SN Computer Science (2021) 2:215
https://doi.org/10.1007/s42979-021-00623-7

ORIGINAL RESEARCH

A Real-Time Prescriptive Solution for Explainable Cyber-Fraud Detection Within the iGaming Industry

David Farrugia[1] · Christopher Zerafa[1] · Tony Cini[1] · Bruno Kuasney[1] · Karen Livori[1]

Received: 7 December 2020 / Accepted: 27 March 2021 / Published online: 15 April 2021

Abstract

This paper presents a real-time fully autonomous prescriptive solution for explainable cyber-fraud detection within the iGaming industry. We demonstrate how our solution facilitates the time-consuming task of player risk and fraud assessment through prescriptive analytics. Our tool leverages machine learning algorithms and advancements in the field of eXplainable AI to derive smarter predictions empowered by local interpretable explanations in real-time. Our best-performing pipeline was able to predict fraudulent behaviour with an average precision of 84.2% and an area under the receiver operating characteristics of 0.82 on our dataset. We also addressed the phenomenon of concept-drift and discussed our empirical and data-driven strategy for detecting and dealing with this problem. Finally, we cover how local interpretable explanations can help adopt a pro-active stance in fighting fraud.

Keywords Machine learning · eXplainable AI · Fraud detection · Prescriptive analytics · iGaming

Introduction

Cyber-fraud is a significant problem which poses severe challenges to all world organisations. A recent study by the Association of Certified Fraud Examiners [2] report that organisations lose an annual 5% of their total revenue to fraudulent activity. Anti-fraud systems have been in constant development to counteract fraud. This is especially true within the iGaming industry where most gaming regulations require operators to comply with stricter anti-fraud measures. As argued by Banks [3], this industry is still susceptible to numerous types of cyber-fraud due to its wide availability and easy access [19]. The rise of transactional anonymity is a factor which helps facilitate fraud [11]. Although for the most part, the types of cyber-fraud remain the same, the criminal approach is continuously changing, rendering this threat even more severe and active. Despite the latter, as reported by McMullan and Rege [19], the industry is still yet to fully mature in literature which investigates the problem of cyber-fraud.

Previous fraud detection investigations include both unsupervised and anomaly detection techniques. We consider the fraudulent activity to be anomalous, and thus, should deviate from normal. Yamanishi et al. [36] implemented a multivariate unsupervised outlier detection technique on medical insurance data which flags outliers using the Hellinger distance. Burge and Shawe-Taylor [5] also used an unsupervised approach based on the Hellinger distance between two recurrent neural networks to predict fraud in telecommunications. Tian et al. [29] used a non-parametric clustering approach for flagging fraudulent behaviours in crowd-sourcing. Cao et al. [6] also use a clustering algorithm to detect malicious accounts in online social networks. Christou et al. [7] perform fraud detection for online games of chance using a clustering approach. Bolton et al. [4] investigated an anomaly detection approach for Peer Group Analysis based on the *t*-statistic. For this research, we had access to a dataset which

The authors would like to thank Gaming Innovation Group for allowing permission to submit this manuscript.

✉ David Farrugia
david.farrugia@gig.com

Christopher Zerafa
christopher.zerafa@gig.com

Tony Cini
tony.cini@gig.com

Bruno Kuasney
bruno.kuasney@gig.com

Karen Livori
karen.livori@gig.com

[1] Gaming Innovation Group, St. Julians, Malta

included labelled instances of previously verified fraudulent players by fraud analysts. Consequently, we decided to structure this problem as a binary classification task with the two classes being: 'fraud' and 'non-fraud'.

In the iGaming business, the most common types of cyber-fraud are money laundering [7], identity theft, and bonus abuse. In money laundering cases, criminals attempt to mask the legitimacy of their funds by depositing, wagering, and finally withdrawing a percentage of their account; thus, creating a complex money trail. In such cases, the gaming operator is not directly faced with a financial hit; however, ensuring mechanisms are in place to protect against such activity is critical. The industry is heavily regulated when it comes to money laundering, and gaming organisations are required to have such measures working. Failure to comply will result in a permanent loss of licence to operate, and subsequently, a reputation hit. Identity theft occurs when criminals gain unauthorised access to a client's account or use stolen payment methods to wager money and bank the winnings. The latter is mostly done through a method called chip dumping, where two or more players (known as a syndicate) intentionally lose to each other to move the funds from one account to another. Even if the operator discovers of the original offender, linking all the players within that syndicate together proves to be a rather tedious and challenging task. The third type, bonus abuse, also use similar methods to exploit promotional content offered by the organisation through chip dumping.

Establishing good governance by taking a pro-active stance against cyber-fraud is vital to ensure that the industry continues to scale and remain a reputable source of entertainment. Most gaming operators rely on manual transaction monitoring conducted by a team of fraud analysts. When considering that a typical gaming operator records millions of daily transactions, it becomes next to impossible to scrutinise every transaction effectively. Inevitably, the innovation of automated solutions is essential.

