Presenting Financial Bankruptcy Risk Prediction Model of Stock and Transborder Companies Using Machine Learning Algorithms

Mohsen Ali
Ph.D. Student in Industrial Management, Rudehen branch, Islamic Azad University. Rudehen, Iran.
Seyed Alireza Mirarab Baygi (corresponding author)
Assistant Professor, Rudehen branch, Faculty Member of Islamic Azad University, Rudehen, Iran.
hossieneghbali@yahoo.com
Nima Farajian
Assistant Professor, Lecturer University of Kashan, Iran.

Abstract:
Bankruptcy or business failure can have a negative impact on both the company itself and the global economy. In this research, the financial bankruptcy risk prediction of stock and transborder companies has been done using machine learning algorithms, Where the ultimate goal is to predict the financial bankruptcy risk of stock exchange and transborder companies. Collective learning is a field of machine learning in which instead of using a model to solve a problem, use multiple models in combination to increase the output estimation power of the model. Each model is retrained using optimal features. As a result, the accuracy of predicting machine learning model by Stacking method, which is one of the strongest techniques of collective learning, to predict financial bankruptcy risk is higher than similar methods.

Keywords: Random Forest, Logistic regression, Financial bankruptcy, Collective Learning, Machine Learning

فصلنامه تحلیل بازار سرمایه. سال دوم، شماره دوم. تابستان ۱۴۰۱. صفحه ۷۹ تا ۹۹.

تاریخ دریافت مقاله:
۱۴۰۱/۰۴/۳۱

تاریخ پذیرش مقاله:
۱۴۰۱/۰۶/۲۴

ارائه مدل پیش بینی ریسک ورشکستگی مالی شرکت های بورسی و فرابورسی با بهره گیری از الگوریتم های یادگیری ماشین

محسن عالی

دانشجوی دکتری مدیریت صنعتی، واحد رودهن، دانشگاه آزاد اسلامی، رودهن، ایران.

سید علیرضا میرعرب بایگی (نویسنده مسئول)

استادیار، واحد رودهن، دانشگاه آزاد اسلامی، رودهن، ایران.

hossieneghbali@yahoo.com

نیما فرجیان

استادیار، مدرس دانشگاه کاشان، ایران.

چکیده:

ورشکستگی یا شکست کسب و کار می تواند تاثیر منفی هم روی خود شرکت و هم اقتصاد جهانی داشته باشد. در این پژوهش ارائه پیش بینی ریسک ورشکستگی مالی شرکت های بورسی و فرابورسی با بهره گیری از الگوریتم های یادگیری ماشین صورت گرفته است، که در آن هدف پیش بینی نهایی ریسک ورشکستگی مالی شرکت های بورسی و فرابورسی است. یادگیری جمعی، حوزه ای از یادگیری ماشین هست که در آن به جای اینکه از یک مدل برای حل یک مسئله استفاده شود، از چندین مدل به صورت ترکیبی استفاده می گردد تا توان تخمین خروجی مدل را بالاتر ببرند. هر مدل با بهره گیری از ویژگی های بهینه مورد آموزش مجدد قرار می گیرد. در نهایت دقت پیش بینی مدل یادگیری ماشین به روش Stacking که یکی از قوی ترین تکنیک های یادگیری جمعی است، برای پیش بینی ریسک ورشکستگی مالی از روش های مشابه بالاتر است.

واژه‌های کلیدی: جنگل تصادفی، رگرسیون لجستیک، ورشکستگی مالی، یادگیری جمعی، یادگیری ماشین.

۱. مقدمه

سرمایه گذاران معمولا می خواهند با پیش بینی امکان ورشکستگی یک شرکت از زوال سرمایه خود جلوگیری کنند. از این رو آنها به دنبال روش هایی هستند که بتوانند به وسیله آن ورشکستگی شرکت ها را پیش بینی کنند. اگر کسی دلیل ورشکستگی شرکت را مطلع شود با برنامه ریزی لازم شرکت را از مرگ حتمی نجات می دهد. بنابراین پیش بینی ورشکستگی، پیش نیاز جلوگیری از ورشکستگی هست. به طور کلی اکثر مردم انتظار دارند که تغییر ناگهانی و بی علت حسابرس، اظهارنظر حسابرس خبره و نسبت های مالی ضعیف به عنوان هشدارهای اولیه از ورشکستگی مورد استفاده قرار گیرد، اما آنها انتظار ندارند که حسابرسی و نسبت های مالی ضعیف به عنوان علائم کافی و قطعی ورشکستگی در نظر گرفته شود. مشخص شدن دلایل دقیق ورشکستگی در هر مورد خاص کار آسانی نیست. در اغلب موارد دلایل متعددی با هم می توانند منجر به ورشکستگی بشوند. تجزیه و تحلیل صورت های مالی مستلزم ابزار و تکنیک هایی می یاشد که تحلیل گران را قادر می سازد صورت های مالی جاری و گذشته را بررسی بکنند به نحوی که عملکرد و وضعیت مالی یک شرکت مورد ارزیابی قرار گیرد و احتمال خطرات آتی و بالقوه برآورد بشود. در این پژوهش ارائه مدل پیش بینی ریسک ورشکستگی مالی شرکت های بورسی و شرکت های فرا بورسی با بهره گیری از الگوریتم های یادگیری ماشین صورت گرفته، که در آن هدف پیش بینی نهایی ریسک ورشکستگی مالی شرکت های بورسی و شرکت های فرا بورسی است.

