Predicción de eventos deportivos

Ayose Iturralde Valencia

Trabajo de fin de grado en Matemáticas
Universidad de Zaragoza

Director del trabajo: Javier López Lorente
11 de julio de 2017

Abstract

This project consists of three chapters and we are going to summarize here the most important ideas of each one. Even though the title is forecasting in sport events, in this paper we are going to focus on football matches in "La Liga".

There are lots of models which predict football matches. Firstly we study Maher's models which affirm that we can predict the number of goals that one team is going to score in a match using:

$$X_{i,j} \sim Poisson(\alpha_i \beta_j \gamma)$$

$$Y_{i,j} \sim Poisson(\alpha_j \beta_i)$$

where $X_{i,j}$ is the number of goals that the local team scores when the team i plays against team j $Y_{i,j}$ is the the number of goals that the away team scores when the team i plays against team j, and α_i,β_i and γ could be interpreted as the attacking coefficient, defending coefficient and local factor respectively. We assume that $X_{i,j}$ anf $Y_{i,j}$ are independent. We compute this variables using maximum likelihood method and we then study the results. Finally we confirm that this method fits really well the numbers of goals. The main purpose of this project is to forecast the final result of a football match in terms of wining, drawing or losing. Thus we compute the probability of each kind of result.Then we tried to make an improvement giving weights to each season in our data base. After a study we could not being able to confirm that these model improve our forecasts. After that we studied some improvement to our model proposed by Dixon and Coles.

After study Maher's model we build a logistic model. This kind of model just give us a probability of a defined success. Therefore first of all we have to define which is a success in our model. We decide to make two different models, in the first model we define the success when the local team wins, and the second when the away team wins. Our model follow the following equation:

$$p_{i,j} = \frac{1}{1 + e^{\beta_0 + \beta_i + \beta'_j}}$$

where $p_{i,j}$ is the likelihood that the defined success happened, β_i is the local coefficient of the team i and β'_i is the away coefficient of the team.

Then we studied the results using ROC curves. As we did in the Maher's model we suggest a improvement giving weights to each season. We studied the new model although we can see clearly that this change improve the model.

Then we have two models to predict football results and we are able to make money with ours studies. We used our forecasts to predict new matches results and we make a betting strategy to make money. We try our betting strategy with our four models and we earn money with three of them. The best model was the logistic model with weights which make us earn almost three times our bank money. Surprisingly the worst model was logistic simple model which make us lose more than 45 % of our bank in one season.

Índice general

Capítulo 1

Introducción

El fútbol es el deporte más seguido del mundo, para comprobarlo basta con ver los espectadores que tuvo la final del mundial de 2014, la cual alrededor de 1000 millones de personas vieron dicho partido, comparado con la inauguración de los juegos olímpicos de Londres en 2012 con 900 millones o la Superbowl con 160 millones muestra la enorme popularidad de este deporte. Es difícil encontrar alguien que no tenga una idea básica sobre este juego de equipo, cuyo objetivo es introducir una pelota dentro de la portería contraria, acción denominada marcar un gol. El equipo que logre más goles al cabo del partido, de una duración de 90 minutos, es el que resulta ganador del encuentro, existiendo la posibilidad de empatar.

Los inicios del fútbol modernos constan del 1863 y desde entonces pocas normas han sido modificadas; su simpleza y las facilidades para practicar este deporte fueron claves para su expansión, primero por Europa y poco a poco llegar a ser el deporte más practicado del planeta.

La idea de que algo que depende de tantas variables como es un partido de fútbol pueda predecirse a través de un modelo matemático puede sonar muy ambiciosa, pero ya existen varios modelos que son capaces de dar unas probabilidades bastantes cercanas a la realidad si tenemos los datos suficientes.

Hoy en día los resultados de temporadas anteriores están a disposición de todo el mundo, existen numerosas paginas web en las que comprobar cualquier dato sobre los partidos del fútbol profesional, desde el resultado, hasta incluso las condiciones climatologicas, pasando por todo tipo de ausencias de jugadores o tiempos de descanso de ambos equipos. Esto nos da la opción de utilizar toda esta información que se encuentra a nuestro alcance para predecir resultados de los próximos partidos. Lo que más puede sorprender a los aficionados de este deporte es que encontraremos modelos capaces de predecir con cierta precisión el ganador de un partido utilizando solo los goles anotados por cada equipo en anteriores partidos.
La mayoría de modelos que intentan predecir resultados deportivos están motivados en la idea de ganar dinero en las apuestas. La profesión de corredor de apuestas es bastante más antigua de lo que se puede llegar a creer ya que desde la Antigua Roma se hacían apuestas en las batallas de gladiadores, aunque donde empezó a ser un negocio lucrativo fue en Reino Unido en las carreras de caballos y poco a poco fue ampliándose a otros deportes hasta llegar al fútbol.

Desde que comenzaron, las apuestas, han ido creciendo su popularidad exponencialmente y actualmente sólo en España movieron mas de 4091 millones de euros en 2015 y este numero sigue creciendo. Por ello no resulta extraño pensar que cada vez más gente esté interesada en predecir el resultado final de un partido de fútbol y utilizar esos cálculos para intentar sacar un beneficio económico.

En la parte principal de este trabajo estudiaremos dos modelos distintos y su capacidad para predecir resultados, se propondrán varias modificaciones y se estudiará si con ellas sería posible mejorar

el modelo de predicción. Los modelos que utilizaremos serán un modelo de Poisson multivariante y un modelo de regresión logística Estos modelos están basados en artículos ya publicados pero este trabajo servirá para comprobar si los estudios pueden aplicarse al futbol de la liga española en la actualidad, ya que puede que con el paso de los años alguna variables que antes se ignoraban ahora sean determinantes a la hora de predecir un resultado.

Una vez estudiados y contrastados estos modelos los aplicaremos en un supuesto caso de apuestas y buscaremos obtener beneficio económico gracias a las casas de apuestas. Para entender correctamente esta parte del trabajo tendremos que introducir unos conceptos básicos sobre casas de apuestas, como son:

1. **Cuota**: Es la cantidad por la que nos multiplicarán la inversión en caso de acertar nuestra apuesta, como es lógico esta nunca sera menor que 1 y cuanto más mayor sea, nos dirá que la casa de apuestas considera que el resultado tiene una probabilidad menor de darse.

2. **Apuesta fija**: son las apuestas en las que la cuota se mantiene constante hasta el inicio del evento.

3. **Spread**: Dada la definición de cuota es claro que para que el juego fuera justo la cuota debería de ser igual al inverso de la probabilidad que se le de a que un suceso ocurra. Veámoslo con el caso de lanzar una moneda: La probabilidad de que salga cara es la misma que la de que salga cruz (0,5) por lo tanto nuestra cuota debería ser $\frac{1}{0,5} = 2$. En la practica esto no es así, ya que las casas de apuestas se guardan un "margen de beneficio.[en] el que si sólo pueden darse dos resultados y a cada uno se le asigna la misma probabilidad la cuota de cada resultadojuga suele rondar el $1,85$ lo que implicaría que se le esta dando una probabilidad de $0,54$ y como podemos ver ya no se cumpliría el axioma de probabilidad que nos dice que la suma de la probabilidad de todos los sucesos es 1 sino que en este caso es $1,08$. Esto asegura a las casas de apuestas un beneficio del $8\,\%$ en el caso ideal en el que haya la misma cantidad de dinero apostado a cada uno de los eventos. A este porcentaje se le denomina Spread y las casas de apuestas deben jugar con él ya que cuanto mayor sea mayores beneficios obtendrán, pero sus cuotas serán menos atractivas para sus clientes.

Ahora que hemos introducido cómo se calculan las cuotas, vamos a concretar un poco el problema que trataremos. Como ya hemos dicho vamos a estudiar partidos de fútbol, en estos eventos pueden darse 3 resultados distintos: Victoria local, victoria visitante o empate. Suponemos que las cuotas son fijas, y el principal objetivo del trabajo será encontrar apuestas favorables que nos permitan sacar un beneficio económico a la larga.

El problema al que vamos a hacer frente en el trabajo consta de dos partes muy diferenciadas, ambas igual de importantes a la hora de sacar una estrategia de apuesta. El primer paso consiste en encontrar un modelo capaz de sacar la probabilidad de que suceda cada uno de los sucesos, en nuestro caso utilizaremos dos modelos un modelo bivariante de Poisson y otro de regresión logística. Una vez calculadas las probabilidades y encontrada una apuesta favorable debemos de decidir la estrategia a seguir para apostar una cantidad de nuestro banco o presupuesto, en nuestro caso será siempre la misma. Esto es muy importante ya que aun encontrando apuestas favorables si nuestra estrategia no es la correcta podemos llegar al estado de banca rota. Veámoslo con un ejemplo:

Si alguien nos ofrece la posibilidad de intentar adivinar cómo caerá una moneda y en caso de acertar nos da el triple de lo apostado, es evidentemente una apuesta favorable, ya que lo justo seria que te pagaran solamente el doble. Pero si apostamos todo nuestro dinero tenemos un $50\,\%$ de probabilidades de arruinarnos en el primer intento. Por este motivo es importante planear una buena estrategia de apuesta que nos permita ganar dinero pero evitando caer en la bancarrota.

Capítulo 2

Modelos de predicción

2.1. Modelo de Maher

2.1.1. Introducción

Uno de los modelos más citados en lo que a predicción de resultados de partidos de fútbol se refiere es el propuesto por Maher (1982) en el que se propone un modelo para predecir el numero de goles que marcará cada equipo. Para ello se toman las siguientes suposiciones:

1. El equipo que juega de local tiene una ligera ventaja frente a su rival. Esto se tiene en cuenta con un factor que llamaremos "factor cancha ".

2. El modelo se encargará de asignar unos valores con los que medir las habilidades de cada equipo. Dichas habilidades se dividirán en dos, la capacidad de meter goles (coeficiente de ataque) y la capacidad de no encajar goles (coeficiente de defensa).

3. El numero de goles marcado por un equipo lo podemos suponer independiente al numero de goles marcado por su equipo rival.

Antes de empezar con la explicación del modelo en si definiremos la notación que usaremos en este capitulo. Definimos $X_{i,j}$ como el número de goles marcados por el equipo local en el partido que enfrenta al equipo i contra el equipo j siendo el equipo i el local. Análogamente definimos $Y_{i,j}$ como el número de goles marcados por el equipo visitante en el mismo partido. La propuesta que hace Maher es la siguiente:

$$X_{i,j} \sim Poisson(\alpha_i \beta_j \gamma)$$

$$Y_{i,j} \sim Poisson(\alpha_j \beta_i)$$

Aquí α_i puede interpretarse como el coeficiente de ataque del equipo i, β_i como el coeficiente de defensa del equipo i y γ como el factor cancha. Como hemos dicho antes supondremos que $X_{i,j}$ e $Y_{i,j}$ son variables independientes. Por la definición que hemos dado es claro que $\alpha_i > 0$, $\beta_i > 0$ y $\gamma > 1$. Además podemos observar que el coeficiente de ataque (α) debe subir la media ya que es la capacidad goleadora, por lo que cuanto mayor sea el α de un equipo más goles meterá por partido. Por el contrario el coeficiente defensivo (β) es la capacidad que tiene el equipo rival para que no le metan gol, por lo que un valor bajo de este nos indica que el equipo en cuestión encajará pocos goles.