Our Contribution

This research makes several contributions to the scientific domain. Firstly, we present a literature review of similar fraud detection solutions both in the iGaming industry and other fields in Sect. 2. We also discuss a pipeline which can predict iGaming fraud with a precision of 84.2%, evaluated on a real-life dataset. We also present a strategy to detect and combat concept-drift, which refers to a changing underlying distribution of the dataset, purely based on data-driven methods. This research extends the predictive component of the machine learning approach and investigates an application for local interpretable explanations to highlight potential fraud indicators per player. We show how our approach results in a prescriptive solution which can predict the likelihood of fraud and list the key indicators supporting that particular prediction. The latter not only allows for a pro-active stance against cyber-fraud but also facilitate the next steps for the fraud analyst, such as requesting further player identification documents or permanently blocking the player.

We structure the rest of the paper as follows. In Sect. 2, we provide a literature review of existing similar solutions. We discuss our pre-processing approach and modelling strategy in Sect. 3. Then, in Sect. 4, we present our results on the real-life dataset and respective observations. Finally, we summarise our findings and present our conclusions and future recommendations in Sect. 5.

Related Work

Several supervised fraud detection studies and solutions exist. In most scenarios, labelled datasets for fraud detection tend to be heavily imbalanced since fraudulent activity tends to be rarer than non-fraudulent [9, 21]. In a classification fraud detection task, we might have to also deal with the class imbalance issue beforehand. Whitrow et al. [35] investigated credit-card fraud. They observed that aggregating transactions over some time (one to three days) helped with dealing with the class distribution problem. The authors observe that the random forest (RF) algorithm yielded the best predictive result when compared to other supervised algorithms, like support vector machine (SVM). RF is an algorithm which builds several decision trees (DT), called an ensemble, and the generates the final prediction by calculating the mode of all decision trees. A DT is a predictive model that maps the training samples into branches and leaves to create a tree-like structure. Dhankhad et al. [9] suggest the grouping of transactions as a means to combat class imbalance along with the utilisation of network-based features. A network-based feature describes some time-dependent variable on a customer, used to generate a 'suspiciousness score'. The authors also recommend using precision, recall, and F1-score as the performance evaluation metrics for tasks with class imbalance. They noted the best performance when they under-sampled the majority class (i.e. non-fraud) as a class balancing strategy. A stacking classifier using a meta logistic regression (LR) estimator, RF, and eXtreme Gradient Boosting (XGB) achieved the best overall results. LR derives conditional probabilities based on a logistic funtion. XGB forms part of the gradient boosting type algorithms. Like RF, gradient boosting models are also ensemble techniques since they construct several weak classifiers in the form of decision trees. The outputs of these weak learners are then generalised using gradient descent optimisation. Phua et al. [24] also investigated the problem of skewed data using public insurance fraud detection dataset. The authors suggest using a stacking classifier

of bagged algorithms based on the C4.5 technique to deal with handling class imbalance. Sahin et al. [27] used a DT with cost-sensitive learning as an application to credit card fraud detection and argue that cost-sensitive learning is the best way to combat class imbalance based on their empirical results.

Li et al. [15] compare a Naïve Bayes (NB) classifier, LR, and an artificial neural network (ANN) on credit card fraud detection. This study observes that on a balanced dataset, the LR model outperformed the other two; however, on an imbalanced dataset, the NB algorithm achieved the best overall performance. Monedero et al. [22] investigated energy consumption fraud using three models based on Pearson correlation, Bayesian networks, and DT. The authors then merge the results of all three models to obtain a final list of potentially fraudulent customers. Coma-Puig et al. [8] later extended the latter and presented an empirical analysis of several machine learning techniques. The authors evaluated the algorithms on their Area Under the receiver operating characteristics (AUROC) after 4-fold cross-validation. The authors found gradient boosting (GB), RF, and AdaBoost with an NB as its base estimator to be the best-performing algorithms. Akhilomen [1] and Kim and Kim [12] also investigate credit card fraud detection using different ANN architectures. Other supervised techniques investigated include a self-organising map (SOM) [23] and combining recency-frequency-monetary (RFM) variables and social network analysis in an RF [31].

Besides the issue of class imbalance, fraudulent activity is also continuously changing, and fraudsters are always discovering newer approaches to remain undetected. In machine learning, we refer to this phenomenon as concept-drift, where the underlying statistical distribution of the data morphs over time; thus, rendering the predictive model to become less accurate by time [34]. Dealing with concept drift requires that the model is kept up-to-date with the latest behavioural trends by performing regular model re-training, i.e. incremental learning. Based on the speed of the data morphism, referred to as the drift rate by Somasundaram and Reddy [28], we can determine the necessary frequency of learning to keep the model relevant. In their study, Somasundaram and Reddy [28] demonstrate how using incremental learning and cost-sensitive estimators can help deal with both class imbalance and concept drift issues. Wang et al. [33] also describe online learning as another viable solution for dealing with concept drift. In online learning, the incoming data samples simultaneously update the learning algorithm as it flows into the model.

