datascience.net

/me

15 000 € à gagner

3533 contributions

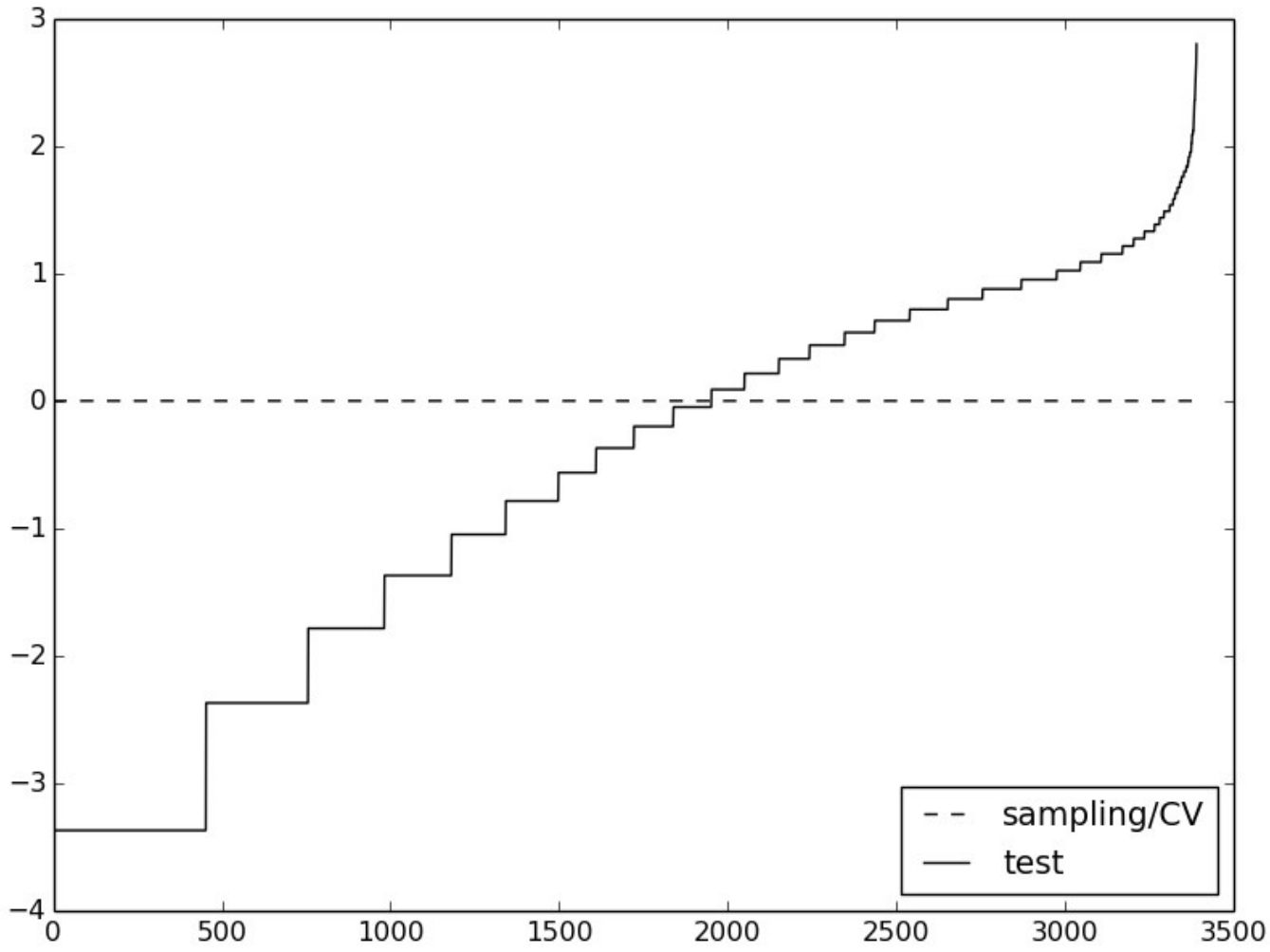
838 participants

terminé

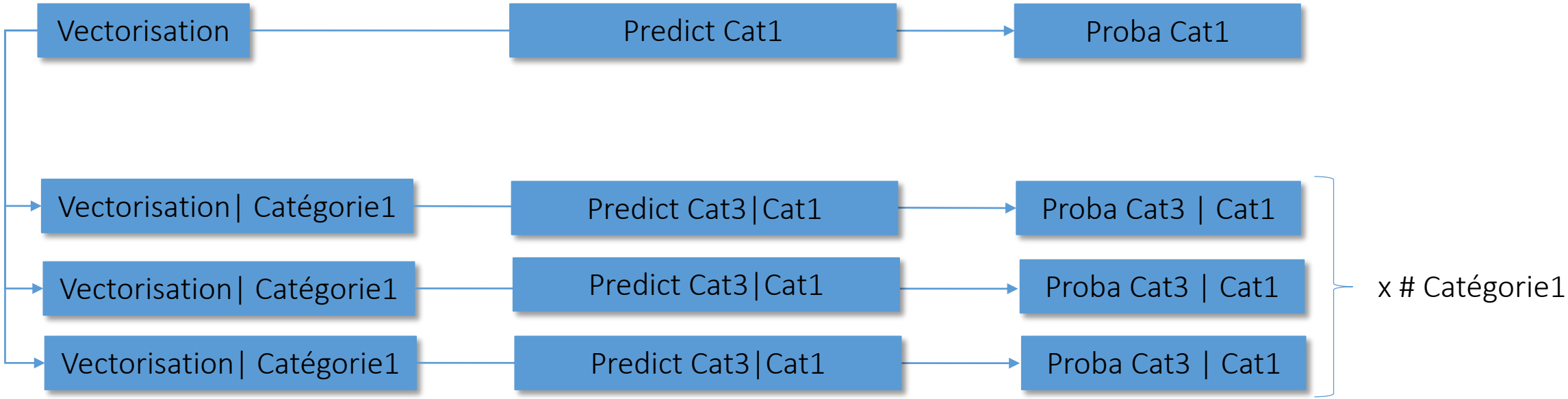
You Know Nothing : « balanced » sampling

