دانشگاه صنعتی امیرکبیر
(پلی‌تکنیک تهران)
دانشکده علوم کامپیوتر

سیستم تشخیص تقلب بیمه خودرو

نگارش
نگار یگانه
تینا حسن نسب

استاد درس
مهسا سعادت

شهریور 1404

چکیده

در این پروژه به بررسی کشف تقلب در بیمه وسایل نقلیه می‌پردازدیم که تشخیص تقلب در بیمه خودرو یکی از چالش‌های اساسی در حوزه مالی است که می‌تواند موجب کاهش خسارات اقتصادی برای شرکت‌های بیمه شود. هدف از این پروژه بهبود سیستم‌های شناسایی تقلب از طریق ابزارهای یادگیری ماشین برای تشخیص تقلب است. با شناسایی و پیاده‌سازی تکنیک‌های یادگیری ماشین متناسب با این حوزه، این پروژه قصد دارد دقت و کارایی سیستم‌های شناسایی تقلب را افزایش دهد. نتایج تحقیق شامل معیارهای کمی مانند دقت، صحت، بازخوانی و امتیاز F1 است که با توجه به تأثیر مستقیم آنها در کاهش ادعاهای تقلبی و بهبود فرآیند تصمیم‌گیری در بخش بیمه وسایل نقلیه تحلیل می‌شوند. این پژوهش به توسعه راهکارهای شناسایی تقلب متناسب کمک می‌کند.

واژه‌های کلیدی:

ویژگی، دقت، امتیاز F1، جنگل تصادفی، طبقه بندی، تشخیص تقلب

صفحه

فهرست مطالب

فصل اول

مقدمه

مقدمه

کلاهبرداری بیمه یکی از نگرانی های اصلی صنعت بیمه خودرو است و ادعاهای تقلبی که تقلبی بخش قابل توجهی از کل ادعاهای دریافتی توسط بیمه گران را تشکیل می دهد سالانه میلیون ها دلار برای بیمه گران هزینه دارد. شناسایی و جلوگیری از ادعاهای متقلبانه برای سلامت مالی بیمه گران و رضایت بیمه گذاران آنها بسیار مهم است. یکی از راه های مبارزه با این مشکل، استفاده از سیستم تشخیص کلاهبرداری ادعای بیمه خودرو است. این سیستم با بهره گیری از فناوری های پیشرفته ای مانند تجزیه و تحلیل داده ها و الگوریتم های یادگیری ماشین، الگوها را در داده های ادعاها شناسایی کرده و نمونه های احتمالی کلاهبرداری را شناسایی می کند. با تجزیه و تحلیل داده هایی مانند تاریخچه ادعا، اطلاعات وسیله نقلیه و راننده و جزئیات تصادف، بیمه گران می توانند فعالیت های مشکوک را شناسایی کرده و تحقیقات بیشتری برای تعیین قانونی یا کلاهبرداری ادعا بررسی کنند.

سیستم شناسایی تقلب در بیمه خودرو با هدف کاهش زیان‌های مالی شرکت‌های بیمه طراحی شده است و به‌طور خودکار ادعاهای مشکوک به تقلب را شناسایی می‌کند. این سیستم با استفاده از تکنیک‌های یادگیری ماشین، ویژگی‌های مختلف مرتبط با ادعاهای بیمه را تحلیل کرده و احتمال وقوع تقلب را پیش‌بینی می‌کند.

با توجه به چالش‌های موجود ، شرکت‌های بیمه در حال بررسی استراتژی‌های مختلفی هستند که شامل اجرای سیستم‌های شناسایی خودکار و استفاده از تکنیک‌های یادگیری ماشین می‌شود. روش‌های سنتی شناسایی تقلب، که مبتنی بر ایجاد قواعد از نشانگرهای شناخته‌شده تقلب هستند، محدودیت‌هایی از خود نشان داده‌اند. ماهیت در حال تغییر طرح‌های تقلبی نیازمند ابزارهای شناسایی هوشمند، انعطاف‌پذیر و مبتنی بر داده است.

پژوهش‌ها و پیشرفت‌های فناوری برای مقابله با پیچیدگی‌های تقلب در بیمه افزایش قابل توجهی داشته است. این پیشرفت‌ها شامل استفاده از روش‌های مبتنی بر داده و هوش مصنوعی است. این رویکردها بر تحلیل و مدل‌سازی روابط پیچیده بین نشانگرهای تقلب و احتمال تقلب تمرکز دارند و هدفشان بهبود شناسایی تقلب با ابزارهایی نیمه‌خودکار، قابل فهم و پاسخگو است.

در این پروژه سعی شده با پیش پردازش مناسب و ازمایش کردن مدل های مختلف و بهبود ان ها و ترکیب برخی از ان ها دقت مدل را در تشخیص تقلب افزایش دهیم .