More recently, researchers are also investigating the topic of explainability to machine learning modelling. Most artificial intelligence (AI) algorithms are black-box machines, meaning that we cannot easily see how the model generates predictions. Even though we measure performance, ultimately, we blindly trust these algorithms and their predictions. The emerging field of eXplainable AI (XAI) is interested in extracting the knowledge and rationale behind a model's particular prediction or set of predictions. The ability to explain a specific prediction is vital, especially when such a prediction motivates or directly invokes another action or reaction [26]. XAI systems would not just help to instil trust in an AI but also act as a medium to understand better what the model believes to be the causal factors of a specific problem. XAI also provides a glimpse of how the model will behave in the future. A model can be either explained globally or locally. A global XAI system attempts to explain to the entire model while a local explanation works by generating the rationale behind every individual sample and subsequent prediction [25]. One of the most recent advancements in local explanations is the local interpretable model-agnostic explanation (LIME) technique, introduced in Ribeiro et al. [26]. Given an incoming sample for prediction, LIME approximates local linear artificial data points (in the neighbourhood of the incoming sample) and performs data perturbation to determine how every feature influences the model's prediction. This process results in prediction explanations consisting of the influential features represented as a mathematical equality.

Lundberg and Lee [17] also tackled the problem of prediction interpretability by presenting SHapley Additive exPlanations (SHAP). Similarly to LIME, SHAP also yields local explanations. Inspired by game theory, SHAP also determines the contribution of every feature to the model's prediction per incoming sample. One of the advantages of LIME and SHAP is their model agnosticism property, meaning that both techniques can work in conjunction with any model. Despite XAI still being in its infancy stage, researchers have already investigated its benefits in relation to the medical domain [10, 16, 20, 30, 32], mass surveillance [10], knowledge graphs [14], and image-based predictive maintenance [13]. Marino et al. [18] also investigated XAI as part of an anomaly detection approach for intrusion detection. To the best of our knowledge, XAI has never been explored as an application for cyber-fraud detection, particularly in the iGaming industry.

Materials and Methods

Data Preparation

Our Risk and Fraud team perform numerous risk assessments daily. Thus, we had access to a historical record of previously confirmed fraudulent cases. We extracted several data points on these players as well as other players not part of the fraudulent list, resulting in a binary classification dataset ('non-fraud' and 'fraud'). Our dataset

consisted of 451,123 players which included a total of 13,591 confirmed fraudsters, with the remaining players (437,532) not being previously flagged for fraud. Further, we filtered out those non-fraud players who had no activity recorded, bringing down the total number of players to 197,733 (184,142 of which not previously flagged as fraud). Besides reducing noise, this process also acted as an under-sampling strategy and helped improve our class imbalance issue slightly. Our fraud class represents 6.87% of the entire dataset. We monitored over 1000 dimensions which we later reduced to 25 features based on the following attributes (we further discuss this process in Sect. 3.3):

1. Multi-session behavioural aggregates
2. Gaming patterns
3. Session identification and geolocation
4. Demographics
5. Payment information

Experimental Setup

For this study, we used Python 3.7 with the Anaconda distribution as our primary development language. We conducted our experiments using Jupyter Notebooks on a 16GB RAM, 64-bit Unix system. To combat over-fitting, we used Stratified k-Fold cross-validation to consider every sample once for validation and $k-1$ times for model training.

Data Pre-processing

We observed that some of our features had missing values, mostly attributed to no player activity concerning that specific data point. To deal with this issue, we imputed all missing values using the median. We justify using the median instead of the mean due to most of our features being highly skewed. We attribute high skewness to differences in behaviour between one player and another as well as the presence of outlier samples. Scaling the dataset without properly handling this characteristic yielded sub-par results in scaling. As part of our scaling strategy, we categorised our features into three groups: Boolean's, scalars, and highly-skewed scalars. For Boolean features, we ensured that they only had either a value of 1 or 0. For the scalar category, we scaled the feature values using Eq. 1. Equation 1 allows us to scale a feature within its respective inter-quartile range. This approach is also robust to outliers. In the case of highly skewed features, we found that scaling using this approach alone did not produce satisfactory results. Therefore, we first applied a logarithmic transformation on the absolute value of the highly-skewed features before using Eq. 1.

$$x^{\tau} = \frac{x_i - Q2(x)}{Q3(x) - Q1(x)} \tag{1}$$

where x^{τ} is the transformed value, x_i is the value of feature x, $Q1(x)$ is the lower quartile of x, $Q2(x)$ is the median of x, $Q3(x)$ is the upper quartile of x.

