Paper Title: A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions

Publisher: ELSEVIER

About:

- The study focuses on the use of data science and machine learning algorithms to detect and classify fraudulent credit card transactions.
- Three machine learning models are examined: logistic regression, random forest, and decision trees.
- The performance of these models is compared to determine their effectiveness in predicting and detecting fraud.
- The random forest model achieves the highest accuracy of 96% in detecting fraudulent credit card transactions.
- The model also has a high area under the curve (AUC) value of 98.9%, indicating its ability to accurately classify transactions.
- Based on these results, the study recommends using the random forest algorithm as the most suitable method for predicting and detecting fraud in credit card transactions.
- The analysis reveals that credit card holders above 60 years old are more susceptible to fraudulent transactions.
- The majority of fraudulent transactions occur during the hours between 22:00 GMT and 4:00 GMT.

Proposed model classification:

- Logistic regression is a machine learning technique used to predict binary outcomes, and it doesn't require normal distribution or correlation among explanatory variables. It's commonly used for detecting financial bankruptcies.
- Decision tree is a non-linear classification technique that splits a sample into smaller subgroups based on explanatory variables. It selects the variable with the strongest correlation to the outcome at each branch. Decision trees can be prone to overfitting but have various applications such as filtering spam emails or predicting vulnerability to viruses.
- Random forests are an extension of the decision tree method. They introduce additional randomness by using bootstrap samples and selecting random subsets of variables at each node. The average of the predictions from all trees is the final output. Random forests are used for analysing complex biological data, video segmentation, image classification, and more.
- Credit card fraud can be categorized into bankruptcy fraud, counterfeit fraud, application fraud, and behavioural fraud. Different machine learning algorithms like logistic regression, naive Bayes, random forest, K-nearest neighbours, gradient boosting, support vector machines, and neural networks have been used to detect fraudulent transactions in different jurisdictions.
- Hybrid models with AdaBoost and majority voting strategies have been developed to identify credit card fraud. Adding noise to the models can improve their performance, and voting systems have been found to be effective in the presence of noise.
- Two different types of random forest models were proposed to analyse the behavioral characteristics of typical and abnormal transactions. These models were tested using data from a Chinese e-commerce company to identify credit card fraud.
- Practical methods for detecting credit card fraud include employing various machine learning algorithms and using resampling techniques such as under-sampling and over-sampling. Random forest, XGBoost, and decision tree models have shown good performance in predicting fraudulent transactions.

- Machine learning algorithms can help detect and classify fraudulent transactions, potentially preventing them from being processed.
- The data used in the study consists of simulated credit card transactions from January 1, 2020, to December 31, 2020. The dataset includes legitimate and fraudulent transactions in the western side of the United States. Data preprocessing techniques such as cleaning, formatting, and handling missing values were applied.
- Feature scaling was performed to keep numeric variables within the same range, and under-sampling was used to handle imbalanced data. Synthetic Minority Oversampling Technique (SMOTE) was mentioned as an alternative method but was not used in this study.

A quick overview of these proposed supervised machine learning formulas for fraud detection is given in the corresponding subsection.

❖ Decision tree (DT):

- Decision trees are non-parametric supervised learning techniques used for classification. They create a tree-like structure based on actual data attributes.
- Decision trees have nodes, edges, and leaf nodes. The root node selects a feature to partition the data into subnodes, and this process is repeated until each training sample is categorized.
- Decision trees have advantages such as not requiring feature scaling, robustness to outliers, and automatic handling of missing values. They are quick to train and effective for classification and prediction.
- The decision tree algorithm uses metrics like the Gini index, information gain, and entropy to determine how to split the root node and classify the data.
- Entropy measures the expected randomness in the features used for splitting, while the Gini index measures the probability of incorrectly identifying a randomly chosen element.
- Information gain quantifies how effectively a variable separates the data into its categories. The goal of decision tree construction is to find the attribute with the highest information gain and lowest entropy.
- These metrics are used to calculate the reduction in uncertainty and gain in information when splitting the data based on specific attributes.

Decision trees are a powerful and interpretable classification method that can be used to make decisions based on data attributes, and different metrics are employed to determine the best splitting points.

❖ Logistic regression (LR):

- Logistic regression is a type of regression analysis used when the dependent variable is binary, such as fraud or not fraud.
- The logistic regression formula involves the natural logarithm of the odds ratio, which is represented as a linear combination of coefficients (α) and independent variables (x).
- The probability (p) that an observation belongs to a specific class (e.g., fraud or not fraud) is determined by the logistic regression model.

- The likelihood of an observation belonging to a class can be expressed as the odds ratio ($odds$), which is the ratio of the probability of the event occurring to the probability of it not occurring.
- Logistic regression uses the sigmoid function to describe the relationship between the response variable and predictor variables.
- The sigmoid function calculates the probability (p) that the response variable belongs to a specific class based on the linear combination of the intercept (α) and coefficients (β) multiplied by the predictor variables (x).
- The relationship between the probability (p) and predictor variables (x) is nonlinear, and the parameters α and β in logistic regression have different interpretations compared to linear regression.
- The logistic curve, represented in a graph, shows the probabilities limited to values between 0 and 1. It can be interpreted in terms of probabilities because of this limitation.

❖ Random forest:

- Random Forest is a supervised machine learning algorithm that uses a group of decision tree models for classification and predictions.
- Each decision tree is a weak learner, but when combined as an ensemble, they improve the accuracy of the model.
- Random Forest employs the bagging method to generate a forest of decision trees. Random vectors are created, and each vector is converted into a decision tree.
- The final classification result is determined by selecting the class that has the majority of votes from all decision trees.
- Random Forest is typically robust and does not require a feature selection process.
- However, it may quickly identify data with a wide range of values and variables with numerous values as fraudulent, which can be a drawback.
- Random Forest is considered one of the most accurate fraud detection algorithms in the financial sector.
- Model performance is evaluated using metrics such as accuracy, precision, recall, specificity, F1-score, and area under the curve (AUC).
- The confusion matrix is used to calculate metrics such as true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).
- Accuracy measures the ratio of correct predictions to the total number of predictions.
- Precision measures the ratio of correctly classified fraud transactions to the total predicted fraud transactions.
- Recall (or sensitivity) measures the ratio of correctly classified fraud transactions to the total number of actual fraud transactions.
- Specificity measures the ratio of correctly classified non-fraud transactions to the total number of non-fraud transactions.
- The F1 score is the weighted mean of precision and recall, providing a balance between the two metrics.
- AUC and the receiver operating characteristic (ROC) curve measure the performance of the model at different threshold values, with values closer to the top-left corner indicating better performance.
- A random classifier would produce points along the diagonal line in the ROC space.

Conclusion:

In this study, three classification models (Logistic Regression, Decision Tree, and Random Forest) were built to categorize online credit card transactions as fraudulent or not. The dataset was balanced using under-sampling to avoid bias towards the majority class. The Random Forest model outperformed the other models, with an AUC value of 98.9% and an accuracy of 96.0%, making it the most suitable for predicting fraud. The analysis revealed that most fraud cases occurred between 10 pm and 5 am, when monitoring might be limited and victims are more likely to be unaware. Cardholders over 60 years old were found to be frequent targets of fraud. Recommendations include prioritizing in-person services for older clients, enhancing security measures during vulnerable hours, and developing robust fraud prevention systems. Future studies can explore other machine learning algorithms and apply the findings to other sectors such as healthcare.