



Introduction à l'Apprentissage Statistique

Food for Thought

M2 Actuariat – ISFA – 2021/2022

Pierrick Piette
Actuaire à Seyna
pierrick.piette@gmail.com

Seyna.

In God we trust, all others bring data
- William Edwards Deming





Unbalanced data

Sampling

Undersampling



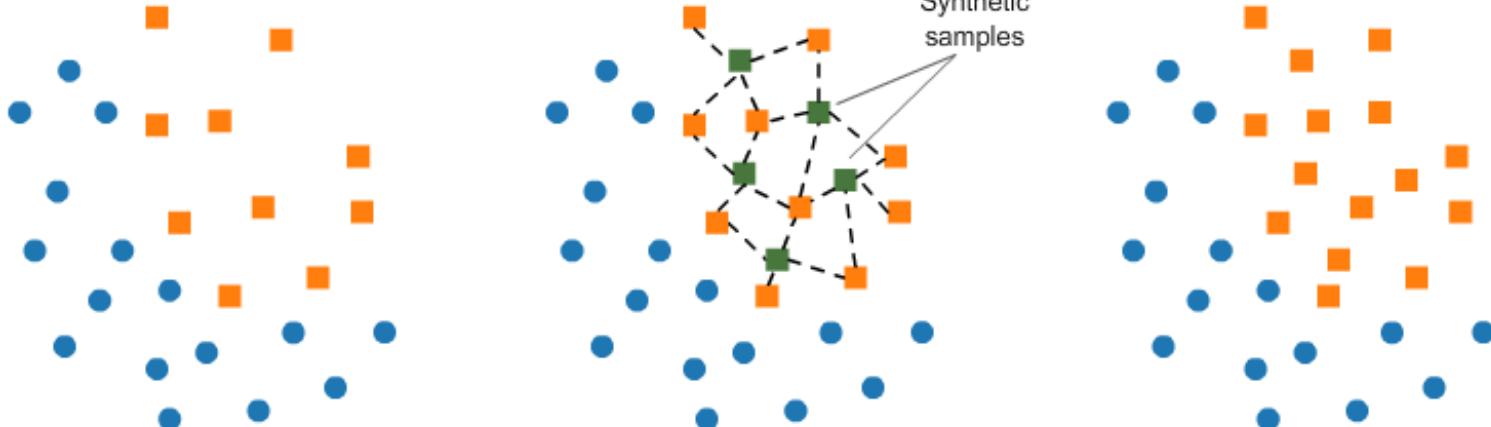
Oversampling



SMOTE

Synthetic Minority Oversampling Technique

- Créer de nouveaux points de la classe minoritaire
- k plus proches voisins sur la classe minoritaire
- Créer un point entre deux voisins



Changement de métrique

Choisir une métrique pénalisante

- Précision $\frac{TP}{TP+FP}$
- Rappel $\frac{TP}{TP+FN}$
- F1 score $2 \times \frac{Précision \times Rappel}{Précision + Rappel}$
- AUC

Choisir cette métrique pour

- le tuning
- la validation
- la fonction de perte (si possible)



