Méthodes de classification supervisée appliquées aux terres agricoles

Tedjani Tarayoun

Insee
Département des Méthodes Statistiques
Division recueil et traitement de l'information

Journées de méthodologie statistique
13 juin 2018

Plan

Contexte

Le SSP et la FNSafer publient le prix à l'hectare des terres et prés non bâtis destinés à conserver, au moment de la transaction, leur vocation agricole

Au niveau des nouvelles régions agricoles (NRA) :

- La moyenne triennale des prix à l'hectare pondérée par la surface
- Les maximums et minimums correspondant au 95e et au 5e centile des prix à l'hectare observés les 3 dernières années

Au niveau des départements :

- La moyenne des prix calculés précédemment au niveau des NRA pondérée par la surface agricole libre des NRA

Au niveau des régions administratives et de la France entière :

- Les prix régionaux et nationaux sont calculés à partir d'indices hédoniques annuels

Contexte

La donnée permettant de savoir si une terre reste ou non à usage agricole est parfois manquante ou erronée

Traitement actuel : utilisation d'un filtre pour éliminer les terres < 70 ares

Variable « Destination du fonds »	Nombre de transactions [2014-2016]	Nombre de transactions Filtrage des terres trop petites [2014-2016]
Dest. inconnue	22 779	5 766
Dest. agricole incertaine	57 978	33 771
Dest. non agricole incertaine	44 034	8 683
Dest. non agricole certaine	22 635	0
Ensemble	147 426	52 238

- Le marché des petites terres est totalement exclu
- Filtre insuffisant : les terres à destination non agricole ne sont pas toutes exclues

Objectifs

Utiliser une méthode d'apprentissage supervisé au lieu d'un filtre (arbitraire) sur la surface

- Établir des modèles de classification à partir de terres dont la destination est connue
- Puis, utiliser ce modèle pour classer les terres dont la destination est inconnue

Créer un nouvel indicateur de la destination :

- Dest. agricole (30 %) : « Dest. agricole incertaine » avec un acquéreur agriculteur en activité (non retraité et moins de 65 ans) ;
- Dest. non agricole (15 %) : « Dest. non agricole certaine » ;
- Dest. inconnue (55 %) : Tout le reste.

