

SPRi Issue Report

2015. 12. 28. (2015-013호)

기계학습 경진대회 활성화 방안 - 빅콘테스트 2015 결과와 시사점

추형석
(hschu@spri.kr)

- 본 보고서는 「미래창조과학부 정보통신진흥기금」을 지원받아 제작한 것으로 미래창조과학부의 공식의견과 다를 수 있습니다.
- 본 보고서의 내용은 연구진의 개인 견해이며, 본 보고서와 관련한 의문사항 또는 수정·보완할 필요가 있는 경우에는 아래 연락처로 연락해 주시기 바랍니다.
 - 소프트웨어정책연구소 연구조정실 추형석 선임연구원(hschu@spri.kr)

《 Executive Summary 》

빅콘테스트 2015는 “야구경기 예측”이라는 현실적인 문제를 사전에 제공된 데이터와 기계학습 방법론을 사용하여 해결하는 경진대회이다. 이번 콘테스트에서는 약 80여 가지의 야구경기 관련 데이터를 제공하고, 딥러닝 방법론에 가산점을 부여하여 기계학습의 활용을 장려했다. 하지만 지금까지 기계학습 관련 콘테스트의 결과를 심도 있게 분석하여 대학생과 일반참가자들이 기계학습을 어떻게 인식하고 활용하는지에 대한 자료는 찾아보기 어렵다. 이에 본 보고서에서는 빅콘테스트 2015에 참여한 103개 팀의 결과자료를 분석하여 참가자들이 문제를 해결하기 위해 어떠한 방법론을 사용하고 어떤 결과를 예측했는지를 제시하고, 시사점을 도출해 보고자 한다.

분석 결과, 참가자들의 대부분이 네 가지 단계(데이터 수집 → 데이터 선별 → 예측 모델링 → 결과)를 거쳐서 문제를 해결했으며, 특히 데이터의 수집과 선별에 중점을 두어 상관관계가 높은 데이터의 추출에 노력을 기울였다. 또한 참가자들의 예측 방법론 사용 분포와 기계학습 적용 비율을 조사한 결과, 전체 103개 팀에서 가장 많이 사용한 예측 방법론은 회귀분석으로 약 43%를 차지했고, 기계학습 적용 비율은 약 72%이었다. 최종 수상한 14팀 중 12팀이 기계학습 방법론을 사용하여 기계학습의 보편적 활용가능성을 입증했으며, 그 외에도 딥러닝의 활용, 상관성이 높은 데이터 선별, 예측모델의 최적화 기법, 여러 가지 예측모델의 동시적용 등이 예측가능성을 높이기 위해 사용되었다.

하지만 콘테스트의 변별력을 더 확보하기 위해서는 문제의 난이도를 높이고 문제에 대한 선행연구 분석으로 구체적인 가이드라인을 제시할 필요가 있다. 또한 국내외 기계학습 경진대회의 장점을 벤치마킹하여 양질의 콘테스트를 위한 지속적인 프로그램 개선을 해야 할 것이다. 이번 콘테스트는 취업연계 프로그램을 통한 기계학습의 저변확대에는 긍정적인 효과가 있었으나, 향후에는 더욱 다양한 유인동기를 제공하여 콘테스트의 질적인 향상을 도모해야 할 것이다.

《 목 차 》

- 1. 개 요 1
 - (1) 배 경 1
 - (2) 대회 개요 및 분석 목적 2

- 2. 분석 방법 7
 - (1) 예측 방법론 관점 7
 - (2) 문제해결 절차 10

- 3. 분석 결과 12
 - (1) 챌린지리그 - 프로야구 승률 예측 12
 - (2) 퓨처스리그 - 프로야구 누적 관객수 예측 16
 - (3) 종합 결과 분석 20

- 4. 시사점 21

- [부 록] 24
- [참고문헌] 30

1. 개 요

(1) 배 경

- 빅데이터와 기계학습을 사용하여 SW를 도구로 활용한 현실적 문제 해결이 가능해짐
 - 과거에는 현실적인 문제 해결의 주체가 분야별 전문가에 한정되어 있었지만, 현재는 방대한 양의 데이터, 즉 **빅데이터**에의 접근성이 용이해지고 이를 분석하여 패턴을 찾아내는 **기계학습**의 활용이 확대됨
 - 빅콘테스트 2015는 “야구경기 예측”이라는 현실적인 문제를 사전에 제공된 데이터와 기계학습 방법론으로 해결하는 경진대회
 - 글로벌 SW 기업들은 빅데이터와 기계학습을 기반으로 한 지능형 SW 개발에 박차를 가하고 있음
 - IBM에서 개발한 지능형 컴퓨터 왓슨(Watson)은 2011년 “퀴즈 쇼 제퍼디(Jeopardy!)”에 참여하여 우승한 뒤 헬스케어, 날씨예보, 클라우드 서비스 등 그 활용분야를 넓혀가고 있음
 - 구글의 딥러닝 프로젝트인 구글 브레인은 약 천만 개의 유튜브 영상 중 고양이를 인식하는데 성공함 (약 70%의 정확도)¹⁾
 - 페이스북의 딥러닝 기반 얼굴인식 기술은 약 4백만 개 이상의 사진을 토대로 사람의 얼굴을 자동으로 인식함 (약 97%의 정확도)
 - 빅콘테스트 2015에서는 “야구경기 예측”에 관련한 양질의 데이터를 제공하고, 기계학습의 활용을 장려함
 - 약 80여 종의 야구경기 관련 데이터 제공
 - 이번 콘테스트에서는 기계학습 방법론 중 특히 딥러닝을 사용한 경우 가산점을 부여하는 평가지표를 마련함²⁾

1) 구글 브레인은 물체(object)가 고양이인지의 여부 정보 없이 학습함

2) 딥러닝은 최근 인공지능의 핵심알고리즘으로 각광받고 있는 방법론으로, 빅데이터에서 패턴을 효율적으로 파악할 수 있는 방법 중 하나임

https://en.wikipedia.org/wiki/Deep_learning

- 더욱이 콘테스트 후원 기업들과의 취업연계 프로그램으로 국내 데이터 과학자 발굴에 긍정적인 역할을 할 것이라 예상

- 빅콘테스트 2015는 올해로 3회 차를 맞는 기계학습 경진대회이나 아직까지 결과에 대한 분석 자료를 찾아보기 어려움
 - 이에 본 보고서에서는 참가자들의 발표자료를 모두 분석하여 참가자들이 어떠한 방법론을 사용하고 어떤 결과를 예측했는지를 제시하고자 함
 - 분석 결과를 토대로 더 양질의 콘테스트를 위한 시사점을 도출하고, 국내외 기계학습 경진대회 사례를 분석하여 향후 방향을 제시하고자 함

(2) 대회 개요 및 분석 목적

□ 대회 개요

- (제목) 빅콘테스트 2015 기계학습 기반 야구경기 예측
 - 참가신청 기간 : 2015. 7. 22 ~ 8. 24
 - * 분석결과 제출 기한은 2015년 9월 6일까지
 - 야구경기 예측 문제는 난이도에 따라 대학생과 일반인이 참가할 수 있는 챌린지리그와 고등학생과 대학생을 대상으로 한 퓨처스리그 두 가지로 나뉨
 - * 국내 프로야구는 1군 선수들이 경쟁을 하는 “챌린지리그”와 2군 선수들로 이루어진 “퓨처스리그” 두 개의 리그로 나뉜 것을 바탕으로 기획함
 - 챌린지리그 : 2015년 프로야구 구단별 승률 예측
 - * 2015년 9월 30일 기준 프로야구 팀 10개 구단별 승률 예측
 - * 사전에 제공된 데이터는 개인 선수별 성적, 연도별 팀 통산 성적 및 전적자료이고, 자세한 설명은 후술함

| 2015. 9. 6 순위/승률 | | | ➔ | 2015. 9. 30 순위/승률 예측 | | |
|------------------|-----|-------|---|----------------------|-----|-----|
| 순 위 | 팀 명 | 승 률 | | 순 위 | 팀 명 | 승 률 |
| 1 | 삼성 | 0.610 | | ? | ? | |
| 2 | NC | 0.580 | | ? | ? | |
| 3 | 두산 | 0.567 | | ... | .. | |
| 4 | 넥센 | 0.553 | | | | |
| 5 | 롯데 | 0.480 | | | | |
| 5 | 한화 | 0.480 | | | | |
| 7 | KIA | 0.475 | | | | |
| 8 | SK | 0.462 | | | | |
| 9 | LG | 0.431 | | | | |
| 10 | KT | 0.366 | | ? | ? | |

데이터+ 기계학습 (약 20경기 예측)

- 퓨처스리그 : 2015년 프로야구 구단별 누적 관객 입장수 예측

* 2015년 9월 30일 기준 프로야구 10개 구단별 누적 관객 입장 수 예측

* 제공된 데이터는 챌린지리그 데이터를 포함하여 2015년 일자별 누적 관객수 데이터가 추가됨

| 2015. 9. 6 누적관객수 | | ➔ | 2015. 9. 30 누적관객수 예측 | |
|------------------|-----------|---|----------------------|-------|
| 팀 명 | 관 객 수 | | 팀 명 | 관 객 수 |
| 삼성 | 514,971 | | ? | |
| 넥센 | 487,562 | | ? | |
| NC | 514,651 | | ... | |
| LG | 1,011,294 | | | |
| SK | 782,133 | | | |
| 두산 | 1,094,381 | | | |
| 롯데 | 796,905 | | | |
| KIA | 679,118 | | | |
| 한화 | 657,385 | | | |
| KT | 634,465 | | ? | |

