابراین مدل جنگل‌های تصادفی روش مناسب‌تری برای استفاده در استراتژی آربیتراژ آماری و کسب سود می باشد.

کلمات کلیدی

جنگل‌های تصادفی، شبکه عصبی عمیق، آربیتراژ آماری

۱- گروه مدیریت و بازرگانی، دانشگاه تبریز، تبریز، ایران. (نویسنده مسئول) Fazlzadeh_Acc@yahoo.com
۲- گروه مدیریت و بازرگانی، دانشگاه تبریز، تبریز، ایران Jafarhaghighat@yahoo.com
۳- گروه مدیریت و بازرگانی، دانشگاه تبریز، تبریز، ایران. F.pourkeivan94@ms.tabrizu.ac.ir
٤- گروه حسابداری، دانشگاه تربیت مدرس، تهران، ایران. alcinbusiness3@gmail.com

مقدمه

بازارهای مالی بر پایه یک قانون تجارت کلی، به صورت خرید با قیمت کم و فروش با قیمت بالا هستند که هدف توسعه استراتژی هایی با ریسک پایین می باشد (دراکوس[۱]، ۲۰۱۶). بازار گردانان در بازارهای مالی با خرید سهام قیمت پایین و فروش با قیمت بالا در بازه های زمانی کوتاه، تولید سود می‌کنند، این فرایند به طور طبیعی رخ می دهد. پرتفوی های بازار می توانند شبیه رفتار سازندگان بازار ساخته شوند (فرنهولز و همکاران[۲]، ۲۰۰۷).

سرمایه گذاری افراد جامعه در بورس میتواند نقش بسیار زیادی در رونق اقتصادی داشته باشد، چون با این کار منابع مالی موردنیاز بنگاه های اقتصادی برای توسعه ی فعالیت ها تامین شده و درنتیجه رونق کسب وکار، اشتغال زایی و رشد و شکوفایی اقتصاد کشور را به دنبال دارد. یکی از اهداف اصلی هر کسی در سرمایه‌گذاری، کسب سود و درآمد است .اگر سرمایه گذاری در بورس به صورت دقیق و آگاهانه باشد از دو روش این هدف تامین میشود :روش اول از طریق دریافت سود نقدی سهام و روش دوم از طریق افزایش قیمت سهام می باشد. اصولاً موفقیت بورس اوراق بهادار و جذابیت آن برای سرمایه گذاران بالقوه از طریق افزایش بازدهی و قیمت سهام شرکت‌های پذیرفته شده در بورس اوراق بهادار امکان‌پذیر می‌شود. (فلاحپور و همکاران،۱۳۹۲) شناسایی موقعیت خریدوفروش سهام و تصمیم‌گیری به موقع برای سرمایه‌گذاران برای کسب سود امری بسیار ضروری است. برای این منظور استراتژی‌های بسیاری توسط افراد پیشنهاد شده‌است. استراتژی آربیتراژ آماری نوعی راهبرد معامله در بازار سرمایه است که روابط قیمتی تعادلی بین سهام را به منظور کسب سود شناسایی می‌کند. ترکیب این استراتژی با آخرین متدهای ماشین یادگیری، معیاری جدید برای جهت دهی تصمیمات سرمایه‌گذاران است.

پیشینه نظری پژوهش

پیشگامان راهبردهای آربیتراژ آماری، معاملات جفت بوده است. معاملات جفتی برای اولین بار در سال۱۹۸۰ توسط یک تیم متشکل از چند فیزیکدان، متخصص کامپیوتر و ریاضیدان که هیچ‌گونه پیش‌زمینه‌ای درزمینه‌ی مالی نداشتند مطرح شد. ایده آن‌ها یافتن قواعدی آماری برای اجرای معاملات آربیتراژی و بهبود مهارت معامله‌گری بود. استراتژی معاملات جفتی توسط گاتو[۳] و همکارانش در سال ۱۹۸۸ در بازار سهام به دلیل نقد شوندگی بالای آن موردبررسی و آزمایش قرار گرفت. آن‌ها این استراتژی را بر روی داده‌های روزانه وال‌استریت در بازه سال‌های ۱۹۶۲ الی ۱۹۹۷ به‌کار گرفتند. بنابر نتایج تحقیقات آن‌ها، افزایش استفاده از استراتژی معاملات جفتی ممکن است باعث بروز مشکلاتی شود زیرا که فرصت‌های معامله با توجه به اشراف بسیاری از آربیتراژگران به این استراتژی و درنتیجه ورود زودهنگام

آن‌ها بلافاصله پس از اندکی انحراف از تعادل کمتر خواهد شد. بخشی از این تجارت متعارف است و در مراجع چندگانه ارائه‌شده است. تجارت جفت استراتژی است که بر اساس قیمت نسبی سهام است و درواقع به ارزش واقعی آن‌ها علاقه‌مند نیست و قیمت نسبی بر اساس این ایده که دو دارایی باهمان ویژگی‌ها می‌تواند در مورد قیمت مشابه به‌عنوان مبنای قانون یک قیمت‌گذاری شود، موردبررسی قرار می‌گیرد. هنگامی‌که در یک زمان معین قیمت سهام متفاوت است، باارزش واقعی نسبت به قیمت واقعی، یک ارزش بیش‌ازحد و دیگری کمتر از حد است. (پول،۲۰۰۷)

الگوریتم جنگل‌های تصادفی را نخستین بار لئو برایمن[۴] و آدل کاتلر[۵] در سال ۲۰۰۱ ایجاد کرده و توسعه دادند. ایده جنگل‌های تصادفی برایمن از کار اولیه امیت و جمن (۱۹۹۷) در انتخاب ویژگی هندسی، از روش تصادفی فضای هو (۱۹۹۸) و روش انتخاب تقسیم به‌صورت تصادفی از دیتریچ (۲۰۰۰) گرفته شده است. روش جنگل‌های تصادفی پس از پیدایش به‌عنوان رقیب جدی برای روش‌هایی مانند boosting (فروند و شاپیر[۶]، ۱۹۹۶) و همچنین ماشین بردار پشتیبان، لحاظ گردیده شد. روش‌های ذکرشده با پیاده‌سازی سریع و آسان، دقت پیش‌بینی بسیار بالایی نیز دارند و می‌توانند تعداد بسیار زیادی از متغیرهای ورودی را بدون بیش برازش، اداره کنند. درواقع، آن‌ها یکی از دقیق‌ترین تکنیک‌های یادگیری عمومی در دسترس در نظر گرفته‌شده بودند. (جیمس و همکاران[۷]، ۲۰۱۳)

