Journal of Human Kinetics volume 77/2021, 245-259 DOI: 10.2478/hukin-2021-0023
Section III – Sports Training

sciendo

Victory Prediction of Ladies Professional Golf Association Players: Influential Factors and Comparison of Prediction Models

by
Jin Seok Chae[1], Jin Park[2], Wi-Young So[3]

This study aims to identify the most accurate prediction model for the possibility of victory from the annual average data of 25 seasons (1993–2017) of the Ladies Professional Golf Association (LPGA), and to determine the importance of the predicting factors. The four prediction models considered in this study were a decision tree, discriminant analysis, logistic regression, and artificial neural network analysis. The mean difference in the classification accuracy of these models was analyzed using SPSS 22.0 software (IBM Corp., Armonk, NY, USA) and the one-way analysis of variance (ANOVA). When the prediction was based on technical variables, the most important predicting variables for determining victory were greens in regulation (GIR) and putting average (PA) in all four prediction models. When the prediction was based on the output of the technical variables, the most important predicting variable for determining victory was birdies in all four prediction models. When the prediction was based on the season outcome, the most important predicting variables for determining victory were the top 10 finish% (T10) and official money. A significant mean difference in classification accuracy was observed while performing the one-way ANOVA, and the least significant difference post-hoc test showed that artificial neural network analysis exhibited higher accuracy than the other models, especially, for larger data sizes. From the results of this study, it can be inferred that the player who wants to win the LPGA should aim to increase GIR, reduce PA, and improve driving distance and accuracy through training to increase the birdies chance at each hole, which can lead to lower average strokes and increased possibility of being within T10.

Key words: *artificial neural network analysis, greens-in-regulation (GIR) increase, putting average (PA), birdies chance, prediction models.*

Introduction

Golf officials, as well as fans, are always interested in the result of each golfing event, and they become aware of it mostly through the press and/or media. The broadcasters and commentators cautiously predict the winner and winning factors, especially, in the Ladies Professional Golf Association (LPGA) majors. Golf fans also judge the result based on the performance of each player. Expert performance-analysis scholars attempt to determine the winning factors and the performance factors that affect the money leader, based on the updated LPGA longitudinal data of many years. It was reported in several research papers that greens in regulation (GIR) and putting average (PA) had higher contributions and were more important that the other factors affecting the average strokes, money leaders or winning (Chae and Park, 2017; Dodson et al., 2008; Finley and Halsey, 2004; Park and Chae, 2016).

In most sports competitions, strategy analysts for each team invest efforts to analyze the records and data of the home and away teams to equip coaching staff with decisive factors that can affect the outcome of the game. These efforts are the same in the LPGA as in various other fields, and skill information such as the length of the game field, types or lay of the land, the level of

[1] - *Measurement and Evaluation in Physical Education and Sports Science, Yongin University, Yongin-si, Republic of Korea.*
[2] - *Department of Human Movement Science, Seoul Women's University, Seoul, Republic of Korea.*
[3] - *Sports and Health Care Major, College of Humanities and Arts, Korea National University of Transportation, Chungju-si, Republic of Korea.*

Authors submitted their contribution to the article to the editorial board.
Accepted for printing in the Journal of Human Kinetics vol. 77/2021 in January 2021.

difficulty of the course, the type of grass and green conditions, weather, and strategy for course targeting, is provided (McGarry et al., 2002). However, recently, prediction and description of the determinant of victory of the team and players, as well as the winner, have been required in sports competitions (Dorsel and Rotunda, 2001; Park and Chae, 2016).

This requirement has reached a level wherein scholars statistically provide winner and rank possibilities employing prediction models on accumulated data (Hayes et al., 2015; Jida and Jie, 2015; Neeley et al., 2009). Chae et al. (2018) used multiple regression analysis, which is a statistical analytical model, for the rank prediction of LPGA players based on the fact that the medal rank of the 2016 Rio Olympic female golf tournament was predicted by multiple regression analysis (Mercuri et al., 2017). The methods of analysis for this type of prediction are usually linear regression analysis, curve estimation, discriminant function analysis, logistic regression analysis, principal component regression analysis, classification tree analysis, and more recently, the frequently used artificial neural network analysis. Classification tree analysis, logistic regression analysis, discriminant analysis, and artificial neural network analysis, in particular, are generally used in quantitative prediction analysis (Agga and Scott, 2015; Cenker et al., 2009; Maszczyk et al., 2012, 2016; Neeley et al., 2009).

The discriminant function analysis is a statistical technique to predict how the individual would behave under given circumstances, based on various characteristics of social phenomena. Several types of supposition should be satisfied when using the discriminant function analysis (Couceiro et al., 2013; Kuligowski et al., 2016; Mieke et al., 2014; Shehri and Soliman, 2015). Classification tree analysis segments the individuals as members of small groups with similar behaviors or conducts stratification based on a certain standard and if the LPGA player will win, fail, or lead in wins (Surucu et al., 2016). Logistic regression analysis is a general linear model, wherein the object variable is a binary variable that is categorical data. Logistic regression analysis has an advantage that there are few constraints for the discrimination variable; however, there exists a regression-analysis–oriented disadvantage that it cannot overcome the interaction effect and the numbers of independent variables (Clark, 2001; Lu, 2017; Sperandei, 2014).

Artificial neural network analysis mimics the human neural–brain system. A typical neural network is composed of three layers, i.e., the input layer, hidden layer, and output layer, which include several neurons (Almassri et al., 2018). The neurons in the hidden layer conduct intermediate treatment if the input nodes receive stimulation, resulting in response from the output nodes. Thus, when using artificial neural network analysis, the predicting variable is applied to the input layer and the dependent variable to the output layer. The hidden layer oversees the intermediary management, and the researcher does not grant a role to a specific observed variable even though the researcher designates the number of hidden layers and neurons.

