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

(where plays were based on a deterministic play-calling strategy) for predicting the outcomes of games, where the probabilities of certain events occurring were based on past performances; using this model, they correctly predicted the outcomes of 58.7% of bowl games over eight seasons. Some other researchers (Ong & Flitman, 1997; Pardee, 1999) have considered the use of neural networks in an attempt to predict the outcomes of future football games. They have built neural network models based on the numerical representation of past/historic information on games/teams for predicting future game outcomes, and demonstrated an improved prediction accuracy (as high as 76%).

Glickman and Stern (1998) worked on predicting the outcome of National Football League games using many variable factors concerning team strength. The variability in team strength, according to the authors, is due to both pre-season changes and changes in the team during regular seasons, such as random injuries, the psychology of the players, changing team members, etc. The authors model these time variant variables in a linear state space framework. The predictive ability of the model built compares favorably with the expert predictions listed in Las Vegas betting lines. In this study, the authors also employed Markov chain Monte Carlo methods for post-modeling analysis of the model parameters. A very important point to be noticed in this work is that they have taken into account not only the during-the-season variation, but also the variation between seasons which affects the team strength.

Standing apart from the complex models being built for football game outcome prediction, Rotshtein, Posner, and Rakityanskaya (2005) worked on simple logical reasoning enhanced with fuzzy logic. The authors use metrics relating to the previous performances of the competing teams as the independent variables in the fuzzy model involving the dependency rules. Then follows the process of reducing the difference between the actual and modeled results using a genetic algorithm, and optimizing the variables for data updating using a neural net. In a discussion of the structure of the authors' model, it is observed that the target variable is not just win or lose, but is divided into five categories, namely high score loss, low score loss, drawn game, low score win, and high score win. This makes the model richer, but is more challenging from the perspective of obtaining reasonably good prediction results.

Joseph, Fenton, and Neil (2006) presented a comparative study of expert-built Bayesian nets and machine-learned Bayesian nets. A Bayesian network serves the purpose of logically laying out the data for decision support and specifically representing the probabilistic relationship between the variables affecting the outcome of the game and the actual outcome of the game. The outcomes in this situation are defined as win, lose or draw. The authors compared the expert-built Bayesian net with machine learning techniques such as a decision tree learner, a naïve Bayesian learner, a data-driven Bayesian net, and a *k*-nearest neighbor learner. They observed that the expert Bayesian net was inferior to the machine learning based Bayesian nets in terms of the simplicity in its build and its reduced use of learning data. Song, Boulier, and Stekler (2007), in another comparative study, compared statistical models to domain experts in predicting the accuracy of football game outcomes. The study suggested that there was no statistically significant difference in accuracy between the two.

Fainmesser, Fershtman, and Gandal (2009) developed "a consistent weighted ranking scheme" for comparing not only the teams which were actually competing with each other, but all of the teams which were participating and competing for the championship. The authors clearly address the issue of a generalized ranking system which considers only the overall win or loss counts, irrespective of the competing teams. The system considers the outcomes of every game played, with respect to the strength of the opponent and schedule, and the existence of a home field advantage, in order to define the weight of the team's performance. According to the authors, losing a game is totally distinguishable from not competing with that particular opponent in that specific game. The authors state that they use the structure of NCAA college football to derive the complete rankings of the teams. The authors compare their CWR scheme with the BCS computer rating schemes based on the prediction accuracy, and find their scheme to be better than the BCS schemes. They give the credit for this to the consistency of their rating scheme, and the ranking function which encapsulates the essential parameters.

Our study differs from the earlier studies in two respects: first, it uses three popular machine learning techniques which are capable of capturing non-linear relationships between the input and output variables (as opposed to traditional statistical models, which are often limited to linear relationships); and, second, it compares the two prediction methodologies (namely classification and regression-based classification) under each of the three machine learning techniques.

3. Methodology

In this research, we follow a popular data mining methodology called CRISP-DM (Cross Industry Standard Process for Data Mining) (Shearer, 2000), which is a six-step process: (1) understanding the domain and developing the goals for the study; (2) identifying, accessing and understanding the relevant data sources; (3) preprocessing, cleaning, and transforming the relevant data; (4) developing models using comparable analytical techniques; (5) evaluating and assessing the validity and utility of the models against each other and against the goals of the study; and (6) deploying the models for use in decision-making processes. This popular methodology provided us with a systematic and structured way of conducting this data mining study, and hence improved the likelihood of obtaining accurate and reliable results.

