MÉTHODES DE CLASSIFICATION SUPERVISÉE APPLIQUÉES AUX TERRES AGRICOLES

Tedjani TARAYOUN

Insee, Direction de la méthodologie et de la coordination statistique et internationale

tedjani.tarayoun@insee.fr

Mots-clés *:* classification, prédiction, apprentissage supervisé, data mining

Résumé

Les prix des terrains agricoles sont publiés chaque année par la Fédération nationale des sociétés d'aménagement foncier et d'établissement rural (FNSafer), ils permettent en outre d'établir le barème de la valeur vénale des terres publié par le ministère de l'Agriculture au Journal Officiel. Ces prix ne relèvent que du seul marché des terres et prés non bâtis acquis dans une perspective d'exploitation agricole. Les informations sur les transactions sont issues des notifications de projets de ventes. Or, la donnée permettant de savoir si un terrain vendu restera, ou non, à usage agricole est erronée. Il est possible, par exemple, qu'entrent dans le champ des parcelles de terres acquises par des personnes désireuses d'améliorer leur cadre de vie ou d'y pratiquer de l'agriculture de loisirs (parcs à chevaux, potagers, etc.).

Notre étude propose d'appliquer des méthodes d'apprentissage supervisée pour imputer la valeur de la destination de la terre lorsqu'elle est inconnue ou incertaine. Les méthodes les plus courantes ont été mises en œuvre : la traditionnelle régression logistique *logit*, les arbres de décision, ainsi que des méthodes plus récentes et plus sophistiquées, les *Support Vector Machine* (séparateurs à vaste marge) et des méthodes d'agrégation de modèle (les forêts aléatoires et le *Boosting*). Ne connaissant par la proportion de destinations erronées, nous rééquilibrons la proportion des destinations agricoles dans la phase d'apprentissage. Ainsi les modèles pourront s'entraîner correctement. Deux méthodes de rééquilibrage sont testées : SMOTE (*Synthetic Minority Over-sampling Technique*) et ROSE (*Random Over-Sampling Examples*).

Les résultats montrent que les méthodes d'agrégation de modèles sont globalement plus performants pour résoudre notre problème de classification, les forêts aléatoires atteignent ainsi un AUC de 0,94. Néanmoins, toutes les méthodes testées ont de bonnes performances. Par ailleurs, les résultats révèlent que le rééquilibrage des modèles améliore la spécificité et la précision au détriment de la sensibilité : les terres à destination non agricole sont un peu mieux identifiées, tandis que celles à destination agricole le sont un peu moins bien.

Abstract

In this paper, we apply classification techniques such as Support Vector Machine, Random Forests and Boosting to predict the agricultural use of the land. Given that the classes to predict are not represented equally, we first re-balance the data using oversampling methods : Synthetic Minority Over-sampling Technique (SMOTE) and Random Over-Sampling Examples (ROSE).

Our results show that ensemble learning methods are overall the best approach to predict the agricultural use of the land (random forests reach an AUC of 0.94).

1. Introduction

Le Service de la Statistique et de la Prospective (SSP) du Ministère de l'Agriculture et de l'Alimentation publie, conjointement avec la Fédération nationale des sociétés d'aménagement foncier et d'établissement rural (FNSafer), le prix à l'hectare des terres et prés non bâtis. Ces prix sont calculés sur la base des valeurs des transactions et ne concernent que les terres et prés agricoles non bâtis destinés à conserver, au moment de la transaction, leur vocation agricole. Les vergers, les vignes et les terres maraîchères sont exclus du champ en raison de la rareté des transactions et la petitesse du marché les concernant.

Deux modes de calcul du prix coexistent suivant le niveau géographique considéré :

- Des moyennes triennales sont calculées au niveau des petites régions agricoles regroupées (zonage infra-départemental) et des départements.
- Des prix régionaux et nationaux sont calculés à partir des indices annuels régionaux définis par des modèles hédoniques.

La difficulté centrale de la mesure du prix est que la variable relative à la destination de la terre est mal codée. Jusqu'à présent, comme approximation, on réduisait le champ au marché de vente des terres et prés de plus de 70 ares, en faisant l'hypothèse que les plus grandes terres conservent un usage agricole. En réalité, ce critère ne garantit pas la destination agricole de la terre (plusieurs sociétés d'aménagement foncier et d'établissement rural ont noté que les prix publiés ne correspondaient pas tous à des prix agricoles). Par ailleurs, en faisant cela, le marché des petites terres agricoles est entièrement exclu du champ alors qu'il correspond à environ 60 % des transactions, soit près de 8 % de la surface vendue chaque année.

C'est ainsi qu'il a été jugé nécessaire de remplacer le filtre des 70 ares et de construire à la place un indicateur pour classer les terres en deux groupes, entre les « terres à destination agricole » et les « terres à destination non agricole ». La littérature statistique fournit plusieurs méthodes, dites d'apprentissage supervisé, pour réaliser ce type de classification. Tuffery (2017) ou encore Friedman, Hastie et Tibshirani (2009) ont établi un panorama de ces méthodes dans leur ouvrage, respectivement « Data Mining et statistique décisionnel, la science des données » et « The Elements of Statistical Learning ». L'apprentissage supervisé consiste à entraîner un modèle de classification à partir d'un échantillon de données où l'on connaît de façon sûre la classe d'appartenance. L'objectif est ensuite d'utiliser ce modèle pour extrapoler les résultats aux observations dont la classe d'appartenance est inconnue.

Le plan de cet article est le suivant. Ce préambule se poursuit par la section 2 « Les données de l'étude ». Il s'agit en particulier d'identifier les terres dont la destination est connue afin ensuite de construire, dans la section 3, l'échantillon d'apprentissage et de test. La section 4 décrit les fondements théoriques des méthodes d'apprentissage mises en œuvre. La section 5 présente les critères de fiabilité utilisés pour comparer les modèles. Enfin, une synthèse des résultats est réalisée dans la section 6.

2. Les données de l'étude

Notre étude s'appuie sur l'ensemble des notifications de ventes que les notaires sont tenus d'adresser aux sociétés d'aménagement foncier et d'établissement rural (Safer) en vertu du code rural et de la pêche maritime. Ces informations couvrent l'ensemble du marché des espaces naturels (agricoles et forestiers).

