Journal of Applied Intelligent Systems & Information Sciences

VOL. 4, ISSUE 1, PP. 60-73, NOVEMBER 2023.
Available at: **https://journal.research.fanap.com/**
DOI: https://doi.org/10.22034/JAISIS.2023.409119.1064

A Novel Hybrid Online Gambling Detection System Using Ensemble Classifiers and Rule-Based Methods

Ali Zare [1,*], Kimia Ghasemzadeh [1], Fatemeh Moslehi [1], Niloofar Haghjoo [1]

[1] *Dotin Research and Studies Center, Datis Arian Qeshm (Dotin), Tehran, Iran.*

Abstract

Detecting fraudulent activities in the banking industry is a challenging task, especially when they involve country-specific restrictions such as unauthorized access to online gambling platforms. Machine learning techniques have emerged as a powerful tool for online gambling detection, but they require a large and reliable labeled dataset to build effective classifiers. This article presents a novel hybrid approach that combines a rule-based system and an ensemble classifier to help and automate gambling transaction detection.
The rule-based system generates labeled data by applying predefined rules to the transactions, and the tree-based ensemble classifier uses the labeled data to learn and differentiate between gambling and non-gambling transactions. Our proposed hybrid approach operates on the principle that both the rule-based system and the machine learning classifier complement each other in a validation cycle and this combination improves the accuracy of the results and produces a robust model. In addition, the ensemble classifier can help detect transactions that were not detected by the rule-based system as they have different and new behaviors compared to this type of fraud.
The hybrid approach was implemented and evaluated using a real-world dataset of an Iranian bank. The results show that LightGBM achieved the best performance with an accuracy of about 97% and an F1 measure of about 96%. The results also demonstrate that the proposed hybrid approach is effective and feasible in illegal gambling detection.
***Keywords*:** Online Gambling Detection, Ensemble Classifiers, Rule-based methods, Tree-based Classifiers

1. Introduction

Gambling is an activity that has been around for centuries and has evolved. It involves wagering some value on an event with an uncertain outcome, with the primary intent of winning additional value or money. The activity requires three elements to be present: consideration (an amount wagered), risk (chance), and a prize. Gambling can take many forms, including casino games, sports betting, lottery games, etc. It is a complex phenomenon that has been studied from various perspectives, including psychology (Pfund et al., 2023; Hagfors et al., 2023; Coelho et al., 2023), sociology (Bond et al., 2023; Manoj Arvind, 2023), and economics (Badji et al., 2023; Chóliz, 2023; Harris et al., 2023; Paterson, 2020). While participation in gambling is a personal choice and can be viewed as harmless entertainment, individuals with a solid understanding of statistics and probability

[*] *Corresponding Author, Email: Ali.zare@dotin.ir*

Received: 01 August 2023; **Accepted**: 11 November 2023; **Published Online**: 12 November 2023

are less susceptible to gambling fallacies (Williams & Connolly, 2006). Gambling has affected both adults and adolescents more than ever before, leading to potential mental health issues and pathological gambling (Floros, 2018). According to a study published in BMC Public Health, responsible gambling messages are widely used as a tool to enable informed choice and encourage appropriate gambling behavior. The study also suggests that gamblers have different levels of risk of developing gambling problems and require various harm-minimization tools and resources (Gainsbury et al., 2018). It is important to note that gambling can have negative consequences on individuals and society as a whole. A conceptual model proposed in another study published in BMC Public Health highlights the social and economic impacts of gambling on gamblers, their significant others, and society (Latvala et al., 2019).

It is important to note that gambling is illegal in certain states of the USA, several European countries like Poland, and almost all Asian countries, including the Islamic Republic of Iran. In these countries, underground and online gambling is a rapidly growing industry that has gained significant attention from researchers in recent years. A study published in the Journal of Gambling Studies highlights the reasons and attitudes of university students toward online gambling and its associated social, economic, and academic implications on their lives. The study found that the ease of making quick money, the anonymous nature of online gambling, and a source of entertainment were the main reasons why participants engaged in online gambling. However, the study also revealed that participants who gambled online ended up becoming depressed, had difficulties sleeping, and barely concentrated in class because of their addictive attitudes toward online gambling (Amoah-Nuamah et al., 2023). Another article published in Frontiers discusses the impact of COVID-19 on online gambling. The article suggests that online casino and bingo gambling appear to be less affected by the COVID-19 crisis, while land-based gambling in these online gamblers appeared to be scarcer (Håkansson, 2020).

In recent years, online gambling has gained immense popularity worldwide and has become a highly profitable industry with an expected CAGR of 11.5% from 2020-2030 (Grandviewresearc Report, 2023). Gambling transaction detection is a critical aspect of anti-money laundering (AML) controls. Casinos and online gambling platforms are cash-intensive facilities that attract criminals looking to launder illicitly obtained funds. These facilities offer gamblers anonymity and the ability to transact large amounts of money at a rapid scale (Tomic, 2022; Gore, 2023). Criminals exploit the anonymity provided by the casino environment by giving the casino incorrect, incomplete, and/or vague identifying information, meaning red flags may not be raised around their identity. To detect such transactions, casinos that are subject to the federal Bank Secrecy Act (BSA) must implement AML programs that include procedures for detecting and reporting suspicious transactions. Online gambling platforms can also implement risk-based fraud detection software solutions that help keep in line with KYC mandates and address risks associated with the online gambling sector specifically.

A study published in the European Journal of Public Health analyzed bank transaction data to investigate the amounts and types of gambling consumed by indebted individuals. The study found that gambling and indebtedness are strongly linked, with casino-type gambling being the most popular among indebted individuals (Marionneau et al., 2023). Transaction monitoring is also an essential part of preventing online gambling platforms from becoming a part of criminal money laundering or the financing of terrorism (Gainsbury, 2012).

