Poornima Mishra, 2024, 12:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

Credit Card Fraud Detection Using Decision Tree Algorithm

Poornima Mishra, Pranjal Dewangan, Sagar Dewangan, Sharmin Ansari, Assistant Professor Abhishek Kumar Dewangan

CSE department,
Shri Shankaracharya Technical Campus, Bhilai (C.G.), India

Abstract- Credit card fraud poses significant challenges to financial institutions, merchants, and consumers alike. As fraudulent activities continue to evolve in sophistication and frequency, the need for robust fraud detection mechanisms becomes imperative. In this study, we propose the utilization of the Decision Tree algorithm as an effective tool for detecting credit card fraud. Decision trees are widely recognized for their simplicity, interpretability, and ability to handle both numerical and categorical data effectively. Leveraging these advantages, we employ a Decision Tree model to analyze historical credit card transaction data, identifying patterns and anomalies indicative of fraudulent behavior.

Keywords- Business-making systems, Decision tree, Information Gain, Fraud detection, Credit Card.

I. INTRODUCTION

With the proliferation of electronic transactions, credit card fraud has become a pervasive threat, costing billions of dollars annually and eroding consumer trust in financial systems. Detecting fraudulent activities in real-time is paramount to mitigate financial losses and protect the interests of both consumers and financial institutions. In response to this challenge, advanced machine learning techniques have emerged as promising tools for fraud detection, offering the ability to analyze vast volumes of transaction data and identify anomalous patterns indicative of fraudulent behavior.

Among these techniques, Decision Tree algorithms have garnered considerable attention due to their simplicity, interpretability, and effectiveness in handling both numerical and categorical data. Decision Trees partition the feature space based on attribute values, iteratively splitting the data into subsets that are increasingly homogeneous with respect to the target variable—in this case, fraudulent or legitimate transactions. This process results in a hierarchical tree-like structure that facilitates intuitive decision- making and allows analysts to trace the logic behind each classification.

The utilization of Decision Tree algorithms for credit card fraud detection offers several advantages over traditional rule-based systems and statistical methods. Firstly, Decision Trees can handle non-linear relationships between features and the target variable, enabling them to capture complex patterns that may elude simpler models. Additionally, Decision Trees are robust to noisy data and can automatically handle missing values without requiring extensive preprocessing. Moreover, the interpretability of Decision Trees allows stakeholders to understand the reasoning behind each classification, facilitating transparency and trust in the model's decisions.

In this paper, we aim to explore the application of Decision Tree algorithms in credit card fraud detection. We will discuss the underlying principles

of Decision Trees, their implementation in fraud detection systems, and the challenges and opportunities associated with their deployment in real-world scenarios. Furthermore, we will conduct empirical experiments to evaluate the performance of Decision Tree models against alternative methodologies, demonstrating their efficacy in detecting fraudulent transactions while minimizing false positives and false negatives.

II. EXISTING SYSTEM

Since, credit card fraud detection (CCFD) systems are an advanced researched area, there are various algorithms and techniques for implementing these systems. One of the earliest systems is the CCFD system using the Markov's model. Various other existing algorithms used in credit cards such as Support vector machine fraud detection systems include cost-sensitive decision trees (CSDT), random forest and more.

CCFD is also proposed using neural networks. Existing systems using neural network get recognition value according to whale swarm optimization algorithm, it uses the Back propagation network for changing the value where the errors were detected. Studies showed the use of GA Feature Selection on Naïve Bayesian Random Forest and SVM for detecting fraudulent transactions.

The research study elaborated on Sequential Behavior Information Processing Using Deep Learning as well as the Markov Transition Field in Online Fraudulent Activities. A method named Attributed Sequence Embedding was displayed, in which various data sets are created using the process. All these techniques have significant drawbacks, such as reduced levels of accuracy, and inefficiency, sometimes categorized as buying regular transactions, and vice versa. The objective of this paper is to find out a new method for detecting fraud, increasing the accuracy and less complexity and time for results. The data set in this paper is based on actual transaction data of European Company, the privacy of which is treated as confidential.

III. PROPOSED SYSTEM

Decision tree technique is statistical data mining technique in which independent and dependent properties are logically expressed in the form of a tree, as illustrated in Fig. 1. The categorization rules derived from the decision tree are if then expressions, and to generate each rule, all tests must pass. Decision trees usually split a complex problem into many simple ones, and use iterations to solve the sub-problems. The tree is a predictive decision support tool that creates mappings of the possible outcomes from different observations. There are numerous prominent classifiers for generating class models from the decision trees. To improve precision and avoid overfitting, During the pruning step, such classifiers create a decision tree and afterwards clean up subtrees from the decision tree. This tree can be created by applying machine learning algorithms to the credit card database, and a multilayer pruned classifier (MLPC). The aim of the Decision

Tree model is to build a small decision tree with high precision. Based on credit card fraud detection, the decision tree comprises two stages. The initial step is to build a decision tree using the training data. The second step is to use decision rules to classify incoming transactions.