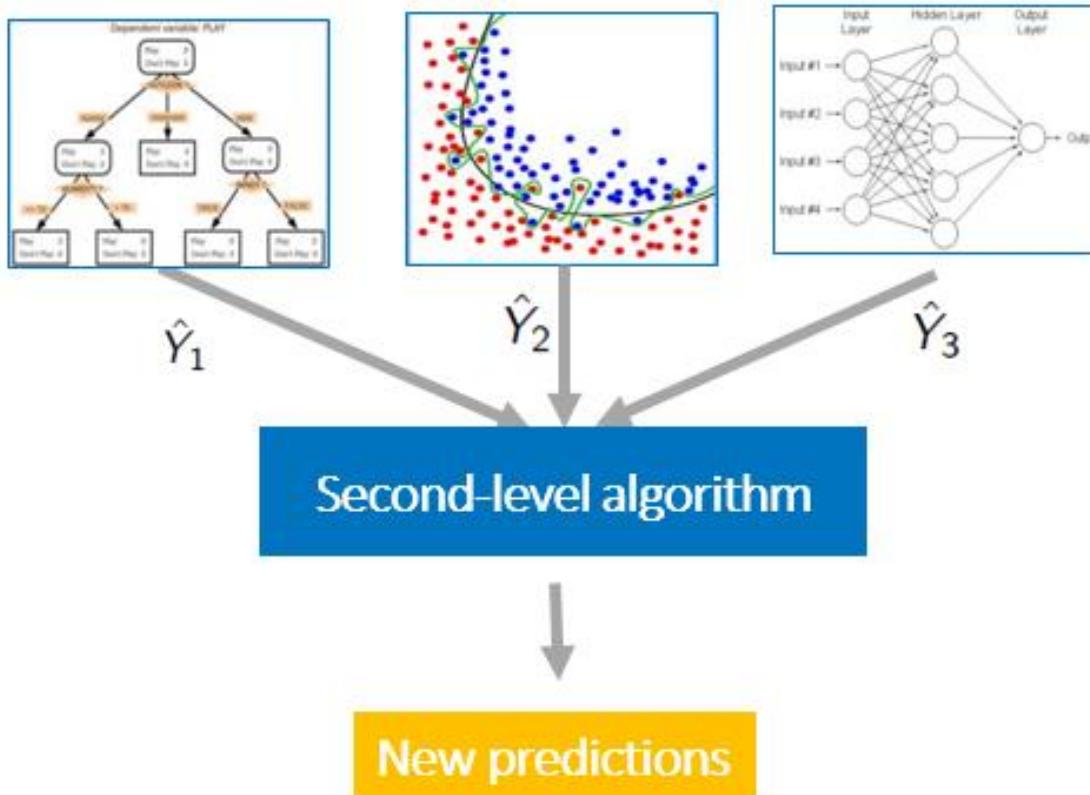
Changement de problématique





Stacking

« Meta » Ensemble Learning





Data science project

Step by step

Google's Rules of Machine Learning

Rule #1 : Don't be afraid to launch a product without machine learning





Compréhension du problème

Quel est le contexte ?

Quel est le but du projet ? (product manager)

Comment le succès est mesuré ?

Quels sont les rendus attendus ?

Quelles sont les contraintes opérationnelles ?

Data Management (1/2)

Data engineering

- Récupération des données
- Organisation des données en bases (data engineer)
- **Dictionnaire de données**

Data cleaning

- Données manquantes
- Données aberrantes
- Attention aux données catégorielles
- Importance du contexte de l'étude

Data sets
in tutorials



Data sets in
the wild



Data Exploration

Statistiques descriptives univariées

- Exploration graphique avant tout
- Variables numériques : Scatter plot
- Variables catégorielles : Box plot

Statistiques multivariées

- Corrélogramme
- Corrélation linéaire et non linéaire
- Heatmap

Data Management (2/2)

Feature engineering

- Création de nouvelles variables
- Utilisation de l'expertise métier
- **Dictionnaire de données**

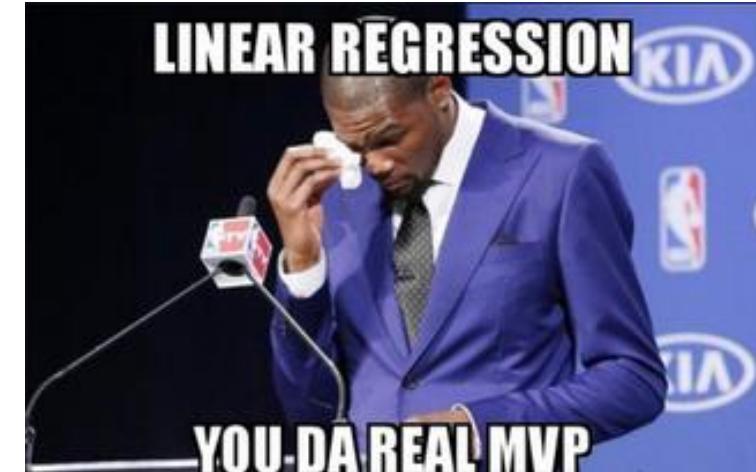
Data enrichment

- Utilisation de sources externes
- Open data ou autres SI internes à l'entreprise
- **Dictionnaire de données**

