



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

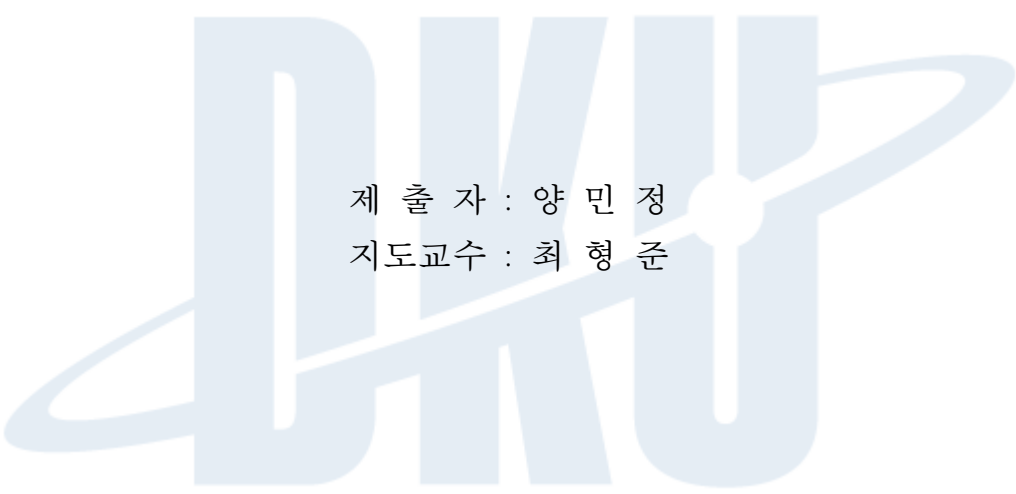
이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

박사학위논문

수영 경기결과 예측을 위한 머신러닝 기법 비교

Comparison of machine learning models for predicting swimming
competition results



제 출 자 : 양 민 정
지도교수 : 최 형 준

2022

체육학과

체육학전공

단국대학교 대학원

수영 경기결과 예측을 위한 머신러닝 기법 비교

Comparison of machine learning models for predicting swimming
competition results

이 논문을 박사학위논문으로 제출함

2022년 12월

단국대학교 대학원






체육학과

체육학전공

양민정

양 민 정의 박사학위 논문을
합격으로 판정함

심 사 일 : 2022. 12. 13.

심사 위원장	권 민 력	
심 사 위 원	이 정 민	
심 사 위 원	이 승 훈	
심 사 위 원	전 용 준	
심 사 위 원	최 형 준	

단국대학교 대학원

(국문초록)

수영 경기결과 예측을 위한 머신러닝 기법 비교

단국대학교 대학원 체육학과

체육학전공

양 민 정

지도교수 : 최 형 준

4차 산업혁명 시대의 주요 정보기술인 인공지능(Artificial Intelligence)을 활용하여 스포츠가 가지는 특성과 종류에 따라 다양한 분석기법의 예측연구가 이루어지고 있다. 또한, 스포츠와 관련된 이용 가능한 데이터의 양이 증가함에 따라 경기결과를 예측하기 위한 지능형 모델개발에 관한 관심이 증가하고 있다. 스포츠경기에서 예측은 경기의 흐름을 제공하고 경기결과에 영향을 미치는 변인을 분석하여 경기운영과 훈련방법을 전략적으로 수립할 수 있게 해주며 경기력 평가 자료를 제시함으로써 미래의 경기력을 향상하는데 필요한 정보를 제공한다. 스포츠 경기결과를 예측하기 위한 머신러닝의 여러 기법 활용은 예측시스템 및 경기결과에 따른 설명력을 보다 구체적으로 파악할 수 있다.

이 연구는 2017년~2021년 전국수영대회 경영 여자 자유형 200m 경기분석 자료(data)를 기반으로 머신러닝을 활용한 경영 경기결과 예측모델을 설계하고, 설계된 예측모델의 성능을 비교·분석하여 경영 경기에 적합한 모델을

제안하는 데 목적이 있다. 또한, 경영 경기결과 예측모델별 경기결과에 영향을 미치는 경기력 변인을 도출하고자 한다. 이 연구의 목적을 달성하기 위해 머신러닝 예측기법인 선형 회귀(Linear Regression), 라쏘 회귀(Lasso Regression), 릿지 회귀(Ridge Regression), 엘라스틱 넷 회귀(Elastic Net Regression), 랜덤 포레스트(Random Forest), 그래디언트 부스팅 머신(GBM: Gradient Boosting Machine), 인공신경망(DNN: Deep Neural Network) 모델을 설계하고 모델별로 예측 성능을 평가하였으며, 경영 경기결과 예측모델별 경기력 변인의 상대적 중요도를 확인하였다.

첫째, 머신러닝 예측모델의 예측 성능을 비교한 결과 Lasso Regression 예측모델이 가장 우수한 것으로 나타났으며 다음으로 Elastic Net Regression 예측모델, Linear Regression 예측모델, Ridge Regression 예측모델, DNN 예측모델, Random Forest 예측모델, GBM 예측모델 순으로 예측력이 나타났다. 둘째, Linear Regression 예측모델, Lasso Regression 예측모델, Ridge Regression 예측모델, Elastic Net Regression 예측모델에 입력된 변인 간 상대적 중요도는 스트로크 구간의 기록 변인인 4lap cleanswim record, 3lap cleanswim record, 2lap cleanswim record 변인에서 높은 비율을 나타냈으며 선형회귀분석을 이용한 예측모델에서는 스트로크 구간의 기록 변인이 모델을 예측하는데 높은 비율을 차지하는 것으로 나타났다. Random Forest 예측모델, GBM 예측모델에 입력된 변인 간 상대적 중요도는 3lap cleanswim speed, 3lap cleanswim record, 4lap cleanswim record 변인에서 높은 비율을 나타냈으며 비선형회귀분석을 이용한 예측모델에서는 스트로크 구간의 속도와 기록이 모델을 예측하는데 높게 기여한 것으로 나타났다. DNN 예측모델은 알고리즘의 특성상 예측에 활용되는 변인이 전체적으로 고르게 기여한 것으로 나타났다. 경영 여자 자유형 200m 경기분석기록(data)을 기반으로 경

영 경기결과 예측을 위한 머신러닝 예측기법의 활용이 가능했으며, Lasso Regression 예측모델이 가장 적합한 것으로 나타났다. 또한, 여자 자유형 200m 경기에서 스트로크 구간의 기록은 최종 기록을 예측하는 주요 요인인 것으로 판단된다.

이 연구는 머신러닝 예측기법을 수영 경영 경기결과 예측에 적용했다는 점에서 의미하는 바가 크다. 추후 머신러닝 예측기법의 스포츠데이터 활용 가능성을 높이고 다양한 분석을 가능하게 할 것으로 기대한다.

주제어: 스포츠경기분석, 수영, 경기결과예측, 머신러닝, 기록스포츠, 스포츠 데이터분석과 융복합

목 차

국문초록	i
목 차	iv
표 목 차	vii
그림목차	ix

I. 서론	1
1. 연구의 필요성	1
2. 연구의 목적	5
3. 연구문제	5
4. 용어의 정의	6
II. 이론적 배경	7
1. 스포츠 데이터분석과 융복합	7
2. 경영 경기분석의 필수요소 및 선행연구	8
3. 스포츠경기결과 예측에 관한 선행연구	12
4. 머신러닝(Machine Learning)	16
5. 머신러닝(Machine Learning) 예측모델	18
1) 회귀분석(Regression Analysis)	18
2) 랜덤 포레스트(Random Forest)	22
3) 그래디언트 부스팅 머신(GBM: Gradient Boosting Machine)	24
4) 인공신경망(ANN: Artificial Neural Network)	26

III. 연구방법	30
1. 연구대상	30
2. 연구절차	32
3. 자료수집 및 자료분석	33
4. 자료처리	44
IV. 연구결과	50
1. 기술통계 및 상관관계 분석	51
2. 머신러닝 예측모델별 분석 결과	56
1) 머신러닝 알고리즘별 최적 예측모델 결정	56
(1) 라쏘 회귀(Lasso Regression) 예측모델	56
(2) 릿지 회귀(Ridge Regression) 예측모델	57
(3) 엘라스틱 넷 회귀(Elastic Net Regression) 예측모델	58
(4) 랜덤 포레스트(Random Forest) 예측모델	59
(5) 그래디언트 부스팅 머신(GBM: Gradient Boosting Machine) 예측모델	61
(6) 인공신경망(Deep Neural Network) 예측모델	63
2) 모델별 예측력 비교	65
3. 경영 경기결과 예측모델 간 경기력 변인의 중요도 분석 결과	74
V. 논의	79
VI. 결론 및 제언	87

참고문헌	89
부 록	95
Abstract	130



표 목 차

〈표 1〉 연구대상의 선정 범위	31
〈표 2〉 연구대상의 특성	31
〈표 3〉 측정 변인의 정의	33
〈표 4〉 경기력 측정 변인	37
〈표 5〉 200m 구간 설정	39
〈표 6〉 구간별 기록 변인의 급내 상관계수	42
〈표 7〉 경영 경기결과 예측을 위한 머신러닝 모델	44
〈표 8〉 하이퍼 파라미터(Hyper parameter)의 범위	46
〈표 9〉 측정 변인별 기술통계	51
〈표 10〉 라쏘 회귀(Lasso Regression) 적합 결과	56
〈표 11〉 릿지 회귀(Ridge Regression) 적합 결과	57
〈표 12〉 엘라스틱 넷 회귀(Elastic Net Regression) 적합 결과	58
〈표 13〉 랜덤 포레스트(Random Forest) 적합 결과	59
〈표 14〉 그래디언트 부스팅 머신(GBM) 적합 결과	61
〈표 15〉 인공신경망(Deep Neural Network) 적합 결과	63
〈표 16〉 머신러닝 모델 간 예측 성능 결과 비교	65
〈표 17〉 선형 회귀(Linear Regression) 모델 예측 결과(단위: 초)	67
〈표 18〉 라쏘 회귀(Lasso Regression) 모델 예측 결과(단위: 초)	68
〈표 19〉 릿지 회귀(Ridge Regression) 모델 예측 결과(단위: 초)	69
〈표 20〉 엘라스틱 넷 회귀(Elastic Net Regression) 모델 예측 결과(단위: 초) 70	
〈표 21〉 랜덤 포레스트(Random Forest) 모델 예측 결과(단위: 초)	71
〈표 22〉 그래디언트 부스팅 머신(GBM) 모델 예측 결과(단위: 초)	72

〈표 23〉 인공지능망(DNN) 모델 예측 결과(단위: 초)	73
〈표 24〉 예측모델의 경기력 변인 간 상대적 중요도 비율	74



그림목차

〈그림 1〉 랜덤 포레스트(Random Forest) 모형	23
〈그림 2〉 그래디언트 부스팅(Gradient Boosting Machine) 모형	25
〈그림 3〉 인공신경망(ANN) 모형	27
〈그림 4〉 역전파 알고리즘 구조	29
〈그림 5〉 연구절차	32
〈그림 6〉 Dartfish Pro S의 태깅 설계	41
〈그림 7〉 Microsoft Excel을 이용해 정리된 분석자료	43
〈그림 8〉 k-fold 교차 검증(k-fold Cross Validation) 모형	49
〈그림 9〉 0.8 이상의 종속변인과 독립변인 간의 상관관계	54
〈그림 10〉 0.1 이하의 종속변인과 독립변인 간의 상관관계	55
〈그림 11〉 선형 회귀(Linear Regression) 예측모델의 실제값과 예측값 상관관계	67
〈그림 12〉 라쏘 회귀(Lasso Regression) 예측모델의 실제값과 예측값 상관관계	68
〈그림 13〉 릿지 회귀(Ridge Regression) 예측모델의 실제값과 예측값 상관관계	69
〈그림 14〉 엘라스틱 넷 회귀(Elastic Net Regression) 예측모델의 실제값과 예측값 상관관계	70
〈그림 15〉 랜덤 포레스트(Random Forest) 예측모델의 실제값과 예측값 상관관계	71
〈그림 16〉 그래디언트 부스팅 머신(GBM) 예측모델의 실제값과 예측값 상관관계	72
〈그림 17〉 인공신경망(DNN) 예측모델의 실제값과 예측값 상관관계	73
〈그림 18〉 예측모델의 경기력 변인 간 상대적 중요도	76

I. 서 론

1. 연구의 필요성

수영은 물속에서 헤엄치는 활동과 그 기술을 사용하는 스포츠를 포괄하는 넓은 범위의 용어이다. 국제수영연맹(FINA: Fédération Internationale de Natation)에서 공인하는 종목은 경영, 다이빙, 하이 다이빙, 아티스틱 스위밍, 오픈워터, 수구이다. 이 중에서 경영 종목은 일정 거리를 정해진 영법으로 헤엄치는 기록경기로 접영, 배영, 평영, 자유형 영법으로 구성되어 있다. 경영의 세부 경기는 접영, 배영, 평영 50m, 100m, 200m, 자유형 50m, 100m, 200m, 400m, 800m, 1500m, 개인혼영 200m, 400m, 혼계영 400m, 계영 400m, 800m 경기로 구분된다. 2019 광주 세계수영선수권대회, 2020 도쿄 올림픽에선 혼성단체전 경기가 추가되었으며 올림픽에서 육상 종목과 함께 많은 메달을 차지할 수 있는 종목이다. 경기 측면에서 경영은 기록 단축이 궁극적 목적이다. 경기의 성공적인 수행은 형태적, 기술적, 체력적, 전략적, 심리적 요인 등으로부터 영향을 받는다(김성진, 2006; 정진배, 2013).

경영 종목의 경기력은 형태적 요인으로 신장, 체중, 체격 등이 있으며(정진배, 2013), 기술적 요인으로 스타트, 턴, 스트로크, 킥, 호흡 등이 있다. 또한 체력적 요인으로 근력, 지구력, 파워, 스피드, 유연성 등이 있고, 전략적 요인으로 페이스(pace) 조절 및 경기 구성이 있다(정철수 등, 2003). 심리적 요인으로 동기, 의지, 자신감 등이 있다(서보영, 2016). 이처럼 다양한 요인으로부터 영향을 받는 종목인 경우 경기력 향상을 위하여 다차원적으로 변인을 고려하는 것이 필요하다(황준일 등, 2012). 경영에 대한 경기력 향상 관련 연구를 살펴보면 생리적(이희창, 2008; 윤미연, 2015; 지무엽, 2017, 박지희, 2020), 심리적(김민석, 2007; 이은경, 2020), 역학적(임승희, 2018; 이재학, 2020)으로 근거한 경기력 분석 연구가 주를 이루고 있다.

최형준(2009, 2010)의 연구를 따르면 경기력 향상을 위해서는 체계적이고 과학적인 훈련방법이 필요하며 보다 객관적인 근거 기반의 경기분석을 통한 피드백이 필

요하다고 역설하였다. 최형준과 정연성(2010)은 객관적으로 스포츠경기를 관찰하고 분석하여 객관화된 데이터를 지도자와 선수에게 제공하는 것이 경기력 향상에 중요한 역할을 한다고 하였으며, Hughes와 Franks(2004)는 스포츠 지도 과정(sport coaching process)을 관찰, 분석, 자료처리, 피드백의 순환과정으로 표현하면서 객관적인 분석을 강조하였다. 따라서 스포츠 경기분석을 통해 경기에서 일어나는 기술과 경기 운영방법, 스포츠경기분석에서 사용되는 기록, 분석기법의 정립 및 효율적인 적용방법 개발 등에 관한 연구가 필요하고 이러한 연구를 바탕으로 우수한 선수가 양성될 수 있도록 해야 한다(김주학, 최형준, 2014).

스포츠경기분석에 관한 선행연구는 수영(윤석훈, 1996; 정철수 등, 2003; 양민정, 2018; 이재승 등, 2019; Cossor, Mason, 2001; Robertson, Hopkins, Anson, 2009; Veiga, Roig, 2016; Simbaña-Escobar 등, 2018; Marinho 등, 2020; Da Silva 등, 2020; Morais 등 2021), 테니스(박해용, 이기청, 2001; 이기봉, 이영석, 이기청, 2004; 김혜진, 박재현, 강상조, 2006; 최형준, 조은형, 김웅준, 한도령, 2011), 농구(박제영, 2003; 김세형, 강상조, 박재현, 김혜진, 2008; 김세중, 허종관, 이강웅, 2010), 축구(최형준, 2009; 임용혁, 임병규, 2006; 이용수, 김용래, 2018), 배구(천영진, 2009, 2019; 황규영, 이종경, 신영철, 2009; 조민정, 박인구, 천영진, 2020), 태권도(최완용, 홍성진, 최형준, 2009; 정광채, 이재봉, 박재현, 2010; 김완수, 양대승, 2018) 등 다양한 종목에서 스포츠경기의 특성에 따라 경기내용을 기록하고, 객관적이고 과학적인 방법에 따라 경기기록을 분석하여 구체적인 경기력 정보를 제공하고 있다. 또한, 4차 산업혁명 시대의 주요 정보기술인 인공지능(Artificial Intelligence)을 활용하여 스포츠경기를 과학적으로 분석하고 평가하기 위한 많은 연구가 이루어지고 있다(최형준, 이운수, 2019). 최형준과 김주학(2006, 2009)은 인공신경망(Artificial Neural Network)을 이용하여 영국 Wimbleton 테니스 대회의 경기결과를 예측하고자 하였으며 테니스 경기결과 예측 시뮬레이터 설계를 위한 기초연구를 하였다. 김주학 등(2007)은 신경망 분석을 이용한 축구경기의 승·패 예측모형을 개발하고자 하였으며 오운학 등(2014)은 KBO 경기데이터를 활용하여 경기의 승패를 예측하기 위해 의사결정나무, 랜덤 포레스트, 신경망 모형, 서포트 벡터 머신, 로지스틱 회귀분석, 선형판별분석, 이차판별분석을

이용하였다. 최혜민 등(2015)의 연구에서는 선형회귀모델과 랜덤 포레스트 모델을 이용하여 경마 경기의 우승 마를 예측하였으며 임정은 등(2017)은 선형회귀, 라쏘 회귀, 릿지 회귀 의사결정나무, 배깅, 랜덤 포레스트, 그래디언트 부스팅, 주성분회귀, K-최근접이웃, 인공신경망 모델을 이용하여 PGA 투어에 출전하는 프로 골프 선수의 경기결과를 예측하는 모델을 제안하였다. 이처럼 해당 스포츠가 가지는 특성과 종류에 따라 다양한 분석기법의 예측연구가 이루어지고 있으며 스포츠와 관련된 이용 가능한 데이터의 양이 증가함에 따라 경기결과를 예측하기 위한 지능형 모델 개발에 관한 관심이 증가하고 있다(Rory, Fadi, 2017).

인공지능 모델을 활용하여 경영 경기결과를 예측한 선행연구를 보면 Edelmann-Nusser 등(2002)은 2000년 시드니 올림픽 200m 여자 배영 결승전 경기결과를 예측하기 위해 인공신경망 모델을 설계하고 선형회귀모델과 성능을 비교하였으며 Maszczyk 등(2012)은 다중회귀분석과 인공신경망 모델을 사용하여 자유형 50m, 자유형 800m 경기의 결과를 예측하였다. Jiang Xie 등(2017)은 수영 경기력을 분류하기 위해서 K-최근접이웃, 서포트 벡터 머신, AdaBoost, 나이브 베이즈 분류, LAD 모델을 사용하였고 기록 예측을 위해서는 이차다항식 회귀, 인공신경망, 서포트 벡터 회귀 모델을 사용하였다. 그러나 기존의 경영 경기결과 예측에 관한 선행연구는 생체 역학 및 인체 측정 요인에 초점을 두고 선수 간의 차이를 밝히려는 연구에 국한되어 있다. 경영 경기는 일정한 경기장의 환경 조건과 각 경기에 할당된 구간(Lap)이 정해져 있어서 구간별 경기력을 관리하는 능력과 페이스(pace) 전략이 경기결과에 영향을 미치기 때문에 기존의 연구방법을 이용한 경영 경기분석은 다차원적으로 경기력을 분석하는데 제한이 있다(Abbiss, Laursen, 2008).

스포츠경기에서 예측은 경기의 흐름을 제공하고 경기결과에 영향을 미치는 변인을 분석하여 경기운영과 훈련방법을 전략적으로 수립할 수 있게 해준다. 또한 경기력 평가 자료를 제시함으로써 미래의 경기력을 향상하는데 필요한 정보를 제공한다(김주학 등, 2007; 김지웅, 2020). 오미애 등(2017)은 최적의 예측모델을 제시하기 위해서는 여러 모델을 비교하여 분석하는 것이 바람직하다고 하였다.

이에 본 연구는 경영 경기결과 예측모델을 설계하여 최적의 경영 경기결과 예측

모델을 제안하고, 예측모델별 경기결과에 영향을 미치는 중요 경기력 변인을 도출하고자 한다. 머신러닝의 여러 기법 활용은 예측시스템 및 경기결과에 따른 설명력을 보다 구체적으로 파악할 수 있다(김주학 등, 2007). 따라서 이 연구에서는 머신러닝 기반 예측기법 모델인 선형 회귀(Linear Regression), 라쏘 회귀(Lasso Regression), 릿지 회귀(Ridge Regression), 엘라스틱 넷 회귀(Elastic Net Regression), 랜덤 포레스트(Random Forest), 그래디언트 부스팅 머신(GBM: Gradient Boosting Machine), 인공신경망(Artificial Neural Network) 모델을 설계하고 예측모델의 성능을 비교·분석하여 경영 경기예에 적합한 모델을 탐색하고자 한다. 이를 통해 머신러닝을 이용한 경영 경기결과 예측의 활용 가능성을 확인하고, 경영 경기력 향상을 위한 새로운 가치를 도출하는 데 활용될 것이다.

2. 연구목적

이 연구는 2017년~2021년 전국수영대회 경영 여자 자유형 200m 경기분석자료(data)를 기반으로 머신러닝을 활용한 경영 경기결과 예측모델을 설계하고, 설계된 예측모델의 성능을 비교·분석하여 경영 경기에 적합한 모델을 제안하는 데 목적이 있다. 또한, 경영 경기결과 예측모델별 경기결과에 영향을 미치는 경기력 변인을 도출하고자 한다.

3. 연구문제

이 연구의 목적을 달성하기 위하여 다음과 같은 연구문제를 설정하였다.

문제 1. 경영 경기결과 예측모델 간 예측 성능에 차이가 있는가?

문제 2. 경영 경기결과 예측모델 간 경기력 변인의 중요도에 차이가 있는가?

4. 용어의 정리

1) 스트로크 수(frequency of stroke): 물속에서 물을 끌어당기는 팔 동작을 스트로크라고 한다. 스트로크 수는 손이 시작하는 순간부터 입수 지점을 1회 스트로크라 하며 레이스를 하는 동안 스트로크 사이클의 수를 의미한다.

2) 스트로크 거리(distance of stroke): 구간별 1회 스트로크에 전진한 평균 거리를 의미한다. (스트로크 거리 = 단위 거리/스트로크 횟수)

3) 스트로크 시간(time duration of stroke): 구간별 1회 스트로크에 걸린 평균 시간을 의미한다. (스트로크 시간 = 구간별 기록/스트로크 횟수)

4) 호흡수(frequency of breath): 머리를 회전하여 호흡하며 레이스를 하는 동안 호흡한 빈도를 의미한다.

5) 속도(speed): 구간별 평균 레이스 속도를 의미한다. (속도 = 단위 거리/구간별 기록)

6) 브레이크아웃(break-out): 스타트, 턴 후 물속에서 15m 이내로 나아가는 동작이다.

7) 랩(lap): 레인을 한번 왕복하는 것을 말한다. (50m=1lap, 100m=2lap, 150m=3lap, 200m=4lap)

II. 이론적 배경

1. 스포츠 데이터분석과 융복합

2016년 세계경제포럼(World Economic Forum, WEF)에서 4차 산업혁명이 소개된 이후 다양한 분야에서 융·복합 사례가 증가하고 있다(한남희, 양도업, 최세희, 2020). 4차 산업혁명은 초연결(Hyper-connected), 초지능(Super-intelligence), 융합(Convergence)을 위해 사물인터넷(IoT: Internet of Things)을 기반으로 데이터가 생성되며 생성된 데이터는 인공지능(AI: Artificial Intelligence)을 통해 해석된다(정현학 등, 2016; 박성건, 황영찬, 2017). 4차산업의 대표적인 기술은 사물인터넷(IoT), 빅데이터(Big Data), 인공지능(AI), 클라우드(Cloud) 등이 있으며 온라인과 오프라인을 결합하는 새로운 사업 모델이 개발되고 있다. 그중에서 빅데이터는 4차 산업혁명이 소개된 이후 다양한 분야에서 중요한 역할을 하고 있으며, 특히 정보통신기술(ICT) 분야에서는 데이터의 활용과 데이터 기반 기술의 중요성이 강조되고 있다. 정보통신기술(ICT)의 발전과 다양한 분야에서 생산되는 데이터는 각 분야의 산업 진흥을 위한 중요한 요소이다. 데이터의 가치는 데이터 수집, 데이터 분석, 데이터 기반 의사결정 등을 통하여 그 가치를 생산한다(남기연, 정현, 2019). 빅데이터는 막대한 데이터의 양을 수집하고, 이를 저장하기 위한 플랫폼, 분석기법 등을 포괄하며 이로부터 새로운 가치를 창출하는 것을 의미한다(북경수, 유재수, 2017).

스포츠 분야에서 빅데이터는 다양한 스포츠의 경기나 훈련에서 나타나는 사건을 분석하는 경기력 분석(performance analysis)을 중심으로 적용되었다. 축구, 필드하키, 럭비 등의 스포츠에서 GPS가 장착된 웨어러블 디바이스와 다각도에서 촬영된 영상으로부터 수집된 정보(이동 거리, 속도, 활동분포, 패스 패턴 등)를 분석하여 경기 전략을 수립하고, 선수교체, 선수들의 컨디션, 부상 관리에 빅데이터 분석 기술을 활용하였다(박성건, 황영찬, 2017). 테니스는 윌블던 테니스 대회에서 2012년부터 IBM과 협력하여 경기 중 나타나는 데이터를 이용하여 실시간 경기분석을 하고 있으며 축적된 데이터를 비교·분석하여 경기결과를 예측하고 있다. 이러한 정보는 선

수, 지도자, 스태프들이 경기력, 전술, 전략 등을 수립하고 예측하는 데 중요하다. 또한, 경기 관객에게 경기 정보와 분석된 정보를 제공하면서 경기의 흥미를 유발한다(남기현, 정현, 2019).

이처럼 기술 발전의 결과로 많은 스포츠에서 사용할 수 있는 데이터가 풍부해지면서 데이터 기반 모델은 스포츠 과학에서 관심을 받고 있다. 데이터의 활용은 경기기록과 트레이닝, 부상 등 경기력 향상을 위해 선수뿐만 아니라 지도자, 스태프 등 다양하게 활용되며 스포츠산업 영역에 새로운 가치를 창출하고 있다.

2. 경영 경기분석의 필수요소 및 선행연구

경영 경기분석은 다양한 전략을 모색할 수 있는 중요한 요소이며 경기 영상을 통해서 경기 전반을 구성하는 경기력을 분석할 수 있다. 경영 경기에서 나타나는 경기력을 평가하기 위해서는 성별, 영법, 세부 경기에 따라 구간을 구분하여 기록, 스트로크 수, 호흡수, 스트로크 거리, 스트로크 시간, 속도 등을 분석해야 한다(Haljand, Saagpakk, 1994; Da Silva 등, 2020). 이를 통해 경기 전략, 기술, 체력 등을 파악하고 비교할 수 있다.

윤석훈(1996)은 경영 100m 경기 시 구간별 운동학적 분석을 하기 위해 출발 구간, 역영 구간, 턴 구간, 종료 구간으로 나누어 결승 진출자와 예선 탈락자의 시간 변인과 속도 변인을 비교 분석하였다. 출발 구간은 경기 출발 신호로부터 선수가 발차기나 스트로크를 시작하는 순간으로 여자 접영경기를 제외하고 결승 진출자와 예선 탈락자 간 통계적으로 유의한 차이가 나타났다고 보고하였다. 역영 구간은 출발 구간이 끝나는 지점부터 방향전환 구간이 시작되는 지점까지의 약 27.7m~32.5m 거리의 구간으로 구분하였으며 남자 평영 경기에서만 통계적 유의한 차이가 나타났다고 제시하였다. 방향전환 구간은 턴 전 7.5m와 턴 후 7.5m의 약 15m로 구분하였으며 모든 경기에서 통계적으로 유의하게 나타났다. 마지막으로 종료 구간은 경기의 마지막 7m~10m 구간으로 나누었으며 남자 자유형, 남자 접영, 여자 자유형, 여자 평영 경기에서 통계적 유의한 차이가 나타났다고 보고하였다. 결론적으로 전체 기록

차이의 비율보다 출발 구간, 턴 구간에서 큰 비율 차이가 나타나며 예선타락 선수는 출발과 턴 기술 훈련에 더 많은 시간을 사용해야 한다고 제안하였다.

정철수 등(2003)은 50m, 100m, 200m 자유형 경기를 출발 구간(start phase: 10m), 스트로크 구간(clean swim phase: 25m 지점 전후의 20m), 턴 구간(turn phase: 반환대 전후의 각 5m), 종료 구간(finish phase: 도착 전 10m)으로 구분하여 기록, 속도, 스트로크 길이, 스트로크 빈도 변인을 분석하였다. 연구결과 모든 경기의 기록 변인에서 유의한 차이가 나타났으며, 출발, 턴, 종료 구간에서는 남자선수가 여자선수보다 속도가 크게 나타났다고 보고하였다. 또한, 스트로크 구간의 평균 스트로크 속도는 모든 경기에서 유의한 차이를 보였고, 평균 스트로크 길이는 경기거리가 증가함에 따라 증가하는 경향을 보였다고 제시하였다. 평균 스트로크 빈도는 경기거리가 증가함에 따라 감소하는 경향이 나타났다. 이 연구는 경영 경기력과 이를 구성하는 세부 변인을 객관적으로 평가할 수 있는 자료이며 전체적인 경기분석과 함께 개개인에 대한 심층분석도 이루어져야 한다는 결론을 도출하였다.

양민정(2018)은 경영 영법별 결승전 진출자를 2그룹(A그룹: 상위 4명, B그룹: 하위 4명)으로 나누어 100m 13구간(start, 0m~5m, 5m~15m, 15m~25m, 25m~35m, 35m~45m, 45m~50m, 50m~55m, 55m~65m, 65m~75m, 75m~85m, 85m~95m, 95m~100m), 200m 25구간(start, 0m~5m, 5m~15m, 15m~25m, 25m~35m, 35m~45m, 45m~50m, 50m~55m, 55m~65m, 65m~75m, 75m~85m, 85m~95m, 95m~100m, 100m~105m, 105m~115m, 115m~125m, 125m~135m, 135m~145m, 145m~150m, 150m~155m, 155m~165m, 165m~175m, 175m~185m, 185m~195m, 195m~200m)의 시간 변인(구간별 기록)과 행동 변인(구간별 스트로크 수, 구간별 호흡수)을 비교 분석하였다. 이 연구에서는 구간별 기록, 구간별 스트로크 수, 구간별 호흡수는 A그룹과 B그룹 간의 경기력 차이를 나타내는 요인으로 나타났으며 경영 경기의 경기력 향상을 위해 더 많은 경기와 연구대상을 확대하여 경기분석이 이루어져야 한다고 보고하였다.

이제승 등(2019)은 한국과 미국 수영선수의 접영경기를 분석하였으며 경기분석을 위해 제99회 전국체육대회와 2018년 Speedo Junior Championships의 접영 200m 경

기 영상을 수집하였다. 총 20명을 대상으로 진행하였으며 10m 기준으로 24개의 구간(출발 구간(0m~15m), 경쟁 구간(15m~25m, 25m~35m, 35m~45m, 65m~75m, 75m~85m, 85m~95m, 115m~125m, 125m~135m, 135m~145m, 165m~175m, 175m~185m, 185m~195m), 턴 구간(45m~50m, 50m~55m, 55m~65m, 95m~100m, 100m~105m, 105m~115m, 145m~150m, 150m~155m, 155m~165m), 대시 구간(185m~195m, 195m~200m))을 설정하였다. 측정 변인은 구간별 기록, 브레이크아웃(break-out) 거리, 스트로크 빈도, 호흡 빈도이며 측정 변인의 차이를 비교하였다. 결론적으로 기록, 브레이크아웃(break-out) 거리, 스트로크 빈도, 호흡 빈도에서 구간별로 한국과 미국 선수의 경기력 차이가 나타났다. 이 연구에서는 경영 경기력 향상을 위해 한국선수만의 훈련법이 다각적으로 연구되어야 하며 수영 종목의 데이터 분석이 활발히 이루어져야 한다고 제언하였다.

Robertson 등(2009)은 9개의 국제 수영 경기에서 상위 16명 결승 진출자의 랩타임을 분석하였다. 영법별(접영, 배영, 평영 100m, 200m, 자유형 100m, 200m, 400m, 개인혼영 200m, 400m) 각 구간의 랩타임과 최종 기록의 관계를 평가했다. 결론적으로 100m 경기의 마지막 랩과 200m~400m 경기의 중간 2개 랩(2lap, 3lap)이 최종 기록과 가장 밀접한 관계가 있다고 하였으며 각 구간의 랩타임을 개선하면 수영 성능이 크게 향상될 수 있다고 하였다. 그러나 가장 빠른 구간과 느린 구간의 랩타임 패턴의 유사성은 전체 속도에 방해하지 않고 랩타임을 개선해야 한다고 시사하였다.

