ترکیب روش های تجمیعی داده کاوی برای کشف تراکنش های تقلب در کارت های اعتباری

سعید بختیاری*، زهرا نصیری، سید محمد صادق حجازی

[۱] گروه فتا، دانشکده اطلاعات، دانشگاه امین، تهران ایران

[۲] گروه کامپیوتر، دانشکده فنی و مهندسی، موسسه آموزش عالی آل طه، تهران، ایران

[۳] گروه کامپیوتر، دانشکده فنی و مهندسی، موسسه آموزش عالی پردیسان، مازندران، ایران

چکیده

استفاده از کارت‌های اعتباری، جهت پرداخت آسان پول از طریق تلفن همراه، اینترنت، دستگاههای خودپرداز و غیره روز به روز گسترده‌تر می شود. در کنار محبوبیت استفاده از کارت‌های اعتباری، مشکلات امنیتی مختلفی مانند تقلب وجود می آید. همان‌طور که روش‌های امنیتی بروز می‌شوند، متقلبان نیز روش‌های خود را بروز می‌کنند که این امر موجب نگرانی بانک‌ها و مشتریان آنها می‌شود. به همین دلیل راه حل های مختلفی جهت تشخیص، پیش‌بینی و پیشگیری از تقلب در کارت های اعتباری حائز اهمیت می‌باشند. یکی از راه‌حل‌ها روش داده‌کاوی و یادگیری ماشین است که افزایش دقت و کارایی یکی از با اهمیت ترین مسائل در این زمینه می باشد. در این مقاله روش‌های Gradient Boosting که زیر مجموعه روش‌های تجمیعی و یادگیری ماشین هستند را بررسی کرده، با اعمال مهندسی ویژگی و با ترکیب روش‌ها نرخ خطا را کاهش و دقت تشخیص را بهبود می‌دهیم. در روش پیشنهادی دو الگوریتم LightGBM و XGBoost را با برخی روش های متداول دیگر مقایسه کرده، سپس آنها را با استفاده از روش‌های تجمیعی میانگین‌گیری ساده و وزن‌دار ترکیب نموده و در نهایت مدل‌ها بوسیله معیارهای AUC و Recall و F1-score و Precision و Accuracy ارزیابی شده اند. مدل پیشنهادی پس از اعمال مهندسی ویژگی با استفاده از روش میانگین‌گیری وزن‌دار به ترتیب برای روش‌های ارزیابی مذکور به اعدادی معادل ۹۵/۰۸، ۹۰/۵۷، ۸۹/۳۵، ۸۸/۲۸ و ۹۹/۲۷ رسیده است. بر این اساس مهندسی ویژگی و میانگین‌گیری وزن‌دار تاثیر به سزایی در بهبود دقت پیش‌بینی و شناسایی داشتند.

واژگان کلیدی: تشخیص تقلب، کارت اعتباری، یادگیری تجمیعی، داده کاوی

Combination of Ensemble Data Mining Methods for Detecting Credit Card Fraud Transactions

Saeid Bakhtiari*, Zahra Nasiri, Mohsen Yazdinejad and Seyed Mohammad Sadegh Hejazi

[1] Department of FATA, faculty of engineering, Amin University, Tehran, Iran

[2] Department of Computer Engineering, faculty of engineering, Ale-Taha Institute of Higher Education, Tehran, Iran

[3] Department of Computer Engineering, faculty of engineering, Pardisan Institute of Higher Education, Mazandaran, Iran

Abstract

As we know, credit cards speed up and make life easier for citizens and bank customers. They can use it anytime and anyplace according to their personal needs, instantly, quickly without worrying about carrying a lot of cash with more security. Together, these factors make credit cards one of the most popular forms of online banking. This reason has led to widespread and increasing use for easy payment for purchases made through mobile phones, the Internet, ATMs, and so on. Despite the popularity and ease of payment with credit cards, various security problems are increasing day by day. One of the most important and constant challenges in this field is fraud detection in credit card transactions all around the world. Due to the increasing

* Corresponding Author

* نویسنده عهده‌دار مکاتبات

security issues in credit cards, fraudsters are also updating themselves. In general, as the popularity of using credit cards grows, more fraudsters are attracted to it, and credit card security comes into play. So naturally, this worries banks and their customers around the world. Meanwhile, financial information acts as the main factor in market financial transactions. For this reason, many researchers have tried to prioritize various solutions for detecting, predicting, and preventing credit card fraud in their research work and provide essential suggestions that have been associated with significant success. One of the practical and successful methods is data mining and machine learning. One of the most critical parameters in fraud prediction and detection in these methods is fraud detection accuracy. This research intends to examine the Gradient Boosting methods, such as LightGBm and XGBoost, a subset of Ensemble Learning and machine learning methods. By combining these methods, we can identify credit card fraud transactions, reduce error rates, and improve the detection process, which in turn increases efficiency and accuracy. This study compared some typical methods like Random Forest, Logistic Regression, and Navie base with LightGBM and XGBoost algorithms. In this paper, we proposed to merge LightGBM and XGBoost using simple and weighted averaging techniques and then evaluate the models using AUC, Recall, F1-score, Precision, and Accuracy. The proposed model provided values of 95.08, 90.57, 89.35, 88.28, and 99.27, respectively. In addition, we developed features by feature engineering techniques and then applied the feature engineering phase to the models. The results show that applying the feature engineering phase to the weighted average approach significantly improved prediction and detection accuracy.

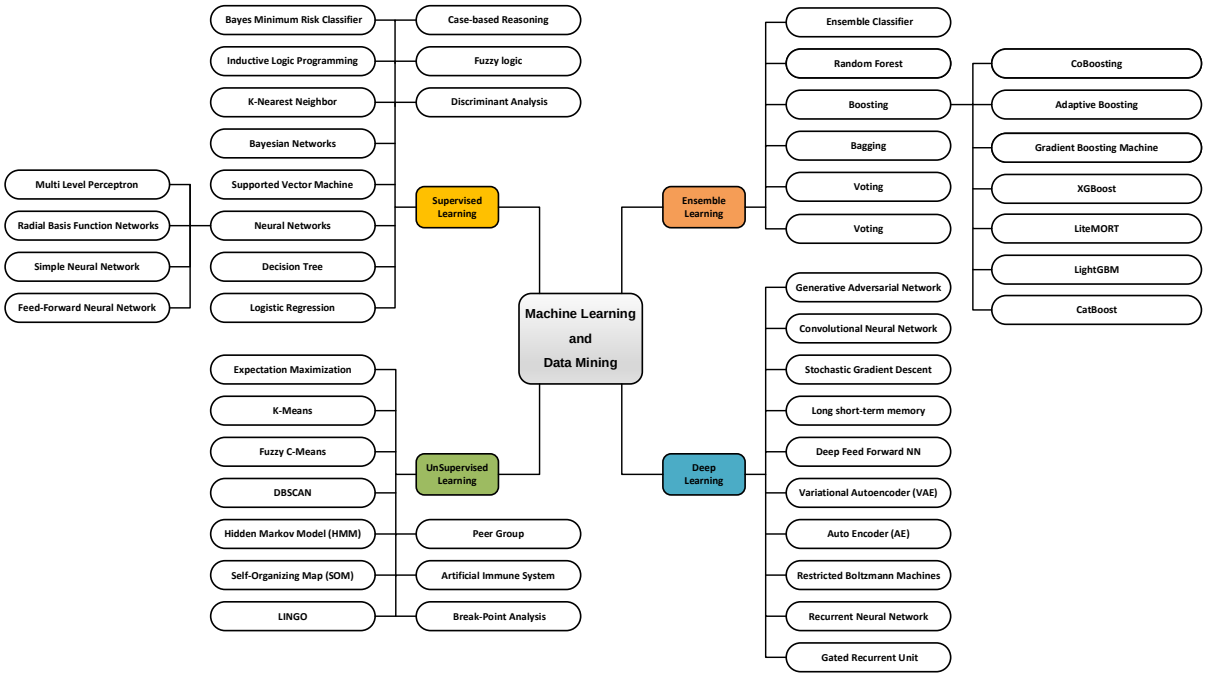
Keywords: **Fraud Detection, Credit Card, Ensemble Learning, Data Mining**

۱- مقدمه

در ایالات متحده، چهار شـــبکه پردازش کارت اعتباری عمده وجـود دارد: **Visa Card**، **Master Card**، **Discover** و **American Express**. شـــبکه های پردازشـــی تعیین کننده دسـتورالعمل های پردازش کارت اعتباری و تسـهیل تراکنش ها برای مشـتریان می باشـند. **Visa Card** و **Master Card** خود کارت اعتباری‌شـان را صـادر نمی کنند، هر زمان کارت اعتباری **Visa** یا **Mastercard** را مشـــاهده کردید بدانید بلنک دیگری وجود دارد که کارت اعتباری را صـــادر می کند. از طرف دیگر شبکه های **American Express** و **Discover** اغلبِ کارت های اعتباری‌شـان را خود صـادر می کنند ولی به بانک های دیگر هم اجازه صدور کارتشان را می دهند. شما می توانید از همه شبکه های پردازشی استفاده کنید اما این را به خاطر داشته باشید در خارج از ایالات متحده، تاجران کمتری پیدا می شـــوند که **American Express** و **Discover** را قبول کننـد. بنـابراین بـا کارتهای متصـل به این نوع شـبکه ها، بیشـتر مشـکل خواهید داشـت. تعدادی از شـبکه های پردازش اصـلی نیز تراکنش های کارت‌بدهی[1] را پردازش می کنند. کارتهای بدهی نیز عملکردی مشـــابه کارتهای اعتباری دارند، اما یک تمایز عمده بین این دو این اسـت که معاملات کارت بدهی به جای یک خط اعتباری از حســـاب بانکی مصـــرف کننده تأمین می شـــود. اغلب این نوع تراکنش ها در ایالات متحده توسط شبکه های ویزا و مسترکارت انجام می شود. بانک ها با ویزا و مسترکارت شریک می شوند تا کارت های بدهی را در دسترس مشتریان خود قرار دهند **[1]**. با افزایش حجم معاملات تجارت الکترونیکی، کلاهبرداری کارت اعتباری روز به روز گسـترده تر می شـود. با تبدیل شـدن تجارت الکترونیکی به جریان اصـــلی و افزایش چند برابری معاملات آنلاین، خطرات امنیتی مرتبط با آن ها به نگرانی های اسـاسـی تبدیل شـده اند. الگوی کلاهبرداری مالی نیز با توسـعه فناوری مدرن تغییر می کند و به ســـرعت افزایش می یابد که برعکس باعث افزایش سطح تقلب در معاملات کارت اعتباری و خسارات زیادی می شـــود. کشـــف تقلب در کارت های اعتباری همواره پیچیده اســـت زیرا رفتار کاربران ثبات ندارد و این تغییراتِ رفتاری پیچیدگی فرآیند را بیشتر می کند.

