Universitat Autònoma de Barcelona

Grau en Estadística Aplicada

Treball de Final de Grau

Eines estadístiques en la modelització de resultats esportius

Autor: Artur Xarles Esparraguera

Tutora: Mercè Farré Cervelló

Juny de 2021

Cerdanyola del Vallès

Agraïments

A la meva tutora, la Mercè Farré, per tot el suport que m'ha donat des de l'inici d'aquest treball. Sense la seva dedicació, els seus consells i les seves correccions aquest projecte no hauria estat possible.

Als meus amics, per les hores de biblioteca compartides que ho han fet tot més senzill.

A la meva família, per donar-me suport en els estudis des de petit i permetre'm estar on soc ara.

Resum

Amb la popularització del futbol en el darrer segle, passant a ser l'esport més reconegut del món, va néixer l'interès per analitzar millor els partits de futbol i, conseqüentment, per intentar predir-ne els resultats. Del mateix interès neix aquest treball, amb l'objectiu principal d'obtenir un model que sigui capaç de fer prediccions dels resultats de partits de futbol amb la major precisió possible. Per dur a terme la modelització s'han utilitzat dades de partits de les cinc principals lligues de futbol europees, dels quals es té el resultat i un conjunt d'informació prèvia al partit. S'han ajustat diversos models distribucionals basats en dues distribucions de Poisson independents i en la distribució de Poisson Bivariant. També s'han ajustat els models anteriors amb la diagonal inflada per ajustar millor els empats. En tots els models s'ha aplicat l'algoritme EM per estimar els paràmetres distribucionals, i un model lineal generalitzat per expressar la relació dels paràmetres amb la informació prèvia. Després de comparar-los s'ha vist que el millor ajust era obtingut pel model basat en dues distribucions de Poisson independents amb la diagonal inflada. Aquest té un percentatge d'encert proper al 50%, similar a l'obtingut amb prediccions comercials. A més, s'han complert dos objectius secundaris del treball, crear un "dashboard"[1] per mostrar les prediccions d'un partit, i simular el tram final de la temporada 2020/21 per les diferents lligues considerades. Tot i assolir els objectius, es podrien considerar petites modificacions dels models proposats per millorar-ne l'ajust i la capacitat predictiva.
Paraules clau: Poisson bivariant, algoritme EM, model lineal generalitzat(GLM), diagonal inflada, selecció del millor model.

Abstract

Football popularization during the last century, becoming the most famous sport in the world, lead to the interest in a better understanding of football matches and, consequently, of trying to predict the results. This project is born from the same interest, where the main objective is to obtain a model capable of making predictions of football matches as well as possible. Data modelling has been done with data proceeding from football matches in the five main European leagues, and contain the final result and some pre-match information. Different distributional models based on two independent Poisson distributions and Bivariate Poisson distribution have been adjusted. An extension of the models of inflating the diagonal has also been considered for a better adjustment of the draws. The EM algorithm has been used in all the models to estimate the distributional parameters. Moreover, a generalized linear model has been considered to express the relationship between the parameters and the previous information. After comparing the different models, the model based on two independent Poisson distributions with the diagonal inflated has been the best. This one has an accuracy close to 50%, similar to the result of commercial predictions. In addition, two secondary objectives have been achieved, creating a dashboard to show the predictions of a match and simulating the different seasons final stretch. Despite achieving the objectives, some small modifications of the models could be considered to increase the predictive capacity.
Keywords: bivariate Poisson, EM algorithm, generalized linear model(GLM), diagonal inflated, best model selection.

[1] "dashboard": Representació visual de la informació més important que es necessita per complir un o més objectius, mostrada en una sola pantalla perquè la informació es pugui controlar amb una mirada.

Índex

Índex de figures

Índex de taules

1 Introducció

Des del segle III aC es té constància de diferents jocs de pilota que es poden assimilar al futbol. Tot i això, no és fins a l'any 1863, any en què es va fundar l'Associació Anglesa de Futbol, que es pot considerar l'inici de la història del futbol.[11] Des de llavors, ha tingut un creixement constant arribant a ser l'esport més popular del món. Amb la seva popularització també va néixer l'interès en intentar entendre millor el funcionament del joc i predir el resultat dels diferents partits. D'aquest mateix interès neix la motivació per dur terme aquest treball, on s'intentarà predir amb el major encert possible el resultat final de diferents partits de futbol. Així doncs, l'objectiu principal del treball consisteix en ajustar un model que sigui capaç de fer prediccions de partits de futbol tan precises com sigui possible.

Per dur a terme aquest estudi, es posa el focus en les cinc principals lligues de futbol europees: Bundesliga (Alemanya), LaLiga (Espanya), Ligue1 (França), Premier League (Anglaterra) i SerieA (Itàlia). Es disposa d'un total de 1683 partits dels quals es té el resultat final i un conjunt de variables que contenen informació prèvia al partit, tant dels jugadors com dels equips que hi participen. Amb aquestes dades s'intenta assolir l'objectiu principal del treball, utilitzant els diferents mètodes proposats a "FiveThirtyEight" [3] i a l'article de Karlis i Ntzoufras [7]. Aquests exposen diferents models per ajustar dades bivariants corresponents a recomptes, és a dir, amb les mateixes característiques que el resultat d'un partit de futbol. Els models que es proposen assumeixen que les dades segueixen dues distribucions de Poisson independents o una distribució de Poisson Bivariant. També consideren una extensió dels models corresponent a inflar la diagonal. Els models s'ajusten mitjançant l'algoritme EM (esperança i maximització) aplicat a les distribucions bivariants anteriorment descrites, els paràmetres de les quals s'enllacen amb les variables regressores (informació prèvia al partit) mitjançant un model lineal generalitzat. Finalment se selecciona el que obté uns millors resultats segons la metodologia de comparació exposada en la Secció 3.6.

A part d'obtenir un model de caràcter predictiu per als partits de futbol, aquest treball també intenta assolir altres objectius secundaris. Aquests son el de crear un "dashboard" per mostrar les prediccions fetes per a un determinat partit amb el model seleccionat i el de simular el tram final de la temporada 2020/21 per a les cinc lligues que s'estudien. Recursos de programació, com ara la paral·lelització o la creació de documents `html` s'han hagut de treballar al llarg de l'elaboració del treball.

Finalment, l'objectiu principal del treball ha estat assolit, obtenint un model basat en dues distribucions de Poisson independents amb la diagonal inflada. Aquest model presenta uns resultats bastant satisfactoris, sent capaç de predir correctament al voltant del 50% de partits. Aquests resultats son lleugerament inferiors als obtinguts per altres empreses de caràcter més professional. També s'han pogut assolir els objectius secundaris de crear un "dashboard" per a les prediccions de partits i simular el tram final de temporada. En aquest darrer cas, les simulacions obtingudes han donat resultats prou similars a la classificació final real de la temporada.

Es mostra al Capítol 2 les característiques del conjunt de dades del que es disposa. Al Capítol 3 hi ha l'explicació detallada de la metodologia que s'ha utilitzat en la realització de l'anàlisi i els resultats que s'han obtingut. Per acabar, al Capítol 4 es presenten els resultats obtinguts, i al Capítol 5 hi ha la discussió i les reflexions finals.

A l'Apèndix del treball s'hi poden trobar les demostracions teòriques dels algoritmes

EM utilitzats per ajustar els diferents models (Apèndix A). També hi ha taules i figures auxiliars (Apèndix B). Tot el codi utilitzat està disponible a l'enllaç `https://github.com/arturxe2/TFG-Artur-Xarles.git`.

2 Dades

Les dades amb les quals es basa aquest treball corresponen a 1683 partits de futbol de les principals lligues europees (Bundesliga, LaLiga, Ligue1, Premier League i SerieA) que s'han disputat entre el 29 de maig de 2020 i el 12 d'abril de 2021. Per a cada un d'ells es disposa dels equips que disputen el partit, el resultat final del partit, expressat en gols per equip i guanyador, i d'un conjunt de variables que contenen informació prèvia al partit dels jugadors i equips que hi participen, és a dir, sabent l'alineació inicial de cada equip. Aquestes variables s'han obtingut a partir de combinacions d'estadístiques disponibles a la pàgina web *Sports Reference*[13]. A continuació, s'enumera tota la informació disponible per a cada partit:

- Noms de l'equip local i visitant.
- Lliga a la qual pertanyen els equips.
- Jornada a la qual correspon el partit.
- Variable categòrica ordinal que expressa on es disputa el partit. Pot ser que es disputi en el camp de l'equip local a porta oberta, amb aforament limitat, a porta tancada o en un camp neutral.
- Per a cada equip, nombre de gols que ha encaixat el porter titular per partit (90 minuts), entre la temporada actual i l'anterior.
- Per a cada equip, proporció de tirs dirigits a porteria que ha aturat el porter titular, entre la temporada actual i l'anterior.
- Per a cada equip, suma de la mitjana entre gols marcats i gols esperats [2] per partit (90 minuts), per cada jugador, entre la temporada actual i l'anterior.
- Per a cada equip, suma de les assistències per partit (90 minuts) realitzades per cada jugador, entre la temporada actual i l'anterior.
- Per a cada equip, suma dels *tackles*[3] i intercepcions[4] per partit (90 minuts) realitzats per cada jugador, entre la temporada actual i l'anterior.
- Per a cada equip, mitjana del nombre de gols marcats per l'equip per partit (90 minuts) amb cadascun dels jugadors sobre el terreny de joc. És a dir, per a cada jugador el nombre de gols marcats per l'equip quan ell estava disputant el partit, en partits de la temporada actual i l'anterior.
- Per a cada equip, mitjana del nombre de gols encaixats per l'equip per partit (90 minuts) amb cadascun dels jugadors sobre el terreny de joc. És a dir, per a cada jugador el nombre de gols encaixats per l'equip quan ell estava disputant el partit, en partits de la temporada actual i l'anterior.
- Per a cada equip, proporció de punts aconseguits en la temporada actual respecte al total possible.
- Per a cada equip, proporció de punts aconseguits en els darrers 5 partits respecte al total possible.

[2] Els gols esperats tenen en compte la probabilitat que un xut acabi en gol, segons un model, en lloc del resultat final del tir. En aquest treball s'utilitzen els gols esperats calculats amb el model de la pàgina web "Sports Reference"[13]

[3] Pilotes robades a un jugador de l'equip contrari.

[4] Pilotes recuperades a través d'interceptar un passe d'un jugador de l'equip rival.

- Per a cada equip, proporció de punts aconseguits en la temporada actual jugant com a local o visitant, depenent de si son locals o visitants en el partit d'interès, respecte al total possible.
- Diferència de punts de l'equip local i visitant respecte la resta de posicions de la classificació.
- Per a cada equip, nombre de gols marcats en el darrer partit disputat entre els dos equips.
- Proporció de punts aconseguits per l'equip local en els darrers 5 partits disputats contra l'equip visitant, respecte al total de punts repartits.
- Resultat final del partit. Representem amb un 1 si ha guanyat l'equip local, un 2 si ha guanyat el visitant i una X si han empatat. (Variable resposta)
- Nombre de gols marcats per l'equip local i visitant en el partit d'interès. (Variable resposta)

3 Metodologia

En aquest apartat s'exposen els diferents mètodes utilitzats en la realització d'aquest treball. S'explica quin tractament s'ha realitzat a les dades inicials, els diferents models proposats per predir les respostes, els mètodes de selecció i de validació de models i, finalment, la metodologia utilitzada per mostrar les prediccions a través de l'elaboració d'un "dashboard" i per simular el tram final de temporada. També es mostren els recursos de programació que s'han utilitzat.

3.1 Tractament de dades

Com a pas previ a modelitzar les dades, s'ha realitzat un tractament de les variables explicatives que utilitzarem per intentar predir el resultat final del partit, per tal de reduir-ne la dimensió i la multicol·linealitat.

El primer pas ha estat transformar les variables relacionades amb la diferència de punts entre els equips i les diferents posicions de la classificació en una de nova per a cada equip, amb valors entre 0 i 1, que expressa la importància del partit per a l'equip. Aquesta variable té en compte les posicions importants en cada lliga, com pot ser la que et permet guanyar el campionat, la que et permet classificar per a una competició europea o no baixar de categoria, i en quina mesura l'equip està disputant aquella posició. Així doncs, fixat un equip, per a cada una de les posicions importants i es calcula el valor següent:

$$\text{IP}i = \begin{cases} 1 - min\{\frac{|difpos(i)|}{ptsR+5} + \frac{ptsR}{ptsT}, 1\} & \text{si } difpos(i) < 0 \text{ i } |difpos(i)| \leq ptsR \\ 0 & \text{si } difpos(i) < 0 \text{ i } |difpois(i)| \geq ptsR \\ 1 - min\{\frac{|difpos(i+1)|}{ptsR+5} + \frac{ptsR}{ptsT}, 1\} & \text{si } difpos(i) \geq 0 \text{ i } |difpos(i+1)| \leq ptsR \\ 0 & \text{si } difpos(i) \geq 0 \text{ i } |difpois(i+1)| \geq ptsR \end{cases}$$

on $difpos(a)$ és la diferència de punts entre l'equip que s'avalua i el que està situat en la posició a, $ptsR$ és el nombre de punts que queden per disputar de la temporada i $ptsT$ representa el nombre total de punts que es disputen.

Finalment, la variable serà per a cada equip el valor IPi màxim prenent per i les diferents posicions importants de cada lliga. D'aquesta manera, la variable representarà en quina mesura l'equip està competint per alguna posició important, d'acord a com estigui de proper a aquella posició i quants punts quedin per disputar-se. Valors alts (propers a l'1) indiquen que el partit és bastant important per a l'equip i valors baixos (propers al 0) indiquen que no ho és.