As previously mentioned, our dataset was highly dimensional with over 1000 features. Besides adding complexity in the model, a high number of features results in an increased risk of over-fitting. We performed a 3-step feature selection strategy. Firstly, we removed all single-valued features since such a feature does not contribute any value to our solution. Secondly, we removed multi-collinear features based on a Pearson correlation coefficient threshold of 0.7, since values greater than 0.7 can be considered as highly correlated. Multi-collinear features tend to increase model complexity without supplementing value to the model, and in some cases, also harms predictive performance. This step drastically reduced our dimensionality. Finally, we trained a base LightGBM (LGB) model for several iterations to extract feature importance. LGB is another model part of the gradient boosting family, similar to XGB. We removed those features which do not contribute to a 99% cumulative importance. This strategy dropped our number of features down to 25.

Motivated by literature, we also attempted to understand whether our observed fraudulent behaviour is prone to concept-drift. We split our dataset into several bins, including 1-year, 6-months, 4-months, 1-month, and 1-week. We used the Mann–Whitney U test to assess whether the features from one bin appear to be from a different distribution when compared to the same feature from another bin. We visualised the percentage of 'significant features', per bin. Furthermore, we also set up an experiment to investigate causal inference between bins using an algorithm based on Bayesian structural time series. The algorithm compares the observed behaviour after the event against the expected behaviour and infers the statistical significance of a causal impact. The algorithm uses the Bayesian-based model for the expected behaviour by predicting posterior behaviour if the event did not occur. For us, the event was simply the end of a particular bin and the commencement of another (i.e. starting of a new year).

Predictive Modelling

Motivated by the literature in Sect. 2, we evaluated the predictive performance of RF, LGB, DT, and LR on our dataset. We used Stratified 10-fold cross-validation to minimise the risk of over-fitting and preserve the class distribution and compare the models on their respective AUROC (the model's ability to distinguish between both classes), precision (the ratio of correct positive predictions to the total positive predictions), Recall (the fraction of positive cases correctly

predicted), and F1-score (the harmonic mean of precision and recall) averages and variances. For the best-performing pipeline, we also evaluated the model's false-alarm rate (using Eq. 2). We also experimented with over-sampling the minority class using Synthetic Minority Over-sampling TEchnique (SMOTE) to balance both classes further. We evaluated all models using their default hyper-parameters (from the scikit-learn Python package). We further performed hyper-parameter tuning using Bayesian optimisation and distributed Tree of Parzen Estimators (TPE) of the best performing model:

$$\mathrm{FR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{2}$$

where FR is the false-alarm rate, FP is the false positives, TN is the true negatives.

Following training and tuning of our best model, we also investigated the concept-drift phenomenon from the perspective of predictive performance decay. With this experiment, we wanted to quantify our drift rate based on an empirical and data-driven approach. We set up an experiment to simulate production and assess how the data-drift evolves. We selected a testing size of x of 3-months and an interval of y of 1-month. We start the experiment by setting our test set equal to the latest x of our data with the remaining months forming part of the training. We repeated the experiment until the training set only included the oldest x. After every iteration, we removed the most recent month based on y from our dataset and repeated the experiment. At every iteration, we recorded the performance metrics. We selected x and y based on previous analysis. We further illustrate this simulation in Fig. 1. We also used the Mann–Whitney U test again to map the statistical significance between the change in the percentage of significant features against the change in predictive performance. This algorithm has a time-complexity of $O(n^2)$ and was executed on a total of 3450 binned observations, which took 668 ms to complete.

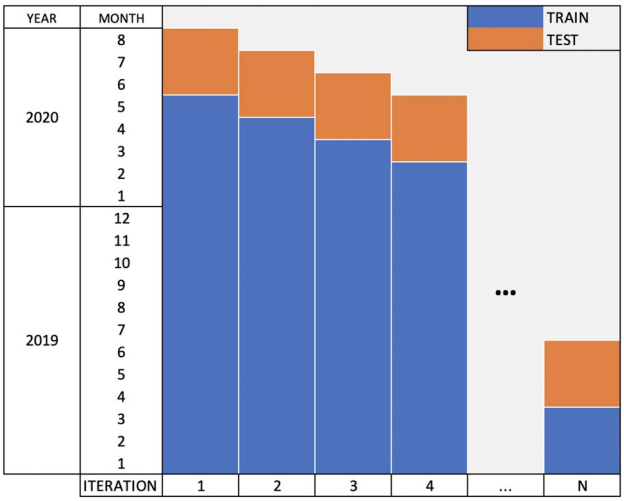

Fig. 1 Schematic of the simulation experiment to quantify the drift-rate

Local Interpretable Explanations

The final part of our pipeline was to extract explanations for every individual prediction. We used locally faithful explanations to determine which features had the most significant contribution for that particular instance and quantify that same contribution. These explanations are represented as feature conditions in the form of mathematical inequalities, and every condition has an attributed relative strength (positive if the contribution was towards the positive class and vice-versa). Our approach was to extract the top 5 features driving the prediction and extract their respective conditions. We filtered out those conditions which had a negative relative strength, primarily because our interest lies in explaining which features are indicative of potentially fraudulent behaviour—answering the question: why does our model think that this player is fraudulent?