پیشگیری از تقلب و کشف تقلب هر دو روش مقابله با تقلب هستند. در پیشگیری از تقلب، هدف اصلی جلوگیری از تقلب و تراکنش های غیرمجاز اسـت. در حالی که در کشـف تقلب، هدف تشـــخیص تراکنش های متقلب از تراکنش های قانونی اسـت. در سـالهای اخیر چندین مطالعه از تکنیک های مختلف داده کاوی برای یافتن راه حل هایی برای این مشـکل اسـتفاده کرده اند. این تکنیک ها مبتنی بر شـبکه عصـبی[2]، یـادگیری عمیق[3]، الگوریتم ژنتیـک، مـدل مـارکوف پنهـان[4]، شـبکه بیزی[5]، درخت تصـمیم[6]، روش خوشـه بندی[7]، سـیسـتم ایمنی مصـــنوعی[8]، ماشـــین بردار پشـــتیبان[9] و داده کاوی که

[1] Debit Card
[2] Neural networks
[3] Deep Learning
[4] Hidden Markov Model
[5] Bayesian Network
[6] Decision Tree
[7] Clustering
[8] Artificial Immune Systems
[9] Support Vector Machine



شکل ۱. روش های یادگیری ماشین

Figure 1. Methods of Machine Learning

شامل طبقه بندی[1]، خوشه بندی، پیش بینی، شناسایی نقاط دور افتاده[2] و رگرسیون[3] می باشد است. یادگیری ماشین کاربردی از هوش مصنوعی است که شامل ابزار مختلف یادگیری است. یادگیری را می توان : یادگیری تحت نظارت[4]، یادگیری بدون نظارت[5] و یادگیری نیمه نظارت[6] قرار داد. الگوریتم های یادگیری ماشینی که معمولاً مورد استفاده قرار می گیرند شامل طبقه بندی بیزی[7]، طبقه بندی درخت تصمیم[8]، رگرسیون خطی[9]، رگرسیون لجستیک[10] و غیره است. شکل ۱ روش های اصلی داده کاوی و یادگیری ماشین را به صورت اجمالی نشان می دهد**[2]** .

در سال ۲۰۲۰ هوانگ **[3]** با استفاده از انتخاب ویژگی[11] در مدل نظارت شده، مدل های آماری خطی و غیرخطی و مدل های یادگیری ماشین از قبیل شبکه عصبی، رگرسیون لاجستیک، درخت تقویت شده[12]، جنگل تصادفی[13] که برمبنای داده های دارایی نیویورک و داده های تراکنش کارت اعتباری است تحقیق خود را ارائه داده است. در ژانویه سال ۲۰۱۶ گانین و همکاران **[4]** یک روش یادگیری نماینده جدید را برای سازگاری دامنه معرفی می کند، این روش در چارچوب معماری شبکه عصبی پیاده سازی می شود. در مقاله **[5]** یک روش شبکه عصبی با استفاده از ۱۰ لایه عمیق انکودر خودکار[14] ارائه شده است و یک مقایسه دقیق با طبقه بندی کننده های مختلف کلاسیک، یعنی درخت تصمیم گیری، ماشین بردار پشتیبان و طبقه بندی گروه های مختلف انجام شده است و از این طبقه بندی ها در مجموعه داده های ناهنجاری تراکنش کارت اعتباری استفاده کرده اند. در **[6]** لبیچات و همکاران با استفاده از استراتژی های سازگاری دامنه در سیستم های کشف تقلب مبتنی بر تراکنش را مورد بررسی قرار دادند. عبدالرضا و همکاران **[7]**، روش های مختلف کشف تقلب را مورد بررسی و طبقه بندی قرار دادند و محدودیت های عمده و دلایل عدم کارآیی روش ها را نیز مطرح کردند.. توسعه روش های مهندسی ویژگی خودکار[15] به یک مسئله مهم برای بهبود کارآیی مدل ها تبدیل شده است. در **[8]** لوکاس و همکاران یک مدل مبتنی بر مهندسی ویژگی خودکار برای کشف تقلب در تراکنش

---

[1] Classification
[2] Outlier
[3] Regression
[4] Supervised
[5] Unsupervised
[6] Semi-Supervised
[7] Bayesian Classifiers
[8] Decision Tree Classifier
[9] Linear Regression
[10] Logistic Regression
[11] Feature Selection
[12] Boostrd Trees
[13] Random Forest
[14] Deep Auto-encoder
[15] Automated Feature Engineering

کارت های اعتباری پیشـــنهاد دادند. اســـتراتژی مهندســـی ویژگی مبتنی بر مدل مارکوف پنهان اســـت. در **[9]** یک مطالعه مقایســـه ای از تکنیک های مبتنی بر شـــبکه های عصـبی، که به مجموعه داده ها اعمال می شـود، انجام شـده است. در **[10]** سایا و کارتا با هدف بررسی مزایای مربوط به اتخاذ استراتژی های کشف تقلب پیشگیرانه[1]، به جای روش های برگشـــت پذیر متعارف، به راه حل هایی که میتولند به ســـمت اجرای مؤثر عملی منجر شـــود پرداختند. کیم و همکاران در **[11]** با در نظر گرفتن عدم توازن نمونه ها در تشـخیص تقلب یک اسـتراتژی یادگیری جدید با نام قهرمان چالشگر[2] با استفاده از یک روش ترکیبی یادگیری تجمیعی[3] و یادگیری عمیق بر روی کشف تقلب در کارت های اعتباری را با استفاده از داده های واقعی انجام داده و پیشنهاد کردند. در **[12]** یک سیستم مبتنی بر یادگیری ماشین تجمیعی به منظور کشف خطر اعتبار که در آن تاجرین با کد های **MCC**[4] نادرست روی سایر قابلیت اطمینان[5] سیستم های امتیازدهی[6] تأثیر میگذارند و می توانند ضـــررهایی را برای بانک ها و سـازمان های صـاحب کارت وارد کنند ارائه دادند. مطالعات مختلفی که در زمینه پیش‌بینی الگوی جرم در گذشـته انجام شـده است نشـان می دهند جرم یک الگوی جغرافیایی را در فضا و زمان نشان می دهد. با بیان این نظریه توسط حاجلا و همکاران در ژانویه ۲۰۲۰ در **[13]** یک رویکرد جدید مبتنی بر خوشـه بندی برای شـناسـایی نقاط مهم[7] برای دسـته های مختلف جرم با اسـتفاده از وقایع تاریخی به عنوان شـاخص و یـک تکنیـک پیش بینی جرم زمـانی و مکـانی مبتنی بر یادگیری ماشـــین همراه با تحلیل نقاط مهم دو بعدی ارائه شـــده اســـت. در ژانویه ۲۰۲۰، راتول و همکاران در **[14]** تحقیقی در مورد جرم و جنایت بر اساس یادگیری ماشین و داده کاوی انجام داد. در این پژوهش از بین الگوریتم های یادگیری تجمیعی علت اســـتفاده ما از الگوریتم های درخت تصـمیم با شـیب تقویت شـده انعطاف پذیری و دقت پیش بینی مناسـب و بالا می باشـد. در بین این الگوریتم ها، روش هایی وجود دارد که سازگار با مقادیر گم شده[8] نیز می باشند. الگوریتم هـای **LightGBM** و **XGBoost** دارای کـارایی مناسب تری نسبت به روش هایی چون **Lasso**، **PCA** و... در انتخاب ویژگی هستند و نسبت به ماشین بردار پشتیبان نیز سـریعتر می باشـند. اما در عین حال با وجود عملکرد بسـیار مناسـب نسـبت به سـایر روش ها نیازمند حافظه بیشـتری نیز هســـتند. در این پژوهش با مقایســـه و ترکیب دو الگوریتم **LightGBM** و **XGBoost** قصد داریم نرخ خطا را کم و دقت را بالا ببریم و همچنین میزان مصـــرف حافظه را تا جایی که ممکن است کاهش دهیم.

۲- کارهای مرتبط

طی دو دهه گذشته سیستم های تجمیعی از رشد فزاینده ای در جامعه هوش محاســـباتی و جامعه یادگیری ماشـــین برخوردار بوده انـد و کـاملا ســـزاوار این توجـه بوده انـد. سـیسـتمهای تجمیعی ثابت کرده اند که در طیف وسـیعی از چالشـــها و کاربردهای دنیای واقعی بســـیار کارآمد و متنوع هســـتند. در اصـــل برای کاهش واریانس -در نتیجه بهبود دقت- در یک سیستم تصمیم گیری خودکار، ایجاد شده اند، سـیسـتم های تجمیعی با موفقیت برای رفع انواع مشـکلات یادگیری ماشـــین مانند انتخاب ویژگی، تخمین اطمینان[9]، ویژگی‌ها با مقادیر گم شـــده، اصـــلاح خطا[10]، داده های نامتوازن و غیره توسعه یافته اند. **[15]**

در این قسـمت تحقیقات انجام شده روی شـناسـایی و پیشـــگیری تقلب با روش های تجمیعی مورد بررســـی قرار گرفته اسـت: در ژوئن ۲۰۲۰ گوتیرز-اسـپینوزا و همکاران در **[16]** کشـــف بازدیدهای جعلی از طریق یادگیری تجمیعی با بررسی مجموعه داده رستوران، تشخیص جعلی بودن را ارائه می دهد. در فوریه ۲۰۲۰ الطیب الطاهر و همکاران در **[17]** یک روش هوشـــمندانه کشـــف تقلب تراکنش کارت های اعتباری را با اسـتفاده از روش **LightGBM** بهینه شـده ارائه داد. در آوریل ۲۰۲۰ آریا و همکاران در **[18]** یک چارچوب یـادگیری تجمیعی عمیق برای کشـــف تقلب کـارت هـای اعتباری جریان داده ایی زمان واقعی[11] با اســـتفاده از روش تجمیعی درخت اضافی[12] همراه با یادگیری عمیق برای بهبود دقت پیش بینی قطعی و اجتناب از بیش برازش با تراکنش های داده های واقعی از یک بانک بزرگ را ارائه داده است. در ژانویه ۲۰۲۰، باگا و همکاران در **[19]** کشـــف تقلب کارت

---

[1] Adopting Preventive Fraud Detection Strategies
[2] Champion-challenger
[3] Ensemble Learning
[4] Merchant Category Code(MCC)
[5] Reliability
[6] Scoring Systems
[7] Hotspot Identification
[8] Missing Values
[9] Reliability Estimation
[10] Error Correction
[11] Real-time Dataflow
[12] Extra Tree Ensemble method

اعتباری با استفاده از خطوط لوله و یادگیری تجمیعی را ارائه دادند. کوماری و همکاران در **[20]** به چند طبقه بندی کننده گروهی[1] مانند **Bagging**، جنگل تصادفی، طبقه بندی از طریق رگرسیون و... پرداخته و آنها را با برخی از طبقه بندی های منفرد و مؤثر مانند **K**-نزدیکترین همسایه، شبکه های بیزی، ماشین بردار پشتیبان، طبقه بند **RBF**، پرسپترون چند لایه، درخت تصمیم مقایسه کرده اند.

