پردیس علوم
دانشکده ریاضی، آمار و علوم کامپیوتر

پیش‌بینی ورشکستگی شرکت‌ها با استفاده از روش‌های داده‌کاوی

نگارنده:

سپیده خواجه حق وردی

استاد راهنما:

دکتر هدیه ساجدی

پایان‌نامه برای دریافت درجه کارشناسی
در رشته علوم کامپیوتر

بهمن ۱۳۹۶

چکیده

در طول قرن گذشته ورشکستگی شرکت‌ها موضوعی مورد علاقه برای پژوهشگران بوده است و هم اکنون نیز از مباحث مهم اقتصادی می‌باشد. در این پروژه با اعمال روش‌های داده کاوی بر روی یک مجموعه‌ی داده اقدام به پیدا کردن مدلی برای پیش‌بینی ورشکستگی کرده ایم. هدف پروژه، پیدا کردن روشی دقیق تر از روش‌های ارائه شده قبلی برای پیشبینی این پدیده است. برای نیل به این هدف، از هفت الگوریتم دسته بند شامل درخت تصمیم، جنگل تصادفی، ماشین بردار پشتیبان، استدلال مبتنی بر حافظه نزدیک‌ترین همسایه و الگوریتم‌های مبنتی بر نظریه بیز شامل گاوسی، برنولی و چندجمله‌ای استفاده شده است. مجموعه داده‌ی مورد استفاده شامل اطلاعاتی است که در طول پنج سال متوالی از بیش از ده هزار شرکت اخذ شده است. این مجموعه در برگیرنده‌ی شصت و چهار خصیصه است که از روی اطلاعات مالی و حسابداری این شرکت‌ها استخراج شده است. داده‌ها در قالب پنج ماتریس مجزا هستند. هر نمونه در این ماتریس‌ها نماینده یک شرکت است و دارای برچسب ورشکستگی یا عدم ورشکستگی می‌باشد. هدف الگوریتم دسته‌بند پیش‌بینی برچسب نمونه ها با استفاده از خصیصه‌های آن‌ها است. در روند پروژه کارایی روش‌های دسته‌بند نامبرده به تفصیل بررسی و با کارایی روش‌های ارائه شده در گذشته مقایسه شده است.

پیش گفتار

توسعه فناوری ها و گسترش صنایع موجب شده است که جوامع بیش از پیش به انرژی های فسیلی و همچنین برق محتاج شوند. شاید پیش‌بینی یک ورشکستگی اهمیت بسیار زیادی در تصمیم‌گیری‌های اقتصادی دارد. یک موقعیت تجاری چه برای یک مجموعه تجاری بزرگ و چه برای یک شرکت کوچک، نه تنها برای صنایع شریک، سرمایه گزاران و شبکه محلی دارای اهمیت است، بلکه بر روی سیاست‌های کلی اقتصادی نیز تاثیر گذار است. بنابراین هزینه‌های بالای اقتصادی و اجتماعی ناشی از ورشکستی شرکت‌ها، توجه محققان را جلب کرده است تا درک بهتری از علل ورشکستی و در نهایت ابزاری برای پیش‌بینی نابسامانی‌های تجاری پیدا کنند.[4][3] در قرن گذشته ورشکستگی شرکت‌ها موضوع جالب توجهی برای پژوهشگران حوزه‌های مالی و اقتصادی بوده است. در سال‌های اخیر مقاله‌های متعددی در حوزه‌ی پیش‌بینی ورشکستگی منتشر شده است که بخش عمده‌ای از آن‌ها از روش‌های شناخته شده دسته‌بندی استفاده کرده‌اند. در الگوریتم‌های دسته‌بند، کل مجموعه داده‌ها به دو قسمت مجموعه داده‌های آموزشی و مجموعه داده‌های آزمایشی تقسیم‌بندی می‌شوند. الگوریتم‌های دسته‌بندی در زمره الگوریتم‌های با ناظر[1] قرار می‌گیرند. چرا که هر رکورد دارای برچسبی مشخص است و هدف الگوریتم یادگیری، یافتن نظم حاکم بر انواع برچسب‌ها بر اساس سایر ویژگی‌های رکوردها می‌باشد. الگوریتم‌های دسته‌بندی شامل دو مرحله آموزش و آزمایش هستند. در مرحله آموزش الگوریتم یادگیرنده براساس مجموعه داده‌های آموزشی یک مدل می‌سازد. در مرحله آزمایش، بر اساس مجموعه داده‌های آزمایشی دقت و کارایی مدل ساخته شده ارزیابی می‌شود. داده‌هایی که در مرحله آزمایش مورد استفاده قرار می‌گیرند در مرحله آموزش و برای ساخت مدل استفاده نشده‌اند. معمولا بین ۲۰ تا ۳۰ درصد داده‌ها به صورت تصادفی برای آزمایش انتخاب شده و باقی داده‌ها برای آموزش و ساخت مدل استفاده می‌شود.[15]

روش معمول دیگری که برای تقسیم دادگان و ارزیابی مدل به کار می‌رود روش اعتبار سنجی متقابل با k حالت[2] است. در این روش، داده ها به K زیرمجموعه افراز می شوند. از این K زیرمجموعه، هر بار یک مجموعه برای آزمایش و $k-1$ مجموعه دیگر برای آموزش به کار می‌روند. این روال K بار تکرار می‌شود و همه داده‌ها دقیقا یک بار برای آموزش و یک بار برای آزمایش به کار می‌روند. در نهایت میانگین

[1] Supervised

[2] k-fold cross validation

نتیجه این K بار اعتبارسنجی به عنوان یک تخمین نهایی برگزیده می‌شود. البته می‌توان از روش های دیگر برای ترکیب نتایج استفاده کرد. مزیت مهم این روش این است که احتمال این که داده‌های قرار گرفته در مجموعه داده‌های آزمایشی بیش از حد آسان باشند و دقت اندازه گیری شده به طور غیر موثری زیاد شود را از بین می‌برد. همچنین احتمال اینکه مجموعه داده‌های آزمایش نسبت به مجموعه داده‌های آموزش سخت تر باشند و دقت به طور غیر موثری کاهش یابد نیز از بین می‌رود. به طور معمول از اعتبارسنجی متقابل با ۱۰ حالت استفاده می شود. در این پروژه برای تقسیم و ارزیابی مجموعه داده‌ها از هر دو روش مذکور استفاده کرده‌ایم. در روش اول داده‌ها به نسبت ۲۰ به ۸۰ به ترتیب به مجموعه داده‌های مورد استفاده برای ارزیابی و مجموعه داده‌های مورد استفاده برای ساخت مدل تقسیم کرده‌ایم. در روش دوم مجموعه داده‌ها با روش اعتبار سنجی متقابل با ۱۰ حالت تقسیم و ارزیابی شده‌اند.[1] مهم‌ترین معیار برای تعیین کارایی یک الگوریتم دسته‌بندی، معیار دقت یا نرخ دسته بندی[۳] است. این معیار دقت کل یک دسته بندی را محاسبه می‌نماید. معیار دقت یا نرخ دسته بندی نشان‌دهنده این حقیقت است که دسته‌بند طراحی شده چند درصد از کل مجموعه رکوردهای آزمایشی را به درستی دسته‌بندی کرده است. اما این معیار همیشه و برای مدل‌های ساخته شده از روی همه‌ی انواع مجموعه داده‌ها مناسب نمی‌باشد. نوعی از مجموعه داده‌ها وجود دارد که به آن مجموعه داده نامتعادل[۴] می‌گویند. در این مجموعه داده‌ها تعداد نمونه‌های مربوط به یک یا چند برچسب نسبت به باقی برچسب ها به طرز مشهودی کمتر است. برای مثال مجموعه داده‌ای که دارای دو برچسب مثبت و منفی هستند و تنها ۱۰درصد داده‌ها دارای برچسب منفی هستند، یک مجموعه داده نامتعادل است. مجموعه داده استفاده شده در این پروژه نیز از جمله مجموعه داده‌های نامتعادل است. چراکه به طور معمول نسبت تعداد شرکت‌هایی که سالانه ورشکست می‌شوند به کل شرکت‌های موجود بسیار کوچک است. لذا نسبت شرکت‌های ورشکسته به شرکت‌هایی که ورشکست نشده‌اند نیز کم است. از همین روی استفاده از معیار دقت به تنهایی برای ارزیابی کارایی مدل‌های معرفی شده کافی نیست و ما نیازمند معیاری هستیم که مستقل از تعداد برچسب‌ها در هر نوع عمل کند. مقاله‌ای که پیش‌تر بر روی مجموعه داده مورد استفاده ما کار کرده است از معیار AUC برای ارزیابی مدل‌ها استفاده کرده است. ما نیز برای اینکه مقایسه بین روش‌های ارائه شده در این مقاله و روش‌های آزمایش شده در پروژه حاضر ممکن شود همین معیار را به عنوان معیار اصلی برگزیده‌ایم. معیار AUC[۵] ۱ نشان دهنده سطح زیر نمودار ROC است. نمودار ROC روشی برای بررسی کارایی دسته‌بندها می‌باشد. هرچقدر عدد AUC مربوط به یک دسته‌بند بزرگ‌تر باشد کارایی نهایی دسته‌بند مطلوب‌تر ارزیابی می‌شود. منحنی‌های ROC رفتار یک دسته‌بند را بدون توجه به توزیع دسته یا هزینه خطا نشان می‌دهند، بنابراین کارایی دسته‌بندی را از این عوامل جدا میکنند، اما در حالی‌که یک منحنی ROC یک تکنیک با ارزش مصورسازی است، در انتخاب دسته‌بندها کمک کمی می‌کند. فقط زمانی که یک دسته‌بند در کل فضای کارایی به وضوح بر دسته‌بند دیگری تسلط یابد می‌توان گفت که بهتر است. به همین دلیل معیار AUC که سطح زیر نمودار ROC

