گزارش پروژه

بررسی روش‌های یادگیری ماشین در تشخیص تقلب مالی

نام نویسنده: امیرحسین رنجبر

چکیده

تشخیص تراکنش‌های مالی تقلبی یکی از مسائل مهم در حوزه یادگیری ماشین است. نامتوازن بودن داده‌ها، تعداد بسیار کم نمونه‌های تقلبی و هزینه بالای تشخیص نادرست از جمله چالش‌های اصلی این مسئله هستند.

در این گزارش، چندین الگوریتم یادگیری ماشین شامل درخت تصمیم، جنگل تصادفی و XGBoost بررسی شده و عملکرد آن‌ها با استفاده از معیارهایی مانند Precision، Recall، F۱score- و PR-AUC ارزیابی می‌شود.

همچنین استفاده از روش‌های Ensemble برای ترکیب چند مدل و بهبود عملکرد سیستم مورد بررسی قرار می‌گیرد.

واژگان کلیدی: یادگیری ماشین، تشخیص تقلب، تراکنش مالی، جنگل تصادفی، XGBoost، Ensemble، داده‌های نامتوازن

۱. مقدمه

افزایش استفاده از سامانه‌های پرداخت الکترونیکی باعث شده است که حجم بسیار زیادی از تراکنش‌های مالی در هر روز ثبت شود. در کنار این افزایش، احتمال وقوع تراکنش‌های تقلبی نیز وجود دارد.

تشخیص خودکار تراکنش‌های تقلبی می‌تواند نقش مهمی در کاهش خسارت‌های مالی داشته باشد. با این حال، مسئله تشخیص تقلب معمولاً با یک مشکل مهم یعنی نامتوازن بودن کلاس‌ها مواجه است.

در بسیاری از مجموعه‌داده‌های واقعی، تعداد تراکنش‌های سالم بسیار بیشتر از تراکنش‌های تقلبی است. بنابراین یک مدل ممکن است دقت بسیار بالایی داشته باشد، اما در تشخیص نمونه‌های تقلبی عملکرد مناسبی نداشته باشد.

به همین دلیل، در این مسئله معیارهایی مانند Recall و PR-AUC اهمیت زیادی پیدا می‌کنند.

۲. معرفی داده‌ها

در این پژوهش از یک مجموعه‌داده شامل تراکنش‌های مالی استفاده شده است. هر نمونه شامل تعدادی ویژگی مربوط به تراکنش و یک برچسب دودویی است.

برچسب صفر نشان‌دهنده تراکنش سالم و برچسب یک نشان‌دهنده تراکنش تقلبی است.

یکی از ویژگی‌های مهم داده، عدم توازن شدید بین دو کلاس است. به همین دلیل، استفاده از Accuracy به‌تنهایی نمی‌تواند معیار مناسبی برای ارزیابی مدل باشد.

داده‌ها پس از انجام مراحل پیش‌پردازش به دو مجموعه آموزش و آزمون تقسیم شدند.

۳. پیش‌پردازش داده‌ها

پیش از آموزش مدل‌ها، داده‌ها بررسی و آماده‌سازی شدند. مراحل پیش‌پردازش شامل بررسی مقادیر گمشده، داده‌های تکراری و آماده‌سازی ویژگی‌ها برای ورود به مدل است.

همان‌طور که در بررسی اولیه داده‌ها مشاهده شد، مجموعه‌داده دارای تعدادی مقدار تکراری بود. از آنجا که وجود داده‌های تکراری می‌تواند بر فرآیند آموزش مدل تأثیر بگذارد، این نمونه‌ها شناسایی و حذف شدند.

تعداد نمونه‌ها و تراکنش‌های تقلبی قبل و بعد از حذف داده‌های تکراری در جدول زیر ارائه شده است.

