# Détection de boiteries chez les vaches à partir de données de suivi

Frédéric Logé[1], Yanis Amirou[2], Florian Bourgey[3],
Sean Hellingman[4], Malo Huard[5] et Solène Thépaut[6]
[1] *Polytechnique,* frederic.logemunerel@gmail.com
[2] *Ecole Normale Supérieure,* yanis.amirou@ens.fr
[3] *Polytechnique,* florian.bourgey@polytechnique.edu
[4] *Wilfrid Laurier University,* hell5140@mylaurier.ca
[5] *Université Paris-Sud,* malo.huard@u-psud.fr
[6] *Université Paris-Sud,* solene.thepaut@u-psud.fr

**Résumé.** La boiterie est une pathologie importante par sa prévalence et ses conséquences, tant sur la santé de la vache qu'économiquement pour l'éleveur. Dans ce travail nous étudions la détectabilité de cette pathologie à partir de données de suivi des vaches, nous indiquant leur activité : rumination, assise/debout, ingestion, etc. L'essentiel de notre travail s'est porté sur la création de variables et sur les modèles prédictifs utilisés. Nous montrons que les données de suivi permettent de prédire la présence d'une boiterie avec une AUC à 80% à l'aide de méthodes de Boosting.

**Mots-clés.** Classification / Apprentissage, Etude de cas, Statistique pour l'industrie

**Abstract.** Lameness is an important pathology because of its prevalence and its consequences, both on the health of the cow and economically for the farmer. In this work we study the detectability of this pathology based on cow monitoring data, indicating their activity : rumination, sitting/standing, ingestion, etc. Most of our work focused on feature engineering and the predictive models used. We show that monitoring data can be used to predict the presence of lameness with an AUC of 80% using Boosting methods.

**Keywords.** Classification / Learning, Case Study, Statistics for industry

# 1 Introduction

## 1.1 Contexte, données, enjeu

Le groupe Seenergi est un acteur majeur du monde agricole. Médria Solutions, une de ses filiales, utilise des colliers enregistrant le comportement des vaches à une fréquence de cinq minutes : ingestion, rumination, repos, debout, sur-activité ou autre activité. En parallèle de cette base de suivi, les éleveurs gèrent les carnets de santé et informations de parage de leurs vaches. En possession de ces données, Julie Dewez et Florine Hardy, représentant le groupe Seenergi à la SEME Orsay 2019, ont posé la question suivante :

**Les boiteries sont-elles détectables au vu des données recueillies ?**

Détecter ces évènements améliorerait le bien-être des animaux, contribuerait à la réduction de l'usage d'antibiotiques et enfin améliorerait les performances de l'élevage de façon globale.

## 1.2 Formalisation de l'objectif

Soit un ensemble de paires $(X_i, Y_i)_{i\in\{1,\dots,n\}}$, où l'index $i$ correspond à une clé (`animal`, `date`), $Y_i$ vaut 1 si à cette date une boiterie a été détectée chez l'animal, 0 sinon, et $X_i$ est un ensemble d'information connue au plus tard à cette date. Dans la suite nous déterminerons $X_i$ à partir des observations des $k$ derniers jours de la base de suivi.

Il s'agit là d'un problème de classification binaire, qui se formalise classiquement comme le problème d'optimisation suivant :

$$\min_{\hat{\mathbb{P}}} \sum_{i\in\texttt{train}} \Big[-Y_i \log(\hat{\mathbb{P}}(Y_i = 1|X_i)) - (1 - Y_i) \log(1 - \hat{\mathbb{P}}(Y_i = 1|X_i))\Big] \tag{1}$$

où $\texttt{train} \subset \{1, \dots, n\}$ correspond aux indexes des éléments de cet ensemble d'apprentissage. Le complément sera cet échantillon de test : $\texttt{test} = \{1, \dots, n\}\backslash\texttt{train}$ qui nous permet d'évaluer la performance des prédicteurs.

## 1.3 Défis

Le premier défi qui se pose est la constitution du jeu de données et particulièrement de la partition `train`/`test`. Nous avons commencé par regrouper les identifications de boiterie entre les bases de Parage et carnet de Santé qui nous fournissent les clés `(animal,date)` pour lesquelles $(Y = 1)$. Puis, nous avons sélectionné au hasard des dates pour des animaux n'ayant jamais eu de boiterie au vu des bases pour constituer les exemples dits "négatifs" $(Y = 0)$. Enfin, pour toutes les clés identifiées, nous avons récupéré les $k = 5$ jours précédents de données de suivi, ainsi que les données historiques du cheptel (nombre de vaches suivies, prévalence des boiteries, etc.). Ces données nous permettront de construire les variables $X$. Du côté informatique, il y a eu également beaucoup de travail pour que les jointures entre tables soient possibles : les clés existent mais certaines ont des préfixes, espaces supplémentaires, suffixes, etc. ; et plusieurs doublons étaient présents.

