Régression logistique - Explication des prédictions

Simon Lebastard

June 2021

1 Introduction

Ce document détaille les fondements théoriques derrière les métriques d'explication des prédictions produites par Signaux Faibles dans le cadre de la prédiction d'entreprises en difficulté.

2 Description du modèle

Le modèle utilisé par Signaux Faibles début 2021 (liste de mars, liste de juin), et dont on cherche à expliquer les prédictions, est une régression logistique multivariée.

La régression logistique prend en entrée un vecteur de K variables, ici appelé x. La régression est paramétrée sur un vecteur de poids (weight) w et un biais (bias, ou offset) b, pour produire un score $f(x)$:

$$\begin{aligned} f(x) &= \sigma(\langle w, x \rangle - b) \\ &= \frac{1}{1 + e^{-(\langle w, x \rangle - b)}} \end{aligned}$$

où σ est la fonction sigmoïde, et $\langle w, x \rangle = \sum_{k=1}^{K} w_k x_k$

Les variables en entrée de cette régression logistique sont normalisées avec un 'sklearn.preprocessing.StandardScaler'.

3 Listage des variables à contribution potentiellement substantielle

La régression logistique étant un modèle non-linéaire, la contribution de la somme de deux variables n'est, en général, pas égale à la somme de leurs contributions. Cependant, la contribution maximale que peut avoir une variable individuelle est limitée et contrôlée par le fait que la dérivée de la régression logistique soit bornée et de valeur absolue maximale en 0.

Plus précisément, le gradient de la régression logistique en $x \in$ est:

$$\begin{aligned}\nabla f(x) &= \nabla\sigma(\langle w, x\rangle - b)w \\ &= f(x)\big(1 - f(x)\big)w \\ &= \frac{e^{-(\langle w,x\rangle - b)}}{\big(1 + e^{-(\langle w,x\rangle - b)}\big)^2} w\end{aligned}$$

La matrice Hessienne associée est:

$$H(x) = \frac{e^{-(\langle w,x\rangle - b)}\big(e^{-(\langle w,x\rangle - b)} - 1\big)}{\big(e^{-(\langle w,x\rangle - b)} + 1\big)^3} ww^{\intercal}$$

On aura une dérivée maximale pour:

$$\begin{aligned} &H(x) = 0 \\ \iff &\langle w, x\rangle - b = 0\end{aligned}$$

Pour l'ensemble des points solutions de cette équation (une droite), on a un gradient maximal égal à:

$$\begin{aligned}\nabla_M f &= \nabla f(x_M) \\ &= \frac{w}{4}\end{aligned}$$

On obtient alors une borne supérieure (atteinte asymptotiquement en 0) sur la contribution que peut avoir une variable sur le score de prédiction. Plus précisément, pour la variable $x_k, k \in [1, K]$ cette borne vaut:

$$C_M(k) = \langle \nabla_M f, x_k^* \rangle$$

où:

$$x_k^* = \begin{pmatrix} 0 & \dots & 0 & x_k & 0 & \dots & 0 \end{pmatrix}$$

On a alors:

$$C_M k = \frac{w_k x_k}{4}$$

La variable x_k ne peut donc contribuer au maximum qu'à un score de $C_M k$. En fixant un seuil θ à partir duquel on juge que la contribution asymptotique maximale de cette variable est importante, on peut alors séléctionner une liste de variables contribuant potentiellement de manière significative à un score élevé. On choisit ainsi de lister les variables k telles que $w_k x_k \geq 4\theta$

4 Calcul des scores radar

Le radar est fait pour positioner chaque groupe de variable par rapport à sa contribution totale potentielle.

On souhaite ainsi produire un score entre 0 et 1 indiquant la situation relative d'un établissement en ce qui concerne un groupe de variable donné.

Observons la distribution des produits $w^\intercal, x \in \mathcal{X}$, où $\mathcal{X}$ est la distribution originale des variables. Comme chaque variable en entrée de la régression logistique est normalisée, la moyenne du produit ci-dessus vaut 0. Pour tout groupe $I = k_1, k_2, ..., k_L$ de L variables, la moyenne de $\sum_{k \in I} w_k x_k$ est également de moyenne nulle.

Pour chaque groupe de variable, on peut donc calculer un score qui positionne l'établissement relativement au reste de la distribution des établissements considérés, en calculant:

$$s_I(x) = \sigma(\sum_{k \in I} w_k x_k)$$

Dans l'application Signaux Faibles, le radar fournit, pour un établissement représenté par ses variables x, l'ensemble des scores de contribution $s_I(x)$ pour tous les groupes I de variables considérés.