The back-propagation algorithm is applied between the input and hidden layers, and hidden and output layers, if the input variable is supplied to the neural network. The connection weight value is adjusted every time to minimize the error between the real value in the unit of the output after applying the back-propagation algorithm and the value calculated by the artificial neural network. The optimum point is investigated by applying the big learning rate from a random point by combining the intensity that affects the direction where the slope is highest by the algorithm (learning rate, $\eta > 0$) and the intensity that affects the direction from the initial to the current direction (moment, $\alpha > 0$) (Chen and Liu, 2014; Smaoui et al., 2018). Artificial neural network analysis is relatively independent of the statistical preconditions and can describe the nonlinear relationships between variables. Therefore, it is preferred over the traditional methods (Chen and Liu, 2014; Maszczyk et al., 2014; Sun and Lo, 2018).

Even though there are many prediction analysis methods, this study aims to investigate performance variables that affect the winning possibilities of players and the degrees of importance of these variables, from the annual data of 25 seasons of the LPGA (1993 to 2017). Moreover, it aims to select the most accurate model from four prediction models (classification tree analysis, logistic regression analysis, discriminant function analysis, and artificial neural network analysis). This study presents a relative comparison of the influence of the predicting

variables in the four prediction models on victory. That is, it tells the performance variable that should be considered for winning, and it can predict the possibility of victory of an individual using an optimum prediction model. The results of this study are expected to show the effect of prior preparation on victory.

Methods

Participants and Data Collection

The data used in this study included LPGA players, falling within the 60th rank (money leaders), from over a period of 25 years from 1993 to 2017; i.e., the annual average value of 1,500 players (60 players multiplied by 25 years). The data were collected from the LPGA homepage (http://www.lpga.com/). Because the data on the LPGA homepage did not collect private identifier information such as telephone numbers, home addresses, social security numbers, etc., ethical approval was not required for this experimental study. The performance variables chosen were those that were being measured and used in the current LPGA analyses. The variables were reconstituted in this study as independent variables (predicting variables), which were continuous variables, and dependent variables (response variables), which were categorical variables (Table 1).

Experimental Approach to the Problem

The data analysis aimed to determine key performance variables that affected the possibility of winning, the variable that was the most significant, and whether the player would win a game or be in the lead in wins. Four prediction models, i.e., classification tree analysis, logistic regression analysis, discriminant function analysis, and artificial neural network analysis, were employed. The most accurate model was selected, according to the purpose of the study.

Procedures

The player's accumulated raw data released by the LPGA were arranged using Microsoft Office Excel 2010 (Microsoft Corporation, Redmond, WA, USA), and the result was deduced using the IBM SPSS 22.0 (IBM Corp., Armonk, NY, USA) statistical program. In the first round of analysis, we used classification accuracy as a basis to find the possibility that a certain player could win the game in the LPGA, using the four prediction models (discriminant function analysis, classification tree analysis, logistic regression analysis, and artificial neural network analysis (multilayer perceptron, MLP)). One-way analysis of variance (ANOVA; post-hoc: least significant difference [LSD] test) was used if there was a mean difference in the classification accuracy of the four prediction models.

The input predicting variables of the four prediction models were divided into skill variables (driving accuracy [DA], driving distance [DD], sand saves [SS], GIR, and PA), skill result variables (birdies, eagles, par3 scoring average [P3A], par4 scoring average [P4A], and par5 scoring average [P5A]), and season outcome variables (official money [OM], scoring average [SA], top 10 finish% [T10], 60-strokes average [60SA], and rounds under par [RUP]). When inputting these predicting variables as dependent variables, they were divided into both two groups (victory/no victory) and three groups (no victory/one victory/multiple victories). From the results of the four prediction models, the standardized discriminant function coefficient, normalization importance or Wald value, which are importance indexes linking the independent variable to the dependent variable, could be obtained. Finally, one-way ANOVA and the post-hoc (LSD) test were conducted to examine the mean difference in the classification accuracy of the four models. Statistical significance was set at 0.05.

Statistical Analyses

In the discriminant function analysis, the function to maximize the group difference of an object based on continuous and discrete variables was deduced, and each participant (player) was classified using Fisher's linear discriminant functions (Mieke et al., 2014; Shehri and Soliman, 2015). It should be known which group, from among the many groups, included each object to be used in this model. When each group was already known, the category to which each object belonged was classified and predicted by calculating the discriminant score of the individuals in each group by finding the discriminant function:

$$\{Z_i = \alpha_0 + \beta_{i1}X_1(GIR) + \beta_{i2}X_2(PA) + \cdots + \beta_{i15}X_{15}\,(SA)\}$$

which could classify each group from the measured variables (Kuligowski et al., 2016; Novak, 2016; Schumm, 2006).

Classification tree analysis was used for classification and prediction by tree-structurally

schematizing the decision-making rule. A decision-making tree consists of a node, body, and stems that connect different nodes. The decision-making pattern is found at the top of the node if it repetitively classifies the node according to the tree structure forming process. Before the analysis using this decision-making tree, decision trees have an assumption that prior to analysis, the type of variable is precisely specified according to the measurement level. That is, it should be analyzed whether the variables have been accurately designated for the measuring levels (Surucu et al., 2016).

The methods of growing the tree are classified according to the characteristics of the data and purpose of decision making into chi-squared automatic interaction detection (CHAID), exhaustive CHAID, classification and regression tree (CRT), and quick, unbiased, and efficient statistical tree (QUEST). The classification accuracy was found to be high for the CRT basic data (Hayes et al., 2015). The tree structure was formed by designating the standard and pattern (decision trees are classified according to the purpose of the analysis and the structure of the data) as well as classifying for the purpose of analysis and data structuring. The decision tree is to select the predicting variable and to set the standard of the category when forming a low node from a single upper node. A pure low node was formed by most efficiently classifying the distribution of the target (dependent) variables. In this case, purity was defined as the degree of including individuals in a certain category of the target (dependent) variable. It set the predicting model according to the analysis result and interpreted by grasping the meaning of certain parts, as the decision-making tree described the relationships between variables as tree structures (Linda et al., 2008; Neeley et al., 2007).

