**Corners Challenge**

Gilad Amzaleg

Mustard Systems – Quantitative Analyst Role

1 - Model used

**Cleaning Regime**

The data provided in the training set was initially read into a Pandas data frame and a strict cleaning regime was applied to the data. All rows containing “NULL” values where removed. The columns “Home Corners” and “Away Corners” data where converted from strings to integers. The columns representing the “HomeTeamId” and “AwayTeamId” data were kept as strings as this is categorical data and if converted to int’s would cause a bias towards larger values of numerical ID’s.

**Model Considerations - Poisson**

When considering a regression model for count data, my initial thought was to use an independent Poisson Distribution to model number of corners “scored” by each team. However, after further investigation I decided this was an overly simplistic model. In a Poisson distribution the mean should equal the variance and as can be seen from the Table 1 this was not the case for this data set.

|  |  |  |
| --- | --- | --- |
|  | Mean | Variance |
| Away Corners | 4.493 | 6.277 |
| Home Corners | 5.726 | 8.088 |

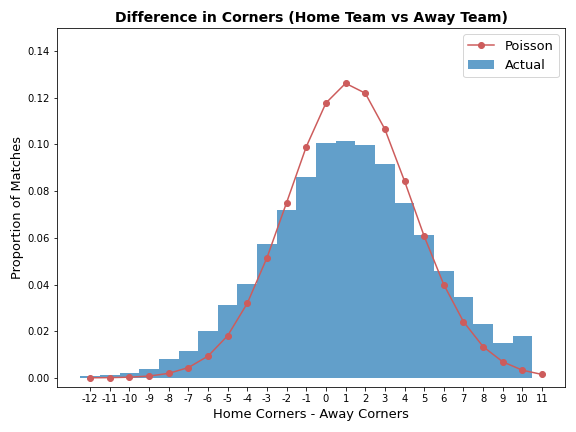
The reason that a Poisson is not sufficient is due to corners not being fully independent of each other. Corners often result in a second/third corner as the ball is returned to play at one extrema of the pitch. Hence a Poisson distribution would not compensate for the overdispersion at the end of tails (fat tail). This can be seen in Figure 1.

Figure - Plot of Poisson Distribution on top of Home/Away Corners

**Model – Negative Binomial**

In order to compensate for the overdispersion a regression model that does not make the equi-dispersion assumption was used, the Negative Binomial (NB).

Before implementing the NB regression model, the data features were manipulated so to better represent the home field advantage and create a usable regression variable matrix. Each game was split up into two separate data events with four corresponding features: Team of interest, opponent, home advantage and corners “scored”. Figure 2 is a schematic example of the feature transformation.

|  |  |  |  |
| --- | --- | --- | --- |
| Home ID | Away ID | Home Corners | Away Corners |
| 1364 | 246 | 13 | 2 |

|  |  |  |  |
| --- | --- | --- | --- |
| Team | Opponent | Corners “Scored” | Home |
| 1364 | 246 | 13 | 1 |
| 246 | 1364 | 2 | 0 |

Figure 2 – Example of data transformation. In the “Home” column 1 represents playing at home and 0 playing away.

As can be seen from Figure 2, we have restructured the data in a way that better represents the home advantage of a team. This restructuring also better represents a team’s record and considers the difficulty of the opponent that they are playing.

As the test data set does not contain any information on goals scored by each team these were not considered features in the predictive model.

To build a NB regression model, the variance parameter α is required. Using the python “Statsmodule” libraryand a method set out by [Cameron & Trivedi](http://faculty.econ.ucdavis.edu/faculty/cameron/racd2/), the value of α was calculated by using a technique they named auxiliary OLS regression without a constant. Initially a Poisson model was fitted and as a result we are able to extract a vector of fitted rates, λ and fit a further regression model that gave us a value of α. Finally, α was used to fit the NB regression model to our data set. Below is a step by step guide to implementing the NB regression model.

**STEP 1:** Fit the Poisson regression model on the data set. This will give us the vector of fitted rates **λ.**

**STEP 2:** Fit the aux OLS regression model on the data set. This will give us the value of α.

**STEP 3:**Use the α from STEP 2 to fit the NB regression model to the data set.

**STEP 4:**Use the fitted NB model to make predictions about expected counts on the test data set.

To assess the efficacy of the model the Pearson Chi2 and Deviance results were investigated. There was considerable improvement of the NB technique over the basic Poisson regression however the results were still considerably inaccurate. However due to time constraints and the scope of the exercise these results were considered acceptable and the NB model was used to predict the number of corners “scored” by a team when facing an opponent either home or away.

2 – Probability Predictions

Now that the mathematical prediction model has been built, the test data was loaded into python and a similar cleaning regime as before was implemented. Due to the order and the method of prediction variables that were used in the model, two separate NB regression were applied to each team for every match. From this we can calculate the probability of various events, such as the home team receiving 0-N number of corners whilst the away team receives 0-N.

The value of N was taken as N=21 as this was the maximum number of corners taken by an individual team either home or away.

