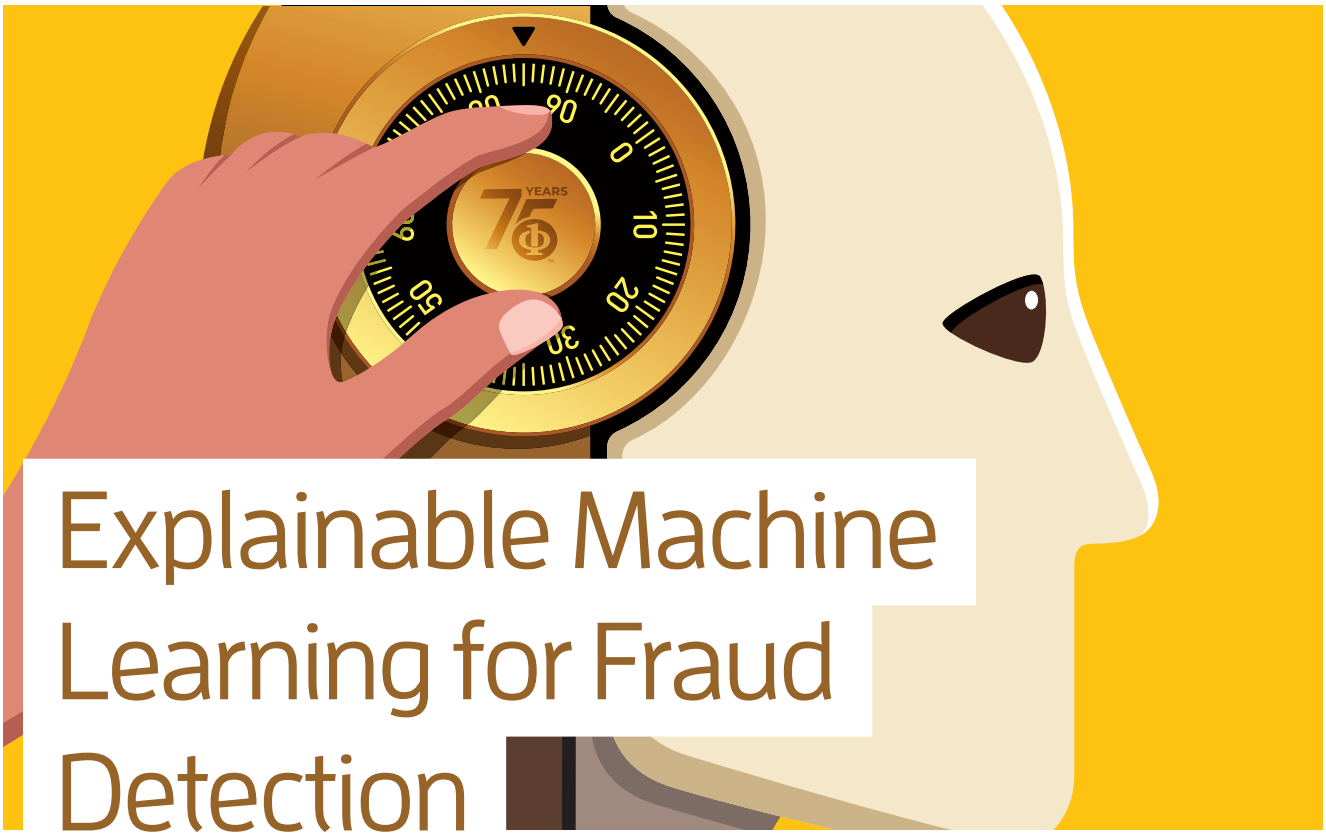

Explainable Machine Learning for Fraud Detection

Ismini Psychoula, Andreas Gutmann, Pradip Mainali, S.H. Lee, Paul Dunphy, and Fabien A.P. Petitcolas,
OneSpan Innovation Centre

The application of machine learning to support the processing of large data sets holds promise in many industries. We explore explainability methods in the domain of real-time fraud detection by investigating the selection of appropriate background data sets and runtime tradeoffs on supervised and unsupervised models.

The digital landscape is constantly evolving and shifting toward the integration of artificial intelligence (AI) and machine learning in core digital service functionalities. Due to the COVID-19 pandemic, this has led to a shift in digital services, transforming these services from being a convenience to a necessity. Many organizations had to transition to online services faster than anticipated. While this creates opportunities for development and growth, it also attracts cybercriminals.

Digital Object Identifier 10.1109/MC.2021.3081249
Date of current version: 24 September 2021

Increases in the hacking of personal accounts and online financial fraud were reported during the strictest times of the lockdown during the COVID-19 outbreak[1] as well as more successful cybercrimes by established and new threat actors that could introduce new crime patterns.[2] Fraud costs businesses and individuals in the United Kingdom £130 billion annually and £3.89 trillion in the global economy.[3] There is an increasing need for financial fraud detection systems that can automatically analyze large volumes of events and transactions, and machine learning-based risk analytics is one of the ways to achieve that. There is no perfect universal rule to distinguish a fraudulent case from a valid one, as fraud

appears in all shapes and sizes and can be indistinguishable from normal cases.

Many approaches have been proposed for automatic fraud detection, from anomaly detection to classical machine learning and modern deep learning models; however, it remains a challenging problem. The work by Varmedja et al.[4] compared logistic regression, naive Bayes, random forests, multilayer perceptron, and artificial neural networks. The authors found that random forests had the best performance. They also state the importance of using sampling methods to address the class imbalance issues that are common in fraud data sets. In a similar study, Thennakoon et al.[5] offer guidance on the selection of optimal algorithms based on four types of fraud.

A recent study[6] proposed discrete Fourier transform conversion to exploit frequency patterns instead of canonical ones. Another recent work[7] explored the use of prudential multiple consensus, which combines the results of several classification models based on the classification probability and majority voting.

