Analyse des données NBA challenge



Auteurs :

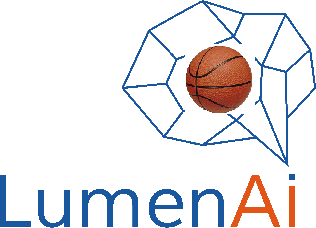
Geoffrey Pouliquen

Kévin Faou

Théophile Gotzorides

Date :

23/01/2019



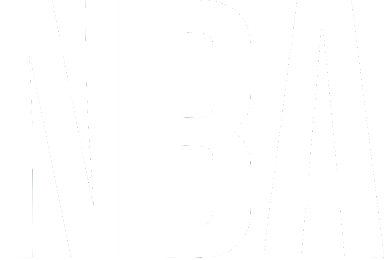


Table des matières

**Introduction2**

**Présentation des variables3**

**Analyse des variables3**

Traitement des variables4

Analyse descriptive multivariée5

Analyse de ScoreMT7

**Régression linéaire8**

Modèle complet8

Modèle hybride basé sur le Cp de Mallows11

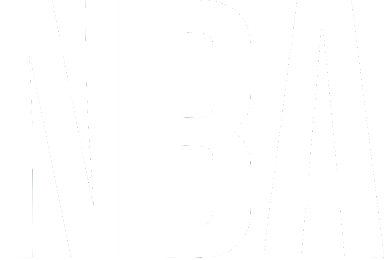
**Régression logistique12**

Etude du modèle complet13

Prédiction du modèle complet13

Sélection de variables d’après le critère AIC15

**Conclusion17**

Introduction

Avec la technologie en perpétuelle évolution l’utilisation et la gestion des données massives, ou big data, est devenu une préoccupation majeure dans beaucoup de domaines. Ces données aussi variées que nombreuses peuvent s’avérer très utiles à condition de savoir s’en servir.

Ici, Lumen AI propose un data challenge dont le but est de prédire le gagnant de match de basket-ball avec des données récoltées lors des deux premiers quart-temps. L’équipe de Lumen AI, une start-up spécialisée dans les données massive, a compilé les statistiques des premières mi-temps de 12576 matchs de basket de la NBA sur ces 15 dernières années. Leur objectif était d’être capable de prévoir, à partir de ces données, le vainqueur d’un match. Le but de notre étude sera donc de répondre le plus précisément possible à la problématique de cette entreprise.

Tout d’abord, nous nous attacherons à décrire le jeu de données pour voir s’il doit être transformé et s’il contient des aberrations ou des erreurs de mesure.

Puis, nous entrerons dans le vif du sujet en tentant de prédire le résultat des matchs d’un échantillon test à partir du score à la mi-temps.

Ensuite, nous essaierons de déterminer le vainqueur à partir d’une partie des données, les 12 et 18 premières minutes du match, grâce à une régression linéaire.

Enfin, nous utiliserons une méthode directe (régression logistique) qui nous permettra de déterminer le gagnant avec un meilleur niveau de précision.

Présentation des données

Les données représentent onze statistiques de match relevées à chaque seconde jusqu’à la mi-temps. De plus, Les statistiques ne sont pas relatives aux performances de chaque équipe mais représentent l’écart entre les deux équipes sur chaque statistique. Par conséquent, les données négatives (resp. positives) signifient un avantage pour l’équipe évoluant à domicile (resp. extérieure).

Les statistiques relevées sont les suivantes :

* Score : Score du match
* Offensive Rebound : Récupération de la balle après un tir manqué de son équipe
* Defensive rebound : Récupération de la balle après un tir manqué de l’équipe adverse
* Offensive foul : Faute offensive
* Defensive foul : Faute défensive
* Assist : Dernière passe avant un tir réussi
* Lost ball : Balle perdue
* Steals : Interception de la balle par l’adversaire
* Bad pass : Passe interceptée ou sortie du terrain
* Block : Tir adverse bloqué
* Miss : Tir manqué

Les données contiennent également le numéro d’identification et le vainqueur de chaque match via les variables ID et Match winner. Concernant la variable Match winner, un 0 représente une victoire de l’équipe jouant à domicile et un 1 une victoire de l’équipe jouant à l’extérieure.

Pour faciliter l’analyse des données, chaque match a été divisé en quatre “huitièmes-de-temps” (HT) de 6 minutes (360 secondes), la moitié d’un quart-temps. Pour chacune des 11 statistiques relevées, les données de chaque seconde de ces HT ont été rassemblées en une seule, représentant la moyenne de la statistique sur ce HT. On conserve cependant la variable Score de la 1440ème seconde qui correspond au score à la mi-temps et que l’on renomme ScoreMT. Ces modifications réduisent le nombre de variables à analyser de 15841 à 47.

Analyse des variables

Echantillon : 12576 matchs

Variables : 47

**Variable qualitative binaire**

Match winner (0 ou 1)