Baseline model

Choix d'un modèle (très) simple

- Régression linéaire / Régression logistique
- Potentiellement CART ou RF
- Attention à garder l'esprit apprentissage statistique



Benchmark

- Pouvoir de prédiction
- Feature Importance
- Premières interprétations et comparaison à la data exploration
- **Faut-il retourner au data management ?**

Modèles complexes (1/2)

Estimation des hyperparamètres

- Cross-validation sur l'ensemble de la base de train
- Utilisation de la data exploration
- Tuning avant la comparaison

Comparaison du pouvoir prédictif

- Cross-validation avec **les mêmes folds** pour tous les modèles
- Fonction de validation en fonction du contexte
- Garder le baseline model dans les comparaisons
- Boxplots sur l'erreur de prédiction si possible

Modèles complexes (2/2)

Comparaison des comportements

- Estimation des modèles sur l'ensemble de la base train
- Feature importances et Shapley values
- Garder le baseline model dans les comparaisons

Choix du modèle

- Pouvoir de prédiction
- Simplicité et interprétabilité
- **Choix en fonction du contexte**

Finalisation du modèle

Feature engineering

- En fonction des premiers résultats d'interprétation
- Forcer certaines variables avec l'expertise métier

Interprétabilité détaillée

- Documentation du modèle
- Compréhension du modèle : global et local
- **Est-ce que je peux l'expliquer au board et aux utilisateurs ?**

Déploiement en production

Développement opérationnel

- Crédit d'une application autour du modèle (software engineer)
- Mise en ligne du modèle (dev ops)

Phase de test

- Vérifier que le modèle fonctionne en monde réel
- A/B testing et mesure de la performance (data analyst)

Mise en production

- Youpi ! ☺
- Mais c'est le début d'autres problèmes...

ML in Production



Data Pipelines

Changement dans les feature encoders

- Changement d'écriture ou ajout de valeur
- Pour l'algorithme les valeurs deviennent **null**
- Dégradation des performances du modèle

Data Leakage

- La variable d'intérêt arrive dans les covariables
- Très bonnes performances en développement
- Mais très mauvais résultats en production

Interactions des modèles

Trop de features, tue les features

- Utiliser beaucoup de covariables améliore la performance
- Par contre, il est difficile et couteux de maintenir la pipeline
- A moyen terme, le modèle devient obsolète

Interaction avec d'autres modèles

- Dans un environnement tech, votre modèle n'est pas seul
- Regarder les performances dans l'ensemble de l'environnement
- Exemple : algorithmes de recommandation

Code et infrastructure chaotiques

Actuaire data scientist, mais pas forcément programmeur

- Pas forcément les « best practice » en code que les développeurs
- Pas les mêmes langages (R, Python, SQL, Java, Spark, ...) : difficile à maintenir
- Sans oublier les nombres magiques en dur
- « Data matures like wine, applications like fish »

Les infrastructures des environnements ne coopèrent pas

- Des mises à jour en dev peuvent avoir de très bons résultats
- Par contre en prod ça peut avoir des conséquences énormes
- Une fois que le système est en production, on ne joue plus au cowboy en dev

Welcome to the real world !

Cercle vicieux autoréalisateur dans le monde réel

- Les prédictions du modèle impacte le monde réel
- Le monde réel impacte les données du modèle, etc.
- Assez difficile à détecter et à contrôler

Pas un monde de bisounours

- Agents économiques adverses essayent de comprendre votre modèle...
- pour faire en sorte que ça leur soit profitable
- Y compris introduire de nouvelles données