Another study published in the Journal of Gambling Studies highlights the importance of detecting anomalous transactions in online gambling platforms (Min et al., 2021). The study proposes a method to detect and prevent these anomalies by conducting a comparative analysis of the results of existing anomaly detection algorithms. The study also suggests that machine learning algorithms such as ANN, SVM, HMM, and clustering can be used for fraud detection in online gambling. In addition, a recent paper proposes a novel system based on Generative Adversarial Networks (GANs) for generating synthetic data to train a supervised classifier for fraud detection. The Synthetic Data Generation GAN (SDG-GAN) framework outperformed density-based oversampling methods and improved the classification performance of benchmark datasets and real-world gambling fraud datasets (Charitou et al., 2021).

Although there has been significant progress in the field of fraud detection in gambling, which aims to capture anomalies in gambling to detect money laundering or fraudulent activity, there are few studies on detecting gambling transactions themselves or gambling organizations. In previous studies, researchers had access to gambling organizations' information, but this is precisely the information that we want to detect. It appears that we want to detect the activities of an organization through its transactions which is a very challenging area due to a lack of information and limited background studies.

In this study, our objective was to differentiate card-to-card transactions that are associated with gambling organizations. As previously mentioned, gambling is illegal in Iran. Therefore, at the outset of the study, our transactions were unlabeled as gambling or non-gambling. To overcome this limitation, we established a two-step algorithm. In the first step, we used specific gambling transactions reported by The Central Bank of Iran (CBI) as a clue. By analyzing the behaviors of the debit cards linked to these transactions, we were able to detect patterns indicative of gambling activities. Using these patterns as a basis, we established rules and designed a rule-based system.

By implementing this system in real-time, we gradually accumulated a significant volume of transactions and cards that were unequivocally connected to gambling activities. In the second step, we employed tree ensemble learning models including Random Forest, XGBoost, and LightGBM to predict the gambling transactions. After examining all three models, we observed that all models are highly accurate. However, LightGBM is slightly better than the other two.

The subsequent sections of the paper are structured as follows: Section 2 outlines our proposed model and the various steps involved in constructing the detection system. Section 3 details our implementation and evaluation procedures for the proposed model, including performance results. Finally, Section 4 presents our conclusions and discussions.

2. PROPOSED MODEL

a. Building the model

Recognizing the limitations of a rule-based system in adapting to new types of gambling activities and detecting those attempting to evade our system, we sought to develop a more robust model. Rule-based systems have drawbacks in fraud detection, as they are not inherently equipped to handle novel forms of fraud, such as online gambling (Amarasinghe et al., 2018).

The process of creating our machine learning classifier involved several steps:

- Data Gathering: After analyzing the CBI report which identified certain debit cards and transactions in the bank as gambling-related, we established a rule-based system to detect and block such transactions in real-time. Subsequently, we collected and compiled data from our implemented system.
- Data Preparation: We preprocessed the gathered data by cleaning, transforming, normalizing data, and handling missing values.
- Feature Selection/Engineering: Drawing upon our knowledge of gambling transactions and extensive analysis, we created relevant features to enhance the model's performance.
- Data Split: We divided the dataset into training and test sets to evaluate the model's performance effectively.
- Model Validation: We validated the model using appropriate evaluation techniques to ensure its accuracy and reliability.
- Hyperparameter Tuning: We fine-tuned the model by optimizing its hyperparameters to enhance its performance further.
- Testing: Finally, we assessed the model's performance using the test set to evaluate its ability to detect and classify gambling transactions accurately.

Tree-based classifiers are often recommended for fraud detection tasks due to their inherent advantages. Decision trees and random forests, which are examples of tree-based classification algorithms, are well-suited for fraud detection because they can handle both categorical and numerical data. Moreover, these algorithms excel at capturing non-linear relationships between variables (Caiola & Reiter, 2010; Breiman, 2001).

Fraudulent behavior often involves intricate patterns and interactions between multiple variables, which traditional linear models may struggle to identify. Tree-based algorithms address this challenge by recursively partitioning the data into smaller subsets based on informative variables until making a final prediction.

Additionally, tree-based classifiers provide transparency and interpretability (Hilas & Sahalos, 2007), which are crucial in fraud detection. The resulting decision tree can be analyzed to understand the key variables and conditions driving the identification of fraudulent behavior. This insight can be leveraged to enhance fraud prevention strategies and inform future investigations.

b. The classifiers

The aim of this part of the study is twofold: First, to highlight the importance and significance of the use of tree-based classifiers in the field of fraud detection in the specific context of the Iranian banking sector; and second, to provide a comprehensive overview of the steps involved in building and training machine learning models specifically tailored to the identification and classification of gambling transactions. In addition, evaluation metrics, including precision, recall, accuracy, and F1 measures, which serve as essential benchmarks for evaluating the performance of machine learning models, were highlighted.

Tree-based classifiers, such as Random Forest, XGBoost, and LightGBM, are commonly used in fraud detection tasks due to their inherent strengths (Priscilla & Prabha, 2020; Ge et al., 2020; Taha & Malebary, 2020; Xuan et al., 2018). These classifiers possess the ability to handle large feature sets, capture non-linear relationships, and provide interpretable results. The following sections provide an in-depth explanation of each classifier:

2.1.1. Decision Trees

Before discussing XGBoost, Random Forest, and LightGBM, it is important to understand how decision trees work. Decision trees are tree-like structures where each node represents a feature, and each edge represents a decision rule. The leaves represent classifications or numerical values. When a new sample is fed into the decision tree, it traverses down the tree based on the values of the features until it arrives at a leaf, which provides the predicted target value.