شدت گرفتن رقابت در عرصه تولید و خدمات موجب گردیده است که بسیاری از شرکت ها ورشکسته شوند و از گردونه رقابت خارج شوند. این امر موجب نگرانی صاحبان سرمایه، مدیران، بستانکاران، و بطور کلی جامعه شده است. ورشکستگی می تواند زیان های هنگفتی را برای سهامداران، مدیران، شرکت ها واقتصاد کشور ایجاد کند، بنابراین انجام تحقیقی که بتواند به حل این مسئله کمک کند، اهمیت دارد. در واقع اگر بتوان از طریق مدلی، وقوع احتمال ورشکستگی در شرکت ها را پیش بینی کرد و پس از آن با علت یابی و استفاده از روش های حل مسئله به اصلاح

امور شرکت پرداخت، می توان از به هدر رفتن ثروت ملی در قالب سرمایه های فیزیکی و انسانی و آثار آن جلوگیری کرد. علاوه بر این، چنین مدلی می تواند راهنمای خوبی برای تصمیم گیرندگان مثل شرکت های سرمایه گذاری، بانک ها و دولت باشد. سرمایه گذاران همواره می خواهند با پیش بینی امکان ورشکستگی یک شرکت، از ریسک سوخت شدن اصل و فرع سرمایه خود جلوگیری کنند. از این رو آن ها، در پی روش هایی هستند که بتوانند به وسیله آن ورشکستگی مالی شرکت ها را تخمین بزنند، زیرا در صورت ورشکستگی، قیمت سهام شرکت ها به شدت کاهش می یابد. در حوزه ادبیات مالی صرف نظر از برداشت های عامیانه، برداشت های متفاوتی از واژه ورشکستگی وجود دارد. گیتمن[1] محققین ورشکستگی را ناشی از سوء مدیریت و فزونی بدهی ها بر دارایی ها می دانند (Gitman, 1998). با توجه به توانایی های مدل یادگیری جمعی و ناشناخته بودن این توانایی ها در بازارهای مالی ایران، تحقیق حاضر برای پیش بینی ورشکستگی از مدل های یادگیری جمعی برای پیش بینی ورشکستگی مالی استفاده می کند.

۲. مبانی نظری و پیشینه پژوهش

۱. مدل رگرسیون لجستیک[2]

معمولا برای بیان شدت رابطه خطی بین دو متغیر کمی، از ضریب همبستگی استفاده می‌شود. همچنین برای نمایش مدل رابطه بین آن دو هم از مدل رگرسیونی استفاده می شود. در این میان یک الگو برای پیش‌بینی متغیر وابسته (Y) براساس متغیر مستقل (X) ایجاد می‌گردد. ولی باید دقت کرد که در مدل ایجاد شده، هر دو متغیر وابسته و مستقل، کمی هستند. همچنین شرط پیوسته بودن این مقدارها هم در روش رگرسیون نهفته است. اما ممکن است بخواهیم رابطه بین یک متغیر مستقل (با مقدارهای پیوسته) را با یک متغیر وابسته با مقدارهای کیفی بسنجیم. در این وضعیت روش عادی

1. Gitman
2. Logistic Regression

رگرسیون خطی جوابگو نخواهد بود و باید از «رگرسیون لجستیک» استفاده شود. رگرسیون لجستیک گاهی یک مورد خاص از مدل خطی عمومی و رگرسیون خطی دیده می شود. مدل رگرسیون لجستیک، بر اساس فرض‌های بسیار متفاوتی (درباره رابطه متغیرهای مستقل و وابسته) از رگرسیون خطی هست. تفاوت مهم این دو مدل را در دو ویژگی رگرسیون لجستیک می‌توان دید. اول توزیع شرطی $y|\vec{x}$ یک توزیع برنولی به جای یک توزیع گوسی می باشد زیرا متغیر وابسته دودویی هست. دوم مقادیر پیش‌بینی احتمالاتی هست و محدود بین بازه صفر و یک و با کمک تابع توزیع لجستیک بدست می‌آید رگرسیون لجستیک احتمال خروجی را پیش‌بینی می نماید. همان طور که می‌دانیم، منظور از رگرسیون خطی، ایجاد رابطه‌ای خطی برحسب پارامتر جهت نمایش ارتباط بین متغیر وابسته و مستقل است. فرم مدل رگرسیون خطی ساده به صورت زیر می باشد:

$$Y=\beta_0+\beta_1X+\epsilon \qquad (۱)$$

همان طور که مشخص هست این رابطه، معادله یک خط است که جمله خطا یا همان ϵ به آن اضافه شده است. پارامترهای این مدل خطی، عرض از مبدا (β_0) و شیب خط (β_1) می باشند. می‌توان $\hat{Y}$ را میانگین مشاهدات برای متغیر وابسته به ازای مقدار ثابت متغیر مستقل در نظر گرفت، اگر $\hat{Y}$ مقدار برآورد برای متغیر وابسته باشد. اگر میانگین را جایگزین با امید ریاضی بکنیم با فرض اینکه میانگین جمله خطا نیز صفر هست، خواهیم داشت:

$$\hat{Y} = E(Y|X = x) = \hat{\beta}_0 + \hat{\beta}_1 x \qquad (۲)$$

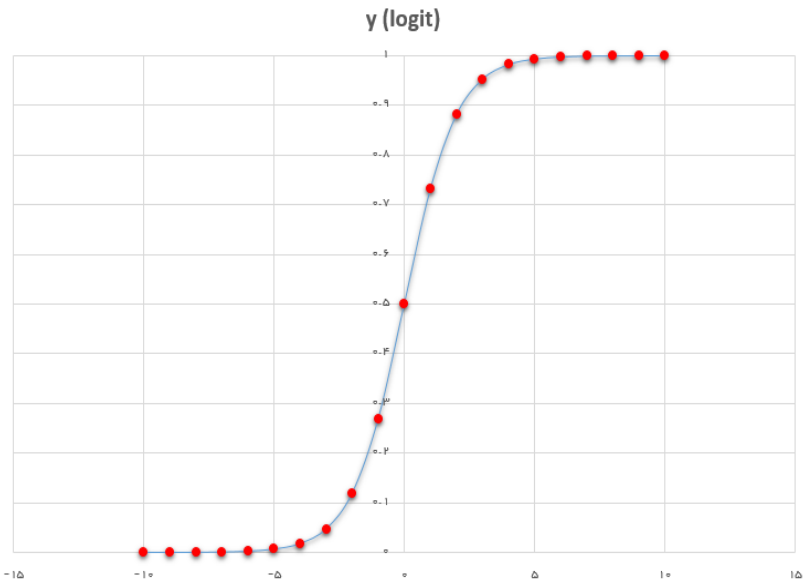
که در آن $E(Y|X=x)$ نشان‌دهنده امید ریاضی (متوسط) شرطی می باشد و همچنین $\hat{\beta}_0$ و $\hat{\beta}_1$ برآوردهای مربوط به هر یک از پارامترها می باشند. اگر مقدار متغیر وابسته (Y)، باینری (دو حالتی) و شامل ۰ و ۱ باشد مشخص هست که دارای توزیع برنولی می باشد و امید ریاضی آن به صورت زیر محاسبه می‌گردد:

$$\hat{Y} = E(Y|X = x) = P(Y = 1|X = x) = p(x) \qquad (۳)$$

به این ترتیب برای متغیر وابسته برنولی، مدل رگرسیون مشخص می‌گردد. مقدار پیش‌بینی برای متغیر وابسته، با احتمال $p(x)$ انجام گرفت. برای مشخص کردن مدل رابطه بین متغیر مستقل و وابسته به جای رابطه خطی، به تابعی نیاز داریم که در حدود ۰ تا ۱ تغییر بکند. در رگرسیون لجستیک احتمال رخداد یک طبقه خاص متغیر وابسته، بر اساس تابع نمایی متغیرهای مستقل برآورد می‌گردد. دامنه این تابع اعداد حقیقی می باشد و برد این تابع بین ۰ و۱ هست (Cramer,2002). نمودار مربوط به تابع بر اساس پارامترهای $b_1=1,\ b_0=0$ در تصویر دیده می‌شود.

$$f(x) = \frac{e^{b_0+b_1x}}{1+e^{b_0+b_1x}} \qquad (۴)$$

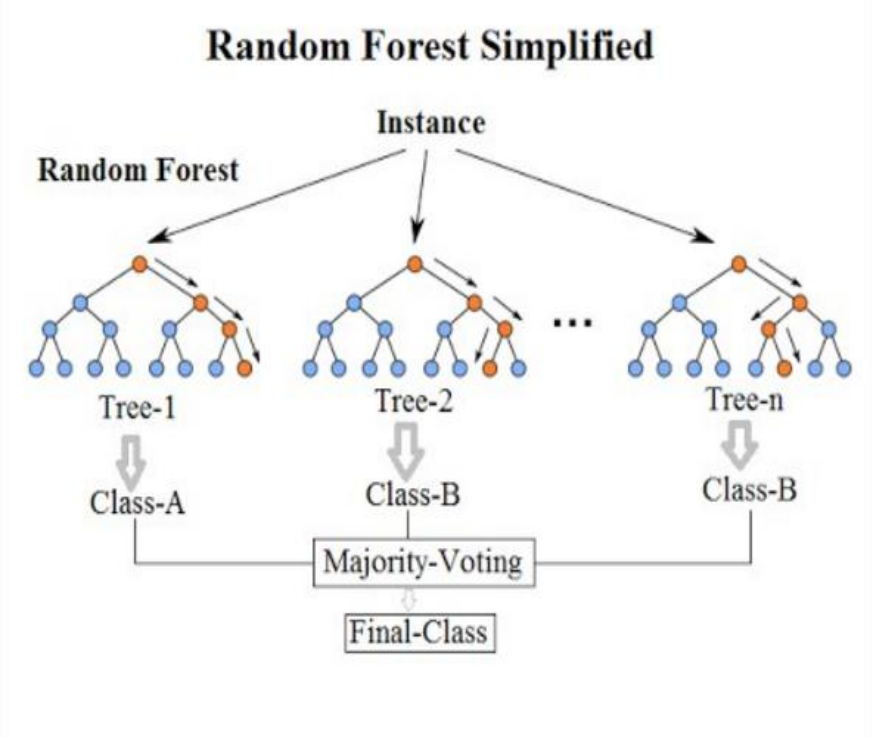
شکل ۱. تابع لجستیک استاندارد



۲. مدل جنگل تصادفی[۳]

جنگل تصـادفی یک الگوریتم یادگیری ماشـین[۴] هست که معمولا نتیجه های خوبی را حتی بدون تنظیم فرا پارامترهایش، فراهم می نماید. یک جنگل تصـــادفی با برازش چندین درخت تصـــمیم و رای‌گیری بین آن ها، نمونه‌های جدید را می تواند پیش‌بینی نماید.