Seguidament, s'ha fet una anàlisi de multicol·linealitat per a les variables explicatives restants a través del VIF (factor d'inflació de la variància). Després de veure les variables més afectades per la multicol·linealitat s'ha decidit aplicar les següents modificacions a la base de dades:

- Eliminar la variable corresponent a la jornada del partit, ja que ja està present en la importància del partit per a cada equip.
- Combinació, a través de la mitjana ponderada, de la variable corresponent als gols encaixats pel porter (amb pes $\frac{1}{11}$) i la que correspon als gols encaixats per l'equip amb els diferents jugadors sobre el terreny de joc (amb pes $\frac{10}{11}$).

- Combinació, a través de la mitjana, de la variable corresponent als gols marcats pels jugadors i la que correspon als gols marcats per l'equip amb els diferents jugadors sobre el terreny de joc.

Amb aquestes modificacions s'aconsegueix reduir el nombre de variables a considerar i la multicol·linealitat entre elles. Tot i que al tractar-se d'un model amb finalitat predictiva no hagués set necessari fer aquesta anàlisi de multicol·linealitat, ja que no afecta la capacitat de predicció, s'ha preferit fer-la per evitar incoherències en els paràmetres del model.

3.2 Distribucions proposades

A continuació s'exposen les diferents distribucions que es proposen per ajustar el resultat final d'un partit, expressat en gols de l'equip local i gols de l'equip visitant.

- **Dues variables Poisson independents**

  Es defineix la distribució de dues variables Poisson independents com:

  Definició: $(X,Y) \sim 2P(\lambda_1,\lambda_2)$, $\lambda_i > 0$, si $X \sim Pois(\lambda_1)$, $Y \sim Pois(\lambda_2)$ i son dues variables aleatòries independents.
  Es pot expressar la seva funció de probabilitat com:

$$f_{2P}(x,y) = P\{X=x, Y=y\} = P\{X=x\} \cdot P\{Y=y\} = \frac{e^{-(\lambda_1+\lambda_2)}\lambda_1^x\lambda_2^y}{x!y!} \quad (1)$$

  Al ser X i Y independents es té que $Cov(X,Y) = 0$. A més, $E[X] = Var(X) = \lambda_1$ i $E[Y] = Var(Y) = \lambda_2$.

- **Poisson Bivariant**

  Es defineix la distribució de Poisson Bivariant com:

  Definició: $(X,Y) \sim BP(\lambda_1,\lambda_2,\lambda_3)$, $\lambda_i > 0$, si $\exists$ X_1, X_2, $X3$ tal que $X_i \sim Pois(\lambda_i)$, independents i $X = X_1 + X_3$ i $Y = X_2 + X_3$.
  Es pot expressar la seva funció de probabilitat com:

$$\begin{aligned} f_{BP}(x,y) &= P\{X=x,Y=y\} = P\{i : (X_1,X_2,X_3) = (x-i, y-i, i)\} \\ &= \sum_{i=0}^{min(x,y)} \frac{e^{-\lambda_1}\lambda_1^{x-i}}{(x-i)!}\frac{e^{-\lambda_2}\lambda_2^{y-i}}{(y-i)!}\frac{e^{-\lambda_3}\lambda_3^i}{i!} \\ &= e^{-(\lambda_1+\lambda_2+\lambda_3)}\lambda_1^x\lambda_2^y \sum_{i=0}^{min(x,y)} \frac{1}{(x-i)!(y-i)!i!}(\frac{\lambda_3}{\lambda_1\lambda_2})^i \\ &= e^{-(\lambda_1+\lambda_2+\lambda_3)}\frac{\lambda_1^x}{x!}\frac{\lambda_2^y}{y!} \sum_{i=0}^{min(x,y)} \binom{x}{i}\binom{y}{i} i!(\frac{\lambda_3}{\lambda_1\lambda_2})^i \quad (2) \end{aligned}$$

  Aquesta distribució permet dependència entre X i Y ja que $Cov(X,Y) = Cov(X_1 + X_3, X_2 + X_3) = Cov(X_1,X_2) + Cov(X_1,X_3) + Cov(X_3,X_2) + Cov(X_3,X_3) = Var(X_3) = \lambda_3$. El paràmetre λ_3 expressa la dependència entre X i Y. De

forma marginal, les variables X i Y queden representades per la suma de dues distribucions de Poisson independents, de manera que $X \sim Pois(\lambda_1 + \lambda_3)$ i $Y \sim Pois(\lambda_2 + \lambda_3)$. En el cas de $\lambda_3 = 0$ la distribució Poisson Bivariant coincideix amb la distribució de dues Poisson independents, ja que (1) = (2) en aquest cas.

3.3 GLM (Model Lineal Generalitzat)

En aquest apartat es presenta el funcionament del model lineal generalitzat (GLM), que servirà per relacionar la variable resposta amb les distribucions proposades i el conjunt de variables explicatives.

Tal com el seu nom indica, el GLM és una generalització del model de regressió lineal ordinari. En aquest es considera una variable resposta Y de distribució Normal, un conjunt de variables regressores $X = (X_1, ..., X_p)$ i un vector de coeficients desconeguts associats a aquestes variables, $\beta = (\beta_0, \beta_1, ..., \beta_p)$. A partir de les covariables i els coeficients, es pot expressar el predictor lineal com $L(X) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$, on $L(X) \in \mathbb{R}$. El model de regressió lineal assumeix que un valor de la variable resposta Y vindrà donat a partir del predictor lineal $L(X)$ i un error, $Y = L(X) + \varepsilon$, on $\varepsilon \sim N(0, \sigma^2)$. Així doncs, $E[Y] = \mu = L(X)$. $L(X)$ es sol expressar com $\mathbf{X}\beta$ on $\mathbf{X}$ és la matriu de disseny $n \times p$ $(1, X_1, ..., X_p)$ amb n el nombre de casos.

A diferència del model de regressió lineal, en el model lineal generalitzat es pot assumir una distribució de Y diferent de la distribució Normal, com per exemple la distribució de Poisson o la de Bernoulli. En aquests casos a vegades no és correcte suposar que $E[Y] = L(X)$, com quan el domini de $E[Y]$ és tan sols un subconjunt de $\mathbb{R}$, per exemple [0, 1]. Per solucionar aquest problema s'utilitza una funció d'enllaç ("link") que estableixi la relació entre el predictor lineal i la mitjana de Y. Establerta la funció "link", $g()$, se suposa $g(E[Y]) = L(X)$.

Un cop s'assumeix la distribució de Y i la funció "link" a utilitzar, s'ha d'estimar el vector de paràmetres desconeguts β a partir d'una mostra de n observacions independents $(y_1, x_1), ..., (y_n, x_n)$. El mètode utilitzat per estimar β és a través de maximitzar la funció de versemblança, és a dir, trobar el conjunt de valors dels paràmetres que maximitzin:

$$\mathcal{L}(\beta | \{(y_i, x_i)\}) = \prod_{i=1}^{n} P(y_i \cap x_i)$$

on P indica la funció de probabilitat en el cas donat.

Aquesta funció variarà segons la distribució que segueixi la variable resposta Y i, en general, s'ha de maximitzar a través de mètodes numèrics. Els estimadors de màxima versemblança tenen la propietat asimptòtica que $\hat{\beta}_j \overset{asym.}{\sim} N(\beta_j, Var(\hat{\beta}_j))$. Un dels mètodes utilitzats per estimar els coeficients és l'algoritme de "Mínims quadrats iterativament ponderats" (IWLS). Aquest algoritme consisteix en els passos següents:

1. Obtenir una estimació inicial pels paràmetres. Una opció és realitzar una regressió lineal entre la transformació de la variable resposta observada a través de la funció "link" i la resta de variables explicatives. A través del mètode de mínims quadrats ordinaris es pot obtenir la primera estimació:

$$\hat{\beta} = (X^t X)^{-1} (X^t g(Y))$$

2. A partir de les estimacions, obtenir els valors predits pel predictor lineal i a partir d'aquests, a través de la inversa de la funció "link", els valors esperats de les respostes, $\hat{\mu}_i = g^{-1}(\mathbf{X}\hat{\beta})$.

3. Construir una variable de treball Z_i que és la diferència entre el valor esperat i el valor observat, i una matriu de ponderació W que té en la diagonal les variàncies dels valors esperats $\hat{\mu}_i$.

4. Fer una regressió de la variable Z a través de les variables explicatives X considerant les ponderacions W, de manera que podem obtenir una estimació a través de mínims quadrats ponderats:

$$\hat{\beta} = (X^t W X)^{-1}(X^t W Z)$$

5. Repetir els passos 2, 3 i 4 iterativament fins que convergeixi o es cumpleixi un criteri de parada.

3.3.1 Regressió de Poisson

En primer lloc, es presenta el funcionament del GLM en el cas concret de suposar que es té una variable resposta univariant tal que $Y \sim Pois(\lambda)$, de manera que $f_P = P(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!}$. Es té que $E[Y] = \lambda \in (0, +\infty)$. Com que $L(X) \in \mathbb{R}$ s'ha d'utilitzar una funció "link" que passi $(0, +\infty) \mapsto (-\infty, +\infty)$. S'utilitza el logaritme neperià, de manera que es pot expressar la mitjana de la distribució a partir del predictor lineal de la manera següent:

$$\ln(\lambda) = L(X) \rightarrow \lambda = e^{L(X)}$$

Així doncs, establerta la distribució de Y i la relació entre la seva mitjana i el predictor lineal a través de la funció "link", es procedeix a estimar els coeficients a través de maximitzar la funció de versemblança. Si es té un conjunt d'observacions $(y_1, x_1), ..., (y_n, x_n)$ la funció és la següent:

$$\begin{aligned}
\mathcal{L}(\beta|\{(y_i, x_i)\}) &= \prod_{i=1}^{n} P(y_i, x_i) = \prod_{i=1}^{n} P(y_i|x_i)P(x_i) \propto {}^{5} \prod_{i=1}^{n} P(y_i|x_i) \\
&= \prod_{i=1}^{n} \frac{(\lambda|x_i)^{y_i}}{y_i!} e^{-(\lambda|x_i)} \propto \prod_{i=1}^{n} (\lambda|x_i)^{y_i} e^{-(\lambda|x_i)} = \prod_{i=1}^{n} e^{L(x_i)y_i} e^{-e^{L(x_i)}} \\
&= \prod_{i=1}^{n} e^{L(x_i)y_i - e^{L(x_i)}} = exp[\sum_{i=1}^{n} (y_i L(x_i) - e^{L(x_i)})]
\end{aligned}$$

Maximitzar aquesta funció equival a maximitzar el seu logaritme, de manera que es pot maximitzar la següent funció de log-versemblança:

$$\ln(\mathcal{L}(\beta|\{(y_i, x_i)\})) = \sum_{i=1}^{n} (y_i L(x_i) - e^{L(x_i)})$$

Com s'ha comentat, l'estimació dels paràmetres es pot fer a través de l'algoritme IWLS entre altres. Un cop estimats es poden utilitzar per fer prediccions de λ, donat el vector de covariables X, a partir d'aplicar la fórmula:

$$\hat{\lambda}|X = e^{L(X)} = e^{\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j X_j} \tag{3}$$

[5] S'eliminen les constants que no depenen dels paràmetres.

3.3.2 Regressió de dues Poisson independents

Si s'assumeix que la variable resposta és bivariant, on $(X, Y) \sim 2P(\lambda_1, \lambda_2)$, el procés per estimar el model consistirà simplement en aplicar dues regressions de Poisson tal com s'ha exposat en la Secció 3.3.1. Si es considera el conjunt de covariables W_1 per estimar la variable X i el conjunt de covariables W_2 per a Y, podrem trobar les estimacions dels paràmetres λ_1 i λ_2 a partir de les estimacions de les regressions:

$$\hat{\lambda}_1 = e^{\hat{\beta}_{1,0} + \sum_{j=1}^{p} \hat{\beta}_{1,j} W_{1,j}}$$
$$\hat{\lambda}_2 = e^{\hat{\beta}_{2,0} + \sum_{j=1}^{p} \hat{\beta}_{2,j} W_{2,j}}$$

Podem ajustar aquest model a partir de la funció `glm2pois` disponible a l'enllaç `https://github.com/arturxe2/TFG-Artur-Xarles.git`.

3.3.3 Regressió de Poisson Bivariant

A continuació, s'assumeix que la variable resposta bivariant segueix la distribució Poisson Bivariant. Es té el conjunt de covariables W i la distribució $(X, Y) \sim BP(\lambda_1, \lambda_2, \lambda_3)$. Segons la distribució proposada en la secció 3.2, $\lambda_i = E(X_i)$ per $i = 1, 2, 3$. Com que es tenen tres paràmetres corresponents a l'esperança de les corresponents distribucions de Poisson, s'han de considerar tres predictors lineals. En tractar-se de tres regressions de Poisson, la funció "link" utilitzada és el logaritme neperià i per tant es té:

$$\ln(\lambda_i) = L_i(W) \rightarrow \lambda_i = e^{L_i(W)}$$

per a valors de $i = 1, 2, 3$.