2.1.2. Random Forest

Random Forest is an ensemble learning method that uses a collection of decision trees to make predictions. It addresses the issue of overfitting and improves generalization by averaging the predictions of individual trees. Each tree in the forest is trained on a random subset of features and data instances, promoting diversity, and reducing variance. At each node, a random subset of features is considered for splitting. The final prediction is obtained by averaging the predictions of all the trees in the forest.

Random Forest is known for its robustness, scalability, and ability to handle high-dimensional data. It can handle missing values and reduce overfitting. However, its main disadvantage is that it can be slow to train on large datasets (Bentéjac et al., 2021).

2.1.3. XGBoost

Extreme Gradient Boosting algorithm (XGBoost) is a powerful and popular gradient-boosting framework that incorporates an ensemble of weak prediction models, such as decision trees, to create a strong predictive model. The algorithm trains trees sequentially and reduces errors made by the previous trees to improve accuracy. XGBoost also uses regularization techniques, such as L1 and L2 regularization, to prevent overfitting and can handle imbalanced datasets well. It supports various types of input data, including sparse input data for tree and linear boosters, and allows the use of customized objective and evaluation functions.

XGBoost is designed for both regression and classification tasks and can handle large and complex datasets. It leverages several features like parallel processing, hardware optimization, cache optimization, and out-of-core computing to ensure efficient computation speed and optimal system performance. In addition, it offers stochastic gradient boosting, which leverages sub-sampling at different levels, to increase efficiency further. XGBoost is a reliable and highly accurate machine learning algorithm that has gained traction in structured data problems due to its speed and accuracy, as well as its ability to handle missing values and complex patterns (Chen & Guestrin, 2016).

2.1.4. LightGBM

LightGBM is a high-performance gradient-boosting framework that uses decision trees for ranking, classification, and regression tasks. It employs the leaf-wise growth technique to grow trees node-by-node, maximizing loss reduction at each node, resulting in faster training times and higher accuracy. Additionally, LightGBM's histogram-based binning technique further enhances its speed and performance by reducing the number of possible split points. It also offers scalability and memory efficiency through Gradient-based One-Side Sampling (GOSS) to prioritize instances with larger gradients. This makes it an excellent choice for handling large and complex datasets, including fraud detection. However, this approach may result in overfitting, which can be handled using the max-depth parameter. Compared to XGBoost, LightGBM builds more complex trees with fewer nodes, making it faster but less robust (Al Daoud, 2019).

2.2. Comparison

All three algorithms - Random Forest (RF), XGBoost, and LightGBM - use decision trees to make predictions and are ensemble methods that combine multiple decision trees. They also have hyperparameters that can be tuned to improve their performance. RF trains trees independently, whereas XGBoost and LightGBM train trees sequentially. While XGBoost uses a regularization term to prevent overfitting, LightGBM uses histogram-based binning to speed up training. Additionally, LightGBM uses a leaf-wise growth strategy, enabling it to create more complex trees with fewer nodes, leading to both higher accuracy and faster training times. In terms of the differences between LightGBM and XGBoost, they include leaf growth, categorical feature handling, and missing values handling, as well as respective feature importance methods. LightGBM uses Gradient-Based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) techniques to achieve high accuracy levels and fast model execution. GOSS retains instances with larger gradients and performs random sampling on instances with smaller gradients, while EFB is a near-lossless method to reduce the number of effective features. On the other hand, XGBoost uses a pre-sorted and histogram-based algorithm for computing the best split and treats nominal variables as numerical ones by default. When it comes to missing values, both algorithms assign them to the side that reduces loss the most in each split.

3. Evaluation

In this section, we describe how we implemented our detection models and evaluated their performance. To compare the performance of different algorithms we must use reliable criteria for measuring detection performance.

3.1. Evaluation metrics

To assess the performance of machine learning models, several evaluation metrics are commonly used. In the context of this project, the following metrics are employed (Townsend, 1971; Erickson & Kitamura, 2021):

3.1.1. Precision

In machine learning classification problems involving multiple classes, a confusion matrix is a performance measurement tool. The confusion matrix, depicted in Table 1, represents the combinations of predicted and actual values using four different categories. It serves as a valuable resource for evaluating various metrics, including Recall, Precision, and Accuracy. By analyzing the information presented in the confusion matrix, it becomes possible to gain insights into the classifier's performance and make informed decisions.

Table 1. The confusion matrix

		Predicted	
		Negative	Positive
ACTUAL	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

According to Table 1, Precision is calculated as below:

$$Precision = True\,Positive/(TruePositive + FalsePositive) \qquad (1)$$

Precision is a metric that assesses the accuracy of a classifier in correctly identifying positive instances, such as gambling debit cards, out of all the instances predicted as positive. It quantifies the classifier's ability to minimize false positives by accurately identifying non-gambling debit cards.

3.1.2. Recall

Recall, also known as sensitivity or true positive rate, measures the proportion of actual positive instances (gambling debit cards) that are correctly identified by the classifier. It highlights the classifier's ability to detect gambling debit cards.

$$Recall = True\,Positive/(TruePositive + FalsePositive) \quad (2)$$

3.1.3. Accuracy

Accuracy represents the overall correctness of the classifier's predictions across all classes. It calculates the ratio of correct predictions to the total number of predictions. While accuracy is an essential metric, it may not be sufficient when dealing with imbalanced datasets, such as fraud detection.

$$Accuracy = TruePositive + True\,Negative/Total\,(TP + TN + FP + FN) \quad (3)$$

3.1.4. F1 Measure

The F1 measure combines precision and recall into a single metric, providing a balanced evaluation of the classifier's performance. It is the harmonic mean of precision and recall and provides a single score that balances both metrics. The F1 measure is particularly useful when the dataset is imbalanced, as it considers both false positives and false negatives.

$$F1_{Measure} = 2 * Recall * Precision/(Recall + Precision) \quad (4)$$

4. Model building

The following steps outline the process of building the machine learning models for detecting gambling transactions:

4.1. Dataset Label Assignment

To construct our dataset, we generated features by analyzing the transactions daily for each target card. This process resulted in a total of 267,309 records, where each record represents the transactions for a specific day.

