NBA 슈팅 및 패스 데이터를 활용한 슈팅 성공 여부 예측 및 팀워크의 네트워크적 가시화와 팀워크-팀 승리의 상관관계 분석

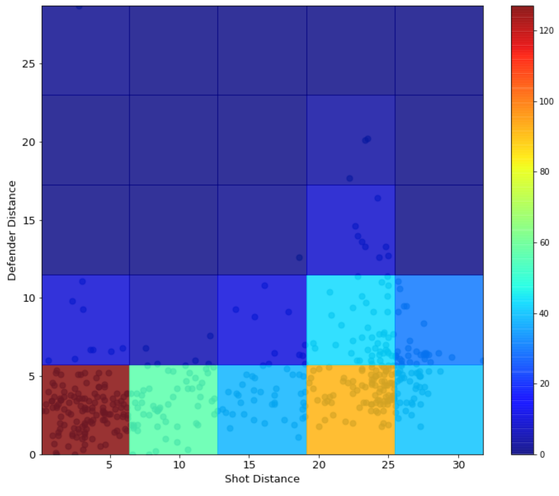
2021320117 배민성 / 고려대학교 정보대학 컴퓨터학과 1학년
데이터과학과 인공지능 최종 프로포절

연구 목적

- 야구에서의 세이버메트릭스나 NBA 팀들의 데이터 분석가 채용 등 프로 스포츠에서 데이터 과학 및 분석의 도입을 통해 선수들의 기량을 향상시키거나 팀의 성적을 개선시키려는 경우가 늘고 있다.
- 이번 연구를 통해 슈팅 성공 여부를 예측하고 슈팅에 영향을 미칠 것으로 예상되는 다양한 변수들의 feature contribution을 통해 선수의 슈팅을 분석하고 슈팅 성공률을 향상시킬 수 있도록 한다.
- 또한 팀 내 선수들 간의 네트워크를 가시화하고 그래프 이론을 활용해 팀워크를 나타낼 수 있는 수학적 지표를 제작하여 팀워크와 승리의 관계를 파악하고 팀의 성적 향상을 이뤄낼 수 있도록 한다.

데이터 수집 및 전처리

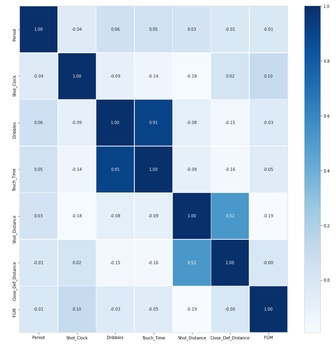
- 슈팅 분석에는 세계적인 데이터 분석 관련 사이트인 kaggle에 업로드되어 있는 NBA 선수들의 2014-2015 시즌 슈팅 및 수비 스탯 CSV 데이터를 사용한다. (https://www.kaggle.com/dansbecker/nba-shot-logs)
- 슈팅 분석에 유의미할 것으로 예측되는 6개의 주요 변수와 슈팅이 성공했는지 실패했는지를 나타내는 열 1개로 데이터를 새롭게 구성하였고, python의 pandas 라이브러리를 이용해 결측치(nan)를 0으로 수정하였다. 최종 데이터의 크기는 (128070, 7)이다.
- 팀워크를 수치화시키기 위한 그래프는 2015년 5월 21일 치뤄졌던 Golden State Warriors와 Houston Rockets의 서부 컨퍼런스 파이널 2차전 경기의 선수 간 패스 데이터를 기반으로 python의 networkx 라이브러리를 통해 제작되었다.
- 패스 데이터는 경기 중 일어난 모든 패스(엔드라인이나 사이드라인에서의 패스 제외, 단 바로 어시스트로 연결된 경우는 포함) 에 대해 패스를 준 선수의 등번호와 패스를 받은 선수의 등번호가 한 줄 씩 적혀 있는 txt 파일이다.



- **Figure 1.** Shooting Heatmap of Stephen Curry

#1. 슈팅 성공 여부 예측

- Python의 scikit-learn 라이브러리를 활용해 모든 학습을 진행하였다.
- 전체 데이터를 train, test 데이터로 (0.85 : 0.15) 나눈 후 train 데이터를 다시 validation을 위한 데이터와 실제 train 데이터로 (0.12 : 0.88) 나누었다. (cross validation)
- 변수 간 다중공선성을 고려하기 위해 feature matrix를 제작하였고, 상관관계가 매우 높은 공 소유 시간 변수와 드리블 횟수 변수 중 하나만 학습에 반영하였다.

