Fraud Detection in an Imbalanced Dataset

Step 1: Observing the database

Samiul Azam

October 27, 2017

Synopsis

In this analysis, I consider the given database ("creditcard.csv" provided by ATB, Calgary) to understand the basic structure of the data, metadata and variables (types and values). In addition, I check for any inconsistency, as well as apply a transformation of the "Time" variable to make it easy-understandable.

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

# Let's see the dimension of the data
dim(data)
```

```
## [1] 284807     31
```

So, there are 284807 observations with 31 variables.

```
# Let see the Variable names
colnames(data)
```

```
##  [1] "Time"   "V1"     "V2"     "V3"     "V4"     "V5"     "V6"
##  [8] "V7"     "V8"     "V9"     "V10"    "V11"    "V12"    "V13"
## [15] "V14"    "V15"    "V16"    "V17"    "V18"    "V19"    "V20"
## [22] "V21"    "V22"    "V23"    "V24"    "V25"    "V26"    "V27"
## [29] "V28"    "Amount" "Class"
```

I need to know the exact variable names of in the dataframe, so that I can use them in the later part of the program.

```
# Summary statistics on each variable (Min, Max, Mean and quantiles)
summary(data)
```

```
##       Time              V1                    V2
##  Min.   :     0   Min.   :-56.40751   Min.   :-72.71573
##  1st Qu.: 54202   1st Qu.: -0.92037   1st Qu.: -0.59855
##  Median : 84692   Median :  0.01811   Median :  0.06549
##  Mean   : 94814   Mean   :  0.00000   Mean   :  0.00000
##  3rd Qu.:139321   3rd Qu.:  1.31564   3rd Qu.:  0.80372
##  Max.   :172792   Max.   :  2.45493   Max.   : 22.05773
##        V3                 V4                    V5
##  Min.   :-48.3256   Min.   :-5.68317   Min.   :-113.74331
##  1st Qu.: -0.8904   1st Qu.:-0.84864   1st Qu.:  -0.69160
```

```
##  Median :  0.1799   Median :-0.01985   Median :  -0.05434
##  Mean   :  0.0000   Mean   : 0.00000   Mean   :   0.00000
##  3rd Qu.:  1.0272   3rd Qu.: 0.74334   3rd Qu.:   0.61193
##  Max.   :  9.3826   Max.   :16.87534   Max.   :  34.80167
##        V6                 V7                 V8
##  Min.   :-26.1605   Min.   :-43.5572   Min.   :-73.21672
##  1st Qu.: -0.7683   1st Qu.: -0.5541   1st Qu.: -0.20863
##  Median : -0.2742   Median :  0.0401   Median :  0.02236
##  Mean   :  0.0000   Mean   :  0.0000   Mean   :  0.00000
##  3rd Qu.:  0.3986   3rd Qu.:  0.5704   3rd Qu.:  0.32735
##  Max.   : 73.3016   Max.   :120.5895   Max.   : 20.00721
##        V9                 V10                V11
##  Min.   :-13.43407   Min.   :-24.58826   Min.   :-4.79747
##  1st Qu.: -0.64310   1st Qu.: -0.53543   1st Qu.:-0.76249
##  Median : -0.05143   Median : -0.09292   Median :-0.03276
##  Mean   :  0.00000   Mean   :  0.00000   Mean   : 0.00000
##  3rd Qu.:  0.59714   3rd Qu.:  0.45392   3rd Qu.: 0.73959
##  Max.   : 15.59500   Max.   : 23.74514   Max.   :12.01891
##       V12                V13                V14
##  Min.   :-18.6837   Min.   :-5.79188   Min.   :-19.2143
##  1st Qu.: -0.4056   1st Qu.:-0.64854   1st Qu.: -0.4256
##  Median :  0.1400   Median :-0.01357   Median :  0.0506
##  Mean   :  0.0000   Mean   : 0.00000   Mean   :  0.0000
##  3rd Qu.:  0.6182   3rd Qu.: 0.66251   3rd Qu.:  0.4931
##  Max.   :  7.8484   Max.   : 7.12688   Max.   : 10.5268
##       V15                V16                V17
##  Min.   :-4.49894   Min.   :-14.12985   Min.   :-25.16280
##  1st Qu.:-0.58288   1st Qu.: -0.46804   1st Qu.: -0.48375
##  Median : 0.04807   Median :  0.06641   Median : -0.06568
##  Mean   : 0.00000   Mean   :  0.00000   Mean   :  0.00000
##  3rd Qu.: 0.64882   3rd Qu.:  0.52330   3rd Qu.:  0.39968
##  Max.   : 8.87774   Max.   : 17.31511   Max.   :  9.25353
##       V18                V19                V20
##  Min.   :-9.498746   Min.   :-7.213527   Min.   :-54.49772
##  1st Qu.:-0.498850   1st Qu.:-0.456299   1st Qu.: -0.21172
##  Median :-0.003636   Median : 0.003735   Median : -0.06248
##  Mean   : 0.000000   Mean   : 0.000000   Mean   :  0.00000
##  3rd Qu.: 0.500807   3rd Qu.: 0.458949   3rd Qu.:  0.13304
##  Max.   : 5.041069   Max.   : 5.591971   Max.   : 39.42090
##       V21                V22                V23
##  Min.   :-34.83038   Min.   :-10.933144   Min.   :-44.80774
##  1st Qu.: -0.22839   1st Qu.: -0.542350   1st Qu.: -0.16185
##  Median : -0.02945   Median :  0.006782   Median : -0.01119
##  Mean   :  0.00000   Mean   :  0.000000   Mean   :  0.00000
##  3rd Qu.:  0.18638   3rd Qu.:  0.528554   3rd Qu.:  0.14764
##  Max.   : 27.20284   Max.   : 10.503090   Max.   : 22.52841
##       V24                V25                V26
##  Min.   :-2.83663   Min.   :-10.29540   Min.   :-2.60455
##  1st Qu.:-0.35459   1st Qu.: -0.31715   1st Qu.:-0.32698
##  Median : 0.04098   Median :  0.01659   Median :-0.05214
##  Mean   : 0.00000   Mean   :  0.00000   Mean   : 0.00000
##  3rd Qu.: 0.43953   3rd Qu.:  0.35072   3rd Qu.: 0.24095
##  Max.   : 4.58455   Max.   :  7.51959   Max.   : 3.51735
##       V27                V28               Amount
```