شکل ۲. جنگل تصادفی



3 . Random Forest
4 . Machine Learning

جنگل تصـــادفی یک الگوریتم یادگیری با نظارت محســـوب می‌گردد. همان طور که از نام آن مشـخص هسـت، این الگوریتم جنگلی را به طور تصـادفی ایجاد می کند. «جنگل» ایجاد شـده در واقع گروهی از درخت‌های تصـمیم[۵] می باشـد. کار ساخت جنگل با استفاده از درخت‌ها معمولا به روش کیسـه‌گذاری[۶] انجام می گردد. ایده اصـلی روش کیسـه‌گذاری این گونه هسـت که ترکیبی از مدل‌های یادگیری، نتایج کلی مدل را افزایش می‌دهند. جنگل تصـادفی دارای فراپارامترهایی همانند درخت تصـمیم یا دسـته‌بند کیسـه‌گذاری[۷] می باشـد. این الگوریتم، به جای جسـت‌وجو به دنبال مهم‌ترین ویژگی‌هـا زمـان تقســیم کردن یـک گره[۸]، بـه دنبال بهترین ویژگی‌ها در میان مجموعه تصـادفی از ویژگی‌ها می‌باشد. این امر منجر به تنوع زیاد و سپس مدل بهتر می‌گردد. اصولا درخت تصـمیمی که بیش از حد عمیق باشـد الگوی دقیق نخواهد داشـت: دچار بیش برارزش می شود، و دارای ســوگیری پایین و واریانس بالا می گردد. برای کیســه گذاری درختان، مجموعه داده را با D نشان می دهیم:

$$D = (x_1 y_1).(x_2 y_2). \dots , (x_n y_n) \qquad (۵)$$

در مسئله رگرسیون مدل نهائی، میانگین همه درخت‌ها می باشد. همچنین با رأی‌گیری بین درختان، در مسئله دسته‌بندی یا Classification به جواب نهائی خواهیم رسید. یکی از مزایای جنگل تصادفی این هست که هم برای دسته‌بندی و هم برای رگرسیون قابل استفاده می باشد و راهکاری مناسب برای دیدن اهمیت نسبی که به ویژگی‌های ورودی اختصاص داده می‌شود هست. یکی از مهم ترین مشکلات در یادگیری ماشین، بیش‌برازش می باشد، ولی اغلب اوقات این مساله به این راحتی که برای دسته‌بند جنگل تصادفی اتفاق می‌افتد، رخ نمی دهد. محدودیت اصلی جنگل تصادفی

5 - Decision Trees
6 - Bagging
7 - Bagging Classifier
8 - Node

این هست که تعداد زیاد درخت‌ها می‌تواند الگوریتم را برای پیش‌بینی‌های دنیا واقعی کنند و غیر موثر کنند.

در صورت کلی، آموزش دادن این الگوریتم‌ها سریع انجام می‌گردد، ولی پیش‌بینی کردن بعد از این که مدل آموزش دید، اندکی کند رخ می دهد. یک پیش‌بینی درست تر احتیاج به درختان بیشتری دارد که منجر به کندتر شدن مدل نیز می گردد. جنگل تصادفی یک ابزار توصیفی نمی باشد، بلکه یک ابزار مدل‌سازی پیش‌بینی می باشد. این یعنی، اگر کاربر دنبال ارائه توصیفی از داده‌های خویش باشد، استفاده از رویکردهای دیگر ترجیح داده می‌شوند (Hastie, 2001).

۳. مدل گرادیان بوستینگ[۹]

روش گرادیان بوستینگ، نیز مانند جنگل تصادفی از درخت های تصمیم «ضعیف» استفاده می‌نماید. تفاوت این دو روش در این هست که در روش گرادیان بوستینگ درخت‌ها یکی بعد از دیگری آموزش داده می‌شوند. هر درخت زیرمجموعه در مرحله اول با داده هایی که به اشتباه توسط درخت قبلی پیش‌بینی شدند آموزش داده می‌شوند. این امر باعث می‌شود مدل بیشتر روی موارد پیچیده متمرکز گردد. تقویت گرادیان یک روش یادگیری ماشین برای مسائل کلاسه بندی و رگرسیون می باشد که یک مدل پیش بینی کننده را به شکل مجموعه ای از مدل های پیش بینی کننده ضعیف ایجاد می نماید. در این پژوهش از روش XGBoost استفاده شده است. همانند دیگر روش‌های تقویتی (بوستینگ)، تقویت گرادیان (گرادیان بوستینگ) ترکیبی خطی از یک سری از مدل‌های ضعیف برای ایجاد یک مدل قوی و کارآمد می باشد (Chen & Guestrin, 2016).

9- XGBoost

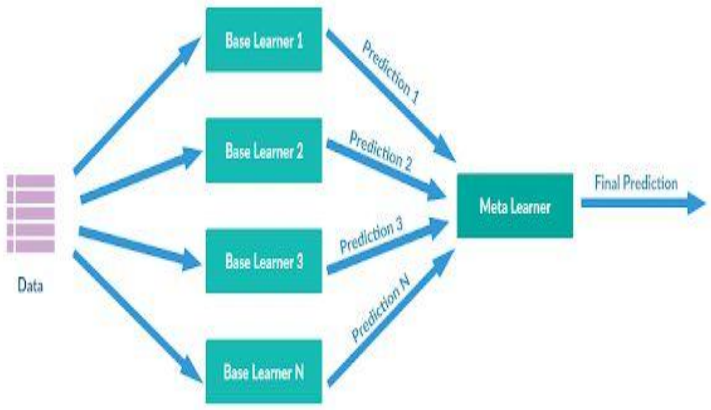
۴. تکنیک stacking

مدل های یادگیری ماشین ترکیبی[10] یا مدل های جمعی، از روش های یادگیری ماشین می باشد، که در این روش چندین مدل، که یادگیری های ضعیف[11] یا مدل های پایه نامیده می شوند، برای حل و به نتیجه رسیدن یک مسئله آموزش داده می شوند و برای داشتن نتایج دقیق تر و بهتر با هم ترکیب می گردند. روش ترکیبی stacking، مدل های پایه را با استفاده از متامدل با یکدیگر ترکیب می نماید. ایده اصلی stacking آموزش چندین مدل پایه متفاوت و ترکیب آنها از طریق آموزش یک متامدل می باشد تا بر اساس پیش بینی های مطرح شده توسط مدل های پایه، پیش بینی نهایی را انجام بدهد. پس برای ساخت مدل stacking به دو مورد نیاز هست:

L مدل پایه جهت آموزش داده ها و یک متامدل جهت ترکیب نتایج (Rokach,2010).