Morais 등(2019)은 2016 유럽 선수권 대회 접영, 배영, 평영, 자유형 100m 남녀 결승 진출자의 출발(start)과 턴(turn) 특성을 조사하였다. 출발(start)은 출발 위치부터 15m 구간으로 구분하였으며 15m 시간(출발 신호부터 15m 표시에 도달하는 시간), 반응 시간, 진입 시간(시작 신호와 손이 물에 들어가는 순간 사이의 시간), 비행시간(발가락이 블록을 떠나는 순간과 손이 물에 들어가는 순간 사이의 시간), 진입 거리(출발 위치부터 손이 물에 들어가는 곳 사이의 거리), 잠영 시간(물에 들어간 순간부터 머리가 수면 위로 나온 시간), 잠영 거리(손이 물속으로 들어가는 지점과 머리가 수면 위로 나온 지점 사이의 거리), 브레이크아웃(break-out) 시간(시작 신호와 수면 위로 나온 머리 사이의 시간), 브레이크아웃(break-out) 거리(출발 위치와 수면

위로 나온 머리 사이의 거리)를 측정 변인으로 하였다. 턴(turn)은 턴 전(turn-in) 5m 부터 턴 후(turn-out) 15m로 구분하였고, 측정 변인은 전체 턴 시간(45m 표시부터 턴 후 15m 표시에 도달하는 시간), 턴 5m 전(turn-in) 시간(머리가 45m 표시에 도달 하고부터 발이 벽에 닿는 시간), 브레이크아웃 거리(벽과 수면 위로 나온 머리 사이의 거리), 브레이크아웃 시간(벽을 만지고부터 머리가 수면 위로 나온 시간), 턴 후 15m(turn-out) 시간(벽을 만지고부터 15m 표시에 도달하는 시간)을 분석하였다. 결론적으로 출발(start) 구간에서 남자 수영선수는 접영 영법이 가장 빠르게 나타났고 여자 수영선수는 자유형 영법이 가장 빠르게 나타났다. 턴(turn) 구간에서는 남자, 여자 수영선수 모두 자유형 영법에서 가장 빠르게 나타났으며 이 연구에서는 출발 (start) 구간과 턴(turn) 구간이 경기에서 총 레이스(race) 시간의 1/3을 차지하며 경기력에 높은 관련성이 있다고 보고하였다.

Da Silva 등(2020)은 2008 베이징올림픽, 2012 런던올림픽, 2016 리우올림픽 자유형 100m, 200m 400m 결승경기에서 109명 수영선수(남자 57명, 여자 52명)의 반응 시간, 구간 시간, 속도 및 최종 시간을 검증하고 비교하였다. 결론적으로 3번의 올림픽 동안 선수의 구간별 기록은 크게 차이가 없었지만 분석된 변인 중 하나인 반응 시간은 세 번의 올림픽 동안 감소하면서 단거리 및 장거리 경기에서 가능한 결정요인으로 나타났다. 이 연구는 생체 역학, 인체 측정 및 생리학적 요인에서 설명하고 확장할 수 있으며 향후 수영선수들의 경기력을 이해하기 위해 다양한 변인과 영법을 이용한 연구를 제안하였다.

Morais 등(2021)은 200m 남자 주니어 수영선수 76명의 레이스 시간, 출발 구간, 스트로크 구간, 턴 구간 및 도착 구간 변인 사이의 랩(lap) 간 안정성과 관계에 대해 평가하였다. 이 연구에서 사용된 출발 구간(출발 위치부터 15m)의 측정 변인은 반응 시간, 비행시간, 진입 시간, 진입 거리, 잠영 시간, 잠영 거리, 잠영 속도(수면 진입부터 수면 위로 나온 머리 사이의 속도), 브레이크아웃 시간, 브레이크아웃 거리, 15m 시간이며 턴 구간(턴 전 5m~턴 후 15m)의 측정 변인은 턴 전 5m 시간, 브레이크아웃 시간, 브레이크아웃 거리, 잠영 속도(벽을 만지고부터 수면 위로 나온 머리 사이의 속도), 턴 후 15m 시간, 전체 턴 시간이다. 스트로크 구간(수영장 중간

30m)의 측정 변인은 30m 구간의 레이스 속도(m/s), 스트로크 시간(Hz), 스트로크 길이(m), 스트로크 인덱스(m^2/s)이며 도착 구간(마지막 5m)의 측정 변인은 도착 전 5m 시간, 도착 전 5m 속도이다. 결론적으로 랩(lap) 간 차이를 보였으며 첫 스트로크 구간과 턴 구간에서도 많은 차이를 보였다. 전반적으로 스트로크 구간, 턴 구간, 출발 구간 및 도착 구간은 자유형 200m 경기의 최종 기록과 높은 상관관계를 나타냈으며 이는 수영선수들이 스트로크 구간뿐만 아니라 다른 구간에서도 경기력을 개선하기 위해 훈련해야 한다고 제안하였다.

3. 스포츠경기결과 예측에 관한 선행연구

김혜진 등(2006)의 연구는 테니스 경기결과 자료를 이용하여 득점과 실점에 따른 경기의 승패 요인을 분석하기 위해 머신러닝 기법의 하나인 의사결정나무 기법을 적용하였다. 이 연구에서 사용된 의사결정나무는 CHAID 알고리즘을 이용하였으며 분석 규칙으로 Maximum Tree Depth는 3으로 설정하였고 최소 사례 수는 parent node에서 5, child node에서 2로 규정하였다. CHAID 방법에서 Splitting과 Merging의 Alpha는 0.05로 설정하였다. 위 구조에서 목표 변인이 맨 위에 있게 되고 각 예측 변인이 계층적으로 위치하는데 예측 변인 중 가장 위쪽에 있는 변인이 가장 영향력(관련성)이 높은 변인으로 승패를 결정짓는 중요한 변인은 리시브 상황에서의 득점으로 나타났다. 의사결정나무의 분석결과를 node 별로 요약하면 승자를 설명하는 가장 이상적인 node는 5번이며, 패자를 가장 잘 설명하는 node는 3번으로 나타났다. 승자가 되기 위해서는 리시브 게임에서 승리하는 것이 무엇보다 중요하며, 하지 않아도 될 자기 실수를 줄여야 한다는 결과를 도출하였다.

최형준과 김주학(2006)은 인공신경망(Artificial Neural Network)을 이용하여 영국 웬블던 테니스 대회 경기결과를 예측하고자 하였다. 적용된 인공신경망은 다층으로 구성되었으며 역전파 학습 과정을 거쳐 출력값을 계산하였다. 결론적으로 14-8-2 신경망과 14-10-2 신경망에서 기존의 예측기법들에 비교해 높은 적중률을 나타냈다. 이 연구에서는 머신러닝 기법 중 인공신경망을 통해 경기결과 예측이 가

능하였지만 정확한 예측을 위해서는 예측의 적중률을 높일 수 있는 요소들을 밝혀내는 것이 중요하다고 제안하였다.

김주학 등(2007)은 신경망 분석을 이용한 축구경기의 승·패 예측모델을 개발하고자 하였다. 승·패 예측모델은 입력층에 32개의 기록요인을 사용했고, 은닉층은 3개, 출력층은 승리, 패배의 두 집단으로 분류하여 예측하였으며 87.5%의 예측률을 나타냈다. 이 연구에서는 경기결과와 승패를 예측하기 위해 승패 관련 기록요소에 대한 설명력, 경기기록의 DB화, 과거 기록의 중요성을 인식해야 한다는 결론을 도출하였다. 또한, 체육 분야에서 머신러닝의 여러 통계기법을 활용한다면 예측시스템 및 경기결과에 따른 설명력과 관련하여 구체적인 내용을 파악할 수 있다고 보고하였다.

오윤학 등(2014)은 KBO 경기데이터를 활용하여 경기의 승패를 예측하기 위해 의사결정나무, 랜덤 포레스트, 신경망 모형, 서포트 벡터 머신, 로지스틱 회귀분석, 선형판별분석, 이차판별분석을 이용하였으며 랜덤 포레스트 분석기법이 가장 우수한 정확도를 나타냈다. 이 연구는 기존의 스포츠분석에서 사용되지 않았던 랜덤 포레스트 기법을 사용하여 우수한 예측력을 나타냈으며 선발투수의 성적이 타자의 성적보다 상대적으로 중요하다는 결과를 도출하였다.

임정은 등(2017)은 선형 회귀, 라쏘 회귀, 릿지 회귀, 의사결정나무, 배깅, 랜덤 포레스트, 그래디언트 부스팅, 주성분 회귀, K-최근접이웃, 인공신경망 모델을 이용하여 PGA 투어에 출전하는 프로 골프 선수의 경기결과를 예측하는 모델을 제안하였다. 선수, 코스, 바람에 대한 측정 변인을 가지고 최종 스코어를 예측하였으며 랜덤 포레스트 모델에서 좋은 예측력을 나타냈다. 추가로 4대 플레이오프 경기에 랜덤 포레스트 모델을 적용하였으며 예측 스코어를 기반으로 상위권 선수의 순위를 50% 이상 예측하며 이 연구의 예측모델은 골프 경기예측에 적합성을 나타내는 것을 확인하였다.

한정섭 등(2022)은 선형 회귀, 라쏘 회귀, 릿지 회귀, XGBoost, LightGBM, 랜덤 포레스트, 서포트 벡터 회귀 모델을 이용하여 KBO 타자의 OPS(On-base Plus Slugging)를 예측하였으며 XGBoost 예측모델이 최적의 OPS 예측 성능을 나타냈다.

이 연구에서 개발된 OPS 예측모델은 데이터가 부족한 타자의 경우 예측 성능이 떨어져 현장에 직접 적용하기에는 어려움이 있으며 향후 연구에서 정확한 데이터 전처리와 추가 분석을 통해 데이터 세트의 구축이 필요하다고 보고하였다.

Edelmann-Nusser 등(2002)은 2000년 시드니 올림픽에서 200m 배영 결승전에 출전한 엘리트 여자 수영선수의 경기결과를 예측하기 위해 인공신경망 모델을 설계하고 기존의 선형 회귀모델과 성능을 비교하였다. 데이터는 시드니 올림픽 전 19개 대회 배영 200m 경기결과와 수영선수의 훈련 기간의 데이터로 구성하였다. 인공신경망 모델은 10개의 입력층, 2개의 은닉층, 1개의 출력층이 있는 MLP를 사용했다. 연구 결과는 MLP 모델의 예측 오차는 0.05를 나타내며 선형 회귀모델보다 정확한 결과를 나타냈다. 이 연구는 결과 예측을 강조하기보다 훈련 프로그램개발을 위해 중요한 변인들을 제공하고 수영 종목에서 머신러닝 예측기법의 활용 가능성을 확인하였다.

Maszczyk 등(2012)은 자유형 50m, 800m 경기의 결과를 예측하기 위해 다중회귀분석과 인공신경망 모델을 사용하였다. 측정 변인은 체력(폐활량), 근력(멀리뛰기), 수영 기술(회전, 활주, 스트로크 당 거리), 인체 측정 변인(손, 발 크기), 특정 거리에서의 속도가 사용되었다. 연구결과 8-4-1 또는 4-3-1로 구성된 다층 퍼셉트론 신경망(MLP) 모델이 초기 훈련, 근력, 체력, 신체 측정에서 최종 경기결과를 다중회귀분석의 모델보다 정확하게 예측하는 것으로 나타났다. 이 연구는 신경망 모델이 수영의 특정 스타일과 거리에 따라 선수를 모집하는 과정뿐만 아니라 미래의 성과를 예측하는 데 사용될 수 있다고 보고하였다.

Maszczyk 등(2014)의 연구는 회귀모델과 인공신경망 모델을 비교하여 창 던지기의 거리를 예측하였다. 자료의 구성은 40개의 훈련 세트, 15개의 검증 세트, 15개의 테스트 세트로 구성하였으며 평균 제곱근 오차가 낮게 나타난 4-3-1 인공신경망 모델을 예측모델로 사용하였다. 연구결과 인공신경망 모델이 회귀모델보다 높은 예측력을 나타냈으며 선수의 경기력을 최적화하기 위해 변인의 정보를 제공하고 개인종목의 스포츠에서 인공신경망의 적합성을 확인하였다.

Jiang Xie 등(2017)은 수영 경기결과를 예측하기 위해 회귀분석과 머신러닝 기법

을 이용하였다. 수영 경기력을 분류하기 위해서는 K-최근접이웃, 서포트 벡터 머신, AdaBoost, 나이브 베이즈 분류, LAD 모델을 사용하였고, 시간 예측을 위해서는 이차다항식 회귀, 인공신경망, 서포트 벡터 회귀 모델을 사용하였다. 각 모델을 분석하여 머신러닝이 수영 최종 시간을 예측할 수 있었으며 각 모델은 영법에 따라 예측 정확도가 달라진다는 결론을 도출하였다. 또한 앙상블 모델과 각 모델의 결과를 비교하였으며 앙상블 모델이 가장 안정적이고 광범위하게 적용될 수 있다고 보고하였다.



4. 머신러닝(Machine Learning)

머신러닝(Machine Learning)은 인공지능 범위 안에 포함되며 다수의 파라미터로 구성된 모델을 이용하여 주어진 데이터로 파라미터를 최적화하는 학습을 한다. 데이터마이닝(data mining)과 혼용되기도 하는데 머신러닝은 컴퓨터가 사전에 프로그래밍되어 있지 않고 데이터로부터 패턴을 학습하여 새로운 데이터에 대해 적절한 작업을 수행하는 알고리즘이나 처리 과정을 의미한다. 반면 데이터 마이닝은 주로 사람에게 어떤 지식을 제공하는 것을 목적으로 하고 있다(Domingos, 2012; 오미애 등, 2017). 머신러닝은 학습 문제의 형태에 따라 지도 학습(supervised learning), 비지도 학습(unsupervised learning) 및 강화 학습(reinforcement learning)으로 구분한다.

지도 학습(supervised learning)은 입력 변인과 출력 변인 사이의 매핑(Mapping)을 학습하는 것이며 입력과 출력이 데이터로 주어질 때 적용한다. 지도학습의 목적은 새로운 입력 변인에 대해 출력 변인을 정확하게 예측할 수 있는 사상함수(mapping function)의 근사치를 만드는 것으로 분류(classification)와 예측(prediction)모델이 있다. 비지도 학습(unsupervised learning)은 입력 변인만 있고 출력 변인은 없는 경우에 적용하며 입력 사이의 규칙성 등을 찾아내는 게 목표이다. 비지도 학습에는 군집(clustering)모델이 있으며 데이터 속에 존재하는 구조를 발견하고 제시하는 역할을 한다. 학습의 결과는 지도학습의 입력으로 사용되거나 전문가에 의해 해석된다. 강화 학습(reinforcement learning)은 지도 학습과 달리 주어진 입력에 대한 출력, 즉 정답 행동이 주어지지 않는다. 대신 일련의 행동의 결과에 대해 보상(reward)이 주어지게 되며 시스템은 이러한 보상을 이용해 학습한다. 강화 학습은 알고리즘이 수행한 결과에 따라 수행 방식을 진화시켜 나가며 이러한 시스템의 예로 로봇이나 게임의 플레이어 등을 들 수 있다. 이 연구는 예측이 목적이기 때문에 지도 학습(supervised learning)의 기법을 살펴보려 한다. 모델 구축을 위해 축적된 데이터 세트(training data)는 독립변인(설명변인)과 종속변인(결과변인)으로 구성되며 이 데이터 세트로 머신러닝 알고리즘을 활용하여 예측모델을 만든다. 만들어진 예측모델을 새로운 데이터 세트(test data)에 적용하여 예측값을 추정할 수 있다. 예측모델로 활

용할 수 있는 기법들은 회귀 분석(Regression Analysis), 판별분석(Discriminant Analysis), 의사결정나무(Decision Tree), 랜덤 포레스트(Random Forest), 인공신경망(Artificial Neural Network), 딥러닝(Deep Learning), 앙상블(Ensemble) 등 매우 많다(오미애 등, 2017). 또한, 종속변인(목표변인)이 범주형인 경우와 연속형인 경우로 나뉜다. 범주형(categorical)은 독립변인을 통해 종속변인의 각 범주에 대한 확률을 예측하는 모델을 만드는 것이 목적이며, 예측모델을 통해 새로운 개체를 분류하고자 한다. 종속변인이 범주형일 때 활용할 수 있는 예측모델은 로지스틱 회귀(Logistic Regression), 판별분석(Discriminant Analysis), 의사결정나무(Decision Tree), 인공신경망(Artificial Neural Network), 서포트 벡터 머신(Support Vector Machine), 자기구성 지도(Self-Organizing Map) 등이 있다. 연속형(continuous)은 독립변인을 통해 종속변인의 값을 예측하는 모델을 만드는 것이 목적이며, 종속변인이 연속형일 때 활용할 수 있는 예측모델은 선형 회귀(Linear Regression), 라쏘 회귀(Lasso Regression: Least Absolute Shrinkage and Selection Operator), 릿지 회귀(Ridge Regression), 엘라스틱넷 회귀(ElasticNet Regression), 랜덤 포레스트(Random Forest), 그래디언트 부스팅 머신(GBM: Gradient Boosting Machine), 인공신경망(Artificial Neural Network), 앙상블(Ensemble) 등이 있다. 예측모델의 성능을 평가하는 모델 평가는 예측(prediction)을 위해 만든 모델이 임의의 모델(random model)보다 예측력이 우수한지, 고려된 다른 모델 중 어느 모델이 가장 우수한 예측력을 보유하고 있는지를 비교·분석하는 과정이라고 할 수 있다. 예측모델의 성능을 평가하는 방법으로는 실제값과 예측값의 결과 데이터가 얼마나 정확하고 오류가 적게 발생하는가에 기반을 두며 분류모델은 정확도(Accuracy), 오차행렬(Confusion Matrix), 정밀도(Precision), 재현율(Recall), ROC AUC 등의 평가지표로 모델을 평가한다. 회귀모델의 경우 오차에 절댓값을 씌운 뒤 평균 오차를 구하거나 오차의 제곱 값에 루트를 씌운 뒤 평균 오차를 구하는 방법과 같이 예측 오차를 가지고 성능 평가하는 평균 제곱 오차(MSE: Mean Square Error), 루트 평균 제곱 오차(RMSE: Root Mean Squared Error), 평균 절대 오차(MAE: Mean Absolute Error) 등이 있다(김은하 등, 2015).

5. 머신러닝(Machine Learning) 예측모델

1) 회귀분석(Regression Analysis)

회귀분석(Regression Analysis)이란 어떤 자료에서 값에 영향을 주는 조건(x)에 대해 각 조건에 대한 영향력을 고려하여 해당 조건에서의 y 를 예측하는 방법이다. 회귀모델에서는 예측치와 실측치의 차이인 오차를 최소화하는 방향으로 회귀계수를 추정한다. 선형 회귀분석(Liner Regression Analysis)은 회귀분석 중 가장 단순한 방법론이며 하나의 독립변인과 하나의 종속변인 간의 관계를 설명한다. 다중회귀분석(Multiple Regression Analysis)은 종속변인은 하나이고 독립변인이 2개 이상인 회귀모델에 대한 분석을 수행하는 방법이다. 다중회귀분석의 기본적인 목표는 다음과 같은 다중회귀 식(수식 1)에서 상수와 계수를 구하는 것이며 오차(수식 3)를 최소화하는 방법으로 회귀계수를 추정하여 회귀 식(수식 2)을 찾는다. (x : 독립변인, y : 종속변인, β : 회귀계수(coefficient), β_0 : y 절편, $\beta_1 \sim \beta_n$: 독립변인의 기울기)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n \quad (\text{수식 1})$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \dots + \hat{\beta}_k x_k \quad (\text{수식 2})$$

$$e_i = y_i - \hat{y}_i \quad (\text{수식 3})$$

다중회귀 모델에서는 독립변인끼리 독립 관계일수록 데이터의 잡음이 없는 양질의 데이터를 이용한 학습이 가능하다. 독립변인 간에 낮은 상관관계를 가질수록 모델의 속도가 빠르고, 변인 간의 중복된 설명력에 따른 다중공선성 문제를 완화해 모델의 정확도를 높일 수 있다. 또한, 회귀분석은 종속변인의 변동성(분산)을 독립변인이 얼마나 잘 설명하는가가 중요하다. 다중회귀분석에서는 각각의 독립변인이 종속변인을 설명하는 변동성(분산)이 클수록 좋은 변인이며 독립변인의 변동성(분산)이 크면 변인 간에 낮은 p -value를 가진다. 일반적으로 p -value는 0.05보다 작은 경우 독립변인이 종속변인에 영향을 미치는 것이 유의미하다고 할 수 있다(서유화, 김은희, 2021).

독립변인들 사이에 다중공선성(multicollinearity)이 존재하는 경우 선형 회귀모델(Linear Regression Model)의 최소제곱 추정값(least square estimate)은 분산이 매우 커지게 되어 추정량으로서의 정도가 나쁘게 된다. 이러한 단점을 보완하기 위해 작은 편향(bias)을 허용하여 회귀계수의 크기를 축소함으로써 모델을 안정화하고 분산을 작게 할 수 있는 계수축소(shrinkage method) 방법을 활용할 수 있다(김은령, 2010). 계수축소법인 shrinkage method의 대표적인 모델은 라쏘 회귀(Lasso Regression), 릿지 회귀(Ridge Regression), 엘라스틱 넷 회귀(Elastic Net Regression) 등이 있다.

라쏘 회귀(Lasso Regression)는 Least Absolute Shrinkage and Selection Operator의 약자로 Tibshirani(1996)가 제안한 방법이다. 릿지 회귀(Ridge Regression)와 같이 회귀계수의 크기에 페널티(penalty)를 부여함으로써 회귀계수의 크기를 축소하는 Shrinkage Method의 방법의 하나로 <수식 4>와 같다. 라쏘 회귀의 경우 상대적으로 영향력이 없는 변인의 회귀계수를 0으로 만들어 변인을 제거함으로써 차원이 축소되며 해석력이 뛰어나다(김홍표, 2018).

릿지 회귀(Ridge Regression)는 Hoerl와 Kennard(1971)가 제안하였으며 회귀계수의 크기에 페널티(penalty)를 부여함으로써 <수식 5>와 같이 회귀계수를 축소하는 방법이다. 릿지 회귀의 경우 회귀계수를 변화시킴으로써 중요하지 않은 변인의 회귀계수를 0에 가까운 값으로 줄이게 되어 예측 정확도(prediction accuracy)를 높여주는

모델을 제공한다. 하지만 회귀계수가 완전히 0이 되지는 않아 라쏘 회귀모델보다 해석이 어렵다.

$$\hat{\beta}_{Lasso} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 + a \sum_{j=1}^n |\beta_j| \quad (\text{수식 4})$$

$$\hat{\beta}_{Ridge} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^n \beta_j^2 \quad (\text{수식 5})$$

라쏘 회귀는 제약식 계수에 $L_1 - norm$ 을 적용한 절댓값을, 릿지 회귀는 $L_2 - norm$ 을 적용한 제곱 값을 사용한다. 따라서 라쏘 회귀는 개별 계수를 0으로 축소할 수 있고 0까지 축소된 변인은 모형 적합에서 제외된다(김홍표, 2018).

엘라스틱 넷 회귀(Elastic Net Regression)는 Zou와 Hastie(2005)에 의해 제안되었으며 변인선택과 추정을 동시에 할 수 있는 기법이다. <수식 6>과 같이 라쏘 회귀와 릿지 회귀 추정의 장점들을 포함한 방법으로 상관관계가 있는 변인들을 모두 선택하여 많은 수의 고차원 변인들에 대한 설명이 가능하다(김홍표, 2019). Zou와 Hastie(2005)에 의해 제안된 별점화 기법인 볼록 결합(convex combination) 형태로 계산되며 상관관계가 있는 변인 중에서 하나의 변인만을 선택하는 라쏘 회귀의 단점을 보완할 수 있다.

$$\tilde{\beta} = \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p x_{ij}\beta_j)^2 + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2 \quad (\text{수식 6})$$

회귀모델의 성능 평가는 결정계수(R^2 : R-squared), 수정된 결정계수(adjusted R^2), 평균 제곱 오차(MSE: Mean Square Error), 루트 평균 제곱 오차(RMSE: Root Mean Squared Error), 평균 절대 오차(MAE: Mean Absolute Error)등의 기준이 사용된다. 결정계수(R^2 : R-squared)는 분산 기반으로 예측 성능을 평가하며 독립변인으로 중

속변인을 설명할 수 변동성을 나타내는 지표로 1에 가까울수록 예측 정확도가 높다. 수정된 결정계수(adjusted R^2)는 독립변인의 수가 증가하면 따라서 증가하는 R^2 (결정계수)을 보완한 값으로 1에 가까울수록 예측 정확도가 높다. 평균 제곱 오차 (MSE: Mean Square Error)는 실제값과 예측값의 차이를 제곱한 평균으로 값이 작을수록 좋으며 루트 평균 제곱 오차(RMSE: Root Mean Squared Error)는 MSE가 오차의 제곱의 구할 때 실제 오차 평균보다 커지는 특성이 있어 MSE에 루트를 취한 값을 나타낸다. 평균절대오차(MAE: Mean Absolute Error)는 실제값과 예측값의 차이를 절댓값으로 변화해 평균한 값을 나타낸다(오미애 등, 2017).



2) 랜덤 포레스트(Random Forest)

랜덤 포레스트(Random Forest)는 결정 나무(Decision Tree)를 학습하여 분류 또는 예측을 출력하는 앙상블(Ensemble) 기법을 사용한 머신러닝 기법의 하나다. 랜덤 포레스트의 가장 큰 특징은 랜덤성(randomness)에 의해 트리들이 조금씩 다른 특성을 갖는다. 대표적인 트리 방법론의 알고리즘인 CART(Classification And Regression Trees) 사용하며 트리의 예측(prediction)들이 비상관화(Decorrelation) 되게 한다. 결과적으로 일반화(generalization) 성능을 향상한다. 또한, 랜덤화(randomization)는 포레스트가 노이즈가 포함된 데이터에 대해서도 강인하게 만들어 준다. 랜덤화는 각 트리의 훈련 과정에서 진행되며 랜덤 학습데이터 추출 방법을 이용한 앙상블 학습법인 배깅(bagging)과 랜덤 노드 최적화(randomized node optimization)가 자주 사용된다. 이 두 가지 방법은 서로 동시에 사용되어 랜덤화 특성을 더욱 증진시킬 수 있다. 배깅은 부트스트랩(Bootstrap) 표본 추출 방식으로 얻어진 복수의 표본에 대해 개별적으로 모델링을 하여 결과값을 결합함으로써 모델의 안정성을 높이는 방법이다(James 등, 2013).

$$\hat{f}_{avg}(x) = \frac{1}{N} \sum_{n=1}^N \hat{f}^{*n}(x) \quad (\text{수식 7})$$

<수식 7>에서와 같이 $\hat{f}^{*n}(x)$ 는 부트스트랩을 표본 추출하여 얻어진 n번째 모델 함수의 결과값이다. 결과값이 연속형 변인일 때는 이 값들이 평균 배깅의 결과값이다. 랜덤 포레스트의 주요 매개변수는 포레스트의 트리 수(결정 나무의 수)와 결정 나무의 최대 허용 깊이이다. 너무 적은 결정 나무를 만들 경우, 유의미한 결과 도출이 어려우므로 충분히 많이 결정 나무를 만들어야 한다. 각각의 결정 나무의 결과에 대하여 회귀분석에서는 평균을, 분류분석에서는 다수결 원칙을 적용하는 방식을 사용한다. 최대 허용 깊이가 얕으면 과소적합(underfitting)이 발생할 수 있고, 최대 허용 깊이가 깊으면 과대적합(overfitting)이 발생할 수 있으므로 적절한 균형 값을 찾

는 것이 중요하다.

랜덤 포레스트의 장점은 높은 정확도를 나타내며 앙상블 기법을 통해 좋은 일반화 성능을 보인다. 또 변수 소거 없이 수천 개의 입력 변수를 다루는 것이 가능하며 변인에 대한 상대적 중요도 지수를 제공한다. 쉽게 학습할 수 있어서 실무에도 활용이 가능한 수준의 모델을 쉽게 만들어 낼 수 있다. 그러나 랜덤 포레스트는 분석하는 데 시간이 오래 걸리고 결정 나무모델의 결과에 대한 설명 가능성이 상실된다는 단점이 있다(김동섭, 2020). <그림 1>은 랜덤 포레스트(Random Forest)의 모형이다.

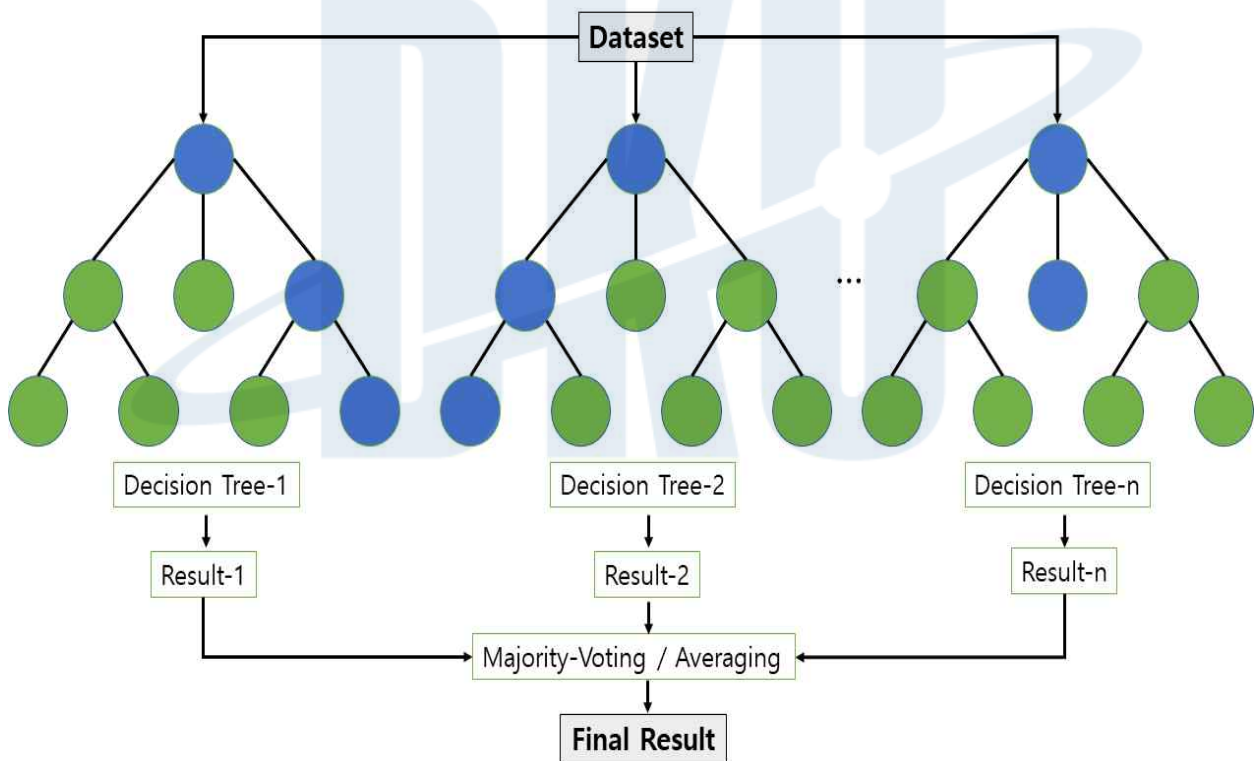


그림 1. 랜덤 포레스트(Random Forest) 모형

3) 그라디언트 부스팅 머신(GBM: Gradient Boosting Machine)

그라디언트 부스팅(GBM: Gradient Boosting Machine)은 전방 학습 앙상블 방법으로 여러 개의 결정 트리를 결합하여 강한 예측모델을 만드는 머신러닝 기법이다(Friedman, 2001). 랜덤 포레스트와 같이 여러 개의 결정 트리를 결합한 방법이지만 그라디언트 부스팅은 최초 모델에서 전체 데이터를 계속 수정하면서 트리를 연속적으로 만들어나간다. 즉 첫 번째 모델에서는 전체 데이터에 같은 가중치를 둔 상태에서 학습마다 나타난 약점을 다음 모델에서는 약점을 보완하여 점점 더 정교해지는 근사치를 통해 좋은 예측 결과를 얻을 수 있다(김민석, 2022). 부스팅(Boosting) 기법의 일종으로 학습 과정에서 경사 하강법(gradient descent)을 이용하며 불균형 데이터에 강점이 있다(Brown, Mues, 2012). 또한 의사결정나무 알고리즘을 통해 약한 모형들을 사용하여 정보획득을 통해 변수의 중요도를 알 수 있어 해석력이 높은 장점이 있다(조현진, 2018). Friedman(2001)이 제시한 그라디언트 부스팅 머신(Gradient Boosting Machine) 알고리즘은 <수식 8>~<수식 9>와 같다. <그림 2>는 그라디언트 부스팅 머신(Gradient Boosting Machine)의 모형이다.

$$\gamma_{im} = - \left[\frac{\alpha L(y_i, F(x_i))}{\alpha F(x_i)} \right] F(x) = F_{m-1}(x) \quad (\text{수식 8})$$

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \quad (\text{수식 9})$$

<수식 8>은 초기 모델로서 상수항만으로 구성되었으며 x 는 독립변인, y 는 종속변인, $L(y, F(x))$ 는 미분 가능한 손실함수(loss function)이다. <수식 8>에 의해 M번 반복 계산된 유사 잔차(pseudo residuals)를 <수식 9>에 적용한 후 γ_m 을 계산하고, <수식 9>와 같이 잔차를 업데이트하는 과정을 M번 반복한다.