The merit of this study is that the process is simpler than the other methods (artificial neural network analysis, discriminant function analysis, regression analysis, and so on), as prediction or classification is described based on the induction rule of the tree structure. In this study, CRTs of four tree-growing methods were used. Homogeneity within nodes was maximized by dividing the parent node for maximum homogeneity of the dependent variable within the child node (Hayes et al., 2015). In the splitting criterion of the classification tree, the status to merge the input variable selection and category when each parent branch formed a child branch was a criterion, and it was processed from the input variable, grasping distribution of the target variable, and child branch forming in sequence (i.e., first from the input variable, then from the grasping distribution, etc.). The degree of classifying the distribution of target variables was measured in terms of the purity or impurity. The purity of the child branch was very high, compared to that of the parent branch. Pruning removed the branch that had high risk of misclassification or inappropriate induction rules.

There is cross-validation and split-sample validation for the validity evaluation. Namely, cross-validation and split-sample validation existed in the assessment of validity. The analytical sample was divided into m (= 2, 3, 4 ...) parts, and the remaining part of the sample was excluded. Thus, each part of the data was used to generate m-1 trees, and 1 was used to evaluate trees. That is, this study used cross-validation that divided analysis samples into parts of m values, made the tree with the rest of certain parts of m values, and conducted model assessment with the remaining one part. Split-sample validation divided the observation samples into training samples (training: 70%) and test samples (test: 30%) and conducted an assessment of the tree with the test samples after making the tree with the training samples. This means that the produced tree, without just being a sample, can perform expended application to a population, which is the origin of the analysis sample. Model assessment could be described with profit charts or risk charts. Namely, the decision tree found the hidden pattern and useful correlations using data and could be used as a reference for decision making in the future, as well as for finding associations between data that were difficult to quantify accurately (Duan et al., 2015).

In logistic regression analysis, variables measured by nominal, ordinal, interval, and ratio scales could be used as independent variables; however, the dependent variables had to be categorical variables that were measured in a nominal scale to analyze and predict whether an individual observation belonged to a certain group. The functional formula of the logistic technique was

$$E(YIX) = P(X) = \frac{e^{b_0+b_1X_1+b_2X_2+\cdots+b_nX_n}}{1+e^{b_0+b_1X_1+b_2X_2+\cdots+b_nX_n}}$$

which was expressed as $P(X)$ when the Y value predicted using X was $E(YIX)$ and $E(YIX)$ had a probability concept when Y was a discrete variable (Pang et al., 2017; Sperandei, 2014). This model was not a linear function, but an S-curve logistic function with an upper limit of 1 and a lower limit of 0, with a problem in analysis as it could not be described as a linear function (Agga and Scott, 2015). The upper and lower limits could be avoided if this probability was converted to logit. The logit relationship with the independent variable can be described by a linear function (Cenker et al., 2009; Zhao et al., 2015)

$$\ln\left(\frac{p}{1-p}\right) = \alpha_0 + \beta_1X_1(DA) + \beta_2X_2(DD) + \beta_3X_3(GIR) + \cdots + \beta_nX_n\ (P5A)$$

resulting in ability possibility for linear regression analysis. Thus, the natural log value in brackets, which is on the left-hand side of this logit linear function is an odds-ratio; p, which is the numerator, is the probability that an individual belongs to a certain group; and $1-p$, which is the denominator, is the probability that an individual does not belong to a certain group. Thus, as a result of calculation using n predicting variables (X) in the right-hand side, the bigger the logit value, the higher is the possibility it belongs to the group (Curtis, 2019).

Artificial neural network analysis, by using learning materials in computers, aims to learn the optimum result, apply that result of learning to new data or conditions, and deduce an expected result such as how a human behaves, through learning (Chen and Liu, 2014). The neural network used in this study was composed of three layers (input layer, hidden layer, and output layer), and each layer included several neurons (Chen and Liu, 2014; Nair et al., 2016). The neurons in the hidden layer received the stimulation (every type of information) from the neurons in the input layer and the linear combination

$$L = \omega_1X_1(DA) + \omega_2X_2(DD) + \cdots + \omega_{10}X_{10}(P5A)\omega$$

was connected as a weighted value. The bigger this linear combination, the higher the activation the neuron received; it was deactivated in the opposite case (Almassri et al., 2018; Nair et al., 2016).

If the degree of this activation value was S, the activation {logistic functions: $S=\frac{e^L}{1+e^L}, (0 \le S \le 1)$} and hyperbolic tangent functions: $\{S = \frac{e^L+e^L}{e^L+e^{-L}}, (-1 \le S \le 1)\}$ were intervened to $S = f\ (L)$, a conversion from L to S to enable S to take a limited range ($0 \le S \le 1, -1 \le S \le 1$). The output node produced the final response by combining signals from the hidden neuron as weighted values. It applied the weighted value combination of the signal when the target variable was continuous, but it was calculated after converting to probability value for softmax conversion (softmax: $O_k = \frac{\exp(L_k)}{\sum_{j=1}^{k}\exp(L_j)}$) to enable all categorical output values to show the probability value when k was categorical value (Nair et al., 2016), where k was the output range index and k was the output range (Li et al., 2017). In this study, the output group was two (victory/no victory) or three (no victory, one victory, multiple victories); thus, k was 2 or 3.

