فصلنامه سیستم‌های پردازشی و ارتباطی چندرسانه‌ای هوشمند
Intelligent Multimedia Processing and Communication Systems (IMPCS)

Summer 2022, 3 (2), 1-9
DOR: 20.1001.1.27832570.1401.3.2.1.1

Received: 21 May 2022
Accepted: 07 July 2022

مقاله پژوهشی

Designing a Hybrid Model for Classification of Imbalanced Data in the Field of Third Party Insurance

Mahnaz Manteqipour[1*], Parisa Rahimkhani[2]

1.PhD., Applied Mathematics (Operation research), Azarbaijan Shahid Madani University, Tabriz, Iran.
(Corresponding Author) mahteqipoor@gmail.com
2. PhD., Numerical Analysis, Alzahra University, Tehran, Iran.
Rahimkhani.parisa@gmail.com

Abstract

Introduction: The major part of Iran's insurance industry portfolio is the field of compulsory civil liability insurance of motor vehicle owners against third parties. Therefore, detecting the behavior of this insurance field will be effective in order to provide better services to the customers of the insurance industry. Predicting the claim rates for insurance policies, based on the features saved for each insurance policy, is one of the problems of the insurance industry that can be solved with the help of data mining techniques. Insurance is designed using the law of large numbers. In simpler words, a sufficient number of insurance policies are issued, and a small part of this number of insurance policies deal with claims. From the sum of the issued insurance premiums, the cost of claims will be compensated. Therefore, the insurance industry is faced with imbalanced data. The imbalances of insurance industry data causes many challenges in data classification. In the field of third-party insurance and in the data set of this research, there are 14 features for every policies and the data imbalance ratio is 1 to 0.0092, which is considered severe imbalanced.

Method: In this research, we deal with the classification of severe imbalanced data in the field of third party insurance. To overcome the problem of imbalanced data, two hybrid models with different architectures based on 5 basic Gaussian Bayes models, support vectors, logistic regression, decision tree and nearest neighbor are designed. First proposed hybrid model is using random sampling from whole dataset and applying a resampling method for classification and second one selects samples from each labels separately and apply a classification model on the whole selected data. The results of these models are compared.

Results: The obtained results show that the proposed hybrid models can predict the occurrence or non-occurrence of traffic accidents better than other data mining algorithms. The popular measures such as precisions and recalls of two proposed hybrid models show that second hybrid model has higher performance. And in ensemble phase, the number of models in simple voting as a hyper parameter can be adjusted based on the company's strategy. Also, the use of decision tree to ensemble basic models to build a combined model provides better results than simple voting of basic models.

Discussion: To do more research on the problem of imbalance data classification more complicated resampling data algorithms could be applied and the results be compared.

Keywords: hybrid model, imbalance data, data mining, third party insurance.

فصلنامه سیستم‌های پردازشی و ارتباطی چندرسانه‌ای هوشمند
Intelligent Multimedia Processing and Communication Systems (IMPCS)
انجمن علمی تجارت الکترونیکی ایران

طراحی مدل ترکیبی برای طبقه‌بندی داده‌های نامتوازن در رشته بیمه شخص ثالث

تاریخ دریافت:۱۴۰۱/۰۲/۳۱ سال سوم، تابستان ۱۴۰۱
تاریخ پذیرش:۱۴۰۱/۰۴/۱۶ شمارهٔ دوم، صص: ۱-۹

مهناز منطقی‌پور[1]*، پریسا رحیم‌خانی[2]

۱. دکترا، ریاضی کاربردی(تحقیق در عملیات)، دانشگاه شهید مدنی آذربایجان، تبریز، ایران.(نویسنده مسئول) mahteqipoor@gmail.com
۲. دکترا، آنالیز عددی، دانشگاه الزهرا، تهران، ایران Rahimkhani.parisa@gmail.com

چکیده

بخش عمده پورتفوی صنعت بیمه کشور ایران را رشته بیمه اجباری مسئولیت مدنی دارندگان وسایل نقلیه موتوری زمینی در مقابل اشخاص ثالث، تشکیل‌داده‌است. توانایی پیش‌بینی وقوع و یا عدم وقوع خسارت به‌ویژه خسارت‌های جانی نه‌تنها برای شرکت‌های بیمه بلکه برای تصمیم‌گیرندگان در حوزه‌های افزایش امنیت جاده‌ها اهمیت بسیاری دارد. به منظور پیش‌بینی برچسب وقوع یا عدم وقوع خسارت از روش‌های طبقه‌بندی استفاده‌می‌شود که درواقع یک مسئله طبقه‌بندی نامتوازن است. این نامتوازن بودن شدید، ناشی از ماهیت کسب و کار بیمه است. نامتوازن بودن داده‌های صنعت بیمه باعث ایجاد چالش‌های بسیاری در تجزیه و تحلیل داده‌های مربوطه می‌شود. در این پژوهش، ما به طبقه‌بندی داده‌های نامتوازن بیمه شخص ثالث در یک شرکت بیمه معتبر می‌پردازیم. در این راستا دو روش ترکیبی برای رفع مشکل نامتوازن بودن داده‌ها براساس ۵ مدل پایه گاوسین بیز، بردارهای پشتیبان، لجستیک رگرسیون، درخت تصمیم، نزدیکترین همسایگی به منظور طبقه‌بندی مؤثرتر داده‌های مربوطه ارائه‌می‌شود. نتایج به‌دست‌آمده نشان‌می‌دهد که مدل‌های ترکیبی ارائه‌شده بهتر از سایر الگوریتم‌های داده‌کاوی برای داده‌های مربوطه جواب‌می‌دهند و استفاده از درخت تصمیم در تجمیع مدل‌های پایه برای ساخت مدل ترکیبی نسبت به رأی‌گیری ساده مدل‌ها نتایج بهتری ارائه‌می‌کند. همچنین ابر پارامتر تعداد مدل‌های لازم در رأی‌گیری بر اساس استراتژی شرکت قابل‌تنظیم است. تعداد ویژگی‌های ثبت‌شده از بیمه نامه‌ها در شرکت‌های بیمه محدود است با تکمیل این ویژگی‌ها به ویژه اضافه‌شدن سوابق رانندگی و سایر ویژگی‌های فردی می‌توان به مدل بهتری دست‌یافت.