فصل دوم مرور ادبیات

در سال‌های اخیر، تشخیص تقلب در بیمه خودرو با بهره‌گیری از داده‌کاوی و یادگیری ماشین به یکی از حوزه‌های پرچالش و پرکاربرد تبدیل شده است. مقاله‌ای منتشرشده در Intelligent Systems with Applications [1] با مرور نظام‌مند ۵۰ مطالعه بین سال‌های ۲۰۱۹ تا ۲۰۲۳، رویکردهای نوین و چالش‌های موجود در این زمینه را بررسی کرده است. رایج‌ترین منابع داده شامل دیتاست carclaims.txtو داده‌های خصوصی شرکت‌های بیمه هستند، اما کمبود داده‌های واقعی و نامتوازن بودن آن‌ها از موانع اصلی محسوب می‌شود .الگوریتم‌هایی مانند Random Forest، Logistic Regression، SVMو XGBoost بیشترین کاربرد را داشته‌اند، در حالی که روش‌های نوآورانه‌تری نظیر یادگیری عمیق (CNN، BERT) ، تحلیل متون باNLP ، و مدل‌های گراف‌محور نیز در حال رشد هستند. برای مقابله با عدم تعادل داده‌ها، تکنیک‌هایی مانند SMOTE و ADASYN به‌کار گرفته شده‌اند .

در پژوهشی دیگر [2] ، با هدف طراحی مدلی دقیق و کم‌هزینه برای شناسایی ادعاهای جعلی بیمه‌ای تحقیقاتی انجام شده است. نویسنده با استفاده از داده‌های عمومی مربوط به بیمه خودرو (شامل ۱۵،۴۲۰ رکورد و ۳۳ ویژگی)، مراحل کامل تحلیل داده را طی کرده است: از پاک‌سازی و تبدیل داده‌ها تا مصورسازی و آموزش مدل‌ها. چهار الگوریتم اصلی مورد بررسی قرار گرفته‌اند: Logistic Regression، Random Forest، K-Nearest Neighbors (KNN)وXGBoost . نتایج نشان داد که مدل Random Forest با دقت ٪۹۸.۵ پس از تنظیم پارامترها، بهترین عملکرد را داشته و KNN نیز با دقت ٪۹۶ عملکرد قابل قبولی ارائه داده است Logistic Regression .عملکرد ضعیف‌تری داشت و حتی با تنظیم‌های بیشتر بهبود چشمگیری نداشت. مقاله به چالش عدم توازن داده‌ها (تنها ٪۶ داده‌ها جعلی بودند) اشاره کرده و برای رفع آن از تکنیک‌های بازنمونه‌گیری مانند SMOTE استفاده شده است .چالش‌های اخلاقی، پیچیدگی مدل‌ها، و نیاز به شفافیت الگوریتم‌ها از دغدغه‌های مهم این حوزه‌اند. مقاله پیشنهاد می‌کند که تحقیقات آینده به سمت استفاده از داده‌های بلادرنگ، مدل‌های قابل تفسیر، و ترکیب روش‌های قاعده‌محور با یادگیری ماشین حرکت کنند.

فصل سوم

روش شناسی

توضیحات

یک راه حل موثر برای مشکل تشخیص تقلب در ادعای بیمه خودرو، توسعه و پیاده سازی یک سیستم تشخیص تقلب است که از فناوری های پیشرفته مانند تجزیه و تحلیل داده ها، الگوریتم های یادگیری ماشین و مدل سازی پیش بینی کننده استفاده می کند. چنین سیستمی می تواند حجم زیادی از داده های ادعاها را تجزیه و تحلیل کند تا الگوها و ناهنجاری هایی را که نشان دهنده تقلب احتمالی است شناسایی کند. در ادامه به توضیح مراحل انجام شده برای دست یابی به مدل مناسب می پردازیم. مجموعه داده ی استفاده شده [3] دارای 33 ستون و 15420 ردیف است. در شکل 1 می توان ویژگی ها و نوع آن ها براساس عددی بودن یا دسته ای بودن مشخص شده است. همچنین در شکل 2 میتوانیم اطلاعات مربوط به ویژگی های عددی را اعم از کمترین مقدار و بیشترین مقدار ، میانگین و چولگی و ... مشاهده کنیم.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15420 entries, 0 to 15419
Data columns (total 33 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Month                 15420 non-null  object
 1   WeekOfMonth           15420 non-null  int64
 2   DayOfWeek             15420 non-null  object
 3   Make                  15420 non-null  object
 4   AccidentArea          15420 non-null  object
 5   DayOfWeekClaimed      15420 non-null  object
 6   MonthClaimed          15420 non-null  object
 7   WeekOfMonthClaimed    15420 non-null  int64
 8   Sex                   15420 non-null  object
 9   MaritalStatus         15420 non-null  object
 10  Age                   15420 non-null  int64
 11  Fault                 15420 non-null  object
 12  PolicyType            15420 non-null  object
 13  VehicleCategory       15420 non-null  object
 14  VehiclePrice          15420 non-null  object
 15  FraudFound_P          15420 non-null  int64
 16  PolicyNumber          15420 non-null  int64
 17  RepNumber             15420 non-null  int64
 18  Deductible            15420 non-null  int64
 19  DriverRating          15420 non-null  int64
...
 31  Year                  15420 non-null  int64
 32  BasePolicy            15420 non-null  object
dtypes: int64(9), object(24)
memory usage: 3.9+ MB
```

شکل 1