The goodness-of-fit of the neural network was obtained by maximizing the corresponding likelihood function using the back-propagation algorithm. Conceptually, this algorithm attempts efficient calculation by combining the learning rate (the intensity in the direction where the slope is the highest) and moment (the intensity in the direction until now) (Jida and Jie, 2015; Smaoui et al., 2018). Namely, the neural-network fitting algorithm was started from a random location, and it actively explored the highest point using a high learning rate at the beginning. It gradually lowered the learning rate to reach the highest point (Sun and Lo, 2018; Xi et al., 2013). This process was repeated at the other locations. The point finally reached by repeating this process dozens of times was not the local highest point, but the global highest point (Nair et al., 2016). It found a weight parameter for which the probability became the maximum. The predicting variable was set to skill (DA, DD, SS, GIR, PA), skill result (birdies, eagles, P3A, P4A, P5A), and season outcome variables (OM, SA, T10, 60SA, RUP), and the dependent variable was categorized to no victory and victory or no victory, one victory, and multiple victories.

Results

Influence of Skill Variable on Achieving Victory

The type of an athlete that belongs to a certain group can be predicted using different models. Namely, it is possible to predict which

athlete will belong to which group using a prediction model. Table 2 solves this problem when it comes to the probability of victory between an LPGA rookie and a veteran. Table 2 categorizes the dependent variables according to victory (Yes/No) from the results of four prediction model tables, when the independent prediction variable was set to a skill variable such as DA, DD, SS, GIR or PA.

This discriminant function was significant as the Wilks'λ test statistic was 0.883 ($p < 0.001$). The classification accuracy of this discriminant function was 74.1% and the importance of the prediction variables was in the order of SS < DA < DD < PA < GIR. The validity evaluation of classification tree analysis, the second model, was described by risk estimates. The misclassification rate was 26.4% and 27.2%, when the classification tree model included training data of the sample and cross-validation, respectively. Namely, this misclassification is a value divided by the misclassified values ((59+335) / 1500), and the total classification accuracy of this model was 73.7%. The importance of the prediction variables was in the order of DA < SS < DD < PA < GIR.

In the goodness-of-fit test of the third model, the binomial logistic regression model, the model was found to be better than the base model, as chi-square (x^2) was 186.83, which was significant ($p < 0.001$). The classification accuracy of this model was 74.2%, and the importance of the predicting variables was in the order of SS < DA < DD < PA < GIR.

The goodness-of-fit of the fourth model, the artificial neural network analysis model, was determined by the area under the curve (AUC), and the model improved as the AUC became closer to 1. The AUC of this study model under the receiver operating characteristic (ROC) curve could fall in two categories: 0.736, a group with winning experience, and 0.736, a group without it. With higher accuracy of prediction, the shape of the ROC curve moved further up from the 45° line. The AUC was the area under the ROC curve, the 45° line was a curve corresponding to the random classification ratio, and the AUC was 0.5.

Thus, the AUC was in the range of 0.5 to 1.0, if it was superior to the random classification, and it became close to 1 for a more accurate model. The probability value was calculated by applying the importance index of each predicting variable to the hyperbolic tangent function between the input and hidden layers. If the hidden layer was formed and the weight coefficient value of the variable that belonged to the hidden layer was applied to the softmax function that was applied between the hidden and output layers, the probability value that corresponded to each category (Yes/No) of the finally calculated dependent variable changed from 0 to 1, and group classification criteria could be applied to the classification standard of the group by estimating the sum of probability to 1.0.

The classification accuracy rate from these repeated processes was 75.3%. The importance of predicting variables in this model was in the order of SS <DD < DA < PA < GIR. To sum up, artificial neural network analysis showed a higher prediction accuracy rate than the other three models; i.e., prediction accuracy rates were as follows: classification tree model (73.7%) < discriminant model (74.1%) < binominal logistic regression model (74.2%) < artificial neural network model (75.3%). Moreover, predicting variables that were most significant for determining victory included GIR and PA in all four prediction models (Table 2).

Influence of Skills on Victory

If an LPGA player needs to determine the possibility of victory in a tour, the results in Table 3 will help solve this problem (or will help provide this information). Table 3 is a result table for the four prediction models, based on the category of victory (Yes/No) and the predicting variable, which is an independent variable composed of the skill variables: birdies, eagles, P3A, P4A, and P5A. The discriminant model discriminated between the groups to which each participant belonged, using the coefficient value of the discriminant function. This discriminant function was significant as the test statistic Wilks'λ was 0.879 ($p < 0.001$).

The classification accuracy of this discriminant function was 74.1% and the importance of the predicting variables was in the order of eagles < P4A < P3A < P5A < birdies. The feasibility study of the second model, the classification tree model, is described by the risk estimate. In the training data of the samples, the misclassification rate of this model was 25.6% and cross-validation showed 26.1% misclassification. Namely, this misclassification was a value divided by misclassified (112+272)/1500, and the total

classification accuracy of this model was 74.4%. The importance of predicting variables was in the order of eagles < P5A < P4A < P3A < birdies. The goodness-of-fit of the binominal logistic regression model was better than that of the base model, as the chi-square value (x^2) was 188.04, which was significant ($p < 0.001$).

The classification accuracy of this model was 74.3%, and the importance of predicting variables was in the order of P4A < eagles < P3A < P5A < birdies. In the artificial neural network analysis goodness-of-fit test, the AUC, which was the area under the ROC curve, could take values in two groups: a group with winning experience (0.733) and a group without any experience of victory (0.733). If it was superior to random classification, the AUC was between 0.5 and 1.0, and the model improved as the AUC increased and reached closer to 1; the AUC was 0.5 for this model. If the hidden layer was formed and the weight coefficient value of the variable that belonged to the hidden layer was applied to softmax function that was applied between the hidden and output layers, the probability value that corresponded to each category (Yes/No) of the finally calculated dependent variable changed from 0 to 1, and could be applied to the classification standard of the group by estimating the sum of probability to 1.0.

