Fraud Detection in an Imbalanced Dataset

Step 2: Exploratory analysis of the variables

Samiul Azam

October 28, 2017

Synopsis

In this analysis, I explore all 30 predictor variables (except "Class" variable) to understand the significance of a predictor in classifying Fraud and non-Fraud transactions.

```
# For Better explanation, I conduct my analysis in separate R files.
# Also, I didn't generate any intermediate CSV file in previous step.
# So, I read the CSV File "creditcard.csv" again and then transform
# the "Time" variable.

# Reading the Dataset again
data <- read.csv(file="creditcard.csv", head=TRUE, sep=",")

# Converting time into 24 hours
data$Time <- floor(data$Time/3600) %% 24
```

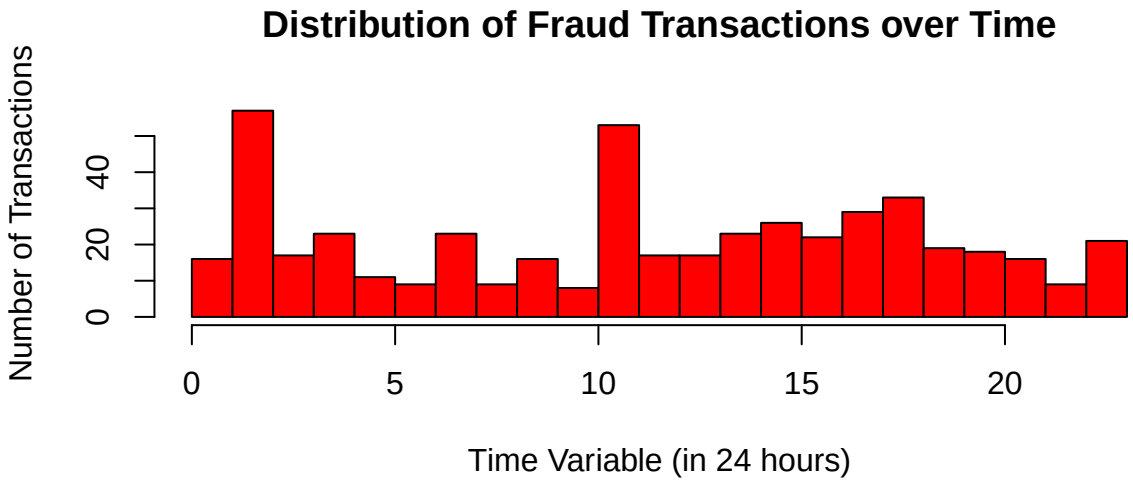
Let's start my exploratory analysis. First split the observations into two data frames based on the "Class" value (1 means Fraud and 0 means non-Fraud).