Another challenging aspect is that, when applying complex models to detect fraud cases, there is no easy way to explain how these methods work and why the model makes a decision. Unlike linear models like logistic regression, where the coefficient weights are easy to explain, there is no simple way to assess the reasons behind a complex machine learning model's or deep neural network's prediction. In particular, for applications with sensitive data or in safety-critical domains, providing effective explanations to the users of the system is paramount[8] and has become an ethical and regulatory requirement in many application domains.[9]

Explainability is not only linked to understanding the inner workings and predictions of complex machine learning models but also to concerns over inherent biases or hidden discrimination and potential harms to privacy, democracy, and other societal values. The General Data Protection Regulation states, in Articles 13, 14, and 22, that data controllers should provide information on "the existence of automated decision making, including profiling" and "meaningful information about the logic involved as well as the significance and the envisaged consequences of such processing for the data subject."

There are several important elements to consider and be able to explain when creating automated decision-making systems:

- The rationale behind the decision should be easily accessible and understandable.
- The system should maximize the accuracy and reliability of the decisions.
- Underlying data and features that may lead to bias or unfair decisions should be identified.
- The context in which the AI is deployed and the impact the automated decision might have for an individual or society should be understood.[10]

Several studies have been proposed to explain models in anomaly-detection settings. Contextual outlier interpretation[11] is a framework designed to explain anomalies spotted by detectors. Situ is another system for detecting and visualizing anomalies in computer network traffic and logs,[12] while Collaris et al.[13] developed dashboards that provide explanations for insurance fraud detected by a random forest algorithm.

Similar studies to this one have also used Shapley additive explanation (SHAP) values for autoencoder explanations,[14] explained network anomalies with variational autoencoders,[15] and compared SHAP values to the reconstruction error of principal component analysis features to explain anomalies.[16] However, so far, there has not been a lot of focus on the impact of the background data set and the runtime implications, which such explanations could have for real-time systems such as fraud.

In this article, we present a case study that explores explanations with two of the most prominent methods, LIME and SHAP, to explain fraud detected by both supervised and unsupervised models. Attribution techniques explain a single-instance prediction by ranking the most important features that affected the generation of it.

LIME[17] approximates the predictions of the underlying black-box model by training local surrogate models to explain individual predictions. Essentially, LIME modifies a single data sample by tweaking the feature values in the simpler local model and observes the resulting impact on the output.

The SHAP[18] method explains the prediction of an instance by computing the contribution of each feature to the prediction using Shapley values based on coalition game theory. Intuitively, SHAP quantifies the importance of each feature by considering the effect each possible feature combination has on the output. The explanations aim to provide insights to experts and end users by focusing on the connections and tradeoffs among the features that affect the final decision, depending on the background data set and the runtime of the explanation method. We focused on black-box explanation methods because they can be applied to most algorithms without being aware of the exact model.

CASE STUDY

We present a financial fraud explainability case study. We use the open source IEEE Computational Intelligence Society (CIS) Fraud Detection data set[19] to provide fraud-detection explanations. The data set provides information on credit card transactions and customer identity—with labels for fraudulent transactions ($Y \in 0,1$). The data set has highly imbalanced classes, with fraud accounting for 3.49% of all transactions.

We experimented with both supervised and unsupervised models and compared their performance on the same data set in terms of the prediction accuracy, reliability of explanation, and runtime. For the supervised models, we used the labels provided in the IEEE CIS Fraud Detection data set[19] to indicate, for each transaction, whether it is fraudulent or genuine during the training phase. However, obtaining labels for each transaction is often not possible, and labeling the data manually or having just clean data is often difficult and time consuming.

Unsupervised methods and representation learning can handle well-imbalanced data sets without requiring labels. In the unsupervised models (autoencoder and isolation forest) we treated the IEEE CIS Fraud Detection data set as unlabeled and used the reconstruction loss and anomaly score, respectively, to detect the fraud cases.

We compared the following models:

- *Naive Bayes*: This probabilistic classifier is simple, highly scalable, and easy to train in a supervised setting.
- *Logistic regression*: This simple and inherently intelligible model shows how much we gain by using more complex models compared to simple ones.
- *Decision trees*: Decision trees offer robust accuracy and inherent transparency when their size is small.
- *Gradient boosted trees*: Tree ensembles are one of the most accurate types of models but also quite complex. We trained the model with 100 estimators, a maximum depth of 12, and a learning rate of 0.002.
- *Random forests*: This classifier uses ensembles of trees to reduce predictive error. We trained the model with 100 estimators.
- *Neural network*: This multilayer perceptron can model nonlinear interactions among the input features. We trained the multilayer perceptron with three hidden layers containing 50 units each with rectified linear unit (ReLU) activation using the Adam optimizer.
- *Autoencoder*: This unsupervised neural network works by using back-propagation and setting the target values to be equal to the inputs. We trained the network with three hidden layers containing 50 units each with ReLU activation, the Adam optimizer, and mean square error as the loss. The reconstruction error measures whether an observation deviates from the rest.
- *Isolation forest*: This unsupervised algorithm uses a forest of decision trees to partition the data. The splits to separate the data are done randomly. The number of splits indicates whether a point is an anomaly. For the training, we used 100 estimators and automatic contamination.

We use only 24 out of the 433 features in the data set. The 24 features were selected to focus on the columns that have some description of their values so that the explanation could be more understandable. These features are "TransactionAMT," the transaction payment amount in U.S. dollars; "ProductCD," the product for each transaction; "Device Type" and "Device Information"; and "card1" to "card6," which show payment card information, such as card type, card category, issuing bank, and country. "P_emaildomain" is the purchaser's email domain, and "R_emaildomain" is the recipient's email domain. "M1" to "M9" indicate matches, such as names on cards and addresses, and "id_x" are numerical features for identity, such as device rating, IP domain rating, and proxy rating.

