

(PRESQUE) 50 NUANCES DE RANDOM FOREST DANS UN CONTEXT BUSINESS

FILIPPO MAZZA
PHD

IDENTIFIEZ ET PRIORISEZ VOS CIBLES



sparklane

A quels principaux enjeux faisons nous face?

#1

Gagner du temps

Ciblez les comptes avec la meilleure probabilité de succès



Financial data analysis
« shallow » learning

#2

Augmenter la conversion

Utilisez les signaux d'affaires pour personnaliser votre message



News NLP analysis
shallow/deep learning

#3

Raccourcir le cycle de vente

Bénéficiez des alertes temps réel pour engager la relation au bon moment



Notre produit : Predict pour identifier les prospects à plus fort potentiel

← Liste 1

Créer le 20/12/2016 15:55

Modifier le 20/12/2016 15:55

LEAD SCORING

★★★★ Opportunités 3/3 Secours

★★★ Leads chauds 100/100 Secours

★★ Leads potentiels 6678 Secours

● Leads froids 0/0 Secours

critères

Tout effacer

NCS, Pensions sociales, Atlas

Département

Date de création

Activité

Forme juridique

Cotation bancaire

Effectifs

Effectifs en croissance

Chiffre d'affaires

Chiffre d'affaires en croissance

Chiffre d'affaires à long terme

Score Financier

Capital

Fonds de roulement

Bénéfice

Bénéfice en croissance

Canaux de vente

secours (16)

signal d'alarme (916)

download

COMPARER

IMPRIMER

EXPORTER

3 opportunités

	Rang	Score	Nom	Ville	APC	RCS	Téléphone	Site web
★★★★	100%	ORANGE	PARIS 15 - 75015	Télécommunications	381127666	0144442222		
★★★★	100%	ATOS INTEGRATION	BOULOGNE - 92100					
★★★★	47%	BOUYGUES	PARIS 8 - 75008					

100 leads chauds

	Rang	Score	Nom	Ville	APC	RCS	Téléphone	Site web
★★★★	100%	PNC AUTOMOBILES	SAINT GERMAIN - 92100					
★★★★	100%	INGENICO GROUP	PARIS 15 - 75015					
★★★★	47%	SE TECHNICS FRANCE SAS	ORLY - 91400					
★★★★	47%	SAS MARIGNAN ELYSEES	PARIS 8 - 75008					
★★★★	47%	GRUPE RH ENVIRONNEMENT	GENEVE - 1200					
★★★★	47%	EUROCLEAR FRANCE	PARIS 9 - 75009	Banque et assurance	342150066	0155345134		
★★★★	47%	ECONOCOM FRANCE	PUTEAUX - 59000	Banque et assurance	305364026	0345473000		
★★★★	47%	ALLEN	BOULOGNE BILLANCOURT - 92100	Informatique et électronique	348657417	014067220		
★★★★	47%	AUDY	SEVRES - 82300	Informatique et électronique	352405707	0144444538		
★★★★	47%	SAP FRANCE	LEVALLOIS PERRET - 92300	Informatique et électronique	379823994	0145177000		

EXEMPLE DE RECOMMANDATION

Par rapport à votre marché, la société Z a un potentiel de 70% de vous intéresser car elle ressemble à vos clients X et Y, et vient de lever des fonds

- ## + résultats

[illegible]

R&D

Panorama du « data lake » France

(on a aussi le UK)

Data lake Sparklane

~ 3 millions de sociétés

~ 11 millions d'établissements

News sociétés

~ 150 000 articles de presse / mois

Historique: ~ 8 millions news



Périmètre client

Périmètre marché

e.g. 10 000 sociétés

Leads

e.g. 100 sociétés



Données spécifiques
du client

Extrait des informations datalake Sparklane

Firmographiques

- Chiffres d'affaires
- Nombre d'employées
- Forme légale, année de création, ...
- Canal de vente
- Présence digitale (social nets, ...)
- ...

News

- Évènement de l'entreprise
- Date

Extraction et traitement à
part, pas ce soir 😊

La R&D chez Sparklane

J. Gosme, D. Cram et F. Mazza

Alexandre Verno
Chef produit



Traitement Automatique de la Langue

INPUT

Sources textuelles (presse, sites de recrutements, informations légales, réseaux sociaux, sites web ...)

OUTPUT

Signaux d'affaires, nomenclature contacts,
Catégorisation intitulés de poste
Amélioration identification entreprises
...



Modèles marchés clients



INPUT

Critères firmographiques des entreprises
(CA, effectif, activité, localisation ...),
signaux d'affaires, variables métier...

OUTPUT

Modèles marchés clients
Analyse anomalies
Imputation valeurs manquantes
...

Le **Machine Learning** est au cœur du produit

Données clients

« ONBOARDING »
*Variables métiers,
informations non
publiques, ...*

Signaux d'affaires



*News concernant la vie des entreprises,
leur évolution, leurs changements, ...
catégorisées et datées*

Données Firmographiques



*Données relatives à la vie
économique de l'entreprise, leur
secteur, leur marché*



Objectifs

du ML pour l'identification de leads

Business

- Modéliser les marchés de chaque client afin de prioriser ses cibles, sur la base de ses anciens clients / prospects
- Proposer des nouvelles cibles
- Expliquer les facteurs qui contribuent aux entreprises proposées
- S'adapter aux changements de la stratégie / du produit vendu / ...
- Suivre l'évolution des performances

R&D

- Supervised learning : probabilité de l'« intérêt d'un prospect »
- Calcul d'importance des variables
- Retraining / active learning
- Métriques (accuracy, ROC, ...)

Contraintes

-> limitations sur le choix des outils

Business

- Dans un temps « raisonnable »
- A partir d'un échantillon potentiellement restreint
- Avec beaucoup de variables et de valeurs pas toujours fiables
- ...



R&D

- Inference time: < 15 ms
 - Model update: < 2 min
- ➡ (cost bounded 😊)
- Nombre d'échantillons inconnus a priori
 - Petites listes au début 1-100 K échantillons
 - Déséquilibre classes : 1:20
 - Nombre de variables inconnu a priori, hétérogène
variables clients, variables payantes, ... en pratique, environ 200
 - Outliers -> difficile de les enlever a priori
 - Gestion des valeurs manquantes

Quelle approche?

(en machine learning, dans notre cas)

Simples

Modèle linéaire

- Interprétation facile
- Plusieurs relations pourraient ne pas être linéaires
e.g. formes juridiques, ...

Arbre de décision

- Ok pour les variables mixtes
- Une seule « stratégie » est modélisée
- Risque d'overfit, surtout avec beaucoup de caractéristiques

...