شکل زیر طرح کلی روش stacking را نشان می دهد:

شکل ۳. روش ترکیبی stacking



10. Ensemble Machine Learning
11. Weak Learner

در نهایت با بهره گیری از ۳ مدل پایه (مدل جنگل تصادفی، مدل رگرسیون لجستیک، XGBoost) و ویژگی های بهینه یک مدل ترکیبی (Stacking) مورد آموزش قرار می گیرد.

جنگل تصادفی یا جنگل‌های تصمیم تصادفی، یک روش یادگیری ترکیبی برای دسته‌بندی، رگرسیون می‌باشند، که بر طبق ساختاری متشکل از شمار زیادی درخت تصمیم، بر روی زمان آموزش و خروجی کلاس‌ها (کلاس‌بندی) یا برای پیش بینی های هر درخت به شکل مجزا، کار می‌کنند. برای درختان تصمیم که در مجموعه آموزشی دچار بیش برازش می‌شوند، جنگل‌های تصادفی مناسب می باشند. عملکرد جنگل های تصادفی معمولا بهتر از درخت تصمیم هست، اما این بهبود عملکرد تقریبا به نوع داده هم بستگی دارد (Hastie, 2001), (Piryonesi & El-Diraby, 2020), (Piryonesi & El-Diraby & Tamer, 2020).

گرادیان بوستینگ توان پیش‌بینی خیلی بالایی دارد که آن را تبدیل به گزینه ایی خوب برای دقت در رویدادهای مختلف می‌کند چرا که هم الگوریتم یادگیری درختی دارد و هم مدل خطی. این الگوریتم حدوداً ۱۰ برابر سریع‌تر از الگوریتم‌های موجود ارتقای گرادیان می باشد و به کاهش مدل‌های بزرگ کمک می‌کند. این الگوریتم شامل تابع‌های عینی مختلف، رگرسیون، رتبه بندی و کلاس بندی است. الگوریتم XGBoost یک کتابخانه تقویت گرادیان توزیع شده بهینه است که با هدف انعطاف‌پذیری، کارآمدی، و قابلیت حمل بالا طراحی شده است. تقویت درخت موازی را فراهم می کند که بسیاری از مشکلات علم داده را به صورت صحیح و سریع برطرف می‌کند. XGBoost در ابتدا به عنوان یک پروژه علمی پژوهشی توسط تیانگی چن به عنوان بخشی از گروه یادگیری عمیق ماشینی توزیع‌شده آغاز گشت (Chen & Guestrin, 2016).

در رگرسیون لجستیک، از تابعی به نام «تابع لجستیک» استفاده می‌شود، به همین علت این روش رگرسیونی، رگرسیون لجستیک نامیده شده است. تابع لجستیک به عنوان مدلی برای پیش‌بینی رشد جمعیت در سال ۱۹۲۰ توسط ریموند پرل و لاول رید دوباره ایجاد و به چاپ رسید که منجر به

استفاده آن در علم آمار گشت. در ابتدا آنها این مدل را برای مدل‌سازی جمعیت ایالات متحده آمریکا به کار گرفته بودند (Cramer,2002). دامنه این تابع لجستیک اعداد حقیقی می باشد و برد این تابع بین صفر و یک می‌باشد.

رگرسیون لجستیک، یک مدل آماری رگرسیون برای متغیرهای وابسته دوسویی مانند بیماری یا سلامت، مرگ یا زندگی می باشد. این مدل را می‌توان به عنوان مدل خطی تعمیم یافته‌ای که از تابع لوجیت به عنوان تابع پیوند استفاده می شود و خطایش از توزیع چندجمله‌ای پیروی می‌کند، به حساب آورد. منظور از دو سویی بودن، رخ داد یک واقعه تصادفی در دو موقعیت ممکن هست. به عنوان مثال خرید یا عدم خرید، ثبت نام یا عدم ثبت نام، ورشکسته شدن یا ورشکسته نشدن و ... متغییرهایی می باشند که فقط دارای دو موقعیت هستند و مجموع احتمال هر یک آن‌ها در نهایت عدد یک خواهد شد. کاربرد این روش در ابتدای ظهور در مورد کاربردهای پزشکی برای احتمال رخداد یک بیماری مورد استفاده قرار می‌گرفت. اما امروزه در تمام زمینه های علمی کاربرد وسیعی یافته اند. مقیاس های پزشکی دیگری که برای ارزیابی شدت بیماری استفاده می شود، توسط رگرسیون لجستیک ساخته شده‌اند (Biondo et al., 2000), (Kologlu et al., 2001), (Le Gall et al., 1993), (Marshall et al., 1995).

جنگل تصادفی به دلیل سادگی آن و قابلیت استفاده (هم برای دسته‌بندی و هم برای رگرسیون) ، یکی از پر کاربردترین الگوریتم‌های یادگیری ماشین محسوب می‌گردد. اولین الگوریتم برای جنگل‌های تصمیم تصادفی را «تین کم هو» با بهره‌گیری از روش زیر فضاهای تصادفی پدید آورد (Ho, 1995), (Ho, 1998). نسخه‌های بعدی توسط لیو بریمن ارتقا یافت (Breiman, 2001).