شبکه عصبی عمیق شبکه عصبی مصنوعی یک سامانه‌پردازشی داده‌ها است که از مغز انسان ایده گرفته و پردازش داده‌ها را به عهده پردازنده‌های کوچک و بسیار زیادی سپرده که به‌صورت شبکه‌ای به‌هم‌پیوسته و موازی با یکدیگر رفتار می‌کنند تا یک مسئله را حل نمایند. در این شبکه‌ها به کمک دانش برنامه‌نویسی، ساختار داده‌ای طراحی می‌شود که می‌تواند همانند نرون عمل کند، که به این ساختار داده نورون گفته می‌شود. بعد با ایجاد شبکه‌ای بین این نورون‌ها و اعمال یک الگوریتم آموزشی به آن، شبکه را آموزش می‌دهند.(شالکوف، ۱۳۸۸)

در این حافظه یا شبکه‌ی عصبی نورون‌ها دارای دو حالت فعال (روشن یا۱) و غیرفعال‌اند (خاموش یا ۰) و هر یال (سیناپس یا ارتباط بین گره‌ها) دارای یک وزن می‌باشد. یال‌های با وزن مثبت، موجب تحریک یا فعال کردن گره غیرفعال بعدی می‌شوند و یال‌های با وزن منفی، گره متصل بعدی را غیرفعال یا مهار (درصورتی‌که فعال بوده باشد) می‌کنند.

به‌طورکلی شبکه‌های عمیق از پشت سر هم قرار دادن شبکه‌های کم‌عمق (مانند شبکه عصبی تک لایه) تشکیل می‌شوند .درنتیجه شبکه‌های عمیق قادر به یادگیری و استخراج ویژگی‌های سلسه مراتبی غیرخطی می‌باشد و می‌تواند الگوهای آماری پیچیده را ذخیره کند. مدل‌های عمیق کارایی خود

را برای کاربردهای مختلف به اثبات رسانده‌اند. شبکه‌های عصبی عمیق دارای بیش از یک لایه مخفی می‌باشند. شبکه عصبی با تعداد زیادی لایه دارای پارامترهای زیادی است، که تشکیل یک مدل انعطاف‌پذیر را می‌دهد. این خاصیت، شبکه عصبی عمیق را قادر می‌سازد روابط پیچیده و غیرخطی بین ورودی و خروجی را مدل کند. یادگیری بهینه شبکه عصبی با تعداد زیادی لایه بسیار مشکل می‌باشد. روش کاهش شیب مقداردهی اولیه وزن‌ها به‌صورت تصادفی روش بهینه‌ای برای یافتن وزن‌های شبکه نیست مگر آنکه وزن‌های شبکه به‌صورت دقیقی مقداردهی اولیه شوند . بنابراین شبکه عصبی عمیق با مقادیر اولیه‌ی تصادفی برای وزن‌ها دقت خوبی بر روی‌داده‌های تست ندارد. (ذوقی و همایون پور، ۱۳۹۳)

اصل استفاده‌شده در یادگیری عمیق، آموزش لایه‌های خارجی با استفاده از یادگیری بدون نظارت است که در هر سطح انجام می‌شود.یک شبکه عصبی عمیق شامل یک‌لایه ورودی، چندین لایه پنهان و یک‌لایه خروجی، تشکیل توپولوژی شبکه است. لایه ورودی مطابق با فضای ویژگی است، بنابراین نورون‌های ورودی به‌عنوان پیش‌بینی کننده‌های بسیاری وجود دارد. لایه خروجی یا یک‌لایه طبقه‌بندی یا رگرسیون برای مطابقت با فضای خروجی است. در واحدهای پایه‌ای چنین مدلی، تمام لایه‌ها از نورون تشکیل‌شده‌اند. در معماری پیش‌فرض کلاسیک، هر نورون در لایه قبلی به‌طور کامل به تمام نورون‌ها در لایه بعدی L+۱ از طریق لبه‌های هدایت‌شده متصل می‌شود، که هر یک از آن‌ها یک وزن خاص است. همچنین، هر نورون در یک‌لایه غیر خروجی از شبکه دارای یک واحد تحت تأثیر است که به‌عنوان آستانه فعال‌سازی آن عمل می‌کند. به همین ترتیب، هر نورون، ترکیب وزن α از خروجی nL نورون در لایه قبلی L را به‌عنوان ورودی دریافت می‌کند.

$$\alpha=\sum_{i=1}^{n_i} w_i x_i + b \qquad \text{رابطه ۱)}$$

با w_i نشان‌دهنده وزن خروجی x_i و b تأثیرات است. ترکیبی وزنی α از فرمول (۱-۲) از طریق برخی از تابع فعال‌سازی f تبدیل می‌شود، به‌طوری‌که سیگنال خروجی $f(\alpha)$ به نورون‌ها در لایه L+1 منتقل می‌شود. (کراووس و همکاران[۸]، ۲۰۱۶)

در یادگیری ماشین ما از اطلاعات برچسب دار شده و بدون برچسب صحبت می‌کنیم، درصورتی‌که در ابتدا به مواردی اشاره می‌کنیم که در آن هدف یا نتیجه شناخته‌شده‌ای را که می‌توان نتیجه خروجی شبکه آن را داشته باشیم، بیان می‌کنیم. در مورد دوم ما چنین هدف خاصی نداریم. با الگوریتم‌های یادگیری عمیق، امکان آموزش شبکه‌های عصبی مصنوعی بدون برچسب یا مقادیر هدف وجود دارد.

پیشینه تجربی پژوهش

در زمینه کاربرد و بررسی آربیتراژ آماری در بورس ایران تحقیق عزیز احمدزاده و همکاران در سال ۱۳۹۳ موجود است، که به تشریح مفهوم و مصادیق آربیتراژ آماری و آزمون قابلیت کاربرد آن در بازار بورس اوراق بهادار تهران طی سال‌های (۱۳۹۱-۱۳۸۰) پرداختند. نتایج مطالعه نشان دهنده آن است که تمام راهبردهای معامله گشتاوری آزمون شده شرایط آربیتراژ آماری را تأمین نموده و همگی از مصادیق آربیتراژ آماری محسوب می‌شوند، در نتیجه می‌توان نتیجه گرفت که آربیتراژ آماری روشی مناسب برای کسب سود از این طریق دلیلی محکم بر ناکارایی بازار سرمایه ایران باشد.

