The Korean Journal of Applied Statistics (2017) DOI: https://doi.org/10.5351/KJAS.2017.30.1.041
30(1), 41–55

Prediction of golf scores on the PGA tour using statistical models

Jungeun Lim[a] · Youngin Lim[a] · Jongwoo Song[a,1]

[a]Department of Statistics, Ewha Womans University

(Received September 21, 2016; Revised December 8, 2016; Accepted December 27, 2016)

Abstract

This study predicts the average scores of top 150 PGA golf players on 132 PGA Tour tournaments (2013–2015) using data mining techniques and statistical analysis. This study also aims to predict the Top 10 and Top 25 best players in 4 different playoffs. Linear and nonlinear regression methods were used to predict average scores. Stepwise regression, all best subset, LASSO, ridge regression and principal component regression were used for the linear regression method. Tree, bagging, gradient boosting, neural network, random forests and KNN were used for nonlinear regression method. We found that the average score increases as fairway firmness or green height or average maximum wind speed increases. We also found that the average score decreases as the number of one-putts or scrambling variable or longest driving distance increases. All 11 different models have low prediction error when predicting the average scores of PGA Tournaments in 2015 which is not included in the training set. However, the performances of Bagging and Random Forest models are the best among all models and these two models have the highest prediction accuracy when predicting the Top 10 and Top 25 best players in 4 different playoffs.

Keywords: PGA tour, golf, average score, linear regression, tree, bagging, gradient boosting, neural network, random forest, KNN, FedExCup

1. 서론

대한골프협회(http://www.kgagolf.or.kr)가 지난 2012년에 이어 [2014 한국 골프 지표]를 발표하였다. 이 지표는 현재 한국의 골프 인구 이용 형태, 골프 활동 유형 및 해외 골프 활동 형태에 대한 내용을 담고 있다. 우리나라 20세 이상 인구 중 619만 명이 골프를 해본 경험이 있는 것으로 조사되었으며 2014년 한 해에만 531만 명이 골프를 하였다. 골프 활동인구가 2012년에 470만 명이었던 것에 비하면 2년 사이에 무려 61만 명이 늘어난 것이다. 성별로는 남자가 71%, 여자가 29%였고, 연령별로는 40대의 골프 인구가 가장 많았다. 자료에 따르면 골프에 대한 흥미와 관심은 계속적으로 증가하는 추세이다. 특히 골프를 하는 이유에 대해서는(복수 응답) 응답자 절반 이상인 53.6%가 '취미 활동을 위해'라고 답

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the ministry of Education, Science and Technology (No. NRF-2015S1A5B6036244).
[1]Corresponding author: Department of Statistics, Ewha Womans University, 52, Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, Korea. E-mail: josong@ewha.ac.kr

했다. 이어 40.6%는 '건강을 위해', 29.9%는 '친분을 위해'라고 응답했다. 이는 골프가 하나의 취미 생활이 되었을 뿐만 아니라 주변인들과 어울리기 위한 문화가 되었음을 나타낸다.

골프와 관련된 연구 또한 국내와 해외 모두 활발하게 이루어지고 있는데 우선 해외 연구 현황을 보면 기술과 운 또는 심리적 압박감 등 다양한 변수들을 통해 PGA 선수들의 점수를 예측하려는 연구들이 많이 진행 되고 있다. Connolly와 Rendleman Jr. (2008) 연구에서는 1998년부터 2001년까지 253명의 PGA 선수들을 데이터로 하여 시간의 흐름에 따른 선수들의 기술과 운의 변화를 랜덤 이펙트 모형을 통해 추정하고 이를 순위에 반영하였다. Hickman과 Metz (2015)는 토너먼트 마지막 홀에서의 퍼팅 성공여부에 따른 상금 변화를 선수의 심리적 압박감으로 두고 보상이 클수록 좋지 않은 성적을 유발하는지에 대해 23,596개의 PGA선수 퍼팅 기록을 이용하여 분석하였다. 이외에도 PGA선수 개인의 기술과 필드의 강도 및 깊이, 랜덤 이펙트 등을 고려하여 PGA 투어 경기의 승리 요소를 분석한 논문도 있다 (Connolly와 Rendleman Jr., 2012). PGA 골프 선수 분석 사이트인 Golf Analytics(https://golfanalytics.wordpress.com)에서는 매달 골프 선수들의 기록에 대한 분석 결과를 홈페이지에 게재하고 있으며 골프선수의 스코어 예측 및 연령 곡선에 대한 연구를 활발히 진행하고 있다.

국내에서는 2006년에 한국 골프학회가 창설되어 2007년부터 한국골프학회지가 간행되기 시작하였다. Lee와 Lee (2014)의 연구에 따르면 한국골프학회지는 2007년부터 2014년까지 156편의 논문을 개제하였고 156편의 논문 중 118편(75.6%)이 자료처리 방법으로 통계기법을 사용하였다. 그러나 연구의 대부분이 일반인(43.5%)을 대상으로 진행되었으며 프로 선수에 대한 논문(10.3%)은 많지 않다. 또한 통계적 기법으로 프로 선수의 골프 스코어를 예측하는 모형에 대한 연구는 거의 이루어 지지 않고 있다.

따라서 본 연구에서는 데이터 마이닝 기법을 이용하여 PGA 투어에 출전하는 프로 골프 선수의 경기 결과를 예측하는 모형을 제안하고자 한다. 앞서 해외 논문들이 운이나 심리적 압박감을 변수로 사용하여 점수에 어떠한 영향을 미치는지에 초점을 맞추었다면, 본 연구에서는 선수 개인의 능력과 경기 환경들을 변수로 사용하여 다양한 데이터 마이닝 모형을 탐색했다는 점 그리고 예측력을 우선으로 했다는 점에서 차이가 있다. 분석은 통계프로그래밍 R (R Development Core Team, 2015)을 이용하여 이루어졌으며 분석에 사용한 모형은 선형회귀모형, 라소 회귀모형(LASSO) (Tibshirani, 1996), 능형 회귀모형 (Hoerl과 Kennard, 1970), 의사결정나무 (Brieman 등, 1984), 배깅(bagging) (Brieman, 1996), 랜덤 포레스트 (Breiman, 2001), 그래디언트 부스팅 (Friedman, 2002; Freund와 Schapire, 1997; Ridgeway, 2012), 주성분회귀(PCR) (Frank와 Friedman, 1993; Stone과 Brooks, 1990), K-최근접이웃방법(KNN) (Cover와 Hart, 1967), 신경망 (Günther와 Fritsch, 2010; Hastie 등, 2009; Park 등, 2011)이다. 선형회귀모형에서는 단계적 선택법(stepwise regression), 모든 가능한 회귀모형(all possible regression)과 같은 변수선택방법을 사용하여 총 11가지 모형을 사용하였다. 위의 모든 방법론은 R에 포함된 다양한 함수와 패키지를 이용하였고 예측력 평가지표로 제곱근평균제곱오차(root mean square error; RMSE)를 이용하였다.

본 논문의 2장에서 골프통계용어와 분석에 사용한 변수에 대해 자세히 설명하고 3장에서 다양한 데이터 마이닝 기법을 이용한 분석 결과와 골프 스코어를 예측하는 최종 모형을 제시하고자 한다. 마지막으로 4장에서는 본 연구의 결과를 요약하며 추후 발전방향에 대해 이야기할 것이다.

