Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/52/20241601

# Model performance comparison and evaluation of bank fraud prevention in simulated datasets

**Zhilu Yang**

Aberdeen institute of data science and artificial intelligence, South China Normal University, Foshan, Guangdong, 528225, China

k24226897@163.com

**Abstract.** The significance of robust fraud detection systems in the banking sector has grown imperative due to the growing prevalence of online transactions. However, the datasets in these particular areas exhibit a greater abundance and diversity. The large amount and variety of data in these domains necessitate the utilization of synthetic sales data, which is derived from real data, as an innovative approach for studying fraud prevention. This study initially derives importance scores for various features through the utilization of random forests. Subsequently, four features that exhibit the highest correlation with fraudulent transactions are selected for further investigation. The training and prediction processes for both random forests and decision tree models are then performed. The study compared the performance of random forests and decision tree models in fraud monitoring using four features. The results indicate that random forests outperform decision trees in terms of accuracy, recall, precision, and F1 scores, with improvements of 0.68%, 0.62%, 0.68%, and 0.65% respectively. These findings provide a comprehensive analysis of the performance comparison between random forests and decision tree models in the context of fraud monitoring.

**Keywords:** random forest algorithm, decision tree algorithm, fraud detection, feature importance score

## 1. Introduction

Fraud has emerged as a significant concern for financial organizations in the contemporary digital era [1]. With the increasing availability of digital payment systems, fraudsters have developed more advanced and discreet techniques to engage in fraudulent activities [2]. Additionally, the growing diversity of data has posed a greater challenge in detecting and preventing fraud. With the increasing availability of digital payment services, fraudsters have developed more intricate and discreet techniques for engaging in fraudulent activities. Simultaneously, the expanding range of data has posed a greater difficulty in detecting instances of fraud. The problem of ensuring secure transactions in the digital era is of significant importance, as evidenced by the substantial financial losses up to hundreds of millions of dollars due to fraudulent activities annually [3].

In light of the intricate nature and delicate nature of fraud detection data in practical scenarios [4], this paper proposes the use of a fabricated sales dataset to facilitate pertinent research and experimentation. While the data used for fraud detection in this study may differ from real-world scenarios, it nevertheless offers useful insights that can be used as a reference for spotting fraud issues in practical settings. While the data used for fraud detection in this study may differ from real-world scenarios, it nevertheless offers useful insights that can be used as a reference for spotting fraud issues in practical settings. Simultaneously, the dataset lacks any information pertaining to actual individuals or entities, hence guaranteeing the absence of privacy concerns associated with its utilization for research purposes.

The objective of this study is to elucidate the progression of fraud detection methodologies, with particular emphasis on the transition towards data-driven and intelligent approaches. Additionally, this research aims to address the existing research void in the field of fraud prevention by employing synthetic sales data. The objective of this study is to elucidate the progression of fraud detection methodologies, with particular emphasis on the transition towards data-driven and intelligent approaches. Additionally, this research aims to address the existing research void in the field of fraud prevention by employing synthetic sales data.

## 2. Introduction of Decision Tree and Random Forest

### *2.1. Decision Tree*

The fundamental premise underlying decision trees is the creation of a hierarchical structure, known as a tree, comprising a root node, internal nodes, and leaf nodes. Each node inside the tree corresponds to a certain feature or attribute determination [5]. The decision procedures employed on the branches of these trees are utilized for the purpose of categorizing or predicting the target variable. In this context, each internal node corresponds to a feature test, while each leaf node reflects the outcome of a feature test. The decision methods employed on these tree branches serve the purpose of categorizing or predicting the target variable. In this context, each internal node corresponds to a feature test, while each leaf node corresponds to a category or numerical result. The creation of decision trees involves a recursive procedure aimed at reducing uncertainty through the careful selection of optimal features for the purpose of partitioning data.

The decision tree is a model that offers interpretability by providing a decision basis and a traceable path. Moreover, decision trees exhibit strong performance in handling extensive datasets and effectively executing feature selection and model generation tasks. Decision trees exhibit a notable attribute of being very resilient to interference, enabling them to efficiently handle noise and outliers within data to a certain degree. Consequently, this characteristic enhances the precision of fraud detection.

### *2.2. Random Forest Algorithm*

Random forests are a machine learning technique that adopts an integrated approach using decision trees [6]. This approach entails the creation of many decision trees on a dataset that is randomly sampled. Notably, during the construction of each tree, a subset of characteristics is randomly selected. This random selection process contributes to the diversity and robustness of the resulting forest. The utilization of a voting mechanism in Random Forest enables the amalgamation of prediction outcomes from individual trees, hence enhancing the model's resilience to outliers and noise.

