1. 김원정(19941125)
2. btfjeong@naver.com
3. 김원정(19941125)/윤지원(19940308)/권순호(19930510)

Online Test - DSC2018

말을 살리기 위한 데이터 분석

2017-07-29

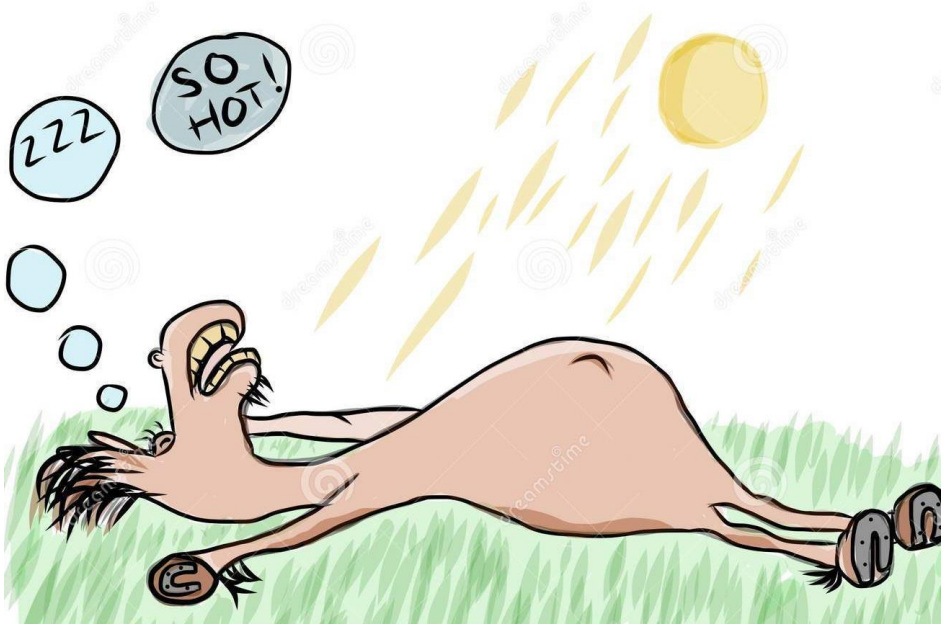
말을 살리기 위한 데이터 분석

권순호, 김원정, 윤지원

1. SUMMARY

BACKGROUND

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



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 극단적 날씨는 말의 건강에 큰 영향을 줄 수 있다. 특히 더운 여름철에는 말의 배앓이(산통)를 조심해야 한다. 말을 죽음에 이르게 하는 가장 흔한 원인으로 요즘 같은 날에는 말에게 조금 더 관심이 필요할 것이다.

산통이란 수의학적인 용어로 쉽게 말해서 배앓이를 말하는 것이다. 말은 큰 체구에 비해 위의 용적이 10리터 정도밖에 되지 않을 정도로 위가 작고 소와 달리 되새김을 할 수 없어 독소가 쉽게 쌓이게 된다. 이러한 특이한 생리구조를 가졌기에 산통이 자주 발생하는데 어떠한 원인에 의해서 발생하더라도 겉으로 드러나는 증상은 대개 비슷하다고 한다. 증상이 심한 경우 수술을 하기도 하고 사망하거나 안락사를 시키기도 한다. 산통이 발생했을 경우 의학적 상태만으로 말의 생존 여부를 예측할 수 있다면 좀 더 빠르게 의학적 조치를 취하고 생존율을 높일 수 있지 않을까?

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.

PROJECT GOAL

1. 말의 생존을 결정하는 결정변수 탐구.
2. 예측 모델링

ANALYSIS MODEL

- Logistic Regression
- OVR Logistic Regression
- Random Forest
- Softmax Regression

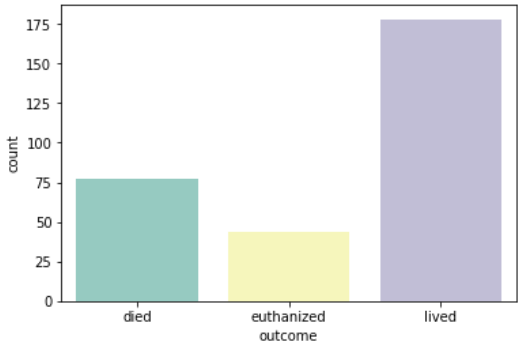
2. DATA EXPLORATION

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 (경주마의 경우 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 가능성이 높거나 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, 복강으로부터 추출하며 흐리거나 혈청이 섞인 것은
    장기가 제대로 작동하지 않음을 시사함.
22. Abdominocentesis total protein: 복강내 총 단백질, 수치가 높을수록 장이 손상되었을 가능성이 큼.
23. Outcome: 결과 (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 비율이다. 이것으로 보아. 산통은 말에게 매우 치명적인 상황인 것을 다시 한번 확인 할 수 있다.