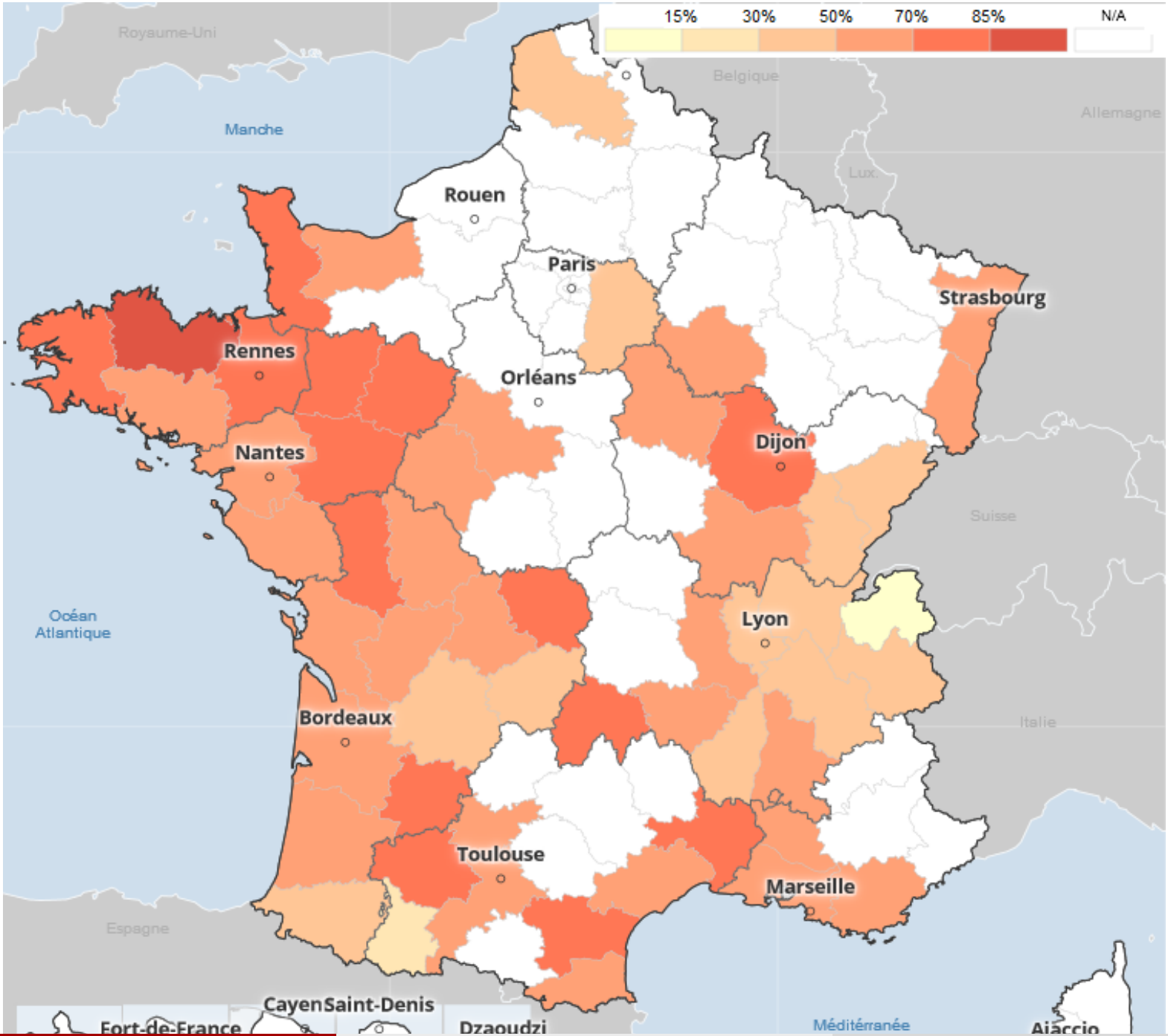
Périmètre de l'étude

Utilisation des données de transactions de 2014, 2015 et 2016

Analyse menée département par département

Analyse limitée aux 56 départements pour lesquels :

- au moins 150 terres restent à usage agricole
- au moins 150 terres changent d'usage