데이터+ 기계학습 (약 20경기 예측)

- (목적) 기계학습을 사용하여 빅데이터 문제 해결을 직접 체험하고 분석하는 기회를 제공함으로써, 현안 문제를 해결할 수 있는 빅데이터 전문가를 발굴하고 취업 연계 지원
- (주최) 소프트웨어정책연구소, 한국정보화진흥원, 한국정보통신진흥협회, KT
 - * (주관) 미래창조과학부, 한국빅데이터연합회
- (평가) 1차 서류심사와 2차 발표심사를 통해 총 14개 참가자에게 수상
 - 1차 서류심사 기준은 예측의 정확도에 중점을 둠
 - * 예측의 정확도는 2015년 9월 30일 기준 실제 결과 값과 예측 값의 제곱평균 제곱근오차(Root Mean Square Error, 이하 RMSE)³⁾로 계산하고, 이것은 예측 항목별 오차의 제곱의 합을 평균한 값으로 예측 값이 얼마나 정확한지 나타낸 지표임
 - * 총 참가자 중 18개 팀을 선별 (각 리그별 9개 팀)
 - 2차 발표심사는 1차 심사 통과자에 한해서 진행
 - * 딥러닝을 사용한 참가자에게는 가산점을 부여하여 변별력 확보

□ 대회 분석 목적 및 방법

- 콘테스트 참여자들의 결과 분석을 통해 기계학습 방법론에 대한 인식과 수준 파악
 - 참가자의 발표자료와 보고서를 토대로 예측에 사용한 방법론, 예측의 정확도, 추가 데이터 수집 내용, 개발환경 등을 분석
 - 기계학습 현황 파악 지표로 예측 방법론*별 사용 분포와 기계학습 적용 비율 도출
 - * 참가자가 주로 많이 사용한 방법론을 기준으로 4가지 분류항목을 도출
 - 기계학습 사용여부와 수상의 상관성 분석

3) $RMSE = \sqrt{\sum_{j=1}^n (y_j - x_j)^2 / n}$, 여기서 x_j 는 예측 값, y_j 는 실제 값 (n 개). RMSE는 예측모델의 정확도를 판단하는 가장 기본적인 척도로 예측의 정밀도를 표현하는데 적합함.

https://en.wikipedia.org/wiki/Root-mean-square_deviation

- 콘테스트 대상 수상자와 우수 참가자 사례 분석
- 향후 콘테스트의 질적인 성장을 위한 보완사항과 제언을 도출함
- 분석 결과를 바탕으로 차회 콘테스트 발전방향 제시
- 국내외 기계학습 관련 콘테스트와 비교 (캐글, 코드스프린트)

□ 대회 세부 사항

- 참가 인원과 자격 <표 1>

<표 1> 빅콘테스트 2015 참가 인원 및 자격

| 구 분 | 팀 수 | 인원수 (명) | 참가자격 |
|-------|-----|---------|-----------------------|
| 챌린지리그 | 58 | 230 | 대학생 이상 누구나 |
| 퓨처스리그 | 45 | 176 | 고등학생, 대학(원)생 (휴학생 포함) |
| 총 계 | 103 | 406 | - |

- 대회 공식 제공 데이터 (스포츠투아이 제공)

- 개인 선수별 성적 (익명 : 1982 ~ 2014, 실명 : 2010 ~ 2014)

* 연도별 투수 29개, 타자 26개 지표 [그림 1]

[그림 1] 개인 선수별 성적 데이터 예시 (타자, 실명)

| | A | B | C | D | E | F | G | H | I | J | K | L | M |
|----|---|----------|-----|----|-----|-----|-----|----|-----|-----|-----|----|-----|
| 1 | | | | | | | | | | | | | |
| 2 | | 구분 | 연도 | 소속 | 경기수 | 타석 | 타수 | 득점 | 안타 | 2루타 | 3루타 | 홈런 | 루타수 |
| 3 | | 강경학(姜景學) | 11 | 한화 | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | | | 14 | 한화 | 41 | 103 | 86 | 11 | 19 | 2 | 3 | 1 | 30 |
| 5 | | 강구성(姜九成) | 13 | NC | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | | 강귀태(姜貴太) | 10 | 넥센 | 97 | 300 | 269 | 24 | 62 | 9 | 0 | 3 | 80 |
| 7 | | | 11 | 넥센 | 33 | 89 | 81 | 4 | 19 | 4 | 1 | 1 | 28 |
| 8 | | | 12 | 넥센 | 12 | 21 | 19 | 2 | 4 | 0 | 0 | 0 | 4 |
| 9 | | 강동우(姜東佑) | 10 | 한화 | 98 | 363 | 309 | 45 | 78 | 8 | 1 | 4 | 100 |
| 10 | | | ★11 | 한화 | 133 | 599 | 518 | 83 | 149 | 13 | 3 | 13 | 207 |
| 11 | | | 12 | 한화 | 76 | 290 | 257 | 35 | 65 | 16 | 1 | 2 | 89 |
| 12 | | | 13 | 한화 | 26 | 60 | 52 | 4 | 11 | 1 | 1 | 0 | 14 |
| 13 | | 강명구(姜明求) | 10 | 삼성 | 55 | 70 | 63 | 17 | 17 | 5 | 0 | 0 | 22 |
| 14 | | | 11 | 삼성 | 89 | 66 | 58 | 22 | 10 | 1 | 0 | 1 | 14 |
| 15 | | | 12 | 삼성 | 72 | 11 | 10 | 16 | 1 | 0 | 0 | 0 | 1 |
| 16 | | | 13 | 삼성 | 55 | 63 | 58 | 22 | 11 | 0 | 1 | 0 | 13 |
| 17 | | | 14 | 삼성 | 21 | 4 | 4 | 8 | 0 | 0 | 0 | 0 | 0 |
| 18 | | | 11 | 삼성 | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 19 | | | 12 | 삼성 | 3 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| 20 | | | 13 | 삼성 | 4 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 21 | | 강민국(姜玟局) | 14 | NC | 6 | 3 | 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| 22 | | 강민호(姜珉鎬) | ★10 | 롯데 | 117 | 465 | 410 | 66 | 125 | 19 | 1 | 23 | 215 |
| 23 | | | ★11 | 롯데 | 124 | 506 | 450 | 63 | 130 | 25 | 2 | 19 | 216 |
| 24 | | | ★12 | 롯데 | 119 | 454 | 400 | 41 | 109 | 21 | 0 | 19 | 187 |
| 25 | | | ★13 | 롯데 | 105 | 405 | 327 | 48 | 77 | 13 | 0 | 11 | 123 |
| 26 | | | 14 | 롯데 | 98 | 360 | 310 | 37 | 71 | 14 | 2 | 16 | 137 |

- 연도별 팀 통산 성적 및 전적자료 (19개)

- * 성적자료 : 통산팀 투수 성적, 통산팀 타격 성적, 팀투수/타자 성적, 평균자책점/타율 순위, 부문별 최다 선수, 구장별 투수/타격 성적, 팀 연도별 주요 부문 성적, 팀 월별 통산 성적, 연도별 구단 변천 및 팀 순위, 구단별 감동 이동상황, 연도별 감독성적
- * 전적자료 : 통산팀간 승패, 팀순위 및 팀간 승패, 전기/후기 승패, 팀 연도별 월별 승패 [그림 2]

[그림 2] 통산팀간 승패 자료 예시

| | A | B | C | D | E | F | G | H | I | J |
|----|---|--------------|------------|---------|---------|------------|-----------|------------|------------|------------|
| 1 | | | | | | | | | | |
| 2 | | 통산팀간승패 | | | | | | | | |
| 3 | | ※1982~2014종합 | | | | | | | | |
| 4 | | 팀 | 삼성 | 넥센 | NC | LG | SK | 두산 | 롯데 | KIA |
| 5 | | 삼성 | ◆ | 78-2-46 | 21-2-9 | 337-10-264 | 141-7-127 | 323-17-269 | 361-14-234 | 319-12-278 |
| 6 | | 넥센 | 46-2-78 | ◆ | 14-0-18 | 76-0-50 | 46-4-76 | 62-1-63 | 60-4-62 | 59-1-66 |
| 7 | | NC | 9-2-21 | 18-0-14 | ◆ | 14-0-18 | 18-0-14 | 12-0-20 | 15-2-15 | 18-1-13 |
| 8 | | LG | 264-10-337 | 50-0-76 | 18-0-14 | ◆ | 123-7-145 | 283-16-310 | 307-16-286 | 277-12-320 |
| 9 | | SK | 127-7-141 | 76-4-46 | 14-0-18 | 145-7-123 | ◆ | 133-5-137 | 155-9-111 | 136-5-134 |
| 10 | | 두산 | 269-17-323 | 63-1-62 | 20-0-12 | 310-16-283 | 137-5-133 | ◆ | 304-13-294 | 300-15-296 |
| 11 | | 롯데 | 234-14-361 | 62-4-60 | 15-2-15 | 286-16-307 | 111-9-155 | 294-13-304 | ◆ | 271-15-325 |
| 12 | | KIA | 278-12-319 | 66-1-59 | 13-1-18 | 320-12-277 | 134-5-136 | 296-15-300 | 325-15-271 | ◆ |
| 13 | | 한화 | 220-10-303 | 64-1-61 | 14-0-18 | 273-18-242 | 108-8-159 | 248-5-278 | 256-15-260 | 229-9-293 |
| 14 | | 쌍방울 | 58-5-101 | ◆ | ◆ | 68-3-93 | ◆ | 66-5-91 | 73-4-85 | 55-7-100 |
| 15 | | 현대 | 210-13-260 | ◆ | ◆ | 207-16-260 | 82-5-62 | 231-10-244 | 232-11-242 | 218-9-258 |