Avancées

Neural nets, SVMs

- Sensibles aux hyperparamètres
- Entraînement plus coûteux
- Peu d'échantillons par rapport aux caractéristiques
- Normalement moins faciles à interpréter

Deep Learning

- Nombre d'échantillons trop petit
- Encore moins facile à interpréter
- Plus coûteux

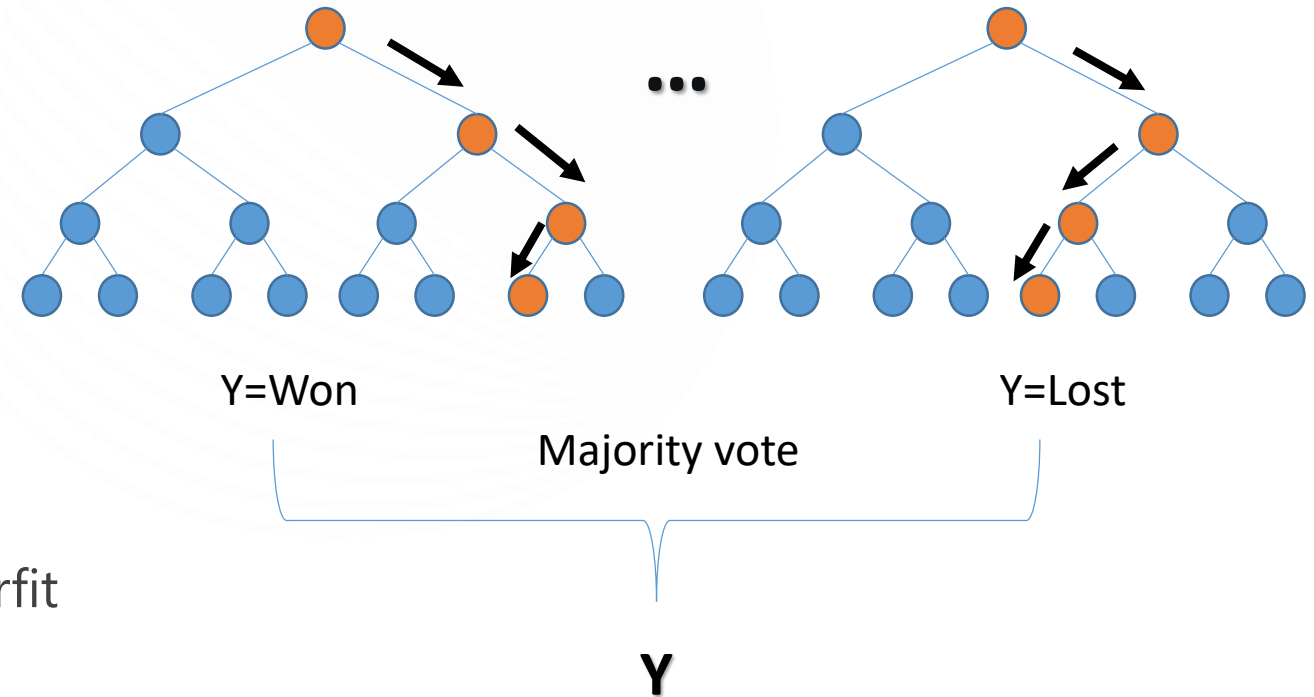
Random Forest

« shallow learning » qui répond à nos contraintes

T.K. Ho, 1995
L. Breiman 2001

- Ensemble d'arbres de décision

- Sortie: vote majoritaire
- Rapide, parallélisable
- Plusieurs type d'arbres, nombre incrémental



- « Features bagging »

bagging + random feature selection

- Réduit la variance du modèle et l'overfit
- Améliore la généralisation

- Plusieurs « outils » intégrés!

« Nuances de RF » ! -> testset intégré OOB, explication de l'importance des variables, ...

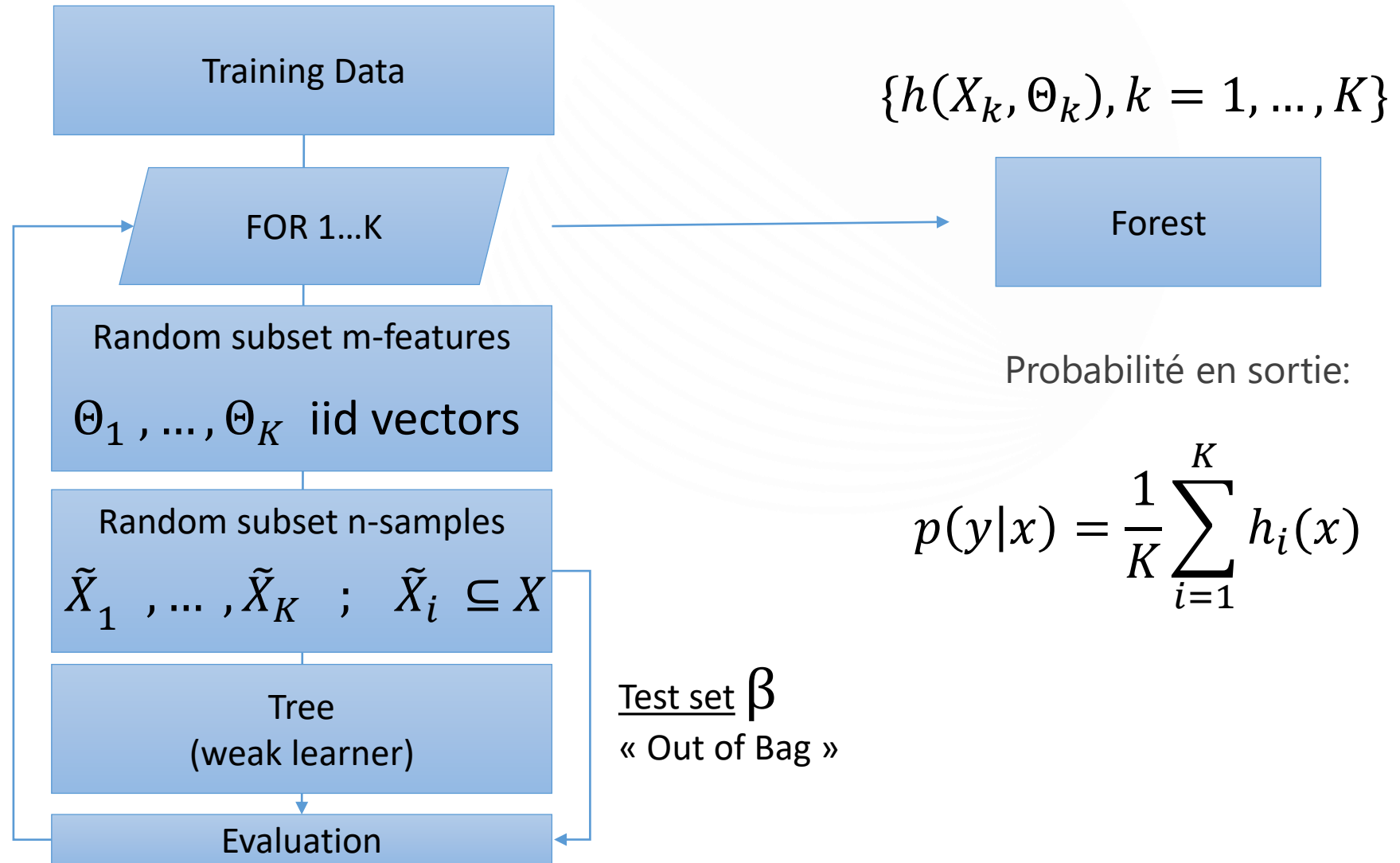
Random Forest

intérêt récemment renouvelé

- Pas seulement pour de la classification/régression, mais aussi pour de l'apprentissage non supervisé [Shi2006]
- Pour évaluer l'importance des variables, surtout dans des datasets avec beaucoup des dimensions [Behnamian2017]
- RF fourni plusieurs approches pour traiter des datasets avec classes déséquilibrées (e.g. « balanced trees ») [More2017]
- Utile en 1^{er} approche pour sonder la modélisation de problèmes big data (business d'entreprise, plus performant que les « vanilla » SVMs) [Weiwei2017]

