

# The categorization challenge organized by Cdiscount on datascience.net in 2015: analysis of the released data set and winning contributions

Yang JIAO<sup>1</sup>, Bruno GOUTORBE<sup>2</sup>, Christelle GRAUER<sup>3</sup>, Matthieu CORNEC<sup>4</sup> et Jérémie JAKUBOWICZ<sup>5</sup>

## TITLE

Le challenge de catégorisation organisé par Cdiscount sur datascience.net en 2015 : analyse du jeu de données mis à disposition et des contributions gagnantes

## RÉSUMÉ

En 2015, Cdiscount a mis la communauté au défi de prévoir la catégorie correcte de ses produits à partir de certains de leurs attributs comme le libellé, la description, le prix ou l'image associée. Les candidats ont eu accès à l'intégralité du catalogue de produits actifs en mai 2015, soit environ 15.8 millions d'items répartis dans 5,789 catégories, hormis une petite partie qui a servi d'ensemble de test. La qualité des données est loin d'être homogène et la répartition des catégories est extrêmement déséquilibrée, ce qui complique la tâche de catégorisation. Les cinq algorithmes gagnants, sélectionnés parmi plus de 3,500 contributions, atteignent un taux de prévisions correctes de 66–68% sur l'ensemble de test. La plupart utilisent des modèles linéaires simples comme des régressions logistiques, ce qui suggère que les étapes préliminaires telles que le pré-traitement du texte, sa vectorisation et le rééchantillonnage des données sont plus cruciales que le choix de modèles non-linéaires complexes. En particulier, les gagnants corrigent tous le déséquilibre des catégories par des méthodes d'échantillonnage aléatoire ou de pondération en fonction de l'importance des catégories. Les deux meilleurs algorithmes se distinguent par leur aggrégation de grands nombres de modèles entraînés sur des sous-ensembles aléatoires des données. Le catalogue de produits est mis à disposition de la communauté de recherche et formation scientifique, qui disposera ainsi de données réelles issues du e-commerce pour étalonner et améliorer les algorithmes de classification basés sur le texte et les images dans un contexte de très grand nombre de classes.

*Mots-clés* : classification, e-commerce, big data, jeu de données public.

---

<sup>1</sup>SAMOVAR, CNRS, Télécom SudParis, Univ. Paris-Saclay, Evry, France, yang.jiao@telecom-sudparis.eu

<sup>2</sup>Cdiscount, Bordeaux, France, bruno.goutorbe@cdiscout.com

<sup>3</sup>Cdiscount, Bordeaux, France, christelle.grauer@cdiscout.com

<sup>4</sup>Cdiscount, Bordeaux, France, matthieu.cornec@cdiscout.com

<sup>5</sup>SAMOVAR, CNRS, Télécom SudParis, Univ. Paris-Saclay, Evry, France, jeremie.jakubowicz@telecom-sudparis.eu

## ABSTRACT

In 2015, Cdiscount challenged the community to predict the correct category of its products from some of their attributes such as their title, description, price or associated image. The candidates had access to the whole catalogue of active products as of May 2015, which accounts for about 15.8 millions items distributed over 5,789 categories, a subset of which served as testing set. The data suffers from inconsistencies typical of large, real-world databases and the distribution of categories is extremely uneven, thereby complicating the classification task. The five winning algorithms, selected amongst more than 3,500 contributions, are able to predict the correct category of 66–68% of the testing set’s products. Most of them are based on simple linear models such as logistic regressions, which suggests that preliminary steps such as text preprocessing, vectorization and data set rebalancing are more crucial than resorting to complex, non-linear models. In particular, the winning contributions all carefully cope with the strong imbalance of the categories, either through random sampling or sample weighting. A distinguishing feature of the two highest-scoring algorithms is their blending of large ensemble of models trained on random subsets of the data. The data set is released to the research and teaching communities, as we hope it will prove of valuable help to improve text and image-based classification algorithms in a context of very large number of classes.

**Keywords:** *classification, e-commerce, big data, public data set.*

## 1 Introduction

E-commerce companies have become major actors of the retail business over the past decade (Turban *et al.*, 2015). As the product catalog of the largest companies now routinely exceeds several millions of distinct items, a large part of which from third-party sellers, and users are less inclined to crawl through pages of results (Spink *et al.*, 2002), a salient yet increasingly tough need consists in filling correctly the products’ characteristics in order to efficiently guide the customers towards the products they desire. It is clear that purely manual procedures are precluded, so one must rely on algorithms based on the description or image of the products.

