The Korean Journal of Applied Statistics (2017) DOI: https://doi.org/10.5351/KJAS.2017.30.1.025
30(1), 25–39

Analysis of cycle racing ranking using statistical prediction models

Gahee Park[a] · Rira Park[a] · Jongwoo Song[a,1]

[a]Department of Statistics, Ewha Womans University

(Received September 7, 2016; Revised October 21, 2016; Accepted October 26, 2016)

Abstract

Over 5 million people participate in cycle racing betting and its revenue is more than 2 trillion won. This study predicts the ranking of cycle racing using various statistical analyses and identifies important variables which have influence on ranking. We propose competitive ranking prediction models using various classification and regression methods. Our model can predict rankings with low misclassification rates most of the time. We found that the ranking increases as the grade of a racer decreases and as overall scores increase. Inversely, we can observe that the ranking decreases when the grade of a racer increases, race number four is given, and the ranking of the last race of a racer decreases. We also found that prediction accuracy can be improved when we use centered data per race instead of raw data. However, the real profit from the future data was not high when we applied our prediction model because our model can predict only low-return events well.

Keywords: cycle racing, linear regression, stepwise regression, logistic regression, random forest, generalized additive model, gradient boosting, ridge regression, lasso regression, principal components regression, important variables

1. 서론

경륜은 우리나라의 경우 7명의 선수가 직선주로가 아닌 333.33m의 경사진 타원형의 경주로(벨로드롬 사이클 트랙, Figure 1.1)를 총 5바퀴(1,691m) 돌면서 기록이 아닌 순위를 겨루는 자전거 경주이다.
경기는 선두고정경주 방식으로 진행된다. 경기 진행시 선두선수의 경우 바람에 의해 불리할 수 있기 때문에 결승선 약 700m전까지 선수들의 앞에서 선수들을 이끌어주는 선두유도원이 경기에 참여하는 방식이다. 결승선 약 700m를 남기고 선두유도원이 퇴피하면 선수들이 스퍼트를 올려 각자 자신들의 주법을 이용해 경주한다. 또한 경주권을 구입하여 여러 가지 승자 적중 방식을 통해 승자를 적중시킨 경우 배당률에 따른 환급금을 받을 수 있는 참여형 레저스포츠이다. 2015년 기준, 경륜의 총매출은 2조 2731억 원에 달하고 입장객은 본장과 장외를 합쳐 554만 명으로 집계되었다 (사행산업통합감독위원회,

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

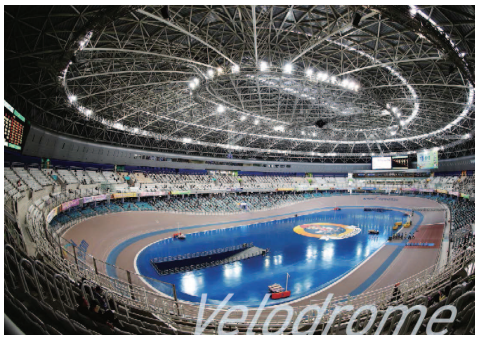

Figure 1.1. Velodrome.

http://static.ngcc.go.kr/). 경륜 경주에 베팅한 매출액의 72%는 고객환급금으로 지급되고, 매출액의 16%는 레저세, 교육세, 농어촌특별세 등으로 국가 및 지방재정 확충에 사용된다. 나머지 12%에서 경주 개최 비용을 제외한 수익금 전액은 국민체육진흥기금, 문화예술진흥기금, 청소년육성기금 등에 사용되고 있다 (경륜경정사업본부, www.kcycle.or.kr). 국민체육진흥공단(http://www.kspo.or.kr)은 경륜, 경정, 체육진흥투표권사업을 통해 레저세 등 제세금과 각종 공익기금 등 1996–2014년까지 총 8조 1천억원 이상의 공익재원을 조성하여 사회에 환원하고 국가재정 안정성에 기여하였다. 이 중 경륜의 공익재원 조성실적이 70.26%를 차지하고 있다.

경륜경기의 선수들은 본인들의 기량에 따라 특선급(SS, S1, S2, S3)과 우수급(A1, A2, A3) 및 선발급(B1, B2, B3)으로 등급을 부여받고, 경기의 편성은 같은 등급에 속한 선수들로 구성되며, 정기등급심사와 특별승강급제도로 6개월마다 한번씩, 또는 불규칙적으로 선수들의 등급변동이 일어난다. 이에 따라 세부등급이 조절되어서 등급 내에서 이동이 일어나는 경우도 있지만 등급자체가 바뀌는 경우도 생긴다.

경륜에 관한 선행연구로는 경륜선수의 신체조건에 관한 연구들 (Cho 등, 2008)과 경륜구매자에 초점을 맞춘 연구들 (Kim과 Kim, 2007) 그리고 경륜과 유사한 경마 경기 우승마 예측 연구 (Choe 등, 2015)가 있어왔지만 경륜순위를 통계적으로 예측하는 모형에 관한 연구는 이루어지지 않았다. 본 연구에서는 데이터 마이닝 기법들을 활용해서 경륜 순위 예측모형을 제시하고자 한다.

현재 국내 경륜경기는 광명, 창원, 부산에서 각각 매주 금, 토, 일 3일 동안 한 회차가 개최되고 있으며, 이 중 2015년 기준 전체 매출액의 55.28%를 차지하고 있는 광명 돔 경륜장에서 개최되는 경기를 분석대상으로 설정하였다. 분석에 사용한 데이터는 2015년 1월부터 2016년 5월까지의 광명에서 개최된 경기자료이고 2015년 1월부터 2016년 4월까지의 경기자료를 train data로 사용해서 예측모형을 만들고, 분석에 사용하지 않은 2016년 5월 경기자료를 test data로 사용해서 결과를 확인한다. 분석은 R을 이용해서 진행하고 (Park 등, 2011) 데이터에 관한 자세한 설명은 2장에서 하도록 한다.

경륜경기의 전체 순위를 정확하게 맞추는 것은 아주 어려운 일이고, 그것보다는 경주권에서 사용되는 승식의 5가지 종류에 해당하는 상위권 3등까지를 맞추는 것이 더욱 의미있다고 판단하여 각 경기의 1등부터 3등까지를 예측하는 모형을 만드는 것을 목적으로 한다. 2장에서는 이를 위해 데이터를 수집하고, 원자료를 분석에 용이하게 변수로 변환하는 과정을 설명한다. 3장에서는 classification 방법과 regression 방법을 이용해서 예측모형들을 만든 후 결과를 비교한다. 또한 순위를 예측하는데 유의한 설명변수들을 살펴보고 반응변수와의 관계를 살펴보도록 한다. 마지막으로 분석에 사용하지 않은 최근 한 달 경기 자료를 이용해서 실제 경륜경기 베팅을 실시하고 결과를 살펴본다. 4장에서는 본연구의 내용을 정리하고 결론을 내리고자 한다.

2. 분석자료 설명

2.1. 용어정의

경륜 베팅 방식으로는 단승식, 연승식, 복승식, 쌍승식, 삼복승식이 있으며, 각 베팅방식은 다음과 같다.