The data set includes behavioral fingerprints, like account login times and failed login attempts as well as how long an account stayed on the page. However, the providers of the data set were not able to elaborate on the meaning of all of the features and correspondence among features and columns due to security terms and conditions.[19]

Table 1 shows the classification results for each of the models. We withheld 20% of the sample for validation. For the implementation of the models, the Scikit-learn (http://scikit-learn.sourceforge.net/) and Keras (https://keras.io/) libraries were used. We evaluated the performance of the models using precision, recall, F1 score, and area under the receiver operating characteristic curve since the data set is highly imbalanced.

TRUSTWORTHINESS OF EXPLANATIONS

To create a benchmark for the explanations, we used a logistic regression classifier to predict fraudulent transactions and measure feature importance

through the coefficient weights. Due to the transparency of the logistic regression model and its wide acceptance among regulatory bodies, we treat the global weights this provides as the ground truth and compare it with the results of attribution methods. Figure 1 presents the global top 10 most important features as determined by logistic regression.

EXPLAINING SUPERVISED MODELS

Attribution techniques, such as LIME and SHAP, explain a single-instance prediction by ranking the most important features that affected the generation of it. To evaluate the performance of LIME and SHAP in fraud detection, we compare and evaluate them by providing explanations with feature importance. We use summary plots that provide an overview of how the values of the same single instance have influenced the prediction for the different models. As *single instance*, we define one fraudulent transaction that is used for all models and experiments.

The experimental results for the supervised models are shown in Figure 2(a) and (b). Both LIME and SHAP produce similar top features with different rankings. Compared with the global features of logistic regression in Figure 1, LIME agrees on seven and SHAP agrees on eight features. However, the ranking of features varies across both methods and all models. We noticed that, on average, SHAP produces explanations closer to the global features of logistic regression in terms of rank.

TABLE 1. The performance results.

Model	Precision	Recall	F1 score	Area under the curve
Naive Bayes	0.543	0.669	0.544	0.663
Logistic regression	0.891	0.533	0.553	0.533
Decision tree	0.762	0.742	0.752	0.706
Random forest	0.84	0.725	0.769	0.688
Gradient boosting	0.88	0.729	0.789	0.709
Neural network	0.795	0.578	0.619	0.581
Autoencoder	0.944	0.767	0.839	0.617
Isolation forest	0.723	0.608	0.664	0.553

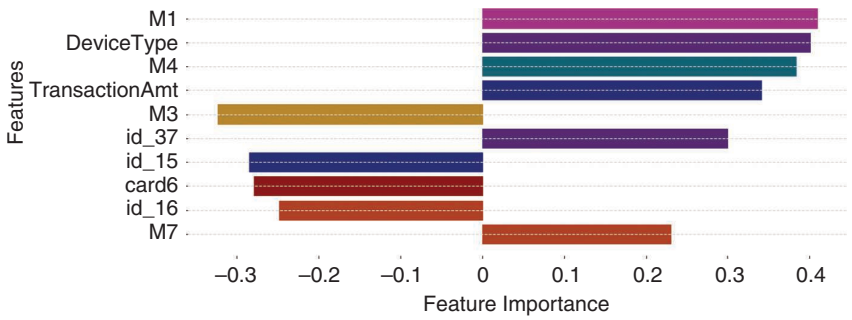

FIGURE 1. The top 10 global features for logistic regression ranked by importance.

EXPLAINING UNSUPERVISED MODELS

In unsupervised methods, particularly for anomaly detection, the result given by a model is not always a probability. In the case of the autoencoder, we are interested in the set of explanatory features that explains the high reconstruction error. In Figure 3(a), we see that SHAP agrees only on four features, with transaction amount and device information the most important features that flagged the transaction as fraud. Figure 3(b) shows the top explanation features for isolation forest with SHAP. We notice that the top features match only four of the explanations of logistic regression and five of the features from the autoencoder.

RELIABILITY OF EXPLANATIONS

The SHAP method requires a background data set as a reference point to generate single-instance explanations. In image processing, for example, it is common to use an all-black image as a reference point, but in financial fraud detection, there is no universal reference point that can be used as a baseline. We explored the impact of different background data sets in fraud-detection explanations and evaluated different reference points that could be used to provide contrasting explanations to fraud analysts.

Figure 4 highlights the differences when using only normal or only fraud

transactions as the reference point. We noticed that models like naive Bayes, logistic regression, and decision trees give more consistent explanations regardless of the background data set, while models like random forest, gradient boosting, and neural networks are more sensitive to the reference point. We also experimented with different background data sets for the autoencoder and isolation forest models. Our findings show that the autoencoder is more robust to changes in the background data set. In the isolation forest model, we find that contributing features remain mostly the same, but their ranking is affected the most.

We noticed that most models are sensitive to the choice of reference, but there is no obvious point that can be used as a reference for fraud detection. By using an intuitive reference point, we can provide a foundation upon which we can produce explanations that are either similar or contrasting (for example, existing blacklisted accounts) to the class we are trying to predict. For each domain, it is important to understand which references are most understandable, trustworthy, and reliable for the end user.