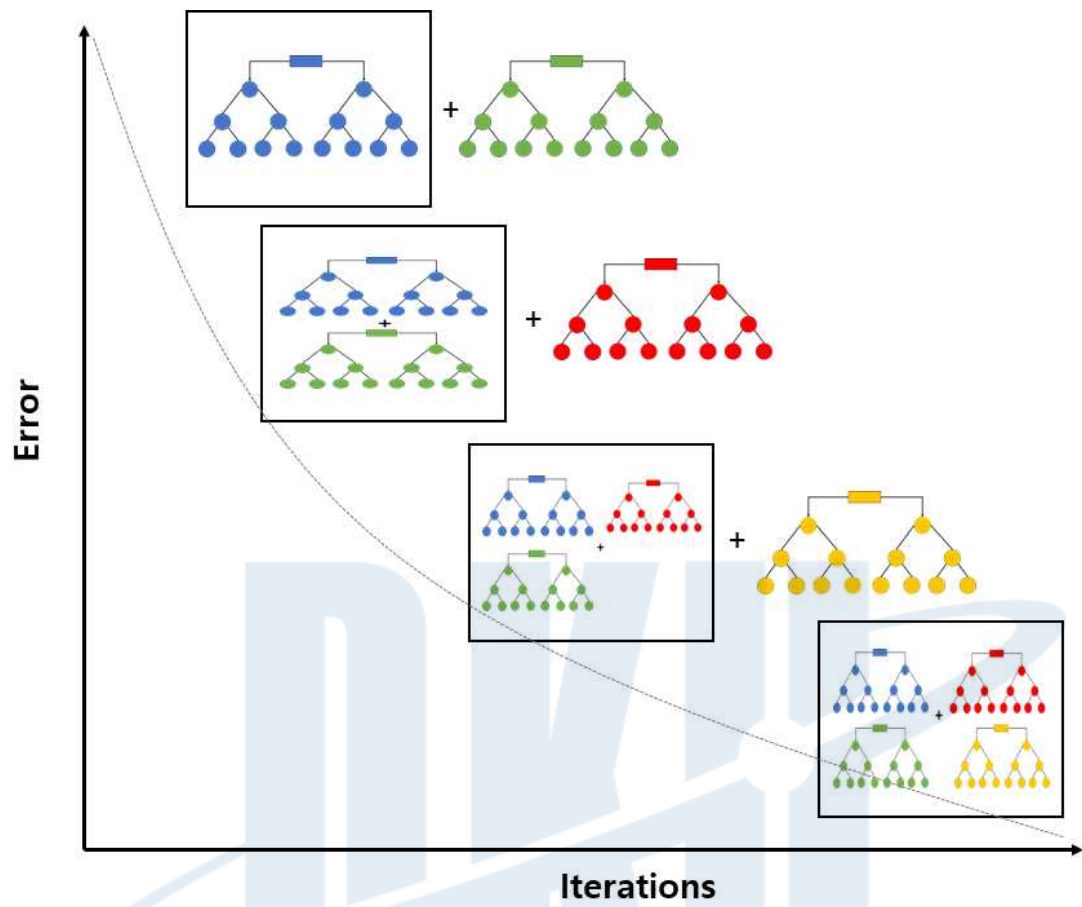


그림 2. 그래디언트 부스팅(Gradient Boosting Machine) 모형

4) 인공신경망(ANN: Artificial Neural Network)

인공신경망(ANN: Artificial Neural Network)은 기계학습 분야에서 연구되고 있는 학습 알고리즘 중 하나로 인간 두뇌 신경의 연결 구조를 모방해 데이터를 네트워크 구조를 거쳐서 처리하도록 만든 시스템이다. 신경망 분석은 복잡한 구조를 가진 자료에서 예측 문제를 해결하기 위해 사용되는 비선형모델(nonlinear model)이다. 자료로부터 반복적인 학습 과정을 거쳐 숨어있는 패턴을 찾아내는 모델링 기법으로 은닉 마디(hidden units)라는 독특한 구성요소를 가진다. 신경망 분석은 입력층(input layer), 은닉층(hidden layer), 출력층(output layer)으로 구성되어 있다. 입력층(input layer)은 입력 값(종속변인)을 받아들이는 계층으로 처리요소는 한 개의 입력 값을 받아서 그대로 다음 계층으로 내보내는 역할을 한다. 은닉층(hidden layer)은 입력층과 출력층 사이에 존재하는 계층으로 입력노드로부터 입력 값을 받아 가중합을 계산하고 전이함수에 적용하여 출력층에 전달한다. 가중치는 연결 강도로 표현되며 랜덤으로 초기에 주어졌다가 예측값을 가장 잘 맞추는 값으로 조정된다. 출력층(output layer)은 인공신경망의 최종 출력 값을 내보내는 층이다. 출력층을 구성하는 출력노드의 개수는 인공신경망 구축에 사용하는 데이터 목표 속성의 개수와 같다(배경태, 김창재, 2016). 신경망 분석에서는 입력층(input layer)과 출력층(output layer)이 반드시 존재해야 한다. 하지만 은닉층(hidden layer)은 없을 수도 있고 여러 개가 존재할 수도 있다. 인공신경망은 구성에 따라 다층 퍼셉트론(MLP: Multi Layer Perceptron), 순환 신경망(RNN: Recurrent Neural Network), 장단기 메모리(LSTM: Long Short Term Memory) 등 종류가 다양하며 여러 개의 은닉층(hidden layer)이 있는 인공신경망을 딥러닝(DNN: Deep Neural Network)이라 부른다. <그림 3>은 인공신경망의 기본 구조에 대해서 나타낸 그림이다.

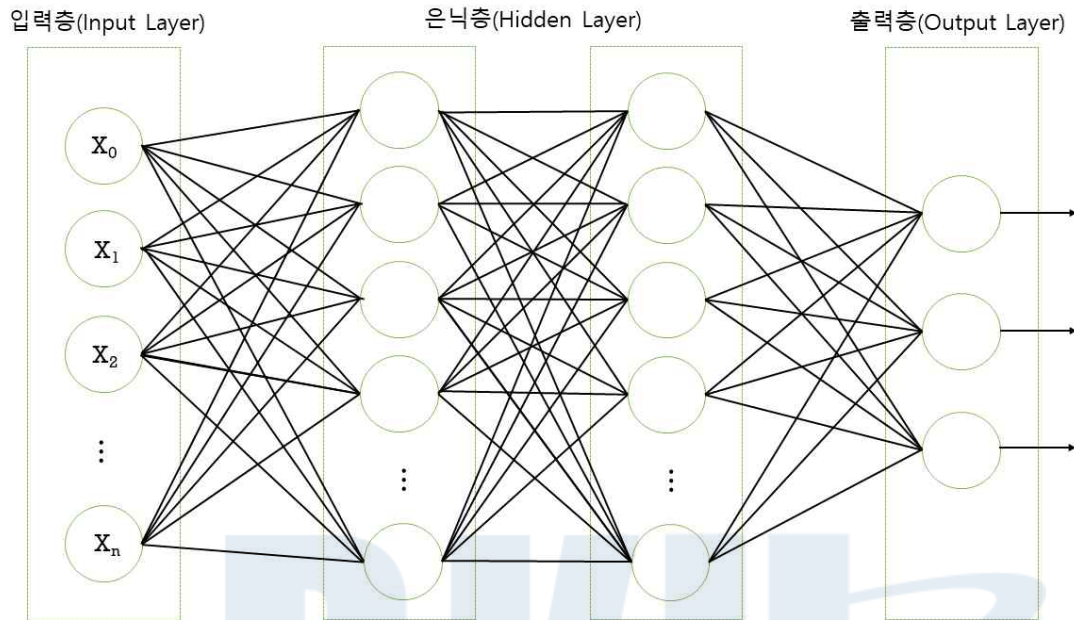


그림 3. 인공신경망(ANN) 모형

인공신경망은 가중치(w)를 고려한 입력 값의 총합에 따라 출력 값을 결정한다.

$\{(x_1, d_1), (x_2, d_2), \dots, (x_n, d_n)\}$ 과 같은 입출력 쌍이 주어졌을 때 입력된 x_n 에 따라 출력된 $y(x_n : w)$ 가 최대한 d_n 과 가까워지도록 w 를 조정하는 것이 신경망 모델에 대한 학습 과정이다. 회귀문제에서는 출력 값(y)과 실제 값(d)의 거리를 최소화하는 것으로 <수식 10>과 같이 제곱 오차(squared error)를 최소화하는 것과 같다(배성완, 2019).

<수식 10>은 <수식 11>과 같이 모든 입력 데이터에 제곱 오차를 합한 후 이를 반으로 나눈 값을 계산하고 최소화시키는 w 를 선택하는 것으로 나타낼 수 있다(손승현, 2021).

$$\|d - y(x : w)\|^2 \quad (\text{수식 10})$$

$$E(w) = \frac{1}{2} \sum_{i=1}^n \|d_i - y(x : w)\|^2 \quad (\text{수식 11})$$

이 연구에서는 역전파 방법(backpropagation)을 사용하여 경사 하강법(gradient descent)으로 훈련된 다층 피드 포워드 인공신경망(Multi Layer Feed Forward Artificial Neural Network)을 기반으로 한다. 다층 피드 포워드는 각각의 입력 뉴런에 가중치 매트릭스 w_{ki} 와 결합하여 은닉층 뉴런에 연결된다. 은닉층의 i 번째 뉴런의 활성화는 <수식 12>와 같다(이태형, 2019).

$$h_i = f(u_i) = f\left(\sum_{k=0}^K w_{ki}x_k\right) \quad (\text{수식 12})$$

$f(u_i)$ 는 입력층과 은닉층 사이의 비선형을 제공하는 연결 함수이다. w_{ki} 는 가중 매트릭스의 (k,i) 번째 가중치, x_k 는 K 번째 입력값을 나타낸다.

h_i 는 <수식 13>과 같이 다시 출력층의 뉴런과 연결되며 그곳에서 다시 가중 매트릭스와 결합하여 최종 출력값을 만든다(이태형, 2019).

$$y_j = f(u'_j) = f\left(\sum_{i=1}^N w'_{ij}h_i\right) \quad (\text{수식 13})$$

인공신경망은 주어진 데이터의 특성을 학습하는데 여러 가지 방법이 있으며 오차를 최소화하는 역전파(backpropagation) 방법을 많이 사용한다. 역전파는 순전파(feedforward) 방법을 통해 입력층의 각 노드에 패턴을 입력하고, 이 신호는 각 노드에서 변환되어 출력층까지 계산과정을 거쳐 신호를 출력하게 된다. 다음으로 예측값과 실제값을 비교하여 오차를 줄여나가는 방향으로 가중치를 반복적으로 조정해 나아간다(Kumar 등, 1997; 김승진, 2022).

역전파를 구하는 과정은 <그림 4>와 같다. E 에 $y = f(x)$ 의 편미분 $\left(\frac{\partial y}{\partial x}\right)$ 을 곱한 값을 하위층으로 전달하는 과정이며 상위층에서 하위층으로 역전파 계산은 미분의 연쇄 법칙(chain rule)에 따라 가능하다(Minsky, Papert, 1969; 배성완, 2019).

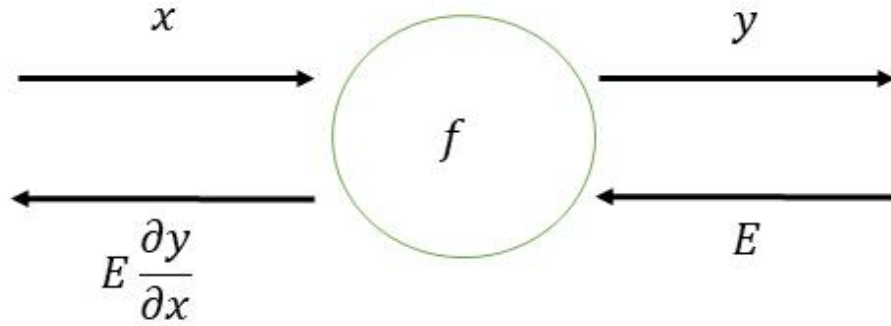


그림 4. 역전파 알고리즘 구조

예측값과 실제값의 오차는 손실함수(loss function)를 사용하여 계산하고 이 오차를 입력층에 도달할 때까지 역방향으로 보낸다. 최종목적은 E 의 함숫값을 0에 가깝게 하는 것이며 마지막으로 오차가 최소화되도록 계산한 그래디언트를 네트워크에 있는 모든 가중치의 매개변수에 반영하는 경사 하강법(gradient descent)을 사용하여 <수식 14>와 같이 가중치를 조정한다(김현주, 2022).

$$\Delta W_{jk}(P) = \beta \times \Delta W_{jk}(P-1) + \alpha \times Y_j(P) \times \delta_k(P) \quad (\text{수식 14})$$

P 번째 학습의 j 번째 뉴런과 k 번째 뉴런 사이에서 나타난 연결 강도의 변화 값 ΔW 는 학습률 α , 관성률 β , k 번째 뉴런의 오차 신호 δ_k , j 번째 뉴런의 출력값 Y_j 를 나타낸다(최형준, 김주학, 2006).

심층신경망(DNN: Deep Neural Network)은 주어진 데이터를 그대로 입력 데이터로 활용하며 데이터 자체에서 중요한 특징을 기계 스스로 학습한다. 비선형 관계를 지닌 문제 해결에 탁월하며 기존의 얕은 학습으로 풀 수 없는 복잡한 문제들을 풀 수 있는 장점이 있다. 하지만 복잡한 학습 과정을 거치기 때문에 모형 구축 시 많은 시간이 소요되며 많은 양의 데이터가 필요하다는 단점이 있다(오미애 등, 2017).

Ⅲ. 연구방법

1. 연구대상

이 연구는 대한수영연맹주관 2017년~2021년 전국수영대회 자유형 200m 경기에 출전한 여자 고등부 35명, 여자 일반부 25명의 선수를 연구대상으로 선정하였다. 연구대상의 선정 범위는 <표 1>과 같으며 총 14개 대회에 출전한 60명 선수의 경기분석자료 164개를 최종 연구자료로 선정하였다.

<표 2>는 이 연구의 특성을 나타냈으며 대한수영연맹에서 제공하는 대회별 선수정보를 참고하였다.

표 1. 연구대상의 선정 범위

구분	연도	대회명
여자 자유형 200m	2017	국가대표 선발대회
		전국체육대회
	2018	대통령배 전국수영대회
		국가대표 선발대회
	2019	전국체육대회
		국가대표 1차 선발대회
	2020	국가대표 2차 선발대회
		전국체육대회
	2021	국가대표 선발대회
		김천 전국수영대회
대통령배 전국수영대회		
한라배 전국수영대회		
전국체육대회		
합계	14개 대회	

표 2. 연구대상의 특성

구분	N	평균 연령(\pm SD)
고등부	35명	17.77(0.83)
일반부	25명	22.70(3.97)
합계	60명	20.71(3.94)

2. 연구절차

이 연구는 수영 경영 여자 자유형 200m 경기를 대상으로 분석자료를 수집하였으며, 문헌고찰을 통해 경기력 변인을 정의하였다. 정의된 경기력 변인에 따라 Dartfish Pro S 영상분석 프로그램을 통해 자료를 수치화하였으며, 수치화된 자료는 Microsoft Excel을 이용하여 최종 기록별로 정리하였다.

경영 경기결과 예측을 위해 머신러닝 기법의 예측모델을 설계하였으며, R (Version 4.2.1) 기반의 R Studio (Version 1.4.1717) 프로그램을 이용하였다. 마지막으로 예측모델의 성능을 평가하여 최적의 경기결과 예측모델을 선정하였다. <그림 5>와 같이 이 연구의 절차를 구성하였다.



그림 5. 연구절차

3. 자료수집 및 자료분석

이 연구는 경영 경기기록을 기반으로 경기결과에 미치는 경기력 변인을 도출하기 위해 선행연구를 기초로 조사하였다. 도출된 측정 변인과 정의는 <표 3>과 같다.

표 3. 측정 변인의 정의

변인 명	정의
record	최종 기록
start record	출발 신호부터 15m 지점에 도달할 때까지 걸린 시간
1lap cleanswim record	50m 구간 중 출발 구간과 턴 구간을 제외한 구간의 기록 (15m~45m)
1lap turn in record	턴 전 5m 표시부터 벽에 닿을 때까지 걸린 시간 (45m~50m)
1lap turn out record	벽면 터치부터 65m 지점에 도달할 때까지 걸린 시간 (50m~65m)
2lap cleanswim record	100m 구간 중 턴 구간을 제외한 구간의 기록 (65m~95m)
2lap turn in record	턴 전 5m 표시부터 벽에 닿을 때까지 걸린 시간 (95m~100m)
2lap turn out record	벽면 터치부터 115m 지점에 도달할 때까지 걸린 시간 (100m~115m)
3lap cleanswim record	150m 구간 중 턴 구간을 제외한 구간의 기록 (115m~145m)
3lap turn in record	턴 전 5m 표시부터 벽에 닿을 때까지 걸린 시간 (145m~150m)
3lap turn out record	벽면 터치부터 165m 지점에 도달할 때까지 걸린 시간 (150m~165m)
4lap cleanswim record	200m 구간 중 턴 구간과 도착 구간을 제외한 구간의 기록 (165m~195m)
finish record	도착 전 5m 표시부터 벽에 닿을 때까지 걸린 시간 (195m~200m)
1lap stroke time	50m 스트로크 구간에서 1회 스트로크 당 걸린 평균 시간
2lap stroke time	100m 스트로크 구간에서 1회 스트로크 당 걸린 평균 시간

변인 명	정의
3lap stroke time	150m 스트로크 구간에서 1회 스트로크 당 걸린 평균 시간
4lap stroke time	200m 스트로크 구간에서 1회 스트로크 당 걸린 평균 시간
finish stroke time	종료 구간에서 1회 스트로크에 걸린 평균 시간
start stroke	출발 후 15m 지점까지 스트로크 빈도
1lap cleanswim stroke	50m 구간 중 출발 구간과 턴 구간을 제외한 구간의 스트로크 빈도 (15m~45m)
1lap turn in stroke	턴 전 5m 표시부터 벽에 닿을 때까지 스트로크 빈도 (45m~50m)
1lap turn out stroke	벽면 터치부터 65m 지점까지 스트로크 빈도 (50m~65m)
2lap cleanswim stroke	100m 구간 중 턴 구간을 제외한 구간의 스트로크 빈도 (65m~95m)
2lap turn in stroke	턴 전 5m 표시부터 벽에 닿을 때까지 스트로크 빈도 (95m~100m)
2lap turn out stroke	벽면 터치부터 115m 지점까지 스트로크 빈도 (100m~115m)
3lap cleanswim stroke	150m 구간 중 턴 구간을 제외한 구간의 스트로크 빈도 (115m~145m)
3lap turn in stroke	턴 전 5m 표시부터 벽에 닿을 때까지 스트로크 빈도 (145m~150m)
3lap turn out stroke	벽면 터치부터 165m 지점까지 스트로크 빈도 (150m~165m)
4lap cleanswim stroke	200m 구간 중 턴 구간을 제외한 구간의 스트로크 빈도 (165m~195m)
finish stroke	도착 전 5m 표시부터 벽에 닿을 때까지 스트로크 빈도 (195m~200m)
start breath	출발 후 15m 지점까지 호흡 빈도
1lap cleanswim breath	50m 구간 중 출발 구간과 턴 구간을 제외한 구간의 호흡 빈도 (15m~45m)
1lap turn in breath	턴 전 5m 표시부터 벽에 닿을 때까지 호흡 빈도 (45m~50m)
1lap turn out breath	벽면 터치부터 65m 지점까지 호흡 빈도 (50m~65m)

변인 명	정의
2lap cleanswim breath	100m 구간 중 턴 구간을 제외한 구간의 호흡 빈도 (65m~95m)
2lap turn in breath	턴 전 5m 표시부터 벽에 닿을 때까지 호흡 빈도 (95m~100m)
2lap turn out breath	벽면 터치부터 115m 지점까지 호흡 빈도 (100m~115m)
3lap cleanswim breath	150m 구간 중 턴 구간을 제외한 구간의 스트로크 빈도 (115m~145m)
3lap turn in breath	턴 전 5m 표시부터 벽에 닿을 때까지 호흡 빈도 (145m~150m)
3lap turn out breath	벽면 터치부터 165m 지점까지 호흡 빈도 (150m~165m)
4lap cleanswim breath	200m 구간 중 턴 구간을 제외한 구간의 호흡 빈도 (165m~195m)
finish breath	도착 전 5m 표시부터 벽에 닿을 때까지 호흡 빈도 (195m~200m)
start speed	출발 신호부터 15m 지점에 도달할 때까지 평균 속도
1lap cleanswim speed	50m 구간 중 출발 구간과 턴 구간을 제외한 구간의 평균 속도 (15m~45m)
1lap turn in speed	턴 전 5m 표시부터 벽에 닿을 때까지 평균 속도 (45m~50m)
1lap turn out speed	벽면 터치부터 65m 지점에 도달할 때까지 평균 속도 (50m~65m)
2lap cleanswim speed	100m 구간 중 턴 구간을 제외한 구간의 평균 속도 (65m~95m)
2lap turn in speed	턴 전 5m 표시부터 벽에 닿을 때까지 평균 속도 (95m~100m)
2lap turn out speed	벽면 터치부터 115m 지점에 도달할 때까지 평균 속도 (100m~115m)
3lap cleanswim speed	150m 구간 중 턴 구간을 제외한 구간의 평균 속도 (115m~145m)
3lap turn in speed	턴 전 5m 표시부터 벽에 닿을 때까지 평균 속도 (145m~150m)
3lap turn out speed	벽면 터치부터 165m 지점에 도달할 때까지 평균 속도 (150m~165m)
4lap cleanswim speed	200m 구간 중 턴 구간과 도착 구간을 제외한 구간의 평균 속도 (165m~195m)

변인 명	정의
finish speed	도착 전 5m 표시부터 벽에 닿을 때까지 평균 속도 (195m~200m)
1lap stroke distance	50m 스트로크 구간에서 1회 스트로크 평균 길이
2lap stroke distance	100m 스트로크 구간에서 1회 스트로크 평균 길이
3lap stroke distance	150m 스트로크 구간에서 1회 스트로크 평균 길이
4lap stroke distance	200m 스트로크 구간에서 1회 스트로크 평균 길이
finish stroke distance	종료 구간에서 1회 스트로크 평균 길이

구간별 경기기록을 분석하기 위해 출발 구간(start phase), 턴 구간(turn phase), 종료 구간(finish phase), 스트로크 구간(clean swim phase)으로 구분하였으며(윤석훈, 1996; 정철수 등, 2003; Marinho 등, 2020), 출발 구간(start phase)은 출발 위치부터 15m 구간으로 총 15m 거리를 나타낸다. 턴 구간(turn phase)은 턴 전(turn-in) 5m부터 턴 후(turn-out) 15m 구간으로 총 60m의 거리를 나타내며, 종료 구간(finish phase)은 도착(touch) 전 5m 구간으로 총 5m의 거리이다. 스트로크 구간(clean swim phase)은 출발, 턴 및 종료 구간을 뺀 나머지 구간으로 구분하였으며 총 120m(15m~45m, 65m~95m, 115m~145m, 165m~195m)로 설정하였다(Morais 등, 2019, 2021; Marinho 등, 2020). 선행연구에 따르면 ‘clean swim phase’를 지칭하는 용어로 스트로크·역영·레이스 구간 등을 혼재하여 사용하고 있으나 본 연구에서는 스트로크 구간이라 지칭하였다. <표 4>는 본 연구의 경기력 측정 변인이며 <표 5>는 200m 경기 구간을 설정한 표이다.

표 4. 경기력 측정 변인

구분	측정 변인
구간별 기록	start record
	1lap cleanswim record
	1lap turn in record
	1lap turn out record
	2lap cleanswim record
	2lap turn in record
	2lap turn out record
	3lap cleanswim record
	3lap turn in record
	3lap turn out record
	4lap cleanswim record
	finish record
독립변인	1lap stroke time
	2lap stroke time
구간별 스트로크 시간	3lap stroke time
	4lap stroke time
	finish stroke time
구간별 스트로크 수	start stroke
	1lap cleanswim stroke
	1lap turn in stroke
	1lap turn out stroke
	2lap cleanswim stroke
	2lap turn in stroke
	2lap turn out stroke
	3lap cleanswim stroke
	3lap turn in stroke
	3lap turn out stroke
	4lap cleanswim stroke
	finish stroke

구분	측정 변인
구간별 호흡수	start breath
	1lap cleanswim breath
	1lap turn in breath
	1lap turn out breath
	2lap cleanswim breath
	2lap turn in breath
	2lap turn out breath
	3lap cleanswim breath
	3lap turn in breath
	3lap turn out breath
	4lap cleanswim breath
	finish breath
구간별 속도	start speed
	1lap cleanswim speed
	1lap turn in speed
	1lap turn out speed
	2lap cleanswim speed
	2lap turn in speed
	2lap turn out speed
	3lap cleanswim speed
	3lap turn in speed
	3lap turn out speed
	4lap cleanswim speed
	finish speed
구간별 스트로크 거리	1lap stroke distance
	2lap stroke distance
	3lap stroke distance
	4lap stroke distance
	finish stroke distance
종속변인	record

단위: record(s), stroke time(s), stroke(frequency), breath(frequency), speed(m/s), stroke distance(m)

표 5. 200m 구간 설정

구분	단위 거리
출발 구간(start phase)	0m~15m
스트로크 구간 (clean swim phase)	15m~45m
턴 구간(turn phase) 턴 전(turn-in) 턴 후(turn-out)	45m~50m(turn-in) 50m~65m(turn-out)
스트로크 구간 (clean swim phase)	65m~95m
턴 구간(turn phase) 턴 전(turn-in) 턴 후(turn-out)	95m~100m(turn-in) 100m~115m(turn-out)
스트로크 구간 (clean swim phase)	115m~145m
턴 구간(turn phase) 턴 전(turn-in) 턴 후(turn-out)	145m~150m(turn-in) 150m~165m(turn-out)
스트로크 구간 (clean swim phase)	165m~195m
종료 구간(finish phase)	195m~200m

또한, 이 연구의 경기결과(최종 기록)와 구간별 랩타임(50m, 100m, 150m, 200m)은 대한수영연맹 공식 사이트와 스포츠다이러리에서 제공하는 대회 결과기록을 참고하여 재작성하였다. 출발 구간(start phase), 턴 구간(turn phase), 스트로크 구간(clean swim phase), 종료 구간(finish phase)의 자료수집을 위해 Digital Video Camera(Sony HDR-PJ820)를 사용하여 경기영상을 수집하였다. 코로나19(COVID-19) 팬데믹(pandemic)으로 인해 경기장 출입이 안 된 2020년, 2021년 경기는 대한수영연맹에서 유튜브 중계로 송출되는 경기영상을 토대로 자료수집을 하였다. 수집된 경기영상은 영상분석 프로그램인 Dartfish Pro S를 이용하여 출발 구간(0~15m), 턴 구간(45m~65m, 95m~115m, 145m~165m), 스트로크 구간(15m~45m, 65m~95m, 115m~145m, 165m~195m), 종료 구간(195m~200m)의 기록, 스트로크 수, 호흡수를 분석하였다. 각 구간(출발 구간, 턴 구간, 종료 구간)은 수영장 레인에 빨간색으로 표시된 구간으로 구분하였다(양민정, 2018; 양민정, 최형준 2019).

이 연구에서 사용된 Dartfish 소프트웨어는 모든 스포츠 분야에서 활용되고 있으며 영상을 기반으로 실시간 혹은 종료 후 경기 상황에서 발생한 변인들을 수치화하거나 전술 형태 및 개인의 움직임 분석할 수 있는 컴퓨터 프로그램이다(강병관, 박성재 2021). 이 연구에서는 Dartfish Pro S 버전의 초시계와 태깅(Tagging) 기능을 이용하여 자료를 수집하였다. 태깅(Tagging) 기능은 경기의 주요 이벤트를 추출하기 위해 사용자가 편리하게 조작할 수 있는 버튼을 생성하고 편집할 수 있게 하는 것이며 키보드 단축키나 마우스를 이용하여 자료를 수집할 수 있다. 또한 경기의 주요 이벤트 영상을 재확인할 수 있으며 수집된 자료는 엑셀 파일로 추출하여 자료를 다양하게 활용할 수 있다(양준석, 박성재, 2016). <그림 6>은 Dartfish Pro S 프로그램을 이용한 태깅 설계 결과이다. 분석과정에서 수영선수 경력 10년 이상의 수영지도자 1명과 함께 자료를 기록하였다.

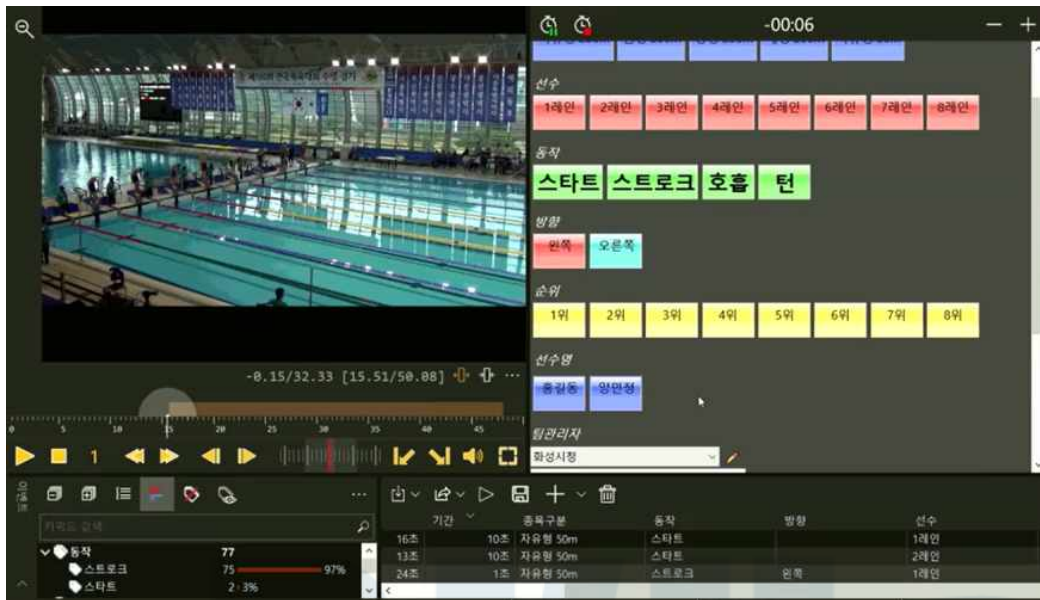


그림 6. Dartfish Pro S의 태깅 설계

측정자 간 신뢰도는 급내 상관계수(ICC: Intraclass Correlation Coefficient)를 실시하였으며 ICC 크기에 따른 해석은 0.40 미만은 좋지 않음(poor), 0.4-0.6은 보통(fair), 0.6-0.75는 좋음(good), 0.75-1.00은 매우 좋음(excellent)으로 해석된다(Cicchetti, 1994). <표 6>은 구간별 기록 변인의 급내 상관계수를 나타낸 표이며 매우좋음(Excellent) 수준으로 나타났다.

측정자의 주관적 경험이 필요한 분석요인은 없지만 오류를 최소화하기 위하여 분석요인의 애매한 상황에서는 경기영상을 반복적으로 시청하며 협의를 거쳐 최종 입력하였다.

표 6. 구간별 기록 변인의 급내 상관계수

변인	ICC	95% CI
start record	.922	.895~.943
1lap cleanswim record	.952	.935~.965
1lap turn in record	.990	.986~.992
1lap turn out record	.980	.973~.985
2lap cleanswim record	.932	.907~.950
2lap turn in record	.934	.910~.952
2lap turn out record	.976	.967~.982
3lap cleanswim record	.956	.940~.968
3lap turn in record	.929	.903~.947
3lap turn out record	.983	.977~.987
4lap cleanswim record	.941	.919~.956
finish record	.982	.976~.987

분석된 자료는 Microsoft Excel(Version 2208)을 이용하여 최종 기록별로 정리하였으며 구간별 기록, 구간별 스트로크 수를 통해 구간별 평균 스트로크 시간, 구간별 평균 스트로크 거리, 구간별 평균 속도를 추가로 기록하였다. 구간별 평균 스트로크 시간과 구간별 평균 스트로크 거리는 출발 구간(start phase)과 턴 구간(turn phase)을 제외한 스트로크 구간(clean swim phase), 종료 구간(finish phase)의 분석내용만 사용하였다. 출발 구간(start phase)과 턴 구간(turn phase)은 브레이크아웃(break-out) 구간으로 물속 동작(dolphin-kick)을 이용해 앞으로 나아가기 때문에 스트로크 수의 차이가 나타난다. <그림 7>은 Microsoft Excel(Version 2208)을 이용해 정리된 분석자료의 예시이다.

	A	E	F	G	H	I	J	K	L	M	N
1	no	record	start_record	1lap_clean	1lap_turn_in_record	1lap_turn_out_record	2lap_clean	2lap_turn_in_record	2lap_turn_out_record	3lap_clean	3lap_turn_in_record
2	1	121.35	6.77	18.22	3.44	8.24	18.89	3.83	8.08	19.02	3.8
3	2	121.55	6.61	18.25	3.69	8.25	18.99	3.67	8.31	18.98	3.9
4	3	122.21	6.79	18.37	3.54	8.2	19.06	3.63	8.31	19.05	3.81
5	4	122.93	6.74	18.52	3.73	8.38	19.09	3.83	8.08	18.92	4.12
6	5	123.22	7.11	18.05	3.44	8.5	18.89	3.67	8.54	19.12	4.19
7	6	123.40	6.81	17.98	3.9	8.51	19.32	3.74	8.61	19.22	4.02
8	7	124.34	7.34	18.15	3.68	8.5	19.05	3.91	8.61	19.42	4
9	8	129.15	6.84	18.25	3.66	8.32	19.62	4.07	8.68	20.48	4.48
10	9	121.23	6.10	18.65	3.66	7.69	18.96	4.33	7.61	19.89	3.79
11	10	121.26	5.97	18.69	3.87	7.77	18.82	4.41	7.7	20.06	3.88
12	11	122.51	6.47	18.15	3.7	7.55	19.09	4.35	7.79	20.15	3.97
13	12	122.64	6.41	18.85	3.99	7.59	19.05	4.41	7.8	19.39	4.26
14	13	124.19	6.47	18.39	3.9	7.74	19.28	4.25	7.9	19.99	4.20
15	14	126.64	6.59	18.71	3.77	8.11	19.32	4.32	8.24	20.52	4.17
16	15	135.62	6.63	19.16	4.11	8.2	20.38	4.71	8.57	22.53	5.31
17	16	122.11	6.77	17.95	3.48	8.24	19.58	3.77	8.45	19.18	3.94
18	17	122.46	6.71	18.11	3.41	8.01	19.58	3.84	8.14	19.42	3.74
19	18	123.18	6.97	18.09	3.5	8.24	19.39	3.7	8.58	19.62	3.7
20	19	123.29	6.94	18.18	3.48	8.17	19.02	3.7	8.41	19.35	3.74
21	20	123.59	6.91	18.55	3.74	8	19.49	3.67	8.31	19.79	3.87
22	21	124.37	7.04	18.65	3.71	6.74	21.05	3.7	8.48	19.65	3.64
23	22	124.98	7.21	18.75	3.67	8.01	19.52	3.77	8.31	20.02	3.77

그림 7. Microsoft Excel을 이용해 정리된 분석자료

5. 자료처리

이 연구의 자료처리는 머신러닝 기법의 경영 경기결과를 예측하기 위해 R(Version 4.2.1) 기반의 R Studio(Version 1.4.1717) 프로그램을 이용하였으며 H2O.ai 오픈 소스 머신러닝 플랫폼을 활용하였다. H2O.ai는 개방형 소스로 일반화된 선형 모델, 그래디언트 부스팅 머신(Gradient Boosting Machine), 딥러닝 등을 포함해 가장 널리 사용되는 통계 및 머신러닝 알고리즘을 지원한다.

