안녕하십니까, 팀 마기꾼들의 조장 이승희입니다.

저희 팀은 “경마 데이터 분석을 통한 경마 경기 결과 예측”을 목표로 프로젝트를 진행하였습니다.

--------------------------------------------------------------------------------------------------------------

목차는  
서론에서는 경마 산업 현황, 예측목표,  
본론에서는 크게 데이터, 모델, 실제예측  
마지막으로 결론에서는 한계 및 의의, 기대방향 순서로 진행됩니다.

--------------------------------------------------------------------------------------------------------------

1922년에 시작된 경마산업은, 올해로 100주년을 맞이하였습니다.

--------------------------------------------------------------------------------------------------------------

긴 역사만큼 항상 2조원을 웃돌았던 매출은 코로나로 인해 85%나 감소된 3천억원을 지나 작년 2천억원대까지 감소하였습니다.

--------------------------------------------------------------------------------------------------------------

그만큼 경마장에 출입하는 인구의 수도 감소하였고, 이에 한국 마사회에서는 지난 100년간의 경마 외에도 다양한 데이터를 축적해 미래사업부를 신설하고, 공공데이터 또한 100% 개방을 완료하여 새로운 가치 창출을 시도하고 있습니다.

이런 한국 마사회의 경향에 발 맞춰 프로젝트를 진행한다면 코로나 이후 위축되었던 경마 산업에도 도움이 될 것이라고 생각했습니다.   
--------------------------------------------------------------------------------------------------------------

다음으로 저희가 주제를 선정하게 된 배경을 말씀드리겠습니다.  
선행 분석을 살펴 본 결과 경마 결과를 예측한 사람들은 거의 없었습니다. 따라서 저희는 경마 데이터를 통해 경기 결과를 성공적으로 예측하고자 이 주제를 선정하게 되었습니다.

저희는 **경마 회차 별 데이터를 수집해 이를 중요한 변수를 중심으로 가공하고, 여러 모델을 시도하여 가장 순위 예측에 적합한 모델을 찾아 경주마가 3위 안에 드는 경우를 분류**하려고 합니다.

--------------------------------------------------------------------------------------------------------------

먼저 한국 마사회에는 서울, 제주, 부산경남 경마 데이터가 있는데 저희는 서울 경마에 대해서 예측하고자 서울 경마 데이터만 수집하였습니다.   
수집한 데이터는 경주 성적표, 부마/모마, 경주마 정보 등 10,000개(만개) 이상의 데이터를 크롤링 하였습니다.

---------------------------- **모델에 사용한 변수 및 전처리 과정 설명:**---------------------------------

다음으로 모델에 사용한 데이터 컬럼들에 대해 말씀드리겠습니다.

경주마 관련 정보로는 부마와 모마, 조교사와 기수의 점수 컬럼과, 말이 달린 레일의 번호를 나타내는 번호컬럼, 순위, 그리고 말의 부담중량, 마체중, 순위점수가 있습니다.

경기 관련 정보로는 경주날짜와 회차 데이터인 경주번호와 거리, 경기등급을 나타내는 군(등급)컬럼과 출전두수, 주로상태 컬럼이 있습니다.

각 변수에 대해서는 뒤에 더 자세히 설명해드리겠습니다.

--------------------------------------------------------------------------------------------------------------

**군(등급):**  
말은 자신의 능력을 수치로 부여 받고 이 수치는 경기 결과에 따라 증가하게 됩니다.  
그렇게 메긴 점수로 말의 등급이 매겨지게 되고,  
경기에 따라 출전할 수 있는 말의 등급이 정해집니다.

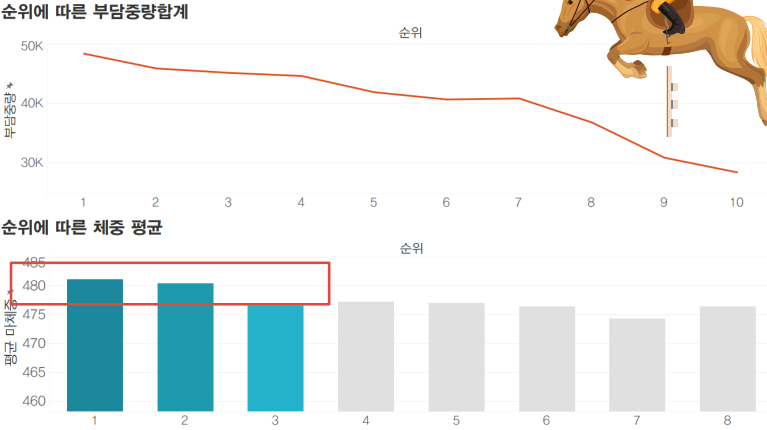
**출전두수:**  
출전두수는 한 경기에 출전하는 말의 수인데, 최소 7마리에서 최대 16마리가 한 경기에 출전합니다.