۳. فرضیه پژوهش

فرضیه سازی فرآیندی است که طی آن پژوهشگر، رابطه احتمالی بین متغیر وابسته و متغیرهای مستقل خود را پیش بینی می کند. پژوهشگر در این مرحله بر اساس تئوری یا تئوری های انتخاب

شده در چارچوب نظری تحقیق به روش قیاسی، فرضیه سازی را انجام می دهد و بر اساس فرضیه های تحقیق به طراحی مدل تحلیلی می پردازد. همان طور که مطرح شد، این پژوهش درصدد این است تا با استفاده از الگوریتم های یادگیری جمعی به پیش بینی ریسک ورشکستگی مالی شرکت های بورسی و فرابورسی بپردازد، بر این اساس فرضیه های مرتبط با این تحقیق بصورت زیر بیان می شوند:

فرضیه : دقت پیش بینی مدل یادگیری ماشین به روش stacking، برای پیش بینی ریسک ورشکستگی مالی از روش های مشابه بالاتر است.

۴. شیوه پژوهش

برای انجام این پژوهش، تمامی شرکت های پذیرفته شده در بورس اوراق بهادار تهران از اسفند ۱۳۸۱ لغایت اسفند ۱۳۹۸ به عنوان جامعه آماری در نظر گرفته می شوند و نمونه آماری از میان این شرکت ها استخراج می گردد. تعداد کل شرکت ها برابر ۱۵۹۴ عدد می باشد. از آنجایی که داده های مورد نیاز ما به صورت سال-شرکت می باشد و هر شرکت امکان دارد در چندین سال مورد مطالعه قرار گرفته باشد لذا تعداد کل داده ها ۱۷۴۷۴ سال- شرکت می باشد.

معمولا از روش های آماری و کمی برای پشتیبانی منطق تئوریک استفاده می کنند. استفاده از مدل های آماری در عمل با محدودیت هایی مواجه است به عنوان نمونه بیور به فرض خطی بودن رابطه بین متغیرها در مدل های تجزیه و تحلیل یک متغیره اشاره می کند (Altman, 1968). آلتمن به سه فرض محدود کننده در مدل های آنالیز تشخیص چند متغیره اشاره دارد که عبارتند از : نرمال بودن توزیع متغیرها، فرض وجود ماتریس توزیع یکنواخت و استفاده از احتمال های پیشین. مدل های شبکه عصبی و هوش مصنوعی نیز خالی از اشکال نبودند (Beaver, 1966). تعیین پارامترها به همراه آموزش دشوار است. طراحی بسیاری از شبکه های عصبی نیازمند داده های زیاد و تکرار زیاد برای آموزش است. در این پژوهش قصد داریم تا از روش های ترکیبی یادگیری ماشین برای

پیش بینی ریسک ورشکستگی مالی شرکت های بورسی و شرکت های فرابورسی بورس اوراق بهادار تهران استفاده کنیم. از روش یادگیری جمعی که یکی از حوزه های جدید یادگیری ماشین می باشد بهره خواهیم جست. یادگیری جمعی، حوزه ای از یادگیری ماشین می باشد که در آن برای حل یک مسئله به جای اینکه از یک مدل استفاده بشود، از چندین مدل به صورت ترکیبی استفاده می کنند تا توان تخمین خروجی مدل را بالاتر ببرند. در این پژوهش از تکنیک stacking که یکی از قوی ترین تکنیک های یادگیری جمعی هست استفاده خواهد شد. در مرحله نخست ۳ مدل پایه پیش بینی شامل مدل رگرسیون لجستیک ، مدل جنگل تصادفی و گرادیان بوستینگ توسط داده ها آموزش خواهند دید. در مرحله دوم مدل نهایی سیستم آموزش می بیند که نقش تصمیم گیرنده نهایی را دارد و در پروسه آموزش یاد می گیرد که هر مدل به چه صورت کار می کند و طبق توان مدل های پایه به هر کدام یک وزنی را مشخص می کند تا در پروسه تصمیم گیری که در این جا همان پیش بینی ورشکستگی مالی می باشد، استفاده کند. با استفاده از این تکنیک یادگیری جمعی، نتایج پیش بینی قابل اطمینان تر خواهند بود.

۵. یافته ها و تحلیل

در این پژوهش برای مدل سازی از ۳ مدل پایه: رگرسیون لجستیک (Logistic Regression)، جنگل تصادفی (Random Forest)، XGBoost استفاده شده است. و در نهایت از مدل ترکیبی stacking بهره گرفتیم. جهت آموزش مدل ها، داده ها به دو قسمت آموزش و تست تقسیم شده است. داده های آموزش، صورت های مالی قبل از سال ۱۳۹۴ می باشد که شامل ۹۱۴۳ داده می باشد. داده های تست، مربوط به صورت های مالی بعد از ۱۳۹۴ می باشد که مشتمل بر ۳۲۶۱ رکورد اطلاعاتی است. در این پژوهش برای جمع آوری داده ها و اطلاعات از روش های کتابخانه ای و میدانی استفاده شده است. مبانی تئوری پژوهش از کتب، مجلات و سایت های تخصصی فارسی و لاتین گردآوری می شود و داده های مالی مورد نیاز با مراجعه به سایت سازمان بورس اوراق بهادار تهران، صورت های مالی شرکت ها و همچنین با استفاده از نرم افزارهای تدبیرپرداز و ره آورد نوین

گردآوری شده اند. در این راستا، سعی می شود که هم متغیرهای حسابداری مبتنی بر ترازنامه، صورت سود و زیان و صورت جریان وجوه نقد و هم متغیرهای بازار استفاده و محتوای اطلاعاتی آنها مد نظر قرار گیرد. بنابراین، ابعاد سودآوری، کارایی، اهرم مالی، نقدینگی، نسبت های مبتنی بر هر سهم، نسبت های مبتنی بر جریان وجوه نقد و نسبت های بازار در نظر گرفته خواهد شد.