The classification accuracy rate from these repeated processes was 75.7%. The importance of predicting variables in this model was in the order of eagles < P3A < P5A < P4A < birdies. To sum up, artificial neural network analysis showed higher prediction accuracy rates than the other three models, as in the discriminant model (74.1%) < binominal logistic regression model (74.3%) < classification tree model (74.4%) < artificial neural network model (75.7%). Moreover, the predicting variable that was most important in determining the victory was found to be birdies in all four predicting models (Table 3).

Influence of the Season Outcome on Victory

The data in Table 4 help a player determine the possibility of victory during the LPGA tour. Table 4 is a result table of the four prediction models and the predicting variable is a season outcome such as OM, SA, T10, 60SA, and RUP. The dependent variable is victory (Yes/No).

The discriminant model discriminated between the groups to which each participant belonged, based on the coefficient value of the discriminant function. This discriminant function was significant as the Wilks' λ test statistic was 0.717 ($p < 0.001$). The classification accuracy of this discriminant function was 78.5% and the importance of the predicting variables was in the order of SA < RUP < 60SA < OM < T10. The evaluation of the validity of the second model, the classification tree model, was described by risk estimates. The misclassification ratio of the model when the sample was training data was 20.3% and cross-validation showed 21.3% misclassification. Namely, this misclassification was a value divided by the wrongly classified (137+167) / 1500, and the total classification accuracy of this model was 79.7%. The importance of predicting variables was in the order of 60SA < RUP < SA < OM < T10. In the goodness-of-fit test of the binominal logistic regression model, the model fit improved compared to the base model as the chi-square (x^2) of the analysis model was 477.262, which was significant ($p < 0.001$).

The classification accuracy of this model was 78.7%, and the importance of the predicting variables was in the order of 60SA < SA < RUP < T10 < OM. In the artificial neural network analysis goodness-of-fit test, the AUC could be in two different groups: a group with winning experience (0.844) and a group without any winning experience (0.844). If it were superior to random classification, the AUC would be between 0.5 and 1.0, and the model improved as the AUC increased and reached closer to 1; the AUC was 0.5 for this model. If the hidden layer was formed and the weight coefficient value of a variable that belonged to the hidden layer was applied to the softmax function between the hidden and output layers, the probability value that corresponded to each category (Yes/No) of the finally calculated dependent variable changed from 0 to 1. Furthermore, this value could be applied to the classification standard of the group by estimating the sum of the probability to 1.0.

The classification accuracy rate from these repeated processes was 80.2%. The importance of predicting variables in this model was in the order of 60SA < RUP < T10 < SA < OM. To sum up, the artificial neural network analysis showed a higher prediction accuracy rate than the other three models, as in the discriminant model (78.5%) < binominal logistic regression model (78.7%) < classification tree model (79.7%) <

artificial neural network model (80.2%). Moreover, predicting variables that were most significant in determining victory were T10 and OM in the discriminant model and classification tree, and OM, T10, and SA in the binominal logistic regression model and artificial neural network model (Table 4).

Test of Mean Difference of Classification Accuracy of Prediction Models

Table 5 shows the best model in terms of the classification accuracy from the four prediction models, showing the mean difference in the classification accuracy of the statistic models, arising from the change in the number of independent variables according to the change in the dependent variable level (2 or 3). The test of mean difference of the classification accuracy ratio was conducted by one-way ANOVA and it was significant ($p < 0.05$). The post-hoc test was necessary to determine the exact difference between the prediction models. The LSD post-hoc test showed that the artificial neural network model had higher classification accuracy than the other three models.

Table 1

Research variables between 1993 and 2017

Independent variable		Dependent variable (winning odds)
Technical variables	1. Driving accuracy (DA) 2. Driving distance (DD) 3. Sand saves (SS) 4. Greens in regulation (GIR) 5. Putting average (PA)	1. No Win, 2. Win
		1. No Win, 2. Win, 3. Multiple Wins
Technical result variables	1. Birdies 2. Eagles 3. Par3Scoring Average (P3A) 4. Par4Scoring Average (P4A) 5. Par5ScoringAverage (P5A)	1. No Win, 2. Win
		1. No Win, 2. Win, 3. Multiple Wins
Season result variables	1. Official money (OM) 2. Scoring average (SA) 3. Top 10 finish% (T10) 4. 60-strokes average (60SA) 5. Rounds under par (RUP)	1. No Win, 2. Win
		1. No Win, 2. Win, 3. Multiple Wins

Table 2

Comparison of importance of skill variables for victory in four predicting models

Discriminant model		Classification tree			Binary logistic regression model		Artificial neural network model		
Wilks' Λ: 0.883 χ^2: 186.4 $df: 5, p < 0.001$		Risk estimate: 0.263 Cross test: 0.272			Model coefficient Test χ^2: 186.83 $df: 5, p < 0.001$		ROC curve Experience of winning Yes : 0.736 No : 0.736		
IV	SDFC	IV	Importance	NI	IV	Wald	IV	Importance	NI
GIR	-1.370	GIR	0.037	100%	GIR	134.12***	GIR	0.447	100%
PA	0.683	PA	0.019	51.3%	PA	56.57***	PA	0.206	46.1%
DD	0.552	DD	0.004	9.9%	DD	27.21***	DA	0.170	37.9%
DA	0.398	SS	0.001	3.7%	DA	16.39***	DD	0.144	32.2%
SS	0.036	DA	0.001	2.2%	SS	0.237	SS	0.034	7.5%

**p < 0.05, **p < 0.01, ***p < 0.001, ROC: receiver operating characteristic, IV: independent variable, SDFC: standardized discriminant function coefficient, NI: normalization Importance, GIR: greens in regulation, PA: putting average, DD: driving distance, DA: driving accuracy, SS: Sand Saves*