IEEE-CIS در زمینه های مختلف هوش مصنوعی و یادگیری ماشین از جمله شبکه های عصبی عمیق، سیستم های فازی و محاسبات تکاملی کار می کند. در این قسمت تعدادی از تحقیقات انجام شده روی دیتاست تشخیص تقلب **IEEE-CIS** مورد بررسی قرار گرفته است: در آوریل ۲۰۲۰ ناجادات و همکاران در **[21]** تحقیق بر روی مجموعه داده های تشخیص تقلب **IEEE-CIS** توسط **Kaggle** با استفاده از مدل های یادگیری ماشین و یادگیری عمیق انجام شده است و مدل جدیدی که مبتنی بر **BiLSTM** و **BiGRU** بود ارائه شده است. همچنین مقایسه شش طبقه بندی کننده یادگیری ماشین شامل: شبکه های بیزی، **Voting**، **Ada Boosting**، جنگل تصادفی، رگرسیون لوجستیک و نتایج حاصل از طبقه بندی کننده های یادگیری ماشین نشان می دهد. در آگوست ۲۰۱۹ گوسویسکا و همکاران در **[22]** با بیان این مسئله که مهمترین بخش انتخاب مدل و تنظیم هایپرپارامتر ارزیابی عملکرد مدل است و با بیان ضعف های مشترک مشهورترین معیارها، مانند **AUC**، **F1-Score**، **ACC** برای طبقه بندی باینری معیار **MAD** و از **RMSE** برای رگرسیون، یا **cross-entropy** برای طبقه بندی چند لایه به منظور حل مسئله روش جدید **EPP**[2] (قدرت پیش بینی کننده مبتنی بر **Elo**) را ارائه می دهد. در مارس ۲۰۲۰ ژلنگ و همکاران در **[23]** یک مدل تشخیص تقلب معامله مبتنی بر **XGBoost** با مهندسی ویژگی و تجسم بر روی مجموعه داده های در رقابت تشخیص تقلب **IEEE-CIS**، **Kaggle** را ارائه دادند. در آوریل سال ۲۰۲۰ دینگ لینگ و همکاران در **[24]** با ارائه مدل مبتنی بر **LightGBM** بر روی مجموعه داده تراکنش تشخیص تقلب **IEEE-CIS**، **Kaggle** نشان دادند.

۳- پیش‌زمینه

۳-۱- الگوریتم های تقویت کننده[3]

تقویت کننده یک طبقه بندی کننده قوی[4] مبتنی بر مجموعه داده های آموزش داده شده طبقه بندی کننده های ضعیف[5] است، و یکی از الگوریتم های موفق برای یادگیری نظارت شده است**[25]** . روشی برای تبدیل مجموعه یادگیرنده های ضعیف به یادگیرنده های قوی است. یک یادگیرنده ضعیف دارای خطای کمتر از ۰/۵ و یادگیرنده قوی دارای خطای نزدیک به ۰ است. خانواده ای از یادگیرندگان ضعیف در کنار هم جمع می شوند تا یک یادگیرنده قوی را تشکیل دهند. سه الگوریتم تقویت کننده که زیاد استفاده می شود: **AdaBoost**، **Gradient Boost** و **XGBoost** هستند**[26]** .

درخت تصمیم با گرادیان تقویتی[6] یک الگوریتم تقویت کننده است که توسط فریدمن ارائه شده است، این الگوریتم از چندین درخت تصمیم تشکیل شده است و برای تولید هر درخت از روش نزول گرادیان[7] استفاده می شود. بر اساس تمام درخت های تصمیم منفرد، بهینه سازی با به حداقل رساندن تابع ضرر[8] به عنوان هدف انجام می شود**[27]** . در گرادیان تقویتی، بسیاری از مدل ها به صورت پیوسته آموزش داده می شوند. هر مدل جدید با استفاده از روش نزول گرادیان به تدریج تابع ضرر را به حداقل می رساند. این مدل پیوسته متناسب با مدل های جدید، تخمین دقیق تری از متغیر پاسخ ارائه می دهد. در اصل این الگوریتم از الگوریتم های چندگانه ضعیف برای تولید الگوریتمی دقیق تر استفاده می کند. استفاده از الگوریتم های گرادیان تقویتی بیشتر بخاطر دقت بالای آنها است**[26]** .

نحوه کار GBM[9] [28]

گرادیان[10] به خطا یا باقیمانده ای گفته می شود که پس از ساخت یک مدل بدست آمده است. تقویت به بهبود اشاره دارد. این روش به عنوان ماشین تقویت گرادیان یا **GBM** شناخته می شود. تقویت گرادیان راهی برای بهبود (کاهش) خطای تدریجی است. برای دیدن نحوه کار **GBM**، فرض کنید یک مدل **M** (که براساس درخت تصمیم است) داریم و

---

[1] Group Classifier
[2] Elo-based Predictive Power(EPP)
[3] Boosting
[4] Strong Classifier
[5] Weak Classifier
[6] Gradient Boosting Decision Tree(GBDT)
[7] Gradient Descent method
[8] Loss Function
[9] Gradient Boosting Machine(GBM)
[10] Gradient

الگوریتم ۱. الگوریتم آموزش LightGBM
Algorithm 1. The training of LightGBM

ورودی: مجموعه آموزشی $\{(x_i . y_i)\}_{i=1}^{N}$

خروجی: بردار $\hat{y}_i^{(t)}$ مدل LightGBM

گام ۱. اولین درخت را با مقدار ثابت مقداردهی اولیه می کنید:

$$\hat{y}_i^{(0)} = f_0 = 0$$

گام ۲. با کمینه سازی تابع هزینه، درخت بعدی را آموزش دهید:

$$f_t(x_i) = \arg\min_{f_t} L\left(y_i , \hat{y}_i^{(t-1)} + f_t(x_i)\right)$$

گام ۳. مدل بعدی را به صورت افزودنی بدست آورید:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i)$$

گام ۴. مراحل ۲ و ۳ را تکرار کنید تا شرط توقف برقرار شود.

گام ۵. مدل اولیه را بدست آورده و به خروجی ارسال کنید:

$$\hat{y}_i^{(t)} = \sum_{t=0}^{M-1} f_t(x_i)$$

Require: input: Training set $\{(x_i . y_i)\}_{i=1}^{N}$

Ensure: output: LightGBM model $\hat{y}_i^{(t)}$

Step 1. Initialize the first tree as a constant:

$$\hat{y}_i^{(0)} = f_0 = 0$$

Step 2. Train the next tree by minimizing the loss function:

$$f_t(x_i) = \arg\min_{f_t} L\left(y_i , \hat{y}_i^{(t-1)} + f_t(x_i)\right)$$

Step 3. Get the next model in an additive manner:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i)$$

Step 4. Repeat the Step 2 and Step 3 until the model reaches the stop condition.

Step 5. Obtain and return the final model:

$$\hat{y}_i^{(t)} = \sum_{t=0}^{M-1} f_t(x_i)$$

جدول ۱. مدل GBM
Table1. GBM Model

گام ها	عملیاتها
Step 1:	$Y = M(x) + error$
Step 2:	$error = G(x) + error2$
Step 3:	$error2 = H(x) + error3$
Step 4:	$Y = M(x) + G(x) + H(x) + error3$

می خواهیم آن را بهبود ببخشیم. مدل را به شرح زیر بیان می کنیم:

همانطور که در جدول ۱ مشاهده می کنید در مرحله۱ **Y** متغیر وابسته است و **M (x)** درخت تصمیم با استفاده از متغیرهای مستقل **x** است. اکنون می خواهیم خطای درخت تصمیم قبلی را پیش بینی کنیم. مرحله ۲ **G (x)** درخت تصمیم دیگری است که سعی می کند خطا را با استفاده از متغیرهای مستقل **x** پیش بینی کند. در مرحله ۳، مشابه مرحله قبل، مدلی ایجاد می کنیم که سعی می کند **error2** را با استفاده از متغیرهای **x** مستقل پیش بینی کند و در نهایت در مرحله ۴ همه با هم ترکیب می شوند.

LightGBM-۳-۲

LightGBM یکی از الگوریتم های تقویت گرادیان است که بر اساس الگوریتم درخت تصمیم گیری است که برای طبقه بندی و بسیاری از کارهای دیگر یادگیری ماشین مورد استفاده قرار می گیرد. این الگوریتم کپسوله ای از انواع داده را فراهم می کند که موجب کاهش مصرف حافظه بر روی اشیاء داده مانند **Numpy**، **Pandas**، **Array** و... می شود. دلیل این امر این است که فقط باید هیستوگرام گسسته را ذخیره کرد. آموزش پیش فرض درخت تصمیم در **LightGBM** استفاده از الگوریتم هیستوگرام است. این گزینه در **XGBoost** نیز موجود است، اما با مقادیر پیش فرض ویژگی های از قبل مرتب شده می باشد. این الگوریتم تنها از الگوریتم های مبتنی بر درخت استفاده می کند. **LightGBM** علاوه بر دقت، دارای کارایی بسیار بالا نیز می باشد**[27]** . این الگوریتم بر اساس الگوریتم های درخت تصمیم گیری استوار است، در حالی که الگوریتم های تقویت کننده دیگر عمق یا سطح درخت را تقسیم می کنند، این الگوریتم برگ درخت را با بهترین تناسب تقسیم می کند. بنابراین وقتی همان برگ در **LightGBM** رشد می کند، الگوریتم های **leaf-wise** می تواند باعث کاهش تلفات بیشتر از الگوریتم **level-wise** شود و از این رو دقت بسیار بهتری حاصل می شود که به ندرت توسط هر یک از الگوریتم های تقویت کننده موجود می توان به دست آورد. در مقایسه با الگوریتم های معمول یادگیری ماشین دارای مزایایی چون آموزش سریع تر، مصرف حافظه کمتر، یادگیری موازی، پردازش داده در مقیاس بزرگ و... است. همچنین **LightGBM** از بسط تیلور تابع هزینه و شروط تنظیم برای کنترل پیچیدگی مدل استفاده می نماید.