Results and Discussion

As discussed in "Predictive Modelling", we evaluated the LGB, RF, DT, and LR techniques on multi-session data using stratified 10-fold cross-validation. We show the obtained averaged results in Table 1.

We demonstrate that the LGB achieved the best performance. The LGB algorithm outperformed all of the other evaluated algorithms on the AUROC, recall, and F1-score. Considering the precision, we note the best result from

Table 1 Performance results of the evaluated techniques

Model	AUROC	Precision	Recall	F1	Time (s)
LGB	**0.818 ± 0.054**	0.842 ± 0.063	**0.644 ± 0.115**	**0.722 ± 0.044**	28.310
RF	0.799 ± 0.042	**0.877 ± 0.054**	0.602 ± 0.083	0.713 ± 0.075	64.730
DT	0.809 ± 0.042	0.620 ± 0.068	0.641 ± 0.085	0.630 ± 0.076	**25.861**
LR	0.732 ± 0.069	0.808 ± 0.082	0.471 ± 0.137	0.591 ± 0.129	135.326

The bold numbers refer to the best in that column

the RF, which also obtained the most consistent (in terms of variance) results across all metrics. Both DT and LR obtained mediocre results when compared to the RF and LGB techniques. In our use-case, we prioritised precision. This inherently excludes DT from any future analysis given that this technique yielded the lowest precision value from the group. Although the LR achieved rather good precision values, overall, the RF and LGB shadowed it on all metrics. We selected LGB as our best-performing pipeline, mainly for two reasons: better AUROC and a better F1. Although RF obtained the best precision values, the LGB approach seems to have yielded a more robust model overall. With a slightly higher AUROC and F1-score, LGB appears the most balanced model out of the two. Further, LGB resulted in an average fraud false-alarm rate of 0.402%. For all algorithms, we tried over-sampling with SMOTE; however, we did not observe any performance improvements. As aforementioned, we performed automated TPE hyper-parameter tuning on LGB which considered rounded uniform 'num_leaves', log uniform 'learning_rate', rounded log uniform 'min_data_in_leaf', uniform 'bagging_fraction', and uniform 'feature_fraction'. We set our hyper-parameter process to perform 50 evaluations, which took about 2 h to complete. The number of evaluations is directly proportional to the execution time and quality of the hyper-parameters.

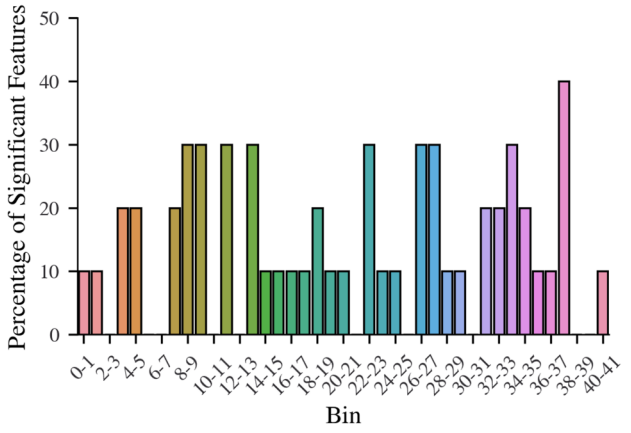

Fig. 2 Percentage of features exhibiting distribution-drift using monthly bins

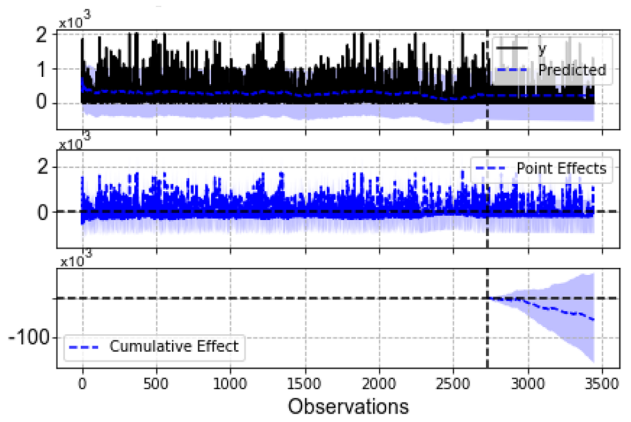

Fig. 3 Causal impact of the start of COVID-19 on the data distribution

Quantifying the Drift-Rate

As our first attempt to quantify the drift-rate, we used the Mann–Whitney *U* test to investigate feature distribution morphism. We tested the latter for different buckets, starting from yearly bins to monthly. We observed that for the broader bins, the majority of the features (60–70% of the features) did show distribution-drifts. This behaviour appears to be tamed when considering the monthly bins, albeit still fluctuating (refer to Fig. 2).

