**프로배구 플레이오프 진출요인 분석**

**스포츠과학과**

**2013310563**

**윤진성**

**목차**

**1. 데이터 소개**

**2. 데이터 수집**

**3. 데이터 전처리**

**4. 데이터 분석**

**5. 결과**

1. Project Scoping(Define Problem)
2. Data Collection
3. EDA
4. Data Preprocessing
5. **Feature Engineering**
6. Modeling
7. Evaluation
8. Project Delivery / Insights

**1. 졸업작품에 대한 설명**

1) 스포츠 경기에서 발생하는 데이터를 통하여 팀들의 왕중왕전을 가르는 플레이오프 진출요인을 분석하는 과제입니다. 현재 프로배구에는 남·녀(7개/6개)팀으로 이루어져 있습니다. 경기는 팀당 4경기씩 6라운드를 거쳐 24경기를 치뤄 경기결과에 따라서 순위가 결정됩니다. 리그에서 결정된 순위로 상위 3개팀(3위팀과 4위팀의 승점이 3점차이 이내일 경우 4팀)이 플레이오프에 진출하여 한번 더 순위를 결정하는 시스템으로 진행되고 있습니다.

7개의 팀들의 수준이 서로 비슷하여 매 리그미다 어떤 팀이 플레이오프에 진출할지 모든 팀들이 플레이오프 진출을 위해 팀별로 전력분석 및 훈련을 진행하고 있습니다. 플레이오프에 진출에 영향을 끼치는 요인을 분석해 봄으로써 플레이오프 진출의 확률을 높이고자 하는 것이 본 프로젝트의 목표입니다.

**2. 데이터 수집**

1) 데이터 분석에 필요한 데이터는 대한배구협회에서 제공하는 시즌별 경기결과의 데이터를 수집하였습니다. 대한배구협회에서 제공하는 데이터는 총94개의 경기요인과 2007년부터 2017년까지의 10년치 데이터를 수집하였습니다. 총 남자는 67개 여자는 52개의 데이터가 수집되었습니다.

한번에 2007년도부터 2017년도까지의 데이터를 크롤링하여 하나의 데이터로 합치려고 하였으나, 중간 중간에 데이터 인터넷 연결 불안으로 인하여 1년치의 데이터를 pickle로 저장한 이후에 데이터 전처리 부분에서 10년치 데이터로 합쳤습니다.

**3. 데이터 전처리**

1) 데이터 병합 및 추가요소 추가

먼저 크롤링을 통하여 수집한 10년치 데이터를 하나로 합칩니다. 플레이오프 진출 여부를 알 수 있도록 시즌별 플레이오프 진출 여부 계산하고, 플레이오프 진출에 영향을 줄 수 있는 공격효율, 최다연승, 최다연패 요소를 계산하여 데이터를 하나로 합칠 때 추가하였습니다.

1차적으로 정리된 경기 데이터는 남자 67개 여자 52개의 경기데이터와 각각 75개의 경기요인이 정리되었습니다. 경기요인은 아래와 같습니다.

|  |  |  |
| --- | --- | --- |
| 대분류 | 개수 | 항목 |
| 종합 | 6개 | 순위/경기수/승/패/세트득실률/점수득실률 |
| 득점 | 4개 | 공격/블로킹/서브/득점 |
| 공격 | 6개 | 시도/성공/공격차단/범실/성공률/효율 |
| 오픈공격 | 5개 | 시도/성공/공격차단/범실/성공률 |
| 시간차공격 | 5개 | “ |
| 이동공격 | 5개 | “ |
| 후위공격 | 5개 | “ |
| 속공 | 5개 | “ |
| 퀵오픈 | 5개 | “ |
| 서브 | 4개 | 시도/성공/범실/세트당평균 |
| 블로킹 | 7개 | 시도/성공/유효블락/실패/범실/어시스트/세트당평균 |
| 디그 | 5개 | 시도/성공/실패/범실/세트당평균 |
| 세트 | 4개 | 시도/성공/범실/세트당평균 |
| 리시브 | 4개 | 시도/정확/범실/세트당평균 |
| 기타 | 5개 | 벌칙/범실/최다연승/최다연패/플레이오프진출 |

2) 데이터 축소

데이터 차원을 줄이는 방법은 특징선택(feature selection)과 특징추출(feature extraction) 두 가지 방법이 있습니다.

특징 선택의 목적은 모든 특징의 부분 집합을 선택해서 간결한 특징 집합을 만드는 것이다. 플레이오프 결과 예측에 영향이 없다고 생각하는 요인과 적대적으로 영향을 끼치는 요인, 중복적으로 수집된 요인을 제거할 수 있다. 이와 같이 원본 데이터에서 불필요한 불필요한 특징들을 제거하는것을 특징선택이라고 한다.

특징 추출은 원본 특징들의 조합으로 새로운 특징을 생성하는 것입니다. 가장 유명한 방법이 주성분분석이 있으며, 데이터들을 가장 잘 설명할 수 있는 주축을 찾아 해당 데이터들을 그 직선에 투영시키는 것입니다.

이번 데이터 분석에서는 특징선택을 이용하여 데이터 축소를 진행하도록 하겠습니다.

2-1) 데이터 제거

플레이오프 진출에 있어서 모든 팀이 비슷한 경기요인과 절대적으로 영향을 주는 요인을 제거합니다.