واژه‌های کلیدی: مدل ترکیبی، داده‌های نامتوازن، داده‌کاوی، بیمه شخص ثالث.

۱.راهبر میزداده‌کاوی، گروه بیمه الکترونیک، پژوهشکده بیمه، ایران.
۲.پژوهشگر میزداده‌کاوی، گروه بیمه الکترونیک، پژوهشکده بیمه، ایران.

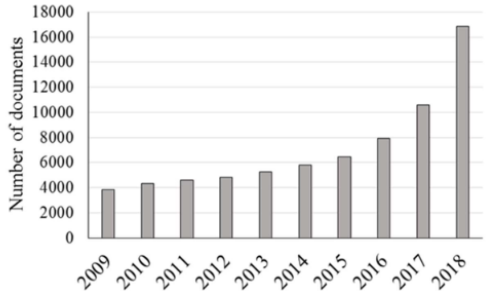
۱. مقدمه

با پیشرفت‌های تکنولوژی و امکان ذخیره‌سازی حجم عظیم اطلاعات، استفاده از داده‌کاوی به عنوان یک فرآیند کشف الگوهای گوناگون در داده‌ها در حال افزایش است[1] و به یک موضوع مهم در حوزه‌های مختلف از جمله صنعت بیمه تبدیل‌شده‌است. رشته بیمه اجباری مسئولیت مدنی دارندگان وسایل نقلیه موتوری زمینی در مقابل اشخاص ثالث که در این مقاله به اختصار رشته بیمه شخص ثالث به آن اطلاق-می‌شود، به لحاظ تبعات اجتماعی و اقتصادی نه‌تنها برای صنعت بیمه که برای آحاد جامعه بسیار با اهمیت است. بر اساس سالنامه آماری ۱۳۹۹ صنعت بیمه، این رشته حدود ۳۳/۶ درصد از کل حق بیمه‌های دریافتی و ۴۰ درصد از کل خسارت‌های پرداختی صنعت بیمه و ضریب خسارت ۱۰۸,۸ درصد را به خود اختصاص‌داده‌است. ازاین‌رو نرخ‌گذاری هوشمند و منصفانه در این رشته بیمه‌ای حائز اهمیت است. از آنجا که داده‌کاوی به کاربردها و اهداف موردنظر محقق، مدیر و یا ناظر توجه‌دارد، تشریح کاربردها و به‌کارگیری آن در تصمیم‌سازی‌های صنعت بیمه مفید و ضروری است.

مهمترین چالش مواجهه با مسئله مدلسازی داده‌های صنعت بیمه، نامتوازن بودن داده‌ها است. داده‌هایی که تعداد اعضای کلاس‌های آن متوازن نیست را مسئله طبقه‌بندی نامتوازن گویند [2]. مسائل طبقه‌بندی بر اساس برچسب ادعای خسارت در صنعت بیمه با توجه به ماهیت کسب وکار بیمه بازرگانی منجر به مسائل طبقه‌بندی نامتوازن می‌گردند. اغلب الگوریتم‌های داده‌کاوی در مواجهه با داده‌های نامتوازن عملکرد خود را ازدست‌می‌دهند و تمام داده‌ها را متعلق به کلاس با برچسب اکثریت پیش‌بینی می‌کنند. از این رو چندین راهکار برای حل این مسائل در مقالات متعدد ارائه‌شده‌است.

به منظور طبقه‌بندی داده‌های نامتوازن، دو رویکرد اصلی وجوددارد: ۱- روش‌های نمونه‌گیری مجدد ۲- روش‌های حساسیت هزینه. روش‌های نمونه‌گیری مجدد به‌دنبال دستیابی به مجموعه داده‌های متوازن هستند و روش‌های حساسیت هزینه، حساسیت تابع زیان مسئله طبقه‌بندی را برای خطای کلاس اقلیت بالامی‌برند تا مینیمم‌سازی تابع زیان منجر به پیش‌بینی همه نمونه‌ها به کلاس حداکثر نشود.

مدل‌های مرسوم یادگیری ماشین با گذر زمان توسعه‌یافته‌اند و مدل‌های ترکیبی و گروهی خلق‌شده‌اند. جنگل تصادفی به‌واقع مدلی ترکیبی است که چندین درخت را ترکیب‌می‌نماید. این روش‌ها به منظور افزایش قدرت مدل‌ها از جنبه‌های محاسباتی، عملکرد، استواری و قدرت مدل‌ها خلق شده‌اند. در مدل‌های ترکیبی، خروجی یک مدل جهت بهبود به مدل بعدی ارسال‌می‌شود و مدل قوی‌تر به این شکل ساخته‌می‌شود و در مدل‌های گروهی، مدل‌های مختلف به صورت مجزا ساخته‌می‌شوند و با استفاده از روش‌های متفاوت خروجی مدل گروهی، تعیین‌می‌شود. در دهه گذشته استفاده از مدل‌های گروهی و ترکیبی به‌سرعت افزایش‌یافته است. نمودار زیر تعداد متون منتشرشده در مورد این مدل‌ها را نمایش-می‌دهد [3].



شکل ۱: افزایش تعداد مقالات مدل‌های ترکیبی در سال‌های اخیر

یک روش ترکیبی در مقاله [4] ارائه‌شد که در آن عدم توازن داده‌ها با استفاده از ترکیب روش کا-معکوس نزدیکترین همسایه و روش ماشین بردار پشتیبان مدیریت شده‌است. نویسندگان در مقاله [5] یک روش ترکیبی بر اساس روش SMOTE، اعتبارسنجی متقابل و جنگل تصادفی برای کشف تقلب در زمینه بیمه درمان ارائه‌دادند. نویسندگان در مقاله [6] روی بهینه‌سازی یادگیرنده برای کشف تقلب در رشته بیمه درمان با استفاده از یک روش نمونه‌گیری مجدد تمرکزکردند.

