

Cdiscount.com

Catégorisation de produits pour le e-commerce

Créer un système automatique de catégorisation des produits à partir de leur description.

Classement

- | | | |
|---|-------------------------------|------------------------|
|  | 1. ➡ (1) Romain Ayres | Score 68,32739% |
|  | 2. ➡ (2) Nicolas Gaude | Score 67,96521% |
|  | 3. ➡ (3) Arnaud de Myttenaere | Score 66,94710% |

[Voir tout le classement](#)

Résumé

Télécharger

Mes contributions

Discuter

Ce projet est terminé.



15 000 €
à gagner



3533
contributions



838
participants



terminé

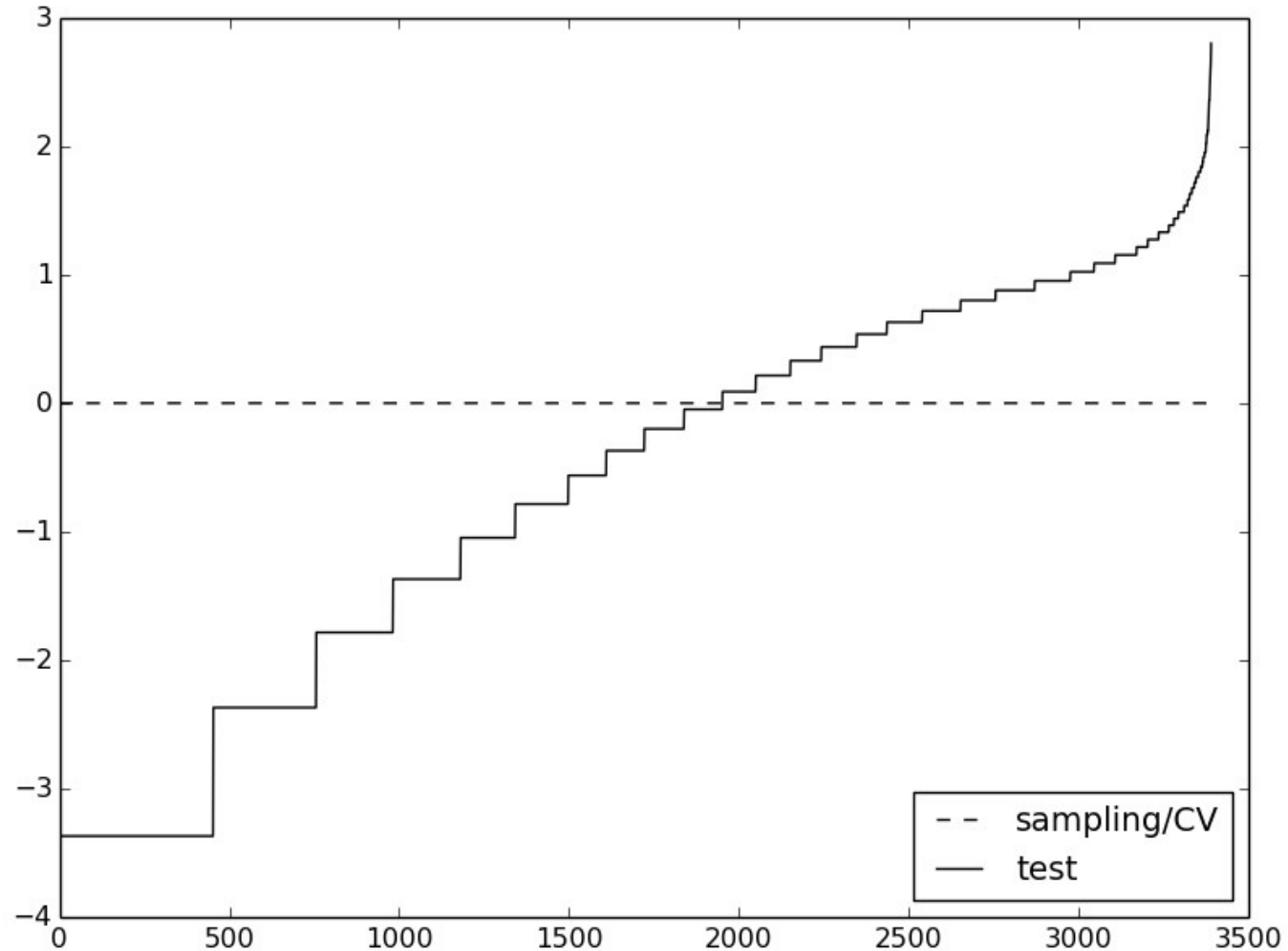


/me

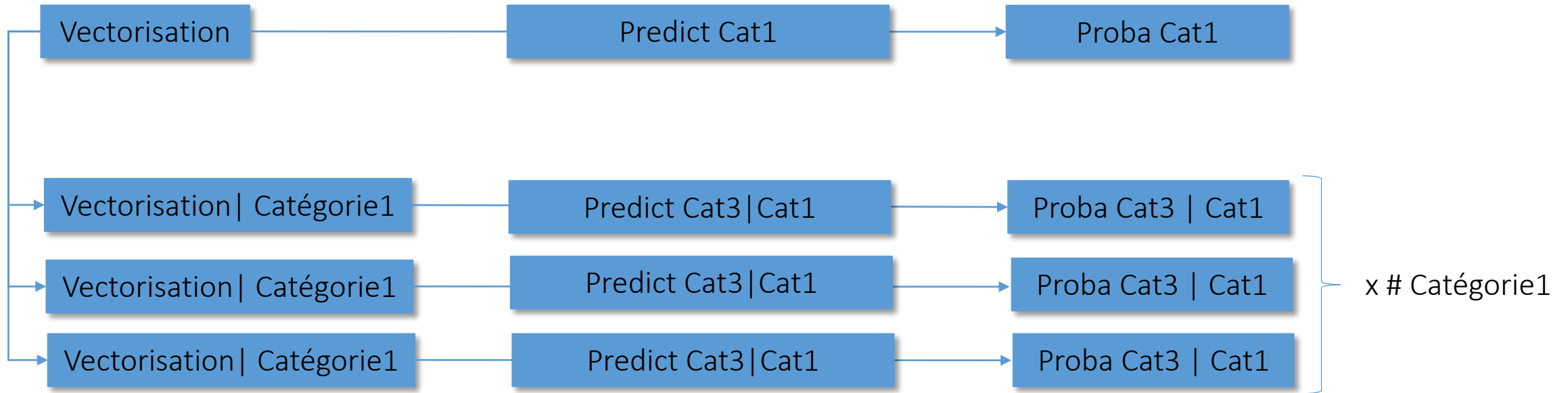
You Know Nothing : « balanced » sampling

Naive prediction « autres livres »

- score_train = 6,5%
- Score_test = 0,02865% (!)



Base Model : Staged Logistic Regression (66,3%)



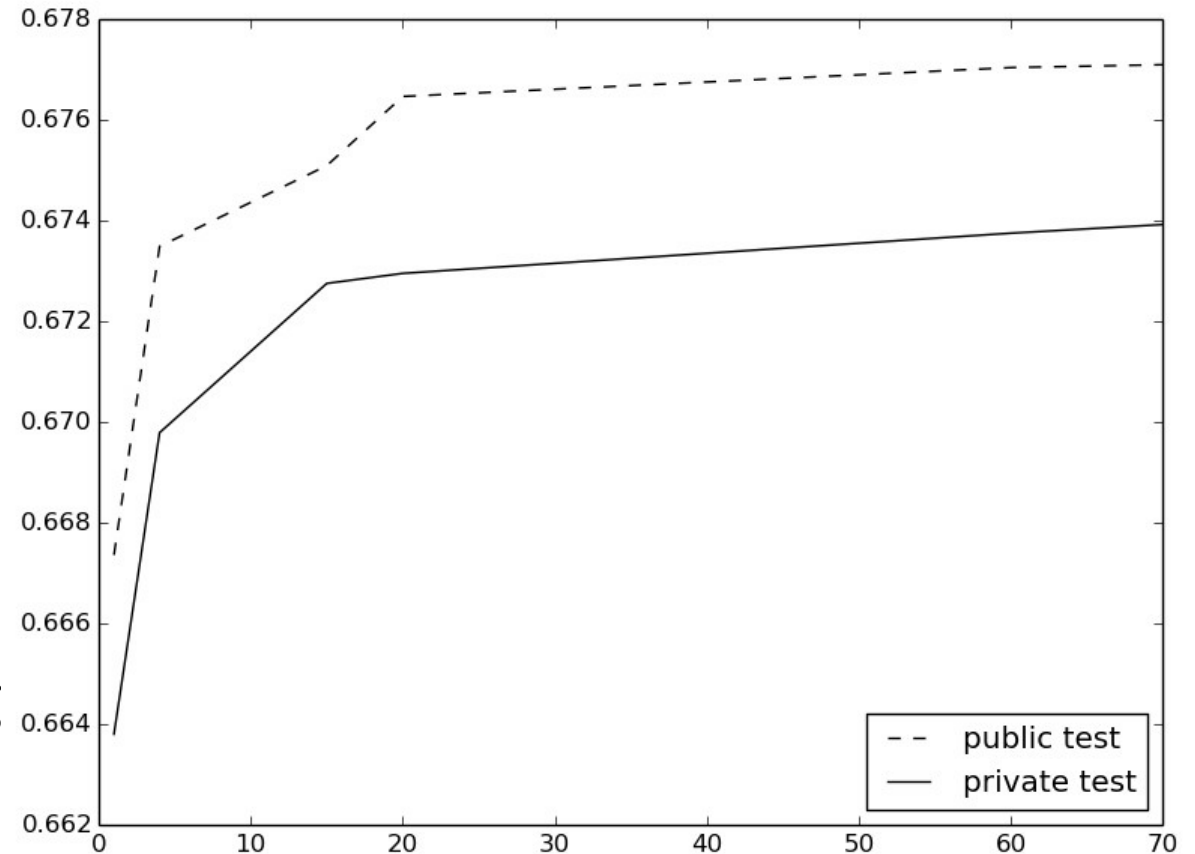
$$\text{Proba Cat3 | Cat1} \cdot \text{Proba Cat1} = \underbrace{\text{Proba Cat1 | Cat3} \cdot \text{Proba Cat3}}$$