- 2015년 구장별 일일 입장객 수 [그림 3]

[그림 3] 구장별 일일 입장객수 자료 예시

| | A | B | C | D | E | F |
|----|-------------------|----|-----|-----|----|--------|
| 1 | 날짜 | 요일 | 홈 | 방문 | 구장 | 관중수 |
| 2 | 경기수 : 396 | | | | | |
| 3 | 경기평균 : 10,346 | | | | | |
| 4 | 경기 합계 : 4,097,087 | | | | | |
| 5 | 2015-03-28 | 토 | 넥센 | 한화 | 목동 | 12,500 |
| 6 | 2015-03-28 | 토 | 롯데 | kt | 사직 | 27,500 |
| 7 | 2015-03-28 | 토 | KIA | LG | 광주 | 22,000 |
| 8 | 2015-03-28 | 토 | 두산 | NC | 잠실 | 21,746 |
| 9 | 2015-03-28 | 토 | 삼성 | SK | 대구 | 10,000 |
| 10 | 2015-03-29 | 일 | 넥센 | 한화 | 목동 | 10,369 |
| 11 | 2015-03-29 | 일 | 롯데 | kt | 사직 | 13,615 |
| 12 | 2015-03-29 | 일 | KIA | LG | 광주 | 13,835 |
| 13 | 2015-03-29 | 일 | 두산 | NC | 잠실 | 15,814 |
| 14 | 2015-03-29 | 일 | 삼성 | SK | 대구 | 8,465 |
| 15 | 2015-03-31 | 화 | LG | 롯데 | 잠실 | 12,277 |
| 16 | 2015-03-31 | 화 | kt | 삼성 | 수원 | 10,886 |
| 17 | 2015-04-01 | 수 | SK | KIA | 문학 | 12,354 |
| 18 | 2015-04-01 | 수 | LG | 롯데 | 잠실 | 14,260 |

- 사전 제공된 데이터 외에 추가로 데이터를 수집하여 사용할 수 있음

2. 분석 방법

(1) 예측 방법론 관점

- 빅콘테스트 2015 대회 문제는 기계학습 기반 야구경기 예측으로 [그림 4]와 같이 문제를 도식화 할 수 있음
- “입력 값 → 모델 → 결과 값” 의 3단계로 추상화

[그림 4] 콘테스트 문제해결 과정 도식화



- 입력 값 : 콘테스트에서 제공한 자료 이외에 필요한 자료를 추가적으로 수집하고 적용함
- 예측 모델 : 입력 값을 바탕으로 결과 값을 추정하는 것으로 기계학습을 비롯한 여러 가지 방법론 사용
- 결과 값 : 콘테스트 문제에서 요구하는 프로야구 팀 별 승률이나 누적 관객수의 예측 값

□ 참가자 팀 별 방법론 분석

- 모든 참가자(103팀)들이 제출한 결과 발표자료와 보고서를 토대로 예측 방법론과 세부내용 정리 (분석 예시 <표 2>, 수상자 분석은 부록참조)

<표 2> 챌린지리그 결과분석 예시

| 팀명 ⁴⁾ | 예측 방법론 | 방법론 분류 | 세부내용 정리 | RMSE | 순위 |
|------------------|---|--------|--|--------|----|
| Team1 | Deep Neural Network (Multi-training) | 인공신경망 | ○ Model - Deep Neural Network (Multi-step MLP) ○ Input/Output Parameters - 타자, 투수의 각 지표별상관관계가 분석 - 타자의 경우 ACF를 사용하여 40타석 고려 ○ Development - input : 타자, 투수의 누적데이터 - hidden : 3 layers - output : 승/패 확률 - Python: Library (scikit-learn, Keras) | 0.0156 | 14 |
| Team2 | Random Forest, CART, Multi-layer Perceptron, Conditional Inference Tree | 기타기계학습 | ○ Models - 네 가지 방법론을 종합하여 최종 승률 예측 ○ Input/Output Parameters - 4가지 유형별 승패 (팀자체, 상대팀별, 요일별, 홈/어웨이) - 4가지 모델의 승패 결과 (0 or 1) ○ Development - 방법론별 정확도를 가중치로 사용함 | 없음 | 없음 |
| ... | | | | | |

- <표 2>의 내용을 바탕으로 예측 방법론 사용 분포와 기계학습 적용 비율 두 가지 지표를 도출함
 - 예측 방법론 사용 분포는 각 팀이 최소 한 번 이상 사용한 방법론을 모두 포함함
 - * 예측 과정이 아닌 데이터 선별과정에서도 방법론이 사용될 수 있으며, 여러 가지 방법론을 적용하여 가장 좋은 예측 모델을 제안한 경우도 해당 방법론 들을 모두 예측 방법론 분포에 포함
 - * 예를 들어 <표 2>의 ‘Team2’ 의 경우 “예측 방법론” 열에 기술된 4가지 방법론을 사용함 (Random Forest, Classification and regression tree, Multi-layer perceptron, Conditional inference tree)
 - 기계학습 적용 비율은 인공신경망, 회귀분석, 기타 기계학습, 통계모델로 분류하여 도출 (<표 3> 참조)

4) 팀명은 익명화 하여 표현함

<표 3> 기계학습 적용 비율 도출을 위한 분류표

| 방법론 분류 | 세부 예측 방법론 (예시) | | 비 고 |
|---------|---|-----|---------------------------------------|
| 인공신경망 | ◦ Deep Neural Network ◦ Deep Belief Network | 딥러닝 | 기계학습 (학습과정이 포함됨) |
| | ◦ Multi-layer Perceptron | | |
| 회귀분석 | ◦ Linear Regression ◦ Logistic Regression ◦ Poisson Regression ◦ Regularized Regression ◦ Auto Regressive Method | | |
| | ◦ Decision Tree ◦ Random Forest ◦ Ensemble Learning ◦ Principal Component Analysis ◦ Support Vector Machine | | |
| | ◦ Pythagorean Expectation ◦ Bradley-Terry Model ◦ Monte Carlo Method ◦ Distribution Analysis ◦ Simple Modeling (Mean, Median) | | |
| | | | |
| | | | |
| 기타 기계학습 | | | 비기계학습 (학습과정이 없이 모델자체로 결과 예측) |
| 통계모델 | | | |

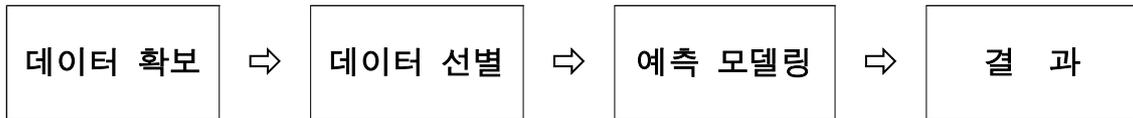
- * 통계모델은 학습과정이 없이 야구통계분야에서 도출된 모델이나 간단한 평균, 중간 값 등을 사용하여 직접 예측결과를 제시함
- * 인공신경망, 회귀분석, 기타 기계학습, 통계모델 4가지 분류는 <표 2>의 “방법론 분류” 열에 표현하고, 이것은 팀 별 “예측 방법론” 열에서 가장 중요도가 높은 것으로 추정
- * 예를 들어 <표 2>의 ‘Team2’ 는 4가지 방법론 중 랜덤포레스트를 가장 비중 있게 사용했으므로 ‘기타 기계학습’ 으로 분류
- * 딥러닝은 인공신경망 분류에 포함됨⁵⁾

5) <표 3>에 제시된 방법 이외에 Recurrent Neural Network, Convolutional Neural Network, Restricted Boltzmann Machine 등이 딥러닝에 포함됨

(2) 문제해결 절차

□ 콘테스트 참가자 103개 팀의 분석 결과 대다수의 참가자가 <표 4>과 같은 네 단계의 절차를 거침

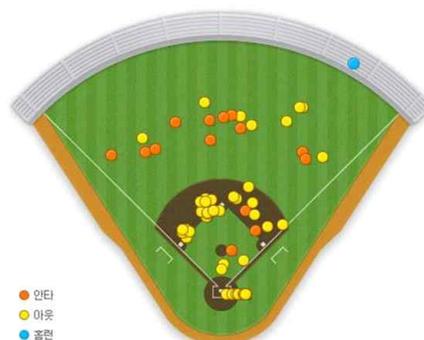
<표 4> 콘테스트 문제해결 절차



① 데이터 확보

- 대부분의 참가자들이 대회에서 제공한 데이터 보다 더 상세한 자료를 수집하고 가공함
- * 제공된 투수/타격 데이터가 연도별 자료이기 때문에, 추가적으로 경기별 데이터를 수집하여 예측 모델에 적용함. 또한 연도별 투수 29개, 타격 26개 지표를 사용하여 미국 메이저리그의 통계수치인 Sabermetrics⁶⁾를 도출함
- * 참가자 중 10개 팀이 웹 크롤링⁷⁾을 사용하여 데이터를 자동으로 수집하고 분류함
- * 특히 경기 당 타자의 타격결과를 기록한 이미지 [그림 5]를 추상화하여 팀의 공격력을 추정하고, 예측모델의 지표로 사용한 경우도 존재함