Algorithme d'apprentissage

L. Breiman 2001



Erreur « Out Of Bag »

OOB

- Le bagging fournit un « test set » β^t pour chaque arbre
- RF utilise ces test sets pour mesurer l'erreur de prédiction
- Les études de Breiman montrent que l'erreur OOB donne une mesure « comme si on utilisait un test set de la même taille que le training set »
 - Ceci nous permet d'avoir une mesure rapide de la validité du modèle
 - Pratique pour de petits datasets
- Dans les cas pratiques cette mesure suffit pour évaluer le modèle
un test set à part est utile si plus de précision est nécessaire

Evaluation

Test set β
« Out of Bag »

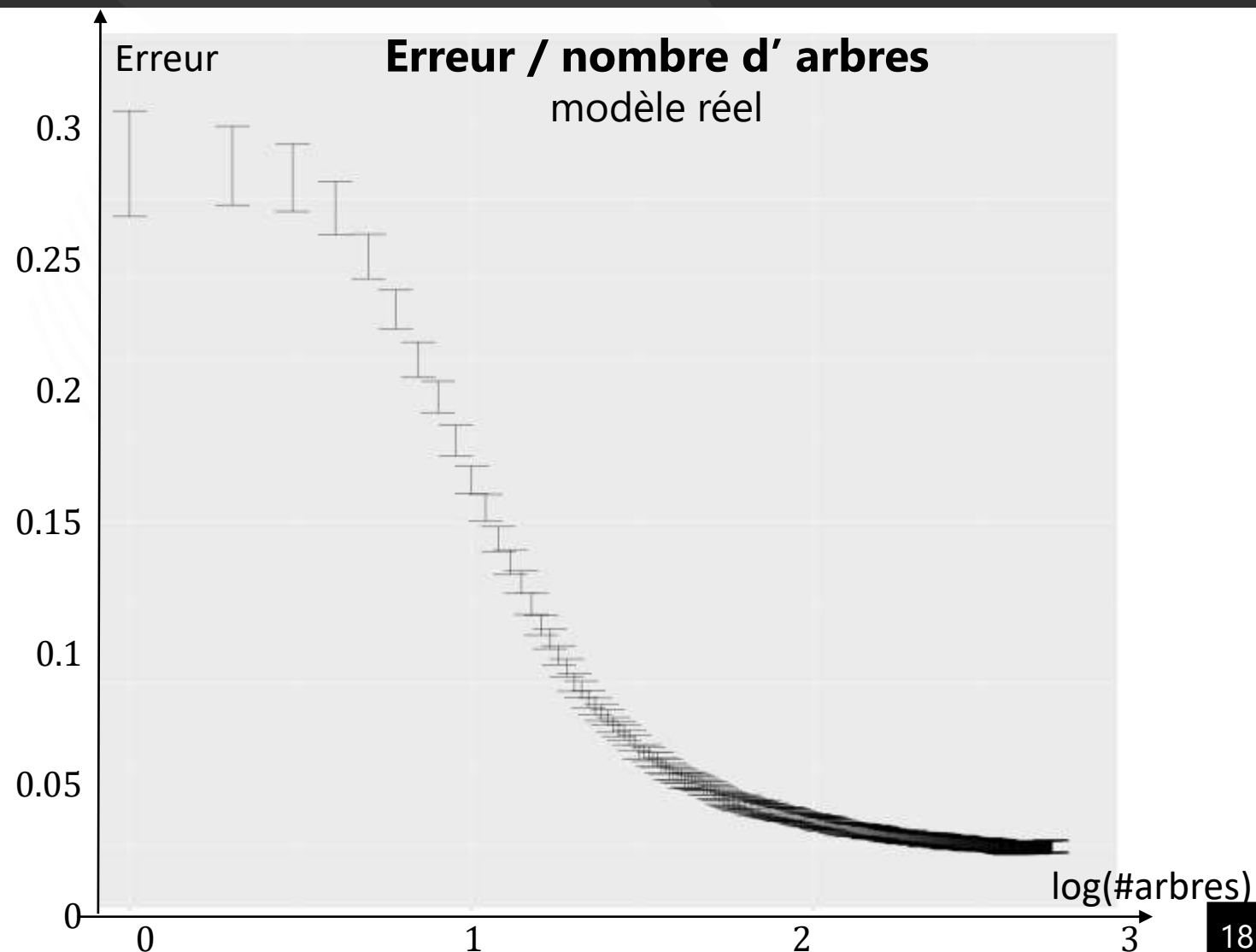
Hyperparamètres du modèle

- Le modèle RF a peu d'hyperparamètres:
(qui jouent un rôle moins important que dans des modèles plus complexes...)
 - ➔ m : nombre de variables aléatoires par arbre
paramètre le plus important, impacte la « diversité des arbres » -> *variance*
 - ➔ k : nombre d'arbres
convergence de l'erreur
 - *depth*: profondeur des arbres (*biais*)
 - n : nombre d'échantillons par arbre
- Optimisation classique (grid-search, cross-validation...)
- Choix suboptimale: compromis entre optimisation et vitesse d'entraînement
marchés similaires -> paramètres suboptimaux ok pour plusieurs modèles

Effet du **nombre d'arbres** sur le taux d'erreur du modèle

Le nombre d'arbres est corrélé avec la capacité de RF de modéliser le dataset

- Plus d'arbres, plus de tirages aléatoires
- ...plus de temps et mémoire requis !
- Ne change pas le *biais* du modèle
- Pas d'overfit!
- Asymptote: dépend de l'information donnée par les variables



Explication du score

importance des variables avec RF

RF permet d'inclure une évaluation de l'importance des variables pendant l'entraînement: *mean decrease accuracy*

Pour chaque arbre, pour chaque variable, on fait une **permutation** des valeurs et on mesure l'accuracy des prédictions

Méthode valide en 1^{ère} approximation mais peut sous-estimer l'importance dans certains cas: [Strobl2007], variables de type hétérogène ou avec beaucoup des catégories