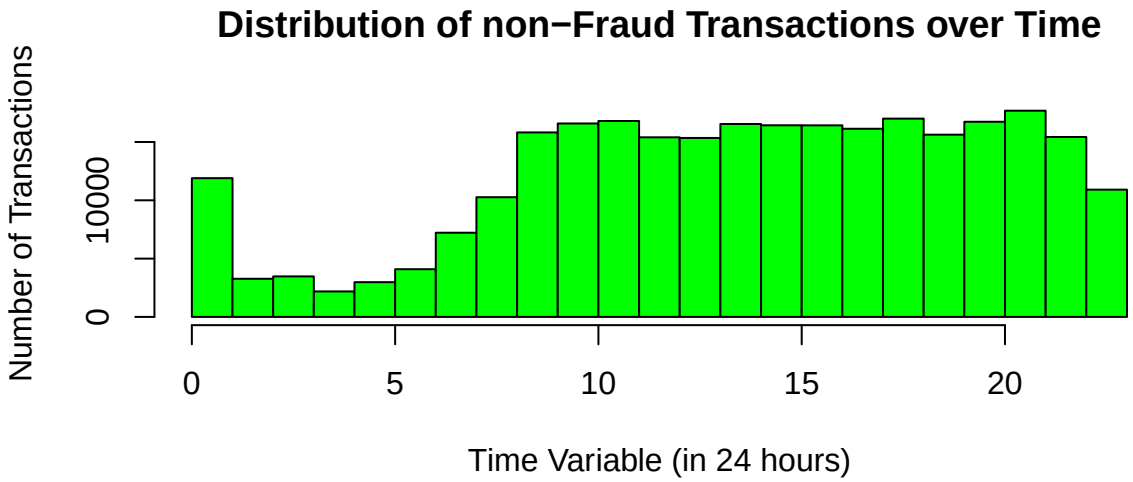
```
# Make two-data frames (one for Fraud, another for non-Fraud)
data_fraud <- data[(data$Class == 1),]
data_non_fraud <- data[(data$Class == 0),]
```

```
#Draw two image in one layout
par(mfrow = c(2, 1))

# Draw the distribution of Fraud Transactions over times
hist(data_fraud[, "Time"], breaks = seq(0, 23, by=1),
     xlab = "Time Variable (in 24 hours)",
     ylab = "Number of Transactions", col = "red",
     main = "Distribution of Fraud Transactions over Time")

# Draw the distribution of non-Fraud Transactions over times
hist(data_non_fraud[, "Time"], breaks = seq(0, 23, by=1),
     xlab = "Time Variable (in 24 hours)",
     ylab = "Number of Transactions", col = "green",
     main = "Distribution of non-Fraud Transactions over Time")
```





We see that two distributions are sufficiently different. So, I think, its important to keep this variable for model learning. Here, we also see that people tends to make lots of transactions from 9AM to 9PM which is expected.

Now, lets see the variable "Amount".

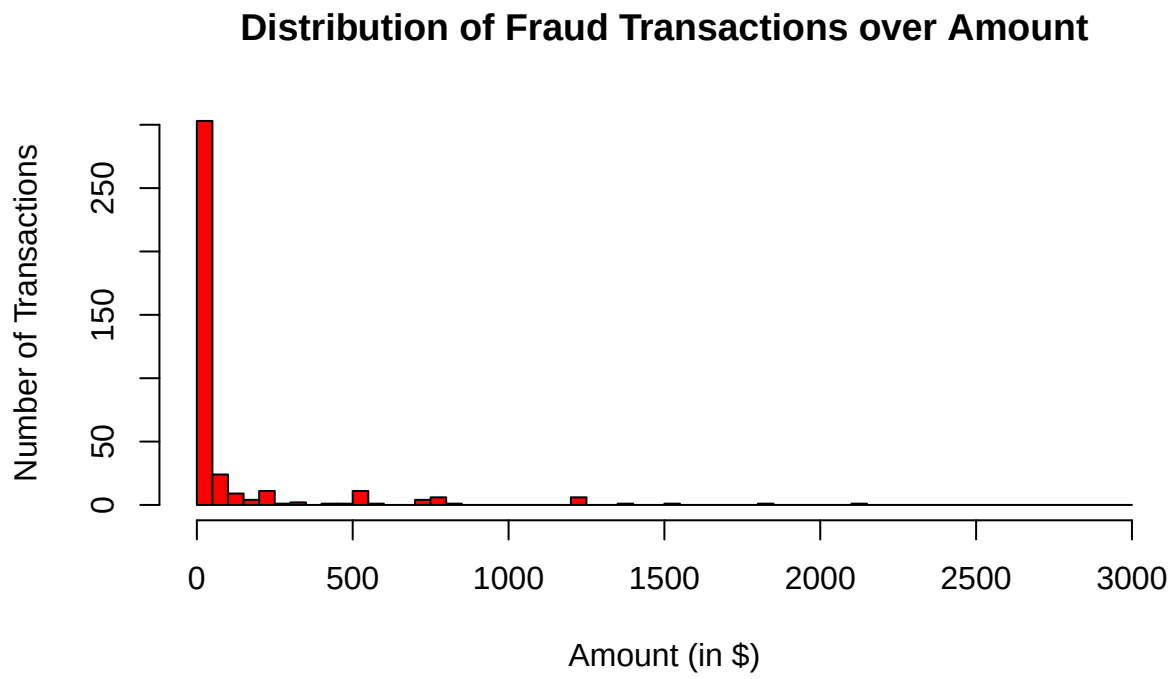
```
#Draw two images in one frame
par(mfrow = c(2, 1))

# See the maximum of "Amount" variable for fraud cases.
summary(data_fraud$Amount)[6]
```

```
## Max.
## 2126
```