The input data of the decision tree are labelled with class labels, such as legitimate or fraudulent.

IV. METHODOLOGY

The system monitors each account individually using appropriate descriptors to identify transactions and flags as legitimate or legitimate. In the course of Decision Tree depicted in Fig. 2, all training examples begin with one node representing the tree dataset at the root node. Each node is split into child nodes in a method-specific binary or multipartition manner. The decision rules are read one by one from the decision table for each transaction that you classify as Match the transaction fields to each decision rule. It first finds an exact match and indicates the matched rule and transaction class of that class. If no match is found,

the highest risk among matching rules is selected and the transaction class is populated by the matched rules of the class. This indicates if a new transaction is a fraud of the same form; the node has been renamed the leaf and is flagged as fraudulent. This model was quick and adaptable. The MLPC approach is utilized as pre-pruning, which stops the tree growth at the pruning level specified before construction. It consists of a tree-top-down. Recursive partitioning and conquest methods. Initially all training examples are maintained on the route. The sample was then split recursively based on the chosen attributes. As the entropy metric, choose the split attribute. The necessary stages were repeated until any of the three conditions.

- All samples from a given node belong to the same class.
- There are no other properties for partitioning.
- No remaining samples were used.

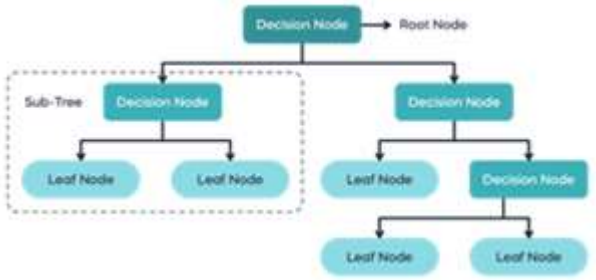

Figure 1: Decision Tree Architecture

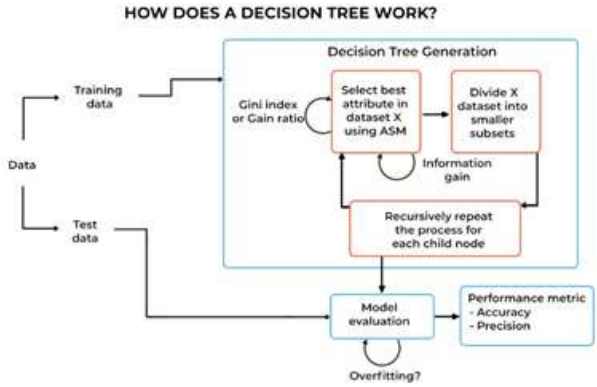

Figure 2: Decision Tree Flow Diagram

1. Decision Tree

Decision trees are a widely used machine-learning algorithm for both classification and regression tasks. They are represented as tree-like structures, with each internal node symbolizing a feature or attribute, each branch signifying a decision based on that feature, and each leaf node denoting the outcome or decision.

Here's an outline of how decision trees work:

Splitting

The tree is constructed by repeatedly dividing the dataset into subsets based on the value of a feature. The chosen feature is determined by criteria such as information gain, Gini impurity, or variance reduction.

Decision Making

At each internal node, a decision is made using the value of a specific feature. This decision dictates which branch to follow.

Leaf Nodes

When a leaf node is reached, a prediction or decision is made based on the majority class (for classification) or the average value (for regression) of the instances in that node.

Stopping Criteria

The process of splitting continues until one of the stopping criteria is fulfilled, such as reaching a maximum depth, a minimum number of samples in a node, or when further splits do not lead to significant improvement in predictive accuracy.

Decision trees have several advantages, including their interpretability, ability to handle both numerical and categorical data, and resistance to overfitting when properly adjusted. However, they can be prone to overfitting if they are not pruned or if the tree is too deep. To enhance the performance of decision trees, techniques such as pruning, ensemble methods (e.g., Random Forests), and boosting (e.g., Gradient Boosting Machines) are frequently employed.

2. Entropy

Entropy is a concept from information theory that is used in machine learning, particularly in decision tree algorithms. It measures the uncertainty or randomness in a dataset and is commonly used to determine the best attribute to split on at each

node. "Entropy Formula: In the context of a binary classification problem (i.e., two classes, such as "yes" and "no"), the entropy (S) of a set S is calculated using the formula:

H(S)= -p1 log2 (p1) – p2 log2 (p2)

where p1 and p2 are the proportions of examples in S belonging to each class."

