데이터 마이닝을 활용한 한국 프로야구 구단의 승패예측과 승률 향상을 위한 전략 도출 연구

Development of Win-Loss Prediction Models and Strategies for Improving Winning Rate of the Korean Professional Baseball Teams Using Data Mining Techniques

저자 (Authors)
김원종, 최연식, 유동희
Kim, Won-jong, Choi, Yeon-sik, Yoo, Dong-hee

출처 (Source)
한국스포츠산업경영학회지 23(3), 2018.6, 88-104(17 pages)
Korean Society For Sport Management 23(3), 2018.6, 88-104(17 pages)

발행처 (Publisher)
한국스포츠산업경영학회
Korean Society For Sport Management

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07478875

APA Style
김원종, 최연식, 유동희 (2018). 데이터 마이닝을 활용한 한국 프로야구 구단의 승패예측과 승률 향상을 위한 전략 도출 연구. 한국스포츠산업경영학회지, 23(3), 88-104

한국스포츠산업경영학회지, 2018, 제23권 3호, pp.88-104
Korean Journal of Sport Management. 2018, Vol. 23, No.3, pp.88-104
http://dx.doi.org/10.31308/KSSM.23.3.6

데이터 마이닝을 활용한 한국 프로야구 구단의 승패예측과 승률 향상을 위한 전략 도출 연구

김원종 · 최연식 · 유동희*(경상대학교)

본 연구에서는 데이터 마이닝 기법을 활용하여 한국 프로야구의 승패예측모형을 구축하는 실험을 진행하였다. 이를 위해, 2017년에 실시된 한국 프로야구 10개 구단의 전체 경기 중 무승부 경기를 제외한 1,418경기에 대한 자료를 사용하였다. 승패예측모형에는 의사결정나무, 베이즈넷, 인공신경망 알고리즘과 앙상블 기법인 배깅과 부스팅이 사용되었으며, 그 결과 배깅 기법에 인공신경망을 적용한 예측모형에서 가장 높은 예측률인 85.18%를 기록하였다. 다음으로 의사결정나무 기반 예측모형을 활용하여 한국 프로야구 전체 구단에 관한 8개의 승패규칙을 도출하였다. 여기에서 승패규칙은 승패예측에 영향을 미치는 주요 요인들인 팀출루율, 팀타율, 피안타, 안타, 타석, 타수로 표현되며, 도출된 규칙을 바탕으로 구단의 승률 향상에 도움을 주는 전략을 제안하였다. 또한 플레이오프 진출 구단과 미진출 구단에 관한 승리규칙을 각각 4개씩 도출하였고 이를 바탕으로 두 집단에 맞춤화된 승률 향상 전략을 제시하였으며, 실제 구단에서 선수를 영입한 방향과의 비교를 통해 제시된 전략의 활용 가능성을 확인하였다.

※ 주요어: 한국프로야구, 데이터마이닝, 승패예측모형, 의사결정나무, 베이즈넷, 인공신경망, 배깅, 부스팅

* e-mail : dhyoo@gnu.ac.kr

Ⅰ. 서 론

1. 연구 배경

2006년 국제야구대회(World Baseball Classic)에서의 4강 진출, 2008년 베이징 올림픽에서의 금메달, 그리고 2009년 국제야구대회에서의 준우승 등 한국야구는 국제대회에서 우수한 성적을 거두고 있다. 이를 바탕으로 한국 프로야구도 중흥기를 맞이하였으며, 국내 스포츠 중 최대관중 수인 800만 관중을 돌파하며 흥행몰이를 하고 있다(한국프로야구위원회, 2017). 1990년 이후로 8개 구단이었던 한국 프로야구에 여러 기업들이 참가를 희망하면서 현재는 10개 구단으로 시즌 경기가 진행되고 있다. 한국 프로야구는 2010년 이후 한국 프로스포츠 중에서 가장 인기 있는 종목이 되었으며 다양한 언론 매체를 통한 접근성도 높아지고 있다. 프로야구 시즌이 시작되면 케이블의 스포츠 채널에서 매일 프로야구 경기가 중계되었고, TV, 인터넷, 모바일 앱 등 다양한 프로야구 중계 플랫폼을 통한 경기 시청 또한 가능해 졌다. 이와 같은 흐름에 따라 이제 프로야구는 단순한 스포츠가 아닌 하나의 문화산업으로 발전하고 있다. 현재 국내의 모든 프로스포츠 산업 중에서 프로야구 산업이 방송 중계권료나 스폰서십 계약금 등에 있어서 가장 큰 시장 규모를 차지하고 있다. 여기에 매일매일 자신의 응원구단의 승패여부를 확인하는 팬들이 있는가 하면, 경기와 관련된 선수들의 각종 지표들을 외우며 어느 구단이 승리할지를 예측하는 팬들도 늘어나고 있다.

이와 같은 프로야구의 성장에 힘입어 많은 구단들이 구단의 수익을 높이기 위한 다양한 비즈니스 활동들을 진행하고 있다. 일반적으로 기업에서 지불하는 스폰서 비용을 제외하고 프로야구의 수익창출에서 가장 큰 비중을 차지하는 것은 관중 동원을 통한 티켓 판매와 관중을 대상으로 한 식음료 및 기념품 판매이다(장경로, 김민철, 2014). 여기에서 구단의 성적이 오를수록 더 많은 관중들이 경기장을 찾게 되어 관중 동원을 통한 수익이 증가하게 된다(이주호, 염준근, 송근혁, 박홍준, 2010). 따라서 각 구단들은 우선적으로 구단의 승률을 높이기 위한 다양한 시도들을 진행하고 있다.

그 중 많은 구단들이 승률 향상을 위한 전략을 마련하기 위해 야구 경기에 관한 데이터를 활용하고 있다(Beneventano, Berger & Weinberg, 2012). 흔히들 야구는 기록의 스포츠라고 한다. 수많은 경기를 치르며 쌓인 기록을 바탕으로 승률을 높이는 전략을 구상할 수 있다. 예를 들어, 미국 프로야구단 '오클랜드 애슬레틱스'의 단장이었던 빌리 빈이 추구한 '머니 볼 이론'에서는 경기 데이터를 분석한 결과를 활용하여 선수들을 적재적소에 배치하여 구단의 승률을 높였다(Lewis, 2003). 이는 이후에 선수의 기록을 수치화한 후 선수를 객관적으로 평가하는 방법론인 세이버메트릭스(sabermetrics)가 개발되는 배경이 되었다. 현재는 대다수의 프로야구 구단들이 세이버메트리션을 고용하여 선수들의 기록을 분석하고 훈련에 활용하며 구단의 승률을 높이는데 중점을 두고 있다.

본 연구에서는 데이터 마이닝 기술을 활용하여 야구 경기에 관한 데이터 분석을 실시하고 한국 프로야구 구단의 승률을 높일 수 있는 실용적인 전략들을 도출하는 방법을 보여주고자 한다. 이를 위해, 2017년 한국 프로야구 시즌에서 발생된 모든 구단들의 경기 데이터를 수집하고자 한다. 분석 방법으로는 데이터 마이닝 기법을 사용하여 한국 프로야구 경기에 관한 승패예측모형을 구축하고자 한다. 승패예측모형은 의사결정나무(decision tree), 베이즈넷(Bayes net), 인공신경망(artificial neural network) 알고리즘과 앙상블 기법인 배깅(bagging)과 부스팅(boosting)을 활용하여 구축하고자 한다. 이를 통해, 한국 프로야구의 승패예측에 가장 높은 예측률을 보이는 예측모형을 개발하고자 한다. 다음으로, 구축된 의사결정나무 기반 예측모형을 활용하여 한국 프로야구 전체 구단의 승패규칙을 도출하고 도출된 규칙을 바탕으로 구단의 승률을 향상시키기 위한 전략을 제시하고자 한다. 또한 플레이오프 진출 구단과 미진출 구단의 승리규칙을 각각 분석하여 두 집단에 맞춤화된 승률 향상 전략을 제시하고, 실제 구단에서 선수를 영입한 방향과의 비교를 통해 제시된 전략의 활용 가능성을 확인하고자 한다.

2. 문헌 연구

프로야구 경기에서 구단의 승률을 높이기 위해서는 과학적인 분석 기법을 활용하여 승패에 영향을 주는 다양한 요인들을 분석하고 분석된 결과를 바탕으로 구단의 전력을 최대치로 높일 수 있는 전략들이 마련되어야 한다. 본 장에서는 한국 프로야구에 있어서 구단들의 승률 향상을 위한 요인들을 분석한 기존 연구들을 조사하였다.

먼저 프로야구 경기의 승패에 영향을 주는 요인을 분석한 연구들을 요약하면 다음과 같다. 조영석, 조용주와 신상근(2007)의 연구에서는 로지스틱회귀분석, 판별분석, 그리고 의사결정나무분석을 통해 경기 승패에 영향을 주는 변수들과 변수들의 우선순위를 분석하였다. 그 결과, 로지스틱회귀분석과 판별분석에서는 안타와 피안타가 중요도가 높은 변수로 분석되었고, 의사결정나무분석에서는 피안타, 피4구, 병살 순으로 중요도가 높은 변수임이 파악되었다. 분석된 결과를 토대로 구단의 승리를 높이기 위해 상대구단보다 더 높은 출루율을 획득할 필요가 있음을 알 수 있었다.

