International Journal of Forecasting 28 (2012) 543–552

Contents lists available at SciVerse ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

A comparative analysis of data mining methods in predicting NCAA bowl outcomes

Dursun Delen [a,*], Douglas Cogdell [b], Nihat Kasap [c]

[a] Spears School of Business, Oklahoma State University, Stillwater, OK, United States

[b] College of Hospitality, Retail, and Sport Management ITS Department, University of South Carolina, Columbia, SC, United States

[c] School of Management, Sabanci University, Istanbul, Turkey

ARTICLE INFO

Keywords:
College football
Knowledge discovery
Machine learning
Prediction
Classification
Regression

ABSTRACT

Predicting the outcome of a college football game is an interesting and challenging problem. Most previous studies have concentrated on ranking the bowl-eligible teams according to their perceived strengths, and using these rankings to predict the winner of a specific bowl game. In this study, using eight years of data and three popular data mining techniques (namely artificial neural networks, decision trees and support vector machines), we have developed both classification- and regression-type models in order to assess the predictive abilities of different methodologies (classification versus regression-based classification) and techniques. In the end, the results showed that the classification-type models predict the game outcomes better than regression-based classification models, and of the three classification techniques, decision trees produced the best results, with better than an 85% prediction accuracy on the 10-fold holdout sample. The sensitivity analysis on trained models revealed that the *non-conference team winning percentage* and *average margin of victory* are the two most important variables among the 28 that were used in this study.

1. Introduction

College football has always been one of the most widely watched sports in the US, with over 50 million in attendance during the course of a single season. It is common to find a college which has a football stadium with a greater seating capacity than the total population of the city in which the college is located. The popularity of American football can be attributed partly to its nature of being ruled by both intricate strategy and physical strength. Because of the physical demands of the game, teams can only play one game a week, and thus they end up playing only 14 competitive games through a season (which includes the end of the season bowl game).

Unlike most other competitive team sports, college football does not follow a playoff system for identifying the national champion in a given season. Instead, the annual national champion is determined by a single game between the two "best" teams, which are selected based on a combination of BCS (bowl championship series), rating formulae, and polls (the tallied votes) of sports writers and football coaches. Of the remaining hundreds of teams, the sixty or more most successful teams are invited to play in one of thirty or more end-of-season bowl games. The selection process of the "successful" teams for these bowl games is also partially based on a highly subjective, and mostly controversial, poll-driven rating and ranking process.

Predicting the outcome of a college football game (or any sports game) is an interesting and challenging problem. Therefore, challenge-seeking researchers among both academics and industry have spent a great deal of effort on forecasting the outcome of sporting events. Large quantities of historic data are available (often publicly available) from different media outlets regarding the structure and

* Corresponding author. 1 918 594 8283; fax: +1 918 594 8281.
E-mail address: dursun.delen@okstate.edu (D. Delen).

0169-2070/$ – see front matter
doi:10.1016/j.ijforecast.2011.05.002

outcomes of sporting events, in the form of a variety of numerically or symbolically represented factors which are assumed to contribute to those outcomes. However, despite the large number of studies in sports (more than 43,000 hits on digital literature databases), only a small percentage of papers has focused exclusively on the characteristics of sports forecasts. Instead, many papers have been written about the efficiency of sports markets. Since most previous betting-market studies have been concerned with economic efficiency (Van Bruggen, Spann, Lilien, & Skiera, 2010), they have not evaluated the actual (or implied) forecasts associated with such events. As it turns out, it is possible to derive a considerable amount of information about the forecasts and the forecasting process from studies that have tested the markets for economic efficiency (Stekler, Sendor, & Verlander, 2010).