Another important tradeoff to consider in real-time systems is the time

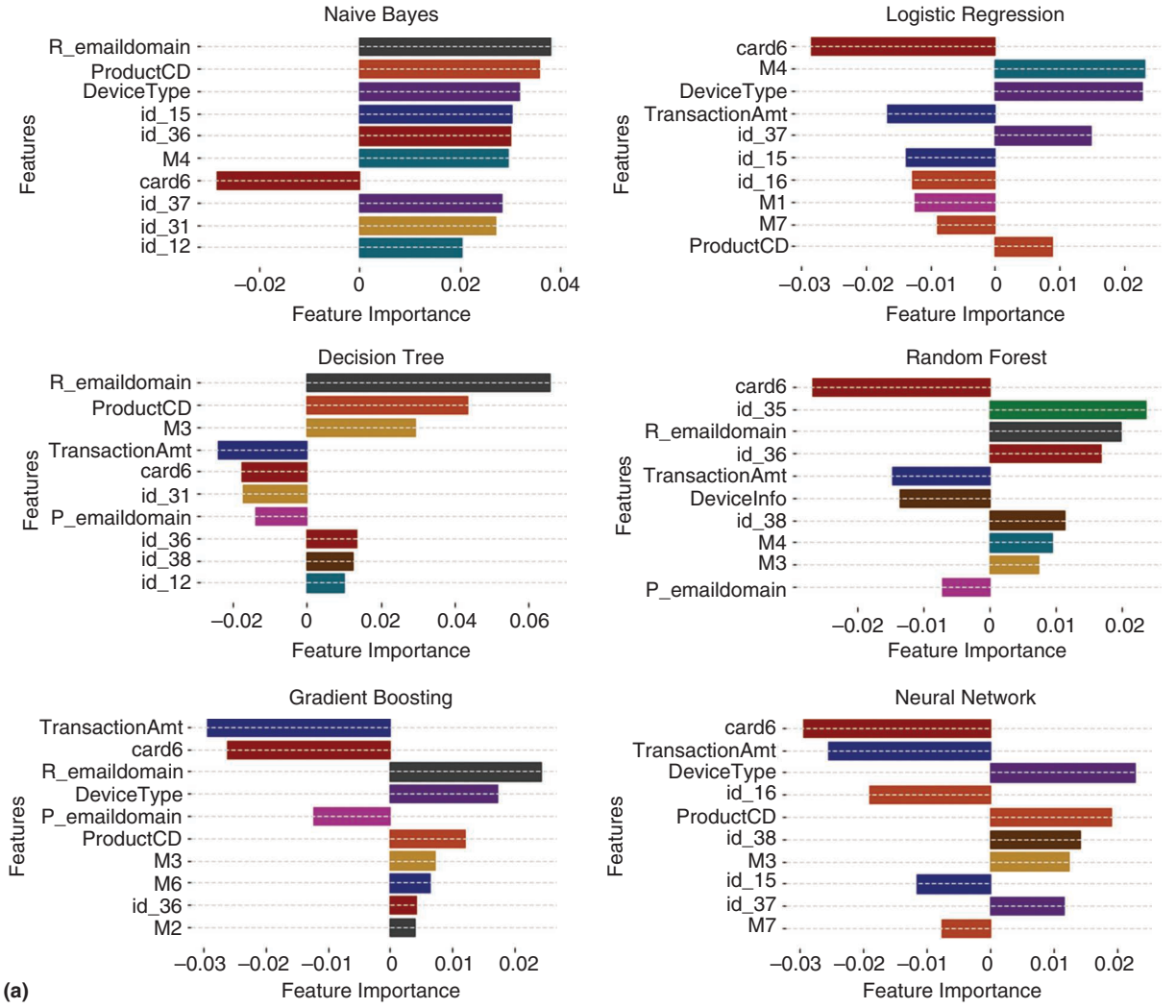

FIGURE 2. The top 10 features ranked by importance for the same normal instance with (a) LIME. (*Continued*)