[3] Classification Accuracy – Rate (CA-CR)

[4] Unbalance Dataset

[5] Area Under the Curve

را نشان می‌دهد می‌تواند نقش تعیین‌کننده‌ای را در معرفی دسته‌بند برتر ایفا کند. مقدار AUC برای یک دسته‌بند که به‌صورت تصادفی، برچسب نمونه مورد بررسی را تعیین می‌کند برابر با ۵/۰ است. همچنین بیشترین مقدار این معیار برابر یک بوده و برای وضعیتی رخ می‌دهد که دسته‌بند ایده‌آل بوده و بتواند کلیه نمونه های مثبت را بدون هرگونه هشدار غلطی تشخیص دهد. معیار AUC بر خلاف دیگر معیارهای تعیین کارایی دسته‌بندها مستقل از آستانه تصمیم گیری دسته‌بند است. بنابراین این معیار نشان‌دهنده میزان قابل اعتماد بودن خروجی یک دسته‌بند مشخص به ازای مجموعه داده‌های متفاوت است. در ادامه تمام هفت روش دسته‌بندی ارائه شده با هر دو معیار دقت و AUC بررسی می‌شوند و همچین هر دو این معیارها هم در قالب تقسیم بندی دادگان به شکل ساده و با نسبت ۸۰ به ۲۰ و هم در قالب اعتبار سنجی متقابل با ۱۰ حالت اعمال می‌شوند. و در نهایت نتایج بدست آمده همگی با هم و با روش های قبلی مقاسیه خواهد شد.

فهرست مطالب

فصل ۱

مجموعه داده ها

۱.۱ شرح مجموعه داده

مجموعه داده ها درباره پیش بینی ورشکستگی شرکت های لهستانی است. داده ها مربوط به بازارهای نوظهور خدمات اطلاعات است که از یک پایگاه داده، حاوی اطلاعات بازارهای در حال ظهور در سراسر جهان، جمع آوری شده است. شرکت های ورشکسته در دوره زمانی بین سال ۲۰۰۰ تا ۲۰۱۲ مورد تجزیه و تحلیل قرار گرفتند و شرکت هایی که ورشکست نشده اند از سال ۲۰۰۷ تا ۲۰۱۳ مورد ارزیابی قرار گرفتند. داده های جمع آوری شده، بر اساس بازه های پیش بینی به پنج دسته تقسیم شده اند؛ اولین سال: داده ها شامل نرخ های مالی از سال اول دوره پیش بینی و دارای برچسب کلاس مربوطه اند که نشان دهنده وضعیت ورشکستگی پس از ۵ سال می باشد. این اطلاعات شامل ۷۰۲۷ نمونه (صورتهای مالی) است که ۲۷۱ شرکت ورشکسته شده و ۶۷۵۶ شرکت در دوره پیش بینی ورشکست نشده اند. دومین سال: داده ها شامل نرخ های مالی از سال دوم دوره پیش بینی و دارای برچسب کلاس مربوطه اند که نشان دهنده وضعیت ورشکستگی پس از ۴ سال است. داده ها شامل ۱۰۱۷۳ نمونه (صورت حساب های مالی) است که ۴۰۰ شرکت ورشکسته شده و ۹۷۷۳ شرکت در دوره پیش بینی ورشکست نشده اند. سومین سال: داده ها حاوی مقادیر مالی از سال سوم دوره پیش بینی و برچسب کلاس مربوطه اند که نشان دهنده وضعیت ورشکستگی پس از ۳ سال است. داده ها حاوی ۱۰۵۰۳ مورد (صورت حساب های مالی) است که ۴۹۵ شرکت ورشکسته شده و ۱۰۰۰۸ شرکت در دوره پیش بینی ورشکست نشده اند. چهارمین سال: داده ها شامل نرخ های مالی از سال چهارم دوره پیش بینی و برچسب کلاس مربوطه اند که نشان دهنده وضعیت ورشکستگی پس از ۲ سال می باشد. داده ها حاوی ۹۷۹۲ نمونه (صورت حساب های مالی) است که ۵۱۵ شرکت بازنشسته شده اند و ۹۲۷۷ شرکت در دوره پیش بینی ورشکست نشده اند. پنجمین سال: داده ها شامل نرخ های مالی از سال پنجم دوره پیش بینی و برچسب کلاس مربوطه اند که نشان دهنده وضعیت ورشکستگی پس از ۱ سال می باشد. داده ها حاوی ۵۹۱۰ نمونه (صورت حساب های مالی) است که ۴۱۰ شرکت ورشکست شده و ۵۵۰۰ شرکت دوره پیشبینی ورشکست نشده اند.[2]

جدول ۱.۱: اطلاعات کلی مجموعه داده.

مشخصات مجموعه داده:	واقعی	تعداد نمونه ها:	۱۰۵۰۳
نوع مسئله:	دسته بندی	تعداد متغیرها:	۶۴
حوزه:	کسب و کار	داده جا افتاده؟	دارد

۲.۱ اطلاعات مجموعه داده

در جدول۱.۱ اطلاعات کلی مربوط به مجموعه داده آورده شده است.

۳.۱ معرفی مفاهیم استفاده شده در متغیرها

۱.۳.۱ کل دارایی

دارایی کل[1] مفهومی مالی است یعنی اموال و حقوقی که منافع آتی قابل تقویم به پول دارند و بر اثر معاملات، عملیات یا رویدادهای مشخص به مالکیت یا تسلط مالکانه یک واحد تجاری درآمده‌اند دارایی آن واحد محسوب می‌شوند. دارایی‌ها عرفاً از دیدگاه اراده در ترازنامه به گروه‌های متمایزی تقسیم می‌شوند .دو گروه که مورد استفاده بیشتری دارند عبارتند از دارایی‌های جاری و دارایی‌های ثابت.[7]

۲.۳.۱ دارایی جاری

دارایی جاری[2] یکی از دسته‌های اصلی حساب‌ها در ترازنامه است که نماینده ارزش کل همه‌ی دارایی‌هایی است که می‌توان انتظار داشت در مدت کمتر از یک سال تبدیل به وجه نقد می‌شوند. دارایی‌های جاری شامل پول نقد و سایر حساب‌های با نقدشوندگی سریع هستند. حساب‌های دریافتنی، موجودی کالا، سهام و سایر اوراق بهاداری که در بازار خرید و فروش می‌شوند جزو دارایی‌های جاری به‌حساب می‌آیند. دارایی جاری به وجوه نقد و سایر دارایی‌هایی که انتظار می‌رود در جریان عملیات جاری واحد، ظرف یک سال تبدیل به وجه نقد گردند، فروخته شوند یا به مصرف رسند، گفته می‌شود. در یک واحد خدماتی کوچک دارایی‌های مذکور علاوه بر وجوه نقد، شامل حساب‌ها و اسناد دریافتنی، موجودی مواد و هزینه‌های پیش‌پرداخت شده است.