قاسمی و فروش باستانی در سال ۱۳۹۱ یک رویکرد کنترل تصادفی برای مسئله‌ی معاملات جفتی مطرح می‌کند. تفاضل لگاریتم قیمت‌های جفت را با یک فرآیند تصادفی پایا مدل کرده وآنرا برای فرموله کردن بهینه سازی سبد استراتژی، براساس مسئله‌ی کنترل تصادفی بکار میبرد. یک استراتژی از رده‌ی آربیتراژ آماری، تحت صندوق سرمایه گذاری پوشش ریسک معاملات جفت می‌باشد.

کراووس و همکاران در سال ۲۰۱۶ در تحقیقی به ترکیب استراتژی آربیتراژ آماری با تکنیک یادگیری ماشین پرداخته‌اند. آنها برای پیش‌بینی قیمت سهام S&P500 از سه الگوریتم شبکه‌های عصبی عمیق، جنگل‌های تصادفی و درخت افزایش حجم استفاده کرده‌اند، همچنین آنها با ترکیب ۳ الگوریتم و طراحی الگوریتم جدید به یک روش با پیش‌بینی دقیق‌تر دست یافتند. آنها با پیش بینی قیمت سهام در بازه زمانی یک روز بعد و مقایسه آنها با متوسط سطح مقطع سهام به یک رتبه بندی برای سهام دست یافتند. در این رتبه بندی K، سهام با بالاترین K در موقعیت خرید و سهام با پایین ترین K در موقعیت فروش قرار می گیرند. با استفاده از این رتبه بندی به تشکیل پرتفوی هایی پرداختند و سود بدست آمده از هر کدام از روش‌های پیش بینی را با هم مقایسه کردند که الگوریتم ترکیبی سپس الگوریتم جنگل‌های تصادفی، دارای بیشترین سود بدست آمده بودند. در مراحل بعدی به مقایسه ریسک این روش با روش تجارت جفت پرداختند و سپس به تفکیک سهام بر اساس صنعت و مقایسه کاربرد این روش در صنایع مختلف پرداختند.

نیکولاس هاگ[۹] در سال ۲۰۰۹ در مقاله ای یک استراتژی آربیتراژ آماری بر اساس ترکیبی از شبکه‌های عصبی المن و ELECTRE III به صورت روش تصمیم‌گیری چند معیاره را توسعه داد. روش او شامل پیش‌بینی، رتبه‌بندی و تجارت می‌باشد.

کراس اخیراً در سال ۲۰۱۶ به بررسی بیش از ۹۰ استراتژی جفت تجارت آربیتراژ آماری با تمرکز بر قیمت گذاری‌های نسبی بین دو یا بیشتر اوراق بهادار پرداخته است. در این مقاله، تحقیقات در زمینه

استراتژی جفت تجارت را به ۵ طبقه تقسیم بندی کرده است. ترکیب استراتژی آربیتراژ آماری با مدل‌های یادگیری ماشین یکی از انواع طبقه بندی موجود می باشد.

سرجیو و همکاران[۱۰] در تحقیقی در سال ۲۰۱۵ یک استراتژی آربیتراژ آماری جدید را بر اساس مدل‌های عامل قیمت پویا معرفی کرده‌اند. هدف در این مقاله بهره‌برداری از خواص معکوس قیمت‌گذاری است. در این مقاله نتایج استراتژی آربیتراژ آماری بر مبنای مدل‌های بازده و قیمت را درسهام بورس S&P500 مقایسه می‌کند و نتایج نشانگر قدرت بیشتر مدل‌های مبتنی بر قیمت در پیش‌بینی می‌باشد.

دراکوس[۱۱] در سال ۲۰۱۶ در مقاله ای یک روش برای ایجاد آربیتراژ آماری در سهام شاخص S&P500 ارائه کرده است. یک دارایی مصنوعی بر اساس رابطه همبستگی سهام با شاخص ساخته شده است. استراتژی تجارت جفت در دوره های مختلف بین S&P500 و دارایی مصنوعی اجرا شده و نتایج مورد ارزیابی قرار گرفته است. معیارهای مختلف نشان داده اند که الگوریتم MultiVariate Kalman آربیتراژ آماری را درریسک پایین و سود بیشتر ایجاد می کند.

ستایش و کاظم نژاد و حلاج در پژوهشی در سال ۱۳۹۵ به پیش بینی بحران مالی شرکت‌های پذیرفته شده در بورس اوراق بهادار تهران با استفاده از طبقه‌بندی‌های غیرخطی جنگل‌های تصادفی پرداخته‌اند. شناسایی ۶۹ متغیر پیش‌بین اولیه از روش انتخاب متغیر ریلیف برای شناسایی متغیرهای پیش‌بین بهینه استفاده کرده‌اند. یافته‌های تجربی مربوط به بررسی ۹۵ شرکت-سال سالم (بدون درماندگی مالی) و ۹۵ شرکت-سال (درمانده‌ی مالی) بیانگر عملکرد بهتر جنگل‌های‌تصادفی نسبت به رگرسیون لجستیک است.

روش شناسی پژوهش

در پژوهش حاضر ۷۰٪ داده ها برای آموزش مدل‌ها انتخاب شدند و ۱۵٪ برای اعتباردهی و باقی‌مانده‌ی داد ها برای تست مدل تقسیم‌بندی شده‌اند. در مدل جنگل‌های تصادفی تعداد درخت‌های طبقه‌بندی جهت رشد در مدل جنگل‌های تصادفی را ۵۰۰ عدد انتخاب کردیم. به‌طورکلی در مدل جنگل‌های تصادفی هر چه تعداد درختان بیشتر باشد، نتایج بهتری به دست می‌آوریم، اما درنهایت این عملکرد به یک مقدار محدود می‌رسد زیرا پس از یک مقدار مشاهدها نشان می‌دهد که با افزایش تعداد درخت، مقدار نوسانات خطا ثابت است. تعداد ویژگی‌هایی که به‌صورت تصادفی برای هر تقسیم درخت اختصاص داده می‌شود را ۴ انتخاب‌شده است و این مقدار، برای شرایطی که با داده ی آفلاین کار می‌شود، مقدار پیش‌فرض برای شروع با مدل جنگل‌های تصادفی است. که با توجه به تعداد core های cpu تنظیم‌شده است. البته در پژوهش حاضر مقادیر ۱۶ و ۶۴ هم تست شد که با توجه به اینکه در مقدار

