Quant Analyst Test - Smartodds

Matthew Williams

March 1, 2021

1.1 and 1.2

The data handling can be seen in the R code.

2.1

I calculated 27.3% of the games in the fitset to be between teams in different conferences.

2.2

For this question I compared the home goal differences (home goals − away goals) for all games in each conference. I found that the mean home goal difference for the eastern conference was 0.523, and for the western conference it was 0.493. I did a t-test to assess whether this was statistically significant. The null hypothesis was that the true difference in means is equal to 0. The alternative hypothesis was that the true difference in means is not equal to 0. I obtained a p-value of 0.765, therefore I failed to reject the null hypothesis. The difference in the means was not seen to be statistically significant.

2.3

A moving average plot for total goals over 20 games can be seen in Figure 1. The plot does appear to be very random, however the highest point in each season is interesting. These points all occur at approximately the same point in the season, approaching the end of the regular season. The increase in goals during this time could be due to teams being more offensive as they try to achieve a good place in the standings.

3.1

I would interpret the parameters in the following way, along with constraints:

$$0 \leq \alpha_i \leq 1 \quad \text{(attacking strength of team i)}$$

$$-1 \leq \beta_i \leq 0 \quad \text{(defensive strength of team i)}$$

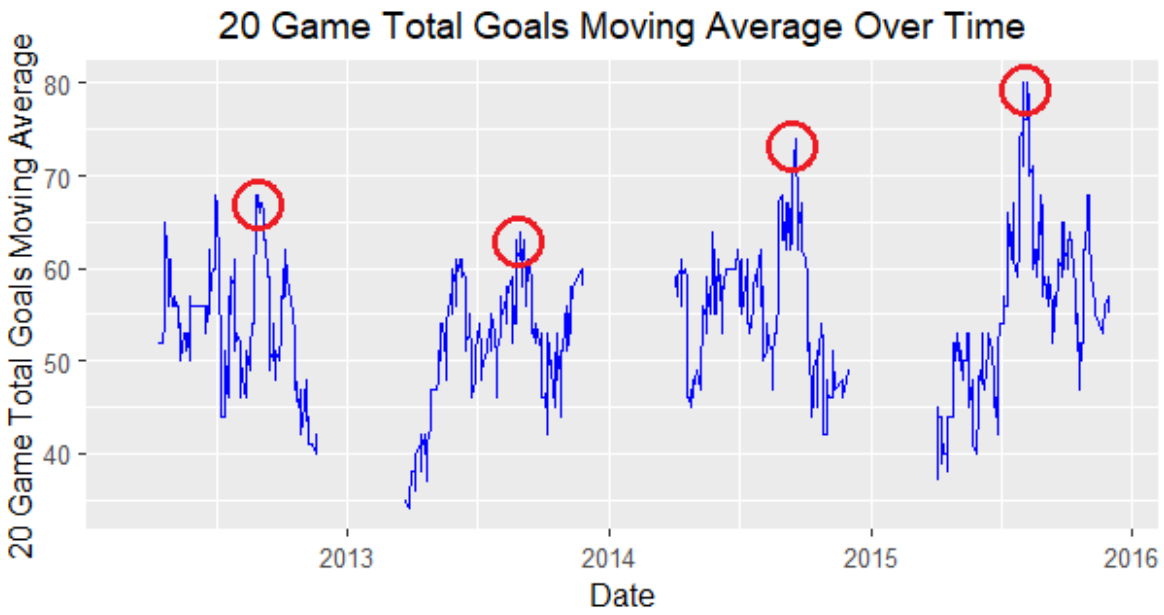

Figure 1: Total goals moving average over 20 matches across time for the fit data set, using only games between teams in the same conference. The peak of the moving average is circled for each season.

$$0 \leq \gamma \leq 1 \quad \text{(constant parameter)}$$
$$0 \leq \eta \leq 1 \quad \text{(home ground advantage parameter)}$$

It makes sense for defensive strengths to be negative in the model, so that as defensive strength increases it makes the scoring rate for the opposition team decrease. The constant parameter in the model allows for a shared random effect in the matches. The home ground advantage parameter takes into account the fact that teams often perform better when playing at home compared to away.