Item		Discriminant model			Classification tree			Binary logistic regression model			Artificial neural network model		
Sample		Predicted			Predicted			Predicted			Predicted		
		No win	Win	Total	No win	Win	Total	No win	Win	Total	No win	Win	Total
n	No Win	979	58	1037	978	59	1037	979	58	1037	958	79	1037
	Win	330	133	463	335	128	463	329	134	463	291	172	463
CA%		74.1%			73.7%			74.2%			75.3%		

No win : A group with no win, Win : A group with more than one win, CA% : classification accuracy%

Table 3

Comparison of importance of skill results for victory in four predicting models

Discriminant model		Classification tree			Binary logistic regression model		Artificial neural network model		
Wilk's Λ : .879 χ^2 : 192.41 df : 5, $p < 0.001$		Risk Source: Training; Cross test Estimate: 0.256; 0.261 Standard error: 0.011; 0.011			Model coefficient test χ^2: 188.04 df: 5, $p < 0.001$		ROC curve Experience of winning Yes : 0.733 No : 0.733		
IV	SDFC	IV	Importance	NI	IV	Wald	IV	Importance	NI
Birdies	0.963	Birdies	0.060	100%	Birdies	105.3***	Birdies	0.362	100%
P5A	0.703	P5A	0.014	24.0%	P5A	18.92***	P5A	0.196	54.2%
P3A	-0.441	P3A	0.009	14.5%	P3A	12.69***	P3A	0.175	48.3%
P4A	-0.282	P4A	0.007	11.2%	P4A	5.21*	P4A	0.171	47.2%
Eagles	0.209	Eagles	0.001	1.5%	Eagles	2.25	Eagles	0.096	26.5%

**p < 0.05, **p < 0.01, ***p < 0.001, ROC: receiver operating characteristic, IV: independent variable, SDFC: standardized discriminant function coefficient, NI: normalization importance, P3A: par3 scoring average, P4A: par4 scoring average, P5A: par5 scoring average*

Item		Discriminant Model			Classification tree			Binary logistic regression model			Artificial neural network model		
Sample		Predicted			Predicted			Predicted			Predicted		
		No win	Win	Total	No win	Win	Total	No win	Win	Total	No win	Win	Total
n	No Win	976	61	1037	925	112	1037	979	58	1037	976	61	1037
	Win	327	136	463	272	191	463	327	136	463	303	160	463
CA%		74.1%			74.4%			74.3%			75.7%		

No Win : A group with no wins, Win : A group with more than one win, CA% : classification accuracy%

Table 4

Comparison of importance of season outcomes for victory in four prediction models

Discriminant model		Classification tree			Binary logistic regression model		Artificial neural network model		
Wilk's Λ : .717 χ^2: 498.09 $df: 5, p < 0.001$		Risk Source: Training; Estimate: 0203; Standard error: 0.010 Source: Cross test; Estimate: 0213; Standard error: 0.011			Model coefficient test χ^2: 477.26 $df: 5, p < 0.001$		ROC curve Experience of winning Yes :0.844 No : 0.844		
IV	SDFC	IV	Importance	NI	IV	Wald	IV	Importance	NI
T10	0.663	T10	0.122	100%	OM	80.380	OM	0.401	100%
OM	0.657	OM	0.104	85.6%	T10	69.294	SA	0.263	65.6%
60SA	-0.147	SA	0.061	50.4%	RUP	2.925	T10	0.193	48.3%
RUP	-0.132	RUP	0.060	49.6%	SA	2.342	RUP	0.075	18.6%
SA	0.055	60SA	0.050	41.4%	60SA	1.916	60SA	0.069	17.1%

**p < 0.05, **p < 0.01, ***p < 0.001, ROC: receiver operating characteristic, IV: independent variable, SDFC: standardized discriminant function coefficient, NI: normalization importance, T10: top 10 finish%, OM: official money, 60SA: 60-strokes average, RUP: rounds under par, SA: scoring average*

Item		Discriminant model			Classification tree			Binary logistic regression model			Artificial neural network model		
Sample		Predicted			Predicted			Predicted			Predicted		
		No win	Win	Total	No win	Win	Total	No win	Win	Total	No win	Win	Total
n	No win	977	60	1037	900	137	1037	960	77	1037	948	89	1037
	Win	262	201	463	167	296	463	243	220	463	208	255	463
CA%		78.5%			79.7%			78.7%			80.2%		

No win: A group with no wins, Win: A group with more than one win, CA%: classification accuracy%

Table 5

Examination of accuracy and mean difference of prediction models

Dependent variable	Independent variable	Discriminant model (%)	Classification tree model (%)	Binary logistic regression model (%)	Artificial neural network model (%)
No Win / Win	Technical variables (5)	74.1	73.7	74.2	75.3
	Technical results (5)	74.1	74.4	74.3	75.7
	Season results (5)	78.5	79.7	78.7	80.2
	Technical variables (5) + Technical results (5)	75.5	74.4	75.5	78.1
	Technical variables (5) + Season results (5)	79.7	79.7	80.3	81.2
	Technical results (5) + Season results (5)	79.1	79.7	79.2	82.8
	Technical variables (5) + Technical results (5) + Season results (5)	79.9	79.7	80.8	84.8
No Win/Win/Wins	Technical variables (5)	71.5	71	71.5	72.2
	Technical results (5)	72.5	72.3	72.2	72.5
	Season results (5)	73.7	74.8	78.7	80.4
	Technical variables (5) + Technical results (5)	72.9	72.3	72.7	77.2
	Technical variables (5) + Season results (5)	74.5	79.7	75.5	81.0
	Technical results (5) + Season results (5)	74.3	79.7	75.7	83.4
	Technical variables (5)+Technical results (5)+Season results (5)	74.7	75.4	75.7	86.0
Mean		75.36	76.18	76.07	79.34
Standard deviation		2.77	3.35	3.02	4.33
	Sum of square	*degree of freedom*	Mean square	F	*p*
Between group	132.291	3	44.097	3.760	0.016
Within group	609.781	52	11.727		
Total	742.071	55			
Post-analysis (least significant difference test)	Discriminant, classification tree, and binary logistic regression < artificial neural network model				