Une fois cela réglé, il reste principalement deux challenges à gérer :

1. Les boiteries sont des évènements assez peu fréquents : nous avons 15% de cas positifs, l'échantillon devrait être rééquilibré.
2. Les données de suivi sont échantillonnées toutes les cinq minutes, soit 288 observations par jour pour chacune des six variables mentionnées dans l'introduction. Vu le faible nombre de cas positifs, il faut trouver un moyen de réduire la dimension.

Nous présentons nos approches pour attaquer ces challenges dans la section suivante, spécifiquement dans les sous-sections 2.3 et 2.2 respectivement.

# 2 Méthodologie

## 2.1 Eléments de littérature

Les porteurs de projet nous ont grandement pourvu en littérature sur les boiteries. Voici quelques observations générales.

Les vaches boiteuses produisent moins de lait et ont une qualité de vie plus basse. Elles sont généralement moins actives et bien qu'elles mangent autant en quantité que les autres elles se déplacent moins souvent aux endroits de rationnement, à cause de la douleur.

Ces observations, que nous avons retrouvées à l'aide de Modèles à Effets Mixtes sur les données fournies, sont des guides précieux pour la construction de variables prédictives pertinentes.

## 2.2 Création de variables

Les cinq jours de données de suivi ont été aggrégées de différentes manières.

**Indicateurs usuels** Moyennes, variances et corrélations entre les six séries ont été calculées sur la période des cinq jours ainsi que sur les 2.5 premiers et 2.5 derniers jours. Nous avons également récupéré les prévalences de boiterie pour chaque animal et son cheptel.

**Création et sélection automatique** Les auteurs du package **python** `tsfresh` proposent (a) une vaste gamme d'aggrégats pour des séries temporelles, et (b) une sélection de ces aggrégats basée sur une procédure de contrôle du taux de fausse détection, de type Benjamini-Hochberg. L'approche est détaillée dans [Christ et al., 2016].

**Représentation en espace d'état** En combinant les six variables, nous nous sommes rendus compte que sept modalités occupaient 90% des observations. Après avoir rajouté une catégorie **Autre** pour les 10% restants, nous nous sommes intéressés aux transitions entre ces huit états et les avons segmentées avec l'aide du package **R** `ClickClust` [Melnykov, 2016].

## 2.3 Echantillonnage

Pour rappel, ce jeu de données contient seulement 15% de cas positifs. A cause de la formulation du problème d'optimisation (cf équation 1), nos classifieurs auront tendance à prédire la class 0 trop souvent.

Nous avons plusieurs façons de gérer ce problème. Certains algorithmes nous permettent de spécifier une pondération à chaque observation dans le problème d'optimisation. Si ce n'est pas possible nous pouvons faire du tirage aléatoire au sein du jeu `train`, pour rééquilibrer la balance entre cas positifs et négatifs. Deux possibilités se présentent :

— du sous-échantillonnage, en enlevant aléatoirement les observations parmi les cas négatifs ;
— et du sur-échantillonnage, en tirant avec remise des observations parmi les cas positifs.

### 2.4 Classifieurs sur étagère

Nous avons considéré quatre classifieurs [Trevor et al., 2009] : la régression logistique, les forêts aléatoires, *Extreme* et *Light Gradient Boosting* et un modèle ensembliste pour terminer, lequel combinera les prédictions des modèles précédents.

## 3 Résultats

Basé sur cet échantillon test, nous avons calculé les prédictions obtenues avec un ensemble de méthodes combinant création de variables, méthode d'échantillonnage et différents classifieurs. Nous obtenons les courbes précision × rappel et ROC de la figure 1. La méthode baseline, qui s'appuie sur la régression logistique et les moyennes des données de suivi, est représentée en marron. La méthode ensembliste est représentée en rose. A rappel fixé, la précision de la méthode ensembliste est systématiquement plus élevée, comme reporté dans la table 1. La même observation est tirée des courbes ROC. Ce sont les méthodes d'arbres combinées au sous-échantillonnage qui ressortent comme étant les plus performantes.

| rappel (%) | 75 | 62.5 | 50 | 37.5 | 25 | 10 |
|---|---|---|---|---|---|---|
| précision (%), baseline | 25 | 26 | 30 | 36 | 44 | 44 |
| précision (%), ensemble | 27 | 36 | 40 | 47 | 52 | 70 |

TABLE 1 – Quelques couples (précision, rappel) extraits de la figure 1 pour la méthode de base et la méthode d'ensemble. *Clé de lecture : pour détecter 10% des boiteries, la méthode baseline fera 56% de faux positifs alors que la méthode d'ensemble n'en fera que 30%.*