2. 분석자료 설명

2.1. 용어 정의

골프 경기는 총 4라운드로 구성되어 있는데 대부분의 경우에 한 라운드당 18개의 홀로 이루어져 있다. 하루에 한 라운드씩 나흘 간 총 72개 홀에서 경기가 진행된다. 대부분의 대회는 2라운드까지 치른 뒤 성

Table 2.1. Description of golf stat terminology

Terminology	Description
Green in regulation	규정타수 그린 온이라는 의미. 파3에서 한번에, 파4에서는 두 번 이내, 파5에서는 3번 이내에 공을 그린 위에 올릴 수 있는 확률
Scrambling	GIR에 실패한 후 파 또는 그 이하(버디, 이글 등)의 스코어를 만드는 능력
Sand save percentage	홀의 성적과 상관없이 단순히 그린주변의 벙커에서 그린에 올린 후 원펏 홀아웃하는 확률을 계산해 랭킹을 매기는 것
Total driving	비거리 랭킹과 드라이빙 정확도 랭킹을 산술적으로 합한 통계로, 두 가지의 랭킹이 높을수록(랭킹수치가 낮을수록) 성적이 좋은 것
Longest drive	한 해 동안 선수가 친 가장 긴 비거리
Driving distance	매 라운드당 2홀을 지정해 페어웨이 킵과 상관없이 드라이빙 거리를 합산해 티샷 수로 나눈 수치로 평균치가 높을수록 상위 랭킹
Driving accuracy	한 선수의 티샷이 페어웨이 킵한 확률을 비율로 표시한 랭킹
Ball striking	총 드라이빙(비거리 + 드라이빙 정확도)랭킹과 GIR 랭킹을 단순히 합한 산술합. 공을 얼마나 멀리 그리고 정확하게 잘 치는가에 대한 통계

적순으로 3라운드 진출자를 가리는데 그 기준 성적에 걸려 통과하지 못하면 컷오프(cut off) 당했다고 한다.

PGA 투어는 Professional Golf Association Tour의 약자로 1968년 미국 프로골프협회에서 독립하여 현재는 프로선수들의 토너먼트 대회를 운영하는 데 주력하고 있다. 해마다 10월 초에 첫 대회를 시작해 다음해 9월 말까지 진행되며 세계 각국 기업들이 스폰서로 나서 해마다 수많은 공식 대회를 치르는데 2015년 정규 시즌에는 52개 대회가 열렸다. 공식 대회 가운데 PGA 챔피언십, 마스터즈 토너먼트(Masters Tournament), US 오픈, 오픈 챔피언십 등 4개 대회를 가리켜 메이저대회라 부르며 여기서 모두 우승한 것을 그랜드슬램(Grand Slam)이라고 한다.

2.2. 자료수집

본 연구에 사용된 자료는 2013년 1월 7일에 개최된 Hyundai Tournament of Champions 경기를 시작으로 2015년 10월 11일 TOUR Championship by Coca-Cola 경기까지 총 132개(2013년 40개, 2014년 45개, 2015년 47개) 경기에 대한 선수 기록을 수집하였다. 본 연구에서는 출전 선수들의 기록을 2013년부터 2015년까지 추적하여 매 경기가 끝날 때마다 그 경기까지의 평균 기록으로 계속 업데이트를 해주었기 때문에 132개 경기에 출전한 모든 선수를 분석하기에는 지나치게 많은 시간이 소요될 뿐만 아니라 PGA 투어에 지속적으로 참여하지 않는 선수들 또한 포함하여 무의미하다고 보았다. 따라서 페덱스컵 포인트 기준 상위 150명 선수만을 분석에 사용하였고 이 중 경기 기록에 점수가 존재하지 않는 171건의 경우는 제거 하였다. 업데이트 방법은 2013년의 선수 기록을 개개인 별로 평균하여 2014년 첫 번째 경기의 설명변수로 사용하였고 2014년 이후 열리는 매 경기마다의 기록으로 새롭게 업데이트 하였다. 2013년부터 2015년까지 총 9,575개 자료 중 2013년의 기록으로 업데이트 해준 2014년과 2015년의 자료만을 분석에 사용하였으며, 2014년의 3,248개 기록을 훈련 자료로, 2015년의 2,956개 기록을 테스트 자료로 설정 하였다.

선수에 대한 정보는 PGA 투어(http://mediaguide.pgatourhq.com)에서 샷 링크 시스템(SHOT LINK SYSTEM)으로 측정한 데이터를 이용하였다. 샷 링크 시스템이란 PGA 투어 대회 대다수의 라운드에서 모든 샷의 정확한 결과를 기록하기 위한 시스템으로 약 350명의 자원봉사자들이 매주 참여하고 있다. 각 코스들은 레이저로 측정되어 페어웨이, 벙커, 워터해저드와 그린의 정확한 윤곽을 포함해 3D 입

체 지도로 만들어진다. 선수들의 경기 기록은 각 조의 기록원이 실시간으로 휴대용 장비에 입력하고 방송국과 투어의 샷트래커(Shot-Tracker)에게 그 정보가 전달된다. 잘못된 데이터의 입력은 투어 측에서 즉각적으로 오류를 잡아내고 교정할 수 있는 시스템을 갖추고 있기 때문에 본 연구에서 사용한 데이터는 신뢰할만하다고 볼 수 있다.

2.3. 변수 설명

골프는 개인적인 플레이 능력뿐만 아니라 날씨 및 경기장 코스까지 점수에 영향을 미치는 스포츠이다. 따라서 다양한 변수들이 골프 스코어에 영향을 미치기 때문에 경기력을 특정 하나로만 설명하기는 어렵다. 본 연구의 목적은 골프 선수에 대한 정보와 코스 정보, 바람에 대한 정보를 가지고 골프 스코어를 예측하는 것이며 반응변수는 해당 경기의 점수로 설정하였다. 분석에 사용된 설명변수는 총 47개로 아래에서 보다 자세히 설명하고자 한다.

2.3.1. 선수 관련 정보

- 연령
  선수 연령은 19세부터 52세까지 고르게 분포해 있으며 30대가 제일 많다(Table 2.3). 골프는 다른 스포츠와 달리 체력이 많이 요구되지 않아 연령이 점수에 크게 영향을 미치지 않다고 알려졌는데 이를 확인하기 위해 연령별로 컷오프 당하는 비율을 계산해 보았다. 그 결과 20대부터 40대까지는 컷오프 당하는 비율이 비슷하며 50대가 되었을 때 그 확률이 조금 증가하는 것을 확인 할 수 있었다. 하지만 그 차이가 0.1 정도로 크지 않고 표본이 1명으로 매우 적다는 것을 고려할 때 연령에 따라 컷오프 당하는 확률 차이는 크지 않다고 보인다. 10대의 경우 PGA Junior League에서, 50대의 경우 Senior PGA 투어인 Champions에서 활동하기 때문에 10대와 50대의 표본이 적은 것으로 판단된다.
- 스코어
  골프 선수들의 평균스코어를 기반으로 하는 모형은 골프 경기의 특성상 1, 2라운드에서 컷오프 당한 선수의 경우 3, 4라운드의 스코어가 존재하지 않으므로 앞의 두 라운드의 평균으로 스코어를 입력해 주었다. 컷을 통과한 선수들의 경우 4라운드 모두 값이 입력되어 있으므로 4라운드의 평균으로 입력하였다. Figure 2.1의 히스토그램과 표를 보면 71점을 중심으로 스코어들이 밀집되어 있으며 비교적 정규분포 형태를 띄고 있는 것을 확인 할 수 있다.

2.3.2. 코스 관련 정보 1934년 마스터스를 창설한 보비 존스는 골프코스에서 가장 중요한 조건은 퍼팅그린의 질이라 하였다. 그만큼 퍼팅 그린의 잔디 질이 스코어에 큰 영향을 미친다는 뜻으로 본 연구를 통해 그린뿐만 아니라 페어웨이의 잔디까지 스코어에 미치는 영향을 통계적으로 확인해 보고자 한다. 그린과 페어웨이 잔디의 단단함 정도를 5단계로 나누어 (soft - medium soft - medium - medium firmness - firmness) 설명변수로 이용하였다.

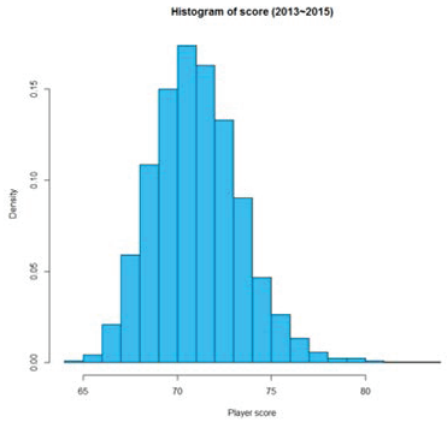
Figure 2.2를 보면 페어웨이의 단단함과 그린의 단단함에 따른 평균 점수의 형태가 매우 비슷하다. 이는 동일한 코스일 경우 페어웨이의 단단함과 그린의 단단함의 정도가 대체로 비슷하기 때문일 것으로 판단된다. 코스가 단단할 경우 공이 매우 빠르게 굴러가고 어디로 튈지 모르기 때문에 스코어가 증가할 수 있으며 반대로 부드러울 경우 공이 느리게 굴러가 원하는 방향으로 가지 않을 수 있다. 그러나 잔디 이외에도 실제 경기 도중 스코어에 영향을 미칠 수 있는 변수가 다양하며 코스의 난이도 또한 잔디와 별개로 스코어에 영향을 미칠 수 있다. 따라서 위의 상자그림만 가지고 잔디의 단단함에 따른 스코어 분포의 특정 패턴을 찾기는 쉽지 않다. 다만 잔디의 단단함 정도에 따른 스코어의 분포에 차이가 있다는 것은 확인할 수 있다.

Table 2.2. Description of variables

Variables	Description
Input variables	
Top.10 ratio	출전한 경기 중 상위 10위 안에 들었던 비율
Top.25 ratio	출전한 경기 중 상위 25위 안에 들었던 비율
Cuts.Made ratio	출전한 경기 중 컷을 안 당한 비율
Cuts.Missed ratio	출전한 경기 중 컷오프 당한 비율
Avg score	이전 경기까지의 총 타수를 총 라운드로 나눈 평균
Player Age	경기 당시 선수나이
FedExCup Points	이전 경기까지의 페덱스컵 포인트 평균
Money	이전 경기까지의 상금액의 평균
Finish Position	이전 경기까지의 선수 등수의 평균
Scoring Avg Tot Adjust	라운드 당 평균타수에서 파를 빼준 값을 이전 경기까지 평균
Eagles	이전 경기까지의 이글 수의 평균
Birdies	이전 경기까지의 버디 수의 평균
Pars	이전 경기까지의 파 수의 평균
Tot Holes Over Par	이전 경기까지 보기, 더블 보기, 더블 보기보다 샷을 더 많이 친 경우의 총합의 평균
Longest Drive	가장 멀리 친 샷의 거리(단위: 야드)의 평균
Ball striking	이전 경기까지의 총 드라이빙 랭킹과 GIR 랭킹의 합의 평균
Drives Over 300 Yards num Drives	이전 경기까지 비거리가 300야드 이상 나온 개수의 평균
Avg dist to hole after app	어프로치 샷을 했을 때 공이 떨어진 지점에서부터 홀까지 남은 거리를 이전 경기까지 평균
Avg fair proxi	페어웨이에서 샷을 했을 때 공이 떨어진 지점에서부터 홀까지 남은 거리를 이전 경기까지 평균
Avg rough proxi	러프에서 샷을 했을 때 공이 떨어진 지점에서부터 홀까지 남은 거리를 이전 경기까지 평균
scram ratio	GIR에 실패한 후 파 또는 그 이하의 스코어를 만든 횟수를 이전 경기까지 평균
Avg scram dist	GIR에 실패한 후 버디를 했을 때 홀까지 남은 거리를 이전 경기까지 평균(단위: 피트)
Going for the Green success	파4에서 첫 번째 샷이 또는 파5에서 두 번째 샷이 그린 위에 올라간 횟수를 이전 경기까지 평균
scram rough ratio	GIR에 실패한 후 러프에서 버디를 성공시킨 비율를 이전 경기까지 평균
scram fringe ratio	GIR에 실패한 후 프린지에서 버디를 성공시킨 비율을 이전 경기까지 평균
scram larg30 ratio	GIR에 실패한 후 홀에서 30야드 떨어진 거리에서 버디를 성공한 비율을 이전 경기까지 평균
scram 20.30 ratio	GIR에 실패한 후 홀에서 20–30야드 떨어진 거리에서 버디를 성공한 비율을 이전 경기까지 평균
scram 10.20 ratio	GIR에 실패한 후 홀에서 10–20야드 떨어진 거리에서 버디를 성공한 비율을 이전 경기까지 평균
scram less10 ratio	GIR에 실패한 후 홀에서 10야드 미만 떨어진 거리에서 버디를 성공한 비율을 이전 경기까지 평균
sandsave ratio	그린사이드 벙커에서 샷을 시도하여 2타 이내로 홀아웃을 한 비율을 이전 경기까지 평균
Tot Hole Outs	그린 또는 프린지가 아닐 때 홀아웃을 한 횟수를 이전 경기까지 평균
Putt Avg GIR Putts	GIR을 성공했을 때의 퍼팅 횟수에 대해 이전 경기까지 평균
Number of One Putt	퍼팅 한 번에 홀인 한 경우를 이전 경기까지 평균
Three Putt Avoidance	3번 이상 퍼팅하여 홀인 한 경우를 이전 경기까지 평균
putt 10 ratio	10피트 이내에서 퍼팅을 시도했을 때 성공한 비율를 이전 경기까지 평균
putt larg10 ratio	10피트 밖에서 퍼팅을 시도했을 때 성공한 비율를 이전 경기까지 평균
Tot Putts Gained	해당 선수의 홀아웃까지의 퍼팅 수를 나머지 PGA 투어 선수들의 홀아웃까지의 평균 퍼팅 수에서 뺀 값을 이전 경기까지 평균
TTL SG T2G	티에서 그린까지 해당 선수가 샷을 친 횟수를 나머지 PGA 투어 선수들의 평균 샷 횟수에서 뺀 값을 이전경기까지 평균
TTL SG Tot	티샷에서부터 홀아웃 할 때 까지 해당 선수가 친 샷의 총 횟수를 나머지 PGA 투어 선수들의 평균 총 샷 횟수에서 뺀 값을 이전 경기까지 평균
Fwy Firmness	페어웨이의 단단함 정도
Grn Firmness	그린의 단단함 정도
Grn Height	그린 존의 잔디 길이(단위: 인치)
Rough Height	러프 존의 잔디 길이(단위: 인치)
Fwy Height	페어웨이 존의 잔디 길이(단위: 인치)
Max INT	전체 4라운드 중 최대 풍속이 속한 구간
Max avg wind int	각 라운드마다 최대 풍속을 구해 평균한 풍속이 속한 구간
Avg wind int	4라운드 전체의 평균 풍속이 속한 구간
Response variable	
Score	해당 경기의 타수의 합계를 라운드로 나눈 평균

Table 2.3. The number of players and cut off rates by player's age

Age	10's	20's	30's	40's	50's	Total
Number of players (2013)	1	41	81	26	1	150
Number of games (2013–2015)	17	2536	5434	1484	104	9575
Cut off ratio	0.23529	0.33477	0.35149	0.35444	0.47115	



Summary	
Min	64.00
1^{st} Qu	69.50
Median	71.50
Mean	71.21
3^{rd} Qu	72.75
Max	84.00

Figure 2.1. Histogram of scores from PGA Tour 2013 to 2015.

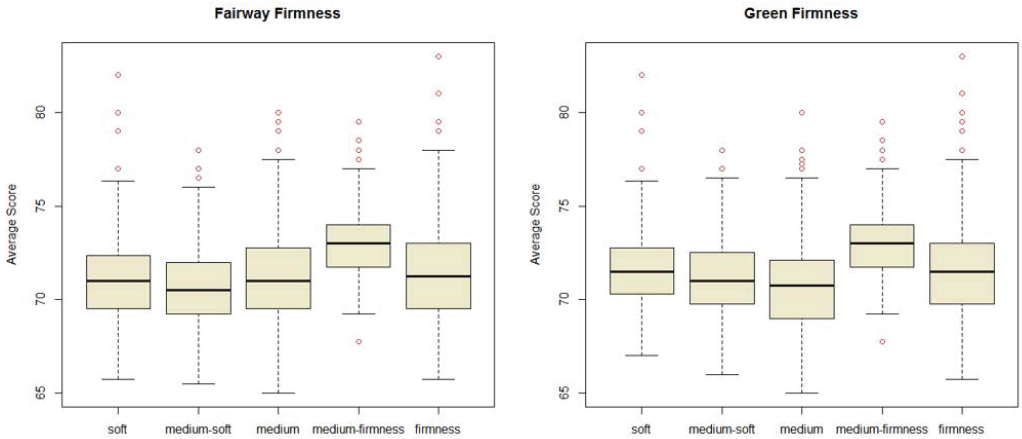


Figure 2.2. Box plots of average scores based on the fairway firmness and green firmness.

2.3.3. 바람 골프는 바람의 영향을 많이 받는 스포츠이다. 본 연구에서는 바람의 강도를 3가지로 나누어 설명변수로 적용하였다. 골프가 대체로 4일 간 열리는 운동이기 때문에 4일 동안의 최대 풍속과 4일 각각의 평균풍속의 최댓값, 그리고 4일 각각의 평균풍속의 평균을 구한 후 보퍼트 풍속 계급표를 참고하여 해당되는 구간 값을 지정해 주었다(Table 2.4).

4일 동안의 최대 풍속과 4일 각각의 평균풍속의 최댓값이 분석 시 유의한 변수로 항상 선택되기 때문에 두 변수에 대해서 Figure 2.3과 같이 상자 그림을 그려보았다. 실제로 바람의 세기가 강해질수록 평균 점수가 증가하는 경향을 보인다.

Table 2.4. Beaufort scale of wind force

Scale	Force rating	Observable land effects	Speed MPH
1	Calm, light air	Vertical smoke	1–3
2	Light breeze	Slight smoke drift	4–7
3	Gentle breeze	Leaves gently rustle	8–12
4	Moderate breeze	Leaves and twigs move	13–18
5	Fresh breeze	Raises paper moves small branches	19–24
6	Strong breeze	Sways large branches	25–31

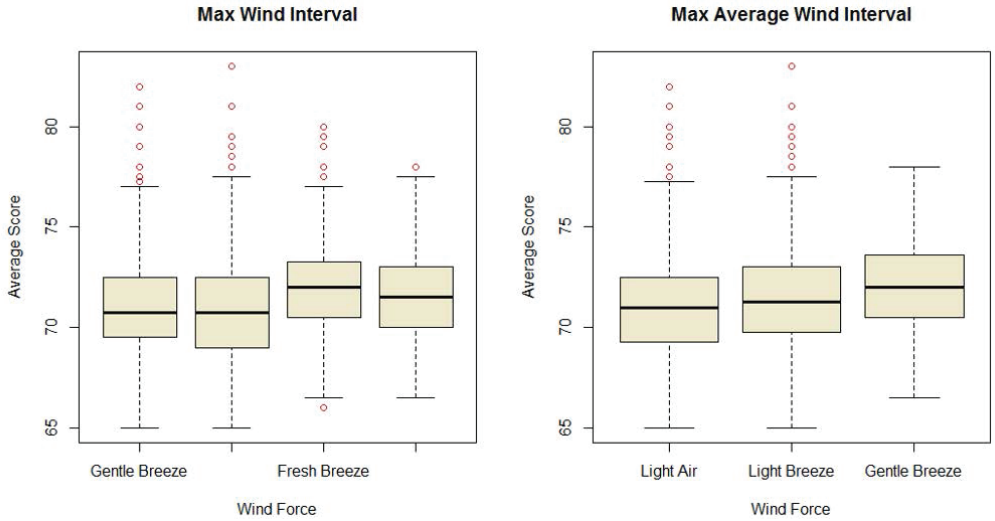


Figure 2.3. Box plots of average scores based on the max wind interval and max average wind interval.

3. 분석결과

이번 장에서는 통계적 분석 기법들을 이용하여 PGA 투어 경기에 출전한 골프 선수들의 평균스코어를 예측하는 모형을 적합해보고 어떠한 모형이 가장 높은 예측률을 주는지와 평균스코어에 영향을 미치는 변수가 무엇인지 파악하고자 한다. 적합된 모형의 예측력을 공정하게 비교하기 위해 10-fold 교차평가(cross validation) 과정을 100번 반복하였다. 앞으로의 분석은 각 모형에 10-fold 교차평가 방법을 적용하여 제곱근평균제곱오차 RMSE $= \sqrt{(\sum(Y_i - \hat{Y}_i)^2/n)}$를 계산하고 예측력을 비교한 후, 각 분석 방법에서 선택된 변수들을 도출하여 공통적으로 선택되는 중요변수들은 어떤 것이 있고 스코어에 어떤 영향을 미치는지 알아보는 것을 중심으로 진행할 것이다. 추가적으로 가장 예측력이 높은 2가지 모델을 가지고 페덱스컵의 4가지 플레이오프의 결과도 예측해 보고자 한다.

3.1. 평균스코어 예측 모형

평균스코어를 예측하기 위한 방법으로 선형회귀분석 방법과 비선형회귀분석 방법 두 가지로 나누어서 분석하였다. 선형회귀분석 방법으로는 라소, 능형회귀, 주성분회귀, 모든 가능한 회귀모형, 단계적 선택법을 사용하였고 비선형회귀분석 방법으로는 신경망 모형, 의사결정나무, 그래디언트 부스팅, 배깅, 랜덤 포레스트, K-최근접이웃방법을 사용하였다. 분석결과 선택된 변수에 대해서는 선형회귀분석과 비선

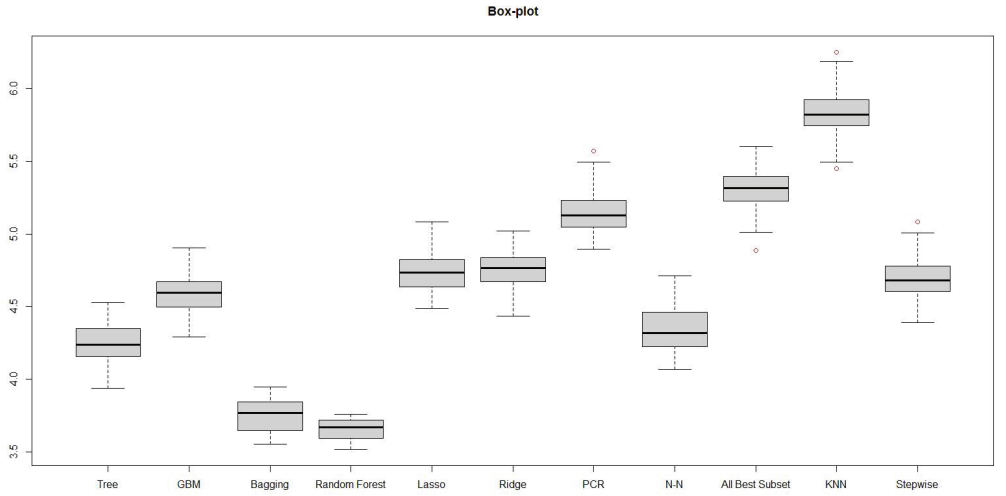


Figure 3.1. Box plots of cross validation errors (MSE).

형회귀분석으로 나누어 3.1.1장과 3.1.2장에 설명하고자 한다. 먼저 훈련용 자료를 이용해 위의 11가지 모형의 예측력을 비교해 보고자 한다. 교차오차의 평균은 2014년의 경기 기록인 훈련 자료를 각각의 방법에 대해 10-fold 교차평가를 100번 모의실험하여 구한 값이며 테스트 오차는 훈련 자료를 이용하여 만든 모형에 2015년의 2,956개 기록을 테스트 자료로 적합하여 구한 오차값이다. 튜닝 모수와 변수선택이 필요한 모형에 대해서는 10-fold 교차 오차가 가장 작은 모수와 변수를 최적으로 선택하였다. 최적의 튜닝모수로 랜덤 포레스트에서는 mtry = 16, 배깅에서는 mtry = 47이었다. 그리고 신경망 모형에서는 hidden = c(2,1), threshold = 0.01, stepmax = $1e+6$이었으며 그래디언트 부스팅에서는 n.trees = 500, shrinkage = 0.01이었다. 변수 선택 방법에서 라소회귀모형에서 shrink factor = 0.587, 능형회귀에서는 df = 10, 최근접이웃방법에서는 k = 19, 주성분회귀에서는 number of direction = 38이었다. 튜닝모수를 추정하는 방법에 대해서는 지면이 부족하기 때문에 Hastie 등 (2009)을 참조하길 바란다.

Figure 3.1과 Table 3.1을 보면 훈련용 자료에서의 교차오차 평균은 랜덤 포레스트와 배깅에서 1.9 정도로 2점이 채 되지 않으며 나머지 모형에서도 2점대로 좋은 결과를 보인다. 테스트 오차의 경우에도 모든 모형에서 2점대의 값을 가져서 훈련용 자료에서의 오차보다 아주 조금 클 뿐이다. 이것은 실제 평균스코어와 예측한 평균스코어가 평균적으로 2점 정도밖에 차이가 안 난다는 뜻으로 본 연구에서 만든 11가지 모형 모두 실제 스코어를 매우 정확하게 추정하고 있음을 알 수 있다. Figure 3.1의 경우에는 차이를 조금 더 잘 볼 수 있도록 제곱근평균제곱오차의 제곱값인 평균제곱오차(MSE)를 사용하여 상자 그림을 작성하였다.

3.1.1. 선형회귀분석(linear regression) 결과 우선 선형회귀모형에서의 결과에 대해 살펴보고자 한다. Table 3.2는 단계적 선택법, 라소(LASSO) 회귀모형, 능형 회귀모형을 통해 추정한 모든 변수들의 계수 값들을 정리한 표이다. 다른 선형회귀모형은 위의 3가지 방법론에 비해 예측오차가 조금 더 컸으므로 Table 3.2에는 정리하지 않았다. 그러나 Table 3.3에서는 중복으로 선택된 변수들을 정리하였는데 여기서는 5가지 선형회귀모형을 모두 반영하였다.

Table 3.2의 모든 회귀계수에 대한 설명은 어렵고 또한 모형마다 다른 값을 가지는 경우가 많으므로 우

Table 3.1. Mean cross validation error and mean test error

	CV error	Test error
Tree model	2.069011	2.336405
Gradient boosting model (GBM)	2.146143	2.251875
Bagging	1.937292	2.242819
Random forests	1.911875	2.177834
LASSO	2.175725	2.346154
Ridge regression	2.179784	2.372952
Principal component regression	2.268374	2.257777
Neural networks	2.103236	2.445885
All possible regression	2.304264	2.346420
K-nearest method	2.416738	2.428211
Stepwise regression	2.172492	2.368452

리는 5가지 모형에서 공통적으로 선택된 변수들을 정리해보았다. Table 3.3은 5가지 선형회귀모형에서 중복으로 선택된 변수들을 계수가 양수일 때와 음수일 때로 나누어 정리한 표이다. 회귀계수의 값이 양수이면 설명변수의 값이 커질수록 평균스코어가 증가한다는 뜻이다. 반대로 음의 계수를 가지는 변수들은 그 값이 증가할수록 평균스코어가 낮아짐을 의미한다. 양의 값을 가지는 회귀 변수로는 Fairway Firmness, Green Height, Max INT, Max average wind interval, Three Putt Avoidance가 선택되었다. 예를 들어 Fairway Firmness의 경우 페어웨이의 단단함 강도가 높을수록 평균스코어가 증가한다. 이는 2.3.2장에서 코스가 단단할 경우 공이 매우 빠르게 굴러가고 어디로 튈지 모르기 때문에 점수가 증가할 수 있다고 분석한 내용과도 일치한다. Green Height도 마찬가지로 그린의 잔디 길이가 길수록 공이 매끄럽게 굴러가기 힘들기 때문에 평균스코어가 증가하는 경향이 있음을 나타낸다. 그리고 최대 풍속과 최대 평균 풍속이 강할수록 평균스코어가 높아지게 되는데 이는 2.3.3장에서 분석한 바와 같다. Three Putt Avoidance는 퍼팅을 3번 이상 한 홀의 개수로 Three Putt Avoidance가 클수록 퍼팅을 3번 이상 한 홀이 많다는 뜻이다. 따라서 평균스코어를 증가시키는 변수이며 양의 계수를 가지게 된다. 반대로 음의 계수를 가지는 변수들은 Total Stroke Gained, Number of One Putt, scrambling 관련 변수들, Longest Drive가 있다. TTL SG Tot(Stroke Gained)는 다른 선수들에 비해 상대적으로 더 성적이 좋았음을 나타내는 변수이며 Number of One Putt은 퍼팅 한 번에 바로 골이 홀에 들어간 수이다. Three Putt Avoidance와는 반대로 Number of One Putt은 그 값이 증가할수록 평균스코어가 낮아지는 음의 계수 값을 지닌다. 또한 GIR에 실패한 후 버디나 이글로 스코어를 만드는 능력을 나타내는 지표인 scrambling 변수들도 음의 계수를 가진다. 즉 게임 회복 능력이 높을수록 높은 scrambling 값을 가지며 평균스코어는 낮아지게 된다. 마지막으로 Longest Drive는 한 해 동안 해당 선수가 친 가장 긴 비거리를 의미하므로 Longest Drive 값이 클수록 공을 멀리 보내는 능력이 뛰어남을 나타낸다. 이는 곧 공을 멀리 보낼 수 있는 선수일수록 평균스코어가 낮은 경향이 있다는 것을 보여준다.

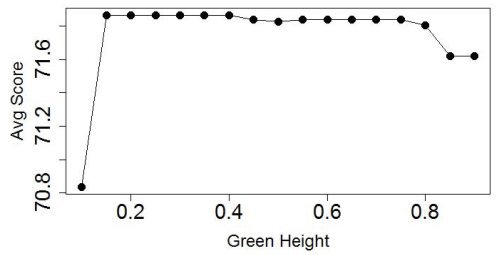
3.1.2. 비선형회귀분석(nonlinear regression) 결과 평균교차오차와 테스트 오차가 가장 낮은 랜덤 포레스트 모형에서 가장 중요도가 높은 세 변수를 선택하여 각 변수의 범위에 따라 평균스코어 예측치가 어떻게 변화하는지 그 양상을 살펴보았다. 선택된 세 가지 변수는 Green Height, Rough Height, Fairway Height로 모두 잔디의 길이와 관련된 변수였으나 평균스코어에 영향을 끼치는 모습은 각기 달랐으며 선형회귀분석을 통해서는 알 수 없었던 흥미로운 정보들을 새롭게 알 수 있었다. 아래의 Figure 3.2의 (a), (b), (c) plot은 특정 변수를 제외한 다른 변수들의 값은 평균으로 두고 특정 변수의 값을 증가시켜 랜덤 포레스트 예측 값이 어떻게 변하는지 관측한 것이다.

Table 3.2. The table of coefficients using stepwise, LASSO, ridge regression models

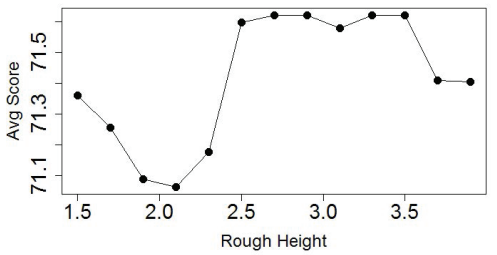
Variables	Stepwise regression	LASSO	Ridge regression
(Intercept)	79.6000	77.6100	85.2585
Top.10 ratio		−0.3230	−0.4414
Top.25 ratio			−1.1025
Cuts.Made ratio		0.6031	3.0284
Cuts.Missed ratio		0.0903	1.7696
Avg score			−0.2134
Player Age	−0.0190	0.0172	0.0212
FedExCup Points	−0.0052	−0.0018	−0.0058
Money	0.0000	0.0000	0.0000
Finish Position		0.0062	0.0259
Scoring Avg Tot Adjust	0.2669	0.2187	0.2014
Eagles			0.2962
Birdies			0.0204
Pars			0.0484
Tot Holes Over Par	0.1809	0.1389	0.3306
Longest Drive	−0.0196	−0.0160	−0.0201
Ball striking			0.0069
Drives Over 300 Yards num Drives			0.0079
Avg dist to hole after app			−0.0334
Avg fair proxi		0.0302	0.0842
Avg rough proxi		0.0152	0.0393
scram ratio			1.6360
Avg scram dist	−0.1266	−0.0682	−0.1345
Going for the Green success	0.3499	0.2922	0.4150
scram rough ratio	−3.6550	−3.1992	−4.0463
scram fringe ratio			0.4938
scram larg30 ratio	−1.5700	−1.4537	−2.0306
scram 20.30 ratio		−0.6602	−1.0100
scram 10.20 ratio	−2.0670	−2.0263	−2.7289
scram less10 ratio	−1.3110	−1.0109	−1.1673
sandsave ratio	2.1710	1.6642	1.8024
Tot Hole Outs	0.8826	0.8271	1.0934
Putt Avg GIR Putts			0.0402
Number of One Putt	−0.1670	−0.1101	−0.1785
Three Putt Avoidance	0.2845	0.2686	0.2080
putt 10 ratio		−1.9408	−4.1178
putt larg10 ratio	8.0000	6.1023	9.5117
Tot Putts Gained	0.1447	0.0779	0.1617
TTL SG T2G		−2.1686	−2.6250
TTL SG Tot	−0.2021	−0.1657	−0.2296
Fwy Firmness	0.3477	0.3103	0.3500
Grn Firmness	−0.1963	−0.1610	−0.2058
Grn Height	0.9529	0.9027	0.9421
Rough Height	−0.2269	−0.2124	−0.2257
Fwy Height	−12.5400	−12.2823	−12.5959
Max INT	0.2576	0.2224	0.2462
Max avg wind int	1.0430	0.9965	1.0613
Avg wind int		−0.0397	−0.0785

Table 3.3. The important variables of linear regression models

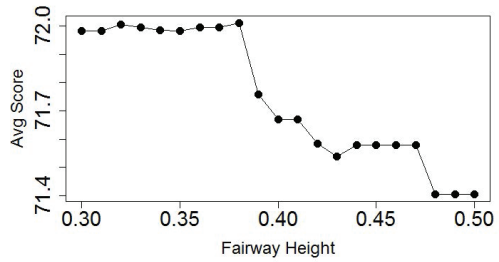
Selected variables (positive)	Selected variables (negative)
Fwy Firmness	TTL SG Tot
Grn Height	Number of One Putt
Max INT	Scrambling variables
Max avg wind int	Longest Drive
Three Putt Avoidance	



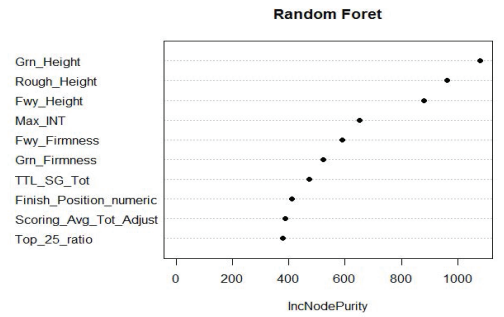
(a) Green Height



(b) Rough Height



(c) Fairway Height



(d) Importance plot

Figure 3.2. Average scores according to green height, rough height and fairway height.