**Variable quantitative discrète**

ScoreMT

**Variables quantitatives continues**

Score

Offensive Rebound

Defensive rebound

Offensive foul

Defensive foul

Assist

Lost ball

Steals

Bad pass

Block

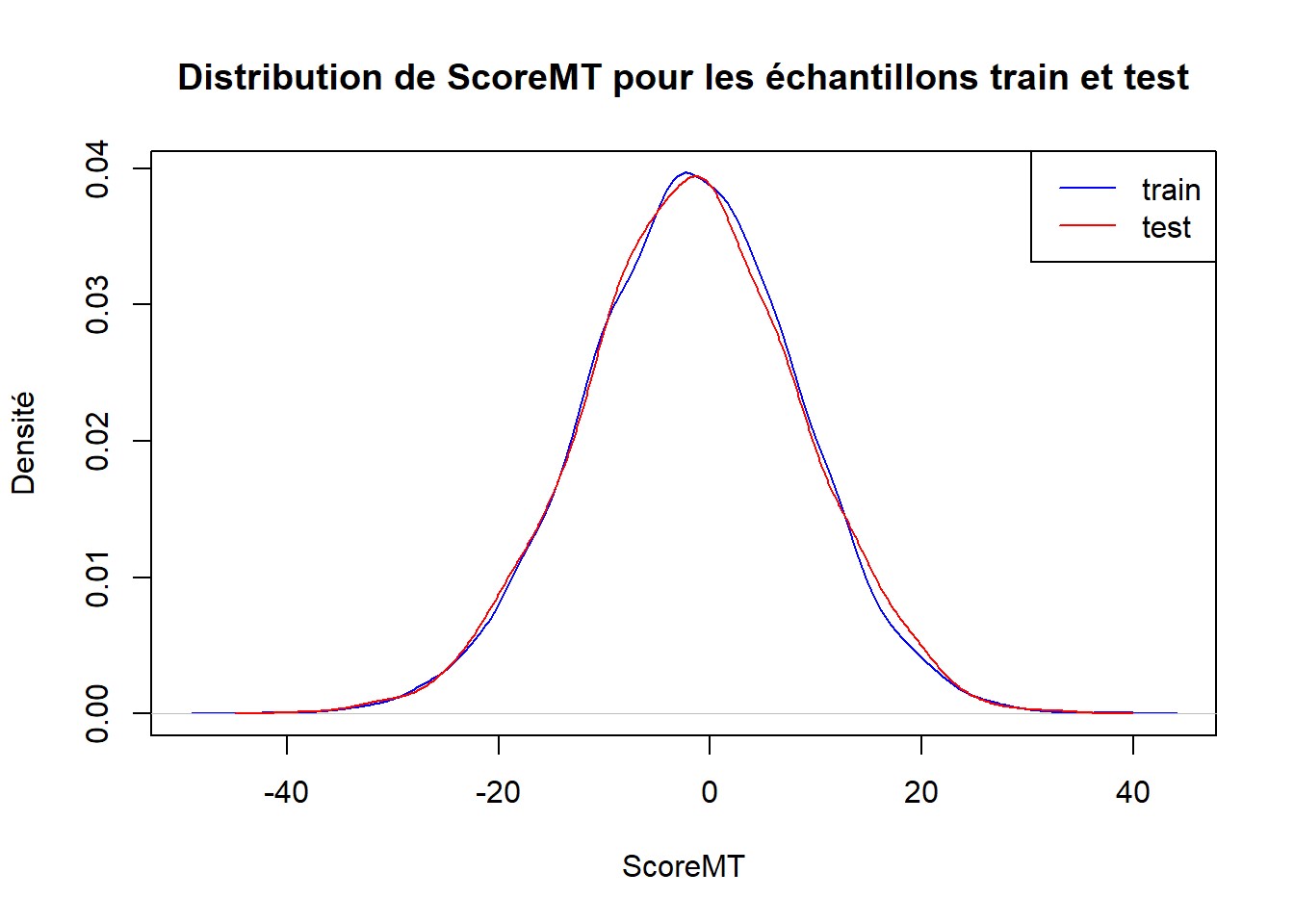
Miss

(On ne considère pas ID comme une variable à analyser)

Traitement des variables

Les statistiques Defensive foul n’ayant que des valeurs nulles, elles sont retirées. On note également que les matchs 1939 et 4058 ont une variable ScoreMT de -97. Cet écart, supérieur au record historique d’écart de points entre deux équipes, est à l’évidence une erreur. Les deux matchs sont donc retirés de l’échantillon.

On crée ensuite, à partir de notre échantillon de 12574 matchs, un échantillon test en retirant au hasard un tiers des matchs de l’échantillon train. On peut voir sur le graphe ci-dessous que la variable ScoreMT possède bien la même distribution dans les deux échantillons.

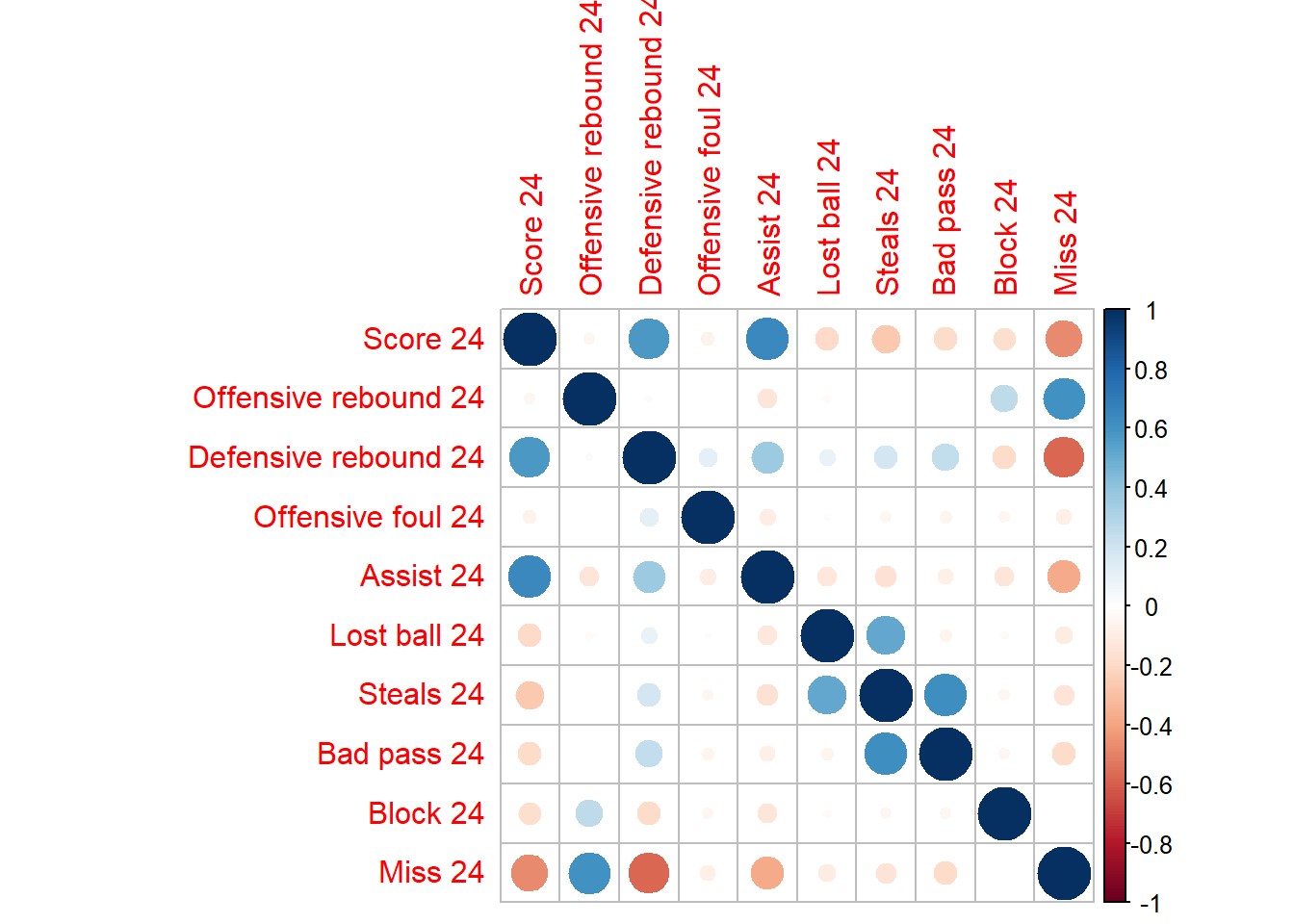


Analyse descriptive multivariée

Afin de déterminer l’influence qu’on les différentes variables les unes sur les autres, on débute cette étude par une analyse descriptive multivariée. On cherche donc les corrélations entre les différentes variables.

On commence par étudier les corrélations entre les variables d’un même HT, le dernier.

