LFD Project#3: Horse race prediction with SVM

Team 10: 김 현, 김 석, 이 은, 한창진

1. Objective and Dataset

본 프로젝트의 목표는 주어진 과거 경마 데이터로 미래의 경마 경주 결과를 예측하는 Support Vector Machine(SVM)을 만드는 것이다. 주어진 데이터셋은 2016.01.02~2020.02.08까지의 경주 결과, 말 정보, 기수 정보, 마주 정보, 조교사 정보가 포함되어 있다. Train set과 Test set 분리는 주어진 뼈대코드대로 진행하였다. (2020.01.19 기준으로 분리) 각 모델의 training에서 GridSearchCV를 활용하여 파라미터를 결정했으며, 최종적으로 테스트의 f1-Score가 높은 모델을 선정하였다.

2. Feature preprocessing

데이터 전처리는 다음과 같이 진행하였다. Categorical variable을 Dummy화하였다. 말의 home 정보에 대해 한/한(포)의 경우 국내, 그 외의 경우 국외로 분류하여 하나의 binary variable로 만들었다. 또한 말의 Gender의 경우 경기 결과에 유의미한 영향을 미친다는 선행연구[1]가 있어 이를 암, 수, 거(총 3개의 값)의 총 2개의 dummy variable로 만들었다.

Null 처리의 경우 선택한 feature들에서 대부분 owner의 owner money에 위치해있었다. 이는 단순히 0으로 처리하는 것보다 다른 owner과 비슷한 수준으로 간주하는 것이 맞다고 판단하여 mean()으로 처리하였다. 이후 RBF에는 [0,1] 사이에 input이 적당하다는 선행 연구에 기반하여, MinMaxScaler를 사용하였다.

또한 추가 feature로 이전 선행연구[1]들에서 총 1,2 횟수보단 이를 총 출전횟수로 나눈 feature들을 사용하여 진행하였기에 횟수 그대로 사용하는 것과 비율로 사용하는 것을 비교해본 결과 횟수 그대로 사용한 결과가 더 좋았다. 또한 해당 말의 이전 연승식 배당률(직전 5개 경기 평균)도 추가 feature로 고려해보았으나 이 역시 유의미한 결과가 아니어서 사용하지 않았다.

3. Feature selection

크게, 말, 기수 등 각각의 정보만을 사용하는 방법과 Extra Trees Classifier(ETC), Recursive Feature Elimination(RFE), Feature ensemble를 이용한 방법을 시도했으나 모두 feature 전체를 쓰는 것이 더 좋은 결과를 보여 사용하지 않았다.

3-1. ETC

linear kernel을 제외하고 다른 svm kernel에 대해선 ETC로 win을 예측하는 classifier을 만든 후 이 feature importance를 기준으로 top 5, 20, 전부 사용하는 방식을 test 해보았는데 전부 사용했을 때 train과 test 모두 더 좋은 결과를 보였다.

3-2. RFE

linear kernel 을 사용하여 svm 을 진행할 경우 각 feature 의 coeeficient 을 이용해 RFE 가 가능하다. 그렇기에 이 때 Sklearn 의 RFECV 를 이용하여 Feature 추출을 시도한 결과 오직, 기수 통산 승률 하나의 Feature 만이 선택되었다. 또한 이 결과가 좋지 않아 이 방법을 사용하지 않았다.

3-3. Feature ensemble

모든 feature 을 쓰고 sklearn 의 bagging classifier 을 사용하여 feature ensemble 을 진행했으나 결과가 좋지 않았다.

3-3. Final feature

race_result: 'lane', 'weight갿
jockey: 'age', 'race_count', 'first', 'second', '1yr_count', '1yr_first', '1yr_second'
owner: 'owner', 'owner_money'
trainer: 'age', 'race_count', 'first', 'second', '1yr_count', '1yr_first', '1yr_second'
horse: 'home', 'gender', 'race_count', 'first', 'second', '1yr_count', '1yr_first', '1yr_second', 'horse_money'

4. Model prediction

4-1. Inrace Test(경기별 상대평가)

경기별로 승자에 속할 확률이 높은 4 마를 class 1 로, 이외의 마를 class 0 으로 할당하는 상대평가를 구현하였다. 각 horse 가 3 등 안에 드는 확률을 예측할 때는 같은 race 에서의 다른 말들의 조합도 중요하다. 그렇기에 한 race 에서 각 horse 들의 win 확률을 상대평가하였다. 이 때 한 경기에서 3 마리~6 마리 까지 win 으로 예측해본 결과, 4 마리에 대해 win 으로 하였을 때 가장 좋은 성능을 보였다. 이는 3 등 안에 무조건 들 것 같은 3 마리의 말에만 배팅하는 것보다 차순위 말까지 고려하는 것이 합리적일 수 있다는 유추를 할 수 있었다.

4-2. Model ensemble

gridsearch 를 통해 parameter 를 test 했고 rbf, poly, sigmoid 에 대해서 최적의 parameter 를 구하고 이를 통해 ensemble 을 시도했으나 결과가 좋지 않아 사용하지 않았다.

4-3. Search parameter

Kernel: linear, poly, sigmoid, rbf / C: 1, 10, 100, 1000 / gamma: 0.001, 0.01, 0.1, 1 에 대해 grid search를 진행하였다.

5. Result

feature ensemble의 경우 test f1-score가 최대 0.5193, model ensemble은 0.517, ETC는 0.537, RFE는 0.45의 결과로 좋지 않았다. 최종 선택한 parameter과 결과는 다음과 같다.

{'C': 10, 'break_ties': False, 'cache_size': 600, 'class_weight': 'balanced', 'coef0': 0.0, 'decision_function_shape': 'ovr', 'degree': 3, 'gamma': 0.1, 'kernel': 'rbf', 'max_iter': -1, 'probability': False, 'random_state': 0, 'shrinking': True, 'tol': 0.001, 'verbose': False}

accuracy: 0.7261029411764706
recall: 0.6382978723404256
f1-score: 0.547112462006079

6. Reference

[1] Choe, H., Hwang, N., Hwang, C., & Song, J. (2015). Analysis of horse races: prediction of winning horses in horse races using statistical models. *Korean Journal of Applied Statistics*, *28*(6), 1133-1146.
[2] Lessmann, S., Sung, M. C., & Johnson, J. E. (2009). Identifying winners of competitive events: A SVM-based classification model for horserace prediction. European Journal of Operational Research, 196(2), 569-577.