۲. مسئله پژوهش

هدف این مقاله ساختن مدلی برای پیش بینی خسارت‌های جانی رشته بیمه شخص ثالث است. این نوع از خسارت‌ها فراوانی اندکی دارند و در عین حال به دلیل آنکه منجر به پرداخت دیه می‌گردند هزینه اقتصادی بالایی ایجادمی‌کنند. علاوه بر هزینه‌های مالی، هزینه‌های جانی غیر قابل جبران هستند و تبعات بسیاری برای مردم جامعه ایجادمی‌کنند. از این رو پیش بینی این نوع از خسارتها برای تصمیم گیرندگان بسیار مفید خواهد بود.

در این پژوهش، پس از استفاده از روش‌های نمونه‌گیری مجدد به منظور حل چالش عدم توازن داده‌ها، مدل‌های درخت تصمیم، ماشین بردار، رگرسیون لجستیک، نزدیکترین کا-همسایگی، مدل گاوسین بیز و جنگل تصادفی برای پیش‌بینی برچسب وقوع و یا عدم وقوع خسارت مورد استفاده قرارگرفتند و همگی عملکرد ضعیفی داشتند. این مدل‌ها در مرحله تست، عمدتاً نمونه‌ها را متعلق به کلاس اکثریت یعنی کلاس بدون خسارت پیش‌بینی‌کردند. از این رو به ساختن مدل‌های ترکیبی روی آوردیم. مدل‌های ترکیبی با هدف تقویت نتایج مدل‌های ساده‌تر ایجادشده‌اند. ایده مدل‌های ترکیبی آن است که از چندین مدل پیش بینی کننده به عنوان مدل پایه به منظور دستیابی به مدلی با عملکرد بهتر استفاده‌می‌شود و مدل نهایی با استفاده از روش‌های متفاوت، جهت تجمیع نتایج مدل‌های پایه، ارائه‌می‌گردد. در این مقاله با تغییر در ساختار مدل ترکیبی به‌دنبال مدلی با نتایج قوی‌تر بوده‌ایم.

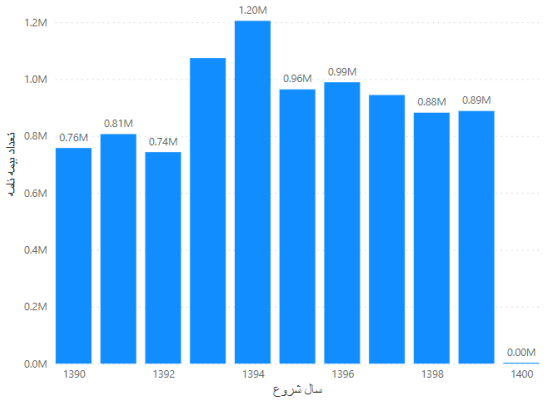
۳. داده‌های تحقیق

داده‌های این پژوهش مربوط به سال‌های ۱۳۹۰ الی ۱۴۰۰ در رشته بیمه شخص ثالث می‌باشند که متعلق به یک شرکت بیمه معتبر در ایران است.

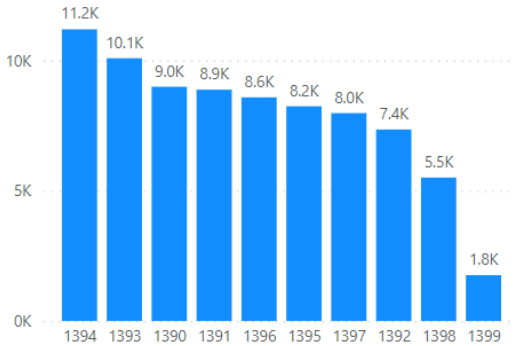
داده‌ها در قالب یک فایل پشتیبان پایگاه داده SQL server 2019 دریافت‌شده‌اند و پس از توصیف و بررسی، بصری‌سازی داده‌ها انجام‌شده‌است. داده‌های پرت حذف‌شده‌اند، پاکسازی داده‌ها انجام‌شده‌است و داده‌ها جهت به‌کارگیری الگوریتم‌های یادگیری ماشین آماده‌شده‌اند.

داده‌ها شامل ۹۲۸۳۳۱۶ بیمه‌نامه شخص ثالث و خسارت‌های مربوطه می‌باشند. تعداد بیمه‌نامه‌ها و تعداد خسارت‌ها به تفکیک سال شروع بیمه نامه مطابق شکل ۲ و شکل ۳ می‌باشد. این داده‌ها ابتدای سال ۱۴۰۰ از پایگاه داده‌های شرکت استخراج‌شده‌است و بنابراین تعداد کمی از بیمه‌نامه‌ها متعلق به سال ۱۴۰۰ می‌باشند.

نرخ نامتوازن بودن داده‌های این پژوهش ۱ به ۰,۰۰۹۲ است. به عبارتی به ازای هر ۱۰۰۰ بیمه‌نامه ۹/۲ بیمه‌نامه با خسارت جانی مواجه‌شده‌اند.



شکل ۲: تعداد بیمه‌نامه‌ها به تفکیک سال شروع بیمه‌نامه



شکل ۳: تعداد خسارت‌های جانی به تفکیک سال شروع

به دلیل تفاوت رفتاری مشاهده‌شده در گروه‌های خودرو محاسبات این مقاله بر داده‌های خودروهای سواری و منحصراً بر خسارت‌های جانی انجام‌شده‌است.

