Analyse des cotes et des résultats de la Ligue 1

Projet Python M2 IMSD

Daniel Kharsa & Samir Lazzali

Novembre 2019

Sommaire

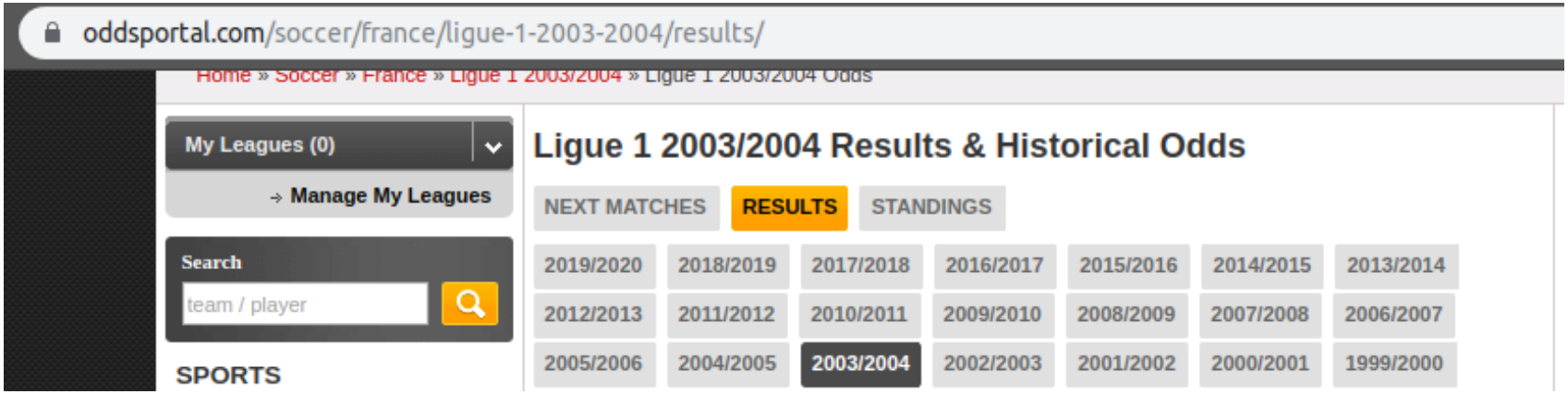
Introduction

- Dans le cadre du cours de Python, nous avons décidé de scraper et d'analyser des données relatives au championnat français de football professionnel (Ligue 1).
- Plus précisément, nous avons étudié les résultats des confrontations ainsi que les cotes attribuées par les bookmakers entre la saison 2004/2005 et la saison 2018/2019.
- Le présent PDF constitue la présentation de notre projet. L'analyse complète et le code du scraping se trouvent dans deux autres fichiers distincts (cf. mail)