To create this probability matrix, we wrapped this process into a function that outputted a matrix (N+1) by (N+1) that calculated the probability of each event based on the original prediction regression. Figure 3 is a schematic representation of an example probability matrix created.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **H0** | **H1** | **H2** | **H3** |
| **A0** | *P(0-0)* | *P(1-0)* | *P(2-0)* | *P(3-0)* |
| **A1** | *P(0-1)* | *P(1-1)* | *P(2-1)* | *P(3-1)* |
| **A2** | *P(0-2)* | *P(1-2)* | *P(2-2)* | *P(3-2)* |
| **A3** | *P(0-3)* | *P(1-3)* | *P(2-3)* | *P(3-3)* |

Figure 3 – Example of a sample (3+1)x(3+1) probability matrix

Figure 3 shows the probability of the home team (columns) and away team (rows) winning a specific number of corners. For example, P(2-3) is the probability of the home team receiving 2 corners and the away team 3.

To apply this to an Over/Under market you can estimate the probability of more or less total corners occurring by summing the upper and lower triangles.

For example, if the line was at 3 corners. The probability of more than 3 corners occurring would be: **P(Over 3) = P(3-1)+P(2-2)+P(1-3)+P(3-2)+P(2-3)+P(3-3).** For exactly 3 corners occurring: **P(At 3)=** **P(3-0) + P(2-1) + P(1-2) + P(0-3)**. A coloured coded representation can be seen in Figure 4.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **H0** | **H1** | **H2** | **H3** |
| **A0** | *P(0-0)* | *P(1-0)* | *P(2-0)* | *P(3-0)* |
| **A1** | *P(0-1)* | *P(1-1)* | *P(2-1)* | *P(3-1)* |
| **A2** | *P(0-2)* | *P(1-2)* | *P(2-2)* | *P(3-2)* |
| **A3** | *P(0-3)* | *P(1-3)* | *P(2-3)* | *P(3-3)* |

Figure 4 – Probabilities of Over/Under/At (line=3 corners). Sum of all Red cells equals probability of less than 3 corners, sum of all green cells equals probability of exactly 3 corners and sum of blue equals probability of more than 3 corners

**Asian Handicap**

When the line is not an integer value there can only be a result of Over/Under (O/U) as you cannot have a fraction of a corner. For all events with a non-integer line, the probability of a draw is added to either the probability of the Over/Under probability, depending on the non-integer line. For example, if a line was set at 3.5 corners to the matrix seen in Figure 4. The probability of under 3.5 goals occurring would be the sum of the Red and Green cells whilst the probability of more than 3.5 goals occurring is only the sum of Blue cells.

When the line is an integer value there is potential for a push bet. This means that if the result is a draw your stake is returned, and no net loss has occurred. This was interpreted as a safe bet and an increased protection against loss, if a draw where to occur. Hence the probability of a draw was added to both the probability of Over and Under for each case as there would be no loss if a draw were to occur. In Figure 4 (line = 3) the new “betting probability” for Over and Under would be Red + Green and Blue + Green, respectively.

3 – Betting Strategy

Now that the probabilities for each event have been calculated, a method of distributing the 342 units across the events was built. The distribution of stake was done in a method that utilises both the probability of the event and the odds given for each event. The aim is to maximise the potential winnings and minimise risk of loss.

To achieve this, a similar approach outlined [here](https://math.stackexchange.com/questions/65877/maximizing-growth-rate-in-betting-on-multiple-events) using fractional **Kelly Criterion** (multiple events) was implemented.

Initially the Expected Value (EV) was calculated for each O/U bet for all events. The EV for a stake of 1-unit is described in following equation:

Where *P(UorO)* is the probability of O/U and *odds* is the supplied odds. As placing an Over and Under bet for the same event will result in at least one loss, either a single Over or Under bet was assigned to each event. Whichever bet had the largest EV was used as that events bet as this metric uses both the probability and the potential winnings (odds) of the event.

The data was then ordered by EV and the “Reserve Rate” (RR) was calculated for the data set.

Where is the sum of each probability bet on and is the sum of each payoff.

The fraction *f* of the total stake that should be placed on a bet was then calculated using following:

A fractional (1/8) Kelly Criterion approach was used as this resulted in an even spread of the total stake across multiple events and hence reducing risk of placing and losing large amounts on few events. Then the value of *f* was summed until *f ≤* 1 and the events, ordered by EV, were chosen as the events to bet on with the *f\**342 used as the stake. The values of stake where then rounded, and the last bet cropped as due to rounding, the total amount of stake exceeded 342.

4 – Conclusion

The following table is an excerpt of the resulting bets and stakes chosen for each match ID event.

|  |  |  |  |
| --- | --- | --- | --- |
| Match ID | | Bet | Stake |
| 4 | U | | 2 |
| 8 | U | | 2 |
| 9 | O | | 4 |
| 14 | O | | 2 |
| 23 | U | | 8 |
| 24 | O | | 3 |
| 25 | U | | 2 |
| 26 | O | | 2 |
| 32 | O | | 2 |
| 33 | U | | 2 |
| 40 | O | | 5 |

A full breakdown of the probabilities calculated for Over, Under and At bet, the bet made and how much stake can be found in the accompanying csv file, “*test-test completed\_GA”*.

In total **116 bets** where made and assuming all bets are successful (without a draw) the projected return on top of stake is estimated to be **333 units**.