# Project Proposal: Credit Card Fraud Detection Using Machine Learning

Introduction Credit card fraud is a significant issue that causes substantial financial losses globally. This project aims to develop a machine learning model to detect fraudulent transactions effectively and in real-time.

**Problem Statement**

* The rise in online transactions has led to an increase in credit card fraud.
* Existing methods are either too slow or not accurate enough to detect fraud.

**Objectives**

* Develop a machine learning model to identify fraudulent transactions with high accuracy.
* Implement the model to process transactions in real-time.

**Methodology**

* Data Collection: Gather historical transaction data, including both fraudulent and legitimate transactions.  
  data source: https://www.kaggle.com/code/gpreda/credit-card-fraud-detection-predictive-models/input
* Data Preprocessing: Clean and preprocess the data for analysis, including handling missing values and feature engineering.
* Model Selection: Implement multiple machine learning algorithms to detect credit card fraud.
* Supervised Learning Algorithms: These algorithms learn from labeled data, making predictions based on input-output pairs.
  + Decision tree
  + Random Forest
  + Regression
* Unsupervised Learning Algorithms: These algorithms explore unlabeled data to find hidden patterns. Techniques include clustering (like K-mean) and dimensionality reduction (like PCA).
  + K-means
  + Principal Component Analysis(PCA)
* Deep Learning Models: These are advanced neural networks with multiple layers. They mimic the human brain to identify patterns in large datasets.
  + Convolution Neural Networks
  + Autoencoders
* Model Training: Train the chosen model(s) using the preprocessed data.
* Model Evaluation: Evaluate the model's performance using metrics such as accuracy, precision, recall, and F1-score.
* Comparison of all Models

**Dataset Definition:**

The **Credit Card Fraud Detection** dataset from Kaggle typically includes transaction data that can be analyzed to identify fraudulent patterns. The dataset may contain the following key columns or types of data:

Data source: https://www.kaggle.com/code/gpreda/credit-card-fraud-detection-predictive-models/input

* **Time**: Elapsed time since the first transaction in the dataset. It is useful for understanding the temporal patterns of transactions.
* **V1 to V28**: These columns represent the result of a PCA (Principal Component Analysis) transformation applied to anonymize the original data. They are crucial for feature engineering and anomaly detection.
* **Amount**: The amount of the transaction, which can be a key indicator of fraudulent behavior when analyzed in conjunction with other variables.
* **Class**: The target variable, where 1 indicates a fraudulent transaction and 0 indicates a legitimate one.
* **Transaction ID**: A unique identifier for each transaction, though this is not typically useful for analysis.
* **Other Derived Features**: You may consider adding derived features, such as transaction frequency, running totals for a customer, or average transaction amounts.

**Anticipated Challenges:**

* **Imbalanced Dataset**: The number of fraudulent transactions is typically far lower than the number of legitimate ones, creating a severe class imbalance. This makes it challenging to train models effectively since most classification algorithms are biased towards the majority class.

With imbalance dataset , there is a risk of overfitting.

* **Feature Engineering**: Since the dataset has anonymized features (V1 to V28), it’s difficult to intuitively understand their meaning and how it will impact fraud detection.
* **Computational Resources:** Running complex models on large datasets, especially with techniques like cross-validation and hyperparameter tuning, can be computationally intensive.

**Expected Outcomes:**

* A machine learning model capable of accurately identifying fraudulent transactions.
* Reduction in the number of fraudulent transactions processed.
* Improved financial security for users.

**Conclusion:**

 This project will leverage machine learning to address the critical issue of credit card fraud, providing a robust and scalable solution to enhance financial security.