D'autres méthodes statistiques plus précises, mais plus coûteuses, (visant à mesurer le *p-value*) ont été proposées (e.g. « Boruta » [Kursa 2010])

$$X_{\pi} = [x_{i,1}, x_{i,2}, \dots, \pi(x_{i,j}), \dots]$$

$$VI^t(x_j) = \frac{\sum_{i \in \beta} I(y_i = \hat{y}_i | X) - \sum_{i \in \beta} I(y_i = \hat{y}_i | X_{\pi})}{|\beta^t|}$$

$$VI(x_j) = \frac{\sum_t VI^t(x_j)}{ntree}$$

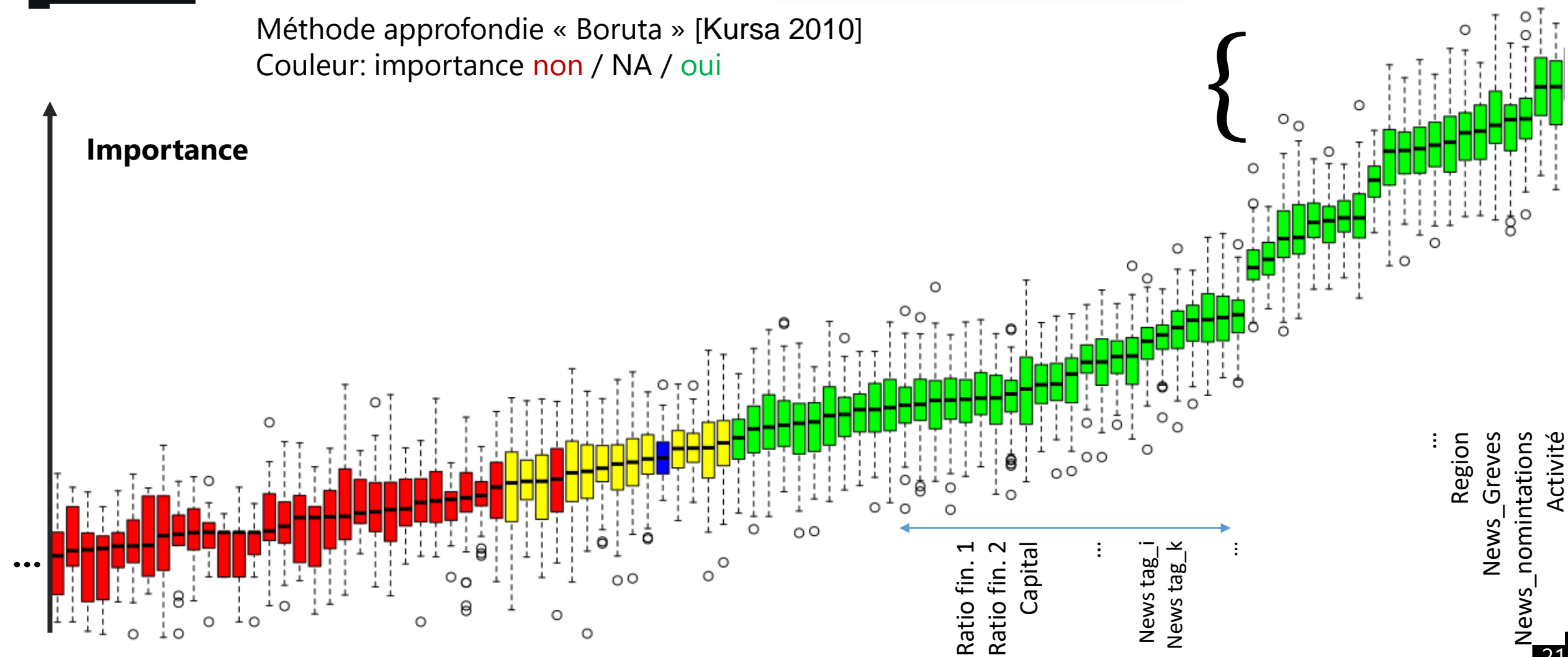
Résultat de cas réel

Extrait importance des variables

R&D

Méthode approfondie « Boruta » [Kursa 2010]

Couleur: importance **non** / NA / **oui**



Résultat de cas réel

importance des variables dans l'interface produit

Business

Avant de les montrer dans la plateforme, elles sont agrégées par catégorie

NB: certaines peuvent avoir une influence négative sur le résultat (e.g. baisse de $P(\text{gagner})$)



FIT SCORE

- Activité
- Coordonnées géographiques
- Date de création
- Flotte (Données embarquées)
- Deal size (Données embarquées)
- Chiffre d'affaires
- Chiffre d'affaires consolidé
- Effectifs consolidés

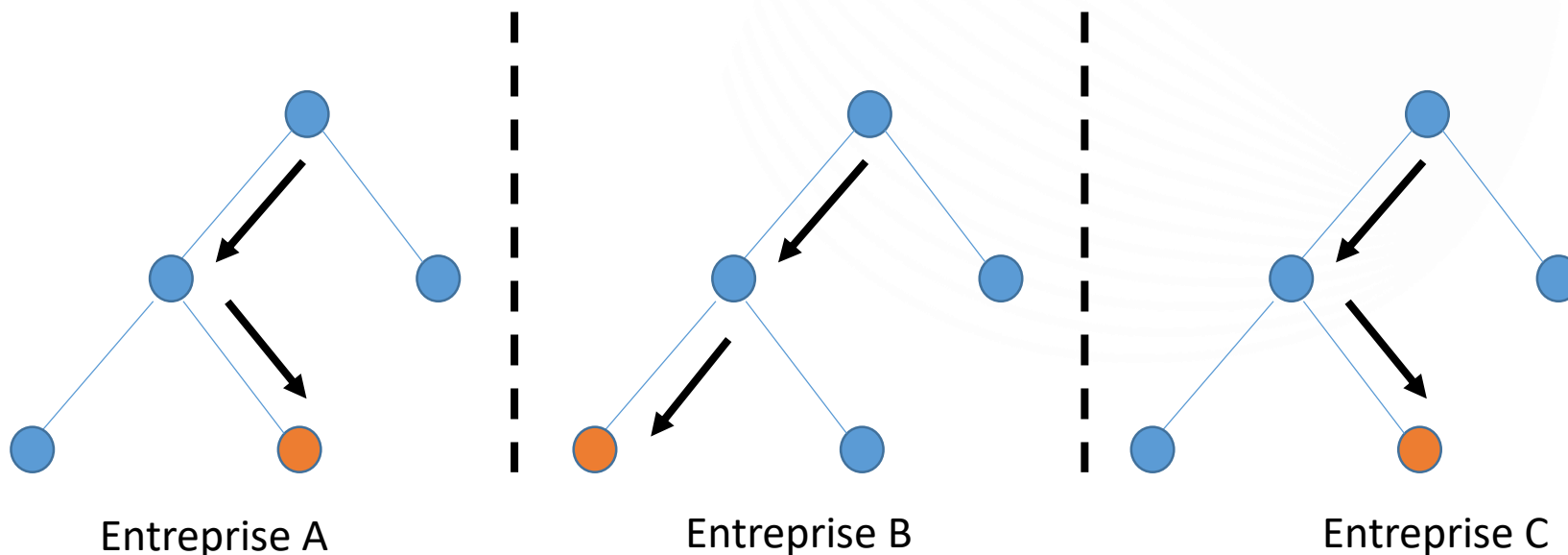


NEED SCORE

- Grèves / Mouvements sociaux (1 an)
- Nominations (1 an)
- Entrée en bourse (6 mois)
- Recrutements (1 an)
- Résultats positifs (1 an)
- Participation / présence salons (1 an)

Explication du score par proximité

RF nous permet de mesurer la proximité des entreprises par rapport au « partage » des *feuilles* des arbres



Exemple pour un même arbre, trois entreprises

L'entreprise A et C « tombent » dans la même feuille:
elles sont plus similaires entre elles que avec B