```
(Index(['WeekOfMonth', 'WeekOfMonthClaimed', 'Age', 'FraudFound_P',
       'PolicyNumber', 'RepNumber', 'Deductible', 'DriverRating', 'Year'],
      dtype='object'),
       WeekOfMonth  WeekOfMonthClaimed      Age  FraudFound_P  PolicyNumber  \
count     15420.00           15420.00 15420.00      15420.00      15420.00
mean          2.79               2.69    39.86          0.06       7710.50
std           1.29               1.26    13.49          0.24       4451.51
min           1.00               1.00     0.00          0.00          1.00
25%           2.00               2.00    31.00          0.00       3855.75
50%           3.00               3.00    38.00          0.00       7710.50
75%           4.00               4.00    48.00          0.00      11565.25
max           5.00               5.00    80.00          1.00      15420.00

       RepNumber  Deductible  DriverRating     Year
count   15420.00    15420.00      15420.00 15420.00
mean        8.48      407.70          2.49  1994.87
std         4.60       43.95          1.12     0.80
min         1.00      300.00          1.00  1994.00
25%         5.00      400.00          1.00  1994.00
50%         8.00      400.00          2.00  1995.00
75%        12.00      400.00          3.00  1996.00
max        16.00      700.00          4.00  1996.00  )
```

شکل 2

در ادامه یک نمودار برای نشان دادن توزیع داده بر اساس ویژگی هدف یعنی تقلب بودن یا نبودن رسم شد که با توجه به ان تنها 6 درصد از داده ها تقلب بودن و این غیرمتوازن بودن شدید را در مجموعه داده نشان می دهد.

دو تا کد پیاده سازی شده است که در کد اول پیش پردازش پیش از تقسیم داده ها به اموزشی و ازمایشی انجام گرفته است. در کد دوم ابتدا تحلیل داده ها و حذف داده های پرت ، سپس تقسیم داده به اموزشی و ازمایشی و در ادامه پیش پردازش و تعریف مدل صورت گرفته است. کد دوم از این جهت صورت گرفته تا اطمینان حاصل شود نشت داده رخ نمی دهد.

3-1- پیش پردازش داده ها

داده ها باید پیش پردازش بشوند که در این مرحله ویژگی های نامربوط حذف شده، اگر داده تکراری وجود داشته باشد حذف می شود، همچنین باید بررسی شود که مقادیر از دست رفته وجود دارد یا خیر.همچنین در این مرحله ویژگی های دسته ای باید رمزگذاری شوند و به عدد تبدیل شوند.

ما در این مجموعه داده هیچ مقدار از دست رفته و هیچ داده تکراری ای نداشتیم. در ویژگی سن برخی از داده ها مقدار صفر دارند که ما آن ها را با میانگین سن ها جایگزین کردیم و همچنین در دو ویژگی ماه و روز ما مقدار صفر داریم که البته ردیف مربوط به ان را حذف کردیم چون فقط یه داده این مقدار را دارد و شامل تقلب نیست پس حذف ان مشکلی ایجاد نمی کند. همچنین ستون پولیسی نامبر به علت عدم اهمیت حذف شد.