Discussion

The purpose of this study was to find the best model, in terms of the classification accuracy, from four prediction models using the annual average performance variable data of LPGA players within the 60th rank, over 25 seasons, and to compare the importance of the predicting variables according to the victory status of the four prediction models (Dodson et al., 2008; McGarry et al., 2002). We found that, first, the artificial neural network model showed a higher prediction rate than the other three models, when the independent variable was a skill variable and the dependent variable was the achievement of victory (Almassri et al., 2018; Jida and Jie, 2015). The prediction rate was in the order of the classification tree (73.7%) < discriminant model (74.1%) < binominal logistic regression model (74.2%) < artificial neural network model (75.3%). The most important predicting variables for determining victory were GIR and PA in all four prediction models.

Second, the artificial neural network model showed a higher prediction rate than the other three models when the independent variable was the skill result and the dependent variable was victory. The prediction rate was in the order of the discriminant model (74.1%) < binominal logistic regression model (74.3%) < classification tree model (74.4%) < artificial neural network model (75.7%). Moreover, the most important predicting variable for determining victory was birdies in all four prediction models.

Third, the artificial neural network model showed a higher prediction rate than the other three models when the independent variable was the season outcome and the dependent variable was victory. The prediction rate was in the order of the discriminant model (78.5%) < binominal logistic regression model (78.7%) < classification tree model (79.7%) < artificial neural network model (80.2%). The most important predicting variables for determining victory were T10 and OM in the discriminant and classification tree models, and OM, T10, and SA in the binomial logistic regression and artificial neural network model. To sum up the above three results, the player who aims for victory in the LPGA should have a chance of birdies at each hole by improving the GIR and PA, driving distance, and driving accuracy among skill variables, lowering the average strokes. This will increase the probability of being within T10 as well as the victory at each competition.

Fourth, the one-way ANOVA was conducted to find the best model in terms of the classification accuracy of the four prediction models and to test the mean difference of the classification accuracy rate rising from the change in the number of independent variables according to the change in the dependent variable level (2 or 3). The LSD post-hoc test showed that the artificial neural network model had higher classification accuracy than the other three models. We can conclude that the artificial neural network model was superior when comparing the classification accuracy rates of the predicting models. This is consistent with the results of another study using neural networks when the sports disciplines considered were basketball, soccer, and tennis (Chae et al., 2018). Future research can supplement the data for predicting variables and quantify the mental strength and teamwork that are difficult to quantify for achieving an optimum harmony of predicting variables.

Conclusions

The first practical implication relates to the prediction of the probability of victory in the LPGA using the artificial neural network model for achieving more meaningful results. The second implication is to arrange the schedule of training based on the DD, DA, GIR, SS, PA, and GIR if the player aims at victory in the LPGA tour. Furthermore, birdies are the most important skill result variable affecting victory as all four prediction models indicated birdies as the most important variable of victory. Thus, more time can be spent establishing a strategy for improving this skill.

Acknowledgements

The authors have no conflicts of interest to declare. This work was supported by a special research grant from Seoul Women's University (2021).

References

Agga GE, Scott HM. Use of generalized ordered logistic regression for the analysis of multidrug resistance data. *Prev Vet Med*, 2015; 121(3-4): 374-379

Almassri AMM, Wan Hasan WZ, Ahmad SA, Shafie S, Wada C, Horio K. Self-calibration algorithm for a pressure sensor with a real-time approach based on an artificial neural network. *Sensors (Basel),* 2018; 18(8), pii: E2561

Cenker E, Ugur B, Mutlu K, Oktay E. Artificial neural network, genetic algorithm, and logistic regression applications for predicting renal colic in emergency settings. *Int J Emerg Med*, 2009; 2(2): 99–105

Chae JS, Park J. 5 years per cycle performances according to the average of the difference between LPGA players and trend analysis. *Kor J Golf Studies*, 2017; 11: 19-33

Chae JS, Park J. So WY. Ranking prediction model using the competition record of ladies professional golf association players. *J Strength Cond Res*, 2018; 32(8): 2363–2374

Chen WB, Liu WC. Artificial neural network modeling of dissolved oxygen in reservoir. *Environ Monit Assess*, 2014; 186(2): 1203-1217

Clark RD 3rd. An analysis of players' consistency among professional golfers: a longitudinal study. *Percept Mot Skills*, 2001; 92(2): 575-585

Couceiro MS, Dias G, Mendes R, Araújo D. Accuracy of pattern detection methods in the performance of golf putting. *J Mot Behav*, 2013; 45(1): 37-53

Curtis D. A weighted burden test using logistic regression for integrated analysis of sequence variants, copy number variants and polygenic risk score. *Eur J Hum Genet*, 2019; 27(1): 114-124

Dodson L, Bisnauth R, James N. Information is power. *Nurs Manag (Harrow)*, 2008; 15(4): 14-19

Dorsel TN, Rotunda RJ. Low scores, top 10 finishes, and big money: an analysis of professional golf association tour statistics and how these relate to overall performance. *Percept Mot Skills*, 2001; 92(2): 575-585

Duan YB, Guo DL, Guo LL, Wei DF, Hou XG. Genetic diversity analysis of tree peony germplasm using iPBS markers. *Genet Mol Res*, 2015; 14(3): 7556-7566

Finley PS, Halsey JJ. Determinants of PGA tour success: an examination of relationships among performance, scoring, and earnings. *Percept Mot Skills*, 2004; 98(3 Pt 1): 1100-1106