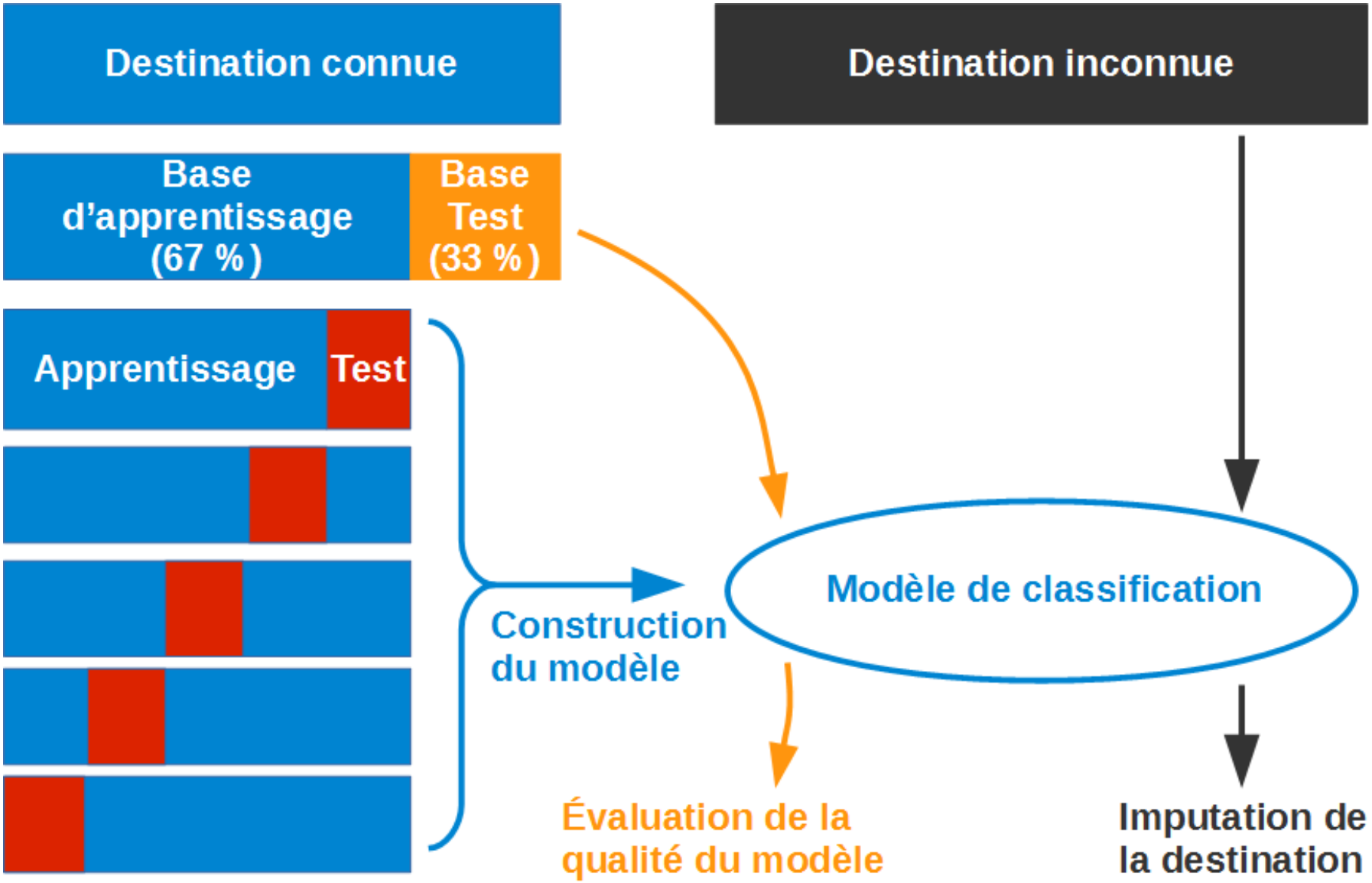

15%
30%
50%
70%
85%
N/A
Royaume-Uni
Manche
Belgique
Allemagne
Lux.
Rouen
Paris
Strasbourg
Rennes
Orléans
Nantes
Dijon
Suisse
Océan Atlantique
Lyon
Bordeaux
Italie
Toulouse
Marseille
Espagne
CayenSaint-Denis
Fort-de-France
Dzaoudzi
Méditérranée
Ajaccio