Explication du score par proximité

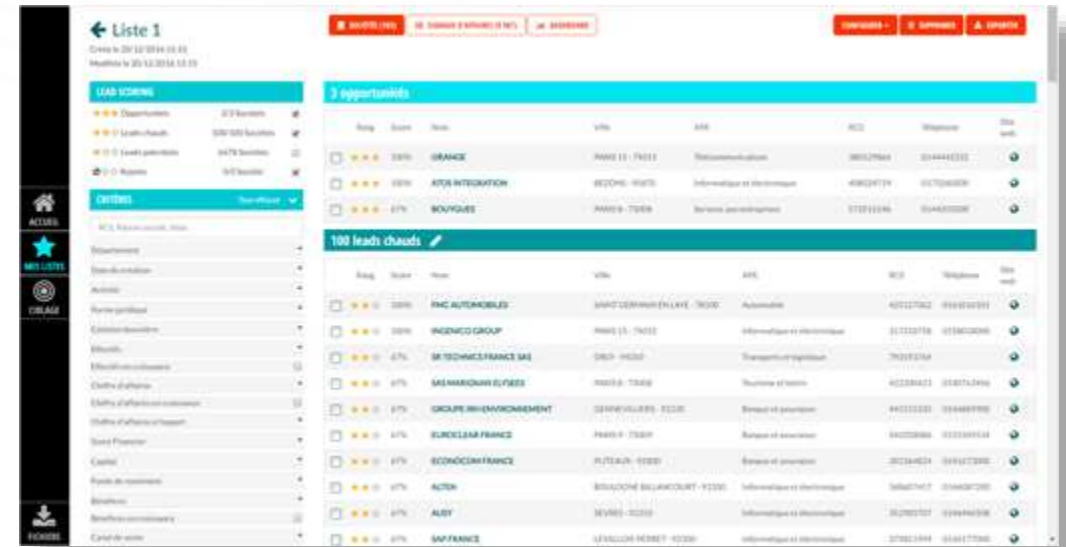
R&D

- En considérant tous les arbres et toutes les entreprises, on obtient une *matrice de proximité*
N.B. : contraintes techniques à considérer pour de gros datasets (mémoire et temps)

- On peut donc trouver les N plus proches entreprises pour un prospect donné

Business

Par rapport à votre marché, la société Z [...] ressemble à vos clients X et Y, [...]



The screenshot shows the Sparklane interface with a sidebar on the left containing navigation icons and a main content area. The main area is titled 'Liste 1' and displays a table of companies with columns for Rank, Score, Name, Vite, Age, SIC, and Téléphone. The table is divided into sections: '100 leads chauds' and 'opportunités'.

Rank	Score	Nome	Vite	Age	SIC	Téléphone	Site web
1	98%	GRANDE	PARIS 15 - 75015	Technocommunication	003327644	0144441333	
2	98%	ADON INTEGRATION	BOISHERY - 91070	Informatique et électronique	0033124719	0170246000	
3	97%	BOUGUES	PARIS 9 - 75009	Services informatiques	075711046	0144300000	
4	98%	PMC AUTOMOBILES	PARIS 15 - 75015	Automobile	0033127062	0144441333	
5	98%	INDENCO GROUP	PARIS 15 - 75015	Informatique et électronique	017221718	0154020000	
6	97%	SE TECHNIQUES FRANCE SAS	BOISHERY - 91070	Transport et logistique	792020104		
7	97%	SALIMARKOV ELITES	PARIS 9 - 75009	Services et services	012388021	0181743496	
8	97%	GRUPPE RHENANOMEMENT	BOISHERY - 91070	Services et services	003312333	0144441333	
9	97%	EUROCLIM FRANCE	PARIS 9 - 75009	Services et services	003308086	012388021	
10	97%	ECONOCOM FRANCE	PARIS 15 - 75015	Services et services	003314424	0144441333	
11	97%	ALBY	BOISHERY - 91070	Informatique et électronique	0033124719	0170246000	
12	97%	ALBY	BOISHERY - 91070	Informatique et électronique	0033124719	0170246000	
13	97%	SAP FRANCE	BOISHERY - 91070	Informatique et électronique	075711046	0144300000	

Résultat exemple du score par proximité

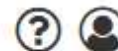


← ELECTRICITE DE FRANCE 🔍

📘 INFORMATIONS

📢 SIGNAUX D'AFFAIRES

+ SUIVIE DANS UNE LISTE



📘 INFORMATIONS

Siège social 📍

22 AVENUE DE WAGRAM
75008 PARIS 8

☎ 01 40 42 22 22 ⓘ

Activité

Energie
Electricité
3511Z - Production d'éle

SIREN

552081317

Capital

1 000 M€

Sociétés similaires

- ⊕ ENGIE
- ✓ EURAZEO
- ✓ FAURECIA
- ✓ EIFFAGE

59% Compte similaire à d'autres succès

Fit



Need



Compte sélectionné



Proximité pour imputer les valeurs manquantes

Business

- On a des valeurs manquantes sur les informations des entreprises, surtout sur les variables fournies par les clients
- On ne doit pas exclure ces entreprises de notre data lake

R&D

- Imputation des valeurs manquantes
 - De solutions rapides existent (médiane, ...)
- RF nous permet d'imputer par proximité : les valeurs de certaines variables sont plus similaires entre échantillons proches
- Plusieurs stratégies proposées:
 1. Pré-imputation par médiane, calcul RF, ré-imputation en considérant seulement les M plus proches échantillons (proximité), répéter N fois [Breiman2003]

Proximité pour imputer les valeurs manquantes

2. Imputation pendant la construction des arbres (adaptative tree, [Ishwaran2008])

3. **MissForest:**

Pré-imputation, calcul RF pour chaque variable avec valeurs manquantes, prédiction de ces valeurs, répéter N fois [Stekhoven2012]

- Surtout pour de datasets avec des variables hétérogènes, l'erreur d'imputation est réduit considérablement par rapport à KNN et MICE
- Est par contre beaucoup plus lent que KNN (env. 5x)
- Parmi les travaux exploratoires en cours chez nous -> opération « backoffice »

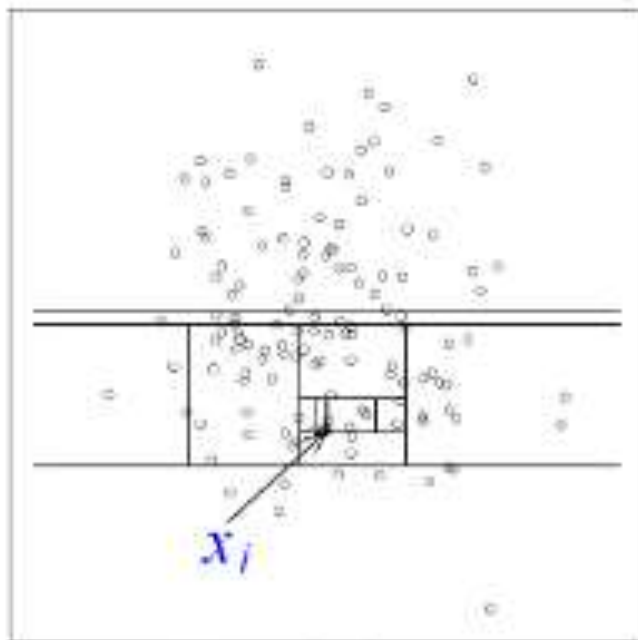
Business ... mais on a aussi des prospects à la marge, biens différents des autres

