**Horse colic 이란 무엇인가?**

산통이란 수의학적인 용어이고 쉽게 말해서 배앓이를 말하는 것이다. 말은 큰 체구에 비해 위의 용적이 10리터 정도밖에 되지 않을 정도로 위가 적으며 말은 소와 달리 되새김을 할 수가 없어서 토해낼 수가 없는 특징을 갖고 있다 이러한 특이한 생리구조를 가졌기에 산통이 자주 발생하는데 대표적인 산통으로는 위식체, 소장감돈, 소장교액, 회맹중첩, 대결장변위 또는 염전, 대결장폐색 등 다양하고, 어떠한 원인에 의해서 발생하더라도 겉으로 드러나는 증상은 대개 비슷하다고 한다. 특히 산통이 심한 경우 수술을 하기도 하고 사망하거나 안락사를 시키기도 한다. 그렇다면 산통이 발생했을 경우 의학적 상태에 따라 말의 생존 여부를 예측 할 수 있을까?

분석에는 UCI에서 제공하는 ‘Horse Colic Data Set’을 이용하였다.

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

**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, 정상: 30-40(경주마의 경우20-25), 극심한 통증 또는 쇼크를 겪으면 올라간다.
6. Respiratory rate : 호흡률, 정상: 8-10, 변동이 심하므로 신뢰성이 없는 변수
7. Temperature of extremities : 손발의 온도 온도가 높으면 직장의 온도가 높다고 볼 수 있고, 낮으면 쇼크의 가능성을 보여줌
8. Peripheral pulse : 맥박, 수치가 낮을경우 관류가 제대로 이루어지지 않음을 의미함.
9. Mucous membranes : 점막, 색에 따라 증상이 다름
10. Capillary refill time : 모세관 재충전 시간
11. Pain : 통증, 심할수록 수술이 필요, 치료를 받은 이력이 있으면 고통지수가 낮게 평가 될 가능성이 있음.
12. Perisralsis : 창자의 운동 상태, 상태가 좋지 않을수록 활동성이 떨어짐
13. Abdominal distension : 복부팽만, **중요한 변수,** 심할수록 내장의 활동성이 떨어지고 고통스러움
14. Nasogastric tube : 비위 관에서 나오는 가스, 말을 불편하게 할 수 있음
15. Nasogastric reflux: 비위관성 역류, 수치가 클수록 또 다른 장폐색을 시사함.
16. nasogastric reflux PH: 비위관성 역류 PH, 정상: 3-4
17. rectal examination – feces: 대변, 배설물이 없다는 것은 장폐색을 시사함.
18. Abdomen : 복부, 굳은 대변은 폐색을 의미하고, 장이 팽창된 것은 수술로 인한 손상을 나타낸다.
19. packed cell volume: 혈액속의 적혈구 수, 정상:30-50, 수치가 높을수록 탈수 가능성이 높아짐
20. total protein: 총 단백질량, 정상:6-7.6, 수치가 높을수록 탈수 가능성이 높아짐
21. abdominocentesis appearance: 복강경 외관, 복강으로부터 추출하며 흐리거나 혈청이 섞인 것은 장기가 제대로 작동하지 않음을 시사함
22. abdomcentesis total protein : 복강내 총 단백질, 수치가 높을수록 탈수 가능성이 높아짐
23. Outcome: 결과(lived, died, was euthanized)
24. surgical lesion: 수술 병변, 수술로 인한 손상이었는지를 나타냄

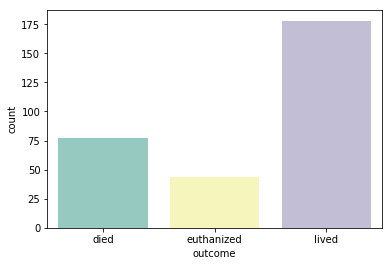
25, 26, 27. type of lesion: 수술 위치, 유형, 하위유형, 특정코드

28. cp data: 병리 자료 여부

**산통이 발생했을 때 말의 생존여부는 어떻게 될까?**

먼저 산통이 발생 했을 때 말의 생존 여부를 확인해보면 약 60%정도의 말이 생존 하지만, 죽거나 안락사 한 비율을 합쳐, 사망한 말의 비율은 40%정도로 굉장히 높은 비율이다. 이것으로 보아. 산통은 말에게 매우 치명적인 상황인 것을 다시 한번 확인 할 수 있다.

