PHASE -1 SUBMISSION

TEAM NO 11

CREDIT CARD FRAUD DETECTION

Problem Definition

The problem is to develop a machine learning-based system for real-time credit card fraud detection. The goal is to create a solution that can accurately identify fraudulent transactions while minimizing false positives. This project involves data pre-processing, feature engineering, model selection, training, and evaluation to create a robust fraud detection system.

Design Thinking

Data Sourcing:

Data sourcing is a critical step in credit card fraud detection. It involves gathering historical transaction data from various sources, such as financial institutions, payment processors, and e-commerce platforms. This data typically includes information on transaction amounts, timestamps, merchant details, and customer profiles. Additionally, dataset provided in the website of Kaggle is taken in hand to carry forward with the processes.

Dataset link: [**https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud**](https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud)

Data Pre-processing:

Data pre-processing is a pivotal phase in credit card fraud detection that focuses on cleaning and preparing the gathered data for analysis. This step involves tasks like handling missing values, removing duplicates, and standardizing data formats. Given the imbalanced nature of fraud detection datasets, where legitimate transactions far outnumber fraudulent ones, techniques like oversampling or undersampling may be employed to balance the classes. Furthermore, feature scaling and normalization help ensure that features contribute equally to the model, preventing any bias. Data pre-processing ensures that the dataset is in a suitable form for subsequent analysis and model development.

Feature Engineering:

Feature engineering is a creative and crucial step in credit card fraud detection. It involves selecting, creating, or transforming features to extract meaningful information from the raw data. Features like transaction frequency, spending patterns, and geographical information can be engineered to improve the model's ability to discriminate between legitimate and fraudulent transactions. Additionally, dimensionality reduction techniques, such as principal component analysis (PCA), can be applied to reduce the computational complexity of the model while preserving essential information. Effective feature engineering plays a vital role in enhancing the model's predictive power.

Model Selection:

Model selection is a pivotal step in credit card fraud detection that involves choosing the most appropriate machine learning algorithms for the task. Commonly used models include logistic regression, decision trees, random forests, support vector machines (SVMs), and neural networks. The choice of model depends on factors like the dataset size, complexity, and interpretability requirements. Ensemble methods, which combine multiple models, are often favored for their ability to capture complex patterns in the data. Model selection is a critical decision as it directly impacts the accuracy and efficiency of fraud detection.

Model Training:

Model training is the process of teaching the selected machine learning model to recognize patterns and make predictions based on the prepared dataset. During training, the model learns to differentiate between legitimate and fraudulent transactions by adjusting its internal parameters. It's essential to split the dataset into training and validation sets to evaluate the model's performance and prevent overfitting. Hyperparameter tuning and cross-validation are used to optimize model performance. The model is trained iteratively until it achieves the desired level of accuracy and generalization.

Model Evaluation:

Model evaluation is the final stage in credit card fraud detection, where the performance of the trained model is assessed using various metrics like precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC). Precision measures the model's ability to correctly identify fraud cases, while recall quantifies its ability to find all fraudulent transactions. The F1-score combines these metrics to provide a balanced assessment of model performance. AUC-ROC evaluates the model's overall ability to discriminate between fraud and non-fraud cases. Regular model re-evaluation is crucial to ensure continued effectiveness in detecting evolving fraud patterns, and adjustments