In order to objectively assess the predictive powers of the different model types, we used a cross-validation methodology, which is a popular statistical technique that is often used in data mining for comparing the predictive accuracies of multiple models. The traditional cross validation methodology splits the data into two mutually exclusive subsets, for training and testing (or three in the case of neural networks, which also include a validation set). This design of a single random split may

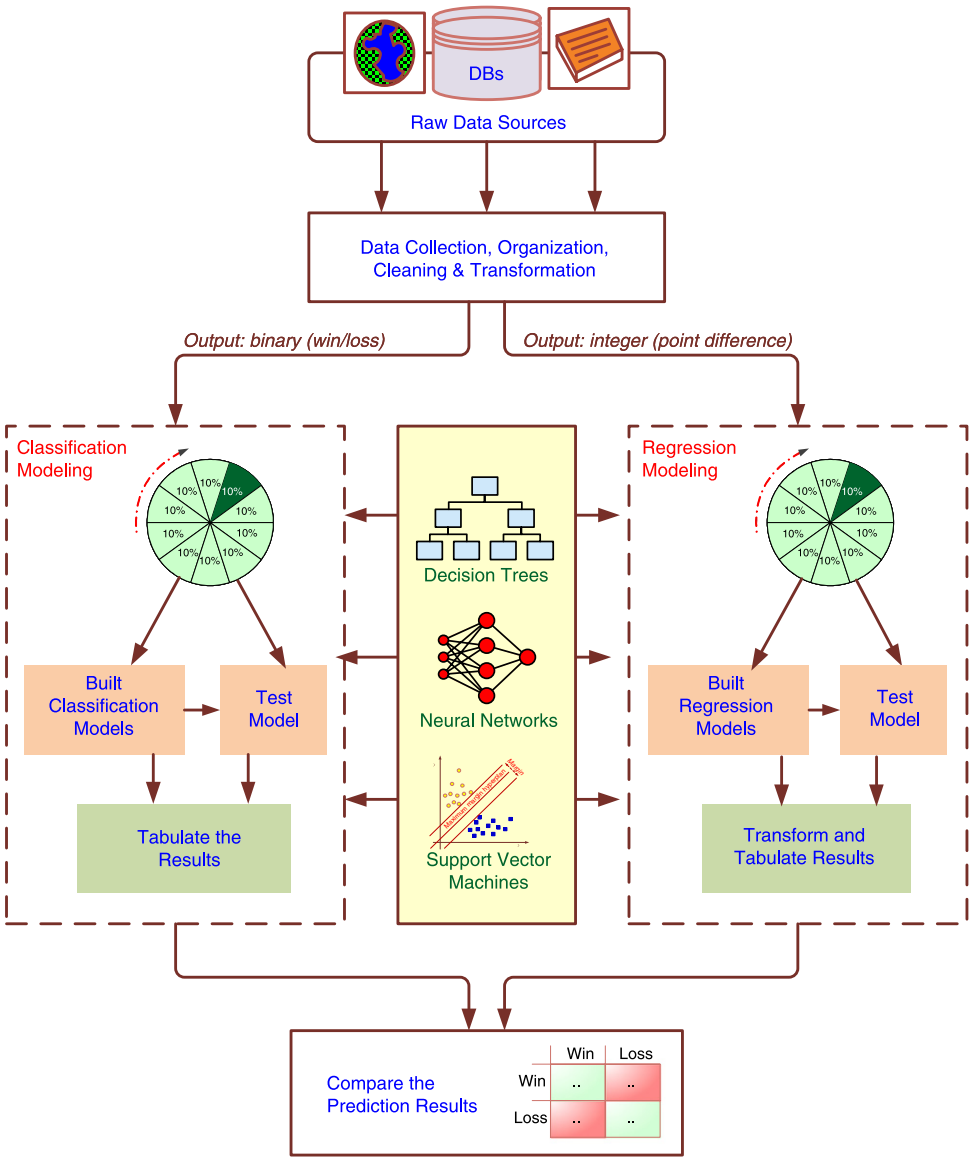

Fig. 1. A graphical representation of the methodology employed in this study.

lead to a non-homogeneous representation in the subsets, especially if the dataset is relatively small. In order to minimize the bias associated with the random sampling of the training and testing data samples, we employed an experimental design method called *k*-fold cross validation. More details of this cross validation methodology are provided in Section 3.3. A graphical representation of the methodology used in this study is shown in Fig. 1.