2.1.2. Obtención de los estimadores

Maher propone estimar los parámetros α_i, β_i y γ por máxima verosimilitud. Teniendo en cuenta todo lo ya mencionado podemos calcular la probabilidad de que el partido acabe con el equipo local marcando x goles y el equipo visitante y:

$$P(X_{i,j}=x, Y_{i,j}=y) = \frac{(\alpha_i\beta_j\gamma)^x e^{-\alpha_i\beta_j\gamma}}{x!}\frac{(\alpha_j\beta_i)^y e^{-\alpha_j\beta_i}}{y!} \qquad x,y=0,1,2...$$

Por lo tanto nuestra función de verosimilitud, si tenemos N equipos distintos y un total de m partidos, es la siguiente:

$$L((\mathbf{x},\mathbf{y}),\alpha_i,\beta_i,\gamma,i=1,...N) = \prod_{k=1}^{m}\frac{(\alpha_{i(k)}\beta_{j(k)}\gamma)^{x_k}e^{-\alpha_{i(k)}\beta_{j(k)}\gamma}}{x_k!}\frac{(\alpha_{j(k)}\beta_{i(k)})^{y_k}e^{-\alpha_{j(k)}\beta_{i(k)}}}{y_k!}$$

Donde $(\mathbf{x},\mathbf{y})) = \{(x_1,y_1),...(x_k,y_k)\}$ es el vector de duplas de los goles marcados en el partido k de nuestra muestra, $i(k)$ es el indice del equipo local en el partido k y $j(k)$ es el indice del equipo visitante en el partido k.

En vez de trabajar con la función de verosimilitud L, es más sencillo trabajar tomando logaritmos, por lo tanto:

$$\begin{aligned}\log L((\mathbf{x},\mathbf{y}),\alpha_i,\beta_i,\gamma,i=1,...N) = \sum_{k=1}^{m} & x_k(\log\alpha_{i(k)}+\log\beta_{j(k)}+\log\gamma)-\alpha_{i(k)}\beta_{j(k)}\gamma - log(x_k!)\\ & +y_k(\log\alpha_{j(k)}+\log\beta_{i(k)})-\alpha_{j(k)}\beta_{i(k)} - log(y_k!)\end{aligned}$$

Para este modelo vamos a utilizar los datos de las últimas tres temporadas de la Liga. Definimos unas variables *dummy* A_i y B_i que nos indican con un 1 si el equipo i ha jugado el partido siendo local o visitante respectivamente o un 0 en caso de que no sea así.

Gracias a nuestras variables *dummy* podemos prescindir de los indices dependientes de k, excepto x_k y y_k. Por todo esto podemos reescribir nuestra función de verosimilitud de la siguiente manera:

$$\begin{aligned}L((\mathbf{x},\mathbf{y}),\alpha_i,\beta_i,\gamma,i=1,...N) = & \prod_{k=1}^{m}\frac{1}{x_k!y_k!}\prod_{i=1}^{N}\prod_{j\neq i}(\alpha_i\beta_j\gamma)^{A_iB_jx_k}e^{-\alpha_i\beta_j\gamma A_iB_j}\\ & (\alpha_j\beta_i)^{y_kA_iB_j}e^{-\alpha_j\beta_iA_iB_j}\end{aligned}$$

Como anteriormente tomamos logaritmos:

$$\begin{aligned}\log L((\mathbf{x},\mathbf{y}),\alpha_i,\beta_i,\gamma,i=1,...N) = \sum_{k=1}^{m}\Big(& -\log(x_k!)-\log(y_k!)+\sum_{i=1}^{N}\sum_{i\neq j}A_iB_jx_k(\log\alpha_i+\log\beta_j+\log\gamma)\\ & -A_iB_j\alpha_i\beta_j\gamma+A_iB_jy_k(\log\alpha_j+\log\beta_i)-A_iB_j\alpha_j\beta_i\Big),\end{aligned}$$

que a la hora de programar será la formula más sencilla que podemos utilizar. Notar también que a la hora de maximizar la función de verosimilitud para estimar los valores de α_i, β_i y γ, el $\log(x_k)$ y el $\log(y_k)$ no influyen. Por los que en la programación los omitiremos.

Antes de calcular la solución debemos asegurarnos de la unicidad de esta. Notar que si $(\alpha'_1,...,\alpha'_{12}, \beta'_1,...,\beta'_{12},\gamma)$ es solución, entonces$(10\alpha'_1,10\alpha'_2,.....,10\alpha'_{27},\frac{\beta'_1}{10},....,\frac{\beta'_{27}}{10},\gamma)$ también lo es. Para solucionar este problema tomamos la decisión de inicializar el coeficiente de ataque del primer equipo que participa en la base de datos con el valor 1, de esta manera el valor de α_1 ya estará fijado y nos podremos asegurar la unicidad de la solución.

Cuando observamos nuestra muestra, lo primero que podemos ver es que se da el caso en el que algunos equipos no se enfrentan entre ellos debido a los ascensos y descensos de cada temporada o, de darse el enfrentamiento, solo lo tenemos una temporada. Para evitar problemas con datos anómalos y

mejorar la precisión de la estimación de los parámetros, tomaremos como restricción en la muestra que sólo computaran los equipos que han conseguido mantenerse en todas las ultimas 3 temporadas. De esta forma solventamos el problema de la falta de datos. Dado que el primer equipo por orden alfabético que se encuentra en nuestra muestra es el Athletic de Bilbao, inicializaremos su coeficiente de ataque en 1 y por tanto $\alpha_1 = 1$.

Obtenemos la solución a través de R y utilizando la función *optim* (en el apéndice se encuentra el programa), los resultados obtenidos son los siguientes:

$\gamma = 1,44$	α_i	β_i
Athletic de Bilbao	1	0,86
Atletico de Madrid	1,34	0,42
Barcelona	2,00	0,61
Celta	1,08	0,99
Espanyol	1,01	1,10
Granada	0,69	1,17
Malaga	0,72	0,80
Real Madrid	2,29	0,85
Real Sociedad	1,11	0,99
Sevilla	1,36	1,07
Valencia	1,25	0,84
Villareal	1,15	0,77

Cuadro 2.1: Estimación de los parámetros α_i, β_i y γ del modelo 2.1

2.1.3. Estudio de los resultados

Una vez estimados los parámetros α_i, β_i y de γ debemos comprobar si el modelo se ajusta a la realidad. Para ello lo primero que haremos será calcular el numero de goles que esperamos que marque cada equipo de local y de visitante, para compararlo con los resultados obtenidos.

Para calcular la media de goles de una temporada, utilizaremos que la suma de variables de Poisson independientes es una Poisson cuya media es la suma de las medias. De esta forma tenemos, por el equipo i, que el numero de goles como local es:

$$\sum_{j\neq i} X_{i,j} \equiv \sum_{j\neq i} Pois\left(\alpha_i\beta_j\gamma\right) = Pois\left(\sum_{j\neq i}\alpha_i\beta_j\gamma\right) = Pois\left(\alpha_i\gamma\sum_{j\neq i}\beta_j\right)$$

Por tanto, el número de goles del equipo local en una temporada en la que se enfrenta a los N equipos restantes tendrá como valor esperado $\alpha_i\gamma\sum_{i\neq j}\beta_j$ y evidentemente la media de tres temporadas simplemente será multiplicar por 3 el resultado obtenido. De forma análoga y eliminando el factor cancha podemos hacer lo mismo con los goles visitantes. Obteniendo la figura 2,1

A simple vista podemos ver que el modelo se ajusta bastante bien a la realidad. En la gráfica de goles anotados como local (derecha) vemos que los equipos que más se alejan son el Celta , Espanyol y Athletic, por contra los equipos que más se ajustan entre lo esperado y lo obtenido son el Sevilla , el Valencia y la Real Sociedad. Son valores muy buenos ya que el coeficiente de correlación lineal es de 0,9759.

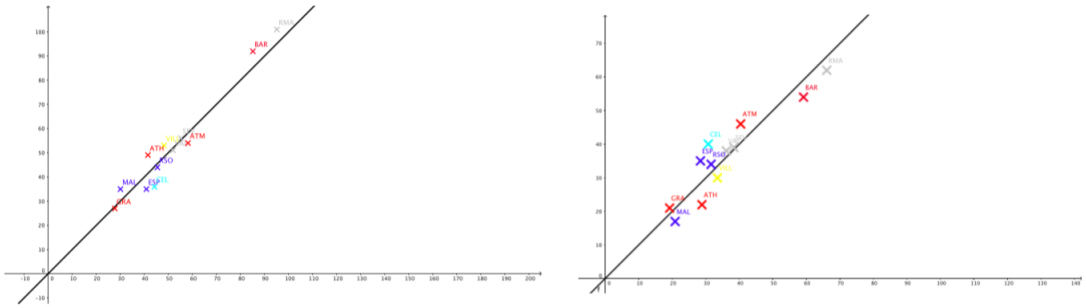


Figura 2.1: Gráfica que representa los goles anotados frente a los goles esperado jugando de visitante durante los 3 años que hemos anotado los datos (Si se desea en el anexo se encuentran las gráficas ampliadas).

En lo que respecta a la gráfica de goles visitante los errores aumentan ligeramente y vuelven a ser el Athletic y el Celta los equipo más alejados de la recta, los equipos en los que nuestro modelo es más preciso también coinciden con los goles locales y son el Sevilla y el Valencia. El coeficiente de correlación en este caso también es bastante cercano a 1 pero un poco peor que en la anterior gráfica, alcanzando el valor de $0,930$.

Cómo hemos supuesto que los goles del local y del visitante son independientes podemos calcular las probabilidades que tiene cada equipo de ganar el partido. Ya hemos mencionado anteriormente cual sería la probabilidad de que un partido acabara con (x,y) es:

$$P(X_{i,j}=x, Y_{i,j}=y) = \frac{(\alpha_i\beta_j\gamma)^x e^{-\alpha_i\beta_j\gamma}}{x!}\frac{(\alpha_j\beta_i)^y e^{-\alpha_j\beta_i}}{y!}$$

Realizando ese calculo para los valores de x e y y sumando la probabilidad de los casos en los que $x > y$ sacamos la probabilidad de que el equipo local gane el partido. Análogamente hacemos lo mismo con el empate y la victoria visitante. La suma de las probabilidades de cada uno de los diferentes partidos que puede jugar un equipo en casa nos da el numero esperado de partidos que gana, empata y pierde durante una temporada y lo mismo sucederá con los que juega como visitante. Volvemos a multiplicar por 3 ya que en nuestra base de datos tenemos 3 temporadas y comparamos con los resultados de la muestra, es de esperar que las gráficas obtenidas tengan una desviación mayor que en el anterior estudio ya que los datos observados toman valores mucho más pequeños. Las gráficas obtenidas son la figura 2.2

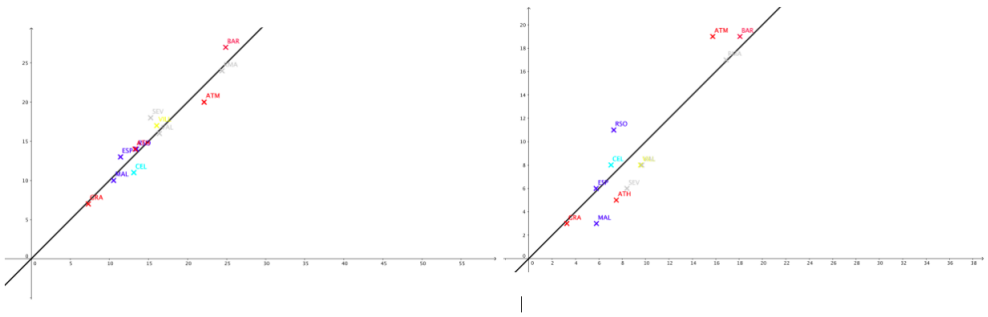


Figura 2.2: Gráfica que representa los partidos ganados frente a los victorias esperadas jugando como visitante durante los 3 años que hemos anotado los datos (Si se desea en el anexo se encuentran las gráficas ampliadas).

