M2 AIC

TC1: Apprentissage

Otto Group Product Classification Challenge

Groupe: Robin DURAZ, Jiaxin GAO

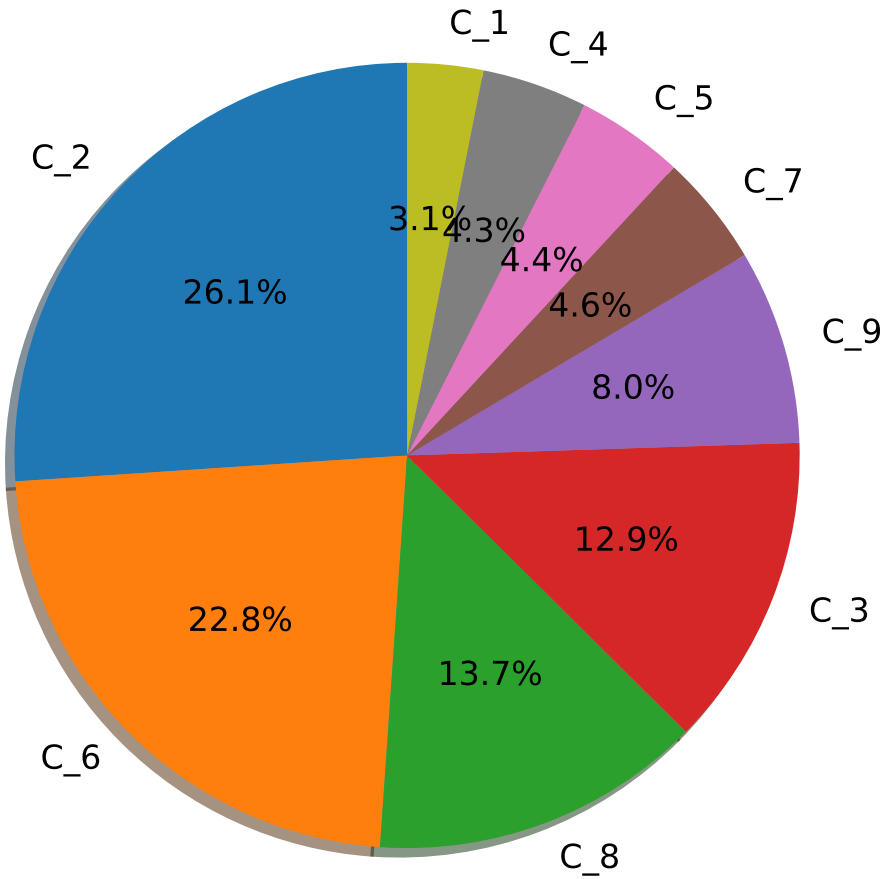
1. Présentation du challenge
2. Les données
   - Analyse des données
   - Prétraitement des données
3. Test modèles individuels
   - Modèles testés
   - Modèles éliminés
   - Processus de tuning
4. Moyenne des modèles (Ensemble averaging)
   - Modèles deux à deux
   - Plus de deux modèles
5. Analyse des résultats - Conclusion

Présentation du challenge

- Challenge sur Kaggle: Une grande communauté proposant des challenges en Machine Learning provenant d'entreprises.
- Otto Group Challenge: Construire un modèle prédictif capable de distinguer des produits en différentes catégories.

Analyse des données

La répartition des données dans les classes est assez inégale



Analyse des données

- Chaque produit dispose de 93 features, ayant toutes des valeurs entières positives.
- Données un peu "sparses" (environ 80% de 0).

Prétraitement des données

- Séparation des données en ensembles de train et de validation
- Standardisation des données

Nous avons testé les modèles suivants, venant tous de la librairie sklearn :

- SVM
- Random Forest
- Multi Layer Perceptron
- XGBoost Classifier
- AdaBoost Classifier
- Extra Trees

Au final, nous avons décidé d'éliminer :

- SVM car trop lent, et moins bon que les autres.
- AdaBoost car trop lent pour une performance correcte.

Tuning sur les modèles gardés pour obtenir les meilleurs performances possibles.
Grid search pour chercher de bonnes valeurs de paramètres.

Modèle	Random Forest	XGBoost	MLP	Extra Trees
Performance	0.475	0.452	0.489	0.463

Table: Score par modèle individuel

Pour encore améliorer nos résultats, nous avons essayé "l'ensemble averaging".
Nous avons donc moyenné les prédictions de nos modèles avec des poids.

Nous avons effectué du grid search pour trouver de bons poids pour moyenner les probabilités.

Combinaison		Performance
Random Forest	XGBoost	0.447
	MLP	0.469
	ET	0.482
XGBoost	MLP	0.45
	ET	0.443
MLP	ET	0.46

Table: Meilleurs résultats avec combinaisons de deux modèles (sur validation set)

Meilleur résultat global avec une combinaison de Random Forest, XGBoost et MLP (0.439)
Essais avec combinaison de 4 modèles globalement décevants, jamais mieux qu'avec 3 modèles.

Pour finir, nous pensons pouvoir améliorer nos résultats avec :

- De meilleurs modèles que ceux de sklearn.
- Des changements au niveau des features.
  - Construction de features.
  - Changements de représentation, comme par exemple prendre des sorties de classifieurs.