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



**<figure 1> outcome countplot**

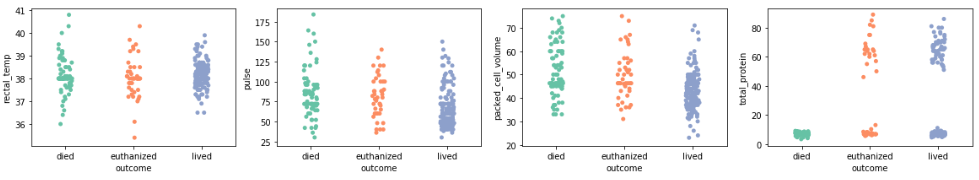
**Data 탐색**

이제 데이터를 본격적으로 탐색하기 위하여 기초 통계량을 조사해 보았다. 연속형 변수의 기초 통계량 표는 아래와 같다.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variables | | Value | | | | | | Missing |
| Mean | Median | Std Dev | Skewness | Kurtosis | range |
| **Rectal\_temperature** | lived | 38.162921 | 38.2 | 0.569344 | 0.036021 | 0.845986 | 36.5-39.9 | 26 |
| died | 38.12337 | 38.1 | 0.925776 | 0.338782 | 0.813765 | 36.0-40.8 | 24 |
| euthanized | 38.04090 | 38.05 | 1.016302 | -0.13436 | 0.631468 | 35.4-40.3 | 10 |
| **Pulse** | lived | 64.00000 | 54.0 | 24.544886 | 1.391464 | 1.315963 | 30.0-150.0 | 12 |
| died | 85.05194 | 86.0 | 31.730982 | 0.796242 | 0.793754 | 30.0-184.0 | 11 |
| euthanized | 81.52272 | 82.0 | 26.001001 | 0.174706 | -0.565632 | 36.0-140.0 | 1 |
| **respiratory rate** | lived | 29.48470 | 24.0 | 18.290945 | 1.683889 | 2.818225 | 8.0-96.0 | 31 |
| died | 32.39936 | 30.0 | 16.424164 | 1.960948 | 5.05072 | 12.0-96.0 | 19 |
| euthanized | 31.01556 | 24.0 | 16.965769 | 1.354574 | 1.323845 | 12.0-80.0 | 8 |
| **nasogastric reflux PH** | lived | 4.684619 | 5.15 | 1.995171 | -0.297241 | -1.343638 | 1.0-7.5 | 154 |
| died | 4.895197 | 5.35 | 1.638067 | -0.584413 | -0.231025 | 1.5-7.5 | 57 |
| euthanized | 4.471913 | 2.0 | 2.200063 | 0.481735 | -1.625962 | 1.0-7.0 | 35 |
| **packed cell volume** | lived | 42.98874 | 42.0 | 7.974895 | 0.724847 | 1.446866 | 23.0-71.0 | 13 |
| died | 51.88908 | 53.0 | 11.402502 | 0.113267 | -0.930693 | 33.0-75.0 | 8 |
| euthanized | 49.96498 | 50.5 | 11.361869 | 0.29834 | -0.643649 | 31.0-75.0 | 8 |
| **total protein** | lived | 25.76657 | 7.5 | 28.452084 | 0.736292 | -1.352775 | 4.5-86.0 | 13 |
| died | 6.829870 | 6.8 | 1.177661 | -0.612426 | 0.539978 | 3.3-9.0 | 12 |
| euthanized | 35.622727 | 56.0 | 30.705228 | -0.116664 | -1.760177 | 5.5-89.0 | 8 |
| **abdomcentesis total protein** | lived | 2.247753 | 2.0 | 1.924404 | 1.046526 | 0.266961 | 0.1-8.0 | 126 |
| died | 2.671429 | 3.6 | 2.13826 | 1.815788 | 3.446602 | 1.4-10.1 | 53 |
| euthanized | 2.209091 | 2.0 | 1.447043 | 2.128424 | 5.384482 | 1.0-7.4 | 19 |

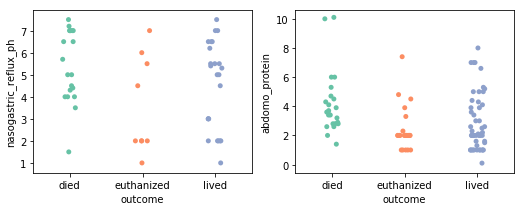
**<Table 1> 연속형 변수의 기초통계량**

가장 먼저 눈에 띈 변수는 total protein이다. 죽은 경우를 보았을 때 평균이 6.8이고 편차는 매우 적었다. 그 외에는 평균은 25.7과 35.6이지만 편차가 매우 높은 것을 확인 할 수 있다. Rectal temperature는 살았을 때 온도 편차가 적고 다른 상황에서는 편차가 큰 것을 볼 수 있다. 정상 온도의 범위를 넘어섰을 때는 몸의 상태가 좋지 못한 경우라고 생각하니 생존여부를 예측하는데 좋은 변수가 될 듯 하다. Pulse와 packed cell volume는 평균과 분산을 비교해 봤을 때 생존 여부에 따라 차이가 나는 것을 확인 하였다. 변수들 사이에 어떤 차이가 있는지 시각적으로 확인하기 위하여 python seaborn 패키지의 stripplot 함수를 이용하여 그래프를 그려보았다.