En lo que respecta a las victorias locales y visitante vemos que los valores se aproximan bastante

bien a la recta como muestran sus coeficientes de correlación 0,9648 y 0,9370 respectivamente. En el caso de la victoria local todos los valores están bastante bien ajustados, en cambio en la recta de victoria visitante vemos como la Real Sociedad si que se aleja un poco del resultado esperado.

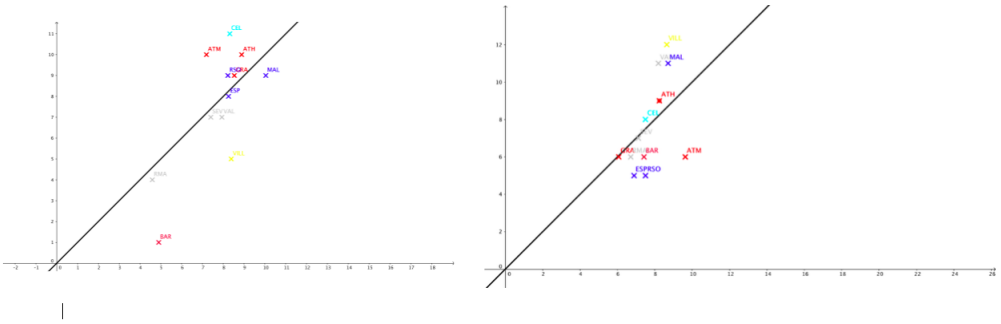


Figura 2.3: Gráfica que representa los partidos empatados frente a los empates esperados como visitante durante los 3 años que hemos anotado los datos (Si se desea en el anexo se encuentran las gráficas ampliadas).

Como es de esperar el empate es el resultado más difícil de predecir. Aunque en las gráficas (representadas en la figura 2.3) los resultados no se alejan tampoco mucho exceptuando algunos valores, como el Barcelona en la gráfica local (derecha). Es importante tener en cuenta que los valores del numero de empates es mucho más bajo que el resto, por lo que una desviación de uno o dos partidos hace que los valores se alejen mucho de la recta.Los indices de correlación son significativamente más bajos que en el resto de gráficas y toman los valores $0,747$ en los empates locales y $0,5289$ en los visitantes.

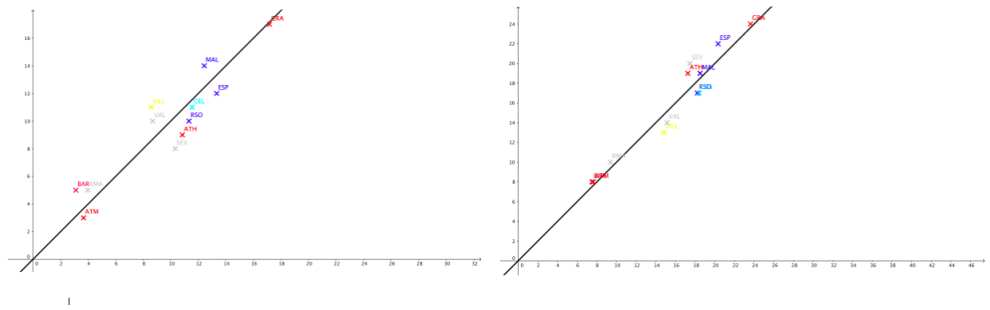


Figura 2.4: Gráfica que representa los partidos perdidos frente a las derrotas esperadas jugando como local durante los 3 años que hemos anotado los datos (Si se desea en el anexo se encuentran las gráficas ampliadas).

En el par de gráficas de la derrota (figura 2.4) es en el único en la que la gráfica visitante (derecha) se acerca más a la recta que la local, aunque solo ligeramente ya que los coeficientes de correlación son 0,900 y 0,9274. Destacar la gran aproximación a la recta de los valores del Granada en ambas gráficas. Si se desea comparar algún valor con más exactitud mirar anexo.

Como última comprobación a nuestro modelo contabilizaremos las veces que da como suceso más probable la vitoria local, el empate o la victoria visitante, a continuación lo contrastaremos con nuestros datos. Dicha comparación la podemos ver en el cuadro 2.2.

Los resultados son favorables al modelo ya que de 396 resultados 223 se aciertan, lo que supone un $56,5\,\%$, también podemos ver que lo que mejor predecimos es la victoria local ya que llegamos a

PREDICCIÓN | RESULTADO	VICTORIA LOCAL	EMPATE	VICTORIA VISITANTE
VICTORIA LOCAL	158 (82,72 %)	72(71,28 %)	50(43,85 %)
EMPATE	2(0,01 %)	1(0,001 %)	0(0,00 %)
VICTORIA VISITANTE	31(16,23 %)	28(27,72 %)	64(56,14 %)
TOTAL	191	101	114

Cuadro 2.2: Se enfrentan la predicción que e modelo considera más favorable con el resultado que se dio en cada partido, al lado de la frecuencia se encuentra el porcentaje por columnas

acertar el $82,7\,\%$ y con la victoria visitante el porcentaje baja hasta el $78,1\,\%$. Donde el modelo falla es a la hora de predecir el empate, pero esto no nos preocupa ya que las casas de apuestas nunca dan como suceso más probable el empate, además en nuestra estrategia de apuestas no solo apostaremos al suceso más probable. Por lo que definitivamente aceptamos que este modelo es válido a la hora de predecir.

2.1.4. Mejora del modelo

Aunque nuestro modelo de resultados aceptables, se pueden implantar mejoras. Una de las criticas que se le hace al modelo de Maher es que no tiene en cuenta la temporada en la que se dan los sucesos. Si el objetivo final es predecir los resultados de esta temporada, es lógico darles un peso mayor a los partidos de temporadas recientes ya que las plantillas serán lo más similares posibles a las actuales. Para ello simplemente damos peso 2 los datos de la segunda temporada y peso 3 a los datos de la temporada más cercana, de esta forma un resultado de la temporada $2015-2016$ tendrá más influencia en los valores de nuestros coeficientes que un partido de la temporada $2013-2014$ a la que se le asigna peso 1. Con este cambio en el modelo los nuevos valores son los del cuadro 2.3.

Al realizar las gráficas y tablas de ajuste en este modelo notamos un ligero empeoramiento en el ajuste a los datos lo cual es lógico pero esperamos que esta estimación nos sea más útil a la hora de predecir.

$\gamma = 1,43$	α	β
Athletic de Bilbao	1	0,84
Atletico de Madrid	1,42	0,40
Barcelona	2,15	0,57
Celta	1,15	0,98
Espanyol	1,03	1,12
Granada	0,74	1,13
Malaga	0,73	0,73
Real Madrid	2,43	0,79
Real Sociedad	1,14	0,89
Sevilla	1,37	1,02
Valencia	1,31	0,81
Villareal	1,20	0,68

Cuadro 2.3: Estimación delos parámetros α_i, β_i y γ del modelo 2.1 con pesos

2.1.5. Otras propuestas de mejora

En su trabajo, Maher, propone varias variaciones del modelo que podrían ser interesantes en algunos casos. Una de las suposiciones que hemos hecho es que el factor cancha es igual para todos los equipos.

Pero podríamos diferenciar los distintos campos teniendo en cuenta que hay estadios en los que el equipo local mejora mucho más su juego que otro. Por lo que una idea para cambiar el modelo sería incluir γ_i en lugar de γ. Llegando más lejos todavía, se podría afirmar que jugar como visitante influye más a unos equipos que a otros, Maher también propuso separar las variables en 4, que serían, coeficiente de ataque como local y visitante y los respectivos coeficientes de defensa como local y visitante. La desventaja de esta variación del modelo es que para realizarlo necesitaríamos muchos más datos para conseguir la convergencia ya que estamos duplicando el numero de parámetros.

Otra idea interesante para mejorar el modelo es la propuesta por Dixon y Coles (1997) en el que no suponen independientes los goles en los casos en los que se anotan pocos goles en el partido, para ellos:

$$P(X_{i,j} = x, Y_{i,j} = y) = \tau_{\lambda,\mu}(x,y)\frac{(\lambda)^x e^{\lambda}}{x!}\frac{(\mu)^y e^{-\mu}}{y!} \qquad x,y = 0,1,2...$$

Donde:

$$\tau_{\lambda,\mu}(x,y) = \begin{cases} 1-\lambda\mu\rho & si \quad x=y=0 \\ 1+\lambda\rho & si \quad x=0, y=1 \\ 1+\mu\rho & si \quad x=1, y=0 \\ 1-\rho & si \quad x=1, y=1 \\ 1 & \text{otro caso} \end{cases}$$

Siendo $\lambda = \alpha_i\beta_j\gamma$, $\mu = \alpha_j\beta_i$ y ρ un valor comprendido en el intervalo $max(-\frac{1}{\lambda}, -\frac{1}{\mu}) \leq \rho \leq min(\frac{1}{\lambda\mu}, 1)$. Notar que para $\rho = 0$ nos encontramos con el modelo que hemos estudiado en este trabajo.

Como podemos comprobar en nuestra muestra, el $56,89\,\%$ de los errores en la predicción se dieron en partidos en los que ningún equipo marcó más de un gol, en el caso del modelo mejorado el porcentaje aumenta hasta $57,14\,\%$ con lo que podemos suponer que esta variación del modelo podría ser interesante para mejorar la predicción de resultados. Por lo que vamos a hacer un estudio un poco más detenido del caso.
En caso de ser independientes se tendría:

$$P(X=0, Y=0) = P(X=0)P(Y=0)$$

Por lo que el producto de las frecuencias relativas debería ser igual (o al menos similar) a la frecuencia del resultado exacto, veamos en la siguiente tabla, propuesta en el trabajos de Dixon y Coles (1997), cuando tenemos suficientes datos el valor del valor del cociente de $\dfrac{P(X=0,Y=0)}{P(X=0)P(Y=0)}$ es muy cercano a 1 lo que no nos sirve para desmentir que las variables sean independientes, por lo que no resultaría interesante complicar el modelo. Como vemos en el cuadro 2.4 los valores en los que tenemos suficientes datos rondan el valor que se encuentra entre paréntesis, que es el cociente del que hemos hablado antes esta en valores muy cercanos a 1. Esto no es una confirmación de que las variables son independientes, pero no nos hace negarlo, por lo que no implementaremos la propuesta de Dixon y Coles a nuestro modelo.