Naive prediction « autres livres »

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



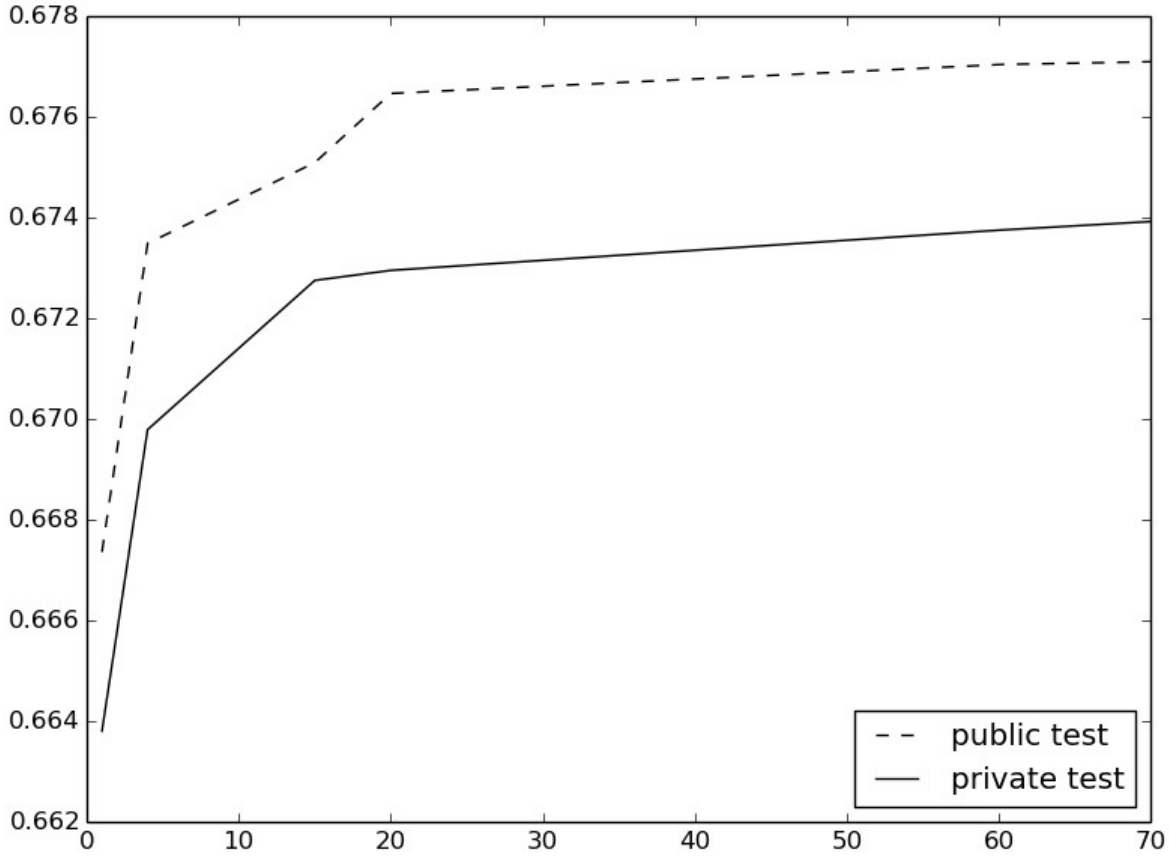
Base Model : Staged Logistic Regression (66,3%)