**<figure 2>**

그래프를 보았을 때 total protein의 분포 차이가 확연히 구분되며 Rectal temperature와 Pulse는 분포는 비슷하나 중앙 집중도의 차이가 보인다. 변수의 분포를 좀 더 자세히 분석하면 좋겟지만, 통계적으로 깊은 이해도를 요하는 문제로 보이므로 다음 기회에 자세히 분석하기로 하자.

그리고 기초통계량에서는 차이가 보이지만 missing value 비율이 상당히 높아 신뢰성이 떨어지는 변수로는 nasogastric reflux PH, abdomcentesis total 2가지 변수가 보인다. 마찬가지로 그래프를 그려 시각화 하였지만 변수의 개수가 너무 적어 좋은 변수라고 보기는 어려울 것 같다.

**D <figure 3>**

범주형 변수의 통계량은 데이터의 카테고리 비율, R-squared, F(p-value)를 구하여 보았다. 범주형 변수의 통계량 표는 아래와 같다.

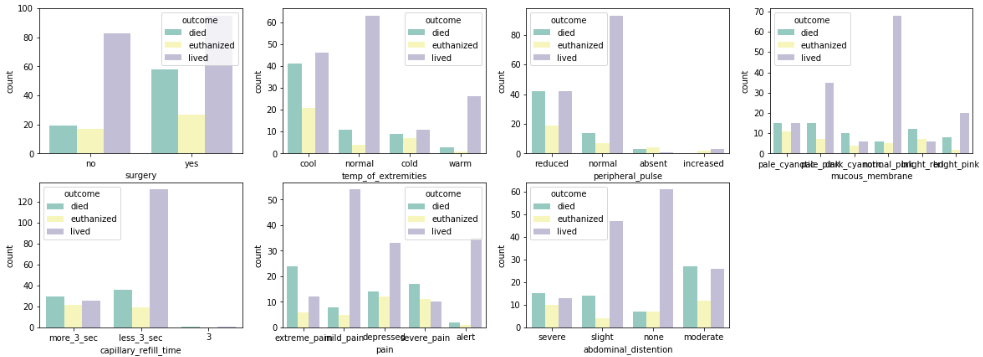
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variables | Value |  | F(P-value) | Missing |
| Surgery(2) | 1=Yes(60%), 2=No(39.6%) | 0.01383 | 0.003633 | 1 |
| Age(2) | 1=adult(92%), 2=young(8%) | 2e-05 | 0.820653 | 1 |
| temperature of extremities(4) | 1=Normal(26%), 2=Warm(10%), 3=Cool(36.3%), 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 2> 범주형 변수의 통계량**

먼저 .F(p-value)값이 낮은 변수로는 Surgery, temperature of extremities, peripheral pulse, capillary refill time.

가 있고 대부분 낮지만 R-squared값이 가장 높은 pain, mucous membranes 변수가 있으며 F(p-value)값은 높고 R-squared값은 낮지만 변수 설명에서 중요하다고 강조된 abdominal\_distention변수까지 총 7개의 변수가 말의 생존 여부를 예측 하는데 좋은 변수가 될 수 있을 것 이라고 생각 된다.

통계량으로는 어떤 차이가 있는지 눈에 보이지 않아 확인하기 위하여 python seaborn 패키지의 countplot 함수를 이용하여 그래프를 그려보았다.

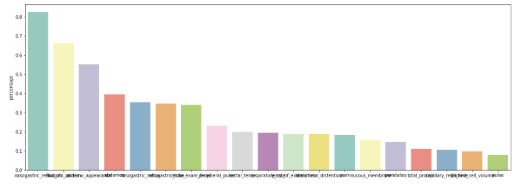


**<figure 4>**

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

**Missing value**

앞서 기초 통계량 표에서 확인 할 수 있듯이 UCI의 ‘Horse Colic Data Set’의 Missing value 비율은 전체 데이터의 30%정도이며, 50 ~ 80%의 missing value를 가진 변수들도 존재한다. 이로 보아 missing value를 잘 대체할 수 있는 통계량을 설정하는 것이 좋은 분석을 하기 위한 포인트가 될 것이다. 하지만 이는 통계적으로 어려운 지식이 필요하니 우리는 간단하게 Missing value 비율이 50%이상인 변수는 제외하고 연속형 변수는 평균값을, 범주형 변수는 최빈값으로 설정하고 분석을 계속해보기로 한다



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

**모델의 평가 기준**

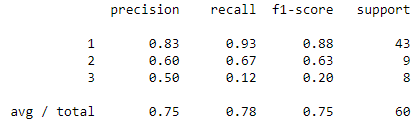
통계량과 그래프를 보아 27개의 변수 중 11개의 변수를 선택하였다. 선택된 변수로 예측 모델을 만들었을 때 어떤 모델이 좋은 모델인지 알 수 있을까? 강의에서 본 precision과 recall중 생명과 관계된 만큼 실제로 사망한 말인데 생존한 말로 예측을 하는 경우에 더 큰 페널티가 존재 해야 한다고 생각되어 Precision을 기준으로 하였다.

**말의 생존여부 예측 모형**

먼저 성능을 비교해보기 위해 python Sklearn패키지의 train\_test\_split함수로 train data set과 test date set을 만든 다음 test data set으로 Precision을 비교해보기로 하였다. 또, Outcome의 결과가 3개인 점을 고려하여 로지스틱 회귀모형 대신 OVS(one vs rest)로지스틱 회귀모형을 만들어 보았다.

**OVR Logistic Regression Model**

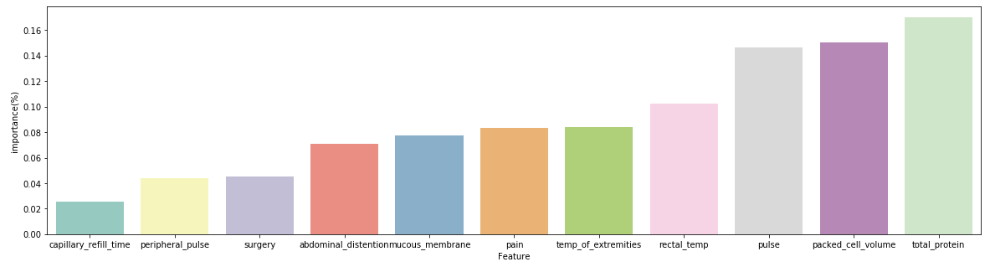
OVS 회귀 모형은 python sklearn 패키지의 OneVsRestClassifier과 LogisticRegresiion함수를 이용하여 만들었다. 변수를 선택하는 과정에서 F(p-value)값을 고려해줬으므로 11개의 변수 그대로 사용하기로 하였고, test dataset에 대한 결과로 아래의 표를 얻었다.



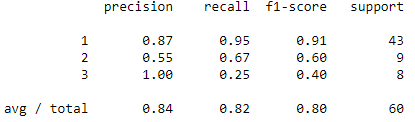
precision 75%가 나왔지만 만족할 만한 점수는 아니다.

**Random Forest Model**

좀더 복잡한 모형을 만들기 위해 기계 학습의 random forest 모형을 만들어 보았다. Python sklearn패키지의 RandomForestClassifier 함수를 이용하여 최대 깊이 30인 100개의 의사결정나무모형이 만들어 졌고 변수들의 중요도를 표현한 그래프는 아래와 같다.



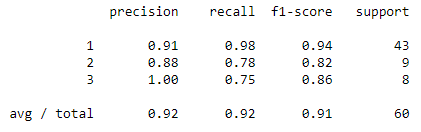
변수를 선택할 때 예상 했던 것처럼 total\_protein, packed\_cell\_volume, pulse, rectal\_temp변수가 생존에 미치는 영향이 크다는 것을 확인 할 수 있다. test dataset에 대한 결과로 아래의 표를 얻었다.



precision 84%로 로지스틱 회귀모형보다 9%나 높게 나왔지만 조금 더 복잡한 모델을 활용하여 precision값을 높혀 보도록 하자.

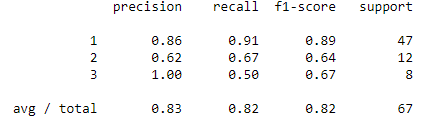
**Softmax Regression**

마지막으로 neural network를 통해 softamx regression 모델을 만들어 보자. Python tensorflow패키지를 이용하여 4-layer, L-th unit : 16, activation function: LeRU인 간단한 네트워크를 만들어 softmax regression 모델을 만들어 보았다.



결과로는 precision 92%가 나왔다. Random forest보다 8%나 높게 나왔지만 overfitting이 의심된다. Missing value처리와 변수 선택을 좀 더 자세하게 했다면 더 좋은 결과가 나왔을 듯 하나, 분석 스케치임을 고려하여 이만 분석을 멈추자.

마지막으로 UCI의 ‘Horse Colic Data Set’에 포함되어 있는 test data를 random forest 모델과 softmax regression모델을 통해 적용시켜 보자

. 

먼저 random forest