**주로상태:**  
주로 상태는 말이 달리는 경기장의 주로 상태를 나타냅니다. 서울의 주로는 모래와 흙을 혼합한 dirt 트랙입니다.  
"건조, 양호, 다습, 포화, 불량"으로 이루어진 범주형 데이터로 0부터 4까지 라벨인코딩을 하였습니다.

**부담중량:**  
부담중량은 기수의 체중과 안장 등의 무게를 합친 것으로 공정하고 박진감 넘치는 경주를 위해 만든 핸디캡입니다. 쉽게 말해 밸런스 패치로 경주에 큰 영향을 끼치는 변수로 생각해주시면 될 것 같습니다. 부담중량은 이전성적 혹은 상금 수득으로 결정되며 최대 상한 60kg을 초과하지 않습니다.

--------------------------------------------------------------------------------------------------------------

그래프를 보시면 부담중량과 순위가 반비례 한다는 것을 알 수 있습니다  
성적이 좋은 말일 수록 부담중량이 증가하게 됩니다. 즉, 패널티를 더 받게 되는 것이죠..



--------------------------------------------------------------------------------------------------------------

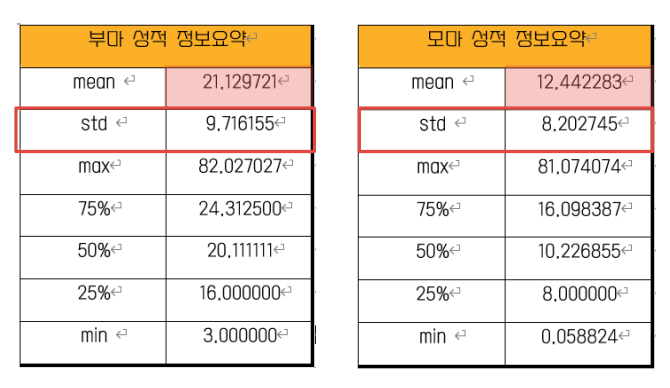
다음으로 저희가 모델을 돌리는 데에 있어 계산을 한 **파생변수에 대해 설명**드리겠습니다.

**혈통(부마 전적, 모마 전적):**경마는 **혈통의 스포츠**라고 할 정도로 혈통이 중요합니다.   
특히 잘 달리는 부마의 혈통이 중요한데 그래서 주로 "능력이 좋은 부마"를 "건강한 모마"와 교배시켜 경마를 위한 말을 생산하기도 합니다. 이를 데이터 안에서도 알 수 있었습니다.

--------------------------------------------------------------------------------------------------------------

부마의 경우 결측값이 3개 밖에 없었지만 모마의 경우는 결측치많앗다

하지만 모마 성적의 분포도를 보면 표준편차가 굉장히 적고 작은 것을 알 수 있기에 부마와 모마데이터 모두 평균으로 결측값을 대체하였습니다.



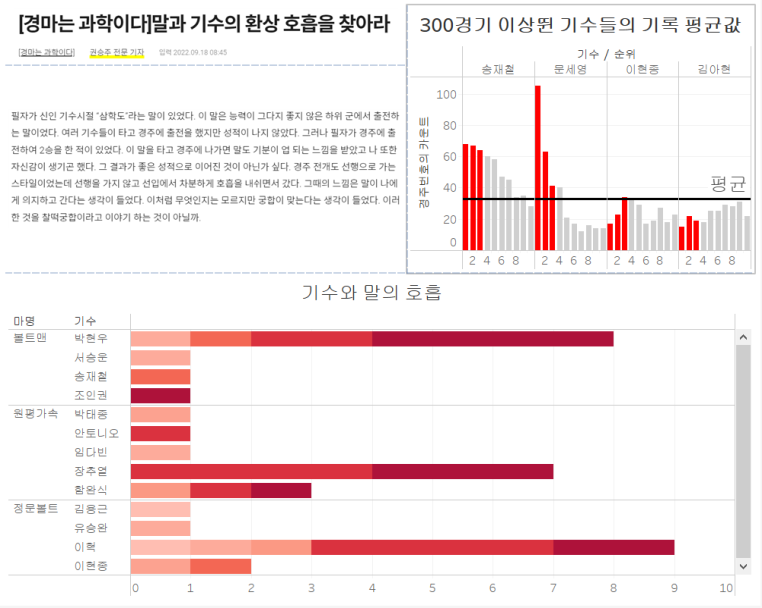
--------------------------------------------------------------------------------------------------------------

**기수, 조교사 전적:**

기수는 말과 함께 경마를 뛰는 사람이고, 조교사는 말을 경마 시스템에 맞게 훈련 시키는 조교와 같은 사람이라고 생각하시면 됩니다. 따라서 한 말의 조교사는 단 한명인 반면 기수는 경주마다 바뀌게 됩니다.