From month to month, we expect around 10–40% of the features to suffer drift. We also note that the highest percentage drift recorded came from bin 37 to 38, which represented early 2020. Since this time coincided with the COVID-19 outbreak, we started to correlate such drifts with this event. To further investigate this behaviour, we visualised the distribution-drift for weekly bins. We observed similar behaviour to that of the monthly bins, and thus, excluded the weekly bins and investigated further using the monthly bins. One added benefit of this data-drift detection strategy is the added coverage of how the data is changing. By monitoring distribution morphism of every individual feature, we were able not just to understand the quantitative severity of the drift but also investigate which features are drifting. Following, we investigated the causal impact on the underlying distribution of the data by two events: the start of the year 2020 and the start of the COVID-19 pandemic (February/March 2020). After both events, we can observe a negative cumulative effect on the data distribution. However, with a level of significance of 0.05, both results are not statistically significant. We show the causal impact of COVID-19 in Fig. 3, which is also representative of what we observed for the 2020 causal impact. Although there is not enough statistical evidence for us to attribute such drift to a particular event, we definitely cannot argue that concept-drift is not present in our dataset. Without training the model on any 2020 data, we noticed poor results (around 50% precision and 1% recall).

We can also confirm the presence of data-drift through the simulation experiment. Similarly to our previous results, Fig. 4 shows a sharp drop in recall at around the 9th (November/December 2019) and 8th (February/March 2020) iterations. The deeper we go into 2020, the worse the

observed recall. Based on these results, we appear to have a drift-rate of approximately 1-month. Hence, we need to perform model re-training at least every month.

Prescriptive Analytics Using XAI

Following the process as discussed in Sect. 3.5, we ended up with a list of explanations attributed to every individual prediction. We visualised these explanations accordingly. We show an example of such visualisations in Fig. 5. In this particular instance, the model correctly predicted this fraudulent player with a probability of 0.98. The model picked up a country mismatch at login (the strongest indicator) and irregular patterns in the bet counts and number of payments as the leading indicators of potentially fraudulent behaviour.

Further, we can understand that the model flagged the login mismatch and betting behaviour because they were > 0, while the total payments were ≤ 0 (based on the scale-transformed values). This process took approximately 2 seconds per instance. Through XAI, we are essentially prescribing courses-of-action to our fraud analysts, saving them time and optimising their workflow. XAI serves as padding between our model and the fraud analysts. Using XAI allows us also to minimise the adverse effects of incorrect predictions. The main drawback of a wrong prediction is eating away time which the fraud analyst could have used to investigate other cases. With the addition of XAI, we are drastically reducing this time by letting the fraud analysts know precisely where to look. Furthermore, these explanations are aiding the fraud analysts in uncovering emerging fraudulent behaviours and discovering other fraudulent patterns, which are not straightforward or manifested. Ultimately, our application of XAI methodology to generate prescriptive predictions equips our fraud analysts with the necessary tools to get a leg up on the fraudulent activity.

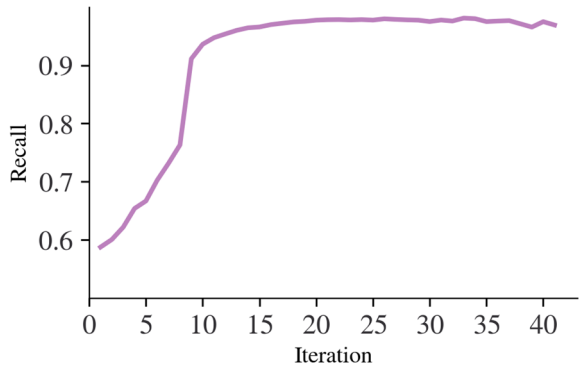

Fig. 4 Recall performance decay per iteration using monthly bins

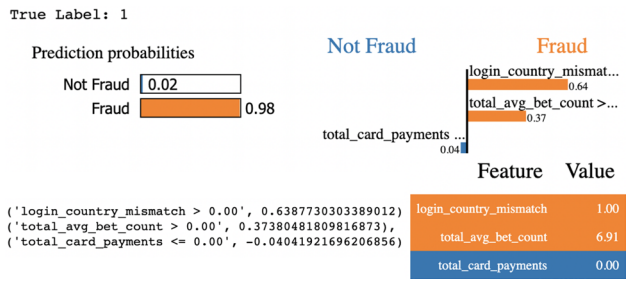

Fig. 5 An example of an explainable prediction

Conclusions

In this research, we investigated the effectiveness of machine learning techniques to flag fraudulent behaviour in the iGaming industry. We demonstrated how one could leverage existing XAI algorithms to explain individual predictions for prescriptive solutions and extract additional knowledge on the causation of cyber-fraud. We had access to a labelled ('fraud' and 'non-fraud') dataset with a sample size of 197,733, where our fraudulent instances represented 6.87% of the entire dataset. We tackled this problem as a binary classification task. We compared pipelines based on the RF, LGB, DT, and LR techniques, evaluated using stratified 10-fold cross-validation. We show that the LGB algorithm achieved an AUROC of 0.818 ($\pm$ 0.054), a precision of 0.842 ($\pm$ 0.063), a recall of 0.644 ($\pm$ 0.115), and an F1-score of 0.722 ($\pm$ 0.044), outperforming the other evaluated models.