Scraping

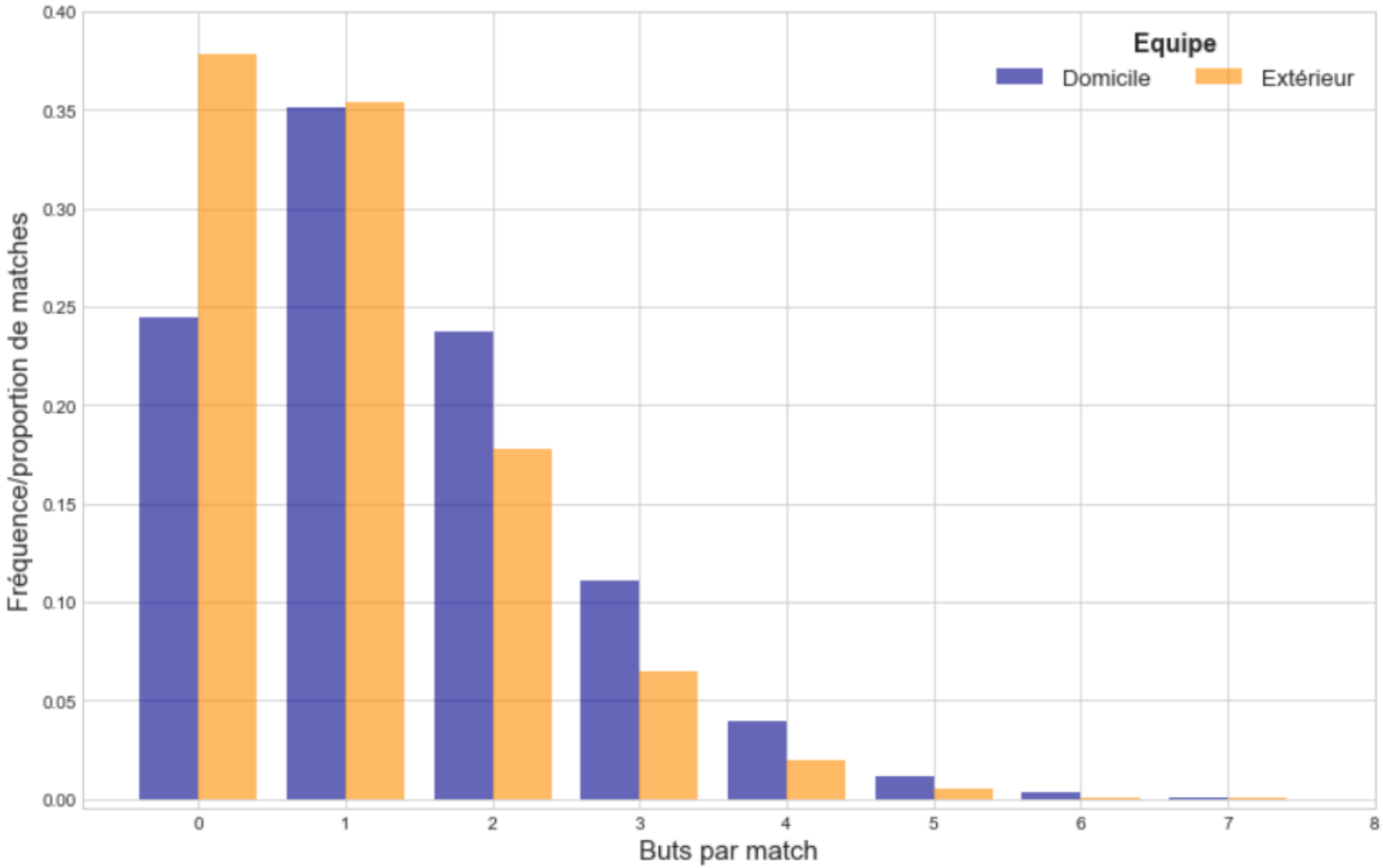
- Nous avons récupéré les données de la Ligue 1 ici.
- Le premier objectif était de récupérer les matches affiché sous forme de tableau. Celui-ci étant chargé de manière asynchrone, nous avons combiné Selenium à Beautifulsoup pour notre scraping.
- L'objectif suivant était de récupérer les matches de toutes les pages pour une saison.