Data	Before	After
Total transactions	284807	284315
Fraudulent transactions	492	473

جدول ۱: مقایسه داده‌ها قبل و بعد از حذف نمونه‌های تکراری

همان‌طور که مشاهده می‌شود، تعداد کل تراکنش‌ها از ۲۸۴۸۰۷ به ۲۸۴۳۱۵ کاهش یافته است. همچنین تعداد تراکنش‌های تقلبی از ۴۹۲ نمونه به ۴۷۳ نمونه رسیده است. بنابراین حذف نمونه‌های تکراری باعث حذف بخشی از تراکنش‌های تقلبی نیز شده است.

داده‌ها به دو مجموعه آموزش و آزمون تقسیم شدند:

$$D = D_{\text{train}} \cup D_{\text{test}}$$

به‌طوری‌که مجموعه آزمون در فرآیند آموزش مدل استفاده نمی‌شود.

در صورت وجود ویژگی‌هایی با مقیاس‌های متفاوت، می‌توان از روش‌های استانداردسازی برای آماده‌سازی داده‌ها استفاده کرد.

۴. روش‌های یادگیری ماشین

در این پژوهش چند الگوریتم مختلف برای تشخیص تراکنش‌های تقلبی بررسی شده‌اند.

۱.۴. درخت تصمیم

درخت تصمیم داده‌ها را با استفاده از مجموعه‌ای از قوانین شرطی به بخش‌های مختلف تقسیم می‌کند. هدف الگوریتم ایجاد تقسیم‌هایی است که نمونه‌های متعلق به کلاس‌های مختلف را تا حد امکان از یکدیگر جدا کنند.

یکی از پارامترهای مهم درخت تصمیم، عمق درخت است. افزایش بیش از حد عمق می‌تواند باعث بیش‌برازش شود. از طرف دیگر، عمق بسیار کم نیز ممکن است باعث کم‌برازش مدل شود.

برای انتخاب مناسب‌ترین تنظیمات مدل، در این پژوهش از روش Search Grid استفاده شد. در این روش، مجموعه‌ای از مقادیر مختلف برای هایپرپارامترهای مدل تعریف شده و ترکیب‌های مختلف آن‌ها به صورت سیستماتیک بررسی شدند.

پارامترهایی مانند عمق درخت (max_depth)، حداقل تعداد نمونه مورد نیاز برای تقسیم یک گره (min_samples_split) و حداقل تعداد نمونه در هر برگ (min_samples_leaf) مورد بررسی قرار گرفتند.

برای ارزیابی ترکیب‌های مختلف از اعتبارسنجی متقابل Strat-ified K-Fold استفاده شد تا نسبت کلاس‌ها در بخش‌های مختلف داده تا حد امکان حفظ شود.

معیار F۱score- به عنوان معیار اصلی انتخاب مدل در نظر گرفته شد؛ زیرا در مسئله تشخیص تقلب، توجه هم‌زمان به Pre-cision و Recall اهمیت زیادی دارد.

در نهایت، ترکیبی از هایپرپارامترها که بالاترین مقدار F۱score- را در فرآیند اعتبارسنجی به دست آورد، به عنوان تنظیمات نهایی مدل انتخاب شد.

۲.۴. جنگل تصادفی

جنگل تصادفی از تعداد زیادی درخت تصمیم تشکیل شده است. هر درخت روی نمونه‌ها و ویژگی‌های متفاوتی آموزش داده می‌شود و در نهایت خروجی درخت‌ها با یکدیگر ترکیب می‌شوند.

احتمال تعلق یک نمونه به کلاس تقلب را می‌توان به صورت زیر در نظر گرفت:

$$P(y=1|x) = \frac{1}{T}\sum_{t=1}^{T} P_t(y=1|x)$$

که در آن T تعداد درخت‌ها است.
مزیت مهم جنگل تصادفی کاهش واریانس و کاهش احتمال بیش‌برازش نسبت به یک درخت تصمیم منفرد است.

