Otto Product Classification

Mariem BOUHAHA Eden BELOUADAH

1er décembre 2017

1 Introduction

Classification is an important field since it's the base of so many real world problems. This project is based on the (Otto product classification) Kaggle Challenge.

In order to make a consistent analysis of products, Otto company organized this challenge. The main goal is to classify the gven product through 10 possible classes.

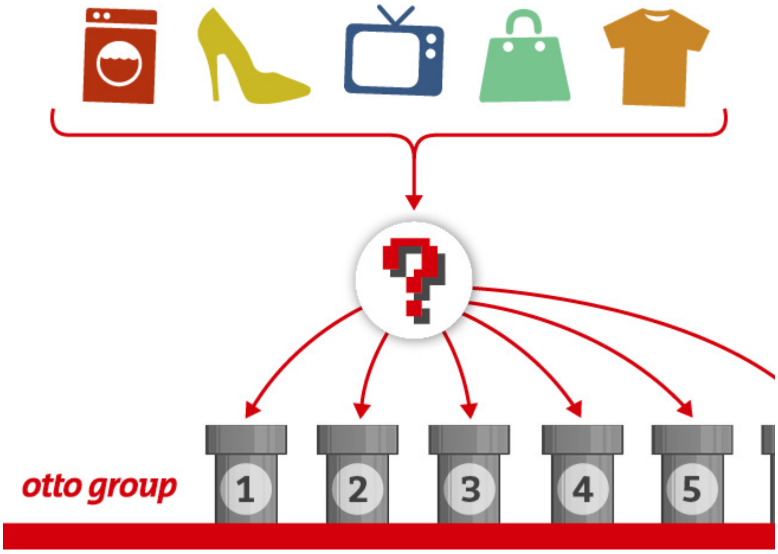
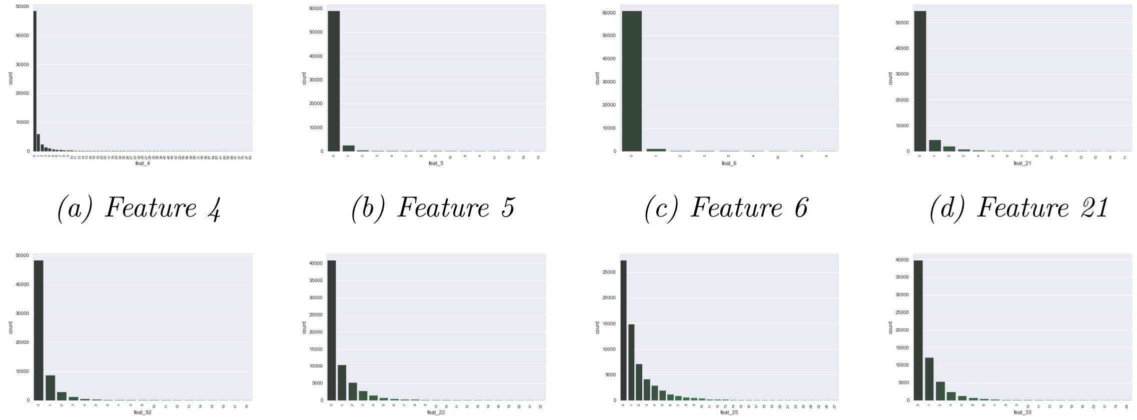


Figure 1. Otto products classification

Tha dataset provided by Otto is composed of 61878 examples. Each example contains 93 features. All features are categorical and each one of them contains so many possible categories that can reach 300.

2 Data analytics

Before going into classification problem, we started by providing some statistics and caracteristics of the dataset. The first remarkable thing about the features categories is that the category 0 is the most dominant one. This is why we notice an unbalacement in categories for all features. The next figure shows examples of features


(a) Feature 4 (b) Feature 5 (c) Feature 6 (d) Feature 21

(e) Feature 92 (f) Feature 22 (g) Feature 25 (h) Feature 33

Figure 2. Categories distribution for some features

The dominance of the zero class is probably due to unknown features values that has been just transformed to 0.

2.1 Features correlation

The correlation study that we did on features reveals that there's no remarkable correlation between them. The more blue is the region, the more correlated the features are.