Une question pratique est la définition du périmètre de l'étude. Pour tenir compte de l'hétérogénéité des départements en termes de marchés fonciers, nous mettons au point des modèles de classification département par département, de façon indépendante. Nous nous sommes de plus limité aux départements présentant au minimum 300 transactions sur une période de 3 ans, entre 2014 et 2016, avec au minimum 150 terres qui restent à usage agricole et 150 terres qui changent d'usage. En effet, pour être pertinente, l'analyse doit reposer sur un nombre suffisant de transactions. *In fine*, 56 départements ont été retenus, couvrant 74 % de l'ensemble des transactions de terres et prés agricoles non bâtis entre 2014 et 2016.

2.1. La variable à prédire

Comme expliqué précédemment, la destination de la terre est erronée dans la base de données, elle mélange en fait les terres qui vont rester à vocation agricole et les achats de loisirs par des particuliers ou ceux de pure convenance. On ne peut considérer comme sûre que la modalité « non agricole certaine » de la destination de la terre (cf. tableau 1).

Puisqu'elles sont bien classées, les terres à vocation « non agricole certaine » vont constituer le groupe *non agricole* sur lequel les modèles de classification vont être entraînés. Pour construire le groupe *agricole*, nous choisissons de nous limiter aux terres dites à destination « agricole incertaine » qui ont été acquises par des agriculteurs en activité, c'est-à-dire non retraités et ayant moins de 65 ans. Ainsi l'échantillon où les terres ont une destination connue se compose de 51 110 terres, dont 62 % qui restent à vocation agricole (cf. tableau 2).

Un comptage du nombre de transaction retenues avec le filtre actuel des 70 ares est également donné dans le tableau 1. Dans le processus actuel, les destinations incertaines et inconnues sont considérées comme des destinations agricoles si les terres ont une surface supérieure à 70 ares.

Tableau 1. Nombre de transactions selon la variable « destination du fonds »

Modalités de la variable erronée « Destination du fonds »	**Nombre de transactions entre 2014 et 2016** France entière	**Nombre de transactions entre 2014 et 2016 avec le filtre des 70 ares** France entière
Destination inconnue	22 779	5 766
Destination agricole incertaine	57 978	33 771
Destination non agricole incertaine	44 034	8 683
Destination non agricole certaine	22 635	0
Ensemble	147 426	52 238

Tableau 2. Nombre de transactions selon la destination du fonds après recodage des modalités

Nouvel indicateur de la destination du fonds	Nombre de transactions entre 2014 et 2016 France entière	Nombre de transactions entre 2014 et 2016 56 départements retenus dans l'étude
Destination agricole	42 185	31 727
Destination non agricole	22 635	19 383
Destination inconnue	82 608	58 648
– initialement inconnue	*22 779*	*14 192*
– initialement loisirs	*44 034*	*34 900*
– initialement agricole avec un acquéreur non agriculteur ou ayant plus 65 ou retraité	*15 793*	*9 556*
Ensemble	147 426	109 758

2.2. Les prédicteurs

Les variables entrant dans les modèles ont été choisies pour leur impact possible sur la destination du fonds. Les notifications transmises par les Safer fournissent le montant de la transaction ainsi que les caractéristiques propres des terres échangées. Ces données sont enrichies de différentes variables de localisation venant de sources externes, comme la distance de la commune à la ville la plus proche, ou encore la part de surface artificialisée dans la commune. L'ensemble des variables caractérisant les terres sont listées dans le tableau 3 ci-dessous.

Tableau 3. liste des variables retenus pour prédire la destination du fonds

Sources	Variables
Terres-d'Europe Scafr	Prix à l'hectare du fonds (en logarithme)
Terres-d'Europe Scafr	Surface du fonds (en logarithme)
Terres-d'Europe Scafr	Nature cadastrale dominante du fonds :'*Terre', 'Pré' ou 'Terre et Pré'*
Terres-d'Europe Scafr	Zonage en petites régions agricoles regroupées, aussi appelé « nouvelles régions agricoles » (zonage infra-départemental) : Bocage, Prairies humides, etc.
Terres-d'Europe Scafr	Part de surface artificialisée dans la commune (en logarithme)
SSP	Nomenclature OTEX principale dans la commune (Orientation technico-économique des exploitations)
DATAR INRA CESAER/ UFC-CNRS ThéMA/ Cemagref DTMA METAFORT	Typologie du littoral : *Littoral artificialisé urbain et périurbain* *Littoral de type rural méditerranéen* *Littoral de type rural atlantique*
DATAR INRA CESAER/ UFC-CNRS ThéMA/ Cemagref DTMA METAFORT	Typologie des montagnes : *La haute et moyenne montagne résidentielle et touristique* *La moyenne montagne agricole ou industrielle* *La montagne urbanisée*
INSEE	Densité de population dans la commune (hab/km²) (en logarithme)
DGCL	Potentiel financier de la commune en euros/hab (élément de mesure de la richesse théorique d'une commune) (en logarithme)
INRA	Distance de la commune à la commune centre du pôle urbain le plus proche

La question du découpage en classes des variables explicatives continues se pose. Il peut en effet être bénéfique de discrétiser une variable explicative si le phénomène que l'on cherche à prédire n'est pas une fonction monotone de cette variable.

Nous avons représenté, ci-dessous, la part de destination agricole par vingtile pour chacune des variables continues. La vocation agricole est visiblement une fonction strictement monotone de la surface de la terre. Cela ne semble en revanche pas être le cas du prix à l'hectare, nous préférons néanmoins conserver cette variable telle quelle. En effet, seules les terres de plus de 10 000 euros à l'hectare et celles de moins de 1000 euros à l'hectare ont un profil particulier, hors ces tranches ne comportent pas suffisamment d'effectifs pour former une classe. Nous discrétisons en revanche les autres variables en regroupant les quantiles qui sont proches du point de vue du pourcentage de terres à vocation agricole : Densité de population : [0 ; 50[, [50 ; 106[et [106 ; +∞[; Distance au pôle urbain : [0 ; 14[, [14 ; 20[et [20 ; +∞[; Part de surfaces artificialisées en % : [0 ; 3[, [3 ; 6[et [6 ; +∞[; Potentiel financier de la commune % : [0 ; 700[, [700 ; +∞[