It is important to highlight that the dataset exhibits an imbalance, with a greater number of normal transactions compared to fraudulent ones. To assign labels to the records, we utilized our rule-based system to identify and flag transactions associated with gambling. These flagged records were labeled as 1, indicating a gambling transaction on the corresponding date, while the remaining records were labeled as 0, representing non-gambling transactions.

This approach allowed us to capture the nuances of daily transactions and create a comprehensive dataset that accounts for the dynamic nature of gambling activities. By incorporating the rule-based system, we ensured that the labeled records accurately reflect the presence or absence of gambling transactions, providing a solid foundation for training, and evaluating our machine learning models.

4.2. Feature Extraction

Distinguishing features, which are indicative of gambling destination transactions, are extracted from the debit card trades. These features may include transaction amounts, locations, timestamps, or other relevant attributes. For a complete description of the attributes, see Table 2. These extracted features are crucial inputs for the subsequent training process.

Table 2. Description of created features

Feature	Description	Row ID

dest_cardidhash	destination's cardid	1
transactiondate	transactiondate	2
min_amou_de	minimum of amounts traded	3
max_amou_de	maximum of amounts traded	4
avg_amou_de	average of amounts traded	5
ATM_cnt_de	number of transactions done using ATM	6
Internet_cnt_de	number of transactions done using the Internet	7
repmax_de	The maximum number of repetitions of a fixed number of transactions deposited to the card	8
repcnt_de	Number of deposit transactions with similar amounts	9
repsum_de	The total amount of similar transactions	10
reprate_de	The ratio of similar transactions to total transactions of a card number	11
repmode_de	The most frequent number of similar transactions	12
cnt_rep_card_de	The number of duplicate source cards	13
totaltr_de	Number of deposit transactions	14
max_rep_card_de	The highest number of transactions with one origin card	15
rate_rep_card_de	The ratio of the number of transactions with duplicate origin to the total number of transactions	16

4.3. Training and Testing Data Split

The dataset is divided into training and testing subsets. In this project, four months is used for training the models, while two months is reserved for testing the models' performance. This split ensures that the models encounter unseen data during testing, which assesses their ability to generalize.

The three tree-based classifiers (Random Forest, XGBoost, and LightGBM) are trained using the labeled training data. The algorithms learn patterns and relationships between the extracted features and the corresponding labels. Training involves optimizing the classifiers' internal parameters to minimize prediction errors and improve accuracy.

4.4. Model Evaluation

After training, the model's performance is evaluated using the reserved test dataset. The evaluation involves measuring Precision, Recall, Accuracy, and F1 measures to assess the models' ability to correctly classify debit card transactions into gambling and non-gambling organizations.

4.5. Fine-Tuning

To optimize the ensemble model further, fine-tuning is performed by adjusting the model's hyperparameters. This process aims to enhance the model's accuracy and generalization capability by finding the optimal configuration.

Fig. 1 shows the schema of our method:

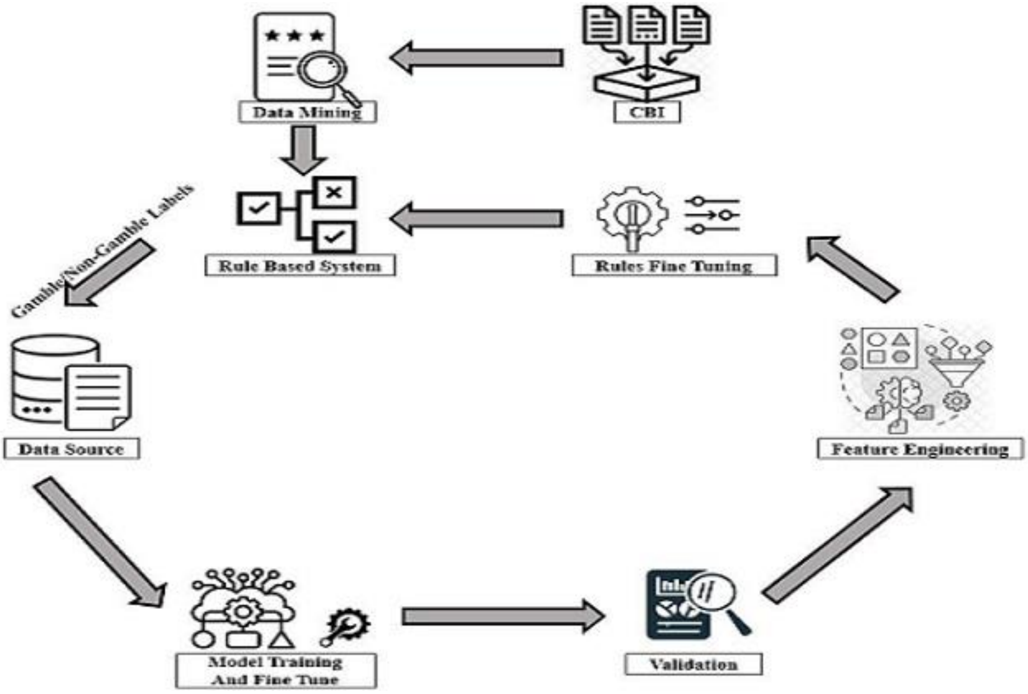

Fig 1. The schema of our proposed model

When it comes to tree-based classifiers, there are a few technical details that are worth mentioning. Firstly, decision trees are constructed by recursively partitioning the data into smaller and smaller subsets based on the values of certain features. Each split is chosen to maximize the information gain or reduction in entropy, to create branches that separate the data into distinct and homogeneous classes.