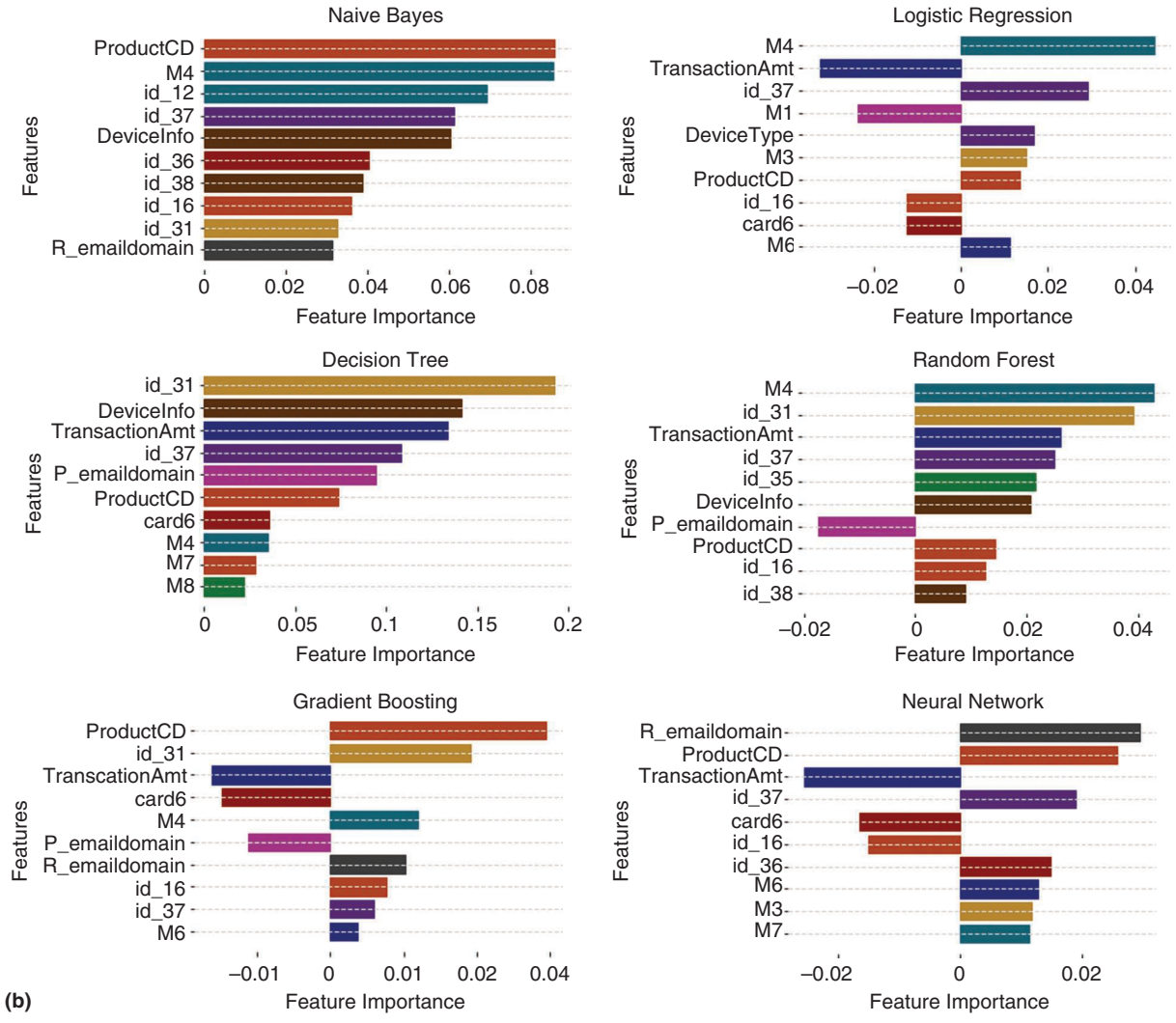

FIGURE 2. (*Continued*) The top 10 features ranked by importance for the same normal instance with (b) SHAP.

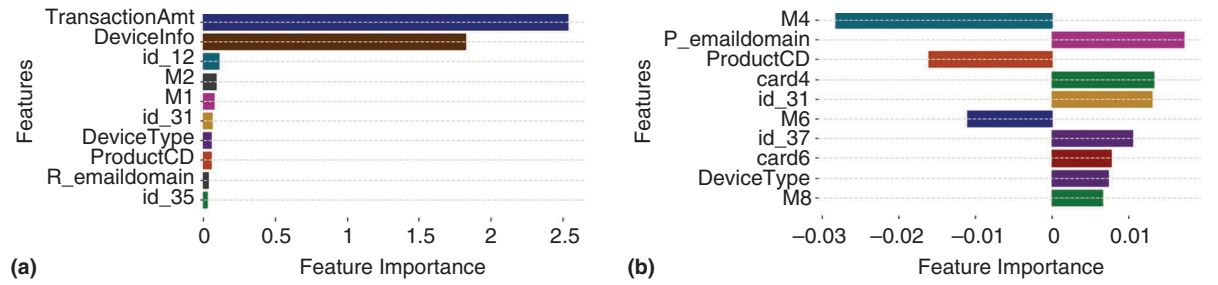

FIGURE 3. The top 10 SHAP features ranked by importance for a single instance with (a) autoencoder and (b) isolation forest.

needed to provide an explanation. As we discussed previously, SHAP depends on background data sets to infer an expected value. For large data sets, it is computationally expensive to use the entire data set, and we rely on approximations (for example, a subsample of the data). However, this has implications for the accuracy of the explanation. Typically, the larger the background data set, the higher the reliability of the explanation.

Table 2 shows the time needed to provide a single-instance explanation with SHAP based on different subsampled background sizes and with LIME. These experiments where run on a Linux server with an Asus TURBO RTX 2080TI with an 11-GB GPU. For fraud-detection systems that operate in real time, the fine tuning of the model and explanation method will affect how many explanations a fraud analyst can receive in time.

Overall, LIME and SHAP are both good approaches to explain models in transaction fraud. In the case of fraud-detection systems, the main concerns for the selection of an explanation method are the tradeoffs among model complexity,