نتایج به شرح زیر می باشد:

۱. نتایج حل با رگرسیون لجستیک :

ارزیابی مدل رگرسیون لجستیک بر روی داده های تست با تمام ویژگی ها

نگاره Evaluation

Class	Precision	Recall
0	0.78	0.84
1	0.84	0.77
AVG/TOTAL	0.81	0.81

نگارهConfusion Matrix

Actual/Prediction	0	1
0	1360	250
1	385	1266

AUC: 0.8057
Accuracy Ratio :0.6998

ارزیابی مدل رگرسیون لجستیک بر روی داده های تست بعد از مهندسی ویژگی ها و Feature Selection

نگاره Evaluation

Class	Precision	Recall

0	0.83	0.84
1	0.84	0.84
AVG/TOTAL	0.83	0.84

نگاره Confusion Matrix

Actual/Prediction	0	1
0	1346	246
1	272	1379

AUC: 0.8991
Accuracy Ratio :0.7987

۱. نتایج حل با جنگل تصادفی :

ارزیابی مدل جنگل تصادفی بر روی داده های تست با تمام ویژگی ها

نگاره Evaluation

Class	Precision	Recall
0	0.75	0.86
1	0.84	0.72
AVG/TOTAL	0.79	0.79

نگاره Confusion Matrix

Actual/Prediction	0	1
0	1379	231
1	470	1181

AUC: 0.7859
Accuracy Ratio :0.7133

ارزیابی مدل جنگل تصـــادفی بر روی داده های تســـت بعد از مهندســـی ویژگی ها و Feature Selection

نگاره Evaluation

Class	Precision	Recall
0	0.82	0.84
1	0.84	0.82
AVG/TOTAL	0.83	0.83

نگاره Confusion Matrix

Actual/Prediction	0	1
0	1354	256
1	296	1355

AUC: 0.9021
Accuracy Ratio :0.8047

۲. نتایج حل با XGBoost :

ارزیابی مدل XGBoost بر روی داده های تست با تمام ویژگی ها

نگاره Evaluation

Class	Precision	Recall
0	0.76	0.84
1	0.83	0.74
AVG/TOTAL	0.79	0.79

نگاره Confusion Matrix

Actual/Prediction	0	1
0	1357	253
1	437	1214

AUC: 0.7890
Accuracy Ratio :0.7223

ارزیـابی مـدل XGBoost بر روی داده هـای تســت بعـد از مهندســی ویژگی ها و Feature Selection

نگاره Evaluation

Class	Precision	Recall
0	0.83	0.83
1	0.84	0.84
AVG/TOTAL	0.84	0.83

نگاره Confusion Matrix

Actual/Prediction	0	1
0	1339	271
1	270	1381

AUC: 0.9029
Accuracy Ratio :0.8060

۳. نتایج حل با Stacking :

ارزیابی مدل Stacking بر روی داده های تسـت با بهره گیری از ویژگی های انتخاب شده توسط هر کدام از مدل های پایه

نگاره Evaluation

Class	Precision	Recall
0	0.85	0.85
1	0.85	0.86
AVG/TOTAL	0.85	0.85

نگاره Confusion Matrix

Actual/Prediction	0	1
0	1365	245
1	239	1412

AUC: 0.9116
Accuracy Ratio :0.8204

مقایسه نتایج این چهار روش:

نگاره مقایسه ایی

روش	Accuracy Ratio
Logistic Regression	0.7987
Random Forest	0.8047
XGBoost	0.8060
Stacking	0.8204

وجه تمایز مدل ترکیبی Stacking ، میزان دقت بالای این روش در پیش بینی ریسک ورشکستگی مالی شرکت های بورسی و فرابورسی می باشد.

۶. بحث و نتیجه گیری

همان گونه که ملاحظه شد در این پژوهش میزان دقت در پیش بینی ریسک ورشکستگی مالی شرکت های بورسی و فرابورسی به روش Logistic Regression برابر است با 0.7987، به روش Random Forest برابر است با 0.8047، به روش XGBoost برابر است با 0.8060، به روش Stacking برابر است با 0.8204، بنابراین روش ترکیبی stacking به نتیجه ی به مراتب بهتری رسیده است. تکنیک stacking به نوعی حالت بهبود یافته شده تکنیک voting می باشد. با بهره گیری از ۳ مدل پایه (Random Forest، XGBoost و Logistic Regression) و ویژگی های بهینه یک مدل ترکیبی (Stacking) مورد آموزش قرار گرفته است که نتایج حاصل میزان دقت بالای مدل ترکیبی Stacking در پیش بینی ریسک ورشکستگی مالی شرکت های بورسی و فرابورسی را

تایید می کند. تکنیک Stacking که یکی از قوی ترین تکنیک های یادگیری جمعی می باشد در این گونه مسائل موفقیت بسیار بالایی از خود نشان می دهد.

البته در تجزیه و تحلیل نسبت‌های مالی، محدودیت‌هایی نیز وجود دارد. مفروضاتی که در حین فرآیند تجزیه و تحلیل مالی باید همواره آن‌ها را مد نظر قرار عبارتند از:

- ❖ تورم وجود ندارد.
- ❖ مانده اقلام ترازنامه در طی دوره ثابت هستند.
- ❖ ارقام صورت سود و زیان به طور یکسان در طی دوره واقع شده‌اند.
- ❖ حساب سازی انجام شده است.
- ❖ حساب آرایی انجام شده است.