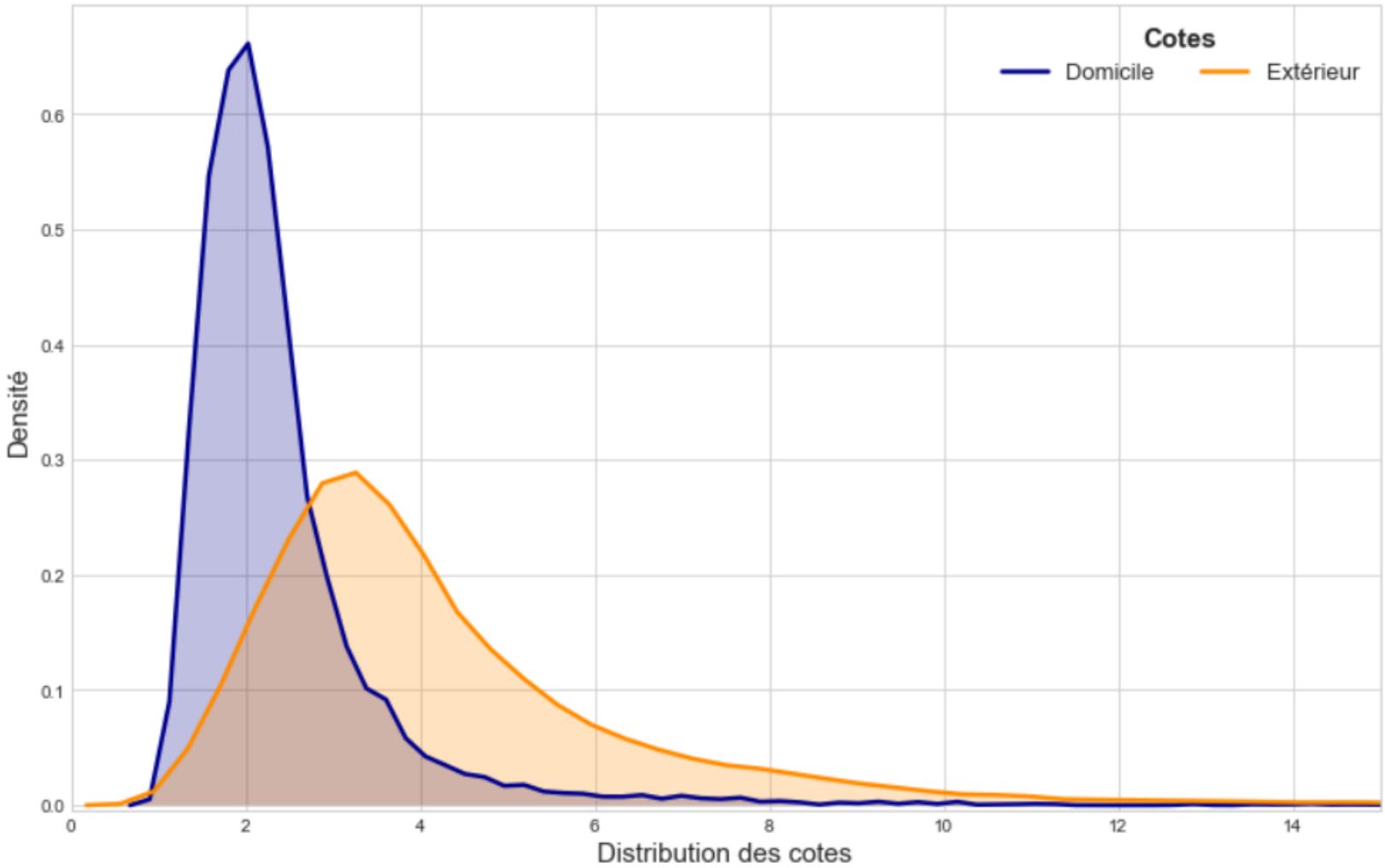
Soccer » France » Ligue 1 2003/2004						
23 May 2004			1	X	2	B's
18:00	**AC Ajaccio** - Metz	**3:1**	1.46	3.80	5.57	3
18:00	Bastia - **Paris SG**	**0:1**	3.06	2.61	2.42	3
18:00	**Le Mans** - Lens	**3:0**	1.88	3.60	3.17	3
18:00	**Lyon** - Lille	**3:0**	1.43	3.62	6.80	3
18:00	**Marseille** - Guingamp	**2:1**	2.24	3.55	2.49	3
18:00	**Nantes** - Sochaux	**3:1**	2.30	3.13	2.70	2
18:00	Nice - Auxerre	**1:1**	2.70	3.17	2.30	3
18:00	**Rennes** - Montpellier	**4:0**	1.43	3.73	6.22	3
18:00	Strasbourg - Toulouse	**0:0**	2.93	3.40	2.03	3
21 May 2004			1	X	2	B's
18:00	Bordeaux - **Monaco**	**1:3**	1.95	3.30	3.25	3


oddsportal.com/soccer/france/ligue-1-2003-2004/results/
Home » Soccer » France » Ligue 1 2003/2004 » Ligue 1 2003/2004 Odds
My Leagues (0)
Manage My Leagues
Search
team / player
SPORTS
Ligue 1 2003/2004 Results & Historical Odds
NEXT MATCHES
RESULTS
STANDINGS
2019/2020
2018/2019
2017/2018
2016/2017
2015/2016
2014/2015
2013/2014
2012/2013
2011/2012
2010/2011
2009/2010
2008/2009
2007/2008
2006/2007
2005/2006
2004/2005
2003/2004
2002/2003
2001/2002
2000/2001
1999/2000

Exploration de la base

- Après nettoyage, la base contient 5686 observations et 14 variables. Une observation correspond à un match ; les variables sont décrites dans un dictionnaire à part (cf. mail).
- L'équipe à domicile inscrit en moyenne plus de buts que l'équipe à l'extérieur (1.4 contre 1) mais les médianes sont égales à 1 $\Rightarrow$ les deux équipes s'affrontant inscrivent au plus 1 but dans plus de 50% des matches $\Rightarrow$ confirme l'analyse faite de la Ligue 1 :la rigueur défensive prime sur l'animation offensive.
- La cote de l'équipe évoluant à domicile est en moyenne inférieure à celle de l'équipe jouant à l'extérieur (2.4 contre 3.7).