Tal com es fa amb la regressió de Poisson, es troba la funció de versemblança a maximitzar:

$$\begin{aligned}
\mathcal{L}(\beta|\{(x,y),w\}) &= \prod_{i=1}^{n} P((x_i,y_i),w_i) = \prod_{i=1}^{n} P((x_i,y_i)|w_i)P(w_i) \propto \prod_{i=1}^{n} P((x_i,y_i)|w_i) \\
&= \prod_{i=1}^{n} e^{-(\lambda_1+\lambda_2+\lambda_3)} \frac{\lambda_1^{x_i}}{x_i!} \frac{\lambda_2^{y_i}}{y_i!} \sum_{j=0}^{min(x_i,y_i)} \binom{x_i}{j} \binom{y_i}{j} j! (\frac{\lambda_3}{\lambda_1\lambda_2})^j \\
&\propto \prod_{i=1}^{n} e^{-(\lambda_1+\lambda_2+\lambda_3)} \lambda_1^{x_i} \lambda_2^{y_i} \sum_{j=0}^{min(x_i,y_i)} \binom{x_i}{j} \binom{y_i}{j} j! (\frac{\lambda_3}{\lambda_1\lambda_2})^j \\
&= \prod_{i=1}^{n} e^{-(e^{L_1(w_i)}+e^{L_2(w_i)}+e^{L_3(w_i)})} e^{L_1(w_i)x_i} e^{L_2(w_i)y_i} \\
&\cdot \sum_{j=0}^{min(x_i,y_i)} \binom{x_i}{j} \binom{y_i}{j} j! \cdot \Big(\frac{e^{L_3(w_i)}}{e^{L_1(w_i)}e^{L_2(w_i)}}\Big)^j \\
&= \prod_{i=1}^{n} e^{-(e^{L_1(w_i)}+e^{L_2(w_i)}+e^{L_3(w_i)})+L_1(w_i)x_i+L_2(w_i)y_i} \\
&\cdot \sum_{j=0}^{min(x_i,y_i)} \binom{x_i}{j} \binom{y_i}{j} j! e^{(L_3(w_i)-L_1(w_i)-L_2(w_i))j}
\end{aligned}$$

La corresponent log-versemblança és:

$$\begin{aligned}\ln(\mathcal{L}(\beta|\{(x_i,y_i),w_i\})) &= \sum_{i=1}^{n} -(e^{L_1(w_i)}+e^{L_2(w_i)}+e^{L_3(w_i)}) + L_1(w_i)x_i + L_2(w_i)y_i \\ &+ \ln\Big(\sum_{j=0}^{min(x_i,y_i)} \binom{x_i}{j}\binom{y_i}{j} j! e^{(L_3(w_i)-L_1(w_i)-L_2(w_i))j}\Big)\end{aligned}$$

Maximitzar aquesta funció a través de mètodes numèrics convencionals és computacionalment bastant costós, ja que conté un sumatori dins d'un logaritme. Així doncs, per obtenir els estimadors de màxima versemblança es proposa el mètode EM (estimació i maximització) presentat a continuació.

- **Algoritme EM**

  L'algoritme EM és un mètode a través del qual es poden obtenir aproximacions dels estimadors de màxima versemblança. S'utilitza principalment en presència de variables latents, és a dir, variables que no s'observen directament en la mostra. El funcionament d'aquest algoritme es basa a anar iterant entre dos passos. El primer pas ("E-step") pretén estimar el valor esperat de la variable latent, mentre que el segon pas ("M-step") maximitza l'ajust dels paràmetres del model amb totes les observacions. Un cop els paràmetres convergeixen s'atura la iteració.

  En el nostre cas, on s'intenta estimar els paràmetres d'una regressió Poisson bivariant es pot veure que hi ha presència de variables latents, ja que s'observa (X,Y) i la distribució té les variables latents (X_1,X_2,X_3). Tot i això, aquestes es poden expressar a través de les dues variables observades i una sola variable latent Z:

  $$\begin{aligned}X_1 &= X - Z \\ X_2 &= Y - Z \\ X_3 &= Z\end{aligned}$$

  Un cop declarada la variable latent, l'algoritme que ens permet trobar una aproximació dels estimadors de màxima versemblança és el següent:

  1. Donar valors inicials dels paràmetres λ_1, λ_2 i λ_3 per a cada observació.
  2. Realitzar els següents dos passos fins que convergeixi:
     (a) E-step: Donat el conjunt de paràmetres $\theta^* = (\lambda_1,\lambda_2,\lambda_3)$ de l'anterior iteració calcular:
     $$z_i = \begin{cases} 0 & \text{si } x\cdot y = 0 \\ \lambda_{3i}\frac{f_{BP}(x_i-1,y_i-1)}{f_{BP}(x_i,y_i)} & \text{altrament}\end{cases}$$
     on f_{BP} és la funció de probabilitat de la distribució Poisson bivariant, de (2).
     (b) M-step: Trobar:
     $$\begin{aligned}\hat{\beta}_1 &= \hat{\beta}(x-z,W) \\ \hat{\beta}_2 &= \hat{\beta}(y-z,W) \\ \hat{\beta}_3 &= \hat{\beta}(z,W)\end{aligned}$$
     on $\hat{\beta}$ és l'estimador de màxima versemblança d'una regressió de Poisson. Estimar λ_1, λ_2 i λ_3 a partir de $\hat{\beta}_1$, $\hat{\beta}_2$ i $\hat{\beta}_3$.

3. Un cop convergeixi ja tindrem estimats els paràmetres $\hat{\beta}_1$, $\hat{\beta}_2$ i $\hat{\beta}_3$ corresponents als coeficients dels predictors lineals $L_1(W)$, $L_2(W)$ i $L_3(W)$.

Cal destacar que es pot tenir un conjunt de covariables diferent per a cada un dels paràmetres, de manera que podem utilitzar diferents variables per estimar λ_1, λ_2 i λ_3.

La demostració de l'algoritme es troba a l'apèndix A.1, i el model es pot ajustar a partir de la funció `glmbpois` disponible a l'enllaç `https://github.com/arturxe2/TFG-Artur-Xarles.git`.

3.4 Extensió dels models

En altres estudis realitzats, entre ells el de "FiveThirtyEight"[3], s'ha observat que al modelar el resultat d'un partit de futbol a través de regressions basades en la distribució de Poisson, el nombre d'empats esperats és inferior al nombre observat. És per això que en aquest treball es considera una extensió dels models proposats, que consisteix a *inflar la diagonal* de les distribucions proposades. Així doncs, es dona una probabilitat superior a valors de X i Y iguals, de manera que la probabilitat que el resultat final del partit sigui empat serà superior a la dels models 2P i BP. A continuació s'expliciten els models amb aquesta extensió.

- **Regressió de dues Poisson independents amb diagonal inflada**

Podem definir la distribució de dues variables Poisson independents amb la diagonal inflada com:

Definició: $(X, Y) \sim 2PDI(\lambda_1, \lambda_2, p, \theta)$ on $\lambda_i > 0$, $p \in [0, 1]$ i $\theta > 0$ amb funció de probabilitat:

$$f_{2PDI}(x, y) = \begin{cases} (1-p) \cdot f_{2P}(x, y; \lambda_1, \lambda_2) & \text{si } x \neq y \\ (1-p) \cdot f_{2P}(x, y; \lambda_1, \lambda_2) + p \cdot f_P(x; \theta) & \text{si } x = y \end{cases} \tag{4}$$

Es té que $E[X] = (1-p) \cdot \lambda_1 + p\theta$ i $E[Y] = (1-p) \cdot \lambda_2 + p\theta$.

En aquest cas també es pot usar una variable latent relacionada amb el paràmetre p, de manera que per estimar els paràmetres del model s'utilitza també l'algoritme EM. Definim aquesta variable latent com $V \sim Bernoulli(p)$. L'algoritme per trobar una aproximació dels estimadors de màxima versemblança és el següent:

1. Donar valors inicials dels paràmetres λ_1, λ_2, p i θ per a cada observació.
2. Realitzar els següents dos passos fins que convergeixi:
   (a) E-step: Donat el conjunt de paràmetres $\theta^* = (\lambda_1, \lambda_2, p, \theta)$ de l'anterior iteració calcular:

$$v_i = \begin{cases} 0 & \text{si } x \neq y \\ \frac{f_P(x_i) \cdot p_i}{(1-p_i) \cdot f_{2P}(x_i, y_i) + p_i \cdot f_P(x_i)} & \text{si } x = y \end{cases}$$

   (b) M-step: Trobar:

$$\hat{\beta}_1 = \hat{\beta}(x, W; (1-v))$$
$$\hat{\beta}_2 = \hat{\beta}(y, W; (1-v))$$

$$\hat{p} = \frac{\sum_{i=1}^{n} v_i}{n}$$

$$\hat{\theta} = \frac{1}{\sum_{i=1}^{n} v_i} \sum_{i=1}^{n} x_i \cdot v_i$$

on $\hat{\beta}(x, W; v)$ és l'estimador de màxima versemblança d'una regressió de Poisson amb vector de pesos v. Estimar λ_1 i λ_2 a partir de $\hat{\beta}_1$ i $\hat{\beta}_2$.

3. Un cop convergeixi ja es tindran estimats els paràmetres $\hat{\beta}_1$ i $\hat{\beta}_2$ corresponents als coeficients dels predictors lineals $L_1(W)$ i $L_2(W)$. També es tindrà l'estimació de $\hat{p}$ i $\hat{\theta}$.

En aquest cas s'utilitza el mateix valor de p i θ per a totes les observacions, però es podria afegir complexitat al problema a partir d'utilitzar una sèrie de covariables per a estimar aquests valors. També es podria utilitzar una distribució diferent de la de Poisson per inflar la diagonal. La demostració de l'algoritme es pot trobar en l'apèndix A.2 i el model es pot ajustar a partir de la funció `glm2poisDI` disponible a l'enllaç `https://github.com/arturxe2/TFG-Artur-Xarles.git`.

- **Regressió de Poisson Bivariant amb diagonal inflada**

Podem definir la distribució de Poisson Bivariant amb la diagonal inflada com:

Definició: $(X, Y) \sim BPDI(\lambda_1, \lambda_2, \lambda_3, p, \theta)$ on $\lambda_i > 0$, $p \in [0, 1]$ i $\theta > 0$ amb funció de probabilitat:

$$f_{BPDI}(x, y) = \begin{cases} (1-p) \cdot f_{BP}(x, y; \lambda_1, \lambda_2, \lambda_3) & \text{si } x \neq y \\ (1-p) \cdot f_{BP}(x, y; \lambda_1, \lambda_2, \lambda_3) + p \cdot f_P(x; \theta) & \text{si } x = y \end{cases} \tag{5}$$

Amb aquesta distribució es té que $E[X] = (1-p) \cdot (\lambda_1 + \lambda_3) + p\theta$ i $E[Y] = (1-p) \cdot (\lambda_2 + \lambda_3) + p\theta$.

Considerant aquesta distribució com a resposta es tenen dues variables latents, una relacionada amb el paràmetre d'inflació p ($V \sim Bernoulli(p)$) i la que ja s'ha considerat en la regressió de Poisson Bivariant (Z). També s'utilitzarà l'algoritme EM per trobar una aproximació dels estimadors de màxima versemblança. Aquest algoritme és el següent:

1. Donar valors inicials dels paràmetres λ_1, λ_2, λ_3, p i θ per a cada observació.
2. Realitzar els següents dos passos fins que convergeixi:
   (a) E-step: Donat el conjunt de paràmetres $\theta^* = (\lambda_1, \lambda_2, \lambda_3, p, \theta)$ de l'anterior iteració calcular:

$$v_i = \begin{cases} 0 & \text{si } x \neq y \\ \frac{f_P(x_i) \cdot p_i}{(1-p_i) \cdot f_{BP}(x_i, y_i) + p_i \cdot f_P(x_i)} & \text{si } x = y \end{cases}$$

$$z_i = \begin{cases} 0 & \text{si } x \cdot y = 0 \\ \lambda_{3i} \frac{f_{BP}(x_i - 1, y_i - 1)}{f_{BP}(x_i, y_i)} & \text{altrament} \end{cases}$$

(b) M-step: Trobar:

$$\hat{\beta}_1 = \hat{\beta}(x, W; (1-v))$$
$$\hat{\beta}_2 = \hat{\beta}(y, W; (1-v))$$
$$\hat{\beta}_3 = \hat{\beta}(z, W; (1-v))$$
$$\hat{p} = \frac{\sum_{i=1}^{n} v_i}{n}$$
$$\hat{\theta} = \frac{1}{\sum_{i=1}^{n} v_i} \sum_{i=1}^{n} x_i \cdot v_i$$

on $\hat{\beta}(x, W; v)$ és l'estimador de màxima versemblança d'una regressió de Poisson amb vector de pesos v. Estimar λ_1, λ_2 i λ_3 a partir de $\hat{\beta}_1$, $\hat{\beta}_2$ i $\hat{\beta}_3$.

3. Un cop convergeixi ja es tindran estimats els paràmetres $\hat{\beta}_1$, $\hat{\beta}_2$ i $\hat{\beta}_3$ corresponents als coeficients dels predictors lineals. També es tindrà l'estimació de $\hat{p}$ i $\hat{\theta}$.

De forma similar a la regressió de dues Poisson independents amb la diagonal inflada s'utilitza el mateix valor de p i θ per a totes les observacions. La demostració de l'algoritme es pot trobar en l'apèndix A.3 i el model es pot ajustar a partir de la funció `glmbpoisDI` disponible a l'enllaç `https://github.com/arturxe2/TFG-Artur-Xarles.git`.