[1] Total assets
[2] Current assets

۳.۳.۱ دارایی ثابت

دارایی ثابت[3] به دارایی مشهودی که در واحد تجاری مورد استفاده قرار می‌گیرد و دارای ماهیتی دائمی یا نسبتاً ثابت است، گفته می‌شود. دارایی‌های مذکور، به استثنای زمین به مرور فرسوده یا در اثر گذشت زمان غیر قابل استفاده می‌شوند. در این صورت اصطلاحاً گفته می‌شود که دارایی‌ها مستهلک شده‌اند.

۴.۳.۱ سود خالص

در تجارت و کسب‌وکار، به آورده یا سود یک نهاد، در طول یک دوره حسابداری اطلاق می‌گردد، که از طریق کسر نمودن هزینه‌های عملیاتی و مالیات از درآمد آن شرکت، محاسبه شده باشد. به بیان ساده‌تر، اگر هزینه‌های عملیاتی و مالیات یک شرکت را در طول یک دوره زمانی مشخص؛ مثلاً یک سال مالی، از درآمدهای آن شرکت کسر نماییم، عدد بدست آمده، نشانگر سود خالص[4] آن شرکت می‌باشد. مثبت یا منفی بودن این عدد، نسبت مستقیم با عملکرد شرکت مورد نظر دارد. همچنین مثبت بودن سود خالص، به افزایش ارزش سهام و در نهایت افزایش حقوق صاحبان سهام می‌انجامد.

۵.۳.۱ سود ناخالص

سود ناخالص[5] به تفاوت میان درآمد و هزینه‌های ساخت محصول یا ارائه خدمات، قبل از کسر هزینه‌های؛ حقوق و دستمزد، مالیات، و بهره اطلاق می‌گردد. سود ناخالص با سود عملیاتی و سود قبل از بهره و مالیات متفاوت می‌باشد. به‌طورکلی درآمد شرکت منهای بهای تمام شده کالای فروش رفته، سود ناخالص گفته می‌شود؛ اما در مبحث حسابداری شخصی، سود ناخالص اشاره به درآمد شخصی کل فرد، قبل از حساب کردن مالیات و سایر کسورات دارد.

۶.۳.۱ بدهی کل

بدهی کل[6]، مجموع بدهی ها بدهی های کلان و تعهدات مالی است که در هر دوره زمانی خاص توسط یک کسب و کار به افراد و سازمان ها داده می شود. مجموع بدهی ها بر روی ترازنامه شرکت گزارش می شود و جزء معیارهای حسابداری عمومی است: دارایی = بدهی ها + سهام.

[3] Fixed assets
[4] Net Profit
[5] Gross Profit
[6] Total liabilities

۷.۳.۱ بدهی کوتاه مدت

بدهی کوتاه مدت[۷]، حسابی است که در بخش بدهی های فعلی ترازنامه شرکت قرار دارد. این حساب از قرض های یک سال گذشته ی شرکت به وجود می آید. این بدهی این حساب معمولا شامل وام های کوتاه مدت بانکی است که توسط شرکت گرفته شده است.

۸.۳.۱ بدهی بلند مدت

بدهی بلند مدت[۸] به بدهی‌ای که سررسیدش طولانی معمولاً بیش از یک سال باشد، بدهی بلندمدت یا بدهی ثابت گفته می‌شود . چنان‌چه سررسید بدهی‌های بلند مدت به یک سال یا کمتر کاهش یابد، به بدهی‌های جاری تبدیل می‌شوند.

۹.۳.۱ بدهی‌های جاری

بدهی جاری[۹] به بدهی‌ای که موعد پرداختش یک سال یا کمتر باشد و از محل دارایی‌های جاری قابل پرداخت باشد گفته می‌شود. از متداول‌ترین بدهی‌های جاری، اسناد پرداختنی و حساب‌های پرداختنی را می‌توان نام برد که دقیقاً عکس اسناد و حساب‌های دریافتنی هستند. سایر بدهی‌های جاری متداول عبارتند از حقوق و سایر هزینه‌های پرداختنی و مالیات بر درآمد واحد.

۱۰.۳.۱ سرمایه ثابت

سرمایه ثابت[۱۰] شامل پول هزینه شده در دارایی های ثابت، یعنی کارخانه، ماشین آلات، زمین و ساختمان ها، مواد خام و هزینه های عملیاتی جانبی (از جمله خدمات خارجی خریداری شده)، و برخی از ضایعات تولیدی (تصادفی هزینه ها) می‌باشد. سرمایه متغیر، در مقابل، به هزینه سرمایه گذاری بر هزینه های کاری تا آنجا که درآمد کارگران را نشان می دهد، اشاره دارد.

۱۱.۳.۱ سرمایه در گردش

سرمایه در گردش[۱۱] مجموعه مبالغی است، که در دارایی‌های جاری یک شرکت، سرمایه‌گذاری می‌شود. اگر بدهی‌های جاری از دارایی‌های جاری یک شرکت کسر گردد، سرمایه در گردش خالص به دست می‌آید.

[7] Short-term liabilities
[8] Long-term liabilities
[9] Current liabilities
[10] Constant capital
[11] Working capital

۱۲.۳.۱ سرمایه سهمی

سرمایه سهمی[۱۲]، به سرمایه‌ای که از راه فروش سهام عادی شرکت تامین می‌شود، اطلاق می‌گردد. سرمایه سهمی اشاره به بخشی ازحقوق صاحبان سهام یک شرکت دارد، که سهام‌داران آن، از طریق معاملات سهام، در قالب وجه نقد، بدست آورده‌اند.

۱۳.۳.۱ وجه نقد

موجودی نقدی[۱۳] نوعی دارایی جاری و عبارت است از پول موجود در واحد، موجودی نزد بانک‌ها اعم از سپرده یا حساب جاری، چک‌ها و حواله‌های بانکی و به طور کلی هر گونه اوراق و وسایل نقل و انتقال پول که مورد قبول بانک‌ها باشد.

۱۴.۳.۱ اوراق بهادار کوتاه مدت

اوراق بهادار کوتاه مدت[۱۴] یک ابزار مالی قابل داد و ستد و مثلی (عوض دار) ای است که دارای ارزش مالی می‌باشد.

۱۵.۳.۱ مطالبات

مطالبات[۱۵] یک تخصیص دارایی است که قابل اجرا برای تمام بدهی ها، معاملات غیرقابل حل و یا سایر وظایف پولی است که مشتریان یا بدهکاران ملزم به پرداخت به شرکت اند.

۱۶.۳.۱ هزینه عملیاتی

هزینه عملیاتی[۱۶]، هزینه‌های جاری یک شرکت به‌منظور راه‌اندازی یک محصول، کسب‌وکار یا سیستم می‌باشد، که در واقع هزینه‌های مربوط به فعالیت عادی و مستمر شرکت است. این هزینه‌ها شامل بهای تمام شده کالای فروش رفته، هزینه فروش، هزینه‌های عمومی و اداری و تحقیق و توسعه هستند. این هزینه‌ها از طریق فعالیت‌های عادی شرکت به وجود می‌آیند. یکی از اهداف شرکت‌ها حداکثر کردن بهره‌وری در مقابل میزان هزینه‌های عملیاتی است. در این مسیر هزینه عملیاتی معیار اصلی نشان‌دهنده بهره‌وری شرکت در طول زمان است.