	0	1	2	3	4	5	6	T
0	30 (0,99)	31 (0,97)	14 (0,85)	7 (1,13)	3 (1,50)	2 (2,25)	1 (4,50)	88
1	44 (0,96)	44 (0,91)	30 (1,21)	10 (1,06)	4 (1,32)	1 (0,74)	0 (0,00)	133
2	30 (0,97)	33 (1,01)	17 (1,01)	8 (1,26)	1 (0,49)	1 (1,10)	0 (0,00)	90
3	15 (0,95)	21 (1,26)	9 (1,05)	0 (0,00)	1 (0,96)	0 (0,00)	0 (0,00)	46
4	12 (1,40)	7 (0,77)	4 (0,86)	2 (1,13)	0 (0,00)	0 (0,00)	0 (0,00)	25
5	2 (0,83)	5 (1,96)	0 (0,00)	0 (0,00)	0 (0,00)	0 (0,00)	0 (0,00)	7
6	3 (2,18)	1 (0,69)	0 (0,00)	0 (0,00)	0 (0,00)	0 (0,00)	0 (0,00)	4
7	0 (0,00)	1 (1,38)	0 (0,00)	1 (7,07)	0 (0,00)	0 (0,00)	0 (0,00)	2
8	0 (0)	0 (0,00)	0 (0,00)	0 (0,00)	0 (0,00)	0 (0,00)	0 (0,00)	0
9	0 (0,00)	1 (2,75)	0 (0,00)	0 (0,00)	0 (0,00) 0	(0,00)	0 (0,00)	1
T	136	144	74	28	9	4	1	

Cuadro 2.4: Frecuencias de cada resultado durante las 3 temporadas, en paréntesis el cociente entre la frecuencia de un resultado y las frecuencias relativas de. los goles del equipo local y visitante por separado

2.2. Modelo de regresión logística

Este segundo modelo que vamos a crear es un modelo de regresión logística. Este tipo de modelos son capaces de dar una probabilidad al éxito y, por tanto, otra probabilidad al fracaso. Estos modelos suelen ser muy utilizados en biomatemáticas, biomedicina o estudios de epidemiología, en los que hay dos casos muy diferenciados, por ejemplo, muerte durante una cirugía, la probabilidad de que un individuo se contagie de una enfermedad, etc... En nuestro caso concreto de los partidos de fútbol pueden darse 3 sucesos por lo que haremos un modelo que intente predecir la victoria local y posteriormente realizaremos un modelo similar para predecir la victoria visitante, para el caso de empate no realizaremos ningún modelo ya que, como hemos comentado en el apartado 2.1 es un resultado algo complejo de predecir. Para ello utilizaremos como muestra los resultados de las últimas 3 temporadas en la que volveremos a hacer la restricción de sólo contar los 12 equipos que han permanecido todas las temporadas en primera división, para así evitar problemas de falta de enfrentamientos debido a ascensos y descensos.

Como ya hemos mencionando antes, nuestra variable puede tomar los valores $Y = 0$ ó $Y = 1$ donde $Y = 1$ implica que el suceso ocurre. Llamando X al vector de predictores $X_1, X_2, ..., X_k$ los modelos de regresión lineal sería de la siguiente forma:

$$\mathrm{E}(Y|X) = X\beta$$

donde β son los pesos que se les da a cada suceso predictor. Notar que

$$\mathrm{E}(Y|X) = \mathrm{Prob}(Y = 1|X) = X\beta$$

Así $\mathrm{Prob}\{Y = 1|X\}$ pueda superar el valor 1 o ser menor que 0. Para afrontar ese problema se propone una técnica clásica en estadística conocida como el modelo de regresión logística (Jobson,2012) el cual propone:

$$\mathrm{Prob}(Y = 1|X) = \frac{1}{1 + e^{-X\beta}}$$

Para un valor de $X_i = (x_{1,i}, ..., x_{n,i})$, tenemos $p_i = P[Y_i = 1|X_i] = \frac{1}{1 - e^{\beta_0 + \beta_1 x_{1,i} + ... + \beta_n x_{n,i}}}$.

Definimos ahora $logit(p_i) = log(\frac{p_i}{1-p_i})$ y se tiene que $logit(p_i) = \beta_0 + \beta_1 x_{1,i} + ... + \beta_k x_{1,i}$.

En las que los se deben calcular los valores de β_i para que el modelo se ajuste a la muestra que tenemos, normalmente se hara mediante el método del estimador máximo verosímil

2.2.1. Presentación del modelo

Existen varios trabajos que comienzan asignando coeficientes relativamente subjetivos a cada equipo sobre distintas facetas del juego y posteriormente al calcular los β_i se da pesos a esas distintas facetas del juego Chinwe y Enoch (2014).

Esta es una forma de trabajar valida, pero en este trabajo consideraremos otra manera de realizar el modelo. Buscaremos que los coeficientes sean solo dos por equipo, un coeficiente indicará la calidad del equipo cuando juega en casa y lo denominaremos çoeficiente localz también calcularemos el çoeficiente visitante"que será el coeficiente que nos indique como de bueno es cada equipo cuando juega de visitante. Para ello nuestra muestra serán unas variables *Dummy* que crearemos en nuestra base de datos. Este tipo de variables son unos indicadores que toman el valor 1 en un caso determinado y el valor 0 en el resto. Nosotros las usaremos las variables que llamaremos A_i a las que nos indique si el equipo i juega de local y otras variables B_i que haran lo propio cuando el equipo i juegue de visitante.

Una vez decidido cuales son nuestras variables predictoras tenemos que asegurarnos de la unicidad de la solución, supongamos que $(\beta_0, \beta_1, ...\beta_{12}, \beta'_1, ..., \beta'_{12})$ es solución siendo los β'_i los coeficientes de visitante del equipo i. Entonces $(\beta_0, \beta_1 + a, ...\beta_{12} + a, \beta'_1 - a, ..., \beta'_{12} - a)$ también es solución por lo que tenemos que inicializar un valor de los coeficientes de local. Una vez inicializado unos de los coeficientes de local los valores de β_i con $i \neq 0$ estarán fijados pero notar que $(\beta_0 + 12a, \beta_1, ...\beta_{12}, \beta'_1 - a, ..., \beta'_{12} - a)$ también es solución por lo que tendremos que fijar también uno de los valores de los coeficientes de visitante. De esta forma fijamos al Athletic de Bilbao sus dos valores con la dupla $(0, 0)$ y una vez hecho esto podemos calcular el resto de valores.

2.2.2. Obtención de coeficientes

En cualquier modelo de regresión, cuantas menos variables haya que ajustar mejor podremos ajustarlas, en nuestro caso tenemos 25 variables (12 coeficientes de ataque, otros 12 de defensa y el β_0) pero hay equipos a los cuales al calcular sus coeficientes de local y visitante el p-valor de dichos coeficientes es demasiado alto, lo que quiere decir que son variables no significativas, dicho de otra manera, se pueden sustituir por 0 ya que no influyen en el modelo. Esto nos ayudará a calcular el resto de variables con mayor precisión.

R tiene la opción de realizar la eliminación de las variables no significativas y volver a calcular el modelo hasta conseguir que todas las variables sean significativas esto se hace con la función *stepAIC* que se encuentra en el paquete *MASS*. Gracias a esta función obtenemos los valores de los coeficientes de local y visitante de todos los equipos que estan presentes en nuestra muestra. Considerando éxito la victoria local los coeficientes son los mostrados en el cuadro 2.5.

Como podemos comprobar Sevilla, Espanyol, Real Sociedad y Athletic tienen exactamente los mismos valores por lo que a la hora de calcular las probabilidades estos cuatro equipos obtendrán los mismos resultados.

En cambio si consideramos éxito la victoria visitante los coeficientes toman los valores del cuadro 2.6.

$\beta_0 = 0,336$	Coeficiente local	Coeficiente visitante
Athletic	0,000	0,000
Atletico	0,000	-1,604
Barcelona	1,612	-1,405
Celta	-0,655	0,000
Espanyol	0,000	0,000
Granada	-1,241	0,628
Malaga	-0,807	0,000
Real Madrid	1,427	-1,136
Real Sociedad	0,000	0,000
Sevilla	0,000	0,000
Valencia	0,000	-0,662
Villarreal	0,000	-0,804

Cuadro 2.5: Estimación de los parámetros β_i y β_i' del modelo 2.2 en el que buscamos calcular la probabilidad de la victoria local

$\beta_0 = -1,086$	Coeficiente local	Coeficiente visitante
Athletic	0,000	0,000
Atletico	-1,515	1,489
Barcelona	-0,881	1,543
Celta	0,000	0,000
Espanyol	0,000	0,000
Granada	0,879	-1,006
Malaga	0,000	-1,134
Real Madrid	-0,928	1,276
Real Sociedad	0,000	0,000
Sevilla	0,000	0,000
Valencia	0,000	0,000
Villarreal	0,000	0,000

Cuadro 2.6: Estimación de los parámetros β_i y β_i' del modelo 2.2 en el que buscamos calcular la probabilidad de la victoria visitante

Una vez obtenidos los coeficientes podemos calcular la probabilidad de que se de el éxito en nuestro cualquier partido, para ello simplemente tendremos que aplicar la siguiente formula

$$p_{i,j} = \frac{1}{1 + e^{\beta_0 + \beta_i + \beta_j'}}$$

Siendo $p_{i,j}$ la probabilidad de que en el partido en el que el equipo i juega de local y el equipo j juega de visitante se de lo que consideramos éxito.

De esta manera podemos realizar los cuadros 2.7 y 2.8 en los que vemos las probabilidades de que el equipo local (en columnas) gane al equipo visitante (en filas), notar que la suma de las probabilidades de los dos modelos podría superar el valor de 1 (aún teniendo en cuenta que el empate es una opción que no tendríamos en cuenta) ya que corresponden a modelos distintos que no guardan relación.

	ATH	ATM	BAR	CEL	ESP	GRA	MAL	RMA	RSO	SEV	VAL	VIL
ATH	X	0,583	0,875	0,421	0,583	0,288	0,384	0,805	0,583	0,583	0,583	0,583
ATM	0,220	X	0,585	0,128	0,220	0,075	0,112	0,453	0,220	0,220	0,220	0,220
BAR	0,256	0,256	X	0,151	0,256	0,090	0,133	0,503	0,256	0,256	0,256	0,256
CEL	0,583	0,583	0,875	X	0,583	0,288	0,384	0,805	0,583	0,583	0,583	0,583
ESP	0,583	0,583	0,875	0,583	X	0,288	0,384	0,805	0,583	0,583	0,583	0,583
GRA	0,724	0,724	0,929	0,577	0,724	X	0,539	0,885	0,724	0,724	0,724	0,724
MAL	0,583	0,583	0,875	0,583	0,288	0,384	X	0,805	0,583	0,583	0,583	0,583
RMA	0,310	0,310	0,693	0,189	0,310	0,115	0,167	X	0,310	0,310	0,310	0,310
RSO	0,583	0,583	0,875	0,583	0,288	0,384	0,805	0,583	X	0,583	0,583	0,583
SEV	0,583	0,583	0,875	0,583	0,288	0,384	0,805	0,583	0,583	X	0,583	0,583
VAL	0,419	0,419	0,783	0,273	0,419	0,173	0,244	0,680	0,419	0,419	X	0,419
VIL	0,385	0,385	0,758	0,245	0,385	0,153	0,218	0,649	0,385	0,385	0,385	X

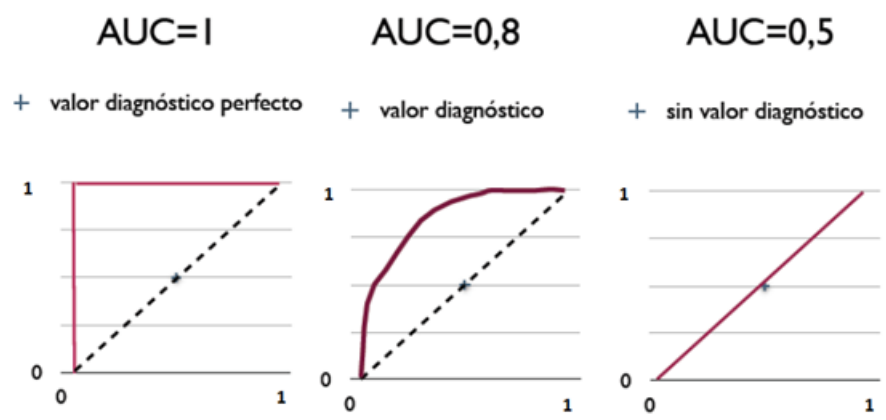
Cuadro 2.7: Probabilidad de la victoria local en los partidos según el modelo 2.2. El equipo local está ordenado por columnas y el visitante por filas.