Equipe
Domicile
Extérieur
Fréquence/proportion de matches
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
1
2
3
4
5
6
7
8
Buts par match


Cotes
Domicile
Extérieur
Densité
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
2
4
6
8
10
12
14
Distribution des cotes

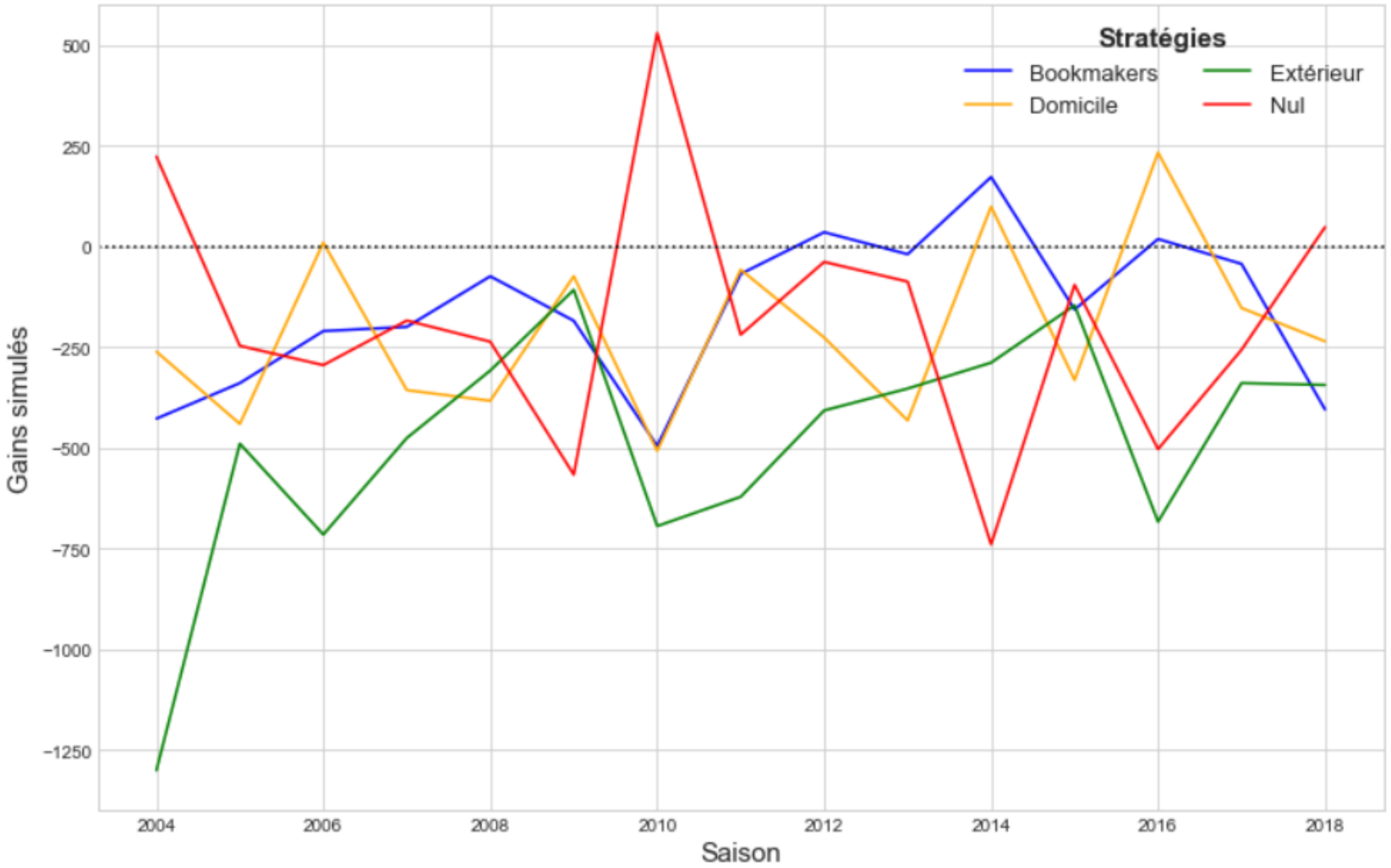
- Nous mettons en place un test non-paramétrique (test de Wilcoxon-Mann-Whitney) afin de tester statistiquement la différence entre les cotes attribuées à l'équipe à domicile et celles attribuées à l'équipe à l'extérieur.
- Plusieurs versions du test existent, nous choisissons celle de car les variables étudiées sont clairement liées (corrélation de Pearson -0.47, corrélation de Spearman $= -0.99$).
- On trouve une $p - value = 0.00$, ce qui nous amène à rejeter l'hypothèse nulle pour tous les niveaux de confiance α.