۴. مدل‌های ترکیبی

پس از پاکسازی و آماده‌سازی داده‌ها و شناخت داده‌ها، تلاش‌نمودیم با استفاده از الگوریتم‌های یادگیری ماشین و داده‌کاوی مدلی برای پیش بینی وقوع خسارت‌ها بسازیم. از آنجا که ظرفیت ویژگی‌ها محدود است، به جای پیش‌بینی نرخ خسارت‌ها ترجیح‌دادیم وقوع یا عدم وقوع خسارت‌ها را بررسی‌کنیم.

ویژگی‌های به‌کاررفته در مدل در جدول زیر آورده‌شده‌است:

جدول ۱: ویژگی‌های بیمه‌نامه‌ها

شمارش	نام متغیر
۱	یت
۲	ذار
۳	ذار
۴	
۵	
۶	
۷	
۸	
۹	
۱۰	رد
۱۱	ت
۱۲	ت
۱۳	ت
۱۴	در

۱۴ ویژگی ذکر شده در جدول ۱: ویژگی‌های بیمه‌نامه‌ها را به منظور پیش‌بینی وقوع یا عدم وقوع خسارت‌های جانی به‌کارگرفته‌ایم. بنابراین قصدداریم برچسبی که مقادیر یک و صفر اختیار می‌کند، پیش‌بینی‌کنیم.

پس از مرحله پاکسازی داده‌ها[7] ، مهمترین چالش این مسئله مواجهه با نامتوازن بودن داده‌ها است.

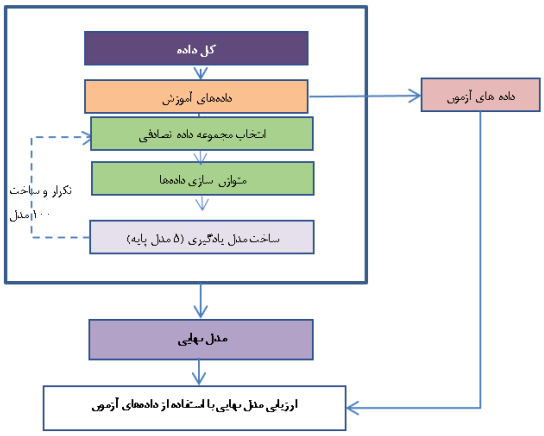
برای طبقه‌بندی داده‌های مذکور ابتدا به پیاده‌سازی الگوریتم درخت تصمیم، جنگل تصادفی، رگرسیون لجستیک، نایو بیز نمودیم [8] . نامتوازن بودن داده‌ها منجر به پیش‌بینی برچسب ۰ به معنی عدم وقوع خسارت برای تمام بیمه‌نامه‌ها شد که این نتیجه ناشی از ناتوانی مدل‌ها در رویارویی با داده‌های نامتوازن است. زیرا مدل‌ها تلاش بر بیشینه نمودن دقت مدل دارند که با پیش‌بینی برچسب کلاس اکثریت برای تمام نمونه‌ها محقق‌می‌شود.

سپس از تکنیک‌های مختلف روش‌های نمونه‌گیری مجدد و روش حساسیت هزینه استفاده‌نمودیم. روش حساسیت هزینه عملکرد مطلوبی نداشت. در روش نمونه‌گیری مجدد ابتدا این روش‌ها را بر داده‌ها به‌کار-بردیم و سپس داده‌ها را به دو بخش آموزش و تست تقسیم‌نمودیم. مدل‌های حاصل بر داده‌های تست دارای عملکرد ۹۹ درصد به بالا داشتند. اما از آنجا که روش‌های نمونه‌گیری مجدد داده‌ها را تغییر-می‌دهند بخشی از داده‌های آموزش در داده‌های تست موجودند و بنابراین تصمیم‌گرفتیم داده‌های آموزش و تست را قبل از به‌کارگیری الگوریتم های نمونه‌گیری مجدد انتخاب‌نمائیم و روش‌های نمونه‌گیری مجدد را صرفاً بر داده‌های آموزش اعمال‌کنیم. با این اقدام نتایج مدل بر داده‌های تست به‌شدت افت‌کرد. از این رو بر آن شدیم تا به منظور بهبود

عملکرد مدل‌ها، این مدل‌ها را به صورت ترکیبی محاسبه و دقت را اندازه‌گیری کنیم.

جهت ساخت مدل ترکیبی از ۵ مدل پایه گاوسین بیز، بردارهای پشتیبان، لجستیک رگرسیون، درخت تصمیم، نزدیکترین همسایگی ۱۰۰ مدل ساخته‌ایم و دقت مدل‌ها را محاسبه‌کرده‌ایم.

۱،۴. ساخت مدل ترکیبی (۱) انتخاب مجموعه داده تصادفی



شکل ۴: معماری مدل ترکیبی ۱

با انتخاب روش Smote به عنوان روش نمونه‌گیری مجدد[2] ، در مدل ترکیبی ۱، ماتریس طبقه‌بندی و معیارهای ارزیابی مدل حاصل با درصد رأی‌گیری ۵۰ ٪ به صورت زیر است:

```
[[857896   7950]
[     0      0]]
```

	precision	recall	f1-score	support
0	0.990818	1.000000	0.995388	857896
1	0.000000	0.000000	0.000000	7950
accuracy			0.990818	865846
macro avg	0.495409	0.500000	0.497694	865846
weighted avg	0.981721	0.990818	0.986249	865846

بر اساس نتایج حاصله بهبودی در عملکرد مدل نسبت به مدل جنگل تصادفی و درخت تصمیم و الگوریتم‌هایی که ذکرشد، حاصل‌نشده‌است. حتی می‌توان ادعاکرد عملکرد مدل ضعیف‌تر نیز شده‌است.

با بررسی عملکرد الگوریتم‌ها متوجه‌شدیم عملکرد سه الگوریتم از دیگر الگوریتم‌ها بهتر است در شکل زیر عملکرد پنج الگوریتم قابل-مشاهده است.