۱۷.۳.۱ استهلاک یا کاهش بها

منظور از استهلاک[۱۷]، ارزش کاهش‌یافته و عمر کوتاه‌شده‌ی کالاهای سرمایه‌ای است که از فرسایش ناشی می‌شود. در معنای گسترده‌تر، ممکن است اشاره باشد به ارزش

[12] Share capital
[13] Cash
[14] Short-term securities
[15] Receivables
[16] Operating expenses
[17] Depreciation

کاهش‌یافته و عمر کوتاه‌شده‌ی هر کالای سرمایه‌ای یا دارایی که در طول دوره‌ی قابل‌توجهی از زمان، منشأ خدماتی قرار گرفته‌است. استهلاک یکی از مواد هزینه‌ای است که مؤسسات اقتصادی آن را به کار می‌برند. به زبان عوام، از تفاوت قیمت خرید و قیمت فروش کالا به دست می‌آید.

۱۸.۳.۱ سود انباشته

سود انباشته[۱۸]، به درصدی از سود یک شرکت سهامی اطلاق می‌شود، که به صورت سود سهام، به سهام‌داران پرداخت نشده است و شرکت مورد نظر، برای سرمایه‌گذاری مجدد در فعالیت‌های اصلی‌اش، یا پرداخت بدهی، آن را نگه داشته است. از اضافه کردن سود خالص یا کسر کردن زیان خالص از سود انباشته ابتدای دوره، سود سهام پرداختنی به سهام‌داران به دست می‌آید.

۱۹.۳.۱ سود قبل از بهره و مالیات

درآمد قبل از بهره و مالیات یا سود قبل از بهره و مالیات[۱۹] معیاری است از سود شرکت که هزینه‌های بهره ومالیات را مستثنی می‌کند. درآمد عملیاتی حاصل تفریق عایدیِ عملیاتی و هزینه‌ی عملیاتی است. گاهی اوقات برای شرکتی که درآمد غیرعملیاتی صفر دارد، درآمد عملیاتی هم‌معنا با درآمد قبل از بهره و مالیات($EBIT$) و سود عملیاتی به‌کار برده می‌شود.

۲۰.۳.۱ سود قبل از بهره، مالیات و استهلاک

سود قبل از بهره، مالیات و استهلاک [۲۰]از اساسا درآمد خالص با بهره، مالیات و استهلاک است که به آن اضافه شده است. $EBITDA$ را می توان برای تجزیه و تحلیل و مقایسه ی سودآوری بین شرکت ها و صنایع مورد استفاده قرار داد زیرا این تأثیرات تأمین مالی و تصمیمات حسابداری را از بین می برد.

۲۱.۳.۱ ارزش دفتری حقوق صاحبان سهام

ارزش دفتری[۲۱] از تقسیم حقوق صاحبان سهام در ترازنامه بر تعداد سهام به دست می‌آید و بر اساس آن در صورت انحلال شرکت و پس از پرداخت تمام بدهی‌های شرکت به ازای هر سهم نصیب سهام‌داران می‌شود.

[18] Retained earnings
[19] EBIT
[20] EBITDA
[21] Book value of equity

۲۲.۳.۱ فروش

فروش[۲۲] به مبلغ کل پولی که کسب و کار از فروش کالاها یا خدمات دریافت می‌کند گویند.

۲۳.۳.۱ حقوق صاحبان سهام

حقوق صاحبان سهام[۲۳] معرف علایق سهام‌داران و صاحبان اصلی شرکت، نسبت به خالص دارایی‌های شرکت است. حقوق صاحبان سهام، باقیمانده منافع مالکین شرکت را در داراییهای شرکت، که پس از کسر بدهیهای آن شرکت، بدست آمده است، نشان می‌دهد. در یک موسسه تجاری، حقوق صاحبان سهام در اصل منافع صاحبان اصلی موسسه را نشان می‌دهد. در حسابداری، حقوق صاحبان سهام، اهدافی چون تعیین سرمایه قانونی و ثبت شده، تعیین منابع سرمایه شرکت و تعیین سود سهامی که می‌تواند، بین صاحبان سهام توزیع شود، را دنبال می‌نماید. در یک شرکت سهامی، حقوق صاحبان سهام مواردی چون: سود و زیان انباشته، سهام سرمایه، کسر سهام و اندوخته‌های شرکت را شامل می‌شود.[5]

۲۴.۳.۱ اقلام فوق‌العاده

یک مورد فوق‌العاده[۲۴] شامل سود ها و زیان های مشمول در صورت درامد از رخدادهای غیر معمول و غیرطبیعی است. اقلام فوق العاده معمولا در یادداشت های صورتهای مالی توضیح داده می‌شوند، و نتیجه رویدادهای غیرقابل پیش بینی و غیر معمول هستند.

۲۵.۳.۱ هزینه‌های مالی

هزینه‌های مالی[۲۵] شامل بهره، مالیات بر درآمد و سایر هزینه هایی است که در مالکیت یا اجاره یک دارایی یا ملک رخ داده است.

۲۶.۳.۱ بهره

پولی که وام‌گیرنده از بابت استفاده کردن از پول‌های وام‌دهنده به او پرداخت می‌کند را بهره[۲۶] می‌گویند. به بیان دیگر، "بهره" بهای پول وام گرفته شده است.

۲۷.۳.۱ موجودی کالا

موجودی کالا[۲۷]، شامل مواد خام و محصولاتی می‌باشد که در انبار نگهداری می‌شود.

[22] Sales
[23] Equity
[24] Extraordinary item
[25] Financial expenses
[26] Interest
[27] Inventory

۲۸.۳.۱ فعالیت های عملیاتی

فعالیت های عملیاتی[۲۸] به فعالیت های اصلی شرکت، مانند تولید، توزیع، بازاریابی و فروش محصول یا خدمات میگویند. فعالیت های اصلی باید به طور کلی اکثریت جریان نقدی شرکت را فراهم کنند و تا حد زیادی تعیین کنند که آیا یک شرکت سودآور است یا خیر.

۲۹.۳.۱ سود فعالیت های عملیاتی

سود ناشی از فعالیت‌های عملیاتی[۲۹].

۳۰.۳.۱ هزینه فروش محصولات

هزینه فروش محصولات[۳۰]، هزینه ای است که یک شرکت برای فروش یک محصول متحمل شده است. این شامل هزینه مواد اولیه و بسته بندی، هزینه های تولید (نیروی کار، خدمات و غیره) و برخی هزینه های حمل و نقل است.

۳۱.۳.۱ گردش موجودی کالا

گردش موجودی کالا[۳۱]، نسبتی است که نشان می‌دهد موجودی کالا و مواد شرکت، در یک بازه مشخص، چند بار بفروش رسیده و جایگزین شده است.

۴.۱ مشخصات متغیرها

در جدول ۲.۱ در ستون اول نام هر متغیر در مجموعه داده و در ستون دوم محتوای متغیر مربوط آورده شده است.

جدول ۲.۱ : مشخصات متغیرها

net profit / total assets	X1
total liabilities / total assets	X2
working capital / total assets	X3
current assets / short-term liabilities	X4
[(cash + short term securities + receivables - short term liabilities) / (operating expenses - depreciation)] * 365	X5
retained earnings / total assets	X6

[28] Operating activities
[29] Profit on operating activities
[30] Cost of products sold
[31] Inventory Turnover

EBIT / total assets	X7
book value of equity / total liabilities	X8
sales / total assets	X9
equity / total assets	X10
(gross profit + extraordinary items + financial expenses) / total assets	X11
gross profit / short-term liabilities	X12
(gross profit + depreciation) / sales	X13
(gross profit + interest) / total assets	X14
(total liabilities * 365) / (gross profit + depreciation)	X15
(gross profit + depreciation) / total liabilities	X16
total assets / total liabilities	X17
gross profit / total assets	X18
gross profit / sales	X19
(inventory * 365) / sales	X20
sales (n) / sales (n-1)	X21
profit on operating activities / total assets	X22
net profit / sales	X23
gross profit (in 3 years) / total assets	X24
(equity - share capital) / total assets	X25
(net profit + depreciation) / total liabilities	X26
profit on operating activities / financial expenses	X27
working capital / fixed assets	X28
logarithm of total assets	X29
(total liabilities - cash) / sales	X30
(gross profit + interest) / sales	X31
(current liabilities * 365) / cost of products sold	X32
operating expenses / short-term liabilities	X33
operating expenses / total liabilities	X34
profit on sales / total assets	X35
total sales / total assets	X36
(current assets - inventories) / long-term liabilities	X37
constant capital / total assets	X38
profit on sales / sales	X39