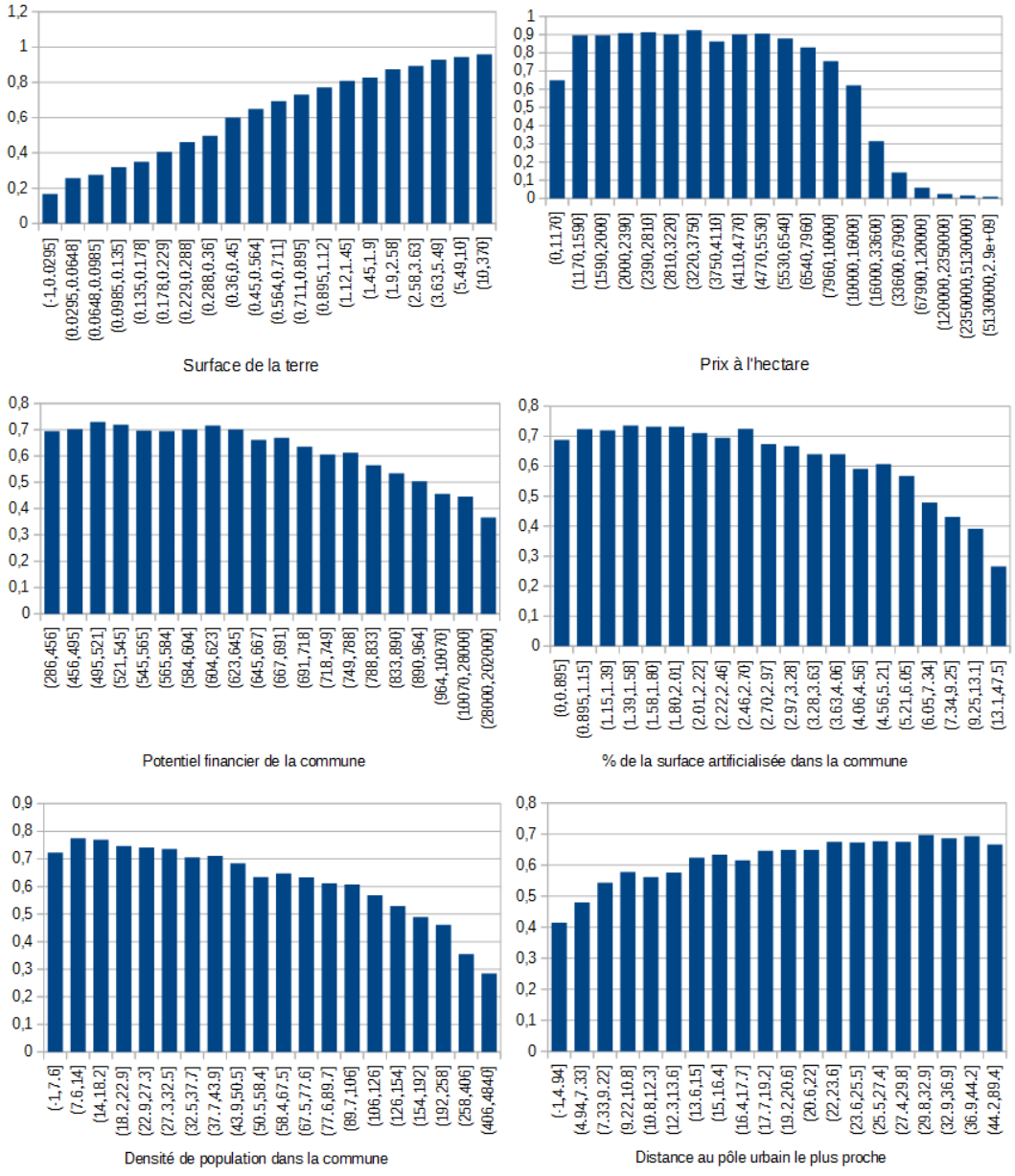

Illustration 1: Part de terres à destination agricole selon les quantiles des variables

3. Échantillon d'apprentissage et de test

L'échantillon où les terres ont une destination connue doit servir au processus d'apprentissage des modèles de classification. Ce processus minimise les erreurs de classement de manière à ce que le modèle puisse être extrapolé à de nouvelles données. On divise l'échantillon où la destination est connue en deux sous-échantillons : un premier dit d'apprentissage et un second dit de test (ou de validation). Le modèle est entraîné sur l'échantillon d'apprentissage et validé sur l'échantillon de test. L'échantillon de test permet de comparer les performances des différents modèles au regard de critères de fiabilité que l'on présentera dans la section 5.

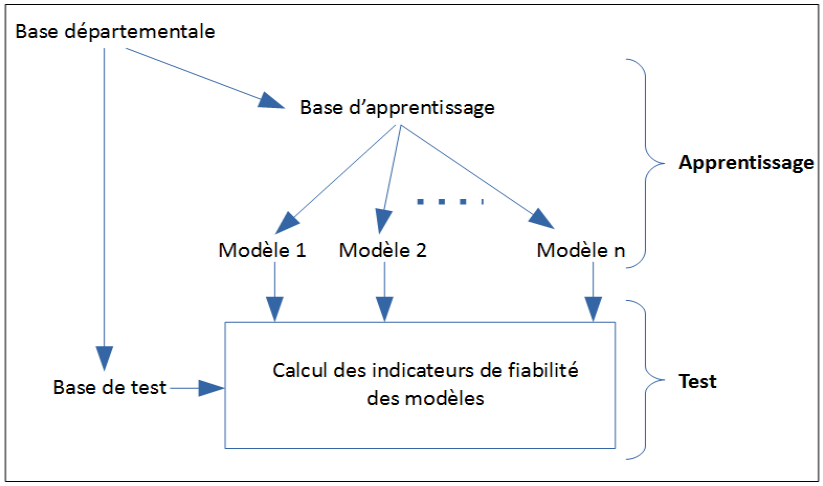

Illustration 2: Schéma d'évaluation des modèles

3.1. Échantillon d'apprentissage

Nous construisons une base d'apprentissage pour chacun des départements du champ de l'étude. Pour cela, nous tirons aléatoirement et sans remise, dans chaque département, 66 % des transactions parmi celles dont on connaît la destination. Puisque dans l'échantillon il y a 300 transactions au minimum par département (dont 150 par destination), il y a donc 200 transactions au minimum par département dans l'échantillon d'apprentissage (dont 100 au minimum par destination).

La plupart des algorithmes d'apprentissage reposent sur l'hypothèse que la base d'apprentissage est un échantillon représentatif de la population sur laquelle le modèle sera appliqué. Or cette hypothèse n'est pas réaliste dans notre cas, il n'y a aucune raison de penser que la part des terres à destination agricole est identique dans l'échantillon d'apprentissage et dans l'échantillon des terres à destination inconnue. On comprend bien que si, en apprentissage, 99 % des terres sont à destination agricole, alors le modèle pourra difficilement faire mieux que le 1 % d'erreurs obtenu en classant automatiquement toutes les terres en destination agricole.

