CONSTRUCTION DE COTE :

L'EXEMPLE DU PARI A HANDICAP AU TOP 14

Loïc Champagne[1], Léo Gerville-Réache[2] & Sebastião Tiarks[3]

[1]Université de Bordeaux, LACES, France, loic.champagne@etude.u-bordeaux2.fr

[2]Université de Bordeaux, CNRS, UMR 5251, France, leo.gerville-reache@u-bordeaux.fr

[3]xStand SAS, Bordeaux, France, sebastiao@rugbystand.com

Résumé : Prévoir l'issue d'une rencontre sportive est un exercice qui mobilise un nombre de plus en plus grand de chercheurs et de parieurs. Un des défis est que chaque type de pari nécessite une modélisation spécifique. Au rugby, il existe 3 paris (marchés) principaux, le joueur pourra prédire le vainqueur/perdant, le nombre de points dans le match ou encore le vainqueur affecté d'un handicap. C'est ce dernier type de pari, le pari à handicap, qui fait l'objet de cette communication.

Mots clé : Pari sportif, rugby, modèle polytomique ordonné, qualité de prévision

Abstract : Forecasting in sports is showing a growing interest not only from the sports specialists but also from the scientific community. One of the challenges is that each proposed market need a specific modelisation. In Rugby, there are 3 main types of markets: the winner/loser market, the total points market and the handicap market. This paper analyses the handicap market.

Key words : Sports betting, rugby, ordered polytomic model, forecast quality

1 Introduction

La prévision de l'issue d'une rencontre sportive présente un intérêt pour tous les amateurs de sports, les spécialistes ou encore les pratiquants eux-mêmes. Chaque individu peut baser sa réflexion sur les informations qu'il possède.

Les victoires, les défaites, les compositions d'équipes, le score, les blessures, le temps de jeu des joueurs, les conditions météorologiques, etc…, sont autant d'informations devenues publiques et disponibles grâce à la professionnalisation et à la médiatisation du sport.

Ainsi, chacun peut aujourd'hui baser sa réflexion par rapport aux « statistiques » de l'équipe ou du joueur qui l'intéresse et induire une vision personnelle de l'issue d'une rencontre. Ensuite, il suffit de se rendre sur un site de paris sportifs et parier. Depuis 2010 en France, des professionnels du sport et de la Statistique se sont intéressés de plus près à la réflexion autour de la prévision des résultats sportifs. En effet, cette année-là, l'Autorité de Régulation des Jeux En Ligne (ARJEL) a délivré ses premiers agréments aux acteurs du pari sportif sur internet, légiférant et démocratisant ainsi cette pratique. Aussi, afin d'obtenir les meilleures prévisions possibles, la réflexion pure et simple, même d'un spécialiste, est le plus souvent insuffisante. Les premières études réalisées sur ce sujet (Moroney en 1956 pour le football par exemple) montrent que la modélisation statistique permet d'obtenir de bons résultats. Ainsi, depuis de nombreuses années, les chercheurs et les parieurs essaient de modéliser les différents paramètres d'une partie, d'un match de football ou encore d'un combat en boxe. Seulement, il n'existe pas de règle quant aux paramètres à modéliser. Quels paramètres influencent le résultat d'un match ? Les paramètres du match en lui-même (résultats passés des deux équipes, classement actuel, etc…)? Les paramètres internes à l'équipe (blessure, ambiance, état d'esprit, motivation, etc…) ? Les paramètres externes (pression des médias, météorologie, public, etc…) ? Les paramètres personnels de chaque joueur d'une équipe (performances, intégration, vie privée, etc…) ?

Dans cette communication, nous présentons quelques résultats d'une collaboration entre l'IMB, l'AMIES et la société xStand sur la prévision de l'écart de points des rencontres du Top 14. Cette étude avait pour but de construire pour chaque rencontre la valeur du pari à Handicap.

2 Paris sur Handicap et l'historique des rencontres

Le pari à Handicap : Un handicap est un nombre que l'on va retrancher au favori d'une rencontre à venir afin de modéliser un "50-50" théorique. Ainsi un handicap de -10,5 signifie que le bookmaker (ou opérateur de paris sportifs) estime à une chance sur deux que l'équipe à domicile gagnera avec plus de 10,5 points d'avance et à une chance sur deux que l'équipe à domicile aura moins de 10,5 points de plus que l'équipe adverse.

Ainsi, les parieurs effectuent un pari « au dessus » ou « en dessous » du handicap du bookmaker.