- 단승식: 1위 선수 1명을 적중시키는 방식
- 연승식: 1, 2위 선수 1명을 적중시키는 방식 (단, 8인 이상의 경우는 3위 이내, 출주 선수 4인 이하는 미발매)
- 복승식: 1, 2위 선수 2명을 순위에 관계없이 적중시키는 방식
- 쌍승식: 1, 2위 선수 2명의 순위를 정확하게 적중시키는 방식
- 삼복승식: 1, 2, 3위 선수 3명을 순위에 관계없이 적중시키는 방식

다음으로 '전법'이란 경륜경주를 함에 있어서 승리를 하기 위해 취하는 주행 스타일로서 선행, 젖히기, 추입, 마크로 총 4가지가 있고 자세한 설명은 다음과 같다.

- 선행: 선두원이 퇴피한 시점부터 마지막 1코너를 접어들기 전에 선두에 나서 주행하는 주법
- 젖히기: 종 주회 1코너를 지난 이후부터 3코너 지점을 지나기 전까지 대열의 중간이나 후미권에 위치해 있다가 단번에 선두권을 넘어서는 주법
- 추입: 선행이나 젖히기 선수의 뒤에서 풍압을 피해 달리다가 마지막 직선 코스에 접어들어 역전을 이루어내는 주법
- 마크: 기본적으로 기량이 부족하거나 상대 선수의 능력이 월등히 강하다고 인정을 할 때 후미를 따르며 2, 3착 승부를 노리는 주법

2.2. 자료수집

본 연구에서 사용된 자료는 2015년 1월 2일부터 2016년 5월 29일까지 광명경륜경기장에서 실시된 경륜경기자료로, 국민체육진흥공단 경륜경정사업본부(www.kcycle.or.kr)에서 제공하는 경주출주표와 경주결과자료를 수집하였다. 경륜경기는 매주 금, 토, 일에 개최되는 일반경륜과 비정기적으로 각 등급별 성적 상위자들이 참가하고 특정타이틀이 걸려있는 대상경륜, 한 해의 마지막 회에 시행되며 한 해 최고의 선수들이 출전하여 경륜 챔피언을 가리는 대회인 그랑프리경륜 등이 있다. 그랑프리경륜의 경우 특선급 선수들만 출전하기 때문에일반경륜과 대상경륜만을 분석대상으로 설정해서 그랑프리경륜인 2015년 12월 18일부터 20일 경기(49회차 2015년도 문화체육관광부 장관배 그랑프리 경륜 대회)자료는 제거하였다. 본 연구에서는 예측모형의 설명변수로 선수 개인별 과거 경기기록들을 사용한다. 그런데 2015년 21기 선수들은 신인선수로서 과거기록이 존재하지 않기 때문에 이 선수들의 첫 경기는 분석대상에서 제외한다. 해당경기는 2015년 26회 경기로 신인선수 10명이 1일차에서 3일차까지 각각 세 번 출전하였고, 신인선수가 포함된 경기는 총 22경기이다. 따라서 이를 제거하고 분석을 진행하였다.

경기결과에는 실격이나 낙차로 인해 순위가 매겨지지 않은 경우가 나타난다. 경주출주표에는 출전선수들의 최근 3회 성적순위가 제공된다. 이를 이용해 Figure2.1에 경기 결과(순위) 별 선수 최근 3회 성적순위/총 선수 수 분포를 경기등급(선발, 우수, 특선)별로 나타내었다. 분포를 살펴보면 1등부터 7등까지는 등수가 낮아질수록 선수의 성적순위도 낮아지는 것을 확인할 수 있다. 하지만 낙차하거나 순위가 매겨지지 않은 경우 선수들의 성적순위 분포가 일정한 패턴을 나타낸다고 보기 힘들다. 이런 값들은 순

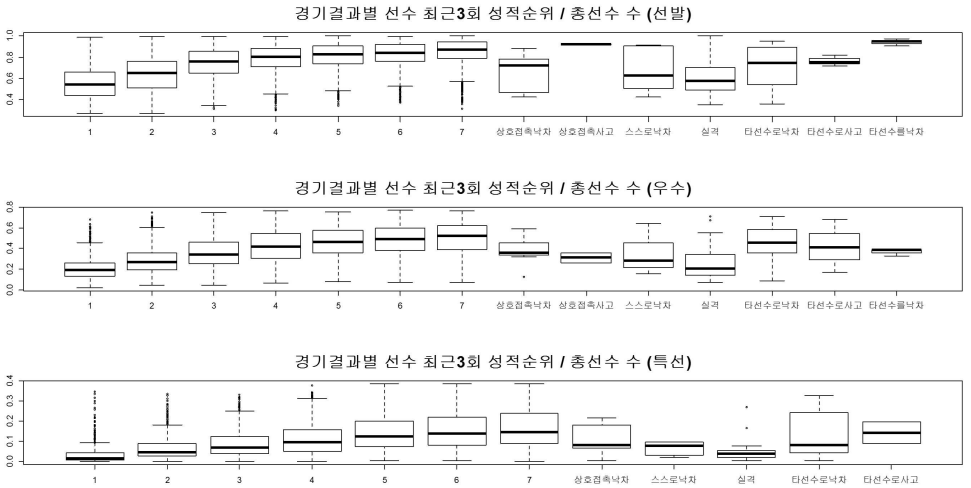

Figure 2.1. Ranking for latest 3 games divided by total number of players.

위를 예측하는데 혼란을 줄 수 있기 때문에 2015년 1월부터 2016년 4월까지의 train data에서 실격이나 낙차인 경우 해당선수들이 포함된 167경기 역시 제거하고 분석을 진행하였다.

따라서 분석에 사용할 자료는 총 19,859개의 자료(총 2,837경기)이고, 그 중 2015년 1월부터 2016년 4월까지의 18,683개의 자료(총 2,669경기)를 train data로, 2016년 5월 동안의 1,176개의 자료(총 168경기)를 test data로 놓고 분석을 진행하였다.

2.3. 변수 설명

자료수집단계에서 수집한 경주출주표와 경주결과자료를 통해 얻을 수 있는 선수의 기수/나이 정보, 경기여건 관련 정보, 선수의 실력관련 정량정보, 선수등급정보, 최근 경기 동향 관련 정보들을 설명변수로 사용한다. 이 중 변환이 필요한 경우 분석에 알맞은 형태로 변환시키고, 파생변수를 생성한다. 변수의 개념정의는 경륜경정사업본부 사이트(www.kcycle.or.kr)를 참고하였다. 추가 설명이 필요한 변수의 경우 아래에 설명을 제시하고, 나머지 변수들은 Table 2.2에서 설명하도록 한다.

2.3.1. 경기여건 관련 정보

- **번호**

경륜경기 출전시 부여되는 번호로 1에서 7까지 부여되며 번호에 따라 다른 색상의 유니폼을 입어서 구분한다. 그 중 4번을 부여받은 선수는 선두유도원 퇴피 전까지 선행해야할 의무가 있다. 선행을 하게 되면 풍압을 피하지 못해 불리할 수 있기 때문에 선수들이 기피하는 경향이 있다. 전체 경기자료에서 각 번호별로 경기결과 1, 2, 3등 비율을 확인해본 결과 Figure 2.2와 같고 4번을 부여 받은 경우 1, 2, 3등 비율이 16.49%로 다른 번호들의 비율이 45%에서 49%정도 인 것에 비해 확연히 낮음을 알 수 있다. 이를 반영하기 위해 부여받은 번호가 4번인지 아닌지를 나타내는 더미변수를 사용한다.