Another important aspect of tree-based classifiers is their ability to handle both categorical and continuous features. Categorical features are typically encoded as binary variables, with each possible category being represented by a separate column in the dataset. Continuous features are split using thresholds, which represent the points at which the feature is divided into two subsets.

Now, coming back to our evaluation results, we can use both feature importance and decision rules generated by our tree-based classifier along with machine learning models to update our rule-based model. By integrating these two approaches, we can leverage the strengths of both methods and improve the accuracy of our system in detecting gamblers.

For instance, we can use the feature importance scores to identify the most relevant features in the dataset, which can then be used to refine the feature set for our rule-based model. We can also examine the decision rules to understand how specific features interact with one another and use this information to create more nuanced rules that capture these complex interactions. Combining the insights gained from the tree-based classifier and rule-based models can help us create a more robust and accurate system for detecting gamblers.

5. Results and Analysis

Our results are divided into two sections. In the first section figures of feature importance based on each classifier are represented. Tree-based models such as random forests, lightGBM, and XGboost provide feature importance as part of their output. The importance of each feature is calculated based on how much it reduces the impurity in the tree.

There are different methods for calculating feature importance, including Gain, Split/Frequency/Weight, and Coverage.

Gain measures the relative contribution of a specific feature within a particular tree, indicating how much relevant information the model gains from that feature for making better predictions. Both XGBoost and LightGBM offer this method. We have used this method to obtain the importance of features.

Split (for LightGBM) and Frequency or Weight (for XGBoost) calculate the relative count of times a feature occurs in all splits of the model's trees. One limitation of this method is that it can be biased when there are many categories in categorical features.

Coverage is another method used to calculate feature importance, which measures the relative number of observations per feature. However, this method is only available in XGBoost.

The figures are generated using SHAP (SHapley Additive exPlanations) which is a library for model interpretability that explains the output of any machine learning model.

Fig. 2 to Fig. 4 show the feature importance of the three-ensemble tree-based methods, LightGBM, XGBoost, and Random Forest. These diagrams show the importance of each feature based on the results of each model. The x-axis represents the average impact on the size of the model output, while the y-axis shows the names of each feature in order of impact and importance.

The main features of these figures with no particular order are the most frequent number of similar transactions, the total amount of similar transactions, and the number of transactions done using the Internet. Although their order in different algorithms is different, they are the top three important features of classifiers. This fact implies that most gamblers use the Internet for money transfers. Moreover, there are predefined amounts for betting that could distinguish between these organizations and legal service providers.