3.1. Data

The sample data for this study are collected from a variety of sports databases available on the web, including jhowel.net, ESPN.com, Covers.com, ncaa.org, and rauzulusstreet.com. The dataset included 244 bowl games, representing a complete set of eight seasons of college football bowl games played between 2002 and 2009. We also included an out-of-sample data set (the 2010–2011 bowl games) for additional validation purposes. Exercising one of the popular data mining rules-of-thumb, we included as much relevant information in the model as we possibly could. Therefore, after an in-depth variable identification and collection process, we ended up with a dataset which included 36 variables, of which the first six were the identifying variables (i.e., the name and year of the bowl game, the home and away team's names and their athletic conferences—see variables 1–6 in Table 1), followed by 28 input variables (which included variables delineating a team's seasonal statistics on offense and defense, game outcomes, team composition characteristics, athletic conference characteristics, and how they fared against the odds—see variables 7–34 in Table 1), and finally, the last two were the output variables (i.e., *ScoreDiff* —the score difference between the home and away teams, represented

Table 1
Description of the variables used in this study.

No	Cat	Variable name	Description
1	ID	YEAR	Year of the bowl game
2	ID	BOWL GAME	Name of the bowl game
3	ID	HOMETEAM	Home team (as listed by the bowl organizers)
4	ID	AWAYTEAM	Away team (as listed by the bowl organizers)
5	ID	HOMECONFERENCE	Conference of the home team
6	ID	AWAYCONFERENCE	Conference of the away team
7	I1	DEFPTPGM	Defensive points per game
8	I1	DEFRYDPGM	Defensive rush yards per game
9	I1	DEFYDPGM	Defensive yards per game
10	I1	PPG	Average number of points a given team scored per game
11	I1	PYDPGM	Average total pass yards per game
12	I1	RYDPGM	Team's average total rush yards per game
13	I1	YRDPGM	Average total offensive yards per game
14	I2	HMWIN%	Home winning percentage
15	I2	LAST7	How many games the team won out of their last seven games
16	I2	MARGOVIC	Average margin of victory
17	I2	NCTW	Non-conference team winning percentage
18	I2	PREVAPP	Did the team appear in a bowl game in the previous year?
19	I2	RDWIN%	Road winning percentage
20	I2	SEASTW	Winning percentage for the year
21	I2	TOP25	Winning percentage against AP top 25 teams for the year
22	I3	TSOS	Strength of schedule for the year
23	I3	FR%	Percentage of games played by freshman class players for the year
24	I3	SO%	Percentage of games played by sophomore class players for the year
25	I3	JR%	Percentage of games played by junior class players for the year
26	I3	SR%	Percentage of games played by senior class players for the year
27	I4	SEASOvUn%	Percentage of times a team went over the O/U[a] in the current season
28	I4	ATSCOV%	Against the spread cover percentage of the team in previous bowl games
29	I4	UNDER%	Percentage of times a team went under in previous bowl games
30	I4	OVER%	Percentage of times a team went over in previous bowl games
31	I4	SEASATS%	Percentage of covering against the spread for the current season
32	I5	CONCH	Did the team win their respective conference championship game?
33	I5	CONFSOS	Conference strength of schedule
34	I5	CONFWIN%	Conference winning percentage
35	O1	ScoreDiff[b]	Score difference (home team score − away team score)
36	O2	WinLoss[b]	Whether the home team wins or loses the game

I1: offense/defense; I2: game outcome; I3: team configuration; I4: against the odds; I5: conference stats; ID: Identifier variables; O1: Output variable for regression models; O2: output variable for classification models.
[a] Over/Under: Whether a team will go over or under the expected score difference.
[b] Output variables: ScoreDiff for regression models and WinLoss for binary classification models.

by an integer—and *WinLoss*—whether the home team won or lost the bowl game, represented by a nominal label).

In the formulation of the dataset, each row (a.k.a. tuple, case, sample, example, etc.) represented a bowl game, and each column stood for a variable (i.e., identifier/input or output type). In order to represent the game-related comparative characteristics of the two opponent teams in the input variables, we calculated and used the differences between the measures of the home and away teams. All of these variable values are calculated from the home team's perspective. For instance, the variable PPG (average number of points a team scored per game) represents the difference between the home team's PPG and the away team's PPG. The output variables represent whether the home team wins or loses the bowl game. That is, if the *ScoreDiff* variable takes a positive integer, then the home team is expected to win the game by that margin; otherwise (that is, if the *ScoreDiff* variable is a negative integer) the home team is expected to lose the game by that margin. In the case of *WinLoss*, the value of the output variable is a binary label, "Win" or "Loss", indicating the outcome of the game for the home team.

3.2. Methods

In this study, three popular prediction techniques are used (and compared to each other): neural networks, decision trees and support vector machines. These prediction techniques are selected because of their capability to model both classification and regression type prediction problems, and also due to their popularity in the recent data mining literature. A brief description of these three modeling techniques follows.