Hayes T, Usami S, Jacobucci R, McArdle JJ. Using Classification and Regression Trees (CART) and random forests to analyze attrition: Results from two simulations. *Psychol Aging*, 2015; 30(4): 911–929

Jida X, Jie C. Design of a Thermoacoustic Sensor for Low Intensity Ultrasound Measurements Based on an Artificial Neural Network. *Sensors (Basel)*, 2015; 15(6): 14788–14808

Kuligowski J, Pérez-Guaita D, Quintás G. Application of Discriminant Analysis and Cross-Validation on Proteomics Data. *Methods Mol Biol*, 2016; 1362:175-184

Li H, Luo M, Zheng J, Luo J, Zeng R, Feng N, Du Q, Fang J. An artificial neural network prediction model of congenital heart disease based on risk factors: A hospital-based case-control study. *Med (Baltimore)*, 2017; 96(6): e6090

Linda F, Michael P, Hsiu-Ju L, Eleni R, Lisa G. Applying classification and regression tree analysis to identify prisoners with high HIV risk behaviors. *J Psychoactive Drugs*, 2008; 40(4): 447–458

Lu K. On logistic regression analysis of dichotomized responses. *Pharm Stat*, 2017; 16(1): 55-63

Maszczyk A, Gołaś A, Czuba M, Krol H, Wilk M, Goodwin J, Stastny P, Kostrzewa M, Zając A. EMG analysis and modelling of the flat bench press using artificial neural networks. *SAJRSPER*, 2016; 38(1): S95-S103

Maszczyk A, Gołaś A, Pietraszewski P, Roczniok R, Zajac A, Stanula A. Application of Neural and Regression Models in Sports Results Prediction. *Procedia - Social Behavior Sci,* 2014; 117: 482-487

Maszczyk A, Roczniok R, Waśkiewicz Z, Czuba M, Mikołajec K, Zajac A, Stanula A. Application of regression and neural models to predict competitive swimming performance. *Percept Mot Skills*, 2012; 114(2): 610-626

McGarry T, Anderson DI, Wallace SA, Hughes MD, Franks IM. Sport competition as a dynamical self-organizing system. *J Sports Sci*, 2002; 20(10): 771-781

Mercuri A, Pagliari M, Baxevanis F, Fares R, Fotaki N. Understanding and predicting the impact of critical dissolution variables for nifedipine immediate release capsules by multivariate data analysis. *Int J Pharm*, 2017; 518(1-2): 41-49

Mieke D, Barbara C, Pascal C, Andry V, Tanneke P, Lieven D. Posture class prediction of pre-peak height velocity subjects according to gross body segment orientations using linear discriminant analysis. *Eur Spine J*, 2014; 23(3): 530–535

Nair VV, Dhar H, Kumar S, Thalla AK, Mukherjee S, Wong JW. Artificial neural network based modeling to evaluate methane yield from biogas in a laboratory-scale anaerobic bioreactor. *Bioresour Technol*, 2016; 217: 90-99

Neeley ES, Bigler ED, Krasny L, Ozonoff S, McMahon W, Lainhart JE. Quantitative temporal lobe differences: autism distinguished from controls using classification and regression tree analysis. *Brain Dev*, 2007; 29(7): 389–399

Novak M. Sex Assessment Using the Femur and Tibia in Medieval Skeletal Remains from Ireland: Discriminant Function Analysis. *Coll Antropol*, 2016; 40(1): 17-22

Pang T, Huang L, Deng Y, Wang T, Chen S, Gong X, Liu W. Logistic regression analysis of conventional ultrasonography, strain elastosonography, and contrast-enhanced ultrasound characteristics for the differentiation of benign and malignant thyroid nodules. *PLoS One*, 2017; 12(12): e0188987

Park J, Chae JS. A study of women's golf performance variables using LPGA data (*in Korean*). *Kor J Golf Studies*, 2016; 10: 79-88

Schumm WR. A discriminant analysis of Whissell's New Testament data: on the statistical trail of the author of Hebrews. *Psychol Rep*, 2006; 98(1): 274-276

Shehri FA, Soliman KE. Determination of sex from radiographic measurements of the humerus by discriminant function analysis in Saudi population, Qassim region, KSA. *Forensic Sci Int*, 2015; 253:138.e1-6

Smaoui S, Ennouri K, Chakchouk-Mtibaa A, Sellem I, Bouchaala K, Karray-Rebai I, Mellouli L. Statistical versus artificial intelligence based modeling for the optimization of antifungal activity against Fusarium oxysporum using streptomyces sp. strain TN71. *J Mycol Med*, 2018; 28(3): 551-560

Sperandei S. Understanding logistic regression analysis. *Biochem Med (Zagreb)*, 2014; 24(1): 12-18

Sun Y, Lo B. An Artificial neural network framework for gait based biometrics. *IEEE J Biomed Health Inform*, 2019; 23(3): 987-998

Surucu M, Shah KK, Mescioglu I, Roeske JC, Small W Jr, Choi M, Emami B. Decision Trees Predicting Tumor Shrinkage for Head and Neck Cancer: Implications for Adaptive Radiotherapy. *Technol Cancer Res Treat*, 2016; 15(1): 139-145

Xi J, Xue Y, Xu Y, Shen Y. Artificial neural network modeling and optimization of ultrahigh pressure extraction of green tea polyphenols. *Food Chem*, 2013; 141: 320-326

Zhao RN, Zhang B, Yang X, Jiang YX, Lai XJ, Zhang XY. Logistic regression analysis of contrast-enhanced ultrasound and conventional ultrasound characteristics of sub-centimeter thyroid nodules. *Ultrasound Med Biol*, 2015; 41: 3102-3108

Corresponding author:

Prof. Wi-Young So
Sports and Health Care Major, College of Humanities and Arts,
Korea National University of Transportation, Chungju-si, Republic of Korea,
E-mail: wowso@ut.ac.kr