Dans nos données, la part de terres qui restent à destination agricole dans la base d'apprentissage varie selon le département, entre 11 % en Haute-Savoie et 86 % dans les Côtes-d'Armor. Nous proposons donc de rééquilibrer l'échantillon d'apprentissage de chaque département. Pour ce faire, nous augmentons artificiellement le nombre de transactions de la classe minoritaire de façon à obtenir à chaque fois 50 % de destinations agricoles. La solution consistant a simplement dupliquer les observations de la classe minoritaire n'a pas été envisagée. Elle causerait en effet un sur-apprentissage qui dégraderait les capacités prédictives des modèles. Deux méthodes plus avancées sont mises en œuvre : SMOTE (*Synthetic Minority Over-sampling Technique*) et ROSE (*Random Over-Sampling Examples*) :

- La méthode SMOTE (*Synthetic Minority Over-sampling Technique*) combine un sur-échantillonnage de la classe minoritaire et un sous-échantillonnage de la classe majoritaire. Pour chaque observation de la classe minoritaire, une observation synthétique est créée en sélectionnant aléatoirement un point sur le segment défini par l'observation et l'un de ses k plus proches voisins (lui-même étant tiré aléatoirement). Pour les variables de type catégoriel, la sélection s'effectue aléatoirement entre la valeur prise par l'observation et celle prise par son voisin. La classe majoritaire peut à l'inverse être réduite au regard de la distribution globale.
- La méthode ROSE (*Random Over-Sampling Examples*) consiste à créer des observations bruitées par une approche « *smooth bootstrap* ». Une observation de la base d'apprentissage est sélectionnée, puis de nouvelles observations sont générées dans le voisinage de cette dernière avec la méthode des noyaux. Cela ne s'applique pas aux variables de type catégoriel, la valeur prise par ce type de variables reste inchangée.

Les graphiques nuages de points suivants permettent de visualiser sur un exemple les modifications de l'échantillon avec la méthode SMOTE et avec la méthode ROSE. L'exemple est celui du département de la Côte d'Or où 75 % de terres vendues sont à destination agricole dans l'échantillon d'origine. On représente l'échantillon dans un plan formé par deux variables (le prix à l'hectare du terrain et la distance à la commune centre du pôle urbain le plus proche) préalablement centrées et réduites.

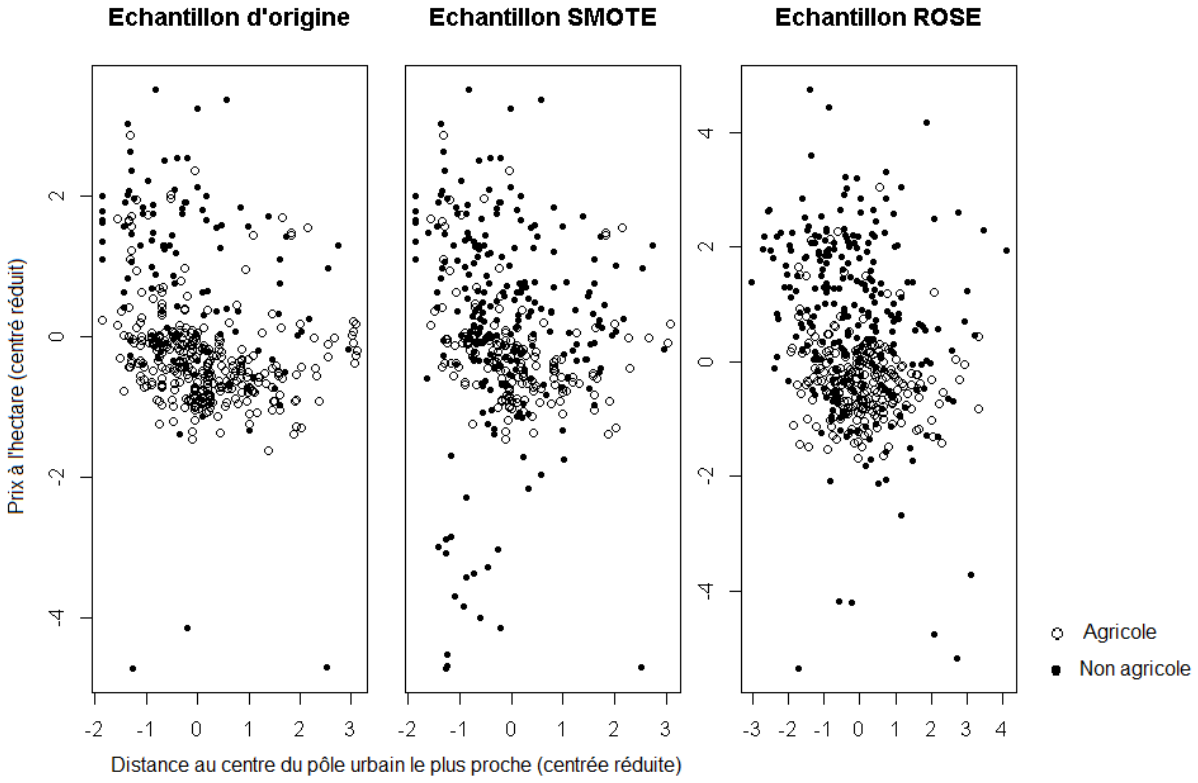

Illustration 3: Modification de l'échantillon par les méthodes SMOTE et ROSE

3.2. Échantillon de test

La base de test est construite département par département à partir du tiers restant des observations dont la destination est connue. Elle contient donc, pour chaque département, au moins 100 observations dont au moins 50 à destination agricole et 50 à destination non agricole. La même base de test est utilisée pour comparer les performances des modèles, que ceux-ci soient entraînés sur la base d'apprentissage brute ou bien sur la base d'apprentissage rééquilibrée SMOTE ou ROSE.

Néanmoins, il existe une asymétrie des classes en test comme nous l'avons vu en apprentissage. Par conséquent, si des modèles font des erreurs sur cette classe, ils ne seraient que très peu pénalisés. Ainsi, pour ne pas donner plus d'importance à l'une ou l'autre des classes, nous préférons réduire la classe majoritaire de manière à obtenir pour chaque département une répartition des classes à parts égales. Un simple tirage aléatoire est réalisé dans la classe majoritaire pour éliminer les observations en trop. En faisant cela, on a toujours dans la base de test au moins 50 observations à destination agricole et 50 observations à destination non agricole dans chaque département.