The Maximum amount in Fraud cases is 2126 $.

```r
# Draw the distribution of Fraud Transactions over Amount.
# As the maximum is 2126$, we take 3000$ as last value in x axis.
hist(data_fraud[ (data_fraud$Amount), "Amount"],
     breaks = seq(0,3000,by=50),
     xlab = "Amount (in $)",
     ylab = "Number of Transactions", col = "red",
     main = "Distribution of Fraud Transactions over Amount")
```




```r
# See the maximum of "Amount" variable for non-fraud cases.
summary(data_non_fraud$Amount)[6]
```

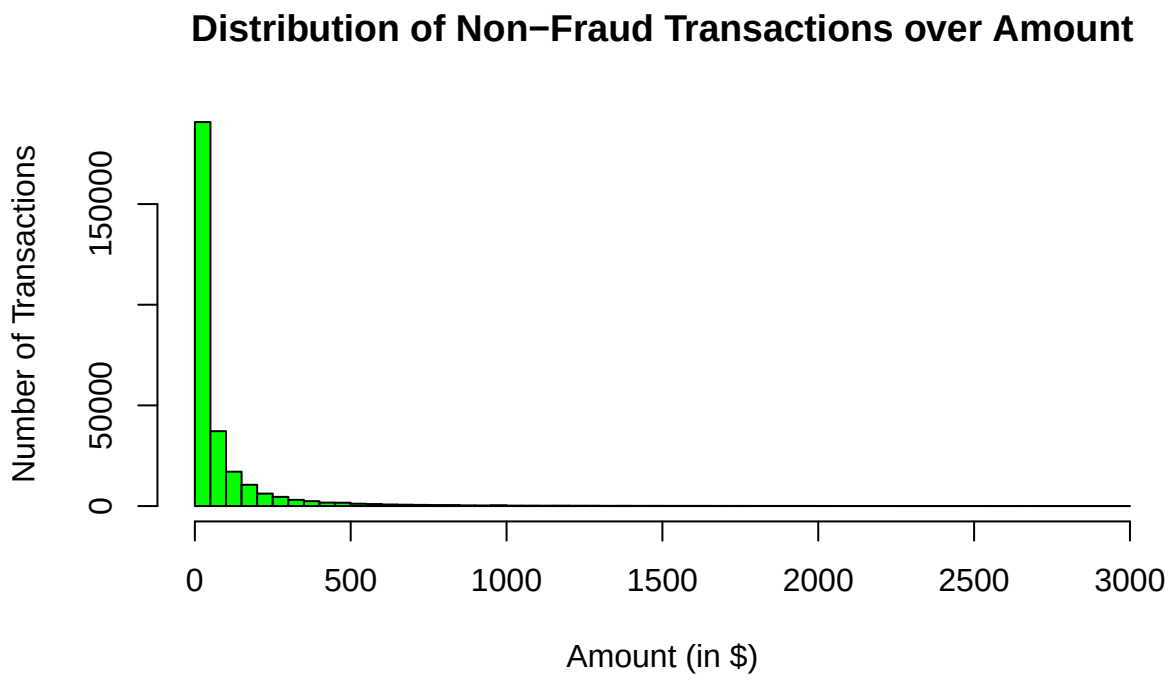
```
##  Max.
## 25690
```

```r
# See the 99.9% quantile
quantile(data_non_fraud$Amount,0.999)
```

```
## 99.9%
##  3000
```

The Maximum amount in non-Fraud cases is 25690 $, which is very high. However, if we see the 99.9% quantile of the non-fraud amount, it is very low (3000 $) than the maximum (25690 $). So, we ignore amounts greater than 3000 $.