Prédicteurs

Sources	Variables
FNSafer	Prix à l'hectare du fonds Surface du fonds Nature cadastrale dominante : 'Terre', 'pré' ou 'Terre et Pré' Nouvelle région agricole (petites régions agricoles regroupées) Part de surface artificialisée dans la commune
DATAR INRA CESAER DTM ThéMA METAFORT	Typologie du littoral -*Littoral artificialisé urbain et périurbain* -*Littoral de type rural méditerranéen* -*Littoral de type rural atlantique* Typologie des montagnes -*Haute et moyenne montagne résidentielle et touristique* -*Moyenne montagne agricole ou industrielle* -*Montagne urbanisée*
Insee	Densité de population dans la commune
SSP	Nomencature OTEX dans la commune
DGCL	Potentiel financier de la commune (richesse théorique de la commune)
INRA	Distance de la commune au pôle urbain le plus proche

Protocole d'apprentissage

Construction de l'échantillon d'apprentissage et de test



Rééquilibrage de l'échantillon d'apprentissage

La proportion de terres qui restent à destination agricole varie selon le département entre 11% et 86% : déséquilibre des classes !

La méthode SMOTE (Synthetic Minority Over-sampling Technique) : création d'observations « interpolées » :

- Sélection d'une observation de la classe minoritaire et calcul de ses k plus proches voisins dans la classe minoritaire
- Création d'une observation aléatoire sur le segment entre une observation et un de ses k-ppv (choisi aléatoirement)

La méthode ROSE (Random Over-Sampling Examples) : création d'observations bruitées :

- Sélection d'une observation et création de nouvelles observations dans son voisinage
- S'apparente à une modélisation de la densité par une méthode à noyau

Critères de fiabilité des modèles

Critères dépendant d'un seuil

Les modèles de classification font intervenir un seuil au-delà duquel on est déclaré positif et au-dessous duquel on est déclaré négatif

$$\text{Sensibilité} = \frac{VP}{VP + FN}$$

$$\text{Rappel} = \frac{VP}{VP + FN}$$

$$\text{Précision} = \frac{VP}{VP + FP}$$

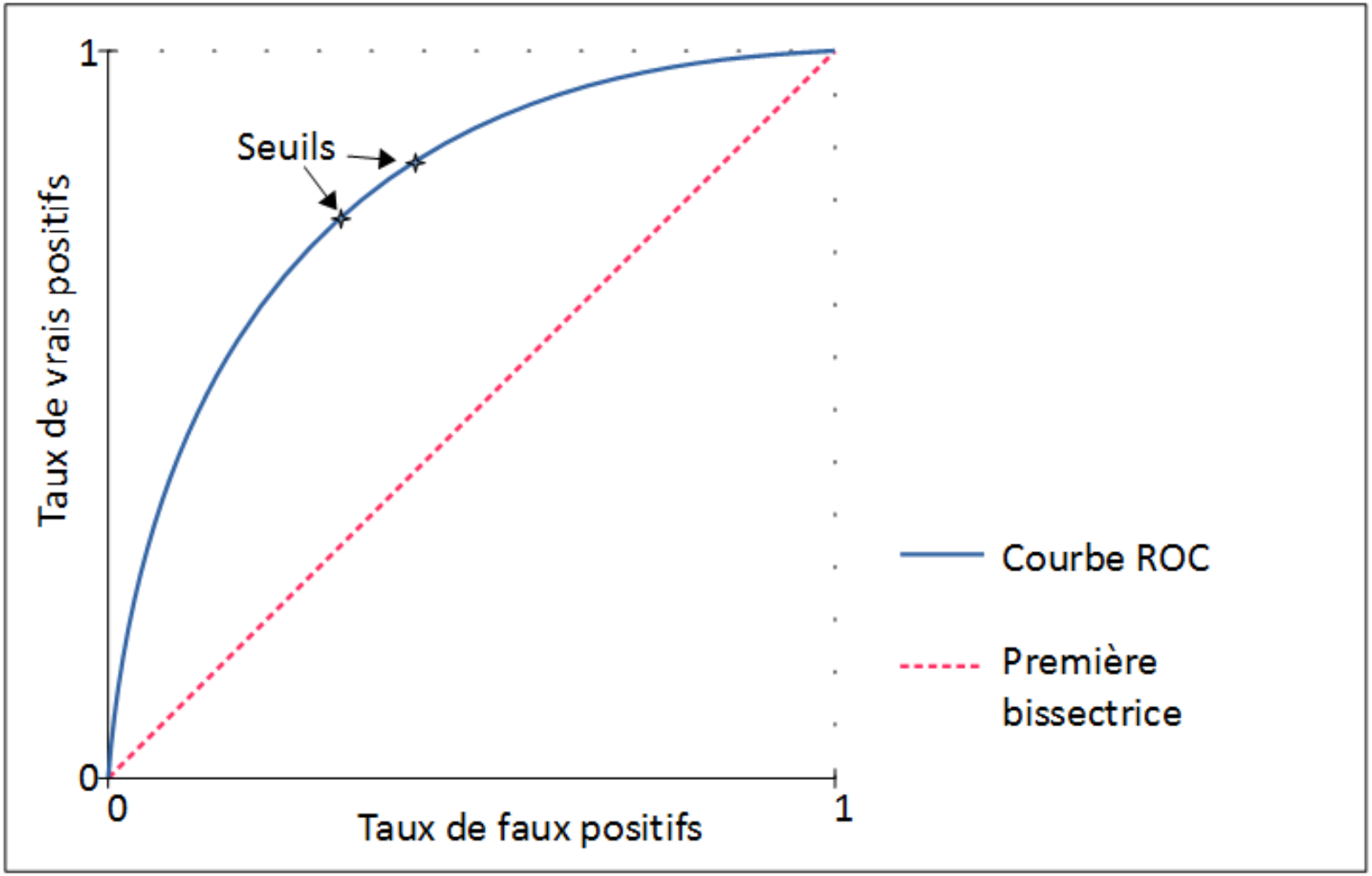
$$\text{Spécificité} = \frac{VN}{VN + FP}$$

