문제정의서(연구계획서)

과제명	축구 승부 예측 모델 개발

조	나 혼자 한다 조
지도교수	이영석 교수님 (서명)
조원	201302366 김규태

1. 연구의 필요성

-연구주제의 배경

이 연구의 주제는 축구에서 두 팀이 경기하여 경기를 통해 발생하는 데이터를 대량으로 수집, 분석하여 두 팀의 승패를 예측할 수 있는 모델을 개발하려는 것이다.

축구는 스포츠 중에서도 오래된 역사를 지닌 스포츠로 기원으로 따지면 1800년대 후반부터 시작되었다. 물론 지금의 현대적인 축구는 1990년대부터로 보고 있으며, 축구 경기를 전문적으로 분석해 데이터를 수집하는 경기 분석 역시 90년대 중반에 처음 시작되었다. 물론 당시에는 양팀의 볼 점유율, 패스 성공률, 코너킥, 파울 등 포괄적인 데이터나 경기 규칙과 관련된 기본적인 데이터만 수집했다. 그 뒤로 축구에 대한 연구가 많아지면서 축구를 분석하기 위한 지표 역시 더 세밀해지고 세분화 되었다. 축구에 대한 데이터 역시 점차 쌓이면서 축구를 수학,통계적인 방법을 적용해서 연구하려는 움직임이 많아졌다.

'The Numbers Game'의 저자 크리스 앤더슨의 말에 따르면 축구는 다른 스포츠에 비해 베팅회사에서 계산한 승리예상팀과 상대 배당률 차이가 50%가 넘는 경우 승리로 이어지는 확률이 약 65%에 달하지만 농구는 80% 이상, 야구, 미식축구 등 다른 스포츠 모두 축구보다 10~15% 이상 승리확률이 높은 것으로 나타났다.

축구계에서 유명한 말인 '공은 둥글다'라는 표현은 실제 경기 결과가 나오기 전까지 축구에서 승부를 함부로 단정짓기 어렵다는 말로 결국 축구는 눈에 보이는 데이터를 통한 수치보다 눈에 보이지 않는 요소, 즉 운에 영향을 받기 쉬운 시장이라는 것이다.

크리스 앤더슨이 조사한 바에 따르면 독일 뮌스터 대학 이론 화학자 호이어의 연구진이 축구 득점 패턴에 보다 중요한 요소가 선수들의 능력, 컨디션 혹은 경기요소인지 아니면 예측하기 어려운 운의 요소인지 독일 분데스리가 20년치 경기를 통계적인 방법을 통해 연구했더니, 두 팀이 각자 3개의 동전을 던져 앞면이 세번 연

달아 나올 확률이 경기에서 득점이 발생할 확률과 유사하다고 한다. 그리고 양팀이 슈팅을 할 횟수는 경기시작전 그 팀의 상태, 팀 스쿼드의 질에 달려 있으며, 슈팅이 골로 연결될 확률이 약 1/8 이라고 한다. 축구 경기에 어떤 팀이 승리할지 가장 우선적 영향 끼치는 것은 운이며, 기술, 컨디션 등은 그 다음이라고 주장했다.
뮌헨 공과대학에서 축구를 과학적으로 접근, 관찰 연구하는 마틴 레임스 박사의 경우 골 장면에서 운이 얼마나 큰 역할을 했는지 평가하기 위해 수년간 약 2500여건의 골 장면을 보고 조사한 결과, 6골 중 1골은 슈팅을 한 선수에게 운이 따라 득점했다는 결론이 나왔다고 한다. 계획된 슈팅이 아닌 제대로 컨트롤 되지않은 상태의 득점을 말하는 것으로 리그별로 차이가 있지만 약 44.4%의 확률로 득점에 행운이 작용했으며, 특히 0-0상황일 때 빈번히 발생했다는 것이다.

-연구개발의 필요성

크리스 앤더슨은 앞의 연구들을 종합해 봤을 때 축구에서 모든 득점의 절반은 눈에도 확연히 식별되는 행운의 요소를 포함해 승리하리라 예상되는 팀이 이길 확률은 약 50% 정도이며 슈팅을 더 많이 한 팀 역시 승리할 확률이 50%가 되지 않는다고 주장했다. 결론적으로 말해 축구는 동전 던지기와 마찬가지며 논리와 우연성이 반반씩 존재한다는 것이다. 그렇다고 축구를 수학, 통계적인 방법으로 연구하는 것이 의미가 없다는 말이 아니다.
경기장에서 일어나는 일들 절반은 우리의 손에 달린 것이 아니지만, 그 나머지 절반은 충분히 통제할 수 있으며 모든 축구팀 역시 통제할 수 있는 영역의 확률을 높이기 위해 최선을 다하는 것이다.