خطا تفاوتی ایجاد نشد، همان مقدار ۴ انتخاب‌شده است. در مدل شبکه عصبی عمیق برای راه‌اندازی این مدل در نرم‌افزار R از پکیج H2O استفاده شد. پس‌ازاینکه مجموعه داده ها از لایه ورودی وارد مدل شدند، وارد لایه پنهان اول می‌شوند. تعداد گره‌ها و تعداد لایه‌های پنهان در پکیج H2O توسط خود ماشین تعیین می‌شود. علت اینکه تعداد گره‌ها توسط ماشین مشخص می‌شود این است که در صورت دخالت ناظر مدل ممکن است دچار Loop بی‌نهایت شود. در مورد تعداد لایه پنهان هم تنها دخالت ناظر تعیین کردن حداکثر تعداد لایه‌ها است که این مقدار حداکثر، بر اساس امکانات سخت‌افزاری و محدودیت‌های زیربنایی مشخص می‌شود. لازم به ذکر است که تعداد لایه‌های پنهان و تعداد گره‌های هر لایه در این مدل مشخص نیست زیرا لایه های پنهان و گره‌های مدل شبکه عصبی به‌صورت جعبه سیاه می‌باشد. در این پژوهش تعداد حداکثر لایه‌ها مقدار ۱۰,۰۰۰ لایه در نظر گرفته شد. هر گره از لایه‌های پنهان یک ترکیب خطی از متغیرهای ورودی می باشند، به این معنی که هرکدام یک معادله‌ی رگرسیونی دارند و برای هر معادله رگرسیونی به تعداد متغیرهای ورودی یک ضریب محاسبه می‌شود. مدل کاربردی این پژوهش full connected می‌باشد. به این صورت که ناظر از بیرون تصمیم نمی‌گیریم که متغیر در چه تعداد از این گره ها وارد شود و ماشین این کار را به‌صورت خودکار انجام می‌دهد. البته ماشین هیچ‌کدام از متغیرها را برای هیچ‌کدام از گره‌ها حذف نمی‌کند، بلکه ضریب آن متغیر را به‌صورت صفر حدی مشخص می‌کند که قابل‌اندازه‌گیری نباشد. تعداد گره‌ها در لایه‌ها باهم برابر نیست اما معمولاً یک حالت صعودی در تعداد گره‌ها برای لایه‌ها وجود دارد، زیرا هدف اصلی کاهش خطا می‌باشد که برای این منظور خطاها در خط های مختلف ترکیبات خطی که از لایه‌های قبلی وجود دارند، تقسیم می‌شوند و در هرکدام از این ترکیبات خطی خطاها به یک سری که قابل‌شناسایی هستند تجزیه می‌شوند و رفته‌رفته خطاها کوچک‌تر می‌شوند. کوچک‌تر شدن خطاها یعنی متغیرهایی که وارد خط رگرسیونی کرده‌ایم، معنی‌دارتر می‌شوند. تعیین تعداد گره‌ها در لایه‌های پنهان توسط روش epoch انجام می‌شود. Epoch به‌صورت خودکار چند مرتبه داده‌های ورودی را وارد کرده خروجی را به دست می‌آورد تا حالت بهینه را به دست آورد و انتخاب کند. حالت بهینه از طریق logloss مشخص می‌شود به این صورت که زمانی که logloss حالت نزولی پیدا کرد، حالت بهینه به‌دست‌آمده است. در پژوهش حاضر آموزش مدل تا یک مرحله بعد از حالت نزولی شدن هم تکرار شد تا از حالت بهینه مطمئن شویم و ۴ بار آموزش مدل با تعداد گره‌های مختلف انجام شد. خروجی‌های لایه پنهان آخر که به‌صورت عدد پیوسته می‌باشد، وارد تابع فعال‌ساز می‌شود. در این پژوهش از تابع فعال‌ساز Maxout استفاده‌شده است. درواقع Maxout مشخص‌کننده‌ی حد آستانه در طبقه‌بندی می‌باشد. این حد آستانه در بازه‌ی ۰ و ۱ تغییر می‌کند، با

epock مورد آزمون‌وخطا قرار می‌گیرد و اعداد موردنظر آزمون می‌شوند تا حد آستانه با کمترین خطا مشخص شود . در پژوهش حاضر حد آستانه برای ورودی های مختلف متفاوت است و در بازه ۰,۴ تا ۰,۶ مشخص‌شده است. بالاتر از حد آستانه در دسته یک و پایین‌تر از آن در دسته‌بندی صفر قرارگرفته است.

جامعه آماری پژوهش حاضر، شامل ۵۰ شرکت برتر بورس اوراق بهادار تهران می‌باشد، که در بازه زمانی ۱۳۹۰ لغایت ۱۳۹۵ به‌صورت سه‌ماهه از طرف بورس اوراق بهادار اعلام می‌شوند و نمونه آماری، شرکت‌هایی انتخاب‌شده است که در همان بازه‌ی زمانی بیش از سه مرتبه از طرف سازمان بورس اوراق بهادار، جزو ۵۰ شرکت برتر بورس معرفی‌شده‌اند، می‌باشد.