Qualité de la prévision : Afin de permettre la critique des résultats (prévisions du bookmaker ou autre modèle statistique) nous avons intégré 9 classes d'écart de points définies comme suivant:

- Classe 1, 2, 3, 4: Victoire de l'équipe à domicile par respectivement [+ de 21]; [de 15 à 21]; [de 8 à 14] et [de 1 à 7] points.
- Classe 5: Match Nul [0].
- Classe 6, 7, 8 et 9: Victoire de l'équipe à l'extérieur par respectivement [de 1 à 7]; [de 8 à 14]; [de 15 à 21] et [+ de 21] points.

L'intérêt de ces différentes classes est non seulement de pouvoir résumer le spectre "écart de points" sur 9 intervalles mais aussi d'introduire un aspect sportif propre au rugby (on remarque que pour basculer d'une classe à l'autre, l'écart de point nécessaire est d'au maximum 7 points soit un essai transformé). Aussi, on évalue la prévision comme étant :

- Excellente : le handicap (avant match) est dans la même classe que l'écart de point réel. En cas de match nul (classe 5), les classes 4 et 6 sont des prévisions également excellentes.
- Bonne: le handicap (avant match) est dans la classe directement en-dessus ou en-dessous de la classe correspondant à l'écart de points réel.
- Mauvaise: les autres cas.

Données à disposition : Le tableau dont nous disposons regroupe les informations de 264 matchs consécutifs, du Top 14, du début de la saison 2012 au 2 novembre 2013. Ce sont les rencontres pour lesquelles nous disposons de l'ensemble des variables (en particulier les variables subjectives qui suivent).

Variables subjectives : Afin de proposer un nouveau modèle de prédiction, ont été intégrés pour chaque match, 3 nouveaux paramètres. Il s'agit:

- De la motivation équipe "domicile" :
    - 0 pour un match important (motivation normale, etc…)
    - 1 pour un match très important (grande motivation, équipe type, etc…)
- De la motivation équipe "visiteuse" :
    - 0 et 1 : comme précédemment
    - -1 : pour une impasse (moins-value de motivation, équipe remaniée, etc…)
- De la météo :
    - Groupe 1: "Temps clair" et "Partiellement couvert" (favorise le jeu ouvert)
    - Groupe 2: "Nuageux avec risque d'averses" (peut nuire au déroulement du jeu)

- Groupe 3: "Pluie fine continue" et "brouillard" (complexifie la maniabilité/visibilité)
- Groupe 4: "Grosse pluie", "orage" et "neige" (conditions extrêmes)

Pour chaque modèle, l'estimation des paramètres est réalisée à l'aide du logiciel R.

3 Le modèle de régression linéaire

Le modèle qui vient naturellement à l'esprit est celui de la régression linéaire multiple (avec un mélange de covariables qualitatives et quantitatives. Le principe est de modéliser directement l'écart de point entre les deux équipes. L'estimation des paramètres du modèle est réalisée par la méthode du maximum de vraisemblance.
Le handicap d'une rencontre entre l'équipe *i* (à domicile) et l'équipe *j* (à l'extérieur) est la prévision obtenue directement par utilisation du modèle selon l'équation suivante :

$$Prévision(i,j) = Intercept + Coeff\ domicile(i) + Coeff\ extérieur(j) + Coeff\ météo \\ + Coeff\ motivation\ domicile + Coeff\ motivation\ extérieur$$

Ici, la prévision est un écart de points qui sera catégorisée pour comparer les différents modèles.

4 Le modèle "attaque-défense"

Dans ce modèle, une dimension sportive entre en jeu avec l'intégration d'un potentiel « offensif » et d'un potentiel défensif « défensif » pour les deux équipes qui se rencontrent. Ils viennent s'ajouter aux variables subjectives. Ce modèle permet d'estimer simultanément le nombre de points marqués par chacune des deux équipes. Les estimations des scores de chaque équipe sont obtenues après l'estimation des coefficients du modèle par maximum de vraisemblance via la procédure de régression linéaire de R. Il faut noter ici que le tableau de données permettant les estimations est particulier. En effet, une rencontre constitue deux lignes du tableau (deux individus statistiques).

Les scores de la rencontre entre l'équipe *i* et l'équipe *j* sont les prévisions obtenues directement par utilisation du modèle selon l'équation suivante :

$$Prévision\ score(i) = Intercept * Domicile(i) + Coeff\ offensif(i) \\ + Coeff\ défensif(j) + Coeff\ météo + Coeff\ motivation$$

$$Prévision\ score(j) = Intercept * Domicile(j) + Coeff\ offensif(j) \\ + Coeff\ défensif(i) + Coeff\ météo + Coeff\ motivation$$

Ici, le paramètre « *Domicile*(*i*) » sera égal à 1 si *i* est l'équipe qui reçoit, et -1 si *i* l'équipe en déplacement (idem pour l'équipe *j*).