3.2.1. Neural networks

Neural networks (NN) are commonly known as biologically inspired mathematical techniques, and are capable of modeling extremely complex non-linear functions (Haykin, 2008). In this study, we used a popular NN architecture known as the multi-layer perceptron (MLP) with back-propagation type supervised-learning algorithm. MLP is capable of producing both classification and regression type prediction models, with the only difference being that the output variable is either nominal or numeric, for classification and regression estimations respectively. MLP is shown to be a strong function approximator for

prediction problems; that is, given the right size and structure, MLP is shown to be capable of learning highly complex nonlinear relationships between the input and output variables (Hornik, Stinchcombe, & White, 1990).

3.2.2. Decision trees

As the name implies, this technique recursively separates observations into branches in order to construct a tree for the purpose of achieving the highest possible prediction accuracy. In doing so, different mathematical algorithms (e.g., information gain, the Gini index, Chi-square statistics, etc.) are used to identify a variable (from the pool of available variables) and the corresponding threshold for that variable in order to split the pool of observations into two or more subgroups. This step is repeated at each leaf node until the entire tree has been constructed. In this study, we choose to use the classification and regression trees (CART or C&RT) which were first developed by Breiman, Friedman, Olshen, and Stone (1984). This decision tree algorithm is capable of modeling both classification and regression type prediction problems.

3.2.3. Support vector machines

Support vector machines (SVM) belong to the family of generalized linear models which aim to achieve a prediction decision (classification or regression) based on a linear combination of features derived from the variables. The input-to-output mapping function in SVM can be either a classification or a regression function. SVM uses nonlinear kernel functions to transform the input data (inherently representing highly complex nonlinear relationships) to a high dimensional feature space in which the input data become more manageable (i.e., linearly representable) than in the original input space (Cristianini & Shawe-Taylor, 2000).

Even though SVM and NN are often assumed to belong to the same family of methods, there are some significant differences between the two. For instance, while the development of NN followed a heuristic path, with applications and extensive experimentation preceding the theory, the development of SVM involved sound theory first, then implementation and experimentations (Hastie, Tibshirani, & Friedman, 2009). A significant advantage of SVM over NN is that while NN suffer from multiple local minima, the solution to an SVM is global and unique (Drucker, Burges, Kaufman, Smola, & Vapnik, 1997). Two other advantages of SVM are that (i) they have a simple geometric interpretation and give a sparse solution (unlike with NN, the computational complexity of SVM does not depend directly on the dimensionality of the input space), and (ii) they use structural risk minimization, while NN use empirical risk minimization (one of the reasons why they often outperform NN in practice, being less prone to overfitting).

3.3. Evaluation criteria

While evaluating and comparing the predictive accuracies of two or more methods, a common practice is to split the data into two subsets for training and validation (or sometimes into three subsets, which also includes testing). Often, 2/3 of the data points are used for model building and 1/3 are used for validation. Such single splits of the dataset are often prone to sampling bias, no matter what type of random sampling technique is used. In order to minimize the bias associated with this random sampling of the training and holdout data samples, we chose to use a k-fold cross validation methodology (Turban, Sharda, & Delen, 2010).

In k-fold cross-validation, the complete dataset is split randomly into k mutually exclusive subsets of approximately equal size. Each classification model is trained and tested k times; each time it is trained on all but one fold and tested on the remaining single fold. The test outcomes of all folds are compiled into a confusion matrix. The cross-validation estimate of the overall accuracy is calculated as the average of the k individual accuracy measures, as in Eq. (1):

$$CV = \frac{1}{k}\sum_{i=1}^{k} A_i, \tag{1}$$

where CV stands for the cross-validation accuracy, k is the number of folds used, and A is the accuracy measure of each fold.

Since the cross-validation accuracy depends on the random assignment of the individual samples to k distinct folds, a common practice is to stratify the folds themselves. In stratified k-fold cross validation, the folds are created in such a way that they contain approximately the same proportion of predictor labels (i.e., classes) as the original dataset. Empirical studies have shown that stratified cross validation tends to generate comparison results with a lower bias and lower variance than regular cross-validation (Olson & Delen, 2008). In this study, the value of k is set to 10 (i.e., the complete set of 244 samples is split into 10 subsets, each having about 25 samples), as is a common practice in predictive data mining applications. A graphical depiction of the 10-fold cross validations is shown in Fig. 2 (Olson & Delen, 2008).

