Quant Analyst Test

Smartodds

The purpose of this test is to give you an opportunity to showcase your approach in manipulating and analyzing data and in implementing, evaluating and using a predictive model. We are interested in your reasoning, the quality and documentation of your code and the clarity with which your model and findings are presented.

What we would like to receive from you is code that we can inspect and run (preferably in R, Python or a similar language), and some form of report of your findings (preferably pdf or html).

1 The data

1.1

Football results are freely available on http://www.football-data.co.uk.

Using only MLS data: Prepare a dataset containing the results of MLS games since the start of the 2012 season. You can find these here: http://www.football-data.co.uk/usa.php. The aim is to fit a model to this dataset up to December 6th 2015 (**the fitset**) in order to make predictions for games after December 6th 2015 and up to October 23rd 2016 (**the simulation set**)

In addition to basic game information, other fields are provided such as odds from various bookmakers. The meanings of the various column names are described in http://www.football-data.co.uk/notes.txt.

1.2

The conference to which each team belongs is currently missing from this dataset.

Add two columns called "home_conference" and "away_conference" with the respective team conferences ("East" or "West"). Be careful as two teams changed conference in this period: Houston Dynamo and Sporting Kansas City are in Eastern conference until the end of the 2014 season but play in Western conference since 2015.

2 Descriptive statistics

In this section, use all games for the 1st question but, for the rest, only use games from the fitset between teams in the same conference.

2.1

What is the percentage of games in the fitset where teams are in different conferences?

2.2

Is the home advantage for the Eastern Conference different from the Western Conference? Is it statistically significant?

2.3

Can you spot any seasonality in the total goals scored (e.g more goals during a certain period of the year)? What could cause the pattern you observe?

3 A model

Let the games be labelled by $k = 1,\ldots,n$. A simple model that could be used to characterise the results of a football match would assume that goals (X_k, Y_k) scored in game k are independent and follow a Poisson distribution:

$$X_k \sim \mathcal{P}ois(\lambda_k) \tag{1}$$
$$Y_k \sim \mathcal{P}ois(\mu_k) \tag{2}$$

with

$$\ln \lambda_k = \alpha_{i(k)} + \beta_{j(k)} + \gamma + \eta/2 \tag{3}$$
$$\ln \mu_k = \alpha_{j(k)} + \beta_{i(k)} + \gamma - \eta/2 \tag{4}$$

where $i(k) \in \{1, ..., n_{teams}\}$ is an index referring to home team in game k, $j(k) \in \{1, ..., n_{teams}\}$ is the corresponding index for away team, λ_k being the home scoring rate for game k and μ_k the away team scoring rate for game k. Please note that team is being treated as a factor variable, so e.g. $(\alpha_1, ...\alpha_{n_{teams}})$ is a vector of coefficients to be estimated, one per team.

3.1

How would you interpret the parameters α_i, β_i, γ and η? Consider what additional constraints or assumptions may be required to make this model identifiable.

3.2

Assuming games are independent of each other, fit this model by maximum likelihood on the fitset and analyse the parameters you obtained.

3.3

How could you handle time dynamics in this model?

4 Model extensions

4.1

Could you suggest 4 extensions/improvements to this basic model that could be carried out without acquiring additional data?

4.2

What additional data do you think could be interesting to consider?

4.3

How would you compare different models and decide which one was best?

5 Expected final points

Using the parameters obtained previously and without simulation:

5.1

Compute the scoring rates λ and μ for each game in the simulation set.

5.2

Compute the probability of home win, draw and home lose for each game in the simulation set.

5.3

Compute the expected table based on expected points of each team in each league on October 23rd 2016 (i.e. after the final match of the 2016 regular season).

5.4

We have provided you with a set of predictions for the simulation set, covering December 6th 2015 up to October 23rd 2016. These include home win/draw/away win probabilities and expected goals. Suppose the predictions provided to you come from model A. Compare the predictive performance of the fitted model with model A. What can you conclude about which performs better with respect to win probabilities and total goals?

6 Ranking probabilities

The MLS league is split into two leagues: "Eastern Conference" and "Western Conference", then the top teams enter the Final Series. We will assume final rankings are sorted by the total points, then goal difference, then goals scored and, if still tied, then the alphabetical order of the teams.

6.1

Write some code that simulates the games from March 6th 2016 to October 23rd 2016 using the models estimates and distribution. The code should return the final (ordered) table for each simulation and each conference.

6.2

Using 1e4 simulations, compute the probability for LA Galaxy to finish in the top 2 in the Western Conference.

6.3

How many simulations would be needed for the 95% confidence interval for the probability estimate in the previous question to have a Monte Carlo error of 0.1%?

7 References

You can find useful information in these articles:

- D. Karlis and I. Ntzoufras
- Dixon Coles