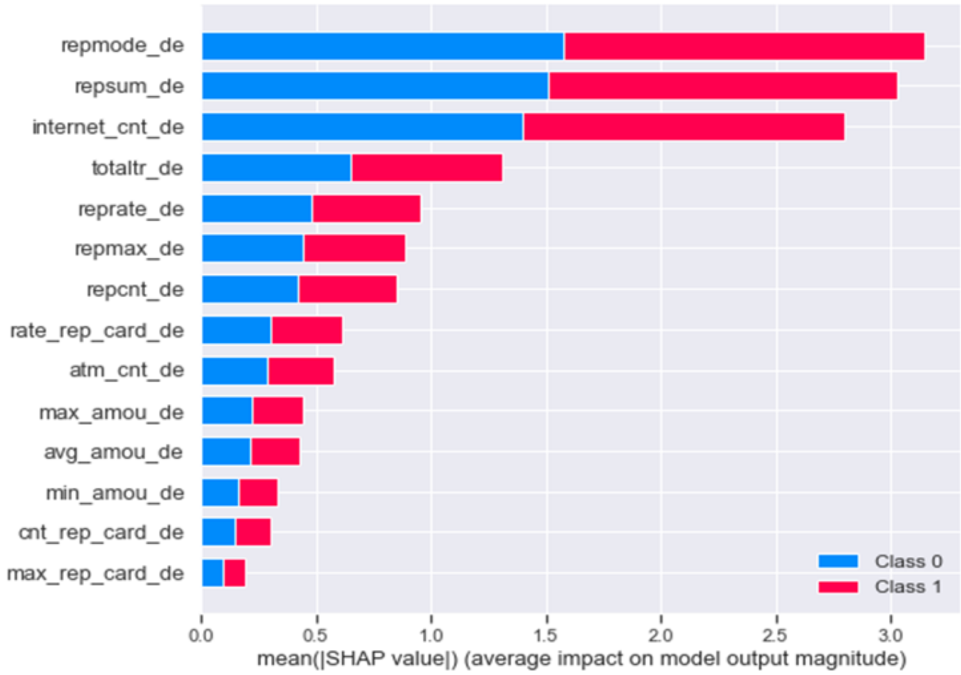

Fig 2. LightGBM Feature Importance

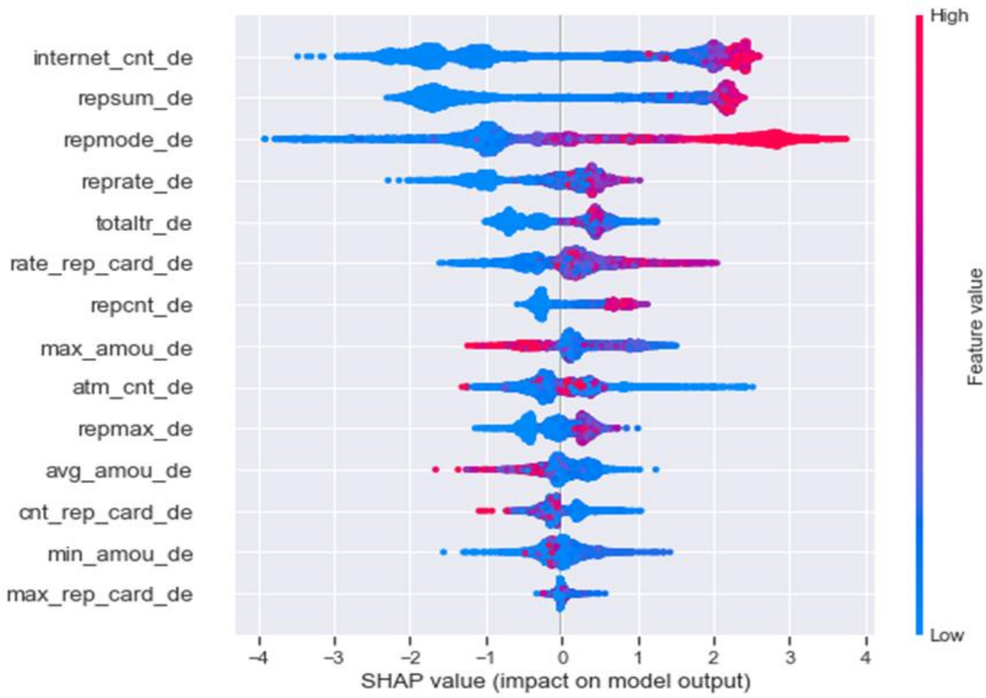

Fig 3. XGBoost feature importance

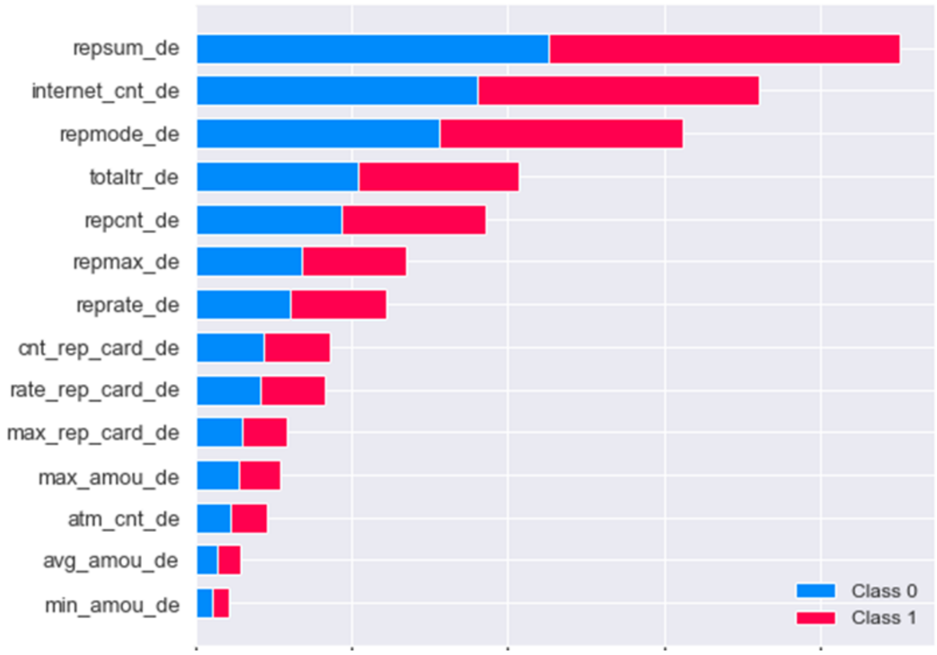

Fig 4. Random Forest feature importance

As previously stated, several metrics can be used to evaluate the performance of a model, including Precision, Recall, Accuracy, and F1 score. These measures provide valuable insights into how well the model is performing in terms of its predictive power. Once the training phase is complete, the model is then tested using a separate test dataset to assess its overall effectiveness. This step is critical in determining whether the model can generalize well to new data and perform consistently across different scenarios. By carefully analyzing the results of these tests, a better understanding of the strengths and weaknesses of models are gained and informed decisions about how to improve them are made.

Table 3. Test Data Set Using Four Measurements

Model	Precision	Recall	Accuracy	F1 measure
LightGBM	96.30%	97.02%	97.65%	96.65%
XGboost	96.82%	96.64%	97.72%	96.72%
Random Forrest	96.08%	95.9%	97.20%	95.99%

Table 3 offers a summary of three different classifier models' performance based on their performance metrics. Precision quantifies how many selected instances are relevant, while recall measures how many relevant instances were selected. F1 measure is the harmonic mean of precision and recall, and accuracy calculates the number of correct predictions over the total number of predictions. To describe the table accurately, it is important to remember that defining a good model based on these metrics is task-dependent. Here are some essential points to consider:

1. If all models have nearly identical precision, recall, F1 scores, and accuracy values, it may indicate that all of them are equally effective for the given problem.

2. If one model has substantially higher precision (at the expense of lower recall) than the other two models, it might be more suitable for tasks in which false positives are more expensive than false negatives.

3. Conversely, if one model has significantly higher recall (at the cost of lower precision) than the other two models, it might be more useful for tasks where false negatives are more detrimental than false positives.