2.3.2. 선수의 실력관련 정량 정보 선수의 이전 경기 실력에 관련된 정량적인 정보들이 경주출주표에 나타나 있는데, 등급 간에 변동이 있는 선수의 경우 현재 등급의 경기결과를 예측하는 데 이전 등급

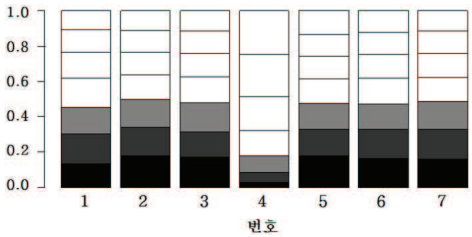

Figure 2.2. Ratio of the first, second, third rank by given number.

Table 2.1. Characteristic score according to each grade

구분	특선급				우수급			선발급		
고유 점수	SS	S1	S2	S3	A1	A2	A3	B1	B2	B3
	102	100	98	96	94	92	90	88	86	84

에서의 경기를 통해 얻어진 정보를 사용하면 선수의 실력을 과소평가 혹은 과대평가할 경우가 생긴다. 경주출주표에 선수의 최근 1회전, 최근 2회전, 최근 3회전 성적이 나타나는데 예를 들어 선발(B)등급에서 승급되어 우수(A)등급으로 등급이동한 선수의 경우 예측하고자 하는 경기는 우수등급선수들과 하는 경기지만 출주표에 나타난 이전 성적들은 선발등급선수들과 한 경기의 성적이 된다. 이런 경우 이전 성적을 그대로 사용하는 것이 해당 선수의 경기결과를 과대평가할 수 있다. 이를 방지하기 위해 현재등급과 같은 등급(선발급, 우수급, 특선급)의 경기에서의 결과만을 선수 실력관련 정량정보로서 사용하기로 하자. 이 경우 문제가 되는 것이 등급변동이 일어난 후 첫 번째 경기인 선수의 경우 최근 3회차가 모두 이전 등급에서의 경기가 된다는 것이다. 이를 해결하기 위한 대체값을 구하기 위해 먼저 train data보다 더 이전 시점인 2014년 6월부터 12월까지의 경기자료를 앞서와 같은 방식으로 수집하였다. 다음으로 수집한 새로운 자료에서 등급이 변경되고 첫 경기인 선수들의 경기결과가 포함된 관측치들을 S에서 A로 강급된 경우, A에서 B로 강급된 경우, B에서 A로 승급된 경우, A에서 S로 승급된 경우 총 4가지로 나누었다. 나눈 각 자료들과 그 경우에 해당하는 train data에서의 등급변동이 일어난 후 첫 번째 경기인 선수의 정보를 $k = 10$인 KNN을 이용해서 경기결과 중위값을 구해서 해당 값을 대체값으로 사용하도록 한다. 등급변동으로 인해 생기는 또 다른 문제는 새로운 등급으로 이동한 선수의 승률, 연대율, 삼연대율 역시 등급변동 후 첫 경기인 경우 예전 등급경기에서만의 결과이기 때문에 이 값을 그대로 사용하기엔 부적절하다는 것이다. 따라서 앞서와 유사하게 2014년 6월부터 12월까지의 경기자료에서 등급이 변경되고 첫 경기인 선수들의 경기결과를 바탕으로 평균적인 승률, 연대율, 삼연대율을 계산하여 대체값으로 사용한다.

- **종합득점**

광명·창원·부산경륜에 출전하여 획득한 최근 3회차 동안의 경주득점의 평균을 나타낸다. 득점을 계산하는 방식은 다음과 같다. 1, 2일차 경기는 매 경기 출전선수에 대하여 Table 2.1에 나타나는 해당 급별 고유점수의 평균점을 중간순위인 4위 경주득점으로 하고 3위 이상은 평균점에 매 순위 2점씩 더하여 계산하며, 5위 이하는 매 순위 2점씩 감하여 순위점수를 계산한다. 3일차 경기는 1, 2일차에 입상기록한 선수별 순위득점의 평균점에 의거하여 1, 2일차와 동일한 방식으로 순위점수를 계산한다. 즉 경기에서 같은 순위를 받았더라고, 경기 구성원들의 세부등급에 따라 다른 경주득점을 얻게 된다.

- **최근성적**

경주출주표에서 최근 3회차 성적, 최근 2회차 성적, 최근 1회차 성적, 금회성적 정보가 주어진다. 이러한 최근 성적 관련된 변수들은 관측치들에 따라 채워진 값의 개수가 다르다는 문제점이 있다. 즉 선

수들마다 결측치로 비워져 있는 값들이 생긴다. 최근 3회차 간의 성적의 경우에는 낙차하거나 실격, 결장한 경우, 후보인 경우, 등급변동이 있어서 이전등급에서의 경기결과여서 사용하지 못하는 경우, 해당 값은 결측치가 된다. 또한 한 회차에도 1일차(금), 2일차(토), 3일차(일) 경기가 있고 선수들은 보통 세 경기를 출전하게 되는데 2일차의 경우 1일차에서 해당 선수의 경기결과를 알 수 있고, 3일차에는 1, 2일차 경기의 성적을 알 수 있다. 즉 3일차경기에서는 1, 2회차 경기결과 정보가 채워져 있지만 1일차 경기에는 1, 2회차 경기결과가 채워져 있지 않다. 다음으로 고려해야할 사항은 주어진 정보들이 여러 회차 동안의 선수 경기성적인데 각각의 이전 경기성적이 발생한 시점이 다 다르다는 것이다. 즉 최근 성적일수록 예측하려고 하는 경기결과에 더 큰 설명력을 가지는 값이라고 볼 수 있다. 최근성적 관련된 값들의 개수가 관측치마다 다른 것과 기간이 다른 것, 두 가지를 모두 고려하면서 정보손실을 최소로 하는 방법으로 최근성적들을 각각의 해당경기날짜들과 현재시점의 경기날짜 사이의 기간에 따른 Exponential decay를 가중치로 사용해서 가중평균을 구하는 방법을 사용하기로 한다. Exponential decay는 $e^{-x_i}/\Sigma e^{-x_i}$로 나타낼 수 있고, 이 때 x_i는 현재경기와 이전 성적이 관측된 경기와의 일수차이다. 최종적으로 최근 3회차 성적, 최근 2회차 성적, 최근 1회차 성적, 금회성적의 가중평균으로 구한 값을 새로운 최근성적변수로 사용하기로 한다.

- **상대전적**

출주표에서 얻을 수 있는 상대전적정보는 직전 연도부터 현재까지 상대되는 두 선수가 동반 출전한 경주의 순위상 승패기록의 누계로서 행렬형태로 주어져 있다. 이를 변수로 사용하기 위해 자신을 제외한 여섯 명의 선수들과의 경기에서 상대선수를 이긴 횟수를 경기 횟수로 나누어 계산한 확률을 모두 더해서 변수로 사용하도록 하자. 상대선수와 경기를 한 적이 없는 경우는 0으로 대체하였다.