- Sur la périodé étudiée, l'évènement le plus probable d'après les bookmakers s'est effectivement réalisé dans seulement 50% des cas.
- Alors que 28% des matches joués ont débouché sur un match nul, les bookmaker ont attribué la cote la plus élevée à cette issue dans moins de 1% des cas (5 matches).
- S'explique par la tendance des bookmakers à surestimer la probabilité de victoire des équipes évoluant à domicile.

- Pour aller plus loin, nous allons tester l'indépendance entre les "prédictions" des bookmakers et les résultats observés avec un test du χ^2.
- On trouve une $p-value = 0.00$, ce qui nous amène à rejeter l'hypothèse nulle pour tous les niveaux de confiance α.
- Mais nous ne pouvons toutefois pas nous fier pleinement à notre test du fait de la faiblesse de certains des effectifs théoriques utilisés lors du calcul de la statistique de test.

Analyse des cotes des bookmakers

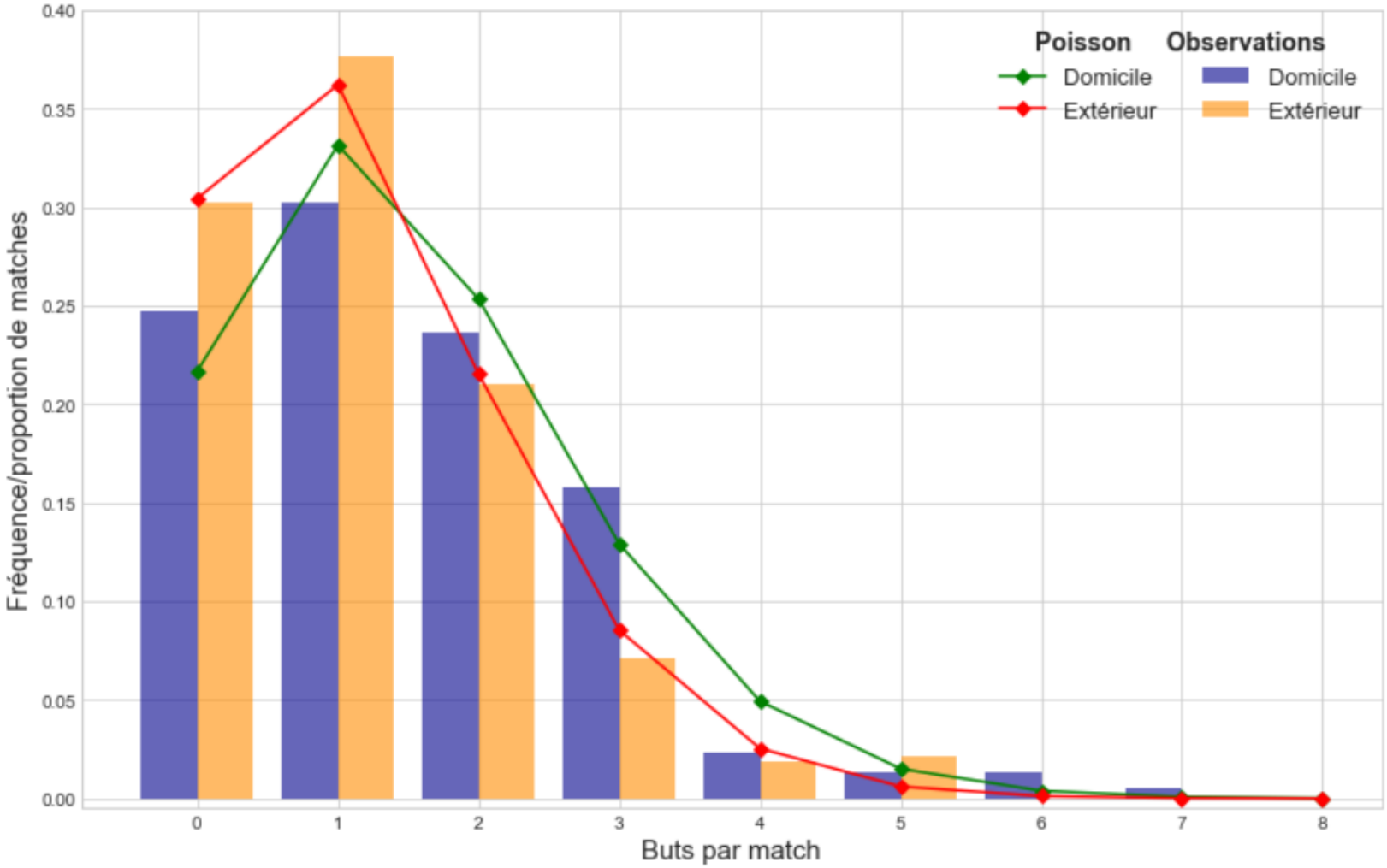
- L'intérêt de cette partie va être d'étudier la justesse des prévisions des bookmarkers.
- Nous allons simplement calculer les gains obtenus pour une mise de 10 euros sur chaque match en se fiant aux prévisions des bookmakers, puis comparer les gains obtenus avec ceux résultant d'une stratégie qui consisterait à toujours parier sur l'équipe à domicile, toujours parier sur l'équipe à l'extérieur, toujours parier match nul.