In 2015, the leading French e-commerce company Cdiscount challenged the community on the datascience.net platform on a simple, real-world question: how can one guess the category of a product from its description, its image and other available attributes? (See <https://www.datascience.net/fr/challenge/20/details>.) The participants had access to Cdiscount’s catalogue of active products, a subset of which had their category hidden to serve as testing set and evaluate the candidates’ algorithms, thus turning the problem into one of supervised classification. Cdiscount released the data set to the public to be used as a practical benchmark and encourage improvements over text and image-based classification algorithms.

The contest was held between May–August 2015 and attracted over 800 participants. In the present paper we describe the underlying data set (Section 2), the challenge and evaluation criteria (Section 3) and the solutions proposed by the winning candidates (Section 4).

Table 1 – *Fields of the data set and examples of products (with associated image). Note that the description can end with an ellipsis.*


Field		Examples
Id	13110226	15572267
Title	Samsung LE32C450	Whirlpool AWOD2850
Category	level 1	1000010900 – TV - vidéo - son
	level 2	1000011032 – TV
	level 3	1000011035 – Téléviseur LCD
Description	Téléviseur LCD 32" (82 cm) HD TV - Triple HDMI - Port USB multimédia - Résolution: 1366 x 768 - Contraste dynamique - Sublimateur de couleur - Dolb...	1000003564 – Electroménager 1000003786 – Gros appareil lavage-séchage 1000003789 – Lave-linge Lave-Linge 8.5 kg - Classe énergétique : A++ - Consommation d'énergie : 240 kWh/an - Consommation d'eau : 10800 Litres/an - Classe d'efficacité à l'essorage: B - 1200 tours/min.
Brand	Samsung	Whirlpool
Seller	Third party	Cdiscount
Price	389.99€	306.49€
Associated image	–	

Table 2 – *Key numbers on the data set.*

15,821,950	products
791,453	products sold by Cdiscount
15,030,497	products sold by third-parties
52	distinct level 1 categories
536	distinct level 2 categories
5,789	distinct level 3 categories
27,982	distinct brands

## 2 Data set

The data set consists of about 15.8 millions of products, which represents virtually the whole catalogue of Cdiscount as of May 2015. Each product is associated with a unique identifier, a three-level category, a title, a description, a brand, a seller (Cdiscount or third party) and a price (Table 1). Some products, owned and sold directly by Cdiscount itself, are also provided with a representative image in jpeg format as additional information.

The total volume of text and image data is about 4 Gb and 1 Gb, respectively. As described hereafter, the data suffers from flaws and inconsistencies typical of large databases involving strong user interaction. Products do not necessarily have a brand, and their description is sometimes cut off, ending in this case with an ellipsis. The price is set to  $-1$  for out of stock products, and can take unrealistically large values. More importantly, the category filled by third-party sellers is not as reliable as that of Cdiscount's products.

As a consequence, the vast majority of the populated categories are not strongly reliable, third-party sellers accounting for almost 95% of the database (Table 2). As can be expected, the  $\sim 5,800$  available categories are strongly unevenly distributed amongst the products: the distribution of the number of products per category approximately follows a power law, which exhibits a long tail of categories containing a large number of products (Fig. 1a). As a matter of fact, about 700 categories hold 90% of the products (Fig. 1b) and the largest one – smartphone covers – contains more than two millions items. Similar trends are observed for the distribution of the  $\sim 28,000$  brands (Fig. 1c) and of the descriptions' vocabulary (Fig. 1d) amongst the products.

It is interesting to focus on the distribution of the attributes amongst the categories (rather than the products), since the former shall be used to predict the latter. Fig. 2 shows that the distribution of the brands and of the vocabulary amongst the categories again approximately follow power laws. In other words, the distributions are characterized by long tails of brands and words that appear in many different categories: unsurprisingly, when taken individually, most of them are uninformative with respect to the product's category.

This section thus illustrates the kind of pitfalls and difficulties encountered when dealing with a large, real-world e-commerce data set of products. In particular, algorithms designed to predict the category have to cope with the strong unevenness of the distribution of the attributes amongst the products and categories as illustrated in Figs 1-2.