2.3.3. 선수의 등급관련 정보

- **현재등급**

선수의 등급변동 후 세부등급까지 포함한 등급을 말한다. 특선급(SS, S1, S2, S3), 우수급(A1, A2, A3), 선발급(B1, B2, B3)으로 총 10개의 등급으로 구분된다.

- **등급변동**

정기 등급 심사나 특별 승강급에 의해 등급변동이 일어나는 자료들이 존재한다. 이 때 선수가 등급변동으로 원래 속해있는 등급경기보다 더 잘하는 선수들이 많은 등급으로 승급되었는지, 원래보다 강급되었는지, 혹은 변동되었는지에 따라 선수들의 경기결과가 영향을 받을 수 있을 것으로 생각되어 승급되었으면 1, 변동 없으면 0, 강급되었으면 -1로 부여하는 범주형 변수를 만들어서 분석에 사용하도록 한다.

분석에 사용할 최종변수들은 Table 2.2와 같다. 선수의 실력관련 정량 정보 변수들 중 상대전적 변수를 제외한 200m기록, 승률, 연대율, 삼연대율, 입상/출전회수, 선행, 젖히기, 추입, 마크, 종합득점, 성적순위, 최근성적의 경우 각 경기들마다 등급이 다르기 때문에 이 값을 그대로 사용하면 각 경기 내에서 순위를 예측하는데 혼선을 줄 수 있다. 예를 들어 종합득점의 경우, 한 경기 내에서는 일반적으로 종합득점이 높은 선수가 높은 순위를 받을 거라고 생각할 수 있다. 하지만 등급별로 경기가 진행되기 때문에 모든 경기들을 고려할 때는 낮은 등급의 경기에서 높은 순위를 받은 선수의 경주득점보다 높은 등급의 경기에서 낮은 순위를 받은 선수의 경주득점이 더 클 수 있다. 이런 경우 순위예측에 혼선을 줄 수 있기 때문에 선수의 실력 관련된 연속형 변수들을 각 경기별로 평균값을 빼서 보정된 변수들을 생성하였다. 편의상 보정된 변수들은 원변수의 이름 끝에 숫자2를 붙여서 사용하도록 한다. 다음 장의 분석에서 보정된 변수들을 사용한 경우(보정된 자료)와 원변수 그대로 사용한 경우(원자료)로 나누어서 예측모형을 제시하고 모형간의 예측률과 선택된 중요변수의 차이를 비교해보도록 하자.

Table 2.2. Description of variables

Variable	Description
Input variables	
기수	선수의 훈련원 입소 기수
나이	선수의 연령
번호	출전시 부여받은 번호가 4번인지 여부
기어배수	페달 안쪽에 있는 큰 기어의 톱니바퀴수를 뒷바퀴에 있는 작은 기어의 톱니바퀴수로 나눈 수치
기어변동	기어배수를 이전경기에 비해 변화했는지 여부 (기어배수를 내렸으면 −1, 변동이 없으면 0, 올렸으면 1)
200m 기록	최근 출전한 회차의 3일(금, 토, 일요일) 경주 기록 중 가장 빠른 기록
승률	당해 연도(당해 연도 첫 출전인 경우 직전 연도)에 1위를 한 회수를 출주한 총 회수로 나눈 수치
연대율	당해 연도(당해 연도 첫 출전인 경우 직전 연도)에 1위, 2위 회수를 더해 출주한 총 회수로 나눈 수치
삼연대율	당해 연도(당해 연도 첫 출전인 경우 직전 연도)에 1위, 2위, 3위 회수를 더해 출주한 총 회수로 나눈 수치
입상/출전회수	직전 반기 첫 회차부터 현재까지 입상(1, 2, 3위)/출전회수 = 입상확률
선행	직전 반기 첫 회차부터 현재까지 '선행'으로 입상한 횟수
입상 젖히기	직전 반기 첫 회차부터 현재까지 '젖히기'로 입상한 횟수
전법 추입	직전 반기 첫 회차부터 현재까지 '추입'으로 입상한 횟수
마크	직전 반기 첫 회차부터 현재까지 '마크'로 입상한 횟수
종합득점	광명·창원·부산경륜에 출전하여 획득한 최근 3회(전전전회차, 전전회차, 전회차) 경주득점의 평균
최근 3회 성적순위	최근 3회차동안의 종합득점으로 구한 순위/전체선수 수
최근성적	선수의 최근 3회차 경기와 1, 2일차 경기가 존재하는 경우 기간에 따른 성적의 가중평균
상대전적	같은 경기에 참여하는 선수들의 상대전적 합산
현재등급	등급 조정 된 후 현재등급
등급변동	등급이 강급했으면 −1, 등급조정이 일어나지 않았으면 0, 승강했으면 1
최근 3회 결승진출회수	선수의 최근 3회차 경기 동안 결승 진출 회수
최근 1회 실격여부	선수의 최근 1회차 3일간 경기에서 실격을 했는지 여부
최근 1회 낙차여부	선수의 최근 1회차 3일간 경기에서 낙차를 했는지 여부
최근 1회 경기 일수차	선수의 최근 1회차 경기일자에서 이번 회차 경기일자의 일수차이
잔여위반점수	개인별 3회차의 위반점을 합산하여 40점 초과시 차 회차에 출전정지 처분을 받게 되며, 40점은 소멸. 잔여점수는 40점을 기준으로 해당선수의 잔여점수를 의미
Response variables	
순위	경주 별 선수가 들어온 순서

3. 분석방법 및 분석결과

본 연구에서는 경륜 선수의 순위 예측 모형의 평가 지표로 경륜 베팅 방식을 이용하였다. 따라서 순위 전체를 맞추기보다는 순서에 상관없이 1위, 1·2위, 1·2·3위를 맞추는 것이 중요하다. 편의상, 순위1위를 맞추는 모형을 순위1 모형, 1·2위를 순서에 상관없이 맞추는 모형을 순위2 모형, 1·2·3위를 순서에 상관없이 맞추는 모형을 순위3 모형이라 하자.

본 연구에서는 순위를 예측하기 위해서 1) 순위를 2-class로 변환한 classification 방법, 2) 순위를 수치로 사용한 Regression 방법을 이용한다. 또한 각 모형에 대해 경기별 순위 예측률을 비교해보고 예측 모형에 포함된 중요 변수에 대해 살펴본다.