Stratégies
Bookmakers
Extérieur
Domicile
Nul
Gains simulés
500
250
0
−250
−500
−750
−1000
−1250
2004
2006
2008
2010
2012
2014
2016
2018
Saison

- Peu importe la stratégie choisie, les gains réalisés auraient été négatifs. difficiles à prédire.
- Logique car une règle de base des paris sportifs est de ne pas parier sur l'ensemble des matches mais plutôt de cibler et suivre attentivement les performances d'un nombre très restreint d'équipes.
- En reconduisant l'analyse par saison, on constate que peu importe la saison considérée, les gains sont négatifs dans l'écrasante majorité des cas.

Analyse à l'aide d'un modèle de Poisson

- La loi de Poisson est un candidat idéal à la modélisation du nombre moyen de but pendant un match.
- La loi de probabilité d'une variable aléatoire X suivant une loi de Poisson est donnée par : $P(X = x) = \frac{e^{-\lambda}\lambda^x}{k!}$, avec λ le nombre moyen de buts inscrits par match dans notre cas.
- On modélise séparemment les buts à domicile et à l'extérieur, puis on compare les probabilités obtenues à l'aide la loi de Poisson avec les fréquences effectivement observées


Poisson
Domicile
Extérieur
Observations
Domicile
Extérieur
Fréquence/proportion de matches
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
1
2
3
4
5
6
7
8
Buts par match

- On observe que l'équipe à domicile inscrit excatement un but dans près de 25% des matches, et que l'équipe à l'extérieur inscrit excatement un but dans près de 30% des matches.
- D'après les lois de Poisson utilisées, la probabilité que l'équipe à domicile inscrive exactement un but est d'environ 22%, et d'environ 30% pour l'équipe à l'extérieur.
- Bien qu'utile, un modèle basé uniquement sur les probabaiblités issues d'une loi de Poisson présente un faible pouvoir explicatif et prédictif $\Rightarrow$ on affine l'analyse en mettant en place une régression de Poisson.

- La modélisation du nombre de buts par match correspond à un des cas où l'utilisation d'un modèle linéaire "classique" n'est pas appropriée. D'où le choix d'une régression de Poisson.
- Pour enrichir l'analyse, nous construisons une variable pour la forme actuelle de l'équipe. Celle-ci va correspondre au nombre de buts inscrits lors des 5 derniers matches.

- L'ensemble des résultats de l'estimation et la méthode d'interprétation sont présents dans le Jupyter notebook.
- Afin de pouvoir interpréter les coefficients retournés par le logiciel, il est nécessaire de leur appliquer la fonction exponentiel. Par exemple, le coefficient associé à *forme* est égal à -0.0074. Lorsqu'on lui applique la fonction exponentiel on obtient $e^{-0.0074} = 0.992$, ce qui signifie qu'en moyenne, pour une hausse égale à 1 de *forme*, on s'attend à ce que la valeur prise par la variable *but* soit **multipliée** par 0.992.

- Concernant les autres coefficients de la régression, on voit que la variable *domicile* a un impact significativement positif ($e^{0.2487} = 1.3$).
- De même, la régression confirme que Lyon, Marseille ou le PSG sont des formations qui marquent en moyenne plus de buts que les autres, et à qui il est en moyenne plus difficile d'en marquer.