왜 이 데이터 셋을 골랐는지

1. missing data가 많다 -> 이를 처리해주는 preprocessing에 대해 공부 가능

2. 생소한 수의학 분야

3. 칼럼의 개수가 많고,,,

Baseline

//어떻게 baseline을 정할지 간단한 설명?

<Preprocessing>

Horse Colic 데이터 셋은 missing data가 많은 데이터 셋이므로, NaN을 먼저 처리해 주어야합니다. NaN의 data type이 String인 경우에는 mode를, Numeric인 경우에는 median을 이용하여 NaN을 채웠습니다. 보다 정확한 학습 결과를 위해 reindex를 해준 이후, outcome을 제외한 String type의 칼럼들에 대해 one\_hot 인코딩을 해주었습니다. 종속변수인 outcome의 경우에는 live, died, euthanized에 대해 각각 1,0 -1의 값을 가지도록 Label 인코딩을 해주었습니다.

<Classification Models>

사용한 모델과 Hyper-parameter은 다음과 같습니다.

1) OneR (based on Decision tree with depth 1)

2) K-Nearest Neighbors

K:3, k:5

3) Linear SVM (Kernel: Linear)

C (Penalty parameter of the error term): 0.025

4) RBF SVM (Kernel: RBF)

C (Penalty parameter of the error term): 0.025

5) Decision Tree

Max-Depth: 10

6) Random Forest

Max-Depth: 10, number of estimators: 300

7) Neural Net

Alpha (L2 Regularization term): 1, hidden layer: (100,50)

8) AdaBoost (Ensemble based on the Decision Tree)

Number of estimators: 300

9) Naive Bayes

Default

해당 Hyper-parameter들은 각 모델에서 기본으로 사용되는 Hyper-parameter들 보다 더 향상된 성능을 위해 약간의 Tuning과정을 거쳤습니다. Baseline이기 때문에 여러 차례의 cross validation로 얻은 결과로 한 모델에 대해 하나의 Hyper-parameter 값을 두어 비교적 간단하게 조정해주어 Classification 모델을 학습하였습니다.

<Evaluation>

10-fold Cross Validation을 이용하여 각 모델의 성능을 평가하였고, 각 모델의 예측 값으로 Accuracy, Precision, Recall, F1 Score을 계산하여 모델을 평가했습니다. 그 중 각 모델을 평가하는 주요 척도로 Accuracy를 이용했습니다.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 Score |
| Random Forest | 0.726 | 0.689 | 0.709 | 0.789 |
| Decision Tree | 0.646 | 0.651 | 0.662 | 0.655 |
| AdaBoost | 0.632 | 0.620 | 0.632 | 0.625 |
| OneR | 0.625 | 0.511 | 0.625 | 0.550 |
| Nearest Neighbors  (K=) | 0.608 | 0.612 | 0.609 | 0.609 |
| RBF SVM | 0.596 | 0.354 | 0.595 | 0.444 |
| Nearest Neighbors  (K=) | 0.593 | 0.591 | 0.592 | 0.589 |
| Linear SVM | 0.589 | 0.439 | 0.589 | 0.452 |
| Neural Network | 0.484 | 0.443 | 0.371 | 0.391 |
| Naïve Bayes | 0.294 | 0.556 | 0.294 | 0.188 |

가장 높은 Accuracy를 가지는 모델은 Random Forest이고, 가장 낮은 Accuracy를 가지는 모델은 Naïve Bayes입니다.

Random Forest 모델은 모델들 가운데 가장 높은 Accuracy 값을 가지고 Precision, Recall, F1 Score 역시 다른 모델보다 높은 값을 가집니다.

Naïve Bayes 모델은 가장 낮은 Accuracy를 가지는 모델이지만, Accuracy, Recall, F1 Score에 비해 상대적으로 높은 Precision을 가집니다. Precision은 True라고 분류했을 때 실제로 True 값을 가지는 비율 (True Positive/True Positive + True False) 이므로, Precision만 유달리 높은 값을 가지는 이유는 실제 True 값을 False로 판단하는 False Negative가 크고, 실제 False 값을 True로 판단하는 False Positive가 작다고 유추할 수 있습니다. 즉, Naïve Bayes 모델은 실제 값이 True이든 False이든 False라고 판단할 확률이 높은 모델입니다.