사용하지 않고 맞추고 싶은 n등까지에 class 1을, 나머지에는 0을 할당하여 2-class로 변환시켜 사용하였다. 또한 모형의 성과를 비교하기 위해 사용될 오분류율은 경기별로 class를 정확히 맞추었으면 0, 그렇지 않으면 1을 할당하여 평균값을 계산한 '경기 오분류율'을 사용한다. 2015년 1월부터 2016년 4월까지의 총 2,669경기 중 임의로 70%의 경기를 train data로 30%의 경기를 validation data로 나누어, train data을 통해 적합시킨 모형으로 validation data에서의 오분류율을 계산한다. 이 과정을 100번 반복하여 계산한 평균 오분류율을 모형 비교 지표로 사용하였다. 이 장의 마지막에서 2016년 5월 동안의 168경기에 최적 모형을 적용하여 도출한 오분류율과 실제 배당률을 적용한 이윤 금액을 계산하여 비

Table 3.1. The important variables of each logistic model

		선택된 변수 (+)	선택된 변수 (−)
순위1	원자료	등급변동(−), 종합득점, 상대전적, 승률	현재등급, 번호, 등급변동(+), 마크, 나이
	보정된 자료	등급변동(−), 종합득점2, 상대전적, 승률2	등급변동(+), 200m기록2, 번호, 최근성적2, 마크2, 나이, 최근 1회 경기 일수차
순위2	원자료	등급변동(−), 상대전적, 종합득점, 연대율	현재등급, 성적순위, 번호, 등급변동(+), 최근성적, 나이, 마크, 최근 1회 경기 일수차
	보정된 자료	등급변동(−), 종합득점2, 상대전적, 연대율2, 승률2	등급변동(+), 번호, 최근성적2, 마크2, 나이, 최근 1회 경기 일수차
순위3	원자료	등급변동(−), 종합득점, 상대전적, 삼연대율	현재등급, 번호, 등급변동(+), 최근성적, 추입, 나이, 최근 1회 경기 일수차
	보정된 자료	등급변동(−), 종합득점2, 상대전적, 연대율2	번호, 등급변동(+), 최근성적2, 나이, 최근 1회 경기 일수차

교해보도록 하자.

3.1. Classification 방법

Classification 방법은 n등까지를 맞추기 위해서 n등까지만 1을 주고 나머지는 0을 주는 방식으로 7-class를 2-class로 변환하여 반응변수로 사용하였다. 모형을 만들 때는 경기를 구분하지 않고 전체자료를 사용하기 때문에 경기별로 1로 예측하는 개수가 n개보다 적거나 n개를 넘는 경우 생길 수 있다. 이를 해결하기 위해서 예측모형을 통해서 class가 1일 확률을 전체 자료에 대해서 구한 다음, 한 경기 내에서 확률이 큰 순서대로 n등까지의 class를 1로 나머지는 0으로 분류하는 방법을 사용한다. Classification 방법으로는 모든 변수를 이용한 로지스틱 회귀모형과 AIC, BIC 기준으로 단계별 변수선택법을 사용해 변수를 선택한 로지스틱 회귀모형, 일반화 가법 모형(GAM), 랜덤 포레스트 (Breiman, 2001), 그래디언트 부스팅 모형(GBM) (Ridgeway, 2006)을 사용하였다.

먼저 순위에 영향을 미치는 변수들을 살펴보기 위해 BIC 기준으로 변수 선택한 로지스틱 회귀모형의 설명변수들을 회귀 계수 부호에 따라 절댓값의 내림차순으로 Table 3.1에 정리해 보았다. 설명변수가 연속형인 경우 회귀계수가 양의 값을 가지면 설명변수의 값이 증가함에 따라 입상할 확률이 높아진다고 해석할 수 있다. Table 3.1에서 양의 효과를 갖는 변수들의 의미를 살펴보면, 1) 등급이 강급될수록 2) 종합득점이 높을수록 3) 상대전적이 높을수록 순위권에 진입할 확률이 높아진다고 할 수 있다. 음의 효과를 갖는 변수들의 의미를 살펴보면, 1) 등급이 승급될수록 2) 나이가 많을수록 3) 최근 1회전 경기를 한지 오래 될수록 4) 번호 4번을 부여받아 선행할수록 5) 최근성적이 나쁠수록 순위권에 진입할 확률이 낮아진다고 할 수 있다.

다음으로 랜덤 포레스트와 그래디언트 부스팅 모형에서 구한 변수 중요도 및 상대적 영향정도를 통해 중요 변수를 살펴보고자 한다. Table 3.2와 Table 3.3을 살펴보면 원자료를 사용한 모형들에서 공통적으로 선택된 중요 변수들은 현재등급과 최근성적이고 보정된 자료를 사용한 모형에서의 중요 변수는 종합득점과 성적순위이다. 경기가 등급(선발, 우수, 특선)별로 이루어지고 경기 내의 상대적인 값이 순위에 큰 영향을 미치기 때문에, 원자료를 사용한 모형의 경우 현재등급이라는 범주형 변수를 통해 등급 간의

Table 3.2. The important variables of each random forest model

		주요변수 10개
순위1	원자료	현재등급, 최근성적, 연대율, 승률, 상대전적, 삼연대율, 최근 3회 결승진출회수, 성적순위, 종합득점, 입상/출전회수
	보정된 자료	종합득점2, 성적순위2, 승률2, 연대율2, 현재등급, 최근성적2, 삼연대율2, 마크2, 입상/출전회수2, 상대전적
순위2	원자료	최근성적, 현재등급, 삼연대율, 연대율, 상대전적, 승률, 종합득점, 성적순위, 입상/출전회수, 최근 3회 결승진출회수
	보정된 자료	종합득점2, 성적순위2, 최근성적2, 연대율2, 삼연대율2, 현재등급, 승률2, 상대전적, 입상/출전회수2, 200m기록2
순위3	원자료	최근성적, 상대전적, 현재등급, 삼연대율, 연대율, 성적순위, 종합득점, 입상/출전회수, 최근 1회 경기 일수차, 승률
	보정된 자료	종합득점2, 성적순위2, 최근성적2, 삼연대율2, 연대율2, 현재등급, 승률2, 상대전적, 입상/출전회수2, 200m기록2

Table 3.3. The relative influence of each gradient boosting model

		상위변수 10개
순위1	원자료	현재등급, 연대율, 등급변동, 최근성적, 최근 3회 결승진출회수, 승률, 상대전적, 종합득점, 성적순위, 입상/출전회수
	보정된 자료	종합득점2, 승률2, 등급변동, 최근성적2, 현재등급, 성적순위2, 연대율2, 상대전적, 마크2, 입상/출전회수2
순위2	원자료	현재등급, 최근 3회 결승진출회수, 최근성적, 상대전적, 등급변동, 연대율, 삼연대율, 승률, 번호, 성적순위
	보정된 자료	종합득점2, 최근성적2, 현재등급, 등급변동, 성적순위2, 승률2, 연대율2, 마크2, 선행2, 상대전적
순위3	원자료	현재등급, 최근성적, 삼연대율, 최근 3회 결승진출회수, 상대전적, 번호, 등급변동, 성적순위, 200m기록, 종합득점
	보정된 자료	종합득점2, 최근성적2, 현재등급, 번호, 등급변동, 추입2, 삼연대율2, 최근 1회 경기 일수차, 나이, 상대전적

차이를 반영하고, 등급별로 나뉘어져 이루어진 최근 경기의 등수를 통해 현재 경기결과 순위를 예측할 수 있다. 이와 달리 보정된 자료를 사용한 모형에서는 해당 경기 출전선수들의 개인성적 기록의 평균으로 빼서 보정하였기 때문에 종합득점, 성적순위와 같이 등급이 높아질수록 값이 커지거나 작아지는 변수들이 보정되어서 경기 내 순위를 예측하는데 큰 도움을 준다. 즉, 경기 내 평균값을 빼주는 보정을 통해 경기 내 출전선수들의 상대적인 기량 차이에 집중하고 있다.