	ATH	ATM	BAR	CEL	ESP	GRA	MAL	RMA	RSO	SEV	VAL	VIL
ATH	X	0,069	0,123	0,252	0,252	0,448	0,252	0,118	0,252	0,252	0,252	0,252
ATM	0,599	X	0,383	0,599	0,599	0,783	0,599	0,372	0,599	0,599	0,599	0,599
BAR	0,612	0,258	X	0,612	0,612	0,792	0,612	0,384	0,612	0,612	0,612	0,612
CEL	0,252	0,069	0,123	X	0,252	0,448	0,252	0,118	0,252	0,252	0,252	0,252
ESP	0,252	0,069	0,123	0,252	X	0,448	0,252	0,118	0,252	0,252	0,252	0,252
GRA	0,110	0,026	0,049	0,110	0,110	X	0,110	0,047	0,110	0,110	0,110	0,110
MAL	0,098	0,023	0,043	0,098	0,098	0,207	X	0,041	0,098	0,098	0,098	0,098
RMA	0,547	0,210	0,334	0,547	0,547	0,744	0,547	X	0,547	0,547	0,547	0,547
RSO	0,252	0,069	0,123	0,252	0,252	0,448	0,252	0,118	X	0,252	0,252	0,252
SEV	0,252	0,069	0,123	0,252	0,252	0,448	0,252	0,118	0,252	X	0,252	0,252
VAL	0,252	0,069	0,123	0,252	0,252	0,448	0,252	0,118	0,252	0,252	X	0,252
VIL	0,252	0,069	0,123	0,252	0,252	0,448	0,252	0,118	0,252	0,252	0,252	X

Cuadro 2.8: Probabilidad de la victoria visitante en los partidos según el modelo 2.2.El equipo local está ordenado por columnas y el visitante por filas.

2.2.3. Análisis de resultados

A simple vista los resultados puedes resultar aceptables, para comprobar si finalmente lo son vamos a realizar lo que se conoce como curva ROC. Una interpretación de este gráfico es la representación de la razón o ratio de verdaderos positivos (VPR = Razón de Verdaderos Positivos) frente a la razón o ratio de falsos positivos (FPR = Razón de Falsos Positivos) también según se varía el umbral de discriminación (valor a partir del cual decidimos que un caso es un positivo). Otra forma de ver este tipo de gráficas es que estamos dibujando la sensibilidad frente a la especificidad. Para comprobar que estas gráficas son aceptables tenemos la siguiente imagen obtenida en el portal web*wikiwand*:



Siendo el AUC el área que se encuentra por debajo de la gráfica, que el modelo tiene un valor de diagnostico aceptable si el AUC>$0,7$. Las gráficas que obtenemos de los modelos de la victoria local y la victoria visitante respectivamente son los siguientes:
Como vemos los valores del AUC están cerca del $0,75$ por lo que el modelo se acepta.

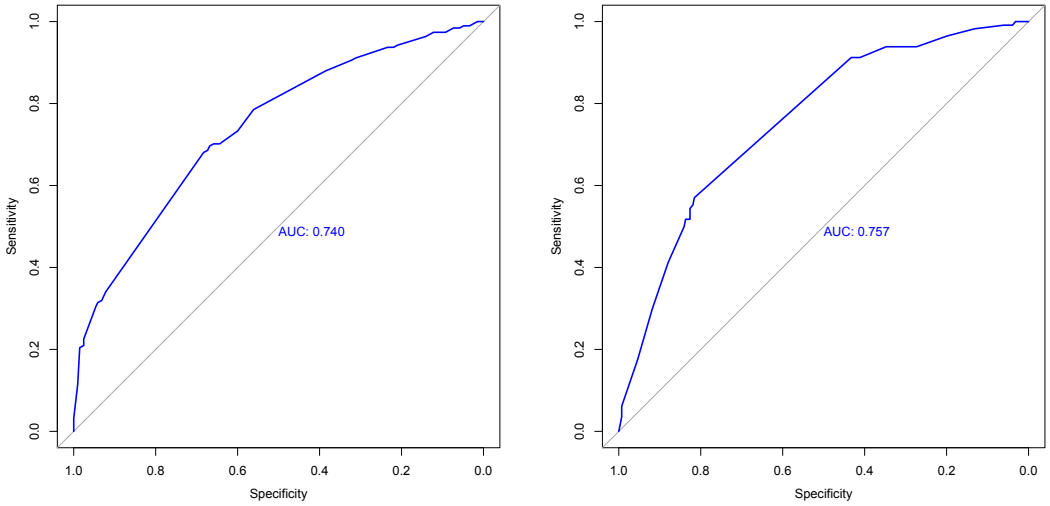


Figura 2.5: Curvas ROC del modelo 2.2. A la izquierda cuando consideramos éxito la victoria local y a la derecha cuando el éxito es la victoria visitante

2.2.4. Mejora de los modelos

Aunque nuestro modelo es considerado bastante aceptable existen algunas mejoras que se pueden plantear. Una de las más lógicas es la ya planteada en el punto 2.1.4 que se basa en la idea de que si un equipo ha hecho una temporada distinta a lo habitual debido a circunstancias externas, estas circunstancias estarán más presentes si la temporada está más cercana al momento actual por lo que asignando peso 2 a la 2^a temporada y peso 3 a la temporada más actual y conservando el peso 1 a la temporada más alejada en el tiempo previsiblemente mejoraremos el modelo. Los resultados para el modelo en el que consideramos éxito a la victoria local son los que se muestran en los cuadros 2.9 y 2.10.
Como vemos los valores son bastante similares y no se producen cambios muy significativos, por lo que el modelo debería de seguir siendo aceptable.
Los valores en el modelo en el cambio en nuestro otro modelo siguen la linea de las variaciones ya mencionadas y los resultados obtenidos se pueden comprobar en los cuadros 2.11 y 2.12:
Las respectivas curvas ROC de los modelos tienen unos valores del AUC muy similares a los anteriores, quizas la predicción de la victoria local sea mejor y de la victoria visitante sea un poco menos fina.

h

$\beta_0 = 0,299$	Coeficiente local	Coeficiente visitante
Athletic	0,000	0,000
Atletico	0,799	-1,725
Barcelona	1,969	-1,751
Celta	0,000	-0,743
Espanyol	0,000	0,000
Granada	-1,164	0,000
Malaga	-0,440	0,000
Real Madrid	1,427	-1,561
Real Sociedad	0,000	-0,602
Sevilla	0,897	0,000
Valencia	0,000	-0,884
Villarreal	0,691	-1,406

Cuadro 2.9: Estimación de los parámetros β_i y β_i' del modelo 2.2 con pesos en el que buscamos calcular la probabilidad de la victoria local.

	ATH	ATM	BAR	CEL	ESP	GRA	MAL	RMA	RSO	SEV	VAL	VIL
ATH	X	0,750	0,906	0,574	0,574	0,249	0,465	0,849	0,574	0,768	0,574	0,729
ATM	0,194	X	0,633	0,194	0,194	0,079	0,134	0,500	0,194	0,371	0,194	0,324
BAR	0,190	0,342	X	0,190	0,190	0,079	0,131	0,494	0,190	0,365	0,190	0,318
CEL	0,391	0,588	0,821	X	0,391	0,165	0,292	0,728	0,391	0,611	0,391	0,561
ESP	0,574	0,750	0,906	0,574	X	0,361	0,465	0,849	0,574	0,768	0,574	0,729
GRA	0,574	0,750	0,906	0,574	0,574	X	0,465	0,849	0,574	0,768	0,574	0,729
MAL	0,574	0,750	0,906	0,574	0,574	0,296	X	0,849	0,574	0,768	0,574	0,729
RMA	0,221	0,386	0,670	0,221	0,221	0,081	0,154	X	0,221	0,410	0,221	0,361
RSO	0,425	0,622	0,841	0,425	0,425	0,187	0,322	0,755	X	0,644	0,425	0,596
SEV	0,574	0,750	0,906	0,574	0,574	0,296	0,465	0,849	0,574	X	0,574	0,729
VAL	0,358	0,553	0,800	0,358	0,358	0,148	0,264	0,699	0,358	0,577	X	0,526
VIL	0,248	0,424	0,703	0,248	0,248	0,094	0,176	0,579	0,248	0,448	0,248	X

Cuadro 2.10: Probabilidad de la victoria local en los partidos según el modelo 2.2 con pesos

$\beta_0 = -1,611$	Coeficiente local	Coeficiente visitante
Athletic	0,000	0,000
Atletico	-1,332	2,001
Barcelona	-0,782	1,9181
Celta	0,000	0,980
Espanyol	0,000	0,000
Granada	0,933	0,000
Malaga	0,000	-0,833
Real Madrid	-0,920	1,840
Real Sociedad	0,000	0,827
Sevilla	0,000	0,000
Valencia	0,000	0,747
Villarreal	0,000	0,749

Cuadro 2.11: Estimación de los parámetros β_i y β_i' del modelo 2.2 con pesos en el que buscamos calcular la probabilidad de la victoria visitante.

	ATH	ATM	BAR	CEL	ESP	GRA	MAL	RMA	RSO	SEV	VAL	VIL
ATH	X	0,050	0,084	0,166	0,166	0,337	0,166	0,074	0,166	0,166	0,166	0,166
ATM	0,596	X	0,403	0,596	0,596	0,790	0,596	0,371	0,596	0,596	0,596	0,596
BAR	0,576	0,264	X	0,576	0,576	0,776	0,576	0,351	0,576	0,576	0,576	0,576
CEL	0,347	0,123	0,196	X	0,347	0,575	0,347	0,175	0,347	0,347	0,347	0,347
ESP	0,166	0,050	0,084	0,166	X	0,337	0,166	0,074	0,166	0,166	0,166	0,166
GRA	0,166	0,050	0,084	0,166	0,166	X	0,166	0,074	0,166	0,166	0,166	0,166
MAL	0,080	0,022	0,038	0,080	0,080	0,181	X	0,033	0,080	0,080	0,080	0,080
RMA	0,557	0,249	0,365	0,557	0,557	0,762	0,557	X	0,557	0,557	0,557	0,557
RSO	0,313	0,108	0,173	0,313	0,313	0,537	0,313	0,154	X	0,313	0,313	0,313
SEV	0,166	0,050	0,084	0,166	0,166	0,337	0,166	0,074	0,166	X	0,166	0,166
VAL	0,297	0,100	0,162	0,297	0,297	0,517	0,297	0,144	0,297	0,297	X	0,297
VIL	0,297	0,100	0,162	0,297	0,297	0,518	0,297	0,144	0,297	0,297	0,297	X

Cuadro 2.12: Probabilidad de la victoria visitante en los partidos según el modelo 2.2 con pesos

