Inferential Statistics for Fraud Detection

This project deals with the problem of building a machine learning model for fraud detection – identifying the fraudulent transactions from a set of credit card transactions, and hence the clients are financial institutions. The dataset used in this project is published by Kaggle, which contains credit card transactions from September 2013, made by European cardholders. This dataset presents transactions that occurred in two days, where there are 492 frauds out of 284,807 transactions. Therefore, the dataset is highly imbalanced as the positive class (frauds) account for 0.172% of all transactions.

It contains only numerical input variables, which are the result of a Principal Component Analysis (PCA) transformation. Unfortunately, due to confidentiality issues, the original features and the detailed background information about the data are not provided. Coded features V1 to V28 are the principal components obtained after applying PCA to the raw dataset, and variable "Amount" represent the transaction amounts. Feature labeled as "Class" is the response variable and it takes value 1 in case of fraud, and 0 otherwise. There are no missing values in the dataset.

The summary statistics for variable "Amount" are in Table 1.

Class	Count	Mean	Std	Min	Max
Fraud	492	122.2	256.7	0.0	2125.9
Non-fraud	284,315	88.3	250.1	0.0	25691.2

Table 1: Summary Statistics for Amount

The t-test for the transaction amounts of the frauds and non-frauds yields the t-statistic is 2.93 and the corresponding p-value 0.0036. Since the p-value is less than 0.05, the null hypothesis that no difference between the means of transaction amounts of frauds and non-frauds, is rejected at the 5% significance level.

Using V1 as the representative for the coded variables, the summary statistics are in Table 2.

Class	Count	Mean	Std	Min	Max
Fraud	492	-4.8	6.8	-30.6	2.1
Non-fraud	284,315	0.0	1.9	-56.4	2.5

Table 2:Summary Statistics for V1

The t-test for the transaction amounts of the frauds and non-frauds yields the t-statistic is -15.63 and the corresponding p-value 5.5×10^{-45}. Since the p-value is less than 0.05, the null hypothesis that no difference between the means of transaction amounts of frauds and non-frauds, is rejected at the 5% significance level.

The plot for the correlation coefficients between the independent variables and the predictor ("Class") is shown below.

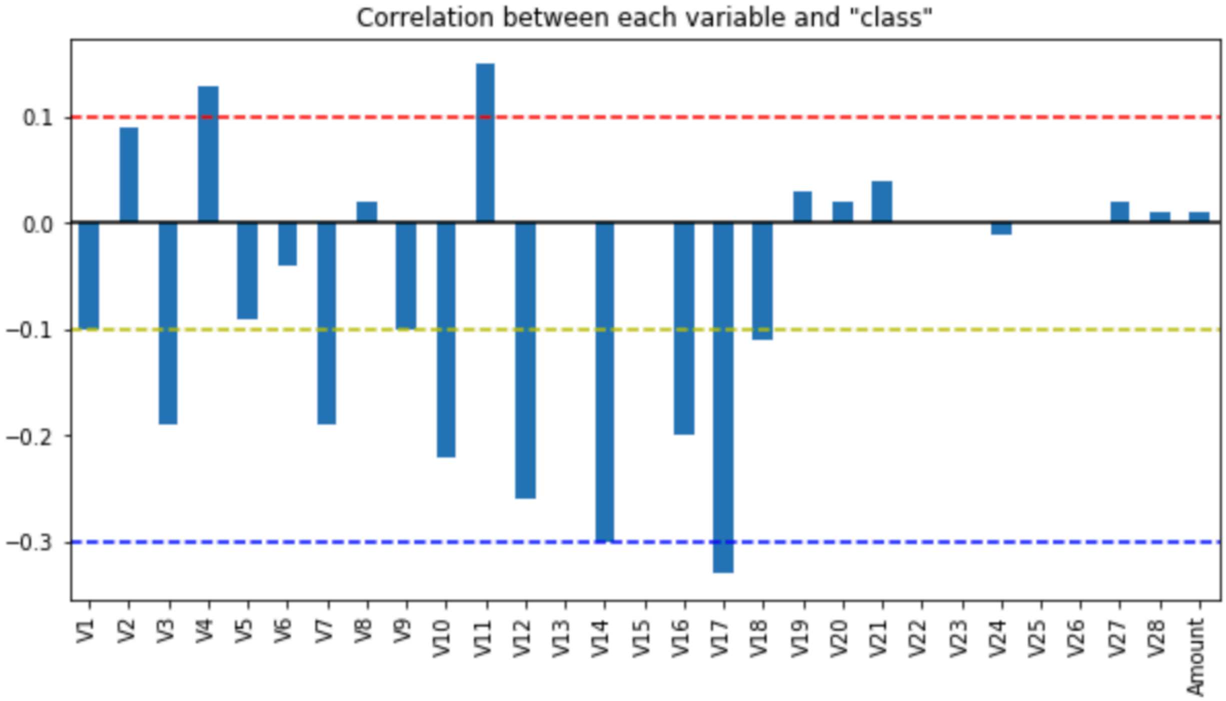

Figure 1: Pearson' s correlation coefficient between each variable and variable "Class"

The dashed lines (0.1, -0.1, -0.3) are markers for the important values used as guidelines for Pearson's correlation coefficient, where [-0.1, -0.3] indicates small negative correlation and [[0.1, 0.3] indicates small positive correlation. The plot suggests that V4 and V11 have small positive correlations with "Class"; V3, V7, V10, V12, V14, V16, and V18 have small negative correlations with "Class"; V17 has medium negative correlations with "Class"; and the remaining variables only have weak correlations with "Class".