랜덤 포레스트모형의 (d) importance plot을 보면 Green Height가 가장 중요한 변수로 선택된다. 제 3.1.1장 선형회귀분석에서 중복적으로 선택된 변수에도 Green Height가 있는 것으로 보아 선형회귀모형과 비선형회귀모형에서 모두 중요하게 선택되는 변수임을 알 수 있다. Table 3.3을 보면 선형회귀분석의 결과 Green Height는 양의 계수를 갖는 변수이며 랜덤 포레스트 모형에서도 마찬가지로 Green Height의 값이 커질수록 평균스코어가 증가하는 양의 증가 형태임을 볼 수 있다. 즉 두 가지 방법론에서 제공하는 정보가 일치함을 알 수 있다. 그러나 Figure 3.2(a) Green Height를 보면 평균스코어가 x값이 0.2일 때를 기준으로 변화 양상이 달라지는 것을 알 수 있다. 전반적으로 보면 양의 방향으로 상승하고는 있지만 그린의 잔디 길이가 0.2인치까지 증가할 때는 평균스코어가 71점에서 72점으로 급격하게 증가하는 반면 0.2인치 이상부터는 거의 증가하지 않고 일정하다. 이는 선형회귀 분석으로는 알 수 없는 새로운 정보이며 그린의 잔디가 0.2인치를 기준으로 평균스코어에 영향을 미치는 정도가 다르다는 매우 흥미로운 정보를 제공한다.

Table 3.4. The results of root mean square error (RMSE) in 4 different playoffs

	Bagging	Random forest
The Barclays	2.705628	2.616469
Deutsche Bank	1.930427	1.906402
BMW Championship	2.093180	2.195377
TOUR Championship	1.766457	1.716561

Figure 3.2(b)의 Rough Height의 경우 감소하다가 증가하고 또 다시 감소하는 형태로 특정 패턴을 지니고 있지 않아 설명하기에 난해하다. 그러나 Figure 3.2(c) Fairway Height의 경우 설명변수의 값이 0.38일 때를 기준으로 평균스코어가 증가하다가 감소한다. Green Height처럼 전반적으로 평균스코어가 양의 방향으로 증가하는 형상과 달리 Fairway Height는 페어웨이의 잔디길이가 0.38인치 미만일 때는 평균스코어가 작은 폭으로 증가하다가 0.38인치 이상부터는 역으로 큰 폭으로 급감한다. 그 결과 페어웨이의 잔디의 길이가 짧을 때보다 길 때 오히려 평균스코어가 더 낮으며 그 폭이 약 0.5점 정도로 차이가 나는 것을 알 수 있다.

3.2. 페덱스컵 투어 - 상위 10명, 상위 25명 예측

매 한해 정규 시즌이 끝나면 페덱스가 마지막으로 PGA 투어 4개 플레이오프 대회인 '더 바클레이스·도이치방크 챔피언십·BMW 챔피언십·투어챔피언십'을 개최하는데 이를 페덱스컵 투어라 한다. 페덱스컵 포인트 기준 상위 125명만이 플레이오프에 출전할 수 있으며, 1차전인 바클레이스 대회를 시작으로 하여 4개 대회를 치르면서 대회 때마다 성적에 따라 선수가 탈락되는 서바이벌 방식이다. 최종전인 투어 챔피언십에는 상위에 랭크된 30명만이 출전할 수 있다.

본 연구에서는 페덱스컵 포인트 기준 상위 150위 선수들을 데이터로 사용하였기 때문에 페덱스컵 4개 경기에 대한 선수 기록을 이번 장에서 추가적으로 예측해 보고자 한다. 또한 만약 평균스코어를 정확하게 예측했다면 순위도 잘 추정할 수 있을 것이라 생각하여 예측 스코어를 낮은 값부터 정렬해 상위권 순위도 얼마나 맞추는지 확인해 보았다.

평균스코어 예측 모형에서 랜덤 포레스트와 배깅 모형의 테스트 오차 값이 가장 작았기 때문에 이 두 가지 모형을 가지고 제곱근평균제곱오차를 계산해보았다. Table 3.4를 보면 모든 경기에서 제곱근평균제곱오차 값이 약 2점 정도로 매우 낮아 추정한 모형의 예측이 상당히 정확함을 알 수 있다. 여기서 구한 예측 스코어 값을 사용하여 아래에서 상위 10명과 상위 25명에 대해 순위를 매겨 보았다. PGA 선수들의 경우 실력차이가 크지 않고 미약한 점수 차이로 랭킹이 나뉘기 때문에 선수 개개인의 순위를 모두 예측하는 것은 너무 어렵다고 판단하였다. 따라서 각 경기의 상위 10명과 상위 25명 선수들을 예측하는데 중점을 두었다. 3차전까지는 상위 10명과 상위 25명을 모두 예측하였고 마지막 4차전에서는 참여 선수가 30명밖에 안되므로 상위 10명만 예측하였다.

Table 3.5는 두 모형에서 예측된 스코어에 따른 순위와 실제 순위를 비교해 상위 10명과 상위 25명 중 몇 명을 맞추었는지 기입한 표이다. 상위 10명을 예측하는 경우 대체로 50%정도를 맞추었으며, 상위 25명을 예측하는 경우 최대 60%까지 맞추기도 했다. 선수들 개개인의 평균스코어를 상당히 작은 오차범위 내로 추정하였기 때문에 동일한 모형을 사용하여 선수의 랭킹을 추정하였을 때도 역시 좋은 결과를 보임을 알 수 있다.

4. 결론

본 연구에서는 PGA 투어에서 제공하는 선수정보 및 코스정보를 사용하여 예측모형을 제시하였다. 다

Table 3.5. The results of root mean square error (RMSE) in 4 different playoffs

		Top 10	Top 25
The Barclays	Bagging	4	9
	Random forest	5	8
BMW Championship	Bagging	3	13
	Random forest	5	15
Deutsche Bank Championship	Bagging	4	14
	Random forest	3	14
TOUR Championship by Coca-Cola	Bagging	7	
	Random forest	7	

양한 데이터 마이닝 모형을 사용하여 PGA 투어에 참여하는 선수들의 평균스코어를 예측하였고, 어떠한 변수들이 스코어에 영향을 미치는지 살펴보았다. 추가적으로 페덱스 플레이오프 4대 경기 데이터를 통해 예측된 스코어에 따른 선수들의 순위 또한 정확하게 추정이 가능한지 확인해 보았다.

평균스코어를 예측하기 위해 의사결정나무, 부스팅, 배깅, 랜덤 포레스트, 라소, 능형회귀, 주성분회귀, 신경망 모형, 모든 가능한 회귀모형, 최근접이웃방법, 단계적 선택법 방법을 이용하였으며 모형 평가 지표로 제곱근평균제곱오차를 사용하였다. 각 모형들에 대한 예측률을 비교해보면 배깅과 랜덤 포레스트에서 가장 좋은 예측률을 보였다.