Isolation Forests

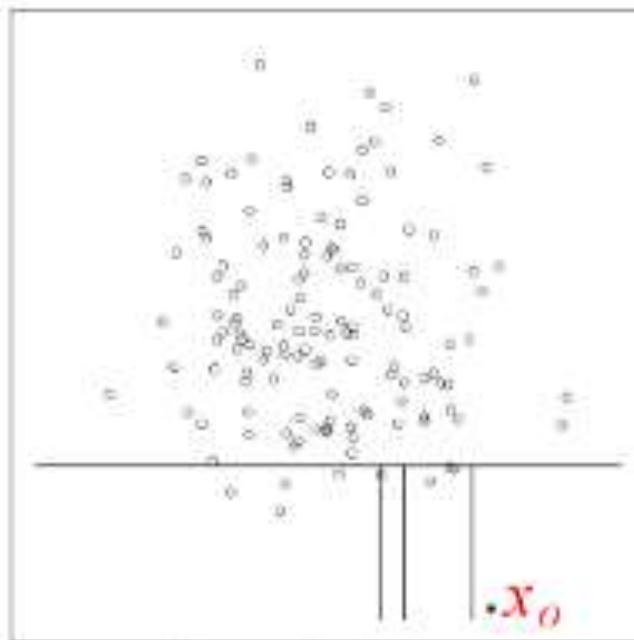
pour trouver les outliers

Liu et Al. 2008

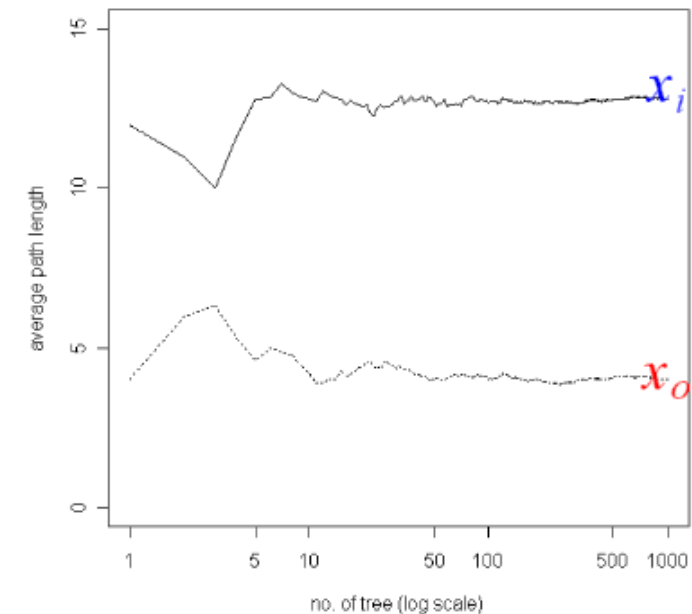
- Isolation Forest: évolution de RF
- Cet algorithme utilise les arbres pour séparer (isoler) les échantillons
- S'il faut des arbres plus étendus pour les séparer, il y a de fortes chances que ces échantillons soient des outliers