الگوریتم ۲. الگوریتم آموزش XGBoost

Algorithm 1. The training of XGBoost

Require: input: Training set $\{(x_i . y_i)\}_{i=1}^{N}$

ورودی: مجموعه آموزشی $\{(x_i . y_i)\}_{i=1}^{N}$

Ensure: output: XGBoost model $\bar{y}_i^{(t)}$

خروجی: بردار $\bar{y}_i^{(t)}$ مدل XGBoost

Step 1. Initialize the first tree as a constant:

گام ۱. اولین درخت را با مقدار ثابت مقداردهی اولیه می کنید:

$$\bar{y}_i^{(0)} = f_0 = 0$$

Step 2. Train the next tree by minimizing the loss function:

گام ۲. با کمینه سازی تابع هزینه، درخت بعدی را آموزش دهید:

$$f_t(x_i) = \arg\min_{f_t} \left(\frac{1}{2} \left[\frac{\left(\sum_{i \in T_L} g_i\right)^2}{\sum_{i \in T_L} h_i + \lambda} + \frac{\left(\sum_{i \in T_R} g_i\right)^2}{\sum_{i \in T_R} h_i + \lambda} - \frac{\left(\sum_{i \in T} g_i\right)^2}{\sum_{i \in T} h_i + \lambda} \right] - \gamma \right)$$

$$g_i = \frac{\partial L\left(y_i , \bar{y}_i^{(t-1)}\right)}{\partial \bar{y}_i^{(t-1)}} \quad , \quad h_i = \frac{\partial^2 L\left(y_i , \bar{y}_i^{(t-1)}\right)}{\partial^2 \bar{y}_i^{(t-1)}}$$

Step 3. Get the next model in an additive manner:

گام ۳. مدل بعدی را به صورت افزودنی بدست آورید:

$$\bar{y}_i^{(t)} = \bar{y}_i^{(t-1)} + f_t(x_i)$$

Step 4. Repeat the Step 2 and Step 3 until the model reaches the stop condition.

گام ۴. مراحل ۲ و ۳ را تکرار کنید تا شرط توقف برقرار شود.

Step 5. Obtain and return the final model:

گام ۵. مدل اولیه را بدست آورده و به خروجی ارسال کنید:

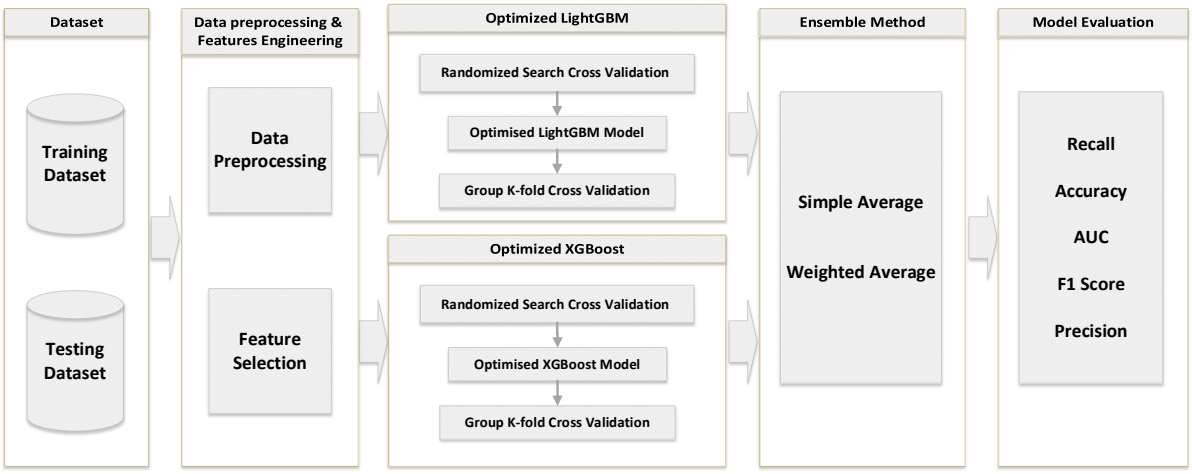
$$\hat{y}_i^{(t)} = \sum_{t=0}^{M-1} f_t(x_i)$$

XGBoost -۳-۳

XGBoost یکی از کارآمدترین روش های پیاده سازی درختان تصمیم گیری با گرادیان تقویتی است و به عنوان یکی از بهترین الگوریتم های یادگیری ماشین شناخته می شود. به طور خاص، این الگوریتم برای بهینه سازی استفاده از حافظه و بهره برداری از قدرت محاسبات سخت افزاری طراحی شده است، **XGBoost** با افزایش عملکرد نسبت به بسیاری از الگوریتم های یادگیری ماشین، زمان اجرا را کاهش می دهد. ایده اصلی تقویت، ساختن زیر درختانی از درخت اصلی است به طوری که هر درخت بعدی خطاهای درخت قبلی را کاهش می دهد. به این ترتیب، زیرشاخه های جدید باقیمانده های قبلی را به منظور کاهش خطای تابع هزینه، به روز می کنند.

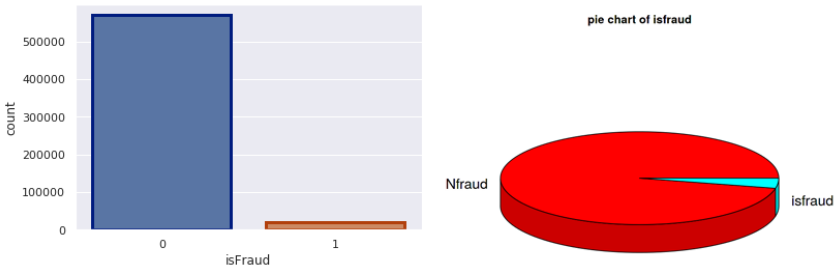
۴- روش پیشنهادی

در این مقاله داده های تراکنش و شناسایی را ادغام کرده و پس از مراحل پیش پردازش و حل چالش داده های گم شده، حل عدم توازن داده، کار بر روی داده های عددی و غیر



شکل ۲. دیاگرام جریان کار الگوریتم پیشنهادی

Figure 2. Proposed Method Flow Diagram



شکل ۳. نمایش عدم توازن متغییر وابسته "isfraud"

Figure 3. plots of "isfraud" unbalanced target variable

عددی و انجام مهندسی ویژگی به آموزش الگوریتم پرداختیم. همان طور که در شکل ۲ مشاهده می نمایید بر اساس الگوریتم **LightGBM** بهینه شده و الگوریتم **XGBoost** بهینه شده بوسیله تنظیم هایپرپارامترها ما مدل را آموزش می دهیم و سپس با استفاده از روشهای میانگین گیری ساده و وزن دار در یادگیری تجمیعی نتایج دو مدل را با روش با هم ترکیب می کنیم تا نتایج نهایی پیش بینی را بدست آوریم و ارزیابی کنیم.

جدول ۲. بهترین پارامترهای الگوریتم LightGBM

Table 2- The best parameters of the LightGBM algorithm

نام پارامتر	محدوده پارامترها	بهترین مقدار هر پارامتر
num_leaves	200 – 600	256
feature_fraction	0.3 – 0.6	0.5
bagging_fraction	0.3 – 0.7	0.4
min_data_in_leaf	40 – 140	80
max_depth	-1 , 5:11	-1
learning_rate	0.002 – 0.01	0.01
reg_alpha	0.01 : 1	0.01
reg_lambda	0.01 : 1	0.01

جدول ۳. بهترین پارامترهای الگوریتم XGBoost

Table 3- The best parameters of the XGBoost algorithm

نام پارامتر	محدوده پارامترها	بهترین مقدار هر پارامتر
max_leaves	50 – 400	72
min_child_weight	0 – 10	2
max_depth	-1 , 5	0
learning_rate	0.002 – 0.05	0.04
reg_alpha	0.01 : 1	0.01
n_estimators	800-1000	800
Subsample	0.7-0.9	0.74
colsample_bytree	0.5-1	0.89

۴-۱- فرآیند مدل سازی

پس از پیش پردازش داده ها، به ترتیب از الگوریتم **LightGBM** و الگوریتم **XGBoost** برای آموزش مدل استفاده می کنیم. مقادیر پارامترها از طریق روش **Randomize Search cross validation** و سعی و خطا با تغییر بازه مقادیر برای به دست آوردن مقادیر بهینه و بهینه سازی مدل به دست می آیند که در جدول ۲ و ۳ مشاهده می نمایید. سپس نتایج پیش بینی شده توسط مدل **LightGBM** و نتایج پیش بینی شده توسط مدل **XGBoost** هر دو وارد مدل تجمیعی می شوند. سرانجام، ترکیب نتایج پیش بینی ها به روش میانگین گیری انجام می شود.

۵- پیاده سازی

۵-۱- دیتاست

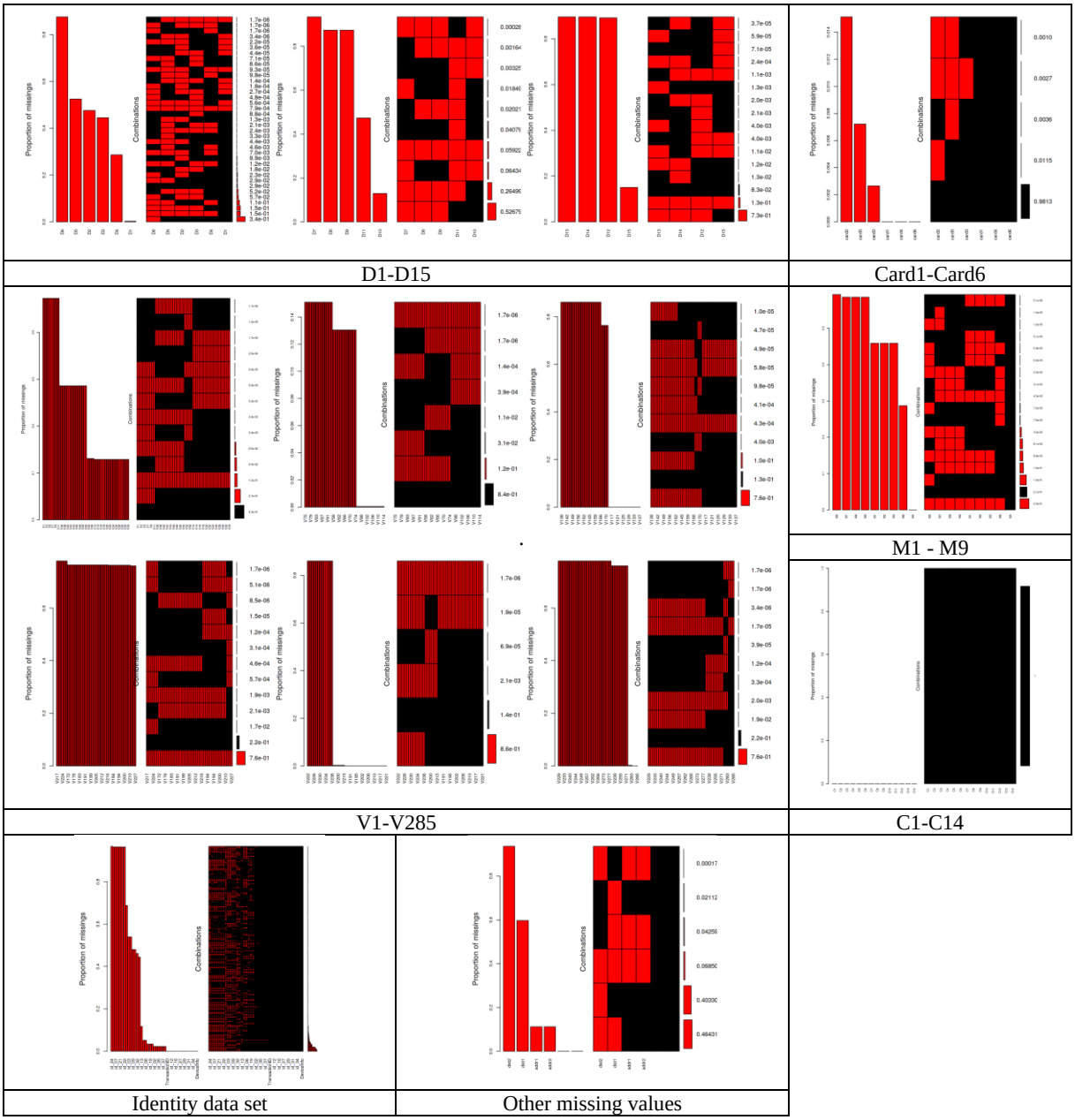
مجموعه داده کشف تقلب کارت های اعتباری **IEEE-CIS** مربوط به سری مسابقات کگل آدرس **https://www.kaggle.com/c/ieee-fraud-detection** است.