۳.۴. XGBoost

XGBoost یکی از الگوریتم‌های قدرتمند مبتنی بر روش Gra-dient Boosting است که از مجموعه‌ای از درخت‌های تصمیم برای ایجاد یک مدل نهایی استفاده می‌کند. برخلاف روش‌هایی مانند جنگل تصادفی که درخت‌ها را به صورت مستقل و موازی آموزش می‌دهند، در XGBoost درخت‌ها به صورت ترتیبی ساخته می‌شوند. به این صورت که هر درخت جدید تلاش می‌کند خطاها و نقاط ضعف مدل‌های قبلی را کاهش دهد.
در ابتدا یک مدل ساده ایجاد می‌شود و سپس درخت بعدی با توجه به خطاهای باقی‌مانده از مدل قبلی آموزش داده می‌شود. این فرآیند به صورت تکراری ادامه پیدا می‌کند و در نهایت خروجی تمام درخت‌ها با یکدیگر ترکیب می‌شود. بنابراین، هر درخت جدید بخشی از خطای مدل قبلی را اصلاح می‌کند.
تابع کلی مدل XGBoost را می‌توان به صورت زیر نمایش داد:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$$

که در آن K تعداد درخت‌های موجود در مدل و f_k تابع مربوط به درخت تصمیم k-ام است. در واقع، پیش‌بینی نهایی مدل از مجموع خروجی تمام درخت‌ها به دست می‌آید.
یکی از ویژگی‌های مهم XGBoost استفاده از تابع هدف شامل تابع خطا و جمله منظم‌سازی است. به صورت کلی می‌توان تابع هدف را به شکل زیر در نظر گرفت:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

که در آن l تابع خطا و Ω جمله منظم‌سازی پیچیدگی درخت‌ها است. وجود جمله منظم‌سازی به کنترل پیچیدگی مدل و کاهش احتمال بیش‌برازش کمک می‌کند.
از پارامترهای مهم XGBoost می‌توان به تعداد درخت‌ها (n_estimators)، عمق درخت (max_depth) و نرخ یادگیری (learning_rate) اشاره کرد. انتخاب مناسب این پارامترها تأثیر مستقیمی بر عملکرد مدل دارد. به همین دلیل، در این پژوهش برای یافتن ترکیب مناسب پارامترها از روش Search Grid استفاده شده است. در Search Grid مجموعه‌ای از مقادیر مختلف برای پارامترهای مدل تعریف شده و ترکیب‌های مختلف آن‌ها با استفاده از اعتبارسنجی متقابل بررسی می‌شوند. در نهایت ترکیبی انتخاب می‌شود که بهترین عملکرد را بر اساس معیار ارزیابی مورد نظر داشته باشد.
از آنجا که مجموعه‌داده مورد استفاده در این پژوهش دارای عدم توازن شدید بین کلاس سالم و تقلبی است، استفاده از پارامتر scale_pos_weight نیز اهمیت زیادی دارد. این پارامتر وزن نمونه‌های مربوط به کلاس مثبت، یعنی تراکنش‌های تقلبی، را در فرآیند یادگیری افزایش می‌دهد.
به طور معمول می‌توان مقدار اولیه این پارامتر را بر اساس نسبت تعداد نمونه‌های کلاس منفی به کلاس مثبت تعیین کرد:

$$\text{scale_pos_weight} = \frac{N_0}{N_1}$$

که در آن N_0 تعداد نمونه‌های سالم و N_1 تعداد نمونه‌های تقلبی است. با افزایش وزن کلاس اقلیت، مدل هنگام آموزش اهمیت بیشتری برای شناسایی تراکنش‌های تقلبی قائل می‌شود. این کار می‌تواند به بهبود Recall و توانایی مدل در شناسایی نمونه‌های تقلبی کمک کند، هرچند ممکن است باعث افزایش تعداد مثبت‌های کاذب و کاهش Precision شود.

در این پژوهش، ترکیب XGBoost با Search Grid و استفاده از scale_pos_weight به منظور یافتن مدل مناسب‌تر برای داده‌های نامتوازن مورد استفاده قرار گرفته است.