1 si Cat1 inclut Cat3

0 si Cat1 n'inclut pas Cat3

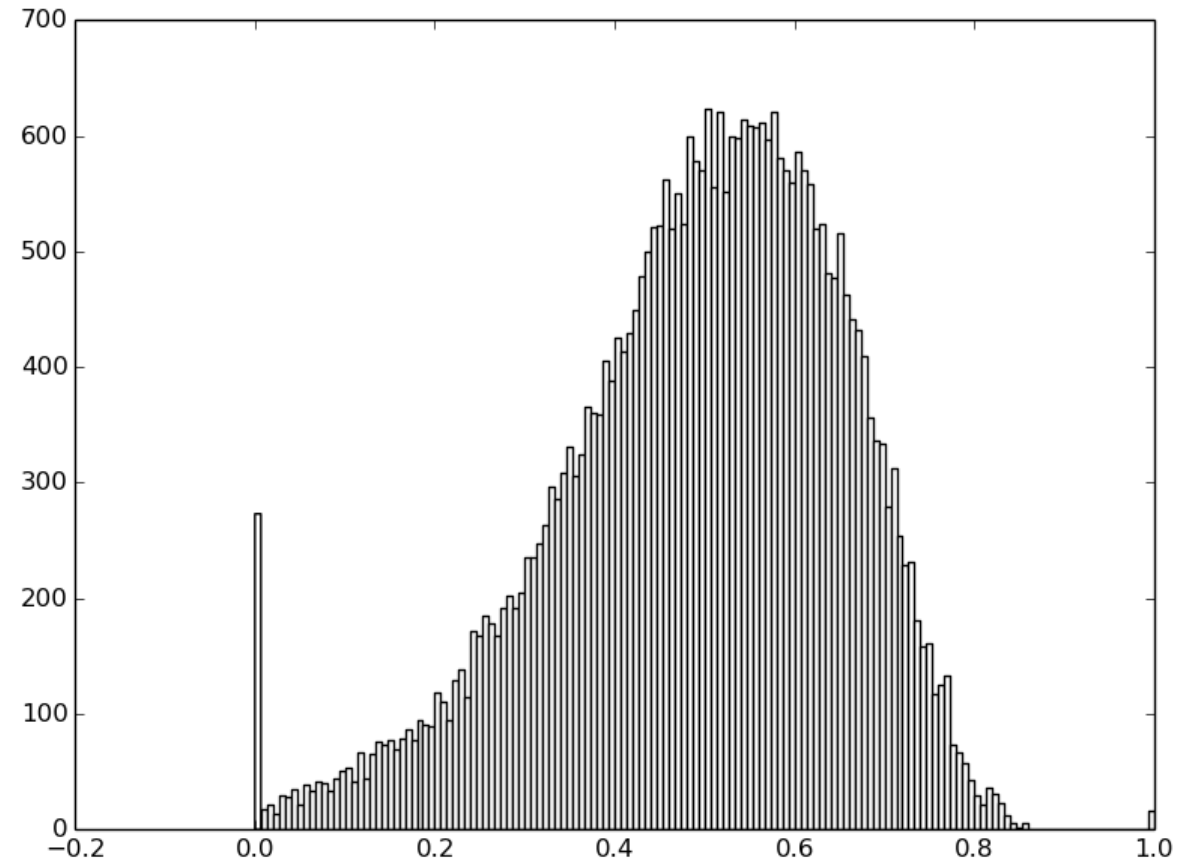
Basic Ensemble : MLE vote (+1,1%)

- Train a lot of base model (~70)
- Mean all proba matrix (35K*6K)
- Guess the cat3 as $\text{argmax}(\text{proba})$
- Next time : Use state of art **Stacking Generalized !** (and xgboost btw)



Duplicate Items (+0.3%)

- Approximate string matching
- Based on regularized **levenshtien** distance : # insert/delete/mutation
- Performance $O(\#items \cdot lenstr \cdot \text{depth index})$: 1000 searches \sim 1M items /mn (low-end cpu and 1GB ram only)



Evidence Rules (+0.3%)

Aka « Fontaine à chocolat » effect

- Fact : Cdiscount had good time building the Cat3 | Cat2 | Cat1 hierarchy
- Guess : Cdiscount used automatic feature word extraction with Decision Tree

- Idea :
$$P_{tp} = P(\text{Categorie3_Name in Libelle, Categorie3} \mid \text{Categorie1})$$
$$P_{fp} = P(\text{Categorie3_Name in Libelle, not Categorie3} \mid \text{Categorie1})$$

If $P_{tp} / (P_{tp} + P_{fp}) \gg \text{score}$, then you might have a guess.