To compare the prediction models, we used three performance criteria: accuracy, sensitivity, and specificity (Eqs. (2)–(4) respectively).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{2}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \tag{3}$$

$$\text{Specificity} = \frac{TN}{TN + FP}, \tag{4}$$

where TP, TN, FP, and FN denote true positive (accurate prediction of the wins), true negative (accurate prediction of losses), false positive (inaccurate prediction of losses as wins), and false negative (inaccurate prediction of wins as losses), respectively. Accuracy, as shown by Eq. (2), measures the proportion of correctly classified games, thus predicting the overall probability of the correct classification. Sensitivity and specificity, as shown by Eqs. (3) and (4) respectively, measure the model's ability to predict the wins and losses separately.

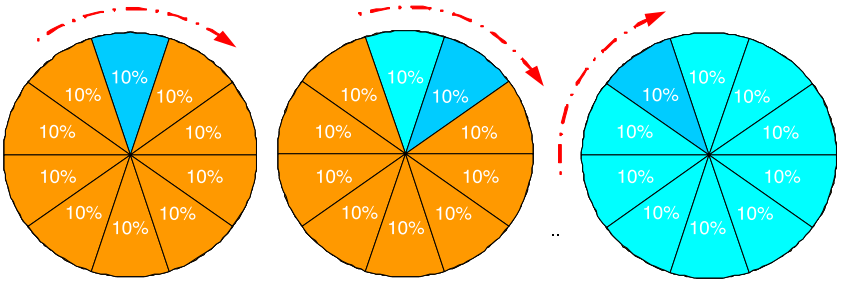

Fig. 2. A pictorial depiction of 10-fold cross validation.

Table 2
Prediction results for the direct classification methodology.

Prediction method (classification*)		Confusion matrix[a]		Accuracy** (%)	Sensitivity (%)	Specificity (%)
		Win	Loss			
ANN (MLP)	Win	92	42	75.00	68.66	82.73
	Loss	19	91			
SVM (RBF)	Win	105	29	79.51	78.36	80.91
	Loss	21	89			
DT (C&RT)	Win	113	21	**86.48**	84.33	89.09
	Loss	12	98			

* The output variable is a binary categorical variable (Win or Loss); the differences were significant.
** $p < 0.01$.
[a] The rows show actual values and the columns show predictions.

Table 3
Prediction results for the regression-based classification methodology.

Prediction method (regression-based*)		Confusion matrix[a]		Accuracy**	Sensitivity	Specificity
		Win	Loss			
ANN (MLP)	Win	94	40	72.54	70.15	75.45
	Loss	27	83			
SVM (RBF)	Win	100	34	74.59	74.63	74.55
	Loss	28	82			
DT (C&RT)	Win	106	28	77.87	76.36	79.10
	Loss	26	84			

* The output variable is a numerical/integer variable (point-diff); the differences were significant.
** $p < 0.01$.
[a] The rows show actual values and the columns show predictions.

4. Results

The prediction results of the three modeling techniques are presented in Tables 2 and 3. Table 2 presents the 10-fold cross validation results of the classification methodology, where the three data mining techniques are formulated to have a binary-nominal output variable (i.e., *WinLoss*). Table 3 presents the 10-fold cross validation results of the regression-based classification methodology, where the three data mining techniques are formulated to have a numerical output variable (i.e., *ScoreDiff*). In the regression-based classification prediction, the numerical output of the models is converted to a classification type by labeling the positive *WinLoss* numbers "Win" and the negative *WinLoss* numbers "Loss", and then tabulating them in the confusion matrixes. Using the confusion matrixes, the overall prediction accuracy, sensitivity and specificity of each model type are calculated and presented in Tables 2 and 3.

As the results indicate, the classification-type prediction methods performed better than the regression-based classification-type prediction methodology. Among the three data mining technologies, classification and regression trees produced better prediction accuracies in both prediction methodologies. Overall, classification and regression tree classification models produced a 10-fold cross validation accuracy of 86.48%, followed by support vector machines, with a 10-fold cross validation accuracy of 79.51%, and neural networks, with a 10-fold cross validation accuracy of 75.00%. Using a t-test, we found that these accuracy values were significantly different at the 0.05 alpha level; that is, decision trees are significantly better predictors of this domain than neural networks or support vector machines, while support vector machines are significantly better predictors than neural networks (see Table 4). For the sensitivity and specificity measures, classification and regression trees once again produced better results than support vector machines or neural networks.

Table 4
Results of the t-test (p-values) for the accuracy measures of all three prediction methods.