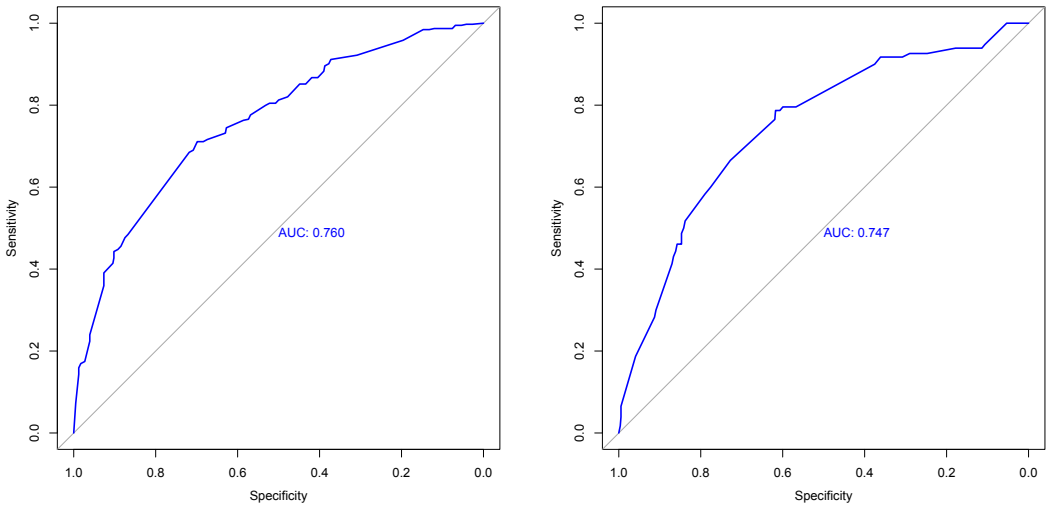


Figura 2.6: Curvas ROC del modelo 2.2 con pesos. A la izquierda cuando consideramos éxito la victoria local y a la derecha cuando el éxito es la victoria visitante

Capítulo 3

Aplicación de los resultados obtenidos en las apuestas

Ahora que tenemos dos modelos que podemos considerar aceptables a la hora de predecir resultados podemos aplicar nuestros modelos para conseguir beneficio económico. A través de las apuestas deportivas podemos encontrar una estrategia que nos asegure una esperanza positiva. El beneficio de una apuesta acertada será $k(c-1)$ siendo c el valor de la cuota y k la cantidad apostada. En caso de fallar la apuesta el beneficio es claramente $-k$. En este trabajo realizaremos una estrategia de apuesta fija, por lo que antes de empezar a apostar debemos decir cual será el valor de k. Calculamos la esperanza del beneficio (B):

$$\mathrm{E}(B) = Pk(c-1) - (1-P)k = k(cP - P - 1 + P) = k(cP-1)$$

Siendo P la probabilidad que le da nuestro modelo a que se de el suceso.

Evidentemente buscamos una esperanza positiva y como $k > 0$ buscamos que:

$$cP - 1 > 0 \Longrightarrow cP > 1 \Longrightarrow P > \frac{1}{c}$$

De esta manera sabemos que el beneficio económico dependerá de dos factores.

1. **La cantidad apostada**: Claramente a más dinero apostado obtendremos más beneficio. Esto es claro en cualquier tipo de apuesta, es decir a mas dinero invertido más dinero se gana, de todas maneras tenemos que tener cuidado en quedarnos si dinero con el que invertir en caso de tener una mala racha inicial por lo que hemos decidido que nuestra cantidad aposta será de 5 euros, teniendo un banco inicial de 100 euros lo que nos permitirá fallar las primeras 20 apuestas antes de quedarnos en bancarrota, situación que consideramos muy poco probable.

2. **La relación entre la cuota dada por la casa de apuestas y la probabilidad de que el suceso ocurra según nuestro modelo:** Como ya hemos visto cuando $P > \frac{1}{c}$ la apuesta es rentable, cuanto mayor sea la diferencia entre el producto entre la cuota y la probabilidad de nuestro modelo sea mayor nuestro beneficio a la larga será mayor. Podemos encontrar el caso en el que haya dos apuestas en un mismo partido que sean rentables, en ese caso cogeremos la apuesta que nos proporcione mayor valor del producto entre la cuota y la probabilidad del suceso.

También puede existir la opción que la casa de apuestas haya realizado unos cálculos de probabilidades similares a los nuestros. En caso de que ningún producto sea mayor que 1 optaremos por no realizar las apuestasEn cuanto a las cuotas, los datos que tenemos consta de una media de 16 casas de apuestas distintas, por lo que en caso de realizar esta simulación en la vida real los beneficios serían mayores ya que buscaríamos siempre apostar en la casa de apuestas con mayor cuota, lo que en caso de acertar nos daría más beneficio y en caso de fallar al estar apostando una cantidad fija, las perdidas serían las

mismas en cualquier casa de apuestas.

Comenzamos comprobando el modelo de regresión logística de la sección 2.2, recordar que este modelo nunca predice un empate. Aún así podría considerar como apuesta rentable ambas opciones por lo que, como hemos dicho antes, tendremos que escoger el mayor valor entre el producto de la probabilidad y a cuota. Con todas esta estrategia, los resultados obtenidos se muestran en la figura 3.1.

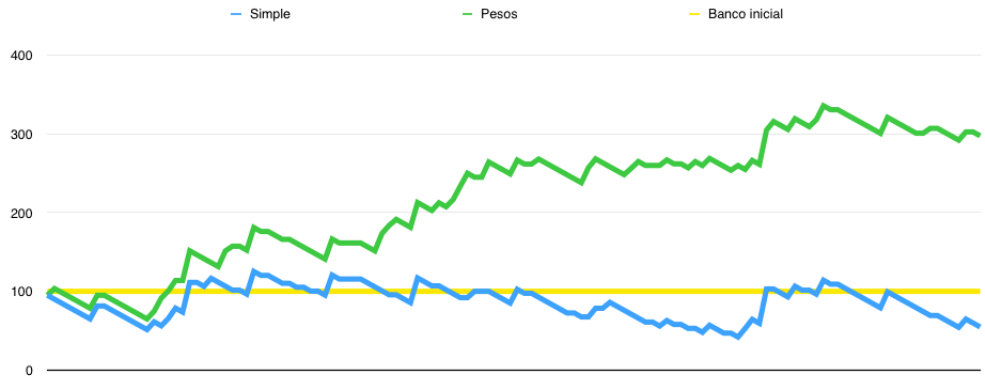


Figura 3.1: Comparación de la estrategia de apuestas utilizando el modelo 2.1 simple y con pesos durante la temporada 2016/2017

Como podemos ver solo una de las dos gráficas está por encima del presupuesto inicial al acabar la temporada, lo cual quiere decir que obtenemos beneficios durante la temporada únicamente si aplicamos la mejora de los pesos. Como datos interesantes comentar que normalmente las apuestas no son al equipo favorito o a cuotas bajas, ya que normalmente el equipo favorito está sobrevalorado según las casas de apuestas. Las cuota más alta acertada ha sido de $8,71$ en ambos casos. En lo que respecta a apuestas acertadas y falladas, el porcentaje se distribuye de la siguiente manera, se aciertan el $20,95\,\%$ de las apuestas realizadas en el caso del modelo simple y el porcentaje sube hasta el $33,91\,\%$ en el caso del modelo con pesos. Ademas el resultado final no es el punto en el que más dinero hemos tenido llegando a tener en un punto de esta temporada llegando a 336 en el caso del modelo con pesos y a 125 en el modelo simple. Por contra, cerca del principio de temporada llegamos a nuestro mínimo en el modelo de pesos que fue de 65 y cerca del final el modelo simple alcanzó mínimos de 42.

Realizamos ahora la misma gráfica pero con el modelo de la sección 2.1, en este modelo podremos realizar apuestas al empate, cosa que suele suceder 1 vez por temporada. Los resultados obtenidos se representan en la figura 3.2.

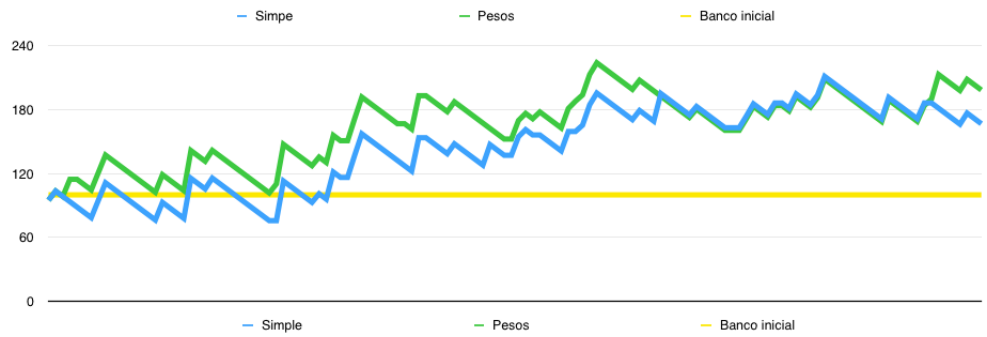


Figura 3.2: Comparación de la estrategia de apuestas utilizando el modelo 2.1 simple y con pesos durante la temporada 2016/2017

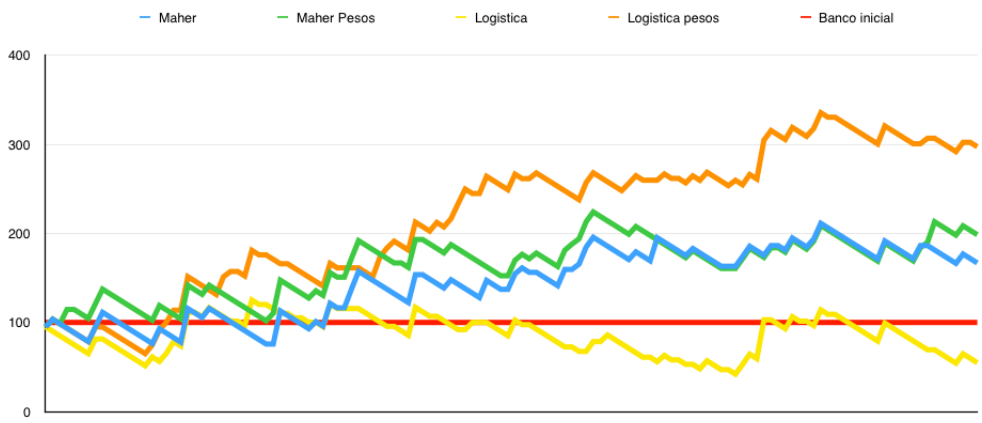


Figura 3.3: Comparación de la estrategia de apuestas utilizando los modelos 2.1 y 2.2 tanto simple como con pesos durante la temporada 2016/2017

Podemos apreciar que el beneficio es menor que en el modelo anterior, pero al igual que antes el modelo de pesos ayuda a la mejora de las predicciones. Esto nos hace proponer otra variación de nuestros modelos, que seria ir añadiendo las jornadas de la temporada en la que apostamos a nuestra muestra para mejorar la precisión de nuestro modelo con peso $n+1$ siendo n el peso de la úlitma temporada registrada en nuestra muestra. Aunque el modelo de regresión logística aporta más beneficios el modelo de Maher aporta unos datos bastante aceptables llegando a acertar una cuota de $7,53$ en ambos modelos y acertando un porcentaje de $26,23\,\%$ del total de las apuestas realizadas en el caso del modelo simple y un $29,03\,\%$ en el caso del modelo con pesos. El punto en el que menos dinero que llegamos a tener es 76 en el caso del modelo simple y un sorprendente 95 en el caso del modelo con pesos, lo que quiere decir que prácticamente toda la temporada hemos estado con beneficios, por contra el punto en el que nuestro beneficio es máximo este es 224 en el caso del modelo con pesos y de 211 en el simple. Al acabar la temporada nuestro banco es de 199 en el caso del modelo con pesos, o lo que es lo mismo, prácticamente el doble del dinero con el que empezamos y 167 en el caso del modelo simple.