[그림 5] 타격결과 이미지



자료 : 네이버 스포츠 게임센터

6) 미국의 야구통계학자 빌 제임스가 제안한 통계학적/수학적 야구 분석 방법론으로 기본적으로 수집된 야구 경기 지표를 통계모델을 사용하여 새로운 의미를 부여함. 예를 들어 BABIP(Batting Average on Balls in Play)는 인플레이로 이어진 타구에 대한 타율을 계산하는 지표임

7) 자동화된 방법으로 월드 와이드 웹(www)을 탐색하고 수집하는 행위를 지칭함

- 문제의 예측 성능을 높이기 위해서 야구 외적인 요소도 고려
 - * 퓨처스리그의 누적 관객수 예측에서는 날씨와 같은 경기 외적인 요소가 큰 관련성이 있기 때문에 참가팀들이 별도로 기상청의 일기예보 등을 토대로 예측 모델에 반영함

② 데이터 선별

- 약 80여 가지의 데이터가 제공되었으나, 목표문제와의 관련도가 적은 지표도 있기 때문에 이를 취사선택할 수 있는 데이터 선별과정 수행
 - * 선형 상관관계(Linear correlation) 분석은 가장 먼저 시도할 수 있는 선별 방법으로, 특정 데이터 항목과 결과 값이 선형적으로 얼마나 관계가 있는지를 나타냄⁸⁾
 - * 예를 들어 팀의 총 득점이 증가할수록 승률이 증가하는 반면에, 팀의 총 홈런수와 승률은 큰 관련성이 없음
 - * 그 밖에, 기계학습의 차원 축소(Dimensionality reduction) 기법으로 데이터를 선별할 수 있는 주성분분석(Principal Component Analysis), 선형/비선형 상관 계수를 도출하는 회귀분석(Regression) 기법 등이 사용됨
- 수상한 팀의 대부분이 세밀한 데이터 선별과정을 거쳤기 때문에, 참가자들이 데이터의 중요성을 잘 인지했다고 볼 수 있음

③ 예측 모델링

- 예측 모델링에 사용된 방법론은 <표 3>와 같이 분류되고, 기계학습의 경우 모델을 학습하는 과정과 예측하는 과정이 구분되는 반면 통계모델은 학습의 과정 없이 바로 결과를 예측함
 - * 기계학습의 경우 주로 인공신경망, 회귀분석을 사용했고 기타 기계학습으로는 결정나무, 지지벡터머신 등이 사용됨
 - * 통계모델은 야구 통계모델을 직접 차용해서 사용하거나, 확률분포 분석과 평균을 사용하여 직접 모델링 한 경우가 포함됨

④ 결과 : 예측 모델을 바탕으로 결과 값 예측

8) 특정 데이터 항목의 값이 증가할수록 결과 값이 증가하거나 감소하는 경향이 있으면 높은 선형 상관 관계가 있음

3. 분석 결과

(1) 챌린지리그 - 프로야구 승률 예측

□ 예측 방법론 분포와 기계학습 적용 비율 도출

○ 총 참가 팀 수는 58개이고, 이 중 분석이 가능한 팀은 52개

* 나머지 6개 팀은 결과물의 설명이 부족하여 분석에서 제외

○ 예측 방법론 분포 [그림 6]과 상위 5개 예측 방법론 <표 5>

- 한 팀이 여러 가지 방법론을 사용한 것도 모두 포함

[그림 6] 챌린지리그 예측 방법론 분포



<표 5 > 챌린지리그 예측 방법론 순위 (상위 5개)

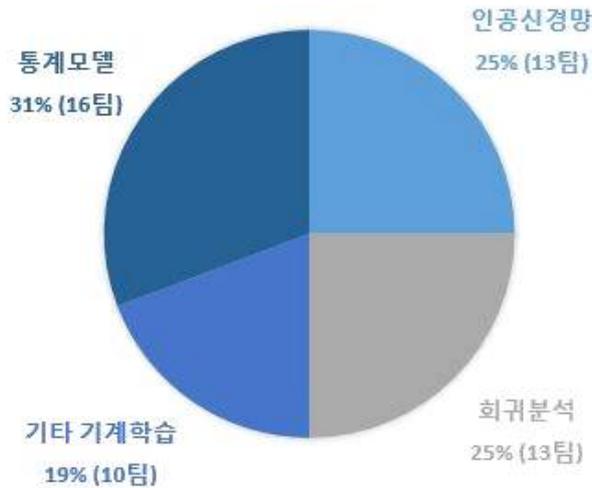
| 예측 방법론 | 사용 횟수 | 사용 비율 |
|---------------------------------------|-------|-------|
| 총 계 | 859) | 100% |
| 선형 회귀분석 (Linear Regression) | 14 | 16.5% |
| 피타고리안 승률 (Pythagorean Expectation) | 13 | 15.3% |
| 다층 퍼셉트론 (Multi-layer Perceptron) | 10 | 11.8% |
| 랜덤 포레스트 (Random Forest) | 7 | 8.2% |
| 심층신경망 (Deep Neural Network) | 6 | 7.1% |
| 기타 방법론 | 35 | 41.1% |

* <표 5>의 5개 방법론은 [부록]에서 간략히 소개

9) 52개 팀이 총 사용한 예측 방법론의 수는 85개이고, <표 4>의 ② 데이터 선별과정과 ③ 예측 모델링과정에서 사용된 방법론을 모두 포함

- <표 3>의 분류방법에 따른 결과는 [그림 7]과 같음
 - 각 팀에서 가장 비중 있게 사용한 예측 방법론을 4가지 분류중 하나로 선택함

[그림 7] 기계학습 적용 비율 (총 52개 팀)



- 분석 결과 챌린지리그 참가자의 69%(36팀)가 기계학습(인공지능경망, 회귀분석, 기타 기계학습)을 사용하여 목표문제를 해결

□ 수상결과 <표 6> 분석

<표 6> 챌린지리그 수상 결과

| 팀 명 | 예측 방법론 | RMSE | 딥러닝 사용여부 | RMSE 순위 (1차 서류심사) | 최종 순위 |
|------------|---|--------|----------|-------------------|-------|
| Challenge1 | Principal Component Analysis Exploratory Data Analysis Multi-step Linear Regression | 0.0091 | - | 1 | 2 |
| Challenge2 | Deep Neural Network | 0.0106 | O | 2 | 1 |
| Challenge3 | Support Vector Machine | 0.0130 | - | 4 | 3 |
| Challenge4 | Linear Regression | 0.0131 | - | 5 | 5 |
| Challenge5 | Linear Regression Regularized Regression Random Forest Support Vector Machine | 0.0143 | - | 6 | 4 |
| Challenge6 | Linear Regression Pythagorean Expectation | 0.0147 | - | 7 | 6 |
| Challenge7 | Pythagorean Expectation | 0.0149 | - | 8 | - |
| Challenge8 | Bradley-Terry Model Pythagorean Expectation | 0.0151 | - | 9 | - |
| Challenge9 | Deep Neural Network (Multi-training) | 0.0156 | O | 14 | 7 |

- 1차 서류심사는 평가지표의 첫 번째 기준인 RMSE와 두 번째 기준인 딥러닝 사용 여부로 평가 (58개 팀 중 9개 팀 선별)
 - 서류심사 통과 9개 팀의 분석내용은 부록 참조
- 최종 수상결과는 RMSE, 발표평가, 기계학습 가산점 세 가지 항목으로 구분하여 평가하고 9개 팀 중 7개 팀 선별
 - RMSE에 가장 높은 가중치가 부여되었으나, 발표평가와 기계학습 가산점으로 인한 최종 순위변동이 있었음
 - * 기계학습을 잘 활용한 ‘Challenge9’ 팀이 수상 순위권에 포함되어 기계학습 가산점이 수상의 결정요소로 작용됨을 입증함
- 대상 수상팀 사례분석

- Challenge2 팀
- RMSE : 0.0106
 - 최종 결과 값인 9월 30일(141경기) 기준 1경기당 승률은 0.007로 10개 팀별 약 1.5경기 예측 오류
 - 9월 6일부터 30일까지 잔여경기는 약 20경기로 잔여경기 대비 예측오류는 7.5%
- 추가 데이터 수집
 - 웹 크롤링을 통한 경기별 상세기록 수집
 - 승률 예측에 가장 중요한 요소로 ① 상대팀과의 역대 전적, ② 최근 경기 결과를 바탕으로 한 팀의 상승/하락세
- 예측 모델
 - 딥 러닝을 사용하여 잔여 경기별 승/패 예측
 - ① 상대팀과의 역대 전적데이터 → 딥러닝 → 승리확률 예측
 - ② 최근 경기 결과 데이터 → 딥러닝 → 승리확률 예측
 - 가중치*① + ②의 모델로 최종 승/패 예측
- 결과 분석
 - 수상의 요인으로는 **상관성이 높은 데이터의 선별**에서 찾을 수 있음
 - **딥 러닝과 앙상블 학습**을 통해 예측 성능을 높임