Moreover, there are two key factors to take into account when applying random forests in the context of fraud detection. Firstly, it is worth noting that Random Forest exhibits several advantages in relation to its integration properties and random feature selection strategies, as highlighted in previous research [7]. Furthermore, the utilization of the voting process in random forests enhances its efficacy in identifying outliers and handling noisy data, hence augmenting the overall dependability of fraud detection.

## 3. Methodology

### *3.1. Data collection and pre-processing*

The dataset utilized in this study comprises synthetic sales data, encompassing several variables such as transaction ID, customer ID, merchant ID, transaction amount, transaction time, card type, transaction location, purchase category, customer age, and a binary indicator denoting fraudulent transactions.

In order to process the features, we employed a technique known as binned coding to categorize customer age. This involved dividing the age of each user into five distinct intervals: (0,18], (18,35], (35,50], (50,70], and (70,+∞). By assigning each customer's age to one of these predefined intervals, we were able to eliminate any extraneous variations that might have been present in the age data, hence enhancing the accuracy of our analysis. Furthermore, in order to facilitate the classification process, data such as card type, location, and purchase category are transformed into numerical representations using a technique known as class label coding.

### *3.2. Experimental setup*

*3.2.1. Feature engineering.* The pre-processed data was subjected to further processing, including feature engineering, in order to extract features specifically relevant to transaction fraud. This was accomplished using the Random Forest algorithm.

The following are the steps: Firstly, it is important to provide a clear definition of the feature matrix, denoted as X, and the target variable, referred to as y. The feature matrix comprises the set of features utilized for training the model, encompassing user identification, quantity, store identification, credit card type, location, purchase type, and age. The target variable denotes the category indicating whether a transaction is fraudulent or not. Subsequently, the feature matrix (X) and target variable (y) are inputted into the Random Forest classifier to derive the importance score associated with each feature.

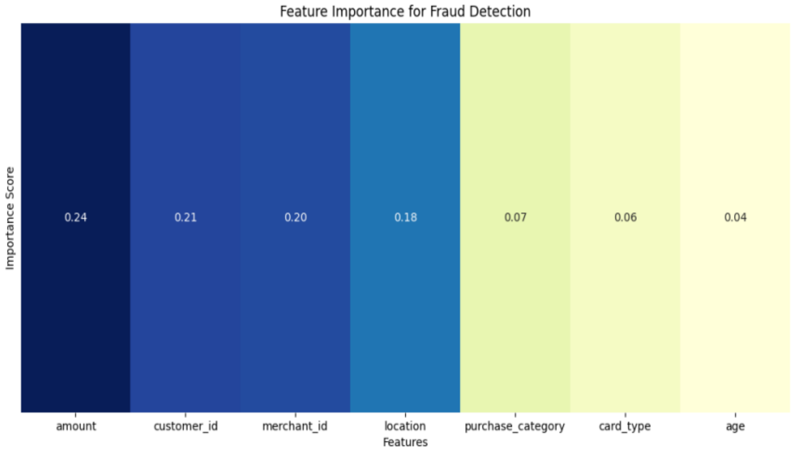

**Figure 1.** feature importance scores

Based on the findings presented in Figure 1, it is evident that the variables of transactions, users, merchants, and locations exhibit the strongest correlation with fraud, with respective correlation coefficients of 0.24, 0.21, 0.20, and 0.18. These results serve as crucial indicators for guiding subsequent model training and prediction processes.

*3.2.2. Model Training and Evaluation Explaining the Training Process of Random Forest and Decision Tree Models.* The process of model training involves several key components, namely the creation of classifiers, data segmentation, model training, and prediction. The feature matrix X and the target variable y are established by incorporating four key metrics pertaining to fraudulent transactions, namely customer ID, transaction amount, merchant ID, and transaction location. These metrics collectively form the feature matrix X, while the target variable y represents the classification of whether a transaction is fraudulent or not. The `train_test_split` function divides the training and test sets in a ratio of 3:1. Following the completion of model training and prediction, a detailed comparison and evaluation of the decision tree and random forest models is conducted. This evaluation is based on performance metrics such as Accuracy, Precision, Recall, and F1-score.

## 4. Results and discussion

The performance evaluation of the two models is presented visually through histograms and heatmaps.