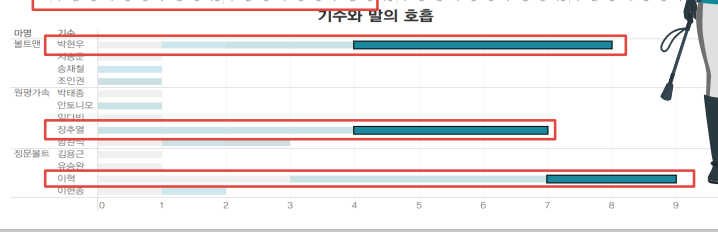
--------------------------------------------------------------------------------------------------------------

말에 탄 기수의 기승능력은 경기 결과에 큰 영향을 미칩니다. 이를 시각화한 그래프를 보시게 되면, 다음과 같습니다.

이는 300회 이상 경기를 뛴 기수들의 순위 개수의 합계를 시각화한 결과로 왼쪽의 두 선수와 오른쪽의 두 선수에 능력에 차이가 있음을 알 수 있습니다.  
여기서 빨간색은 1등에서 3등까지의 카운트 횟수를 나타내며, 4등 이하는 회색으로 나타냈습니다.  
왼쪽의 두 기수는 3등안에 든 횟수가 총합 200회 가까이 하였고, 오른쪽 두 기수는 3등안에 든 횟수가 총합 100회 미만인 것을 알 수 있습니다.

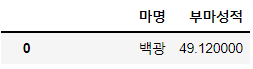
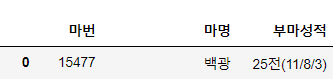
--------------------------------------------------------------------------------------------------------------

말은 사람과 교감을 하는 동물입니다. 따라서 *말과 사람이 만났을 때 나타나는 시너지의 효과도 경마 결과에 큰 영향을 끼칩니다.*  
아래 그래프를 보시면, 같은 말을 탄 4-5명의 기수들의 성적을 나타낸 그래프 입니다.  
진한 초록색은 3등안에 든 횟수의 누적합이고, 연한 색은 4등 이하의 누적합입니다.  
이런 결과는 기수의 개인 능력뿐만아니라 말과 기수의 조합 또한 중요하다는 것을 알 수 있습니다.



--------------------------------------------------------------------------------------------------------------

**스코어링화한 방법 설명**:

  
전적 데이터는 가운데 보시는 것과 같이 25전, 11,8,3이 써져있습니다. 전 앞에 쓰여있는 숫자는 총경기 출전수이고 괄호 안에 슬래쉬로 구분된 세개의 수가 각각 1등,2등,3등을 한 총 횟수입니다. 이때 단순히 평균을 내게 되면 1등을 3번한 말과 3등을 3번한 말의 평균이 같게 됩니다.

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++  
그래서 등수 별로 가중치를 두어 스코어링 시켜 다음과 같이 변환된 결과를 확인할 수 있습니다.

++++++++++++++++++++++++ **전처리 과정추가설명**++++++++++++++++++++++++++

저희가 모델에 넣은 데이터 프레임 구조에 대해 설명드리겠습니다.

모델을 훈련 시킬 때 한 레코드에 한 경기의 모든 말의 데이터가 모두 들어가도록 했습니다.

*Ex) 예를 들어 12마리가 출전하는 경기에서 주축으로 학습시키고자 하는 1번 말의 정보와 그 옆에는 2번부터 12번 말의 정보를 옆으로 쌓아주어 파라미터 수를 늘렸습니다.*

여기서 옆으로 더 쌓아준 다른 말들의 정보는 경기데이터를 제외한 나머지 데이터를 넣었습니다.

이렇게 만든 데이터프레임에 결측값이 생기는 경우가 발생하였는데,  
바로 12마리가 뛴 한 경기에서 말의 정보가 1마리밖에 존재하지 않았던 경우입니다.  
이런 경우는 다른 11마리의 말이 은퇴한 경우였기 때문에 정보가 존재하지 않았습니다.  
이런 경우가 다수 존재하여, **한 경기 내에서 정보가 존재하는 말의 수가 6개 미만인 경우는 분석에서 제외**하였습니다

--------------------------------------------------------------------------------------------------------------  
그래서 6~12마리의 데이터 중 빈 데이터에 대해 **null값을 다음과 같이 평균으로 대체**해주었습니다.