En la figura 3.3 vemos los cuatro modelos unidos. Como podemos ver más o menos pasado el ecuador de la temporada tenemos unas jornadas en las que nuestros resultados son un poco anómalos, esto puede ser debido a las competiciones europeas que tienen los equipos importantes durante este periodo, en el caso del modelo de regresión logística con pesos en dicho periodo se mantiene estable, mientras que el resto de modelos comienzan a perder dinero. Una vez pasado este periodo los modelos se recuperan volviendo a bajar en las ultimas jornadas en las que hay equipos que no se juegan nada en dichos partidos y por tanto también se registran datos anómalos. Esto es muy difícil de medir objetivamente pero las casas de apuestas lo tienen cuenta. Veamos lo que sucede con un ejemplo, si el equipo A según nuestro modelo es muy superior al equipo B, pero el equipo B depende de ese partido para mantener la categoría y el equipo al equipo A no le influye el resultado en sus aspiraciones deportivas, la casa de apuestas dará como favorito al equipo B y lo lógico es que gane dicho equipo, nuestro modelo considerará rentable la cuota del equipo A y tendremos altas probabilidades de perder dicha apuesta.
Como conclusión final del trabajo podemos sacar que hemos sido capaces de sacar un beneficio del $296\,\%$ gracias a nuestras predicciones durante una única temporada y apostando una cantidad fija del $5\,\%$ de nuestro banco inicial. Para aumentar el beneficio aún más se proponen dos opciones:

- Aumentar el porcentaje de nuestro banco inicial a apostar, asumiendo el riesgo que con forme más aumentemos dicho porcentaje más probable será llegar a la bancarrota. No se recomienda superar el $20\,\%$.
- Mejorar la predicción de nuestros modelos, actualizando nuestra muestra con los partidos de las jornadas en las que hemos apostado con su respectivo peso que será $n+1$ siendo n el numero de

temporadas que tenemos en la muestra, de esta manera nuestro modelo será más preciso y nos permitirá acertar más apuestas y como consecuencia ganar más dinero.

Como hemos visto, a la hora de realizar apuestas los modelos con pesos han resultado mejores que los modelos simples. Al conseguir beneficios podríamos llegar a pensar que hemos encontrados unos modelos que son mejor que los que utilizan las casas de apuestas, esto no tiene por que ser cierto del todo, para ver si nuestros modelos de pesos son mejores que los de las casas de apuestas compararemos los resultados de esta temporada en la siguiente tabla.

PREDICCION RESULTADO	VICTORIA LOCAL	EMPATE	VICTORIA VISITANTE	
VICTORIA LOCAL	53	25	16	MODELO
	52	65	53	CASAS DE APUESTAS
EMPATE	0	0	1	MODELO
	0	0	0	CASAS DE APUESTAS
VICTORIA VISITANTE	9	5	24	MODELO
	10	6	28	CASAS DE APUESTAS

Cuadro 3.1: Tabla que enfrenta el resultado más posible dado por nuestro modelo o la casa de apuesta a lo observado a posteriormente.

Como podemos ver nuestros modelos son bastante similares a la hora de aciertos y errores respecto a las casas de apuestas. Lo que hace que nuestro modelo nos permita ganar dinero es que encuentra los partidos en los que algún resultado (aunque sea poco probable) no esta correctamente ajustado y esto unido a que realizamos una gran cantidad de apuestas durante la temporada nos da un beneficio económico. Por lo que pese a no tener un modelo mejor somos capaces de sacarle rentabilidad.

3.1. Conclusiones finales

- Los modelos estadísticos son útiles para predecir los resultados de fútbol en la primera división del fútbol español.
- La incorporación de pesos en los modelos, dando un importancia mayor a los partidos recientes, mejora los resultados.
- Aunque nuestro modelos no presentan una gran mejora global respecto a las predicciones de las casas de apuestas estos son suficientemente buenos para obtener beneficio económico apostando en aquellos partidos en los que detectemos que las casas de apuestas han hecho una mala predicción.

Apendice: Codigo R

3.2. Obtencion de estimadores del modelo de Maher

```
data1<-read.csv("/Users/ayoseiturralde/Desktop/TFG/base de datos/Hoja 1-Resultados de
    La Liga desde la temporada 2013-2014.csv", header=T,sep=";")%lectura de la tabla
lm.loss<- function(par) {
a.3 <- par[1] %declaracion de variables
a.4 <- par[2]
a.5 <- par[3]
a.6 <- par[4]
a.7 <- par[5]
a.8 <- par[6]
a.9 <- par[7]
a.10 <- par[8]
a.11 <- par[9]
a.12 <- par[10]
a.13 <- par[11]
a.14 <- par[12]
a.15 <- par[13]
a.17 <- par[14]
a.18 <- par[15]
a.19 <- par[16]
a.20 <- par[17]
a.21 <- par[18]
a.22 <- par[19]
a.23 <- par[20]
a.24 <- par[21]
a.25 <- par[22]
a.26 <- par[23]
a.27 <- par[24]
b.1 <- par[25]
b.2 <- par[26]
b.3 <- par[27]
b.4 <- par[28]
b.5 <- par[29]
b.6 <- par[30]
b.7 <- par[31]
b.8 <- par[32]
b.9 <- par[33]
b.10 <- par[34]
b.11 <- par[35]
b.12 <- par[36]
b.13 <- par[37]
b.14 <- par[38]
b.15 <- par[39]
b.17 <- par[40]
```

```
b.18 <- par[41]
b.19 <- par[42]
b.20 <- par[43]
b.21 <- par[44]
b.22 <- par[45]
b.23 <- par[46]
b.24 <- par[47]
b.25 <- par[48]
b.26 <- par[49]
b.27 <- par[50]
g <- par[51]
log.likelihoods <- -(1-(a.3+a.4+a.5+a.6+a.7+a.8+a.9+a.10+a.11+a.12+a.13+a.14+a.15+a.17
+a.18+a.19+a.20+a.21+a.22+a.23+a.24+a.25+a.26+a.27))^data1[,8]
*a.3^data1[,9]*a.4^data1[,10]*a.5^data1[,11]*a.6^data1[,12]
*a.7^data1[,13]*a.8^data1[,14]*a.9^data1[,15]*a.10^data1[,16]
*a.11^data1[,17]*a.12^data1[,18]*a.13^data1[,19]*a.14^data1[,20]
*a.15^data1[,21]*a.17^data1[,23]*a.18^data1[,24]*a.19^data1[,25]
*a.20^data1[,26]*a.21^data1[,27]*a.22^data1[,28]*a.23^data1[,29]
*a.24^data1[,30]*a.25^data1[,31]*a.26^data1[,32]*a.27^data1[,33]
*b.2^data1[,35]*b.3^data1[,36]*b.4^data1[,37]*b.5^data1[,38]*b.6^data1[,39]
*b.7^data1[,40]*b.8^data1[,41]*b.9^data1[,42]*b.10^data1[,43]
*b.11^data1[,44]*b.12^data1[,45]*b.13^data1[,46]*b.14^data1[,47]
*b.15^data1[,48]*b.17^data1[,50]*b.18^data1[,51]*b.19^data1[,52]
*b.20^data1[,53]*b.21^data1[,54]*b.22^data1[,55]*b.23^data1[,56]
*b.24^data1[,57]*b.25^data1[,58]*b.26^data1[,59]*b.27^data1[,60]*g

+data1[,5]*(data1[,8]*log(1-(a.3+a.4+a.5+a.6+a.7+a.8+a.9+a.10+a.11+a.12
+a.13+a.14+a.15+a.17+a.18+a.19+a.20+a.21+a.22+a.23+a.24+a.25+a.26+a.27))
+data1[,9]*log(a.3)+data1[,10]*log(a.4)+data1[,11]*log(a.5)
+data1[,12]*log(a.6)+data1[,13]*log(a.7)+data1[,14]*log(a.8)+
data1[,15]*log(a.9)+data1[,16]*log(a.10)+data1[,17]*log(a.11)+
data1[,18]*log(a.12)+data1[,19]*log(a.13)+data1[,20]*log(a.14)+
data1[,21]*log(a.15)+data1[,23]*log(a.17)+data1[,24]*log(a.18)+
data1[,25]*log(a.19)+data1[,26]*log(a.20)+data1[,27]*log(a.21)+
data1[,28]*log(a.22)+data1[,29]*log(a.23)+data1[,30]*log(a.24)+
data1[,31]*log(a.25)+data1[,32]*log(a.26)+data1[,33]*log(a.27)+
data1[,35]*log(b.2)+data1[,36]*log(b.3)+data1[,37]*log(b.4)+data1[,38]*log(b.5)
+data1[,39]*log(b.6)+data1[,40]*log(b.7)+data1[,41]*log(b.8)+data1[,42]*log(b.9)
+data1[,43]*log(b.10)+data1[,44]*log(b.11)+data1[,45]*log(b.12)+data1[,46]
*log(b.13)+data1[,47]*log(b.14)+data1[,48]*log(b.15)+data1[,50]*log(b.17)
+data1[,51]*log(b.18)+data1[,52]*log(b.19)+data1[,53]*log(b.20)
+data1[,54]*log(b.21)+data1[,55]*log(b.22)+data1[,56]*log(b.23)+
data1[,57]*log(b.24)+data1[,58]*log(b.25)+data1[,59]*log(b.26)+
data1[,60]*log(b.27)+log(g))

-(1-(a.3+a.4+a.5+a.6+a.7+a.8+a.9+a.10+a.11+a.12+a.13+a.14+a.15+a.17
+a.18+a.19+a.20+a.21+a.22+a.23+a.24+a.25+a.26+a.27))^data1[,35]*a.3^data1[,36]
*a.4^data1[,37]*a.5^data1[,38]*a.6^data1[,39]*a.7^data1[,40]*a.8^data1[,41]
*a.9^data1[,42]*a.10^data1[,43]*a.11^data1[,44]*a.12^data1[,45]*a.13^data1[,46]
*a.14^data1[,47]*a.15^data1[,48]*a.17^data1[,50]*a.18^data1[,51]*a.19^data1[,52]
*a.20^data1[,53]*a.21^data1[,54]*a.22^data1[,55]*a.23^data1[,56]*a.24^data1[,57]
*a.25^data1[,58]*a.26^data1[,59]*a.27^data1[,60]*b.2^data1[,8]*b.3^data1[,9]
*b.4^data1[,10]*b.5^data1[,11]*b.6^data1[,12]*b.7^data1[,13]*b.8^data1[,14]
*b.9^data1[,15]*b.10^data1[,16]*b.11^data1[,17]*b.12^data1[,18]*b.13^data1[,19]
*b.14^data1[,20]*b.15^data1[,21]*b.17^data1[,23]*b.18^data1[,24]*b.19^data1[,25]
*b.20^data1[,26]*b.21^data1[,27]*b.22^data1[,28]*b.23^data1[,29]*b.24^data1[,30]
*b.25^data1[,31]*b.26^data1[,32]*b.27^data1[,33]
```

```
+data1[,6]*(data1[,35](1-(a.3+a.4+a.5+a.6+a.7+a.8+a.9+a.10+a.11+a.12+a.13+a.14
+a.15+a.17+a.18+a.19+a.20+a.21+a.22+a.23+a.24+a.25+a.26+a.27))+data1[,36]*log(a.3)
+data1[,37]*log(a.4)+data1[,38]*log(a.5)+data1[,39]*log(a.6)+data1[,40]*log(a.7)
+data1[,41]*log(a.8)+data1[,42]*log(a.9)+data1[,43]*log(a.10)+data1[,44]*log(a.11)
+data1[,45]*log(a.12)+data1[,46]*log(a.13)+data1[,47]*log(a.14)
+data1[,48]*log(a.15)+data1[,50]*log(a.17)+data1[,51]*log(a.18)
+data1[,52]*log(a.19)+data1[,53]*log(a.20)+data1[,54]*log(a.21
)+data1[,55]*log(a.22)+data1[,56]*log(a.23)+data1[,57]*log(a.24)
+data1[,58]*log(a.25)+data1[,59]*log(a.26)+data1[,60]*log(a.27)+data1[,8]*log(b.2)
+data1[,9]*log(b.3)+data1[,10]*log(b.4)+data1[,11]*log(b.5)+data1[,12]*log(b.6)
+data1[,13]*log(b.7)+data1[,14]*log(b.8)+data1[,15]*log(b.9)+data1[,16]*log(b.10)
+data1[,17]*log(b.11)+data1[,18]*log(b.12)+data1[,19]*log(b.13)+
data1[,20]*log(b.14)+data1[,21]*log(b.15)+data1[,23]*log(b.17)+
data1[,24]*log(b.18)+data1[,25]*log(b.19)+data1[,26]*log(b.20)+data1[,27]*log(b.21)
+data1[,28]*log(b.22)+data1[,29]*log(b.23)+data1[,30]*log(b.24)+
data1[,31]*log(b.25)+data1[,32]*log(b.26)+data1[,33]*log(b.27)) %funcion
deviance <- -2 * sum(log.likelihoods)
return(log.likelihoods)
}
parameter.fits <- optim(par = v,
fn = lm.loss, hessian = T
)
%minimiza la funcion
parameter.fits
```

