안녕하십니까, 팀 마기꾼들의 조장 이승희입니다.

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

경마는 주식과 같이 셀 수 없이 많은 요인들로 인해 결과가 만들어집니다. 하지만 경마를 예측한 사람들은 거의 없었습니다. 따라서 저희는 경마 데이터를 통해 경기 결과를 성공적으로 예측하고자 주제를 선정하게 되었습니다.

1922년에 시작된 경만산업은, 올해로 100주년을 맞이했습니다. 긴 역사만큼 항상 2조원을 웃돌았던 순매출은 코로나로 인해 85%감소된 3천억원을 지나 작년 2천억원대까지 감소하였습니다. 그만큼 경마장에 출입하는 인구의 수도 비례해 감소하였습니다.

이에 한국 마사회에서는 지난 100년간의 경마, 축산, 등 다양한 분야의 데이터를 축적해 데이터 활용을 위한 미래사업부를 신설하고, 공공데이터 또한 100% 개방을 완료하여 빅데이터로 새로운 가치 창출을 시도하고 있습니다.

+경마 산업이 코로나 이전까지 사행성 사업 중 1위로 매출액이 제일 높고 규모가 큰 사업이었습니다. 따라서 코로나로 위축이 된 경마 산업의 부흥을 기대하며 이에 맞춰 저희 프로젝트는 새로운 경마 경기에 대해 새로운 예측 지표가 될 수 있다고 생각합니다.

목차는 ~~,~~,~~ 순서로 진행됩니다.

먼저 한국 마사회에서 서울,제주, 부산경남 경마데이터 중 서울 경마만을 수집하였습니다.
수집한 데이터는 경주 성적표, 부마/모마, 경주마 정보 등 10,000개(만개)의 데이터를 크롤링 하였습니다.

모델에 사용한 변수 및 전처리 과정 설명:
먼저 모델에 사용한 컬럼들과 전처리 과정에 대해 설명해드리겠습니다

군(등급):
말은 자신의 능력을 수치로 부여받고 경기 결과에 따라 증가하게 됩니다.
그렇게 메긴 점수로 말의 등급이 매겨지게 되고,
한 경기에는 같은 등급의 말들이 경쟁을 하게 되는 겁니다.

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 모델링을 위해 라벨인코딩을 하였습니다.

부담중량:

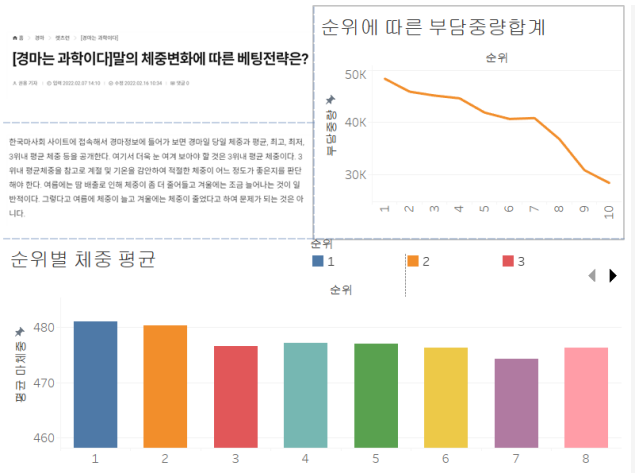
부담중량은 기수의 체중과 안장 등의 무게를 합친 것으로 공정하고 박진감 넘치는 경주를 위해 만든 핸디캡입니다. 쉽게 말해 밸런스 패치로 경주에 큰 영향을 끼치는 변수로 생각해주시면 될 것 같습니다.

부담중량은 이전성적 혹은 상금 수득으로 결정되며 성적이 좋은 말일 수록 부담중량이 증가하게 됩니다.

즉, 패널티를 더 받게 되는 것이죠.

최대 상한 60kg을 초과하지 않으며 이 또한 말의 레이팅 등급에 따라 다르게 주어집니다.

다음과 같은 사진을 보면 부담중량과 순위가 연관이 있다고 볼 수 있습니다.