이제 각 모형을 비교하기 위해서 validation data에서의 평균 오분류율(표준편차)을 계산하여 Table 3.4에 나타내었다. 결과를 살펴보면 순위1 모형은 평균적으로 약 60% 정도의 예측률을 가지고, 순위2 모형은 40%, 순위3 모형은 30% 정도의 예측률을 보이는 것을 알 수 있다. Random Guess일 경우 각각 14, 4.7, 2.8%의 예측률을 가지는 것을 감안하면 모든 모형에서 예측률이 좋게 나타난다. 자세히 들여다보면 먼저 원자료를 사용한 모형과 보정된 자료를 사용한 모형의 오분류율을 비교했을 때 보정된 자료를 사용한 모형의 오분류율이 모든 경우에서 더 낮게 나타났다. 순위1, 2, 3 모형 안에서 각 방법별로 예측률 차이가 크진 않지만 순위1 모형에서는 BIC 기준으로 변수 선택한 로지스틱 회귀모형이, 순위2 모형은 원자료를 사용한 경우 AIC 기준으로 변수 선택한 로지스틱 회귀모형, 보정된 자료를 사용한 경우 일반화가법모형에서 오분류율이 가장 낮았다. 순위3 모형은 원자료를 사용한 경우 BIC 기준으로 변수선택한 로지스틱 회귀모형에서, 보정된 자료를 사용한 경우 일반화가법모형에서 오분류율이 가장 낮았다.

Table 3.4. Average misclassification rates in validation data

		Logistic	Logistic-AIC	Logistic-BIC	GAM	Random forest	GBM
순위1	원자료	0.3943 (0.0119)	0.3934 (0.0127)	**0.3891** (0.0123)	0.4050 (0.0127)	0.4276 (0.0158)	0.4133 (0.0143)
	보정된 자료	0.3818 (0.0111)	0.3833 (0.0130)	**0.3798** (0.0125)	0.3872 (0.0157)	0.3914 (0.0150)	0.3873 (0.0142)
순위2	원자료	0.5916 (0.0130)	**0.5892** (0.0122)	0.5897 (0.0122)	0.5929 (0.0149)	0.6347 (0.0141)	0.6017 (0.0141)
	보정된 자료	0.5833 (0.0119)	0.5837 (0.0125)	0.5887 (0.0127)	**0.5819** (0.0144)	0.5946 (0.0153)	0.5881 (0.0142)
순위3	원자료	0.6943 (0.0127)	0.6930 (0.0115)	**0.6902** (0.0109)	0.6925 (0.0132)	0.7237 (0.0140)	0.6928 (0.0151)
	보정된 자료	0.6886 (0.0132)	0.6906 (0.0114)	0.6893 (0.0108)	**0.6781** (0.0130)	0.6932 (0.0133)	0.6894 (0.0132)

하지만 순위1, 2, 3 모형 모두 로지스틱회귀와 일반화가법모형 간의 오분류율 차이가 크지 않기 때문에 이 경우 설명이 쉽고 간단한 BIC 기준으로 변수선택한 로지스틱 회귀모형을 최종모형으로 선택하여도 무방하다.

예측률 자체는 높더라도 실제 배당률이 높은 경기의 순위 예측과 배당률이 낮은 경기의 순위 예측은 알 수 없으므로 어느 모형이 더 높은 배당을 얻을지는 장담할 수 없다. 그러므로 이 장의 마지막에서는 로지스틱회귀와 일반화가법모형 모두를 사용하여 배당률을 적용해보고자 한다.

3.2. Regression 방법

Regression 방법에서는 경기결과 순위가 순서적인 의미가 있기 때문에 수치 그대로를 반응변수로 사용하였다. 앞서 Classification 방법에서 원자료와 보정된 자료를 사용한 모형간의 비교를 하였을 때 보정된 자료에서의 오분류율이 더 낮았기 때문에 Regression 방법에서는 보정된 자료만을 이용해서 분석을 진행한다. Regression 방법을 이용해서 예측치를 구하면 Classification 방법에서와 마찬가지로 순위 예측치가 1등부터 7등까지의 정수로 도출되지 않으므로 한 경기 내에서 순위 예측치가 가장 작은 순서대로 n등까지는 1을, 나머지는 0을 부여한다. 그런 다음 다시 분류문제로 돌아와서 앞 절에서와 마찬가지로 경기 오분류율을 구해서 모형간의 평가지표로 활용한다. Regression 방법으로는 All Subset, Ridge (Hoerl과 Kennard, 1970), Lasso (Tibshirani, 1996), 주성분 회귀 (Hastie 등, 2001), 랜덤 포레스트를 사용하였다.

Regression 방법도 마찬가지로 순위에 영향을 미치는 변수들을 살펴보기 위해 All Subset회귀모형의 설명변수들을 회귀계수 부호에 따라 절댓값의 내림차순으로 Table 3.5에 정리하였다. 회귀계수가 음의 값을 가지면 해당 설명변수의 값이 커질수록 순위의 값 자체가 작아져서 순위가 높아진다고 해석할 수 있다. 해석해보자면, 등급이 강급될수록 종합득점이 높을수록 연대율이 높을수록 순위가 높아지고, 번호 4번을 부여받을수록 승급될수록 최근성적의 순위가 낮을수록 순위가 낮아진다. 또한 현재등급은 범주형 변수이기 때문에 (범주 개수 $-$ 1)개 만큼의 더미 변수를 생성한다. 각 모형에서 선택된 현재등급의 더미 변수를 살펴보면 각 등급(선발, 우수, 특선) 내에서 높은 세부 등급 순서대로 회귀계수의 절댓값이 줄어드는 것을 살펴볼 수 있다. 즉, 등급 내에서 높은 세부 등급에 속할수록 순위가 높아지는 것을 알 수 있다.

Table 3.5. The important variables of each linear model

	선택된 변수 (+)	선택된 변수 (−)
순위1	번호, 200M기록2, 등급변동(+), 최근 1회 낙차 여부, 최근 1회 실격 여부, 최근성적2, 최근 3회 결승진출회수, 나이, 추입2, 최근 1회 경기 일수차	현재등급SS, 등급변동(−), 종합득점2, 현재등급A1, 현재등급S1, 현재등급B1, 현재등급S2, 현재등급A2, 상대전적, 승률2, 연대율2
순위2	번호, 등급변동(+), 최근성적2	등급변동(−), 종합득점2, 연대율2
순위3	번호, 200M기록2, 등급변동(+), 최근 1회 낙차 여부, 최근 1회 실격 여부, 최근성적2, 최근 3회 결승진출회수, 추입2, 젖히기2, 나이, 최근 1회 경기 일수차, 마크2, 선행2	기어배수, 현재등급SS, 성적순위2, 현재등급A1, 등급변동(−), 종합득점2, 현재등급S1, 현재등급B1, 현재등급S2, 기어변동(+), 현재등급A2, 현재등급B2, 현재등급S3, 상대전적, 현재등급A3, 기수, 승률2, 연대율2, 삼연대율2, 잔여위반점수

Table 3.6. Average misclassification rates in validation data using regression models

	All Subset	Ridge	Lasso	PCR	Random Forest
순위1	0.3901 (0.0147)	0.3934 (0.0137)	0.3885 (0.0144)	0.3910 (0.0141)	0.3952 (0.0133)
순위2	0.5880 (0.0136)	0.5862 (0.0144)	0.5906 (0.0139)	0.5903 (0.0129)	0.5893 (0.0126)
순위3	0.6822 (0.0137)	0.6848 (0.0141)	0.6866 (0.0121)	0.6854 (0.0147)	0.6900 (0.0121)

각 회귀 모형을 통해 validation data에서의 평균 오분류율(표준편차)을 계산한 결과, 순위1 모형은 Lasso회귀모형, 순위2 모형은 Ridge회귀모형, 순위3 모형은 All Subset회귀모형이 가장 오분류율이 낮게 나타났다 (Table 3.6). 각 회귀 방법별로 차이가 크지 않기 때문에 가장 간단하고 해석이 쉬운 All Subset회귀모형을 최종 모형으로 선택한다. 하지만 앞서와 마찬가지 이유로 마지막 배당률 적용은 모든 회귀 방법을 사용하도록 한다.