4. Les modèles de classification

Dans cette section nous présentons brièvement les méthodes de classification mises en œuvre.

4.1. Régression logistique

La régression logistique, réputée comme étant fiable, est la première méthode de classement que nous avons envisagée. La difficulté ici est de détecter les variables à inclure au modèle afin que ses capacités prédictives soient les meilleures. Nous proposons deux modèles *logit*, l'un obtenu à partir des variables présentées au début de l'étude, l'autre obtenu après avoir découpé en classes certaines de ces variables comme nous l'avons vu. Dans les deux cas une sélection ascendante *forward* basée sur l'AIC (*Akaike information criterion*) est pratiquée.

4.2. *Support Vector Machine* (SVM)

Les « *support vector machine* », appelé aussi « machines à vecteurs de support » ou encore « séparateurs à vaste marge » sont des outils récents faisant l'objet d'un grand intérêt théorique. 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 : il s'agit alors de résoudre un problème d'optimisation sous contraintes.

En pratique, il n'est pas toujours possible d'avoir une séparation linéaire parfaite des classes. On autorise donc des erreurs de classement et on introduit un paramètre de pénalisation des erreurs. Ce paramètre doit être réglé afin de trouver un bon compromis entre l'ajustement et la robustesse. Avec un paramètre trop grand, le modèle sera trop précis sur les données et donc se comportera de manière instable sur de nouvelles données. À l'inverse, un paramètre trop petit ne pénaliserait pas suffisamment les erreurs.

Par ailleurs, on peut rechercher une séparation linéaire dans un espace de dimension plus grande muni d'un produit scalaire. L'exemple de la figure ci-dessous montre comment une séparation non linéaire en dimension 2 peut devenir linéaire si on transpose les données dans un espace de dimension 3. La transformation ϕ associée est ici $x=(x_1, x_2) \rightarrow \Phi(x)=(x_1^2, \sqrt{2}\,x_1 x_2, x_2^2)$

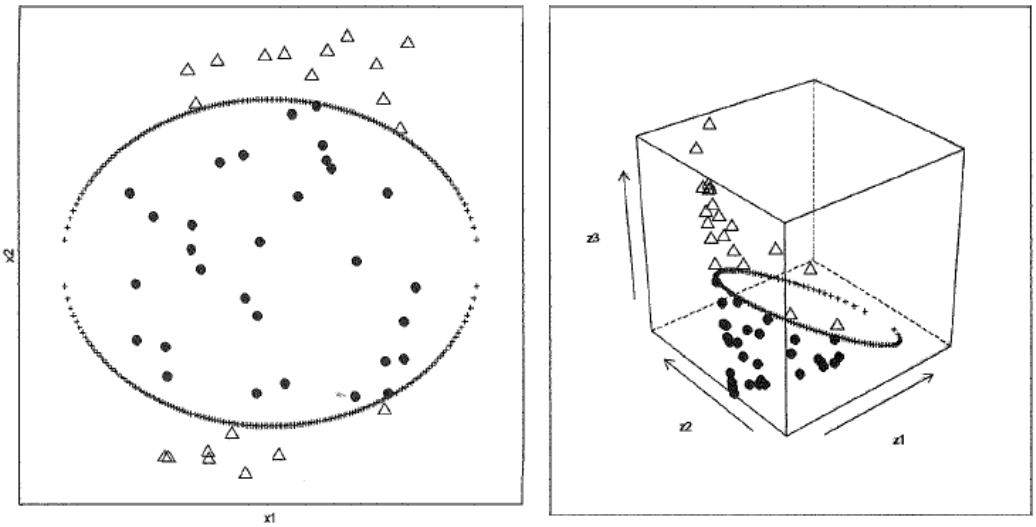

Illustration 4: Exemple de changement d'espace d'une dimension 2 à une dimension 3 (source: Stéphane Tufféry [1])

Tout l'intérêt de cette méthode est de ne pas avoir besoin de connaître expressément la transformation **ϕ**. En effet, lorsqu'on pose le problème d'optimisation, on s'aperçoit que **ϕ** intervient uniquement sous la forme d'un produit scalaire $\Phi(x)\cdot\Phi(x')$, lequel est défini par une fonction noyau $k(x,x')$ continue, symétrique et semi-définie-positive. Il est possible d'utiliser n'importe quelle fonction noyau vérifiant ces propriétés. Parmi celles couramment utilisées, il y a par exemple :

- Le noyau polynomial : $k(x,x')=(x\cdot x'+c)^n$ où *n* est un paramètre à régler ;
- Le noyau radial gaussien : $k(x,x')=\exp(-\gamma\|x-x'\|^2)$ où γ est un paramètre à régler ;
- Le noyau radial laplacien : $k(x,x')=\exp(-\gamma\|x-x'\|)$ où γ est un paramètre à régler ;

Pour notre étude, nous mettons en œuvre le noyau linéaire ainsi que le noyau radial gaussien. Des modèles SVM sont construits en faisant varier le paramètre de pénalisation et le paramètre γ[1]. Le meilleur modèle « SVM linéaire » et le meilleur modèle « SVM noyau radial gaussien » sont ensuite sélectionnés par validation croisée sur l'échantillon d'apprentissage.

4.3. 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 ». À la différence des méthodes précédentes qui cherchent une séparation sous forme d'hyperplan, l'arbre de décision cherche à diviser l'échantillon en rectangles emboîtés les mieux différenciés par rapport à la variable d'intérêt, ici la destination de la terre.

Concrètement, il faut trouver, pour chaque division de l'échantillon, une variable *X* permettant de définir deux nouveaux rectangles, c'est-à-dire une variable *X* et une condition de la forme :

- $X<x_0$ si *X* quantitative ;
- $X\in\{X_{i1},\dots,X_{ij}\}$ si *X* qualitative avec pour modalités $X_1\dots X_k$ et $\{X_{i1},..,X_{ij}\}\subset\{X_1,..,X_k\}$.

Le partage optimal sera celui dont les rectangles sont les plus purs, c'est-à-dire les plus homogènes possibles au sens de la variable d'intérêt[2]. Le critère de pureté que nous avons optimisé est le critère de Gini. Il aurait été possible d'utiliser d'autres critères tels que l'entropie.

Pour obtenir l'arbre final, servant de modèle pour l'extrapolation, il convient *in fine* d'élaguer l'arbre construit. En effet, cet arbre risque d'être sur-ajusté à l'échantillon d'apprentissage, décrivant exagérément bien l'échantillon d'apprentissage mais ayant de faibles capacités de généralisation. Pour élaguer l'arbre construit, on calcule un taux d'erreur de prédiction par validation croisée pour différentes tailles de l'arbre : l'arbre est alors élagué au niveau offrant l'erreur minimale.