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

[[[[[[[[********** **순위점수**, **거리**(100으로 나눔), [**부마, 모마, 조교사, 기수**]의 전적 **********]]]]]]]]

(큰 분류로 전적변수(부마,모마,기수,조교사)와 아닌 것으로 나눴습니다)
(나눈 이유는 전적변수들의 전처리 과정이 같기 때문입니다) **to 효은언니**

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.
따라서 부마와 모마의 전적 데이터를 이용해 각각의 혈통을 점수화 시켰습니다.
점수화하는 방법에 대해서는 뒤에서 더 자세히 설명하겠습니다.

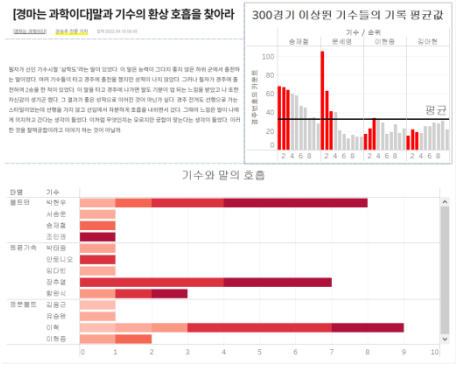
	마번	마명	부마성적		마번	마명	모마성적
0	15477	백광	25전(11/8/3)	0	9519	선봉대감	22전(3/3/3)

기수, 조교사 전적:
말은 사람과 교감을 하는 동물입니다. 따라서 *말과 사람이 만났을 때 나타는 시너지의 효과도 경마 결과에 큰 영향을 끼칠 것이라는 가설을* 세우게 되었고, 이를 수치화 시켰습니다.

이때 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.

	기수명	통산전적		조교사명	통산전적
0	김귀배	4550전(326/326/386)	0	강성오	703전82/81/60

이때 저희는 기수에 집중해 시각화를 한 결과 300회 이상 경기를 뛴 기수들간에 전적에 차이가 있음을 발견했습니다.



이와 같이 경기를 많이 뛰었음에도 1,2,3등을 많이 한 기수와 그렇지 못한 기수가 있다는 것을 알 수 있습니다.
그래서 부마,모마와 같이 기수,조교사 전적 데이터 또한 같은 방법으로 전처리하여 점수화 시킨 후 모델에 학습시켰습니다.

스코어링화한 방법 설명:

	마번	마명	부마성적
0	15477	백광	25전(11/8/3)

전적 데이터는 [X0전(X1/X2/X3)]. --→ [총경기출전수(1등/2등/3등)]
전 앞에 쓰여있는 숫자가 총경기 출전수이고 괄호 안에 슬래쉬로 구분된 세개의 수가 각각 1등,2등,3등을 한 총 횟수입니다.
이를 각 등수별로 가중치를 두어 점수화 시켰습니다.
(1등 * 3 + 2등 * 2 + 3등 * 1) / 총출전수 --> 이렇게

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

다음으로 좀 더 자세한 **전처리 과정**(**null값 처리**)을 설명하겠습니다.

1. 부마 모마 결측값 처리

경마의 경우 부마의 혈통이 중요하다고 하였는데, 이때 부마의 경우 결측값이 3개 정도로 거의 없어 평균대체를 사용하였습니다.

모마의 경우는 전체 약 1300개 중 170개의 데이터가 결측치로, 단순히 임신을 위한 목적으로 쓰인 경우라고 생각됩니다.

하지만 모마 성적의 분포도를 보면 표준편차가 굉장히 적고 작은 것을 알 수 있기에 이 또한 평균으로 결측값을 대체하였습니다.

```
In [346]: df_fm['모마성적'].describe()
Out[346]: count    1311.000000
          mean       12.442283
          std         8.202745
          min         0.058824
          25%         8.000000
          50%        10.226855
          75%        16.098387
          max        81.074074
          Name: 모마성적, dtype: float64
```

++

데이터프레임 구조 설명:

마지막으로 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 (부담중량, 마체중, 순위점수, 조교사/기수/부마/모마의 전적)데이터를 넣었습니다.
처음에는 모든 정보를 넣었다가, 1마리의 말을 제외한 다른 말들의 등수를 넣어주면 자동으로 남은 등수를 알게 되어 **과적합**이 되는 문제를 발견해,
경기 데이터에 고유한 수치는 다 넣고, **경기 순위에 직접적으로 영향을 주는 변수는 제거**하였습니다.