[그림 8] Challenge2 팀 발표자료



○ 우수 참가자 사례분석

- Challenge9 팀
- RMSE : 0.0156
 - 프로야구 팀별 약 2.2경기 예측 오류
 - 잔여경기(약 20경기) 대비 예측 오류는 11%
- 인공지능망 모델링
 - 야구의 계층적 특성을 반영함
(1구 → 타석 → 회 → 경기)
 - 타자의 타석별 지표를 통하여 각 회별로 나올 수 있는 결과를 종합한 뒤 최종 승리 확률을 도출 [그림 9]

[그림 9] Challenge9 팀 발표자료

- 예측 모델
 - 다층 퍼셉트론을 2중으로 설계함
 - ① 타자의 과거 성적과 경기 외부요인을 종합하여 각 타자가 네 번의 타순에서 성적(performance)을 도출
 - ② 타자의 성적에서 도출된 값으로 각 회별 결과를 예측하고, 이를 바탕으로 최종 승리확률을 예측 함
- 결과 분석
 - 인공지능망이 가지고 있는 계층적 특성을 가장 잘 반영하여 모델링 함
 - 전반적으로 기계학습에 대한 이해도가 높음

○ 수상결과를 종합해 보면 기계학습 방법론을 사용한 참가자들이 대체로 예측 성능이 뛰어났으나 더 간단한 통계모델로도 예측이 가능함

- 켈린지리그에서 수상한 7팀 중 6팀이 기계학습 방법론 사용
- 피타고리안 승률(부록 참조)은 야구 통계학자가 직접 개발한 모델로 실제 승률과 매우 높은 상관관계가 있기 때문에 예측의 정확도 측면에서 이점을 가짐

* 실제로 1차 서류심사를 통과한 9개 팀 중 3개 팀이 피타고리안 승률을 사용함

* 피타고리안 승률과 같이 실제 결과와 밀접한 관계가 있는 모델은 참고모델로 제시하여 기준점을 부여하는 등의 개선이 필요함

(2) 퓨처스리그 - 프로야구 누적 관객수 예측

□ 예측 방법론 분포와 기계학습 적용 비율 도출

- 총 참가 팀 수는 45개이고 모두 분석 대상임
- 예측 방법론 분포 [그림 10]과 상위 5개 예측 방법론 <표 7>

[그림 10] 퓨처스리그 예측 방법론 분포



<표 7 > 퓨처스리그 예측 방법론 순위 (상위 5개)

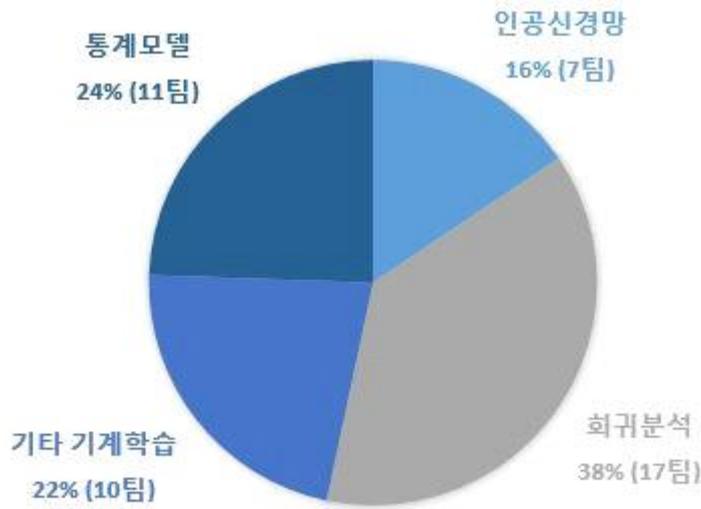
| 예측 방법론 | 사용 횟수 | 사용 비율 |
|-------------------------------------|-------|-------|
| 총 계 | 6910) | 100% |
| 선형 회귀분석 (Linear Regression) | 18 | 26.1% |
| 평균 모델 (Mean) | 9 | 13.0% |
| 랜덤 포레스트 (Radom Forest) | 9 | 13.0% |
| AR 방법 (Auto Regressive Method) | 5 | 7.25% |
| 다층 퍼셉트론 (Multi-layer Perceptron) | 5 | 7.25% |
| 기타 방법론 | 23 | 33.4% |

* <표 6>의 5개 방법론 중 챌린지리그와 중복된 것을 제외하고 [부록]에서 간략히 소개

10) 45개 팀이 총 사용한 예측 방법론의 수는 69개

○ <표 3>의 분류방법에 따른 결과는 [그림 11]과 같음

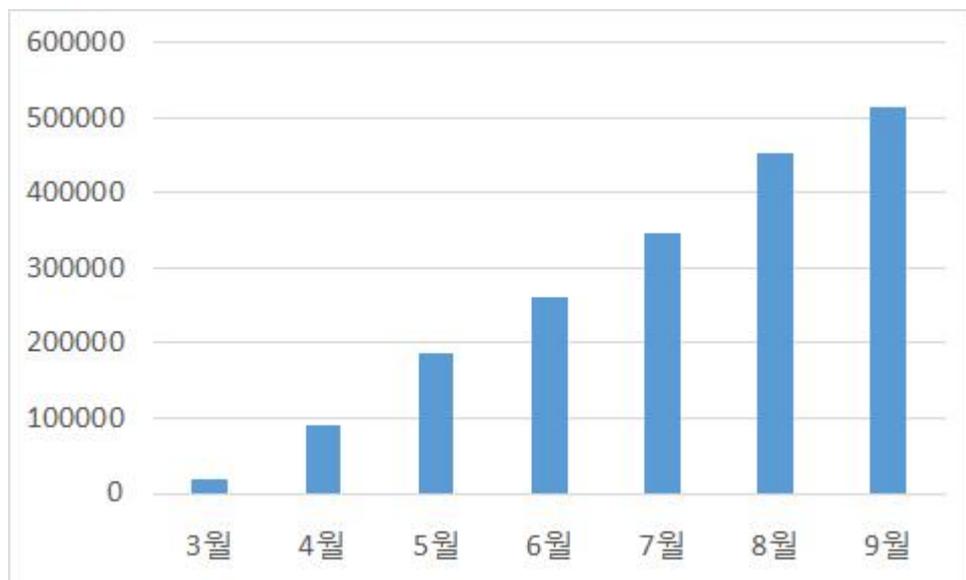
[그림 11] 기계학습 적용 비율 (총 45개 팀)



- 분석 결과 챌린지리그 참가자의 76%(34팀)가 기계학습(인공지능경망, 회귀분석, 기타 기계학습)을 사용하여 목표문제를 해결
- 기계학습 방법론 중 회귀분석의 사용 비율이 가장 큰 이유는 누적 관객수 자체가 시간에 따라 증가하는 경향을 갖기 때문

* [그림 12]는 삼성구단의 월 별 총 누적관객수를 나타낸 것으로 선형적으로 증가함을 알 수 있음

[그림 12] 2015년 삼성 구단의 월별 누적 관객수 (명)



□ 수상결과 <표 8> 분석

<표 8> 퓨처스리그 수상 결과

| 팀 명 | 방법론 | RMSE | 딥러닝 사용여부 | RMSE 순위 (1차 서류심사) | 최종 순위 |
|----------|--|--------|-------------|----------------------|-------|
| Futures1 | Linear Regression | 14,523 | - | 1 | 2 |
| Futures2 | Linear Regression | 15,335 | - | 2 | 1 |
| Futures3 | Linear Regression | 16,575 | - | 3 | 3 |
| Futures4 | Mean | 16,790 | - | 4 | 5 |
| Futures5 | Mean | 16,976 | - | 5 | - |
| Futures6 | Restricted Boltzmann Machine Stacked Autoencoder | 17,372 | 0 | 7 | 4 |
| Futures7 | Random Forest | 17,479 | - | 8 | 6 |
| Futures8 | Linear Regression | 17,533 | - | 9 | - |
| Futures9 | Random Forest | 17,727 | - | 10 | 7 |

- 평가지표는 챌린지리그와 동일하고 1차 서류심사를 통과한 9개 팀에 대한 세부내용은 부록 참조
- 최종 평가 결과역시 RMSE가 가장 큰 평가지표로 사용되었으나, 기계학습 방법론을 사용한 팀에 대한 가중치를 부과하여 변별력을 높임
- 대상 수상팀 사례분석

- **Futures2 팀**
- RMSE : 15,335명
 - 9월 30일 기준 10개 구단의 총 누적 관객수는 7,172,865명이고, 예측해야할 9월 관객수는 1,146,989명으로 이 기간 중 관객 10,000명은 약 0.87% 비중을 차지
 - RMSE는 구단별 예측이 평균적으로 15,335명이 차이가 나는 것을 표현하며, 이는 예측한 9월 관객수 대비 약 1.3%의 오차를 가짐
- 데이터 수집 및 선별
 - 날씨의 특수성 반영 (기상정보)
 - 주중과 주말, 공휴일의 특성 반영
 - 홈/원정 여부, 구단별 관중 동원력

◦ 예측 모델

- 선형 회귀분석으로 일자별 관중 수 예측
- 기상정보 활용 가능 여부에 따라서 모델을 구분함

① 9/6 ~ 9/15 : 팀 정보, 요일, 주말 여부, 기온, 강수여부

② 9/15 ~ 9/30 : 팀 정보, 요일, 주말여부

◦ 결과 분석

- 여러 가지 선형모형을 적용하여 예측에 사용할 최적의 모수를 선택함

[그림 13] Futures2 팀 발표자료

◦ 우수 참가자 사례분석

◦ Futures6 팀

◦ RMSE : 17,372명

- 프로야구 팀별 약 1.5% 9월 관객수 예측 오류

◦ 데이터 선별

- 입력 값 총 19개 (날씨, 요일, 홈/원정, 경기시장, 구장 등)
- 예측 값은 9월 구장별 관객수

◦ 예측 모델

- Deep Belief Network와 Stacked Auto-Encoder를 사용하여 모델링 한 뒤 성능이 더 좋은 Deep Belief Network 선택
- 과적합(Overfitting)문제를 해결하는 Validation 기법을 사용
 - * Training(학습) → Validation(검증) → Test(테스트)
- 은닉층(Hidden Layer)는 총 3개의 계층을 사용했고, 각 층별 노드 개수는 100개

◦ 결과 분석

- 인공신경망의 계층적 특성을 잘 살리지는 못했으나 Validation 기법을 사용함으로써 예측 성능을 높임
- 타 팀과는 다르게 알고리즘을 직접 구현함 (C++ 사용)

[그림 14] Futures6 발표자료

11) 인공신경망에서 과적합 문제(Overfitting Problem)는 학습 시 학습데이터에 크게 의존하게 되어 발생하는 문제로, 학습데이터의 예측성능은 매우 높지만 테스트데이터의 예측성능은 크게 떨어지는 현상을 말함

- 퓨처스리그의 문제는 챌린지리그 보다 상대적으로 쉽고, 누적 관객수라는 지표 자체가 강한 선형성을 가지기 때문에 선형회귀분석을 사용해도 좋은 예측성능을 보임
 - 문제의 난이도가 낮으면 간단한 모델로도 예측이 가능하기 때문에 참가자들의 변별력이 떨어지므로, 난이도를 높이는 방향이 적절
- * 데이터 자체에서 결과 값과 강한 상관관계가 있는 문제는 적절하지 않음

(3) 종합 결과 분석

- 예측 방법론 분석 결과, 참가자 97개 중 기계학습을 사용한 팀은 72%(70팀)이고, 기계학습 방법론 중에서는 회귀분석이 43%, 인공신경망과 기타 기계학습이 각각 28.5%를 차지함
- 문제의 난이도 관점에서 수상 팀 1등과 7등의 예측성능 차이는 수상 팀 평균 대비 챌린지리그 48.6%, 퓨처스리그 19.4%
 - 각 리그 수상팀의 1등과 7등의 RMSE 차이를 수상팀의 평균으로 나눈 것이 1등과 7등의 실력차이임
 - 챌린지리그 수상팀의 평균 RMSE는 0.0134이고 1등과 7등의 RMSE차이가 0.0065으로 48.5%, 퓨처스리그 수상팀의 평균 RMSE는 16,543명이고 1등과 7등의 RMSE 차이는 3,204명으로 19.4%임
 - 따라서 챌린지리그는 1등과 7등의 실력차이가 퓨처스리그 보다는 크다는 것은 챌린지리그 문제의 난이도가 더 높았다는 점을 시사함
- 이번 콘테스트에서 가산점을 둔 “딥러닝”은 수상의 요인으로 작용했으나 예측 방법론 사용 분포의 관점에서는 많은 비중을 차지하지 못함
 - 챌린지리그에서는 딥러닝을 사용한 팀이 1등을 했지만 단 두 팀만이 딥러닝을 사용함 (<표 6> 참조)
 - 퓨처스리그에서는 단 한 팀이 딥러닝을 사용했고, 그 성능 또한 선형회귀 분석에 미치지 못함

4. 시사점

- 빅콘테스트 2015 분석 결과 보편적으로 기계학습의 예측성능이 뛰어난
 - 참가자의 72%가 기계학습법을 사용하여 프로야구 경기예측 문제 해결
 - 최종 수상을 받은 14개 팀 중 12개 팀이 기계학습 방법론을 사용
 - 기계학습 방법론 중 43%(30팀)가 회귀분석을 사용했고, 이 중 대부분이 선형 회귀분석임
 - 퓨처스리그에서는 누적 관객수 예측 문제 자체가 가지고 있는 선형성 때문에 [그림 12] 17개 팀이 회귀분석을 사용함
 - 챌린지리그에서도 13개 팀이 회귀분석을 사용했는데, 이 부분이 시사하는 바는 높은 상관관계를 갖는 데이터를 선별하여 예측 모델을 단순화 한 것으로 판단됨
 - 선형 회귀분석을 예측모델로 사용한 팀은 1차 서류통과 팀 18개 중 6개로 높은 예측성능을 보였기 때문에, 선형 회귀분석이 문제해결에 있어서 간단하지만 효율적인 모델임
 - 콘테스트에서 가산점을 부여한 딥러닝 방법론은 예측 방법론 분포의 관점에서 챌린지리그 6회, 퓨처스리그 4회가 사용됨¹²⁾
 - 수상팀 14팀 중 딥러닝을 사용한 비율은 21%(3팀)으로 콘테스트의 '딥러닝 사용시 가산점 부여'가 참가자들에게 큰 동기가 됐다고 보기 어려움
 - * 그 이유는 문제자체가 복잡하지 않기 때문에, 선형회귀분석이나 통계모델과 같이 간단한 모델을 사용해도 어느 정도 예측이 가능함
 - 수상팀 중 딥러닝을 적용한 팀은 총 3팀으로 이 중 'Challenge9'팀은 인공신경망의 계층적 특성을 사용하여 딥러닝을 가장 잘 응용함

12) 챌린지리그 6회 (Deep Neural Network 6회), 퓨처스리그 4회 (Deep Neural Network 3회, Restricted Boltzmann Machine 1회)

- 딥러닝은 학습 신경망의 구조에 따라 그 성능이 좌우되고, 어떻게 효율적인 신경망을 구축할 것인가에 대한 변수가 많기 때문에 높은 예측성능을 확보하기 어렵다는 점이 있음
 - * 국내외 딥러닝 관련 강의 자료를 활용하여 효율적인 신경망 구축을 위한 가이드라인을 제시할 필요가 있음

- 향후 콘테스트에서는 문제에 대한 더 구체적인 제한사항과 난이도 조절이 필요함
 - 야구경기 승률을 예측한 챌린지리그에서는 문제의 난이도가 적절한 것으로 판단되나 결과 값과 큰 상관관계를 갖는 모델에 대한 제한이 필요함
 - 피타고리안 승률 모델을 기준으로 제공하여 참가자들이 적용할 모델의 정확도를 판단
 - 퓨처스리그는 상대적으로 난이도가 쉽기 때문에 변별력을 높이기 위한 문제를 고려해야함

- 콘테스트 참가자의 대부분이 데이터의 중요성을 인지하고 추가적인 데이터 확보와 데이터 선별에 큰 비중을 둠
 - 수상자의 대부분이 참신한 데이터 수집과 세밀한 데이터 선별과정을 거침
 - 10개 팀이 웹 크롤링을 사용하여 데이터를 수집하였으며, 여러 가지 상관관계분석을 통해 의미 있는 데이터를 선별
 - 수상팀 중 5개 팀이 웹 크롤링을 사용했으므로, 추가적인 데이터 수집이 수상에 긍정적인 역할을 함

- 빅콘테스트의 향후 발전 방안
 - 예측 문제의 범위와 난이도 조정으로 콘테스트의 질적 향상 도모
 - 자연과학 분야에서는 아직까지 경향성이나 법칙이 발견되지는 않았으나 매우 중요하고 방대한 데이터를 처리해야하는 난제가 많음
 - * 대학과 연계하여 가장 예측을 잘 한 팀과 공동으로 논문을 작성하는 등 참가자 개인의 실적을 쌓을 수 있는 동기 부여

- 또한 빅데이터를 기반으로 한 패턴 분석 등 기업에서 사업적으로 중요한 예측문제들 역시 산재해 있음
 - * 기업에서 실제 문제를 직접 제안하고 이를 가장 성공적으로 해결한 팀에게 취업기회를 부여하는 등의 연계프로그램으로 윈-윈 체계 마련
- 국내외 기계학습 기반 콘테스트에서 향후 방향 모색
 - 기계학습 기반 문제해결 오픈 플랫폼인 캐글¹³⁾은 전 세계의 기업이 상금을 동반한 문제를 내고, 이를 해결한 개인이나 팀에게 상금 수여 (2015년 11월 기준 약 41만 명의 데이터 과학자 인력 풀 확보)
 - * 대표적인 문제로 2012년 병원에서 불필요한 진료 여부 예측문제(Heritage health prize)에 3백만 달러의 상금을 제시함
 - * 마이크로소프트의 동작인식 장치인 키넥트의 성능을 높이는 문제와 유럽입자연구소(CERN)의 힉스입자와 관련한 문제 등 다양한 분야를 포괄함
 - 국내에서는 SK 플래닛의 코드스프린트¹⁴⁾가 다양한 문제를 제시함
 - * 2015년 7월에 실시된 문제는 두 가지로 로보코드 챌린지 게임과 VOD 추천
 - * 문제의 결과에 따라 1~3등에게는 상품이 수여되고, SK 플래닛 입사 시 가산점을 부여함
 - 우리나라에서도 캐글과 같은 시스템 도입을 적극적으로 고려하여 현실적인 기업의 문제를 해결하는 방향을 모색해야 함
 - * 지식기반의 SW 중심사회는 현실적인 문제를 해결하고, 지능형 SW의 핵심인 데이터 과학자를 육성해야 실현될 수 있음

13) Kaggle 홈페이지, <https://www.kaggle.com>

14) SK 플래닛 코드스프린트 홈페이지, <http://codesprint.skplanet.com/2015>

[부 록]

1. 챌린지리그 / 퓨처스리그 방법론 소개

< 선형회귀분석 >

- 선형회귀(Linear Regression) 분석은 종속변수와 한 개 이상의 독립변수(설명변수)와의 선형상관관계를 모델링하는 기법
- 한 개의 독립변수를 사용할 경우 단순 선형회귀, 두 개 이상의 독립변수를 사용할 경우에는 다중 선형회귀라고 지칭함
 - 이 관계를 일반화하면 주어진 데이터 집합 $\{y_j, x_{j1}, x_{j2}, \dots, x_{jp}\}_{j=1}^N$ 에 대하여 다음과 같은 식으로 나타낼 수 있음

$$y_j = \sum_{k=1}^p a_k x_{jk} + b_j, \quad j = 1, \dots, N$$

- 여기서 y_j 는 종속변수, x_{jk} ($k = 1, \dots, p$)는 p 개의 독립변수를 나타내고 총 데이터의 수는 N 개
- 위 식에서 미지수는 p 개의 선형 기울기 a_k 와 N 개의 절편 b_j 이고, 일반적으로 미지수보다 데이터가 많은 overdetermined system이기 때문에 이 시스템을 푸는 해법에 따라 해가 무수히 많을 수 있음
- 시스템 해법의 전통적인 방법에는 최소제곱법(Least Squares Method)이 있음
- 선형회귀분석은 독립변수와 종속변수가 선형적인 상관관계가 있다는 가정이 있어야 성공적인 예측 모델로 사용할 수 있으므로, 독립변수의 선택이 가장 중요함

< 피타고리안 승률 >

- 피타고리안 승률(Pythagorean Expectation)은 미국의 야구 통계학자 빌 제임스가 제안한 공식으로 승률이 팀의 총 득점과 실점으로 표현될 수 있음
 - “피타고리안”이라는 단어는 득점과 실점의 관계가 피타고라스의 정리와 유사한 형태를 가지기 때문
 - 구체적인 식은 다음과 같음

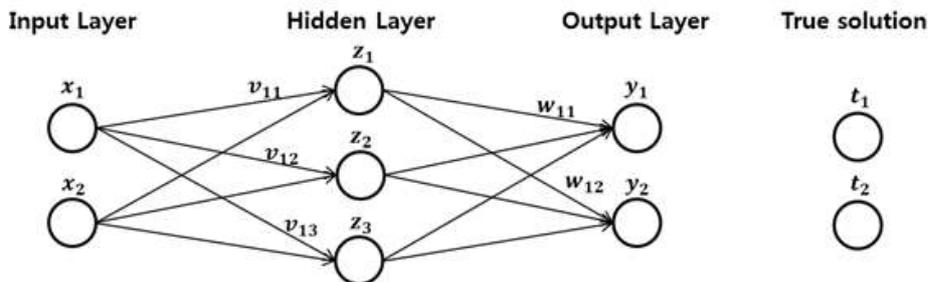
$$\text{승률} = \frac{\text{총 득점}^n}{\text{총 득점}^n + \text{총 실점}^n}$$

- 여기서 n 은 보통 2를 나타내지만 리그마다 조금씩 다를 수 있음 (우리나라 프로야구는 1.8 ~ 1.85 사이의 수치임)
- 피타고리안 승률은 실제 승률과 매우 높은 상관관계를 가지고 있으므로, 콘테스트의 많은 참가자들이 프로야구 승률예측에 사용함

< 다층퍼셉트론 / 심층인공신경망 >

- 다층퍼셉트론(Multi-layer Perceptron)과 심층인공신경망(Deep Neural Network)은 기계학습의 인공신경망(Artificial Neural Network) 기법에서 가장 널리 사용되는 방법
 - 다층퍼셉트론과 심층인공신경망은 네트워크의 구조적인 측면에서 큰 차이가 없지만, 신경망을 학습하는 과정에서 그 차이점이 존재함
 - 다층퍼셉트론의 경우 지도학습(Supervised Learning)으로 신경망을 학습하는데 여기서 “Vanishing Gradient”¹⁵⁾문제가 발생하기 때문에, 이것을 극복하고자 한 학습 방법이 심층인공신경망에 적용된 비지도학습(Unsupervised Learning)임
 - 인공신경망의 가장 큰 장점은 비선형관계를 모델링 할 수 있다는 점이지만, 경험에 의한 최적화가 필수적이므로 이에 소요되는 비용이 상당할 수 있음
- 인공신경망으로 예측모델을 구축하는 단계는 크게 학습(train)과 시험(test) 두 단계로 나누어짐
 - 인공신경망의 구조는 입력층(Input Layer), 은닉층(Hidden Layer), 결과층(Output Layer)로 구성되며 은닉층의 개수와 해당 은닉층의 노드개수를 정함으로써 신경망을 구축할 수 있음
 - 이 신경망을 학습(train)시키는 방법 중, 지도학습의 경우 일반적으로 오차역전파법(Error Back-propagation Method)을 사용하며 이것은 신경망의 오차를 최소화하는 가중치 갱신을 통해 이루어짐
 - 다층퍼셉트론을 예로 들어보면 다음과 같은 그림으로 표현되는데, 이 구조에서 미지수는 가중치인 v 와 w 임

[그림 15] 다층퍼셉트론 구조 예시



- 층과 층 사이는 활성화함수가 존재하며 이것은 정보의 전파도 혹은 전파확률을 나타냄. 사용되는 함수는 시그모이드, 탄젠트하이퍼볼릭, 가우시안분포 등이 있음

15) Vanishing Gradient문제는 지도학습의 오차역전파법으로부터 제기된 문제로, 오차를 미분한 값이 빈번한 연쇄법칙(Chain rule)에 의해서 사라지게(vanish)되는 것을 뜻함

https://en.wikipedia.org/wiki/Vanishing_gradient_problem

< 랜덤포레스트 >

- 랜덤포레스트(Random Forest)는 앙상블 학습 방법의 일종으로, 다수의 결정트리를 학습시키고 이로부터 분류나 예측을 출력하는 기계학습 기법임
 - 결정나무(Decision Tree)기법은 그 결과와 성능의 변동 폭이 크다는 결점을 가지고 있기 때문에 이를 극복하는 것이 이 방법론이 부상하게 된 동기임
 - 랜덤포레스트의 가장 핵심적인 특징은 임의성(Randomness)에 의해서 서로 다른 특성을 갖는 여러 개의 결정트리로 구성된다는 점이고, 이에 따라 결과의 일반화 성능을 향상시킴
- 랜덤포레스트의 학습방법 중 가장 일반적인 방법은 배깅(Bagging)을 사용한 방법으로 이것은 부트스트랩을 통해 여러 개의 트리를 학습시키고 하나의 포레스트로 결합하는 과정을 말함
 - 결정나무는 작은 편향과 큰 분산을 갖는 경향이 있기 때문에 과적합의 위험이 존재함. 부트스트랩 과정은 편향의 크기는 유지하면서 분산을 감소시키기 때문에 더 나은 성능을 확보할 수 있음
 - 추가적으로 랜덤포레스트 기법을 사용하여 변수의 중요도를 측정할 수 있음

< AR 방법 >

- AR(Auto Regressive) 방법은 시계열 분석(Time series analysis)의 한 방법으로 미래의 예측 값을 과거의 데이터로 추정하는 것을 말함
 - AR 방법은 과거 데이터의 수 p 에 따라 $AR(p)$ 라고 정의하며 이것을 수식으로 나타내면 다음과 같음

$$X_t = c + \sum_{j=1}^p \phi_j X_{t-j} + \epsilon_t$$

- 여기서 X_t 는 t 시점에서의 시계열 데이터를 나타내며 ϕ_j 는 시계열 모수임
- 콘테스트에서는 누적 입장 관객수 문제를 시계열 분석으로 처리함

< 평균 모델 >

- 평균 모델은 이번 콘테스트에서 직접 모델링을 수행한 것을 말하며, 그 모델링의 기법이 평균이나 중간값 등 간단한 지표로부터 도출됨
 - 예를 들어, 누적 관객수 예측 문제에서 구단의 요일별, 홈/원정별 관객수의 평균과 가중치를 바탕으로 최종 누적 관객수를 예측한 경우에 평균 모델에 포함
 - 상대적으로 다른 기법들 보다 매우 간단하나, 평균모델을 사용한 팀의 RMSE 값이 고급 기계학습을 사용한 팀보다 낮은 경우도 있었음