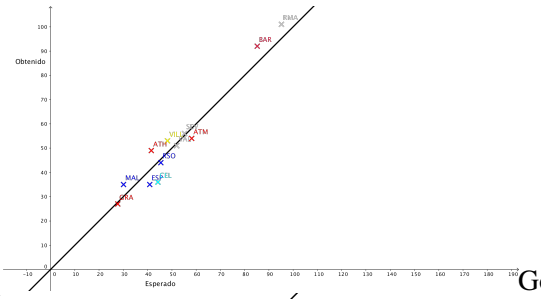
3.3. Definir el modelo de regresion logistica

```
regresion<-glm(formula = data1[, 63] ~ data1[, 10] + data1[, 11] + data1[,
13] + data1[, 18] + data1[, 20] + data1[, 24] + data1[, 27] +
data1[, 28] + data1[, 29] + data1[, 31] + data1[, 33] + data1[,
37] + data1[, 38] + data1[, 40] + data1[, 45] + data1[, 47] +
data1[, 51] + data1[, 54] + data1[, 55] + data1[, 56] + data1[,
58] + data1[, 60], family = binomial(logit))

stepAIC(regresion)
```

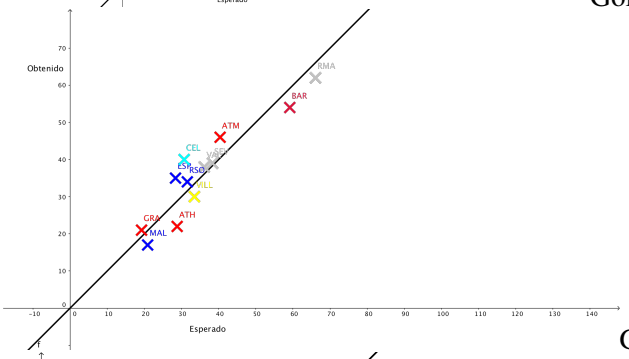
En la que la salida de R en el Ãºlimo paso es la siguiente:

```
Call:  glm(formula = data1[, 61] ~ data1[, 10] + data1[, 11] + data1[,
20] + data1[, 24] + data1[, 27] + data1[, 29] + data1[, 33] +
data1[, 37] + data1[, 38] + data1[, 40] + data1[, 54] + data1[,
55] + data1[, 58] + data1[, 60], family = binomial(logit))

Coefficients:
(Intercept) data1[, 10] data1[, 11] data1[, 20] data1[, 24]
0.2996      0.7997      1.9698      -1.1646     -0.4401
data1[, 27] data1[, 29] data1[, 33] data1[, 37] data1[, 38]
1.4272      0.8979      0.6910      -1.7258     -1.7518
data1[, 40] data1[, 54] data1[, 55] data1[, 58] data1[, 60]
-0.7432     -1.5611     -0.6026     -0.8843     -1.4060

Degrees of Freedom: 791 Total (i.e. Null); 777 Residual
Null Deviance:  1097
Residual Deviance: 918.2  AIC: 948.2
```

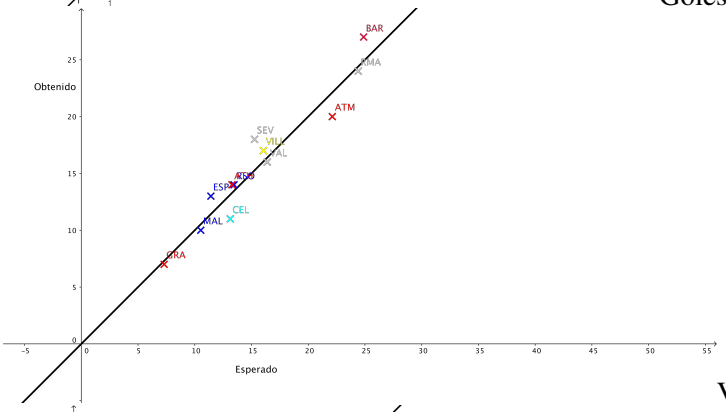
3.4. Graficas ampliadas de las figuras 2.1, 2.2, 2.3 y 2.4



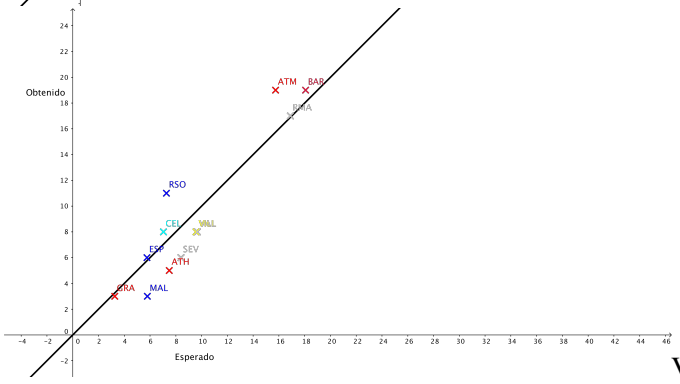
Goles de local



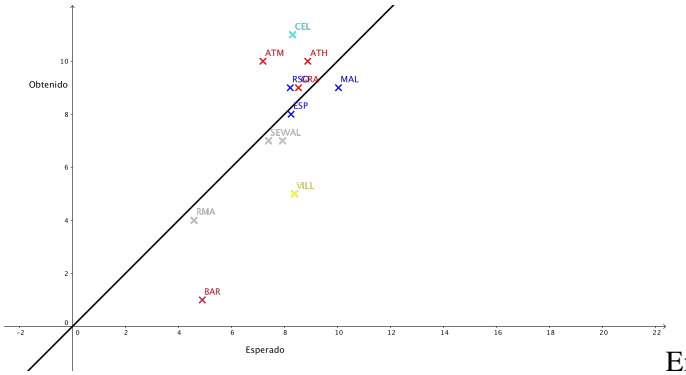
Goles de visitante



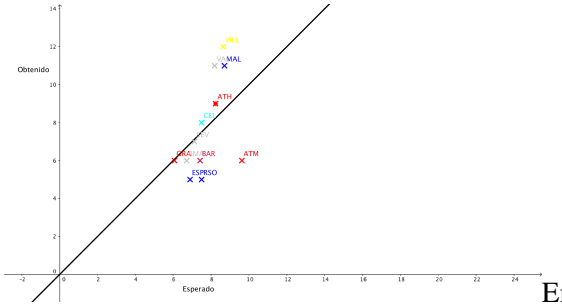
Victoria de local



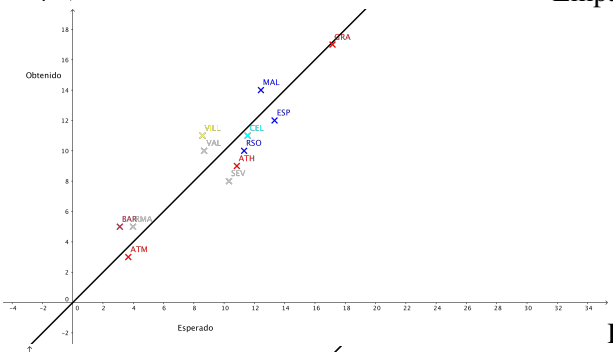
Victoria de visitante



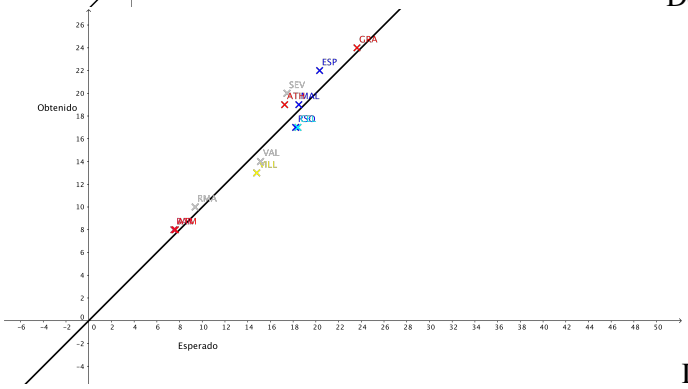
Empate de local



Empate de visitante



Derrota de local



Derrota de visitante

Bibliografía

[1] JOBSON, JOHN, *Applied multivariate data analysis: Volume I: Regression and Experimental Desing*, Springer Science & Business media, 2012.

[2] I. CHINWE Y N. ENOK, An Improved Prediction Systemfor Football a Match Result , *IOSR Journal of Engineering* **12** (4) (2014), 12–20.

[3] M. MAHER, Modelling association football scores, *Statica Neerlandica* **36.3** (1982), 109–118.

[4] M. DIXON Y S. COLES, Modelling association football scores ans inefficiencies in betting market, *Journal of the Royal Statcal Society: Series C(Applied Statistics)* **46.2** (1997), 265–280.

[5] X. ROBIN, N. TURCK, A. HAINARD, N. TIBERTI, F. LISACEK, JC. SANCHEZ Y M. MÜLLER*pROC:: an open-source package for R and S+ to analyze and compare ROC curves*(2011),BMC Bioinformatics

[6] W. N. VENABLES AND B. D. RIPLEY*MASS: Modern Applied Statistics with S* (2002),http://www.stats.ox.ac.uk/pub/MASS4