추가적으로 오윤학, 김한, 윤재섭과 이종석(2014)의 연구에서는 다양한 데이터 마이닝 알고리즘을 활용하여 승패예측모형을 수립하였는데, 그 결과 랜덤포레스트 알고리즘을 사용한 예측모형의 예측률이 가장 높은 것으로 확인되었다. 또한 승패에 영향을 주는 중요 요인들을 파악하기 위해 의사결정나무, 랜덤포레스트, 로지스틱회귀분석을 통해서 각각의 요인들을 점수로 변환하여 비교한 결과 이닝 당 삼진, 구단의 평균연봉, 투수의 평균 자책점 순으로 중요도가 높음이 분석되었다. 여기에서 구단의 승률을 높이기 위해 구단의 연봉이 높아야 한다는 점과 투수의 성적이 중요하다는 점을 이야기 했다.

장진희와 문춘걸(2014)은 구단 승률에 영향을 주는 요인을 공격력, 수비력, 작전 스타일과 수행 능력, 감독의 지도력, 선수층 두께와 선수구성 특성 등으로 구분하고 고정효과 패널회귀모형으로 데이터를 분석하였다. 그 결과, 출루율, 장타율, 이닝 당 출루허용이 중요 요인으로 파악되었으며, 승률에 있어서 투수의 능력이 야수의 수비력 보다 중요하다고 언급하였다.

한편, 선행 연구에서는 포스트시즌 진출 구단과 미진출 구단의 경기력을 분석한 연구들도 진행되었다. 구체적으로 채진석과 엄한주(2010)의 연구에서는 포스트시즌 진출여부에 영향을 주는 요인들을 투수력, 타력, 득점력, 기동력, 그리고 수비력 관점에서 구분하여 분석하였다. 두 집단의 평균을 비교하기 위해 t-test를 실시하였고, 포스트시즌 진출에 영향력이 높은 변수로 방어율, 세이브, 출루율, 출루율과 장타율의 합(OPS), 그리고 타율이 분석되었다. 이를 통해 구단의 포스트시즌 진출여부의 예측은 타력 및 득점력과 투수력 영역만으로도 충분히 높은 적중률을 얻을 수 있으며 수비력은 미미한 기여를 하는 것을 알 수 있었다.

또한 채진석, 조은형과 엄한주(2010)의 연구에서는 포스트시즌 진출에 관한 예측모형을 여러 통계적 방법을 사용하여 구축하였는데, 인공신경망 기반의 예측모형에서 예측변인으로 세이브, 득점, 실점, 피희생타를 사용하였을 때 가장 높은 예측률을 기록하였다.

한국 프로야구 경기에서의 득점에 영향을 주는 요인을 분석한 연구들도 진행되었다. 조영석과 조용주(2005)는 OPS와 득점간의 관계를 분석하였다. 먼저 OPS를 구성하는 출루율과 장타율을 기준으로 군집분석을 실시한 결과 출루율과 장타율 모두가 가장 높은 군집이 경기당 득점도 가장 높았다. 다음으로 상관관계 분석하였는데 OPS가 타율, 장타율, 출루율에 비해 득점과 높은 상관관계를 보여주었다. 이를 토대로 경기당 득점을 추정하는 주요 변수로 OPS가 활용될 수 있음을 알 수 있었다.

아울러 신상근, 이경준, 조영석과 박찬근(2010)은 특정 구단의 사례를 중심으로 득점 여부에 영향을 주는 요인을 분석하였다. 로지스틱회귀분석에서는 한 이닝 당 득점에 가장 영향을 미치는 요인으로는 3번째 타자의 3루타, 1번째 타자의 3루타, 그리고 2번째 타자의 3루타 순으로 주요 영향요인이 분석되었고, 의사결정나무분석에서는 1번째 타자의 1루타, 1번째 타자의 2루타, 1번째 타자의 4구 3번째 타자의 1루타 순으로 파악되었다. 이를 활용하여 구단의 득점을 높이기 위한 효과적인 출루 전략의 수립을 기대할 수 있다.

김창권과 진서훈(2014)은 회귀분석을 통하여 득점과 실점이 승패에 어떤 영향을 주는지를 분석하였다.

그 결과, 득점과 관련된 요인인 장타율이 1할 증가하면 11.1승의 효과가 있고 출루율이 1할 증가할 경우 29.6승의 효과가 있다고 언급하였다. 또한 실점관련 요인인 이닝 당 베이스 허용수가 1할 증가하면 4.4패의 효과가 발생한다고 밝혔다.

지금까지 살펴본 문헌들의 내용 통해, 구단의 승률 향상에 영향을 주는 요인들은 투수력, 타력, 수비력, 득점 등과 같이 경기 결과와 직접 관련이 있는 요인과 평균 연봉, 감독 연차 등과 같은 기타 요인으로 구분할 수 있다. 여기에서 투수력 부분에서는 피안타가 낮을수록, 타력 부분에서는 출루율이 높을수록 승리할 확률이 높다는 결과가 도출되었다. 또한 데이터의 분석 시기에 따라 투수력과 타력이 승패에 영향을 주는 영향력의 정도가 변하는 것을 파악할 수 있었다.

기존 연구와 본 연구와의 차이점을 요약하면 다음과 같다. 기존 연구에서는 주로 통계적 기법을 활용하여 승패에 영향을 주는 요인들의 영향력을 분석하거나 데이터 마이닝 분석을 통해 승패예측모형을 구축하는 것에 초점을 둔 연구들이 진행되었다. 따라서 현재 한국 프로야구 데이터를 분석한 연구 중에서 앙상블 기법 및 언더샘플링과 같은 정교한 데이터 마이닝 분석 기법을 활용하여 프로야구의 승패예측모형을 구축한 연구가 없으며, 분석 결과를 토대로 구단의 승률 향상 전략을 제안한 연구라는 점에서 기존 연구와의 차별성을 가진다.

Ⅱ. 연구방법

1. 연구 진행 과정

본 연구의 진행과정을 요약하면 <그림 1>과 같다. 먼저 '실험 1'에서는 한국 프로야구 전체 구단에 관한 데이터를 활용하여 가장 성능이 좋은 승패예측모형을 발견하고, 한국 프로야구에 관한 승패규칙을 도출하고자 한다. 이를 위해, 2017년에 진행된 한국 프로야구 전체 구단의 경기 데이터를 수집하고자 한다. 전처리(preprocessing) 과정을 통해 분석에 사용될 목표변수와 독립변수들을 결정한다. 다음으로 변수 선택(feature selection) 과정을 통해 목표변수에 영향을 주

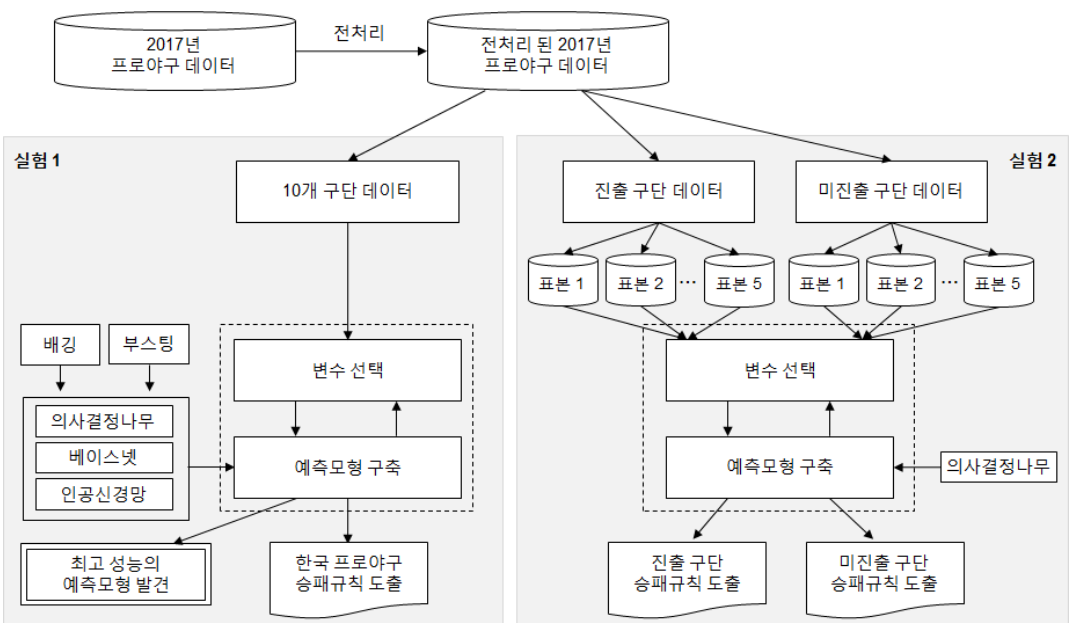


그림 1. 연구 진행 과정

는 독립변수들을 최종 선택하고, 의사결정나무, 베이즈넷, 인공신경망 알고리즘과 앙상블 기법인 배깅과 부스팅을 사용하여 여러 형태의 승패예측모형을 구축한다. 이렇게 구축된 여러 예측모형들 중 가장 예측률이 높은 예측모형을 발견하고자 한다. 또한, 앞서 구축된 예측모형들 중, 규칙의 설명력이 좋은 의사결정나무 기반 예측모형의 결과를 활용하여 한국 프로야구 전체 구단에 관한 승패규칙을 도출하고자 한다.

