Project proposal

Team member: Duyen Su, Mugdha Dixit, Pragya Verma, Yeeyin Lim, Priyanka Patil

Credit fraud detection

1. Description of data source and weblink:

The data is coming from Kaggle website: <https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud> with the license <https://opendatacommons.org/licenses/dbcl/1-0/>

This dataset includes transactions made by credit cards in September 2023 by European card holders.

1. Number of records and number of attributes with description of each attribute

The dataset contains 284,807 records (transactions) with total of 31 features(columns).

The features are ‘Time’, V1 to V28, ‘Amount’, and ‘Class’. Feature ‘Time’ contains the second elapsed between each transaction and the first transaction in the dataset. Feature ‘Amount’ is the transaction amount. Feature ‘Class’ classify whether the transaction is fraud (1 means fraud). Features V1, V2,..,V27,V28 are numerical value and are principle component analysis (PCA). The information background of V1 to V29 are hidden due to the confidentiality issue.

1. Some general statistics of the data set (we also attached the python analysis file for more information)

A screenshot of a graph

Description automatically generated

This transaction dataset includes 284,807 records (transactions) that occurred in 2 days, with 492 frauds. This fraud accounts for almost 0.2% of the data and this makes the data is unbalanced.

Sum of null value in the dataset:

A number and numbers on a white background

Description automatically generated

1. Tools/Methods:

We will use Logistic Regression method with Python Sklearn

1. Detail description of what problems/questions:

The question is that we will try to find the model to predict the “Class” whether or not the transaction is fraud.