Classification			Regression-based classification		
	SVM	DT		SVM	DT
ANN	0.00037	0.00000	ANN	0.00585	0.00000
SVM		0.00000	SVM		0.00004

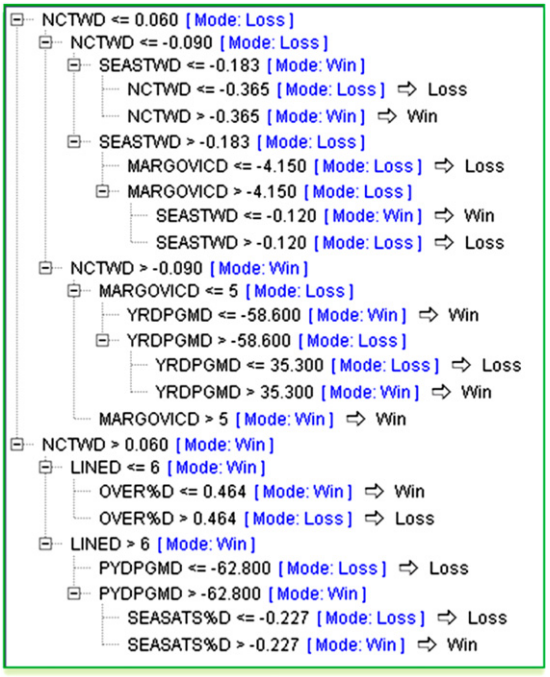

Fig. 3. A classification-type decision tree output for this prediction domain.

Besides being the best predictors in the domain (for the data used), classification and regression trees are preferable to the other two machine learning techniques, for two main reasons: (1) classification and regression tree outcomes are more transparent: they produce easily readable, understandable and digestible models in the form of trees; and (2) classification and regression tree models are easier to convert into mathematical expressions (i.e., converting decision trees into series of nested if-then rules) and integrate into decision support systems. Fig. 3 illustrates one of the classification-type decision trees produced in this study in the form of an indented list. As can be seen, the classification and regression tree algorithm identifies the most discernable variables and their split values, in order to contract branches recursively in a divide-and-conquer manner until the leaf nodes (the predictions) are purified/homogenized to a satisfactory level.

4.1. Scoring the 2010–2011 bowl games

In order to assess the predictive ability of the three classification-type prediction models further, we collected data on the latest NCAA bowl season, namely the 2010–2011 bowl games. The three model types are re-developed using all of the available data for the previous eight years (i.e., from the 2002–2003 bowl season to the 2009–2010 bowl season), then tested/scored based on the 2010–2011 bowl game data. The results are presented in Table 5.

The prediction results for the 2010–2011 bowl games turned out to be somewhat lower than (but quite comparable to) those obtained from the 10-fold cross validation. Out of the 35 bowl games, ANN predicted 25 of them correctly (giving a 71.43% prediction accuracy), SVM predicted 26 correctly (giving a 74.29% prediction accuracy), and finally, DT predicted 29 accurately (giving an 82.86% prediction accuracy).

As an additional comparison and as a benchmark, we looked at the predictions of the betting markets. We converted the betting line values for all 35 bowl games for the same season (2010–2011 bowl games) into win/loss predictions. After this conversion, we found that, out of the 35 games, only 24 of them were predicted accurately by the betting lines. Compared to the 29 correct predictions made by the decision tree, it seems that the decision tree algorithm was more accurate, at least for this comparison criterion. From the ease of deployment (and scoring) perspective, the decision tree method is preferable to the other two methods once again, because it can be converted into a series of if-then rules and can easily be integrated into a scoring system.

5. Discussion and conclusion

The results of the study show that the classification-type models predict the game outcomes better than regression-based classification models. Of the three classification techniques, classification and regression trees produced the best results, with a prediction accuracy better than 86% on the 10-fold holdout sample, followed by support vector machines (79.51% prediction accuracy) and neural networks (75.00% prediction accuracy). For other assessment metrics (i.e., sensitivity and specificity), we see once again that classification and regression trees produce better results than either support vector machines or neural networks. Even though these results are specific to the application domain and data used in this study, and therefore should not be generalized beyond the scope of the study, they are still exciting because not only are decision trees the best predictors, they are also better in understanding and deployment than the other two machine learning techniques employed in this study.

In order to understand the relative importances of the factors used in the study, we conducted a sensitivity analysis on trained prediction models, where we measured the comparative importance of the input variables in predicting the output. That is, a sensitivity analysis measures the relative importance of a variable based on the difference in modeling performance with and without the inclusion of a variable (i.e., the sensitivity of a specific predictor variable is the error of the prediction model without the predictor variable divided by the error of the model with the predictor variable) (Saltelli, Chan, & Scott, 2008). After normalizing, combining and consolidating

Table 5
Prediction results for the test dataset (2010–2011 bowl season).