مرحله دیگری که برای پیش پردازش انجام دادیم رمزگذاری ویژگی های دسته ای بود که با توجه به تعداد دسته های ویژگی یا ترتیبی بودن آن ویژگی روش مناسب برای هر کدام از ویژگی های دسته ای انتخاب شد. همچنین ویژگی های دوره ای مانند ماه و روز برایشان از انکدینگ سایکلیک استفاده شد. ویژگی هایی که ترتیب ندارند و تعداد دسته های کمی دارند از روش وان هات انکدینگ استفاده دشه است و برای ویژگی هایی که ترتیب داستند از اوردینال انکدینگ استفاده کردیم. از باینری انکدینگ هم برای برخی ویژگی هایی که به علت تعداد زیاد دسته هایشان از وان هات استفاده نکردیم ، بهره بردیم.

3-2- تحلیل اکتشافی داده ها

در ابتدا نمودار مربوط به ماتریس همبستگی ویژگی های عددی جهت درک بهتر رسم شده است اگر در این جدول دو تا ویژگی با ضریب همبستگی بالا وجود داشته باشد یکی از ان ها را حذف میکنیم که وجود نداشت. در ادامه برای درک بهتر ارتباط هر یک از ویژگی های عددی با هدف باکس پلات هایی رسم شد و همچنین برای درک ارتباط بین ویژگی های دسته ای با هدف هم نمودار هایی رسم کردیم . که از طریق این نمودار ها به خوبی می توان توزیع دسته های ویژگی های مختلف را در دو دسته ی هدف مشاهده کرد. در ادامه به کمک باکس پلات ها داده های پرت را شناسایی کرده و به کمک روش ای کیو ار داده های پرت را حذف کردیم. در ویژگی های سن و مالیات پذیر داده های پرت مشاهده شده بود که ان ها را حذف کردیم سپس جهت اطمینان دوباره نمودار ها را رسم کردیم تا مطمئن شویم داده های پرت حذف شده اند. همچنین نمودار های میله ای برای رسم توزیع ویژگی ها استفاده کردیم.

3-3- تقسیم داده ها به دو مجموعه ازمایشی و اموزشی و بررسی مدل ها

ابتدا مجموعه داده را به مجموعه اموزشی و ازمایشی به نسبت 70 به 30 تقسیم میکنیم. سپس فقط ویژگی های عددی مربوط به داده ی ترین را استاندارد سازی میکنیم. در کد دوم که پیش پردازش بعد از تقسیم داده صورت گرفته است، در انکدینگ نیز تنها از داده های اموزشی استفاده می شود چون ممکن است مدل مقادیری را ببیند که فقط در تست وجود دارند و به‌طور غیرمنصفانه‌ای از این اطلاعات بهره‌مند شود.

3-3-1-استفاده از لیزی پردیکت

از کتابخانه لیزی پردیکت جهت مشاهده عملکرد مدل های مختلف استفاده شد تا بتوانیم از طریق ان مدل های بهتر با دقت بالاتر را شناسایی کنیم و برروی آن ها کار کنیم. با توجه به نموداری که درابتدا گفتیم و اینکه مقادیر مربوط به دقت بالا بود ولی مقادیر مربوط به دقت بالانس برای بسیاری از مدل ها در حدود 0.5 بود . از اسموت جهت بالانس کردن داده ها استفاده کردیم و دوباره لیزی پردیکت را اجرا کردیم سپس بر اساس ان چند مدل طبقه بندی انتخاب شد شامل : جنگل تصادفی، ایکس جی بی ، ال جی بی ام و اکسترا تریز .

3-3-2-جنگل تصادفی

الگوریتم جنگل تصادفی یک تکنیک قدرتمند مبتنی بر یادگیری درخت در حوزه یادگیری ماشین است. این الگوریتم با ایجاد تعدادی درخت تصمیم‌گیری در طی مرحله آموزش کار می‌کند. هر درخت با استفاده از یک زیرمجموعه تصادفی از داده‌ها و ارزیابی یک زیرمجموعه تصادفی از ویژگی‌ها در هر بخش ساخته می‌شود. این تصادفی بودن باعث ایجاد تنوع میان درخت‌های فردی شده و خطر بیش‌برازش را کاهش داده و عملکرد پیش‌بینی کلی را بهبود می‌بخشد.