본 연구 역시 앞에서의 연구를 통해 축구에서의 완벽한 승률 예측은 한계가 있다는 것을 충분히 인지하지만, 마찬가지로 그 나머지 영역의 경우 우리가 통제 가능하지만 아직 알아내지 못한 미지의 잠재성 있는 부분이 충분히 남아있다고 생각한다. 초창기에 축구 분석을 위해 사용된 데이터 지표는 볼 점유율, 패스 성공률 , 슈팅 횟수 등 큰 틀에서 바라보는 것이 대부분이었지만 오늘날 축구 경기 분석을 전문으로 하는 업체가 많아지면서 축구를 바라보는 시선 역시 더욱 디테일해지고 세분화 되었다. 지금은 패스 하나하나의 길이, 슈팅을 시도한 위치 등 기록할 수 있는 거의 모든 것들을 기록하고, 선수 들의 몸에 별도의 장치를 부착하여 피지컬 데이터까지 수집하여 구단에서 선수 관리에 만전을 기하고 있는 실정이다.

이를 통해 앞으로도 지금 밝혀진 것보다 더 복잡하고 발견해내지 못한 변수들이 충분히 많다고 여기기에 아직 우리에게 남아 있는 통제 가능한 영역을 넓히기 위해 승부 예측 모델 개발을 진행하려고 한다.

-국내 외 연구 현황

이와 관련해 축구 경기의 승률 예측을 위한 연구 역시 축구를 전문적으로 분석하기 시작한 90년대 중반부터 시작됐으며, 대표적 연구로 (Dixon and Coles)이 1997년에 진행한 연구가 있다. 축구 경기에서 나오는 골의 수는 홈팀과 원정팀이 각각 다른 포아송 분포를 따른다는 가정하에 경기의
승률을 추정했으며, 2000년대 들어 국내 역시 김주학 외(2007)는 독일 월드컵을 대상으로 신경망을 이용한 축구경기 승패예측모형을 개발하였으며, 이해용(2012)이 포아숭 분포를 이용한 축구경기의 승률 예측에 관한 연구를 진행, 대체적으로 해외에서 먼저 연구되었던 포아송 분포를 따라 승률 예측을 진행했으나 가장 최근에는 김형원(2020)이 익스트림 그래디언트 부스팅 알고리즘에 기반한 축구경기 예측을 통해 인공신경망을 통한 승부 예측 모델을 개발하는 등 머신러닝 모델을 활용하여 승부예측 모델을 개발하는 것으로 모델 개발방법 역시 다양해지고 있다.

그러한 연구들을 보면서 연구 결과들을 활용해 충분히 의미 있는 데이터들을 발견해 낼 수 있으며, 그러한 데이터들을 활용하거나 새로 발굴하여 충분히 성능 좋은 승부 예측 모델이 개발 가능하다는 확신이 생겼다.

2. 연구의 목표 및 내용

앞서 언급했던 축구 승부 예측 모델 연구들은 포아송 모델을 필두로 통계적인 분석 방법부터 시작하여
최근에는 머신러닝 모델을 이용한 예측 연구가 주를 이루고 있다.
축구는 토너먼트를 제외하고 일반적으로 승,무,패 세가지의 결과로 이루어져 있기에 머신러닝 역시 분류
모델을 사용하여 나이브 베이즈 모델, SVM, 랜덤 포레스트, 다항 로지스틱 회귀 모델 등 다양한 분류
알고리즘 등을 활용하여 연구가 진행되어 왔다.

연구를 위해 사용된 데이터 역시 구체적으로는 골과 코너킥, 유효슈팅을 이용하거나 경기 데이터 외에 외적인 변수인 지역간 온도차, 경기 시 강수량 등의 데이터도 이용했으며, 대체로 지난 시즌의 경기 데이터를 이용했다고 하였다.

본 연구는 우선 가장 최근에 진행한 김형원(2020)이 익스트림 그래디언트 부스팅 알고리즘에 기반한 축구경기 예측에서 인공신경망, XGBoost 를 이용해 약 53%에서 59%의 적중률을 가진 모델을 만든 점을 참고하여 최소 이보다 더 향상된 적중률을 가진 모델을 개발하는 것을 기본 목표로 삼았다.