3.5 Finalitat dels models proposats

La finalitat dels models proposats és la d'obtenir una predicció del resultat per als diferents partits. Aquesta predicció pot ser expressada a través de la probabilitat que els equips local i visitant marquin un nombre determinat de gols, que ve determinada per les funcions de probabilitat dels diferents models distribucionals (1), (2), (4) i (5). També es pot mostrar la predicció com la probabilitat de que guanyi l'equip local (Resultat = 1), que guanyi l'equip visitant (Resultat = 2) o que hi hagi un empat (Resultat = X). Es poden calcular les probabilitats pels diferents models de la manera següent:

1. **Distribució 2P**:

   Es tenen dues variables Poisson independents $X \sim Pois(\lambda_1)$ i $Y \sim Pois(\lambda_2)$, de manera que la diferència de les quals $X - Y = Z$ segueix una distribució Skellam[12] amb paràmetres λ_1 i λ_2. D'aquesta manera es poden obtenir les probabilitats següents:

$$P(Res. = 1) = P(Sk.(\lambda_1, \lambda_2) > 0) = \sum_{z=1}^{\infty} P(Z = z)$$
$$P(Res. = X) = P(Sk.(\lambda_1, \lambda_2) = 0) = P(Z = 0)$$
$$P(Res. = 2) = P(Sk.(\lambda_1, \lambda_2) < 0) = \sum_{z=-\infty}^{-1} P(Z = z)$$

2. **Distribució BP**:

Es tenen dues variables $X = X_1 + X_3$ i $Y = X_2 + X_3$ tal que $X_i \sim Pois(\lambda_i)$, de manera que la diferència de les quals $X - Y = X_1 - X_2 = Z$ segueix una distribució Skellam amb paràmetres λ_1 i λ_2. D'aquesta manera es poden obtenir les probabilitats següents:

$$P(Res. = 1) = P(Sk.(\lambda_1, \lambda_2) > 0) = \sum_{z=1}^{\infty} P(Z = z)$$
$$P(Res. = X) = P(Sk.(\lambda_1, \lambda_2) = 0) = P(Z = 0)$$
$$P(Res. = 2) = P(Sk.(\lambda_1, \lambda_2) < 0) = \sum_{z=-\infty}^{-1} P(Z = z)$$

3. **Distribució 2PDI**:

   En aquest cas, amb probabilitat p la diferència serà segur 0, mentres que amb probabilitat $(1-p)$ aquesta seguirà una distribució Skellam de la mateixa manera que en la distribució de dues Poisson independents. Així doncs, podem obtenir les probabilitats següents:

$$P(Res. = 1) = (1-p) \cdot P(Sk.(\lambda_1, \lambda_2) > 0) = (1-p) \cdot \sum_{z=1}^{\infty} P(Z = z)$$
$$P(Res. = X) = (1-p) \cdot P(Sk.(\lambda_1, \lambda_2) = 0) + p = (1-p) \cdot P(Z = 0) + p$$
$$P(Res. = 2) = (1-p) \cdot P(Sk.(\lambda_1, \lambda_2) < 0) = (1-p) \cdot \sum_{z=-\infty}^{-1} P(Z = z)$$

4. **Distribució BPDI**:

   En aquest cas, amb probabilitat p la diferència serà segur 0, mentres que amb probabilitat $(1-p)$ aquesta seguirà una distribució Skellam de la mateixa manera que en la distribució d'una Poisson Bivariant. Així doncs, podem obtenir les probabilitats següents:

$$P(Res. = 1) = (1-p) \cdot P(Sk.(\lambda_1, \lambda_2) > 0) = (1-p) \cdot \sum_{z=1}^{\infty} P(Z = z)$$
$$P(Res. = X) = (1-p) \cdot P(Sk.(\lambda_1, \lambda_2) = 0) + p = (1-p) \cdot P(Z = 0) + p$$
$$P(Res. = 2) = (1-p) \cdot P(Sk.(\lambda_1, \lambda_2) < 0) = (1-p) \cdot \sum_{z=-\infty}^{-1} P(Z = z)$$

3.6 Selecció del model

Un cop proposats diversos models (dues Poisson independents, Poisson Bivariant, dues Poisson independents amb la diagonal inflada i Poisson Bivariant amb la diagonal inflada), es procedeix a seleccionar el que té més bons resultats. En primer lloc, s'han seleccionat diferents conjunts de variables per a cada model distribucional (2P, BP, 2PDI, BPDI) segons diferents metodologies proposades en la Secció 3.6.1, i s'ha escollit el que obtenia uns millors resultats per a cada model. Seguidament, s'han comparat els quatre models amb les variables ja seleccionades. Per fer-ho, s'han utilitzat els mètodes proposats en la Secció 3.6.2. Tot aquest procés de selecció s'ha realitzat utilitzant un 80% de les dades de les quals es disposen, de manera que el 20% restant s'utilitzarà per avaluar el model final.

3.6.1 Selecció de variables

Per seleccionar diversos conjunts de variables s'han utilitzat dues metodologies molt similars, el "Forward Stepwise" i el "Backward Stepwise". Aquestes es basen en anar afegint o eliminant variables una a una a partir d'un criteri de decisió. S'ha de tenir en compte que per als models de regressió de dues variables Poisson independents (amb la diagonal inflada o sense) es poden tenir dos conjunts de variables diferents, i per als models de regressió d'una Poisson Bivariant (amb la diagonal inflada o sense) se'n poden tenir tres de diferents. A continuació es mostra el funcionament d'aquests dos procediments:

- **"Forward Stepwise"**

  En aquest cas es pot trobar el conjunt de variables a partir dels següents passos:

  1. Començar amb un model sense variables (només una constant).
  2. Afegir variables una a una segons un criteri de decisió.
  3. Parar quan es compleixi certa condició o quan s'hagin afegit totes les variables al model.

  Els criteris de decisió que es consideren amb les respectives condicions per deixar d'afegir variables son els següents:

  – AIC (criteri d'informació d'Akaike): s'afegeix la variable que proporciona un model amb un menor AIC on:

  $$AIC = 2k - 2\ln(L)$$

  on k és el nombre de paràmetres del model i L la versemblança.
  Es deixa d'afegir variables quan fer-ho augmenta l'AIC, respecte al model amb menor AIC, més que certa quantitat ("treshold"). Finalment es selecciona entre tots els subconjunts de variables el que té menor AIC.

  – BIC (criteri d'informació Bayesià): s'afegeix la variable que proporciona un model amb un menor BIC on:

  $$BIC = k\ln(n) - 2\ln(L)$$

  on k és el nombre de paràmetres del model, n el nombre d'observacions utilitzades per ajustar-lo i L la versemblança.
  S'afegeixen variables fins que fer-ho augmenta el BIC, respecte al model amb menor BIC, més que certa quantitat ("treshold"). Finalment es selecciona el subconjunt de variables que té un menor BIC.

  – p-valor: s'afegeix la variable que té un p-valor menor al afegir-la al model anterior. S'afegeixen variables fins que el p-valor de totes les possibles variables a afegir son superiors a un "treshold".

- **"Backward Stepwise"**

  S'obté el conjunt de variables a partir dels següents passos:

  1. Començar amb un model amb totes les variables possibles
  2. Treure variables una a una segons un criteri de decisió.

3. Parar quan es compleixi certa condició o quan s'hagin eliminat totes les variables del model.

Els criteris de decisió i les condicions per deixar d'eliminar variables que es consideren son els mateixos que en el cas del "Forward Stepwise".

A partir d'aquests dos mètodes s'obté, per a cada model, un total de 6 conjunts de variables, tres obtinguts a partir del "Forward Stepwise" i tres a partir del "Backward Stepwise", utilitzant els diferents criteris de decisió.

3.6.2 Comparació de models

Per comparar diferents models entre sí i seleccionar el que obté uns millors resultats ho farem a través de validació encreuada amb k iteracions. D'aquesta manera, es dividirà la base de dades en k subconjunts i per a cada un d'ells s'ajustarà el model amb la resta de $k-1$ subconjunts i s'avaluarà amb ell. Per avaluar-lo s'utilitzen les mètriques següents:

- *"Accuracy"*: Proporció de resultats (1, X o 2) correctament predits. Amb els diferents models es pot obtenir la probabilitat que succeeixi cada resultat possible, tal com s'indica en la Secció 3.5, i s'escull com a predit el que té una probabilitat major.
- *"Cross-Entropy Loss"* $= -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} log(\hat{p}_{ij})$ on N és el nombre total d'observacions, M el nombre de possibles resultats, y_{ij} té valor 1 si el resultat final de l'observació i és la classe j i 0 altrament, i $\hat{p}_{ij}$ és la probabilitat predita pel model de què el resultat de l'observació i sigui la classe j. Interessen valors baixos de la mètrica, ja que indiquen que la probabilitat predita $\hat{p}_{ij}$ és més alta.
- *"MSE"* $= \frac{1}{N}\sum_{i=1}^{N}(X_i - \hat{X}_i)^2 + (Y_i - \hat{Y}_i)^2$ on X_i i Y_i son el nombre de gols marcats en l'observació i per l'equip local i l'equip visitant respectivament, i $\hat{X}_i$ i $\hat{Y}_i$ son la predicció de gols marcats per l'equip local i l'equip visitant en l'observació i, segons el model.

Per a cada mètrica, es calcula la seva mitjana en les k iteracions per obtenir per a cada model una estimació del valor real de les mètriques.

Finalment, per seleccionar el model amb millors mètriques tenint en compte les tres, per a cada una s'ordenen els models segons quins tenen millors resultats (valors més alts en "Accuracy" i més baixos en "Cross-Entropy Loss" i en "MSE") i n'obtenim el rang. S'escull el model amb un rang mitjà més baix.

A la Taula 1 podem veure un exemple de com s'escolliria el model amb la metodologia comentada a partir de les mètriques obtingudes a través de validació encreuada amb k iteracions.

	Acc.	C-E Loss	MSE	Rang Acc.	Rang C-E Loss	Rang MSE	Rang mitjà
Model 1	0.55	0.95	2	1	2	1	1.33
Model 2	0.45	1	2.15	3	3	2	2.67
Model 3	0.50	0.9	2.20	2	1	3	2

Taula 1: Exemple de selecció de model a partir de diferents mètriques obtingudes a través de validació encreuada amb k iteracions.

S'observa que el model amb un rang mitjà més petit és el primer, de manera que en aquest exemple s'escolliria com el model amb millors resultats.

3.7 Validació del model

Un cop escollit el model que presenta uns millors resultats d'acord amb la metodologia presentada en la Secció 3.6, es procedeix a validar-lo. Per fer-ho s'observen els resultats que obté el model en predir el 20% de dades que no s'utilitzen per escollir el model. D'aquestes es calcula l'"Accuracy", la "Cross-Entropy Loss" i el "MSE".

Seguidament es comparen les prediccions fetes pel nostre model amb les fetes per la casa d'apostes bet365. Aquestes es poden obtenir a través de les quotes que ofereixen per cada resultat, que es troben en la pàgina web "Football-Data.co.uk"[10].

Per últim, s'avalua el model en diferents franges de jornades, per tal d'observar si a l'inici de la temporada, on certes variables son més volàtils i menys fiables, la capacitat de predicció del model és inferior a altres franges de la temporada. També s'avalua en les diferents lligues, per tal de veure si en alguna d'elles els partits son més fàcils de predir que en les altres.

3.8 Elaboració d'un "dashboard" per analitzar prediccions

Havent escollit el model final i havent-lo validat, es poden fer prediccions per a partits nous. Cada una d'aquestes es pot mostrar a través d'un "dashboard", de manera que es representa visualment la informació més important dels partits. La informació a mostrar és en primer lloc els dos equips que disputen el partit, els gols esperats per a cada equip segons el model, els punts esperats, la probabilitat de cada resultat (1, X o 2), la probabilitat de que els dos equips marquin almenys un gol, la probabilitat de que cada equip marqui un nombre determinat de gols, la distribució de la suma de gols dels dos equips i les probabilitats de diferents resultats expressats en gols de l'equip local i l'equip visitant. Aquesta visualització es crea amb `R Markdown`.

3.9 Simulació del tram final de la temporada

També es pot utilitzar el model final per a simular el tram final de la temporada (a partir del 13 d'abril de 2021) per a les diferents lligues, per tal de veure quina probabilitat té cada equip d'acabar en cada posició. Es fa a través de simulacions de Monte Carlo, de manera que per a cada una es segueixen els següents passos:

1. Es considera la classificació a dia 13 d'abril i el model ajustat amb els partits fins aquesta data.
2. S'assumeix que les variables corresponents a estadístiques dels jugadors son les mateixes que en el darrer partit disputat per aquell equip, i que no varien fins a final de temporada.
3. Per a cada jornada restant:
   (a) Es simula el resultat dels diferents partits d'aquella jornada.
   (b) S'actualitza la classificació segons els resultats simulats.
   (c) S'actualitzen les variables relacionades amb la classificació dels equips.
4. S'obté la posició final de cada equip en aquella simulació.

Realitzant n simulacions es pot obtenir una estimació de la probabilitat de cada equip d'acabar en cada posició, a partir de la proporció de simulacions en les quals l'equip ha finalitzat en aquella posició, respecte al total de simulacions.

3.10 Recursos de programació

En la realització d'aquest treball s'ha utilitzat el llenguatge de programació `R`, a través de la interfície `RStudio`. Part de la feina d'aquest treball ha consistit en la creació de diferents funcions, tant per a l'ajust dels diversos models distribucionals proposats com per dur a terme els mètodes de selecció de variables i la comparació de models. Per a aquestes funcions ha set rellevant l'ús de recursos de programació específics per fer més eficient el codi. Concretement, funcions de paquets que incrementen la velocitat.

- `speedglm`: Aquest paquet s'ha utilitzat per incrementar la velocitat de l'ajust de la regressió de Poisson, a través de la funció amb el mateix nom. D'aquesta manera s'ha reduit el temps de computació en l'ajust dels diferents models a pràcticament la meitat.