ابتدا مدل جنگل تصادفی را پیاده سازی و امتیاز F1 ان برای مقدار هدف 1 بسیار پایین بود . جهت بهبود ان از استانه بهینه استفاده کردیم که حدود 0.21 ان را بهبود داد . با استفاده از نمودار هایی که رسم شد می توان درک بهتری از انتخاب بهترین استانه پیدا کرد. سپس جهت بهبود بیشتر این مدل از

هایپرپارامتر ها استفاده کردیم که بهترین پارامتر ها برای مدل جهت کسب امتیاز بیشتر پیدا شود و پس از ان بار دیگر از استانه بهینه برای مدل بهبود یافته استفاده کردیم که به امتیاز 0.28 رسید.

3-3-3- ایکس جی بوست

ایکس جی بوست یک کتابخانه محبوب برای تقویت گرادیان برای آموزش جی پی یو، محاسبات توزیع شده و موازی سازی است. دقیق است، به خوبی با انواع داده ها و مشکلات سازگار است، مستندات عالی دارد و به طور کلی استفاده از آن بسیار آسان است.

این مدل برروی داده های اموزشی اجرا شد و پیش بینی برروی داده های تست انجام شده و ارزیابی ها صورت گرفت و مجدد برای این مدل هم هایپرپارامتر و تعیین استانه بهینه انجام شد. پیش از تعیین استانه بهینه امتیاز مدل 0.19 بود و پس از تعیین استانه 0.28 است.

3-3-4- ال جی بی ام

یک الگوریتم یادگیری ماشین بر پایه تکنیک گرادیان بوستینگ است که به‌طور خاص برای مدل‌سازی سریع و کارآمد روی داده‌های بزرگ و پیچیده طراحی شده است. این الگوریتم به دلیل استفاده از تکنیک‌های پیشرفته ، سرعت آموزش بالاتری دارد و در عین حال دقت بالایی را ارائه می‌دهد.

مشابه مدل های قبلی تمام مراحل انجام شد و پس از بهبود دادن پارامتر ها و مدل و انتخاب استانه مناسب امتیاز مدل به 0.31 رسید.

3-3-5- اکسترا تریز

این الگوریتم یکی از تکنیک‌های یادگیری ماشین است که بر اساس مجموعه‌ای از درخت‌های تصمیم کار می‌کند و برای مسائل رگرسیون و دسته‌بندی مناسب است. این الگوریتم شبیه جنگل تصادفی است اما تفاوت‌های کلیدی دارد که آن را متمایز می‌کند. برخلاف جنگل تصادفی که بهینه‌ترین نقاط تقسیم را برای هر ویژگی انتخاب می‌کند، این الگوریتم به‌صورت تصادفی نقاط تقسیم را انتخاب می‌کند. این امر

موجب افزایش سرعت آموزش می‌شود. با افزایش تصادفی‌سازی، مدل کمتر به داده‌های آموزشی حساس می‌شود و احتمال اورفیت کاهش می‌یابد.

مشابه مدل های قبلی مراحل انجام شد و پس از بهینه سازی و مشخص کردن استانه بهینه امتیاز به 0.25 رسید.

3-4- کراس ولیدیشن

کراس ولیدیشن یکی از تکنیک‌های اساسی در یادگیری ماشین است که برای ارزیابی عملکرد مدل و بهبود تعمیم‌پذیری آن استفاده می‌شود. هدف اصلی این روش این است که مدل را بر روی داده‌های متفاوت آزمایش کند. روش کار به این صورت است : داده‌ها به چند بخش معمولاً k بخش، تقسیم می‌شوند. این بخش‌ها به نام فولد شناخته می‌شوند. مدل روی $k-1$ بخش آموزش داده می‌شود و بخش باقی‌مانده به‌عنوان داده‌های آزمون استفاده می‌شود. این فرآیند k بار تکرار می‌شود، به‌طوری‌که هر بخش دقیقاً یک‌بار به‌عنوان داده آزمون استفاده می‌شود. نتایج حاصل از هر تکرار مثلاً دقت، F1-Score، یا هر متریک دیگری میانگین‌گیری می‌شوند تا عملکرد کلی مدل محاسبه شود.

برای ارزیابی عملکرد مدل‌ها در شناسایی تقلب در بیمه خودرو، از تکنیک کراس ولیدیشن با ۵ بخش استفاده شد. معیار ارزیابی مدل‌ها امتیاز F1بود که نشان‌دهنده توازن بین دقت و بازخوانی است.