경영 경기결과를 예측을 위한 머신러닝 모델로는 선형 회귀(Linear Regression), 라쏘 회귀(Lasso Regression), 릿지 회귀(Ridge Regression), 엘라스틱 넷 회귀(Elastic Net Regression), 랜덤 포레스트(Random Forest), 그래디언트 부스팅 머신(GBM: Gradient Boosting Machine), 인공신경망(DNN: Deep Neural Network)을 실시하였다.

<표 7>은 경영 경기결과 예측을 위한 머신러닝 모델을 나타낸 표이다.

표 7. 경영 경기결과 예측을 위한 머신러닝 모델

구분	모델명
1	선형 회귀(Linear Regression)
2	라쏘 회귀(Lasso Regression)
3	릿지 회귀(Ridge Regression)
4	엘라스틱 넷 회귀(Elastic Net Regression)
5	랜덤 포레스트(Random Forest)
6	그래디언트 부스팅 머신(GBM: Gradient Boosting Machine)
7	인공 신경망(DNN: Deep Neural Network)

머신러닝 방법은 하이퍼 파라미터(Hyper parameter)에 따라 모델의 성능이 달라진다. 하이퍼 파라미터는 머신러닝 학습 과정에서 변수값 범위를 지정하기 위한 매개변수이다. 하지만 하이퍼 파라미터를 결정하는 절대적인 값은 존재하지 않으며 적정 하이퍼 파라미터값을 구하기 위해 그 값을 변화시키면서 오차가 최소화되는 파라미터를 찾는 것이 일반적이다(김지웅, 2020). 이를 위해 연구자가 직접 시행착오 끝에 하이퍼 파라미터를 결정하거나, 그리드 서치(Grid Search), 랜덤 서치(Random Search) 등 컴퓨터가 설정된 파라미터 값의 범위 내에서 오차가 최소화되는 하이퍼 파라미터를 찾는 방법을 활용한다(배성완, 2019).

이 연구에서는 훈련 데이터를 학습하는 과정에서 학습모델별 파라미터의 탐색 범위를 <표 8>처럼 설정하고 그리드 서치(Grid Search)를 적용하여 하이퍼 파라미터를 결정하였다. 그리드 서치(Grid Search)는 교차 검증을 기반으로 모델에 사용되는 파라미터를 차례로 입력하면서 최적의 파라미터를 도출하는 방안을 제공한다(김은미, 2020). <수식 15>처럼 하이퍼 파라미터가 변화됨에 따라 검증 데이터의 평균 제곱 오차(MSE: Mean Square Error)가 가장 낮은 모델을 최종 예측모델로 선정하였다.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y}_i)^2 \quad (\text{수식 15})$$

표 8. 하이퍼 파라미터(Hyper parameter)의 범위

모델명	Hyper parameter	범위	정의
Lasso Regression	α	0.1, 1	회귀계수 값의 크기 제어
Ridge Regression	λ	0.1, 1	회귀계수 값의 크기 제어
Elastic Net Regression	α, λ	0.1, 1	라쏘 회귀와 릿지 회귀를 적용하여 회귀계수 값의 크기 제어
Random Forest	ntrees	50, 100, 200	빌드할 트리 수
	max depth	1, 20	최대 트리 깊이
	min rows	1, 10	리프(R에서)에 대한 최소 관찰 수
	col sample rate per tree	0.3, 1, 0.05	트리당 컬럼 샘플 레이트
	sample rate	0.3, 1	행 샘플링 속도
GBM	ntrees	50, 100, 200	빌드할 트리 수
	max depth	1, 20	최대 트리 깊이
	min rows	1, 5, 10, 20, 50, 100	리프(R에서)에 대한 최소 관찰 수
	learn rate	0.001, 0.01, 0.001	학습률
	col sample rate	0.3, 1, 0.05	컬럼 샘플링 비율(y축)
	col sample rate per tree	0.3, 1, 0.05	트리당 컬럼 샘플 레이트
	sample rate	0.3, 1, 0.05	행 샘플링 속도
DNN	huber alpha	0.0, 0.5, 0.9, 1.0	Huber/M-회귀에 대해 원하는 분위 수(2차 손실과 선형 손실 사이의 임계값)
	L1	0, 0.5	안정성을 추가하고 일반화를 개선하기 위해 L1 정규화 많은 가중치의 값을 0(기본값)

모델명	Hyper parameter	범위	정의
			으로 설정
	L2	0, 0.5	안정성을 추가하고 일반화를 개선하기 위해 L2 정규화 많은 가중치의 값을 더 작은 값으로 설정
	rho	0.9, 0.95, 0.99, 0.999	적응 학습률 시간 감쇠 계수
	epsilon	1e-10, 1e-8, 1e-6, 1e-4	0으로 나누는 것을 방지하기 위해 적응 학습률 시간 평활화 인수 지정
	epochs	10, 50, 100	데이터 세트를 반복할 횟수
	hidden	(30, 5), (50, 25), (10, 10)	히든 레이어(은닉층) 크기
	input dropout ratio	0, 0.1, 0.2	일반화를 개선하기 위해 입력 레이어 드롭아웃 비율
	loss	Absolute, Quadratic, Huber	손실 함수
	activation	Rectifier, Maxout, Tanh, RectifierWithDropout, MaxoutWithDropout, TanhWithDropout	활성화 함수

k-fold 교차 검증은 <그림 8>과 같이 전체 데이터를 k개의 겹으로 나눈 뒤 k-1개 겹들을 훈련 데이터(training data)로, 나머지 한 개의 겹을 검증 데이터(test data)로 활용하여 모델의 성능을 검증하는 방법이다. 이는 훈련 데이터 전체를 효율적으로 사용하고 모델이 특정 데이터 세트에 과적합 되는 것을 방지하여 일반화된 모델이 구축될 수 있다(이용성 2022). 본 연구에서는 10-fold 교차 검증을 하였으며 교차 검증 모델은 전체 데이터의 80%를 학습데이터로 20%는 검증 데이터로 사용되었다. 머신러닝 모델별 최종 예측모델이 결정되면 모델별 유효성 검증을 시행하였다. 10개의 교차 유효성 검사 모델 각각은 학습데이터 중 20%에 대해 예측을 수행하고 20% 검증 데이터의 예측 성능과 비교하였다. 각 예측모델의 성능 평가는 평균 제곱 오차(MSE: Mean Square Error), 평균 제곱근 오차(RMSE: Root Mean Square Error)의 평가지표를 활용하였다. RMSE는 MSE가 오차의 제곱의 구할 때 실제 오차 평균보다 커지는 경향이 있어서 MSE와 함께 모델의 정확성을 평가하고 비교하였다. RMSE는 <수식 16>처럼 <수식 15>의 MSE 수식에 루트를 취한 값이다. MSE 및 RMSE는 값이 작아질수록 예측값과 실제값의 격차가 감소한다는 의미이며 이는 모델의 정확성이 높아진다는 것으로 해석할 수 있다(배성완 2019).

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y}_i)^2} \quad (\text{수식 16})$$

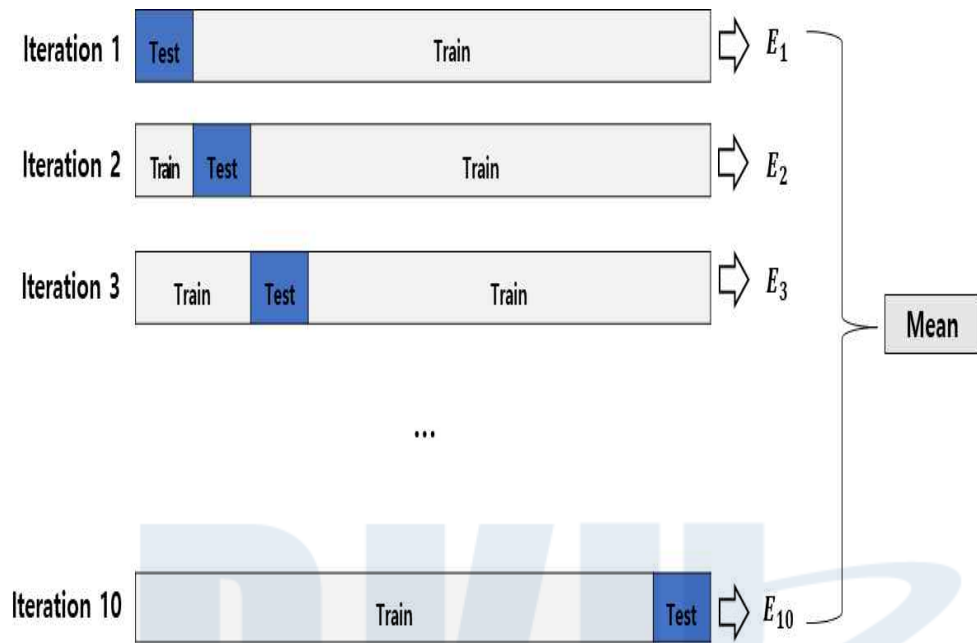


그림 8. k-fold 교차 검증(k-fold Cross Validation) 모형

예측력 비교를 위해 경영 경기결과와 실제값(최종 기록)과 머신러닝 예측모델을 적용한 경기결과 예측값의 오차와 오차율을 제시하였다. 오차는 <수식 17>과 같이 계산되었고, 오차율은 <수식 18>과 같이 계산되었다.

$$\text{오차} = \text{실제값} - \text{예측값} \quad (\text{수식 17})$$

$$\text{오차율}(\%) = (\text{오차} / \text{실제값}) \times 100 (\%) \quad (\text{수식 18})$$

또한 예측모델별 실제값과 예측값의 상관관계(R^2)를 그래프로 나타내 모델의 예측력을 확인하였다. 마지막으로 경영 경기결과 예측의 주요 예측 요인을 도출하기 위해 모델별로 예측 변수의 중요도(variable importance)를 도출하였다.

IV. 연구결과

이 연구는 2017년~2021년 전국수영대회 경영 여자 자유형 200m 경기분석기록(data)을 기반으로 경영 경기결과 예측을 위한 머신러닝 예측기법인 선형 회귀(Linear Regression), 라쏘 회귀(Lasso Regression), 릿지 회귀(Ridge Regression), 엘라스틱 넷 회귀(Elastic Net Regression), 랜덤 포레스트(Random Forest), 그래디언트 부스팅 머신(GBM: Gradient Boosting Machine), 인공신경망(DNN: Deep Neural Network) 모델을 설계하고 최적의 경영 경기결과 예측모델을 제안하기 위해 각 모델의 성능을 평가하였다. 또한, 경영 경기예측 모델별 경기결과에 영향을 미치는 경기력 변인을 도출하였다.

1. 기술통계 및 상관관계 분석

2017년~2021년 전국수영대회 경영 여자 자유형 200m 경기분석기록(data)의 측정 변인별 탐색적 분석을 위해 기초통계와 상관분석을 하였다.

기초통계 결과는 <표 9>와 같이 최소값(Min), 최대값(Max), 평균(Mean), 표준편차(Standard Deviation) 값으로 나타냈다.

표 9. 측정 변인별 기술통계

variable	N	Min	Max	Mean	SD
record	164	118.41	143.26	125.27	3.97
start record	164	5.97	7.98	6.91	0.30
1lap cleanswim record	164	17.38	21.23	18.33	0.52
1lap turn in record	164	3.3	4.15	3.65	0.18
1lap turn out record	164	6.74	10.28	8.33	0.39
2lap cleanswim record	164	18.29	22.38	19.38	0.65
2lap turn in record	164	3.02	4.71	3.81	0.24
2lap turn out record	164	7.61	9.8	8.55	0.39
3lap cleanswim record	164	18.29	23.09	19.99	0.82
3lap turn in record	164	2.91	5.31	3.91	0.25
3lap turn out record	164	7.55	10.12	8.70	0.46
4lap cleanswim record	164	18.52	23.59	20.26	0.95
finish record	164	2.97	4.69	3.45	0.30
1lap stroke time	164	0.54	1.17	0.78	0.10
2lap stroke time	164	0.56	0.78	0.68	0.04
3lap stroke time	164	0.52	0.87	0.73	0.05
4lap stroke time	164	0.59	0.86	0.74	0.05
finish stroke time	164	0.6	0.97	0.72	0.05
start stroke	164	1.00	7.00	3.98	1.38
1lap cleanswim stroke	164	24.00	33.00	26.95	1.69
1lap turn in stroke	164	2.00	4.00	3.45	0.51
1lap turn out stroke	164	3.00	12.00	8.73	1.22
2lap cleanswim stroke	164	22.00	33.00	26.49	1.78
2lap turn in stroke	164	2.00	5.00	3.46	0.54
2lap turn out stroke	164	6.00	12.00	9.20	1.14
3lap cleanswim stroke	164	24.00	34.00	27.02	1.70
3lap turn in stroke	164	2.00	5.00	3.51	0.54

3lap turn out stroke	164	7.00	12.00	9.57	1.17
4lap cleanswim stroke	164	22.00	34.00	27.97	1.68
finish stroke	164	3.00	6.00	4.46	0.55
start breath	164	0.00	3.00	0.82	0.78
1lap cleanswim breath	164	4.00	15.00	11.28	2.44
1lap turn in breath	164	1.00	3.00	1.68	0.49
1lap turn out breath	164	2.00	6.00	4.40	0.87
2lap cleanswim breath	164	7.00	16.00	12.52	1.56
2lap turn in breath	164	1.00	3.00	1.76	0.47
2lap turn out breath	164	3.00	6.00	4.76	0.76
3lap cleanswim breath	164	8.00	16.00	12.99	1.40
3lap turn in breath	164	1.00	3.00	1.82	0.47
3lap turn out breath	164	2.00	7.00	4.79	0.92
4lap cleanswim breath	164	6.00	17.00	12.87	2.10
finish breath	164	0.00	3.00	0.98	0.70
start speed	164	1.88	2.51	2.17	0.09
1lap cleanswim speed	164	1.41	1.72	1.63	0.04
1lap turn in speed	164	1.2	1.51	1.37	0.07
1lap turn out speed	164	1.45	2.22	1.80	0.08
2lap cleanswim speed	164	1.34	2.19	1.55	0.07
2lap turn in speed	164	0.5	1.65	1.31	0.10
2lap turn out speed	164	1.53	1.97	1.75	0.08
3lap cleanswim speed	164	1.29	1.64	1.50	0.06
3lap turn in speed	164	0.94	1.71	1.28	0.08
3lap turn out speed	164	1.48	1.98	1.72	0.09
4lap cleanswim speed	164	1.27	1.61	1.48	0.07
finish speed	164	1.06	1.68	1.45	0.12
1lap stroke distance	164	0.83	1.66	1.14	0.14
2lap stroke distance	164	0.9	1.25	1.12	0.07
3lap stroke distance	164	0.9	1.36	1.14	0.08
4lap stroke distance	164	0.88	1.25	1.11	0.07
finish stroke distance	164	0.88	1.36	1.08	0.07

단위:record(s), stroke time(s), stroke(frequency), breath(frequency), speed(m/s), stroke distance(m)

최종 기록(record)의 평균은 125.27(2분 5초 27)초이며 최소 118.41(1분 58초 41)초에서 최대 143.26(2분 23초 26)초로 나타났다. 구간별 기록의 평균을 보면 start record는 6.91초, 1lap cleanswim record는 18.33초, 1lap cleanswim record는 18.33초, 1lap turn in record는 3.65초, 1lap turn out record는 8.33초, 2lap cleanswim record는 19.38초, 2lap turn in record는 3.81초, 2lap turn out record는 8.55초, 3lap

cleanswim record는 19.99초, 3lap turn in record는 3.91초, 3lap turn out record는 8.70초, 4lap cleanswim record는 20.26초, finish record는 3.45초가 나타났다. 경기를 진행하면서 기록이 느려지는 경향을 보이며 마지막 스트로크 구간(4lap cleanswim record) 기록의 차이가 가장 크게 나타났다. 구간별 평균 스트로크 시간은 1lap stroke time은 0.78초, 2lap stroke time은 0.68초, 3lap stroke time은 0.73초, 4lap stroke time은 0.74초, finish stroke time 0.72초로 나타났다. 구간별 평균 스트로크 수는 start stroke는 3.98개, 1lap cleanswim stroke는 26.95개, 1lap turn in stroke는 3.45개, 1lap turn out stroke는 8.73개, 2lap cleanswim stroke는 26.49개, 2lap turn in stroke는 3.46개, 2lap turn out stroke는 9.20개, 3lap cleanswim stroke는 27.02개, 3lap turn in stroke는 3.51개, 3lap turn out stroke는 9.57개, 4lap cleanswim stroke는 27.97개, finish stroke는 4.46개가 나타났다. 구간별 평균 호흡수는 start breath는 0.82개, 1lap cleanswim breath는 11.28개, 1lap turn in breath는 1.68개, 1lap turn out breath는 4.40개, 2lap cleanswim breath는 12.52개, 2lap turn in breath는 1.76개, 2lap turn out breath는 4.76개, 3lap cleanswim breath는 12.99개, 3lap turn in breath는 1.82개, 3lap turn out breath는 4.79개, 4lap cleanswim breath는 12.87개, finish breath는 0.98개가 나타났다. 구간별 평균 속도는 start speed는 2.17m/s, 1lap cleanswim speed는 1.63m/s, 1lap turn in speed는 1.37m/s, 1lap turn out speed는 1.80m/s, 2lap cleanswim speed는 1.55m/s, 2lap turn in speed는 1.31m/s, 2lap turn out speed는 1.75m/s, 3lap cleanswim speed는 1.50m/s, 3lap turn in speed는 1.28m/s, 3lap turn out speed는 1.72m/s, 4lap cleanswim speed는 1.48m/s, finish speed는 1.45m/s로 나타났다. 구간별 평균 스트로크 거리는 1lap stroke distance는 1.14m, 2lap stroke distance는 1.12m, 3lap stroke distance는 1.14m, 4lap stroke distance는 1.11m, finish stroke distance는 1.08m로 나타났다.

상관관계 분석은 연속적 속성을 갖는 두 변인 간 상호 연관성에 대한 기술통계 정보를 제공한다. <그림 9>는 종속변인 record(최종 기록)가 절댓값 상관계수 기준 0.8 이상의 변인들에 대한 상관관계를 나타내는 값이며, <그림 10>은 절댓값 상관계수 기준 0.1 이하의 변인들에 대한 상관관계를 나타내는 값이다.

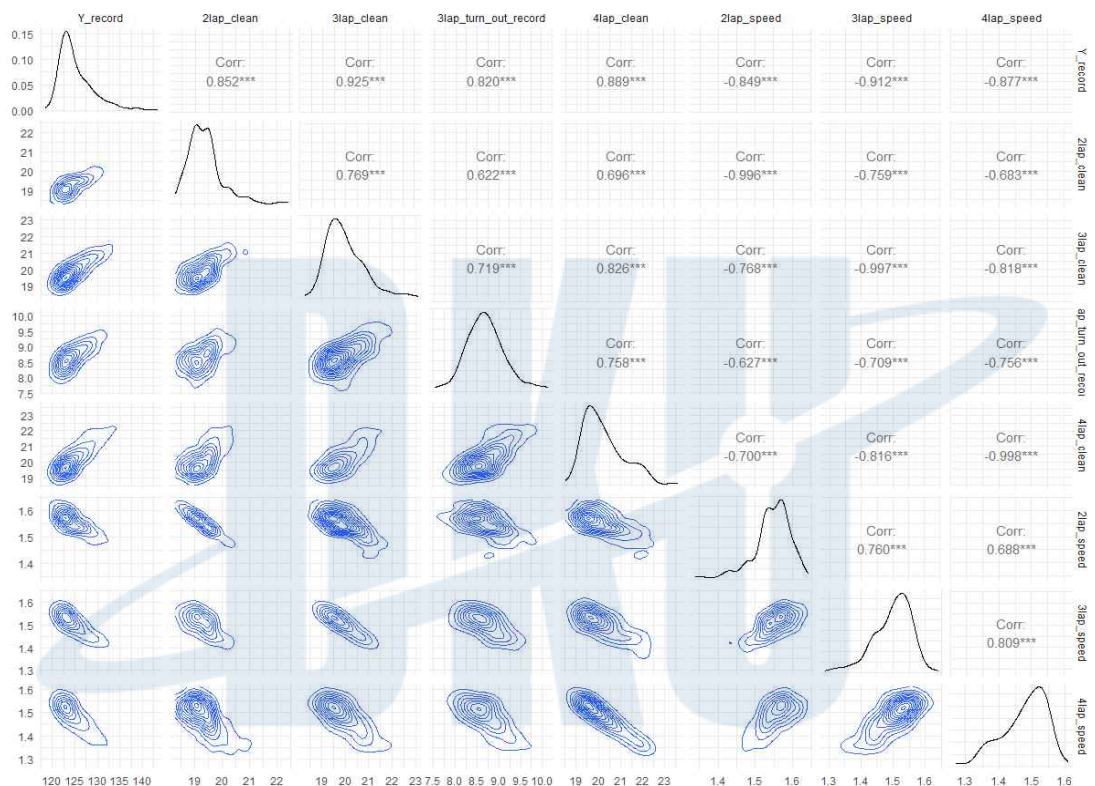


그림 9. 0.8 이상의 종속변인과 독립변인 간의 상관관계

종속변인 record(최종 기록)는 독립변인 3lap cleanswim record $r=0.925$, $p<0.01$, 2lap cleanswim record $r=0.852$, $p<0.01$ 3lap turn out record $r=0.820$, $p<0.01$, 4lap cleanswim record $r=0.889$, $p<0.01$, 2lap cleanswim speed $r=-0.849$, $p<0.01$, 3lap cleanswim speed $r=-0.912$, $p<0.01$, 4lap cleanswim speed $r=-0.877$, $p<0.01$ 의 매우 높은 상관관계가 나타나는 것으로 볼 수 있다.

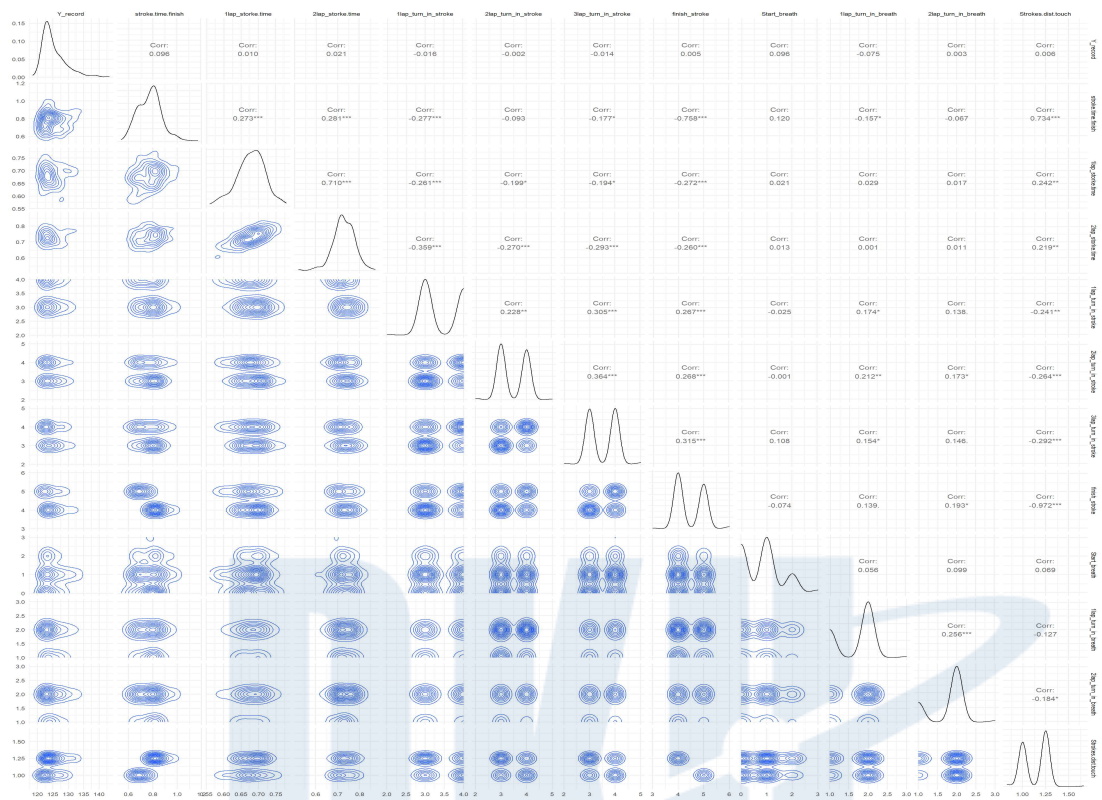


그림 10. 0.1 이하의 종속변인과 독립변인 간의 상관관계

종속변인 record(최종 기록)는 독립변인 stroke time finish $r=0.096$, 1lap stroke time $r=0.010$, 2lap stroke time $r=0.021$, 1lap turn in stroke $r=-0.016$, 2lap turn in stroke $r=-0.002$, 3lap turn in stroke $r=-0.014$, finish stroke $r=0.005$, start breath $r=0.096$, 1lap turn in breath $r=-0.075$, 2lap turn in breath $r=0.003$, finish stroke distance $r=0.006$ 의 매우 낮은 상관관계가 나타나는 것으로 볼 수 있다.

2. 머신러닝 예측모델별 분석 결과

1) 머신러닝 알고리즘별 최적 예측모델 결정

(1) 라쏘 회귀(Lasso Regression) 예측모델

Lasso Regression 모델을 최적화하기 위한 하이퍼 파라미터는 alpha로 지정하였다. alpha는 0.1~1로 범위를 설정하여 36번의 Grid Search를 통해 MSE 값이 최소가 되는 모델을 최적 모델로 결정하였다. <표 10>은 Lasso Regression 모델의 MSE 값이 낮은 상위 5개 모델을 나타낸 결과표이다. alpha: 0.575일 때 MSE 값이 0.0009의 값을 나타내며 Lasso grid model 1이 최종 모델로 선택되었다.

Lasso Regression 모델은 하이퍼 파라미터의 변화에도 MSE 값에 변화가 없는 것으로 나타났다. 따라서 Grid Search를 통해 나타난 결과값으로 Lasso Regression의 최적 모델로 결정하였다.

표 10. 라쏘 회귀(Lasso Regression) 적합 결과

no	Hyper parameter	Model ids	MSE
1	alpha: 0.5757	Lasso grid model 1	0.0009
2	alpha: 0.3130	Lasso grid model 10	0.0009
3	alpha: 0.7575	Lasso grid model 11	0.0009
4	alpha: 0.4342	Lasso grid model 12	0.0009
5	alpha: 0.8989	Lasso grid model 13	0.0009

(2) 릿지 회귀(Ridge Regression) 예측모델

Ridge Regression 모델을 최적화하기 위한 하이퍼 파라미터는 lambda로 지정하였다. lambda는 0.1~1로 범위를 설정하여 36번의 Grid Search를 통해 MSE 값이 최소가 되는 모델을 최적 모델로 결정하였다. <표 11>은 Ridge Regression 모델의 MSE 값이 낮은 상위 5개 모델을 나타낸 결과표이다. lambda: 0.0302일 때 MSE 값이 0.0162로 가장 작은 값을 나타내며 Ridge grid model 2가 최적 모델로 선택되었다.

표 11. 릿지 회귀(Ridge Regression) 적합 결과

no	Hyper parameter	Model ids	MSE
1	lambda: 0.0302	Ridge grid model 2	0.0162
2	lambda: 0.0504	Ridge grid model 4	0.0192
3	lambda: 0.0908	Ridge grid model 8	0.0241
4	lambda: 0.1110	Ridge grid model 18	0.0266
5	lambda: 0.1716	Ridge grid model 22	0.0349

(3) 엘라스틱 넷 회귀(Elastic Net Regression) 예측모델

Elastic Net Regression 모델을 최적화하기 위한 하이퍼 파라미터는 alpha, lambda로 지정하였다. alpha는 0.1~1, lambda는 0.1~1 범위를 설정하여 36번의 Grid Search를 통해 MSE 값이 최소가 되는 모델을 최적 모델로 결정하였다. <표 12>는 Elastic Net Regression 모델의 MSE 값이 낮은 상위 5개 모델을 나타낸 결과표이다. alpha: 0.5757, lambda: 0.0302일 때 MSE 값이 0.0090으로 가장 작은 값을 나타내며 Ela grid model 1가 최적 모델로 선택되었다.

표 12. 엘라스틱 넷 회귀(Elastic Net Regression) 적합 결과

no	Hyper parameter	Model ids	MSE
1	alpha: 0.5757 lambda: 0.0302	Ela grid model 1	0.0090
2	alpha: 0.8383 lambda: 0.0504	Ela grid model 16	0.0156
3	alpha: 0.8585 lambda: 0.0504	Ela grid model 2	0.0156
4	alpha: 0.2120 lambda: 0.0908	Ela grid model 32	0.0203
5	alpha: 0.0908 lambda: 0.0901	Ela grid model 12	0.0204

(4) 랜덤 포레스트(Random Forest) 예측모델

Random Forest 모델을 최적화하기 위한 하이퍼 파라미터는 col sample rate per tree, max depth, min rows, ntrees, sample rate로 지정하였다. col sample rate per tree는 0.3, 1, 0.05, max depth는 1, 20, min rows는 1, 10, ntrees는 50, 100, 200, sample rate는 0.3, 1 범위를 설정하여 36번의 Grid Search를 통해 MSE 값이 최소가 되는 모델을 최적 모델로 결정하였다. <표 13>은 Random Forest 모델의 MSE 값이 낮은 상위 5개 모델을 나타낸 결과표이다. col sample rate per tree: 0.45, max depth: 16, min rows: 1, ntrees: 200, sample rate: 0.6일 때 MSE 값이 1.084로 가장 작은 값을 나타내며 rf grid model 9가 최적 모델로 선택되었다.

표 13. 랜덤 포레스트(Random Forest) 적합 결과

no	Hyper parameter	Model ids	MSE
1	col sample rate per tree: 0.45 max depth: 16 min rows: 1 ntrees: 200 sample rate: 0.6	rf grid model 9	1.084
2	col sample rate per tree: 0.9 max depth: 16 min rows: 3 ntrees: 100 sample rate: 0.9	rf grid model 35	1.128
3	col sampl rate per tree: 0.5 max depth: 5 min rows: 3 ntrees: 50 sample rate: 0.8	rf grid model 27	1.172

no	Hyper parameter	Model ids	MSE
4	col sample rate per tree: 0.6 max depth: 17 min rows: 3 ntrees: 50 sample rate: 0.55	rf grid model 11	1.336
5	col sample rate per tree:0.95 max depth: 20 min rows: 3 ntrees: 50 sample rate: 0.4	rf grid model 6	1.501

(5) 그래디언트 부스팅 머신(GBM: Gradient Boosting Machine) 예측모델

GBM 모델을 최적화하기 위한 하이퍼 파라미터는 ntree, max depth, min row, learn rate, sample rate, col sample rate, col sample rate per tree로 지정하였다. ntree는 50, 100, 200, max depth는 1, 20, min row는 1, 5, 10, 20, 50, 100, learn rate는 0.001, 0.01, 0.001, sample rate는 0.3, 1, 0.05, col sample rate는 0.3, 1, 0.05, col sample rate per tree는 0.3, 1, 0.05로 범위를 설정하여 36번의 Grid Search를 통해 MSE 값이 최소가 되는 모델을 최적 모델로 결정하였다. <표 14>는 GBM 모델의 MSE 값이 낮은 상위 5개 모델을 나타낸 결과표이다. col sample rate: 0.75, col sample rate per tree: 0.3, learn rate: 0.009, max depth: 6, min rows: 10, ntrees: 200, sample rate: 0.75일 때 MSE 값이 2.108로 가장 작은 값을 나타내며 gbm grid model 17이 최적 모델로 선택되었다.

표 14. 그래디언트 부스팅 머신(GBM) 적합 결과

no	Hyper parameter	Model ids	MSE
1	col sample rate: 0.75 col sample rate per tree: 0.3 learn rate: 0.009 max depth: 6 min rows: 10 ntrees: 200 sample rate: 0.75 col sample rate: 0.65 col sample rate per tree: 0.55 learn rate: 0.01	gbm grid model 17	2.108
2	max depth: 14 min rows: 5 ntrees: 200 sample rate: 0.95	gbm grid model 27	3.448

no	Hyper parameter	Model ids	MSE
3	col sample rate: 0.6	gbm grid model 19	3.610
	col sample rate per tree: 0.65		
	learn rate: 0.01		
4	max depth: 9	gbm grid model 35	3.678
	min rows: 5		
	ntrees: 100		
5	sample rate: 0.85	gbm grid model 8	3.696
	col sample rate: 0.9		
	col sample rate per tree: 0.4		
	learn rate: 0.005		
	max depth: 6		
	min rows: 5		
	ntrees: 200		
	sample rate: 0.75		
	col sample rate: 0.65		
	col sample rate per tree: 0.6		
	learn rate: 0.007		
	max depth: 7		
	min rows: 20		
	ntrees: 200		
	sample rate: 0.6		

(6) 인공신경망(DNN: Deep Neural Network) 예측모델

DNN 모델을 최적화하기 위한 하이퍼 파라미터는 activation, epochs, epsilon, hidden, huber alpha, input dropout ratio, L1, L2, loss, rho로 지정하였다. activation은 Rectifier, Maxout, Tanh, RectifierWithDropout, MaxoutWithDropout, TanhWithDropout, epochs는 10, 50, 100, epsilon은 1e-10, 1e-8, 1e-6, 1e-4, hidden은 (30, 5), (50, 25), (10, 10), huber alpha는 0.0, 0.5, 0.9, 1.0, input dropout ratio는 0, 0.1, 0.2, L1은 0, 0.5, L2는 0, 0.5, loss는 Absolute, Quadratic, Huber, rho는 0.9, 0.95, 0.99, 0.999로 범위를 설정하여 60번의 Grid Search를 통해 MSE 값이 최소가 되는 모델을 최적 모델로 결정하였다. <표 15>는 DNN 모델의 MSE 값이 낮은 상위 5개 모델을 나타낸 결과표이다. activation: MaxoutWithDropout, epochs: 103.41356, epsilon: 0.00010, hidden : 50, 25, huber alpha: 0.0, input dropout ratio: 0.1, L1: 0, L2: 0, loss: Quadratic, rho: 0.999일 때 MSE 값이 0.2572로 가장 작은 값을 나타내며 dnn grid model 48이 최적 모델로 선택되었다.