전처리 과정 설명: (데이터프레임 자체(?) 전처리)

모델을 돌리기 위한 데이터 프레임을 구성하 때 결측값이 생기는 경우가 있었습니다.
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.
그럼에도 불구하고 각 경기마다 출전두수, 즉 출전하는 말의 수가 달라 데이터를 쌓을 때 null값이 생기는 문제가 발생하였습니다.
그래서 6~12마리의 데이터 중 빈 데이터에 대해 **null값을 다음과 같은 임의로 대체**해주었습니다.
마번:40000 / 순위:17 / 기록:max+10 / 1F,G3F,G1F기록:max+1 / 부담중량,마체중:mean / 순위점수:0 / 기수,조교사,모마,부마:mean

---

Y값

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

모델선택과정:

데이터 결측치를 처리를 하였으나 특성상 처리된 null값이 많아 의미 없는 피처를 탈락시키는 트리 구조의 모델이 좋다고 판단했습니다.
따라서 tree베이스인 모델로 학습할 경우 성능이 좋을 것이라는 가설을 세웠고, tree베이스 light GBM, XGBoost, RandomForest 이 3가지 모델을 학습시켜보았습니다.

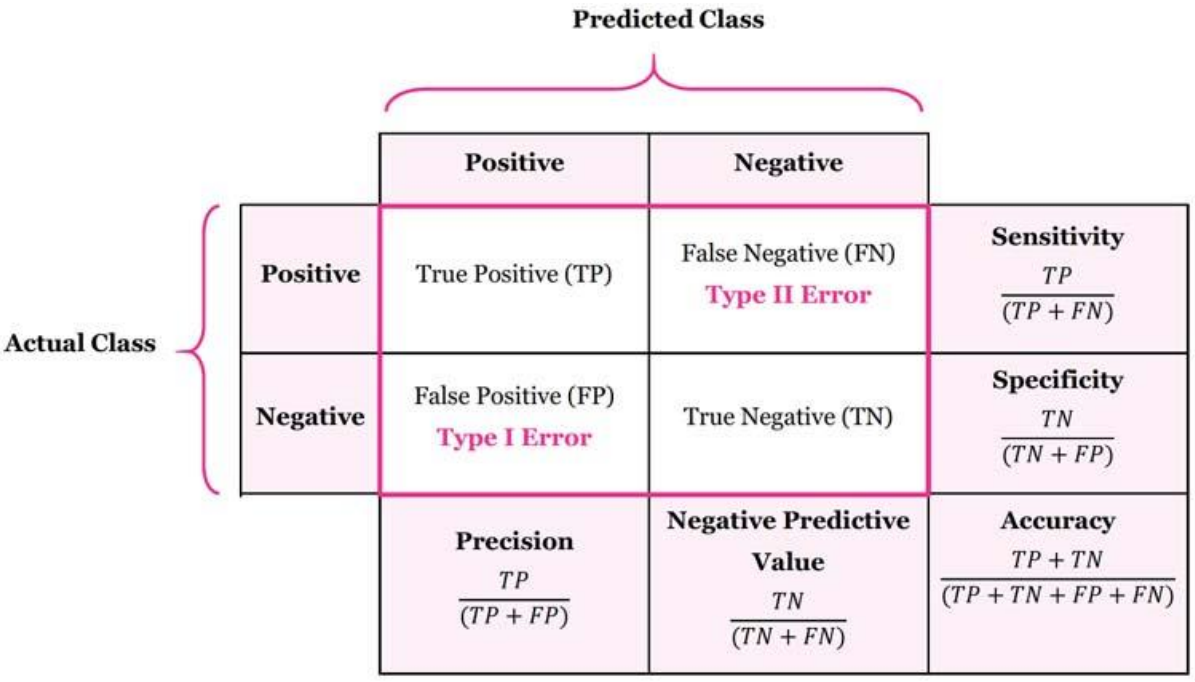
또한 변수가 많아 optuna를 사용하였습니다.

