회귀 분석을 통한 경마 순위 예측 모형

허태성[O], 송민섭[*], 고동수[*]
[O]인하공업전문대학 컴퓨터정보공학과,
[*]인하공업전문대학 컴퓨터정보공학과
e-mail: tshur@inhatc.ac.kr[O], magnet9805@naver.com[*], sky7ds@naver.com[*]

A Model for Predicting Horse Racing Ranking by Regression Analysis

Hur Tai-sung[O], Song Min Seob[*], Ko Dong Su[*]
[O]Dept. of Computer Science Engineering, Inha Technical College,
[*]Dept. of Computer Science Engineering, Inha Technical College

● 요 약 ●

본 논문에서는 국내 합법 사행산업의 가장 큰 비중을 차지하는 경마에 대한 데이터 분석 모델을 제공하여 건전한 국민 여가 스포츠로 인식 개선을 제안한다. 고배당을 강조하는 경마 예측론이 성행하며 경마가 스포츠가 아닌 도박에 가깝다는 부정적 이미지를 개선하고자 부모마의 수득 상금을 이용한 순위 분석 모델을 제공한다. 현재 국내 경마 경기는 서울, 부산, 제주에서 개최되며, 이 중 서울 지역 경마 데이터를 분석 데이터로 하였다. 분석에 이용한 데이터는 2019년 3월부터 2022년 3월까지의 경주 성적, 경주마 정보, 부모마 수득상금을 이용하였다. 분석에는 선형 회귀 모형, 랜덤 포레스트 회귀 모형 (Breiman, 2001)을 이용하였다. 분석은 Python 을 이용하였으며, Python에서 제공하는 다양한 라이브러리를 이용하여 크롤링, 전처리, 분석하였다.

키워드: 랜덤 포레스트 회귀(Random Forest Regression), 크롤링(Crawling)

I. Introduction

2019년 사행산업통합감독위원회에서 제공하는 자료에 따르면 매출액을 기준으로 경마는 국내 합법 사행산업의 매출액의 32.4% 비중인 가장 큰 비중을 차지했다. 그러나 2020년 코로나 영향으로 체육진흥 투표권(38%)과 복권(42.1%)에 뒤를 이은 8.5% 비중을 차지하며 3위에 머물렀다. 파이썬을 이용한 선형회귀, 랜덤포레스트 회귀 분석으로 경마 순위 예측 모델을 제공하고 이를 통해 참여자들의 정보 활용성이 높아지고 합리적인 예측이 가능해짐으로써 경마가 건전한 국민 여가 스포츠로서의 기능을 할 수 있도록 한다.

II. Preliminaries

자료 수집에 있어 본 연구에 필요한 데이터 수집 방법은 크롤링이다. 크롤링에서 정적 url 부분은 Python 라이브러리인 BeautifulSoup를 이용하며 동적 url 부분은 Selenium을 이용하여 Xpath와 css Selector 정보를 추출한다[1]. 본 연구에서 진행 할 분석 방법은 부모마 수득상금에 따른 순위 예측으로 회귀 분석이 필요하며 분석에는 최근 선형 회귀 모형, 랜덤포레스트 모형을 이용하는 추세이다[2].

III. The Proposed Scheme

본 연구에 사용할 데이터는 한국마사회(www.kra.co.kr)에서 제공하는 경주마 정보, 부모마 정보, 부모마 수득상금 데이터이며 크롤링을 이용해 수집한 데이터를 Pandas와 Mysql Database에 저장하며 데이터 병합을 완료하여 총 30370개의 데이터를 수집하였다.

출전 취소 등 순위 column의 결측치가 있는 639개의 행을 제거했으며, 부모마 수득 상금 중 원화 단위가 아닌 USD, JPY 등의 해외 수득 상금은 서울외국환중개 기준 2019년~2021년의 평균 환율 데이터를 적용해 원화 단위로 변경했다. 회귀 분석에 사용할 라이브러리인 scikit-learn 라이브러리에선 모든 feature를 수치화 해야하기 때문에 문자열 제거 후 float 타입으로 변환하였다. Fig. 1. 과같이 전처리가 끝난 데이터는 csv 형식으로 저장 및 관리하였다.