اسامی این شرکت‌ها عبارتند از: اخابر، بترانس، پارسان، تایرا، حفاری، حکشتی، خبهمن، خپارس، خزامیا، خساپا، خکاوه، خودرو، رانفور، رمپنا، سفارش، شاراک، شبدیز، شبهرن، شپدیس، شپنا، شخارک، شفن، شیراز، شیران، فاذر، فاراک، فخاس، فخوز، فملی، فولاد، کچاد، کگل، واتی، وامید، وانصار، وبانک، وبشهر، وبملت، وپاسار، وتجارت، وساپا، وسینا، وصندوق، وغدیر، وکار، ولساپا، ومعادن، ونفت، ونوین، فاسمین

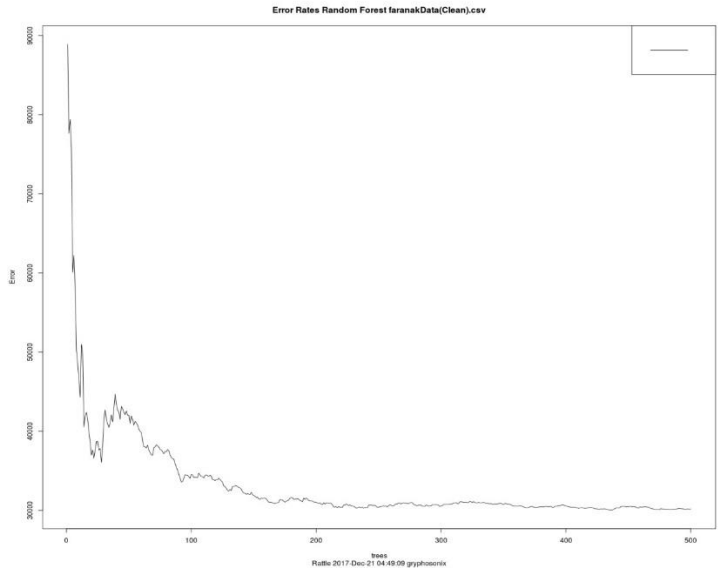
پژوهش به دنبال بررسی پاسخ این سوال اصلی است که آیا دسته‌بندی سهام بر اساس موقعیت خرید یا فروش برای استراتژی آربیتراژ آماری با الگوریتم جنگل‌های تصادفی دارای خطای کمتری نسبت به الگوریتم شبکه عصبی عمیق می‌باشد؟ با توجه به پیشینه تحقیق از روش‌های یادگیری ماشین برای پیش‌بینی قیمت سهام و درزمینه آربیتراژ آماری استفاده‌شده است و با در نظر گرفتن نتایج پژوهش‌های پیشین مبنی بر اینکه روش‌های یادگیری ماشین، عملکرد بالایی درزمینه ی پیش‌بینی و کسب سود در استراتژی آربیتراژ آماری دارند، می‌توان نتیجه گرفت که ترکیب روش‌های یادگیری ماشین و استراتژی آربیتراژ آماری روش مناسبی برای استفاده درزمینه‌ی خریدوفروش سهام جهت تشخیص موقعیت خریدوفروش می‌باشد. در میان روش‌های یادگیری ماشین جنگل‌های تصادفی که یک روش ترکیبی از درختان تصمیم می‌باشد، به‌عنوان یکی از دقیق‌ترین روش‌ها معرفی‌شده (کراووس و همکاران ۲۰۱۷) و همچنین روش شبکه عصبی عمیق که نوعی شبکه عصبی بسیار پارامتریزه شده است که اجازه‌ی انتزاع ویژگی‌ها ورودی را می‌دهد، امروزه درزمینه‌های مختلف به نتایج قابل‌توجهی دست‌یافته است. (تاکوچی و همکاران ۲۰۱۳). با توجه به عدم استفاده از استراتژی آربیتراژ آماری به‌صورت ترکیب‌شده با متدهای یادگیری ماشین و به‌منظور آزمون کارایی روش‌های جنگل‌های تصادفی و شبکه عصبی عمیق، فرضیه‌ی این پژوهش در رابطه با مقایسه خطای طبقه‌بندی سهام بر اساس موقعیت خریدوفروش برای استراتژی آربیتراژ آماری تبیین شد. مدل سازی توسط ۴۲ متغیر اندیکاتوری و ۲۰ متغیر بازده به عنوان متغیر

ورودی انجام گرفت. با جمع‌آوری داده‌ها از منابع، مجموعه داده تشکیل گردیده است و فضای ویژگی لازم برای متغیر پاسخ به این صورت تعریف‌شده است: یک پاسخ دوتایی به‌عنوان متغیر پاسخ به‌صورت (۰و۱) تعریف‌شده است که اگر R_{t+1} (بازده روز بعد) برای هر سهم S از میانه ی مقطعی بازده‌های کل سهم‌های موجود در نمونه آماری بیشتر باشد، متغیر پاسخ برابر یک و درغیراینصورت برابر صفر خواهد بود. خروجی یک به معنی موقعیت خرید و خروجی صفر به معنی موقعیت فروش سهم موردنظر می‌باشد. معیار مقایسه عملکرد مدل‌ها دقت طبقه بندی سهام است که با معیار AUC مقایسه شده‌اند.در رابطه با قلمرو زمانی پژوهش نیز، به دلیل اینکه برای آموزش مدل‌ها بـه‌داده‌های زیادی نیاز است، بازه زمانی از ابتدای سال ۱۳۹۰ تا پایان سال ۱۳۹۵ انتخاب شده است.

یافته های پژوهش

پس از مرحله‌ی پیش‌پردازش داده‌ها، در مدل جنگل‌های تصادفی ۷۰٪ داده‌های جمع‌آوری‌شده به‌عنوان داده‌های آموزش مدل و ۱۵٪ به‌عنوان داده‌ی اعتباردهی و باقی‌مانده داده‌ها به‌عنوان داده‌های تست مدل تقسیم‌بندی شده‌اند. در ادامه تعداد مناسب درخت و تعداد متغیر برای هرکدام از درخت‌ها با توجه به خطای محاسبه‌شده اندازه‌گیری شده است.

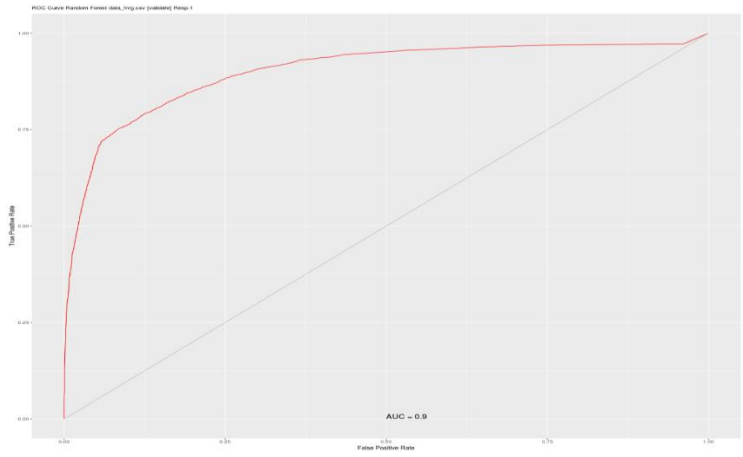
نمودار ۱- نمودار بررسی تعداد درخت در مدل جنگل‌های تصادفی



منبع: یافته‌های پژوهش

که محور x مقدار خطا و محور y تعداد درخت را نشان می‌دهد. با توجه به این نمودار مشاهده می‌شود که با افزایش تعداد درخت‌ها تا ۳۰۰ تقریباً از نوسانات مقدار خطا کاسته شده و مدل همگرا می‌شود، اما در این مدل برای اطمینان بیشتر، تعداد درخت‌ها تا ۵۰۰ افزایش داده شد و درنهایت تعداد درخت ۵۰۰ در نظر گرفته‌شده است. مقدار Mtry هم به‌صورت پیش‌فرض ۴ در نظر گرفته‌شده است.