La prévision de l'écart de score est alors obtenue par différence de la prévision du score de l'équipe à l'extérieur et de l'équipe à domicile. Si *i* joue à domicile, on a :

$$Prévision\ (i,j) = Prévision\ score(j) - Prévision\ score(i)$$

Ici, la prévision est un écart de point qui sera catégorisée pour comparer les différents modèles.

5 Le modèle polytomique ordonné (MPO)

Le MPO est une généralisation de la régression logistique pour laquelle la variable à expliquer est multinomiale ordonnée. Pour notre étude, ce sont les catégories "écarts de score" des rencontres qui sont modélisées. Les différentes catégories sont celles qui ont été présentées en partie 2. Ainsi, dans le tableau de données, chaque écart de points observé est classé dans l'une des 9 classes correspondante.

Le MPO modélise une variable à expliquer Y prenant ici k modalités ordonnées (ici, la classe d'écart de points entre les 2 équipes).

Pour présenter ce modèle, on se place dans un premier temps dans le cas d'une seule variable explicative X, et on introduit plusieurs seuils $\alpha_1, \ldots, \alpha_{k-1}$ tels que :

$$(Y|X = x) = \begin{cases} 1 \; si \; Y' < a_1 \\ j \; si \; a_{j-1} \leq Y' < a_j, \qquad j = 2, \ldots, k-1 \\ k \; si \; Y' \geq a_{k-1} \end{cases}$$

avec

$$Y' = \beta_1 x + \epsilon$$

Le choix de la fonction de répartition logistique conduit au modèle :

$$logit \; P(Y \leq j|X = x) = a_j - \beta_1 x, \qquad j = 1, \ldots, k-1$$

Si on est en présence de p variables explicatives (ici équipe domicile, équipe extérieure, motivation à domicile, motivation à l'extérieur et météo), le modèle devient :

$$logit \; P(Y \leq j|\boldsymbol{X} = \boldsymbol{x}) = \alpha_j - \beta_1 X_1 - \cdots - \beta_p X_p, \qquad j = 1, \ldots, k-1$$

Ou encore

$$P(Y \leq j|\boldsymbol{X} = \boldsymbol{x}) = \frac{\exp\left(\alpha_j - \beta_1 X_1 - \cdots - \beta_p X_p\right)}{1 + \exp\left(\alpha_j - \beta_1 X_1 - \cdots - \beta_p X_p\right)}$$

A travers une telle modélisation, seule la constante diffère suivant les différents niveaux de Y. L'estimation par maximum de vraisemblance des paramètres du modèle a été réalisée avec le logiciel R. Le tableau de données est ici le tableau standard où chaque rencontre constitue une ligne (un individu statistique).

Le résultat de la fonction ci-dessus donne l'estimation de la probabilité que l'écart de points corresponde à une classe inférieure ou égale à la classe 1, à la classe 2, 3, 4 […] à 9. L'estimation de la probabilité que l'écart de points corresponde à une classe inférieure ou égale à la classe 9 est de 1 car c'est la dernière classe possible.

Afin d'estimer la probabilité d'appartenir à une classe, on travaille de la manière suivante :

- Classe 1 : nous reportons le résultat trouvé au calcul précédent (car il n'y a aucune classe inférieure à 1). Nous obtenons l'estimation de la probabilité que l'écart de points se situe dans la classe 1.
- Classe 2 : on soustrait l'estimation de la probabilité que l'écart réel soit inférieur ou égal à la

classe 1 à la probabilité qu'il soit inférieur ou égal à la classe 2. On obtient l'estimation de la probabilité que l'écart se situe dans la classe 2.

On réitère cette opération pour toutes les classes.

Parmi ces probabilités, le maximum correspondra à la meilleure prévision que nous pouvons obtenir avec ce modèle. Ce maximum pourrait être l'estimation de la classe de l'issue du match. Cependant, ce maximum est relativement volatile. Nous préférons définir la prévision via l'estimation de "l'espérance de la classe". Le principe est de calculer la somme pondérée par les probabilités estimées des classes (*c*). La prévision de la classe est alors l'entier le plus proche de cette "espérance". Pour résumer, on a :

$$logit\left(P\left(Y_{i,j} \leq c|\boldsymbol{X_{ij}} = \boldsymbol{x}\right)\right) = Intercept(c) + Coeff\ dom(i) + Coeff\ ext(j) + Coeff\ météo + Coeff\ motivation\ domicile + Coeff\ motivation\ extérieur,$$

puis,

$$P\left(Y_{i,j} = c|\boldsymbol{X_{ij}} = \boldsymbol{x}\right) = P\left(Y_{i,j} \leq c|\boldsymbol{X_{ij}} = \boldsymbol{x}\right) - P\left(Y_{i,j} \leq c-1|\boldsymbol{X_{ij}} = \boldsymbol{x}\right)$$

enfin,

$$Prévision(i,j) = Arrondi\left(\sum_{c=1}^{k} c * P\left(Y_{i,j} = c|\boldsymbol{X_{ij}} = \boldsymbol{x}\right)\right)$$