Further, we tuned the hyper-parameters of the LGB model using Bayesian optimisation methods. We also addressed the phenomenon of concept-drift through an empirical and data-driven strategy which confirmed the presence of data-drift. We concluded that the drift-rate was around the 1-month mark, suggesting a monthly re-training for the model to remain reliably updated. Using this approach, we were also able to understand how the players' behaviour drifts. We also extended the predictive component of our work by leveraging the emerging field of XAI to generate prescriptive solutions through locally faithful explainable predictions. One limitation of this research is the manually labelled dataset, which might have introduced bias and human-error in our analysis. Nonetheless, given that the same team that labelled this data will be using our solution, the effect of this limitation is diminished. In our future work, we will be investigating the relevance of online learning to combat the concept-drift problem further.

Funding This work was funded and supported by Gaming Innovation Group.

Availability of data and material Due to the confidential nature of this work, the data used cannot be made publicly available.

Code availability Due to the confidential nature of this work, we cannot share the source code. All modules used here fall under open-source license. Source code is available within references presented.

Declarations

Conflict of interest All authors are with Gaming Innovation Group.

References

1. Akhilomen J. Data mining application for cyber credit-card fraud detection system. In: Perner P (ed) Advances in data mining. applications and theoretical aspects, lecture notes in computer science. Springer, Berlin; 2013. pp. 218–228. https://doi.org/10.1007/978-3-642-39736-3_17.
2. Association of Certified Fraud Examiners: ACFE Report to the Nations | 2020 Global Fraud Study. Technical report, Association of Certified Fraud Examiners. 2020. http://www.acfe.com/report-to-the-nations/2020/.
3. Banks J. Online gambling and crime: a sure bet? ETHICOMP J. 2012.
4. Bolton RJ, Hand DJ, H DJ. Unsupervised profiling methods for fraud detection. In: Proceedings of credit scoring and credit control VII, 2001. pp. 5–7.
5. Burge P, Shawe-Taylor J. An unsupervised neural network approach to profiling the behavior of mobile phone users for use in fraud detection. J Parallel Distrib Comput. 2001;61(7):915–25.
6. Cao Q, Yang X, Yu J, Palow C. Uncovering large groups of active malicious accounts in online social networks. In: Proceedings of the 2014 ACM SIGSAC conference on computer and communications security, CCS '14, pp. 477–488. Association for Computing Machinery, New York, NY, USA. Event-place: Scottsdale, Arizona, USA. 2014. https://doi.org/10.1145/2660267.2660269.
7. Christou IT, Bakopoulos M, Dimitriou T, Amolochitis E, Tsekeridou S, Dimitriadis C. Detecting fraud in online games of chance and lotteries. Expert Syst Appl. 2011;38(10):13158–69.
8. Coma-Puig B, Carmona J, Gavaldà R, Alcoverro S, Martin V. Fraud detection in energy consumption: a supervised approach. In: 2016 IEEE international conference on data science and advanced analytics (DSAA), 2016. pp. 120–129. https://doi.org/10.1109/DSAA.2016.19.
9. Dhankhad S, Mohammed E, Far B. Supervised machine learning algorithms for credit card fraudulent transaction detection: a comparative study. In: 2018 IEEE international conference on information reuse and integration (IRI), 2018. pp. 122–125. https://doi.org/10.1109/IRI.2018.00025.
10. Hossain MS, Muhammad G, Guizani N. Explainable AI and mass surveillance system-based healthcare framework to Combat COVID-I9 like pandemics. IEEE Netw. 2020;34(4):126–132 **(Conference Name: IEEE Network)**
11. Karpoff JM. The future of financial fraud. J Corp Finance. 2020;101694.
12. Kim MJ, Kim TS. A neural classifier with fraud density map for effective credit card fraud detection. In: Yin H, Allinson N, Freeman R, Keane J, Hubbard S, editors. Intelligent data engineering and automated learning—IDEAL 2002, lecture notes in computer science. Springer, Berlin; 2002. pp. 378–383. https://doi.org/10.1007/3-540-45675-9_56.
13. Krishnamurthy V, Nezafati K, Stayton E, Singh V. Explainable AI framework for imaging-based predictive maintenance for automotive applications and beyond. Data-Enabled Discov Appl. 2020;4(1):7.
14. Lecue F. On the role of knowledge graphs in explainable AI. Semant Web. 2020;11(1):41–51 **(Publisher: IOS Press)**.
15. Li Z, Zhang H, Masum M, Shahriar H, Haddad H. Cyber fraud prediction with supervised machine learning techniques. In: Proceedings of the 2020 ACM southeast conference, ACM SE '20. Association for Computing Machinery, New York, NY, USA. 2020. pp. 176–180. https://doi.org/10.1145/3374135.3385296. Event-place: Tampa, FL, USA.
16. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee SI. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2(1):56–67.
17. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, editors. Advances in neural information processing systems, vol. 30. Curran Associates, Inc. 2017. pp. 4765–4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.
18. Marino DL, Wickramasinghe CS, Manic M. An adversarial approach for explainable AI in intrusion detection systems. In: IECON 2018—44th annual conference of the IEEE Industrial Electronics Society; 2018. pp. 3237–3243. https://doi.org/10.1109/IECON.2018.8591457**(ISSN: 2577-1647)**.
19. McMullan JL, Rege A. Online crime and internet gambling. J Gambl Issues. 2010;24:54–85.
20. Meacham S, Isaac G, Nauck D, Virginas B. Towards explainable AI: design and development for explanation of machine learning predictions for a patient readmittance medical application. In: Arai K, Bhatia R, Kapoor S, editors. Intelligent computing, advances in intelligent systems and computing. Springer International Publishing, Cham. 2019. pp. 939–955. https://doi.org/10.1007/978-3-030-22871-2_67.
21. Melo-Acosta GE, Duitama-Muñoz F, Arias-Londoño JD. Fraud detection in big data using supervised and semi-supervised learning techniques. In: 2017 IEEE Colombian conference on communications and computing (COLCOM); 2017. pp. 1–6. https://doi.org/10.1109/ColComCon.2017.8088206.
22. Monedero I, Biscarri F, León C, Guerrero JI, Biscarri J, Millán R. Detection of frauds and other non-technical losses in a power utility using Pearson coefficient, Bayesian networks and decision trees. Int J Electric Power Energy Syst. 2012;34(1):90–8.
23. Olszewski D. Fraud detection using self-organizing map visualizing the user profiles. Knowl Based Syst. 2014;70:324–34.
24. Phua C, Alahakoon D, Lee V. Minority report in fraud detection: classification of skewed data. SIGKDD Explor Newsl. 2004;6(1):50–9.
25. Rai A. Explainable AI: from black box to glass box. J Acad Market Sci. 2020;48(1):137–41.
26. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, San Francisco, CA, USA, August 13–17, 2016, 2016. pp. 1135–1144.
27. Sahin Y, Bulkan S, Duman E. A cost-sensitive decision tree approach for fraud detection. Expert Syst Appl. 2013;40(15):5916–23.
28. Somasundaram A, Reddy S. Parallel and incremental credit card fraud detection model to handle concept drift and data imbalance. Neural Comput Appl. 2019;31(1):3–14.