다음으로 '실험 2'에서는 한국 프로야구 전체 구단의 데이터를 플레이오프 진출 구단과 미진출 구단으로 구분한 뒤, 의사결정나무 알고리즘을 활용하여 플레이오프 진출 구단과 미진출 구단에 관한 승패규칙을 각각 도출하고자 한다.

2. 분석 데이터

본 연구에서는 Statiz(http://www.statiz.co.kr)를 통해 한국 프로야구 경기에 관한 자료를 수집하였다. 2017년도에는 10개의 프로야구 구단이 경기를 진행했고, 구단 별 144경기를 가졌기 때문에 총 1440경기에 대한 데이터가 수집되었다. 그 중 무승부 22경기를 제외한 1418경기에 대한 데이터를 분석하고자 하였다. 수집된 데이터는 승패여부를 포함하여 매 경기의 각종 지표인 차전, 상대, 홈/어웨이, 타석 등과 같은 25개의 변수들의 값을 포함 하고 있다(<표 1>의 전처리 전 사용변수 참고).

표 1. 실험에서 사용한 변수 정보

종류	변수명	변수 설명	전처리 전 사용변수	전처리 후 사용변수	변수 번호
제거변수	차전	경기 수	○		F1
제거변수	상대	경기의 상대 구단	○		F2
독립변수	홈/어웨이	홈구장/어웨이 구장	○	○	F3
독립변수	타석	타자가 배터박스에 들어온 수	○	○	F4
독립변수	타수	타자가 타석에 들어서서 타격을 완료한 횟수 (볼넷, 희생 번트, 타격 방해 등은 타수에 포함되지 않음)	○	○	F5
제거변수	득점	타자가 홈에 들어온 점수	○		F6
독립변수	안타	타자가 베이스에 나아갈 수 있도록 공을 치는 일	○	○	F7
독립변수	홈런	타자가 홈 베이스까지 살아서 돌아올 수 있도록 친 안타	○	○	F8
제거변수	타점	타자가 안타 등으로 자기 구단에 얻게 한 점수	○		F9
독립변수	볼넷	투수가 타자에게 스트라이크가 아닌 볼을 네 번 던지는 일	○	○	F10
독립변수	사구	투수가 던진 공이 타자의 몸에 닿는 일	○	○	F11
독립변수	삼진아웃	타자가 세 번 스트라이크를 당하여 그대로 아웃되는 것	○	○	F12
독립변수	땅볼아웃	땅 위를 굴러가도록 치거나 찬 공	○	○	F13
독립변수	플라이아웃	타자가 쳐서 땅에 한 번도 닿지 않고 야수가 잡아 아웃되는 것	○	○	F14
독립변수	투구 수	투수가 던진 공의 수	○	○	F15
독립변수	병살	두 사람의 주자를 한꺼번에 아웃시키는 일	○	○	F16
독립변수	잔루	공격 구단과 수비 구단이 교체할 때에 주자가 득점에 성공하지 못하고 베이스에 남아 있는 일	○	○	F17
독립변수	피안타	투수가 맞은 안타의 개수	○	○	F18
제거변수	실점	수비 구단이 잃은 점수	○		F19
제거변수	자책점	투수의 잘못으로 상대 구단에 준 점수	○		F20
독립변수	피볼넷	투수가 타자에게 내준 볼넷의 개수	○	○	F21
독립변수	피사구	투수가 타자에게 내준 몸에 맞는 공의 수	○	○	F22
독립변수	탈삼진	투수가 상대 타자에게 잡은 삼진의 수	○	○	F23
독립변수	피홈런	투수가 맞은 홈런의 개수	○	○	F24
독립변수	팀타율	구단 전체의 안타수를 타격수로 나눈 백분율		○	F25
독립변수	팀출루율	타자가 베이스로 나간 횟수를 백분율로 나타낸 수치		○	F26
목표변수	승/패	구단의 승/패	○	○	F27

3. 목표변수와 독립변수

한국 프로야구 승패에 대한 예측모형을 구축하기 위해 승패 여부를 나타내는 '승/패' 변수를 목표변수로 선택하였다. 전체 분석 경기 중, 승리 경기와 패배 경기 수는 각각 709경기로 그 비율은 동일하다.

독립변수 선정을 위해 앞서 문헌연구에서 조사된 승패 관련 변수들이 고려되었고, 그 중 전문가 회의를 통해 승패 여부에 영향을 줄 것으로 판단되는 20개의 독립변수를 최종 결정하였다(<표 1>의 전처리 후 사용변수 참고). 즉, 분석에 불필요하다고 판단된 '차전'과 '상대' 변수와 독립변수의 값이 목표변수의 값과 거의 동일한 의미를 지니는 '득점', '타점', '실점' '자책점' 변수도 제외하였다. 그리고 기존의 '안타'의 개수를 '타수'로 나눈 '팀타율' 변수와 '안타', '볼넷', '사구'의 합을 '타수'로 나눈 '팀출루율' 변수를 기존의 변수들로부터 파생시켜 독립변수에 추가하였다.

4. 변수 선택

전처리 후 선정된 독립변수들 중에는 승패예측모형을 구축하는데 아무런 도움을 주지 못하는 것들이 존재할 수 있다. 목표변수의 예측에 큰 도움이 되지 않음에도 불구하고 예측모형 구축에 이용된다면 분류의 정확성을 저하시킬 수도 있기 때문에 이러한 변수들을 제거한 후 예측모형을 구축하는 것이 바람직하다(Dash & Liu, 1997). 이와 같이, 전처리 후 선정된 독립변수들 중 목표변수 분류에 불필요한 변수는 제외하고 중요한 역할을 하는 변수들만을 선택하는 과정을 변수 선정이라고 한다. 본 연구에서는 변수 선정 알고리즘으로 이득비(gain ratio)를 사용하여 독립변수들의 중요도를 평가하였고, 래퍼(wrapper) 방법 중 역방향 제거 방식을 사용하여 예측모형에 포함될 독립변수들을 최종 선택하였다. 여기에서 역방향 제거란 매회 가장 중요성이 낮은 변수를 하나씩 제거한 후 예측모형을 구축하는 방식으로 중요도가 가장 높은 변수 하나가 남을 때까지 그 과정이 진행되며, 전 단계에서 탈락된 변수의 영향을 제거한 상태에서 현재의 변수들 사이의 중요도가 반복되어 계산된다(Witten & Frank, 2005). 이 과정을 통해 변수의 개수만큼 예측모형이 구축되며, 구축된 여러 개의 예측모형들 중 가장 높은 정확도를 보인 변수들의 조합을 최종 변수로 선택하게 된다.

5. 전체 구단의 승패예측모형 구축

본 연구에서는 오픈소스 데이터 마이닝 툴인 Weka ver.3.8을 이용하여 프로야구 전체 구단의 승패에 관한 예측모형을 구축하였다. 예측모형 구축에는 의사결정나무, 베이즈넷, 인공신경망 알고리즘이 사용되었고, 분석 데이터들을 각각 70%의 학습데이터와 30%의 검증데이터 비율로 분할하여 실험을 진행하였다. 여기에서 학습데이터는 예측모형을 학습시키기 위해 사용되는 데이터이며, 검증데이터는 학습데이터를 통해 구축된 예측모형의 성능을 평가할 때 사용되는 데이터이다. 전처리 후 선택된 20개의 독립변수들 중 중요도가 낮은 변수를 하나씩 제거하면서 예측모형의 예측률을 계산하였다. 그 결과 각 알고리즘마다 역방향 제거를 통한 20개의 예측모형을 구축하였다. 다음으로 앙상블 기법인 배깅과 부스팅에 이 세 알고리즘을 적용하여 예측모형을 구축하는 과정에서도 역방향 제거를 진행하였다. 그 결과, '실험 1'에서는 최종적으로 180개의 예측모형이 구축되었다.

6. 플레이오프 진출 구단과 미진출 구단의 승패 예측모형 구축

본 연구에서는 구단의 특성이 반영된 맞춤화된 승률 향상 전략을 제안하기 위하여 구단의 특성을 플레이오프 진출 구단과 미진출 구단으로 구분하여 각 구단의 승리규칙을 분석하고자 한다.

플레이오프 진출 구단과 미진출 구단에 관한 승리규칙이 도출되는 과정은 <그림 1>의 '실험 2'와 같다. 전반적으로 실험이 진행되는 과정은 '실험 1'과 비슷하지만 분석 데이터로부터 표본을 5개씩 추출하는 부분에서 차이를 보인다. 그 이유는 진출 구단과 미진출 구단의 목표변수인 '승/패' 속성의 클래스 분포가 달라 표본 추출을 통한 데이터 균형화(data balancing)

작업이 필요하기 때문이다. 즉, 목표변수 내에 특정 클래스에 속하는 관측대상이 다른 클래스에 속하는 관측대상 보다 많을 경우, 관측대상이 많은 클래스를 중심으로 학습이 많이 이루어지기 때문에 특정 클래스만을 잘 분류하는 편향된 예측모형이 구축될 수 있다(Chawla, 2005). 따라서 예측모형을 구축하기 전 목표변수 내에 존재하는 클래스들의 비율을 맞추는 데이터 균형화 작업을 진행해야 한다.