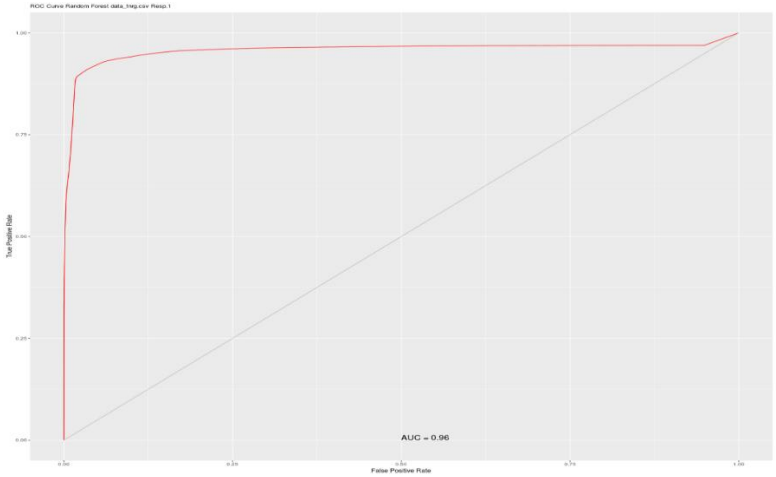
نمودار۲- ROC curve برای متغیرهای اندیکاتوری در مدل جنگل‌های تصادفی



منبع: یافته‌های پژوهش

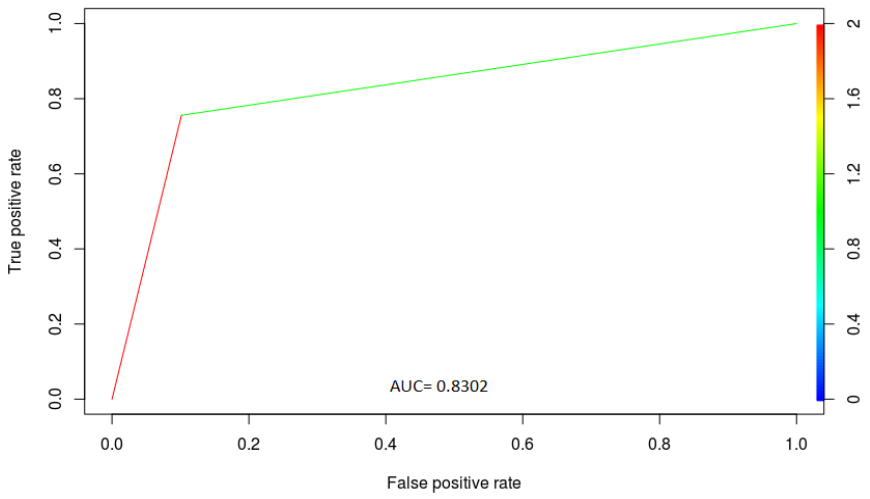
در این نمودار محور x نشان‌دهنده خروجی‌هایی است که به‌درستی طبقه‌بندی‌شده‌اند و نمودارy نشان‌دهنده خروجی‌هایی است که نادرست دسته‌بندی‌شده‌اند و در این نمودار این دو مقدار در مقابل هم رسم شده است. سطح زیر این نمودار مقدار AUC را نشان می‌دهد. که همان‌طور که مشاهده می‌شود AUC برابر ۰,۹ است.

نمودار۳- ROC curve برای تمامی داده‌ها در مدل جنگل‌های تصادفی



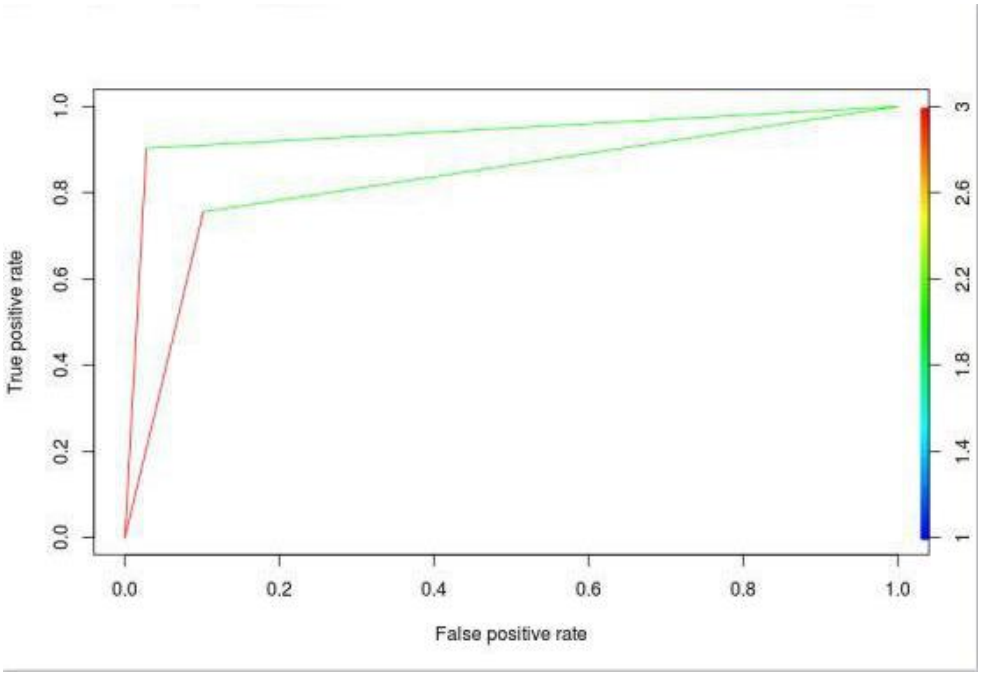
منبع: یافته‌های پژوهش

نمودار (۳) نشان‌دهنده ROC curve در صورت واردکردن تمامی متغیرها است. همان‌طور که مشاهده می‌شود مقدار AUC ۰,۰۴ افزایش‌یافته است. پس با مقایسه‌ی نمودارهای (۴-۴) و(۴-۳) می‌توان نتیجه گرفت که واردکردن متغیرهای بازده باعث بهبود مدل شده است. مقدار AUC ۰,۹۶ بیانگر این است که در طبقه‌بندی سهام، مدل جنگل‌های تصادفی دارای تنها ۴ درصد خطا است.