۵. عدم توازن کلاس‌ها

در مسئله تشخیص تقلب، تعداد نمونه‌های کلاس تقلب معمولاً بسیار کمتر از نمونه‌های سالم است.
فرض کنیم:

$$N_0 \gg N_1$$

که در آن N_0 تعداد نمونه‌های سالم و N_1 تعداد نمونه‌های تقلبی است.
در چنین شرایطی، یک مدل ممکن است تقریباً تمام نمونه‌ها را سالم پیش‌بینی کند و همچنان Accuracy بسیار بالایی به دست آورد.
بنابراین معیارهای دیگری مانند Precision، Recall و F۱-score اهمیت بیشتری پیدا می‌کنند.
Precision به صورت زیر تعریف می‌شود:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall نیز برابر است با:

$$\text{Recall} = \frac{TP}{TP + FN}$$

و F۱-score به صورت زیر محاسبه می‌شود:

$$F1 = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

۶. ارزیابی مدل‌ها

برای مقایسه مدل‌ها از ماتریس درهم‌ریختگی و معیارهای Pre-cision، Recall، F۱-score و PR-AUC استفاده می‌شود.
ماتریس درهم‌ریختگی به شکل زیر است:

$$\begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix}$$

در مسئله تشخیص تقلب، مقدار FN اهمیت زیادی دارد؛ زیرا هر FN نشان‌دهنده یک تراکنش تقلبی است که توسط مدل به عنوان تراکنش سالم شناسایی شده است.
بنابراین در بسیاری از کاربردهای تشخیص تقلب، افزایش Re-call می‌تواند یکی از اهداف اصلی بهینه‌سازی مدل باشد.

۷. بهینه‌سازی هایپرپارامترها

به منظور بهبود عملکرد مدل‌ها، علاوه بر آموزش اولیه، فرآیند بهینه‌سازی هایپرپارامترها نیز انجام شد. هایپرپارامترها پارامترهایی هستند که پیش از فرآیند آموزش مدل تعیین می‌شوند و مستقیماً از داده‌های آموزشی یاد گرفته نمی‌شوند.
در این پژوهش از روش Search Grid برای بررسی سیستماتیک ترکیب‌های مختلف هایپرپارامترها استفاده شد. در این روش، تمام ترکیب‌های مشخص‌شده از مقادیر هایپرپارامترها بررسی شده و عملکرد هر ترکیب با استفاده از اعتبارسنجی متقابل ارزیابی شد.
برای مدل درخت تصمیم، پارامترهایی مانند max_depth، min_samples_split و min_samples_leaf مورد بررسی قرار گرفتند.
از آنجا که داده‌های مورد استفاده نامتوازن هستند، معیار Ac-curacy به عنوان معیار اصلی انتخاب مدل در نظر گرفته نشد. در عوض، F۱-score به عنوان معیار اصلی Search Grid انتخاب شد تا تعادل مناسبی بین Precision و Recall ایجاد شود.

به صورت کلی، فرآیند انتخاب بهترین مدل را می‌توان به شکل زیر نمایش داد:

$$\theta^* = \arg\max_{\theta \in \Theta} F۱(\theta)$$

که در آن Θ مجموعه تمام ترکیب‌های هایپرپارامترها و θ^* ترکیب بهینه‌ای است که بیشترین مقدار score۱F- را به دست می‌آورد.
پس از یافتن بهترین ترکیب هایپرپارامترها، مدل نهایی با استفاده از این تنظیمات روی مجموعه آموزش آموزش داده شد و عملکرد آن روی مجموعه آزمون مستقل ارزیابی شد.

۸. روش Ensemble

برای بهبود عملکرد می‌توان خروجی چند مدل را با یکدیگر ترکیب کرد.
برای مثال، احتمال تقلب حاصل از سه مدل را می‌توان به شکل زیر ترکیب کرد:

$$P_{\text{ensemble}} = w_۱ P_{\text{DT}} + w_۲ P_{\text{RF}} + w_۳ P_{\text{XGB}}$$

به‌طوری‌که:

$$w_۱ + w_۲ + w_۳ = ۱$$

برای مثال می‌توان از ترکیب زیر استفاده کرد:

$$P_{\text{ensemble}} = ۰/۲ P_{\text{DT}} + ۰/۴ P_{\text{RF}} + ۰/۴ P_{\text{XGB}}$$

پس از محاسبه احتمال نهایی، می‌توان با استفاده از یک آستانه مناسب کلاس نهایی را تعیین کرد:

$$\hat{y} = \begin{cases} ۱ & P_{\text{ensemble}} \geq \tau \\ ۰ & P_{\text{ensemble}} < \tau \end{cases}$$

انتخاب مقدار τ تأثیر مستقیمی بر Precision و Recall مدل دارد.