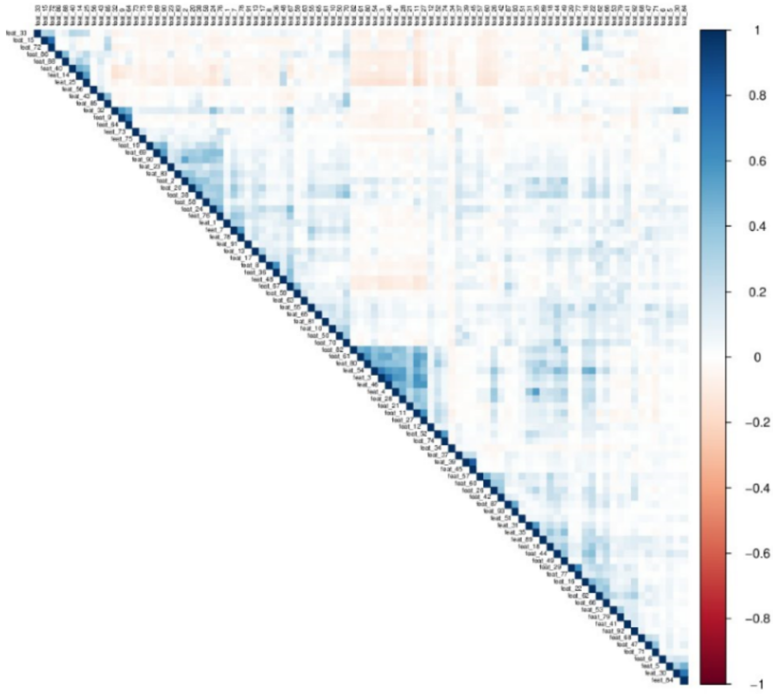


Figure 3. Correlation between features

2.2 Classes distribution

As the features categories are not balanced, the problem classes are not balanced too ! We see that some classes figure most of time, but this in not the case for the other classes. This makes classification problem more difficult.

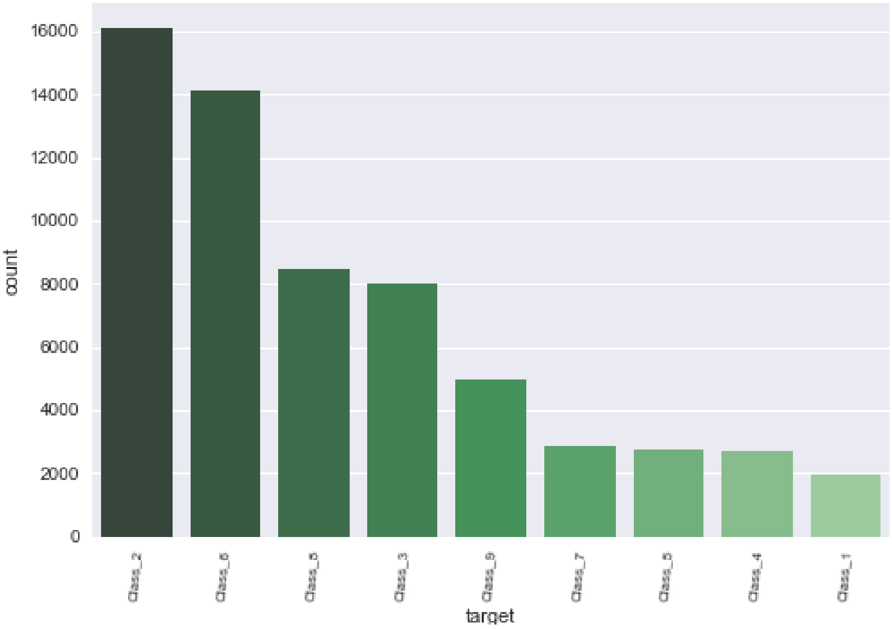
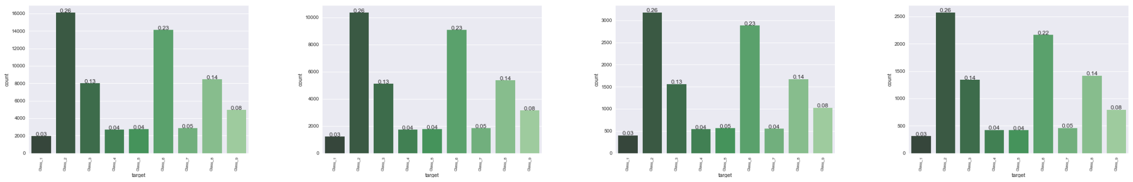


Figure 4. Classes distribution in the dataset

3 Classification process

Because the classes are not balanced, we aimed to devide the dataset in such a way that we keep the same pourcentage of each classe in both train and test sets. The figures below show the result of classes distribution before and after deviding the dataset.



(a) Full dataset

(b) Train set

(c) Test set

(d) Validation set

Figure 5. Classes distribution for different sets

4 Conclusion

This is a conclusion