*Graphe des corrélations du 4ème HT*

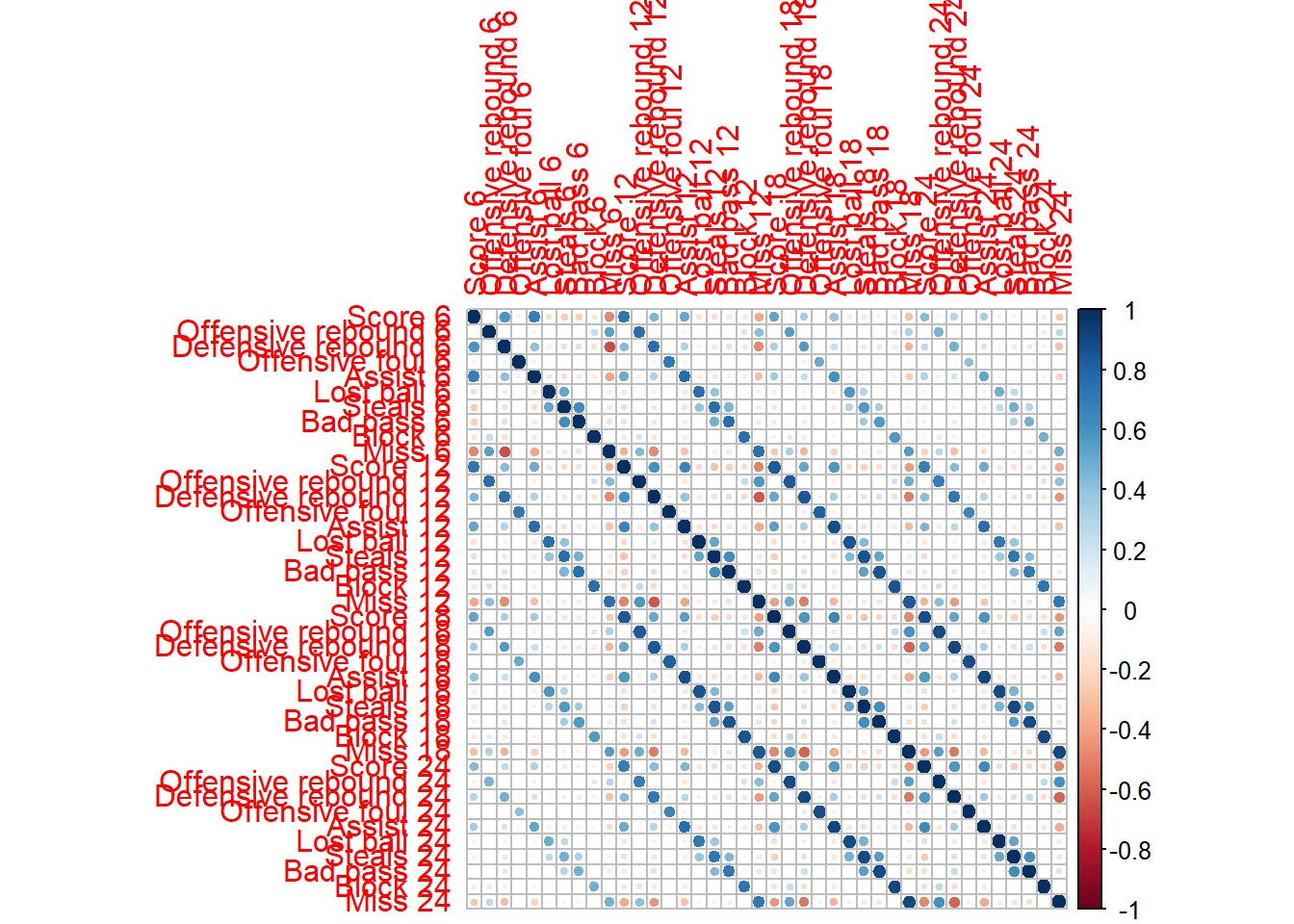


Analysons ces corrélations :

* Score et Assist : Le score est logiquement corrélé aux nombres de passes décisives. La corrélation n’est pas de 1 car il est possible de marquer un panier suite à une action solitaire ou grâce à un lancer franc.
* Score et Miss : Moins une équipe manque de tir, plus elle marque.
* Miss et Offensive rebound : Plus une équipe rate de tir, plus elle est susceptible de récupérer la balle suite à un tir raté.
* Miss et Defensive rebound : Un tir manqué permet à l’équipe adverse de récupérer la balle.
* Steals et Bad pass : Une mauvaise passe entraîne souvent une interception.

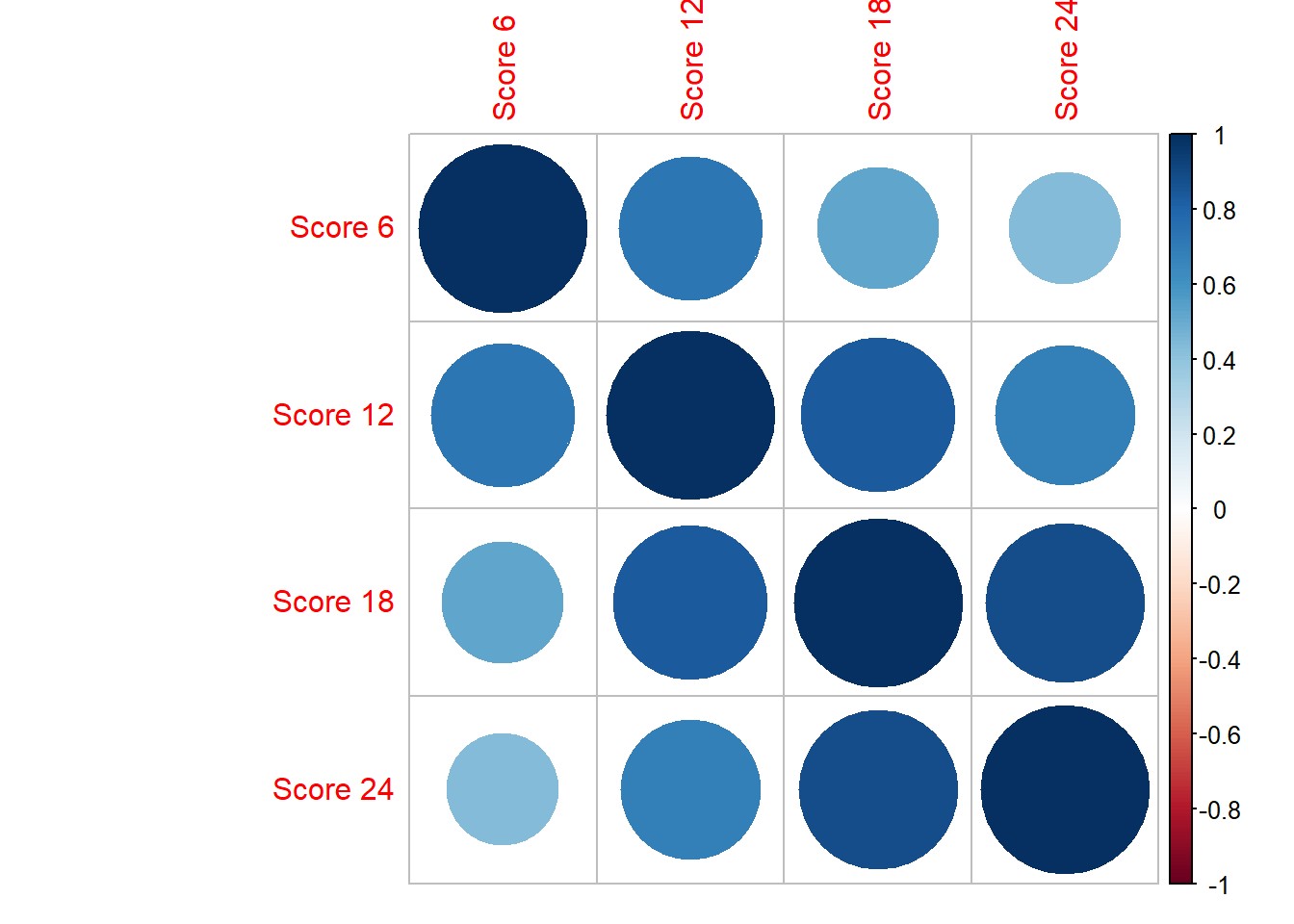
La corrélation entre Score et Defensive rebound est plus mystérieuse. Elle peut cependant s’expliquer par le fait que récupérer la balle après un tir manqué de l’équipe adverse permet de lancer une action ou une contre-attaque. En effet, au basket, la plupart des actions aboutissent à un panier. De plus, la contre-attaque est une phase de jeu au cours de laquelle il est beaucoup plus facile de marquer.

