Globant technical test

Jeison Mesa

16 abril, 2021

Introduction

In the workflow for data science professionals, it is important to achieve process automation. Automating processes allows to control human errors made by performing tasks manually. It is proposed to download files using web scraping. In this case, R is connected to the information available for each of the data sets in the web page. It is to recognise the structure of the information. It was possible to identify two types of data structure xls and csv files, for the set of csv files a totally different pattern was found to the xls format files, for this situation only the xls format files were keep. The next step is to have a logical file name. In this case, having files for different months, it is proposed to work with the following format "PCT_year_month" for all the files. This is because the structure presented on the website seemed rather "messy."

The additional step is done to identify the columns that are found for all the files. Resulting in the following variables:

```
colnames(clean_data)
```

```
##  [1] "date"               "TRANS VAT DESC"     "ORIGINAL GROSS AMT"
##  [4] "MERCHANT NAME"      "CARD NUMBER"        "TRANS CAC CODE 1"
##  [7] "TRANS CAC DESC 1"   "TRANS CAC CODE 2"   "TRANS CAC DESC 2"
## [10] "TRANS CAC CODE 3"   "year"               "month"
## [13] "day"                "wday"
```

As can be seen, a date column has been generated in ymd format; in addition to this, year, month and day columns were generated separately in order to have a more universal date format, it is also proposed to calculate the day of the week. In economics, it is usual to find different relationships for weekdays compared to weekends in financial variables.

Quality Control

The inferences that can be by a statistical model are largely influenced by the quality of the information, i.e. we must be able to guarantee that the information was verified and validated; for this purpose the following is generally checked:

1. **Problems associated with the scale of the variable**

Regardless of the type of software that we want to use, it is necessary that at the moment of loading the information those variables that by their typology are strictly numerical, the software can identify it because if we adjust a statistical model with a variable that is numerical and the software detects it as categorical we will reach erroneous conclusions (this is only an example of the many cases that can occur) or vice versa when giving numerical values to categorical variables the software detects them as numerical, in this case it is necessary to perform a transformation.

2. **Missing values**

It is important to be able to check which variables have a high percentage of missing values, since many of the optimisations of statistical models require having the complete data matrix. In this sense, if the information provided presents missing values for some of the variables, a process of imputation of values must be carried out (by some statistical technique ML, pca + NIPALS, KNN, etc...). But it is important to recognize the adjacent stochastic process that arises from credit card transactions. In essence, it is possible to consider a time series with irregular observations over time (Eyheramendy, Elorrieta, and Palma 2016). This is a new approach and can be analyzed by kalman filters and state space representation. In my master thesis I am proposing a regression approach with autocorrelated errors and irregular observations over time, such as credit transactions.

Figure 1: Valores Faltantes

Assuming that traditional methods will be used, the greatest concentration of missing data is found in the "TRANS VAT DESC" variable (Figure 1), and for this reason it will not be used in the statistical inferences to be developed. But it is important to recognize the adjacent stochastic process that arises from credit card transactions. In essence, one can consider a time series with irregular observations over time.

3. **Problems due to the Characteristic of your Variables**.

In many cases we cannot assume that the information is correct, since there may be typing problems or problems due to "outliers" (e.g. imagine a negative precipitation is not possible). When the problem related to the outlier is trivial, i.e. we can know with certainty that it is a value that cannot occur, we can choose to induce a missing value for that data. On the other hand, in order to identify possible outliers it is necessary to use descriptive or inferential statistical techniques; for the descriptive option a box plot can be performed and for the inferential case a probability distribution can be fitted and those values which are not under the distribution curve can be determined.

The Figure 2 shows possible outliers for some transactions. In this case we have a record that may be due to a return of around £5000000 and a value per purchase of more than £250000. These are points that we will later evaluate if they can truly be considered as outliers or fraud issues.

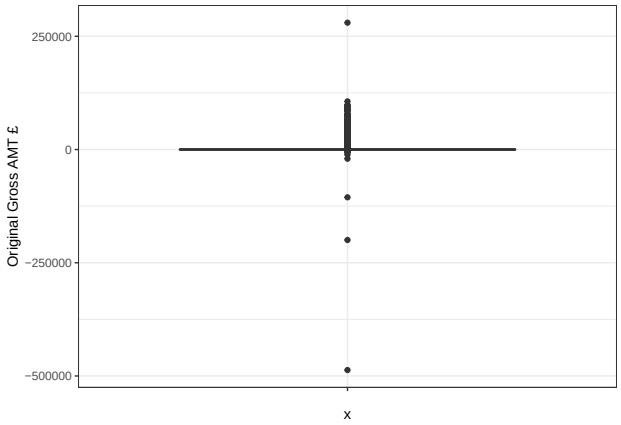

Figure 2: Gross Outlier Detection

Why is it necessary to develop a deeper analysis? It is necessary because we cannot consider outliers under the assumption of the interquartile ranges of the boxplot. Other types of outlier distributions such as the Gumbel distribution or non-parametric methods can be considered.

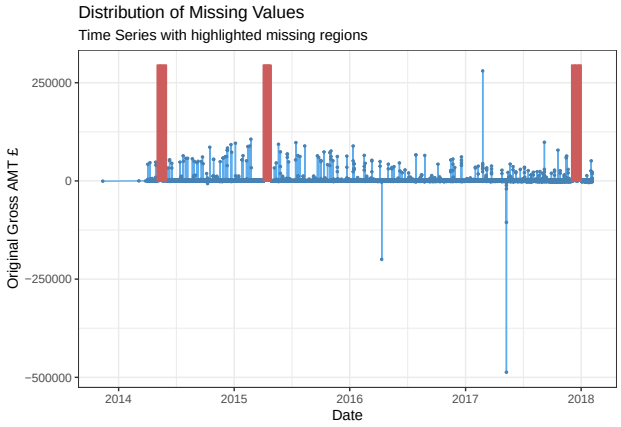

Figure 3: Time Serie Gross AMT

The figure 3 shows the dynamics associated with the transactions carried out. In particular, some rather "strange" points can be observed according to the records. The red areas indicate two cases: no transactions were made or no information was found for those periods.

References

Eyheramendy, Susana, Felipe Elorrieta, and Wilfredo Palma. 2016. "An Autoregressive Model for Irregular Time Series of Variable Stars." *Proceedings of the International Astronomical Union* 12 (S325): 259–62. https://doi.org/10.1017/S1743921317000448.