- `parallel`: Aquest paquet s'ha utilitzat per a reduir el temps de computació a través de paral·lelitzar diverses tasques. Es fa servir per obtenir, en l'ajust dels models, una estimació de la desviació típica dels paràmetres mitjançant "bootstrap", un mètode de remostreig per aproximar la distribució d'un estadístic [1]. També s'utilitza a l'hora de fer validació encreuada i en els mètodes de selecció de variables "Forward Stepwise" i "Backward Stepwise".

A més, també s'ha utilitzat el llenguatge `R Markdown`, que permet la creació de documents en els quals s'hi pot introduir codi `R`. S'ha fet servir per a la realització del "dashboard" en format HTML.

El codi de tot el treball es pot obtenir a través de l'enllaç `https://github.com/arturxe2/TFG-Artur-Xarles.git`.

4 Resultats

En aquest apartat es poden veure els resultats obtinguts en el procés de selecció del model, la forma que té el model final, la validació feta amb ell, les prediccions que es poden fer i la simulació del tram final de temporada per a les diferents lligues.

4.1 Selecció del model

En primer lloc, s'obtenen els diferents conjunts de variables que es tenen en compte en els diferents models distribucionals a través dels mètodes proposats en la Secció 3.6.1 i es comparen entre ells amb la metodologia explicada en la Secció 3.6. Els resultats obtinguts es mostren a l'Apèndix en la Taula 6. Per al model que considera la distribució de dues Poisson independents com a variable resposta, el conjunt de variables seleccionat és el que ve donat pel mètode "Forward Stepwise" amb el p-valor com a criteri de decisió. En el cas del model amb la distribució Poisson Bivariant com a resposta, el conjunt de variables ve donat pel mètode "Forward Stepwise" amb l'AIC com a criteri de decisió. Pel model que té la distribució de dues Poissons independents amb la diagonal inflada, el conjunt de variables és el que s'obté amb el mètode "Backward Stepwise" amb l'AIC o el p-valor com a criteri de decisió, ja que en els dos casos s'obté el mateix conjunt. Per últim, quan es considera la distribució Poisson Bivariant amb la diagonal inflada, el conjunt de variables que es selecciona és el que ve donat pel mètode "Forward Stepwise" amb l'AIC com a criteri de decisió.

Havent seleccionat els conjunts de variables per als diferents models distribucionals, es procedeix a comparar-los entre ells. Els resultats obtinguts es mostren en la Taula 2.

Model	Acc.	C-E Loss	MSE	Rang Acc.	Rang C-E Loss	Rang MSE	Rang mitjà
2P	0.5233	0.9929	2.6079	4.0	3.0	1.0	2.67
BP	0.5277	0.9934	2.6104	1.0	4.0	3.0	2.67
2PDI	0.5247	0.9927	2.6099	2.5	2.0	2.0	2.17
BPDI	0.5247	0.9923	2.6108	2.5	1.0	4.0	2.50

Taula 2: Resultats de la comparació a través de validació encreuada amb 40 iteracions dels 4 models distribucionals proposats amb respectius conjunts de variables seleccionats. Es mostra de color verd el que obté uns millors resultats.

El model que presenta uns millors resultats és el de dues Poisson independents amb la diagonal inflada. Així doncs, aquest és el que es selecciona com a model final. En la Taula 3 es poden observar les variables que conté i els coeficients estimats.

	Coeficient	Sd	p-valor	Significació
Variables per X:				
(Intercept)	-0.2460	0.1887	0.1925	-
Gols local	0.3516	0.1000	0.0004	***
Assistències local	0.2774	0.1259	0.0276	*
Assistències visitant	-0.3302	0.0895	0.0002	***
Gols encaixats local	-0.1573	0.0872	0.0711	.
Gols encaixats visitant	0.3045	0.0764	0.0001	***
Classificació local	0.2841	0.1706	0.0958	.
Punts com a local equip local	-0.2801	0.1327	0.0347	*
Importància partit visitant	-0.2594	0.1060	0.0144	*
Variables per Y:				
(Intercept)	-0.6919	0.2440	0.0046	**
Gols local	-0.1406	0.0686	0.0404	*
Gols visitant	0.4415	0.0605	0.0000	***
"Tackles" i intercepcions visitant	0.0241	0.0094	0.0104	*
Gols encaixats local	0.3832	0.0789	0.0000	***
Gols encaixats visitant	-0.2913	0.0863	0.0007	***
Importància partit local	-0.3056	0.1079	0.0046	**

Taula 3: Resultat de l'ajust del model final corresponent a la regressió de dues Poisson independents. Es mostren les variables seleccionades per predir X i Y juntament amb els valors estimats dels paràmetres, la desviació típica, el p-valor i si la variable és o no significativa.

Tot i que l'objectiu d'aquest model no és explicatiu, es poden destacar alguns aspectes de les estimacions obtingudes. En primer lloc, amb un nivell de significació de 0.10 totes les variables incorporades al model son significatives. A més, el valor dels coeficients s'adequa bastant al que es podria esperar, de manera que variables relacionades amb els gols que marca l'equip local, les assistències que fa o els gols que encaixa l'equip visitant fan que augmenti l'esperança de gols de l'equip local a mesura que elles augmenten. En canvi, altres variables relacionades amb les assistències de l'equip visitant o com d'important és el partit per al visitant, al augmentar, disminueixen l'esperança de gols de l'equip local. De forma similar passa en el cas de l'esperança de gols per a l'equip visitant.

4.2 Validació del model

Un cop seleccionat el model final, es procedeix a validar-lo tal com es proposa en la Secció 3.7. El primer pas és veure quins resultats obté amb dades completament noves i comparar-los amb els resultats obtinguts per bet365. En la Taula 4 es mostren les mètriques corresponents a aquestes noves observacions.

	Accuracy	Cross-Entropy Loss	MSE
Model	0.4866	1.0150	2.7489
bet365	0.4940	0.9989	-

Taula 4: Validació del model amb noves observacions. Es mostren els valors de les diferents mètriques proposades al fer les prediccions per a noves observacions. També es mostren les mètriques per a les prediccions realitzades per bet365.

Els resultats son lleugerament pitjors al valor mitjà obtingut a través de validació

encreuada. També son una mica pitjors que les mètriques obtingudes per bet365, però al ser una diferència relativament petita, es pot considerar que els resultats obtinguts son bastant satisfactoris.

En la Figura 1 es mostra com evolucionen les diferents mètriques en diferents èpoques de la temporada, tant pel nostre model com per les prediccions fetes per bet365.

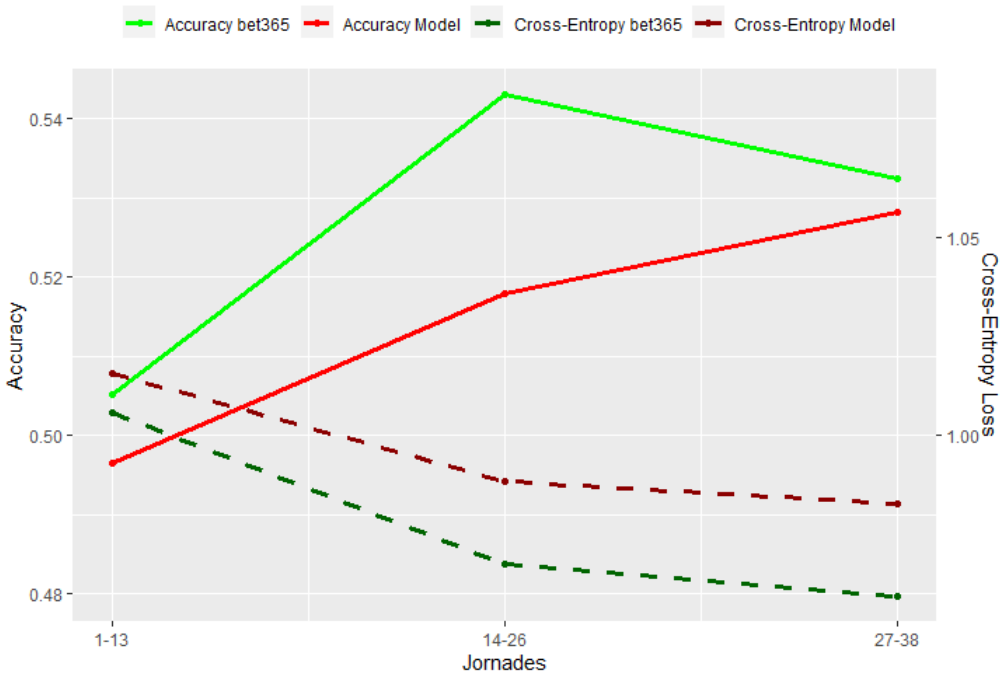

Figura 1: Evolució de l'"Accuracy" i la "Cross-Entropy Loss" per a diferents grups de jornades, tan pel model seleccionat com per les prediccions de bet365.

En el gràfic es pot veure que el model obté millors resultats per a trams de temporada més avançats. Passa el mateix amb les prediccions fetes per bet365, on a l'inici de la temporada els resultats no son tan bons com al final. Aquest fet podria ser degut a que a l'inici de la temporada hi ha més incertesa sobre el nivell d'alguns equips que no pas a final de temporada, on ja s'han pogut observar molts partits de cada equip. També es pot veure que en tot moment el rendiment del model proposat és una mica inferior al de la casa d'apostes.

Per últim, en la Figura 2 es mostra l'"Accuracy" per a les diferents lligues de les que es disposa.

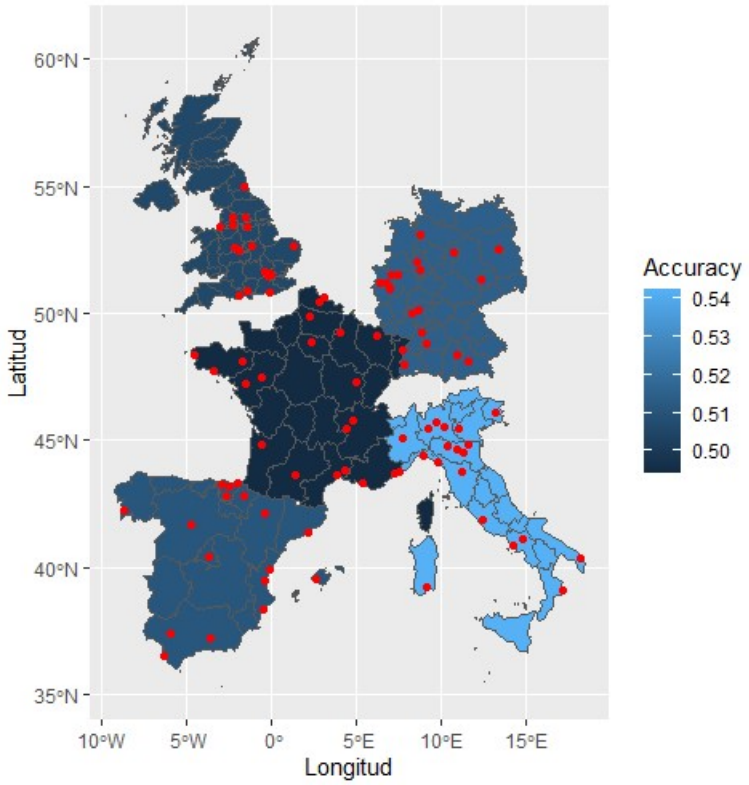

Figura 2: "Accuracy" del model per a les diferents lligues que estudiem. També es mostren amb punts vermells totes les ciutats que tenen almenys un equip de futbol en aquestes lligues.

La lliga en la qual és més fàcil predir correctament els resultats amb el model és la italiana, és a dir, la Serie A. Pel contrari, la que té més incertesa i per tant és més difícil fer-hi bones prediccions és la francesa, la Ligue 1.

Així doncs, podem concloure que el model ajustat dona uns resultats bastant satisfactoris i per tant queda correctament validat.

4.3 "Dashboard" per visualitzar prediccions

Un cop seleccionat i validat el model, es pot utilitzar per predir nous partits. Per aquests, un cop estimats els paràmetres, es pot crear el "dashboard" que es mostra en la Figura 3. Les prediccions estan fetes per al partit disputat el 7 de maig del 2021 entre el Leicester City i el Newcastle.