```
Cross-validating: Random Forest
Random Forest -> F1 Scores: [0.8747816  0.99355504 0.99535604 0.99330587 0.99355836]
Random Forest -> Mean f1: 0.9701, Std f1: 0.0477
Cross-validating: Extra Trees
Extra Trees -> F1 Scores: [0.92558783 0.98896021 0.99305019 0.99125064 0.98896587]
Extra Trees -> Mean f1: 0.9776, Std f1: 0.0260
Cross-validating: LightGBM
LightGBM -> F1 Scores: [0.81645181 0.99689119 0.99792531 0.998187   0.99870298]
LightGBM -> Mean f1: 0.9616, Std f1: 0.0726
Cross-validating: XGBoost
XGBoost -> F1 Scores: [0.81608842 0.99714952 0.99818418 0.99534643 0.99766537]
XGBoost -> Mean f1: 0.9609, Std f1: 0.0724

Cross-Validation Results:
              Mean f1  Std f1
Random Forest     0.97    0.05
Extra Trees       0.98    0.03
LightGBM          0.96    0.07
XGBoost           0.96    0.07
```

شکل 3

نتایج در شکل 3 نشان داده شده است.

3-5-مدل های ترکیبی

مدل‌های **یادگیری تجمعی** مانند گروهی از کارشناسان متنوع عمل می‌کنند که برای تصمیم‌گیری با یکدیگر همکاری می‌کنند . تصور کنید گروهی از دوستان با مهارت‌های مختلف با هم روی یک پروژه کار می‌کنند. هر دوست در یک حوزه خاص توانایی برتری دارد و با ترکیب نقاط قوتشان، راه‌حلی قوی‌تر از آنچه هر فرد به تنهایی می‌تواند ارائه دهد، ایجاد می‌کنند.

به‌طور مشابه، در یادگیری تجمعی، مدل‌های مختلف (اغلب از یک نوع یا انواع مختلف) با هم همکاری می‌کنند تا عملکرد پیش‌بینی را بهبود بخشند. این رویکرد بر استفاده از خرد جمعی گروه تمرکز دارد تا محدودیت‌های فردی را برطرف کرده و تصمیمات آگاهانه‌تری در وظایف مختلف یادگیری ماشین اتخاذ کند. این مدل‌ها به دلیل قدرت ترکیب نقاط قوت مدل‌های مختلف، در بسیاری از کاربردهای یادگیری ماشین مورد استفاده قرار می‌گیرند.

در نهایت سعی شد با ترکیب چهار مدل و وزن دادن به آن ها به یک مدل ترکیبی برسیم وآن را ارزیابی کنیم.

تلفیق مدل‌ها یکی از روش‌های پیشرفته برای بهبود عملکرد مدل‌های یادگیری ماشین است که با ترکیب چندین مدل، خطای کلی کاهش یافته و دقت پیش‌بینی افزایش می‌یابد. در پروژه ، از Voting Classifierیا Stackingبرای تلفیق مدل‌ها استفاده شد. در روش وتینگ ، خروجی چندین مدل ترکیب می‌شود (به صورت سخت یا نرم) و تصمیم نهایی گرفته می‌شود. و در استکینگ ، خروجی پیش‌بینی مدل‌های پایه به عنوان ورودی یک مدل نهایی (Meta-Model) استفاده می‌شود. در نهایت این مدل های ترکیبی با چهار مدل پایه مقایسه شدند. بهترین نتیجه مربوط به روش وتینگ با 0.32 امتیاز بود.

فصل چهارم

نتایج

نتایج نهایی بدست امده برای مدل ها را می توان در شکل 4 و 5 مشاهده کرد.