4. If one model has higher accuracy than the other two models, it might suggest that it is better at correctly classifying a balanced dataset. However, if the dataset is imbalanced, accuracy alone may not be enough to evaluate model performance effectively.

By considering these factors and analyzing each model's performance metrics, one can gain insight into which model might be the most effective for your specific problem. In this particular case, all three models exhibit similar metric values. However, the second model performs slightly better in all four metrics.

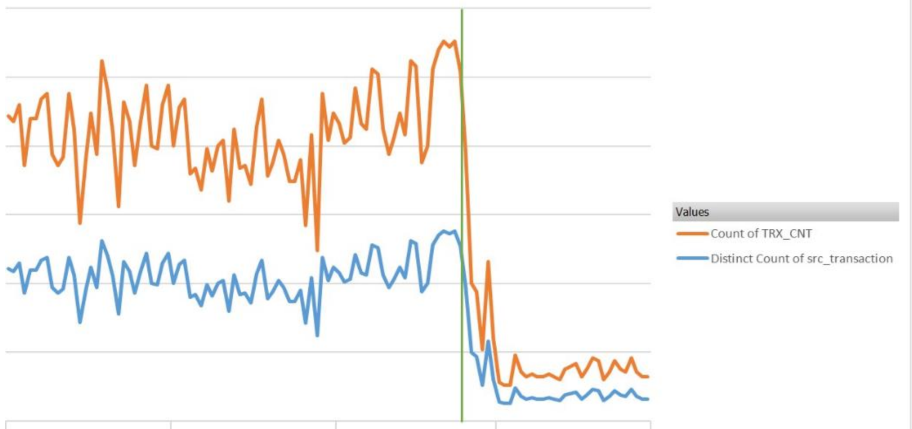

Fig 5. The number of gambling debit cards and merchant numbers before and after the application of the proposed model

6. CONCLUSION

In this paper, we presented a hybrid approach to detect gambling transactions using a combination of rule-based systems and machine learning classifiers. The goal was to address the limitations of rule-based systems in handling new types of fraud and enhance the effectiveness of fraud detection. A rule-based system was developed to identify debit cards involved in gambling transactions. This initial system helped label transactions as either gambling or non-gambling. However, rule-based systems have inherent drawbacks when it comes to adaptability to new types of fraud, including online gambling. Updating rules in such systems can be expensive and time-consuming, making them less effective in the long run. To overcome these limitations, machine learning classifiers were employed. The output of the rule-based system, which consisted of labeled gambling transactions, was utilized as training data for three tree-based classifiers: Random Forest, XGBoost, and LightGBM. Tree-based classifiers are known for their ability to handle complex data and capture non-linear relationships, making them suitable for fraud detection tasks.

By integrating an ensemble classifier with the rule-based system, the model became more robust in identifying gambling transactions. The classifiers not only provided accurate predictions but also allowed for rule extraction, enabling updates to the rule-based system as new patterns and types of gambling transactions emerged. This combination of techniques offered a comprehensive and adaptive approach to fraud detection.

The effectiveness of the developed model was evidenced by the significant decrease in the number of gambling transactions reported by the Central Bank of Iran. Our results shed new light on the power of combining rule-based systems and machine learning classifiers in addressing the challenges of fraud detection, specifically in the context of online gambling transaction detection. The approach presented here not only improved the accuracy, adaptability, and robustness of the detection model but also made the bank whose data were used in this study one of the safest banks in Iran for gambling transactions.

The present study has made a significant contribution to the field of banking by successfully training a well-performing model using real banking data, which has not yet been accomplished by any other bank in Iran. The novelty of this work lies in the fact that it is the first time such an achievement has been made in the country. The model's performance is a testament to the effectiveness of our approach, which can be attributed to the use of real banking data. Our findings have important implications for the banking industry, as they demonstrate that real banking data can be used to train models that outperform those trained on synthetic data. This is particularly relevant in Iran, where the use of real banking data has been limited due to privacy concerns and regulatory restrictions.

Future research must investigate the impact of time and memory on the model. For instance, deploying recurrent neural networks (RNNs) or time series analysis would be a viable option to enhance the results even further, given that gambling organizations' activities are not one-time events, and they will continue their activities in the same manner until they feel they are going to be visible. Additionally, combining graph models with neural networks is another area that remains unexplored in this field. This could be a promising approach to improve the accuracy of gambling detection models.

The present study is subject to certain limitations, including the absence of comprehensive information from all banks across the country. This lack of information restricts our ability to improve the detection of all gambling transactions. However, we have improved the precision of our detection model, which represents the proportion of detected correct gambling transactions. Furthermore, time complexity is another important factor to consider when developing a gambling detection model that is capable of detecting suspicious transactions in less than 150 milliseconds.

REFERENCES

Al Daoud, E. (2019). Comparison between XGBoost, LightGBM, and CatBoost using a home credit dataset. International Journal of Computer and Information Engineering, 13(1), 6-10.

Amarasinghe, T., Aponso, A., & Krishnarajah, N. (2018, May). Critical analysis of machine learning-based approaches for fraud detection in financial transactions. In Proceedings of the 2018 International Conference on Machine Learning Technologies (pp. 12-17).

Amoah-Nuamah, J., Agyemang-Duah, W., Mensah, B., & Opoku, B. (2023). University students' reasons and attitudes towards online gambling and its implication on their lives. Journal of Gambling Studies, 39(1), 203-224.

Badji, S., Black, N., & Johnston, D. W. (2023). Economic, Health and Behavioural Consequences of Greater Gambling Availability. Economic Modelling, 123, 106285.

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54, 1937-1967.