원래 여러 모델을 Voting해서 결정하기로 하였으나 다른 모델들의 과적합이 너무 심해서 XGBoost 모델만을 사용하기로 결정하였습니다.

~~아직 수정 중~~뭐가 좋은지 나오면 그거 따라서 추가 예정

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.



모델 성능 향상

전처리 했던 전적 관련 변수들 중, (부마,모마)를 제외한 (기수,조교사)의 경우 기수 조교사에 따른 경기 횟수의 차이가 커 가중치 처리를 했을 때 값 간의 차이가 컸습니다.
따라서 기수, 조교사의 전적 점수화시킨 변수를 minmax정규화를 해주었습니다.

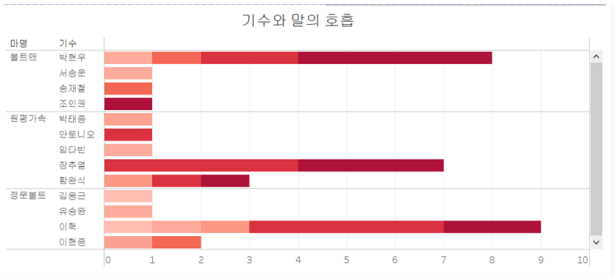
2	곽영효	2455.087712

8	문병기	40.107692

아쉬운 점 :

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

(모델 성능 향상에 대해 아쉬운 점)
첫째로,
전적 데이터 중 기수의 전적 데이터와 말 과의 조합에 가중치 부여하는 변수를 만들지 못한 것이 아쉽습니다. 같은 말임에도 기수에 따라 좋은 성적을 내는 말이 있어 더 큰 가중치를 부여한다면 모델의 성능을 향상 시킬 수 있을 것이라 예상됩니다.



둘째로,
같은 전적 데이터 중 부마의 혈통이 모마의 혈통보다 더 중요하다는 가설에 기반하여 더 가중치를 둔다면 모델의 성능을 향상 시킬 수 있을 것이라 예상됩니다.

https://www.weeklytrade.co.kr/m/content/view.html?&section=1&no=67025&category=160

프로젝트 기대방안(?):

경마장에서는 경마 순서를 맞추는 마권을 파는데 이때 맞추고자 하는 등수에 따라 승식이 7가지

가 있습니다.
저희 프로젝트를 기반으로 다음 경마를 예측한다면,
경마에 가서 마권을 구입하실 때 **연승식**을 구매하시면 될 것 같습니다.

연승식 - 고른말이 1, 2, 3 등 안에 들어오면 적중한다.

+정보+

[승식 - 마권을 구입할 때 승식을 골라 구매 할 수 있다. 승식은 7개로 나뉜다.]

단승식 - 고른 말이 1등으로 들어오면 적중된다.
연승식 - 고른말이 1, 2, 3 등 안에 들어오면 적중한다. 맞히기 가장 쉽고 배당금이 가장 적은 편이다.
복승식 - 말 2마리를 골라 1, 2등으로 들어와야 하며 순서는 상관 없다.
쌍승식 - 말 2마리를 골라 1, 2등으로 들어와야 하며 순서까지 맞춰야 적중한다. 복연승식 - 말 2마리를 골라 3등 이내에 모두 들어오면 오면 적중한다.
삼복승식 - 말 3마리를 골라 1, 2, 3등으로 모두 들어와야하며 순서는 상관없다.
삼쌍승식 - 1,2,3 등을 순서대로 맞히는 승식이다. 맞추긴 어렵지만 배당이 가장 높은 편이다.

발표시간 : 20분 이내로 무조건