# Otto Group Product Classification Challenge

*Classify products into the correct category*

By

ZHUFENG LI, HAO LIU


Computer Science Department of the Faculty of Science

UNIVERSITY PARIS SUD

NOVEMBER 5, 2018

# 1. Introduction

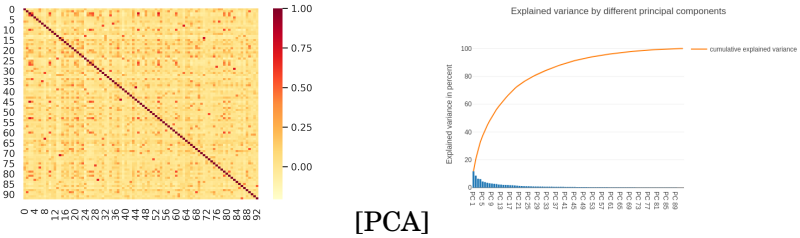
For this competition, we are provided a dataset with 93 features for more than 200,000 products. The objective is to build a predictive model which is able to distinguish between our main product categories. Submissions are evaluated using the multi-class logarithmic loss. Each product has been labeled with one true category. For each product, you must submit a set of predicted probabilities (one for every category). The formula is then,

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} log(p_{ij})$$

where N is the number of products in the test set, M is the number of class labels, log is the natural logarithm, $y_{ij}$ is 1 if observation $i$ is in class $j$ and 0 otherwise, and $p_{ij}$ is the predicted probability that observation $i$ belongs to class $j$. More description on the following link

# 2. Data analysis

We will study the correlation between all these variables. We plotted the correlation matrix which is a table showing correlation coefficients between sets of variables. In this matrix, +1 is the case of a perfect direct (increasing) linear relationship (correlation), -1 is the case of a perfect decreasing (inverse). We also did principal component analysis on all the features. A distribution of variance of all the principal components is given below. The total variance is the sum of variances of all individual principal components.




[Correlation matrix] [PCA]

Figure 1: Figures side by side

## 3. Machine learning models

Several different methods of machine learning are used in this project. They are decision tree, random forest, naive bayes, k nearest neighbors, support vector machines, and neural network. After testing and fine tuning hyperparameters, we benchmarked their performances with kaggle submission score. As it shows in the table below TABLE 1, a finely tuned nueral network gave us the best score w.r.t the others. Submissions made by k nearest neighbors are worse than others in general. The results of SVM are acceptable. However due to the complexity of SVM, it took quite a while to predict $p_{ij}$. Decision tree performed very poorly on this challenge. Because, it can only provide a categorical prediction instead of a probability. Thus, it has a huge log loss. Random forest gave a very good score, especially when it comes to computational efficiency.

| Comparison of performance of models | | | |
|---|---|---|---|
| Model | Score | Model | Score |
| Neural network (Complexe ver.) | 0.46422 | KNN(100) | 0.69295 |
| Random forest | 0.47818 | SVM(poly) | 0.78819 |
| Neural Network (Simple ver.) | 0.48763 | SVM(linear) | 0.78844 |
| SVM(rbf) | 0.57218 | KNN(1000) | 0.81102 |

TABLE 1. Comparison of performance

## 4. Results

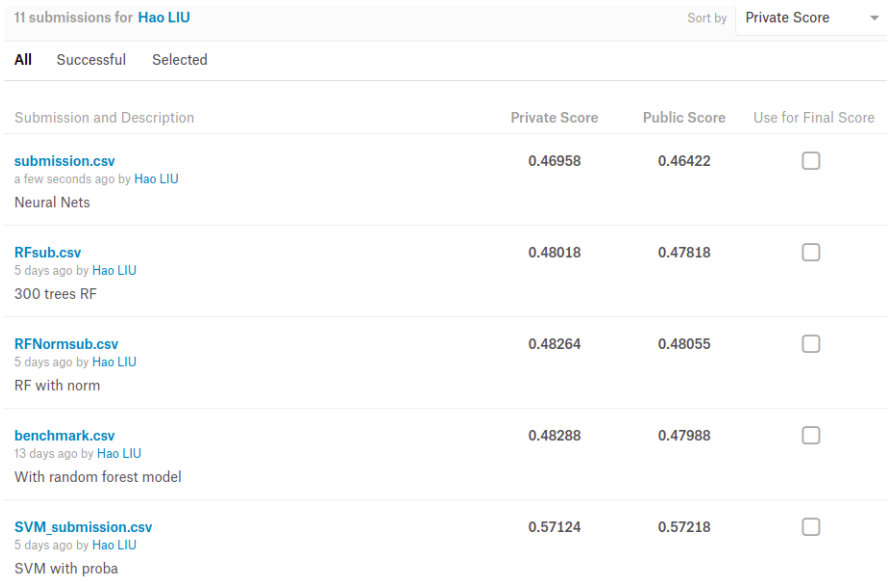



FIGURE 2. Results from submission board