표 15. 인공신경망(Deep Neural Network) 적합 결과

no	Hyper parameter	Model ids	MSE
1	activation: MaxoutWithDropout epochs: 103.41356 epsilon: 0.00010 hidden : 50, 25 huber alpha: 0.0 input dropout ratio: 0.1 L1: 0, L2: 0 loss: Quadratic rho: 0.999	dnn grid model 48	0.2572
2	activation: Maxout epochs: 103.41356 epsilon: 0.00010 hidden : 30, 5	dnn grid model 35	0.2655

no	Hyper parameter	Model ids	MSE
3	huber alpha: 0.9	dnn grid model 19	0.3376
	input dropout ratio: 0.2		
	L1: 0, L2: 0.5		
4	loss: Absolute	dnn grid model 22	0.5912
	rho: 0.999		
	activation: Maxout		
5	epochs: 10.63871	dnn grid model 13	0.5956
	epsilon: 0.00010		
	hidden : 10, 10		
4	huber alpha: 1.0	dnn grid model 22	0.5912
	input dropout ratio: 0.1		
	L1: 0, L2: 0.5		
5	loss: Quadratic	dnn grid model 13	0.5956
	rho: 0.990		
	activation: Maxout		
4	epochs: 52.14393	dnn grid model 22	0.5912
	epsilon: 0.00000		
	hidden : 30, 5		
5	huber alpha: 1.0	dnn grid model 13	0.5956
	input dropout ratio: 0.1		
	L1: 0, L2: 0.0		
4	loss: Quadratic	dnn grid model 22	0.5912
	rho: 0.90		
	activation: Rectifier		
5	epochs: 103.41356	dnn grid model 13	0.5956
	epsilon: 0.00010		
	hidden : 50, 25		
4	huber alpha: 0.0	dnn grid model 13	0.5956
	input dropout ratio: 0.2		
	L1: 0, L2: 0.5		
5	loss: Absolute	dnn grid model 13	0.5956
	rho: 0.990		
	activation: Maxout		

2) 모델별 예측력 비교

이 연구에 적용된 머신러닝 예측모델은 선형 회귀(Linear Regression), 라쏘 회귀(Lasso Regression), 릿지 회귀(Ridge Regression), 엘라스틱 넷 회귀(Elastic Net Regression), 랜덤 포레스트(Random Forest), 그래디언트 부스팅 머신(GBM), 인공신경망(DNN)이며, 머신러닝 예측모델별 Grid Search를 통해 하이퍼 파라미터를 확인하여 최적의 예측모델을 결정하였다. 결정된 최종 모델들은 10-fold 교차 유효성 검증 통해 모델별 예측 성능을 평가하고 예측력을 비교하였다. <표 16>은 각 머신러닝 예측모델의 유효성 검증(tset)을 통해 산출된 MSE, RMSE와 학습 과정 중 검증 데이터를 통해(validation) 산출된 MSE, RMSE 값이 낮은 순서대로 나타냈다.

표 16. 머신러닝 모델 간 예측 성능 결과 비교

Model	validation		test	
	MSE	RMSE	MSE	RMSE
Lasso Regression	0.00098	0.03143	0.00090	0.02683
Elastic Net Regression	0.00906	0.09520	0.00784	0.07981
Linear Regression	0.01282	0.11322	0.01217	0.10771
Ridge Regression	0.01623	0.12742	0.0147	0.11541
DNN	0.29266	0.50723	0.23553	0.46703
Random Forest	1.09031	1.04418	0.9560	0.9090
GBM	2.10870	1.45213	1.86647	1.22909

note: validation은 훈련 데이터 중 검증 데이터를 통해 산출된 MSE 및 RMSE이며, test는 최종 모델에 유효성 검증 데이터를 통해 산출된 MSE 및 RMSE임.

<표 16>과 같이 경영 경기결과 예측모델은 Lasso Regression 모델의 MSE(0.00090) 및 RMSE(0.02683) 값이 가장 낮게 나타나 다른 예측모델보다 예측 성능이 우수한 것으로 나타났다. 다음으로 Elastic Net Regression, Linear Regression, Ridge Regression, DNN, Random Forest, GBM 순으로 예측 성능을 나타냈다. GBM 모델은 예측모델 중 MSE(1.86647), RMSE(1.22909) 값이 가장 커 예측력이 가장 낮은 것으로 나타났다. DNN, Random Forest, GBM은 Lasso Regression, Elastic Net Regression, Linear Regression, Ridge Regression의 선형 회귀 예측모델보다 MAE 및 RMSE 값이 다소 크지만 유사한 수준으로 모델별 예측 성능을 나타낸다.

실제값(최종 기록)과 머신러닝 예측모델을 적용한 경영 여자 자유형 200m 경기결과 예측값의 차이를 확인하기 위해 실제값의 평균(Mean)과 표준편차(SD), 예측모델별 최종 모델에서 도출된 예측값의 평균(Mean)과 표준편차(SD), 평균 오차(Mean Error), 평균 오차율(Mean Error Rate)(%)을 나타냈다. 본 장에서는 <표 17>~<표 23>과 같이 예측모델별 도출된 예측값의 평균에 대한 값을 나타냈으며 이 연구에서 사용된 164개 경기결과에 대한 실제값과 예측값의 기록 차이를 나타낸 결과는 부록에 첨부하였다.

<그림 11>~<그림 17>은 예측모델별 실제값과 예측값 간의 상관관계를 그래프로 나타냈다.

표 17. 선형 회귀(Linear Regression) 모델 예측 결과(단위: 초)

Model	Actual record Mean(SD)	Predicted record Mean(SD)	Mean Error	Mean Error Rate(%)
Linear Regression	125.27(3.96)	125.26(3.97)	0.004	0.003

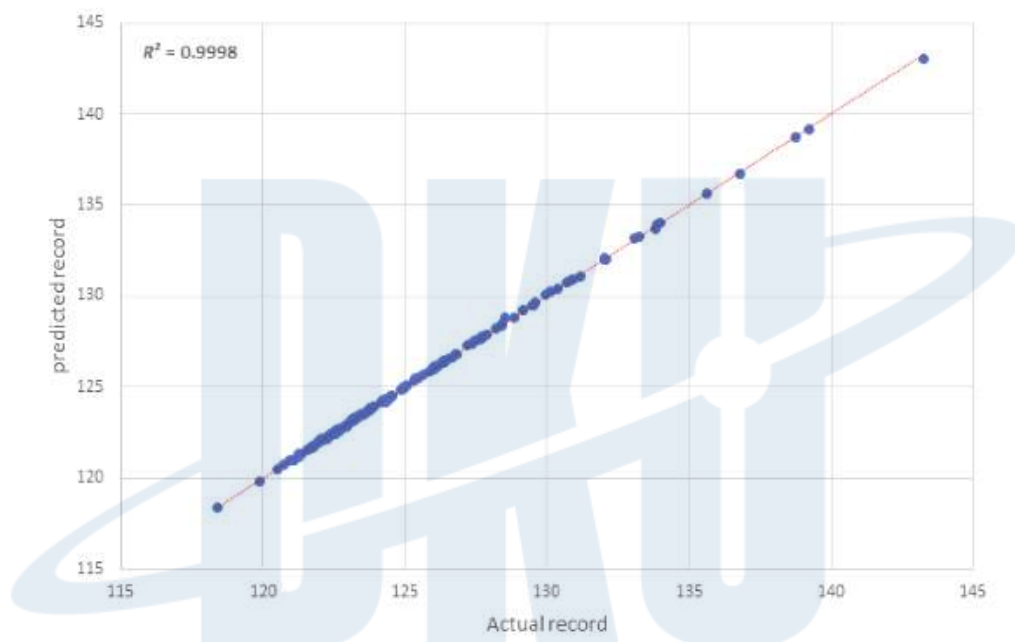


그림 11. 선형 회귀(Linear Regression) 예측모델의 실제값과 예측값 상관관계

<표 17>과 같이 Linear Regression 예측모델에서 도출된 예측값의 평균은 125.26(3.97)초로 나타났으며 실제값과 평균 0.004초 차이가 나타났다. 평균 오차율은 0.003%로 최소 -0.2%부터 최대 0.14%로 나타났다(부록의 모델별 결과표 참고). 또한, <그림 11>과 같이 실제값과 예측값의 상관관계를 분석하면 R^2 값이 0.9998로 이는 예측값이 실제값을 99.98% 수준으로 적합도를 가지고 있는 것으로 나타났다.

표 18. 라쏘 회귀(Lasso Regression) 모델 예측 결과(단위: 초)

Model	Actual record Mean(SD)	Predicted record Mean(SD)	Mean Error	Mean Error Rate(%)
Lasso Regression	125.27(3.96)	125.26(3.96)	0.004	0.003

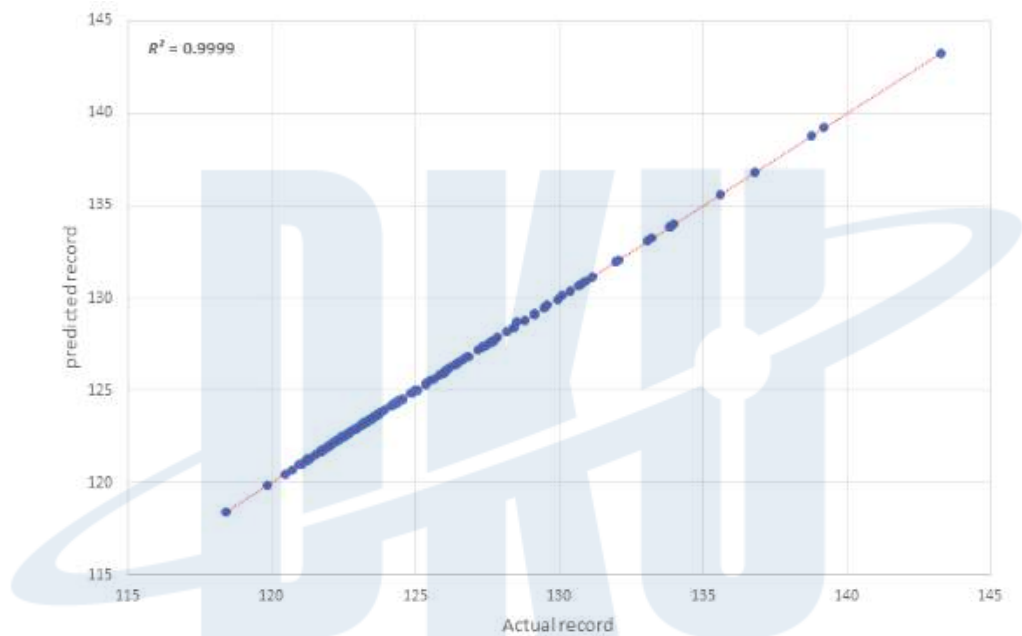


그림 12. 라쏘 회귀(Lasso Regression) 예측모델의 실제값과 예측값 상관관계

<표 18>과 같이 Lasso Regression 예측모델에서 도출된 예측값의 평균은 125.26(3.96)초로 나타났으며 실제값과 평균 0.004초 차이가 나타났다. 평균 오차율은 0.003%로 최소 -0.14%부터 최대 0.03%로 나타났다(부록의 모델별 결과표 참고). 또한, <그림 12>와 같이 실제값과 예측값의 상관관계를 분석하면 R^2 값이 0.9999로 이는 예측값이 실제값을 99.99% 수준으로 적합도를 가지고 있는 것으로 나타났다.

표 19. 릿지 회귀(Ridge Regression) 모델 예측 결과(단위: 초)

Model	Actual record Mean(SD)	Predicted record Mean(SD)	Mean Error	Mean Error Rate(%)
Ridge Regression	125.27(3.96)	125.26(3.95)	0.005	0.003

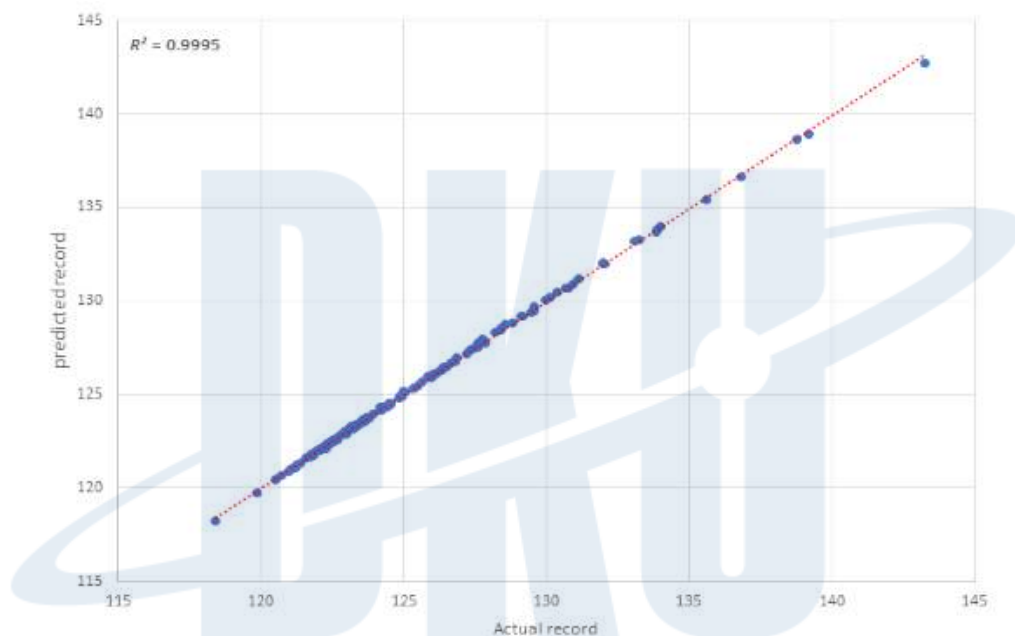


그림 13. 릿지 회귀(Ridge Regression) 예측모델의 실제값과 예측값 상관관계

<표 19>와 같이 Ridge Regression 예측모델에서 도출된 예측값의 평균은 125.26(3.95)초로 나타났으며 실제값과 평균 0.005초 차이가 나타났다. 평균 오차율은 0.003%로 최소 -0.17%부터 최대 0.36%로 나타났다(부록의 모델별 결과표 참고). 또한, <그림 13>과 같이 실제값과 예측값의 상관관계를 분석하면 R^2 값이 0.9995로 이는 예측값이 실제값을 99.95% 수준으로 적합도를 가지고 있는 것으로 나타났다.

표 20. 엘라스틱 넷 회귀(Elastic Net Regression) 모델 예측 결과(단위: 초)

Model	Actual record Mean(SD)	Predicted record Mean(SD)	Mean Error	Mean Error Rate(%)
Elastic Net Regression	125.27(3.96)	125.26(3.93)	0.004	0.003

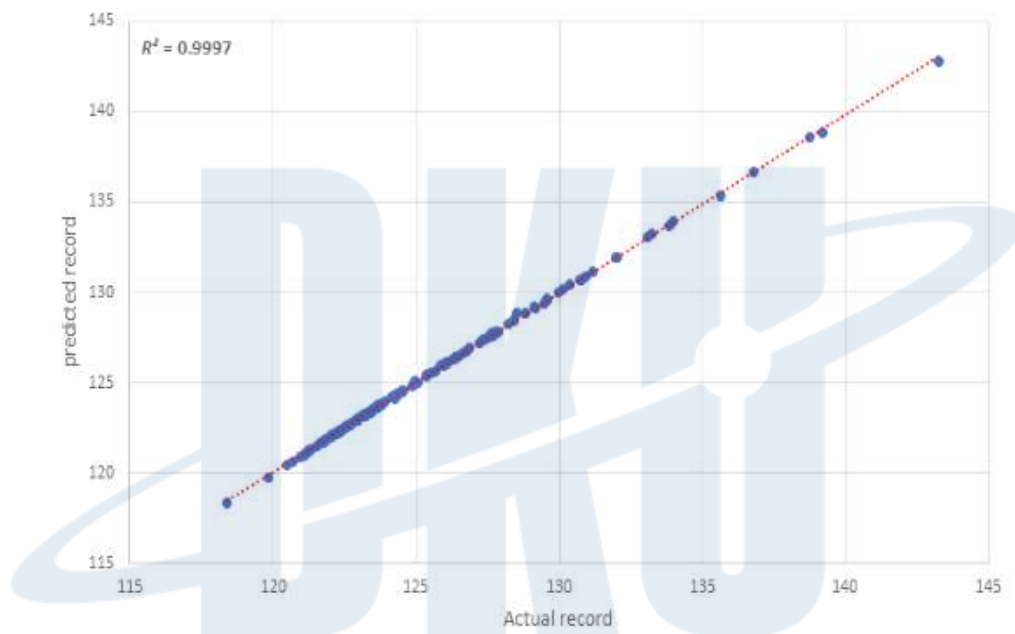


그림 14. 엘라스틱 넷 회귀(Elastic Net Regression) 예측모델의 실제값과 예측값 상관관계

<표 20>과 같이 Elastic Net Regression 예측모델에서 도출된 예측값의 평균은 125.26(3.93)초로 나타났으며 실제값과 평균 0.004초 차이가 나타났다. 평균 오차율은 0.003%로 최소 -0.22%부터 최대 0.35%로 나타났다(부록의 모델별 결과표 참고). 또한, <그림 14>와 같이 실제값과 예측값의 상관관계를 분석하면 R^2 값이 0.9997로 이는 예측값이 실제값을 99.97% 수준으로 적합도를 가지고 있는 것으로 나타났다.

표 21. 랜덤 포레스트(Random Forest) 모델 예측 결과(단위: 초)

Model	Actual record Mean(SD)	Predicted record Mean(SD)	Mean Error	Mean Error Rate(%)
Random Forest	125.27(3.96)	125.28(3.80)	-0.016	-0.017

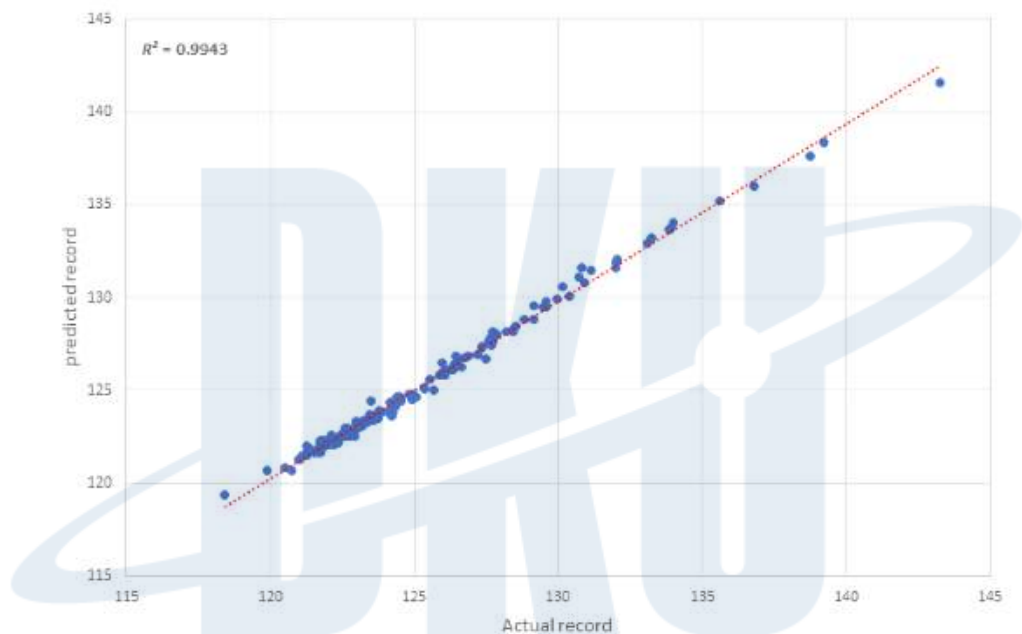


그림 15. 랜덤 포레스트(Random Forest) 예측모델의 실제값과 예측값 상관관계

<표 21>과 같이 Random Forest 예측모델에서 도출된 예측값의 평균은 125.28(3.80)초로 나타났으며 실제값과 평균 -0.016초 차이가 나타났다. 평균 오차율은 -0.017%로 최소 -0.81%부터 최대 1.18%로 나타났다(부록의 모델별 결과표 참고). 또한, <그림 15>와 같이 실제값과 예측값의 상관관계를 분석하면 R^2 값이 0.9943으로 이는 예측값이 실제값을 99.43% 수준으로 적합도를 가지고 있는 것으로 나타났다.

표 22. 그래디언트 부스팅 머신(GBM) 모델 예측 결과(단위: 초)

Model	Actual record Mean(SD)	Predicted record Mean(SD)	Mean Error	Mean Error Rate(%)
GBM	125.27(3.96)	125.27(3.02)	0.002	-0.021

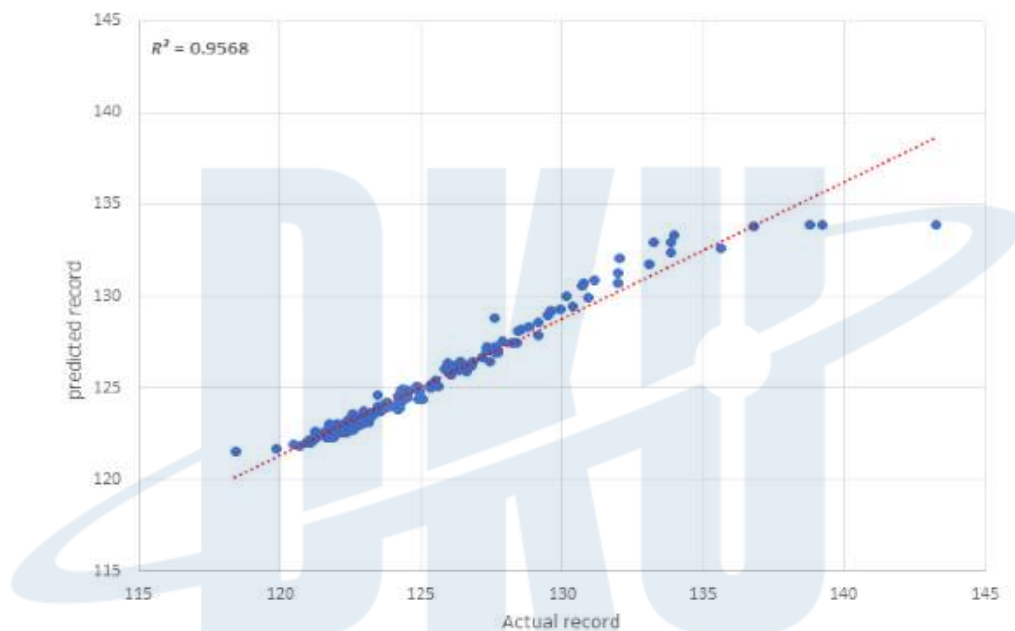


그림 16. 그래디언트 부스팅 머신(GBM) 예측모델의 실제값과 예측값 상관관계

<표 22>와 같이 GBM 예측모델에서 도출된 예측값의 평균은 125.27(3.02)초로 나타났다으며 실제값과 평균 0.002초 차이가 나타났다. 평균 오차율은 -0.021%로 최소 -2.65%부터 최대 6.52%로 나타났다(부록의 모델별 결과표 참고). 또한, <그림 16>과 같이 실제값과 예측값의 상관관계를 분석하면 R^2 값이 0.9568로 이는 예측값이 실제 값을 95.68% 수준으로 적합도를 가지고 있는 것으로 나타났다.

표 23. 인공신경망(DNN) 모델 예측 결과(단위: 초)

Model	Actual record Mean(SD)	Predicted record Mean(SD)	Mean Error	Mean Error Rate(%)
DNN	125.27(3.96)	125.06(3.85)	0.209	0.164

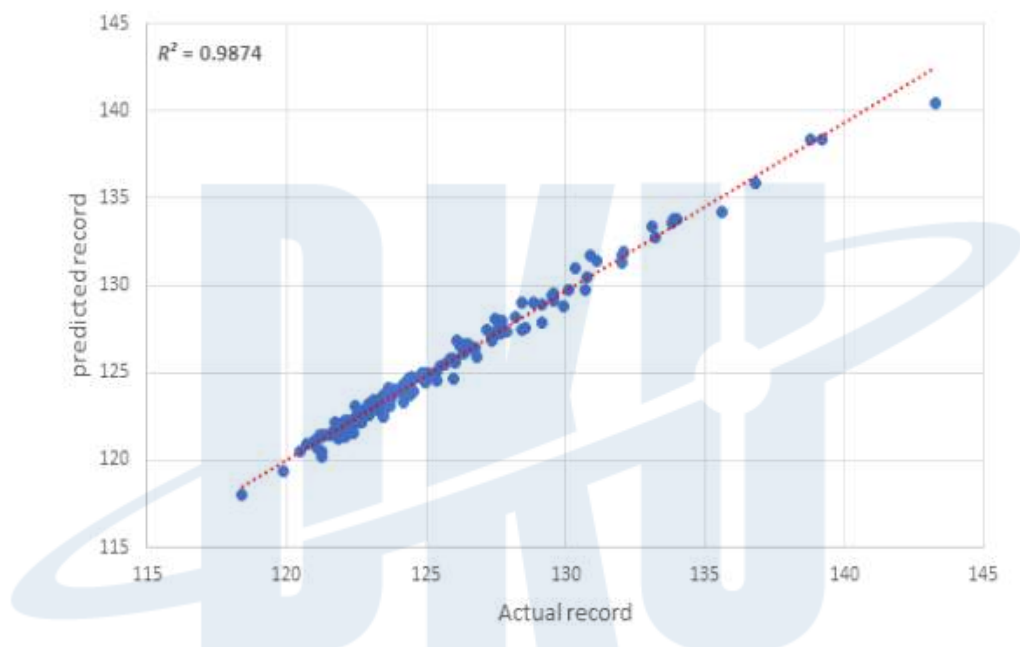


그림 17. 인공신경망(DNN) 예측모델의 실제값과 예측값 상관관계

<표 23>과 같이 DNN 예측모델에서 도출된 예측값의 평균은 125.06(3.85)초로 나타났다. 실제값과 평균 0.209초 차이가 나타났다. 평균 오차율은 0.164%로 최소 -0.63% 부터 최대 1.98%로 나타났다(부록의 모델별 결과표 참고). 또한, <그림 17>과 같이 실제 값과 예측값의 상관관계를 분석하면 R^2 값이 0.9874로 이는 예측값이 실제값을 98.74% 수준으로 적합도를 가지고 있는 것으로 나타났다.

3. 경영 경기결과 예측모델 간 경기력 변인의 중요도 분석

경영 경기결과 예측모델의 경기력 변인 간 상대적 중요도를 확인하였다.

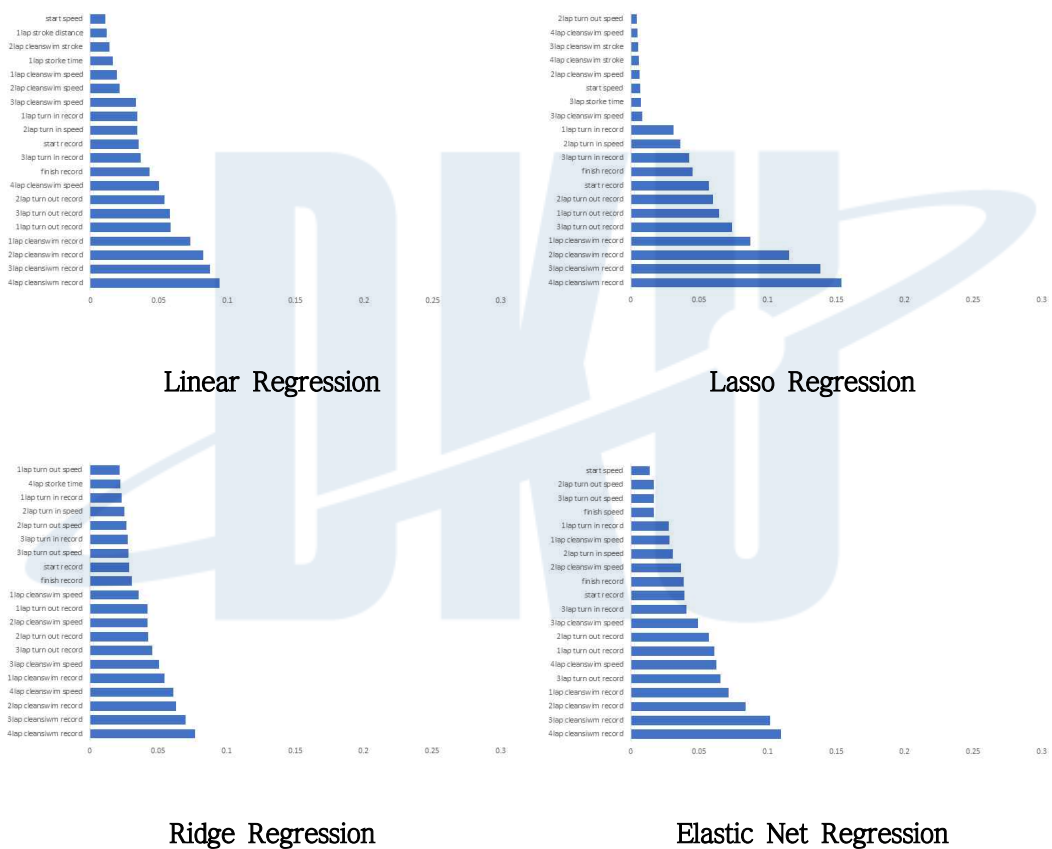
<표 24>는 선형 회귀(Linear Regression) 예측모델, 라쏘 회귀(Lasso Regression) 예측모델, 릿지 회귀(Ridge Regression) 예측모델, 엘라스틱 넷 회귀(Elastic Net Regression) 예측모델, 랜덤 포레스트(Random Forest) 예측모델, 그래디언트 부스팅 머신(GBM) 예측모델, 인공신경망(DNN) 예측모델의 변인 간 상대적 중요도를 퍼센트 단위의 비율로 변환하여 나타냈다.