3.3. 배당률 이용 결과

우리는 앞에서 제시한 순위 예측모형들을 이용하여 모형적합에 사용하지 않은 test data인 2016년 5월에 광명에서 실시된 총 168경기에 대하여 한 경기당 10,000원씩 모의베팅을 실시하였다. 이윤은 세금 공제 전 금액으로 168개의 경기에 대해 경기를 정확하게 예측한 경우에는 10,000원 × (배당률 − 1), 맞추지 못한 경우에는 −10,000원으로 계산하여 합을 구하였다. 순위1, 2, 3 모형의 test data에서의 오분류율은 Table 3.7에 나타내었고, 예측모형으로 구한 예측 순위로 각각에 해당하는 승식인 단승식, 복승식, 삼복승식에 베팅하였을 때 얻은 이윤은 Table 3.8에 나타내었다.

베팅 결과를 살펴보면 먼저 순위1 모형에서는 오분류율이 가장 낮게 나타난 모형은 보정된 자료를 사용해서 AIC 기준으로 변수 선택한 로지스틱 회귀모형, All Subset, Ridge, Lasso, PCR 모형이고, 값은 0.3571이다. 즉, 64.29%를 정확하게 예측하는 것이다. 이 중 All Subset, Ridge, Lasso, PCR 모형은 이윤이 −16.9로 음수가 나왔다. 다시 말해 168만원을 베팅해서 1,511,000원만큼을 환급받게 된다. 순위2 모형에서 경기 오분류율이 가장 낮은 모형은 All Subset회귀모형이고 오분류율 값은 0.5595이다. 경기 순위1·2등을 정확하게 맞추는 확률이 44.05%인 것이다. 복승식에 베팅한 결과 이윤이 가장 크게

Table 3.7. Misclassification rates in test data

		순위1 모형	순위2 모형	순위3 모형
Classification -원자료	Logistic	0.3690	0.6250	0.7202
	Logistic-AIC	0.3750	0.6250	0.7262
	Logistic-BIC	0.3631	0.6190	0.7202
	GAM	0.3810	0.6250	**0.7083**
Classification -보정된 자료	Logistic	0.3631	0.5833	0.7262
	Logistic-AIC	**0.3571**	0.5893	0.7321
	Logistic-BIC	0.3750	0.5833	0.7202
	GAM	0.3690	0.5833	0.7500
Regression	All Subset	**0.3571**	**0.5595**	0.7262
	Ridge	**0.3571**	0.5893	0.7202
	Lasso	**0.3571**	0.5714	0.7202
	PCR	**0.3571**	0.5714	0.7321
	Random Forest	0.3810	0.5714	0.7262

Table 3.8. Betting profits in test data

		단승식	복승식	삼복승식	총이윤
Classification -원자료	Logistic	**−16.9**	−30.7	−30.6	−78.2
	Logistic-AIC	−23.0	−30.7	−35.6	−89.3
	Logistic-BIC	−19.9	−23.0	−30.0	−72.9
	GAM	−20.4	−34.5	**−14.9**	−69.8
Classification -보정된 자료	Logistic	−19.5	−20.5	−36.2	−76.2
	Logistic-AIC	−18.2	−22.1	−36.8	−77.1
	Logistic-BIC	−23.0	−6.4	−19.5	−48.9
	GAM	−19.7	−17.6	−52.5	−89.8
Regression	All Subset	**−16.9**	−2.9	−30.9	−50.7
	Ridge	**−16.9**	−23.0	−21.6	−61.5
	Lasso	**−16.9**	−11.8	−21.6	−50.3
	PCR	**−16.9**	−17.8	−33.6	−68.3
	Random Forest	−17.8	**12.1**	−28.8	−34.5

나타난 모형은 오분류율이 가장 낮게 나타난 All Subset회귀모형이 아닌 랜덤 포레스트(회귀)이고 이윤이 121,000원 발생하였다. 오분류율이 가장 낮은 모형과 이윤이 가장 큰 모형이 다른 이유는 오분류율이 낮아도 어떤 배당률을 가지는 경기를 맞췄느냐에 따라 이윤은 달라지기 때문이다. 마지막으로 순위3 모형에서 경기 오분류율이 가장 낮은 모형은 원자료를 이용한 일반화가법모형(GAM)이고 오분류율은 0.7083이므로 29.17%를 정확하게 맞추는 것이다. 이 경우에도 이윤금액은 음수로 나타났다.

순위1, 2, 3 모형에서 모두 Random guess에 비해 맞춘 확률이 높지만 전체 이윤이 음수가 나온 이유를 파악하기 위해 각 순위 모형에서 가장 좋은 성과를 낸 모형의 결과들을 살펴보았다. Table 3.9는 해당 모형에서 맞춘 경기의 수와 배당률의 평균과 중앙값을 나타낸 것이다.

배당률의 경우 예를 들어 배당률이 1이면 베팅에 지불한 금액만큼 배당금을 받게 되는 것이므로 수익은 0이 된다. 배당률이 2이면 베팅에 지불한 금액의 두 배를 배당금으로 받게 되고 수익은 베팅에 지불한 금액만큼이 된다. Table 3.9를 살펴보면 순위1 모형의 경우 168경기 중 108경기를 맞추었지만 맞춘 경기의 배당률들이 1에 가까운 값들이다. 즉, 맞추어도 얻는 이윤이 0에 가까운 작은 값들인데 60경기를 맞추지 못해서 손실이 60만큼 생겨 이윤이 음수가 나온다. 순위2 모형의 경우 순위1 모형보다는 맞춘

Table 3.9. Dividend rate comparison of correct and incorrect predictions

총168경기		맞춘 경기의 배당률			못 맞춘 경기의 배당률		
		경기수	평균	중앙값	경기수	평균	중앙값
순위1 모형	All Subset	108	1.3991	1.2	60	7.4383	4.9
순위2 모형	Random Forest-회귀	74	2.4338	1.8	94	16.0160	7.9
순위3 모형	GAM-원자료	49	3.1245	2.4	119	29.0336	9.3

경기의 배당률 평균과 중앙값이 2에 가깝게 높아졌지만 못 맞춘 경기의 배당률이 더 높아졌다. 거기에 다 맞춘 경기의 수까지 74경기로 줄어들면서 74경기를 맞추면서 얻는 이익보다 94경기를 맞추지 못해서 잃는 손실이 더 큰 경우가 많아서 랜덤포레스트(회귀)를 제외한 나머지 모형에서는 총 이윤이 음수가 나온다. 랜덤포레스트(회귀)의 경우 다른 회귀모형들과 오분류율은 비슷한데 배당률이 높은 경기를 더 많이 맞춰서 이윤이 양수가 나온 것으로 볼 수 있다. 마지막으로 순위3 모형은 맞춘 경기의 배당률이 높아졌지만 못 맞춘 경기의 배당률이 더 큰 폭으로 상승했고, 맞춘 경기의 수 역시 49경기로 못 맞춘 119경기보다 현저히 작기 때문에 손실이 발생하였다고 볼 수 있다.