플레이오프에 진출한 상위 다섯 구단의 관측대상의 수는 '승리'가 403경기이고 '패배'가 307경기이기 때문에 목표변수의 클래스 분포가 '승리' 방향으로 많이 편향되어 있다. 마찬가지로 플레이오프에 미진출한 하위 다섯 구단의 관측대상의 수는 '승리'가 319경기이고 '패배'가 393경기이기 때문에 목표변수의 클래스 분포가 '패배' 방향으로 많이 편향되어 있다. 이를 해결하기 위해, 본 연구에서는 각각 5개의 표본을 추출할 때 언더샘플링(under-sampling)을 통해 클래스의 비율을 맞추는 작업을 추가로 진행하였다. 그 결과 플레이오프 진출 구단은 표본별 614경기(승리: 307경기, 패배: 307경기), 플레이오프 미진출 구단은 표본별 638경기(승리: 319경기, 패배: 319경기)로 이루어진 데이터 셋을 최종 데이터 셋으로 구성하여 분석을 진행하였다.

'실험 1'과 동일한 방식으로 역방향 제거를 실시하여 표본별로 20개의 의사결정나무 기반 예측모형을 구축하였다. '실험 2'의 실험 목적은 승률 향상을 위한 전략 도출이기 때문에 다른 알고리즘들과 앙상블 기법은 사용되지 않았다. 그 결과, '실험 2'에서는 100개의 플레이오프 진출 구단의 승패예측모형과 100개의 미진출 구단의 승패예측모형이 구축되었다.

Ⅲ. 연구결과

1. 전체 구단의 승패예측모형 결과

<표 2>는 의사결정나무, 베이즈넷, 그리고 인공신경망을 통해 구축된 승패예측모형들의 예측 결과를 보여준다. 여기에서 변수 순서가 높을수록 중요도가

표 2. 의사결정나무, 베이즈넷, 인공신경망 기반 예측모형의 예측률 비교

변수 순서	의사결정나무 기반 예측모형		베이즈넷 기반 예측모형		인공신경망 기반 예측모형	
	제거변수	예측률(%)	제거변수	예측률(%)	제거변수	예측률(%)
1	F14	78.12	F14	72.71	F14	83.29
2	F16	78.12	F16	72.71	F16	81.18
3	F3	79.29	F3	72.71	F3	81.18
4	F13	80.24	F13	72.94	F13	80.71
5	F17	80.24	F17	72.94	F17	83.53
6	F23	78.35	F23	72.94	F23	80.47
7	F22	79.06	F22	72.24	F22	81.18
8	F11	79.29	F11	72.24	F11	79.76
9	F21	79.53	F21	72.47	F21	80.94
10	F10	78.35	F10	70.82	F10	83.53
11	F12	78.82	F12	70.35	F12	82.82
12	F15	78.82	F15	70.12	F15	82.82
13	F8	78.82	F8	72.00	F8	82.12
14	F24	79.76	F24	71.76	F24	80.71
15	F5	80.24	F5	69.65	F5	80.00
16	F4	79.53	F4	71.76	F4	79.06
17	F7	78.35	F7	72.00	F7	79.53
18	F18	77.41	F18	75.06	F18	78.82
19	F25	71.76	F25	67.76	F25	70.35
20	F26	67.53	F26	67.53	F26	68.94

높은 변수를 의미하며, 1번 변수부터 20번 변수 순으로 변수가 제거되면서 구축된 예측모형의 예측률을 보여주고 있다. <표 2>를 보면 의사결정나무 기반 예측모형은 6가지 변수인 F26, F25, F18, F7, F4, F5를 사용했을 때에 최고 예측률인 80.24%를 기록하였다. 베이즈넷 기반 예측모형은 3가지 변수인 F26, F25, F18을 사용했을 때 최고 예측률인 75.06%가 나타났고, 인공신경망 기반 예측모형의 경우 16가지 변수인

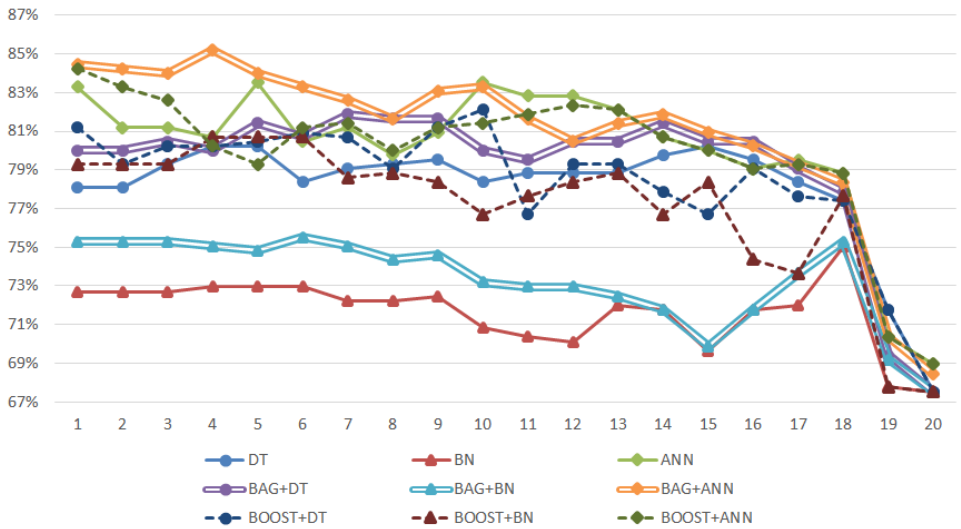


그림 2. 전체 구단의 승패예측에 관한 모형별 예측률 도식화

표 3. 배깅 작업을 통한 예측모형의 예측률 비교

변수 순서	배깅에서의 의사결정나무 기반 예측모형		배깅에서의 베이즈넷 기반 예측모형		배깅에서의 인공신경망 기반 예측모형	
	제거변수	예측률(%)	제거변수	예측률(%)	제거변수	예측률(%)
1	F14	80.00	F14	75.29	F14	84.47
2	F16	80.00	F16	75.29	F16	84.24
3	F3	80.47	F3	75.29	F3	84.00
4	F13	80.00	F13	75.06	F13	85.18
5	F17	81.41	F17	74.82	F17	84.00
6	F23	80.71	F23	75.53	F23	83.29
7	F22	81.88	F22	75.06	F22	82.59
8	F11	81.65	F11	74.35	F11	81.65
9	F21	81.65	F21	74.59	F21	83.06
10	F10	80.00	F10	73.18	F10	83.29
11	F12	79.53	F12	72.94	F12	81.65
12	F15	80.47	F15	72.94	F15	80.47
13	F8	80.47	F8	72.47	F8	81.41
14	F24	81.41	F24	71.76	F24	81.88
15	F5	80.47	F5	69.88	F5	80.94
16	F4	80.47	F4	71.76	F4	80.24
17	F7	79.06	F7	73.65	F7	79.29
18	F18	77.88	F18	75.29	F18	78.35
19	F25	69.41	F25	69.18	F25	70.35
20	F26	67.53	F26	67.53	F26	68.47

표 4. 부스팅 작업을 통한 예측모형의 예측률 비교

변수 순서	부스팅에서의 의사결정나무 기반 예측모형		부스팅에서의 베이즈넷 기반 예측모형		부스팅에서의 인공신경망 기반 예측모형	
	제거변수	예측률(%)	제거변수	예측률(%)	제거변수	예측률(%)
1	F14	81.18	F14	79.29	F14	84.24
2	F16	79.29	F16	79.29	F16	83.29
3	F3	80.24	F3	79.29	F3	82.59
4	F13	80.24	F13	80.71	F13	80.24
5	F17	80.47	F17	80.71	F17	79.29
6	F23	80.94	F23	80.71	F23	81.18
7	F22	80.71	F22	78.59	F22	81.41
8	F11	79.06	F11	78.82	F11	80.00
9	F21	81.18	F21	78.35	F21	81.18
10	F10	82.12	F10	76.71	F10	81.41
11	F12	76.71	F12	77.65	F12	81.88
12	F15	79.29	F15	78.35	F15	82.35
13	F8	79.29	F8	78.82	F8	82.12
14	F24	77.88	F24	76.71	F24	80.71
15	F5	76.71	F5	78.35	F5	80.00
16	F4	79.06	F4	74.35	F4	79.06
17	F7	77.65	F7	73.65	F7	79.29
18	F18	77.41	F18	77.65	F18	78.82
19	F25	71.76	F25	67.76	F25	70.35
20	F26	67.53	F26	67.53	F26	68.94

F26, F25, F18, F7, F4, F5, F24, F8, F15, F12, F10, F21, F11, F22, F23, F17을 사용했을 때에 최고 예측률인 83.53%를 보여주었다. 따라서 세 알고리즘 중에서는 인공신경망 기반 예측모형에서 가장 높은 예측률을 기록한 것을 알 수 있었다.