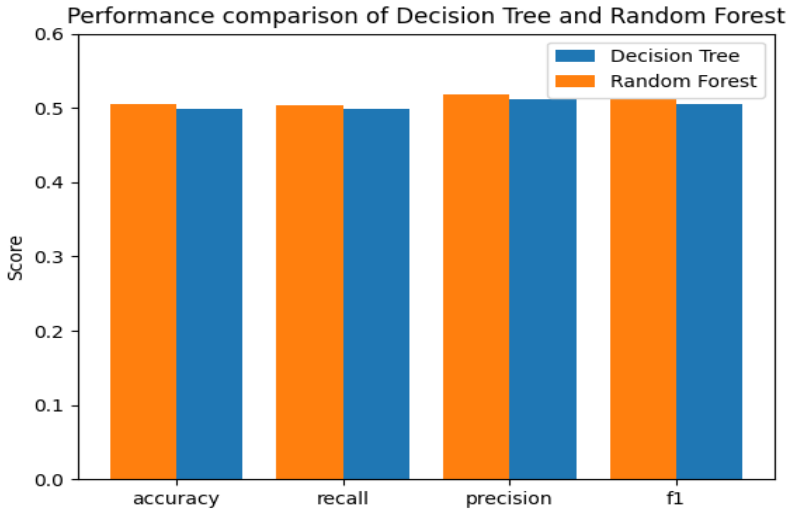

**Figure 2.** Histogram comparing the performance of decision trees and random forests

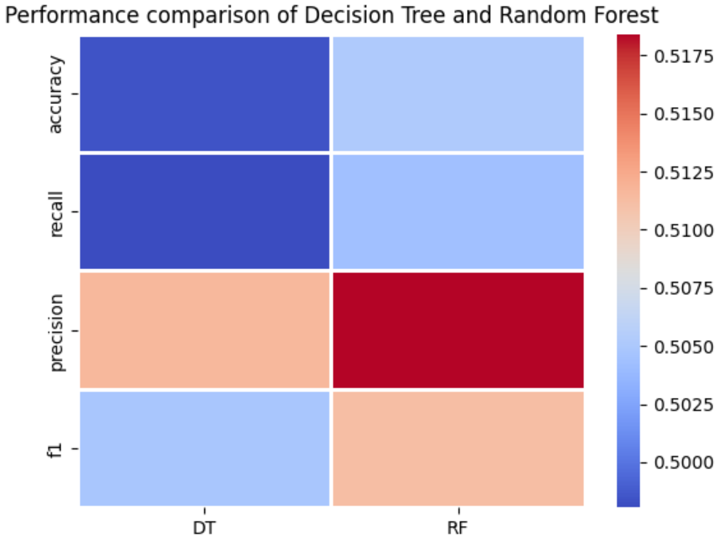

**Figure 3.** Heat map of decision tree and random forest hybrid matrix

The visual representation of the data in figure 2 and figure 3 reveals that there is no discernible difference in the effectiveness of Random Forest and Decision Tree algorithms in predicting fraudulent transactions. However, it can be observed that Random Forest outperforms Random Forest in terms of accuracy, recall rate, precision, and f1 score. In relation to accuracy, the decision tree model achieves a value of 0.4984, while the random forest model achieves a slightly higher value of 0.5052. In terms of recall, the decision tree model attains a value of 0.4980, whereas the random forest model achieves a slightly higher value of 0.5042. Regarding precision, the decision tree model achieves a value of 0.5116, while the random forest model achieves a slightly higher value of 0.5184. Lastly, in terms of the f1-score metric, the decision tree model obtains a value of 0.5047, whereas the random forest model achieves a slightly higher value of 0.5112. In terms of f1 scores, the random forest model has superior performance compared to the decision tree model, exhibiting improvements of 0.68%, 0.62%, 0.68%, and 0.65% in each respective case.

Upon conducting a comprehensive examination of the experimental findings, a thorough analysis elucidates the underlying factors that contribute to the superior performance of Random Forest in the realm of fraudulent transaction prediction, as compared to that of a solitary decision tree. The utilization of the integrated learning framework in random forests facilitates enhanced model stability by leveraging the collective outcomes of many decision trees. In the context of predicting fraudulent transactions, it is common for the dataset to exhibit intricate patterns and noise. This characteristic poses a challenge for a single decision tree model, as it becomes more susceptible to overfitting. The incorporation of randomization in the Random Forest algorithm serves to mitigate the potential issue of overfitting, hence enhancing its performance when confronted with novel, unseen data.