	A	B	C
1	ranked	fmoney	mmoney
2	1	574419755	41484000
3	2	1220375040	69370000
4	3	3588156497	256115000
5	4	574419755	123741851
6	5	59884471	57878000
7	6	4464612000	95013592
8	7	2106636213	156911000
9	8	1826932981	307463200
10	9	2014956840	197913668
11	10	3588156497	0
12	11	3707930092	0
13	12	1636444032	22381561

Fig. 1. Processed data for analysis

분석에 들어가기에 앞서 2019월 3월~2021년도 데이터를 train set, 2022년 1월~3월 데이터를 test set으로 분리하였다[3]. 첫번째로, 부모마 총 수득 상금에 따른 순위를 선형 회귀로 분석하였다. 선형 회귀 시 상관관계가 없었으며 수득 상금을 상.중.하 단계로 binning 하여 선형 회귀, Fig. 2. 을 보면 6단계로 binning 하여 선형 회귀 한 결과도 상관관계가 없다는 결론이 나왔다[4]. 두번째로, 부마 수득 상금에 따른 순위를 선형 회귀로 분석 하였으며 이 또한 상관관계가 없다. 세번째로, 모마 수득 상금에 따른 순위를 선형 회귀 분석 결과 상관관계가 높진 않지만 유의미 하다는 결론이 나왔으며 랜덤 포레스트 분석 결과 Fig. 3.의 결과 하단부, 0.22의 상관관계가 도출됐다[5].

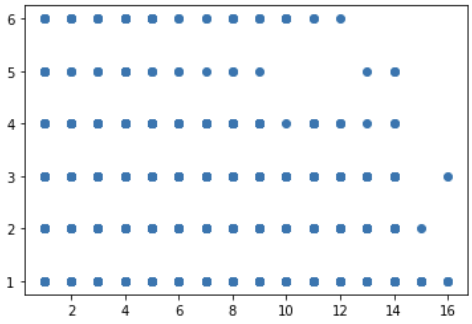

Fig. 2. Total amount of prize money by binning step 6

```
clf = RandomForestClassifier(n_estimators=100, max_depth=100,random_state=0)
clf.fit(X,Y)

predict4 = clf.predict(X_test)
print('RandomForest accuracy : ', accuracy_score(Y_test,predict4))
```

```
RandomForest accuracy :  0.2265193370165746
```

Fig. 3. Random Forest based on mother's horse prize money

IV. Conclusions

부모마 수득 상금과 순위와의 상관관계가 높진 않지만 모마 수득 상금에 따른 순위는 랜덤 포레스트 결과 0.22의 상관관계가 도출된 것으로 보아 부마에 비해 모마와 경주마의 성적이 더 관계가 있는 것으로 보인다. 순위를 예측하는 데 있어서 기수 성적, 경주 말의 성적뿐만 아니라 다양한 데이터를 활용해 분석 모델을 제공함으로써 경마 순위 예측에 도움을 줄 수 있으나 본 연구를 토대로 향후 다양한 분석 방법을 적용할 예정이다.

REFERENCES

[1] Won-Seob Lee, Jea-Moon Shin, Ji-Ho Lim, Dan-I Kim, Kyung-Ill. (2017). Subject oriented crawlling method

[2] Hyemin Choe, Nayoung Hwang, Chankyoung Hwang, Jongwoo Song. (2015). Analysis of Horse Races: Prediction of Winning Horses in Horse Races Using Statistical Models.

[3] Team KUBIG. (2018). Analysis and prediction of horse racing performance according to weather.

[4] Kevin H. Knuth. (2006). Optimal Data-Based Binning for Histograms

[5] Breiman, L. (2001). Random forests, Machine Learning, 45, 5-32.