Entropy is highest when the dataset is evenly split between classes, indicating maximum uncertainty, and lowest when all examples in the dataset belong to the same class, indicating no uncertainty. In building a decision tree, the goal is to find the attribute that best splits the dataset into subsets that are as pure as possible. Information gain is used to measure the effectiveness of a particular attribute in reducing entropy and is calculated as the difference between the entropy of the parent node and the weighted average of the entropies of the child nodes after the split. Decision tree algorithms, such as ID3, C4.5, and CART, use entropy (or related measures like Gini impurity) to determine the best attribute to split on at each node. The attribute that maximizes information gain is chosen for the split. The decision tree recursively splits the dataset based on the selected attributes until a stopping criterion is met. This process aims to create a tree that minimizes entropy and accurately predicts the target variable. By using entropy as a measure of uncertainty and information gain as a criterion for splitting, decision trees can effectively learn from data and make predictions in classification tasks.

3. Information Gain

"Information gain quantifies the effectiveness of an attribute in reducing uncertainty (entropy) in the dataset. It's calculated as the difference between the entropy of the parent node H(S) and the weighted sum of entropies of the child nodes after splitting on that attribute

IG(S,A) = H(S) - ∑v € Values(A) (| Sv | / | S |) H(Sv)

where :

IG(S,A) = is the information gain by splitting set S attribute A. Values(A) = represents the possible values of attribute A.

|Sv | = is the number of examples in set S for attribute A has value v.

|S| = is the total number of examples in set S."

V. RESULT AND ANALYSIS

The data that has been collected undergoes pre-processing before the modeling phase commences. As previously mentioned, the distribution of data with respect to the classes is highly imbalanced. The time frame that was used to construct our sample comprised 978 fraudulent records and 22 million normal ones, with a ratio of approximately 1:22500. Therefore, stratified sampling is employed to under sample the normal records, enabling the models to learn the characteristics of both the normal and fraudulent records' profiles. This is achieved by identifying the variables that are most effective in distinguishing between fraudulent and legitimate transactions, and then using these variables to create stratified samples of the legitimate records. Subsequently, these stratified samples of the legitimate records are combined with the fraudulent ones to form three samples with varying fraudulent to normal record ratios. The first sample set has a ratio of one fraudulent record to one normal record, the second one has a ratio of one fraudulent record to four normal ones, and the last one has a ratio of one fraudulent to nine normal ones. The factors that shape the card usage profile and the techniques applied in building the model significantly impact fraud detection systems. Our goal in identifying the variables used to create the data-mart is to differentiate the fraudulent card usage profile of criminals from the legitimate card usage profile of cardholders. We will only discuss the type of variables used, but due to privacy, confidentiality, and security concerns, we are unable to disclose the full list of variables. A variable refers to a specific level of deviation from an individual's average statistics. There are five main types of variables: all transaction statistics, regional statistics, sectorial statistics, daily amount

statistics, and daily number of transactions statistics.

Fraud detection systems must evolve dynamically to stay ahead. Future research could explore online learning techniques that enable models to adapt to changing fraud patterns in real-time, leveraging streaming data and incremental updates to improve detection performance. Despite the effectiveness of complex machine learning models, they often lack interpretability, limiting stakeholders' ability to understand and trust the decisions made by these systems. Future research could focus on developing techniques to enhance the interpretability and explainability of Decision Tree-based fraud detection models, enabling stakeholders to comprehend the rationale behind each decision and facilitating regulatory compliance.

Table 1: Input Parameters and Classifiers model

Classifier Model	Parameter
C&RT	Impure. Measure: Gini Max.Surrogates: 10 Tree_depth: 6
Classifier Model	Parameter
C&RT	Impure.Measure: Gini Max.Surrogates: 10 Tree_depth: 6

The process of creating a decision tree for classification problems involves two stages: first, using a training dataset to develop the decision tree; second, applying the decision tree to each element to identify elemental groups. The table below illustrates this concept with an example of ten articles about fraud and related information, represented as dataset S. In this instance, the ID3 algorithm is employed to generate a decision tree for credit card fraud classification.

1. Predection Model

Machine learning prediction models are algorithms that identify patterns in historical data and use them to make predictions or decisions on new, unseen data. These models are a fundamental element of numerous machine learning applications, ranging from basic linear regression models to intricate deep learning systems.

```
# Prepare your input data with the correct feature names
input_data = [[2, 9800, 170136, 160296]]  # Replace with your actual input data

# Make predictions
predictions = model.predict(input_data)
print(predictions)
```

```
['no fraud']
```

"Transaction Type" typically refers to the nature or category of a financial transaction. In the context of banking, finance, or e- commerce, transactions can be classified into different types based on their purpose or characteristics. The classification of transactions is essential as it helps in the identification of various types of transactions and the tracking of their sources and uses.