--------------------------------------------------------------------------------------------------------------

특히 (기수,조교사)의 경우 보시는 것과 같이 점수들 간의 값의 차이가 큰 것을 발견하였고, 이는 기수마다 경기를 한 횟수의 차이가 크기 때문임을 발견하였습니다.   
기수마다 4000번 넘게 뛴 기수가 있는 반면, 50회 미만의 경기를 뛴 기수도 있었기에 값간의 차이가 큰 점수화 변수를 minmax정규화를 해주었습니다.





--------------------------------------------------------------------------------------------------------------

**Y값 선택**

분류 모델을 사용해 y값인 등수를 1~3등을 1로 두고 나머지를 0으로 두었습니다.

--------------------------------------------------------------------------------------------------------------

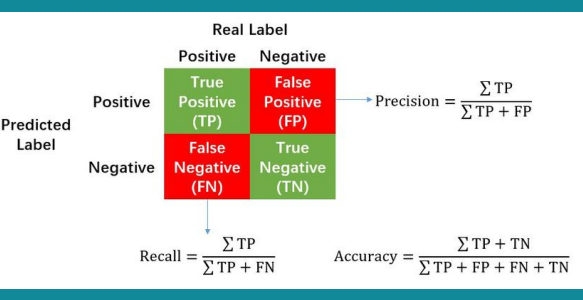
**모델선택과정:**

데이터 특성상 결측치를 처리를 하였으나 null값이 많아 의미 없는 피처를 탈락시키는 트리 구조의 모델이 좋다고 판단했습니다.  
tree베이스 RandomForest, XGBoost, light GBM이 3가지 모델을 학습시켜보았고, 변수가 많아 optuna를 사용하였습니다.

--------------------------------------------------------------------------------------------------------------

**모델 성능 확인 지표 선택 과정: (**Precision)

저희 프로젝트의 목적은 경마 경기 결과 예측입니다.  
따라서 우승할 경마에 대한 예측이 얼마나 잘 맞는지가 중요합니다.  
따라서 **우승(Positive)이라 예측한 것(TP+FP)중에서 얼마나 우승을 많이 했는지(TP)가 중요**합니다.  
따라서 Precision 을 모델 성능 확인 지표로 선택하였습니다.



--------------------------------------------------------------------------------------------------------------

**분석결과:**

다음으로 분석결과입니다.  
첫행은 베이스 라인으로 모델을 돌렸을 때 나온 결과입니다.  
가운데 행은 옵튜나로 해본 결과로 train데이터에서 모두 70% 가량 나왔지만,   
test에서 LightBGM의 결과가 가장 유의미하게 나왔습니다.

옵튜나로 해본 결과 트레인이 모두 70%가량 나왔지만 test결과에서 실제 데이터에 적용해본 결과 1위를 예측한 것을 볼 수 있습니다.

--------------------------------------------------------------------------------------------------------------

**실제예측:**

실제 데이터를 넣었을 때 RF는 애매하게 0과 1을 분류하였으나 예측결과를 0으로 모두 예측하였고,

XGB는 모든 것이 0으로 편향된 문제가 있었습니다.

마지막으로 LGBM은 1로 예측한 것이 1등으로 나왔으나 3위안에 들것이라고 예측을 한 개수가 그 외에 없어서 2,3위는 아예 예측하지 못했습니다.

--------------------------------------------------------------------------------------------------------------

**모델선택:**

원래 여러 모델을 Voting해서 결정하기로 하였으나 다른 모델들의 과적합이 너무 심해서 **light GBM** 모델만을 사용하기로 결정하였습니다.

--------------------------------------------------------------------------------------------------------------

**한계:**

모델에 대한 이해 부족으로 인한 파라미터 및 하이퍼 파라미터 수정과정의 난항이 있었습니다.

--------------------------------------------------------------------------------------------------------------

**결론:**

1. 데이터 수집 과정 자체에서 중복되어 사용되는 데이터가 많아서 과적합될 수 밖에 없는 구조 2. 모델에 대한 이해 부족으로 인한 파라미터 및 하이퍼 파라미터 수정과정의 난항

--------------------------------------------------------------------------------------------------------------

**의의:**

그래도 한정적인 변수로 어느정도 유의미한 예측을 할 수 있었음. 데이터 종류가 많아지고, 조금 더 정교한 파라미터 수정이 있다면 명확한 예측이 가능할 것이라는 가능성 발견 ->죽어가던 경마 산업의 관심 제고

--------------------------------------------------------------------------------------------------------------

**발전방안:**

1. 데이터를 더 수집- INDEPENDENT한 FEATURE를 더 수집  
2. 과적합이 그나마 덜 된 LGBM 파라미터 수정을 더 발전시킨다

**발표시간 : 20분 이내로 무조건**