پس از ساخت ۵۰۰ مدل، عملکرد مدل‌ها با استفاده از معیار F1-score به صورت زیر بوده‌است:

شکل ۵: مقایسه عملکرد مدل ها بر اساس معیار F1

همان‌طور که مشاهده می‌شود ضعیفترین پیش‌بینی‌ها متعلق به روش گاوسین بیز است. روش نزدیکترین همسایگی، پیش‌بینی‌هایی بهتر از پیش‌بینی‌های تصادفی دارد و به ترتیب سه مدل رگرسیون لجستیک، بردار پشتیبان و درخت تصمیم بهترین عملکرد را داشته‌اند. بنابراین مدل ترکیبی را با استفاده از این سه مدل می‌سازیم. از این رو مجدد مدل را بررسی می‌کنیم و این بار مدل‌های پایه را شامل این سه مدل در نظر می‌گیریم.

عملکرد مدل ترکیبی ۱ با استفاده از سه مدل پایه، در صد مرتبه نمونه گیری منجر به ۳۰۰ مدل شده است که ترکیب آنها با درصد رای گیری ۵۷ درصد، عملکرد زیر را داشته است:

```
[[848001   7976]
[ 9760    109]]
```

	precision	recall	f1-score	support
0	0.990682	0.988622	0.989651	857761
1	0.011045	0.013482	0.012142	8085
accuracy			0.979516	865846
macro avg	0.500863	0.501052	0.500896	865846
weighted avg	0.981534	0.979516	0.980523	865846

مطابق ماتریس طبقه‌بندی و معیارهای ارزیابی مدل نمایش‌داده‌شده فوق، می‌توان گفت اندکی نتایج بهتر است اما همچنان ضعیف است. زیرا تنها ۱/۳٪ از کل خسارت‌ها شناسایی‌شده‌اند و از کل خسارت‌های پیش-بینی شده توسط مدل ۱/۱ ٪ درست بوده‌اند.

عملکرد الگوریتم به ازای انتخاب مدل SVC به عنوان مدل پایه با درصد رای‌گیری ۵۲٪ به صورت زیر بوده‌است:

```
[[580978   5289]
[276783   2796]]
```

	precision	recall	f1-score	support
0	0.990979	0.677319	0.804663	857761
1	0.010001	0.345826	0.019439	8085
accuracy			0.674224	865846
macro avg	0.500490	0.511572	0.412051	865846
weighted avg	0.981818	0.674224	0.797331	865846

مشاهده‌می‌شود که حدود ۳۴,۵٪ از یک‌های شناسایی‌شده درست بوده‌اند. در عین حال ۵۲۸۹ بیمه‌نامه بدون خسارت به عنوان بیمه نامه‌های دارای خسارت جانی شناسایی‌شده‌اند. بنابراین به‌کارگیری این مدل با کاهش ۳۲,۳ درصدی پرتفو منجر به کاهش ۳۴,۵٪ از

مهناز منطقی پور، پریسا رحیم خانی

خسارت‌های جانی می‌شود که می‌تواند برای شرکت سودآور باشد. اما کاهش ۳۲,۲ درصدی پرتفو برای کمتر شرکتی قابل‌قبول است.

سایر ترکیب‌های مدل‌های یادگیری از ۵ مدل به عنوان مدل‌های پایه عملکرد ضعیفتری داشته‌اند. از ذکر نتایج آن صرف نظر می‌کنیم. در مدل ترکیبی دوم نحوه انتخاب نمونه داده‌ها را در هر تکرار تغییرداده‌ایم که در بخش بعد توضیح داده‌شده‌است.

۴. ۲. مدل ترکیبی ۲ (با نمونه‌گیری مجزا از هر کلاس)

مراحل انجام کار به صورت زیر طراحی شده‌است:

تبدیل متغیرهای اسمی به متغیرهای عددی

جداسازی داده‌ها به داده‌های آزمون و تست

هم مقیاس نمودن داده‌ها بدون استفاده از داده‌های آزمون و صرفاً تبدیل داده‌های آزمون با استفاده از الگوی هم مقیاسی داده‌های آموزش

جداسازی داده‌های آزمون بر اساس برچسب‌های دارای خسارت جانی، ۱ و بدون خسارت جانی ۰

مراحل ۵ و ۷ ، ۱۰۰ مرتبه تکرارشده‌اند:

انتخاب ۱۰۰۰۰ نمونه تصادفی با برچسب ۰ و انتخاب ۱۰۰۰۰ نمونه تصادفی با برچسب ۱

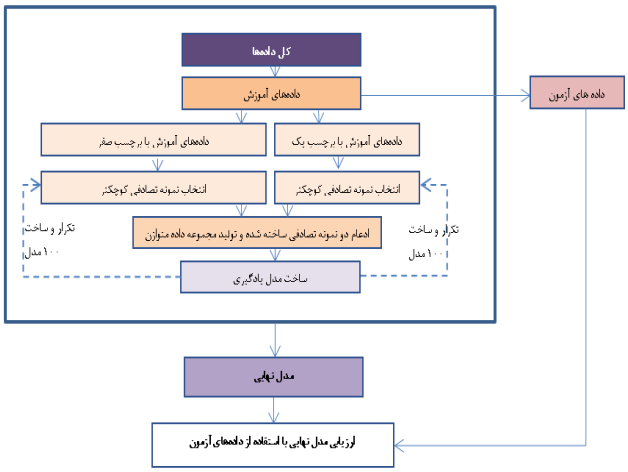
ادغام نمونه‌های برچسب‌ها و ایجاد مجموعه داده نمونه

محاسبه مدل مربوطه بر اساس مجموعه داده نمونه ۲۰۰۰۰ تایی

خروجی مدل نهایی که بر اساس رأی‌گیری تمام مدل‌های ساخته‌شده در مرحله ۷ حاصل‌می‌شود.