(current assets - inventory - receivables) / short-term liabilities	X40
total liabilities / ((profit on operating activities + depreciation) * (12/365))	X41
profit on operating activities / sales	X42
rotation receivables + inventory turnover in days	X43
(receivables * 365) / sales	X44
net profit / inventory	X45
(current assets - inventory) / short-term liabilities	X46
(inventory * 365) / cost of products sold	X47
EBITDA (profit on operating activities - depreciation) / total assets	X48
EBITDA (profit on operating activities - depreciation) / sales	X49
current assets / total liabilities	X50
short-term liabilities / total assets	X51
(short-term liabilities * 365) / cost of products sold)	X52
equity / fixed assets	X53
constant capital / fixed assets	X54
working capital	X55
(sales - cost of products sold) / sales	X56
(current assets - inventory - s-t-liabilities) / (sales - gross profit - depreciation)	X57
total costs /total sales	X58
long-term liabilities / equity	X59
sales / inventory	X60
sales / receivables	X61
(short-term liabilities *365) / sales	X62
sales / short-term liabilities	X63
sales / fixed assets	X64

فصل ۲

آزمایش‌ها

همانطور که گفته شد در این پروژه از هفت دسته بند درخت تصمیم، جنگل تصادفی، ماشین بردار پشتیبان، استدلال مبتنی بر حافظه نزدیک‌ترین همسایه و الگوریتم‌های مبنتی بر نظریه بیز شامل گاوسی، برنولی و چندجمله ای استفاده شده است. پارامترهای هر دسته‌بندی به سه حالت مختلف مقداردهی شده و آزمایش شده اند. به جز روش الگوریتم مبتنی بر نظریه بیز گاوسی امکان مقداردهی در حالت‌های مختلف برایش فراهم نبود. در زیر مقداردهی پارامترهای هر دسته‌بندی آمده است.

۱.۲ درخت تصمیم

در زیر جدول مربوط به مقادیر پارامترهای درخت تصمیم در سه آزمایش انجام شده، آمده است.

جدول ۱.۲: پارامترهای درخت تصمیم و مقداردهی آن‌ها

$Parameters$	$Experimnet1$	$Experimnet2$	$Experimnet3$
c	1	1	1
$kernel$	$'rbf'$	$'Linear'$	$'Poly'$
$degree$	3	3	3
$gamma$	$'auto'$	$'auto'$	$'auto'$
$coef0$	0	0	0
$shrinking$	$TRUE$	$TRUE$	$TRUE$
$probability$	$FALSE$	$FALSE$	$FALSE$
tol	0.001	$1.00E-04$	$1.00E-07$
$sizecache$	200	200	200
$classweight$	$None$	$None$	$None$
$verbose$	$FALSE$	$FALSE$	$FALSE$
$maxiter$	-1	-1	-1
$decisionfunction_shape$	$'ovr'$	$'ovr'$	$'ovr'$
$randomstate$	$None$	$None$	$None$

۲.۲ جنگل تصادفی

در جدول ۲.۲ مقادیر پارامترهای جنگل تصادفی در سه آزمایش انجام شده، آمده است.

جدول ۲.۲: پارامترهای جنگل تصادفی و مقداردهی آن‌ها

$Parameters$	$Experimnet1$	$Experimnet2$	$Experimnet3$
$estimators\ n$	20	20	50
$criterion$	$'gini'$	$'gini'$	$'gini'$
$maxdepth$	2	2	2
$minsamples_split$	2	2	2
$minsamples_leaf$	1	1	1
$leaf\ fraction\ minweight$	0	0	0
$maxfeatures$	$'auto'$	$'auto'$	$'auto'$
$nodes\ maxleaf$	30	30	15
$decrease\ minimpurity$	0	0	0
$split\ minimpurity$	$None$	$None$	$None$
$bootstrap$	$TRUE$	$TRUE$	$TRUE$
$oobscore$	$FALSE$	$FALSE$	$FALSE$
$njobs$	1	1	1
$state\ random$	0	0	0
$verbose$	0	0	0
$start\ warm$	$FALSE$	$FALSE$	$FALSE$
$weight\ class$	$None$	$None$	$None$

۳.۲ الگوریتم مبتنی بر نظریه بیز گاوسی

در جدول ۳.۲ پارامتر الگوریتم مبتنی بر نظریه بیز گاوسی و مقدار آن در تنها آزمایش انجام شده، آمده است.

جدول ۳.۲: تنها پارامتر الگوریتم مبتنی بر نظریه بیز گاوسی و مقدار آن

$Parameters$	$Theonlyexperimnet$
$Priors$	$None$

۴.۲ الگوریتم مبتنی بر نظریه بیز برنولی

در جدول زیر مقادیر پارامترهای الگوریتم مبتنی بر نظریه بیز برنولی در سه آزمایش انجام شده، آمده است.

جدول ۴.۲: پارامترهای الگوریتم مبتنی بر نظریه بیز برنولی و مقداردهی آن‌ها

$Parameters$	$Experimnet1$	$Experimnet2$	$Experimnet3$
$alpha$	1	0.5	0.6
$binarize$	0	0	0
$prior\ fit$	$TRUE$	$FALSE$	$TRUE$
$prior\ class$	$None$	$None$	$None$

۵.۲ الگوریتم مبتنی بر نظریه بیز چندجمله‌ای

در جدول ۳.۲ پارامترهای الگوریتم مبتنی بر نظریه بیز چندجمله‌ای و مقادیرآن ها در سه آزمایش انجام شده، آمده است.

جدول ۵.۲: پارامترهای الگوریتم مبتنی بر نظریه بیز چندجمله‌ای و مقداردهی آن‌ها

$Parameters$	$Experimnet1$	$Experimnet2$	$Experimnet3$
$alpha$	0.5	0.5	0.7
$prior\ fit$	$TRUE$	$FALSE$	$TRUE$
$prior\ class$	$None$	$None$	$None$

۶.۲ استدلال مبتنی بر حافظه نزدیک‌ترین همسایه

در جدول زیر پارامترهای استدلال مبتنی بر حافظه نزدیک‌ترین همسایه و مقادیرآن ها در سه آزمایش انجام شده، آمده است.

جدول ۶.۲: پارامترهای استدلال مبتنی بر حافظه نزدیک‌ترین همسایه و مقداردهی آن‌ها

$Parameters$	$Experimnet1$	$Experimnet2$	$Experimnet3$
$neighbors\ n$	3	7	12
$weights$	$'uniform'$	$'uniform'$	$'uniform'$
$algorithm$	$'auto'$	$'auto'$	$'auto'$
$leaf_size$	30	30	30
p	2	2	2
$metric$	$'minkowski'$	$'manhattan'$	$'euclidean'$
$params\ metric$	$None$	$None$	$None$
$jobs\ n$	1	1	1

۷.۲ ماشین بردار پشتیبان

در زیر جدول مربوط به مقادیر پارامترهای ماشین بردار پشتیبان در سه آزمایش انجام شده، آمده است.