4-1- مقایسه بین مدل‌ها:

همان‌طور که در شکل ۴ نشان داده شد، الگوریتم‌های LGBM و XGBoost بهترین عملکرد را در معیار F1 به دست آوردند (حدود 0.31 و 0.28). این مسئله طبیعی است، زیرا این مدل‌ها از خانواده گرادیان بوستینگ بوده و توانایی بالایی در یادگیری الگوهای پیچیده و غیرخطی دارند .در مقابل، الگوریتم‌های Random Forest و Extra Trees عملکرد ضعیف‌تری داشتند (حدود 0.25–0.26). دلیل این موضوع می‌تواند به حساسیت کمتر آن‌ها نسبت به داده‌های نامتوازن و قدرت کمترشان در بهینه‌سازی آستانه‌های تصمیم‌گیری مربوط باشد.

4-2- تأثیر نامتوازن بودن داده‌ها بر نتایج:

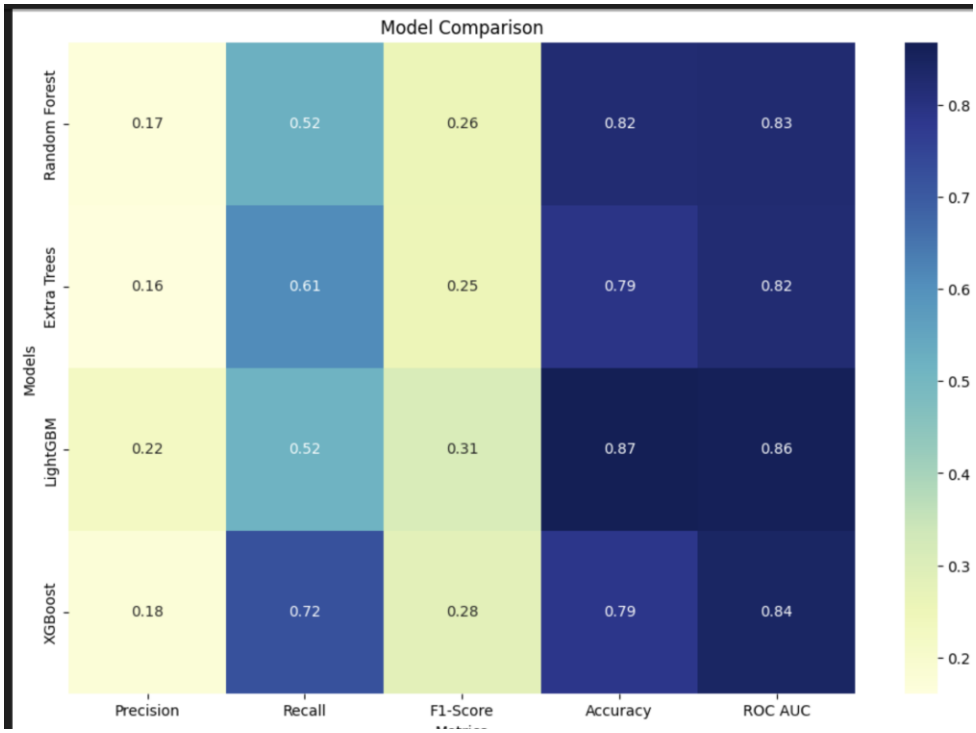
از آنجا که فقط حدود ۶٪ داده‌ها مربوط به کلاس تقلب بودند، بسیاری از مدل‌ها دچار سوگیری به سمت کلاس اکثریت شدند. استفاده از روش SMOTE در افزایش نمونه‌های کلاس اقلیت نقش مؤثری در بهبود امتیاز F1 داشت. این نشان می‌دهد که مدیریت درست داده‌های نامتوازن یکی از عوامل کلیدی در موفقیت مدل‌ها است.

4-3- بررسی مدل‌های ترکیبی:

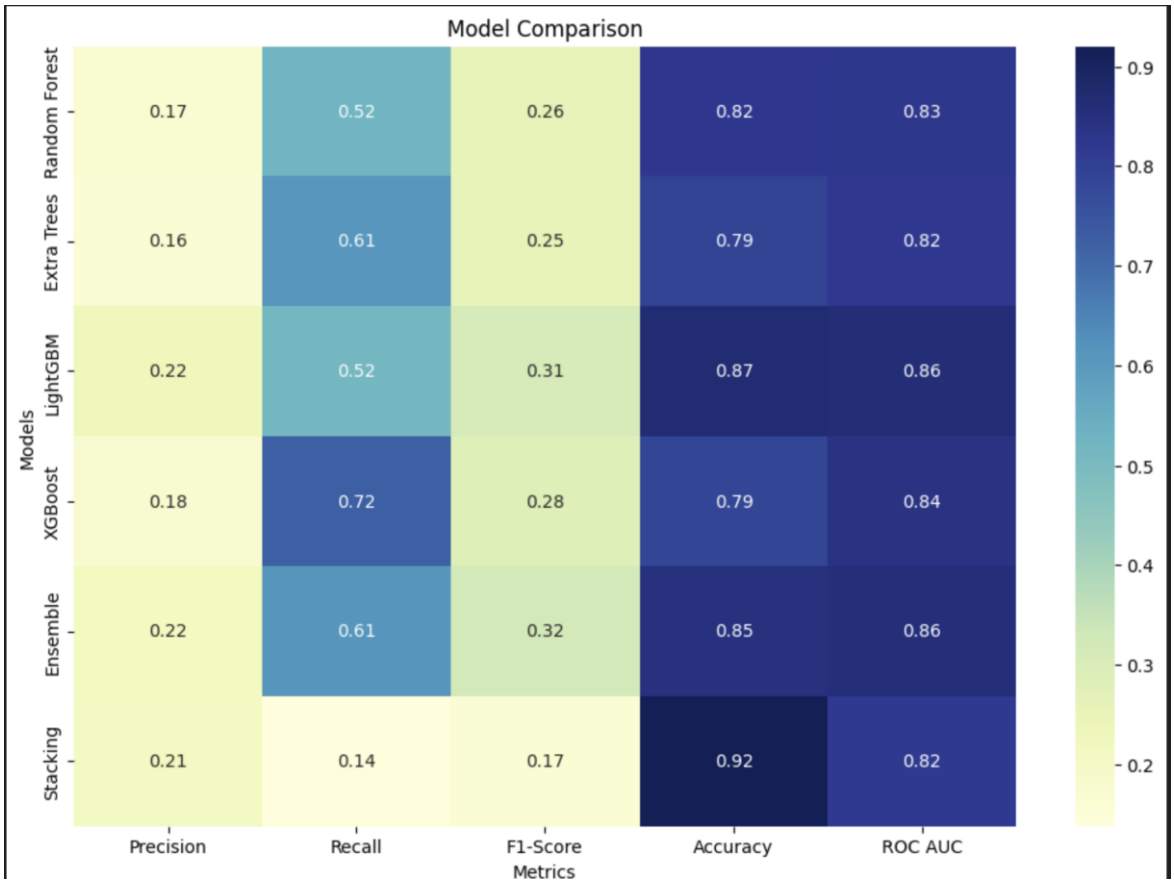
همان‌طور که در شکل ۵ مشاهده می شود، روش ترکیبی مبتنی بر Voting توانست از مدل‌های پایه عملکرد بهتری کسب کند (حدود 0.32). این افزایش هرچند کوچک به نظر می‌رسد، اما نشان می‌دهد که استفاده از خرد جمعی مدل‌ها می‌تواند به افزایش پایداری و دقت پیش‌بینی کمک کند. در مقایسه، روش‌های استکینگ یا استفاده مجدد از مدل‌های پایه به‌عنوان ورودی برای متا مدل، بهبود قابل‌توجهی نسبت به بهترین مدل‌های منفرد نداشتند. علت این موضوع می‌تواند به محدودیت حجم داده و پیچیدگی ذاتی ویژگی‌ها برگردد.

4-4- برداشت کلی:

مدل‌های گرادیان بوستینگ XGBoost و LGBM برای این مسئله مناسب‌تر هستند .استفاده از مدل‌های ترکیبی توانست کمی عملکرد را بالاتر ببرد، که نشان‌دهنده پتانسیل رویکردهایEnsemble است .همچنان نیاز به بهبود وجود دارد، خصوصاً درRecall ، زیرا شناسایی حداکثری تقلب‌ها در حوزه بیمه از نظر تجاری بسیار حیاتی است.



شکل 4


Model Comparison
Random Forest
Extra Trees
LightGBM
XGBoost
Ensemble
Stacking
Models
0.17 0.52 0.26 0.82 0.83
0.16 0.61 0.25 0.79 0.82
0.22 0.52 0.31 0.87 0.86
0.18 0.72 0.28 0.79 0.84
0.22 0.61 0.32 0.85 0.86
0.21 0.14 0.17 0.92 0.82
Precision
Recall
F1-Score
Accuracy
ROC AUC
Metrics
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2

شکل 5

فصل پنجم

جمع‌بندی و نتیجه‌گیری و پیشنهادات

در نتیجه، یک سیستم تشخیص کلاهبرداری در ادعای بیمه خودرو برای شناسایی ادعاهای متقلبانه و جلوگیری از متحمل شدن خسارت توسط شرکت بیمه ضروری است. این سیستم از یادگیری ماشین و تجزیه و تحلیل داده ها برای شناسایی الگوها و ناهنجاری ها در داده های ادعا استفاده می کند .