Bond, A. J., Widdop, P., & Myers, J. (2023). Gambling lifestyles: The importance of social capital and diverse networks. Leisure Studies, 1-15.

Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

Caiola, G., & Reiter, J. P. (2010). Random Forests for Generating Partially Synthetic, Categorical Data. Trans. Data Priv., 3(1), 27-42.

Charitou, C., Dragicevic, S., & Garcez, A. D. A. (2021). Synthetic data generation for fraud detection using gans. arXiv preprint arXiv:2109.12546.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM sigkdd International Conference on knowledge discovery and data mining (pp. 785-794).

Chóliz, M. (2023). Crisis, What Crisis? The Effect of Economic Crises on Spending on Online and Offline Gambling in Spain: Implications for Preventing Gambling Disorder. International Journal of Environmental Research and Public Health, 20(4), 2909.

Coelho, S. G., Keough, M. T., Hodgins, D. C., Shead, N. W., Parmar, P. K., & Kim, H. S. (2023). Loot box purchasing is associated with gambling and problem gambling when controlling for known psychological risk factors of gambling. Addiction Research & Theory, 31(4), 260-269.

Erickson, B. J., & Kitamura, F. (2021). Magician's Corner: 9. Performance metrics for machine learning models. Radiology: Artificial Intelligence, 3(3), e200126.

Floros, G. D. (2018). Gambling disorder in adolescents: prevalence, new developments, and treatment challenges. Adolescent health, medicine and therapeutics, 43-51.

Gainsbury, S. (2012). Internet gambling: Current research findings and implications.

Gainsbury, S. M., Abarbanel, B. L., Philander, K. S., & Butler, J. V. (2018). Strategies to customize responsible gambling messages: a review and focus group study. BMC Public Health, 18, 1-11.

Ge, D., Gu, J., Chang, S., & Cai, J. (2020, April). Credit card fraud detection using lightgbm model. In 2020 International Conference on E-commerce and Internet Technology (ECIT) (pp. 232-236). IEEE.

Gore, A. (2023). An assessment of AML risks linked to accepting crypto-payments in the gambling sector (A regulator's guide).

Hagfors, H., Vuorinen, I., Savolainen, I., & Oksanen, A. (2023). A longitudinal study of gambling motives, problem gambling, and need frustration. Addictive Behaviors, 144, 107733.

Håkansson, A. (2020). Impact of COVID-19 on online gambling–a general population survey during the pandemic. Frontiers in Psychology, 11, 568543.

Harris, S., Pockett, R. D., Dighton, G., Wood, K., Armour, C., Fossey, M., ... & Dymond, S. (2023). Social and economic costs of gambling problems and related harm among UK military veterans. BMJ Mil Health, 169(5), 413-418.

Hilas, C. S., & Sahalos, J. N. (2007). An application of decision trees for rule extraction towards telecommunications fraud detection. In Knowledge-Based Intelligent Information and Engineering Systems: 11th International Conference, KES 2007, XVII Italian Workshop on Neural Networks, Vietri sul Mare, Italy, September 12-14, 2007. Proceedings, Part II 11 (pp. 1112-1121). Springer Berlin Heidelberg.

Latvala, T., Lintonen, T., & Konu, A. (2019). Public health effects of gambling–debate on a conceptual model. BMC Public Health, 19, 1-16.

Marionneau, V. K., Lahtinen, A. E., & Nikkinen, J. T. (2023). Gambling among indebted individuals: an analysis of bank transaction data. European Journal of Public Health, ckad117.

Min, M., Lee, J. J., Park, H., & Lee, K. (2021). Detecting anomalous transactions via an IOT-based application: A machine learning approach for horse racing betting. Sensors, 21(6), 2039.

"Online Gambling Market Size & Trends Analysis Report, 2030." https://www.grandviewresearch.com/industry-analysis/online-gambling-market (accessed Jun. 10, 2023).

P., Manoj Arvind. (2023). Expositing the Legality of Gambling in India: A Relook through the Spectrum of Pound's Social Engineering. *Indian J. Integrated Rsch. L.*, *3*, 1.

Paterson, M., Taylor, M., & Gray, M. (2020). Trajectories of social and economic outcomes and problem gambling risk in Australia. Social Indicators Research, 148, 297-321.

Pfund, R. A., King, S. A., Forman, D. P., Zech, J. M., Ginley, M. K., Peter, S. C., ... & Whelan, J. P. (2023). Effects of cognitive behavioral techniques for gambling on recovery defined by gambling, psychological functioning, and quality of life: A systematic review and meta-analysis. Psychology of Addictive Behaviors.

Priscilla, C. V., & Prabha, D. P. (2020, August). Influence of optimizing XGBoost to handle class imbalance in credit card fraud detection. In 2020 third International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 1309-1315). IEEE.

Taha, A. A., & Malebary, S. J. (2020). An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine. IEEE Access, 8, 25579-25587.

Tomic, S. (2022). Regulatory Approach to Anti-Money Laundering in Online Gambling in the UK. In Financial Technology and the Law: Combating Financial Crime (pp. 47-65). Cham: Springer International Publishing.

Townsend, J. T. (1971). Theoretical analysis of an alphabetic confusion matrix. Perception & Psychophysics, 9, 40-50.

Williams, R. J., & Connolly, D. (2006). Does learning about the mathematics of gambling change gambling behavior? Psychology of Addictive Behaviors, 20(1), 62.

Xuan, S., Liu, G., Li, Z., Zheng, L., Wang, S., & Jiang, C. (2018, March). Random forest for credit card fraud detection. In 2018 IEEE 15th International Conference on Networking, sensing, and Control (ICNSC) (pp. 1-6). IEEE.