جدول ۷.۲: پارامترهای ماشین بردار پشتیبان و مقداردهی آن‌ها

$Parameters$	$Experimnet1$	$Experimnet2$	$Experimnet3$
c	1	1	1
$kernel$	$'rbf'$	$'Linear'$	$'Poly'$
$degree$	3	3	3
$gamma$	$'auto'$	$'auto'$	$'auto'$
$coef0$	0	0	0
$shrinking$	$TRUE$	$TRUE$	$TRUE$
$probability$	$FALSE$	$FALSE$	$FALSE$
tol	0.001	$1.00E-04$	$1.00E-07$
$cachesize$	200	200	200
$classweight$	$None$	$None$	$None$
$verbose$	$FALSE$	$FALSE$	$FALSE$
$maxiter$	-1	-1	-1
$shape\ function\ decision$	$'ovr'$	$'ovr'$	$'ovr'$
$state\ random$	$None$	$None$	$None$

فصل ۳

نتایج

نتایج گرد آمده در ادامه این بخش را در چهار جدول نمایش داده شده است. برای هر دسته‌بند محتوای جداول مشابه و به قرار زیر است. هر چهار جدول پنج ستون دارد، هر ستون بیان متعلق به مجموعه داده مربوط به یکی از سال‌ها می‌باشد، همانطور که گفتیم این مجموعه داده‌ها مجزا می‌باشند و دسته‌بندها روی این پنج مجموعه به طور جداگانه اعمال شده‌اند. جدول اول دقت دسته‌بند را نشان می‌دهد.سطر اول جدول دقت دسته‌بند بر روی مجموعه داده آموزش و سطر دوم جدول دقت دسته‌بند بر روی مجموعه داده آزمایش را نشان می‌دهد. جدول دوم AUC دسته بند را نشان می‌دهد. سطر اول جدول AUC بر روی مجموعه داده‌های آموزشی و سطر دوم AUC بر روی مجموعه داده‌های آزمایشی. جدول سوم میانگین($mean$) و انحراف معیار(std۱) AUC دسته بند را بر روی $10 - foldcrossvalidation$ نشان می‌دهد. جدول چهارم میانگین و انحراف معیار دقت دسته بند را بر روی $10 - foldcrossvalidation$ نشان می‌دهد. برای تست روش های ذکر شده، از کتابخانه $Scikit - Learn$ پایتون نسخه ۱.۱۹.۰ استفاده شده است. قبل از اعمال روش ها، تمام مجموعه داده ها را عادی‌سازی۱ شده‌اند.

۱.۳ درخت تصمیم

جدول ۱.۳: دقت درخت تصمیم بر روی مجموعه های آزمایش و آموزش

Decision tree	Year1	Year2	Year3	Year4	Year5
Accuracy on Train set	1.00000	1.00000	1.00000	1.00000	1.00000
Accuracy on Test set	1.00000	1.00000	1.00000	.999489	.999154

جدول ۲.۳: AUC درخت تصمیم بر روی مجموعه های آزمایش و آموزش

Decision tree	Year1	Year2	Year3	Year4	Year5
AUC on Train set	1.00000	1.00000	1.00000	1.00000	1.00000
AUC on Test set	1.00000	1.00000	1.00000	.994737	.994118

جدول ۳.۳: میانگین و انحراف معیار AUC درخت تصمیم در $10-foldcrossvalidation$

Decision tree	Year1	Year2	Year3	Year4	Year5
Mean:	.950037	.950026	.950000	.950027	.950610
std:	.099926	.099949	.100000	.099946	.098790

جدول ۴.۳: میانگین و انحراف معیار دقت درخت تصمیم در $10-foldcrossvalidation$

Decision tree	Year1	Year2	Year3	Year4	Year5
Mean:	.950000	.950000	.950000	.950000	.950000
std:	.143477	.143418	.142377	.141362	.138834

۲.۳ جنگل تصادفی

جدول ۵.۳: دقت جنگل تصادفی بر روی مجموعه داده‌های آموزش و آزمایش

RF	Year1	Year2	Year3	Year4	Year5
Accuracy on Train set	.991640	.973830	.966203	.974726	.996193
Accuracy on Test set	.992176	.973464	.964779	.969883	.993238

جدول ۶.۳: AUC جنگل تصادفی بر روی مجموعه داده‌های آموزش و آزمایش

RF	Year1	Year2	Year3	Year4	Year5
AUC on Train set	.893182	.670279	.645885	.764286	.972308
AUC on Test set	.892157	.649351	.606383	.689474	.952941

پ

جدول ۷.۳: میانگین و انحراف معیار AUC جنگل تصادفی در $10-foldcrossvalidation$

RF	Year1	Year2	Year3	Year4	Year5
Mean:	.964882	.972850	.976227	.97294	.981672
std:	.083204	.059101	.051220	.05550	.040714

جدول ۸.۳: میانگین و انحراف معیار دقت جنگل تصادفی در $10 - foldcrossvalidation$

RF	Year1	Year2	Year3	Year4	Year5
Mean:	.944330	.947998	.955342	.958740	.939770
std:	.104376	.051541	.019721	.017502	.132835

۳.۳ الگوریتم مبتنی بر نظریه بیز گاوسی

جدول ۹.۳: دقت الگوریتم مبتنی بر نظریه بیز گاوسی بر روی مجموعه داده‌های آموزش و آزمایش

GNB	Year1	Year2	Year3	Year4	Year5
Accuracy on Train set	.848097	.837326	.851719	.906050	.9321065
Accuracy on Test set	.852774	.838329	.856735	.902501	.9256128

جدول ۱۰.۳: AUC الگوریتم مبتنی بر نظریه بیز گاوسی بر روی مجموعه داده‌های آموزش و آزمایش

GNB	Year1	Year2	Year3	Year4	Year5
AUC on Train set	.916595	.901946	.906747	.922288	.548901
AUC on Test set	.895311	.897271	.909802	.918797	.547477

جدول ۱۱.۳: میانگین و انحراف معیار AUC الگوریتم مبتنی بر نظریه بیز گاوسی در $10 - foldcrossvalidation$

GNB	Year1	Year2	Year3	Year4	Year5
Mean:	.913811	.915947	.916807	.908977	.951885
std:	.155572	.149216	.140058	.131956	.035142

جدول ۱۲.۳: میانگین و انحراف معیار دقت الگوریتم مبتنی بر نظریه بیز گاوسی در $10 - foldcrossvalidation$

GNB	Year1	Year2	Year3	Year4	Year5
Mean:	.853707	.868241	.854198	.847961	.940618
std:	.300435	.287565	.301615	.303847	.019198

۴.۳ الگوریتم مبتنی بر نظریه بیز برنولی

جدول ۱۳.۳: دقت الگوریتم مبتنی بر نظریه بیز برنولی بر روی مجموعه داده‌های آموزش و آزمایش

BNB	Year1	Year2	Year3	Year4	Year5
Accuracy on Train set	.940946	.668633	.900988	.893796	.928511
Accuracy on Test set	.942390	.664373	.903379	.910158	.927303

جدول ۱۴.۳: AUC الگوریتم مبتنی بر نظریه بیز برنولی بر روی مجموعه داده‌های آموزش و آزمایش

BNB	Year1	Year2	Year3	Year4	Year5
AUC on Train set	.520000	.550000	.513000	.522000	.507000
AUC on Test set	.526670	.551109	.518475	.528222	.504972

جدول ۱۵.۳: میانگین و انحراف معیار AUC الگوریتم مبتنی بر نظریه بیز برنولی در $10 - foldcrossvalidation$

BNB	Year1	Year2	Year3	Year4	Year5
Mean:	.750000	.488000	.474000	.573000	.563000
std:	.358266	.062240	.199020	.142166	.119807

جدول ۱۶.۳: میانگین و انحراف معیار دقت الگوریتم مبتنی بر نظریه بیز برنولی در $10 - foldcrossvalidation$

BNB	Year1	Year2	Year3	Year4	Year5
Mean:	.941094	.690598	.901276	.897070	.928269
std:	.004445	.086559	.015338	.015573	.002412

۵.۳ الگوریتم مبتنی بر نظریه بیز چندجمله‌ای

جدول ۱۷.۳: دقت الگوریتم مبتنی بر نظریه بیز چندجمله ای بر روی مجموعه داده‌های آموزش و آزمایش

MNB	Year1	Year2	Year3	Year4	Year5
Accuracy on Train set	.786909	.766802	.760800	.761042	.807107
Accuracy on Test set	.794452	.763145	.764874	.767739	.804734

جدول ۱۸.۳: AUC الگوریتم مبتنی بر نظریه بیز چندجمله ای بر روی مجموعه داده‌های آموزش و آزمایش

MNB	Year1	Year2	Year3	Year4	Year5
AUC on Train set	.889115	.878582	.874406	.873752	.896434
AUC on Test set	.893358	.876915	.876931	.877951	.894809