본 연구에서는 배깅 기법에 앞서 언급된 세 알고리즘을 적용하여 예측모형들을 구축하는 실험을 진행하였다. 배깅은 분석 데이터로부터 무작위로 표본을 추출한 뒤 표본 별 예측모형을 구축하고, 다수의 예측모형에서 분류한 클래스를 사용하는 기법을 말한다. <표 3>은 배깅 작업을 통한 예측모형들의 예측률을 보여주고 있다. 배깅에서의 의사결정나무 기반 예측모형은 14가지 변수를 사용했을 때에 최고 예측률인 81.88%를 기록하였고, 베이즈넷 기반 예측모형은 15가지 변수를 사용했을 때에 최고 예측률인 75.53%를 보여주었으며, 인공신경망 기반 예측모형에서는 17가지 변수를 사용했을 때에 최고 예측률인 85.18%를 얻을 수 있었다. 이를 통해, 배깅 작업에서도 인공신경망 기반 예측모형이 가장 높은 예측률을 보여주었다.

또한 부스팅 기법에 앞서 언급된 세 알고리즘을 적용하여 예측모형들을 구축하는 실험을 진행하였다. 부스팅은 분석 데이터가 보유한 각 개체들에게 처음에는 동일한 가중치를 부여하여 표본을 만들지만 예측을 반복하면서 잘못 분류된 개체들에게 높은 가중치를 부여하여 새롭게 만들어지는 표본에 포함될 확률을 높이고, 표본 별 예측모형을 구축한 뒤 예측모형에서 분류한 클래스에 대하여 예측률의 가중평균을 반영하여 그 값이 높은 클래스를 사용하는 기법을 의미한다. <표 4>는 부스팅 작업을 통한 예측모형들의 예측률을 보여주고 있다. 부스팅에서의 의사결정나무 기반 예측모형은 11가지 변수를 사용하였을 때 최고 예측률인 82.12%를 기록하였고, 베이즈넷 기반 예측모형은 15가지 변수를 사용했을 때 최고 예측률인 80.71%를 보여주었으며, 인공신경망 기반 예측모형은 모든 변수를 사용했을 때에 최고 예측률인 84.24%를 얻을 수 있었다. 따라서 부스팅 작업에서도 인공신경망 기반 예측모형이 가장 높은 예측률을 기록하였다.

앞서 살펴본 실험 결과들을 종합해 보면 <그림 2>

와 같다. 의사결정나무 기반 예측모형과 베이즈넷 기반 예측모형은 부스팅 작업을 시행했을 때 각각의 최고 예측률인 82.12%와 80.71%를 얻을 수 있었다. 인공신경망 기반 예측모형의 경우 배깅 작업을 함께 실시할 때 최고 예측률인 85.18%를 기록하였다. 따라서 본 연구에서 제시한 배깅 기법에 인공신경망을 적용한 예측모형을 활용할 경우 한국 프로야구 승패예측에 대한 높은 예측 결과를 기대해 볼 수 있을 것으로 판단된다.

2. 전체 구단의 승패규칙

일반적으로 의사결정나무로 구축된 예측모형의 경우, 의사결정나무를 구성하는 분류 규칙들이 이해하기 쉬운 형태로 만들어지기 때문에 규칙을 통한 결과 해석이 용이한 장점이 있다(Witten & Frank, 2005). 그러나 인공신경망과 베이즈넷의 경우 블랙박스의 형태로 해석부분이 계산되기 때문에 규칙 간의 관계를 파악하기에는 무리가 있다. 따라서 앞서 진행된 실험에서 배깅 기법에 인공신경망을 적용한 예측모형이 가장 높은 예측률을 보여주었지만 본 연구에서는 의사결정나무 기반 예측모형의 결과를 활용하여 구단의 승률 향상을 위한 전략을 도출하고자 한다.

<표 2>에서 제시된 의사결정나무 기반 예측모형들 중, 예측률이 가장 높은 예측모형의 분류규칙을 활용하여 프로야구 전체 구단의 승패규칙을 도출하고자 한다. 본 연구에서는 6가지 변수인 팀출루율, 팀타율, 피안타, 안타, 타석, 타수를 사용하여 의사결정나무를 구축하였을 때 가장 높은 예측률을 얻을 수 있었다. <그림 3>은 의사결정나무 기반 예측모형에서 도출된 분류 규칙들을 보여주며, 의사결정나무를 간결하게 표현하기 위해 승패규칙들 중 예측률이 높은 상위 8개의 규칙들만 그림에 표시하였다. 이러한 상위 8개의 규칙들은 총 1,123경기의 승/패를 설명해 주며, 이는 전체 1,440경기 중 무승부 22경기를 제외한 1,418경기의 79.20%에 해당되는 수이다.

3. 플레이오프 진출 구단과 미진출 구단의 승리규칙

<표 5>는 의사결정나무 기반 승패예측모형에서 예측한 플레이오프 진출 구단의 표본별 예측 결과를 보여준다. 그 결과, 표본 1에서 F25, F26, F18, F7, F4, F10, F5, F12, F24, F8의 변수를 사용하여 예측모형을 구축할 경우 최고 예측률인 80.86%를 얻을 수 있었다.

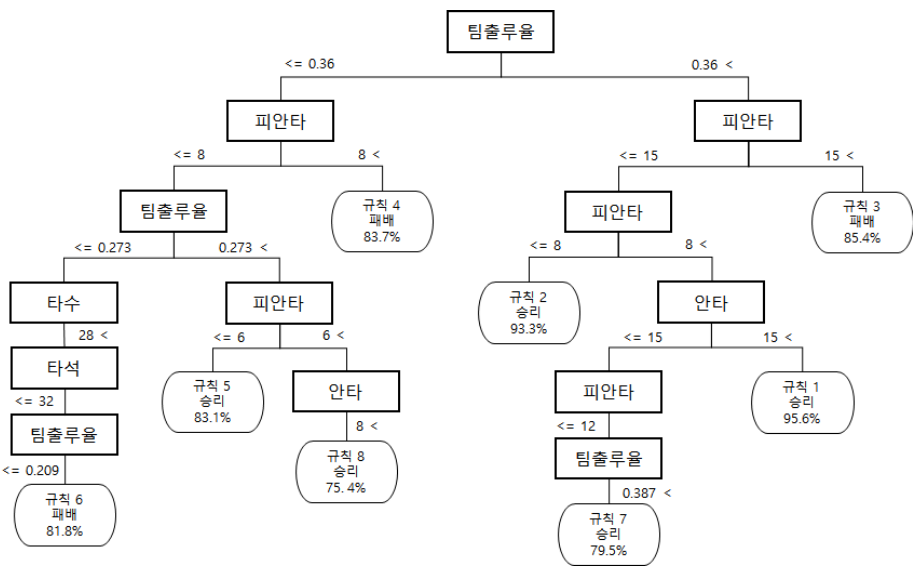


그림 3. 한국 프로야구 전체 구단의 승리규칙 도식화

표 5. 플레이오프 미진출 구단의 표본 분석 결과

변수 순서	표본 1		표본 2		표본 3		표본 4		표본 5	
	변수	예측률(%)	변수	예측률(%)	변수	예측률(%)	변수	예측률(%)	변수	예측률(%)
1	F14	77.74	F14	77.18	F14	83.87	F14	75.58	F14	80.18
2	F17	75.39	F17	78.18	F16	83.87	F17	75.58	F17	80.65
3	F3	75.71	F13	79.18	F17	83.87	F13	75.58	F3	80.65
4	F13	75.71	F11	77.18	F3	83.87	F3	75.58	F13	80.65
5	F22	76.02	F3	77.90	F13	83.87	F11	75.58	F22	78.80
6	F12	76.18	F10	78.37	F22	83.87	F10	75.58	F11	78.80
7	F10	76.02	F22	77.74	F11	83.87	F22	75.58	F12	78.80
8	F11	76.18	F23	78.53	F10	83.87	F21	75.58	F10	78.80
9	F23	76.33	F12	77.59	F15	83.87	F15	75.58	F23	78.80
10	F15	76.49	F15	77.59	F12	83.87	F8	77.88	F21	75.12
11	F21	75.55	F21	75.55	F8	83.87	F12	77.88	F15	75.12
12	F24	73.82	F8	75.86	F23	83.87	F23	77.88	F8	80.18
13	F5	73.51	F5	75.71	F21	82.95	F24	76.04	F16	77.88
14	F8	73.67	F24	72.73	F24	79.72	F16	76.04	F5	72.81
15	F16	72.88	F4	73.67	F5	79.72	F4	76.04	F18	74.65
16	F4	74.29	F16	73.98	F4	82.03	F18	72.81	F24	70.05
17	F18	68.50	F18	71.16	F18	70.05	F7	72.81	F4	66.82
18	F7	68.97	F7	71.63	F7	70.05	F5	70.51	F7	66.82
19	F25	69.28	F25	72.63	F26	72.81	F26	70.51	F25	66.82
20	F26	69.28	F26	73.63	F25	72.81	F25	70.51	F26	66.82