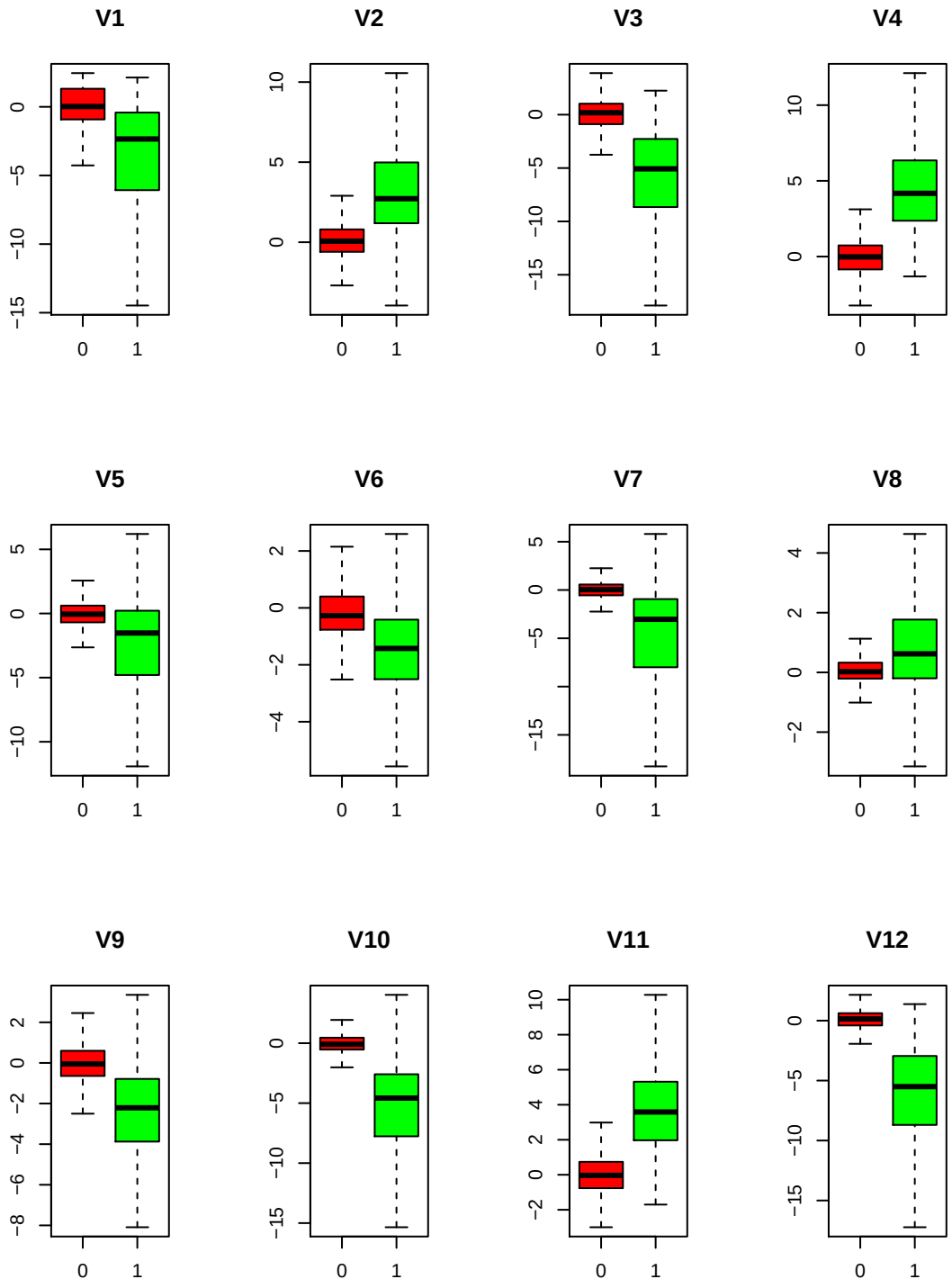
```r
# Draw the distribution of non-Fraud Transactions over Amount.
hist(data_non_fraud[ (data_non_fraud$Amount <= 3000), "Amount"],
     breaks = seq(0,3000,by=50),
     xlab = "Amount (in $)",
     ylab = "Number of Transactions", col = "green",
     main = "Distribution of Non-Fraud Transactions over Amount")
```



