1. 안녕하세요 야구선수 연봉예측모델에 대해 발표를 맡게 된 이준규입니다.
2. 저희조는 다음과 같이 발표를 진행할 예정입니다. 주제선정 배경, 데이터 수집 및 전처리, 모델 설명 및 결과, 결론
3. 저희가 이 주제를 하게 된 이유는 다음과 같습니다. 혹시 이번에 손흥민 골 영상 보신 분 있으십니까? 와 대박이었죠. 손흥민이 연봉을 108억 받는다고 합니다. 류현진은 약 280억을 받고요. 축구나 할걸 그랬어요?
4. 스포츠 선수의 연봉이 높다는 것은 즉 그 선수의 실력, 위치 등을 알려주는 지표이며, 또한 팬 입장에서도 한 선수의 연봉은 자부심이 되기도 합니다.
5. 다음 그래프는 한국 내에서 스포츠별 관중 순위입니다. 저희는 스포츠 중에 한국에서 가장 인기가 많은 프로야구를 대상으로 연봉 예측을 실시했습니다. 가장 팬도 많고 재밌는 스포츠중 하나이죠.
6. 데이터 수집은 스텟티즈라는 전문 야구 통계 사이트에서 크롤링을 통해 가져왔습니다. 스텟티즈에서는 연봉 데이터와 선수의 통계를 함께 제공하죠. 투수와 타자 중에서는 데이터 숫자가 많은 타자 데이터로 선정하였습니다. 즉 타자 데이터는 14~17년, 연봉 데이터는 15~18년을 수집하였습니다. 왜 연봉 데이터와 타자 데이터를 다르게 수집했느냐? 일반적으로 15년 연봉은 14년 타자의 통계 데이터를 보고 작성되기 때문에 연봉이 1년 밀리는게 맞다고 판단했습니다.
7. 실제 데이터는 다음과 같고
8. 수집한 데이터는 다음과 같습니다. 이름 WAR, 게임 수 타석 등 총 29개의 변수가 있고, 트레인 데이터는 14년~16년과 연봉을 합쳐 총 627개의 열을, test data는 208개를 수집하였습니다. 이 두 데이터는 이름을 기준으로 join하였고, 모든 선수의 연봉 데이터가 공개된 것이 아니기 때문에 연봉 데이터가 존재하지 않는 선수들은 제외하였습니다.
9. As
10. 데이터를 살펴보니 전처리해야할 것이 있었습니다. 부상 등의 이유로 타자 데이터에서 NA가 있는 타자들이 있었습니다. 예를 들어 이 손승락 친구는 연봉을 7억이나 받았는데 게임은 4경기 밖에 뛰지 않아 정확한 데이터를 제공받기 힘들었습니다. 따라서 게임이 10이상인 선수들만 필터링 했습니다. 또한 타석 타수, 득점, 안타와 같은 수치는 정수임에 반면 타율, 출루, 장타, OPS는 0에서 1사이의 실수 값입니다. 이 둘은 스케일이 다르기 때문에 타수와 득점, 안타와 같은 변수들은 타석으로 나누어 정규화를 진행시켜줬습니다.
11. (칠판으로 가며) 저희가 데이터의 분포도를 박스 플롯으로 표현했는데, 다음과 같이 나왔습니다. 일반적으로 아웃라이어를 선별할 때, (삼분위 수 - 1분위 수) X1.5 를 계산한다고 합니다. 그 값은 4억 2250만원이고, 저희는 그 값 이상이면 스타플레이어, 그 값 이하이면 일반플레이어로 구분하게되었습니다.
12. 이제 저희가 자체적으로 구상한 연봉 예측 모델에 대해 설명드리도록 하겠습니다. 이 모델은 총 3가지 모델을 사용하게 되는데, 아웃라이어, 만약 타자 데이터가 들어오게 된다면, Outlier Classiffier 모델을 통해 그 선수가 스타 플레이어인지, 일반 플레이어인지 결정하게 됩니다. 만약 그 선수가 스타 플레이어 라면 스타플레이어 모델을 통해 스타플레이어의 순위를 예측하고, 일반 플레이어라면 총 8개의 클래스로 나누어진 연봉 구간에서 어디에 속할 지 배정받게 됩니다.
13. 자 조금 더 자세히 설명해보도록 하겠습니다. 만약 타자 데이터가 들어오게 된다면, 아웃라이어 클래시파이어를 통해 이 선수가 스타플레이어이면 1, 일반플레이어이면 0으로 결정하게 됩니다. 이러한 모델은 히든 레이어 3층의 딥러닝 모델로 구성했고, validation 정확도 0.93, test accuracy 0.85의 매우 높은 정확도를 보였습니다.