### 3 Description of the challenge

In 2015, Cdiscount offered a simple challenge on the datascience.net platform based on the data set described in the previous section: given a list of product attributes (title, description, brand, seller and price, see Table 1), what is its correct category? (See <https://www.datascience.net/fr/challenge/20/details>.) A subset of 35,065 products, the category of which was hidden, served to evaluate the prediction algorithms proposed by the candidates. This testing set was built internally by the data scientists of Cdiscount, in order to ensure the procedure to be free from selection bias. Within each category, we selected at random samples amongst products sold by Cdiscount, as the category filled by third-party sellers is considered as not reliable: we believed that the evaluation would be more faithful by doing so (however this fact was not disclosed to the participants). As an evaluation metric, we simply used the proportion of correct predictions:

$$\text{score} = \frac{1}{N} \sum_{i=1}^N \begin{cases} 1 & \text{if } \hat{c}_i = c_i \\ 0 & \text{else} \end{cases}, \quad (1)$$

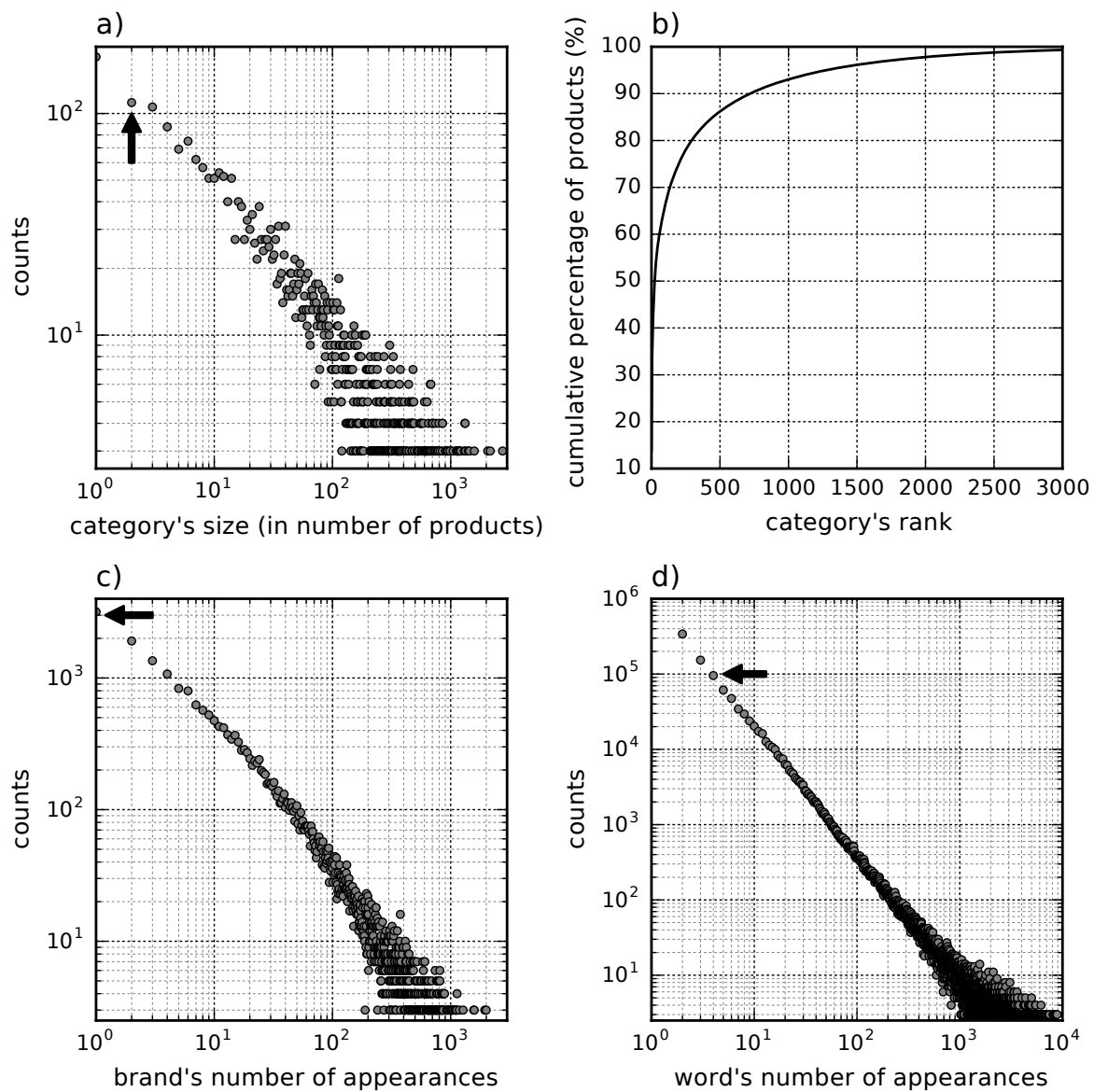


Figure 1 – (a) Size distribution of the categories: size corresponds to the number of products belonging to a category (e.g., the point shown by an arrow indicates that there are slightly more than 100 categories which contain only 2 products). (b) Cumulative percentage of products held by the categories, sorted by size (largest categories first). (c) Recurrence distribution of the brands: recurrence corresponds