۹. نتایج

در این بخش نتایج حاصل از اجرای مدل‌های مختلف یادگیری ماشین بررسی می‌شوند.
معیارهای اصلی مورد استفاده شامل Precision، Recall، ۱F-score و PR-AUC هستند.
در مسئله تشخیص تقلب، معیار Recall اهمیت ویژه‌ای دارد؛ زیرا نشان می‌دهد چه نسبتی از تراکنش‌های تقلبی واقعاً توسط مدل شناسایی شده‌اند.
همچنین می‌توان عملکرد مدل‌ها را با استفاده از ماتریس درهم‌ریختگی بررسی کرد.
نتایج به دست آمده برای هر مدل باید بر اساس داده‌های واقعی آزمایش ثبت و مقایسه شوند.
استفاده از Search Grid باعث شد تنظیمات مناسب‌تری برای هایپرپارامترهای مدل انتخاب شود. مدل انتخاب‌شده بر اساس بالاترین score۱F- حاصل از فرآیند اعتبارسنجی انتخاب شد و سپس روی مجموعه آزمون مستقل مورد ارزیابی قرار گرفت.

۱۰. بحث

نتایج نشان می‌دهد که ارزیابی مدل در مسائل نامتوازن باید فراتر از Accuracy انجام شود.
در این مسئله، افزایش Recall می‌تواند به معنای شناسایی تعداد بیشتری از تراکنش‌های تقلبی باشد، هرچند ممکن است باعث افزایش تعداد مثبت‌های کاذب شود.
از طرف دیگر، افزایش بیش از حد Recall ممکن است Precision را کاهش دهد. بنابراین انتخاب آستانه مناسب اهمیت زیادی دارد.
همچنین استفاده از Ensemble می‌تواند با ترکیب نقاط قوت مدل‌های مختلف، عملکرد کلی سیستم را بهبود دهد.
بهینه‌سازی هایپرپارامترها با استفاده از Search Grid نیز می‌تواند از انتخاب تصادفی پارامترها جلوگیری کرده و به یافتن تنظیمات مناسب‌تر برای مدل کمک کند.

۱۱. نتیجه‌گیری

در این گزارش، مسئله تشخیص تراکنش‌های تقلبی با استفاده از روش‌های مختلف یادگیری ماشین بررسی شد.
مدل‌های درخت تصمیم، جنگل تصادفی و XGBoost از جمله روش‌های مورد بررسی بودند.
همچنین برای یافتن تنظیمات مناسب مدل، روش Grid Search همراه با Cross-Validation K-Fold Stratified مورد استفاده قرار گرفت. score۱F- به عنوان معیار اصلی انتخاب بهترین ترکیب هایپرپارامترها در نظر گرفته شد.
نتایج نشان می‌دهد که در داده‌های نامتوازن، معیارهایی مانند Recall و PR-AUC برای ارزیابی مدل اهمیت ویژه‌ای دارند.
همچنین ترکیب مدل‌های مختلف در قالب روش‌های Ensemble می‌تواند راهکاری مناسب برای افزایش توانایی سیستم در تشخیص نمونه‌های تقلبی باشد.
در ادامه می‌توان با بهینه‌سازی هایپرپارامترها، انتخاب آستانه مناسب و استفاده از روش‌های پیشرفته‌تر، Ensemble عملکرد سیستم را بهبود داد.

مراجع

[۱] Breiman, L. "Random Forests," Machine Learning, ۲۰۰۱.

[۲] Chen T. and Guestrin, C. "XGBoost: A Scalable Tree Boosting System," Proceedings of the ۲۲nd ACM SIGKDD, ۲۰۱۶.

[۳] Han, J. and Kamber M. Pei, J. Data Mining: Concepts and Techniques.