CREDIT CARD FRAUD DETECTION

UNDER GUIDANCE OF
MR. GOKUL YENDURI

TEAM MEMBERS

21MIC7171- MOHAMMAD NUZHAT KULSUM

21MIC7191- SHAIK AFIFA ALIYA

21MIC7147- GUNDA KARTHEEK

SUMMER INTERNSHIP

5- YEAR INTEGRATED MTECH

SOFTWARE ENGINEERING

21ST BATCH

DEPARTMENT OF SCOPE

Title of the Project	Credit Card Fraud Detection
Software Requirements	**Operating System:** Windows XP or later **Tools**: Google Colab, Kaggle, Visual studio code. **Languages Used :** Python
Hardware Requirements	**Processor:** Intel i5 or later RAM: 512GB RAM **Hard Disk:** PC with 20GB or more

AIM:

The aim of this project is to develop a reliable credit card fraud detection system using machine learning model and XAI for identifying credit card fraud. This involves utilizing huge data and various data preprocessing, feature and model evaluation techniques to effectively identify fraudulent transactions**.**

GOAL OF OUTCOME:

The goal is to create a fraud detection system that not only accurately identifies fraudulent transactions in real-time but also provides clear explanations for each decision, helping to build trust and improve decision-making in financial markets.

Abstract

As the internet becomes more accessible, businesses are expanding their online offerings. Additionally, since e-commerce websites have grown in popularity, people and companies in the financial industry are increasingly depending on internet marketing managers to run their operations. The number of credit card fraud cases has risen due to the growing popularity of online banking and

shopping. The systematic operation of the current fraud detection system might be interrupted by fraudsters using any means possible to overcome this problem using Machine learning algorithms and statistical methods have been used in credit card fraud detection to work past these restrictions. These methods are predicated on the analysis of important variables, like the customer's transaction history and account information, in addition to transaction-related data, like the transaction amount, location, and time by identifying and filtering fraudulent activity in real time, this research aims to reduce fraud manipulation by creating an effective fraud detection system through the use machine learning techniques that adapt to changing patterns of customer behaviour. The techniques include Logistic Regression Explainable Artificial Intelligence (XAI). These models have demonstrated positive outcomes in identifying fraudulent transactions by learning data patterns and improving fraud detection efficiency. Overall, credit card fraud detection is an important field of research in the financial industry, with an opportunity for improving fraud detection rates and minimize financial losses.

THE MODULES THAT ARE INVOLVED:

1. Data Collection
2. Data Preprocessing
3. Model selection and Training
4. Logistic regression
5. XG boost
6. Decision tree
7. Random forest
8. SHAP
9. Model Evaluation and Validation
10. Evaluation Metrics

DATA COLLECTION :

Data collection is a crucial step in any research process, as it involves gathering information that will be used to answer research questions, test hypotheses, and evaluate outcomes. There are several methods of data collection, each with its own advantages and disadvantages.

DATA PREPROCESSING :

Data preprocessing involves cleaning and transforming raw data to prepare it for analysis. This includes handling missing values, normalizing data, encoding categorical variables, and feature extraction. The goal is to improve data quality and ensure it is suitable for training machine learning models, ultimately enhancing model performance and accuracy.

MODEL SELECTION AND TRAINING :

Model selection and training are fundamental steps in the development of machine learning systems. These steps involve choosing the appropriate model for your task, training the model on your data, and evaluating its performance.

LOGISTIC REGRESSION :

Logistic Regression is a statistical method used for binary classification problems, where the outcome is a binary variable it has two possible outcomes. Despite its name, logistic regression is actually a linear model for classification rather than regression.

XG BOOST :

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. It implements machine learning algorithms under the Gradient Boosting framework.

SHAP :

SHAP is a game-theoretic approach to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from cooperative game theory and their related extensions.

DECISION TREE :

A decision tree is a supervised learning algorithm used for both classification and regression tasks. It is a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility.

RANDOM FOREST :

Random Forest is an ensemble learning method used for classification, regression, and other tasks. It operates by constructing multiple decision trees during training and outputting the mode of the classes (classification) or mean prediction (regression) of the individual trees.

MODEL EVALUATION AND VALIDATION :

Model evaluation and validation are essential steps in the machine learning workflow to ensure that a model performs well on unseen data and is not overfitting or underfitting.

EVALUATION METRICS :

Evaluation metrics are used to assess the performance of machine learning models. The choice of metric depends on the type of problem classification, regression and the specific requirements of the application.