2. 챌린지리그 1차 서류심사 통과 참가자 분석

| 팀명 | 예측 방법론 | 방법론 분류 | 세부내용 정리 | RMSE | 순위 |
|------------|---|--------|---|--------|----|
| Challenge1 | Principal Component Analysis, Exploratory Data Analysis, Multi-step Linear Regression | 기타기계학습 | <ul style="list-style-type: none"> o Model <ul style="list-style-type: none"> - Principal Component Analysis : 주성분분석 - Exploratory Data Analysis : 자료분석 - Multi-step Linear Regression : 추정 o Input/Output Parameters <ul style="list-style-type: none"> - 10년간 팀별 투타 성적(53개), 연도/월별 승패(4개) o Development <ul style="list-style-type: none"> - PCA로 주성분분석후 EDA로 자료 선별 - 다중 회귀분석으로 최종 승률 도출 | 0.0091 | 1 |
| Challenge2 | Deep Neural Network | 인공신경망 | <ul style="list-style-type: none"> o Model <ul style="list-style-type: none"> - Deep Neural Network - 앙상블 학습을 통한 신뢰성 제고 o Input/Output Parameters <ul style="list-style-type: none"> - 투타지표, 상대전적, 분위기 등 8개 지표 o Development <ul style="list-style-type: none"> - 웹 크롤링 - R : 딥러닝 패키지 사용 | 0.0106 | 2 |
| Challenge3 | Support Vector Machine | 기타기계학습 | <ul style="list-style-type: none"> o Model <ul style="list-style-type: none"> - Support Vector Machine Time Series o Input/Output Parameters <ul style="list-style-type: none"> - 기대승률 한 가지의 지표 사용 o Development <ul style="list-style-type: none"> - 기대승률을 시계열로 처리하여 hyperplane을 찾음 (예측 값을 다음 예측에 적용) | 0.0130 | 4 |
| Challenge4 | Linear Regression | 회귀분석 | <ul style="list-style-type: none"> o Model <ul style="list-style-type: none"> - Linear Regression - 다중공선성 기법 : 데이터 선별 o Input/Output Parameters <ul style="list-style-type: none"> - 상관관계 분석으로 데이터 선별 - 경기당 실점, OPS로 모델 구축 o Development <ul style="list-style-type: none"> - R을 활용하여 잔여경기 승수 도출 | 0.0131 | 5 |
| Challenge5 | Linear Regression, Regularized Regression, Random Forest, Support Vector Machine | 회귀분석 | <ul style="list-style-type: none"> o Model <ul style="list-style-type: none"> - Lasso (Least Squares) - Ridge (Tikhonov Regularization) - Elastic Net (Regularized Regression) - Random Forest, SVM o Input/Output Parameters <ul style="list-style-type: none"> - 팀(3개), 투수(11개), 타자(16개) 정보사용 - P value로 유효한 데이터 분류 o Development <ul style="list-style-type: none"> - 각 모델별로 가장 좋은 결과를 선택 (Lasso) | 0.0143 | 6 |

| | | | | | |
|------------|--|-------|--|--------|----|
| Challenge6 | Linear Regression, Pythagorean Expectation | 회귀분석 | <ul style="list-style-type: none"> o Model <ul style="list-style-type: none"> - 선형회귀분석, 피타고리안 승률 o Input/Output Parameters <ul style="list-style-type: none"> - 상대전적 승률, 후반기 승률, 피타고리안 승률 o Development <ul style="list-style-type: none"> - 3가지 입력값을 토대로 선형회귀모델 구축 | 0.0147 | 7 |
| Challenge7 | Pythagorean Expectation | 통계 모델 | <ul style="list-style-type: none"> o Model <ul style="list-style-type: none"> - 피타고리안 승률 - 득실점 예측 (회귀분석) o Input/Output Parameters <ul style="list-style-type: none"> - 경기 스코어 | 0.0149 | 8 |
| Challenge8 | Bradley-Terry Model, Pythagorean Expectation | 통계 모델 | <ul style="list-style-type: none"> o Model <ul style="list-style-type: none"> - Bradley-Terry 모델, 피타고리안 승률 o Input/Output Parameters <ul style="list-style-type: none"> - 상대전적 승률 -> 잔여경기 승률 | 0.0151 | 9 |
| Challenge9 | Deep Neural Network (Multi-training) | 인공신경망 | <ul style="list-style-type: none"> o Model <ul style="list-style-type: none"> - Deep Neural Network (Multi-step MLP) o Input/Output Parameters <ul style="list-style-type: none"> - 타자, 투수의 각 지표별상관관계가 분석 - 타자의 경우 ACF를 사용하여 40타석 고려 o Development <ul style="list-style-type: none"> - input : 타자, 투수의 누적데이터 - hidden : 3 layers - output : 승/패 확률 - Python: Library (scikit-learn, Keras) | 0.0156 | 14 |

3. 퓨처스리그 1차 서류심사 통과 참가자 분석

| 팀명 | 예측 방법론 | 방법론 분류 | 세부내용 정리 | RMSE | 순위 |
|----------|---|---------|---|--------|----|
| Futures1 | Linear Regression | 회귀분석 | <ul style="list-style-type: none"> o Input/Output Parameters <ul style="list-style-type: none"> - 요일, 날씨, 메르스 기간 o Development <ul style="list-style-type: none"> - 모수의 적합성 검증 | 14,523 | 1 |
| Futures2 | Linear Regression | 회귀분석 | <ul style="list-style-type: none"> o Input/Output Parameters <ul style="list-style-type: none"> - 일자별 관중현황, 기상정보, 홈원정, 요일 o Development <ul style="list-style-type: none"> - 여러 선형모델을 비교함 | 15,335 | 2 |
| Futures3 | Linear Regression | 회귀분석 | <ul style="list-style-type: none"> o Input/Output Parameters <ul style="list-style-type: none"> - 요일, 홈원정, 관중수, 메르스 고려 o Development <ul style="list-style-type: none"> - R을 사용한 웹 크롤링 | 16,575 | 3 |
| Futures4 | Mean | 통계 모델 | <ul style="list-style-type: none"> o Input/Output Parameters <ul style="list-style-type: none"> - 홈원정, 날씨, 선발, 요일, 분위기 o Development <ul style="list-style-type: none"> - 요소별 가중치를 평균치 등을 사용 | 16,790 | 4 |
| Futures5 | Mean | 통계 모델 | <ul style="list-style-type: none"> o Input/Output Parameters <ul style="list-style-type: none"> - 홈원정, 요일별 관람객수, 대전인기도 배울 o Development <ul style="list-style-type: none"> - 평균을 사용하여 간단한 모델 제시 - 예측 웹페이지 제작 | 16,976 | 5 |
| Futures6 | Restricted Boltzmann Machine, Stacked Autoencoder | 인공신경망 | <ul style="list-style-type: none"> o Input/Output Parameters <ul style="list-style-type: none"> - 요일, 홈원정, 구장, 시기, 날씨 o Development <ul style="list-style-type: none"> - 신경망 학습 시 Validation기법 적용 - RBM의 예측 성능이 더 좋았음 - C++로 직접 알고리즘 구현 | 17,372 | 7 |
| Futures7 | Random Forest | 기타기계 학습 | <ul style="list-style-type: none"> o Input/Output Parameters <ul style="list-style-type: none"> - 요일, 구장, 날씨, 경기외 정보 o Development <ul style="list-style-type: none"> - 웹 크롤링 - 모수별 중요도 평가 | 17,479 | 8 |
| Futures8 | Linear Regression | 회귀분석 | <ul style="list-style-type: none"> o Input/Output Parameters <ul style="list-style-type: none"> - 요일, 홈원정, 구장, 강수량 o Development <ul style="list-style-type: none"> - 선형모델 적용 | 17,533 | 9 |
| Futures9 | Random Forest | 기타기계 학습 | <ul style="list-style-type: none"> o Input/Output Parameters <ul style="list-style-type: none"> - 월, 요일, 홈/원정, 구장, Median 변수 특성 고려 o Development <ul style="list-style-type: none"> - 웹 크롤링 | 17,727 | 10 |

[참고문헌]

1. 국외문헌

Taigman, Y., Yang, M., Ranzato, M. A., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on (pp. 1701-1708). IEEE.

2. 기 타

IBM Watson Homepage. <http://www.ibm.com/smarterplanet/us/en/ibmwatson/>

Kaggle Homepage. <https://www.kaggle.com/>

SK Code Sprint. (2015). <http://codesprint.skplanet.com/2015>

Wikipedia. AR Model. https://en.wikipedia.org/wiki/Autoregressive_model

Wikipedia. Deep Learning. https://en.wikipedia.org/wiki/Deep_learning

Wikipedia. Google Brain. https://en.wikipedia.org/wiki/Google_Brain

Wikipedia. Linear Regression. https://en.wikipedia.org/wiki/Linear_regression

Wikipedia. Pythagorean Expectation. https://en.wikipedia.org/wiki/Pythagorean_expectation

Wikipedia. Random Forest. https://en.wikipedia.org/wiki/Random_forest

주 의

1. 이 보고서는 소프트웨어정책연구소에서 수행한 연구보고서입니다.
2. 이 보고서의 내용을 발표할 때에는 반드시 소프트웨어정책연구소에서 수행한 연구결과임을 밝혀야 합니다.