$$F - mesure = 2 \times \frac{\text{Rappel} \times \text{Précision}}{\text{Rappel} + \text{Précision}}$$

	Prédit : agricole	Prédit : non agricole
Réel : agricole	Vrai positif (VP)	Faux négatif (FN)
Réel : non agricole	Faux positif (FP)	Vrai négatif (VN)

Critères de fiabilité des modèles

Critères indépendant du seuil : l'aire sous la courbe ROC (AUC)



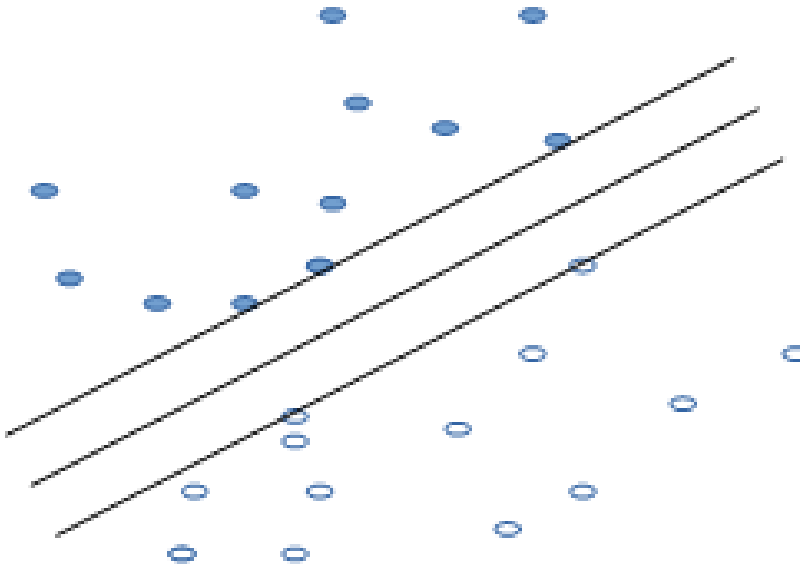
Modèles de classification

Les modèles de classification mis en œuvre sont :

- Régression logistique logit - Sélection ascendante en minimisant l'AIC
- Support Vector Machine - Linéaire
- Support Vector Machine - Radial gaussien
- Arbre de décision
- Forêt aléatoire
- Boosting (d'arbres de décision)