to the number of products associated with a brand (e.g., the point shown by an arrow indicates that more than 3,000 brands are represented by a single product in the catalogue). (d) Recurrence distribution of the words of the vocabulary used in descriptions: recurrence corresponds to the number of products wherein a word of the vocabulary appears (e.g., the point shown by an arrow indicates that about  $10^5$  words of the vocabulary appear in exactly 4 distinct products).

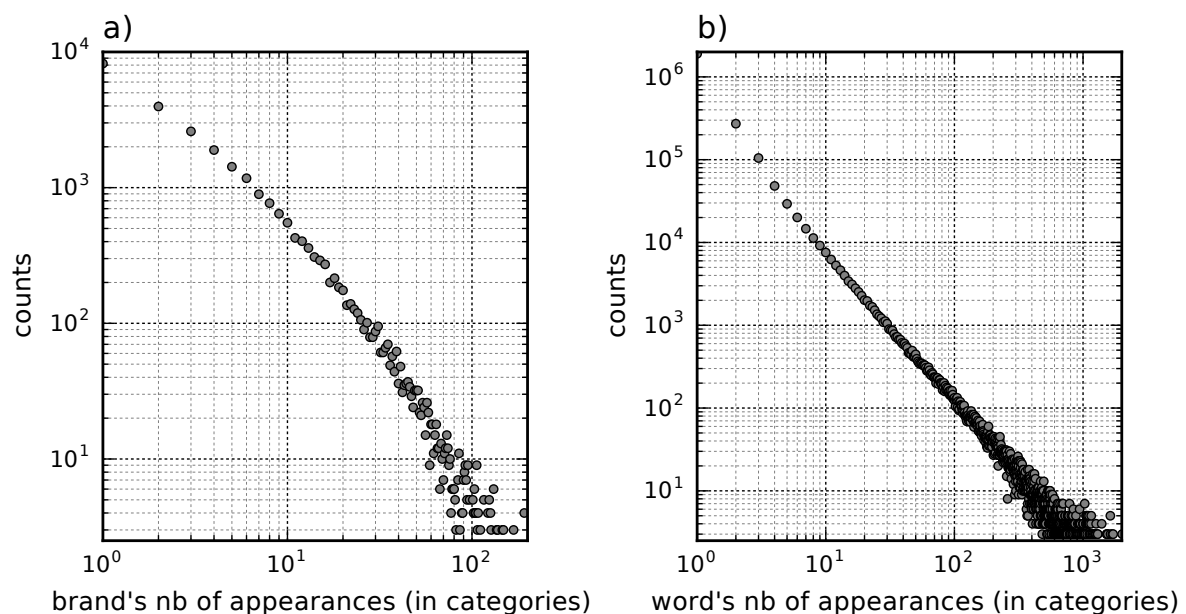


Figure 2 – (a) Recurrence distribution of the brands: recurrence corresponds to the number of distinct categories containing at least one product associated with a given brand. (b) Recurrence distribution of the words of the vocabulary used in product descriptions: recurrence corresponds to the number of distinct categories containing at least one product wherein a given word of the vocabulary appears.

with  $N$  the size of the testing set and  $\hat{c}_i, c_i$  the predicted and correct category of product  $i$ , respectively. Note that, in order to build up a testing set not too biased towards the most popular categories, no more than 20 products may belong to the same category. The resulting distribution of categories amongst the products (Fig. 3) consequently strongly differs from that of the whole data set (Fig. 1a). Although this was not mentioned on the website of the competition, the participants soon realized the difference in the distributions by trial and error, and consequently resorted to rebalancing techniques (see next section).

