Development and validation of an odds and winner prediction model for sports lottery using data mining techniques

Contact details

Name author(s): Chi-Wen Chen, ABD (1), Tian-Shyug Lee, PhD (2), Chiung-Cha Chen, PhD (3) & Yin-Feng Chen, PhD (4)
Institution(s) or organisation(s): Graduate Institute of Business Administration, Fu Jen Catholic University (1); Graduate Institute of Management, Fu Jen Catholic University (2); Office of Physical Education, Fu Jen Catholic University (3); Department of Kinesiology, Texas Women's University (4)
City and country: Taipei, Taiwan (1, 2, 3); Denton, Texas, USA (4)
Email address for correspondence: 055341@mail.fju.edu.tw

Aim of paper and research questions

The purpose of this study was (a) to use data mining techniques to develop an odds and winner prediction model for sports lotteries, (b) to determine the important variables affecting the odds and winner models, (c) to compare the strengths and weaknesses of the selected odds and winner predictors, and (d) to validate the odds and winner prediction models through virtual sports lottery betting.

The research questions were:

1. Will the odds and winner prediction by using data mining techniques (Discriminant Analysis, Logistic Regression, Artificial Neural Networks, and Multivariate Adaptive Regression Splines) for sports lottery have greater probability of winning?
2. What are the important variables affecting the odds and winner prediction?

Literature review

A sports lottery is associated with odds. Odds are correlated with winning probability. In order to increase the probability of winning, sports lottery players must have enough knowledge of the game such as the rules of the game, the strength and weakness of a team and/or player, and the record of winning and losing. Previous studies developed odds prediction for games. For instance, Forrest, Goddard, and Simmons validated the odds prediction model for English football (2005). However, there is no research on the odds prediction using data mining technique for sports lotteries.

Data mining is a powerful method (a) to analyse ambiguous data in a large database, (b) to extract hidden predictive information automatically, and (c) to discover the long term relevant patterns. This type of process is also referred to as knowledge discovery in database (Piatetsky-Shapiro, 1993; Matheus, Chan & Piatetsky-Shapiro, 1993; Fayyad, Piatetsky-Shapiro & Symth, 1996). Additionally, the major contribution of the data mining technique is to investigate the valuable hidden information from database and to induce a structural model for use by companies as a decision support system (Chen & Chen, 2001). Data mining techniques have been applied for many studies (Grupe & Owrang, 1995; Berry & Linoff, 1997; Cabena, Hadjinaian, Stadler, Verhees & Zanasi, 1997). This study used data mining technique to develop and validate the odds and winner prediction models for sports lottery.

Research design and data analysis

The odds and winner prediction model for sports lottery was based on the season records of Yankee and Red Sox from 2006 to 2008 and developed using data mining techniques, such as Discriminant Analysis (DA), Logistic Regression Analysis (LRA), Artificial Neural Networks (ANNs) and Multivariate Adaptive Regression Splines (MARS). In addition, this model was validated through the use of virtual sports lottery betting for 2009 season games of Yankee and Red Sox.

Results

The findings of this study were:

1. The whole correct classification rate of DA was 55.56%. The root mean square error (RMSE) of home betting odds was 0.047115, and the RMSE of away betting odds was 0.048428;
2. The whole correct classification rate of LR was 66.67%. The RMSE of home betting odds was 0.170423, and the RMSE of away betting odds was 0.168253;
3. The whole correct classification rate of ANNs was 72.22%. The RMSE of home betting odds was 0.000058, and the RMSE of away betting odds was 0.000040;
4. The whole correct classification rate of MARS was 72.22%. The RMSE of home betting odds was 0.069329, and the RMSE of away betting odds was 0.083143;
5. The important factors affecting the odds and winner prediction were:
   (1) DA: average earned runs of starting pitcher (SP) of home team, average home runs of SP of home team, average home runs of SP of away team, and average runs of SP of away team at the last game;
   (2) LRA: average earned runs of SP of home team, average earned runs of SP of away team, average home runs of SP of home team, and average home runs of SP of away team at the last game;
   (3) ANNs: all variables;
   (4) MARS: average strikeouts of SP of away team at the last game.

Discussion and conclusion

The proposed odds and winner prediction model in this study was an effective method to predict the odds and winner. The results showed that the whole correct classification rate of ANNs was 72.22% with the minimum RMSE and had greater probability to win comparing with the odds which the sports lottery operators provided. By using the data mining techniques, the gamblers would get the objective analysis to increase the probability of winning. Therefore, the researchers recommended that using ANNs to predict the odds for sports lotteries could deliver a greater chance to win.

References

Berry, M.J.A., & Linoff, G.S. (1997). *Data mining techniques: For marketing, sales, and customer support*. New York: Wiley Computer.

Cabena, P., Hadjinaian, P., Stadler, R., Verhees, J., & Zanasi, A. (1997). *Discovering data mining from concept to implementation*. New Jersey: Prentice Hall PTR.

Chen, C.W., & Chen, H.Y. (2001). The application of data mining in sports marketing. *Paper presented at the symposium of Sports Technology*, Taipei, Taiwan.

Fayyad, U.M., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Mag.*, *17*, 37-54.

Forrest, D., Goddard, J., & Simmons, R. (2005). Odds-settrs as forecasters: The case of English football. *International Journal of Forecasting*, *21*, 551-564.

Guape, F.H., & Owrang, M.M. (1995). Database mining discovering new knowledge and cooperative advantage. *Information Systems Management*, *12*, 26-31.

Matheus, C.J., Chan, P.K., & Piatetsky-Shapiro, G. (1993). Systems for knowledge discovery in databases. *IEEE trans. On Knowledge Discovery and Data Engineering*, *5(6)*, 903-913.

Piatetsky-Shapiro, G. (1993). An overview of knowledge discovery in database: Recent progress and challenges. *Rough Sets, Fuzzy Sets and Knowledge Discovery - Proceedings of the International Workshop on Rough Sets and Knowledge Discovery*, *93*, 1-10.