Support Vector Machine

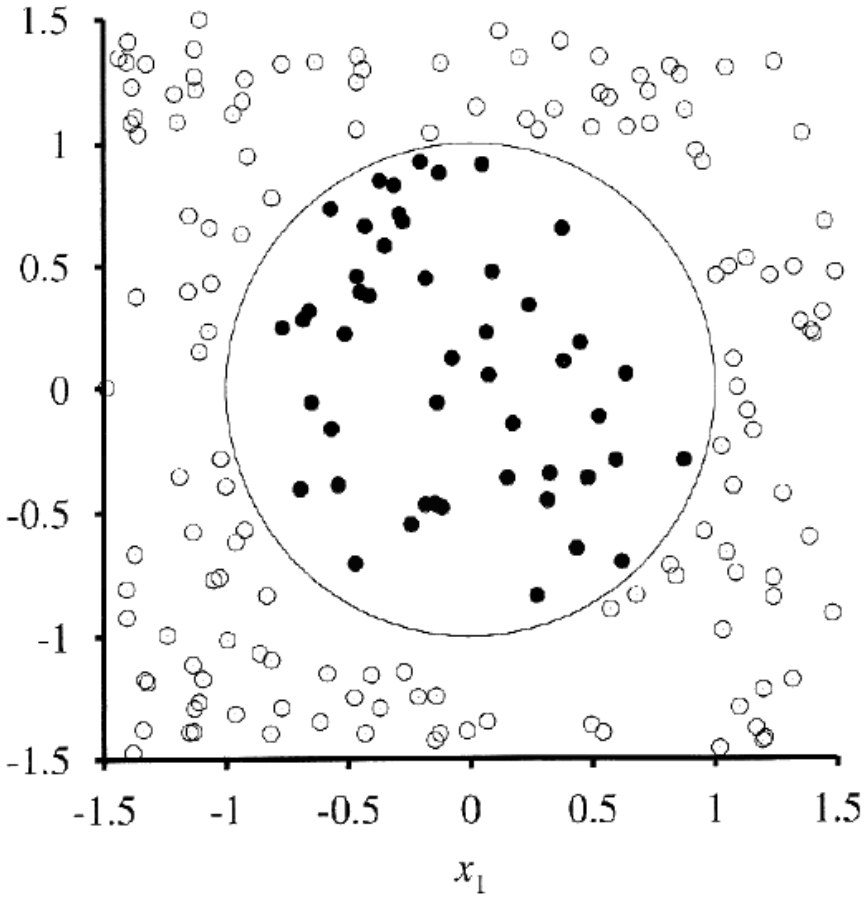
Le principe de base est la recherche, lorsqu'elle existe, d'une séparation linéaire des classes sous la forme d'un hyperplan. Cet hyperplan est optimal s'il maximise la marge entre les classes.

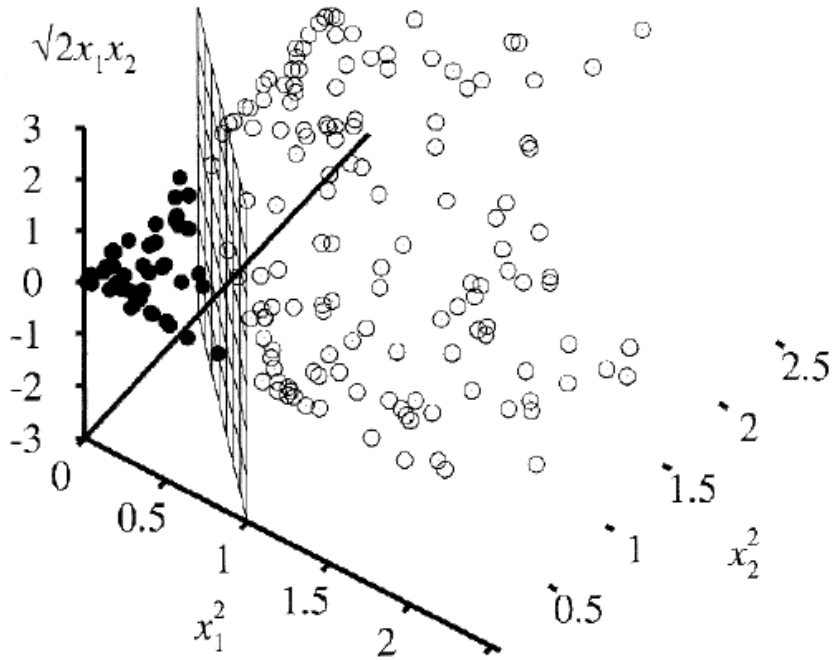


En pratique, on autorise des erreurs de classement et on introduit un paramètre de pénalisation des erreurs

Support Vector Machine

On peut aussi rechercher une séparation linéaire dans un espace de dimension plus grande.





Support Vector Machine

Il n'y a pas besoin de connaitre expressément la transformation : elle intervient dans le problème d'optimisation uniquement sous la forme d'un produit scalaire, défini par une fonction noyau k(x,x').