(a) Isolating x_i

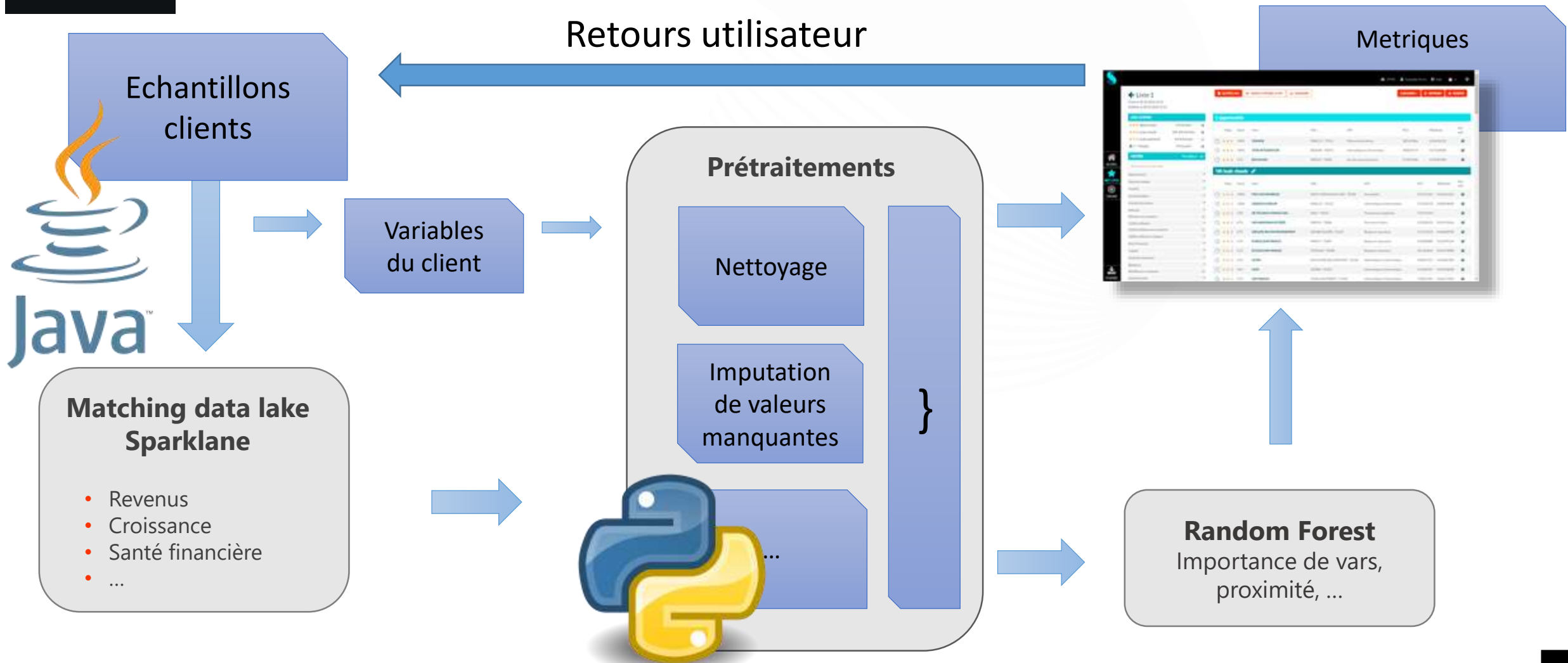


(b) Isolating x_o



(c) Average path lengths converge

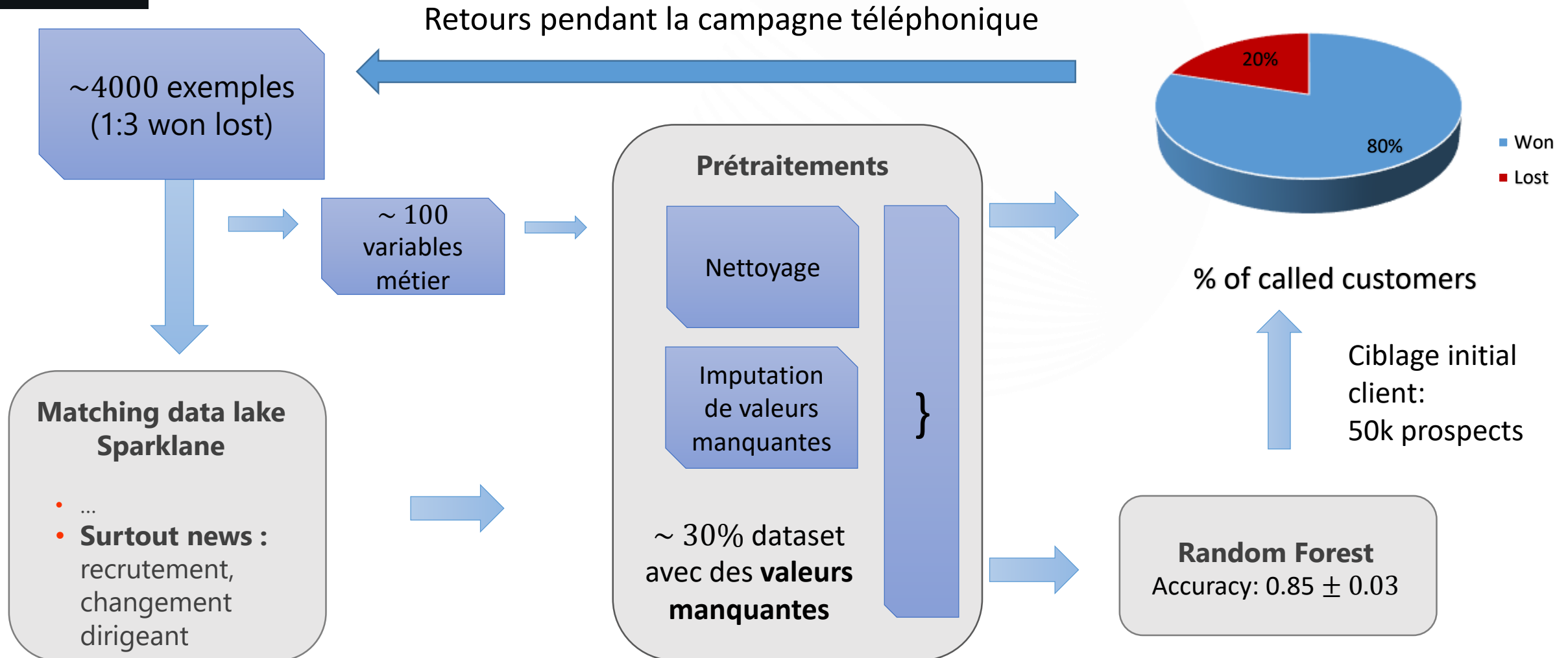
Chaîne de traitements (simplifiée) dans notre cas d'usage



Exemple de résultat

Éditeur solutions logiciel prospecte un *upsell*

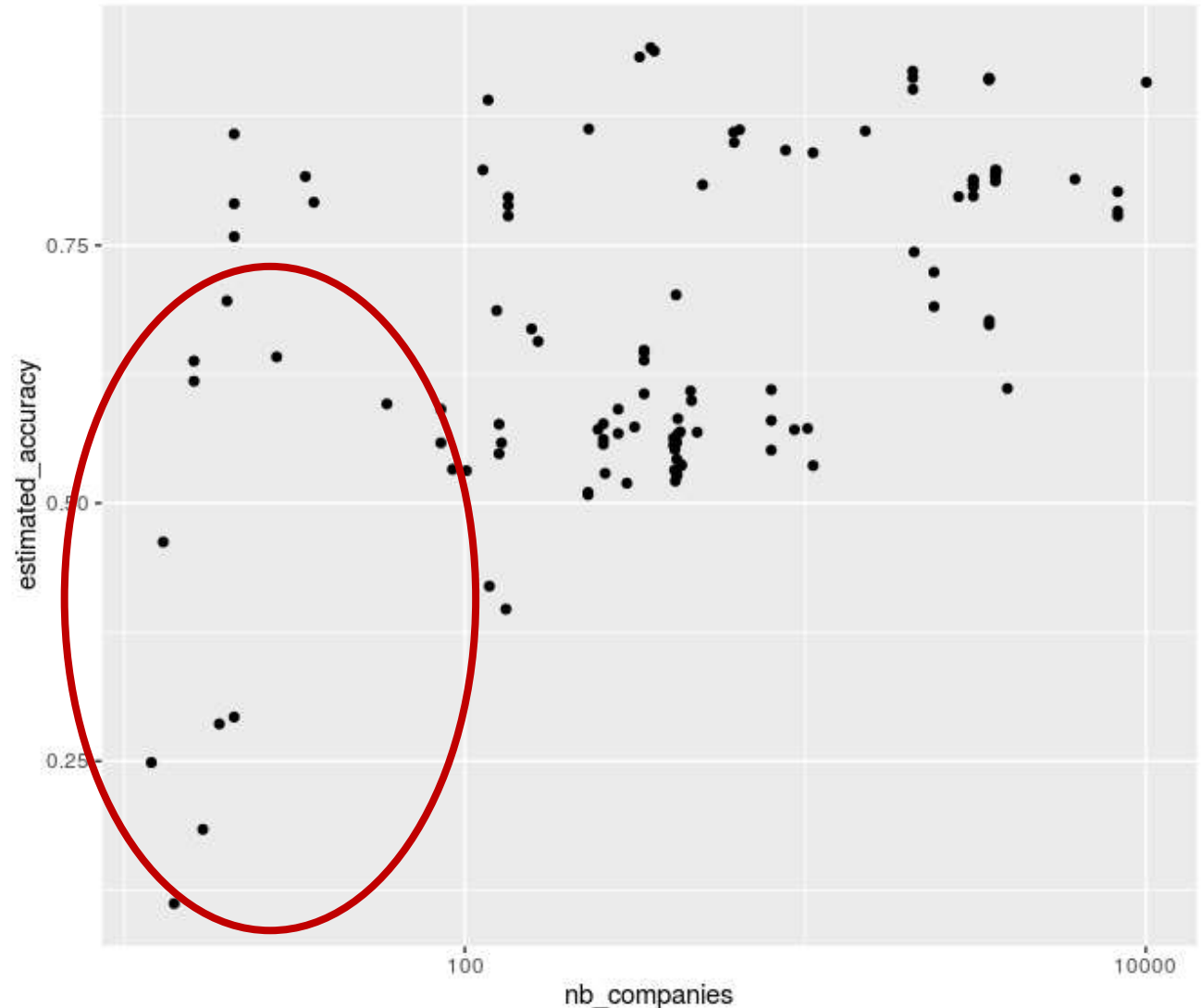
~ 30%
d'efficacité
relative en plus



Résultats: OOB modèles

Q4 2017

- Relation avec la taille du jeu d'entraînement
- Plus grande variation avec peu d'entreprises
- Certains clients ont **très peu d'échantillons** pour l'entraînement!
- A faire: corrélation avec « onboarding »



Evolution du modèle

ou réponse au problème précédent

Business

- L'utilisateur pourrait ne pas avoir accès à l'historique des ventes ou avoir peu de cas clients à proposer
 - Le marché de l'utilisateur et sa stratégie commerciale évoluent avec le temps
-