در زمینه های مختلف هوش مصنوعی و یادگیری ماشین کار می کند، از جمله شبکه های عصبی عمیق ، سیستم های فازی ، محاسبات تکاملی. امروز آنها با شرکت خدمات پرداخت برتر جهان، **Vesta Corporation** همکاری می کنند و به دنبال بهترین راه حل ها برای صنعت پیشگیری از کلاهبرداری هستند. شرکت وستا مجموعه داده این مسابقه را ارائه داده است. شرکت وستا پیشگام راه حل های پرداخت تضمینی تجارت الکترونیکی است. وستا در سال ۱۹۹۵ تاسیس شد و در روند معاملات پرداخت کاملاً تضمینی کارت غیر موجود (**CNP**) برای صنعت ارتباطات راه دور پیشگام بود. از آن زمان ، وستا بصورتی پایدار و محکم توانمندی های علم داده و یادگیری ماشین را در سرتاسر جهان گسترش داده و جایگاه خود را در راس پرداخت های تجارت الکترونیکی تقویت کرده است. امروز وستا معاملات سالانه بیش از ۱۸ میلیارد دلار را تضمین می کند.

مجموعه داده شـــامل چهار مجموعه که دو مجموعه داده **Train** و **Test** برای **Transaction data** و دو مجموعه داده **Train** و **Test** تست مربوط به **identity data** می باشد اســت. حجم بالای مجموعه داده و وجود ویژگی های بسـیار زیاد و متنوع اعم از عددی و غیرعددی باعث می شـــد به فضای بالایی برای حافظه نیاز داشـته باشـیم و از چالش های مهم کار محسوب می شد.

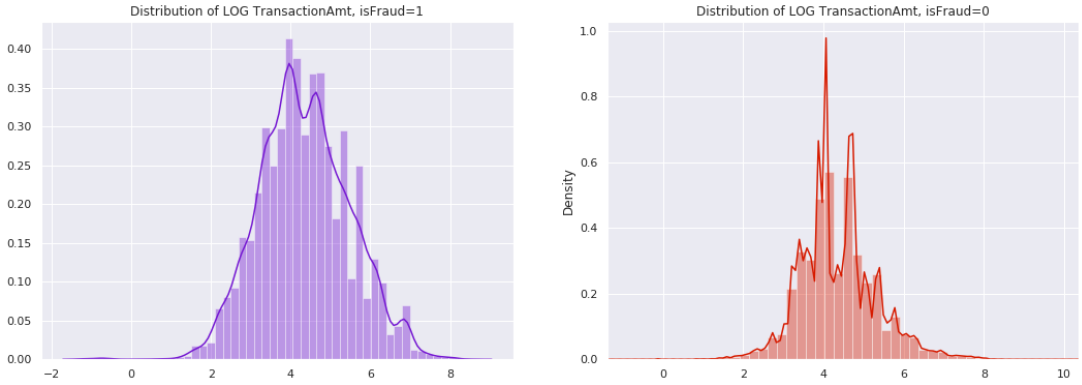
ویژگی **"isFruad"** برچسـب کلاس حاصـل اسـت که در صـورت تقلب معادل **"fruad"** و در صـورت سـالم بودن معادل **"Notfruad"** نشـــان داده می شـــود. همانطور که در شـکل ۳ دیده می شـود توزیع متغیر پاسـخ بسـیار نامتوازن اســـت. فقط ۳/۵٪ از تراکنش ها در مجموعه داده به عنوان تقلب تعیین شـــدند و بقیه به عنوان تراکنش های ســـالم شـــناخته می شـــوند که در نمودار **barplot** اعداد در محور عمودی نشـــان دهنده تعداد تراکنش ها و محور افقی (۰و۱) به ترتیب نشان دهنده تراکنش سالم و تقلب می باشد.

مجموعه داده دارای مقدار زیادی مقادیر گم شـده می باشـد، زمانی که دو مجموعه داده ترکیب می شـود در ۴۱۱ ویژگی از کل ۴۳۳ ویژگی مقادیر از دســـت رفته داریم. بیش از ۴۷٪ ویژگی ها بالای ۷۰٪ مقدار از دســـت رفته دارند.



شکل ۴. نمودار aggrplot از ویژگی های دارای مقادیر گم شده

Figure 4. aggrplot of missing values features



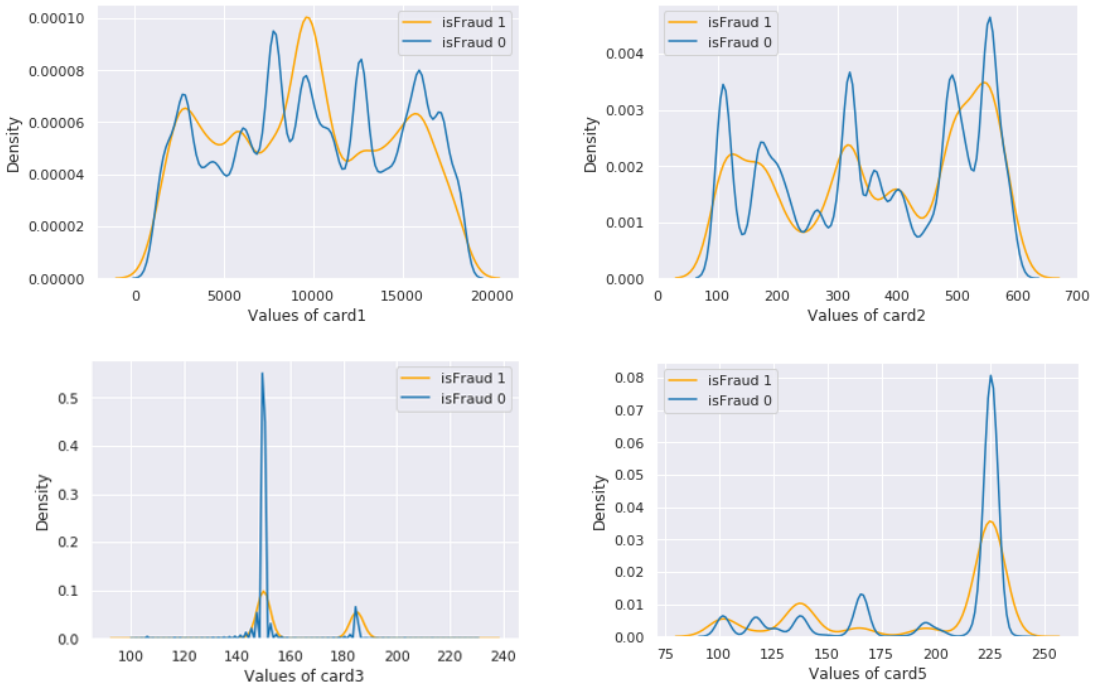
شکل ۵. نمودار توزیع لگاریتم ویژگی TransactionAMT

Figure 5. Distribution of log TransactionAMT

ویژگی هایی که بیش از ۹۹٪ مقدار از دست رفته داشته باشند را می توانیم به طور کامل حذف نماییم. در شکل ۴ مقادیر گمشده ویژگی ها را مشاهده می نمایید. به علت حجم بالا داده ها را تفکیک کرده و بررسی انجام شد. همانطور که مشاهده می کنید نمودار **aggregate plot** اطلاعات زیادی در مورد میزان داده های گم شده می دهد. رنگ سیاه شامل داده های مشاهده شده و رنگ قرمز شامل داده های گم شده می باشد. اعدادی که سمت راست ملاحظه می نمایید از تقسیم تعداد مشاهده هر حالت بر کل نمونه ها می باشد. برای درک بهتر مثالی را بیان می کنیم. در نمودار **other missing value** نشان می دهد در ۴۶/۴۳۱٪ کل نمونه ها **dist1** و **dist2** دارای مقادیر گم شده می باشند و باقی یعنی ویژگی های **addr1** و **addr2** دارای مقادیر مشاهده شده می باشند.

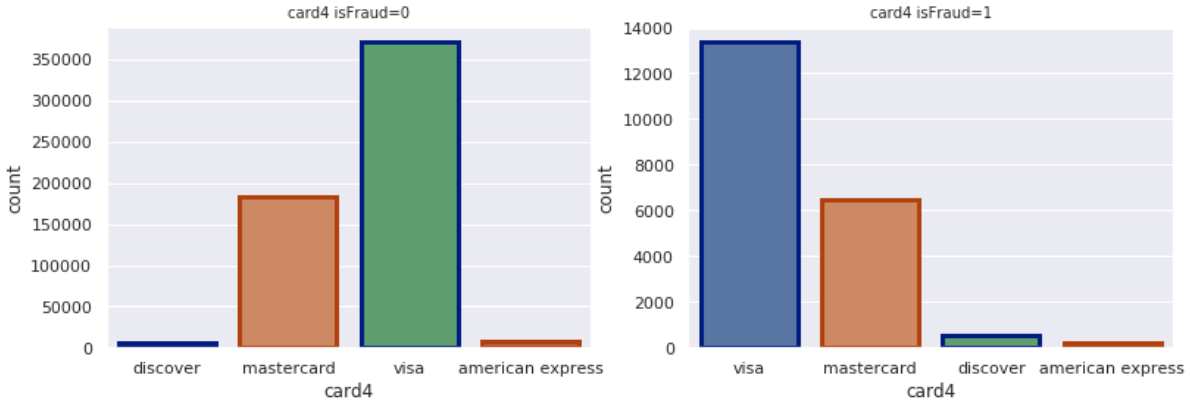
۲-۵- داده های تست و آموزش

مجموعه داده را روی یک لپ تاپ با پردازنده **intel core2Duo** و حافظه ۴ گیگابایت اجرا شد اما به علت بالا بودن حجم مجموعه داده ها اجرای الگوریتم ممکن نبود، بنابراین مجموعه داده ها را در فضاهای **Python** و **R** در بستر ابری کگل که ۱۶ گیگ رم و ۱۰۰ گیگ هارد در اختیار می گذاشت اجرا شده است. همچنین در این سرویس می توان در حلت های مختلف **TPU**، **GPU** و **CPU** الگوریتم را اجرا نمود که آزمایشات در حالت **GPU** انجام شده است.