جدول۱۹.۳: میانگین و انحراف معیار AUC الگوریتم مبتنی بر نظریه بیز چندجمله ای در $10 - foldcrossvalidation$

MNB	Year1	Year2	Year3	Year4	Year5
Mean:	.994375	.996853	.997512	.996853	.996621
std:	.155572	.149216	.140058	.131956	.035142

جدول ۲۰.۳: میانگین و انحراف معیار دقت الگوریتم مبتنی بر نظریه بیز چندجمله ای در $10 - foldcrossvalidation$

MNB	Year1	Year2	Year3	Year4	Year5
Mean:	.762257	.746707	.743053	.743219	.782572
std:	.381441	.392696	.394734	.393439	.364826

۶.۳ استدلال مبتنی بر حافظه نزدیک‌ترین همسایه

جدول ۲۱.۳: دقت استدلال مبتنی بر حافظه نزدیک‌ترین همسایه بر روی مجموعه داده‌های آموزش و آزمایش

KNN	Year1	Year2	Year3	Year4	Year5
Accuracy on Train set	.999110	.998894	.998095	.998595	.998730
Accuracy on Test set	.998577	.996560	.997144	.996937	.997464

جدول ۲۲.۳: AUC استدلال مبتنی بر حافظه نزدیک‌ترین همسایه بر روی مجموعه داده‌های آموزش و آزمایش

KNN	Year1	Year2	Year3	Year4	Year5
AUC on Train set	.990817	.993488	.987156	.990274	.990769
AUC on Test set	.980392	.973260	.983295	.978411	.987780

جدول ۲۳.۳: میانگین و انحراف معیار AUC استدلال مبتنی بر حافظه نزدیک‌ترین همسایه در $10 - foldcrossvalidation$

KNN	Year1	Year2	Year3	Year4	Year5
Mean:	.964302	.968089	.962905	.962678	.967435
std:	.071871	.072264	.073890	.071447	.075505

جدول ۲۴.۳: میانگین و انحراف معیار دقت استدلال مبتنی بر حافظه نزدیک‌ترین همسایه در $10 - foldcrossvalidation$

KNN	Year1	Year2	Year3	Year4	Year5
Mean:	.949578	.949658	.949576	.949495	.949751
std:	.142312	.142921	.141542	.141078	.135812

۷.۳ ماشین بردار پشتیبان

جدول ۲۵.۳: دقت ماشین بردار پشتیبان بر روی مجموعه داده‌های آموزش و آزمایش

SVM	Year1	Year2	Year3	Year4	Year5
Accuracy on Train set	.961224	.964000	.992146	.994511	.994289
Accuracy on Test set	.963727	.965602	.990481	.994385	.995773

جدول ۲۶.۳: AUC ماشین بردار پشتیبان بر روی مجموعه داده‌های آموزش و آزمایش

SVM	Year1	Year2	Year3	Year4	Year5
AUC on Train set	.504545	.54644	.917706	.955548	.962736
AUC on Test set	.963727	.965602	.990481	.994385	.995773

جدول ۲۷.۳: میانگین و انحراف معیار AUC ماشین بردار پشتیبان در $10 - foldcrossvalidation$

SVM	Year1	Year2	Year3	Year4	Year5
Mean:	1.00000	1.00000	1.00000	1.00000	.999867
std:	0.00000	0.00000	0.00000	0.00000	.000399

جدول ۲۸.۳: میانگین و انحراف معیار دقت ماشین بردار پشتیبان در $10 - foldcrossvalidation$

SVM	Year1	Year2	Year3	Year4	Year5
Mean:	.917993	.91742	.945105	.946434	.945522
std:	.130785	.134406	.139726	.140611	.136292

فصل ۴

تفسیر نتایج

همانطور که گفته شد، تا کنون روش‌های متعددی برای پیش‌بینی ورشکستگی شرکت‌ها استفاده شده است. مقاله‌ای که پیش‌تر مجموعه داده مورد استفاده ما را به کار گرفته است، روش‌هایی را که که در گذشته برای حل این مساله پیشنهاد شده بودند روی مجموعه داده حاضر آزمایش کرده و نتایج حاصل را در جدولی مانند جدول ۱.۴ ارائه کرده است. در سطر آخر بهترین نتیجه گزارش شده از نظر میانگین AUC بر روی $10 - foldcrossvalidation$ قرار داده شده است. در جدول ۲.۴ نتایج حاصل از آزمایش‌های ما در این پروژه با بهترین نتیجه از نتایج قبلی مقایسه شده است.[8 – 14]

در جدول ۲.۴ نتایج هفت دسته‌بند آزمایش شده در این پروژه با بهترین نتیجه از نتایج قبلی مقایسه شده است. از بین این هفت دسته‌بند چهار دسته‌بند جنگل تصادفی، الگوریتم مبتنی بر نظریه بیز چندجمله‌ای، استدلال مبتنی بر حافظه نزدیک‌ترین همسایه و ماشین بردار پشتیبان نتایج بهتری داشته اند. شایان ذکر است که راه حل ارائه شده در این پروژه فقط شامل معرفی دسته‌بند نمی‌شود چراکه بعضی از این دسته‌بندها در گذشته نیز استفاده شده‌اند. آنچه که در پروژه پیشنهاد می‌کنیم دسته‌بند و مقداردهی مشخص پارامترهای آن می‌باشد که در فصل دوم گفته شد.

همانطور که پیش‌تر گفته شد برای مجموعه داده‌های حاضر که از جمله مجموعه داده‌های نامتعادل قلمداد می‌شوند معیار دقت معیار خوبی برای سنجش کارایی دسته‌بند نمی‌باشد. لذا ما نیز از معیار AUC که معیار مناسب‌تری برای این قبیل مجموعه داده‌ها می‌باشد استفاده کرده‌ایم. اما در این مقطع که چهار دسته‌بند با AUC بسیار مطلوب هستیم می‌توان برای مقایسه بین این چهار دسته‌بند از معیار دقت که در فصل قبل نتایج آن به تفصیل آورده شده استفاده کرد.

با توجه به جدول ۳.۴ و جدول ۲.۴ در می‌یابیم که دو دسته‌بند ماشین بردار پشتیبان و الگوریتم مبتنی بر نظریه بیز چند جمله‌ای که بر اساس معیار AUC عملکرد بسیار خوبی دارند بر اساس معیار دقت عملکرد ضعیف‌تری نسبت به بقیه دارند. می‌توان نتیجه گرفت که این دو دسته‌بند (به‌ویژه دسته‌بند الگوریتم مبتنی بر نظریه بیز چندجمله‌ای) تمایل بیشتری به پیش‌بینی برچسب ورشکستگی دارند. از

جدول ۱.۴: نتایج گزارش شده از آزمایشات انجام شده قبلی بر روی مجموعه داده‌ها

	1stYear		2ndYear		3ndYear		4thYear		5thYear	
	MN	STD	MN	STD	MN	STD	MN	STD	MN	STD
LDA	.639	.083	.660	.037	.688	.030	.714	.063	.796	.041
MLP	.543	.042	.514	.042	.548	.041	.596	.049	.699	.059
JRip	.523	.030	.540	.025	.535	.022	.538	.026	.654	.049
CJRip	.745	.112	.774	.073	.804	.054	.799	.070	.778	.035
J48	.717	.059	.653	.068	.701	.062	.691	.076	.761	.049
CJ48	.658	.047	.652	.047	.618	.061	.611	.025	.719	.046
LR	.620	.065	.513	.042	.500	.000	.500	.000	.632	.119
CLR	.704	.065	.671	.032	.714	.034	.724	.041	.821	.037
AB	.916	.020	.850	.029	.861	.023	.885	.031	.925	.026
AC	.916	.023	.849	.022	.859	.022	.886	.015	.928	.023
SVM	.502	.006	.502	.006	.500	.000	.500	.000	.505	.006
CSVM	.578	.040	.517	.064	.614	.040	.615	.034	.716	.039
RF	.851	.044	.842	.028	.831	.031	.848	.027	.898	.035
XGB	.945	.033	.917	.027	.922	.025	.935	.024	.951	.024
XGBE	.953	.024	.941	.019	.929	.049	.940	.027	.954	.018
EXGB	.959	.018	.944	.021	.940	.032	.941	.025	.955	.019