Classification method		Confusion matrix[a]		Overall accuracy (%)
		Win	Loss	
Neural networks (MLP)	Win	15	3	71.43
	Loss	7	10	
Support vector machines (RBF)	Win	14	4	74.29
	Loss	5	12	
Decision trees (C&RT)	Win	15	3	**82.86**
	Loss	3	14	

[a] The rows shows actual values and the columns show predictions.

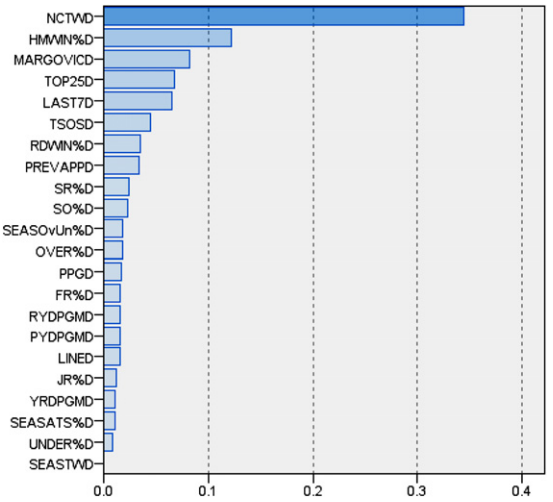

Fig. 4. Relative variable importance based on the consolidated sensitivity analysis results.

the sensitivity analysis results of all classification and regression models, the following input variables made it to the top of the list (presented in ranking order): NCTW (non-conference team winning percentage), HMWIN (home win percentage), MARGOVIC (average margin of victory during the current season), TOP25 (success against the top 25 teams during the current season), and LAST7 (success in the last seven games of the season). It is somewhat surprising that none of the "against the odds" variables made it to the top five. The ordered list of variable importance is presented in a horizontal bar-chart in Fig. 4, where the size of the horizontal bars represents the relative importance of each variable with respect to the rest of the predictive variables.

The results obtained herein should be interpreted within the scope of the study. The use of other prediction techniques and/or other variables may produce somewhat different results. In order to be able to comment further on the generalizability of the findings in this study, more elaborate experimentations with much larger data sets and prediction techniques are required.

The main directions for future research following this study include (a) the enrichment of the variable set (e.g., identifying and including more input variables, representing variables in different forms for better expressiveness, etc.), (b) the employment of other classification and regression methods and methodologies (the use of other prediction techniques such as rough sets, genetic algorithm based classifiers, etc., and the use of ensemble models), (c) experimentation with seasonal game predictions (which may need a combination of static and time series variable identifications), and (d) experimentation with other college and professional sports predictions.

References

Bassett, G. W. (1997). Robust sports ratings based on least absolute errors. *The American Statistician*, *51*, 99–105.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth & Brooks Books & Software.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. London: Cambridge University Press.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. In *Advances in neural information processing systems*: *Vol. 9* (pp. 155–161). MIT Press.

Fainmesser, I., Fershtman, C., & Gandal, N. (2009). *A consistent weighted ranking scheme with an application to NCAA college football rankings*. CEPR discussion papers 5239.

Farlow, S. J. (1984). A computer program for ranking athletic teams. *International Journal of Mathematical Education in Science and Technology*, *15*(6), 697–702.

Glickman, M. E., & Stern, H. S. (1998). A state-space model for national football league scores. *Journal of the American Statistical Association*, *93*, 25–35.

Harville, D. A. (1980). Predictions for national football league games via linear-model methodology. *Journal of the American Statistical Association*, *75*(371), 516–524.

Harville, D. A. (2003). The selection or seeding of college basketball or football teams for postseason competition. *Journal of the American Statistical Association*, *98*, 17–27.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction* (2nd ed.) New York: Springer Publishing.

Haykin, S. (2008). *Neural networks and learning machines* (3rd ed.) New Jersey: Prentice Hall.

Hornik, K., Stinchcombe, M., & White, H. (1990). Universal approximation of an unknown mapping and its derivatives using multilayer feedforward network. *Neural Networks*, *3*, 359–366.

Joseph, A., Fenton, N. E., & Neil, M. (2006). Predicting football results using Bayesian nets and other machine learning techniques. *Knowledge-Based Systems*, *19*(7), 544–553.

Olson, D., & Delen, D. (2008). *Advanced data mining techniques*. New York, NY: Springer.

Ong, E. S., & Flitman, A. M. (1997). Using neural networks to predict binary outcomes. In *Proceedings of the 1997 international conference on neural information processing and intelligent information systems* (pp. 427–431). Beijing, New Jersey, USA: IEEE, ISBN/ISSN: 7-80003-410-O/TP-17.