표 6. 플레이오프 진출 구단의 표본 분석 결과

변수 순서	표본 1		표본 2		표본 3		표본 4		표본 5	
	변수	예측률(%)	변수	예측률(%)	변수	예측률(%)	변수	예측률(%)	변수	예측률(%)
1	F14	79.90	F14	77.69	F14	71.29	F14	70.33	F14	76.56
2	F13	79.90	F17	77.69	F17	73.68	F17	70.33	F13	76.56
3	F11	79.90	F13	77.85	F13	73.68	F13	70.33	F11	76.56
4	F16	79.90	F11	78.34	F11	73.68	F11	69.86	F16	76.56
5	F23	79.90	F16	77.85	F16	73.68	F16	69.86	F17	76.56
6	F3	79.90	F3	78.01	F3	73.68	F3	69.86	F23	76.56
7	F22	79.90	F22	77.36	F23	73.68	F22	70.81	F3	76.56
8	F17	78.95	F23	76.55	F21	73.68	F23	66.51	F22	76.56
9	F21	78.95	F15	76.55	F12	73.68	F15	69.38	F21	76.56
10	F15	80.86	F21	76.55	F15	73.68	F12	69.38	F15	76.56
11	F8	80.86	F5	77.36	F22	73.68	F21	69.38	F12	76.56
12	F24	79.91	F8	78.66	F24	73.21	F5	71.77	F11	75.12
13	F12	78.95	F12	77.04	F8	73.21	F24	70.81	F24	71.77
14	F5	78.95	F4	75.90	F10	75.60	F4	70.81	F5	71.77
15	F10	76.56	F24	76.06	F4	75.12	F10	75.12	F10	75.12
16	F4	76.56	F10	75.90	F5	75.12	F8	74.64	F4	75.12
17	F7	75.60	F7	76.22	F18	75.60	F7	72.25	F18	75.12
18	F18	80.38	F26	77.04	F7	66.51	F18	72.73	F7	67.94
19	F26	71.29	F18	78.34	F26	66.51	F26	67.46	F26	66.03
20	F25	71.29	F25	69.71	F25	66.51	F25	66.03	F25	67.94

<표 6>은 의사결정나무 기반 승패예측모형에서 예측한 플레이오프 미진출 구단의 표본별 예측 결과를 나타낸다. 그 결과, 표본 3에서 F25, F26, F7, F18, F4, F5, F24, F21, F23의 변수를 사용하여 예측모형을 구축할 경우 최고 예측률인 83.87%를 기록하였다.

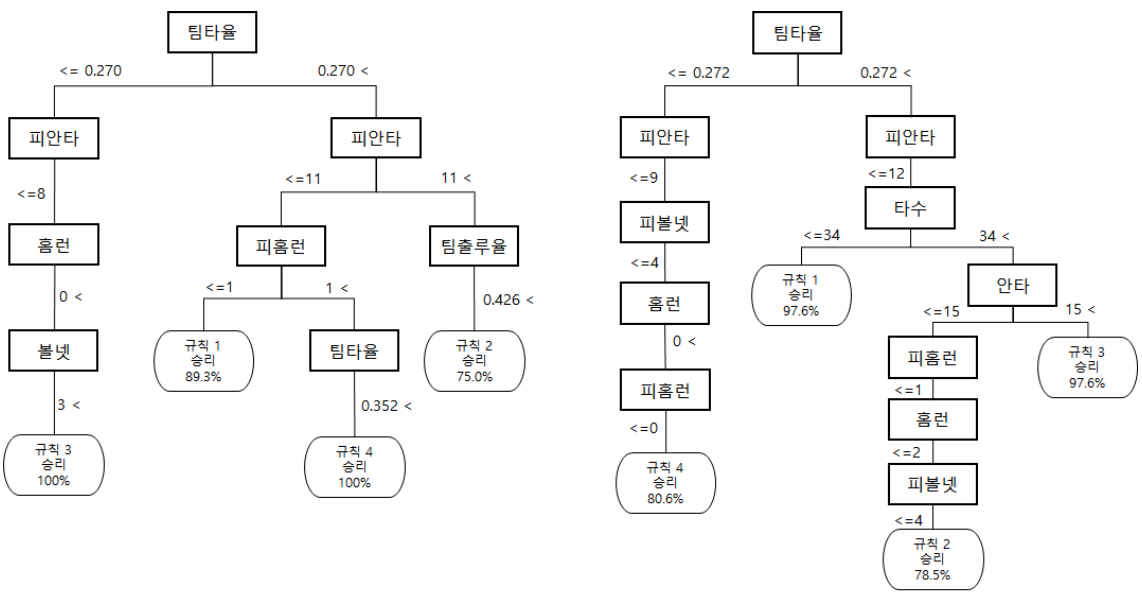
본 연구에서는 승패예측모형에서 도출된 승패규칙 중, 승리규칙을 중심으로 진출 구단과 미진출 구단의 승률 향상 전략을 제시하고자 한다. 따라서 플레이오프 진출 구단의 승리규칙은 진출 구단의 표본 1에서 도출하였고, 플레이오프 미진출 구단의 승리규칙은 미진출 구단의 표본 3에서 도출하였다.

<그림 4>는 플레이오프 진출 구단과 미진출 구단의 예측모형에서 도출된 분류 규칙들 중 승리를 가장 많이 예측한 상위 4개의 규칙들을 보여준다. 여기에서 진출 구단에 관한 4개의 승리규칙들은 전체 307회 승리 중 78.82%에 해당되는 242회의 승리를 설명하며, 미진출 구단에 대한 4개의 승리규칙들은 전체 319회 승리 중 70.53%에 해당되는 225회의 승리를 설명한다.

Ⅳ. 논 의

본 장에서는 <그림 3>에서 표현된 규칙을 중심으로 한국 프로야구 구단의 승률 향상을 위한 전략을 제시하고 그 의미를 논의하고자 한다. <표 7>은 <그림 3>에서 보여준 한국 프로야구 전체 구단의 승패규칙에 관한 세부 내용을 설명하고 있다.

전체 구단의 승리규칙을 살펴보면 다음과 같다. 첫 번째, 규칙 1을 보면 팀출루율이 0.36 이상이고 피안타가 8~15개 사이이고 안타수가 16개 이상, 즉 상대 구단보다 안타수를 더 많이 쳐냈을 시에 승리할 확률이 95.6%로 가장 높은 승리 확률을 보여주었다. 두 번째, 규칙 2를 살펴보면 팀출루율이 0.36 이상이고 피안타가 8개 이하일 때에 승리 확률이 93.3%라는 것을 확인할 수 있다. 세 번째, 규칙 5를 통해서는 팀출루율이 0.273~0.36 사이일 때에는 우선적으로 피안타를 6개 이하로 낮춰야 승리할 확률이 높은 것을 알 수 있다. 네 번째, 규칙 7은 피안타가 9~12개 사이이고 안타수가 15개 이하, 즉 한 경기에서 상대 구단의 안타수보다 같거나 적을 시에는 구단 출루율이 0.387



a) 진출 구단　　b) 미진출 구단

그림 4. 플레이오프 진출 구단과 미진출 구단의 승리규칙 도식화

표 7. 전체 구단 승패규칙

규칙	내용	클래스	경기수	확률
1	팀출루율 > 0.36 & 피안타 <= 15 & 피안타 > 8 & 안타 > 15	승리	68	95.6%
2	팀출루율 > 0.36 & 피안타 <= 15 & 피안타 <= 8	승리	226	93.3%
3	팀출루율 > 0.36 & 피안타 > 15	패배	48	85.4%
4	팀출루율 <= 0.36 & 피안타 > 8	패배	512	83.7%
5	팀출루율 <= 0.36 & 피안타 <= 8 & 팀출루율 > 0.273 & 피안타 <= 6	승리	89	83.1%
6	팀출루율 <= 0.36 & 피안타 <= 8 & 팀출루율 <= 0.273 & 타수 > 28 & 타석 <= 32 & 팀출루율 <= 0.209	패배	11	81.8%
7	팀출루율 > 0.36 & 피안타 <= 15 & 피안타 > 8 & 안타 <= 15 & 피안타 <= 12 & 팀출루율 > 0.387	승리	112	79.5%
8	팀출루율 <=0.36 & 피안타 <= 8 & 팀출루율 > 0.273 & 피안타 > 6 & 안타 > 8	승리	57	75.4%

보다 높아야 승리할 확률이 높아지는 것을 알 수 있다. 다섯 번째, 규칙 8은 팀출루율이 0.273~0.36 사이일 때, 피안타를 6개 이상으로 허용하였을 때에는 안타를 9개 이상 쳐야 함을 알 수 있으며, 이를 통해 상대보다 더 많은 안타를 쳐야 승리할 확률이 높아짐을 알 수 있다.

전체 구단의 패배규칙은 다음과 같다. 첫 번째, 규칙 3은 팀출루율이 0.36 이상이지만 피안타가 15개를 넘을 경우 패배할 확률이 85.4%로 가장 높은 패배 확률을 보여주었다. 두 번째, 규칙 4는 팀 출루 0.36 이하이고 피안타가 8개보다 많으면 패배할 확률이 83.7%라는 것을 알 수 있다. 세 번째, 규칙 6은 피안타를 8개 이하로 낮추는 수비적인 야구를 하더라도 팀출루율이 0.209 이하의 저조한 출루율이라면 승리하기 어렵다는 것을 보여주고 있다.