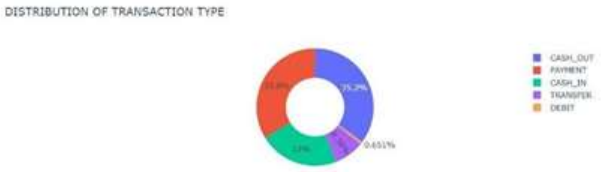

Figure 3: Distribution of Transaction Types

2. Analysis

Decision tree approaches tend to outperform support vector machine (SVM) approaches in solving the problem at hand. However, as the size of the training data sets grows larger, the accuracy performance of SVM-based models improves to match that of decision tree-based models. Despite this, the number of frauds caught by SVM models remains significantly lower than that of decision tree methods, particularly the C&RT model.

In this context, financial institutions can employ credit card fraud detection models to evaluate transaction information against historical profile patterns to predict the likelihood of fraud for a new transaction, thereby providing a scientific basis for authorization mechanisms.

Additionally, resources can be focused on more suspicious transactions to reduce overall fraud levels.

As future work, other data mining algorithms such as various versions of Artificial Neural Networks (ANN) and logistic regression may be employed to construct new classification models on the same real- world dataset. The performance of these new models will then be compared with those presented in this paper, considering additional performance metrics beyond prediction accuracy.

Figure 4: Kaggal Dataset and Output

V. FUTURE SCOPE

Ensemble techniques such as Random Forests and Gradient Boosting can enhance the accuracy and robustness of fraud detection systems. Future research could explore the integration of these techniques with Decision Trees to build more sophisticated systems capable of handling diverse and evolving fraud patterns.

The effectiveness of fraud detection models relies heavily on the quality and relevance of features used for analysis. Future studies could focus on developing novel feature engineering techniques tailored specifically for credit card fraud detection, as well as automated methods for feature selection to identify the most informative attributes.

Imbalanced datasets pose challenges for traditional machine learning algorithms, and future research could investigate advanced sampling techniques, cost-sensitive learning approaches, or algorithmic modifications to address class imbalance effectively and improve the detection of fraudulent activities. As fraudsters continuously adapt their tactics

V. CONCLUSION

We utilized Kaggle's credit card dataset to evaluate the effectiveness of various supervised machine-learning models in predicting fraudulent transactions. To determine the best model, we used accuracy, sensitivity, and time as criteria. Accuracy was not used because it is not sensitive to class imbalances and does not provide a clear answer. We examined KNN, Naive Bayes, decision tree, K-means, and Random Forest models. Our results showed that the Decision Tree classification model is the most suitable choice as it is both accurate and time-sensitive. Although the Random Forest model has slightly higher sensitivity than the Decision Tree model, we selected the Decision Tree model because the Random Forest model takes an excessively long time to process the data. Decision trees are recommended for negative detection because they provide fast predictions.

Acknowledgement

We appreciate the utilization of decision tree algorithms in the development of our credit card fraud detection system. Decision trees served as a fundamental component in our approach to identifying suspicious transactions and minimizing the risk of fraudulent activity. We extend our gratitude to the pioneers in machine learning and data science, whose research and contributions paved the way for the application of decision trees in fraud detection, enabling us to safeguard financial transactions and protect our customers' interests. We extend our gratitude to the reviewers for their valuable suggestions and modifications that have significantly enhanced the quality of our paper.

REFERENCES

1. V. Dheepa, R. Dhanapal, "Analysis of Credit Card Fraud Detection Methods", International Journal of Recent Trends in Engineering, Vol. 2, no. 3, pp 126 – 128, 2019.
2. S. Dzomira, "Fraud prevention and detection," Prevention, vol. 6, no. 14, 2015.
3. Y. Sahin, E. Duman, "Detecting Credit Card Fraud by Decision Trees and Support Vector

Machines", Proceeding of International Multi-Conference of Engineering and Computer Webology, Volume 18, Number 4, 2021

4. Y. Sahin, S. Bulkan, and E. Duman, "A cost-sensitive decision tree approach for fraud detection," Expert Systems with Applications, vol. 40, no. 15, pp. 5916–5923, 2013.
5. Snehal Patil, Harshada Somavanshi, Jyoti Gaikwad, Amruta Deshmane, Dept of Computer Engg JSPM'S BSIOTR Pune, India "Credit Card Fraud Detection Using Induction Algorithm".