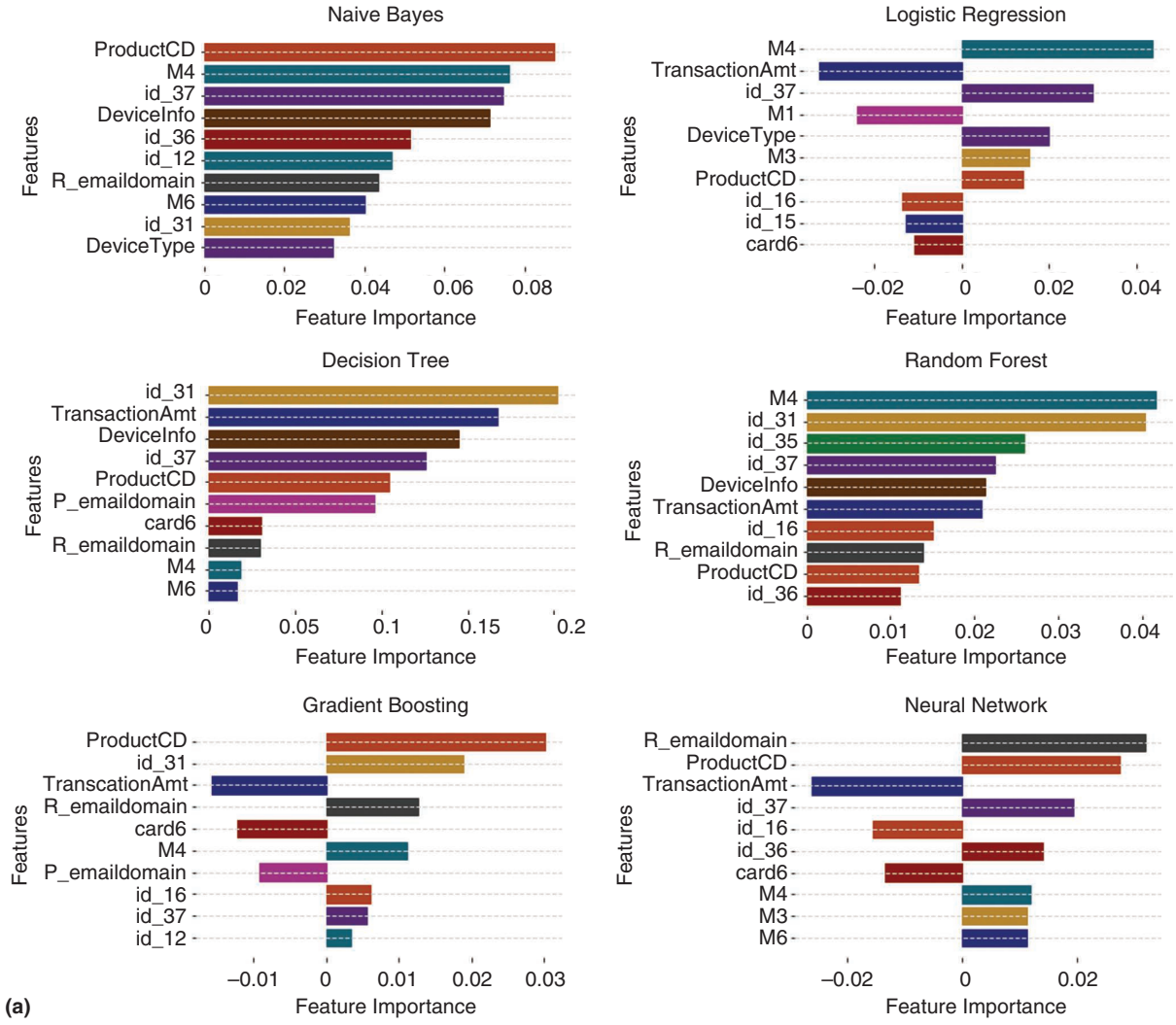

FIGURE 4. The top 10 features for a single instance with SHAP using as background only (a) normal transactions. (*Continued*)

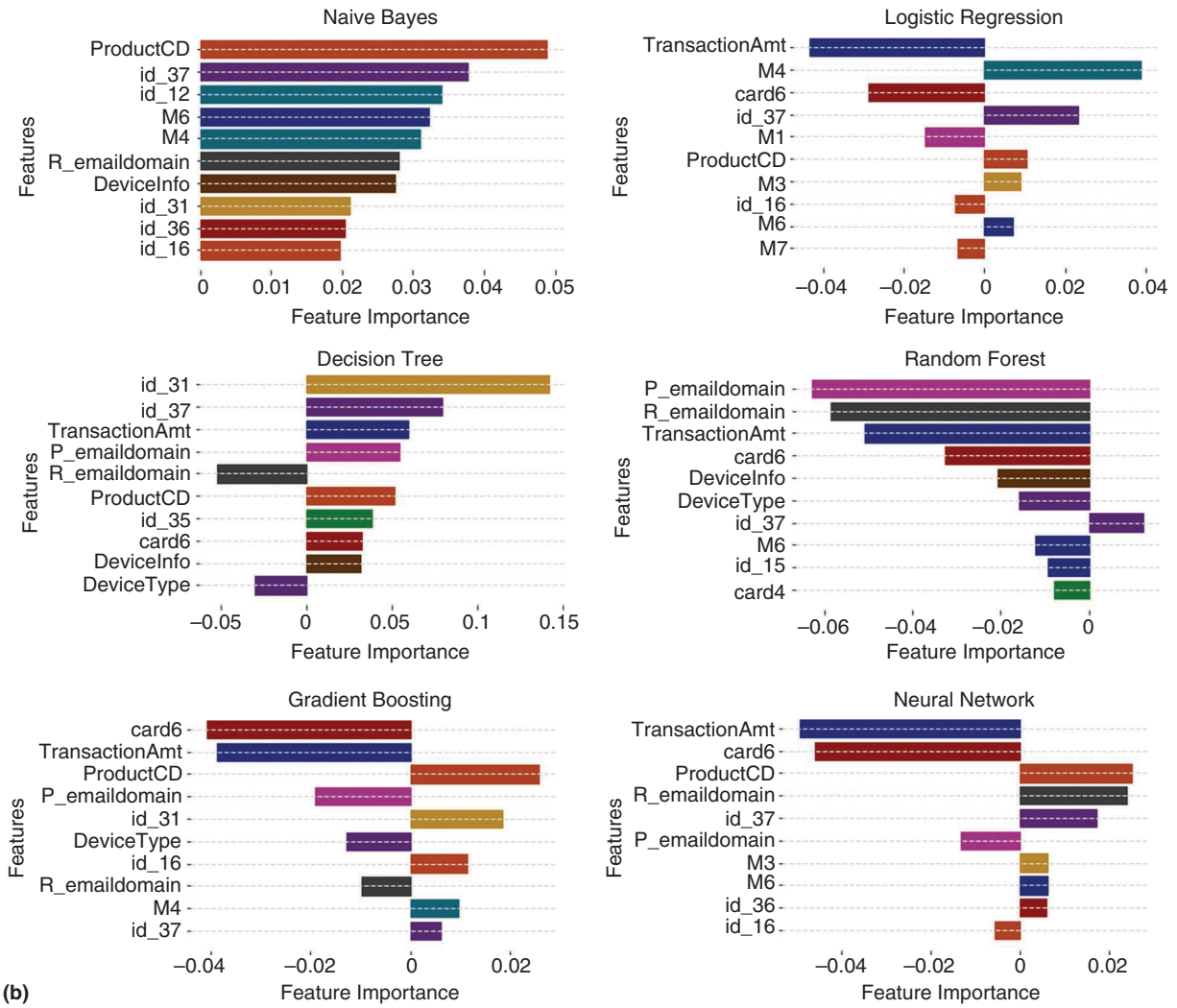

FIGURE 4. (*Continued*) The top 10 features for a single instance with SHAP using as background only (b) fraud transactions.

explanation reliability, and runtime. In Table 2, we documented that LIME is much faster in providing single-instance explanations than SHAP. However, this can come at the cost of reliability.

Fraud data sets contain mixed types of variables with many categorical, numerical, and text data. In cases where there are a lot of categorical variables, like in the fraud data set used in this article, using LIME results in high discrepancies between the predicted value and explanation because of perturbation of the method used to create the approximate model. We found that one of the combinations that provided the best tradeoff was to select SHAP with a background data set of sample size 600. In this case, the mean prediction of the explanation was closest to the mean value of the model, while the runtime overhead was close to that of LIME.