شکل ۶. برخی از متغیرهای categorical، به علت داشتن مقادیر منحصر به فرد زیاد رفتاری مشابه متغیرهای عددی دارند.

Figure 6. numerical-like behavior of some categorical features (Card1, Card2, Card3, and Card5)



شکل ۷. مقدار تراکنش های نرمال و تقلب در شبکه پردازشی بر اساس ویژگی Card4

Figure 7. The amount of fraud and normal transactions in processing networks based on Card4

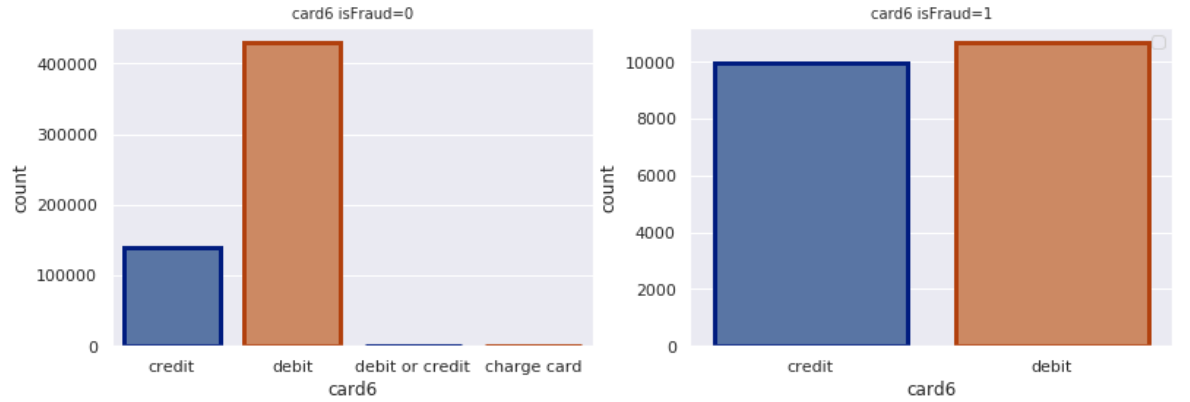
داده های آموزش در **identity dataset** شامل ۱۴۴۲۳۳ نمونه و داده های آزمون آن شامل ۱۴۱۹۰۷ نمونه و ۴۰ ویژگی است. داده های آموزش **Transaction dataset** شامل ۵۹۰۵۴۰ نمونه و ۳۹۳ ویژگی و داده های آزمون آن شامل ۵۰۶۶۹۱ نمونه است. در ادامه به بررسی برخی از ویژگی ها می پردازیم.

در شکل ۵ توزیع مبلغ تراکنش به صورت لگاریتم ویژگی (**TransactionAMT**) در نمودارها نشان داده شده است که با توجه به نمودارها می توان نتیجه گرفت که تراکم میانگین مبلغ در تراکنش های تقلب بیشتر از تراکنش های عادی است یعنی این میزان در مبالغ رو به بالا بیشتر دیده می شود و این نوعی هشدار برای جا به جایی با مبالغ بالا می باشد. در نتیجه کلاهبرداران تمرکز بیشتری روی مبالغ بالا دارند که بدین صورت امنیت آن نیز به همان میزان باید توسط شبکه های پردازشی و بانک ها تامین شود.

ویژگی های **card1** تا **card6** نشان دهنده اطلاعات کارت اعم از نوع کارت، گروه کارت، بانک صادر کننده کارت، کشور و... می باشند که از نوع **categorical** هستند. همانطور که در شکل ۶ مشاهده می نمایید، برخی از این ویژگیها به علت داشتن مقادیر منحصر به فرد بسیار، رفتاری مشابه مقادیر پیوسته دارند مانند **card1** و **card2** که به ترتیب دارای ۱۳۵۵۳ و ۵۰۰ مقدار منحصر به فرد می باشند. محور عمودی میزان تراکم و محور افقی مقادیر منحصر به فرد را نشان می دهد.

ویژگی **card4** شبکه پردازشی تراکنش ها را نشان می دهد. با توجه به شکل ۷ **visa card** و **master card** دارای بیشترین تراکنش تقلب و **discover** و **american express** دارای کمترین میزان تراکنش تقلب می باشند که به همان میزان در شبکه های پردازشی **visa card** و **master card** بیشترین تقلب صورت گرفته است. قسمت عمده ایی از آن به محبوبیت این شبکه ها برمی گردد که بیشتر از شبکه های دیگر استفاده می شوند.

ویژگی **card6** انواع کارت های اعتباری را نشان می دهد. همان طور که در شکل ۸ مشاهده می نمایید، میزان



شکل ۸. تعداد تراکنش ها بر اساس ویژگی Card6

Figure 8. The amount of transactions based on Card6

استفاده **debit** و بعد از آن **credit** به ترتیب بسیار بالا است. و به همان میزان تراکنش تقلب فقط در کارت های **debit** و **credit** صورت پذیرفته است که **debit card** دارای تقلب بیشتری نسبت به **credit card** می باشد ولی با این وجود میتوان گفت **credit card** با توجه به حجم کمتر تراکنش ها نسبت به **debit card** بیشتر در معرض خطر میباشد و شاید این بدین معنی باشد که **debit card** پروتکل های امنیتی بیشتری را رعایت می کند.

۳-۵- بحث

همانطور که گفته شد از معیارهای مختلفی استفاده شده است اما در مورد **AUC** توضیحاتی بیان شده است و دلیل آن خواست مسابقه بر این معیار بوده است . بر اساس خواست مسابقه در تعیین معیار , یکی از معیارهای ارزیابی را **AUC (Area Under the ROC Curve)** قرار دادیم. **AUC** نشان دهنده سطح زیر نمودار **ROC (Receiver Operating Characteristic)** می‌باشد که هر چه مقدار این عدد مربوط به یک طبقه بند بزرگتر باشد کارایی نهایی طبقه بند مطلوب‌تر ارزیابی می‌شود. به عبارت دیگر ، به حداکثر رساندن **ROC AUC** به حداکثر رساندن همبستگی رتبه هدف و پیش بینی است. می توان **ROC curve** را نیز به صورتی که در فرمول های ۱ و ۲ نشان دادیم، بیان کرد.

$$ROC\ AUC = \frac{Cov\,(y\,, rank\,(\mu))}{Cov\,(y\,, rank\,(y))} * 0.5 + 0.5 \qquad (۱)$$

$$ROC\ AUC = \frac{Cov\,(rank\,(y), rank\,(\mu))}{Cov\,(rank\,(y), rank\,(y))} * 0.5 + 0.5 \qquad (۲)$$

استراتژی ارزیابی **Group Kfold cross validation** است. برای یافتن بهترین پارامترها از روش **Randomize Search cross validation** و سعی و خطا با تغییر بازه مقادیر تعیین کردیم.

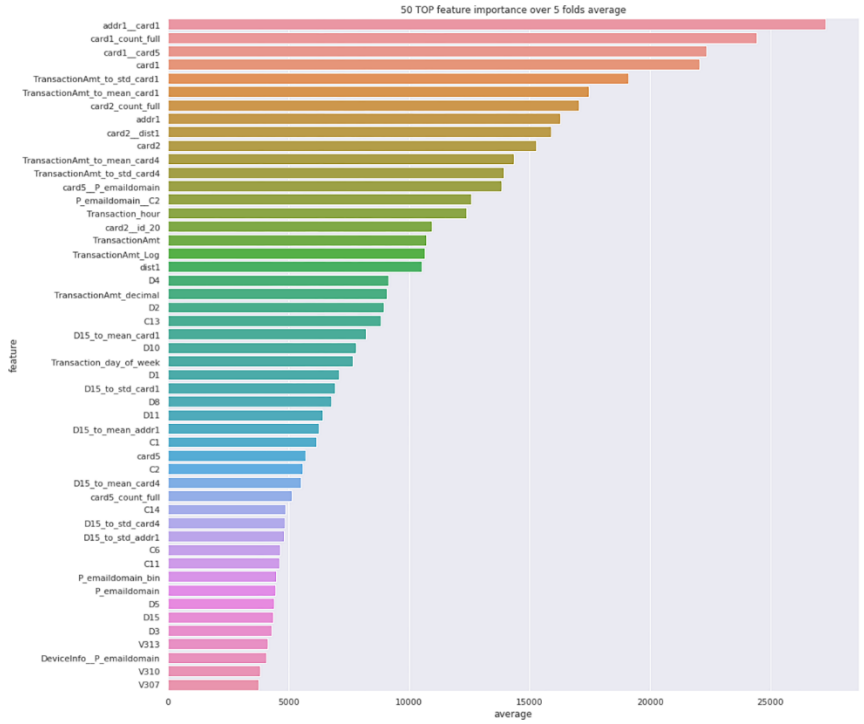
یکی از چالش های مجموعه داده نامتوازن بودن آن است که شدیدا مشاهده می شود. استفاده از پارامتری به نام **scale_pos_weight** که وزن کلاس مثبت را تعیین می کند تا چالش داده های نامتوازن را رفع نماید. مقدار پیشفرض این پارامتر برابر ۱ است. مقدار ۱ یعنی داده ها متوازن هستند. مقدار این پارامتر با روش های مختلفی محاسبه می شود. مقدار مناسب این پارامتر را همانطور که در فرمول ۳ مشاهده می نمایید براساس تعداد کل نمونه ها تقسیم بر تعداد نمونه های مثبت منهای یک محاسبه کردیم. این پارامتر در الگوریتم های پیشنهادی موجود است.

$$\frac{Total\ Samples}{Positive\ Samples} - 1 \qquad (۳)$$

در مجموعه داده انواع مختلفی از ویژگی ها از جمله غیرعددی وجود دارد که این نوع داده ها را مورد بررسی قرار داده و با استفاده از **Label Encoding** نیاز خود برای رفع این چالش را حل کردیم. البته می توان گفت **Label Encoding** تنها راه انکود کردن نیست. می توان از روش های دیگر نظیر **one-hot encoding** نیز استفاده کرد. تفاوت روش ها در موقعیت استفاده از آن است. اغلب در شبکه های عصبی از **one-hot encoding** بیشتر استفاده می شود و برای درخت معمولا از **Label Encoding** استفاده می شود که موقعیتی در جایگاه درختی است.