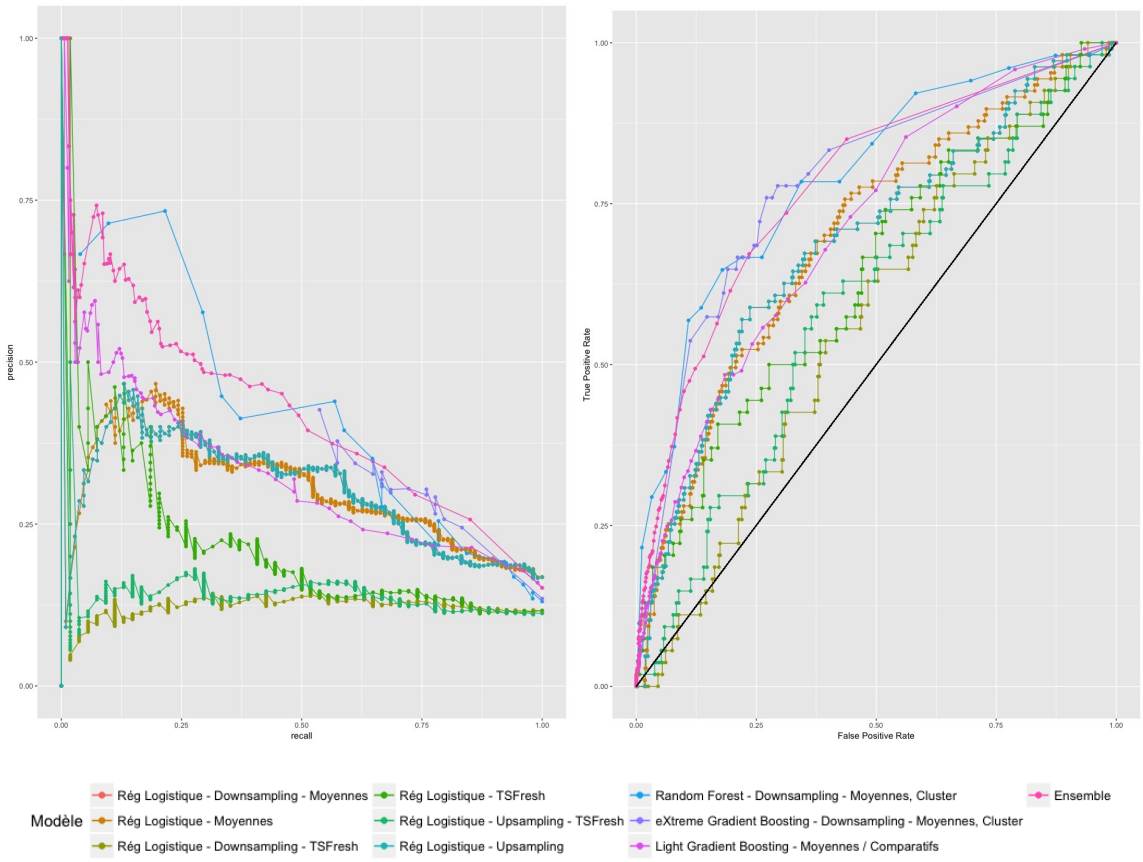


FIGURE 1 – Courbes précision × rappel (gauche) et ROC (droite) des modèles testés. Par rapport à la méthode de base (marron, "Rég Logistique - Moyennes"), la méthode ensembliste (en rose, "Ensemble") performe bien mieux. Pour référence, l'aire sous la courbe est estimée à 70% pour la méthode baseline contre 80% pour la méthode d'ensemble.

# 4 Dicussion

Au vu des résultats obtenus lors de cette étude, il est clair que les données de suivi contiennent bien un signal prédictif des boiteries et que la création de variables, le sous-échantillonnage et l'utilisation de modèles d'arbres augmentent la détectabilité de ces évènements. Cela étant, le porteur de projet avaient des attentes de l'ordre de (90%, 50%) pour le couple (précision, rappel), ce que nous n'atteignons pas (cf table 1). Nous proposons donc les pistes d'améliorations suivantes :

• Auditer la qualité de la donnée pour assurer que les modèles soient correctement évalués.

• Au sujet de la création de variables, trois autres approches sont possibles : estimer un modèle statistique des séries de suivi et en récupérer les paramètres ; décomposer les signaux de suivi sur une base de Haar et en extraire les coefficients ; appliquer un auto-encodeur variationnel.

• Concernant les modèles prédictifs, nous proposons de tester d'autres modèles, notamment les SVM.

# 5 Remerciements

Ce travail a été réalisé au cours de la Semaine Etudes Maths-Entreprises d'Orsay, en janvier 2019. Nous adressons nos remerciements aux organisateurs de cette semaine passionnante.

Nous tenons également à remercier Julie Dewez et Florine Hardy, représentantes du groupe Seenergi, qui nous ont fourni les données et la documentation nécessaire pour travailler et qui nous ont accordé leur temps pour nous expliquer en détail la problématique.

Enfin, nous souhaitons remercier Gilles Stoltz, qui a été notre mentor au cours de cette SEME, pour son suivi et ses conseils.

# Références

[Christ et al., 2016] Christ, M., Kempa-Liehr, A. W., and Feindt, M. (2016). Distributed and parallel time series feature extraction for industrial big data applications. *CoRR*, abs/1610.07717.

[Melnykov, 2016] Melnykov, V. (2016). ClickClust : An R package for model-based clustering of categorical sequences. *Journal of Statistical Software*, 74(9) :1–34.

[Trevor et al., 2009] Trevor, H., Robert, T., and JH, F. (2009). The elements of statistical learning : data mining, inference, and prediction.