La figure suivante représente l'arbre obtenu pour un des départements, la Haute-Garonne. Il ne fait intervenir que le prix et la surface. Son interprétation est simple : les terres dont le prix à l'hectare est inférieur à 13 000 euros sont classés en « destination non agricole ». Parmi les terres de plus de 13 000 euros à l'hectare, ce sont celles qui ont une surface supérieure à 16 ares qui sont classées en agricoles, ainsi que celles qui ont une surface inférieure à 16 ares avec un prix à l'hectare supérieur à 1050 euros.

[1]Dans le cas linéaire, les paramètres de pénalisation testés sont : 0,001 / 0,01 / 0,1 / 1 / 10 ; Dans le cas du noyau radial gaussien, les paramètres de pénalisation testés sont : 0,1 / 0,2 / 0,3 /… / 10 et les paramètres γ testés : 0,001 / 0,01 / 0,1 / 1.

[2]Un rectangle est partagé que s'il possède suffisamment d'observations. Le nombre d'observations minimal doit être fixé par l'utilisateur. Nous avons testé 10 et 20 observations par rectangle.

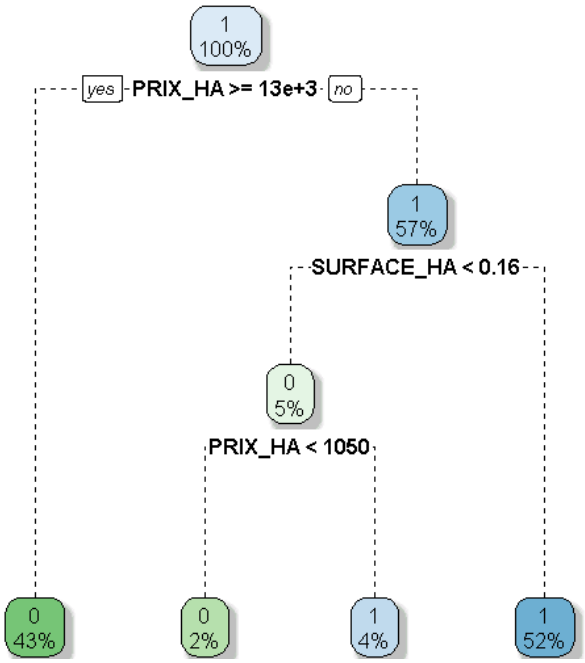

Illustration 5: Arbre de décision pour la Haute-Garonne (package R rpart.plot)

4.4. Forêt aléatoire

Les arbres de décision ne peuvent pas toujours être généralisés correctement, en raison notamment des seuils qui sont très dépendants de l'échantillon d'apprentissage sur lequel ils ont été estimés. Un raffinement de la méthode consiste à construire plusieurs arbres de façon aléatoire afin d'obtenir une forêt aléatoire.

Concrètement, la méthode consiste à générer un grand nombre d'échantillons bootstrap à partir de l'échantillon d'apprentissage[3] et à construire un arbre de décision sur chaque échantillon. De plus, chaque nœud de chaque arbre est construit de façon aléatoire : la variable qui permet de partager le nœud n'est pas cherchée sur l'ensemble des variables, mais sur un sous-ensemble tiré aléatoirement[4]. Par ailleurs, aucun des arbres construits n'est élagué et, *In fine*, la prédiction est obtenue en agrégeant tous les arbres au vote majoritaire.

L'intérêt de cette méthode est de réduire l'erreur en réduisant la variance. Le biais n'est pas réduit puisque la « moyenne » de tous les arbres a la même espérance que chacun des arbres.

Par ailleurs, les forêts aléatoires fournissent une mesure de l'importance des variables dans la classification :

- Tout d'abord on mesure pour chaque arbre l'erreur *Out Of Bag* (erreur sur les observations qui n'appartiennent pas à l'échantillon sur lequel l'arbre a été construit) ;
- On permute aléatoirement les valeurs de la variable explicative sur les observations *Out Of Bag* ;
- On mesure une nouvelle fois le taux d'erreur, et on regarde la différence. Plus elle est élevée et plus la variable était importante dans la construction de l'arbre.
- Cette différence est normalisée et sa valeur moyenne sur la forêt est calculée.

Le diagramme suivant présente cette mesure pour un département particulier, la Haute-Garonne.

[3]Nous avons généré 1000 échantillons pour garantir la convergence du taux d'erreur.
[4]Les préconisations de la littérature sont d'utiliser un nombre de variables égal à la partie entière de la racine carré du nombre total de variables, soit 3 dans notre cas.

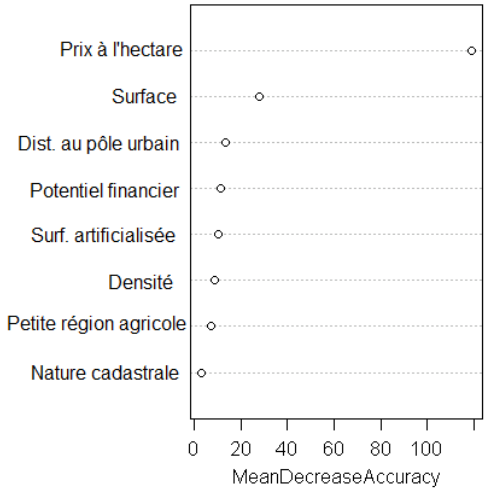

Illustration 6: Mesure de l'importance des variables par la foret aléatoire (en Haute-Garonne)

4.5. Boosting d'arbres

Une dernière méthode, le boosting, a été mise en œuvre. Il s'agit aussi d'une méthode d'agrégation de modèles, mais à la différence des forêts aléatoires, elle permet de réduire à la fois la variance et le biais de prévision.

Le boosting consiste à appliquer un même classifieur, ici un arbre de décision, sur un échantillon d'apprentissage adaptatif. C'est un processus itératif qui donne, à chaque étape, plus de poids aux observations mal prédites à l'étape précédente.