Furthermore, random forests possess a notable advantage in effectively handling datasets with a substantial number of characteristics. In the context of identifying fraudulent transactions, it is common to encounter a multitude of transaction features. However, relying solely on a single decision tree for feature selection during the splitting process may introduce bias towards certain features, leading to suboptimal model performance on other features. Random Forest enhances the model's capacity to comprehend intricate data relationships by including information from a greater number of features through the random selection of feature subsets for splitting during the training of each decision tree.

In general, the modest enhancements in accuracy, recall, precision, and F1 score observed in Random Forest can be attributed to its inherent traits of integrated learning and resilience to a high number of features. The aforementioned features enable Random Forest to effectively capture intricate patterns in fraudulent transactions, hence enhancing the overall performance of the model.

Furthermore, the incorporation of parallelization in Random Forest expedites the training procedure of the model to a certain degree [8]. On the other hand, the training procedure of decision trees is sequential, posing challenges in effectively leveraging computational resources, particularly when confronted with extensive datasets that may result in prolonged training durations.

## 5. Conclusion

The objective of this study is to examine the efficacy of Random Forest and Decision Tree models in the context of fraud detection, while also offering a comprehensive comparison analysis. Based on empirical investigations and rigorous statistical analysis, it has been shown that the random forest model demonstrates superior accuracy, recall, precision, and F1 score in the context of fraud detection. This finding underscores the substantial performance advantage of the random forest model in this particular task. The discovery holds significant practical implications within the financial sector, as even a marginal improvement in the precision and comprehensiveness of fraud detection can have a pivotal impact on mitigating losses in extensive financial transactions.

Although this study offers valuable insights into the field of fraud detection, it is important to acknowledge certain limitations that could potentially affect the outcomes of our research. Although the synthetic dataset used in our work serves as an experimental tool, it does not fully capture the intricacies of complex fraud behaviors observed in the real world. Furthermore, given the study's methodology and data collecting characteristics, the primary emphasis lies in the training and evaluation of offline models. It is important to acknowledge that the real-time performance of this approach for fraud detection tasks has not been thoroughly examined.

This work offers valuable insights into the comparative performance of random forest and decision tree models in the context of fraud detection. However, it is crucial to acknowledge that there are other potential avenues that warrant further investigation in future scholarly endeavors.

Potential avenues for future research could encompass an exploration of integration methodologies for diverse models, employing enhanced integration approaches and more resilient feature engineering to more effectively capture and depict patterns and relationships within transactional data. Furthermore, it is crucial to explore applications in the realm of fraud detection, since they play a significant role in facilitating prompt interventions to mitigate fraudulent activities.

This study offers an initial investigation into the field of fraud detection. However, future research can delve deeper into enhancing the performance of the model to effectively address evolving fraud behaviors and meet the specific requirements of various applications. By doing so, the practical effectiveness and applicability of fraud detection systems can be significantly enhanced.

## Acknowledgement

I would like to extend my sincere appreciation to Ishita Biswas, the creator of the dataset utilized in this research. The dataset offered useful insights and fundamental data for this study, which helped in establishing a robust groundwork for study's endeavor.

## References

[1] Zhang, Yue. "Research on Fraud and Cash-Out Risk Prevention of Credit Cards in the Internet Finance Background." Modern Finance, 2019(11): 35-36.

[2] Yang, Huo. "Research on Internet Financial Default Fraud Risk Events." Economic Research Reference, 2016(63): 28-34.

[3] Zhuo, Shangjin. "Internet Finance Fraud also Needs Strict Crackdown." Financial Times, 2014-06-13(004).

[4] Ding, Shuangsi. "Research on Internet Financial Fraud Behavior Identification Based on Big Data." Doctoral Dissertation, Capital University of Economics and Business, 2016.

[5] Yang, Xuebing, & Zhang, Jun. (2007). "Decision Tree Algorithm and Its Core Technologies." Computer Technology and Development, 17(1), 43-45.

[6] Biau, G., & Scornet, E. (2016). "A Random Forest Guided Tour." Test, 25, 197-227.

[7] Yao, Xu, Wang, Xiaodan, Zhang, Yuxi, & Quan, Wen. (2012). "Overview of Feature Selection Methods." Control and Decision, 27(2), 161-166.

[8] Gu, Xingbo, Wen, Qi, Shi, Xiaowen, & Liu, Yan. (2016). "Parallel Computing Methods and Applicable Conditions of Random Forest." Practical Preventive Medicine, 23(2), 129-132.