경마 데이터 기반 경마 순위 예측을 위한 종합적인 분석과 모델링

이경재*
*중앙대학교

*kyungjae.lee@ai.cau.ac.kr

Comprehensive Analysis and Modeling for Horse Racing Ranking Prediction Based on Racing Data

Kyungjae Lee*
*Chung-Ang Univ.

요 약

최근 스포츠 예측 및 분석 분야에서는 경마 데이터의 활용이 주목받고 있다. 본 연구는 경마 데이터를 종합적으로 분석하고 경주 결과를 예측하기 위해 머신 러닝 모델을 활용하고자 한다. 중복 데이터 문제를 해결하고 다양한 특성을 추가함으로써 정확한 예측 모델을 구축하는 것이 주요 목표이다. 중복 데이터 처리, 경마마 등급 및 총 상금과 같은 새로운 특성 도입, t-SNE 를 통한 데이터 분포 시각화 등을 통해 예측 정확도를 향상시켰다. 또한, Plackett-Luce 및 Bradley-Terry 모델을 적용하여 모델을 최적화했다. 정확도 및 Top-3 정확도를 활용한 성능 평가 결과, 새로운 특성의 추가로 인해 상당한 성능 향상 확인되었다.

Ⅰ. 서 론

최근 들어 스포츠 예측 및 분석은 다양한 분야에서 큰 주목을 받고 있으며, 그 중에서도 경마 데이터의 활용은 중요한 연구 주제 중 하나로 부각되고 있다. 본 연구에서는 경마 데이터를 종합적으로 분석하고, 머신 러닝 모델을 활용하여 경마 경주 결과를 예측하는 방법에 대한 연구를 수행하였다. 이 연구의 목적은 경마 데이터를 효과적으로 활용하여 경주 결과를 예측하는 머신 러닝 모델을 개발하고자 함에 있다. 특히, 데이터의 중복 문제와 다양한 특성을 고려하여 정확한 예측 모델을 구축하는 것을 목표로 한다.

Ⅱ. 본론

2.1 데이터 크롤링과 이슈 해결

본 연구에서는 다양한 데이터 소스로부터 경마 데이터를 크롤링하였다. 수집한 데이터에서 다양한 특성들을 고려하여 모델의 예측 성능을 향상시켰다. 예를 들어, 경주마의 등급, 총 상금, 경주 조건, 기수의 성적 등이 특성으로 추가되었다. 자세한 특성은 아래와 같다.

raceResultDetail.rank: 순위를 나타내는 정수 (int32)
raceResultDetail.raceNo: 경주 번호 (int32)
raceResultDetail.name: 선수 이름 (문자열)
raceResultDetail.country: 국가 (범주형)
raceResultDetail.sex: 성별 (범주형)
raceResultDetail.age: 나이 (float32)
raceResultDetail.raceWeight: 경주 중량 (float32)
raceResultDetail.winRate: 승률 (float32)
raceResultDetail.placeRate: 단승률 (float32)
raceResultDetail.raceHorseProfile.sex: 경주마의 성별 범주형)
raceResultDetail.raceHorseProfile.hometown: 경주마의 고향 (범주형)
raceResultDetail.raceHorseProfile.birth: 경주마의 출생일 (datetime64[ns])
raceResultDetail.raceHorseProfile.age: 경주마의 나이 (float32)
raceResultDetail.raceHorseProfile.color: 경주마의 색상 (범주형)
raceResultDetail.raceRiderProfile.totalWinRate.total: 기수의 총 승률 (float32)
raceResultDetail.raceRiderProfile.totalWinRate.firstWinCount: 기수의 1 등 승리 횟수 (float32)
raceResultDetail.raceRiderProfile.totalWinRate.secondWinCount: 기수의 2 등 승리 횟수 (float32)
raceResultDetail.raceRiderProfile.totalWinRate.thirdWinCount: 기수의 3 등 승리 횟수 (float32)
raceResultDetail.raceRiderProfile.totalWinRate.fourthWinCount: 기수의 4 등 승리 횟수 (float32)
raceResultDetail.raceRiderProfile.totalWinRate.fifthWinCount: 기수의 5 등 승리 횟수 (float32)
raceResultDetail.raceRiderProfile.recentWinRate.total: 기수의 최근 총 승률 (float32)
raceResultDetail.raceRiderProfile.recentWinRate.firstWinCount: 기수의 최근 1 등 승리 횟수 (float32)
raceResultDetail.raceRiderProfile.recentWinRate.secondWinCount: 기수의 최근 2 등 승리 횟수 (float32)
raceResultDetail.raceRiderProfile.recentWinRate.thirdWinCount: 기수의 최근 3 등 승리 횟수 (float32)
raceResultDetail.raceRiderProfile.recentWinRate.fourthWinCount: 기수의 최근 4 등 승리 횟수 (float32)
raceResultDetail.raceRiderProfile.recentWinRate.fifthWinCount: 기수의 최근 5 등 승리 횟수 (float32)
raceResultDetail.horseWeight: 경주마의 체중 (float32)

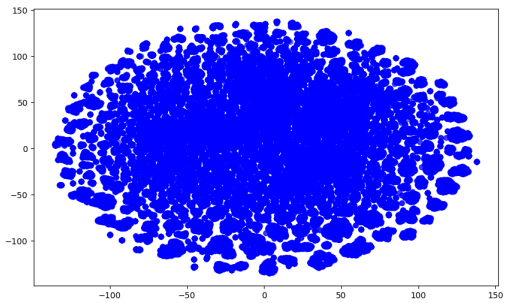
raceResultDetail.horseWeightDiff: 경주마의 체중 변화 (float32)
raceRecordDetail.rank: 순위 (int32)
raceRecordDetail.raceNo: 경주 번호 (범주형)
raceRecordDetail.furlong3fg: 3Furlong 타임 (float32)
raceRecordDetail.furlong1fg: 1Furlong 타임 (float32)
raceRecordDetail.raceRecord: 경주 기록 (float32)
raceDate: 경주 일자 (문자열)
round: 라운드 (int32)
category: 카테고리 (범주형)
grade: 등급 (범주형)
distance: 거리 (float32)
courseCondition: 코스 상태 (범주형)
weather: 날씨 (범주형)
region: 지역 (범주형)

위의 특성들 외에도 새로운 feature 들을 고려하여 모델을 구성하였다. 새롭게 추가된 feature 들은 다음과 같다.

raceResultDetail.totalPrize: 해당 경주의 총 상금 (float32)
grade: 경주 등급을 나타내는 범주형 변수 (category)

2.3 t-SNE 를 활용한 데이터 분포 확인

t-SNE [1] 를 통해 데이터의 분포를 시각적으로 확인하였다. 각 feature 들이 어떻게 분포되어 있는지를 확인함으로써 모델 학습에 적절한 feature 를 선정하는 데 도움을 주었다. 선정된 feature 를 활용한 t-SNE 분포는 아래와 같다. 여러 데이터가 널리 분포된 것을 확인 할 수 있다.

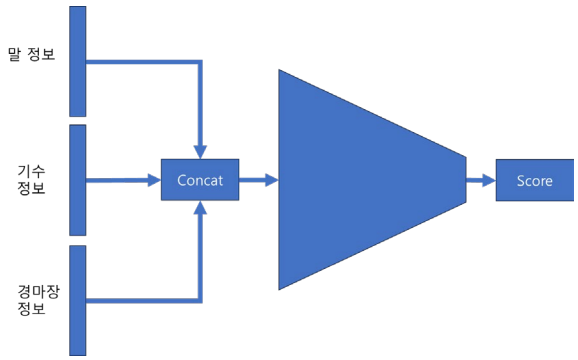


2.4 Plackett-Luce Model 및 Bradley-Terry Model 적용

경마 예측 모델을 위해 Plackett-Luce Model [1] 및 Bradley-Terry Model [2] 을 활용하였다. 각 말과 기수의 상호 작용을 반영하여 가상의 점수를 예측하고, 학습을 통해 모델을 최적화하였다. 확률모델은 아래와 같다.

$$p((i_1, i_2, \cdots, i_N); \boldsymbol{w}) = \prod_{j=1}^{N} \frac{\exp(w_{i_j})}{\sum_{l=j}^{N} \exp(w_{i_l})}$$

ik 들은 각 말의 등수를 의미한다. wi 들은 각 말들의 잠재 점수를 의미하며 이를 뉴럴네트워크를 활용하여 예측하도록 하였다. 주어진 경주 순위에 대해 위의 확률 모델을 계산하고 최대우도를 갖도록 뉴럴네트워크의 파라미터를 학습하였다. 네트워크의 구조는 아래와 같다.



2.5 성능 평가

모델의 성능을 평가하기 위해 정확도 (Accuracy)와 상위 3 개 예측 중 정확도 (Top-3 Accuracy)를 계산하였다. 이전에 비해 새로운 feature 들의 추가로 정확도가 향상되었음을 확인하였다.

기존 feature 사용 시: Acc 20.61, Top-3 Acc 24.11
totalPrize 추가 시: Acc 24.24, Top-3 Acc 30.92
grade 추가 시: Acc 34.19, Top-3 Acc 46.6

Ⅲ. 결론

본 연구에서는 다양한 경마 데이터를 활용하여 경주 결과를 예측하는 종합적인 분석을 수행하였다. 중복 데이터 처리, 추가 특성 고려, t-SNE 를 통한 데이터 분포 확인, 그리고 특정 모델의 적용 등을 통해 모델의 예측 성능을 향상시켰다. 그러나 본 연구에서는 아직 해결되지 않은 문제들이 존재한다. 향후에는 더 다양한 특성들을 고려한 모델의 개발 및 성능 평가가 필요하며, 데이터의 동적인 변화에 대응할 수 있는 안정적인 크롤링 방법에 대한 연구가 필요하다. Acc 와 Top-3 Acc 를 고려하여 평가 지표를 확장하고 모델의 예측 능력을 더욱 심층적으로 분석할 필요가 있다.

참 고 문 헌

[1] Van der Maaten, Laurens, and Geoffrey Hinton. "Visualizing data using t-SNE." *Journal of machine learning research* 9.11 (2008).

[2] Maystre, Lucas, and Matthias Grossglauser. "Fast and accurate inference of Plackett- Luce models." *Advances in neural information processing systems* 28 (2015).

[3] Hunter, David R. "MM algorithms for generalized Bradley-Terry models." *The annals of statistics* 32.1 (2004): 384-406.