En analysant les corrélations entre les variables de tous les HT, on peut voir l’impact du temps sur les corrélations entre les variables.



*Graphe des corrélations entre toutes les variables*

On peut voir sur le graphe ci-dessus que les corrélations sont identiques, bien que moins fortes, au cours de tous les HT. L’impact d’une statistique à un instant t sur sa valeur future est de plus en plus faible à mesure que le match progresse. Cette analyse est confirmée par le graphe des corrélations entre les 4 variables Score. Pour prédire le vainqueur du match, et donc le score final, il semble nécessaire de s’appuyer sur les données relatives au dernier HT plutôt que sur les données des premières minutes du match.



*Graphe des correlations des variables Score*

On constate également que la variable la plus corrélée au score d’un HT est le score du HT précédent. On va donc s’intéresser à la variable ScoreMT qui représente le score à la mi-temps, c’est-à-dire le score le plus proche, sur l’axe des temps, du score final.

Analyse de ScoreMT

|  |  |  |  |
| --- | --- | --- | --- |
|  | 0 | | 1 |
| [-44,-20] | 0.991 | | 0.009 |
| [-19,-13] | 0.912 | | 0.088 |
| [-12,-9] | | 0.823 | 0.177 |
| [-8,-6] | | 0.742 | 0.258 |
| [-5,-3] | | 0.647 | 0.353 |
| [-2,2] | | 0.518 | 0.482 |
| [3,5] | | 0.415 | 0.585 |
| [6,7] | | 0.334 | 0.666 |
| [8,11] | | 0.229 | 0.771 |
| [12,39] | | 0.118 | 0.882 |
| Somme | | 0.565 | 0.435 |

Afin de mieux étudier l’impact du score à la mi-temps sur le résultat final du match, on regroupe les valeurs de ScoreMT en différentes classes. Ce regroupement est fait en fonction de la probabilité de victoire de l’équipe qui mène à la mi-temps comme le montre le tableau ci-dessus.

Le tableau à droite donne les chances de victoire de chaque équipe en fonction du score à la mi-temps. On peut voir que, hormis les matchs où le score est très serré, ScoreMT semble être la variable adéquate pour prédire le vainqueur du match.

On regroupe cette fois ScoreMT en trois classes : l’équipe à domicile mène, le match est serré et l’équipe à l’extérieure mène. On peut voir alors que miser sur l’équipe en tête à la mi-temps, si le score n’est pas trop serré, permet de déterminer le vainqueur du match avec une probabilité, particulièrement élevée, de 76,5%.

Les matchs indécis, particulièrement imprévisibles, représentent 20% de l’échantillon.

*Tables des effectifs et des probabilités de ScoreMT en fonction de Match winner*

|  |  |  |  |
| --- | --- | --- | --- |
|  | 0 | 1 | Somme |
| [-44,-3] | 3091 | 791 | 3882 |
| [-2,2] | 839 | 782 | 1621 |
| [3,39] | 762 | 2036 | 2798 |
| Somme | 4692 | 3609 | 8301 |

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| [-44,-3] | 0.796 | 0.204 |
| [-2,2] | 0.518 | 0.482 |
| [3,39] | 0.272 | 0.728 |
| Somme | 0.565 | 0.435 |

Si on tient compte des matchs indécis, alors on peut prévoir le résultat du match à seulement 72%.

*Table des probabilités de ScoreMT en fonction de Match winner*

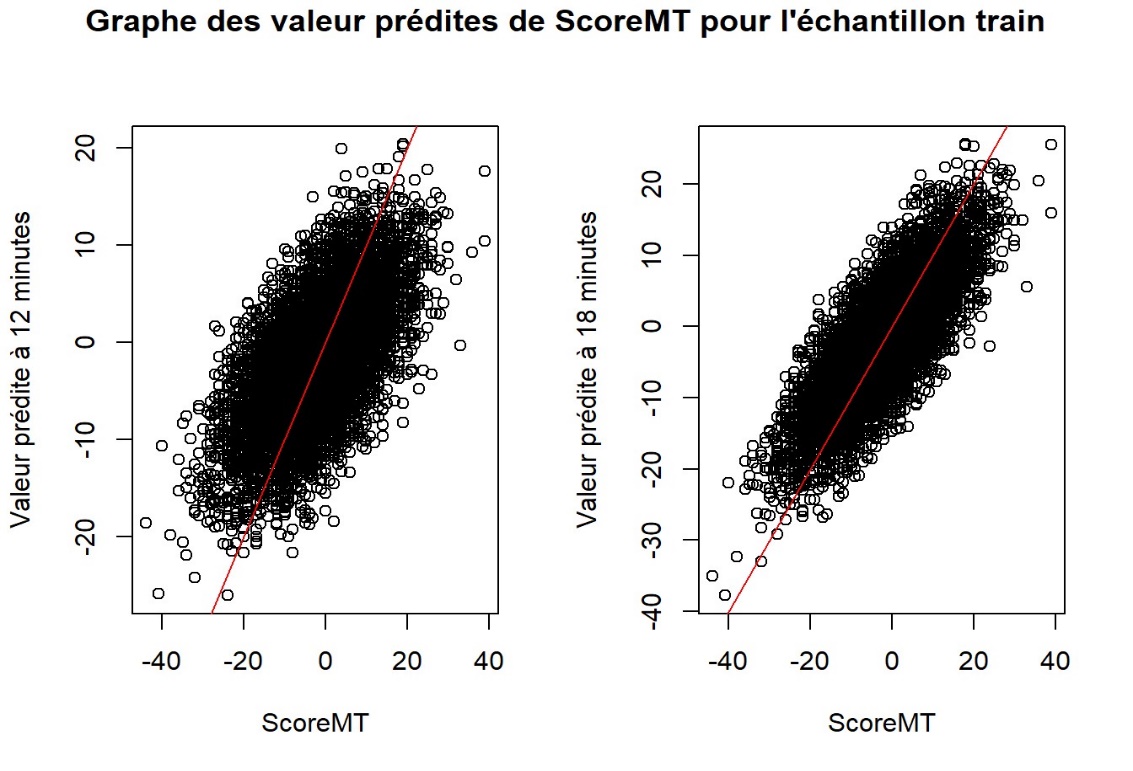
|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| [-44,1] | 0.729 | 0.271 |
| [2,39] | 0.291 | 0.709 |
| Somme | 0.565 | 0.435 |

Régression linéaire

Modèle complet

Ayant constaté l’importance de l’écart de points à la mi-temps dans le résultat du match, nous allons essayer de prédire cette valeur à l’aide d’une régression linéaire. Nous le ferons avec deux jeux de données différents, l’un contenant les valeurs du premier quart-temps et l’autre les données après 18 minutes de jeu. Cela nous permettra de montrer l’évolution de la qualité de notre prédiction en fonction du temps de jeu écoulé.