29. Tian T, Zhu J, Xia F, Zhuang X, Zhang T. Crowd fraud detection in internet advertising. In: Proceedings of the 24th international conference on world wide web, WWW '15. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE. Event-place: Florence, Italy; 2015. pp. 1100–1110. https://doi.org/10.1145/2736277.2741136.
30. Tosun AB, Pullara F, Becich MJ, Taylor DL, Fine JL, Chennubhotla SC. Explainable AI (xAI) for anatomic pathology. Adv Anat Pathol. 2020;27(4):241–50.
31. Van Vlasselaer V, Bravo C, Caelen O, Eliassi-Rad T, Akoglu L, Snoeck M, Baesens B. APATE: a novel approach for automated credit card transaction fraud detection using network-based extensions. Decis Support Syst. 2015;75:38–48.
32. Wang D, Yang Q, Abdul A, Lim BY. Designing theory-driven user-centric explainable AI. In: Proceedings of the 2019 CHI conference on human factors in computing systems, CHI '19. Association for Computing Machinery, New York, NY, USA; 2019. pp. 1–15. https://doi.org/10.1145/3290605.3300831.
33. Wang S, Minku LL, Yao X. A systematic study of online class imbalance learning with concept drift. IEEE Trans Neural Netw Learn Syst. 2018;29(10):4802–21.
34. Webb GI, Hyde R, Cao H, Nguyen HL, Petitjean F. Characterizing concept drift. Data Min Knowl Discov. 2016;30(4):964–94.
35. Whitrow C, Hand DJ, Juszczak P, Weston D, Adams NM. Transaction aggregation as a strategy for credit card fraud detection. Data Min Knowl Discov. 2009;18(1):30–55.
36. Yamanishi K, Takeuchi Ji, Williams G, Milne P. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. Data Min Knowl Discov. 2004;8(3):275–300.