R&D

- Plusieurs entraînements du modèle au fur et à mesure
 - RF a l'avantage d'être rapide
- Possibilité: remplacer une partie des arbres
- Possibilité: apprentissage « actif » avec un retour utilisateur:
 - on construit un modèle initiale
 - on demande à l'utilisateur d'étiqueter les échantillons (*prospection*) qui sont les plus informatifs pour mettre à jour les seuils décisionnels
 - on entraîne et on réitère

Active Learning basé sur l'incertitude

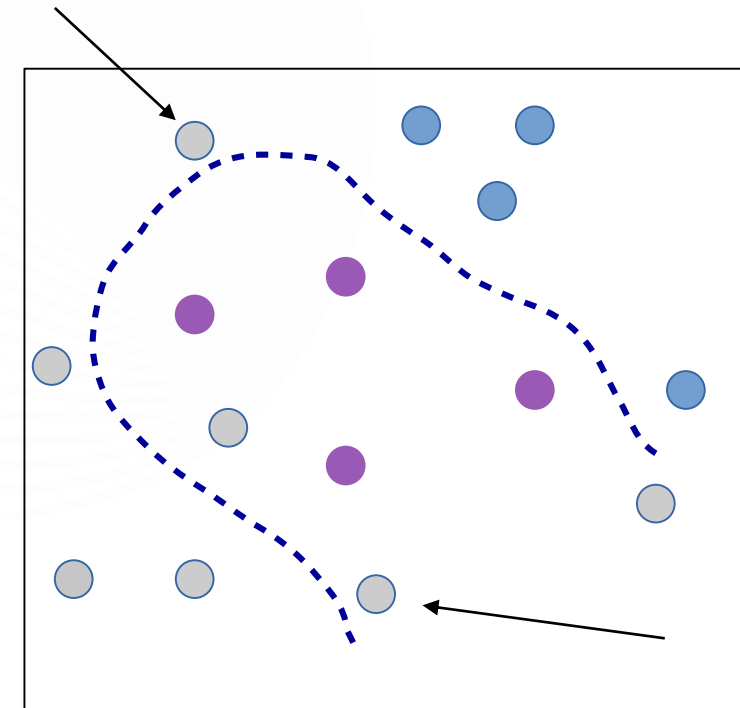
[Scheffer 2001]

- Online learning: nouveaux échantillons au fur et à mesure
- Deux phases: exploration et exploitation
- Exploration: on propose à l'utilisateur l'entreprise

$$\operatorname{argmax}_x (\varphi_{\text{uncertainty}})$$

$$\varphi_{\text{uncertainty}} = P_{\theta}(Y_{\text{won}}|X) - P_{\theta}(Y_{\text{Lost}}|X)$$

- Très utile aussi si l'utilisateur n'a aucun échantillon! (bootstrap)



Violet: entreprises intéressantes
Bleu: entreprises non intéressantes
Candidates: flèches

Conclusions

et pistes pour la suite

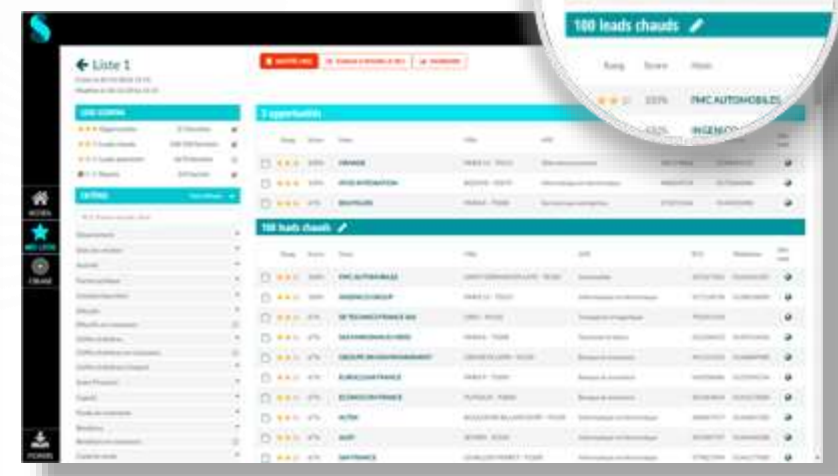
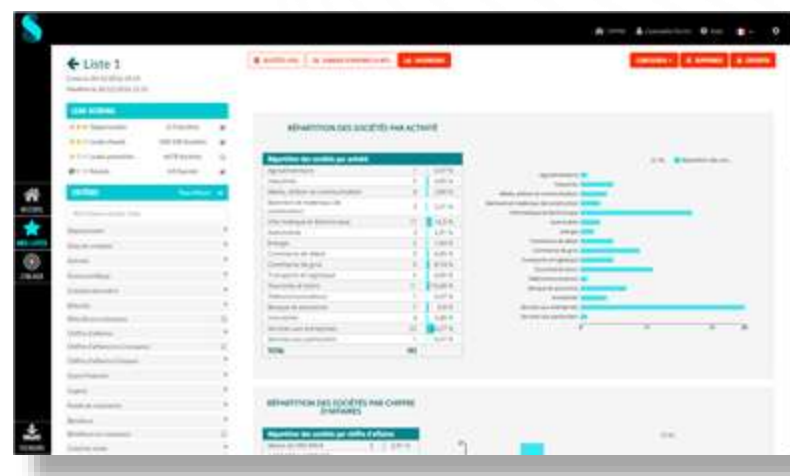
Random Forest

- Méthode presque « out of the box »
- Rapide, peu coûteuse
- Moins sensible aux hyperparamètres que d'autres méthodes
- Fourni des « nuances » utiles
- Plus simple à interpréter au niveau générale que d'autres méthodes
- Justesse légèrement plus faible que d'autres modèles ...mais à moindre coût

NEXT:

- Comparaisons plus approfondies avec des autres approches (pour les cas avec suffisamment d'échantillons)
- Imputation MissForest

Merci pour
votre attention



Suivez nous sur



www.sparklane-group.com

References

- Breiman L., « Random Forests », research report of University of Berkeley, 2001
- Breiman L., « Setting up, using and understanding Random Forests » , 2003
- Tao Shi et al., « Unsupervised Learning with Random Forest Predictors », Journal of Computational and Graphical Statistics, Vol.15 N.1, 2006
- Strobl C. et al., « Bias in random forest variable importance measures: illustrations, sources and a solution », BMC Bioinformatics, 2007
- Liu F. et al., « Isolation Forest », IEEE International Conference on Data Mining, 2008
- Kursa M.B., Rudnicki W.R., « Feature selection with the Boruta Package », Journal of Statistical Software, Vol 36 issue 11, 2010
- Stekhoven et al., « Missforest-non-parametric missing value imputation for mixed-type data », Bioinformatics, Vol28 n.1, 2012

References

- Criminisi A. et al., « Decision Forests for classification, regression , density estimation, manifold learning and semi-supervised learning », Microsoft Research technical report 114, 2011
- Behnamian A. et al., « A Systematic approach for variable selection with Random Forests: achieving stable variable importance values », IEEE Geoscience and Remote Sensing, V14:11, 2017
- More A.S., Rana D.P., « Review of Random Forest classification techniques to resolve data imbalance », IEEE Conference on Intelligent Systems and Information Management, 2017
- Weiwei L. et al., « An ensemble Random Forest algorithm for insurance big data analysis », IEEE journal open access, 2017