نمودار (۴) ROC curve برای مدل شبکه عصبی عمیق است. که سطح زیر این نمودار نشان‌دهنده مقدار AUC برای مدل شبکه عصبی عمیق است. مساحت زیر نمودار ۸۳,۰۲ محاسبه‌شده است که نشان‌دهنده‌ی دقت مدل به‌اندازه ۰,۸۳ است.

برای اینکه صحت و دقت طبقه بندی دو مدل مقایسه شود، از نمودار ROC[۱۲] استفاده شده است. نمودار ROC برای بررسی کارایی طبقه بندی ها به کار برده شده است. در این منحنی دوبعدی نرخ تشخیص صحیح دسته مثبت (True Positive Rate - TPR) روی محور Y و به‌طور مشابه نرخ تشخیص غلط دسته منفی (False Positive Rate - FPR) روی محور X رسم شده است. هرچه سطح زیر نمودار ROC بیشتر باشد نشان از بالاتر بودن دقت دسته بندی دارد. در واقع نرخ تشخیص دسته مثبت نسبت به نرخ تشخیص غلط دسته منفی بالاتر است.



شکل۱- نمودار مقایسه ROC مدل های شبکه عصبی عمیق و جنگل های تصادفی

شکل ۱ برای نشان دادن نمودار ROC دو مدل می‌باشد. همانطور که ملاحظه می‌شود، منحنی مدل شبکه عصبی عمیق پایین‌تر از مدل جنگل‌های تصادفی قرار گرفته است که همین باعث کمتر شدن سطح زیر نمودار شبکه عصبی نسبت به مدل جنگل‌های تصادفی می‌شود.

جدول۱- مقایسه AUC مدل جنگل های تصادفی و شبکه عصبی عمیق

مدل	AUC
جنگل‌های تصادفی	۰,۹۶
شبکه عصبی عمیق	۰,۸۳

جدول ۱ مقدار دقیق AUC هریک از مدل‌ها را نشان می‌دهد. اعداد قید شده در جدول از محاسبه سطح زیر نمودار ROC محاسبه شده‌اند. همانطور که ملاحظه می‌شود مقدار خطای مدل جنگل‌های تصادفی ۰,۴ محاسبه شده است که نشان می دهد مدل از ۱۰۰ دسته بندی فقط ۴ دسته منفی اشتباه انجام می‌دهد.

نتیجه‌گیری و پیشنهادها

تصمیم‌گیری برای اتخاذ موقعیت خرید یا فروش برای سهام امری بسیار مهم و حیاتی است و می‌تواند منجر به سودآوری بالایی گردد و نقش مهمی در اخذ تصمیمات فعالان بازار سهام دارد. هدف این پژوهش هم معرفی استراتژی سودآوری برای سرمایه‌گذاران است تا با استفاده از این متدها به موقعیت سودآور در بازار دست یابند. با توجه به یافته‌های تحقیق، مقایسه خطای طبقه بندی مدل‌های جنگل‌های تصادفی و شبکه عصبی عمیق نشان داد مدل جنگل‌های تصادفی طبقه بندی قوی‌تری نسبت به مدل شبکه عصبی عمیق انجام داده‌است. به طور کلی دقت طبقه‌بندی سهام در موقعیت خرید و فروش بر اساس استراتژی آربیتراژ آماری مدل جنگل‌های تصادفی سیزده درصد بالاتر بدست آمد و این مدل دارای نتایج بهتری نسبت به مدل شبکه عصبی عمیق در پژوهش حاضر بوده است و می تواند به عنوان بهترین مدل برای ترکیب با استراتژی آربیتراژ آماری مورد استفاده قرار گیرد. البته همان‌طور که گفته شد هردو مدل برای جامعه آماری پژوهش حاضر، با متغیرهای ورودی مورد آزمون قرار گرفت، که هردو دارای ضریب AUC بالای ۸۰٪ بوده است. با این وجود نتایج نشان می‌دهند که مدل جنگل‌های تصادفی روش مناسب‌تری برای استفاده در استراتژی آربیتراژ آماری می‌باشد. استراتژی آربیتراژ آماری که یک استراتژی برای کسب سود می‌باشد که می‌تواند برای هر دارایی که اجازه یا امکان خریدوفروش دارد، اعمال گردد که تمرکز ما در پژوهش حاضر روی بورس اوراق بهادار بوده است. با توجه به نتایج تحقیق مدل‌های ارائه شده می‌تواند جایگزین مناسب روش‌های فعلی موجود در نرم‌افزارهای پیش‌بینی قیمت آتی سهام در بورس اوراق بهادار در راستای بهبود عملکرد تصمیم‌گیری کارگزاران و سرمایه‌گذاران باشند و به عنوان سرور در وب سایت قرار گرفته و جهت راهنمایی سریع برای سرمایه‌گذاران مبنی بر اتخاذ موقعیت خرید یا فروش برای سهام خود مورد استفاده قرار گیرند.

فهرست منابع

۱) احمدزاده، عزیز و یاوری، کاظم و صالح‌آبادی، علی و عیسیایی تفرش، محمد. (۱۳۹۳). آزمون آربیتراژ آماری در بورس اوراق بهادار تهران، پژوهش‌ها و سیاست‌های اقتصادی، شماره ۷۰، صص۲۶۸-۲۴۷.

۲) افشاری، حسین. (۱۳۸۲). بررسی ساختاری قابلیت پیش‌بینی قیمت سهام در بورس اوراق بهادار تهران، بررسی‌های حسابداری و حسابرسی، سال دهم، شماره ۳۲، صص ۱۰۳-۱۲۶.