CHALLENGES AND OPPORTUNITIES

The development of explainable machine learning methods still faces some research, technical, and practical challenges, particularly in anomaly-detection methods. One of the main challenges is the prevalence of imbalanced data sets. The benign transactions influence explanations more than the fraudulent ones. SHAP can help by

specifying the background data set or reference point depending on the goal of the explanation.

Besides, to be able to trust the explanations received, we need to be able to evaluate them. This, however, has proven to be a challenging point for explanations. One way to evaluate an explanation is by checking if the explanation was sufficient to allow an end user to achieve his or her task. That is, in the fraud example, the explanation should include enough details and features that will allow a fraud analyst to effectively decide if the transaction was flagged correctly.

It is equally important to be able to identify situations where the AI system is not performing as it should, and human intervention or collaboration between human experts and the AI is needed. To create effective human–computer collaboration, we need to understand and be transparent about the capabilities and limitations of the AI system.[20]

Another challenging aspect of particular interest to the financial domain is the confidential and private nature of the data. Financial data sets contain sensitive personal and corporate information that should be protected. This usually means that the data sets are anonymized, and even the feature names can be changed in the format of "M1" or "id-1," as we have seen in the case study.

In these cases, it is very difficult to explore the data. A lot of time needs to be spent on masking, unmasking, reverse engineering, and deciding whether to include or exclude confidential features since the model cannot be transparent about them and provide explanations that are understandable to an end user.

There are also contextual factors that need to be taken into account when presenting explanations. In a perfect case, the explanations provided by a machine learning method would be identical to human understanding and match with the ground truth. However, that is not usually the case: the explanations provided by the methods analyzed in this article might be clear to expert data scientists but may not be as easily understood by fraud analysts or end users. How an explanation is presented to the end user (for example, visualizations; textual explanations; or numerical, rule-based, or mixed approaches) can determine how effective the explanations are in helping the user understand the inference process and output of the model.

TABLE 2. The runtime for single-instance explanation (in seconds), where *s* is size of the subsampled background data set.

Model	SHAP (*s* = 600)	SHAP (*s* = 1,000)	SHAP (*s* = 4,000)	LIME
Naive Bayes	4.32	7.03	30.61	4.38
Logistic regression	3.78	6.43	26.38	4.43
Decision tree	3.88	6.23	27.55	4.42
Random forest	22.66	35.67	221.74	4.55
Gradient boosting	119.98	193.31	241.8	5.19
Neural network	6.34	11.1	33.78	4.44
Autoencoder	9.26	14.66	73.88	—
Isolation forest	39.11	71.97	318.59	—

A common issue in fraud-detection methods is a large false-positive rate (that is, transactions falsely classified as fraud). We can use explanations to verify whether the features in anomaly detection are indeed making sense and are what we would expect. Explaining an anomaly-detection model in critical domains is equally important with the model's prediction accuracy, as it enables end users to understand and trust these predictions and act upon them.

One of the main advantages of explaining anomalies is being able to differentiate between detecting fraudulent anomalies and detecting rare but benign events, which could be domain specific, from genuine users. By presenting explanations for the outliers found in financial fraud detection, we can reduce the time and effort needed by fraud analysts to manually inspect each case.

Furthermore, it is common in fraud detection to experience data shifts, for example, changing spending patterns in certain times of the year or due to unforeseen circumstances like the COVID-19 pandemic. Most fraud-detection algorithms rely on unsupervised learning or anomaly detection and reinforcement learning. In cases like these, we cannot be sure what the algorithm learns because the data shifts can lead to concept drifts, that is, the model predicts something different than its original purpose. Explainable AI can indicate if there is any data or concept drift of the model. It also makes it easier to improve and debug the models

as well as reuse them without having to learn and figure out the biases from the start each time they are updated.

Another case where explainable AI could help in fraud detection is adversarial behavior. In adversarial machine learning, an adversary inserts specific instances in a machine learning model knowing that this will affect its learning and cause it to misclassify certain instances. For example, a cybercriminal could insert perturbed instances in the data set and affect the fraud score assigned to transactions.

Explainable AI is one means to enhance protection against adversarial attacks. Detecting such attacks is not trivial, and explainable AI can have a great impact in assisting in the detection of such manipulation, giving companies and end users more trust in the machine learning inferences.

For advanced machine learning algorithms to be successfully adopted in the financial domain, model explainability is necessary to address regulatory requirements and ensure trust in the results. The relevant literature proposed several methods to detect and explain anomalies in different settings, from network traffic to insurance fraud. However, we found that the exploration of the reliability and practical considerations of real-time systems was limited.

In this work, we provide insights on the tradeoffs for explanations of financial fraud decisions. We extend the current literature by exploring different reference points and comparing the performance of methods for real-time fraud systems. Using a transparent logistic regression model as the ground truth, we find that attribution methods are reliable but can be sensitive to the background data set, which can lead to different explanation models. Thus, choosing an appropriate background is important and should be based on the goals of the explanation.

We also found that, while SHAP gives more reliable explanations, LIME is faster. In real-time systems, it is not always feasible to explain everything. We must balance the deployability of the models and explanation methods with the time needed for a human and the likelihood of fraud. It may be beneficial to use a combination of both methods, where LIME is utilized to provide real-time explanations for fraud prevention, and SHAP is used to enable regulatory compliance and examine the model accuracy in retrospect.