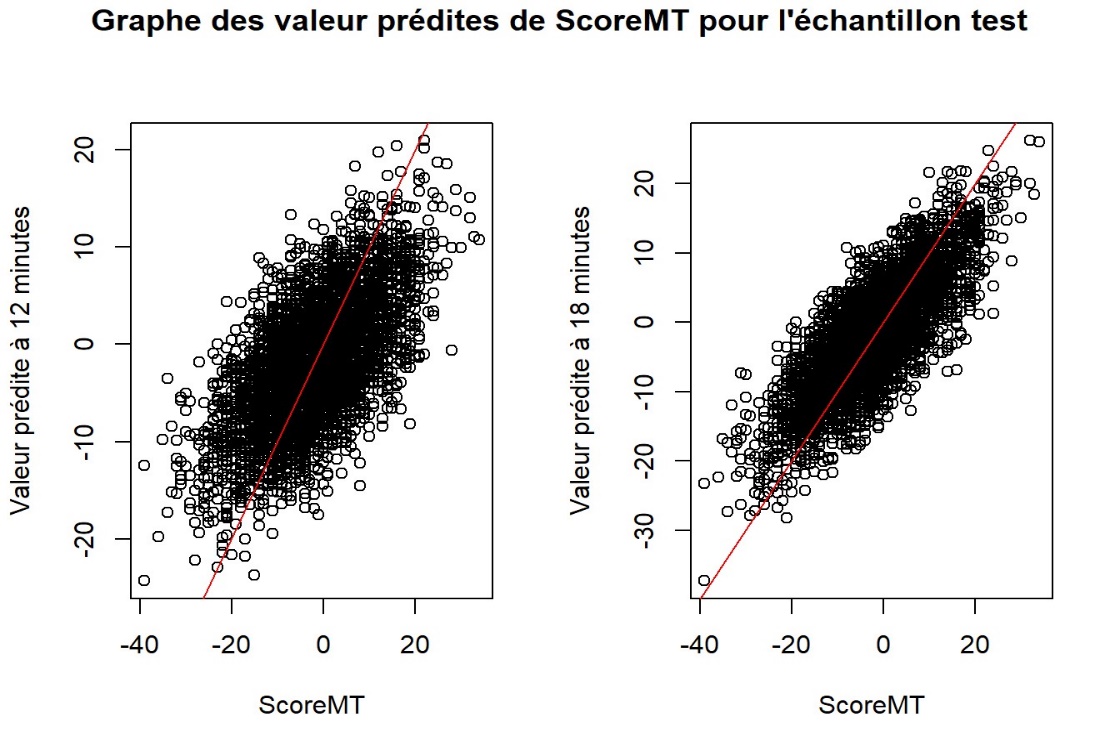
Pour cela, dans les deux cas, on ajuste le modèle complet par moindres carrés sur l’échantillon d’apprentissage puis on compare les valeurs ajustées aux valeurs prédites par le modèle pour les échantillon d’apprentissage et de test.



Erreurs de prédictions pour les deux premiers HT :

Erreur ajustée : 63.64

Erreur prédiction : 64.46



Erreurs de prédictions pour les trois premiers HT :

Erreur ajustée : 39.218

Erreur de prédiction : 38.919

Les erreurs de prédiction sont proches des erreurs d’ajustement ce qui nous permet de valider notre régression. On est donc capable de prédire l’écart de point à la mi-temps de façon convenable à partir des données des deux ou trois premiers HT.

**Estimation des valeurs de Match winner**

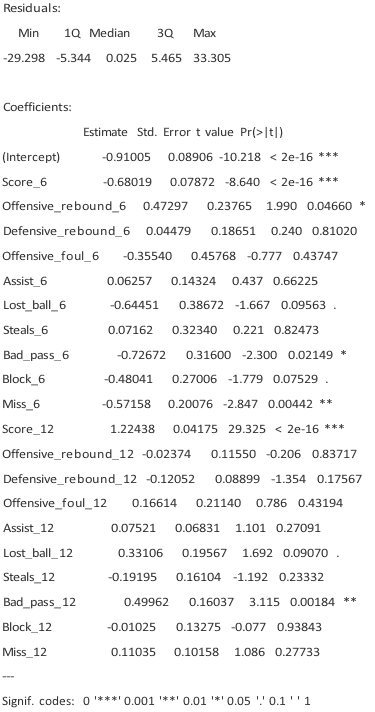
A partir des valeurs prédites de ScoreMT pour l’échantillon test, on estime le vainqueur de chaque match en utilisant la règle établie dans la première partie. Si l’équipe à l’extérieure mène de plus d’un point (ScoreMT > 1) alors on suppose qu’elle va remporter le match. Sinon, on suppose que c’est l’équipe à domicile qui va l’emporter. On estime alors le nombre de bonnes prédictions en comparant ces valeurs aux valeurs de Match winner de notre échantillon test.

*Table des prédictions*

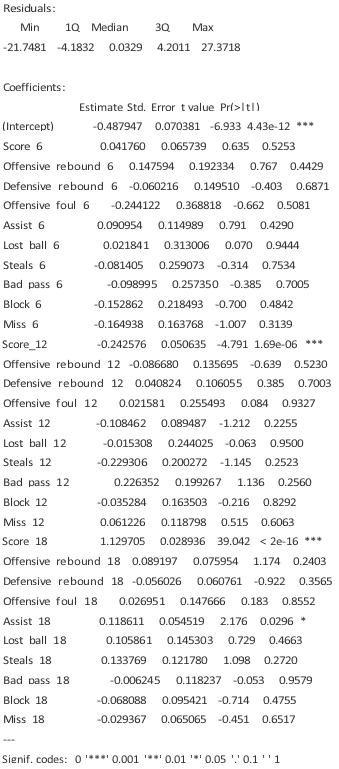
|  |  |
| --- | --- |
| **Régression** | **Bonnes prédictions** |
| 12 min | 63,7% |
| 18 min | 67,9% |

**Significativité des variables**

Dans le but d’améliorer la prédiction, on teste la significativité des variables du modèle de régression.