مدل را با استفاده از طبقه بندی کننده های **LightGBM** و **XGBoost** ، یکبار با مدل **LightGBM**، یکبار با مدل **XGBoost** و یکبار به صورت ترکیبی با استفاده از **Averaging methode** اجرا کردیم. در اجرای اول، با استفاده از ویژگی **feature_importance** الگوریتم **LightGBM**، میزان اهمیت ویژگی ها بررسی شد که در شکل ۹ می توانید مشاهده کنید. ۵۰ مورد از با اهمیت ترین و موثرترین ویژگی های دیتاست را براساس ۵ بار اجرا آورده ایم .ویژگی هایی که وزن پایین تری داشتند(ویژگی هایی که میزان اهمیت آنها نزدیک به صفر بود) را حذف کرده و مجددا الگوریتم را اجرا نمودیم.

نتایج اجرای الگوریتم ها با استفاده از استراتژی ارزیابی **Group Kfold** با ۵ فولد با معیارهای **AUC**، **Accuracy**، **Percision**، **Recall** و **F1 score** قبل و بعد از عملیات مهندسی ویژگی ها را در جدول ۴ تا ۷ مشاهده می نمایید. در الگوریتم تجمیعی به روش میانگین گیری وزن دار که روش توسعه یافته میانگین گیری ساده است به مدلی که دارای ارزیابی بهتری میباشد وزن بیشتری اختصاص داده میشود، به دلیل ارزیابی بهتر الگوریتم **XGBoost** وزن ۰/۶ و به الگوریتم **LightGBM** وزن ۰/۴ را اختصاص داده شده است. نکته حائزاهمیت در وزن دهی که باید به آن توجه داشت این است که مجموع وزن های مدل باید یک باشد. همان طور که در جدول ۴ تا ۷ مشاهده می شود، نتایج قبل از انجام مهندسی ویژگی دارای مقادیر کمتری نسبت به مقادیر پس از مهندسی ویژگی است و این بدین معناست که انتخاب ویژگی های تاثیرگذار تا چه اندازه می تواند مهم



شکل ۹. مهمترین ویژگی ها

Figure 9. The most important features

باشـد. همچنین ترکیب تجمیعی **LightGBM** و **XGBoost** به روش میانگین گیری سـاده و وزن دار پس از مهندسـی ویژگی ها با معیار مسـابقه یعنی **AUC** ۹۴/۶۹ و ۹۵/۰۸ نتایج بهتری را در مقایسه با حالت های دیگر نشان می دهد. در جدول ۸ مقایسه ای بین روش های مذکور و برخی روش های متداول دیگر مانند شبکه های بیزی، جنگل تصادفی و رگرسیون لوجستیک انجام شده است، همان طور که مشاهده می شـود، روش تجمیعی میانگیری وزن دار پس از اعمال مهندسی ویژگی ها بهترین مقادیر ارزیابی را دارا می باشد.

جدول ۴. نتایج پیاده سازی مدل LightGBM

Table 4. The Result of LightGBM

	بدون استفاده از مهندسی ویژگی					با استفاده از مهندسی ویژگی				
Fold	***AUC***	***Acc.***	***Recall***	***Prec.***	***F-Meas.***	***AUC***	***Acc.***	***Recall***	***Prec.***	***F-Meas.***
1	91.14	98.10	83.82	57.05	67.89	92.24	98.23	85.95	59.07	70.02
2	92.19	98.90	84.91	87.34	86.11	93.09	99.07	86.60	89.86	88.20
3	93.32	99.10	87.06	88.87	87.95	94.24	99.21	88.87	90.09	89.47
4	93.60	98.89	87.87	84.39	86.10	95.12	99.21	90.69	89.33	90.01
5	93.11	98.97	86.81	83.98	85.37	94.00	99.10	88.53	85.85	87.17
Avg	92.67	98.79	86.09	80.33	82.68	93.74	98.96	88.13	82.84	84.97

جدول ۵. نتایج پیاده سازی مدل XGBoost

Table 5. The Result of XGBoost

	بدون استفاده از مهندسی ویژگی					با استفاده از مهندسی ویژگی				
Fold	***AUC***	***Acc.***	***Recall***	***Prec.***	***F-Meas.***	***AUC***	***Acc.***	***Recall***	***Prec.***	***F-Meas.***
1	93.13	98.28	87.72	59.70	71.05	94.90	98.38	91.25	60.79	72.97
2	93.31	99.09	87.02	90.10	88.54	94.37	99.19	89.13	90.51	89.82
3	94.24	99.21	88.87	90.09	89.47	94.98	99.28	90.33	90.54	90.44
4	92.96	99.05	86.35	89.06	87.68	94.81	99.21	90.04	89.75	89.89
5	94.13	99.11	88.77	85.98	87.36	95.16	99.26	90.74	88.31	89.51
Avg	93.55	98.95	87.75	82.99	84.82	94.84	99.06	90.30	83.98	86.53

جدول ۶. نتایج پیاده سازی مدل میانگین گیری ساده

Table 6. The Result of Simple Average Method

	بدون استفاده از مهندسی ویژگی					با استفاده از مهندسی ویژگی				
Fold	***AUC***	***Acc.***	***Recall***	***Prec.***	***F-Meas.***	***AUC***	***Acc.***	***Recall***	***Prec.***	***F-Meas.***
1	93.56	99.12	87.72	78.39	82.79	94.37	99.18	89.31	79.27	83.99
2	93.84	99.14	88.08	90.31	89.18	94.32	99.20	89.03	90.85	89.93
3	94.69	99.25	89.77	90.28	90.02	95.21	99.31	90.79	90.85	90.82
4	93.72	99.12	87.87	89.42	88.64	94.22	99.17	88.85	89.98	89.41
5	94.78	99.23	90.00	88.12	89.05	95.35	99.29	91.11	88.73	89.91
Avg	94.12	99.17	88.69	87.30	87.94	94.69	99.23	89.82	87.94	88.81

جدول ۷. نتایج پیاده سازی مدل میانگین گیری ویزن دار

Table 7. The Result of Weighted Average Method

	بدون استفاده از مهندسی ویژگی					با استفاده از مهندسی ویژگی				
Fold	***AUC***	***Acc.***	***Recall***	***Prec.***	***F-Meas.***	***AUC***	***Acc.***	***Recall***	***Prec.***	***F-Meas.***
1	93.92	99.15	88.42	78.78	83.32	94.90	99.22	90.37	79.76	84.73
2	93.95	99.15	88.29	90.43	89.34	94.65	99.23	89.66	91.14	90.40
3	94.81	99.26	90.00	90.40	90.20	95.56	99.34	91.46	91.15	91.31
4	93.83	99.13	88.09	89.55	88.81	94.55	99.21	89.50	90.28	89.89
5	94.91	99.25	90.25	88.25	89.24	95.72	99.33	91.85	89.07	90.44
Avg	94.28	99.19	89.01	87.48	88.18	95.08	99.27	90.57	88.28	89.35

جدول ۸. مقایسه نتایج نهایی چهار مدل پیشنهادی و سه روش متداول

Table 8. the comparison of total results of four proposed methods and other three common methods

	بدون استفاده از مهندسی ویژگی					با استفاده از مهندسی ویژگی				
Method	***AUC***	***Acc.***	***Recall***	***Prec.***	***F-Meas.***	***AUC***	***Acc.***	***Recall***	***Prec.***	***F-Meas.***
Naïve base	86.87	95.53	74.62	77.44	70.85	87.60	98.43	75.98	78.86	76.91
Random Forest	88.83	98.51	78.44	78.40	78.42	90.37	98.65	81.42	82.54	81.97
Logistic Regression	89.92	97.21	79.85	89.84	81.79	91.36	98.81	83.36	82.31	82.83
LightGBM	92.67	98.79	86.09	80.33	82.68	93.74	98.96	88.13	82.84	84.97
XGBoost	93.55	98.65	87.75	82.99	84.82	94.84	99.06	90.30	83.98	86.53
Simple Average	94.12	**99.21**	88.69	87.30	87.94	94.69	99.23	89.82	87.94	88.81
Weighted Average	**94.28**	99.19	**89.01**	**87.48**	**88.18**	**95.08**	**99.27**	**90.57**	**88.28**	**89.35**

۶- نتیجه گیری

شناسایی تقلب در کارت های اعتباری به عنوان یک مسئله جدی برای سازمان های مالی مانند بانک ها و شرکت های کارت اعتباری شناخته شده است. با کشف سریع تراکنش متقلبانه می توان از خسارات هنگفت جلوگیری کرد. در صنعت کارت اعتباری، استانداردهای ثابتی برای توسعه مدل کشف تقلب به عنوان مجموعه ایی از مدل های متنوع وجود داشت. در این پژوهش مطالعه ایی بین مدل های **XGBoost** و **LightGBM** با معیارهای ارزیابی **AUC**، **Accuracy**، **Percision**، **Recall** و **F1- score** انجام شده است تا مشخص شود (کدام مدل ها در داده های تراکنش های حجیم دنیای واقعی عملکرد بهتری نسبت به مدل های دیگر دارند.) مدل یادگیری تجمیعی به روش میانگین گیری ساده و میانگین گیری وزن دار برای توسعه و مقایسه دو مدل معرفی شده است.

مجموعه داده کشف تقلب کارت های اعتباری مربوط به مسابقه **IEEE-CIS fraud detection** از سایت کگل گرفته شده است و روی آن تحقیق و آزمایش لازم انجام شد. در این تحقیق برای کاهش مصرف حافظه مصرفی از تابع **reduce_mem_usage** استفاده شده است که عملکرد مناسبی در کاهش مصرف حافظه از خود نشان داد. در راهکار پیشنهادی، بوسیله سعی و خطا نتیجه به دست آمده نشان داد بهترین نتیجه برای داده های گم شده زمانی به دست می آید که به جای پرداختن به داده های گم شده و برآورد آن، از خود الگوریتم های سازگار با داده های گم شده یعنی **LightGBM** و **XGBoost** استفاده شود. همچنین تعداد ویژگی ها بسیار زیاد بودند که با استفاده از مهندسی ویژگی ها و استفاده از ویژگی **feature_importance** الگوریتم **LightGBM** در راستای انتخاب ویژگی ها مدل کارآمدتری به دست آمد.