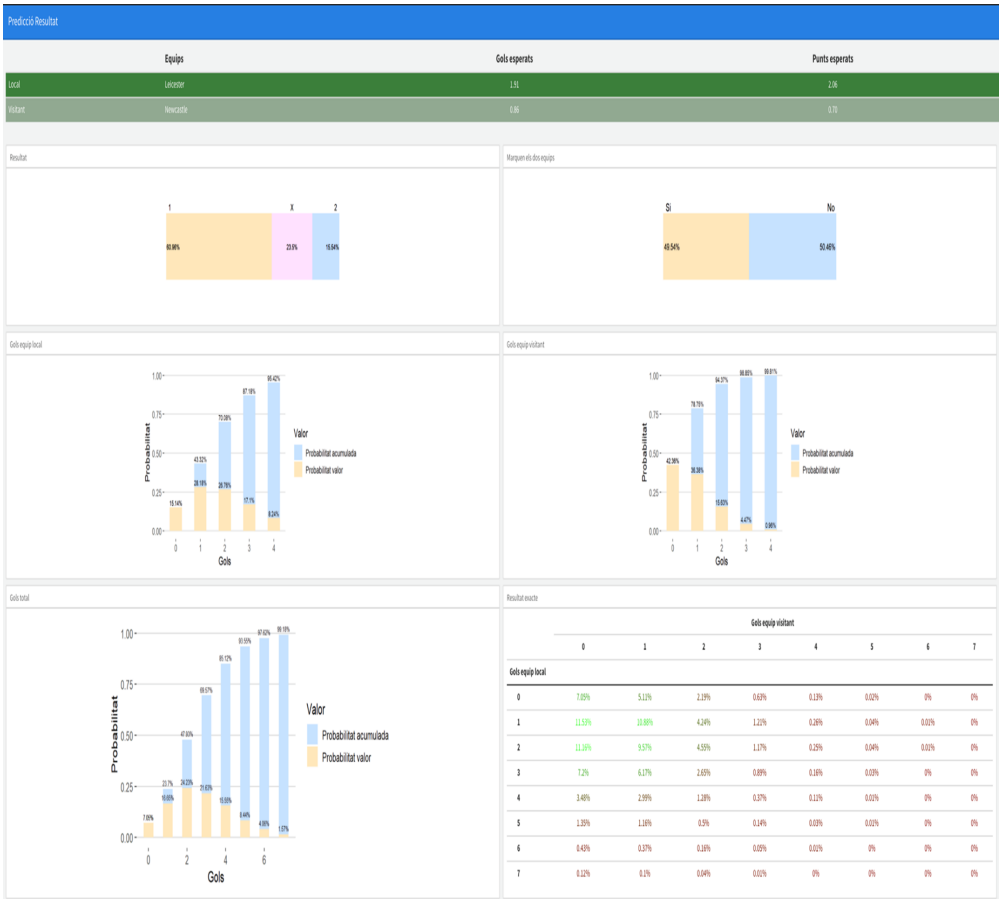

Figura 3: "Dashboard" per mostrar les prediccions fetes pel model sobre un partit en concret. Per a cada un es mostra a la part superior els equips que disputen el partit, l'esperança de gols i l'esperança de punts. També es mostra la probabilitat que succeeixi cada resultat, que els dos equips marquin almenys un gol, que cada equip marqui un nombre determinat de gols, la distribució de la suma de gols dels dos equips i les probabilitats de diferents resultats expresats en gols de l'equip local i l'equip visitant.

El "dashboard" es pot observar amb millor resolució en l'Apèndix B.3, i es pot obtenir, juntament amb el codi utilitzat per a crear-lo, en l'enllaç `https://github.com/arturxe2/TFG-Artur-Xarles.git`.

4.4 Simulació del tram final de la temporada

Finalment, es poden calcular les prediccions per al tram final de la temporada seguint la metodologia presentada en la Secció 3.9. Aquestes es mostren a través d'una taula en la qual hi ha la probabilitat de cada equip d'acabar en cada posició de la classifiació. Per a la lliga espanyola es mostra la simulació realitzada en la Taula 5.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Atlético Madrid	25.4	27.9	41.4	5.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Real Madrid	37.1	36.1	22.9	3.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Barcelona	37.2	33.1	26.2	3.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Sevilla	0.3	2.9	9.5	86.6	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Real Sociedad	0.0	0.0	0.0	0.5	56.5	28.0	13.2	1.5	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Betis	0.0	0.0	0.0	0.1	24.0	34.2	33.2	6.6	1.5	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Villarreal	0.0	0.0	0.0	0.1	18.4	33.5	39.4	6.1	1.8	0.5	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Granada	0.0	0.0	0.0	0.0	0.0	1.7	2.9	20.0	22.4	20.0	14.4	9.1	4.9	3.1	1.4	0.1	0.0	0.0	0.0	0.0
Levante	0.0	0.0	0.0	0.0	0.1	1.1	4.3	23.7	21.6	16.0	14.0	9.3	5.5	3.3	0.7	0.4	0.0	0.0	0.0	0.0
Celta Vigo	0.0	0.0	0.0	0.0	0.3	0.6	2.5	13.0	17.3	16.6	16.2	13.4	10.1	6.0	2.6	1.1	0.3	0.0	0.0	0.0
Athletic Club	0.0	0.0	0.0	0.0	0.0	0.8	3.7	18.6	16.4	17.8	16.9	11.1	8.3	3.7	2.1	0.3	0.3	0.0	0.0	0.0
Cádiz	0.0	0.0	0.0	0.0	0.0	0.0	0.1	3.9	7.1	10.1	10.7	15.0	16.9	16.1	10.3	5.7	2.4	1.6	0.1	0.0
Valencia	0.0	0.0	0.0	0.0	0.0	0.0	0.2	3.2	4.7	9.0	12.4	17.6	18.6	15.9	9.7	6.7	1.2	0.5	0.3	0.0
Osasuna	0.0	0.0	0.0	0.0	0.0	0.1	0.5	3.1	6.0	7.6	10.8	14.2	16.2	19.1	12.7	5.3	3.3	0.9	0.2	0.0
Getafe	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.6	1.3	2.7	6.0	9.6	12.5	22.7	19.1	12.1	8.9	3.5	0.7
Huesca	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.2	0.3	1.1	3.4	7.9	12.6	16.9	21.0	17.7	12.5	6.2
Valladolid	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.3	0.9	1.7	4.0	5.6	13.9	18.1	20.3	19.9	10.3	4.9
Elche	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.4	1.2	2.2	4.9	7.0	14.1	18.3	20.9	20.9	9.8
Alavés	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.1	1.1	2.2	7.6	14.0	17.1	28.2	29.4
Eibar	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.8	2.1	4.6	6.8	12.5	24.0	49.0

Taula 5: Probabilitat de cada equip de la lliga espanyola d'acabar la temporada en cada posició, d'acord al mètode de simulació proposat, expressada en percentatge. Estan calculades havent-se disputat les primeres 30 jornades de lliga. De color verd es mostra la posició final real de cada equip. L'ordre dels equips ve donat per la posició que ocupava cada equip en el moment de fer les prediccions.

Amb una freqüència prou elevada, la posició final dels equips es correspon a una de les que tenia més probabilitat de succeir d'acord a la simulació realitzada. Les simulacions fetes per a la resta de lligues es troben en les Taules 7, 8, 9 i 10 de l'Apèndix B.2.

Es pot veure que amb les prediccions fetes a dia 13 d'abril de 2021, en 3 de les 5 lligues el campió ha estat l'equip que tenia més probabilitat de ser-ho d'acord a les simulacions. Pel que fa a la darrera posició de la lliga, en 4 de les 5 lligues l'últim classificat ha estat el que tenia més probabilitat de ser-ho.

5 Discussió

Tal i com es plantejava a la introducció, s'han estudiat diverses distribucions modelitzar resultats esportius, concretament de partits de futbol. Concretament, s'han considerat dues Poissons correlacionades (Poisson bivariant), dues Poissons independents i les extensions amb diagonal inflada, per recollir millor la incidència dels empats en la predicció dels resultats (gols local, gols visitant) d'una *training sample* de diverses lligues Europees. En el marc dels models lineals generalitzats s'ha recollit la informació que diverses variables pre-partit poden aportar al resultat final.

Dels diferents models proposats, el que presenta uns millors resultats és el que s'obté a través de la distribució de dues Poisson independents amb la diagonal inflada, considerant com a variables explicatives les que se seleccionen a través del mètode "Backward Stepwise" amb l'AIC o el p-valor com a criteris de decisió. A l'hora d'ajustar aquest model amb les dades d'entrenament s'obtenen uns coeficients que semblen coherents, almenys el signe, amb la idea prèvia que es té de l'hipotètic sentit de la influència de les variables sobre el resultat. A més, aquest model és capaç de predir correctament un 48.66% dels resultats en els partits que s'han utilitzat per a la validació. Aquest rendiment és una mica inferior a l'obtingut per la casa d'apostes bet365, que és del 49.40%. Tot i això, al ser una diferència relativament petita podem considerar que el model final satisfà les expectatives del treball. D'aquesta manera, s'ha complert l'objectiu principal del treball d'obtenir un model per predir el resultat de partits de futbol de forma satisfactòria.

També s'ha vist que els partits corresponents a l'inici de la temporada son més difícils de predir, tant pel model proposat com pel de la casa d'apostes. Comparant les diferents lligues, els partits corresponents a la lliga italiana son els més fàcils de predir amb el nostre model, mentre que els de la lliga francesa son els més complicats.

Amb el model proposat es poden fer prediccions de nous partits de futbol. Aquestes es poden mostrar a cada partit amb el "dashboard" creat. També s'han utilitzat aquestes prediccions per simular els trams finals de temporada de les cinc lligues de les quals es disposa de dades. La posició final real dels equips és de les que tenen més possibilitats de succeir segons les probabilitats del model en la majoria de casos.

Tot i els bons resultats de l'estudi, aquest treball té certs aspectes que es podrien millorar. En primer lloc, es podria intentar perfeccionar el model obtingut. Per fer-ho es podrien tenir en compte altres variables predictores interessants a priori, com ara el nombre de dies de descans que han tingut els jugadors entre el partit anterior i l'actual o el nombre de xuts per partit per cada equip, que podrien ajudar a predir millor el resultat final. A més, es podrien considerar interaccions entre les diferents variables, relacions no necessàriament lineals o considerar models separats per a cada lliga. També es podrien utilitzar una sèrie de variables regressores per estimar els paràmetres corresponents a inflar la diagonal, per mirar de millorar l'ajust.

Així doncs, tot i haver-hi aspectes a millorar en el treball, es pot dir que s'han complert els objectius que es plantejaven a l'inici d'aquest, obtenint un model amb prediccions suficientment bones considerant la gran incertesa inherent als resultats d'un partit de futbol.

Per últim, cal destacar que els models distribucionals i la metodologia proposada també es poden aplicar a altres conjunts de dades, en funció dels diferents interessos

i necessitats, sempre que siguin dades bivariants corresponents al recompte d'esdeveniments. En aquest cas, s'ha optat per treballar a partir dels resultats de les lligues de fútbol per l'interès de l'autor en el món dels esports i per la quantitat de dades accessibles en aquest àmbit.

Referències

[1] Alabert A. & Borràs R. (2020). *Simulació i Remostreig* [apunts de classe].

[2] Ato, M. & Losilla, J.M. et al. (2005). *Modelo Lineal Generalizado,* Espanya: Grupo ModEst & Edicions a Petició, SL.

[3] Boice, J. (2020). *FiveThirtyEight, How Our Club Soccer Predictions Work.* Disponible a: `https://fivethirtyeight.com/methodology/how-our-club-soccer-predictions-work/`.

[4] Enea M. (2021). *speedglm: Fitting Linear and Generalized Linear Models to Large Data Sets. R package version 0.3-3.* Disponible a: `https://CRAN.R-project.org/package=speedglm`.

[5] J. Faraway, J. (2015). *Linear Models with R*, Nova York, Estats Units: CRC Press.

[6] Karlis, D. & Ntzoufras, I. (2003). *Bayesian and Non-Bayesian Analysis of Soccer Data using Bivariate Poisson Regression Models.* Disponible a: https://www.betgps.com/betting-library/Karlis-Ntzoufras-Presentation-Bayesian-and-Non-Bayesian-Analysis-of-Soccer-Data.pdf.

[7] Karlis, D. & Ntzoufras, I. (2005). *Bivariate Poisson and Diagonal Inflated Bivariate Poisson Regression Models in R.* DOI: 10.18637/jss.v014.i10

[8] R Core Team (2020). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. Disponible a: `https://www.R-project.org/`.

[9] (2020). *Cross-Entropy Loss Function.* Disponible a: `https://towardsdatascience.com/cross-entropy-loss-function-f38c4ec8643e`.

[10] (2021). *Football Data.* Disponible a: `http://www.football-data.co.uk/data.php`.

[11] (2021). *Historia del fútbol.* Disponible a: `https://es.wikipedia.org/wiki/Historia_del_f%C3%BAtbol#Escuelas_brit%C3%A1nicas`.

[12] (2021). *Skellam distribution.* Disponible a: `https://en.wikipedia.org/wiki/Skellam_distribution`.

[13] (2021). *Sports Reference.* Disponible a: `https://www.sports-reference.com/`.

A Demostracions teòriques

A.1 Demostració EM BP

Es poden trobar els dos passos de l'algoritme EM per a la regressió Poisson Bivariant de la manera següent:

1. E-step:
   Estimar la variable latent Z a partir de les variables observades X i Y i del conjunt de paràmetres $\theta = (\lambda_1, \lambda_2, \lambda_3)$ en el pas anterior (θ^*), a partir de l'esperança de la Llei $Z|X, Y, \theta^*$.

$$\begin{aligned} f(Z|X,Y,\theta^*) &= \frac{f(X,Y|Z,\theta^*)f(Z|\theta^*)}{f_{BP}(X,Y|\theta^*)} \\ &= \frac{f(X_1 = X - Z|\theta^*)f(X_2 = Y - Z|\theta^*)f(Z|\theta^*)}{f_{BP}(X,Y|\theta^*)} \end{aligned}$$

   Nota: Condicionar respecte θ^* simplement ens indica que es tenen unes estimacions per als paràmetres λ_1, λ_2 i λ_3 en la iteració anterior.