14. 저희 수업에서는 딥러닝을 배우지 않았으니 조금 더 자세히 설명해보자면,
15. 만약 들어온 데이터가 스타플레이어라고 한다면 바로 스타플레이어 모델을 거치게 됩니다. 스타 플레이어 모델에서는 새로운 평가기준인 Rank\_score을 제시하게 되는데, 다음과 같습니다. 만약 이대호 선수가 들어옵니다. 실제 랭크는 1이고, 이 모델을 거친 예측 랭크가 11이라고 해봅시다. ~~~, 정말 다양한 모델을 시도해보았고, PCA를 실행한 다중 선형회귀 모델이 가장 성능이 좋았습니다.
16. 자 여기서 PCA란 무엇이냐. 저희가 배우긴 배웠지만 혹시 기억이 안나실분들을 위해 그냥 일반적으로 선형회귀를 긋게 되면 이렇게 나옵니다. 얼핏봐서는 선형을 이루고 있는 것처럼 보이지만 변수들 간의 공분산성 때문에 의미없는 회귀분석이 되어버립니다. 따라서 이러한 공분산성을 해결해주기 위해 pca를 실시했고, 누적 분산이 99퍼센트가 되는 점까지 변수 7개를 통해 새로운 회귀식을 만들어냈습니다. 그리고 그 값을 통해 순위를 산출해 냈습니다.
17. 왼쪽 그림이 결과값입니다. 보시면 정상호씨는 실제 랭크가 24위에 예측 랭크가 28위로 4계단 차이나게 됩니다. ~~~해서 Rank\_score 최저 점수 10점대가 나오게 되었습니다. 하지만 우측 그림같이 예측 랭크와 실제 랭크가 매우 많이 차이나는 경우도 있습니다. 이는 통계적으로는 매우 잘 했을 수 있으나, 과소평가되어 연봉이 적게 예측된 것일 수도 있습니다. 실제로 저기 있는 김재환 선수와 같은 경우에는 2018년에 홈런왕, 타점왕을 획득하며 연봉이 확 뛰게 됩니다.
18. 그 다음은 일반플레이어 모델입니다. 일반플레이어 모델은 일단 데이터의 분포에 따라 최적 군집화 시켜주는 비지도 학습인 NBCLUST 기법을 사용했습니다. 총 8개의 군집이 나왔으며, 각 클래스당 2000~3000만원 사이의 구간을 지니고 있습니다.
19. 이것도 정말 다양한 모델을 사용해 XG부스트 분류의 스코어가 가장 좋았습니다. 하지만 가장 좋았던 점수가 0.35, test\_accuracy는 0.24로 현저한 성능적 오류를 발생시켰습니다. 이유에 대해 곰곰히 생각해본 결과 적은 레코드에 비해 분류할 class가 상대적으로 많은 것이 문제였습니다. 추후 데이터셋 보강 혹은 발전된 모델을 사용해 성능을 향상시키는 것이 목적입니다.
20. 결과적으로 이 프로젝트를 진행하면서 매우 많은 모델의 차이를 알 수 있었고, 머신러닝과 매우 가까워질 수 있었으며, 다양한 상황에 대처할 수 있게되었습니다. 또한 스타플레이어를 구별해주는 outlier 모델은 상당한 정확도를 보여줌으로써 스타플레이어인지 아닌지에 대한 정보를 참고적으로 제공해 줄 수 있게 되었습니다.
21. 다음은 한계입니다. 롯데 이대호와 같은 경우 메이저리그에서 금의환향하며 4년 100억 계약을 맺지만, 통계적으로 그에 못 미치는 성적을 보여주며 과대평가되었다고 할 수 있는 선수가 되어 버렸습니다. 이처럼 통계만으로는 보여지지 않는 수치를 계산할 수 없다는 단점이 있습니다. 또한 아웃라이어와 스타 플레이어에서는 만족할 만한 성능을 재현했지만 일반 플레이어 모델의 성능이 너무 낮다는 단점이 있어 개선이 필요합니다.
22. 마지막으로 2020년 연봉을 예측하며 마치겠습니다. 일반 플레이어 모델은 성능의 제한으로 실행하지 않았고 스타 플레이어들에 대해서만 순위를 매겨보았습니다. 1위 강백호, 2위 박병호, 3위 김하성 등 이렇게 순위가 나왔으며, 이제서야 연봉 협상이 진행되고 있기 때문에 결과는 맞는지 확인할 수 없습니다. 하지만 3위인 김하성이 이번 시즌에 골든 글러브를 받았고, 페르난데스는 최다안타왕, 박병호와 강백호는 원래 잘하던 선수임으로 이번 연봉 협상에서 충분히 가능성 있는 순위임을 다시 한 번 확인할 수 있었습니다. 감사합니다.