모든 팀이 동일한 경기요인은 : ‘경기수’, ‘벌칙’

절대적인 영향을 끼치는 ‘순위’, ’승’, ’패’, ’세트득실률’, ’점수득실률’

또한 득점 카테고리가 공격, 블로킹, 서브로 얻은 득점이므로 중복 데이터이며 득점은 3개의 파트의 총 합이므로 득점파트를 제거합니다.

위의 데이터를 제거한 이후에도 64개의 경기요인이 남습니다. 여전히 경기요인의 개수가 데이터의 개수보다 많으므로 경기요인을 더 줄여야 합니다.

64개의 데이터에서 배구 경기의 중요한 요소인 서브/리시브/공격/블로킹/연승결과 데이터로 축소하여 16개 요인으로 축소하였다.

**4. 데이터 예측**

프로배구 플레이오프 진출 여부에 관하여 분석할 기법은 지도학습 기반의 분류 알고리즘을 통하여 플레이오프 진출 여부에 대해서 예측하고 그에 따른 속성의 가중치를 확인하는 방법으로 진행할것이다.

분류 알고리즘의 종류는 아래와 같이 있습니다.

* Boosted Trees
* Decision Tree
* K-Nearest Neighbor(KNN)
* Naïve Bayesian Theorem
* Neural Network
* Linear Regression
* Logistic Regression
* Random Forest
* Support Vector Machine(SVM)

위 알고리즘들 중 Decision Tree, Logistic Regression, Random Forest, Support Vector Machine을 사용하였다.

Decision tree 알고리즘은 의사결정나무의 분류모형을 따라 결과를 도출하는 알고리즘이다. 의사결정나무 모델은 2개의 셋으로 나눔으로써 데이터들을 분류하고, 잎세 노드에서 분류에 대한 결과를 보여줌으로써 인풋으로 들어온 데이터에 대한 결과를 도출해 낸다. 16개의 경기요인을 Decision tree 알고리즘을 통하여 플레이오프 진출 예측 모델을 만들고 그 모델에 따라서 진출여부를 예측하고 모델에 사용된 경기요인을 통하여 플레이오프 진출에 영향을 끼치는 요인을 확인할 수 있습니다.

Logistic regression은 다양한 입력변수들을 sigmoid함수값을 이용하여 이분법의 결과값을 도출하는 알고리즘입니다. Sigmoid 함수를 통하여 나온 odd 값을 일정 값 이상일 경우 1을 미만일 경우 0을 도출하도록 분류하여 2항분류를 도출할 수 있습니다. 스포츠경기에서 여러가지 변수값을 가지고 sigmoid함수에 대입하여 계산한 후 일정 odd값을 기준으로 하여 플레이오프 진출여부를 확인할 수 있으며, sigmoid함수에 계산될 때 곱해지는 theta값을 통하여 입력변수들의 가중치를 확인할 수 있습니다.

Support Vector Machine은 지도학습의 한 방법으로써 데이터를 분석하고 패턴을 파악하여 분류분석을 할 수 있는 개념이다. Decision tree처럼 예제는 여러 개의 vector로 이루어져 있어서 이것은 Rn차원이다. SVM은 데이터 들을 가장 큰 마진을 갖도록 분류하는 hyperplane을 찾는 것이다. Hyperplane를 기준으로 데이터는 2개의 class로 나뉘어진다. 이 Hyperplane을 일정 거리에 위치한 training-data들의 vector을 support vector라고 하며 support vector와 hyperplane간의 거리가 커야 새로 들어온 데이터에 대한 에러확률이 줄어든다. SVM은 선형분리가 되지 않는 데이터에 대해서는 더 높은 차원으로 매핑하여 데이터를 쉽게 분리할 수 있는 hyperplane을 찾아 데이터를 분리하기에 데이터 분류 알고리즘 중에서는 가장 많이 쓰이는 알고리즘으로 알려져 있다. 플레이오프 진출과 미진출로 2개의 계층값으로 결과가 나와야 하는 이번 연구와 hyperplane값을 통하여 경기요인의 가중치를 확인해 볼 수 있음으로써 본 연구를 위한 적절한 알고리즘으로 판단된다.

Random\_forest는 무작위로 변수들을 추출하여, 의사결정나무를 만들고, 만들어진 의사결정나무들로 테스트 데이터 예측하여 가장 많이 나온 결과를 투표를 통하여 도출하는 알고리즘이다. 기존의 의사결정나무는 한번 분기할 때 엔트로피가 낮아지도록 하는 변수를 선택하여 분기 시킨다. 하지만 이때 트레이닝 데이터에 대해 과적합이 발생하여 새로 들어온 테스트 데이터에 대해서는 예측률이 감소할 수 있게 된다. 따라서 전체의 변수들 중에서 임의의 변수를 랜덤하게 추출하여 의사결정나무를 만든다. 이렇게 만들어진 의사결정나무에 테스트 데이터를 넣고 가장 많이 나온 결과를 반환하는 알고리즘이다. 의사결정나무 모델은 모든 변수를 예측에 사용하면 과적합이 발생하여 테스트 데이터에 대한 예측률이 감소하고, 모든 변수를 사용하지 않으면 해당 변수에 대한 가중치를 알기 힘들다. 따라서 최대한 많은 변수들을 사용하여 의사결정모델을 만들고 그 모델들을 사용하여 테스트 데이터의 결과값을 앙상블기법을 통해 도출함으로써 예측률도 상승시켜 예측률과 변수 가중치를 얻는데 있어서 적절한 방법이라고 할 수 있다.