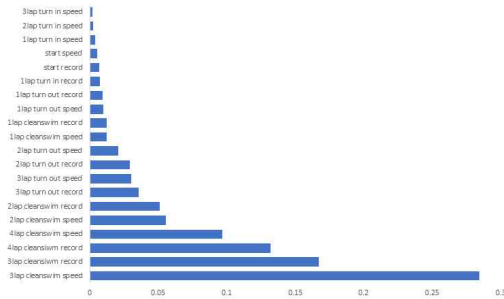
표 24. 예측모델의 경기력 변인 간 상대적 중요도 비율

variable	Models						
	LR	Lasso	Ridge	Elastic Net	RF	GBM	DNN
start record	3.53%	5.68%	2.87%	3.92%	0.68%	1.23%	1.76%
1lap cleanswim record	7.29%	8.73%	5.44%	7.13%	1.22%	5.34%	1.32%
1lap turn in record	3.45%	3.11%	2.3%	2.78%	0.72%	0.41%	2.16%
1lap turn out record	5.88%	6.47%	4.19%	6.12%	0.92%	0.99%	1.56%
2lap cleanswim record	8.25%	11.54%	6.32%	8.37%	5.09%	5.41%	1.38%
2lap turn in record	3.46%	3.61%	2.51%	3.07%	0.23%	0.08%	0.82%
2lap turn out record	5.43%	6.02%	4.27%	5.68%	2.94%	2.53%	1.59%
3lap cleanswim record	8.76%	13.86%	6.98%	10.16%	16.72%	20.26%	1.03%
3lap turn in record	3.7%	4.24%	2.78%	4.07%	0.14%	0.11%	2.6%
3lap turn out record	5.8%	7.39%	4.58%	6.56%	3.54%	3.4%	1.71%
4lap cleanswim record	9.47%	15.39%	7.69%	10.97%	13.2%	13.29%	0.79%
finish record	4.34%	4.49%	3.06%	3.87%	0.09%	0.14%	0.7%
1lap stroke time	1.63%	0.2%	0.1%	0%	0.13%	0.03%	1.49%
2lap stroke time	0.91%	0.14%	0.73%	0%	0.08%	0.03%	3.23%
3lap stroke time	0.03%	0.73%	1.67%	0%	0.11%	0.34%	1.98%
4lap stroke time	1.02%	0.29%	2.24%	0%	0.07%	0.18%	1.84%
finish stroke time	0.86%	0.1%	0.43%	0%	0.09%	0.05%	2.31%
start stroke	0.02%	0.03%	0.05%	0%	0.06%	0.04%	0.98%
1lap cleanswim stroke	0.28%	0.23%	0.77%	0%	0.06%	0.04%	2.57%
1lap turn in stroke	0.07%	0.03%	0.02%	0%	0.02%	0.01%	2.16%
1lap turn out stroke	0.41%	0.06%	0.06%	0%	0.12%	0.37%	2.7%
2lap cleanswim stroke	1.38%	0.01%	0.52%	0%	0.05%	0.06%	2.24%
2lap turn in stroke	0.07%	0.07%	0.13%	0%	0.02%	0.01%	0.94%

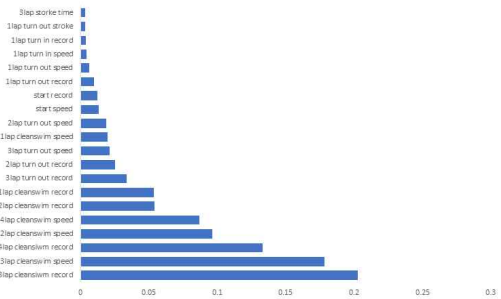
variable	Models						
	LR	Lasso	Ridge	Elastic Net	RF	GBM	DNN
2lap turn out stroke	0.17%	0.05%	0.15%	0%	0.16%	0.26%	0.87%
3lap cleanswim stroke	0.55%	0.52%	0.93%	0%	0.04%	0.03%	2.08%
3lap turn in stroke	0.02%	0.06%	0.22%	0%	0.04%	0.02%	1.7%
3lap turn out stroke	0.54%	0.01%	0.01%	0%	0.08%	0.08%	2.92%
4lap cleanswim stroke	0.18%	0.56%	1.41%	0%	0.06%	0%	1.93%
finish stroke	0.76%	0.18%	0.17%	0%	0.04%	0.01%	1.56%
start breath	0.14%	0.09%	0.11%	0%	0.03%	0.01%	1.22%
1lap cleanswim breath	0.09%	0.09%	0.18%	0%	0.08%	0.01%	2.49%
1lap turn in breath	0.04%	0.03%	0.19%	0%	0.02%	0%	1.94%
1lap turn out breath	0.17%	0.09%	0.11%	0%	0.04%	0%	4.08%
2lap cleanswim breath	0.03%	0.17%	0.04%	0%	0.039%	0.01%	2.72%
2lap turn in breath	0.26%	0.02%	0.1%	0%	0.02%	0%	1.52%
2lap turn out breath	0.16%	0.16%	0.04%	0%	0.07%	0.01%	1.05%
3lap cleanswim breath	0.06%	0.21%	0.01%	0%	0.1%	0.02%	2.43%
3lap turn in breath	0.2%	0.04%	0.03%	0%	0.01%	0%	1.64%
3lap turn out breath	0.11%	0.13%	0.06%	0%	0.13%	0.1%	3.43%
4lap cleanswim breath	0.02%	0.03%	0.21%	0%	0.06%	0.01%	2.75%
finish breath	0.16%	0.04%	0.22%	0%	0.02%	0%	2.03%
start speed	1.1%	0.69%	1.93%	1.37%	0.51%	1.34%	0.81%
1lap cleanswim speed	1.96%	0.1%	3.58%	2.83%	1.25%	1.97%	3.07%
1lap turn in speed	0.36%	0.02%	1.07%	0.64%	0.38%	0.45%	2.35%
1lap turn out speed	0.23%	0.05%	2.18%	0.97%	0.97%	0.63%	3.27%
2lap cleanswim speed	2.16%	0.61%	4.2%	3.67%	5.56%	9.62%	2.47%
2lap turn in speed	0.2%	0.21%	1.43%	1.09%	0.14%	0.07%	1.65%
2lap turn out speed	1.06%	0.44%	2.66%	1.66%	2.07%	1.87%	1%
3lap cleanswim speed	3.33%	0.81%	5.06%	4.92%	28.47%	17.82%	0.96%
3lap turn in speed	0.03%	0.17%	1.24%	0.55%	0.2%	0.18%	0.72%
3lap turn out speed	0.95%	0.13%	2.81%	1.68%	3.02%	2.12%	0.98%
4lap cleanswim speed	5%	0.48%	6.08%	6.24%	9.68%	8.67%	0.84%
finish speed	1.04%	0.38%	1.92%	1.69%	0.13%	0.29%	1.24%
1lap stroke distance	1.21%	0.39%	0.46%	0%	0.06%	0.01%	0.78%
2lap stroke distance	0.47%	0.21%	0.24%	0%	0.05%	0.01%	0.49%
3lap stroke distance	0.02%	0.25%	0.12%	0%	0.03%	0.01%	0.34%
4lap stroke distance	0.84%	0.27%	0.66%	0%	0.08%	0.01%	0.22%
finish stroke distance	0.34%	0.01%	0.48%	0%	0.04%	0.02%	0.6%
합계	100%	100%	100%	100%	100%	100%	100%

<그림 18>은 선형 회귀(Linear Regression) 예측모델, 라쏘 회귀(Lasso Regression) 예측모델, 릿지 회귀(Ridge Regression) 예측모델, 엘라스틱 넷 회귀(Elastic Net Regression) 예측모델, 랜덤 포레스트(Random Forest) 예측모델, 그래디언트 부스팅 머신(GBM) 예측모델, 인공신경망(DNN) 예측모델에서 상대적 중요도가 높은 상위 20 개의 변인을 그래프로 나타냈다.





Random Forest



GBM



DNN

그림 18. 예측모델의 경기력 변인 간 상대적 중요도

Linear Regression 예측모델은 4lap cleanswim record 변인이 9.47%로 가장 높은 중요도를 나타냈으며 다음으로 3lap cleanswim record 변인 8.76%, 2lap cleanswim record 변인 8.25%를 나타냈다. Lasso Regression 예측모델은 4lap cleanswim record 변인이 15.39%로 가장 높은 중요도를 나타냈으며 다음으로 3lap cleanswim record 변인 13.86%, 2lap cleanswim record 변인 11.54%를 나타냈다. Ridge Regression 예측모델은 4lap cleanswim record 변인이 7.69%로 가장 높은 중요도를 나타냈으며 다음으로 3lap cleanswim record 변인 6.98%, 2lap cleanswim record 변인 6.32%를 나타냈다. Elastic Net Regression 예측모델은 4lap cleanswim record 변인이 10.97%로 가장 높은 중요도를 나타냈으며 다음으로 3lap cleanswim record 변인 10.16%, 2lap

cleanswim record 변인 8.37%를 나타냈다. Random Forest 예측모델은 3lap cleanswim speed 변인이 28.47%로 가장 높은 중요도를 나타냈으며 다음으로 3lap cleanswim record 변인 16.72%, 4lap cleanswim record 변인 13.20%를 나타냈다. GBM 예측모델은 3lap cleanswim record 변인이 20.26%로 가장 높은 중요도를 나타냈으며 다음으로 3lap cleanswim speed 변인 17.82%, 4lap cleanswim record 변인 13.29%를 나타냈다. DNN 예측모델은 1lap turn out breath 변인이 4.08%로 가장 높은 중요도를 나타냈으며 다음으로 3lap turn out breath 변인 3.43%, 1lap turn out speed 변인 3.27%를 나타냈다.



V. 논의

이 연구는 2017년~2021년 전국수영대회 경영 여자 자유형 200m 경기분석자료(data)를 기반으로 머신러닝을 활용한 경영 경기결과 예측모델을 설계하고, 설계된 예측모델의 성능을 비교·분석하여 경영 경기에 적합한 모델을 제안하는 데 목적이 있다. 또한, 경영 경기결과 예측모델별 경기결과에 영향을 미치는 경기력 변인을 도출하고자 한다. 이 연구의 목적을 달성하기 위해 머신러닝 예측기법인 선형 회귀(Linear Regression), 라쏘 회귀(Lasso Regression), 릿지 회귀(Ridge Regression), 엘라스틱 넷 회귀(Elastic Net Regression), 랜덤 포레스트(Random Forest), 그래디언트 부스팅 머신(GBM: Gradient Boosting Machine), 인공신경망(DNN: Deep Neural Network) 모델을 설계하고 모델별로 예측 성능을 평가하였다. 이에 본 장에서는 선행연구를 바탕으로 앞장에서 제시된 연구결과의 의미를 경영 경기결과 예측모델 간 성능 비교, 경영 경기결과 예측모델 간 경기력 변인의 중요도 비교로 구분하여 다음과 같이 논의하고자 한다.

1. 경영 경기결과 예측모델 간 성능 비교

경영 경기결과 예측을 위해 머신러닝 예측기법인 선형 회귀(Linear Regression), 라쏘 회귀(Lasso Regression), 릿지 회귀(Ridge Regression), 엘라스틱 넷 회귀(Elastic Net Regression), 랜덤 포레스트(Random Forest), 그래디언트 부스팅 머신(GBM: Gradient Boosting Machine), 인공신경망(DNN: Deep Neural Network) 모델을 생성하였다. 모델별 오차가 최소화되는 하이퍼 파라미터를 결정하기 위해 그리드 서치(Grid Search)를 이용하였고, k-fold 교차 검증을 통해 모델별 예측 성능을 비교 분석하였다.

Linear Regression 예측모델의 성능 평가 결과 MSE 값이 0.01217, RMSE 값은

0.10771이 나타났다. Lasso Regression 예측모델의 최적 하이퍼 파라미터는 alpha 값이 0.5757일 때 MSE가 0.0009로 나타났으며 성능 평가 결과 MSE 값은 0.00090, RMSE 값은 0.02683으로 나타났다. Ridge Regression 예측모델의 최적 하이퍼 파라미터는 lambda 값이 0.0302일 때 MSE가 0.0162로 나타났으며 성능 평가 결과 MSE 값은 0.0147, RMSE 값은 0.11541로 나타났다. Elastic Net Regression 예측모델의 최적 하이퍼 파라미터는 alpha 값이 0.5757, lambda 값이 0.0302일 때 MSE가 0.0090으로 나타났으며 성능 평가 결과 MSE 값은 0.00784, RMSE 값은 0.07981로 나타났다. 김홍표(2018)는 PGA Tour 선수의 평균 스코어에 영향을 미치는 요소를 연구하기 위해 Lasso Regression과 Elastic Net Regression을 이용하였다. 이 연구는 97개의 독립변인을 대상으로 변인 선택(Feature Selection)이 가능한 Lasso Regression과 Elastic Net Regression의 적합성과 예측성을 확인하였으며 검증 결괏값이 0.89 수준으로 나타났다. 모델의 최적화를 위해 하이퍼 파라미터인 alpha, lambda 값을 변화하며 진행하였으며 하이퍼 파라미터값이 증가함에 따라 제거되는 변인은 많아졌으나 변인이 줄어들어도 적합도나 예측력에는 큰 변화가 없게 나타났다. 이 연구에서는 다중공선성의 검토 없이 다중공선성을 해결하여 결괏값을 나타낸 Lasso Regression 모델의 유용성을 제안하였다. 이처럼 본 연구에서도 위와 같은 양상의 결과가 도출되었다. 최혜민 등(2015)의 연구에서는 경마 경기의 우승 마를 예측하기 위해 순위기반의 분류분석과 기록기반의 예측분석을 하였다. 분류분석은 로지스틱회귀와 랜덤 포레스트 모델을 이용하였고, 예측분석은 선형회귀와 랜덤 포레스트 모델을 이용하였다. 연구결과 분류분석의 경우 랜덤 포레스트 모델의 예측력이 더 높게 나타났으며 예측분석의 경우 선형회귀모델의 예측력이 높게 나타났다. 본 연구에서 분석된 Random Forest 예측모델의 결과는 하이퍼 파라미터값이 col sample rate per tree 0.45, max depth 16, min rows 1, ntrees 200, sample rate 0.6일 때 MSE 값이 가장 낮게 나타났으며 성능 평가 결과 MSE 값은 0.9560, RMSE 값은 0.9090으로 선형회귀모델보다 낮은 예측값을 보였다. GBM 예측모델의 결과는 하이퍼 파라미터값이 col sample rate 0.75, col sample rate per tree 0.3, learn rate 0.009, max depth 6, min rows 10, ntrees 200, sample rate 0.75일 때 MSE 값이 가장 낮게 나타났으며

성능 평가 결과 MSE 값은 1.86647, RMSE 값은 1.22909로 나타났다. DNN 예측모델의 결과는 하이퍼 파라미터값이 epochs 103.41356, epsilon 0.00010, hidden 50, 25, huber alpha 0.0, input dropout ratio 0.1, L1 0, L2 0, loss는 Quadratic, rho는 0.999, activation은 MaxoutWithDropout일 때 MSE 값이 가장 낮게 나타났으며 성능 평가 결과 MSE 값은 0.23553, RMSE 값은 0.46703으로 나타났다.

임정은 등(2017)은 Linear Regression, Lasso Regression, Ridge Regression, Decision Tree, Bagging, Random Forest, GBM, 주성분회귀, K-최근접 이웃, 인공신경망 모델을 이용하여 PGA 투어에 출전하는 프로 골프 선수의 경기결과를 예측하는 모델을 제안하였다. 연구결과 Random Forest 모델에서 가장 낮은 MSE 값이 나타났으며 Bagging 모델, Decision Tree 모델, 인공신경망 모델, GBM 모델, Linear Regression 모델, Lasso Regression 모델, Ridge Regression 모델, 주성분회귀모델, K-최근접이웃 모델 순으로 예측력이 나타났다. 이철원 등(2021)은 냉동기 에너지소비량을 예측하기 위해 Random Forest 모델과 ANN을 이용하여 결과를 도출하였다. Random Forest 모델과 ANN 모델 모두 출력값과 비교적 상관관계가 낮은 독립변인 일지라도 독립변인의 개수를 증가시킴으로써 예측 성능을 향상할 수 있었다. 하지만 독립변인을 증가시킬 때 인공신경망 모델이 Random Forest 모델보다 더 우수한 예측 성능을 나타내며 독립변인 증가에 따른 예측 성능의 개선 효과는 인공신경망 모델이 더 우수하다고 보고하였다. 이처럼 본 연구에서 분석된 GBM 모델은 Random Forest 모델과 DNN 모델보다 예측 성능이 낮게 나타났으며 DNN 모델은 Random Forest 모델보다 예측 성능이 높게 나타난 것을 확인할 수 있다. DNN 모델은 비선형적이지만 매끄러운 결정경계를 만들기 때문에 Random Forest 모델과 GBM 모델보다 높은 예측 성능을 나타낸 것으로 사료된다.

머신러닝 예측모델의 예측 성능을 비교한 결과 Lasso Regression 예측모델의 MSE(0.00090) 및 RMSE(0.02683)가 가장 낮게 나타나 다른 예측모델보다 예측 성능이 우수한 것으로 나타났다. 다음으로 Elastic Net Regression 예측모델 MSE 0.00784, RMSE 0.07981, Linear Regression 예측모델 MSE 0.01217, RMSE 0.10771, Ridge Regression 예측모델 MSE 0.0147, RMSE 0.11541, DNN 예측모델 MSE

0.23553, RMSE 0.46703, Random Forest 예측모델 MSE 0.9560, RMSE 0.9090, GBM 예측모델 MSE 1.86647, RMSE 1.22909 순으로 나타났다. 모델별 머신러닝 예측 성능을 비교한 결과 비선형회귀모델보다 선형회귀모델의 MSE, RMSE 값이 상대적으로 낮게 나타나며 예측 성능이 높은 것으로 나타났다.

이경문과 황규백(2017)의 연구는 프로야구 선수 연봉을 예측하기 위해 Linear Regression 모델과 Random Forest 모델을 이용하였으며 연구결과 Random Forest 모델이 메이저 리그 시즌 기록 데이터와 연봉 데이터를 기반으로 연봉예측을 하는데 선형회귀분석보다 좋은 결과를 나타낸다고 보고하였다. 한정섭 등(2022)은 Linear Regression, Lasso Regression, Ridge Regression, XGBoost, LightGBM, Random Forest, Support Vector Regression의 머신러닝 기법을 이용하여 KBO 타자의 OPS(On-base Plus Slugging)를 예측하였다. 연구결과 XGBoost 모델에서 R^2 은 0.99740, MSE는 0.00460의 값을 나타내며 머신러닝 모델 중 가장 좋은 성능을 나타냈다. 다음으로 Random Forest, Support Vector Regression, LightGBM, Ridge Regression, Linear Regression, Lasso Regression 모델 순으로 나타났다. 하지만 Wiseman(2016)와 박수민 등(2021)의 선행연구에서는 선형회귀모델에서 높은 예측력을 나타냈다. Wiseman(2016)의 연구는 Linear Regression, Neural Network Regression, Bayesian Linear Regression, Decision Forest Regression, Boosted Decision Tree Regression을 적용하여 PGA 우승 점수를 예측하였으며 모델의 정확도를 평가하기 위해 R^2 값과 MSE 값을 사용하였다. 연구결과 Linear Regression($R^2=0.594$, MSE=2.64)과 Bayesian Linear Regression($R^2=0.592$, MSE=2.64)이 가장 좋은 성능을 나타낸다고 보고하였다. 박수민 등(2021)의 연구는 서울시 아파트 매매가격지수 예측을 위해 Linear Regression, Random Forest, XGBoost, LSTM 모델을 활용하였고 모델의 성능을 평가하기 위해 RMSE 값을 사용하였다. 연구결과 선형회귀모델에서 RMSE 값이 가장 낮게 나타나며 예측력이 가장 좋은 것으로 나타났다. 이러한 결과는 자료의 특성이 연속성을 나타내고 있으면 전통적인 분석기법인 선형회귀모델에서 좋은 예측력을 나타내는 본 연구결과를 뒷받침해준다.

위 내용을 종합해 볼 때 머신러닝 예측기법을 이용한 연구가 다양한 주제로 이루

어지고 있으며 연구결과도 선행연구마다 차이를 나타내고 있다. 이는 분석자료 (data)의 양과 특성, 하이퍼 파라미터 설정에 따라 분석결과가 달라질 수 있음을 의미한다. 따라서 경영 경기결과 예측모델의 예측 성능은 높은 수준을 나타내고 있지만 본 주제의 결과만으로 예측모델의 예측력이 우수하고 단정하는 것은 신중할 필요가 있다고 판단된다.



2) 경영 경기결과 예측모델 간 경기력 변인의 상대적 중요도 비교

이 연구에서 설계된 경영 경기결과 예측모델에 입력된 경기력 변인 간 상대적 중요도를 비교 분석하였다. Linear Regression 예측모델은 4lap cleanswim record 변인이 9.47%로 가장 높은 중요도를 나타냈으며 다음으로 3lap cleanswim record 변인이 8.76%, 2lap cleanswim record 변인이 8.25%를 나타냈다. Lasso Regression 예측모델은 4lap cleanswim record 변인이 15.39%로 가장 높은 중요도를 나타냈으며 다음으로 3lap cleanswim record 변인이 13.86%, 2lap cleanswim record 변인이 11.54%를 나타냈다. Ridge Regression 예측모델은 4lap cleanswim record 변인이 7.69%로 가장 높은 중요도를 나타냈으며 다음으로 3lap cleanswim record 변인이 6.98%, 2lap cleanswim record 변인이 6.32%를 나타냈다. Elastic Net Regression 예측모델은 4lap cleanswim record 변인이 10.97%로 가장 높은 중요도를 나타냈으며 다음으로 3lap cleanswim record 변인이 10.16%, 2lap cleanswim record 변인이 8.37%를 나타냈다. 선형회귀분석을 이용한 예측모델에서는 스트로크 구간의 기록 변인이 모델을 예측하는데 높은 비율을 차지하는 것으로 나타났다.

비선형회귀분석 모델인 Random Forest 예측모델은 3lap cleanswim speed 변인이 28.47%로 가장 높은 중요도를 나타냈으며 다음으로 3lap cleanswim record 변인이 16.72%, 4lap cleanswim record 변인이 13.20%를 나타냈다. GBM 예측모델은 3lap cleanswim record 변인이 20.26%로 가장 높은 중요도를 나타냈으며 다음으로 3lap cleanswim speed 변인이 17.82%, 4lap cleanswim record 변인이 13.29%를 나타냈다. DNN 예측모델은 1lap turn out breath 변인이 4.08%로 가장 높은 중요도를 나타냈으며 다음으로 3lap turn out breath 변인이 3.43%, 1lap turn out speed 변인이 3.27%를 나타냈다. DNN 예측모델은 분석과정에서 많은 변인을 필수적 다루기 때문에 예측에 활용되는 변인이 전체적으로 고르게 기여한 것으로 나타났다. DNN 예측모델을 제외한 Random Forest 예측모델과 GBM 예측모델은 스트로크 구간의 속도와 기록이 모델을 예측하는데 높게 기여한 것으로 나타났다. 이 연구에서 구간별

속도는 구간별 기록에 의해 결정되었다. 즉 구간기록이 빠르면 구간 속도도 빠른 것으로 나타난다. 경영 경기결과 예측모델에서 예측 성능이 가장 높게 나타난 Lasso Regression 예측모델은 스트로크 구간의 기록 변인이(4lap cleanswim record, 3lap cleanswim record, 2lap cleanswim record, 1lap cleanswim record) 49.52%의 중요도를 나타내며 모델을 예측하는데 높은 비율을 차지하는 것으로 나타났다.

Morais 등(2021)은 경영 경기에서 모든 스트로크 구간의 변인(1lap, 2lap, 3lap, 4lap)이 최종 기록에 긍정적이고 중요한 관계가 있다고 보고하였으며 이는 30m(15m~45m 사이의 거리) 구간을 이동하는데 많은 시간을 소요할수록 최종 기록이 더 길어지는 것을 의미한다고 하였다. 또한 4번째 스트로크 구간은 최종 기록과 강한 관계를 보이고 마지막 구간(4lap)의 경기력과 가장 높은 연관성을 나타내는 구간은 3번째 구간(3lap)이라고 하였다. 본 연구의 상관관계 결과에서도 4번째 스트로크 구간기록 변인이 $r=0.889$, $p<0.01$ 을 나타내며 최종 기록과 높은 관계를 나타내는 것을 확인할 수 있었다. 따라서 여자 자유형 200m 경기에서의 스트로크 구간은 최종 기록을 결정하는 주요 요인이며 특히 3번째, 4번째 스트로크 구간은 경기결과를 예측하는 데 중요한 것으로 판단된다.

최근 수영 훈련에서는 스트로크 구간뿐만 아니라 출발, 턴, 종료 구간의 훈련에 집중하고 있는 것을 알 수 있다(Veiga, Roig, 2016; Simbaña-Escobar 등, 2018; Marinho 등, 2020). 경영 경기결과 예측모델에서 턴 구간의 기록은 스트로크 구간의 기록 다음으로 상대적 중요도를 나타내고 있다. Morais 등(2022)의 연구에서는 턴 구간에서 많은 시간을 소비한 선수는 최종 기록이 좋지 않다고 보고하였으며 Marinho 등(2020)은 200m 경기를 레이스 하는데 브레이크아웃(break-out)을 늘려 턴 구간을 최적화하여 스트로크 구간에서의 체력을 절약할 수 있다고 제안하였다. 따라서 턴 구간을 개선하는 것은 최종 기록을 단축하는데 중요할 수 있다고 판단된다.

자유형 200m 수영선수들은 유사한 전략으로 경기를 운영하는 것으로 나타났으며 이것은 종목의 제약에 따라 가능한 최고의 경기력을 나타내고 유지하기 위해 각 구간에서 선수의 개별 특성에 맞게 조정되어야 한다(Huot-Marchand 등, 2005; Morais

등, 2022). 따라서 선수의 특성을 기반으로 주요 대회 준비 및 경기 전략을 맞춤화할 수 있는 훈련 프로그램을 설계해야 한다고 사료된다.



Ⅵ. 결론 및 제언

이 연구는 2017년~2021년 전국수영대회 경영 여자 자유형 200m 경기분석자료(data)를 기반으로 머신러닝을 활용한 경영 경기결과 예측모델을 설계하고, 설계된 예측모델의 성능을 비교·분석하여 경영 경기에 적합한 모델을 제안하는 데 목적이 있다. 이를 통해 머신러닝을 이용한 경영 경기결과 예측의 활용 가능성을 확인하고자 하였다. 이 연구의 목적을 달성하기 위해 머신러닝 예측기법인 선형 회귀(Linear Regression), 라쏘 회귀(Lasso Regression), 릿지 회귀(Ridge Regression), 엘라스틱 넷 회귀(Elastic Net Regression), 랜덤 포레스트(Random Forest), 그래디언트 부스팅 머신(GBM: Gradient Boosting Machine), 인공신경망(DNN: Deep Neural Network) 모델을 설계하고 모델별로 예측 성능을 평가하였다. 또한, 경영 경기 예측모델 간 경기력 변인의 상대적 중요도를 확인하였다.

첫째, 머신러닝 예측모델의 예측 성능을 비교한 결과 Lasso Regression 예측모델이 가장 우수한 것으로 나타났으며 다음으로 Elastic Net Regression 예측모델, Linear Regression 예측모델, Ridge Regression 예측모델, DNN 예측모델, Random Forest 예측모델, GBM 예측모델 순으로 예측력이 나타났다.

둘째, Linear Regression 예측모델, Lasso Regression 예측모델, Ridge Regression 예측모델, Elastic Net Regression 예측모델에 입력된 변인 간 상대적 중요도는 스트로크 구간의 기록 변인인 4lap cleanswim record, 3lap cleanswim record, 2lap cleanswim record 변인에서 높은 비율을 나타냈으며 선형회귀분석을 이용한 예측모델에서는 스트로크 구간의 기록 변인이 모델을 예측하는데 높은 비율을 차지하는 것으로 나타났다. Random Forest 예측모델, GBM 예측모델에 입력된 변인 간 상대적 중요도는 3lap cleanswim speed, 3lap cleanswim record, 4lap cleanswim record 변인에서 높은 비율을 나타냈으며 비선형회귀분석을 이용한 예측모델에서는 스트로크 구간의 속도와 기록이 모델을 예측하는데 높게 기여한 것으로 나타났다. DNN

예측모델은 분석과정에서 많은 변인을 필수적 다루기 때문에 예측에 활용되는 변인이 전체적으로 고르게 기여한 것으로 나타났다.

결론적으로 경영 여자 자유형 200m 경기분석기록(data)을 기반으로 경영 경기결과 예측을 위한 머신러닝 예측기법의 활용이 가능했으며, Lasso Regression 예측모델이 가장 적합한 것으로 나타났다. 또한, 여자 자유형 200m 경기에서 스트로크 구간의 기록은 최종 기록을 예측하는 주요 요인인 것으로 판단된다.

이 연구는 머신러닝 예측기법을 수영 경영 경기결과 예측에 적용했다는 점에서 의미하는 바가 크다. 추후 머신러닝 예측기법의 스포츠데이터 활용 가능성을 높이고 다양한 분석을 가능하게 할 것으로 기대한다. 하지만 머신러닝은 근본적인 한계점을 가지고 있으며 학습을 기반으로 분석이 이루어지기 때문에 분석자료(data)의 양과 특성에 따라 머신러닝의 적용이 어려울 수 있다. 또한 머신러닝을 활용하기 위해서는 하이퍼 파라미터를 설정해야 하는데 하이퍼 파라미터를 결정하는 명확한 기준이 없어 본 연구에서 적용된 하이퍼 파라미터가 최적의 하이퍼 파라미터라고 단정 지을 수 없으며 하이퍼 파라미터의 설정에 따라 분석결과가 달라질 수 있는 한계점을 가지고 있다.

경영 경기결과는 0.01초로 경기결과가 바뀌기 때문에 다차원적인 변인들의 고려가 필요하다는 선행연구에 따라 가능한 많은 경기력 변인을 연구에 포함하였다. 하지만 경영 경기는 하나의 경기력 요인이 아닌 다양한 요인들로부터 영향을 받기 때문에 본 연구에서 나타난 경영 경기결과에 영향을 미치는 경기력 변인이 가장 중요한 요인이라고 결정하기에는 무리가 있다.

따라서 추후 연구에서는 자료의 범위를 넓혀 다양한 경기분석자료의 확보와 머신러닝 모델마다 하이퍼 파라미터의 인과관계를 연구하여 예측기법 모델을 최적화하기 위한 연구가 진행되어야 한다. 또한 경영 경기의 영법, 거리별로 경기력에 영향을 주는 변인들에 대한 추가 발굴을 통하여 본 연구결과와 비교할 필요가 있으며 경영 경기분석자료를 기반으로 본 연구에서 다뤄지지 않은 머신러닝 모델을 활용한 지속적인 연구가 필요하다.

참고문헌

- 강병관, 박성제(2021). 배드민턴 남자단식 경기패턴 분석. 코칭능력개발지, 23(3), 180-189.
- 김민석(2007). 스프린터 수영선수의 시합 전 경쟁불안이 경기력에 미치는 영향. 미간행 석사학위논문, 동아대학교 대학원.
- 김민석(2022). 층화추출을 활용한 기계학습 사례연구. 미간행 석사학위논문, 고려대학교 대학원.
- 김동섭(2020). 기계학습 모형의 설명가능성에 관한 연구 : 미국 주택담보대출 자료를 중심으로. 미간행 박사학위논문, 건국대학교 대학원.
- 김완수, 양대승(2018). 태권도 경기규칙 개정에 따른 선취득점 시 경기결과 분석. 국기원태권도연구, 9(2), 237-254.
- 김성진(2006). 경륜 경기력 영향 심리요인의 구조 탐색. 한국스포츠심리학회, 12(3), 91-103.
- 김세중, 허종관, 이강웅(2010). 한국프로농구경기 2000~2009시즌 PLAY-OFF 진출팀의 경기 자료 데이터를 이용한 경기분석. 한국체육과학회지, 19(4), 1405-1411.
- 김세형, 강상조, 박재현, 김혜진(2008). 한국프로농구 경기기록 분석에 의한 승패결정요인. 한국체육측정평가학회지, 10(1), 1-12.
- 김승진(2022). 인공지능을 이용한 광산란 미세먼지 측정방법 신뢰도 향상 연구. 미간행 박사학위논문, 세종대학교 대학원.
- 김은하, 추병주, 최현수, 오미애, 김용대, 김경범, 이창수, 신동민, 주명원, 위세아(2015). 사회보장정보시스템을 활용한 복지 사각지대 발굴방안 연구. 사회보장정보원.
- 김지웅(2020). 머신러닝을 활용한 핸드볼 경기결과 시각화 및 예측력 비교. 미간행 박사학위논문, 국민대학교 대학원.
- 김주학, 노갑택, 박종성, 이원희(2007). 신경망분석을 이용한 축구경기 승, 패 예측모형 개발-2006 독일 월드컵 대회를 중심으로. 체육과학연구, 18(4), 54-63.
- 김주학, 최형준(2009). 테니스 경기결과 예측 시뮬레이터 설계를 위한 기초연구. 한국체육학회지, 48(4), 593-601.
- 김현주(2022). 딥러닝을 이용한 일별 천연가스 수요 예측. 미간행 석사학위논문, 성균관대학교 대학원.
- 김홍표(2018). LASSO 회귀분석을 활용한 PGA선수 스코어에 영향을 미치는 요인 분석. 미간행 석사학위논문, 연세대학교 공학대학원.
- 김혜진, 박재현, 강상조(2006). 테니스 경기의 득점 및 실점상황에 의한 승패결정 경기내용분석. 한국체육측정평가학회지, 8(2), 43-57.
- 남기연, 정현(2019). 스포츠 빅데이터 활용을 위한 방안. 스포츠엔터테인먼트와 법, 22(4), 85-101.
- 박성진, 황영찬(2017). 4차 산업혁명의 스포츠 현장 적용을 위한 탐색적 연구. 한국체육학회지,

56(4), 397-413.

북경수, 유재수(2017). 4차 산업혁명에서 빅데이터. Communications of the Korean Institute of Information Scientists and Engineers, 35(6), 29-39.

박제영(2003). 측정평가: 2002-2003 시즌 한국프로 농구경기의 승패 요인 분석. 한국체육학회지, 42(5), 793-800.

박수민, 이연재, 박주현, 박주아, 임진섭, 김현희(2021). 머신러닝과 딥러닝을 이용한 부동산 지수 예측 모델 비교. 한국정보처리학회 학술대회논문집, 28(2), 1156-1159.

박지희(2020). 블록 주기화(Block Periodization) 프로그램이 엘리트 수영 선수들의 체력 및 경기력에 미치는 영향. 미간행 박사학위논문, 제주대학교 대학원.

서보영(2016). 수영선수들의 인지된 경기력 개념 구조 탐색. 미간행 박사학위논문, 성균관대학교 일반대학원.

박해용, 이기청(2001). 테니스 남자단식 경기의 승패 요인 분석. 한국유산소운동과학회지, 5(2), 37-46.

배경태, 김창재(2016). 인공신경망의 은닉층 최적화를 통한 농산물 가격예측 모델. 한국정보기술학회 논문지, 14(12), 161-169.

배성원(2019). 머신 러닝을 이용한 주택 가격 예측력 비교. 미간행 박사학위논문, 단국대학교 대학원.

서유화, 김은희(2021). 기계학습을 이용한 유선 액세스 네트워크의 에너지 소모량 예측 모델. 한국정보전자통신기술학회 논문지, 14(1), 14-21.

손승현(2021). 머신러닝 기반 공동주택 개발사업의 분양가 예측 모델. 미간행 박사학위논문, 경희대학교 대학원.

양민정(2018). 경영 수준별 경기내용 비교. 미간행 석사학위논문, 단국대학교 대학원.

양민정, 최형준(2019). 경영 100m에서 나타난 경기력 수준별 비교: 제98회 전국체육대회 경기 중심으로. 한국체육과학회지, 28(6), 1173-1186.

오미애, 최현수, 김수현, 장준혁, 진재현, 천미경(2017). 기계학습 (Machine Learning) 기반 사회보장 빅데이터 분석 및 예측모형 연구. 한국보건사회연구원.

오운학, 김한, 윤재섭, 이종석(2014). 데이터마이닝을 활용한 한국프로야구 승패 예측모형 수립에 관한 연구. 대한산업공학회지, 40(1), 8-17.

양준석, 박성제(2016). 평행좌표계(Parallel Coordinates)를 활용한 배드민턴 단식경기 시각화 연구. 체육과학연구, 27(3), 631-643.

윤미연(2015). 청소년 수영선수의 taper가 혈중 스트레스 호르몬, 염증성 사이토카인 및 경기력에 미치는 영향. 미간행 박사학위논문, 동아대학교.

윤석훈(1996). 경영 100M 경주시 구간별 운동학적 분석. 미간행 석사학위논문, 한국체육대학교 대학원.

이기봉, 이영석, 이기청(2004). 측정평가: 국내 남자 테니스 단식 경기의 승패 요인과 득점 과정 분석.