در این پروژه، با هدف شناسایی تقلب در بیمه خودرو، ابتدا داده‌ها پیش‌پردازش شدند تا کیفیت داده‌ها افزایش یابد و الگوهای مهم شناسایی شوند. سپس با استفاده از روش‌های بازنمونه‌گیری مانند اسموت، عدم‌تعادل موجود در داده‌ها برطرف گردید. برای تحلیل بهتر، از الگوریتم‌های مختلف یادگیری ماشین مانند جنگل تصادفی ، ایکس جی بوست، لایت ال جی بی و اکسترا تریز استفاده شد و عملکرد آن‌ها با استفاده ازکراس ولیدیشن ارزیابی گردید.

دامنه آینده یک سیستم تشخیص کلاهبرداری ادعای بیمه خودرو امیدوار کننده است. با پیشرفت های مداوم در یادگیری ماشین و تجزیه و تحلیل داده ها، سیستم را می توان برای بهبود دقت و کارایی آن بیشتر کرد. می توان آن را بر روی مجموعه داده های بزرگتر آموزش داد تا توانایی آن را در تعیین الگوهای ظریفی که می تواند نشان دهنده تقلب باشد، بهبود بخشد. همچنین، این سیستم را می توان با منابع داده خارجی مانند رسانه های اجتماعی و پایگاه داده های اجرای قانون ادغام کرد تا بینش جامع تر و درک بهتری از زمینه ادعاها ارائه دهد.

منابع و مراجع

[1] A. I. Alrais, "Fraudulent insurance claims detection using machine learning," Master's Project, Rochester Institute of Technology, Rochester, NY, USA, 2022. [Online]. Available: https://repository.rit.edu/theses/11376

[2] G. Schrijver, D. K. Sarmah, and M. El-Hajj, "Automobile insurance fraud detection using data mining: A systematic literature review," Intell. Syst. Appl., vol. 21, Art. no. 200340, Mar. 2024, doi: 10.1016/j.iswa.2024.200340.

[3] Vehicle Insurance Claim Fraud Detection

[4] A. Bhishek, "Car.ly - VehicleInsurance Claim Fraud Detection," GitHub repository, 2022. [Online]. Available: https://github.com/abhisheks008/Car.ly/blob/main/Model/Car.ly%20-%20VehicleInsurance%20Claim%20Fraud%20Detection.ipynb

[5] S. P. Yeo, "Vehicle-insurance-fraud-detection," GitHub repository, 2022. [Online]. Available: https://github.com/SiewPingYeo/Vehicle-Insurance-Fraud-Detection

[6] "XGBoost," GeeksforGeeks, 2023. [Online]. Available: https://www.geeksforgeeks.org/xgboost/

[7] "Random forest algorithm in machine learning," GeeksforGeeks, 2023. [Online]. Available: https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/

[8] "LightGBM (Light Gradient Boosting Machine)," GeeksforGeeks, 2023. [Online]. Available: https://www.geeksforgeeks.org/lightgbm-light-gradient-boosting-machine/

Abstract