From the distributions of transactions over amounts, we see both of them are identical (both fraud and non-fraud transactions are usually very low in amount: 1 $ to 150 $). So, I think it's not a good predictor for fraud detection.

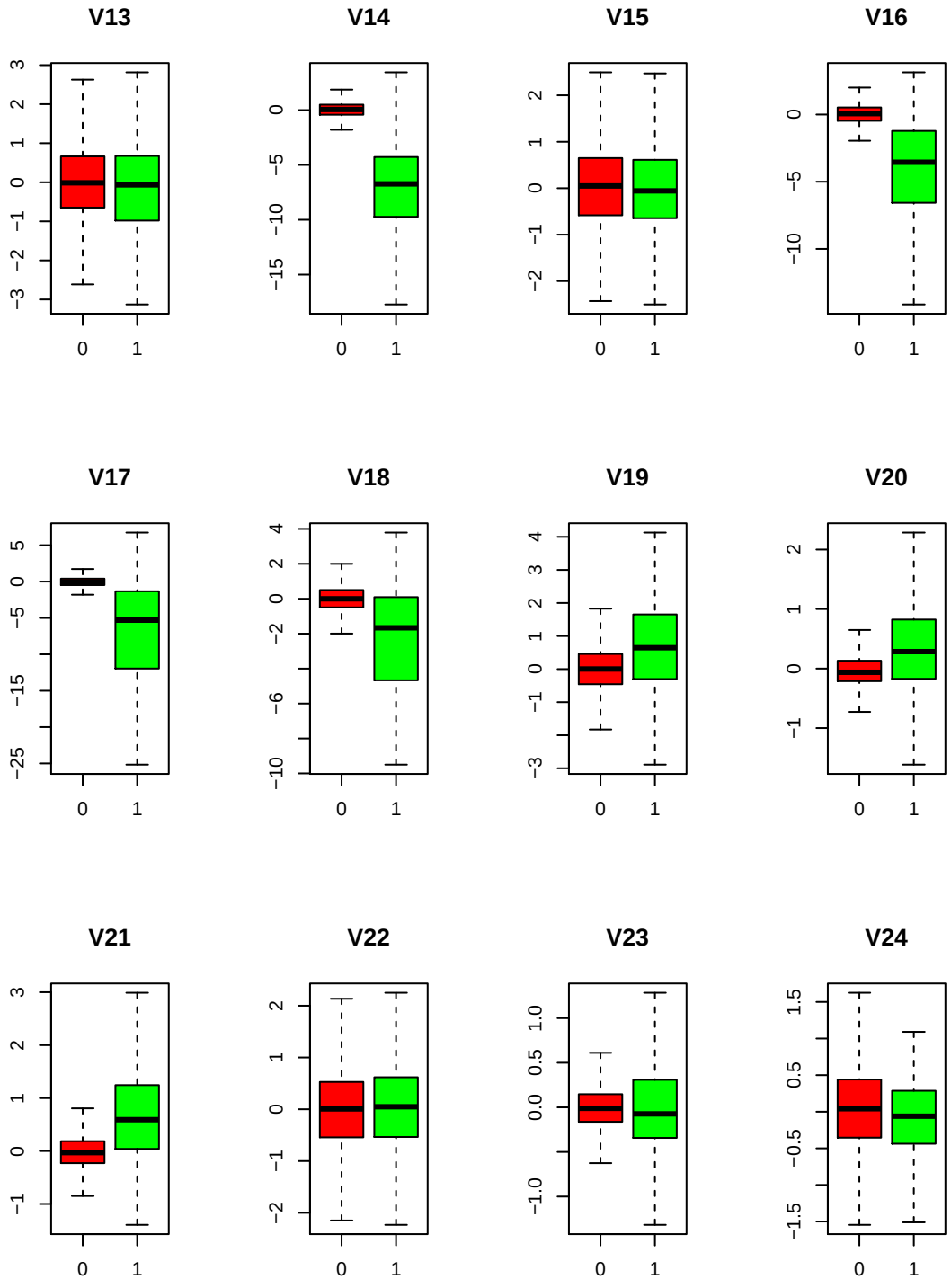
Now lets explore the PCA transformed variables "V1" to "V28" using Boxplots.

```r
par(mfrow = c(3, 4))
for (i in 1:12){# Display Box Plots of the variables V1 to V12
      var_name <- paste("V",as.character(i),sep = "")
      boxplot(data[,var_name] ~ Class, data = data, outline = FALSE,
              col = c("red","green"), main = var_name)
}
```