3.2

The likelihood of a given model is:

$$\begin{aligned} \mathrm{L} &= \mathrm{L}(\lambda_1, \lambda_2, ..., \lambda_n, \mu_1, \mu_2, ..., \mu_n, x_1, x_2, ..., x_n, y_1, y_2, ..., y_n) \\ &= \prod_{k=1}^{n} \left(e^{-\lambda_k} \frac{1}{x_k!} \lambda_k^{x_k} \right) \left(e^{-\mu_k} \frac{1}{y_k!} \mu_k^{y_k} \right) \end{aligned}$$

and the log likelihood is:

$$\log(\mathrm{L}) = \sum_{k=1}^{n} -\lambda_k - \log(x_k!) + x_k \log(\lambda_k) - \mu_k - \log(y_k!) + y_k \log(\mu_k)$$

where:

$$\lambda_k = e^{\alpha_{i(k)} + \beta_{j(k)} + \gamma + \frac{\eta}{2}}$$

$$\mu_k = e^{\alpha_{j(k)} + \beta_{i(k)} + \gamma - \frac{\eta}{2}}$$

I obtained the optimal α_i, β_i, γ and η parameters by maximum likelihood. This was done my maximising the log-likelihood. The fitted α_i and $-\beta_i$ parameters for each team, along with their ranks can be seen in Table 1. The negative β_i values are used, since a higher $-\beta_i$ value is better defensively for a team. It can be seen from the table that Los Angeles Galaxy have the best attacking strength parameter, along with the fourth highest defensive strength. Sporting Kansas City have the best defensive strength parameter value, however they only have the 16th highest attacking strength rating. It is worth noting that in this table there are 21 teams, whilst in the simulation set there are only 20 teams. This is because Chivas USA disbanded in 2014.

The fitted γ parameter was 0.330 for a shared random effect. The fitted η parameter was 0.382. This means that a teams' goals rate parameter will be $e^{0.382} = 1.465$ times higher if they play a given team at home compared to away. Also, the goals rate parameter for the opposition will be $e^{-0.382} = 0.682$ times the value it would be if they played that team away.

3.3

Time dynamics could be handled in the model by modifying the likelihood function to be the pseudo-likelihood function:

$$\begin{aligned} \mathrm{L}_t &= \mathrm{L}(\lambda_1, \lambda_2, ..., \lambda_n, \mu_1, \mu_2, ..., \mu_n, x_1, x_2, ..., x_n, y_1, y_2, ..., y_n) \\ &= \prod_{k=1}^{n} \left[\left(e^{-\lambda_k} \frac{1}{x_k!} \lambda_k^{x_k} \right) \left(e^{-\mu_k} \frac{1}{y_k!} \mu_k^{y_k} \right) \right]^{\phi(t-t_k)} \end{aligned}$$

whereby t is the time point at which the likelihood is being calculated and t_k is the time that game k was played. The function ϕ could be implemented as:

$$\phi(t) = \exp(-\xi t)$$

with $\xi > 0$. By doing this, observations in the past are exponentially down weighted. This then takes into account that the form of teams change over time, meaning that more emphasis should be given to recent results when fitting the model parameters.

i	Team	α_i	α_i Rank	$-\beta_i$	$-\beta_i$ Rank
1	Chicago Fire	0.681	10	0.623	16
2	Chivas USA	0.244	21	0.556	19
3	Colorado Rapids	0.641	13	0.741	11
4	Columbus Crew	0.729	9	0.644	15
5	DC United	0.532	19	0.734	12
6	FC Dallas	0.830	5	0.789	7
7	Houston Dynamo	0.631	14	0.732	13
8	Los Angeles Galaxy	0.998	1	0.908	4
9	Montreal Impact	0.631	15	0.609	17
10	New England Revolution	0.643	12	0.759	9
11	New York City	0.790	6	0.484	20
12	New York Red Bulls	0.893	2	0.750	10
13	Orlando City	0.578	17	0.363	21
14	Philadelphia Union	0.518	20	0.696	14
15	Portland Timbers	0.859	4	0.769	8
16	Real Salt Lake	0.733	8	0.984	2
17	San Jose Earthquakes	0.760	7	0.907	5
18	Seattle Sounders	0.861	3	0.929	3
19	Sporting Kansas City	0.607	16	0.996	1
20	Toronto FC	0.548	18	0.567	18
21	Vancouver Whitecaps	0.663	11	0.892	6

Table 1: All fitted α_i and $-\beta_i$ parameters for each team, along with their ranks.

4.1

Here are four suggestions for extensions/improvements to the basic model that could be carried out without acquiring additional data:

1. A separate home ground advantage parameter could be used for each team.

2. The γ parameter could be replaced by a correlation variable that has special cases for low scoring games, whereby results are believed to be more closely correlated.

3. A conjugate method could be used, whereby the strength parameters, or rather the exponentials of them, have Gamma prior distributions. This would allow model parameters to be updated quickly after a new observation is made, rather than re-optimising them again.

4. Since the dates of all the matches are available, an extra parameter could be added based on the time since the previous game. Teams may not perform as well if they've not had much rest since their last match.

4.2

I think that the most interesting additional data to consider would be the players that played in each game. If key players could be identified for each team, whereby the chances of their team winning are higher with them in the team, then predictions could be tailored based on the likelihood of these players being available. Expected goals would also be interesting to consider. The model parameters could be based on these values instead of actual observed goals. Expected goals aren't as random as actual goals, and less influenced by luck or 'flukes'. That is because these values are based on numerous chances created in each game. Other factors that could be interesting to consider are the weather and motivation for each side, such as if it is a local derby or if they have another big game coming up.

4.3

I would use a scoring method to compare different models to decide which one was best. These include the Brier Score, Log Score and Ranked Probability Score. These scoring methods show how well the model predicts data out of sample, against actual observed values. I would choose the Ranked Probability Score since it allows for ordinal data. This can take into account that a draw is closer to a home win in football than an away win is. Also, when assessing total goals for a model, it can take into account that a prediction of 3 goals is closer to 2 goals than 10 goals for instance.

The equation to calculate Ranked Probability Score is

$$\text{RPS} = \frac{1}{2n}\sum_{t=1}^{n}\sum_{k=1}^{2}(\sum_{j=1}^{k}(z_{j,t} - P_{j,t}))^2$$

The notation $z_{j,t}$ represents whether or not the result for game t was j. The possible values for j are $[1, 0, 0]$ (Home win), $[0, 1, 0]$ (Draw) or $[0, 0, 1]$ (Away win). If $z_{j,t}$ is 1, it means that the result was j, or if it is 0 it means that this was not the result. $P_{j,t}$ is the model's probability for the result of game t to be j.

5.1 and 5.2

The scoring rates λ and μ for each game in the simulation set can be seen in the R code. The Skellam distribution was used to find the probabilities. This gives the probability distribution of the difference in observations between two independent Poisson random variables.

5.3

The expected tables after the 2016 regular season were calculated using expected points totals from the fitted model. Table 2 shows the expected overall table. Table 3 shows the expected eastern table. Table 4 shows the expected western table.

Position	Conference	Team	Expected Points
1	West	Los Angeles Galaxy	53.06
2	East	New York Red Bulls	51.90
3	West	Seattle Sounders	50.83
4	West	Real Salt Lake	49.55
5	West	San Jose Earthquakes	48.82
6	West	Portland Timbers	48.56
7	West	FC Dallas	48.45
8	West	Sporting Kansas City	47.64
9	East	New England Revolution	47.46
10	West	Vancouver Whitecaps	47.16
11	East	Columbus Crew	47.01
12	East	Chicago Fire	45.89
13	East	New York City	45.86
14	East	DC United	45.08
15	East	Montreal Impact	44.69
16	West	Colorado Rapids	44.37
17	East	Philadelphia Union	44.16
18	West	Houston Dynamo	43.98
19	East	Toronto FC	43.00
20	East	Orlando City	40.73

Table 2: Expected overall table based on expected points for each team on October 23rd 2016.

5.4

I have calculated the Ranked Probability Scores for the fitted model to be 0.208 for full time results and 0.1006 for total goals. I also calculated the ranked probability scores for model A to be 0.201 for full time results and 0.1005 for total goals. The scores for model A are slightly lower in both cases, meaning a better performance than the fitted model against the simulation set.

6.1

A method for simulating the games in the simulation set can be seen in the R code.

6.2

I calculated the probability for LA Galaxy to finish in the top 2 in the Western Conference for the 2016 regular season to be 63.3%. This was done using 1e4 simulations for the matches in the simulation set.

Position	Team	Expected Points
1	New York Red Bulls	51.90
2	New England Revolution	47.46
3	Columbus Crew	47.01
4	Chicago Fire	45.89
5	New York City	45.86
6	DC United	45.08
7	Montreal Impact	44.69
8	Philadelphia Union	44.16
9	Toronto FC	43.00
10	Orlando City	40.73

Table 3: Expected eastern table based on expected points for each team on October 23rd 2016.

Position	Team	Expected Points
1	Los Angeles Galaxy	53.06
2	Seattle Sounders	50.83
3	Real Salt Lake	49.55
4	San Jose Earthquakes	48.82
5	Portland Timbers	48.56
6	FC Dallas	48.45
7	Sporting Kansas City	47.64
8	Vancouver Whitecaps	47.16
9	Colorado Rapids	44.37
10	Houston Dynamo	43.98

Table 4: Expected western table based on expected points for each team on October 23rd 2016.

6.3

I calculated the number of simulations needed for the 95% confidence interval for the probability estimate in the previous question to have a Monte Carlo error of 0.1% to be 892,004.

This was calculated using the formula:

$$N = \frac{(1-\hat{p})\hat{p}\left(\phi^{-1}(\frac{1+\gamma}{2})\right)^2}{\epsilon^2}$$

where $\phi^{-1}(x)$ is the quantile function of the standard normal distribution, $\gamma = 0.95$ is the confidence level and $\epsilon = 0.001$ is the Monte Carlo error. $\hat{p}$ is 63.3% as calculated in the previous question.