REFERENCES

1. D. Buil-Gil, F. Miró-Llinares, A. Moneva, S. Kemp, and N. D. iaz-Castaño, "Cybercrime and shifts in opportunities during COVID-19: A preliminary analysis in the UK," *Eur. Soc.*, vol. 23, pp. S47–S59, July 2020. doi: 10.1080/14616696.2020.1804973.
2. A. V. Vu, J. Hughes, I. Pete, B. Collier, Y. T. Chua, I. Shumailov, and A. Hutchings, "Turning up the dial: The evolution of a cybercrime market through set-up, stable, and covid-19 eras," in *Proc. ACM Internet Measurement Conf.*, 2020, pp. 551–566.
3. "The financial cost of fraud 2019." Crowe. https://www.crowe.com/uk/croweuk/-/media/Crowe/Firms/Europe/uk/CroweUK/PDF-publications/The-Financial-Cost-of-Fraud-2019.pdf
4. D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic, and A. Anderla, "Credit card fraud detection-machine learning methods," in *Proc. 2019 18th Int. Symp. INFOTEH-JAHORINA (INFOTEH)*, pp. 1–5. doi: 10.1109/INFOTEH.2019.8717766.
5. A. Thennakoon, C. Bhagyani, S. Premadasa, S. Mihiranga, and N. Kuruwitaarachchi, "Real-time credit card fraud detection using machine learning," in *Proc. 2019 9th Int. Conf. Cloud Comput., Data Science Eng. (Confluence)*, pp. 488–493. doi: 10.1109/CONFLUENCE.2019.8776942.
6. R. Saia and S. Carta, "A frequency-domain-based pattern mining for credit card fraud detection," in *Proc. 2nd Int. Conf. Internet of Things, Big Data Security*, 2017, pp. 386–391. doi: 10.5220/0006361403860391.
7. S. Carta, G. Fenu, D. R. Recupero, and R. Saia, "Fraud detection for e-commerce transactions by employing a prudential multiple consensus model," *J. Inform. Security Appl.*, vol. 46, pp. 13–22, June 2019. doi: 10.1016/j.jisa.2019.02.007.
8. A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52,138–52,160, Sept. 2018. doi: 10.1109/ACCESS.2018.2870052.
9. "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (General Data Protection Regulation)," Official Journal of the European Union, Apr. 2016. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679
10. "The impact of the General Data Protection Regulation (GDPR) on

ABOUT THE AUTHORS

ISMINI PSYCHOULA is a research scientist at the OneSpan Innovation Centre, Cambridge, CB3 0FA, U.K. Her research interests include machine learning, privacy-enhancing technologies, explainability, and trustworthy artificial intelligence. Psychoula received a Ph.D. in computer science from De Montfort University. Contact her at ismini.psychoula@onespan.com.

ANDREAS GUTMANN is a researcher at the OneSpan Innovation Centre, Cambridge, CB3 0FA, U.K. His research interests include user and transaction authentication, user experience, privacy-enhancing technologies, and financial technologies and services. Gutman received a Ph.D. in computer science from University College London. Contact him at andreas.gutmann@onespan.com.

PRADIP MAINALI is a principal researcher at the OneSpan Innovation Centre, Cambridge, CB3 0FA, U.K. His research interests include privacy-preserving machine learning, deep learning, computer vision, and parallel computing on multi-/many-core platforms. Mainali received a Ph.D. in electrical engineering from KU Leuven. Contact him at pradip.mainali@onespan.com.

S.H. LEE is a principal researcher at the OneSpan Innovation Centre, Cambridge, CB3 0FA, U.K. Her research interests include trustworthy artificial intelligence and adaptive machine learning for fraud detection. Lee received a Ph.D. in engineering from the University of Cambridge. Contact her at sharon.lee@onespan.com.

PAUL DUNPHY is a principal researcher at the OneSpan Innovation Centre, Cambridge, CB3 0FA, U.K. His research interests include user-centered security and privacy issues and future digital identity infrastructures. Dunphy received a Ph.D. from Newcastle University. Contact him at paul.dunphy@onespan.com.

FABIEN A.P. PETITCOLAS is manager of the OneSpan Innovation Centre, Brussels, 1853, Belgium. His research interests include information hiding and security issues related to identity management and user authentication. Petitcolas received a Ph.D. in computer science from the University of Cambridge. Contact him at fabien.petitcolas@onespan.com.

artificial intelligence." European Parliament. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530_EN.pdf

11. N. Liu, D. Shin, and X. Hu, "Contextual outlier interpretation," 2017, arXiv:1711.10589. .
12. J. R. Goodall et al., "Situ: Identifying and explaining suspicious behavior in networks," *IEEE Trans. Vis. Comput. Graph.*, vol. 25, no. 1, pp. 204–214, 2018. doi: 10.1109/TVCG.2018.2865029.
13. D. Collaris, L. M. Vink, and J. J. van Wijk, "Instance-level explanations for fraud detection: A case study," 2018, arXiv:1806.07129.
14. L. Antwarg, B. Shapira, and L. Rokach, "Explaining anomalies detected by autoencoders using SHAP," 2019, arXiv:1903.02407.
15. Q. P. Nguyen, K. W. Lim, D. M. Divakaran, K. H. Low, and M. C. Chan, "Gee: A gradient-based explainable variational autoencoder for network anomaly detection," in *Proc. 2019 IEEE Conf. Commun. Network Security (CNS)*, pp. 91–99. doi: 10.1109/CNS.2019.8802833.
16. N. Takeishi, "Shapley values of reconstruction errors of PCA for explaining anomaly detection," in *Proc. 2019 Int. Conf. Data Mining Workshops (ICDMW)*, pp. 793–798. doi: 10.1109/ICDMW.2019.00117.
17. M. T. Ribeiro, S. Singh, and C. Guestrin, "why should I trust you?" explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining*, 2016, pp. 1135–1144.
18. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Advances Neural Inform. Process. Syst.*, 2017, pp. 4765–4774.
19. "IEEE-CIS fraud detection." Kaggle. https://www.kaggle.com/c/ieee-fraud-detection
20. S. Amershi et al., "Guidelines for human-AI interaction," in *Proc. 2019 Chi Conf. Human Factors Comput. Syst.*, pp. 1–13.