선형회귀분석을 이용한 모형들 중에서 중복으로 선택된 변수들을 살펴보면 페어웨이의 단단함과 그린의 잔디 길이, 평균최대풍속 등이 평균스코어증가에 영향을 미치는 것을 알 수 있었다. 즉, 페어웨이가 딱딱하면 공이 원하는 대로 굴러가지 않을 수 있고 그린에 있는 잔디의 길이가 길면 스윙을 할 때 방해요소가 되기도 한다. 뿐만 아니라 풍속이 높아지면 볼이 제대로 날아가지 않아 평균스코어를 높이는 요인이 된다. 반면 다른 선수들에 비해 더 성적이 좋음을 나타내는 변수인 Stroke gained 값이 크면 선수들의 능력이 좋은 것이므로 평균스코어가 낮아지게 되며, 평균적으로 퍼팅을 한 번에 성공시키는 경우가 많고, GIR 실패 후 버디나 이글로 점수를 스코어를 낼 때에도 평균스코어는 낮아진다. 한 해 동안 선수가 친 가장 긴 비거리를 나타내는 Longest drive도 스코어 감소에 영향을 미친다는 결과를 얻을 수 있었다.

추가적으로 분석한 4대 플레이오프 경기의 경우 예측률이 가장 좋았던 배깅과 랜덤 포레스트 모형을 사용하여 선수들의 평균스코어를 예측하였다. 이 예측 스코어를 기반으로 하여 상위권 선수의 순위를 예측했을 때에도 50%이상을 맞추는 좋은 결과를 보였다.

위와 같은 흥미로운 결과들을 도출해 내기 위해 많은 시간과 노력을 투자해야 했는데 골프스코어를 예측하는 것이 본 연구의 목표이기 때문에 매 경기가 열릴 때마다 그 경기 전까지 선수들의 모든 기록을 업데이트 시켜야 했고 이를 9,757개의 데이터에 모두 적용하였다. 매 경기를 하나씩 업데이트 하는 것은 굉장히 힘든 작업이었으므로 PGA 투어 선수들 중 상위 150명(페덱스컵 포인트 기준)만을 대상으로 분석을 했다는 한계점이 있다. 이를 보완하여 모든 선수들의 기록을 매 경기마다 업데이트시켜 분석에 사용한다면 기존 예측모형보다 더 좋은 결과를 얻을 수 있을 것이라고 생각된다.

References

Breiman, L. (1996). Bagging predictors, *Machine Learning*, **24**, 123–140.

Breiman, L. (2001). Random forests, *Machine Learning*, **45**, 5–32.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984)., *Classification and Regression Trees*, Chapman

and Hall, New York.

Connolly, R. A. and Rendleman Jr., R. J. (2008). Skill, luck and streaky play on the PGA tour, *Journal of The American Statistical Association*, **103**, 74–88.

Connolly, R. A. and Rendleman Jr., R. J. (2012). What it takes to win on the PGA tour (If your name is "Tiger" or if it isn't), *Interfaces*, **42**, 554–576.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification, *IEEE Transactions on Information Theory*, **13**, 21–27.

Frank, I. and Friedman, J. (1993). A statistical view of some chemometrics regression tools (with discussion), *Technometrics*, **35**, 109–148.

Freund, Y. and Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences*, **55**, 119–139.

Friedman, J. (2002). Stochastic gradient boosting, *Computational Statistics & Data Analysis*, **38**, 367–378.

Günther, F. and Fritsch, S. (2010). Neuralnet: training of neural networks, *The R Journal*, **2**, 30–38.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*, Springer, New York.

Hickman, D. C. and Metz, N. E. (2015). The impact of pressure on performance: evidence from the PGA tour, *Journal of Economic Behavior & Organization*, **116**, 319–330.

Hoerl, A. and Kennard, R. (1970). Ridge regression: biased estimation for nonorthogonal problems, *Technometrics*, **12**, 55–67.

Lee, H. W. and Lee, S. H. (2014). Analysis on the trend of domestic studies on golf : focusing on the Korean Journal of Golf Studies, *Korean Journal of Golf Studies*, **8**, 77–84.

Park, C., Kim, Y., Kim, J., Song, J., and Choi, H. (2011). *Datamining using R*, Kyowoo, Seoul.

R Development Core Team. (2015). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.

Ridgeway, G. (2012). Generalized Boosted Models: A guide to the gbm package.

Stone, M. and Brooks, R. (1990). Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression, *Journal of the Royal Statistical Society Series B (Methodological)*, **52**, 237–269.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society B (Methodological)*, **58**, 267–288.

PGA 투어의 골프 스코어 예측 및 분석

임정은[a] · 임영인[b] · 송종우[c,1]

[a]이화여자대학교 통계학과

(2016년 9월 21일 접수, 2016년 12월 8일 수정, 2016년 12월 27일 채택)

요 약

최근 골프는 많은 사람들의 취미 생활로서 자리를 잡아가고 있으며 골프와 관련된 연구도 다양하게 이루어지고 있다. 본 연구에서는 데이터 마이닝 기법을 사용하여 PGA 투어에 참여하는 선수들의 평균스코어를 예측하고 스코어에 유의한 영향을 미치는 변수들을 제시하고자 한다. 그리고 추가적으로 4개의 PGA 투어 플레이오프에 대해 상위 10명, 상위 25명의 선수들을 예측하는 것을 목표로 한다. 우리는 다양한 선형/비선형 회귀분석 방법을 이용하여 평균스코어를 예측하는데, 선형회귀분석 방법으로는 단계적 선택법, 모든 가능한 회귀모형, 라소(LASSO), 능형회귀, 주성분회귀분석을 사용하였으며 비선형회귀분석 방법으로는 트리(CART), 배깅, 그래디언트 부스팅, 신경망 모형, 랜덤 포레스트, 최근접이웃방법(KNN)을 사용하였다. 대부분의 모형에서 공통적으로 선택된 변수들을 살펴보면 페어웨이의 단단함와 그린의 풀의 높이, 평균최대풍속이 높을수록 선수들의 평균스코어는 높아지며 반대로 한 번에 퍼팅을 성공시키는 횟수와 그린적중률 실패 후 버디나 이글로 점수를 만드는 scrambling 변수들, 그리고 공을 멀리 보낼 수 있는 능력을 나타내는 longest drive는 그 값이 높아짐에 따라 선수들의 평균스코어가 낮아지는 경향이 있음을 알 수 있었다. 11가지 모형 모두 테스트 데이터인 2015년 경기 결과를 예측하는데 낮은 오류율을 보였으나 배깅과 랜덤 포레스트의 예측률이 가장 좋았으며 두 모형 모두 상위 10명과 상위 25명의 랭킹을 예측할 때 상당히 높은 적중률을 보였다.

주요용어: PGA 투어, 골프, 평균스코어, 선형회귀모형, 의사결정나무, 배깅, 그레디언트 부스팅, 인공신경망, 랜덤 포레스트, 최근접이웃방법

이 논문은 2015년도 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2015S1A5B6036244).

[1]교신저자: (03760) 서울특별시 서대문구 이화여대길 52, 이화여자대학교 통계학과.
E-mail: josong@ewha.ac.kr