- 한국체육학회지, 43(3), 903-911.
- 이경문, 황규백 (2017). 선형 회귀 및 랜덤 포레스트를 이용한 개인 기록 기반 프로야구 선수 연봉 예측. 한국정보과학회 학술발표논문집, 1842-1844.
- 이태형(2019). 인공지능망을 활용한 주택가격지수 예측에 관한 연구: 서울 주택 가격 지수를 중심으로. 미간행 박사학위논문, 중앙대학교 대학원.
- 임승희(2018). 수영 트랙 스타트 동작 시 양발 스탠스에 따른 운동학적 비교분석. 미간행 박사학위논문, 충남대학교 대학원.
- 이용성(2022). 딥러닝 및 관련 기법을 활용한 철근 가격 장단기예측. 미간행 박사학위논문, 건국대학교 대학원.
- 이용수, 김용래(2018). 2018 러시아 월드컵 아시아 예선경기 중 한국 축구 국가대표팀의 포지션 및 15분 단위 경기력 비교 분석. 한국체육과학회지, 27(1), 825-839.
- 이은경(2020). 고등학교 수영선수의 슬럼프 극복을 위한 심리기술훈련 사례연구. 미간행 박사학위논문, 영남대학교 대학원.
- 임정은, 임용인, 송종우(2017). PGA 투어의 골프 스코어 예측 및 분석. 30(1), 41-55.
- 이재학(2020). 대학 수영선수 발목관절 가동범위와 하지 근력이 돌핀 킥에 미치는 영향. 미간행 박사학위논문, 차의과학대학교 일반대학원.
- 이제승, 이태현, 최현수, 이성노(2019). 한국과 미국 수영선수의 경영 경기력 비교: 접영을 중심으로. 한국스포츠학회지, 17(3), 1177-1188.
- 이철원, 성남철, 최원창. (2021). Python 을 이용한 냉동기 에너지소비량 예측 모델의 성능 개선 및 비교 평가. 한국건축친환경설비학회 논문집, 15(3), 252-264.
- 이희창(2008). 수영선수들의 체력요인에 의한 경기력 결정요인 분석. 미간행 박사학위논문, 우석대학교.
- 정광채, 이재봉, 박재현(2010). 태권도 경기 운영형태에 따른 경기내용분석. 대한무도학회지, 12(3), 89-100.
- 정진배(2013). 주니어 골프선수의 운동몰입 및 스포츠 자신감이 경기력 향상에 미치는 영향. 한국체육과학회지, 22(6), 49-60.
- 정철수, 이병두, 육현철, 이웅기, 김현준, 은선덕, 이영석, 김용운(2003). 50m, 100m, 200m 수영 자 유형 종목의 경기분석. 한국체육학회지, 42(4), 729-737.
- 정현학, 최영임, 이상원(2016). 4차 산업혁명과 보건산업 패러다임의 변화. KHIDI BRIEF, 215, 1-28.
- 조민정, 박인구, 천영진(2020). 남자 프로 배구 경기 중 듀스 상황의 세트포인트에서 득점 유형 분석. 한국체육과학회지, 29(1), 823-832.
- 조현진(2018). 머신러닝을 활용한 특허 품질 예측. 미간행 석사학위논문, 서울과학기술대학교.
- 지무엽(2017). 엘리트 수영선수의 경기력 향상을 위한 코어트레이닝 기반 지상훈련 프로그램의 효과.

- 미간행 석사학위 논문, 한국체육대학교 대학원.
- 천영진(2009). 남자 프로배구 경기에서 득점 과정을 이용한 경기분석. 한국사회체육학회지, 38(2), 1305-1312.
- 천영진(2019). 남자 프로 배구 경기 기록이 V-리그 시즌 승률에 미치는 영향. Journal of The Korean Data Analysis Society, 21(1), 375-383.
- 최완용, 홍성진, 최형준(2009). 태권도 경기 공격 패턴 분석. 체육과학연구, 20(4), 767-777.
- 최형준, 김주학(2006). 인공 신경망(Artificial Neural Network)을 이용한 2005년도 월드컵 테니스 회의 경기결과 예측에 관한 연구. 한국체육학 회지, 45(3), 459-467.
- 최형준(2009). 2002, 2006년 축구 월드컵대회를 통한 경기력 분석에 관한 연구. 한국체육측정평가 학회지, 11(2), 41-51.
- 최형준(2010). 테니스 그랜드슬램 대회의 선수 수준별 경기력 비교. 한국체육학회지, 49(3), 385-393.
- 최형준, 정연성(2010). 배드민턴 혼합복식의 승패 및 랠리 상황별 경기분석. 한국체육측정평가학회지, 12(3), 103-113.
- 최형준, 조은형, 김웅준, 한도령(2011). 테니스 경기분석을 위한 주요 분석인자의 일관성 탐색. 한국체육측정평가학회지, 13(3), 65-75.
- 최형준, 이윤수(2019). 축구 월드컵대회의 경기기록 기반 경기결과 예측. 한국체육과학회지, 28(1), 1317-1325.
- 최형준, 양민정(2020). 인공지능을 활용한 자유형 100m 경기분석 자료의 거리 구간, 선수 및 구간 내 순위에 대한 분류 예측 비교. 한국체육과학회지, 29(3), 989-998.
- 최혜민, 황나영, 황찬경, 송종우(2015). 서울 경마 경기 우승마 예측 모형 연구. 응용통계연구, 28(6), 1133-1146.
- 한정섭, 정다현, 김성준. (2022). 머신러닝을 활용한 빅데이터 분석을 통해 KBO 타자의 OPS 예측. 차세대융합기술학회논문지, 6(1), 12-18.
- 황준일, 어수주, 김효식(2012). 경영선수들의 상지 근력과 스트로크 변인이 수영 속도에 미치는 영향. 한국체육과학회지, 21(4), 1079- 1088.
- Abbiss, C. R., & Laursen, P. B. (2008). Describing and understanding pacing strategies during athletic competition. Sports medicine, 38(3), 239-252.
- Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications, 39(3), 3446-3453.
- Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological assessment, 6(4), 284.
- Cossor, J., & Mason, B. (2001). Swim start performances at the Sydney 2000 Olympic Games.

In ISBS-Conference Proceedings Archive.

- Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10), 78-87.
- Da Silva, J. K., Enes, A. A., Sotomaior, B. B., Barbosa, M. A. R., De Souza, R. O., & Osiecki, R. (2020). Analysis of the performance of finalist swimming athletes in Olympic games: reaction time, partial time, speed, and final time. *Journal of Physical Education and Sport*, 20(2), 539-545.
- Edelmann-Nusser, J., Hohmann, A., & Henneberg, B. (2002). Modeling and prediction of competitive performance in swimming upon neural networks. *European Journal of Sport Science*, 2(2), 1-10.
- Escobar, D. S., Hellard, P., Pyne, D. B., & Seifert, L. (2018). Functional role of movement and performance variability: Adaptation of front crawl swimmers to competitive swimming constraints. *Journal of applied biomechanics*, 34(1), 53-64.
- Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.
- Haljand, R., & Saagpakk, R. (1994). Swimming competition analysis of the European Sprint Swimming Championship. LEN, Stavanger.
- Hopkins, W. G., & Hewson, D. J. (2001). Variability of competitive performance of distance runners. *Medicine and Science in Sports and Exercise*, 33, 1588-1592.
- Hughes, M., & Franks, I. M. (2004). *Notational Analysis of Sport 2nd Edition-better systems for improving coaching and performance*. London: E. & FN Spon.
- Huot-Marchand, F., Nesi, X., Sidney, M., Alberty, M., & Pelayo, P. (2005). Swimming: Variations of stroking parameters associated with 200 m competitive performance improvement in top-standard front crawl swimmers. *Sports Biomechanics*, 4(1), 89-100.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.
- Kumar, G. P., & Venkataram, P. (1997). Security management architecture for access control to network resources. *IEE Proceedings-computers and Digital Techniques*, 144(6), 362-370.
- Minsky, M., & Papert, S. (1969). *An introduction to computational geometry*. Cambridge tiass., HIT, 479, 480.
- Maszczyk, A., Roczniok, R., Czuba, M., Zajac, A., Waśkiewicz, Z., Mikołajec, K., & Stanula, A. (2012). Application of regression and neural models to predict competitive swimming performance. *Perceptual and Motor Skills*, 114(2), 610-626.

- Maszczyk, A., Goł aś, A., Pietraszewski, P., Rocznik, R., Zając, A., & Stanula, A. (2014). Application of neural and regression models in sports results prediction. *Procedia-Social and Behavioral Sciences*, 117, 482-487.
- Marinho, D. A., Barbosa, T. M., Neiva, H. P., Silva, A. J., & Morais, J. E. (2020). Comparison of the start, turn and finish performance of elite swimmers in 100 m and 200 m races. *Journal of sports science & medicine*, 19(2), 397.
- Morais, J. E., Marinho, D. A., Arellano, R., & Barbosa, T. M. (2019). Start and turn performances of elite sprinters at the 2016 European Championships in swimming. *Sports Biomechanics*, 18(1), 100-114.
- Morais, J. E., Barbosa, T. M., Forte, P., Pinto, J. N., & Marinho, D. A. (2021). Assessment of the inter-lap stability and relationship between the race time and start, clean swim, turn and finish variables in elite male junior swimmers 200m freestyle. *Sports Biomechanics*, 1-14.
- Morais, J. E., Barbosa, T. M., Lopes, T., & Marinho, D. A. (2022). Race level comparison and variability analysis of 100 m freestyle sprinters competing in the 2019 European championships. *International Journal of Performance Analysis in Sport*, 1-14.
- Robertson, E. Y., Pyne, D. B., Hopkins, W. G., & Anson, J. M. (2009). Analysis of lap times in international swimming competitions. *Journal of Sports Sciences*, 27(4), 387-395.
- Simbaña-Escobar, D., Hellard, P., & Seifert, L. (2018). Modelling stroking parameters in competitive sprint swimming: Understanding inter-and intra-lap variability to assess pacing management. *Human Movement Science*, 61, 219-230.
- Veiga, S., & Roig, A. (2016). Underwater and surface strategies of 200m world level swimmers. *Journal of Sports Sciences*, 34(8), 766-771.
- Wiseman, O. (2016). Using Machine Learning to Predict the Winning Score of Professional Golf Events on the PGA Tour (Doctoral dissertation, Dublin, National College of Ireland).
- Xie, J., Xu, J., Nie, C., & Nie, Q. (2017). Machine learning of swimming data via wisdom of crowd and regression analysis. *Mathematical Biosciences & Engineering*, 14(2), 511.

부록

<부록 1> 실제값과 예측모델별 도출된 예측값의 기록 차이 결과

선형 회귀(Linear Regression) 모델 예측 결과(단위: 초)

Linear Regression				
no	Actual record	Predicted record	Error	Error Rate(%)
1	121.35	121.28	0.07	0.057
2	121.55	121.56	-0.01	-0.008
3	122.21	122.16	0.05	0.04
4	122.93	122.95	-0.02	-0.016
5	123.22	123.23	-0.01	-0.008
6	123.4	123.4	0	0
7	124.34	124.34	0	0
8	129.15	129.2	-0.05	-0.038
9	121.23	121.15	0.08	0.065
10	121.26	121.19	0.07	0.057
11	122.51	122.43	0.08	0.065
12	122.64	122.63	0.01	0.008
13	124.19	124.22	-0.03	-0.024
14	126.64	126.66	-0.02	-0.015
15	135.62	135.59	0.03	0.022
16	122.11	122.05	0.06	0.049
17	122.46	122.45	0.01	0.008
18	123.18	123.22	-0.04	-0.032
19	123.29	123.28	0.01	0.008
20	123.59	123.62	-0.03	-0.024
21	124.37	124.34	0.03	0.024
22	124.98	125.01	-0.03	-0.024
23	120.71	120.72	-0.01	-0.008
24	121.09	121.01	0.08	0.066
25	121.85	121.86	-0.01	-0.008
26	122.97	122.88	0.09	0.073
27	122.97	122.9	0.07	0.056
28	123.1	123.13	-0.03	-0.024
29	124.16	124.19	-0.03	-0.024
30	124.23	124.27	-0.04	-0.032
31	119.86	119.78	0.08	0.066

Linear Regression				
no	Actual record	Predicted record	Error	Error Rate(%)
32	122.28	122.27	0.01	0.008
33	122.37	122.38	-0.01	-0.008
34	122.71	122.67	0.04	0.032
35	123.16	123.19	-0.03	-0.024
36	123.73	123.82	-0.09	-0.072
37	124.9	124.95	-0.05	-0.04
38	125.06	125.1	-0.04	-0.031
39	118.41	118.35	0.06	0.05
40	122.36	122.31	0.05	0.04
41	122.9	122.87	0.03	0.024
42	123.32	123.31	0.01	0.008
43	123.4	123.4	0	0
44	123.44	123.42	0.02	0.016
45	123.67	123.67	0	0
46	127.7	127.69	0.01	0.007
47	121.2	121.11	0.09	0.074
48	121.73	121.69	0.04	0.032
49	121.75	121.76	-0.01	-0.008
50	121.78	121.73	0.05	0.041
51	122.57	122.51	0.06	0.048
52	122.57	122.52	0.05	0.04
53	122.78	122.77	0.01	0.008
54	123.19	123.19	0	0
55	121.03	120.99	0.04	0.033
56	121.66	121.68	-0.02	-0.016
57	121.8	121.76	0.04	0.032
58	122.27	122.28	-0.01	-0.008
59	122.34	122.32	0.02	0.016
60	122.52	122.58	-0.06	-0.048
61	123.49	123.47	0.02	0.016
62	126.78	126.79	-0.01	-0.007
63	120.49	120.45	0.04	0.033
64	121.83	121.82	0.01	0.008
65	122.69	122.61	0.08	0.065
66	123.09	123.05	0.04	0.032
67	123.13	123.14	-0.01	-0.008
68	124.45	124.35	0.1	0.08
69	124.95	124.94	0.01	0.008
70	125.36	125.41	-0.05	-0.039

Linear Regression				
no	Actual record	Predicted record	Error	Error Rate(%)
71	120.95	120.98	-0.03	-0.024
72	121.25	121.3	-0.05	-0.041
73	121.74	121.71	0.03	0.024
74	122.95	122.94	0.01	0.008
75	123.69	123.66	0.03	0.024
76	123.78	123.77	0.01	0.008
77	125.35	125.35	0	0
78	127.6	127.6	0	0
79	122.06	122.12	-0.06	-0.049
80	122.27	122.18	0.09	0.073
81	122.66	122.67	-0.01	-0.008
82	122.96	122.85	0.11	0.089
83	124.25	124.26	-0.01	-0.008
84	125.97	125.94	0.03	0.023
85	126.41	126.33	0.08	0.063
86	126.52	126.52	0	0
87	122.05	122.07	-0.02	-0.016
88	122.18	122.16	0.02	0.016
89	123.3	123.35	-0.05	-0.04
90	123.66	123.61	0.05	0.04
91	123.73	123.77	-0.04	-0.032
92	124.17	124.15	0.02	0.016
93	124.49	124.4	0.09	0.072
94	124.51	124.5	0.01	0.008
95	124.38	124.39	-0.01	-0.008
96	126.84	126.77	0.07	0.055
97	128.21	128.24	-0.03	-0.023
98	129.49	129.49	0	0
99	131.99	132.08	-0.09	-0.068
100	133.08	133.17	-0.09	-0.067
101	133.84	133.73	0.11	0.082
102	133.98	134.04	-0.06	-0.044
103	136.81	136.72	0.09	0.065
104	138.76	138.73	0.03	0.021
105	139.2	139.17	0.03	0.021
106	124.37	124.34	0.03	0.024
107	122.53	122.43	0.1	0.081
108	123.29	123.29	0	0
109	123.5	123.47	0.03	0.024

Linear Regression				
no	Actual record	Predicted record	Error	Error Rate(%)
110	125.51	125.49	0.02	0.015
111	126.36	126.38	-0.02	-0.015
112	123.45	123.44	0.01	0.008
113	124.32	124.22	0.1	0.08
114	126.19	126.21	-0.02	-0.015
115	127.35	127.36	-0.01	-0.007
116	127.59	127.62	-0.03	-0.023
117	127.87	127.87	0	0
118	129.55	129.44	0.11	0.084
119	130.7	130.74	-0.04	-0.03
120	125.86	125.86	0	0
121	126.37	126.41	-0.04	-0.031
122	127.61	127.59	0.02	0.015
123	127.66	127.68	-0.02	-0.015
124	128.43	128.38	0.05	0.038
125	129.14	129.26	-0.12	-0.092
126	130.13	130.21	-0.08	-0.061
127	130.91	130.95	-0.04	-0.03
128	131.15	131.11	0.04	0.03
129	143.26	143.05	0.21	0.146
130	123.77	123.77	0	0
131	123.91	123.94	-0.03	-0.024
132	124.55	124.53	0.02	0.016
133	125.96	125.94	0.02	0.015
134	126.04	126.08	-0.04	-0.031
135	127.33	127.38	-0.05	-0.039
136	128.82	128.83	-0.01	-0.007
137	128.54	128.8	-0.26	-0.202
138	129.59	129.69	-0.1	-0.077
139	130.37	130.44	-0.07	-0.053
140	130.79	130.82	-0.03	-0.022
141	131.99	132.03	-0.04	-0.03
142	132.05	132.03	0.02	0.015
143	133.88	133.92	-0.04	-0.029
144	122.59	122.48	0.11	0.089
145	124.53	124.49	0.04	0.032
146	124.84	124.86	-0.02	-0.016
147	129.95	130.03	-0.08	-0.061
148	133.24	133.24	0	0

Linear Regression				
no	Actual record	Predicted record	Error	Error Rate(%)
149	122.01	122.01	0	0
150	123.56	123.51	0.05	0.04
151	124.27	124.3	-0.03	-0.024
152	124.92	124.92	0	0
153	126.05	126.1	-0.05	-0.039
154	126.07	126	0.07	0.055
155	127.19	127.26	-0.07	-0.055
156	127.68	127.72	-0.04	-0.031
157	121.94	121.96	-0.02	-0.016
158	122.42	122.39	0.03	0.024
159	123.22	123.24	-0.02	-0.016
160	123.27	123.3	-0.03	-0.024
161	125.65	125.69	-0.04	-0.031
162	127.47	127.54	-0.07	-0.054
163	127.74	127.78	-0.04	-0.031
164	128.41	128.5	-0.09	-0.07
Mean	125.27	125.26	0.004	0.003

라쏘 회귀(Lasso Regression) 모델 예측 결과(단위: 초)

Lasso Regression				
no	Actual record	Predicted record	Error	Error Rate(%)
1	121.35	121.33	0.02	0.016
2	121.55	121.53	0.02	0.016
3	122.21	122.2	0.01	0.008
4	122.93	122.91	0.02	0.016
5	123.22	123.21	0.01	0.008
6	123.4	123.39	0.01	0.008
7	124.34	124.32	0.02	0.016
8	129.15	129.12	0.03	0.023
9	121.23	121.21	0.02	0.016
10	121.26	121.26	0	0
11	122.51	122.49	0.02	0.016
12	122.64	122.63	0.01	0.008
13	124.19	124.19	0	0
14	126.64	126.64	0	0
15	135.62	135.62	0	0
16	122.11	122.13	-0.02	-0.016
17	122.46	122.47	-0.01	-0.008
18	123.18	123.18	0	0
19	123.29	123.29	0	0
20	123.59	123.59	0	0
21	124.37	124.35	0.02	0.016
22	124.98	125	-0.02	-0.016
23	120.71	120.72	-0.01	-0.008
24	121.09	121.08	0.01	0.008
25	121.85	121.83	0.02	0.016
26	122.97	122.97	0	0
27	122.97	122.96	0.01	0.008
28	123.1	123.09	0.01	0.008
29	124.16	124.17	-0.01	-0.008
30	124.23	124.23	0	0
31	119.86	119.86	0	0
32	122.28	122.28	0	0
33	122.37	122.37	0	0
34	122.71	122.72	-0.01	-0.008
35	123.16	123.15	0.01	0.008
36	123.73	123.74	-0.01	-0.008
37	124.9	124.88	0.02	0.016

Lasso Regression				
no	Actual record	Predicted record	Error	Error Rate(%)
38	125.06	125.04	0.02	0.015
39	118.41	118.39	0.02	0.016
40	122.36	122.35	0.01	0.008
41	122.9	122.89	0.01	0.008
42	123.32	123.31	0.01	0.008
43	123.4	123.41	-0.01	-0.008
44	123.44	123.41	0.03	0.024
45	123.67	123.69	-0.02	-0.016
46	127.7	127.68	0.02	0.015
47	121.2	121.2	0	0
48	121.73	121.71	0.02	0.016
49	121.75	121.76	-0.01	-0.008
50	121.78	121.78	0	0
51	122.57	122.57	0	0
52	122.57	122.57	0	0
53	122.78	122.78	0	0
54	123.19	123.2	-0.01	-0.008
55	121.03	121.03	0	0
56	121.66	121.67	-0.01	-0.008
57	121.8	121.81	-0.01	-0.008
58	122.27	122.26	0.01	0.008
59	122.34	122.34	0	0
60	122.52	122.5	0.02	0.016
61	123.49	123.48	0.01	0.008
62	126.78	126.78	0	0
63	120.49	120.49	0	0
64	121.83	121.84	-0.01	-0.008
65	122.69	122.68	0.01	0.008
66	123.09	123.09	0	0
67	123.13	123.12	0.01	0.008
68	124.45	124.44	0.01	0.008
69	124.95	124.96	-0.01	-0.008
70	125.36	125.36	0	0
71	120.95	120.96	-0.01	-0.008
72	121.25	121.25	0	0
73	121.74	121.74	0	0
74	122.95	122.95	0	0
75	123.69	123.66	0.03	0.024

Lasso Regression				
no	Actual record	Predicted record	Error	Error Rate(%)
76	123.78	123.77	0.01	0.008
77	125.35	125.32	0.03	0.023
78	127.6	127.58	0.02	0.015
79	122.06	122.07	-0.01	-0.008
80	122.27	122.25	0.02	0.016
81	122.66	122.66	0	0
82	122.96	122.95	0.01	0.008
83	124.25	124.25	0	0
84	125.97	125.94	0.03	0.023
85	126.41	126.41	0	0
86	126.52	126.53	-0.01	-0.007
87	122.05	122.06	-0.01	-0.008
88	122.18	122.19	-0.01	-0.008
89	123.3	123.3	0	0
90	123.66	123.65	0.01	0.008
91	123.73	123.73	0	0
92	124.17	124.17	0	0
93	124.49	124.48	0.01	0.008
94	124.51	124.5	0.01	0.008
95	124.38	124.35	0.03	0.024
96	126.84	126.84	0	0
97	128.21	128.2	0.01	0.007
98	129.49	129.49	0	0
99	131.99	131.99	0	0
100	133.08	133.08	0	0
101	133.84	133.82	0.02	0.014
102	133.98	133.98	0	0
103	136.81	136.81	0	0
104	138.76	138.77	-0.01	-0.007
105	139.2	139.2	0	0
106	124.37	124.37	0	0
107	122.53	122.52	0.01	0.008
108	123.29	123.28	0.01	0.008
109	123.5	123.5	0	0
110	125.51	125.52	-0.01	-0.007
111	126.36	126.35	0.01	0.007
112	123.45	123.43	0.02	0.016
113	124.32	124.32	0	0

Lasso Regression				
no	Actual record	Predicted record	Error	Error Rate(%)
114	126.19	126.19	0	0
115	127.35	127.35	0	0
116	127.59	127.58	0.01	0.007
117	127.87	127.87	0	0
118	129.55	129.54	0.01	0.007
119	130.7	130.69	0.01	0.007
120	125.86	125.86	0	0
121	126.37	126.36	0.01	0.007
122	127.61	127.61	0	0
123	127.66	127.64	0.02	0.015
124	128.43	128.41	0.02	0.015
125	129.14	129.14	0	0
126	130.13	130.12	0.01	0.007
127	130.91	130.89	0.02	0.015
128	131.15	131.17	-0.02	-0.015
129	143.26	143.25	0.01	0.006
130	123.77	123.75	0.02	0.016
131	123.91	123.92	-0.01	-0.008
132	124.55	124.54	0.01	0.008
133	125.96	125.94	0.02	0.015
134	126.04	126.02	0.02	0.015
135	127.33	127.33	0	0
136	128.82	128.82	0	0
137	128.54	128.72	-0.18	-0.14
138	129.59	129.59	0	0
139	130.37	130.37	0	0
140	130.79	130.77	0.02	0.015
141	131.99	131.99	0	0
142	132.05	132.03	0.02	0.015
143	133.88	133.88	0	0
144	122.59	122.56	0.03	0.024
145	124.53	124.51	0.02	0.016
146	124.84	124.84	0	0
147	129.95	129.94	0.01	0.007
148	133.24	133.22	0.02	0.015
149	122.01	122.01	0	0
150	123.56	123.54	0.02	0.016
151	124.27	124.27	0	0

Lasso Regression				
no	Actual record	Predicted record	Error	Error Rate(%)
152	124.92	124.91	0.01	0.008
153	126.05	126.05	0	0
154	126.07	126.06	0.01	0.007
155	127.19	127.19	0	0
156	127.68	127.66	0.02	0.015
157	121.94	121.95	-0.01	-0.008
158	122.42	122.42	0	0
159	123.22	123.23	-0.01	-0.008
160	123.27	123.28	-0.01	-0.008
161	125.65	125.6	0.05	0.039
162	127.47	127.45	0.02	0.015
163	127.74	127.73	0.01	0.007
164	128.41	128.38	0.03	0.023
Mean	125.27	125.26	0.004	0.003

릿지 회귀(Ridge Regression) 모델 예측 결과(단위: 초)

Ridge Regression				
no	Actual record	Predicted record	Error	Error Rate(%)
1	121.35	121.34	0.01	0.008
2	121.55	121.56	-0.01	-0.008
3	122.21	122.21	0	0
4	122.93	122.99	-0.06	-0.048
5	123.22	123.18	0.04	0.032
6	123.4	123.37	0.03	0.024
7	124.34	124.3	0.04	0.032
8	129.15	129.18	-0.03	-0.023
9	121.23	121.18	0.05	0.041
10	121.26	121.2	0.06	0.049
11	122.51	122.49	0.02	0.016
12	122.64	122.57	0.07	0.057
13	124.19	124.34	-0.15	-0.12
14	126.64	126.67	-0.03	-0.023
15	135.62	135.41	0.21	0.154
16	122.11	122.08	0.03	0.024
17	122.46	122.5	-0.04	-0.032
18	123.18	123.31	-0.13	-0.105
19	123.29	123.26	0.03	0.024
20	123.59	123.69	-0.1	-0.08
21	124.37	124.35	0.02	0.016
22	124.98	125.16	-0.18	-0.144
23	120.71	120.68	0.03	0.024
24	121.09	121.01	0.08	0.066
25	121.85	121.82	0.03	0.024
26	122.97	122.88	0.09	0.073
27	122.97	122.87	0.1	0.081
28	123.1	123.2	-0.1	-0.081
29	124.16	124.34	-0.18	-0.144
30	124.23	124.21	0.02	0.016
31	119.86	119.71	0.15	0.125
32	122.28	122.25	0.03	0.024
33	122.37	122.28	0.09	0.073
34	122.71	122.73	-0.02	-0.016
35	123.16	123.2	-0.04	-0.032
36	123.73	123.75	-0.02	-0.016
37	124.9	124.98	-0.08	-0.064

Ridge Regression				
no	Actual record	Predicted record	Error	Error Rate(%)
38	125.06	125.07	-0.01	-0.007
39	118.41	118.2	0.21	0.177
40	122.36	122.4	-0.04	-0.032
41	122.9	122.87	0.03	0.024
42	123.32	123.41	-0.09	-0.072
43	123.4	123.33	0.07	0.056
44	123.44	123.43	0.01	0.008
45	123.67	123.71	-0.04	-0.032
46	127.7	127.81	-0.11	-0.086
47	121.2	121.12	0.08	0.066
48	121.73	121.75	-0.02	-0.016
49	121.75	121.74	0.01	0.008
50	121.78	121.81	-0.03	-0.024
51	122.57	122.51	0.06	0.048
52	122.57	122.61	-0.04	-0.032
53	122.78	122.79	-0.01	-0.008
54	123.19	123.23	-0.04	-0.032
55	121.03	120.98	0.05	0.041
56	121.66	121.67	-0.01	-0.008
57	121.8	121.72	0.08	0.065
58	122.27	122.27	0	0
59	122.34	122.38	-0.04	-0.032
60	122.52	122.49	0.03	0.024
61	123.49	123.46	0.03	0.024
62	126.78	126.75	0.03	0.023
63	120.49	120.44	0.05	0.041
64	121.83	121.83	0	0
65	122.69	122.61	0.08	0.065
66	123.09	123.04	0.05	0.04
67	123.13	123.25	-0.12	-0.097
68	124.45	124.36	0.09	0.072
69	124.95	124.91	0.04	0.032
70	125.36	125.31	0.05	0.039
71	120.95	120.86	0.09	0.074
72	121.25	121.23	0.02	0.016
73	121.74	121.64	0.1	0.082
74	122.95	122.97	-0.02	-0.016
75	123.69	123.76	-0.07	-0.056

Ridge Regression				
no	Actual record	Predicted record	Error	Error Rate(%)
76	123.78	123.74	0.04	0.032
77	125.35	125.3	0.05	0.039
78	127.6	127.56	0.04	0.031
79	122.06	122.08	-0.02	-0.016
80	122.27	122.11	0.16	0.13
81	122.66	122.58	0.08	0.065
82	122.96	122.96	0	0
83	124.25	124.29	-0.04	-0.032
84	125.97	125.91	0.06	0.047
85	126.41	126.43	-0.02	-0.015
86	126.52	126.47	0.05	0.039
87	122.05	122.02	0.03	0.024
88	122.18	122.14	0.04	0.032
89	123.3	123.34	-0.04	-0.032
90	123.66	123.6	0.06	0.048
91	123.73	123.75	-0.02	-0.016
92	124.17	124.19	-0.02	-0.016
93	124.49	124.45	0.04	0.032
94	124.51	124.51	0	0
95	124.38	124.33	0.05	0.04
96	126.84	126.98	-0.14	-0.11
97	128.21	128.31	-0.1	-0.077
98	129.49	129.43	0.06	0.046
99	131.99	132.01	-0.02	-0.015
100	133.08	133.23	-0.15	-0.112
101	133.84	133.69	0.15	0.112
102	133.98	134	-0.02	-0.014
103	136.81	136.66	0.15	0.109
104	138.76	138.65	0.11	0.079
105	139.2	138.9	0.3	0.215
106	124.37	124.35	0.02	0.016
107	122.53	122.53	0	0
108	123.29	123.31	-0.02	-0.016
109	123.5	123.42	0.08	0.064
110	125.51	125.45	0.06	0.047
111	126.36	126.31	0.05	0.039
112	123.45	123.46	-0.01	-0.008
113	124.32	124.3	0.02	0.016

Ridge Regression				
no	Actual record	Predicted record	Error	Error Rate(%)
114	126.19	126.19	0	0
115	127.35	127.39	-0.04	-0.031
116	127.59	127.69	-0.1	-0.078
117	127.87	127.79	0.08	0.062
118	129.55	129.49	0.06	0.046
119	130.7	130.69	0.01	0.007
120	125.86	125.94	-0.08	-0.063
121	126.37	126.4	-0.03	-0.023
122	127.61	127.78	-0.17	-0.133
123	127.66	127.71	-0.05	-0.039
124	128.43	128.45	-0.02	-0.015
125	129.14	129.21	-0.07	-0.054
126	130.13	130.21	-0.08	-0.061
127	130.91	130.91	0	0
128	131.15	131.16	-0.01	-0.007
129	143.26	142.74	0.52	0.362
130	123.77	123.73	0.04	0.032
131	123.91	123.95	-0.04	-0.032
132	124.55	124.56	-0.01	-0.008
133	125.96	125.91	0.05	0.039
134	126.04	126.12	-0.08	-0.063
135	127.33	127.36	-0.03	-0.023
136	128.82	128.82	0	0
137	128.54	128.77	-0.23	-0.178
138	129.59	129.68	-0.09	-0.069
139	130.37	130.46	-0.09	-0.069
140	130.79	130.68	0.11	0.084
141	131.99	132.03	-0.04	-0.03
142	132.05	131.99	0.06	0.045
143	133.88	133.86	0.02	0.014
144	122.59	122.53	0.06	0.048
145	124.53	124.44	0.09	0.072
146	124.84	124.8	0.04	0.032
147	129.95	130.05	-0.1	-0.076
148	133.24	133.25	-0.01	-0.007
149	122.01	122.02	-0.01	-0.008
150	123.56	123.52	0.04	0.032
151	124.27	124.27	0	0

Ridge Regression				
no	Actual record	Predicted record	Error	Error Rate(%)
152	124.92	124.99	-0.07	-0.056
153	126.05	126.04	0.01	0.007
154	126.07	126.07	0	0
155	127.19	127.21	-0.02	-0.015
156	127.68	127.79	-0.11	-0.086
157	121.94	121.98	-0.04	-0.032
158	122.42	122.4	0.02	0.016
159	123.22	123.24	-0.02	-0.016
160	123.27	123.29	-0.02	-0.016
161	125.65	125.69	-0.04	-0.031
162	127.47	127.49	-0.02	-0.015
163	127.74	127.95	-0.21	-0.164
164	128.41	128.57	-0.16	-0.124
Mean	125.27	125.26	0.005	0.003

엘라스틱 넷 회귀(Elastic Net Regression) 모델 예측 결과(단위: 초)