در مجموع در الگوریتم پیشـــنهادی جهت بررســـی و افزایش عملکرد مدل، دو الگوریتم **XGBoost** و **LightGBM** با اســـتفاده از روش های میانگین گیری ســـاده و وزن دار یادگیری تجمیعی ترکیب شـده اسـت و سـپس با مدل های **Logistic Regression، Random Forest، Naïve base** مقایســه انجام شــد که در این میان مدل پیشــنهادی نتیجه بهتری در این مجموعه داده داشت.

۷- کارهای آتی

مهندســـی ویژگی ها اهمیت بالایی در رســـیدن به عملکرد مناســب ایفا می کند. با بررســی دقیق تر ویژگی ها و مبحث انتخاب ویژگی های جدید می توان به مجموعه ای دقیق تر و با اهمیت‌تر رســـید که با هزینه کمتر دقت بیشـــتری را به ارمغان آورد. همچنین اســـتفاده و ترکیب مدل های دیگر هم از نظر تعداد مدل ها و هم از نظر نوع الگوریتم آموزشی مانند شبکه های عصبی عمیق ممکن است دقت و عملکرد مدل را افزایش دهد. از طرفی یادگیری افزایشـــی و خودکار نیز یک کار مهم است که در آینده می توان به آن پرداخت. برای کنار آمدن با رانش مفهوم، مدل باید الگوهای تقلبی را از جریان معاملات به طور مداوم و بدون فراموش کردن دانش موجود بیاموزد. در حالی که یادگیری در مورد داده های بســـیار نامتوازن در محیط یادگیری اســـتاتیک بررســـی شـــده اســـت، یادگیری از جریان داده های غیر ثابت قابل بررســـی اســـت. برای یک فرایند یادگیری کاملاً خودکار، علاوه بر یادگیری مـدل، عـوامـلـی مـانـنـد اســـتـخـراج داده، تـغـیـیـر شـــکل(**transformation**)، پیش پردازش و ارزیابی مدل نیز بـایـد خودکـار بـاشـــد. همچنین روش های دیگر یادگیری تجمیعی نیز ممکن است در بهبود عملکرد مدل موثر باشـند که می توان در آینده به آنها پرداخت.

8- References ۸- مراجع

[1] t. o. c. cart. [Online]. Available: [https://www.thebalance.com/key-differences-between-visa-mastercard-discover-and-american-express-4588450#citation-4.

[2] "Performance Evaluation of Credit Card Fraud Transactions using Boosting Algorithms," *International Journal of Electronics Communication and Computer Engineering*, vol. 10, no. 6, pp. 262-270, 2019.

[3] J. Huang, "Credit Card Transaction Fraud Using Machine Learning Algorithms," in *2019 International Conference on Education Science and Economic Development (ICESED 2019)*, 2020.

[4] Y. Ganin, E. Ustinova, H. Ajakan and P. Germain, "Domain-Adversarial Training of Neural Networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2030-2096, 2016.

[5] M. Raza and U. Qayyum, "Classical and deep learning classifiers for anomaly detection," *2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, pp. 614-618, 2019.

[6] B. Lebichot, Y.-A. L. Borgne, L. He-Guelton, F. Oblé and G. Bontempi, "Deep-learning domain adaptation techniques for credit cards fraud detection," in *NNS Big Data and Deep Learning conference*, Cham, 2019.

[7] A. A. Abdulrazaq, M. B. Abdulrazaq, I. J. Umoh and E. A. Adedokun, "Fraud Detection in Credit Card and Application of VAT Clustering Algorithm: A Review," in *2019 2nd International Conference of the IEEE Nigeria Computer Chapter (NigeriaComputConf)*, 2019, October.

[8] Y. Lucas, P.-E. Portier, L. Laporte, L. He-Guelton, O. Caelen, M. Granitzer and S. Calabretto, "Towards automated feature engineering for credit card fraud detection using multi-perspective HMMs," *Future Generation Computer Systems*, vol. 102, pp. 393-402, 2020.

[9] I. Sadgali, N. Sael and F. Benabbou, "Comparative Study Using Neural Networks Techniques for Credit Card Fraud Detection," *The Proceedings of the*

[18] M. H. S. G. Arya, "DEAL–'Deep Ensemble ALgorithm'Framework for Credit Card Fraud Detection in Real-Time Data Stream with Google TensorFlow.," *Smart Science*, vol. 8, no. 2, pp. 71-83, 2020 Apr 2.

[19] S. A. G. N. G. A. G. Bagga, "Credit Card Fraud Detection using Pipeling and Ensemble Learning," *Procedia Computer Science*, vol. 173, pp. 104-112, 2020 Jan 1.

[20] P. Kumari and S. P. Mishra, "Analysis of credit card fraud detection using fusion classifiers," *Computational Intelligence in Data Mining*, pp. 111-122, 2019.

[21] H. Najadat, O. Altiti, A. A. Aqouleh and M. Younes, "Credit Card Fraud Detection Based on Machine and Deep Learning," in *11th International Conference on Information and Communication Systems (ICICS),IEEE*, 2020.

[22] G. Alicja, M. Bakala, K. Woznica, M. Zwolinski and P. Biecek, "EPP: interpretable score of model predictive power.," *arXiv*, p. preprint arXiv:1908.09213, 2019 Aug 24.

[23] Z. Yixuan, J. Tong, Z. Wang and F. Gao, "Customer Transaction Fraud Detection Using Xgboost Model," in *International Conference on Computer Engineering and Application (ICCEA), IEEE*, 2020 Mar 18.

[24] D. J. G. S. C. a. J. C. Ge, "Credit Card Fraud Detection Using Lightgbm Model.," in *International Conference on E-Commerce and Internet Technology (ECIT), IEEE*, 2020 Apr 22.

[25] J. Choi, B. Jeong, Y. Park, J. Seo and C. Min, "AN OPTIMAL BOOSTING ALGORITHM BASED ON NONLINEAR CONJUGATE GRADIENT METHOD," *Journal of the Korean Society for Industrial and Applied Mathematics*, vol. 22, no. 1, pp. 1-13, 2018.

Third International Conference on Smart City Applications, 2019.

[10] R. Saia and S. Carta, "Evaluating the benefits of using proactive transformed-domain-based techniques in fraud detection tasks," *uture Generation Computer Systems*, vol. 93, 2019.

[11] E. Kim, J. Lee, H. Shin, H. Yang, S. Cho, S.-k. Nam and e. al, "Champion-challenger analysis for credit card fraud detection: Hybrid ensemble and deep learning," *Expert Systems with Applications*, vol. 128, pp. 214-224, 2019.

[12] C.-H. Su, F. Tu, X. Zhang, B.-C. Shia and T.-S. Lee, "A ENSEMBLE MACHINE LEARNING BASED SYSTEM FOR MERCHANT CREDIT RISK DETECTION IN MERCHANT MCC MISUSE," *Journal of Data Science*, vol. 17, no. 1, pp. 81-106, 2019.

[13] G. M. C. A. R. Hajela, "A Clustering Based Hotspot Identification Approach For Crime Prediction," *Procedia Computer Science*, vol. 167, pp. 1462-1470, 2020.

[14] R. Md and A. Rab, "A Comparative Study on Crime in Denver City Based on Machine Learning and Data Mining.," *arXiv preprint arXiv:2001.02802*, 2020 Jan 9.

[15] R. Polikar, Ensemble Learning, M. Y. Zhang C., Ed., Boston, Massachusetts: Springer, 19 January 2012.

[16] L. F. A. A. S. N. K. S. J. D. R. S. Gutierrez-Espinoza, "Fake Reviews Detection through Ensemble Learning.," *arXiv preprint arXiv:2006.07912*, 2020 Jun 14.

[17] A. A. a. S. J. M. Taha, "An Intelligent Approach to Credit Card Fraud Detection Using an Optimized Light Gradient Boosting Machine.," *IEEE Access 8*, vol. 8, pp. 25579-25587, 2020 Feb 3.

[26] D. Kavya and K. Chitharanjan, "Performance Evaluation of Credit Card Fraud Transactions using Boosting Algorithms," *International Journal of Electronics Communication and Computer Engineering*, vol. 10, no. 6, pp. 262-270, 2019.

[27] Y. Liang, W. Jiyu, W. Wei, C. Yujun, Z. Biliang, C. Zhenkun and L. Zhenzhang, "Product marketing prediction based on XGboost and LightGBM algorithm," *the 2nd International Conference on Artificial Intelligence and Pattern Recognition*, pp. 150-153, 2019.

[28] V. K. Ayyadevara, "Gradient Boosting Machine," *Pro Machine Learning Algorithms*, pp. 117-134, 01 July 2018.

[29] P. KHANDELWAL, "Which algorithm takes the crown: Light GBM vs XGBOOST?," 12 June 2017. [Online]. Available: https://www.analyticsvidhya.com/blog/2017/06/which-algorithm-takes-the-crown-light-gbm-vs-xgboost/.

[30] S. Mittal and S. Tyagi, "Computational Techniques for Real-Time Credit Card Fraud Detection.," *Handbook of Computer Networks and Cyber Security*, pp. 653-681, 2020.

دکتر سعید بختیاری عضو هیئت علمی و استادیار دانشگاه امین و همچنین مشاور امنیتی سازمان های دولتی و بانکی است. در سال ۲۰۰۹، وی لیسانس مهندسی نرم افزار را از دانشگاه آمل و در سالهای ۲۰۱۱ و ۲۰۱۶ ، به ترتیب مدرک کارشناسی ارشد و دکترای خود را در زمینه امنیت شبکه و اطلاعات از **UTM** مالزی دریافت کرد. فعالیت های مورد علاقه وی رمزنگاری و داده کاوی است.

Saeid_bakhtiarii@yahoo.com

زهرا نصیری داوطلب دکترای هوش مصنوعی و فارغ التحصیل مهندسی نرم افزار رایانه در دانشگاه آل طاها در تهران، ایران است. وی چندین سال تجربه برنامه نویسی هوش مصنوعی و یادگیری ماشین دارد. یادگیری ماشین، داده کاوی، بهینه سازی و رایانش ابری از جمله علایق تحقیقاتی وی است. وی دارای اعتبار حرفه ای در زمینه علوم داده از وزارت علوم، تحقیقات و فناوری است.

Lnasiri007@gmail.com

سید محمد صادق حجازی کارشناس ارشد مهندسی نرم افزار رایانه در دانشگاه پردیسان در مازندران، ایران است. وی چندین سال تخصص برنامه نویسی و تدریس هوش مصنوعی و یادگیری ماشین دارد. یادگیری ماشین، یادگیری عمیق، تحلیل داده، داده کاوی، بهینه سازی، رایانش ابری و اینترنت اشیاء از جمله علایق تحقیقاتی وی است.

Sadegh.hejazi@hotmail.com