6 Discussion

Les différents modèles présentés dans les sections précédentes ont fait l'objet d'une estimation avec le logiciel R sur les 264 rencontres de la base de données. La qualité des différents modèles est résumée dans le tableau ci-après, selon la règle d'évaluation précisée dans la partie 2.

Modèles / Résultats	Bookmakers	MPO sans Subjectif	Modèle Attaque-Défense Subjectif	Modèle Linéaire Subjectif	MPO Subjectif
Nombre de matchs observés	264	264	264	264	264
Prévisions « excellentes »	82	73	78	112	124
	31%	28%	30%	42%	47%
Prévisions « bonnes »	101	103	75	80	90
	38%	39%	28%	30%	34%
TOTAL « prévisions acceptables »	183	176	153	192	214
	69%	67%	58%	73%	81%

Le modèle polytomique ordonné subjectif (MPOS) est celui, parmi les modèles construits, qui donne la meilleure qualité globale. Avec 81% de prévisions acceptables, ce modèle purement statistique, fait bien mieux que le modèle "Attaque-Défense". En effet, bien que le modèle "Attaque-Défense" se fonde sur une théorie sportive basée sur les notions de potentiel offensif et

défensif, ce modèle souffre ici de deux écueils. Dans un premier temps, la modélisation n'est pas basée sur l'écart de point mais sur les points marqués de chaque équipe. De plus l'indicateur de qualité utilisé est basé sur la proportion de bonne prévision dans les classes d'écart. Il est donc normal que le MPOS soit ici meilleur. Cela met en évidence que l'objectif de la modélisation est un élément essentiel du choix et de la pertinence de la modélisation

La comparaison MPO - MPOS montre également que l'inclusion de variables subjectives améliore grandement les performances de la prédiction. En effet, l'issue des rencontres est très dépendante de la situation des équipes dans le championnat au moment de la rencontre. Ces variables traduisent, d'une certaine manière, cette dépendance temporelle.

La comparaison avec la qualité du handicap bookmakers (notre référence) est plus délicate. En effet, les handicaps bookmakers sont réalisés avant la rencontre alors que ceux réalisés par les autres modèles le sont a posteriori (l'ensemble des rencontres a été utilisé pour l'estimation de ces modèles). Une validation sur de futures rencontres sera nécessaire pour comparer honnêtement ces modèles aux handicaps proposés par le bookmaker.

Bibliographie

[1] Bradley & Terry (1952). *Rank analysis of incomplete block designs*. Biometrika, Vol. 39, pages 324-345.

[2] Caron F. & Doucet A. (2010). *Efficient Bayesian Inference for Generalized Bradley-Terry Models*. Journal of Computational and Graphical Statistics, Vol. 21, Issue 1, 2012.

[3] Coulom R. (2010). *Jeux et sports : le problème des classements*. Pour la Science n°393, pages 20-27.

[4] Foulley JL. (2012). *Tentative d'évaluation et de classement des 16 équipes de l'Euro 2012*. w3.jouy.inra.fr/unites/miaj/public/.../applibugs.12_12_20.jlfoulley.pdf.

[5] Hébert BP. (1998). *Régression avec une variable dépendante ordinale : comparaison de la performance de deux modèles logistiques ordinaux et du modèle linéaire classique à l'aide de données simulées*. Thèse, Bibliothèque nationale du Canada.

[6] Karlis D. & Ntzoufras I. (2003). *Analysis of sports data by using bivariate Poisson models*. Journal of the Royal Statistical Society: Series D (The Statistician), Vol. 52, Issue 3, pages 381–393.

[7] Langville A. & Meyer CD. (2012). *Who's one ? : The Science of Rating and Ranking*. Princeton University Press.

[8] Louvière L. (2008). *Régression sur variables catégorielles*. Support de cours, Université de Rennes2.

[9] Massey K. (1997). *Statistical Models Applied to the Rating of Sports Teams.* Bluefield College, 1997 - 74.207.231.132.