```
##  Min.   :-22.565679   Min.   :-15.43008   Min.   :    0.00
##  1st Qu.: -0.070840   1st Qu.: -0.05296   1st Qu.:    5.60
##  Median :  0.001342   Median :  0.01124   Median :   22.00
##  Mean   :  0.000000   Mean   :  0.00000   Mean   :   88.35
##  3rd Qu.:  0.091045   3rd Qu.:  0.07828   3rd Qu.:   77.17
##  Max.   : 31.612198   Max.   : 33.84781   Max.   :25691.16
##      Class
##  Min.   :0.000000
##  1st Qu.:0.000000
##  Median :0.000000
##  Mean   :0.001728
##  3rd Qu.:0.000000
##  Max.   :1.000000
```

Here, we see the range of the data for each variable. Variables "V1" to "V28" are the PCA transformed anonymized data. As they are PCA transformed, all of them have zero means. If we see the mean of the "Class" variable, we can say that only 0.173% of transactions are the fraud.

```
# Let see the types of the variables with few examples.
str(data)
```

```
## 'data.frame':    284807 obs. of  31 variables:
##  $ Time  : num  0 0 1 1 2 2 4 7 7 9 ...
##  $ V1    : num  -1.36 1.192 -1.358 -0.966 -1.158 ...
##  $ V2    : num  -0.0728 0.2662 -1.3402 -0.1852 0.8777 ...
##  $ V3    : num  2.536 0.166 1.773 1.793 1.549 ...
##  $ V4    : num  1.378 0.448 0.38 -0.863 0.403 ...
##  $ V5    : num  -0.3383 0.06 -0.5032 -0.0103 -0.4072 ...
##  $ V6    : num  0.4624 -0.0824 1.8005 1.2472 0.0959 ...
##  $ V7    : num  0.2396 -0.0788 0.7915 0.2376 0.5929 ...
##  $ V8    : num  0.0987 0.0851 0.2477 0.3774 -0.2705 ...
##  $ V9    : num  0.364 -0.255 -1.515 -1.387 0.818 ...
##  $ V10   : num  0.0908 -0.167 0.2076 -0.055 0.7531 ...
##  $ V11   : num  -0.552 1.613 0.625 -0.226 -0.823 ...
##  $ V12   : num  -0.6178 1.0652 0.0661 0.1782 0.5382 ...
##  $ V13   : num  -0.991 0.489 0.717 0.508 1.346 ...
##  $ V14   : num  -0.311 -0.144 -0.166 -0.288 -1.12 ...
##  $ V15   : num  1.468 0.636 2.346 -0.631 0.175 ...
##  $ V16   : num  -0.47 0.464 -2.89 -1.06 -0.451 ...
##  $ V17   : num  0.208 -0.115 1.11 -0.684 -0.237 ...
##  $ V18   : num  0.0258 -0.1834 -0.1214 1.9658 -0.0382 ...
##  $ V19   : num  0.404 -0.146 -2.262 -1.233 0.803 ...
##  $ V20   : num  0.2514 -0.0691 0.525 -0.208 0.4085 ...
##  $ V21   : num  -0.01831 -0.22578 0.248 -0.1083 -0.00943 ...
##  $ V22   : num  0.27784 -0.63867 0.77168 0.00527 0.79828 ...
##  $ V23   : num  -0.11 0.101 0.909 -0.19 -0.137 ...
##  $ V24   : num  0.0669 -0.3398 -0.6893 -1.1756 0.1413 ...
##  $ V25   : num  0.129 0.167 -0.328 0.647 -0.206 ...
##  $ V26   : num  -0.189 0.126 -0.139 -0.222 0.502 ...
##  $ V27   : num  0.13356 -0.00898 -0.05535 0.06272 0.21942 ...
##  $ V28   : num  -0.0211 0.0147 -0.0598 0.0615 0.2152 ...
##  $ Amount: num  149.62 2.69 378.66 123.5 69.99 ...
##  $ Class : int  0 0 0 0 0 0 0 0 0 0 ...
```

