Group No. 47

ML CASE STUDY

INTRODUCTION:

Credit card fraud is a significant concern for financial institutions, businesses, and consumers. Detecting fraudulent transactions accurately and swiftly is essential to minimize financial losses and protect customers. Machine learning techniques, particularly decision trees, have proven effective in identifying fraudulent activities.

The primary objectives of this case study are as follows:

a. Develop a Credit Card Fraud Detection System using decision trees.
b. Implement Gini impurity and entropy as impurity measures for decision tree nodes.
c. Employ repeated k-fold cross-validation to evaluate and fine-tune the model's performance.

We will be using Kaggle Dataset: Credit Card Fraud Detection. This datasetpresents transactions that occurred in two days, with 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all trades. It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, … V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependent cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

ABSTRACT:

Credit card fraud is a pervasive problem that imposes significant financial losses on both financial institutions and cardholders. Machine learning models have emerged as a powerful tool to combat this issue by detecting fraudulent transactions in real-time. This abstract provides an overview of a credit card fraud detection machine learning model. In this study, we present a comprehensive approach to building an effective credit card fraud detection system using machine learning techniques.

REVIEW OF EXISTING MODEL:

Decision Tree Model:

The core of the fraud detection system will be built using decision tree algorithms, particularly the CART (Classification and Regression Trees) algorithm. Decision trees are apt for this task because they can handle both numerical and categorical data effectively. In this context, two impurity measures will be implemented: Gini impurity and entropy.

The Gini impurity measure quantifies the probability of misclassifying a randomly chosen element from the dataset. It guides the decision tree's node-splitting process by minimizing the Gini impurity at each step. Alternatively, entropy measures the information gain at each node. It aims to maximize the information gain by selecting attributes that lead to the most distinct separation between classes.

Weakness:

1. Overfitting: Decision trees have a tendency to overfit the training data, especially when they are deep and complex. This means that they can capture noise in the data and generalize poorly to new, unseen data. Techniques like pruning and setting a maximum depth can help mitigate this issue.
2. High Variance: Decision trees are sensitive to small variations in the training data, leading to different tree structures with slight changes in the dataset. This makes them unstable and can result in high variance.
3. Lack of Smoothness: Decision trees create step-like, piecewise constant predictions, which may not be suitable for problems that require smooth, continuous functions.
4. Biased towards Dominant Classes: In classification tasks with imbalanced datasets, decision trees tend to favour the majority class, as they partition the data to maximize purity. This can lead to poor performance on minority classes unless addressed through techniques like class weighting or resampling.

Repeated k-Fold Cross-Validation:

To ensure the robustness and reliability of the model, repeated k-fold cross-validation will be employed. This technique involves repeatedly splitting the dataset into k subsets (folds) and training/testing the model on different combinations of these folds. Repeated k-fold cross-validation helps mitigate the risk of overfitting and provides a more accurate estimation of the model's generalization performance.

Weakness:

1. K-fold cross-validation requires training and evaluating the model k times (where k is the number of folds). For large datasets or complex models, this can be computationally expensive and time-consuming. The performance metrics (e.g., accuracy, F1-score) obtained from different folds can vary. This variance in results can make it challenging to confidently assess model performance, especially when the dataset is small.
2. In some cases, there may be subtle data leakage issues. For instance, if preprocessing steps (e.g., feature scaling) are performed before cross-validation, information from the test set can unintentionally influence the training data.
3. For time-series data, traditional K-fold cross-validation may not preserve the temporal order of data points. This can be a problem when the order of events matters, as it often does in time-series forecasting or anomaly detection tasks.

Results and Discussion:

The study will present the results of the Credit Card Fraud Detection System, including performance metrics, confusion matrices, and ROC curves. It will also discuss the strengths and weaknesses of using Gini impurity and entropy as impurity measures in decision trees, analyzing their respective contributions to the model's performance. Furthermore, the importance of different features in fraud detection will be examined, shedding light on which aspects of transactions are most indicative of fraud. The benefits of repeated k-fold cross-validation in ensuring the model's robustness will be highlighted.

CONCLUSION:

In conclusion, this study aims to provide a comprehensive approach to Credit Card Fraud Detection using decision trees, Gini impurity, entropy impurity measures, and repeated k-fold cross-validation. The results and insights obtained from this research will contribute to the development of more effective fraud detectionsystems, with potential future improvements and extensions to enhance the accuracy and reliability of such models.

PLAGIARISM REPORT: