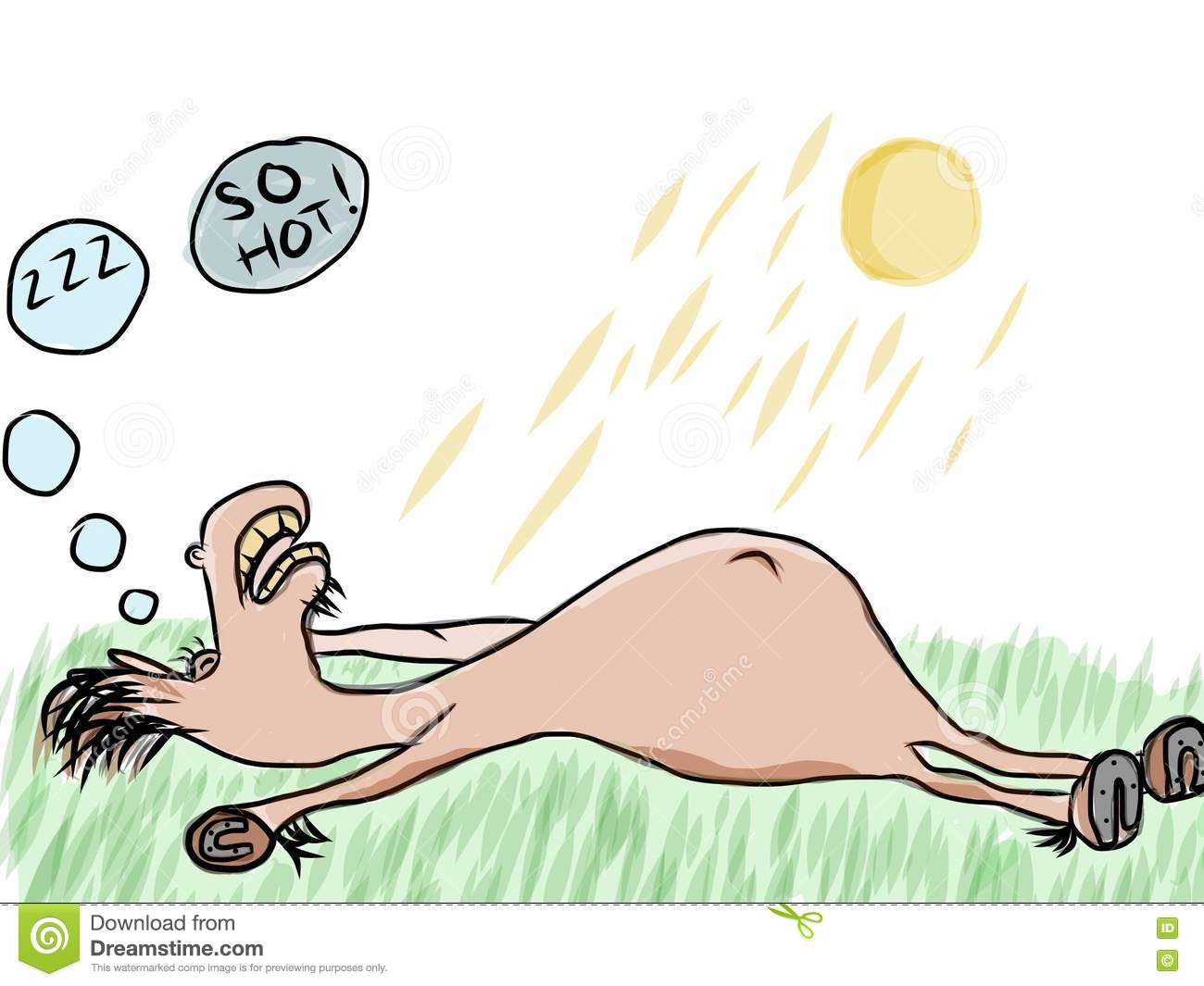
**BACKGROUND**

요즘같이 갑자기 더워진 날씨에 사람만큼 견디기 힘든 건 말도 마찬가지일 것입니다.

출처: https://thumbs.dreamstime.com/z/hot-summer-very-high-quality-hand-drawing-illustration-picture-you-can-see-some-funny-horse-lying-ground-day-you-73715118.jpg

이런 극단적인 날씨는 말의 건강에 큰 영향을 줄 수 있습니다. 특히 더운 여름철에는 말의 산통을 조심해야 합니다. 말을 죽음에 이르게 하는 가장 흔한 원인 중 하나이므로 요즘 같은 날에는 말에게 조금 더 관심이 필요할 것 입니다.

산통이란 수의학적인 용어로, 쉽게 말해서 배앓이를 말하는 것 입니다. 말은 큰 체구에 비해 위의 용적이 10리터 정도 밖에 되지 않을 정도로 위가 작고 소와 달리 되새김을 할 수 없어 독소가 쉽게 쌓이게 됩니다. 이렇게 특이한 생리구조를 가졌기에 산통이 발생하는 원인이 다양하고 흔하게 발병합니다. 하지만 어떠한 원인에 의해서 발생하더라도 겉으로 드러나는 증상이 비슷하여 적절한 조치가 어렵다고 합니다. 특히 증상이 심한 경우 수술을 하더라도 사망하거나 해결을 하지 못하여 안락사 시키기도 합니다. 산통이 발생했을 경우 의학적 상태만으로 원인을 예측할 수 있다면 좀더 빠르고 정확한 의학적 조치가 조기에 취할 수 있어 말의 고통과 생존율을 낮추는데 큰 기여를 할 수 있을 것 입니다.

**DATA DESCRIPTION**

- 데이터 출처: UC Irvine Machine Learning Repository (<https://archive.ics.uci.edu/ml/datasets/Horse+Colic>)

- 데이터 설명: 말의 과거 수술이력, 검진 결과로 이루어진 데이터.

- 데이터 개수: 299 row X 34 variable

**PROJECT GOAL**

- 산통 발생에 대한 원인 탐구

- 말의 생존을 결정하는 결정변수 탐구.

- Missing value의 적절한 처리로 예측 능력 극대화.

- 효과적인 범주형 자료 탐구

**변수 이름과 설명**

1. Surgery: 수술 여부

2. Age: 나이(6개월 이상부터 성인 말)

3. Hospital Number: 말에게 부여되는 고유 번호

4. Rectal temperature: 직장의 온도(°C) / 정상: 37.8°C

5. Pulse: 심장박동, 심장상태를 대변해줌. 극심한 통증 또는 순환 쇼크를 겪으면 올라간다. / 정상: 30-40 (경주마 20-25)

6. Respiratory rate: 호흡률, 변동이 심함. / 정상: 8-10

7. Temperature of extremities: peripheral circulation의 지표. 손발의 온도 온도와 상관관계 존재, 낮으면 쇼크 가능성.

8. Peripheral pulse: 맥박, 수치가 낮을 경우 관류가 제대로 이루어지지 않음을 의미.

9. Mucous membranes: 점막, 색에 따라 증상이 다름. Circulation 상태를 추측할 수 있는 지표.

10. Capillary refill time: 모세관 재충전 시간. refill시간이 길어질수록 circulation이 제대로 이루어지지 않음.

11. Pain: 통증, 심할수록 수술이 필요, 치료를 받은 이력이 있으면 pain지수가 낮게 평가 될 가능성이 있음.

12. Peristalsis: 장기 운동성, 팽창하거나 유독할수록 활동성이 떨어짐

13. Abdominal distension: 복부팽만, 심할수록 내장의 활동성이 떨어지고 고통스러움.

14. Nasogastric tube: 비위관에서 나오는 가스 양. 말을 불편하게 할 수 있음.

15. Nasogastric reflux: 비위관성 역류, 수치가 클수록 다른 창자로부터의 혈류의 방해가 있음.

16. nasogastric reflux PH: 비위관성 역류 PH / 정상: 3-4

17. rectal examination – feces: 대변, 배설물이 없다는 것은 장 폐색을 시사함.

18. Abdomen: 복부, 굳은 대변은 폐색을 의미, 장이 팽창된 것은 수술로 인한 손상을 나타냄.

19. Packed cell volume: 혈액속의 적혈구 수, 수치가 높을수록 탈수 또는 혈류 순환의 문제가 있음. / 정상: 30-50,

20. Total protein: 총 단백질량, 수치가 높을수록 탈수 가능성이 높아짐 / 정상: 6-7.6,

21. Abdominocentesis appearance: 복강경 외관, 흐리거나 혈청이 섞인 것은 장기가 제대로 작동하지 않음을 시사.

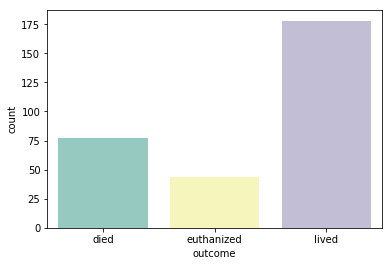
22. Abdominocentesis total protein: 복강내 총 단백질, 수치가 높을수록 장이 손상되었을 가능성이 큼.

23. Outcome: 생존 여부 (lived, died, was euthanized)

**산통이 발생했을 때 말의 생존여부는 어떻게 될까?**

먼저 산통이 발생 했을 때 말의 생존 여부를 확인해보면 약 60%정도의 말이 생존 하지만, 죽거나 안락사 비율을 합쳐, 사망한 말의 비율은 40%정도로 굉장히 높은 비율입니다. 산통이 말에게 매우 치명적인 상황인 것을 다시 한번 확인 할 수 있습니다. 안락사는 치료과정에서 수의사의 판단에 따라 시행됩니다.

|  |  |  |  |
| --- | --- | --- | --- |
| outcome | Lived | Died | Was euthanized |
| Count | 178 | 77 | 44 |
| percentage | 59.3% | 25.7% | 14.7% |



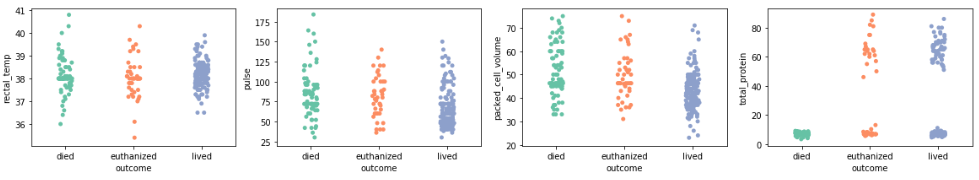
<figure 1> outcome countplot

**죽은 말과 생존말의 차이점은?**

산통이 발생하였지만 어떤 말은 생존이 가능하였고 어떤 말은 죽음을 피할 수 없게 되었습니다. 산통이 발생하는 원인은 매우 다양하다고 합니다. 말 들에게 어떤 차이가 있는지 알아보고 원인에 대하여 탐구 해보겠습니다. 다음 기초 통계량과 그래프를 통해 가설을 설정할 있었습니다.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variables | | Value | | Missing |
| Mean | SD |
| **Rectal temperature** | lived | 38.16 | 0.57 | 26 |
| died | 38.12 | 0.93 | 24 |
| euthanized | 38.04 | 1.02 | 10 |
| **Pulse** | lived | 64.00 | 24.54 | 12 |
| died | 85.05 | 31.73 | 11 |
| euthanized | 81.52 | 26.00 | 1 |
| **Respiratory rate** | lived | 29.48 | 18.29 | 31 |
| died | 32.40 | 16.42 | 19 |
| euthanized | 31.02 | 16.97 | 8 |
| **Nasogastric reflux PH** | lived | 4.68 | 2.00 | 154 |
| died | 4.90 | 1.64 | 57 |
| euthanized | 4.47 | 2.20 | 35 |
| **packed cell volume** | lived | 42.99 | 7.97 | 13 |
| died | 51.89 | 11.40 | 8 |
| euthanized | 49.96 | 11.36 | 8 |
| **Total protein** | lived | 25.77 | 28.45 | 13 |
| died | 6.83 | 1.18 | 12 |
| euthanized | 35.62 | 30.71 | 8 |
| **abdominocentesis total protein** | lived | 2.25 | 1.92 | 126 |
| died | 2.67 | 2.14 | 53 |
| euthanized | 2.21 | 1.45 | 19 |

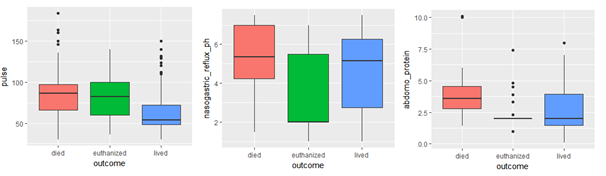
<Table 2> 연속형 변수의 기초통계량



**<figure 2> packed cell volume, total protein scatter plot**

Packed cell volume은 수치가 증가할수록 혈류 순환이 잘 안되거나 탈수증상이 나타나는 것을 의미합니다. 일정 수치를 넘어갔을 때 다수의 말이 생존하지 못하였습니다. Total protein또한 수치가 증가할수록 탈수가 더 심해짐을 의미하는 변수입니다. 하지만 죽은 말의 total protein은 6.82±1.18로, 정상범위를 거의 넘어가지 않았음을 확인할 수 있었습니다. 하지만 생존한 말과 안락사 당한 말들 중 total protein 수치가 월등히 높은 경우도 있습니다. **즉, 탈수가 말의 생존의 직접적인 원인이 아니고 혈류 순환 문제가 장기 문제를 야기하였고 그로 인해 탈수증상과 사망률이 높아졌다고 가설을 세웠습니다**. 그래프를 통해 total protein과 nasogastric reflux pH 사이에 음의 상관관계가 있을 것이라 유추하였고 실제로 Pearson correlation test를 실행한 결과 coefficient -0.72라는 강한 상관관계가 나타났습니다.

**가설: 산통이 발생했을 때 탈수증상이 있었던 말들의 사망 원인은 obstruction 으로 인한 circulatory problem이다.**

****

**<figure 3> pulse, nasogastric reflux ph, abdominocentesis protein scatter plot**

혈류 순환과 관련된 변수들을 추가로 탐색해 보았습니다. Box plot으로 볼 때 죽은 말이 평균적으로 pulse가 높은 편이었습니다. Pulse는 순환 쇼크가 발생하면 높아집니다. Nastogastric ph또한 정상범위를 벗어난 수치를 보였습니다. 어느 정도 산성은 상태가 균을 죽이고 정상적인 활동성을 의미하지만, 너무 높은 수치는 해로운 균을 죽이기 힘든 상태이며 낮은 수치는 장기 손상이 야기하여 말이 고통 받습니다. 이를 보아 obstruction으로 인해 순환계통에 문제가 생겨 산의 분비가 정상적이지 않음을 유추할 수 있었습니다. 이에 정말로 위장이 손상된 것인지 abdominocentesis total protein을 통해 확인해본 결과 수치가 높음을 확인할 수 있었습니다. 따라서 회귀 분석을 통해 말의 죽음과 위 세 개 변수가 유의하게 관련이 있는지 알아보았습니다. 분석에 앞서 missing value에 대한 이슈가 있습니다. 본 데이터는 결측치가 무작위하게 나타났고, 변수를 삭제할 경우 분석에 어려움이 생길 것으로 보여 단순대체법으로 결측치를 대체하여 모델을 만들어 보았습니다. 이후 변수 특성에 따라 correlation의 coefficient를 이용하는 방법, 단순대체법 모두 시도하면 좋을 것 입니다.

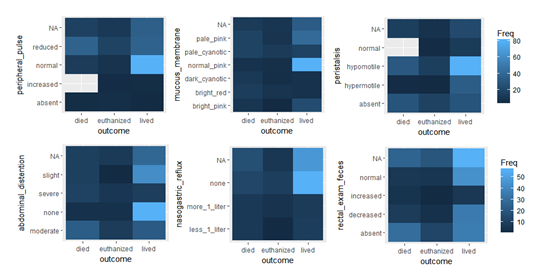
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variables | Pulse | | | Nasogastric reflux PH | | | Abdominocentesis total protein | | |
| lived | died | euthanized | lived | died | euthanized | lived | died | euthanized |
| Missing | 8.0% | | | 82.3% | | | 66.2% | | |

**<table2> Missing value percentage of three variable**

세 변수의 회귀 식은 아래와 같고 예측율은 74.9%였습니다. 충분히 유의미하다고 볼수 있습니다.

**Logit = 6.34 – 0.02 \* pulse – 0.56 \* abdominocenthesis\_total\_protein - 0.41 \* nasogastric\_reflux\_ph**

다음은 말들이 장 폐색으로 circulation이 안되고 장 활동성이 떨어지고 장기가 손상되어 사망했다는 가설을 설명해 줄 범주형 변수들입니다.



|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Most Frequent value by outcome** | | |
|  | **lived** | **euthanized** | **died** |
| **Peripheral pulse** | Normal | Reduced | Reduced |
| **Mucous membrane** | Normal pink | Pale cyanotic | Pale cyanotic,  Pale pink |
| **Peristalsis** | Hypomotile | Absent | Hypomotile |
| **Abdominal distension** | None | Moderate | Moderate |
| **Nasogastric reflux** | None | None | None |
| **Rectal exam feces** | Normal | Absent | Absent |

형광으로 표시한 부분은 장 활동성이 적고 순환이 잘 안됨을 잘 설명해준 값입니다. 범주형 변수의 통계량은 데이터의 카테고리 비율, R-squared, F(p-value)를 구하여 보았습니다. 범주형 변수의 통계량 표는 아래와 같습니다.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variables | Value |  | F(P-value) | Missing |
| Surgery(2) | 1=Yes(60%), 2=No(39.6%) | 0.01383 | 0.003633 | 1 |
| Age(2) | 1=adult(92%), 2=young(8%) | 2e-05 | 0.820653 | 1 |
| temperature of extremities(4) | 1=Normal(26%), 2=Warm(10%), 3=Cool(36.3%), 4=Cold(9%) | 0.11672 | 0.004919 | 56 |
| peripheral pulse(4) | 1=normal(38.3%), 2=increased(1.7%), 3=reduced(34.3%), 4=absent(2.7%) | 0.09993 | 0.073029 | 69 |
| mucous membranes(6) | 1=normal pink(16.3%), 2=bright pink(10%), 3=pale pink(19.3%), 4=pale cyanotic(13.7%), 5=bright red / injected(8.3%), 6=dark cyanotic(6.7%) | 0.1223 | 0.321139 | 47 |
| capillary refill time(2) | 1= <3 seconds(62.7%), 2= 3 seconds(26%), 3?(이상치) | 0.10031 | 0.038865 | 32 |
| Pain(5) | 1=alert, no pain(12.7%), 2=depressed(19.7%), 3=intermittent mild pain(22.3%), 4= intermittent severe pain(13%), 5=continuous severe pain(14%) | 0.14579 | 0.053172 | 55 |
| Peristalsis(4) | 1=hypermotile(13%), 2=normal(5.3%), 3=hypomotile(42.7%), 4=absent(24.3%) | 0.07706 | 0.835977 | 44 |
| abdominal distension(4) | 1=none(25.3%), 2=slight(21.7%), 3=moderate(21.7%), 4=severe(12.7%) | 0.08303 | 0.171248 | 56 |
| nasogastric tube(3) | , 1=none(23.7%), 2=slight(34%), 3=significant(7.7%) | 0.00157 | 0.749853 | 104 |
| nasogastric reflux(3) | 1=none(40%), 2= >1 liter(11.7%), 3= <1 liter(13%) | 0.03661 | 0.493985 | 106 |
| rectal examination – feces(4) | 1=normal(19%), 2=increased(4.3%), 3=decreased(16.3%), 4=absent(26.3%) | 0.04509 | 0.472205 | 102 |
| Abdomen(5) | 1=normal(9.3%), 2=other(6.3%), 3=firm feces in the large intestine(4.3%), 4=distended small intestine(14.3%), 5=distended large intestine(26.3%) | 0.03441 | 0.822744 | 118 |
| abdominocentesis appearance(3) | 1=clear(13.7%), 2=cloudy(16%), 3=serosanguinous(15.3%) | 0.05046 | 0.611320 | 165 |
| surgical lesion(2) | 1=Yes(18.7%), 2=No(81.3%) | 0.04374 | 0.810743 | 1 |
| type of lesion |  |  |  |  |
| cp data(2) | 1=yes(66%), 1=no(33%) | 0.00123 | 0.297407 | 1 |

**<Table 3> Statistics of Categorical Variables**

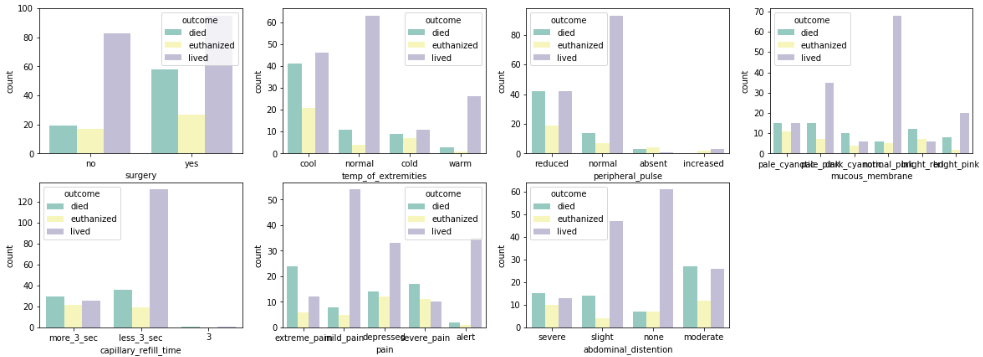
**말의 생존 여부를 예측하기 위해 사용한 변수(7개)**

1. F(p-value)값이 낮은 변수인 **surgery, temperature of extremities, peripheral pulse, capillary refill time**

2. R-squared값이 가장 높은 **pain, mucous membranes**

3. F(p-value)값은 높고 R-squared값이 낮음에도, 변수 설명에서 중요하다고 강조된 **abdominal distention**

통계량적 차이를 시각적으로 확인하기 위하여 python seaborn 패키지의 countplot 함수를 이용하여 그래프를 그려보았습니다.

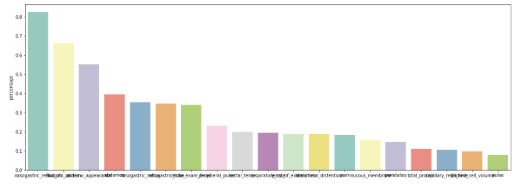


**<figure 4>**

변수의 상태 별로 생존 여부 경향이 다르게 나타나는 걸로 보아 말의 생존 여부를 예측하는데 좋은 변수가 될 것 같습니다. 특히 pain, mucous membrane, abdominal distention변수들은 그래프 변동이 심한 것으로 보아 예측에 영향을 많이 끼칠 것 같다는 생각이 듭니다.

**Missing value**

앞서 언급했듯 UCI의 ‘Horse Colic Data Set’의 Missing value 비율은 전체 데이터의 30%정도이며, 50 ~ 80%의 missing value를 가진 변수들도 존재합니다. 이로 보아 missing value를 잘 대체할 수 있는 통계량을 설정하는 것이 좋은 분석을 하기 위한 포인트가 될 것입니다. 우선, 앞에서와는 다르게 가장 단순한 방법을 사용해 보았다. Missing value 비율이 50%이상인 변수는 제외하고 연속형 변수는 평균값을, 범주형 변수는 최빈값으로 설정하고 분석을 계속해보기로 하였습니다. 추후 여러 방법을 적용하여 본 데이터를 가장 잘 설명할 수 있는 방법을 심층적으로 탐구해야 합니다.



**<figure 6> missing value percentage plot**

**모델의 평가 기준**

통계량과 그래프를 보아 27개의 변수 중 11개의 변수를 선택하였다. 선택된 변수로 예측 모델을 만들었을 때 어떤 모델이 좋은 모델인지 알 수 있을까? precision과 recall중 생명과 관계된 만큼 사망한 경우에 더욱 민감하여야 하기에 recall을 기준으로 하였습니다.

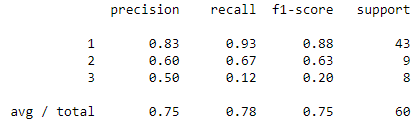
**말의 생존여부 예측 모형**

먼저 성능을 비교해보기 위해 python Sklearn패키지의 train\_test\_split함수로 train data set과 test date set을 만든다음 train data set으로 학습시키고 test data set으로 지표를 비교해보기로 하였습니다. 또, Outcome의 결과가 3개인 점을 고려하여 로지스틱 회귀모형 보다 OVS(one vs rest)로지스틱 회귀모형을 만들어 보았습니다.

**OVR Logistic Regression Model**

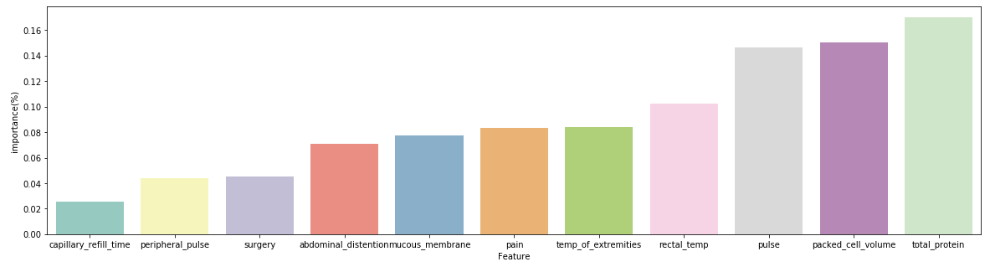
OVR회귀모형은 python sklearn 패키지의 OneVsRestClassifier과 LogisticRegression함수를 이용하여 만들었습니다.

변수를 선택하는 과정에서 F(p-value)값을 고려해줬으므로 11개의 변수 그대로 사용하기로 하였고, test dataset에 대한 결과로 아래의 표를 얻었습니다. Precision 75% Recall 78%지만 3(사망)의 경우 recall이 12%로 좋은 모델이 아닙니다.

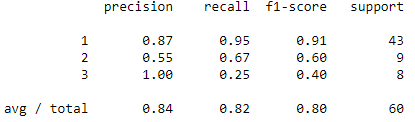


**Random Forest Model**

변수 중요도를 다른 관점으로도 확인해보기 위하여 기계 학습의 random forest 모형을 만들어 보았습니다. Python sklearn패키지의 RandomForestClassifier 함수를 이용하여 최대 깊이 30인 100개의 의사결정나무모형이 만들어 졌고 변수들의 중요도를 표현한 그래프는 아래와 같습니다.

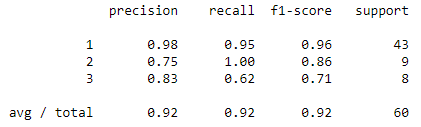


변수를 선택할 때 예상 했던 것처럼 total\_protein, packed\_cell\_volume, pulse, rectal\_temp변수가 생존에 미치는 영향이 크다는 것을 다시 한번 확인 할 수 있었습니다. test data set에 대한 결과로는 precision 84%, Recall 82%가 나왔습니다. 특히 사망한 말의 Recall이 더욱 향상 되었습니다. 회귀모형보다 좋은 성능 이지만 다른 방법도 적용해보겠습니다.



**MLP**

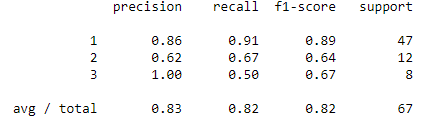
마지막으로 Neural Network를 통해 Softamx MLP 모델을 만들어 보겠습니다. Python tensorflow 패키지를 이용하여 4-layer, L-th unit : 16, activation function: LeRU인 간단한 모델을 만들어 보았습니다.



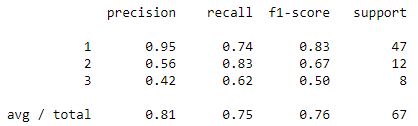
만족스러운 결과가 나왔습니다. parameter tuning과 optimization algorithm을 다르게 적용해보고 Missing value처리와 변수 선택을 좀 더 자세하게 했다면 더 신뢰성 있는 결과가 나왔을 듯 하나, 이후 더욱 정확한 탐구를 한 뒤 모델링 하는 것이 좋아 보입니다.

**Test data prediction**

이제 UCI의 ‘Horse Colic Data Set’에 포함되어 있는 test data를 모델에 적용시켜 보겠습니다. 먼저 random forest model입니다. 앞서 테스트 해본 결과보다 좋은 결과가 나왔습니다. 특히 사망한 경우는 dev set보다 좋습니다.



다음으로 가장 성능이 좋았던 MLP model에 적용시켜보았습니다. 앞서 테스트 해본 값과 차이가 심한걸로 보아 일반화가 필요해 보입니다. 하지만 중요하게 생각하는 사망한 경우의 recall은 random forest model보다 좋습니다.



혈류 순환과 장기 관련 변수로 모델을 만들어본 결과 어느 정도 신뢰도 있는 결과가 나온 걸로 보아 분석 스케치의 방향이 옳아 보입니다. 최종적으로 변수들의 상호관계를 적용한 model들의 앙상블을 선택하면 가장 좋은 성능을 발휘 할 수 있지 않을까 생각됩니다.