   L'esperança de la llei per calcular el valor esperat de la variable latent Z és la següent:

$$\begin{aligned} E[Z|X,Y] &= \sum_{i=0}^{min(x,y)} i\frac{f_1(x-i)f_2(y-i)f_3(i)}{f_{BP}(x,y)} \\ &= \frac{1}{f_{BP}(x,y)} \sum_{i=1}^{min(x,y)} i\frac{e^{-\lambda_1}\lambda_1^{(x-i)}}{(x-i)!}\frac{e^{-\lambda_2}\lambda_2^{(y-i)}}{(y-i)!}\frac{e^{-\lambda_3}\lambda_3^{(i)}}{i!} \\ &= \frac{\lambda_3}{f_{BP}(x,y)} e^{-(\lambda_1+\lambda_2+\lambda_3)}\lambda_1^{(x-1)}\lambda_2^{(y-1)} \sum_{i=1}^{min(x,y)} \frac{\lambda_1^{-(i-1)}\lambda_2^{-(i-1)}\lambda_3^{i-1}}{(x-i)!(y-i)!(i-1)!} \\ &= ^{6} \frac{\lambda_3}{f_{BP}(x,y)} e^{-(\lambda_1+\lambda_2+\lambda_3)}\lambda_1^{(x-1)}\lambda_2^{(y-1)} \\ &\cdot \sum_{k=0}^{min(x-1,y-1)} \frac{(\frac{\lambda_3}{\lambda_1\lambda_2})^k}{(x-k-1)!(y-k-1)!k!} \end{aligned}$$

   Com que $e^{-(\lambda_1+\lambda_2+\lambda_3)}\lambda_1^{(x-1)}\lambda_2^{(y-1)} \sum_{k=0}^{min(x-1,y-1)} \frac{(\frac{\lambda_3}{\lambda_1\lambda_2})^k}{(x-k-1)!(y-k-1)!k!} = f_{BP}(x-1, y-1)$, es té:

$$E[Z|X,Y] = \lambda_3 \frac{f_{BP}(x-1,y-1)}{f_{BP}(x,y)}$$

   Es pot apreciar que en el cas que x o y sigui 0, $f_{BP}(x-1, y-1)$ serà 0 i per tant es pot estimar el valor de la variable latent Z a partir de l'expressió següent:

$$z = E[Z|X,Y] = \begin{cases} 0 & \text{si } x \cdot y = 0 \\ \lambda_3 \frac{f_{BP}(x-1,y-1)}{f_{BP}(x,y)} & \text{altrament} \end{cases}$$

[6] S'utilitza $k = i - 1$

2. M-step
   Havent estimat el valor de Z i sabent els valors observats de X i Y es pot trobar fàcilment els valors de X_1, X_2 i X_3. En tractar-se de tres Poisson independents d'acord amb la distribució Poisson Bivariant, el pas de maximització consisteix simplement a fer una regressió de Poisson per a les tres variables. Així doncs, s'ha de calcular:

$$\hat{\beta}_1 = \hat{\beta}(x - z, W)$$
$$\hat{\beta}_2 = \hat{\beta}(y - z, W)$$
$$\hat{\beta}_3 = \hat{\beta}(z, W)$$

   on $\hat{\beta}(x, W)$ son els paràmetres estimats per màxima versemblança d'una regressió de Poisson amb resposta x i matriu de dades W.

   A partir de $\hat{\beta}_1$, $\hat{\beta}_2$ i $\hat{\beta}_3$ es pot estimar λ_1, λ_2 i λ_3, valors necessaris en l'E-step.

A.2 Demostració EM 2PDI

Es poden trobar els dos passos de l'algoritme EM per a la regressió de dues Poisson independents amb la diagonal inflada de la manera següent:

1. E-step:
   Estimar la variable latent V a partir de les variables observades X i Y i del conjunt de paràmetres $\theta = (\lambda_1, \lambda_2, p, \theta)$ en el pas anterior (θ^*), a partir de l'esperança de la Llei $V|X, Y, \theta^*$.

$$\begin{aligned} f(V|X, Y, \theta^*) &= \frac{f(X, Y|V, \theta^*) f(V|\theta^*)}{f(X, Y|\theta^*)} \\ &= \frac{f_{2PDI}(X, Y|V, \theta^*) f_{Bern.}(V|p^*)}{f_{2PDI}(X, Y|\theta^*)} \end{aligned}$$

   Condicionar respecte θ^* simplement ens indica que es tenen unes estimacions per als paràmetres λ_1, λ_2, p i θ en la iteració anterior. L'esperança de la llei per estimar el valor de V és la següent:

$$\begin{aligned} v &= E[V|X, Y, \theta^*] = 1 \cdot \Big(\frac{f_{2PDI}(x, y|V = 1, \theta^*) \cdot f_{Bern.}(V = 1|\theta^*)}{f_{2PDI}(x, y|\theta^*)} \Big) \\ &+ 0 \cdot \Big(\frac{f_{2PDI}(x, y|V = 0, \theta^*) \cdot f_{Bern.}(V = 0|\theta^*)}{f_{2PDI}(x, y|\theta^*)} \Big) \\ &= \begin{cases} 0 & \text{si } x \neq y \\ \frac{f_P(x|\theta^*) p^*}{f_{2PDI}(x, y|\theta^*)} & \text{si } x = y \end{cases} \end{aligned}$$

2. M-step
   Al ser $E[V|X, Y, \theta^*]$ una estimació de p per a cada observació, el pas de maximització consisteix en ajustar regressions de Poisson amb pesos donats per v. Així doncs, s'ha de calcular:

$$\hat{\beta}_1 = \hat{\beta}(x, W; (1-v))$$
$$\hat{\beta}_2 = \hat{\beta}(y, W; (1-v))$$

on $\hat{\beta}(x, W; v)$ son els paràmetres estimats per màxima versemblança d'una regressió de Poisson amb resposta x, matriu de dades W i vector de pesos v.

Per als paràmetres p i θ, si no considerem covariables per ajustar-los es poden maximitzar a partir de la mitjana de v i la mitjana ponderada de x:

$$\hat{p} = \frac{\sum_{i=1}^{n} v_i}{n}$$
$$\hat{\theta} = \frac{1}{\sum_{i=1}^{n} v_i} \sum_{i=1}^{n} x_i \cdot v_i$$

A partir de $\hat{\beta}_1$ i $\hat{\beta}_2$ es pot estimar λ_1 i λ_2, valors necessaris en l'E-step.

A.3 Demostració EM BPDI

Es poden trobar els dos passos de l'algoritme EM per a la regressió d'una Poisson Bivariant amb la diagonal inflada de la manera següent:

1. E-step:
   Estimar les variable latents Z i V a partir de les variables observades X i Y i del conjunt de paràmetres $\theta = (\lambda_1, \lambda_2, \lambda_3, p, \theta)$ en el pas anterior (θ^*), a partir de l'esperança de les Lleis $Z|X, Y, \theta^*$ i $V|X, Y, \theta^*$. Per a la variable latent Z, es pot realitzar tal com s'ha fet en l'algoritme EM per a una distribució Poisson Bivariant (A.1), i per a la variable latent V tal com s'ha fet en l'algoritme EM per a una distribució de dues Poisson independents amb la diagonal inflada (A.2). S'obtenen les estimacions:

$$v = \begin{cases} 0 & \text{si } x \neq y \\ \frac{f_P(x) \cdot p}{(1-p) \cdot f_{BP}(x,y) + p \cdot f_P(x)} & \text{si } x = y \end{cases}$$
$$z = \begin{cases} 0 & \text{si } x \cdot y = 0 \\ \lambda_3 \frac{f_{BP}(x-1, y-1)}{f_{BP}(x,y)} & \text{altrament} \end{cases}$$

2. M-step
   Havent estimat el valor de Z i sabent els valors observats de X i Y es pot trobar fàcilment els valors de X_1, X_2 i X_3. En tractar-se de tres Poisson independents d'acord amb la distribució Poisson Bivariant, el pas de maximització consisteix simplement a fer una regressió de Poisson per les tres variables amb pesos establerts per v. Així doncs, s'ha de calcular:

$$\hat{\beta}_1 = \hat{\beta}(x - z, W; (1-v))$$
$$\hat{\beta}_2 = \hat{\beta}(y - z, W; (1-v))$$
$$\hat{\beta}_3 = \hat{\beta}(z, W; (1-v))$$

on $\hat{\beta}(x, W; v)$ son els paràmetres estimats per màxima versemblança d'una regressió de Poisson amb resposta x, matriu de dades W i vector de pesos v.

Per als paràmetres p i θ, si no considerem covariables per ajustar-los es poden maximitzar a partir de la mitjana de v i la mitjana ponderada de x:

$$\hat{p} = \frac{\sum_{i=1}^{n} v_i}{n}$$
$$\hat{\theta} = \frac{1}{\sum_{i=1}^{n} v_i} \sum_{i=1}^{n} x_i \cdot v_i$$

A partir de $\hat{\beta}_1$, $\hat{\beta}_2$ i $\hat{\beta}_3$ es pot estimar λ_1, λ_2 i λ_3, valors necessaris en l'E-step.

B Material complementari

B.1 Taula selecció variables

		Acc.	C-E Loss	MSE	Rang Acc.	Rang C-E Loss	Rang MSE	Rang mitjà
2P	Variables 1	0.5214	0.9926	2.6136	3.0	2.0	3.0	2.67
	Variables 2	0.5185	0.9959	2.6195	5.5	5.5	5.5	5.50
	Variables 3	0.5244	0.9931	2.6110	1.0	4.0	1.0	2.00
	Variables 4	0.5214	0.9926	2.6136	3.0	2.0	3.0	2.67
	Variables 5	0.5185	0.9959	2.6195	5.5	5.5	5.5	5.50
	Variables 6	0.5214	0.9926	2.6136	3.0	2.0	3.0	2.67
BP	Variables 1	0.5274	0.9934	2.6161	1.0	3.0	1.0	1.67
	Variables 2	0.5185	0.9958	2.6199	5.5	5.5	2.5	4.50
	Variables 3	0.5200	0.9935	122.1481	4.0	4.0	4.0	4.00
	Variables 4	0.5221	0.9934	$4.27 \cdot 10^{83}$	3.0	2.0	6.0	3.67
	Variables 5	0.5185	0.9958	2.6199	5.5	5.5	2.5	4.50
	Variables 6	0.5230	0.9925	$4.92 \cdot 10^{27}$	2.0	1.0	5.0	2.67
2PDI	Variables 1	0.5214	0.9930	2.6132	2.0	3.0	3.0	2.67
	Variables 2	0.5185	0.9964	2.6194	5.5	5.5	5.5	5.5
	Variables 3	0.5214	0.9935	2.6136	4.0	4.0	4.0	4.0
	Variables 4	0.5214	0.9930	2.6132	2.0	1.5	1.5	1.67
	Variables 5	0.5185	0.9964	2.6194	5.5	5.5	5.5	5.5
	Variables 6	0.5214	0.9930	2.6132	2.0	1.5	1.5	1.67
BPDI	Variables 1	0.5222	0.9926	2.6139	2.0	2.0	1.0	1.67
	Variables 2	0.5185	0.9959	2.6198	4.5	5.5	2.5	4.17
	Variables 3	0.5222	0.9928	282.5469	3.0	3.0	4.0	3.33
	Variables 4	0.5184	0.9934	$+\infty$	6.0	4.0	6.0	5.33
	Variables 5	0.5185	0.9959	2.6198	4.5	5.5	2.5	4.17
	Variables 6	0.5230	0.9925	$1.29 \cdot 10^{28}$	1.0	1.0	5.0	2.33

Taula 6: Resultats de la comparació a través validació encreuada amb 20 iteracions dels 4 models distribucionals proposats, utilitzant diferents conjunts de variables. Per cada model, es mostra de color verd el conjunt de variables que obté un millor ajust.

B.2 Taules simulació final temporada

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Bayern Munich	94.3	5.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
RB Leipzig	5.7	87.0	6.2	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Wolfsburg	0.0	2.0	29.6	46.1	21.4	0.8	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Eint. Frankfurt	0.0	5.2	59.8	29.3	5.4	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Dortmund	0.0	0.1	4.2	22.2	60.2	10.3	2.3	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Leverkusen	0.0	0.0	0.2	1.1	8.2	46.4	26.9	13.2	3.2	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Union Berlin	0.0	0.0	0.0	0.0	1.2	11.5	19.5	28.8	24.4	12.9	1.5	0.2	0.0	0.0	0.0	0.0	0.0	0.0
Monchengladbach	0.0	0.0	0.0	0.2	3.0	24.1	33.1	24.3	11.3	3.7	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Stuttgart	0.0	0.0	0.0	0.0	0.5	6.4	15.6	23.8	31.7	19.7	2.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Freiburg	0.0	0.0	0.0	0.0	0.1	0.2	2.4	8.8	24.9	45.5	14.8	2.5	0.7	0.1	0.0	0.0	0.0	0.0
Augsburg	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.7	4.0	22.9	31.8	18.6	12.9	5.4	2.3	0.4	0.0
Hoffenheim	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	2.4	11.6	35.7	27.2	12.5	7.1	2.3	0.7	0.1	0.0
Werder Bremen	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.3	0.4	9.6	14.9	21.2	24.5	18.7	9.0	1.3	0.0
Mainz 05	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	3.6	7.3	16.1	19.5	26.4	19.8	7.2	0.0
Hertha BSC	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	1.2	7.3	11.8	22.7	21.4	17.6	11.7	6.2	0.0
Arminia	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	1.8	3.4	5.8	10.0	18.1	31.0	29.8	0.0
Colonia	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.8	2.4	4.5	11.5	25.5	54.7	0.3
Schalke 04	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	99.7

Taula 7: Probabilitat de cada equip de la lliga alemana d'acabar la temporada en cada posició, d'acord al mètode de simulació proposat, expressada en percentatge. Estan calculades havent-se disputat les primeres 28 jornades de lliga. De color verd es mostra la posició final real de cada equip. L'ordre dels equips ve donat per la posició que ocupava cada equip en el moment de fer les prediccions.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Inter	97.8	2.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Milan	0.0	17.3	28.4	27.6	18.7	7.5	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Juventus	0.1	17.3	33.6	27.0	14.7	6.9	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Atalanta	2.1	59.8	23.8	9.5	4.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Napoli	0.0	2.0	7.9	18.6	30.1	34.0	7.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Lazio	0.0	1.4	6.3	16.1	28.0	36.6	11.5	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Roma	0.0	0.0	0.0	1.2	4.5	14.2	79.0	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Sassuolo	0.0	0.0	0.0	0.0	0.0	0.0	0.4	36.2	43.0	14.5	4.6	1.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Hellas Verona	0.0	0.0	0.0	0.0	0.0	0.0	0.8	51.8	31.3	10.6	3.8	1.0	0.6	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Sampdoria	0.0	0.0	0.0	0.0	0.0	0.0	0.0	7.3	14.7	31.6	20.5	13.9	5.9	3.2	2.2	0.5	0.2	0.0	0.0	0.0
Bologna	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.2	5.7	19.5	25.0	17.6	11.9	6.9	6.3	3.9	1.0	0.0	0.0	0.0
Udinese	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	3.0	13.0	19.5	19.7	17.7	11.4	7.5	4.7	2.5	0.4	0.0	0.0
Genoa	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.3	2.4	6.3	10.3	14.4	16.7	16.8	17.5	11.7	3.3	0.2	0.0
Spezia	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	1.0	3.6	9.0	13.7	17.0	17.7	16.2	11.5	8.4	1.7	0.0	0.0
Fiorentina	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.8	2.9	6.4	12.1	16.2	17.3	18.1	13.7	9.6	2.3	0.2	0.0
Benevento	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	1.3	4.6	6.1	12.7	15.3	22.4	26.4	9.7	1.0	0.0
Torino	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	1.4	3.6	6.0	9.2	12.5	15.6	18.2	22.9	9.2	1.2	0.0
Cagliari	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.6	1.4	1.7	6.0	13.2	50.8	25.5	0.7
Parma	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.3	1.6	4.0	22.2	64.9	6.8
Crotone	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.4	7.0	92.5

Taula 8: Probabilitat de cada equip de la lliga italiana d'acabar la temporada en cada posició, d'acord al mètode de simulació proposat, expressada en percentatge. Estan calculades havent-se disputat les primeres 30 jornades de lliga. De color verd es mostra la posició final real de cada equip. L'ordre dels equips ve donat per la posició que ocupava cada equip en el moment de fer les prediccions.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Lille	36.5	33.1	21.0	9.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Paris S-G	48.4	30.4	15.4	5.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Monaco	4.7	13.4	28.3	53.4	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Lyon	10.4	23.1	35.3	31.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Lens	0.0	0.0	0.0	0.2	44.7	28.7	19.8	5.9	0.6	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Marseille	0.0	0.0	0.0	0.0	35.2	32.9	21.3	7.5	2.6	0.4	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Rennes	0.0	0.0	0.0	0.0	15.5	26.4	31.8	18.8	6.0	1.1	0.3	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Montpellier	0.0	0.0	0.0	0.0	4.2	9.7	20.5	40.0	15.1	7.0	2.0	0.9	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Nice	0.0	0.0	0.0	0.0	0.1	0.7	3.3	11.5	24.6	22.9	15.4	9.0	6.3	4.2	1.7	0.3	0.0	0.0	0.0	0.0
Metz	0.0	0.0	0.0	0.0	0.1	1.2	1.8	8.4	23.9	22.4	17.5	11.5	6.7	4.0	2.1	0.3	0.1	0.0	0.0	0.0
Angers	0.0	0.0	0.0	0.0	0.0	0.3	1.0	2.4	11.2	15.3	18.5	19.1	14.9	8.9	4.6	3.3	0.4	0.1	0.0	0.0
Reims	0.0	0.0	0.0	0.0	0.0	0.1	0.5	3.5	8.6	18.1	18.8	16.7	15.7	9.1	6.0	2.1	0.8	0.0	0.0	0.0
Saint-Etienne	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.2	2.5	5.5	10.8	14.2	14.9	18.2	17.1	12.9	2.6	0.1	0.0	0.0
Strasbourg	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	2.5	2.5	7.0	10.6	15.2	18.5	20.1	13.5	7.0	2.2	0.6	0.0
Bordeaux	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	1.4	2.5	5.3	8.8	13.4	16.1	20.6	22.9	6.7	2.1	0.1	0.0
Brest	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	1.0	2.2	4.2	8.6	11.0	16.8	19.5	24.8	8.2	2.5	0.8	0.0
Lorient	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.2	1.1	3.0	5.2	11.3	40.4	25.9	12.7	0.0
Nimes	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.1	0.4	1.7	4.7	19.6	37.6	35.7	0.0
Nantes	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.8	1.4	3.9	14.2	29.5	50.1	0.0
Dijon	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0

Taula 9: Probabilitat de cada equip de la lliga francesa d'acabar la temporada en cada posició, d'acord al mètode de simulació proposat, expressada en percentatge. Estan calculades havent-se disputat les primeres 32 jornades de lliga. De color verd es mostra la posició final real de cada equip. L'ordre dels equips ve donat per la posició que ocupava cada equip en el moment de fer les prediccions.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Manchester City	98.4	1.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Manchester Utd	1.6	90.3	6.0	1.7	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Leicester City	0.0	4.0	36.1	24.7	18.3	10.8	5.3	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
West Ham	0.0	0.5	12.8	19.6	23.3	23.1	15.3	4.1	1.1	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Chelsea	0.0	2.0	26.8	22.3	19.7	17.3	7.7	3.1	0.9	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Liverpool	0.0	1.5	16.0	26.2	24.0	18.8	8.9	3.9	0.5	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Tottenham	0.0	0.1	2.0	4.8	10.7	17.4	29.7	21.7	8.1	4.1	1.2	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Everton	0.0	0.0	0.3	0.6	2.5	7.2	17.9	28.9	18.9	13.8	7.9	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Arsenal	0.0	0.0	0.0	0.1	0.9	2.7	8.5	18.2	31.0	21.4	13.3	3.6	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Leeds United	0.0	0.0	0.0	0.0	0.1	1.4	3.3	9.7	21.1	29.4	28.0	5.6	1.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Aston Villa	0.0	0.0	0.0	0.0	0.2	1.2	3.4	8.8	16.6	25.9	31.0	11.2	1.5	0.1	0.1	0.0	0.0	0.0	0.0	0.0
Wolves	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	1.6	3.6	11.3	43.8	20.4	10.2	5.6	2.3	0.5	0.0	0.0	0.0
Crystal Palace	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.7	2.7	11.0	26.6	25.7	19.3	9.6	3.9	0.3	0.0	0.0
Southampton	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	3.5	14.3	26.8	22.4	15.7	12.1	4.4	0.1	0.0	0.0
Brighton	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.7	5.2	13.2	19.1	23.2	23.5	13.3	1.6	0.1	0.0
Burnley	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	2.2	6.4	14.4	22.2	25.5	24.4	4.7	0.1	0.0
Newcastle Utd	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.9	3.4	7.8	13.2	23.8	42.0	8.1	0.7	0.0
Fulham	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.7	2.7	8.9	60.3	26.5	0.6
West Brom	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	2.6	24.7	69.1	3.1
Sheffield Utd	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	3.5	96.3

Taula 10: Probabilitat de cada equip de la lliga anglesa d'acabar la temporada en cada posició, d'acord al mètode de simulació proposat, expressada en percentatge. Estan calculades havent-se disputat les primeres 31 jornades de lliga. De color verd es mostra la posició final real de cada equip. L'ordre dels equips ve donat per la posició que ocupava cada equip en el moment de fer les prediccions.

B.3 Figures "dashboard" ampliades

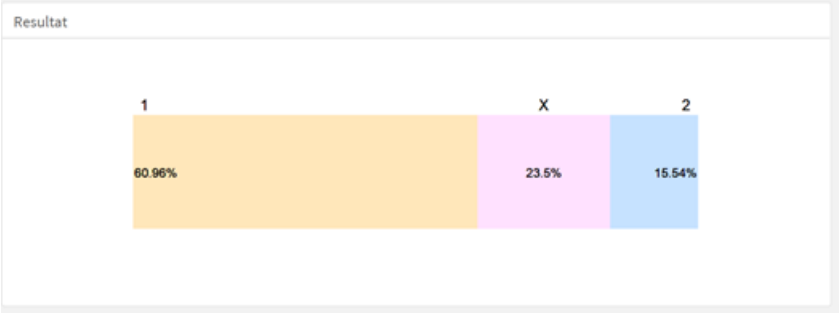

Figura 4: Figura superior esquerra del "Dashboard". Es mostra la probabilitat de succeir de cada resultat (1, X, 2) d'acord al model final.

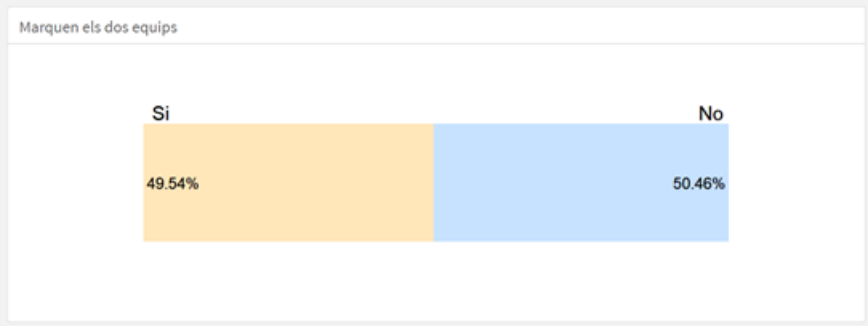

Figura 5: Figura superior dreta del "Dashboard". Es mostra la probabilitat de que els dos equips marquin almenys un gol, juntament amb la seva complementària.

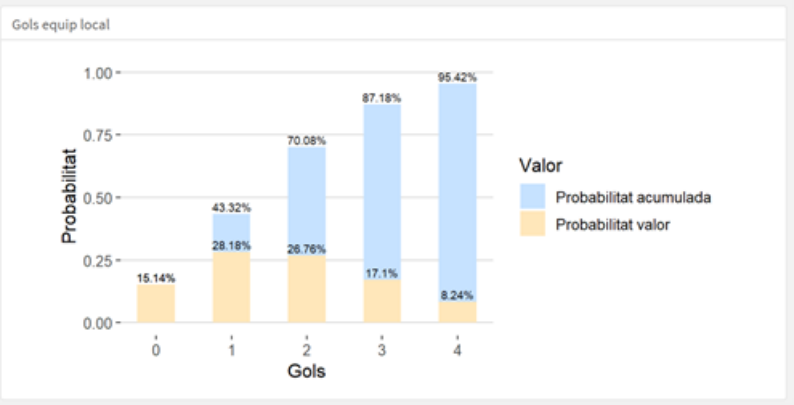

Figura 6: Figura central esquerra del "Dashboard". Es mostra la distribució de probabilitat de gols de l'equip local i la distribució acumulada.

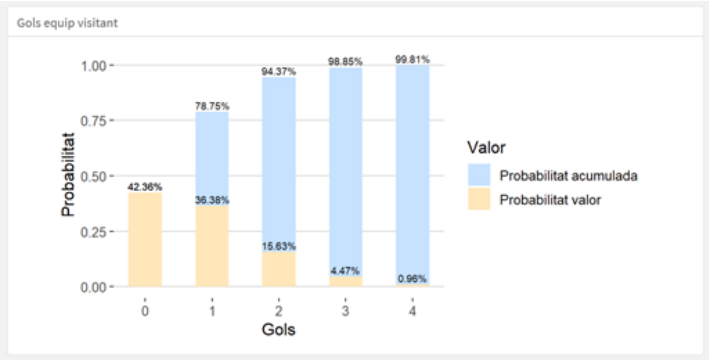

Figura 7: Figura central dreta del "Dashboard". Es mostra la distribució de probabilitat de gols de l'equip visitant i la distribució acumulada.

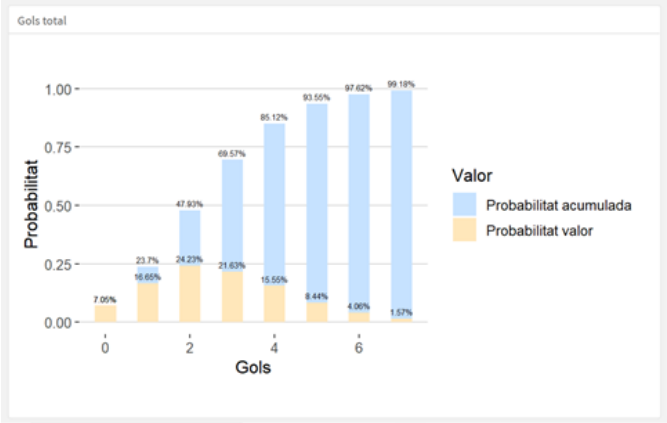

Figura 8: Figura inferior esquerra del "Dashboard". Es mostra la distribució de probabilitat de gols totals i la distribució acumulada.

Resultat exacte

	Gols equip visitant							
	0	1	2	3	4	5	6	7
Gols equip local								
0	7.05%	5.11%	2.19%	0.63%	0.13%	0.02%	0%	0%
1	11.53%	10.88%	4.24%	1.21%	0.26%	0.04%	0.01%	0%
2	11.16%	9.57%	4.55%	1.17%	0.25%	0.04%	0.01%	0%
3	7.2%	6.17%	2.65%	0.89%	0.16%	0.03%	0%	0%
4	3.48%	2.99%	1.28%	0.37%	0.11%	0.01%	0%	0%
5	1.35%	1.16%	0.5%	0.14%	0.03%	0.01%	0%	0%
6	0.43%	0.37%	0.16%	0.05%	0.01%	0%	0%	0%
7	0.12%	0.1%	0.04%	0.01%	0%	0%	0%	0%

Figura 9: Figura inferior dreta del "Dashboard". Es mostra la probabilitat de succeir de diferents resultats, expresats en gols de l'equip local i visitant. De color verd es mostren els resultats més probables.