$$\text{Proba Cat3}|\text{Cat1} \cdot \text{Proba Cat1} = \underbrace{\text{Proba Cat1}|\text{Cat3}}_{\text{1 si Cat1 inclut Cat3}\\ \text{0 si Cat1 n'inclut pas Cat3}} \cdot \text{Proba Cat3}$$

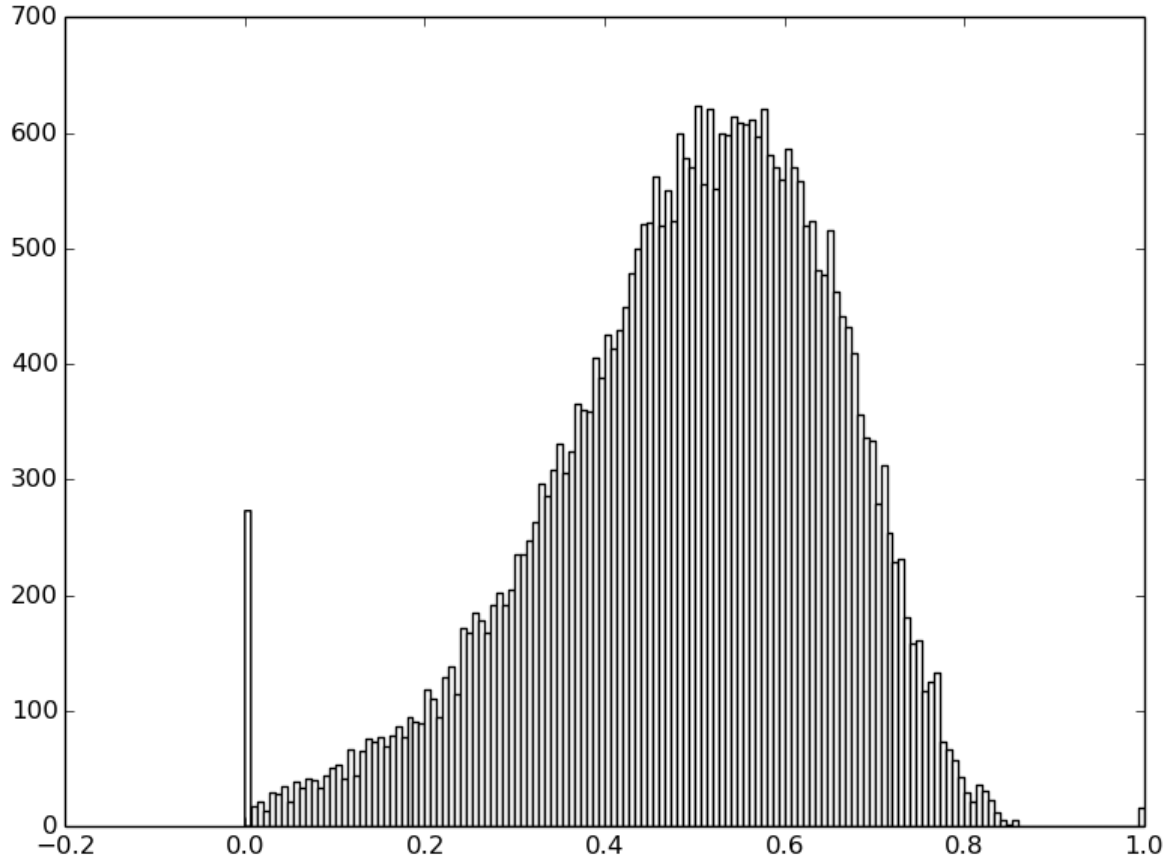
Basic Ensemble : MLE vote (+1,1%)

- Train a lot of base model (~70)
- Mean all proba matrix (35K*6K)
- Guess the cat3 as argmax(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 . lenstr . depth index) : 1000 searches ~ 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 :

```
Ptp=P(Categorie3_Name in Libelle, Categorie3 |Categorie1)
Pfp=P(Categorie3_Name in Libelle, not Categorie3 |Categorie1)
```

If Ptp/(Ptp+Pfp) >> score, then you might have a guess.