앞서 살펴본 승패규칙의 주요 특성들을 통해 한국 프로야구 전체 구단의 승률 향상을 위한 전략을 다음과 같이 제안하고자 한다. 첫 번째, 우선적으로 팀의 승리를 높이기 위해 팀출루율을 0.36 이상으로 끌어올리는 전략이 필요하다. 이는 팀출루율이 0.36 이상인 것을 전제로 하는 4개의 규칙 중에서 3개의 규칙이 승리규칙인 점, 그리고 3개의 패배규칙 중 2개 규칙의 첫 전제조건이 팀출루율 0.36 이하인 점을 통해서 알 수 있다. 팀출루율을 높이기 위해 구단에서는 선수 영입 조건으로 타율이 높은 타자와 선구안이 높은 타자를 우선 영입 대상으로 고려해야 한다. 또한, 빠른 발로 내야 안타나 상대 구단의 수비 실책을 유도하는 선수, 그리고 수비수들이 반응하지 못할 정도로 빠른 타구 속도를 가진 선수도 고려해 볼 수 있다.

두 번째, 상대 팀보다 안타의 개수를 더 많이 쌓는 전략을 마련해야 한다. 이는 강타자를 영입함으로써 구단의 안타 수를 높이는 방법으로 해결할 수 있다. 그러나 구속과 변화구가 좋아 삼진 능력이 뛰어난 투수, 제구력이 좋아 볼넷을 허용하지 않는 투수, 뛰어난 수비력으로 상대 구단의 안타성 타구를 잡아내는 선수를 영입하는 것으로도 해결할 수 있다.

세 번째, 구단 측면에서는 구단의 부족한 부분을 보완하는 방향 또는 구단의 강점을 극대화하는 방향으로 구단 운영 전략을 수립해야 한다. 예를 들어, 뛰어난 투수진을 보유하고 있지만 수비진이 약하여 상대 구단의 출루율을 높이는 구단의 경우 시즌 후 수비력 보강을 위한 새로운 선수 영입이나 수비 훈련을 강화하는 것이 필요하다. 투수에 비해 타자가 빈약하여 점수를 내지 못해 패하는 구단의 경우는 시즌 후에 타율 및 출루율을 어느 정도 보장하는 타자를 영입하는 것을 고려해 볼 수 있다. 또한 뛰어난 타선에 비해 중간계투진이 약해 경기 후반이 불안한 구단은 불펜보강을 위해 뛰어난 선발투수를 중간계투로 돌리는 방법 등의 구단 운영을 시도해 볼 수 있다.

다음으로 본 연구에서는 <그림 4>에서 제시된 플레이오프 진출 구단과 미진출 구단의 승리규칙의 차이점을 토대로 두 집단에 맞춤화된 승률 향상 전략을 제시하고 그 의미에 대해 살펴보고자 한다. <표 8>은 <그림 4>에서 보여준 플레이오프 진출 구단과 미진출

표 8. 플레이오프 진출 구단과 미진출 구단의 승리규칙

구분	규칙	내용	클래스	경기수	확률
진출	1	팀타율 > 0.270 & 피안타 <= 11 & 피홈런 <= 1	승리	178	89.3%
	2	팀타율 > 0.270 & 피안타 > 11 & 팀출루율 > 0.426	승리	32	75.0%
	3	팀타율 <= 0.270 & 피안타 <= 8 & 홈런 > 0 & 볼넷 > 3	승리	18	100%
	4	팀타율 > 0.270 & 피안타 <= 11 & 피홈런 > 1 & 팀타율 > 0.352	승리	14	100%
미진출	1	팀타율 > 0.272 & 피안타 <= 12 & 타수 <= 34	승리	85	97.6%
	2	팀타율 > 0.272 & 피안타 <= 12 & 타수 > 34 & 안타 <= 15 & 피홈런 <= 1 & 홈런 <= 2 & 피볼넷 <= 4	승리	65	78.5%
	3	팀타율 > 0.272 & 피안타 <= 12 & 타수> 34 & 안타 > 15	승리	44	97.6%
	4	팀타율 <= 0.272 & 피안타 <= 9 & 피볼넷 <= 4 & 홈런 > 0 & 피홈런 <= 0	승리	31	80.6%

구단의 승리규칙에 관한 세부 내용을 설명하고 있다.

플레이오프 진출 구단의 승리규칙은 다음과 같다. 첫 번째, 규칙 1을 통해 진출을 위해서는 타율이 0.270을 넘기는 것을 목표로 해야 하며, 피안타의 개수를 11개 이하로 낮추고 홈런을 1개 이하로 허용해야함을 알 수 있다. 두 번째, 규칙 2를 살펴보면 팀타율이 0.270을 넘긴 상태에서 피안타를 11개 이상으로 허용할 시 구단 출루율은 0.426 보다 높아야 한다. 세 번째, 규칙 4처럼 만약에 홈런을 2개 이상 허용한다면 팀타율은 0.352 보다 높아야 한다. 네 번째, 규칙 3과 같이 타율이 0.270이 안될 때에는 피안타의 개수를 8개 이하로 낮춤과 동시에 홈런은 1개 이상 그리고 볼넷은 4개 이상을 얻을 수 있는 전략이 필요하다.

다음으로 플레이오프 미진출 구단의 승리규칙은 다음과 같다. 첫 번째, 규칙 1을 통해 미진출 구단에서도 승리를 위해서 팀타율 0.272를 넘겨야함을 알 수 있으며, 동시에 피안타의 개수를 12개 이하로만 허용하며 타수를 34타수 이하로 들어가게끔 해야 한다. 두 번째, 규칙 3은 팀타율이 0.272를 넘기고 피안타의 개수가 12개 이하지만 타수가 34타석 이상으로 늘어난다면 안타를 15개 이상으로 쳐야 함을 뜻한다. 세 번째, 만약 규칙 3과 동일한 조건에서 안타를 15개 이상 치지 못할 경우에는 규칙 2처럼 피홈런, 홈런, 볼넷허용과 같은 여러 가지 상황들이 고려되어야 함을 알 수 있다. 네 번째, 규칙 4와 같이 팀타율이 0.272를 넘지 못한 경우에는 피안타를 9개 이하로 줄이고 볼넷 허용을 4개 이하, 홈런을 1개 이상 치고 피홈런을 0개로 억제하는 전략이 필요함을 알 수 있다.

앞서 살펴본 플레이오프 진출 구단과 미진출 구단의 승리규칙을 비교해보면 다음과 같다. 첫째, 승리구단의 승리규칙이 미진출 구단의 승리규칙 보다 간단하다. 그 이유는 진출 구단의 승리규칙의 최다 승리 회수가 178승으로 전체 승리 회수인 307승의 58%를 차지하는 것에 비해, 미진출 구단은 승리규칙의 최다 승리 회수는 85승으로 전체 승리회수인 319승의 27%에 불과하기 때문이다.

둘째, 팀타율의 최소 요구치가 진출 구단과 미진출 구단이 각각 0.270, 0.272로 비슷한 수치를 보인다. 이를 통해, 구단 타율 요인만으로는 플레이오프의 진출 여부를 판가름하기가 어려움을 알 수 있다.

셋째, 진출 구단의 규칙 1과 미진출 구단의 규칙 1, 3을 서로 비교해보면 유사한 팀타율과 피안타를 기록하더라도 진출 팀은 수비력 변수인 피홈런을 필요로 하고, 미진출 팀은 공격력 변수인 타수와 안타를 고려하는 것을 볼 수 있다. 진출 구단은 승리 요건에 수비적인 면을 우선순위로 둔다면 미진출 팀은 공격적인 면을 우선순위로 두어야 함을 알 수 있다.

넷째, 진출 구단의 규칙 1, 규칙 4와 미진출 구단의 규칙 1, 2, 3을 묶어서 비교하였을 때 진출 구단의 규칙들은 공통 변수인 팀타율이나 피안타를 제외하면 피홈런이라는 수비력 변수만 고려하면 되는 것을 알 수 있다. 반대로 미진출 구단의 규칙들을 보면 공통으로 포함되는 변수들을 제외하고도 타수, 홈런, 안타 등 진출 구단에 비해 공격과 수비 양면에서 고려해야

할 변수의 수가 많아져 승리의 조건이 더 까다로워지는 것을 확인할 수 있다.

다섯째, 미진출 구단의 규칙 2와 규칙 4를 보면 진출 구단에서는 고려하지 않아도 되는 변수인 피볼넷이 있다. 이를 통해, 미진출 구단의 투수의 제구력이 좋지 못함을 알 수 있고 상대 구단에게 불필요한 출루를 제공하여 점수를 내줄 확률이 높아짐을 알 수 있다.