در صورت مقایسه با سایر شرکت‌ها تمام شرکت‌ها از رویه های یکسان حسابداری استفاده کرده‌اند. برای تعیین اینکه کدام یک از نسبت‌های مالی برای پیش بینی‌ها و ارزیابی‌ها مفید تر بوده و به طور معمول در مطالعات بازار به کار گرفته می‌شوند، هیچ‌گونه ملاک مطلق برای اهمیت متغیرهای تأثیرگذار در آن نسبت‌ها و نیز میزان اهمیت متغیرهای خاص در جهت تأیید یا مخالفت با آمارهای مختلف مورد بررسی، وجود ندارد. تصمیم گیری همواره یک امر فردی بوده و سرمایه گذار با توجه به سلیقه و میل و اشتیاق خود نسبت به یک موضوع خاص، جهت رسیدن به هدف مورد نظر خود تصمیم گیری می‌نماید. طبیعتاً مدیران شرکت هم جهت بهبود روش‌های دستیابی به ثروت تلاش می‌کنند اما آن‌ها هم با توجه به سطح دانش و آگاهی خود در نهایت با توجه به سلیقه خود به اتخاذ تصمیم می‌پردازند.

در پایان پیشنهادهای زیر را می توان برای زمینه های پژوهش های آتی ارائه نمود:

محققین می توانند به جای تکنیک Stacking، از دیگر روش های یادگیری جمعی در این پژوهش استفاده کنند. شاید عملکرد بهتری را در حل این گونه مسائل داشته باشند.

در این پژوهش تکنیک Stacking بر روی داده های متعلق به برخی شرکت ها در بازه ۱۳۸۱ تا ۱۳۹۸ اعمال شده است، محققین می توانند از بازه زمانی دیگر نیز بهره بگیرند.

فهرست منابع

Altman, e i. (1968). "financial ratios, discriminant analysis and prediction of corporate bankruptcy" , journal of finance, 23, pp.589-609.

Beaver, w. (1966) “financial ratios as predictors of failure “, journal of accounting research (supplement), 4, pp.71-102.

Biondo, S.; Ramos, E.; Deiros, M.; Ragué, J. M.; De Oca, J.; Moreno, P.; Farran, L.; Jaurrieta, E. (2000). "Prognostic factors for mortality in left colonic peritonitis: A new scoring system". Journal of the American College of Surgeons. 191 (6): 635–42. doi:10.1016/S1072-7515(00)00758-4. PMID 11129812.

Breiman L (2001). "Random Forests". Machine Learning. 45 (1): 5-32. doi: 10.1023/A:1010933404324. ISSN 0885-6125.

Chen, Tianqi; Guestrin, Carlos (2016). "XGBoost: A Scalable Tree Boosting System". In Krishn apuram, Balaji; Shah, Mohak; Smola, Alexander J.; Aggarwal, Charu C.; Shen, Dou; Rastogi, Rajeev (eds.). Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016. ACM. pp. 785–794.

Cramer, J. S. (2002-12-01). "The Origins of Logistic Regression". Rochester, NY:5. doi: 10.2139/ssrn.360300.

Gitman I.J. (1998), Principle of Managerial Finance, *Working paper, New York, Harper Collins College.*

Hastie, Trevor. (2001). The elements of statistical learning: data mining, inference, and prediction: with 200 full-color illustrations. Tibshirani, Robert., Friedman, J. H. (Jerome H.). New York: Springer. ISBN 0-387-95284-5. OCLC 46809224.

Ho, Tin Kam (1995). Random Decision Forests (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282. Archived from the original (PDF) on 17 April 2016. Retrieved 5 June 2016.

Ho TK (1998). "The Random Subspace Method for Constructing Decision Forests" (PDF). IEEE Transactions on Pattern Analysis and Machine Intelligence.20(8):832–844.doi:10.1109/34.709601

Kologlu, M.; Elker, D.; Altun, H.; Sayek, I. (2001). "Validation of MPI and PIA II in two different groups of patients with secondary peritonitis". Hepato-Gastroenterology. 48 (37): 147–51. PMID 11268952.

Le Gall, J. R.; Lemeshow, S.; Saulnier, F. (1993). "A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multi center study". JAMA. 270 (24):2957–63. Doi:10.1001/jama.1993.03510240069035. PMID 8254858.

Marshall, J. C.; Cook, D. J.; Christou, N. V.; Bernard, G. R.; Sprung, C. L.; Sibbald, W. J. (1995). "Multiple organ dysfunction score: A reliable descriptor of a complex clinicaloutcome". Critical Care Medicine. 23 (10): 1638–52. doi:10. 1097/00003246-19951000000007.PMID 7587228.

Piryonesi, S. M.; El-Diraby, T. E. (2020) [Published online: December 21, 2019]. "Data Analytics in Asset Management: Cost-Effective Prediction of the Pavement Condition Index". Journal of Infrastructure Systems. 26 (1). doi: 10.1061/(ASCE)IS.

Piryonesi, S. Madeh; El-Diraby, Tamer E. (2020) "Role of Data Analytics in Infrastructure Asset Management: Overcoming Data Size & Quality Problems". Journal of Transportation Engineer ing, Part B:Pavements.146(2):04020022.Doi: 10.1061/jpeodx.0000175.ISSN 2573-5438

Rokach, L. (2010). "Ensemble-based classifiers". Artificial Intelligence Review.