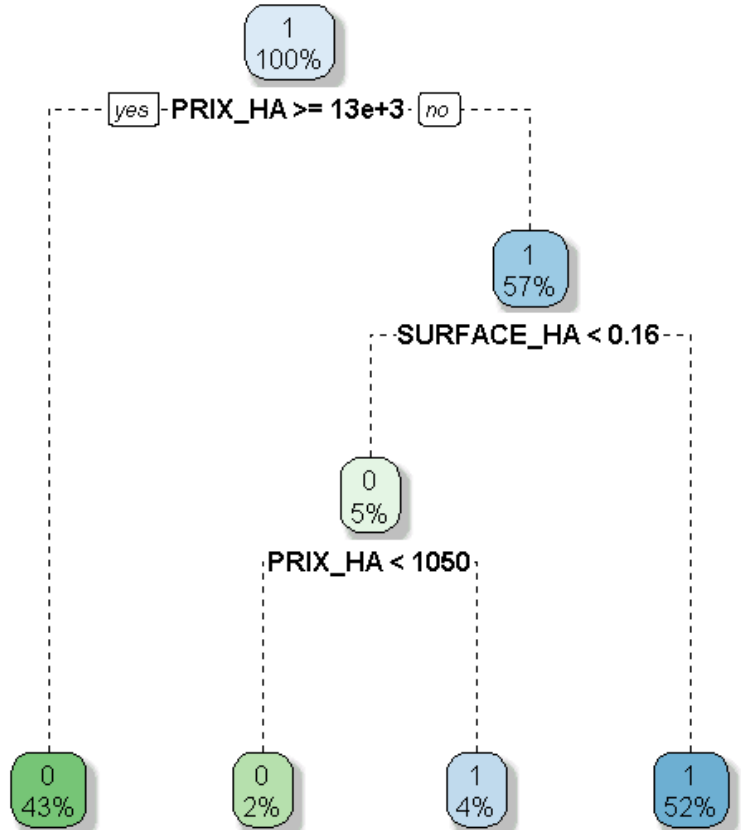
Les noyaux les plus utilisés sont :

- noyau linéaire : $k(x, x') = x \cdot x'$
- noyau radial gaussien : $k(x, x') = e^{-\gamma \|x - x'\|^2}; \gamma \geq 0$
- noyau radial laplacien : $k(x, x') = e^{-\gamma \|x - x'\|}; \gamma \geq 0$
- noyau polynomial :
  $k(x, x') = (x \cdot x' + c)^d$; avec d le dégré du polynome

Arbre de décision

L'arbre de décision (CART) est un processus récursif de divisions binaires de l'échantillon en sous-ensemble de plus en plus « purs » par rapport à la variable d'intérêt.

- Chaque division est réalisée à partir d'une variable explicative et d'une condition sur cette variable
- Le critère de pureté utilisé est la concentration de Gini
- Une fois construit, l'arbre est élagué au niveau offrant l'erreur de prédiction minimale (validation croisée)



Forêts aléatoires

La forêt aléatoire est une méthode d'agrégation de modèles, elle réduit l'erreur de prédiction de l'arbre de décision en réduisant la variance

- On génère un grand nombre d'échantillons bootstrap à partir de l'échantillon d'apprentissage
- On construit un arbre de décision sur chaque échantillon, mais chaque nœud de chaque arbre est construit de façon aléatoire : la variable pour partager l'échantillon est cherchée sur un sous-ensemble des variables tiré aléatoirement (3 sur 11 ici)
- La prédiction est obtenue au vote majoritaire

Boosting

Le boosting (Adaboost) est aussi une méthode d'agrégation de modèles, elle permet de réduire à la fois la variance et le biais de prédiction

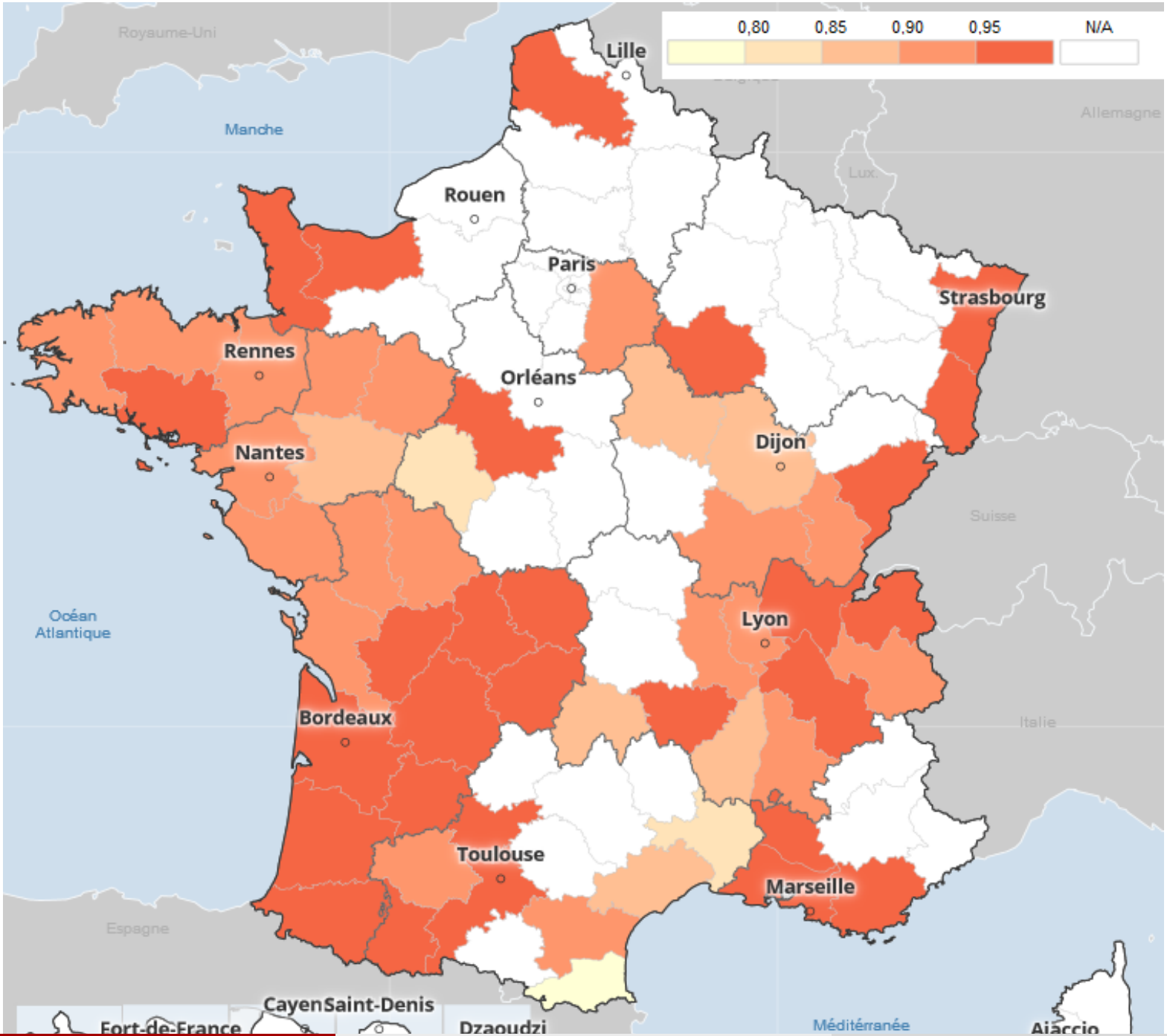
- Chaque modèle est une version adaptative du précédent en donnant plus de poids, lors de l'estimation suivante, aux observations mal ajustées
- On force ainsi l'apprentissage à se concentrer sur les cas les plus difficiles à prédire
- Le programme calcule automatiquement le nombre d'itérations optimal par validation croisée