به عبارتی با فرآیند فوق اگر ۵ مدل یادگیری به عنوان مدل‌های پایه انتخاب‌شوند، در پایان مرحله ۷، الگوریتم طراحی شده ۵۰۰ مدل محاسبه‌شده‌است. خروجی مدل نهایی از رأی‌گیری نتایج تمام مدل‌ها حاصل خواهدشد.

معماری این مدل ترکیبی در شکل ۶ خلاصه‌شده‌است:



شکل ۶: معماری مدل ترکیبی ۲

به منظور انتخاب مجموعه داده از هر کلاس صفر و یک، برای ساخت هر مدل ۱۰۰۰۰ نمونه انتخاب‌شده‌است و بنابراین مجموعه داده برای ساخت هر مدل دارای ۲۰۰۰۰ نمونه می‌باشد. برای هر مجموعه داده پنج مدل یادگیری اجرامی‌شود. مدل‌های یادگیری انتخاب‌شده شامل درخت تصمیم، رگرسیون لجستیک، نایو بیز گاوسین، بردارهای پشتیبان، کا- نزدیکترین همسایگی می‌باشند. بنابراین پس از پایان برنامه ۵۰۰ مدل خواهیم‌داشت. در ادامه نتایج اجرای این برنامه را شرح خواهیم‌داد.

جدول ۲: تعداد نمونه‌ها به تفکیک برچسب‌های صفر و یک

مجموعه داده	تعداد نمونه‌ها	تعداد نمونه‌های هر کلاس	
کل داده‌ها	۴۳۲۹۲۳۰	۰	۴،۲۸۹،۵۹۲
		۱	۳۹،۶۳۸
داده‌های آموزش	۳۴۶۳۳۸۴	۰	۳۴۳۱۸۳۱
		۱	۳۱۵۵۳
داده‌های آزمون	۸۶۵۸۴۶	۰	۸۵۷،۷۶۱
		۱	۸۰۸۵

ماتریس طبقه‌بندی مدل حاصل از رای‌گیری این پنج مدل با نسبت رای گیری ۵۶٪ به صورت زیر است. (به این معنی که اگر بیش از ۵۶٪ از مدل‌ها برچسب یک را پیش‌بینی‌کرده‌اند، پیش‌بینی مدل نهایی برچسب ۱ است):

```
                                [[641593   5857]
                                 [216168   2228]]
              precision    recall  f1-score   support

           0   0.990954 0.747986 0.852496    857761
           1   0.010202 0.275572 0.019675      8085

    accuracy                      0.743574    865846
   macro avg   0.500578 0.511779 0.436085    865846
weighted avg   0.981796 0.743574 0.844719    865846
```

که نتایج فوق به این معنی است که ۲۷٪ از کل خسارت‌های جانی به‌درستی دارای خسارت جانی پیش‌بینی‌شده‌اند و۷۴٪ از بدون خسارت‌ها نیز به‌درستی بدون خسارت پیش بینی‌شده‌اند. با پیاده سازی این روش برای صدور بیمه‌نامه‌ها با کاهش ۲۵/۲ درصدی پرتفو، خسارت‌ها ۲۷/۵٪ کاهش‌خواهندیافت. این نتایج در مقایسه با مدل‌های ساده غیرترکیبی که نمونه‌گیری مجدد Smote انجام‌شده‌بود عملکرد بهتری است. حتی از ۱۰۰ مدل ترکیبی SVC نیز عملکرد بهتر است.

با کاهش درصد رأی‌گیری به ۵۵٪ نتایج به صورت زیر است. اختلاف رأی‌گیری ۵۵٪ و ۵۶٪، به اندازه اثرگذاری ۵ مدل از ۵۰۰ مدل است. در این مدل، ۷۱ درصد از خسارت‌ها به‌درستی شناسایی‌شده‌اند و پرتفو ۶۹,۴ کوچک‌شده‌است. در شرایطی که استراتژی شرکت، کوچک‌شدن پرتفوی بیمه‌نامه‌های رشته بیمه شخص ثالث و کاهش ضریب خسارت باشد به‌کارگیری چنین مدلی مناسب است.

```
                                [[260470   2315]
                                 [597291   5770]]
              precision    recall f1-score   support

           0   0.303663 0.991191 0.464898    262785
           1   0.713667 0.009568 0.018883    603061

    accuracy                      0.307491    865846
   macro avg   0.508665 0.500379 0.241890    865846
weighted avg   0.589231 0.307491 0.154249    865846
```

به عنوان مثال شرکتی که در رشته بیمه شخص ثالث عملکرد خوبی ندارد، نهاد ناظر می‌تواند آن شرکت را صرفاً برای صدور بیمه‌نامه‌های بدون خسارت پیش‌بینی شده توسط مدل فوق مجاز بداند.

بنابراین به نظر می‌رسد متوازن‌سازی داده‌ها با روش تصادفی ساده به شکل الگوریتم فوق کارایی بهتری دارد.

۱،۲،۴. تغییر تعداد نمونه های هر برچسب در مدل ترکیبی۲

به منظور اصلاح عملکرد مدل فوق در هر بار نمونه‌گیری از کلاس یک، ۲۰۰۰۰ نمونه و از کلاس صفر ، ۳۵۰۰۰ نمونه انتخاب‌می‌کنیم. اجرای این برنامه ۳۵۰۱۳۲,۱۰۹۳۷۵ ثانیه معادل ۹۷,۲۵ ساعت برابر با حدود ۴ شبانه روز به طول انجامیده‌است.