The participants were able to submit their predictions and get the evaluation score and ranking in real-time. In order to avoid over-fitting the testing set, (1) a limited number of submissions were allowed per day and (2) the real-time, public scores were calculated on only half of the testing set. The final evaluation, based on the whole testing set, did not alter significantly the scores and ranking, which suggests that the submitted algorithms indeed avoided over-fitting the testing set.

The challenge attracted 838 participants, mostly from France. We know little about their sociological characteristics, but we believe that most were students or junior data scientists. The candidates submitted a total of 3,533 contributions. The five highest-scoring submissions were sent to a jury, which made the final ranking based on the score, quality and originality of the proposed solutions. The following section briefly describes the winning contributions, which received money prizes between 500€ and 9,000€.



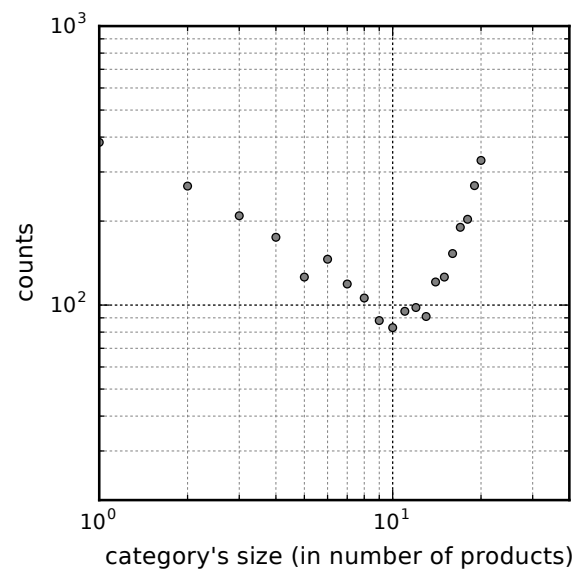
Y. Jiao *et al.*

Figure 3 – *Size distribution of the categories in the testing set: size corresponds to the number of products belonging to a category.*

Table 3 – *Summary of the winning contributions.*

Rank	Score	Language/library	Method(s)
#1	68.3%	Python/scikit-learn	Logistic regression with stochastic gradient descent + multinomial naive Bayes + passive aggressive classifier
#2	68.0%	Python/scikit-learn	Two-stage logistic regression
#3	66.9%	Python, R, Vowpal Wabbit	Linear classifier with square loss function
#4	66.3%	Python/PIL, C++, Dataiku	Logistic regression
#5	66.3%	R/ConText	Three-stage convolutional neural network

## 4 Analysis of the winning contributions

The winning algorithms, mostly coded in Python or R, are able to predict the correct category of 66–68% of the products in the testing set (Table 3). The four best algorithms use linear models, mostly with a logistic loss function (Walker and Duncan, 1967). Interestingly, the square loss function also gives good results (contribution #3), although it is known to lack robustness against outliers. Two other linear classifiers appear in the winning contribution, namely, the passive aggressive classifier (Crammer *et al.*, 2006) and the naive Bayes method (Zhang, 2004) with multinomial distribution of the features. The only non-linear model is the convolutional neural network (Johnson and Zhang, 2015b, 2015a), which is used by candidate #5.

## 4.1 On input features

All the candidates concatenate at least the title, brand and description of the products to build input features. Candidate #2 applies larger weights to the title and the brand. Because of the large range of values it takes and the errors it contains, the price seems more delicate to include, but it nevertheless appears as input in two contributions (#1 and #3). In order to tackle the above-mentioned issues, the winner considers that values above 10,000 are wrong and divides them by 1000, and uses as input the interval to which the price belongs, which takes a limited number of values. Only one candidate (#4) fully integrates the images, by associating with each product the category of the three nearest neighbours of its image, weighted by their inverse distance, as an additional input feature. Curiously, candidate #1 finds that simply appending a piece of text describing the image's geometry (rectangular or not rectangular) significantly improves the categorization of books.