여섯째, 두 집단 모두 팀타율이 0.270을 넘지 못하는 진출구단의 규칙 3과 미진출 구단의 규칙 4를 보면, 승리를 위해 진출 구단은 홈런, 볼넷을 요구하며 미진출 구단은 피볼넷, 피홈런, 홈런을 요구하고 있다. 앞서 팀타율이 0.270을 넘을 경우 진출 구단은 수비력 변수를 그리고 미진출 구단은 공격력 변수를 고려하였다면, 팀타율이 0.270을 넘지 않을 경우에는 진출 구단은 공격력 변수를 그리고 미진출 구단은 수비력 변수를 고려함을 알 수 있다.

지금까지 살펴본 진출 구단과 미진출 구단의 승리 규칙에 대한 비교분석 결과와 현재 미진출 구단에서 자신의 전력강화를 위해 실제 선수를 영입한 방향과 비교해 보고자 한다. 타자의 승리기여도(Wins Above Replacement) 순위를 보면 1위에서 10위까지는 진출 구단의 클린업트리오(팀의 3, 4, 5번 타자)가 차지하고 있고 그 중 7위만이 미진출 구단 선수였는데 이 선수 혼자 이번 시즌 홈런순위에 2위를 차지하면서 미진출 구단에서 독보적인 타격력을 보여주었다. 이 구단은 홈런 2위의 타자를 통하여 미진출 구단의 약점인 득점권에서의 타격을 보완하려 하였음을 알 수 있다. 즉, 상대적으로 비슷한 수의 안타를 치더라도 득점권에 강한 타자가 미진출 구단에 필요함을 알 수 있으며 실제로 미진출 구단 중 한 구단은 이번 시즌 중에 상대구단의 4번 타자를 영입해서 2017년 시즌 후반기에 우수한 공격력을 보여주었다. 또한 시즌이 끝난 후 미진출 구단 중 세 구단은 메이저리그에서 돌아온 거포타자들을 자유계약선수로 데려와 전력보강을 하였고 나머지 한 구단은 진출 구단의 클린업트리오 1명을 자유계약선수로 데리고 왔다.

수비 방면에서는 미진출 구단은 외국인 투수를 통하여 팀의 투수문제를 해결한 것을 확인할 수 있다. 이는 메이저리그에서 뛰던 선수를 영입한 모 구단이나 예전 국내 경기에서 수준급 경기력을 보여주었던 외국인 투수를 다시 영입한 것을 통해 알 수 있다. 또한 넓은 수비범위를 자랑하는 선수를 영입하여 구단의 수비 능력을 올려 구단의 피안타를 억제하고자 한 것을 알 수 있다. 이를 통해 분석결과가 실제 구단의 선수 영입 전략과 관련이 있다는 것이 확인되었다. 이는 곧 구단의 성적 향상과 관중 동원을 통한 구단의 수익 증대에도 영향을 줄 것을 기대해 볼 수 있다.

V. 결론 및 제언

본 연구에서는 한국 프로야구 구단의 경기력 향상을 위한 방안의 일환으로, 프로야구 승패규칙을 도출한 후 이를 바탕으로 승률 향상 전략을 수립하여 보다 효율적인 경기 전략과 선수 구성 방안을 제안하고자 하였다. 이를 위해 데이터 마이닝 기법을 활용하여 프로야구의 경기 결과를 예측하는 승패예측모형들을 구축하였다. 그 결과 배깅 기법에 인공신경망을 적용한 예측모형에서 최고 예측률인 85.18%를 얻을 수 있었다. 이때 승패에 영향을 주는 요인으로는 땅볼아웃, 잔루, 피사구, 탈삼진, 사구, 피볼넷, 볼넷, 삼진아웃, 투구 수, 홈런, 피홈런, 타수, 타석, 안타, 피안타, 팀타율, 팀출루율임을 알 수 있었다.

그 후, 구축된 의사결정나무 기반 예측모형을 활용하여 한국 프로야구 전체 구단의 승패규칙을 도출하였다. 여기에서 승패규칙은 승패예측에 영향을 미치는 주요 요인들인 팀출루율, 팀타율, 피안타, 안타, 타석, 타수로 표현되며, 도출된 규칙을 바탕으로 구단의 승률 향상에 도움을 주는 전략을 제안하였다. 또한 플레이오프 진출 구단과 미진출 구단의 승리규칙을 각각 도출하여 두 집단에 맞춤화된 승률 향상 전략을 제시하였고, 제시된 전략이 실제 구단의 선수 영입 전략과도 관련 있음을 확인하였다. 따라서 향후 본 연구에서 제안한 분석 방법과 전략 도출 방법을 프로야구 구단들이 참고할 경우 프로야구 경기에 관한 보다 실용적이고 다양한 관점의 전략 수립이 가능해 질 것이며

그에 따른 구단의 성적 향상과 이윤 창출에도 영향을 줄 것으로 기대된다.

데이터 마이닝 분석의 특성상 분석 결과는 수집된 데이터에 영향을 받게 된다. 본 연구의 경우 2017년 한 해의 경기만을 분석하였기 때문에 연구 결과의 내용이 2017년 경기에만 한정된다. 또한 분석에서는 경기 내적인 요소들만 고려하였고 감독의 성향, 팀 분위기, 연승 및 연패 등과 같은 외적인 요소들을 반영하지 못한 점이 본 연구의 한계점으로 인식된다.

참고문헌

김창권, 진서훈 (2014). 프로야구 기록을 통한 승리 요인에 관한 연구. **한국자료분석학회**, **16(1)**, 211-220.

신상근, 이경준, 조영석, 박찬근 (2010). 각 이닝 선두 세 타자의 출루 능력이 득점에 미치는 영향-롯데자이언츠를 중심으로. **한국자료분석학회**, **12(1)**, 563-572.

오윤학, 김한, 윤재섭, 이종석 (2014). 데이터마이닝을 활용한 한국프로야구 승패예측모형 수립에 관한 연구. **대한산업공학회지**, **40(1)**, 8-17.

이주호, 염준근, 송근혁, 박홍준 (2010). 프로야구 관중수의 결정요인에 대한 연구. **한국자료분석학회**, **12(6)**, 3507-3517

장경로, 김민철 (2014). 국내 프로야구구단의 손익계산서를 활용한 수익-지출구조 연구. **한국체육학회**, **53(3)**, 357-369

장진희, 문춘걸 (2014). 한국 프로야구의 구단 승률에 대한 분석. **한국스포츠산업경영학회지**, **19(3)**, 17-31.

조영석, 조용주 (2005). 한국 프로야구에서 OPS와 득점에 관한 연구. **한국자료분석학회**, **7(1)**, 221-231.

조영석, 조용주, 신상근 (2007). 한국프로야구에서 승패 추정에 관한 연구. **한국자료분석학회**, **9(1)**, 501-510.

채진석, 조은형, 엄한주 (2010). 프로야구 포스트시즌 진출 예측을 위한 통계적 모형 비교. **한국체육측정평가학회지**, **12(1)**, 33-48.

채진석, 엄한주 (2010). 프로야구구단의 성적변화 추이와 상대적 전력 비교평가. **체육과학연구**, **21(1)**, 956-973.

한국프로야구위원회 (2017). 2017년 한국프로야구연감.

Beneventano, P., Berger, P. D., & Weinberg, B. D. (2012). Predicting Run Production and Run Prevention in Baseball: The Impact of Sabermetrics. *International Journal of Business, Humanities and Technology*, 2(4), 67-75.

Chawla, N. V. (2005). *Data Mining for Imbalanced Datasets: An Overview*. In: Maimon, O., Rokach, L. (Eds). Data Mining and Knowledge Discovery Handbook. Boston, MA: Springer.

Dash, M., & Liu, H. (1997). Feature selection for classification. *Intelligent Data Analysis*, 1(3), 131-156.

Lewis, M. M. (2003). *Moneyball: the Art of Winning an Unfair Game*. New York: W.W. Norton.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann Publishers.

ABSTRACT

Development of Win-Loss Prediction Models and Strategies for Improving Winning Rate of the Korean Professional Baseball Teams Using Data Mining Techniques

Kim, Won-jong · Choi, Yeon-sik · Yoo, Dong-hee*(Gyeongsang National University)

This study conducted an experiment to develop win-loss prediction models for the Korean professional baseball league using data mining techniques. To this end, we used data on 1,418 games from all games played by the ten Korean professional baseball teams in 2017, except draw games. We developed win-loss prediction models using not only a decision tree, Bayse net, and artificial neural network algorithms, but also ensemble methods, such as bagging and boosting. As a result, first, we found that the artificial neural network-based prediction model using the bagging method reported the best accuracy (85.18%). Second, we derived eight win-loss rules for entire teams from the decision tree-based prediction model. These rules consist of six influential factors: team on base average, team batting average, hit by opponent, hit, plate appearances, and at bat. Using the derived rules, we proposed helpful strategies for improving the winning rate. Third, we derived four winning rules for both playoff teams and non-playoff teams; using the rules we proposed customized strategies for improving the winning rate of the two different groups. Finally, we confirmed the feasibility of the proposed strategies by comparing non-playoff teams' actual player recruitment strategies.

※ Key words : Korean professional baseball, data mining, win-loss prediction model, decision tree, Bayes net, artificial neural network, bagging, boosting.

논문투고일 : 2018. 04. 30
심 사 일 : 2018. 06. 02
심사완료일 : 2018. 06. 22