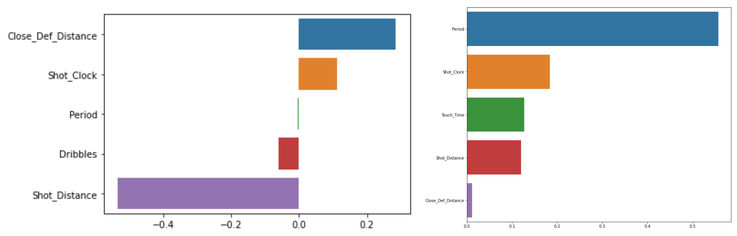


- **Figure 2.** Feature Matrix of Shooting Prediction Data

- Logistic Regression과 Random Forest 두 개의 방법을 통해 학습을 진행하였다.
- 슈팅 성공 여부에 미치는 영향을 정확히 비교하기 위해 학습 시 모든 변수들은 StandardScaler를 통해 평균을 0, 표준편차를 1로 조정하였다.
- Logistic Regression의 경우 학습 시 validation set을 통해 먼저 모델의 정확도를 측정하고, 정확도가 가장 높은 feature 조합을 정하고, validation set과 train set을 다시 합쳐 최종적으로 학습을 진행시켜 test set으로 평가를 진행했다.
- Random Forest의 경우 같은 방식으로 학습을 진행하였고 Grid Search 방식을 통하여 최적의 parameter setting을 찾았다.

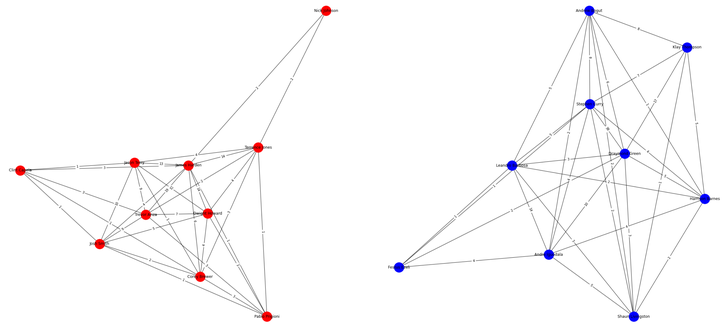
Model	Standard Model	Logistic Regression	Random Forest
Accuracy_val (%)	54.9	61.1	62.1
Accuracy_test (%)	54.8	60.9	62.1

- **Table 1.** Accuracy of models



- **Figure 3.** Feature Importance of Logistic Regression(Left) / Random Forest(Right)

#2. 네트워크 가시화 및 상관관계 분석



- **Figure 4.** Pass Network of Houston Rockets (Left) / Golden State Warriors (Right)

- 텍스트 데이터를 활용하여 edge에 두 선수 간의 패스 횟수를 가중치로 부여하는 가중치가 있는 무방향 그래프를 인접행렬로 제작하였고, networkx 라이브러리를 통해 시각화했다.
- 각 선수들을 나타내는 node에는 대표적 선수 개인 평가 스탯인 해당 경기의 선수 별 BPM (리그 평균선수대비 보정 코트 마진)을 가중치로 부여하였다.
- 임의의 weighted graph에 대하여 해당 그래프의 연결성을 나타내는 지표인 CI (connectivity index) 를 다음과 같이 정의한다.

- 임의의 weighted graph G 위의 어떤 path P에 대해 P의 strength는 **P를 구성하는 모든 edge들의 가중치 중 최소값**으로 정의한다.
- 임의의 weighted graph G 위의 어떤 node u, v에 대해 두 node의 strength of connectedness CONN(u, v)는 **u와 v 사이의 모든 path들의 strength 중 최댓값**으로 정의한다.
- 이때 CI는 다음 수식과 같이 정의한다. (w(node)는 node의 가중치)

$$CI(G) = \sum_{u,v \in V(G)} w(u)w(v)CONN_G(u,v)$$

- 다음 수식을 통해 두 팀의 CI를 계산한 결과, HOU의 CI는 약 -4872, GSW의 CI는 -1448로 승리 팀인 GSW의 CI가 매우 월등히 큰 것을 볼 수 있다.

결론

- 슈팅 성공 여부 예측의 경우 Random Forest 기반 모델은 약 62%의 정확도로 슈팅의 성공을 예측해내며, 이는 기준 모델보다 높은 수치이다. 또한 Logistic Regression에서의 feature importance의 타당한 해석이 가능하고, 이에 따라 각 변수의 기여도를 알아볼 수 있었다. 그러나 Random Forest에서는 타당한 해석이 어려우며, 정확도의 추가적 향상이 필요해보인다.
- 팀워크와 팀 승리 간 상관관계 분석의 경우 해당 경기에 대해 수학적으로 제시된 지표 CI가 승리 팀에게서 높게 나온 것을 긍정적으로 볼 수 있으나, 그 표본이 1경기에 불과하므로 추가적인 연구가 필요해 보인다. 또한 BPM은 음수 값이 가능한 스탯이므로 선수 개인 평가 스탯 중 양수 값만 가능한 스탯을 node의 가중치로 선택할 경우 CI 값이 양수만 나올 것이므로 그 값을 더 직관적으로 해석 가능할 것이다.

References

[1] N. Jicy and Sunil Mathew. **Strong connectivity index of weighted graphs**. 2019. Indian Institute of Technology Calicut, Department of Mathematics.