*Régression sur les deux premiers HT*



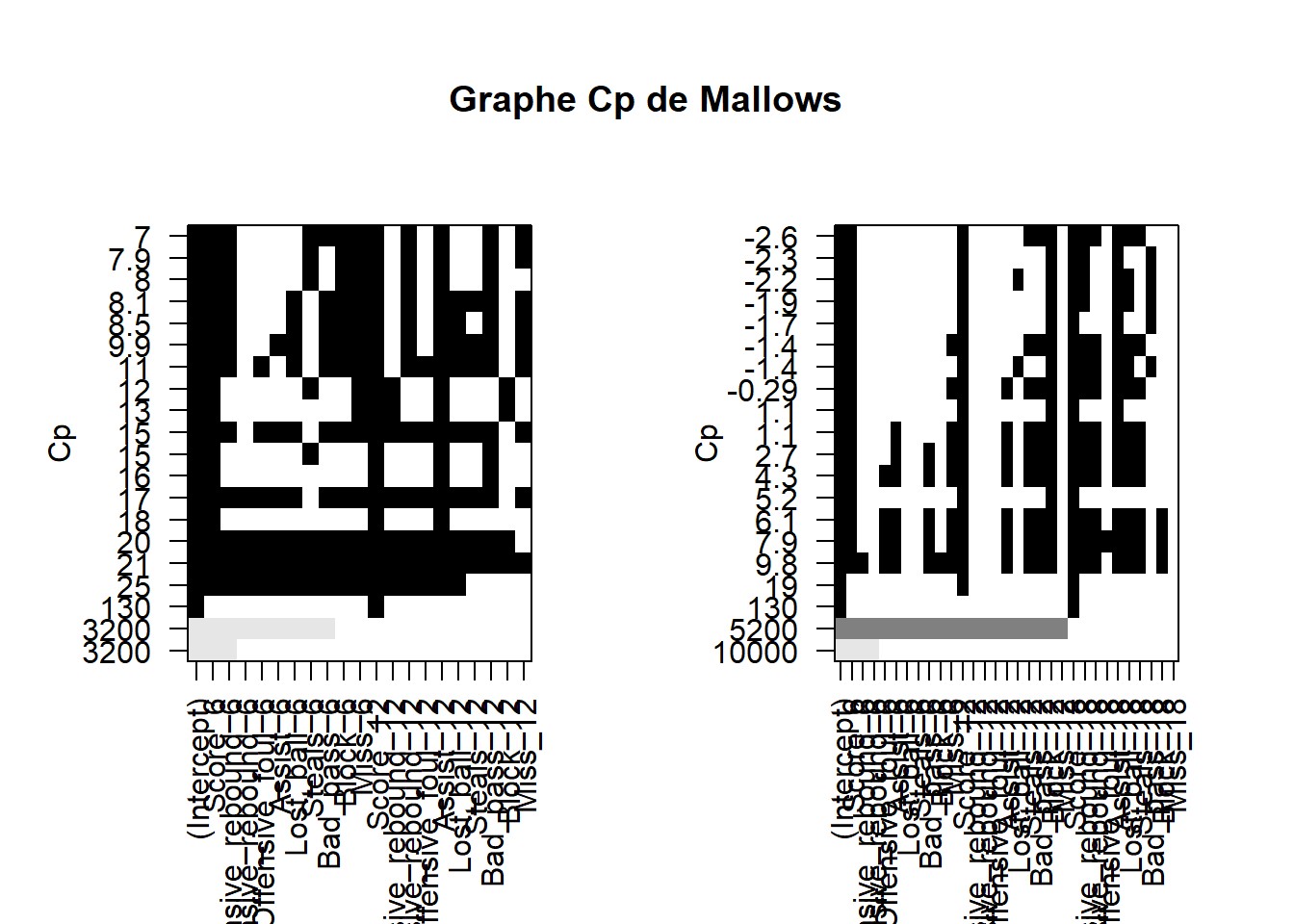
*Régression sur les trois premiers HT*

On constate rapidement que la plupart des variables ne sont pas significatives (p-valeur > 0,05). On va donc essayer de simuler ScoreMT avec un modèle sélectionnant seulement un nombre réduit de variables pour voir si la prédiction de Match winner sera meilleure.

Modèle hybride basé sur le Cp de Mallows

On utilise maintenant un autre modèle pour simuler ScoreMT, le modèle hybride basé sur le Cp de Mallows. Les graphes ci-dessous montrent les variables à sélectionner que l’on affiche après.

|  |  |
| --- | --- |
| 12 min | 18 min |
| Score 6  Offensive rebound 6  Steals 6  Bad pass 6  Block 6  Miss 6  Score 12  Defensive rebound 12  Assist 12  Bad pass 12  Miss 12 | Score 6  Score 12  Steals 12  Bad pass 12  Block 12  Score 18  Offensive rebound 18  Defensive rebound 18  Assist 18  Lost ball 18  Steals 18 |



Variables récupérées pour les deux simulations :

Le modèle sélectionné par le Cp de Mallows est validé par le calcul des erreurs de prédiction sur l’échantillon test.

Erreurs de prédictions pour les deux premiers HT

Erreur de prédiction CP : 64.509

Erreur de prédiction régression : 64.469

Erreurs de prédictions pour les deux premiers HT

Erreur de prédiction CP : 38.921

Erreur de prédiction régression : 38.919

**Comparaison entre le modèle de régression et le modèle hybride**

En estimant les valeurs de Match winner à partir des valeurs de ScoreMT prédites par le modèle hybride, on obtient des résultats similaires au modèle de régression linéaire classique.

*Tables des bonnes prédictions*

|  |  |  |
| --- | --- | --- |
|  | Régression | Mallows |
| 12 min | 63,7% | 63,8% |
| 18 min | 67,9% | 67,9% |

Régression logistique

Dans les deux parties précédentes, nous avons vu qu’il était possible de prévoir le résultat d’un match à différents niveaux de probabilité avec une règle de décision simple basé sur le score à la mi-temps (réel ou simulé) . Notre taux de réussite est actuellement d’environ 72%.

On cherche désormais à savoir s’il est possible d’obtenir une meilleur prédiction de Match winner à l’aide d’une nouvelle méthode, la régression logistique.

Nous allons dans cette partie effectuer des régressions logistiques afin de trouver le meilleur modèle capable d’expliquer le gagnant du match(variable Match winner) en fonction des covariables des 2 premiers quart-temps.

Nous regarderons tout d’abord le modèle avec toutes les covariables afin de tester si des variables sont pertinentes puis nous chercherons le meilleur modèle au sens du critère AIC en sélectionnant les variables. Enfin nous terminerons avec de la prédiction sur la variable Match winner.

# Etude du modèle complet

On mène une régression logistique (famille binomiale) car la variable réponse prend des valeurs binaires (0 et 1). On s’intéresse ici au modèle composé de toutes lesvariablesexplicatives ou modèle complet.

*D*0 − *D*1 = 2326.32

q1−α(χ249) = 66.34

On regarde le test suivant :

***H0*** *:* *Aucune variable n’explique la réponse*

vs***H1*** *: Il existe au moins une variable explicative pertinente*

Ici, nous avons : *D*0 − *D*1 > q1−α(χ249) donc on rejette ***H0***.

Il existe au moins une variable qui permet d’avoir un modèle qui explique mieux Match winner que le modèle nul.