4. 결론

본 연구에서는 경륜 경기의 순위 예측을 위해 국민체육진흥공단 경륜경정사업본부에서 제공하는 경주출주표와 경주결과자료를 사용하여 변수를 생성하였다. 예측 방법으로는 경기 순위를 2-class로 변환한 Classification 방법과 순위를 수치로 사용한 Regression 방법을 사용하였고 각 모형의 오분류율을 비교해보고 중요 변수도 살펴보았다.

순위를 2-class로 변환하여 사용한 Classification 방법에서는 모든 변수를 사용한 로지스틱 회귀와 AIC, BIC 기준 단계별 변수선택법을 이용한 로지스틱 회귀, 일반화가법모형, 랜덤 포레스트, 그래디언트 부스팅 모형을 사용하였다. BIC 기준 단계별 변수선택법을 이용한 로지스틱 회귀를 통해 1) 등급이 강급될수록 2) 종합득점이 높을수록 3) 상대전적이 높을수록 순위권에 진입할 확률이 높아지고, 1) 승급될수록 2) 나이가 많을수록 3) 최근 1회전 경기를 한지 오래 될수록 4) 번호 4번을 부여받아 선행할수록 5) 최근성적이 나쁠수록 순위권에 진입할 확률이 낮아짐을 확인할 수 있었다. 또한 랜덤 포레스트와 그래디언트 부스팅 모형에서 원자료를 사용했을 때 중요 변수는 등급 간의 차이를 조정할 필요가 없는 현재등급과 등급 별로 경기가 이루어진 최근성적 등수이고, 보정된 자료를 사용했을 때 중요 변수는 등급별로 값의 범위가 달라지는 종합득점과 성적순위라는 사실을 확인할 수 있었다.

다음으로 순위를 수치로 사용한 Regression 방법에서는 All Subset, Ridge, Lasso, 주성분 회귀, 랜덤 포레스트 모형을 적합시켜 보았다. All Subset회귀모형의 회귀 계수를 통해 1) 강급될수록 2) 종합득점이 높을수록 3) 연대율이 높을수록 순위가 높아지고, 1) 승급될수록 2) 번호가 4번일수록 3) 최근성적의 순위가 낮을수록 순위가 낮아진다는 것을 확인할 수 있었다. 또한 등급 내에서 높은 세부 등급에 속할수록 순위가 높아지는 것을 알 수 있었다.

Classification 방법과 Regression 방법에서의 모형별 예측률을 비교해 보았을 때, 모든 모형에서 나쁘지 않은 예측률을 보였다. 실제 최근 한 달 경기를 예측해보았을 때도 예측률은 train data에서의 결과와 비슷하였지만, 금액을 베팅한 결과 손해가 발생하였다. 보통 최종 배당금이 높은 경기는 베팅을 적게 받은 선수가 승리하는 경우이다. 따라서 실제 베팅에서 손해가 발생한 이유는 우리가 모형을 만들 때 배당률은 고려하지 않고, 경기 오분류율을 줄이는 방향으로 모형들을 적합시켰고 만들어진 모형의 유의한 변수들이 대부분 선수들의 이전 경기 성적 관련된 변수들이기 때문에 배당률이 큰 경기를 잘 예측하기는 힘들기 때문이다. 나아가서 사람들의 실제 관심이 경기 순위를 잘 예측하는 것이 아니라 배당금이 높은

경기를 잘 예측하는 것이라면 배당률을 모형에 반영하여 배당금을 높이는 방향으로 분석을 진행해 보는 것도 좋을 것이라고 생각된다.

References

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

Cho, H. C., Kang, S. K., and Kim, J. K. (2008). Relationship of lower extremity factors, 200m record and Wingate anaerobic power in racing and competitive cyclists, *Korean Journal of Sport Science*, **19**, 9–20.

Choe, H., Hwang, N., Hwang, C., and Song, J. (2015). Analysis of horse races : prediction of winning horses in horse races using statistical models, *Korean Journal of Applied Statistics*, **28**, 1133–1146.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning; Data Mining, Inference and Prediction*, Springer, New York.

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

Kim, B. S. and Kim, J. S. (2007). Relationship between customer satisfaction and patterns of ticket purchase of cycle racing customers, *Journal of Sport and Leisure Studies*, **30**, 203–211.

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

Ridgeway, G. (2006). Generalized boosted models: a guide to the gbm package, Available from: http://cran.r-project.org/web/packages/gbm

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

통계적 예측모형을 활용한 경륜 경기 순위 분석

박가희[a] · 박리라[a] · 송종우[a,1]

[a]이화여자대학교 통계학과

(2016년 9월 7일 접수, 2016년 10월 21일 수정, 2016년 10월 26일 채택)

요 약

최근 경륜은 2015년도 기준, 5백만 명 이상의 많은 사람들이 참여하고 2조를 넘어선 매출을 발생시키는 대중적인 레저스포츠로서 자리 잡고 있다. 본 연구의 목적은 다양한 통계적 분석기법을 사용하여 경륜경기의 순위를 예측하고, 순위에 유의한 영향을 미치는 변수들을 파악하는 데에 있다. 다양한 Classification 방법과 Regression 방법들을 적용하여 순위예측모형을 만들고 비교분석하였다. 대부분의 모형에서 공통적으로 선택된 변수들을 살펴보면, 등급이 강급될수록, 종합득점이 높을수록 순위가 높아지며 반대로 등급이 승급될수록, 번호 4번을 부여받을수록 그리고 최근성적의 순위가 낮을수록 순위가 낮아지는 것을 알 수 있었다. 또한, 선수의 실력과 관련된 연속형 변수들을 각 경기별로 평균값을 빼서 보정한 자료와 원자료를 사용하여 모형을 적합시킨 결과 모든 모형에서 보정된 자료를 사용하였을 때 더 낮은 오분류율을 보였다. 마지막으로 분석에 사용하지 않은 최근 한 달 경기결과를 예측해서 베팅했을 때 모든 경우에 예측률은 높았지만 큰 이익을 거두지 못했는데 그 이유는 낮은 배당률을 가진 경기의 결과만을 잘 예측했기 때문이다.

주요용어: 경륜, 단계적 회귀분석, 로지스틱 회귀모형, 랜덤 포레스트, 일반화 가법 모형, 그래디언트 부스팅, All Subset회귀모형 Ridge회귀모형, Lasso회귀모형, 주성분회귀모형, 주요변수

이 논문은 2015년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2015S1A5B6036244).
[1]교신저자: (03760) 서울특별시 서대문구 이화여대길 52, 이화여자대학교 통계학과.
E-mail: josong@ewha.ac.kr