ML Project Preparation

Hongru He [ML Project Group 2]

01/25/2026

CPSC5310 26WQ

1. What is the project about?

This project aims to build a **machine learning model for credit card fraud detection**.

The goal is to classify each transaction as either **legitimate** or **fraudulent** based on transaction-related features.

This is a **binary classification problem** with a strong real-world constraint: fraud cases are **extremely rare**, which makes accurate detection challenging.

Example problem:

- Given a credit card transaction, predict whether it is fraudulent (Class = 1) or legitimate (Class = 0).
- Minimize false negatives (missed fraud) while controlling false positives (incorrectly flagged transactions).

Success will be measured not only by accuracy, but by metrics such as precision, recall, F1-score, and ROC-AUC, with particular emphasis on recall for the fraud class due to the imbalance in the dataset.

2. What data features might be used?

The dataset contains **284,807 transactions** and **31 columns**, all numeric.

Input Features

- **V1–V28**
  - PCA-transformed features derived from original transaction attributes
  - Continuous numerical variables
  - Linearly uncorrelated due to PCA, but not necessarily independent
- **Time**
  - Seconds elapsed since the first transaction in the dataset
  - Represents relative transaction timing
- **Amount**
  - Transaction amount
  - Continuous numerical feature
  - Requires scaling or transformation due to skewness

Among all the columns above, we should find the most related features via methods like calculating the correlation with the target variable.

Target Variable

- **Class**
    - 0: Legitimate transaction
    - 1: Fraudulent transaction

This dataset has **severe class imbalance**, with fraud accounting for approximately **0.17%** of all transactions.

3. What would be your first step?

The first step would be **exploratory data analysis (EDA)** and problem clarification, including:

- Understanding feature distributions and scales
- Examining class imbalance
- Comparing fraud vs non-fraud patterns
- Identifying appropriate evaluation metrics (e.g., precision, recall, F1-score, PR-AUC instead of accuracy)

At the same time, we would define a **baseline model** (e.g., logistic regression with class weights) to establish a reference point for later improvements.

This step ensures that modeling decisions are informed by the data's characteristics and real-world constraints.