## 4.2 On preprocessing and the dirty details

As can be expected when dealing with large chunks of text data with potential inconsistencies, the candidates have to apply a variety of preprocessing techniques before the vectorization step. These usually include lower-casing, removal of stop words, conversion to plain ASCII text and word stemming. Some candidates additionally remove numbers, or replace them with generic strings such as `|NUMBER|` or `|DIGIT|`. Candidate #1 also prepends the preposition “for” (in French, “pour”) to every following word in sentences where this preposition appears, in order to better differentiate accessories from the products to which they are associated. For example, for the product #1963634 whose title starts with `BASEUS CABLE LIGHTNING FORMAT CLE USB POUR IPHONE IPAD IPOD`, the tokens `POUR_IPHONE`, `POUR_IPAD` and `POUR_IPOD` are appended to the text. Candidate #4 applies the same technique to a wider set of prepositions for the same reason, and adds the token `START_BY_<WORD>` to the text, where `<WORD>` is the first meaningful word of the title or of the description (i.e. not the brand, not a number...): the rationale behind this processing is that the beginning of the text often allows guessing the product's category. For example, for the product #5360298 whose title starts with `DRONE X46 2,4GHZ`, the token `START_BY_DRONE` is appended to the text.

## 4.3 On representation

The vectorization step is then most often realized with the tf-idf statistic (Spärck Jones, 1972), wherein the concatenated text associated with a product is converted to a vector whose  $i^{\text{th}}$  element is proportional to the number of appearances of the  $i^{\text{th}}$  token of the corpus within the product's text, and offset by the frequency of the token in the corpus. Depending on the algorithms, tokens can be words (unigrams) or sequences of two consecutive words (bigrams). A simple word count is also applied to some of the models in contribution #1. Candidate #5 takes on a different approach that partly preserves the order of the words, wherein the text is split in successive regions of 15–20 contiguous words, and a word count is applied to each region.



## 4.4 On reweighting

In order to cope with the unevenness of the distribution of the categories outlined in section 2, most of the candidates resort to some form of stratified sampling (Cochran, 1953): in other words, subsets of the catalogue are randomly selected as training sets, with a limit of a few hundreds products per category and with replacement oversampling for underrepresented categories. Candidates #1 and #2 repeat this subsetting procedure several thousands of times and blend the predictions from the resulting ensemble of models, which we assume to be a key ingredient to their success. The winning algorithm actually goes a step further by (1) random parametrising several processing steps applied to the subsets (e.g., tf-idf or word count, word stemming or not, unigrams or bigrams. . .) and (2) including three families of classifiers in the ensemble of models (Table 3). Only candidate #3 chooses not to re-sample the catalog of products, but rather assigns them weights inversely proportional to the frequency of appearance of their categories.

As for the better reliability of the category of Cdiscount products (section 2), it is an information only two candidates take advantage of (#1 and #5): the former candidate specifically trains models on Cdiscount or third-party products and assigns them different weights in the final blend; the latter candidate explicitly gives priority to Cdiscount products in the stratified sampling step described in the previous paragraph.

## 4.5 On the use of hierarchy

Finally, candidates #2 and #5 use the three-level hierarchical structure of the categorization (see Table 1) to reduce ambiguity between categories belonging to different branches, by performing classifications by stage. The idea is to successively predict the category across the levels of the hierarchical tree from top to bottom. Candidate #2 trains logistic classifiers to get the probabilities of belonging to the first-level categories,  $P(\text{product} \in \text{cat}_1)$ , and the conditional probabilities of belonging to the third-level ones,  $P(\text{product} \in \text{cat}_3 \mid \text{product} \in \text{cat}_1)$ , then applies the classical chain rule to estimate the desired marginal probabilities  $P(\text{product} \in \text{cat}_3)$  as:

$$P(\text{product} \in \text{cat}_3 \mid \text{product} \in \text{cat}_1) \cdot P(\text{product} \in \text{cat}_1). \quad (2)$$

Candidate #5 goes through the three levels of categories and trains one neural network per category of a given level to predict the category of the next level.

## 5 Teaching materials

We believe that the data set released to the community and the highlights of the distinguishing features of the winning contributions will prove of valuable help not only to the scientific research community, but also to statistics educators. Students should indeed face more often real-world, unclean data sets (Mandran and Stoltz, 2017) which are not abundant, particularly in a context of extremely large number of classes. As a matter of fact, it is clear from the winning contributions how crucial cleaning and preprocessing steps are: these encompass lower-casing, ASCII-ization, processing of stop words and

prepositions, stemming, detection of outliers, rebalancing... It is also obvious that, in addition to strong statistical skills, hands-on coding capabilities in a high-level, dynamic language such as Python or R are essential to perform well within a data-based competition, in order to efficiently work out those processing steps and quickly extract meaningful insights from the data.