Pardee, M. (1999). *An artificial neural network approach to college football prediction and ranking*. Technical Paper. University of Wisconsin, Electrical and Computer Engineering Department.

Purucker, M. (1996). Neural network quarterbacking. *IEEE Potentials*, *15*(3), 9–15.

Rotshtein, P., Posner, M., & Rakityanskaya, A. B. (2005). Football predictions based on a fuzzy model with genetic and neural tuning. *Cybernetics and Systems Analysis*, *41*(4), 619–630.

Saltelli, A., Chan, K., & Scott, E. M. (2008). *Sensitivity analysis*. New York, NY: Wiley Publishing.
Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of Data Warehousing*, *5*, 13–22.
Song, C.-U., Boulier, B. L., & Stekler, H. O. (2007). The comparative accuracy of judgmental and model forecasts of American football games. *International Journal of Forecasting*, *23*(3), 405–413.
Stefani, R. T. (1980). Improved least squares football, basketball, and soccer predictions. *IEEE Transactions on Systems, Man and Cybernetics*, *10*, 116–123.
Stefani, R. T. (1987). Applications of statistical methods to American football. *Journal of Applied Statistics*, *14*(1), 61–73.
Stekler, H. O., Sendor, D., & Verlander, R. (2010). Issues in sports forecasting. *International Journal of Forecasting*, *26*, 606–621.
Stern, H. S. (1995). Who's number 1 in college football? …and how might we decide? *Chance Magazine*, *8*(3), 7–14.
Trono, J. A. (1988). A deterministic prediction model for the American game of football. *ACM SIGSIM Simulation Digest*, *19*(1), 26–53.
Turban, E., Sharda, R., & Delen, D. (2010). *Decision support and business intelligence systems* (9th ed.) New Jersey: Prentice Hall.
Van Bruggen, G. H., Spann, M., Lilien, G. L., & Skiera, B. (2010). Prediction markets as institutional forecasting support systems. *Decision Support Systems*, *49*, 404–416.
West, B. T., & Lamsal, M. (2008). A new application of linear modeling in the prediction of college football bowl outcomes and the development of team ratings. *Journal of Quantitative Analysis in Sports*, *4*(3), 1–19.

Dursun Delen is an Associate Professor of Management Science and Information Systems in the Spears School of Business at Oklahoma State University (OSU). He received his Ph.D. in Industrial Engineering and Management from OSU in 1997. Prior to his appointment as an Assistant Professor at OSU in 2001, he worked for a privately-owned research company, Knowledge Based Systems Inc., in College Station, Texas, as a research scientist for five years, during which time he led a number of decision support and other information systems related research projects funded by federal agencies such as DoD, NASA, NIST and DOE. His research has appeared in various major journals, including *Decision Support Systems*, *Communications of the ACM*, *Computers and Operations Research*, *Computers in Industry*, *Journal of Production and Operations Management*, *Artificial Intelligence in Medicine*, and *Expert Systems with Applications*, among others. He recently published two books (*Advanced Data Mining Techniques* with Springer, in 2008; and *Decision Support and Business Intelligence Systems* with Prentice Hall, in 2010). He is an associate editor for the *International Journal of RF Technologies*, and serves on the editorial boards of the *Journal of Information and Knowledge Management*, *International Journal of Intelligent Information Technologies*, *International Journal of Service Sciences*, and *Journal of Emerging Technologies in Web Intelligence*. His research interests are in decision support systems, data and text mining, expert systems, knowledge management, business intelligence and enterprise modeling.

Douglas Cogdell has been working as a Network Administrator for the College of Hospitality, Retail, and Sport Management and Entertainment at the University of South Carolina since 2006. He received his MPA in Public Administration from the University of South Carolina in 2005 and his B.S. in Political Science from Lander University in 2001. He attained his Microsoft Certified IT Professional (MCITP) certification recently, in 2010. His research interests are in the analytic modeling of market efficiencies and inefficiencies with regard to the betting industry.

Nihat Kasap is an Assistant Professor at the Faculty of Management, Sabanci University, Istanbul, Turkey. Dr. Kasap holds a Ph.D. in business with an MIS major from the University of Florida, and B.S. and M.S. degrees from Middle East Technical University, Ankara, Turkey, and the University of Florida, respectively. His research focuses on pricing and QoS issues in telecommunication networks, mobile technologies in E-government services, heuristic design and optimization, data mining, and machine learning. His work has appeared in *Operations Research Letters*, decision support systems journals and the proceedings of numerous international and national conferences. Research projects which he has participated in have been awarded funding from The Scientific and Technological Research Council of Turkey (TUBITAK) and the Ministry of Industry and Commerce.