Elastic Net Regression				
no	Actual record	Predicted record	Error	Error Rate(%)
1	121.35	121.32	0.03	0.024
2	121.55	121.54	0.01	0.008
3	122.21	122.23	-0.02	-0.016
4	122.93	122.98	-0.05	-0.04
5	123.22	123.21	0.01	0.008
6	123.4	123.37	0.03	0.024
7	124.34	124.3	0.04	0.032
8	129.15	129.16	-0.01	-0.007
9	121.23	121.23	0	0
10	121.26	121.21	0.05	0.041
11	122.51	122.47	0.04	0.032
12	122.64	122.6	0.04	0.032
13	124.19	124.25	-0.06	-0.048
14	126.64	126.66	-0.02	-0.015
15	135.62	135.38	0.24	0.176
16	122.11	122.09	0.02	0.016
17	122.46	122.49	-0.03	-0.024
18	123.18	123.27	-0.09	-0.073
19	123.29	123.26	0.03	0.024
20	123.59	123.68	-0.09	-0.072
21	124.37	124.38	-0.01	-0.008
22	124.98	125.06	-0.08	-0.064
23	120.71	120.64	0.07	0.057
24	121.09	121	0.09	0.074
25	121.85	121.85	0	0
26	122.97	122.9	0.07	0.056
27	122.97	122.96	0.01	0.008
28	123.1	123.13	-0.03	-0.024
29	124.16	124.25	-0.09	-0.072
30	124.23	124.17	0.06	0.048
31	119.86	119.76	0.1	0.083
32	122.28	122.26	0.02	0.016
33	122.37	122.29	0.08	0.065
34	122.71	122.71	0	0
35	123.16	123.16	0	0
36	123.73	123.8	-0.07	-0.056
37	124.9	124.94	-0.04	-0.032

Elastic Net Regression				
no	Actual record	Predicted record	Error	Error Rate(%)
38	125.06	125.03	0.03	0.023
39	118.41	118.33	0.08	0.067
40	122.36	122.35	0.01	0.008
41	122.9	122.91	-0.01	-0.008
42	123.32	123.35	-0.03	-0.024
43	123.4	123.38	0.02	0.016
44	123.44	123.48	-0.04	-0.032
45	123.67	123.69	-0.02	-0.016
46	127.7	127.65	0.05	0.039
47	121.2	121.16	0.04	0.033
48	121.73	121.68	0.05	0.041
49	121.75	121.7	0.05	0.041
50	121.78	121.78	0	0
51	122.57	122.51	0.06	0.048
52	122.57	122.56	0.01	0.008
53	122.78	122.78	0	0
54	123.19	123.22	-0.03	-0.024
55	121.03	120.98	0.05	0.041
56	121.66	121.69	-0.03	-0.024
57	121.8	121.79	0.01	0.008
58	122.27	122.3	-0.03	-0.024
59	122.34	122.38	-0.04	-0.032
60	122.52	122.58	-0.06	-0.048
61	123.49	123.53	-0.04	-0.032
62	126.78	126.72	0.06	0.047
63	120.49	120.47	0.02	0.016
64	121.83	121.83	0	0
65	122.69	122.63	0.06	0.048
66	123.09	123.13	-0.04	-0.032
67	123.13	123.18	-0.05	-0.04
68	124.45	124.49	-0.04	-0.032
69	124.95	124.97	-0.02	-0.016
70	125.36	125.43	-0.07	-0.055
71	120.95	120.91	0.04	0.033
72	121.25	121.24	0.01	0.008
73	121.74	121.75	-0.01	-0.008
74	122.95	122.96	-0.01	-0.008
75	123.69	123.78	-0.09	-0.072

Elastic Net Regression				
no	Actual record	Predicted record	Error	Error Rate(%)
76	123.78	123.8	-0.02	-0.016
77	125.35	125.36	-0.01	-0.007
78	127.6	127.56	0.04	0.031
79	122.06	122.05	0.01	0.008
80	122.27	122.24	0.03	0.024
81	122.66	122.64	0.02	0.016
82	122.96	122.91	0.05	0.04
83	124.25	124.28	-0.03	-0.024
84	125.97	125.99	-0.02	-0.015
85	126.41	126.42	-0.01	-0.007
86	126.52	126.47	0.05	0.039
87	122.05	122.09	-0.04	-0.032
88	122.18	122.16	0.02	0.016
89	123.3	123.35	-0.05	-0.04
90	123.66	123.64	0.02	0.016
91	123.73	123.78	-0.05	-0.04
92	124.17	124.19	-0.02	-0.016
93	124.49	124.48	0.01	0.008
94	124.51	124.55	-0.04	-0.032
95	124.38	124.38	0	0
96	126.84	126.9	-0.06	-0.047
97	128.21	128.24	-0.03	-0.023
98	129.49	129.42	0.07	0.054
99	131.99	131.97	0.02	0.015
100	133.08	133.09	-0.01	-0.007
101	133.84	133.69	0.15	0.112
102	133.98	133.97	0.01	0.007
103	136.81	136.63	0.18	0.131
104	138.76	138.54	0.22	0.158
105	139.2	138.87	0.33	0.237
106	124.37	124.39	-0.02	-0.016
107	122.53	122.51	0.02	0.016
108	123.29	123.33	-0.04	-0.032
109	123.5	123.51	-0.01	-0.008
110	125.51	125.52	-0.01	-0.007
111	126.36	126.36	0	0
112	123.45	123.46	-0.01	-0.008
113	124.32	124.3	0.02	0.016

Elastic Net Regression				
no	Actual record	Predicted record	Error	Error Rate(%)
114	126.19	126.21	-0.02	-0.015
115	127.35	127.41	-0.06	-0.047
116	127.59	127.64	-0.05	-0.039
117	127.87	127.84	0.03	0.023
118	129.55	129.52	0.03	0.023
119	130.7	130.69	0.01	0.007
120	125.86	125.93	-0.07	-0.055
121	126.37	126.41	-0.04	-0.031
122	127.61	127.69	-0.08	-0.062
123	127.66	127.7	-0.04	-0.031
124	128.43	128.45	-0.02	-0.015
125	129.14	129.19	-0.05	-0.038
126	130.13	130.18	-0.05	-0.038
127	130.91	130.88	0.03	0.022
128	131.15	131.1	0.05	0.038
129	143.26	142.75	0.51	0.355
130	123.77	123.77	0	0
131	123.91	123.94	-0.03	-0.024
132	124.55	124.57	-0.02	-0.016
133	125.96	125.97	-0.01	-0.007
134	126.04	126.1	-0.06	-0.047
135	127.33	127.38	-0.05	-0.039
136	128.82	128.84	-0.02	-0.015
137	128.54	128.83	-0.29	-0.225
138	129.59	129.62	-0.03	-0.023
139	130.37	130.41	-0.04	-0.03
140	130.79	130.73	0.06	0.045
141	131.99	131.97	0.02	0.015
142	132.05	131.96	0.09	0.068
143	133.88	133.8	0.08	0.059
144	122.59	122.6	-0.01	-0.008
145	124.53	124.52	0.01	0.008
146	124.84	124.83	0.01	0.008
147	129.95	129.99	-0.04	-0.03
148	133.24	133.22	0.02	0.015
149	122.01	122.03	-0.02	-0.016
150	123.56	123.56	0	0
151	124.27	124.27	0	0

Elastic Net Regression				
no	Actual record	Predicted record	Error	Error Rate(%)
152	124.92	124.94	-0.02	-0.016
153	126.05	126.05	0	0
154	126.07	126.07	0	0
155	127.19	127.21	-0.02	-0.015
156	127.68	127.74	-0.06	-0.046
157	121.94	121.92	0.02	0.016
158	122.42	122.38	0.04	0.032
159	123.22	123.2	0.02	0.016
160	123.27	123.29	-0.02	-0.016
161	125.65	125.65	0	0
162	127.47	127.43	0.04	0.031
163	127.74	127.84	-0.1	-0.078
164	128.41	128.47	-0.06	-0.046
Mean	125.27	125.26	0.004	0.003

랜덤 포레스트(Random Forest) 모델 예측 결과(단위: 초)

Random Forest				
no	Actual record	Predicted record	Error	Error Rate(%)
1	121.35	121.84	-0.49	-0.403
2	121.55	121.65	-0.1	-0.082
3	122.21	122.21	0	0
4	122.93	122.54	0.39	0.317
5	123.22	123.34	-0.12	-0.097
6	123.4	123.41	-0.01	-0.008
7	124.34	124.11	0.23	0.184
8	129.15	128.85	0.3	0.232
9	121.23	121.63	-0.4	-0.329
10	121.26	122	-0.74	-0.61
11	122.51	122.52	-0.01	-0.008
12	122.64	122.54	0.1	0.081
13	124.19	123.63	0.56	0.45
14	126.64	126.24	0.4	0.315
15	135.62	135.22	0.4	0.294
16	122.11	122.58	-0.47	-0.384
17	122.46	122.58	-0.12	-0.097
18	123.18	123.32	-0.14	-0.113
19	123.29	123.29	0	0
20	123.59	123.4	0.19	0.153
21	124.37	124.36	0.01	0.008
22	124.98	124.78	0.2	0.16
23	120.71	120.73	-0.02	-0.016
24	121.09	121.5	-0.41	-0.338
25	121.85	122.08	-0.23	-0.188
26	122.97	122.91	0.06	0.048
27	122.97	123.31	-0.34	-0.276
28	123.1	123.05	0.05	0.04
29	124.16	123.94	0.22	0.177
30	124.23	123.87	0.36	0.289
31	119.86	120.73	-0.87	-0.725
32	122.28	122.13	0.15	0.122
33	122.37	122.4	-0.03	-0.024
34	122.71	122.52	0.19	0.154
35	123.16	123.09	0.07	0.056
36	123.73	123.51	0.22	0.177
37	124.9	124.53	0.37	0.296

Random Forest				
no	Actual record	Predicted record	Error	Error Rate(%)
38	125.06	124.64	0.42	0.335
39	118.41	119.37	-0.96	-0.81
40	122.36	122.18	0.18	0.147
41	122.9	122.87	0.03	0.024
42	123.32	123.29	0.03	0.024
43	123.4	123.31	0.09	0.072
44	123.44	123.46	-0.02	-0.016
45	123.67	123.67	0	0
46	127.7	128.19	-0.49	-0.383
47	121.2	121.52	-0.32	-0.264
48	121.73	121.63	0.1	0.082
49	121.75	122.16	-0.41	-0.336
50	121.78	121.86	-0.08	-0.065
51	122.57	122.55	0.02	0.016
52	122.57	122.68	-0.11	-0.089
53	122.78	122.62	0.16	0.13
54	123.19	123.32	-0.13	-0.105
55	121.03	121.31	-0.28	-0.231
56	121.66	121.93	-0.27	-0.221
57	121.8	122.09	-0.29	-0.238
58	122.27	122.24	0.03	0.024
59	122.34	122.38	-0.04	-0.032
60	122.52	122.51	0.01	0.008
61	123.49	124.47	-0.98	-0.793
62	126.78	126.79	-0.01	-0.007
63	120.49	120.8	-0.31	-0.257
64	121.83	122.33	-0.5	-0.41
65	122.69	122.88	-0.19	-0.154
66	123.09	123.15	-0.06	-0.048
67	123.13	123.18	-0.05	-0.04
68	124.45	124.53	-0.08	-0.064
69	124.95	124.68	0.27	0.216
70	125.36	125.14	0.22	0.175
71	120.95	121.27	-0.32	-0.264
72	121.25	121.5	-0.25	-0.206
73	121.74	122.27	-0.53	-0.435
74	122.95	122.97	-0.02	-0.016
75	123.69	123.7	-0.01	-0.008

Random Forest				
no	Actual record	Predicted record	Error	Error Rate(%)
76	123.78	123.92	-0.14	-0.113
77	125.35	125.06	0.29	0.231
78	127.6	127.76	-0.16	-0.125
79	122.06	122.34	-0.28	-0.229
80	122.27	122.44	-0.17	-0.139
81	122.66	122.95	-0.29	-0.236
82	122.96	123.23	-0.27	-0.219
83	124.25	123.9	0.35	0.281
84	125.97	126.47	-0.5	-0.396
85	126.41	126.86	-0.45	-0.355
86	126.52	126.69	-0.17	-0.134
87	122.05	122.48	-0.43	-0.352
88	122.18	122.1	0.08	0.065
89	123.3	123.33	-0.03	-0.024
90	123.66	123.56	0.1	0.08
91	123.73	123.66	0.07	0.056
92	124.17	124.38	-0.21	-0.169
93	124.49	124.67	-0.18	-0.144
94	124.51	124.45	0.06	0.048
95	124.38	124.62	-0.24	-0.192
96	126.84	126.82	0.02	0.015
97	128.21	128.17	0.04	0.031
98	129.49	129.52	-0.03	-0.023
99	131.99	131.61	0.38	0.287
100	133.08	132.9	0.18	0.135
101	133.84	133.64	0.2	0.149
102	133.98	134.03	-0.05	-0.037
103	136.81	135.99	0.82	0.599
104	138.76	137.6	1.16	0.835
105	139.2	138.36	0.84	0.603
106	124.37	124.36	0.01	0.008
107	122.53	122.65	-0.12	-0.097
108	123.29	123.39	-0.1	-0.081
109	123.5	123.49	0.01	0.008
110	125.51	125.57	-0.06	-0.047
111	126.36	126.12	0.24	0.189
112	123.45	123.71	-0.26	-0.21
113	124.32	124.39	-0.07	-0.056

Random Forest				
no	Actual record	Predicted record	Error	Error Rate(%)
114	126.19	126.14	0.05	0.039
115	127.35	127.38	-0.03	-0.023
116	127.59	127.63	-0.04	-0.031
117	127.87	127.99	-0.12	-0.093
118	129.55	129.77	-0.22	-0.169
119	130.7	131.14	-0.44	-0.336
120	125.86	125.79	0.07	0.055
121	126.37	126.44	-0.07	-0.055
122	127.61	127.81	-0.2	-0.156
123	127.66	127.6	0.06	0.046
124	128.43	128.18	0.25	0.194
125	129.14	129.55	-0.41	-0.317
126	130.13	130.61	-0.48	-0.368
127	130.91	130.78	0.13	0.099
128	131.15	131.49	-0.34	-0.259
129	143.26	141.56	1.7	1.186
130	123.77	123.8	-0.03	-0.024
131	123.91	123.82	0.09	0.072
132	124.55	124.56	-0.01	-0.008
133	125.96	125.95	0.01	0.007
134	126.04	126.07	-0.03	-0.023
135	127.33	127.32	0.01	0.007
136	128.82	128.81	0.01	0.007
137	128.54	128.49	0.05	0.038
138	129.59	129.52	0.07	0.054
139	130.37	130.05	0.32	0.245
140	130.79	131.63	-0.84	-0.642
141	131.99	131.88	0.11	0.083
142	132.05	132.04	0.01	0.007
143	133.88	133.76	0.12	0.089
144	122.59	122.99	-0.4	-0.326
145	124.53	124.55	-0.02	-0.016
146	124.84	124.82	0.02	0.016
147	129.95	129.91	0.04	0.03
148	133.24	133.22	0.02	0.015
149	122.01	122.07	-0.06	-0.049
150	123.56	123.6	-0.04	-0.032
151	124.27	124.39	-0.12	-0.096

Random Forest				
no	Actual record	Predicted record	Error	Error Rate(%)
152	124.92	124.73	0.19	0.152
153	126.05	125.84	0.21	0.166
154	126.07	125.81	0.26	0.206
155	127.19	126.96	0.23	0.18
156	127.68	127.42	0.26	0.203
157	121.94	122.06	-0.12	-0.098
158	122.42	122.56	-0.14	-0.114
159	123.22	123.08	0.14	0.113
160	123.27	123.34	-0.07	-0.056
161	125.65	125	0.65	0.517
162	127.47	126.74	0.73	0.572
163	127.74	127.62	0.12	0.093
164	128.41	128.29	0.12	0.093
Mean	125.27	125.28	-0.016	-0.017

그래디언트 부스팅 머신(GBM) 모델 예측 결과(단위: 초)

GBM				
no	Actual record	Predicted record	Error	Error Rate(%)
1	121.35	122.38	-1.03	-0.848
2	121.55	122.54	-0.99	-0.814
3	122.21	122.75	-0.54	-0.441
4	122.93	123.02	-0.09	-0.073
5	123.22	123.58	-0.36	-0.292
6	123.4	123.63	-0.23	-0.186
7	124.34	124.28	0.06	0.048
8	129.15	127.82	1.33	1.029
9	121.23	122.43	-1.2	-0.989
10	121.26	122.65	-1.39	-1.146
11	122.51	123.01	-0.5	-0.408
12	122.64	122.78	-0.14	-0.114
13	124.19	123.85	0.34	0.273
14	126.64	125.88	0.76	0.6
15	135.62	132.58	3.04	2.241
16	122.11	122.75	-0.64	-0.524
17	122.46	122.94	-0.48	-0.391
18	123.18	123.59	-0.41	-0.332
19	123.29	123.63	-0.34	-0.275
20	123.59	123.71	-0.12	-0.097
21	124.37	124.95	-0.58	-0.466
22	124.98	124.81	0.17	0.136
23	120.71	121.86	-1.15	-0.952
24	121.09	122.02	-0.93	-0.768
25	121.85	122.65	-0.8	-0.656
26	122.97	123.13	-0.16	-0.13
27	122.97	123.74	-0.77	-0.626
28	123.1	123.33	-0.23	-0.186
29	124.16	124.17	-0.01	-0.008
30	124.23	123.93	0.3	0.241
31	119.86	121.69	-1.83	-1.526
32	122.28	122.78	-0.5	-0.408
33	122.37	122.79	-0.42	-0.343
34	122.71	122.92	-0.21	-0.171
35	123.16	123.07	0.09	0.073
36	123.73	123.94	-0.21	-0.169
37	124.9	124.39	0.51	0.408

GBM				
no	Actual record	Predicted record	Error	Error Rate(%)
38	125.06	124.38	0.68	0.543
39	118.41	121.55	-3.14	-2.651
40	122.36	122.58	-0.22	-0.179
41	122.9	123.11	-0.21	-0.17
42	123.32	123.63	-0.31	-0.251
43	123.4	123.73	-0.33	-0.267
44	123.44	123.83	-0.39	-0.315
45	123.67	124.08	-0.41	-0.331
46	127.7	126.91	0.79	0.618
47	121.2	122.24	-1.04	-0.858
48	121.73	122.34	-0.61	-0.501
49	121.75	122.88	-1.13	-0.928
50	121.78	122.34	-0.56	-0.459
51	122.57	122.69	-0.12	-0.097
52	122.57	123.01	-0.44	-0.358
53	122.78	122.95	-0.17	-0.138
54	123.19	123.51	-0.32	-0.259
55	121.03	122.16	-1.13	-0.933
56	121.66	122.34	-0.68	-0.558
57	121.8	122.59	-0.79	-0.648
58	122.27	122.7	-0.43	-0.351
59	122.34	122.63	-0.29	-0.237
60	122.52	123.02	-0.5	-0.408
61	123.49	124.64	-1.15	-0.931
62	126.78	126.21	0.57	0.449
63	120.49	121.91	-1.42	-1.178
64	121.83	122.97	-1.14	-0.935
65	122.69	123.38	-0.69	-0.562
66	123.09	123.43	-0.34	-0.276
67	123.13	123.6	-0.47	-0.381
68	124.45	124.59	-0.14	-0.112
69	124.95	124.86	0.09	0.072
70	125.36	125	0.36	0.287
71	120.95	122.03	-1.08	-0.892
72	121.25	122.27	-1.02	-0.841
73	121.74	123	-1.26	-1.034
74	122.95	123.3	-0.35	-0.284
75	123.69	123.98	-0.29	-0.234

GBM				
no	Actual record	Predicted record	Error	Error Rate(%)
76	123.78	124.19	-0.41	-0.331
77	125.35	125.17	0.18	0.143
78	127.6	128.83	-1.23	-0.963
79	122.06	122.85	-0.79	-0.647
80	122.27	122.99	-0.72	-0.588
81	122.66	123.53	-0.87	-0.709
82	122.96	123.48	-0.52	-0.422
83	124.25	124.14	0.11	0.088
84	125.97	126.32	-0.35	-0.277
85	126.41	126.47	-0.06	-0.047
86	126.52	126.25	0.27	0.213
87	122.05	123.04	-0.99	-0.811
88	122.18	122.56	-0.38	-0.311
89	123.3	123.55	-0.25	-0.202
90	123.66	123.92	-0.26	-0.21
91	123.73	123.9	-0.17	-0.137
92	124.17	124.57	-0.4	-0.322
93	124.49	124.86	-0.37	-0.297
94	124.51	124.48	0.03	0.024
95	124.38	124.52	-0.14	-0.112
96	126.84	126.46	0.38	0.299
97	128.21	127.44	0.77	0.6
98	129.49	128.95	0.54	0.417
99	131.99	130.73	1.26	0.954
100	133.08	131.72	1.36	1.021
101	133.84	132.34	1.5	1.12
102	133.98	133.32	0.66	0.492
103	136.81	133.83	2.98	2.178
104	138.76	133.84	4.92	3.545
105	139.2	133.91	5.29	3.8
106	124.37	124.5	-0.13	-0.104
107	122.53	123.04	-0.51	-0.416
108	123.29	123.56	-0.27	-0.218
109	123.5	123.79	-0.29	-0.234
110	125.51	125.43	0.08	0.063
111	126.36	125.96	0.4	0.316
112	123.45	123.95	-0.5	-0.405
113	124.32	124.55	-0.23	-0.185

GBM				
no	Actual record	Predicted record	Error	Error Rate(%)
114	126.19	126.22	-0.03	-0.023
115	127.35	127.25	0.1	0.078
116	127.59	127.09	0.5	0.391
117	127.87	127.51	0.36	0.281
118	129.55	129.02	0.53	0.409
119	130.7	130.58	0.12	0.091
120	125.86	126.04	-0.18	-0.143
121	126.37	126.34	0.03	0.023
122	127.61	127.26	0.35	0.274
123	127.66	127.15	0.51	0.399
124	128.43	128.1	0.33	0.256
125	129.14	128.54	0.6	0.464
126	130.13	129.96	0.17	0.13
127	130.91	129.95	0.96	0.733
128	131.15	130.88	0.27	0.205
129	143.26	133.91	9.35	6.526
130	123.77	124.17	-0.4	-0.323
131	123.91	123.94	-0.03	-0.024
132	124.55	124.78	-0.23	-0.184
133	125.96	126.04	-0.08	-0.063
134	126.04	125.72	0.32	0.253
135	127.33	127.05	0.28	0.219
136	128.82	128.31	0.51	0.395
137	128.54	128.14	0.4	0.311
138	129.59	129.19	0.4	0.308
139	130.37	129.44	0.93	0.713
140	130.79	130.7	0.09	0.068
141	131.99	131.23	0.76	0.575
142	132.05	132.05	0	0
143	133.88	132.91	0.97	0.724
144	122.59	123.6	-1.01	-0.823
145	124.53	124.82	-0.29	-0.232
146	124.84	125.1	-0.26	-0.208
147	129.95	129.29	0.66	0.507
148	133.24	132.94	0.3	0.225
149	122.01	122.68	-0.67	-0.549
150	123.56	123.74	-0.18	-0.145
151	124.27	124.83	-0.56	-0.45

GBM				
no	Actual record	Predicted record	Error	Error Rate(%)
152	124.92	124.97	-0.05	-0.04
153	126.05	125.74	0.31	0.245
154	126.07	125.76	0.31	0.245
155	127.19	126.71	0.48	0.377
156	127.68	126.95	0.73	0.571
157	121.94	122.29	-0.35	-0.287
158	122.42	123.24	-0.82	-0.669
159	123.22	123.41	-0.19	-0.154
160	123.27	123.65	-0.38	-0.308
161	125.65	125.06	0.59	0.469
162	127.47	126.41	1.06	0.831
163	127.74	127.01	0.73	0.571
164	128.41	127.48	0.93	0.724
Mean	125.27	125.27	0.002	-0.021

인공신경망(DNN) 모델 예측 결과(단위: 초)

DNN				
no	Actual record	Predicted record	Error	Error Rate(%)
1	121.35	121.48	-0.13	-0.107
2	121.55	121.44	0.11	0.09
3	122.21	122.16	0.05	0.04
4	122.93	122.59	0.34	0.276
5	123.22	123.43	-0.21	-0.17
6	123.4	123.26	0.14	0.113
7	124.34	124.19	0.15	0.12
8	129.15	127.88	1.27	0.983
9	121.23	120.19	1.04	0.857
10	121.26	120.52	0.74	0.61
11	122.51	122.29	0.22	0.179
12	122.64	122.17	0.47	0.383
13	124.19	123.27	0.92	0.74
14	126.64	126.46	0.18	0.142
15	135.62	134.23	1.39	1.024
16	122.11	122.28	-0.17	-0.139
17	122.46	122.01	0.45	0.367
18	123.18	123.17	0.01	0.008
19	123.29	123.02	0.27	0.218
20	123.59	123.54	0.05	0.04
21	124.37	123.98	0.39	0.313
22	124.98	124.53	0.45	0.36
23	120.71	120.88	-0.17	-0.14
24	121.09	120.71	0.38	0.313
25	121.85	121.25	0.6	0.492
26	122.97	122.66	0.31	0.252
27	122.97	122.77	0.2	0.162
28	123.1	123.42	-0.32	-0.259
29	124.16	123.74	0.42	0.338
30	124.23	123.67	0.56	0.45
31	119.86	119.36	0.5	0.417
32	122.28	121.75	0.53	0.433
33	122.37	121.56	0.81	0.661
34	122.71	122.15	0.56	0.456
35	123.16	122.89	0.27	0.219
36	123.73	123.56	0.17	0.137
37	124.9	125	-0.1	-0.08

DNN				
no	Actual record	Predicted record	Error	Error Rate(%)
38	125.06	124.98	0.08	0.063
39	118.41	117.98	0.43	0.363
40	122.36	122.14	0.22	0.179
41	122.9	122.63	0.27	0.219
42	123.32	123.09	0.23	0.186
43	123.4	123.33	0.07	0.056
44	123.44	122.69	0.75	0.607
45	123.67	123.38	0.29	0.234
46	127.7	127.94	-0.24	-0.187
47	121.2	121.45	-0.25	-0.206
48	121.73	121.65	0.08	0.065
49	121.75	122.18	-0.43	-0.353
50	121.78	121.3	0.48	0.394
51	122.57	122.66	-0.09	-0.073
52	122.57	122.79	-0.22	-0.179
53	122.78	122.88	-0.1	-0.081
54	123.19	123.31	-0.12	-0.097
55	121.03	121.11	-0.08	-0.066
56	121.66	121.68	-0.02	-0.016
57	121.8	121.78	0.02	0.016
58	122.27	121.97	0.3	0.245
59	122.34	121.79	0.55	0.449
60	122.52	122.63	-0.11	-0.089
61	123.49	122.48	1.01	0.817
62	126.78	126.41	0.37	0.291
63	120.49	120.53	-0.04	-0.033
64	121.83	121.67	0.16	0.131
65	122.69	122.24	0.45	0.366
66	123.09	123.19	-0.1	-0.081
67	123.13	123.39	-0.26	-0.211
68	124.45	123.67	0.78	0.626
69	124.95	124.46	0.49	0.392
70	125.36	125.12	0.24	0.191
71	120.95	121.04	-0.09	-0.074
72	121.25	121.38	-0.13	-0.107
73	121.74	121.76	-0.02	-0.016
74	122.95	123.22	-0.27	-0.219
75	123.69	123.07	0.62	0.501

DNN				
no	Actual record	Predicted record	Error	Error Rate(%)
76	123.78	123.66	0.12	0.096
77	125.35	124.58	0.77	0.614
78	127.6	127.21	0.39	0.305
79	122.06	121.29	0.77	0.63
80	122.27	122.23	0.04	0.032
81	122.66	122.28	0.38	0.309
82	122.96	122.89	0.07	0.056
83	124.25	123.89	0.36	0.289
84	125.97	124.7	1.27	1.008
85	126.41	126.65	-0.24	-0.189
86	126.52	126.65	-0.13	-0.102
87	122.05	121.9	0.15	0.122
88	122.18	122.08	0.1	0.081
89	123.3	123.21	0.09	0.072
90	123.66	124.12	-0.46	-0.371
91	123.73	123.52	0.21	0.169
92	124.17	124.38	-0.21	-0.169
93	124.49	124.74	-0.25	-0.2
94	124.51	124.54	-0.03	-0.024
95	124.38	123.89	0.49	0.393
96	126.84	125.89	0.95	0.748
97	128.21	128.16	0.05	0.038
98	129.49	129.4	0.09	0.069
99	131.99	131.35	0.64	0.484
100	133.08	133.39	-0.31	-0.232
101	133.84	133.62	0.22	0.164
102	133.98	133.79	0.19	0.141
103	136.81	135.82	0.99	0.723
104	138.76	138.4	0.36	0.259
105	139.2	138.39	0.81	0.581
106	124.37	124.66	-0.29	-0.233
107	122.53	122.25	0.28	0.228
108	123.29	123.2	0.09	0.072
109	123.5	123.58	-0.08	-0.064
110	125.51	125.38	0.13	0.103
111	126.36	126.14	0.22	0.174
112	123.45	123.7	-0.25	-0.202
113	124.32	124.31	0.01	0.008

DNN				
no	Actual record	Predicted record	Error	Error Rate(%)
114	126.19	126.56	-0.37	-0.293
115	127.35	127.1	0.25	0.196
116	127.59	127.56	0.03	0.023
117	127.87	127.32	0.55	0.43
118	129.55	129.53	0.02	0.015
119	130.7	129.78	0.92	0.703
120	125.86	125.77	0.09	0.071
121	126.37	126.25	0.12	0.094
122	127.61	127.5	0.11	0.086
123	127.66	127.25	0.41	0.321
124	128.43	127.44	0.99	0.77
125	129.14	128.94	0.2	0.154
126	130.13	129.71	0.42	0.322
127	130.91	131.74	-0.83	-0.634
128	131.15	131.39	-0.24	-0.182
129	143.26	140.42	2.84	1.982
130	123.77	123.78	-0.01	-0.008
131	123.91	124.06	-0.15	-0.121
132	124.55	123.88	0.67	0.537
133	125.96	125.81	0.15	0.119
134	126.04	125.61	0.43	0.341
135	127.33	126.87	0.46	0.361
136	128.82	128.99	-0.17	-0.131
137	128.54	127.55	0.99	0.77
138	129.59	129.11	0.48	0.37
139	130.37	130.95	-0.58	-0.444
140	130.79	130.45	0.34	0.259
141	131.99	131.74	0.25	0.189
142	132.05	131.88	0.17	0.128
143	133.88	133.77	0.11	0.082
144	122.59	122.81	-0.22	-0.179
145	124.53	124.61	-0.08	-0.064
146	124.84	124.93	-0.09	-0.072
147	129.95	128.78	1.17	0.9
148	133.24	132.8	0.44	0.33
149	122.01	121.89	0.12	0.098
150	123.56	123.26	0.3	0.242
151	124.27	124.08	0.19	0.152

DNN				
no	Actual record	Predicted record	Error	Error Rate(%)
152	124.92	124.74	0.18	0.144
153	126.05	125.64	0.41	0.325
154	126.07	126.82	-0.75	-0.594
155	127.19	127.42	-0.23	-0.18
156	127.68	127.83	-0.15	-0.117
157	121.94	122.04	-0.1	-0.082
158	122.42	123.09	-0.67	-0.547
159	123.22	122.92	0.3	0.243
160	123.27	123.4	-0.13	-0.105
161	125.65	125.47	0.18	0.143
162	127.47	128.08	-0.61	-0.478
163	127.74	127.92	-0.18	-0.14
164	128.41	128.98	-0.57	-0.443
Mean	125.27	125.06	0.209	0.164

(Abstract)

Comparison of machine learning models for predicting swimming competition results

Yang Minjung

Department of Physical Education

Graduate School of Dankook University

Advisor: Prof. Choi Hyongjun

Using artificial intelligence, which is a major information technology in the era of the 4th industrial revolution, various analysis techniques are being used for predictive research according to the characteristics and types of sports. In addition, as the amount of available data related to sports increases, interest in developing intelligent models to predict game results is increasing. In sports games, prediction provides variables that affect the flow of the game, game results, allows strategic establishment of game management and training methods, and presents performance evaluation data to provide information needed to improve performance in the future. By predicting the results of a sports game using various machine learning techniques, the prediction system and explanatory ability according to the results of the game can be more specifically identified.

The purpose of this study is to design a swimming competition result prediction model using machine learning based on the 2017-2021 national swimming competition women's freestyle 200m competition analysis data, and compare the performance of the designed prediction model to find a

model suitable for swimming competition. In addition, the performance variables that affect the competition results of each swimming competition result prediction model are searched. To achieve the purpose of this study, machine learning prediction techniques such as Linear Regression, Lasso Regression, Ridge Regression, Elastic Net Regression, Random Forest, Gradient Boosting Machine (GBM) and Deep Neural Network (DNN) models were designed and prediction performance was evaluated for each model. In addition, the relative importance of performance variables for each swimming competition result prediction model was confirmed.

First, as a result of comparing the prediction performance of machine learning prediction models, Lasso Regression prediction model was found to be the best, followed by the Elastic Net Regression prediction model, Linear Regression prediction model, Ridge Regression prediction model, DNN prediction model, Random Forest prediction model and GBM predictive model showed predictive accuracy in order. Second, the relative importance of the variables entered in the Linear Regression prediction model, Lasso Regression prediction model, Ridge Regression prediction model, and Elastic Net Regression prediction model was high in the 4lap clean swimming record, 3lap clean swimming record, and 2lap clean swimming record. In the prediction model using linear regression analysis, it was confirmed that the variables recorded in the stroke section accounted for a high proportion of the model prediction. The relative importance of the variables entered into the Random Forest prediction model and GBM prediction model showed high in the 3lap cleanswim speed, 3lap cleanswim record, and 4lap cleanswim record variables. In the predictive model using nonlinear regression analysis, it was found that the speed and record of the stroke section contributed highly to predicting the model. Due to the nature of the algorithm, the DNN prediction model showed that the variables used for prediction contributed evenly throughout.

Based on the swimming women's freestyle 200m competition analysis record (data), it was possible to use a machine learning prediction technique to predict the results of the swimming competition, and the Lasso

Regression prediction model was found to be the most suitable. In addition, it was confirmed that the record in the stroke section in the women's 200m freestyle was a major predictor of the final record.

This study is significant in that it applied machine learning prediction techniques to predict swimming results. In the future, it is expected to increase the possibility of using sports data of machine learning prediction techniques and enable various analyses.

Keywords: Sports performance analysis, Swimming, Prediction of competition results, Machine Learning, Record sports, Sports data analysis and convergence