نتایج مدل بر داده‌های آزمون، به صورت زیر است:

به ازای رأی‌گیری ۵۰ درصدی (رأی‌گیری با سهم ۲۵۰ مدل):

```
                                [[850010   7862]
                                 [ 7838    136]]
              precision    recall f1-score   support

           0   0.990863 0.990835 0.990849    857872
           1   0.017004 0.017055 0.017030      7974

    accuracy                      0.981867    865846
   macro avg   0.503934 0.503945 0.503940    865846
weighted avg   0.981894 0.981867 0.981881    865846
```

رأی‌گیری ۳۳۰ مدل:

```
                                [[632771 225101]
                                 [ 5175   2799]]
              precision    recall f1-score   support

           0   0.991888 0.737605 0.846053    857872
           1   0.012282 0.351016 0.023733      7974

    accuracy                      0.734045    865846
   macro avg   0.502085 0.544311 0.434893    865846
weighted avg   0.982866 0.734045 0.838480    865846
```

رأی‌گیری ۳۳۵ مدل:

```
                                [[714535 143337]
                                 [ 6196   1778]]
              precision    recall f1-score   support

           0   0.991403 0.832916 0.905275    857872
           1   0.012252 0.222975 0.023228      7974

    accuracy                      0.827298    865846
   macro avg   0.501828 0.527945 0.464252    865846
weighted avg   0.982386 0.827298 0.897152    865846
```

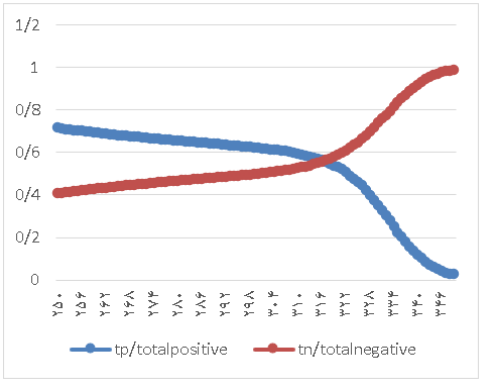
رأی‌گیری ۳۴۰ مدل:

```
                                [[786508 71364]
                                 [ 7047    927]]
              precision    recall f1-score   support

           0   0.991120 0.916813 0.952519    857872
           1   0.012823 0.116253 0.023098      7974

    accuracy                      0.909440    865846
   macro avg   0.501971 0.516533 0.487809    865846
weighted avg   0.982110 0.909440 0.943960    865846
```

جدول ۳: تعداد مدل در رای گیری و نسبت پیشبینی‌ها

	ری	تباه	ت
۱	۳۴۰	۸/۳	۱۱/۶
۲	۳۳۵	۱۶/۷	۲۲/۳
۳	۳۳۰	۲۶/۲	۳۵/۱
۴	۲۵۰	۵۹	۷۱

نکته قابل‌توجه آن است که با افزایش تعداد نمونه‌های هر مدل از ۲۰۰۰۰ نمونه به ۵۵۰۰۰ نمونه عملکرد مدل نهایی بهتر شده‌است. به منظور شناخت مدل به نمودار زیر توجه‌کنید. با افزایش تعداد مدل‌ها در رای‌گیری برای تشخیص خسارت‌ها، تعداد صفرهایی که به‌درستی شناخته‌می‌شوند، افزایش و تعداد خسارت‌های تشخیص‌داده‌شده کاهش-می‌یابد.



شکل ۷: نسبت صفرهای درست و یک‌های درست به تفیک تعداد مدل‌های رأی‌گیری

شکل فوق بررسی عملکرد مدل نهایی با پارامتر تعداد رای‌گیری، نمودار آبی نسبت تعداد پیشبینی‌های یک درست به تعداد اعضای کلاس یک‌ها و نمودار قرمز نسبت تعداد پیش‌بینی‌های صفر درست به تعداد اعضای کلاس صفر را نشان می‌دهد.

در شکل، مشاهده‌می‌شود که به منظور انتخاب تعداد مدل‌های مورد نیاز رأی‌گیری، لازم است استراتژی شرکت لحاظ‌شود. اگر استراتژی شرکت کاهش خسارت پرداختی است می‌تواند از تعداد مدل‌های کمتری برای رأی‌گیری استفاده‌کند و در شرایطی که بزرگ بودن پورتفو برای شرکت در اولویت است از تعداد مدل‌های کمتری برای رأی‌گیری استفاده-می‌شود. به عنوان مثال؛ انتخاب تعداد ۳۳۵ مدل برای رأی‌گیری، ۲۲ درصد از بیمه‌نامه‌های منجر به خسارت را به‌درستی پیش‌بینی‌می‌کند و

این موضوع با هزینه شناسایی نادرست ۱۶٪ از بیمه‌نامه‌های بدون خسارت به عنوان بیمه‌نامه‌های دارای خسارت همراه است. واضح است که هزینه ۲۲٪ از خسارت‌ها بیشتر از مجموع حق بیمه ۱۶٪ از بیمه‌نامه‌ها است.

بررسی نتایج مدل ترکیبی با کاهش مدل‌ها به کمتر از ۵ مدل پایه اولیه، نتایج بهتری ارائه کرد و عملکرد مدل نهایی کاهش‌یافت.

۲،۲،۴. جایگزینی روش رأی‌گیری مدل ترکیبی۲ با استفاده از مدل‌های یادگیری ماشین

با به کارگیری مدل درخت تصمیم بر خروجی ۵۰۰ مدل، عملکرد مدل حاصل به صورت زیر است:

```
                                [[428947      2]
                                 [ 3788    186]]
              precision    recall  f1-score   support

           0   0.991246  0.999995  0.995602    428949
           1   0.989362  0.046804  0.089380      3974

    accuracy                       0.991246    432923
   macro avg   0.990304  0.523400  0.542491    432923
weighted avg   0.991229  0.991246  0.987283    432923
```

این مدل بر داده‌های تست ساخته‌شده‌است و به نظر می‌رسد استفاده از درخت تصمیم بر خروجی نهایی مدل‌ها می‌تواند عملکرد مدل را بهبود بخشد. ۴/۶٪ از کل خسارت‌ها و ۹۹٪ از صفرها به‌درستی، شناسایی شده‌اند. ۹۸٪ از یک‌های شناسایی‌شده صحیح‌اند. استفاده از این مدل در انتخاب ریسک، با کاهش کمتر از ۱٪ از پرتفو، ۴/۶٪ از خسارت‌ها شناسایی‌می‌شود که منطقی است و برای شرکت‌ها قابل‌پیاده‌سازی می‌باشد.