جدول ۲.۴: نتایج آزمایش‌های انجام شده در پروژه در مقایسه با بهترین نتیجه تا کنون

	1stYear		2ndYear		3ndYear		4thYear		5thYear	
	MN	STD	MN	STD	MN	STD	MN	STD	MN	STD
EXGB	.959	.018	.944	.021	.940	.032	.941	.025	.955	.019
DT	.950	.100	.950	.100	.950	.100	.950	.100	.951	.099
RF	.965	.083	.973	.059	.976	.051	.973	.055	.982	.041
GNB	.914	.156	.916	.149	.917	.140	.909	.132	.952	.035
MNB	.994	.004	.997	.002	.998	.001	.997	.002	.997	.002
BNB	.750	.358	.488	.062	.474	.199	.573	.142	.563	.120
KNN	.964	.072	.968	.072	.963	.074	.963	.071	.967	.076
SVM	1.00	.000	1.00	.000	1.00	.000	1.00	.000	1.00	.000

جدول ۳.۴: مقایسه نتایج آزمایش‌های سنجش دقت بر $10 - fold cross validation$

	1stYear		2ndYear		3ndYear		4thYear		5thYear	
	MN	STD	MN	STD	MN	STD	MN	STD	MN	STD
RF	.944	.104	.948	.052	.955	.020	.959	.018	.940	.133
MNB	.762	.381	.747	.393	.743	.395	.743	.393	.783	.365
KNN	.950	.142	.950	.143	.950	.142	.949	.141	.950	.136
SVM	.918	.131	.917	.134	.945	.140	.946	.141	.946	.136

بین این چهار دسته بند، همه بجر استدلال مبتنی بر حافظه نزدیک‌ترین همسایه یا KNN، هرچه به سمت مجموعه داده‌های سال آخر نزدیک تر می‌شویم نتایج شان بهتر می‌شود و در واقع هر چه مجموعه داده آسان‌تر می‌شود، دقت و AUC آنها نیز بهتر می‌شود. از این امر می‌توان نتیجه گرفت که KNN نسبت به سختی و آسانی مجموعه داده حساسیت کمتری دارد. برای زمانی که ما از مجموعه داده به عنوان یک نمونه استفاده می‌کنیم و هدف اصلی معرفی یک دسته‌بند مناسب برای مسائل مشابه است، مانند حال، حساس نبودن دسته‌بند به سختی و آسانی مجموعه‌داده یک مزیت مهم حساب می‌شود

فصل ۵

نتیجه گیری

این پروژه روش‌های تازه‌ای برای پیشبینی ورشکستگی شرکت‌ها بر اساس فاکتورهای مالی آن‌ها ارئه می‌کند. در این مسیر مجموعه‌ای از اطلاعات حسابداری و مالی شرکت‌های لهستانی در طول پنج سال مورد استفاده قرار گرفت تا با ساخت مدل‌های دسته‌بند بر اساس این اطلاعات، وضعیت ورشکستگی یا عدم ورشکستگی این شرکت‌ها پیشبینی شود. کارایی و اعتبار الگوریتم‌های استفاده شده در این پروژه با روش‌های قبلی مقایسه شد و دیدیم که از بین این روش‌ها چهار روش جنگل تصادفی، الگوریتم مبتنی بر نظریه بیز چندجمله‌ای، استدلال مبتنی بر حافظه نزدیک‌ترین همسایه و ماشین بردار پشتیبان بهتر از همه روش‌های ارائه شده قبلی عمل‌کردند. از بین این چهار روش، روش استدلال مبتنی بر حافظه نزدیک‌ترین همسایه به دلیل ثبات بیشتر در مجموعه داده‌های مختلف، روش بهتری ارزیابی می‌شود. با وجود اینکه از مجموعه داده خاصی برای سنجش روش‌ها استفاده شد اما این آزمایش محدود به این مجموعه داده نمی‌شود و بر روی هر جامعه از شرکت‌های درحال فعالیت که قصد پیش‌بینی ورشکستگی آنها را داریم، با ساخت ماتریسی مشابه با ماتریس‌های ارائه شده در معرفی مجموعه داده، می‌توان این جامعه از شرکت‌ها را با الگوریتم‌های ارائه شده در این پروژه و پارامترهای تعیین شده پیشبینی کرد.

منابع

[1] Kohavi, Ron (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection". Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. San Mateo, CA: Morgan Kaufmann. 2 (12): 1137–1143.

[2] Zieba, M., Tomczak, S. K., Tomczak, J. M. (2016). Ensemble Boosted Trees with Synthetic Features Generation in Application to Bankruptcy Prediction. Expert Systems with Applications.

[3] Fabio Sartori, Alice Mazzucchelli, Angelo Di Gregorio. Bankruptcy forecasting using case-based reasoning: The CRePERIE approach. Expert Systems with Applications, Volume 64, 1 December 2016, Pages 400-411.

[4] Philippe du Jardin. Dynamics of firm financial evolution and bankruptcy prediction. Expert Systems with Applications, Volume 75, 1 June 2017, Pages 25-43.

[5] Altman, 1968, E.I. Altman. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy.The Journal of Finance, 23 (1968), pp. 589-609.

[6] Ho, 1995,T.K. Ho. Random decision forests. Document analysis and recognition, 1995., proceedings of the third international conference on, IEEE (1995), pp. 278-282.

[7] E.I. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy", The Journal of Finance, 23 (1968), pp. 589-609

[8] B. Back, T. Laitinen, K. Sere, "Neural networks and genetic algorithms for bankruptcy predictions", Expert Systems with Applications, 11 (1996), pp. 407-413

[9] W.W. Cohen, "Fast effective rule induction", Proceedings of the twelfth international conference on machine learning (1995), pp. 115-123

[10] J.R. Quinlan, "C4.5: Programs for machine learning", Morgan Kaufman Publishers, San Francisco, CA, USA(1993)

[11] Y. Freund, R.E. Schapire, et al. "Experiments with a new boosting algorithm", Icml (1996), pp. 148-156

[12] W. Fan, S.J. Stolfo, J. Zhang, P.K. Chan, "Adacost: Misclassification cost-sensitive boosting", Icml (1999), pp. 97-105

[13] C. Cortes, V. Vapnik, "Support-vector networks", Machine learning, 20 (1995), pp. 273-297

[14] T.K. Ho, “Random decision forests”, Document analysis and recognition, 1995., proceedings of the third international conference on, IEEE (1995), pp. 278-282

[۱۵] داده کاوی کاربردی/محمد صنیعی آباده، سینا محمودی، محدثه طاهرپرور، ویراست۲.نشر نیاز دانش.

Abstract

During the last century, bankruptcy of companies has been a topic of interest to researchers and is now also an important economic topic. In this project, we are trying to find a model for predicting bankruptcy by applying data mining classification methods. The purpose of the project is to find a more precise method than previously proposed methods to predict this phenomenon. In order to achieve this goal, seven algorithms including decision tree, random forest, support vector machine, K-Nearest neighbor, and algorithms based on Naïve Bayes theory including Gaussian, Bernoulli, and multinomial have been used. The used dataset, includes information that has been collected over more than 10,000 companies over a period of five consecutive years. This collection includes 64 attributes that are derived from financial and accounting information of these companies. Dataset is divided into five matrixes. Each sample in these matrixes represents a company with a label of bankrupted or not-bankrupted. The goal of the algorithm is to classify sample labels using their attributes. In the process of the project, the efficiency of the mentioned methods is examined in detail and compared with the efficiency of the methods presented in the past.

Faculty of Science
School of mathematics, statistics and computer science

Bankruptcy Prediction of companies by Data Mining methods

Author:

Sepideh Khajeh Haghverdi

Supervisor:

Dr. Hedieh Sajedi

A thesis submitted to Graduate Studies Office
in partial fulfillment of the requirements for the degree of
B.Sc.in Computer Science

Februry 2018