C'est plus précisément l'algorithme Discrete AdaBoost que nous avons mis en œuvre. Il est implémenté en R dans le package *ada* :

1. Tout d'abord on initialise les poids des *N* observations de l'échantillon d'apprentissage $p_i = 1/N$;
2. Puis, pour m=1 à M (M est le nombre d'itérations) :
   - On ajuste le classifieur $f_m(x)$ sur un sous-échantillon de l'échantillon d'apprentissage dans lequel les observations sont tirées (avec remise) selon les poids. Ce classifieur vaut +1 si l'observation x_i est bien classée, et -1 sinon.
   - On calcule le taux d'erreur ϵ_m en tenant compte du poids.
   - On calcule le poids α_m du classifieur : $\alpha_m = \log((1-\epsilon_m)/\epsilon_m)$.
   - On peut avoir à multiplier α_m par un paramètre de pénalisation λ afin de réduire diminuer l'intensité du mécanisme adaptatif. La littérature utilise couramment un paramètre de pénalisation valant 0,1 ou 0,01. Plus ce paramètre est petit, plus le risque de sur-apprentissage est faible, mais plus la convergence est longue[5].
   - On multiplie les poids p_i de chaque observation par $\exp(\alpha_m)$.
3. Le classifieur final est : $sign(\sum_m \alpha_m f_m(x))$.

Le programme calcule automatiquement le nombre d'itérations optimal par la méthode Out Of Bag.

[5]Compte tenu des temps de calcul élevés, nous avons utilisé paramètre de pénalisation égal à 0,1.

5. Fiabilité d'un modèle

Une fois les modèles de classification entraînés sur l'échantillon d'apprentissage, nous les appliquons sur l'échantillon de test pour comparer leurs performances. Nous décrivons ci-dessous les critères de fiabilité utilisés dans l'étude.

5.1. Matrice de confusion

Traditionnellement, la performance d'un modèle de classification est mesurée en utilisant des critères dérivés de la matrice de confusion (tableau 4). Ainsi, pour chaque observation de notre base que l'on doit classer, quatre scénarios sont possibles :

- Vrai positif (VP) : la terre est classée correctement en destination agricole ;
- Vrai négatif (VN) : la terre est classée correctement en destination non agricole ;
- Faux positif (FP) : la terre est classée à tort en destination agricole ;
- Faux négatif (FN) : la terre est classée à tort en destination non agricole.

Tableau 4. Nombre de transactions selon la variable mal codée « destination du fonds »

	Prédit : agricole	Prédit : non agricole
Réel : agricole	VP	FN
Réel : non agricole	FP	VN

On définit alors la sensibilité comme étant la proportion de vrais positifs correctement identifiés par le modèle et la spécificité comme étant la proportion de vrais négatifs correctement identifiés par le modèle :

$$\textbf{\textit{Sensibilité}} = \frac{\textbf{\textit{VP}}}{\textbf{\textit{VP}} + \textbf{\textit{FN}}}$$

$$\textbf{\textit{Spécificité}} = \frac{\textbf{\textit{VN}}}{\textbf{\textit{VN}} + \textbf{\textit{FP}}}$$

Par ailleurs, des indicateurs que l'on appelle « rappel » et « précision » permettent de mesurer la qualité de la classification en se concentrant sur l'une des deux classes jugée plus importante, dans notre cas la destination « agricole ». Le rappel est égal à la sensibilité tandis que la précision est la proportion de vrais positifs parmi ceux identifiés comme positifs.

$$\textbf{\textit{Rappel}} = \textbf{\textit{Sensibilité}} = \frac{\textbf{\textit{VP}}}{\textbf{\textit{VP}} + \textbf{\textit{FN}}}$$

$$\textbf{\textit{Précision}} = \frac{\textbf{\textit{VP}}}{\textbf{\textit{VP}} + \textbf{\textit{FP}}}$$

L'indicateur synthétique généralement associé est la F-mesure, la moyenne harmonique de ces deux grandeurs.

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

5.2. La courbe ROC

Les modèles de classification font généralement intervenir une mesure quantitative *s* au-delà de laquelle on est déclaré positif et au-dessous de laquelle on est déclaré négatif. En faisant varier ce seuil *s*, on fait varier la sensibilité et la spécificité.

Graphiquement, on représente la courbe ROC (Receiver Operating Characteristic) sous la forme d'une courbe qui donne le taux de vrais positifs (la sensibilité) en fonction du taux de faux positifs (1 – Spécificité). Elle est obtenue en faisant varier le seuil *s :* plus on l'augmente, plus sensibilité diminue et plus la spécificité augmente (cf. illustration 2). En principe la courbe ROC est située au-dessus de la première bissectrice, laquelle correspond à un classement au hasard des observations. Le point (0 ; 1) du graphique représente le scénario idéal où toutes les observations positives sont bien classées et aucune observation négative est mal classée.

L'indicateur synthétique associé à la courbe ROC est la surface située sous la courbe, c'est l'AUC (*Aera Under the Curve*). Un modèle de classification est performant si l'AUC est proche de 1. À l'inverse, un modèle de classification n'est pas discriminant si l'AUC est proche de 0,5.

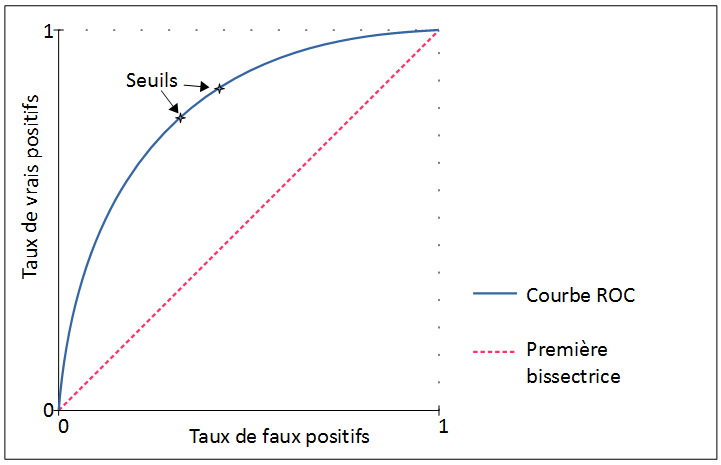

Illustration 7: Courbe ROC

6. Résultats

Les modèles sont construits à partir du jeu de variables présenté au tableau 3 de la section 2. Un jeu de variable supplémentaire, utilisant les variables discrétisées, est testé pour les méthodes Logit et SVM. Il est inutile d'utiliser les variables discrétisées pour les arbres de décision (et les agrégations d'arbres), car ils savent découper des variables en classes.

Les résultats sont rassemblés dans le tableau 5. Les valeurs du tableau sont les moyennes obtenues sur les 56 départements.

Tableau 5. Synthèse de la performance des modèles.