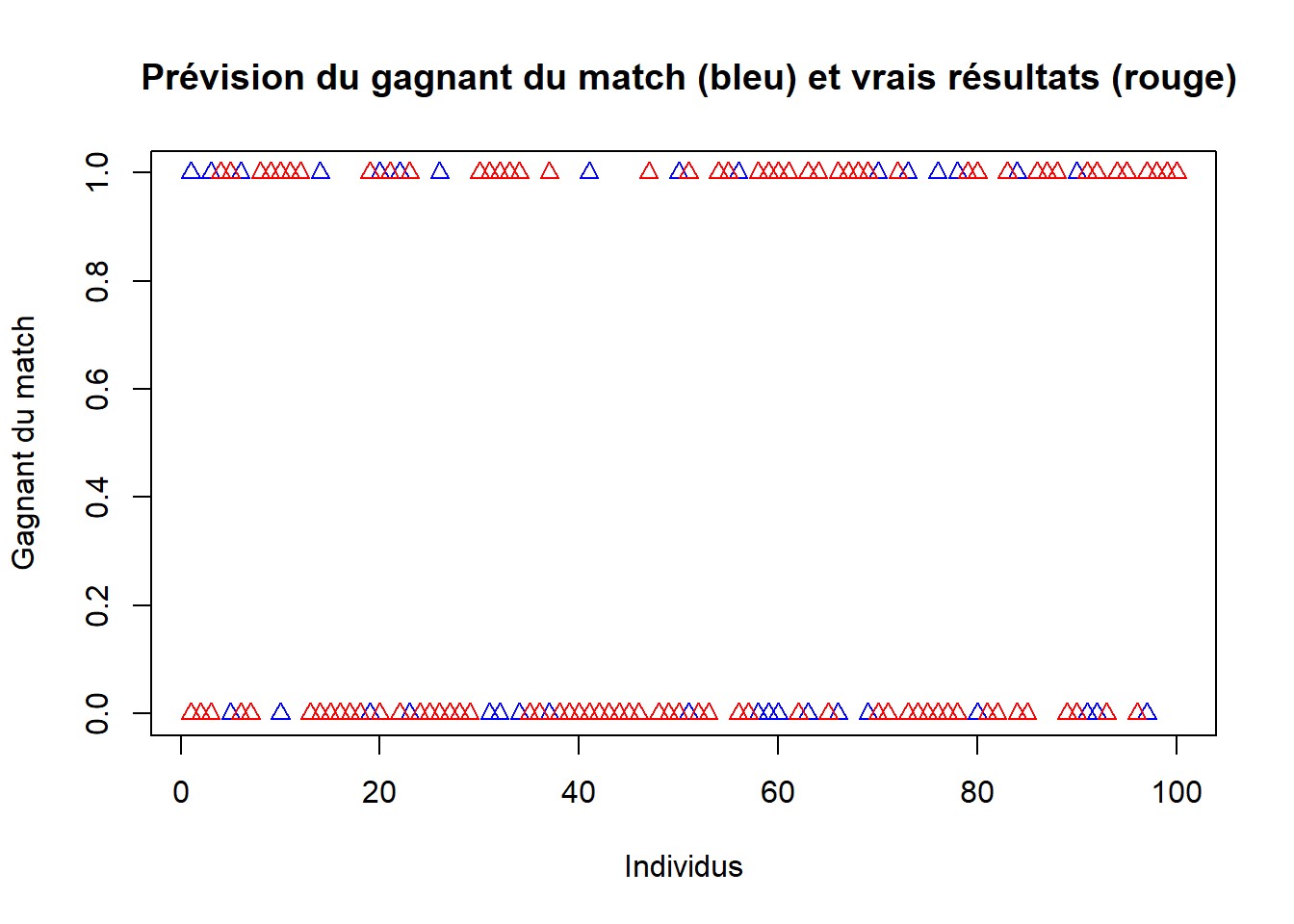
## Prédiction à l’aide du modèle complet

Comme le modèle complet est meilleur que le modèle avec seulement l’intercept, on effectue une prédiction du gagnant du match sur l’échantillon de test.

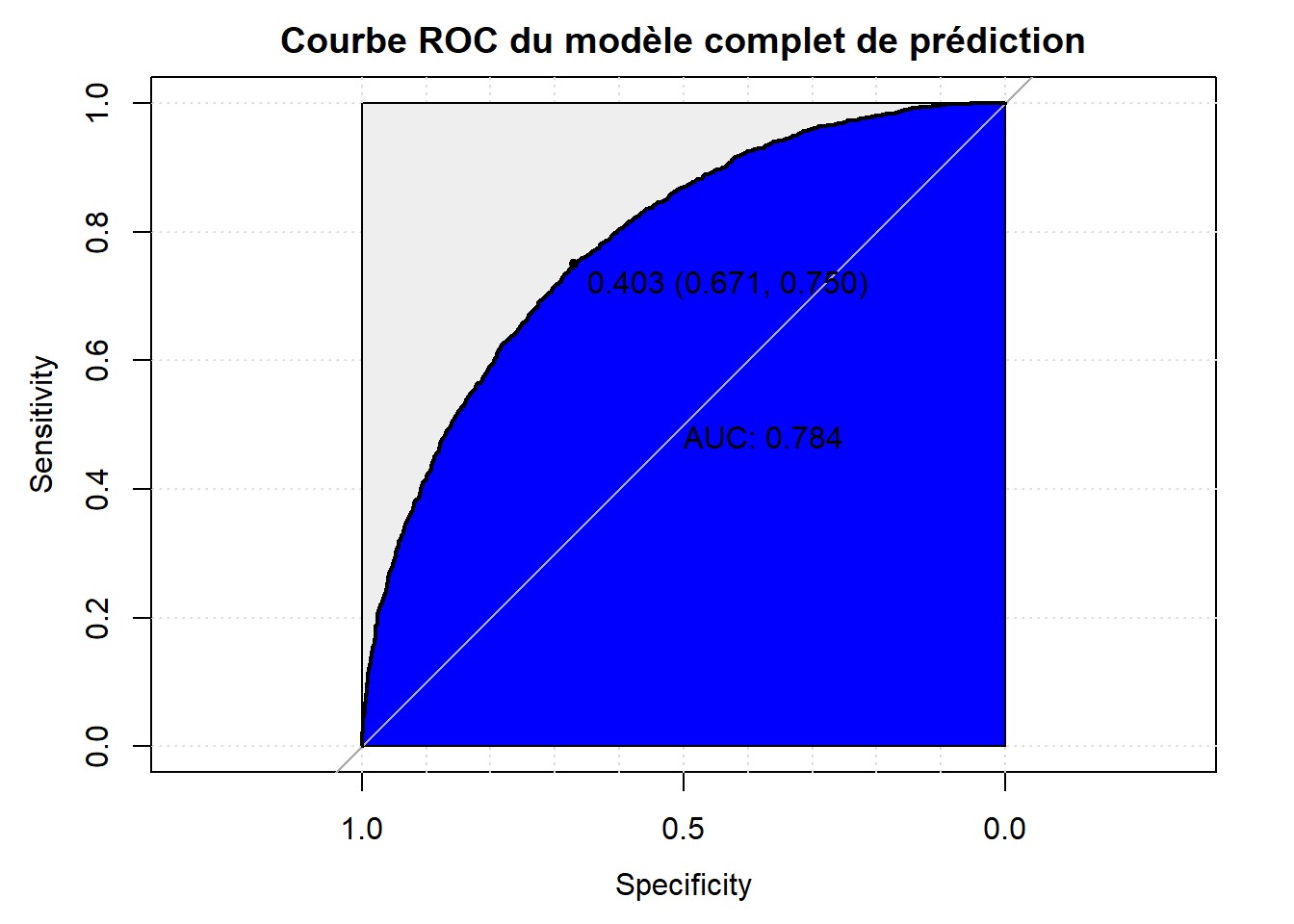
*Table de confusion*

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 1880 | 518 |
| 1 | 714 | 1161 |

On obtient 3041 sur 4273 de bonnes prédictions contre 1232 mauvaises, soit une prédiction juste dans 71,2 % des cas.