In order to provide a starting point to educators and students, a basic code is provided as accompanying material to this article, which shows how to load and manipulate the data set. A simple nearest neighbour classifier is also included, which achieves a precision of 48% on the testing set. Then, the ensemble of tricks employed by the winners and summarized in the previous section should provide guidelines to experiment optimization steps from that sample code, e.g., in a context of practical work. Specifically, students could start exploring the effect of taking stop words in to account, using  $n$ -grams of size larger than 1 or setting up logistic regressions (as many winners did) in place of the nearest neighbour classifier.

## 6 Conclusion

In this paper we gave statistical insights into the catalog of products of Cdiscount in order to highlight the kind of pitfalls and difficulties a classification algorithm applied to a real-world data set has to cope with. Specifically, the potential inconsistencies of the products' attributes, the varying reliability of the data, the large number of categories and the extreme imbalance of their distribution obviously complicate the classification task.

The five winning contributions of the datascience.net challenge are able to predict the correct category of 66–68% of the products in the testing set. Most of the algorithms are based on simple linear classifiers; in particular, the logistic regression appears in three contributions. When dealing with noisy, imperfect data, the preliminary processing steps thus appear to be more crucial than the choice of a complex, non-linear classifier. Another factor could be the bad scalability of such algorithms with respect to the number of classes, which is extremely large in our case. Preprocessing steps include text processing, vectorization and rebalancing of the training data. The last point is particularly salient and was tackled by all the winning candidates, either through random stratified sampling to set up balanced training sets or by weighting training samples by the inverse of their categories' frequency of appearance. A distinguishing feature of the two most accurate algorithms is their training of ensemble of thousands of models on random subsets of the data, whose predictions are then averaged to get the final predicted category: we thus assume this to be a key ingredient to their success.

The whole data set is released to the public. The availability of a large, real-world catalogue of products with associated images and text attributes, together with benchmark results from the most accurate models to date, should prove of valuable help to the scientific community in order to improve over existing text and image-based classification algorithms in a context of very large number of classes.

## Supplementary material

The data set and sample code described in this article are released to the public and can be obtained by contacting any of the authors affiliated with Cdiscount. Alternatively, the following mailing list may be used: [datascience@cdiscout.com](mailto:datascience@cdiscout.com).

## Acknowledgements

We thank the datascience.net team for hosting and helping us organize the 2015 categorization challenge. Two anonymous reviewers provided constructive comments which helped improve and clarify this manuscript.

## References

- [1] Cochran, W.G. (1953), *Sampling techniques*, John Wiley, Oxford.
- [2] Crammer, K., O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer (2006), Online passive-aggressive algorithms, *Journal of Machine Learning Research*, 7, 551-585.
- [3] Johnson, R. and T. Zhang (2015a), Semi-supervised convolutional neural networks for text categorization via region embedding. In *Advances in Neural Information Processing Systems*, 919-927.
- [4] Johnson, R. and T. Zhang (2015b), Effective use of word order for text categorization with convolutional neural networks. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL'15)*, number 2011, 103-112.
- [5] Mandran, N. et G. Stoltz (2017), Entretien avec Nadine Mandran : comment enseigner la pratique métier ?, *Statistique et Enseignement*, 8(1), 45-57.
- [6] Spärk Jones, K. (1972), A statistical interpretation of term specificity and its application in retrieval, *Journal of Documentation*, 28(1), 11-21.
- [7] Spink, A., B.J. Jansen, D. Wolfram, and T. Saracevic (2002), From e-sex to e-commerce: Web search changes, *IEEE Computer*, 35(3), 107-109.
- [8] Turban, E., D. King, J. Lee, T. Liang, and D. Turban (2015), *Electronic commerce: A managerial and social networks perspective*, Springer.
- [9] Walker, S.H. and D.B. Duncan (1967), Estimation of the probability of an event as a function of several independent variables, *Biometrika*, 54(1867), 167-179.
- [10] Zhang, H. (2004), The optimality of naive Bayes. In *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference*, Miami Beach.