Modèle	Rééquilibrage	Sensibilité	Spécificité	Précision	F-mesure	AUC
LOGIT	Non	0,91	0,74	0,79	0,84	0,90
	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	0,90
	ROSE	0,84	0,80	0,81	0,83	0,88
	SMOTE	0,84	0,82	0,82	0,83	0,89
SVM LINEAIRE	Non	**0,93**	0,73	0,79	0,85	0,90
	ROSE	0,87	0,79	0,81	0,84	0,89
	SMOTE	0,87	0,82	0,83	0,85	0,89
SVM LINEAIRE (variables discrétisées)	Non	**0,93**	0,73	0,79	0,85	0,90
	ROSE	0,87	0,80	0,81	0,84	0,89
	SMOTE	0,86	0,82	0,83	0,84	0,89
SVM NOYAU 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	0,90
SVM NOYAU RADIAL GAUSSIEN (var. discrétisées)	Non	**0,93**	0,77	0,81	0,86	0,92
	ROSE	0,84	0,78	0,80	0,82	0,87
	SMOTE	0,83	0,81	0,82	0,82	0,90
ARBRE DE DECISION	Non	0,92	0,80	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 ALEATOIRE	Non	**0,93**	0,83	0,85	**0,89**	**0,94**
	ROSE	0,90	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	0.90
	SMOTE	0,89	0,86	0,87	0,88	0,92

Les résultats révèlent en premier lieu une bonne capacité de prédiction de nos modèles, avec un AUC entre 0,87 et 0,94. Les terres à destination agricole sont en général mieux identifiée que celles à destination non agricole, la sensibilité est en moyenne de 0,88 et la spécificité est en moyenne de 0,80. On trouve en outre que l'arbre de décision est toujours amélioré par les techniques d'agrégation de modèles (le *boosting* et les forêts aléatoires). Ce sont d'ailleurs ces techniques qui obtiennent globalement les meilleurs scores pour prédire la destination de la terre. Nous avons représenté ci-dessous leur courbe ROC sur un département en particulier, la Haute-Garonne. Dans cet exemple, l'AUC est de 0,97 pour la forêt aléatoire et de 0,96 pour le boosting.

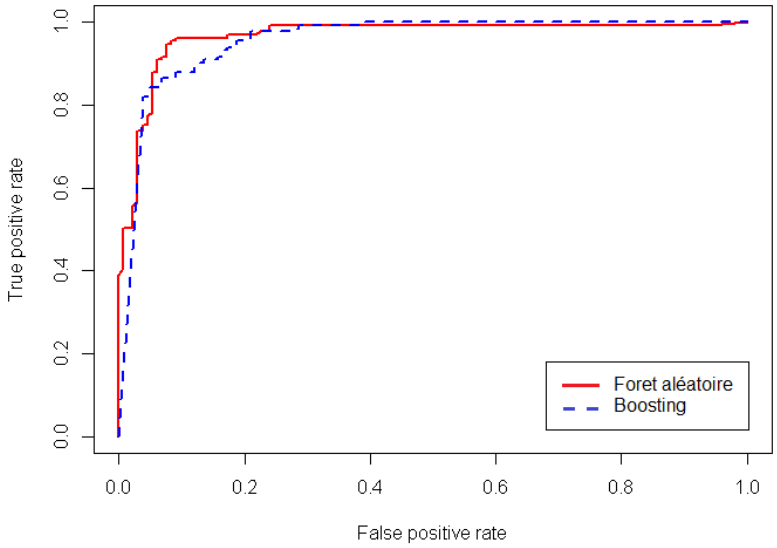

Illustration 8: Courbe ROC pour le département de la Haute Garonne

Par ailleurs, pour pouvoir publier des prix les plus justes possibles, il convient de ne pas classer en « destination agricole » les terres susceptibles d'avoir un autre usage (i.e. les faux positifs). L'analyse montre que le rééquilibrage SMOTE améliore le taux de vrais positifs parmi ceux identifiés comme positifs (critère de précision). Le rééquilibrage ROSE ne semble, en revanche, pas adapté : il dégrade la sensibilité et l'AUC, et ne fait guère augmenter les autres critères.

Enfin, le découpage en classes de certaines variables continues ne semble pas avoir modifié les capacités prédictives des modèles logistiques et des SVM. Cela peut s'expliquer par le fort pouvoir discriminant des deux variables que nous n'avons pas modifiées : le prix à l'hectare et la surface.

7. Conclusion et perspectives

Nous avons mis en œuvre des méthodes d'apprentissage supervisée pour imputer la valeur de la destination de la terre lorsqu'elle est inconnue ou incertaine : la traditionnelle régression logistique *logit*, les arbres de décision, ainsi que des méthodes plus récentes et plus sophistiquées, les *Support Vector Machine* (séparateurs à vaste marge) et des méthodes d'agrégation de modèle (les forêts aléatoires et le *Boosting*).

Finalement, ce sont les méthodes d'agrégation d'arbres qui ont eu les meilleurs résultats. Ces dernières permettent en particulier de corriger les problèmes de variance qui rendent instables les prédictions de l'arbre de décision. Les *Support Vector Machine* et les modèles logistiques ont donné de moins bons résultats. Poursuivre plus avant l'examen des variables explicatives pourrait peut-être permettre d'améliorer leurs capacités prédictives. D'autres fonctions noyaux pourraient également être testés pour les *Support Vector Machine.*

Rappelons que nous envisageons de mettre en production une de ces méthodes. Or, la complexité et temps de calcul souvent long des *Support Vector Machine* et du *Boosting* ne justifient pas, au vu des résultats, leur utilisation.

Par ailleurs, nous avons montré l'intérêt de rééquilibrer l'échantillon avec l'approche SMOTE (*Synthetic Minority Over-sampling Technique)*. Elle améliore la spécificité et la précision, même si elle dégrade la sensibilité.

Il s'agira ensuite d'analyser dans quelle mesure ces résultats restent valables pour les départements que nous avons exclus du périmètre de l'étude. Il est toutefois prévu, dans les cas qui posent des difficultés (par exemple si le nombre de transactions est insuffisant) de faire appel à des dires d'expert pour établir les prix.

Bibliographie

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

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

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

[4] Cornuéjols, « une nouvelle méthode d'apprentissage : Les SVM. Séparateurs à vaste marge », 2002

[5] Menardi, Torelli, « Training and assessing classification rules with unbalanced data », 2014

[6] « Le prix des terres – Analyse des marchés fonciers ruraux 2016 », *publication FNSafer*, 2017