On constate que les prédictions coïncident assez souvent avec le vrai résultat. Nous avons seulement affiché les 100 premiers individus pour avoir un graphique lisible.

On s’intéresse maintenant à la courbe ROC.

On retrouve la qualité de la prédiction, en effet plus l’aire sous la courbe (AUC) est proche de 1, meilleure est la prédiction. Ici nous avons une AUC = 0.784 ce qui est assez satisfaisant.

Le seuil optimal (probabilité avec laquelle on attribue la valeur 1 au vecteur prédisant le gagnant du match) se lit sur la courbe et donne le meilleur seuil pour lequel la prédiction est la plus fiable.

## Sélection de variables d’après le critère AIC

On utilise la fonction stepAIC avec une recherche mixte (mélange de méthode forward et backward) afin de trouver le meilleur modèle au sens du critère AIC, puis on prédit de la même manière le vainqueur du match.

Voici les variables retenues pour expliquer le modèle obtenu par minimisation du critère AIC:

Offensive foul 6

Lost ball 6

Steals 6

Block 6

Assist 6

Bad pass 12

Block 12

Score 18

Defensive rebound 18

Score 24

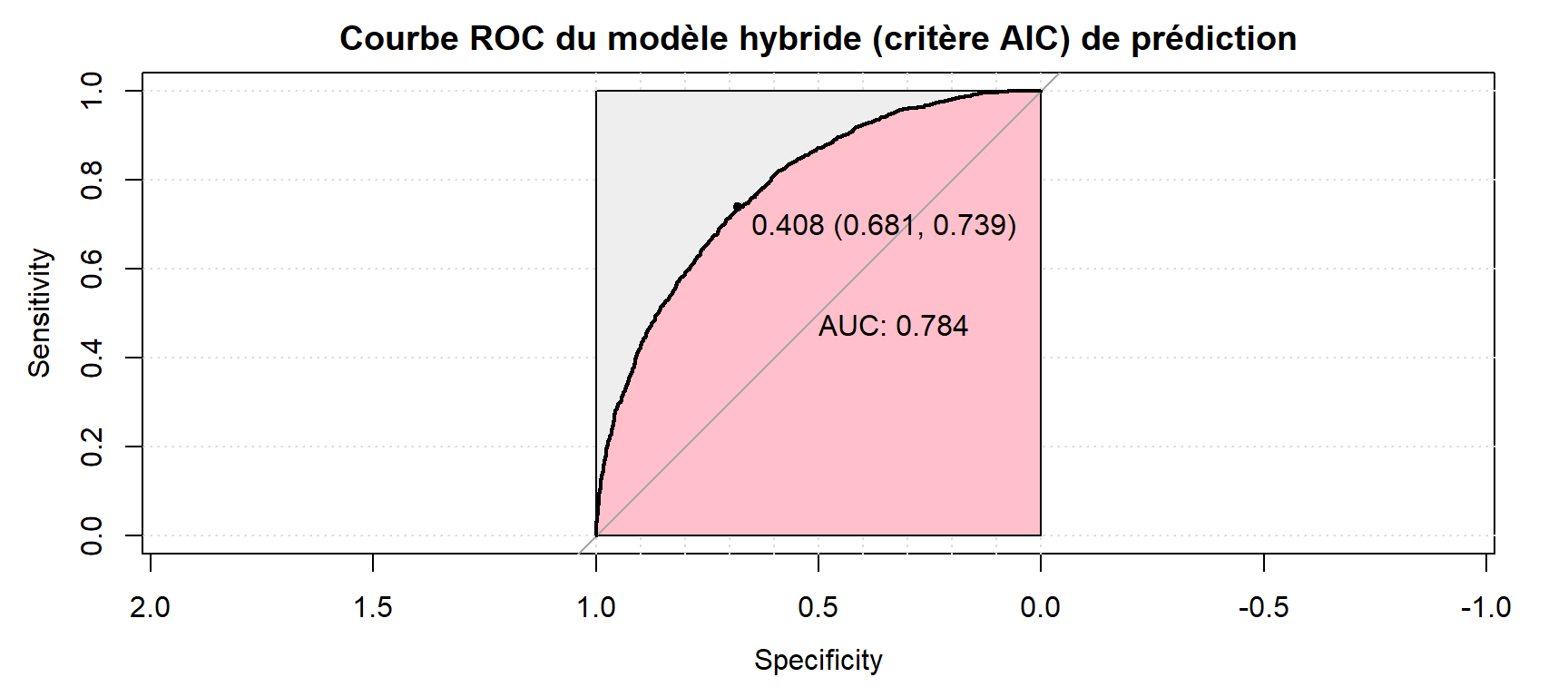
Defensive rebound 24

Lost ball 24

Bad pass 24

On obtient 3041 sur 4273 de bonnes prédictions contre 1232 mauvaises, soit une prédiction juste dans 71.2 % des cas. On retrouve un résultat très similaire à la prédiction précédente sauf qu’ici nous avons seulement 13 variables explicatives contre 44 dans le modèle complet et nous obtenons la même prédiction à très peu de choses près. Donc dans le cadre de la prédiction nous pouvons affirmer qu’à peu près 75% des variables explicatives n’apportent pas (ou très peu) d’information supplémentaire.

On retrouve, logiquement, une courbe ROC similaire :



Conclusion

En analysant le jeu de données, on a donc pu établir qu’il était possible de prévoir à 72% le résultat d’un match de NBA à partir des données de la première mi-temps du match en utilisant une règle de décision simple : si l’équipe visiteuse mène de plus d’un point à la mi-temps, elle va gagner sinon c’est l’équipe à domicile qui l’emportera.

En se basant sur cette règle, on a également pu prédire le vainqueur d’un match bien avant la mi-temps avec des taux de réussites (63% à 12 minutes / 68% à 18 minutes) inférieurs. Une méthode de prédiction plus complexe, la régression logistique, donne des résultats à peu près équivalents.

De plus, en consultant la page du challenge NBA, où sont recensés les meilleurs taux de prédictions, déterminés avec des outils fait pour l’analyse de données massives, on voit que les résultats sont meilleurs que les nôtres de seulement quelques pourcents (environ 75% pour les meilleurs). Cela semble indiquer qu’il sera difficile d’améliorer la prédiction sans collecter de nouvelles données ou d’améliorer celles du jeu actuel.

Tout d’abord, on a pu voir que les données relatives aux fautes défensives (Defensive foul) étaient manquantes. Des informations sur les deux équipes (favori, forme du moment, joueurs absents, …etc) pourraient également être utiles.

Avec plus de temps, d’autres méthodes comme RIDGE ou LASSO pourraient être utilisées pour améliorer la qualité de prédiction.