Most of the variables are real numbers. Only "Class" variable is integer. Most importantly, there is no non-numeric (string) data.

```
#Let see is there any missing values inside any variable or column
colSums(is.na(data))
```

```
##   Time     V1     V2     V3     V4     V5     V6     V7     V8     V9
##      0      0      0      0      0      0      0      0      0      0
##    V10    V11    V12    V13    V14    V15    V16    V17    V18    V19
##      0      0      0      0      0      0      0      0      0      0
##    V20    V21    V22    V23    V24    V25    V26    V27    V28 Amount
##      0      0      0      0      0      0      0      0      0      0
##  Class
##      0
```

Great, there is no missing values. If there are missing values, then I need to replace them with the mean/median value of the corresponding variable.

```
# Count number of positive classes (Fraud transactions)
sum(data$Class == 1)
```

```
## [1] 492
```

Only 492 transactions are fraud. Rest of the 284315 transactions are genuine. So, the data is highly skewed/imbalanced.

```
# Let see the first 100 observations of the Time variable.
head(data$Time, n = 100)
```

```
##   [1]  0  0  1  1  2  2  4  7  7  9 10 10 10 11 12 12 12 13 14 15 16 17 18
##  [24] 18 22 22 23 23 23 23 24 25 26 26 26 26 27 27 29 29 32 32 33 33 34 34
##  [47] 34 34 35 35 35 36 36 36 37 38 39 39 40 41 41 41 41 42 42 44 44 44 44
##  [70] 46 46 46 47 48 48 49 49 49 50 50 51 52 52 53 54 55 55 56 56 59 59 60
##  [93] 60 62 64 64 64 67 67 68
```

"Time" variable is monotonically increasing sequence of data, as it's the elapsed time (in seconds) from the start point of the data collection phase.

```
# Let see the time frame of the data in hours
summary(data$Time)[6]/3600 # 1 hour is 3600 seconds
```

```
## Max.
##   48
```

That means, all these transactions are gathered in two days of time frame.

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   15.00   14.05   19.00   23.00
```

For easy interpretability, the "Time" variable is transformed into 24 hours (0 - 23). It will help to co-relate the fraud transactions with the hours of a day. Let assume that data collection starts at 00:00 AM.

After this analysis, we have got the basic understanding of the data and the variables. We conclude that the data is well-structured and tidy. In the nest step, I will do exploratory analysis of the variables to get some intuitions on the importance of individual variables in Fraud detection.