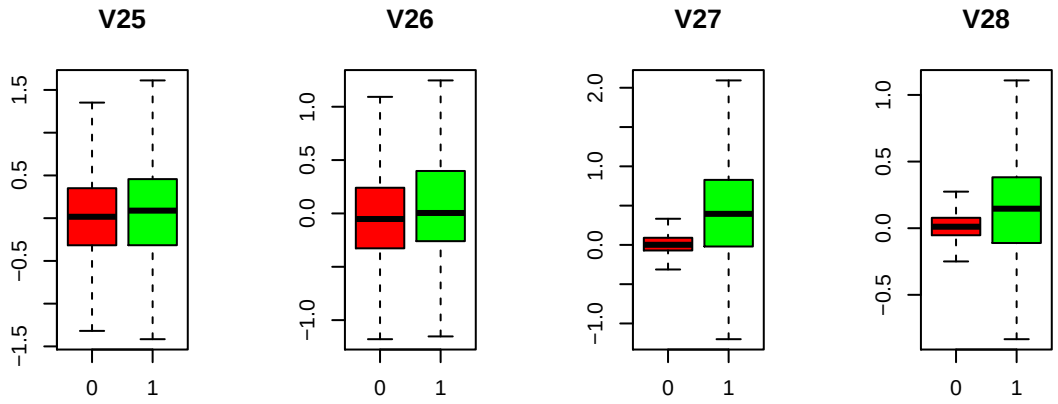



```
par(mfrow = c(3, 4))
for (i in 13:24){ # Display Box Plots of the variables V13 to V24
      var_name <- paste("V",as.character(i),sep = "")
      boxplot(data[,var_name] ~ Class, data = data, outline = FALSE,
```

```
                col = c("red","green"), main = var_name)
}
```



```
par(mfrow = c(1, 4))
for (i in 25:28){ # Display Box Plots of the variables V25 to V28
      var_name <- paste("V",as.character(i),sep = "")
      boxplot(data[,var_name] ~ Class, data = data, outline = FALSE,
              col = c("red","green"), main = var_name)
}
```



Boxplot provides a simple way to display data distribution for different groups. Here, groups are the Fraud (0 or Red box) and Non-fraud (1 or Green Box). We can ignore variables V13, V15, V22, V23, V24, V25, V26 and V28 as there are significant overlapping (less separability) between Fraud and non-Fraud transactions.

Finally, the selected variables (after this analysis) are: V1 to V12, V14, V16 to V21, V27 and Time.

(Total 21 variables are selected out of 30)