روش‌های SVC ، جنگل تصادفی وکا- نزدیکترین همسایگی در این رویکرد عملکرد ضعیفتری داشتند.

۵. نتیجه‌گیری

مدل‌های ترکیبی عملکرد بهتری نسبت به الگوریتم‌های جنگل تصادفی، درخت تصمیم، ماشین بردار پشتیبان و رگرسیون لجستیک به همراه روش‌های مختلف متوازن‌سازی در این پژوهش داشته‌اند.

در این مقاله، دو مدل ترکیبی برای تعیین وقوع یا عدم وقوع خسارت در رشته بیمه شخص ثالث ارائه‌شده‌است. در هر دو مدل ترکیبی از پنج روش طبقه‌بندی شامل نزدیکترین همسایگی، نایو بیز، درخت تصمیم، لجستیک رگرسیون و نزدیکترین همسایگی به عنوان مدل پایه استفاده شده‌است. در هر دو مدل ترکیبی پیشنهادی، ۱۰۰ نمونه از داده‌های انتخاب شده‌است و مدل‌های پایه با استفاده از این صد نمونه در مرحله آموزش، آموزش داده‌شده‌اند. بنابراین هر مدل ترکیبی دارای ۵۰۰ مدل طبقه‌بندی پایه است. در مدل ترکیبی اول انتخاب داده‌های نمونه برای ساخت هر مدل پایه به صورت تصادفی انتخاب‌شده‌اند و نمونه‌های انتخاب‌شده با استفاده از روش smote متوازن شده‌اند. در مدل ترکیبی دوم نمونه داده‌ها، صد مرتبه به صورت متوازن از هر برچسب به صورت جداگانه انتخاب ترکیب‌شده‌اند و مدل‌های پایه بر هریک از این صد نمونه داده آموزش‌دیده‌اند. مقایسه نتایج دو مدل ترکیبی مذکور نشان‌می‌دهد مدل دوم عملکرد بهتری دارد. به عبارتی انتخاب داده‌ها به صورت متوازن از هر برچسب مدل‌های بهتری برای پیش‌بینی می‌سازد.

همچنین روش‌های متعددی برای تجمیع مدل‌های پایه وجوددارد. روش رأی‌گیری ساده در این مقاله مورد استفاده قرارگرفته‌است و نتایج نشان‌می‌دهد تعداد مدل‌های پیش‌بینی‌کننده برچسب یک که بر چسب اقلیت است می‌تواند به عنوان ابر پارامتر مدل ترکیبی لحاظ‌شود. این ابر پارامتر بر اساس استراتژی شرکت در حفظ حجم پرتفوی شرکت و یا کاهش ضریب خسارت قابل‌تنظیم است.

لازم به ذکر است ویژگی‌های ثبت‌شده در پایگاه داده‌های یک شرکت بیمه می‌تواند افزایش‌یابد از جمله سوابق تخلفات رانندگی و ویژگی‌هایی مانند شغل راننده می‌تواند در بهبود نتایج مدل‌ها تاثیرچشم‌گیری داشته‌باشد.

۶. مطالعات آتی

به منظور دستیابی به مدلی قویتر برای فائق‌آمدن بر مسئله طبقه‌بندی داده‌های نامتوازن لازم است در ساخت مدل ترکیبی سایر روش‌های نمونه‌گیری مجدد را نیز به‌کارگرفت و یا از ترکیب آن‌ها استفاده‌کرد. همچنین نحوه تجمیع مدل‌های پایه نیز موضوعی است که می‌تواند به تفصیل مورد بحث و بررسی قرارگیرد و از ابتکارات و تکنیک‌های بیشتری برای تجمیع مدل‌ها استفاده‌شود. از جمله استفاده از الگوریتم مورچگان در تجمیع درختان می‌تواند مورد استفاده قرارگیرد [9].

سپاسگزاری

در پایان از مدیران واحد طرح و توسعه شرکت بیمه البرز که ما را در انجام مطالعهٔ حاضر حمایت‌کردند، به‌ویژه جناب آقای عابدینی، سرکار خانم کمالخانی و سرکار خانم ابوطالبی صمیمانه سپاسگزاریم.

Reference

[1] K. P. Murphy., Probabilistic Machine Learning: An Introduction, MIT Press, 2022.

[2] A. Fernández, . S. García and M. Galar, R, Learning from Imbalanced Data Sets, Springer, 2018.

[3] S. Ardabili , A. Mosavi and . A. R. Varkonyi-Koczy, "Advances in Machine Learning Modeling Reviewing Hybrid and Ensemble Methods," *Preprints* , 2019.

[4] G. . G. Sundarkumar and V. Ravi, "A novel hybrid under sampling method for mining unbalanced datasets in banking and insurance," *Engineering Applications of Artificial Intelligence*, vol. 37, p. 368–377, 2015.

[5] S. I. V. Shamitha, S. K. Shamitha and V. Ilango, "A hybrid technique for health insurance fraud detection on highly imbalanced dataset," *International Journal of Innovative technology and exploring engineering (IJITEE)*, vol. 8, no. 11, pp. 2278-3075, 2019.

[6] S. Kotekani and I. Velchamy, "An Effective Data Sampling Procedure for Imbalanced Data learning on

health insurance fraud detection, CIT," *Journal of Computing and Information Technology,*, vol. 28, no. 4, p. 269–285, (2020)..

[7] J. Brownlee, Data Preparation for machine learning, Jason Brownle, 2020.

[8] A. Géron, Hands-on Machine learning with scikit-learn, keras, tensorflow, Beijing, Boston, Farnham, Sebastopol, Tokyo: O'Reilly Media, Inc, 2019.

[9] J. Kozak, Decision Tree and ensemble learning based on ant colony algorithm, Katowice, Poland: Springer, 2019.