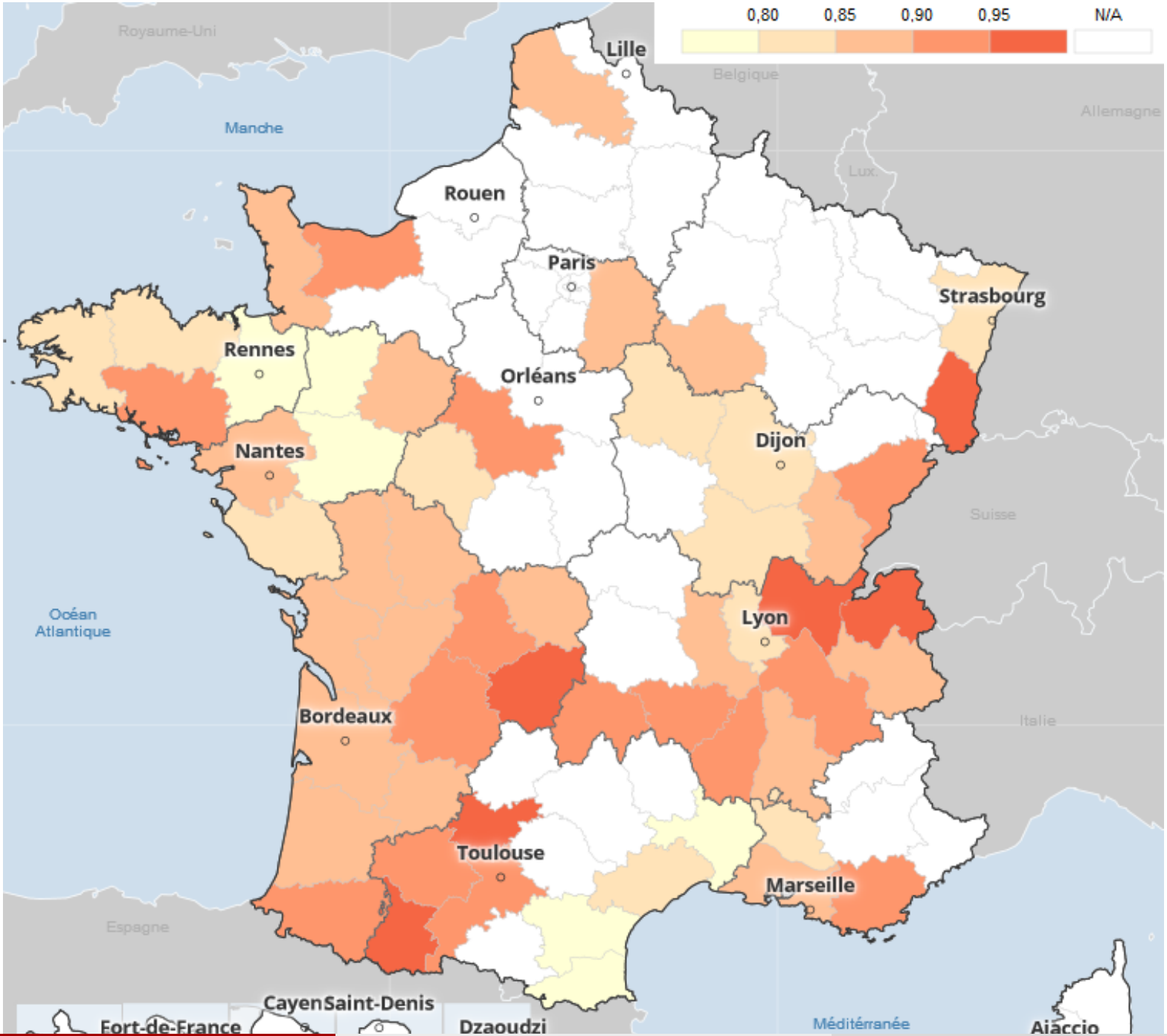
Résultats

Modèle	Équilibrage	Sensibilité	Spécificité	Précision	F-mesure	AUC
Logit	Non	0,91	0,74	0,79	0,84	0,9
	ROSE	0,85	0,81	0,82	0,83	0,89
	SMOTE	0,84	0,82	0,82	0,83	0,89
Logit Variables discrétisées	Non	0,91	0,74	0,79	0,84	0,9
	ROSE	0,84	0,8	0,81	0,83	0,88
	SMOTE	0,84	0,82	0,82	0,83	0,89
SVM Linéaire	Non	0,93	0,73	0,79	0,85	0,9
	ROSE	0,87	0,79	0,81	0,84	0,89
	SMOTE	0,87	0,82	0,83	0,85	0,89
SVM Linéaire Variables discrétisées	Non	0,93	0,73	0,79	0,85	0,9
	ROSE	0,87	0,8	0,81	0,84	0,89
	SMOTE	0,86	0,82	0,83	0,84	0,89

Résultats (suite)

Modèle	Équilibrage	Sensibilité	Spécificité	Précision	F-mesure	AUC
SVM Radial gaussien	Non	0,93	0,77	0,81	0,86	0,92
	ROSE	0,86	0,79	0,81	0,83	0,89
	SMOTE	0,85	0,83	0,83	0,84	0,9
SVM Radial gaussien Var. discrétisées	Non	0,93	0,77	0,81	0,86	0,92
	ROSE	0,84	0,78	0,8	0,82	0,87
	SMOTE	0,83	0,81	0,82	0,82	0,9
Arbre de décision	Non	0,92	0,8	0,83	0,87	0,87
	ROSE	0,91	0,79	0,82	0,86	0,87
	SMOTE	0,87	0,86	0,86	0,87	0,89
Foret aléatoire	Non	0,93	0,83	0,85	0,89	0,94
	ROSE	0,9	0,83	0,85	0,87	0,92
	SMOTE	0,89	0,87	0,88	0,88	0,94
Boosting	Non	0,92	0,81	0,84	0,87	0,92
	ROSE	0,88	0,81	0,83	0,85	0,9
	SMOTE	0,89	0,86	0,87	0,88	0,92


0,80
0,85
0,90
0,95
N/A
Royaume-Uni
Manche
Allemagne
Lux.
Suisse
Italie
Océan Atlantique
Espagne
Méditerranée
Lille
Rouen
Paris
Strasbourg
Rennes
Orléans
Nantes
Dijon
Lyon
Bordeaux
Toulouse
Marseille
CayenSaint-Denis
Fort-de-France
Dzaoudzi
Ajaccio


0,80
0,85
0,90
0,95
N/A
Royaume-Uni
Manche
Belgique
Allemagne
Lux.
Suisse
Italie
Océan Atlantique
Espagne
Méditerranée
Lille
Rouen
Paris
Strasbourg
Rennes
Orléans
Dijon
Nantes
Lyon
Bordeaux
Toulouse
Marseille
Fort-de-France
CayenSaint-Denis
Dzaoudzi
Ajaccio

Conclusion

Bilan :

- La forêt aléatoire donne de meilleurs résultats
- La méthode SMOTE a en particulier permis d'améliorer le critère de précision (les vrais positifs parmi ceux identifiés comme positifs)
- La complexité et les temps de calcul des SVM et du Boosting ne justifient pas, au vu des résultats, leur utilisation pour prédire la destination de la terre
- Néanmoins, les SVM et les modèles logistiques pourraient donner de meilleurs résultats en examinant plus finement les variables explicatives ou en utilisant d'autres fonctions noyaux pour les SVM

Pour la suite :

- Étendre l'analyse aux départements exclus de l'étude
- Identifier les cas qui posent des difficultés : dans ce cas faire appel à des dires d'expert pour établir les prix.

Bibliographie

Cornuéjols, « Une nouvelle méthode d'apprentissage : Les SVM. Séparateurs à vaste marge », 2002.

Friedman, Hastie, Tibshirani, « The Elements of Statistical Learning », 2009.

Menardi, Torelli, « Training and assessing classification rules with unbalanced data », 2014.

Tufféry, « Data mining et statistique décisionnelle : l'intelligence des données », 2017.

Tufféry, « Modélisation prédictive et apprentissage statistique avec R », 2017.