۳) الاکشمی پای، ویجی و سکاران، راجا. (۱۳۹۱). شبکه های عصبی، منطق فازی و الگوریتم ژنتیک: ترکیب و کاربرد، مترجم محمود کشاورزمهر، نشر نوپردازان.

۴) الوانی، سیدمهدی و حسین‌پور، داود. (۱۳۸۰). بررسی ساختاری قابلیت پیش‌بینی قیمت سهام در بورس اوراق بهادار تهران، بررسی‌های حسابداری و حسابرسی، سال دهم، شماره ۳۲، صص ۱۰۳-۱۲۶.

۵) امیری، مقصود و بکی حسکوئی، مرتضی و بیگلری کامی، مهدی. (۱۳۹۲). رتبه‌بندی شرکت‌های تولیدی در بورس اوراق بهادار تهران با استفاده از مدل‌های تصمیم‌گیری با معیارهای چندگانه و شبکه عصبی مصنوعی، فصلنامه علمی پژوهشی دانش سرمایه‌گذاری، دوره دوم، شماره هفتم، صص۷۳-۸۴ .

۶) پارساییان، علی و جهان خوانی، علی. (۱۳۷۵). بورس اوراق بهادار، چاپ دوم، تهران، انتشارات دانشگاه تهران.

۷) جهانخانی، علی و صفاریان، امیر. (۱۳۸۲). واکنش بازار سهام نسبت به اعلان سود برآوردی هر سهم در بورس اوراق بهادار تهران، تحقیقات مالی، شماره ۱۶، صص۶۱-۸۲.

۸) جونز، چارلزپی. (۱۳۸۴). مدیریت سرمایه‌گذاری، ترجمه تهرانی، رضا و نوربخش، عسگر، انتشارات نگاه دانش،چاپ دوم.

۹) نبوی چاشمی، سید علی و معماریان، عرفان و شعبانی ورنامی، محمد. (۱۳۹۲). انتخاب سبد سهام بهینه با استفاده از شبکه عصبی مصنوعی، اولین کنفرانس ملی حسابداری و مدیریت، شیراز، موسسه بین المللی آموزشی پژوهشی خوارزمی.

۱۰) نوروزی، حسین و اصغری مقدم، اصغر و ندیری، عطاالله. (۱۳۹۴). تعیین مناطق آسیب‌پذیر آبخوان دشت ملکان به نیترات با استفاده از روش جنگل تصادفی، فصلنامه محیط‌شناسی، دوره ۴۱، شماره ۴، صص ۹۲۳-۹۴۲.

۱۱) نوری جلیانی، محمد حسن و امجدی فرد، رؤیا. (۱۳۹۱). استفاده از جنگل‌های تصادفی جهت پیش‌بینی شاخص کل بازار بورس ایران، پایان‌نامه کارشناسی ارشد علوم کامپیوتر، دانشگاه تربیت معلم تهران.

۱۲) نوری، سحر و جلیانی،کرامت و محمد، کاظم. (۱۳۹۰). آنالیز جنگل‌های تصادفی: یک روش آماری مدرن برای غربالگری در مطالعات با بعد بالا و کاربرد آن در یک مطالعه همبستگی ژنتیکی جمعیت پایه، مجله دانشگاه علوم پزشکی خراسان شمالی، جلد ۳، شماره ۵، صص ۹۳-۱۰۱.

۱۳) هال، جان. مبانی مهندسی مالی و مدیریت ریسک. ترجمه: سجاد سیاح و علی صالح‌آبادی، (۱۳۸۴). تهران، انتشارات رایانه تدبیر پرداز، چاپ اول.

14) Bialkowski, J. P., Bohl, M. T., Kaufmann, P., & Wisniewski, T. P. (2011). "Do Mutual Fund Białkowski, J., Etebari, A., & Wisniewski, T. P. (2012). "Fast profits: Investor sentiment and stock returns during Ramadan", Journal of Banking & Finance,36(3), 835-845.

15) BREIMAN, LEO. (2001). "Random Forests", Machine Learning, 45, 5–32.

16) Cao, X. L., Premachandra, I. M., Bhabra, G. S., & Tang, Y. P. (2009). "Firm size and the pre-holiday effect in New Zealand", International Research Journal of Finance and Economics, 32, 171-187.

17) Cheng, J.H. Chen, H.P. & Lin, Y.M. (2010). "A hybrid forecast marketing timing model based on probabilistic neural network, rough set and C4.5", Expert Systems with Applications, 37, 1814–1820.

18) Drakos, S. (2016). " Statistical Arbitrage in S&P500", journal of mathematical Finance 6,166-177.

19) Fernholz, R. and Maguire, C. jr. (2007). "The statistics of statistical arbitrage" Financial analyst journal 63(5) 761-782.

20) Genuer, R, J and Poggi, M. and Malot, C. (2010). "variable selection using random forests", pattern recognition Letters, 31(14):p.2225-2236.

21) Huck, N. (2009). "Pairs selection and outranking: An application to the S&P 100 index", European journal of operational research, 196(2):819-825.

22) Huck, N. (2010). "Pairs trading and outranking: The multi-step-ahead forecasting case", European journal of operational research, 207(3):1702-1716..

23) Huck, N. (2015). "Pairs trading: does volatility timing matter", Applied Economics, 47(57):6239-6256..

24) Keivin, P, M. (2012). “Machine Learning: A Probabilistic Perspective”, the MIT press Cambridge.

25) Khaidem, L., Saha, S. & Dey, S. R. (2016). “Predicting the direction of stock market prices using random forest”, CoRR, abs/1605.00003.

26) Krauss, C. (2015). “Statistical arbitrage pairs trading strategies: Review and outlook IWQW” Discussion Paper Series, University of Erlangen-nurnberg.

27) Krauss, C. and Anh Do, x and Hck, N. (2016). “Deep neural networks, gradient-boosted trees, random forest: Statistical arbitrage on S&P 500”, European journal of operational research, 256(2):689-702.

28) Krauss, Christopher(2016), “Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500”, European Journal of Operational Research, No. 03

یادداشت ها :

۱ Drakos, S.
۲ Fernholz
۳ Gatev
٤ Leo Breiman
٥ Adele Cutler
٦ Freund and Schapire
۷ James
۸ Krauss
۹ Nicolas hock
۱۰ Sergio
۱۱ Drakos
۱۲ Receiver Operating Characteristic