Bowl games are very important for colleges, both financially (bringing in millions of dollars of additional revenue) and for recruiting highly regarded high school athletes for their football programs. The teams that are selected to compete in a given bowl game split a purse, the size of which depends on the specific bowl (some bowls are more prestigious and have higher payouts for the two teams), and therefore securing an invitation to compete in a bowl game is the main goal of any division I-A college football program. The decision makers in the bowl games are given the authority to select and invite successful bowl-eligible (teams that have six wins against their Division I-A opponents in that season) teams (as per the ratings and rankings) which will play an exciting and competitive game, attract fans of both schools, and keep the remaining fans tuned in via a variety of media outlets for advertising (West & Lamsal, 2008).

Every year, people either casually (i.e., recreational office pools for bragging rights) or somewhat seriously (i.e., wagering/betting for monetary gain) put their knowledge of the game on the line in an attempt to accurately predict the outcomes of the bowl games. The emotional and highly dynamic nature of college football, coupled with the selection process, which aims to bring together equally rated opponents from different conferences (which often have not played each other in the recent past), makes this prediction even more challenging and exciting. Many statisticians and quantitative analysts have explored ways to quantify the variables of a college football bowl game numerically and/or symbolically, and to use these variables in a wide variety of models for predicting the outcome of a game (Stekler et al., 2010). As can be seen in the literature review section, many of these studies rely on the ranking-based selection, and even though some have claimed to have met with limited success, many have reported the difficulty of this prediction problem.

In this paper, we report on a data mining study where we used eight years of bowl game data, along with three popular data mining techniques (decision trees, neural networks and support vector machines), to predict both the classification-type outcome of a game (win versus loss) and the regression-type outcome (projected point difference between the scores of the two opponents). The rest of the paper is organized as follows. The next section provides a review of the relevant literature in this prediction domain. Section 3 describes the methodology (i.e., the data, prediction model types and evaluation methods used in the study), followed by Section 4, which provides the prediction results. Finally, Section 5 summarizes the study, discusses the findings, and indentifies the limitations and future research directions.

2. Literature review

The literature on college football has concentrated mainly on two particular themes: the development of ratings and rankings of the teams, and the prediction of game outcomes (probably more importantly). While some of these studies have focused on the accuracy and fairness of the rating and/or ranking schemas, others have developed these rankings and used them for the purpose of predicting the outcome of a specific game. Since this study is about the prediction of bowl game outcomes, this review therefore excludes the body of literature which is dedicated to developing and/or criticizing the subjective nature of the poll-driven rating and/or ranking methods.

Many studies have used methods based on various forms of least squares estimation, where the parameters are formulated (as various statistics of the competing teams) using linear models to predict game outcomes. These studies include those of Stefani (1980), who, among other predictors, incorporated the home field advantage into least squares ratings; Farlow (1984), who developed a linear model for calculating ratings that can be used for the prediction of game outcomes; Stefani (1987), who discussed additional applications of least square methods in the prediction of future game outcomes; Stern (1995), who used a linear combination of variables representing past performances to predict the outcomes of future games; Purucker (1996), who used four variables—yards gained, rushing yards gained, turnover margin and time of possession—to predict the game outcome; Bassett (1997), who proposed the use of least absolute errors rather than least squares estimation, in order to reduce the influence of outliers on prediction model development; and Harville (2003), who proposed a modified least squares approach which incorporated the home field advantage and removed the influence of the margin of victory on ratings, identified seven key attributes of any ranking system, and showed that the ratings based on the modified least squares approach had a reasonably good predictive accuracy. Most recently, West and Lamsal (2008) used a combination of team defense and offence statistics to predict the game outcome, reporting a prediction accuracy of 59.4% (explaining only 22% of the variance), which is somewhat similar to the prediction accuracies reported by other similar studies.

Several other studies have been dedicated to more inclusive methods of predicting the outcomes of future games, using a variety of past information to predict future outcomes. For instance, Harville (1980) included results from previous seasons and information other than the point spread to develop ratings for teams in future seasons and predict the outcomes of future games using linear mixed models. Trono (1988) proposed a probabilistic model based on the simulated outcomes of individual plays