다만 적중률 뿐만이 아니라 승부 예측에 사용되는 데이터 역시 기본적으로는 이전 시즌의 데이터를 이용하겠지만, 단지 수집된 데이터를 그대로 사용하는 것이 아닌 그동안 밝혀진 축구의 승패에 영향을 끼치는 요인 등을 활용하여 데이터 역시 이에 맞게 새롭게 가공하여 사용하려고 한다.

기존에 이미 다양한 머신러닝 모델을 사용한 승부 예측 연구가 진행된 만큼 어떤 머신러닝 모델을 사용하는지 보다 기존에 사용하지 않았던 데이터를 사용하여 축구의 승패에 영향을 끼치는 요인을 밝혀내고, 축구 분석에 도움이 되는 방향으로 연구를 진행하려고 한다.

3. 연구의 추진전략 및 방법

앞서 연구 목표 및 내용에서 말한 바와 같이 기존에 이미 다양한 머신러닝 모델을 사용하여 연구를 진행한만큼

특정한 머신러닝 모델을 이용하기 보다 우선 모델에 입력할 데이터 선정 및 수집에 초점을 맞춰 그 후에 다양

한 머신러닝 모델에 데이터를 적용해 나오는 적중률 간의 차이를 보려고 한다.

머신러닝 모델도 기본적으로 사용되는 데이터의 질이 좋아야 더 높은 성능을 보일 수 있는 만큼 먼저 승패에 영향을 끼치는 보다 중요한 데이터들을 수집하면 기존의 머신러닝 모델이나 데이터 패턴 발견을 통해 충분히 더 성능 좋은 모델을 개발할 수 있다고 생각한다.

따라서 예측 모델에 사용할 데이터 선정을 위해 승패에 대한 관점으로 축구에 대해 다시 정의하여 접근하였다. 축구란 '상대보다 골을 많이 넣거나 적게 실점하면 이기는 스포츠' 라는

정의를 시작으로, 승패를 예측하는데 가장 기본적인 데이터는 상대보다 골을 많이 넣을 수 있는지, 적게 실점할 수 있을지 아는 것이다. 따라서 득점율/실점율 을 기본 데이터로 삼아 상대보다 먼저 득점할 수 있는지, 득점한다면 득점한 점수를 지킬 수 있는지, 상대에게 실점해도 다시 득점할 수 있는지 여부 등을 예측할 수 있는 승리하는 팀, 패배하는 팀의 특징을 찾아 확률적인 관점에서 분석해 승패에 영향을 미치는 다양한 데이터를 수집하려고 한다.

이를 통해 최종적인 데이터 선정 후 데이터 패턴 분석을 통해 최종 선정한 데이터 들 간의 영향도를 분석하여 알고리즘을 개발한다.

4. 연구 팀의 구성 및 과제 추진 일정

연구진은 연구를 진행하는 본인과 담당 지도교수이신 이영석 교수님, 자문 역할을 해주실 하석재님으로 구성되어 진행한다.

연구는 지난 19년 2학기부터 본 주제를 선정해 진행하였으며, 2학기 동안 승부 예측 모델 개발을 위해 앞서 진행된 연구들을 조사하여 승부 예측 모델 연구가 어떤 방식과 데이터를 이용하여 진행되었는지 보고

기본적으로 접할 수 있는 축구 경기 데이터를 수집하여

축구 승부예측에 관한 책, 축구를 수학,확률적인 관점으로 분석한 책들을 참고하여 의미있는 데이터를 선정하는데 중점을 두었다.

이번 학기부터 앞서 선정한 승부예측 모델의 틀을 잡아줄 중심적인 데이터를 바탕으로 패턴을 분석, 기본적인 모델을 프로토타입으로 만들어 적중률을 시험한 후, 앞서 반영하지 않았던 다른 데이터들을 추가로 선정하여 기존 모델에 반영해 평가한다.

그래서 해당 모델에 가장 적합한 데이터들을 최종적으로 선정한 후, 여러 후보군의 머신러닝 모델을 선정하여 적중률을 평가해 최종 모델을 선정한다.

선정한 최종 승부 예측 모델을 실제로 구현할 User Interface를 개발한다.

※ 반드시 5쪽 ~ 6쪽 분량으로 작성

– 참고문헌(Reference)

Dixon, M. J, Coles, S. C. (1997) Modelling association football scores and inefficiencies in the football betting market, Applied Statistics, 46, 265-280.

이해용(2012).포아송분포를 이용한 축구 경기의 승률 예측에 관한 연구

Chris Anderson(2013).The NumBers GAME

김형원(2020).익스트림 그래디언트 부스팅 알고리즘에 기반한 축구 경기 예측