<figure 1> outcome countplot

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

죽은 말과 생존말의 차이점

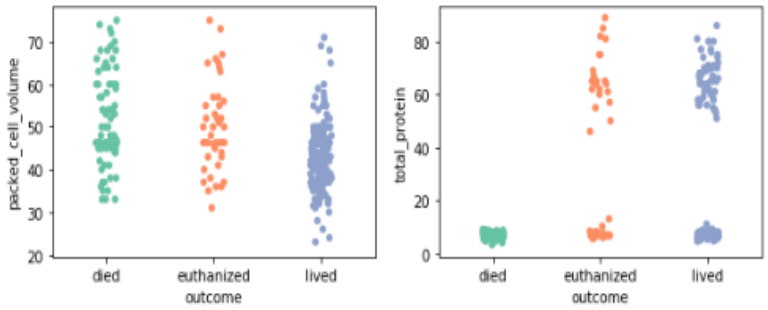
산통이 발생하였지만, 어떤 말은 생존이 가능하였고 어떤 말은 죽음을 피할 수 없었다. 어떤 차이가 있는지 알아보기로 하였다. 다음 기초 통계량 표를 통해 가설을 설정할 있었다.

1) 연속형 변수 탐구

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

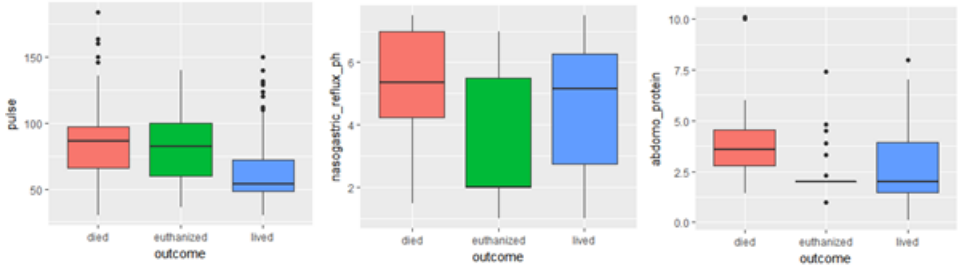
가설: 말의 죽음의 원인은 탈수가 아닌 장기 혈류 순환과 관련된 문제다.



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

Packed cell volume은 수치가 증가할수록 순환이 잘 안되거나 탈수증상이 나타나는 것을 의미한다. 평균으로 살펴보았을 때, 수치가 높은 말이 생존하지 못하였다. Total protein도 수치가 증가할수록 탈수가 더 심해짐을 의미하는 변수이다. 하지만 죽은 말의 total protein은 6.82±1.18로, 정상범위를 거의 넘어가지 않았음을 확인할 수 있었다. 반대로 안락사 당한 말은 total protein 수치가 월등히 높았다. 즉, 어떤 의미에서 탈수 증세가 안락사 결정에 영향을 미쳤다는 뜻이다. 그래프를 통해 total protein과 nasogastric reflux pH 사이에 음의 상관관계가 있을 것이라 유추하였고 실제로 Pearson correlation test를 실행한 결과 coefficient -0.72라는 강한 음의 상관관계가 나타났다. 하지만 샘플이 매우 적어 신뢰성이 떨어지는 변수이므로, 다른 변수를 통해 탐색을 하는 것이 좋을 것이다.

말의 죽음은 obstruction으로 인한 circulatory problem 때문인가



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

Pulse는 순환 쇼크가 발생하면 높아진다. Box plot으로 볼 때 죽은 말이 평균적으로 pulse가 높은 편이었다. Nastogastric ph또한 정상범위를 벗어나 높은 수치를 보였다. 어느정도 산성인 상태가 균을 죽이고 정상적인 활동성을 의미하지만, 평균적으로 높은 PH를 띄고 있어 장 활동이 원활하게 이루어지지 않으며 해로운 균을 죽이기 힘든 상태인 것이다. obstruction으로 인해 산이 분비 되지 않음을 예측할 수 있다. 마지막으로 정말로 위장이 손상된 것인지 abdominocentesis total protein을 통해 확인해 보았다. 수치가 높음을 확인할 수 있다.

Logistic Regression Model

로지스틱 회귀 분석을 통해 말의 죽음과 위 세 개 변수가 유의하게 관련이 있는지 알아보았다.

로지스틱 회귀 분석에 앞서 missing value에 대한 이슈가 있었다. 본 데이터는 결측치가 무작위하게 나타났고, 변수를 삭제할 경우 분석에 어려움이 생길 것으로 보여 다중대입법, 단순대입법을 사용하여 결측치를 대체하여 모델을 두 가지 만들어 보았다. (R mice 패키지, glm함수 사용)
이후 기회가 된다면 변수 특성에 따라 correlation의 coefficient를 이용하는 방법, 단순대입법 모두 사용하여 볼 것이다.

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

• 단순대입법(mean)
Logit = 6.34 - 0.03*pulse - 0.35*abdominocenthesis_total_protein
예측율 72.2%

• 다중대입법(multiple imputation, MI)
Logit = 6.34 - 0.02*pulse - 0.56*abdominocenthesis_total_protein-0.41*nasogastric_reflux_ph
예측율은 74.9%

예상대로 세 변수 모두 음의 상관관계를 가졌다. 다중대입법이 세 변수 모두 유의하게 나왔으며 예측율도 더 높게 나왔다.

결론: 다음 변수들은 obstruction으로 인한 말의 죽음에 어느정도 기여를 하였다.

2) 범주형 변수 탐구

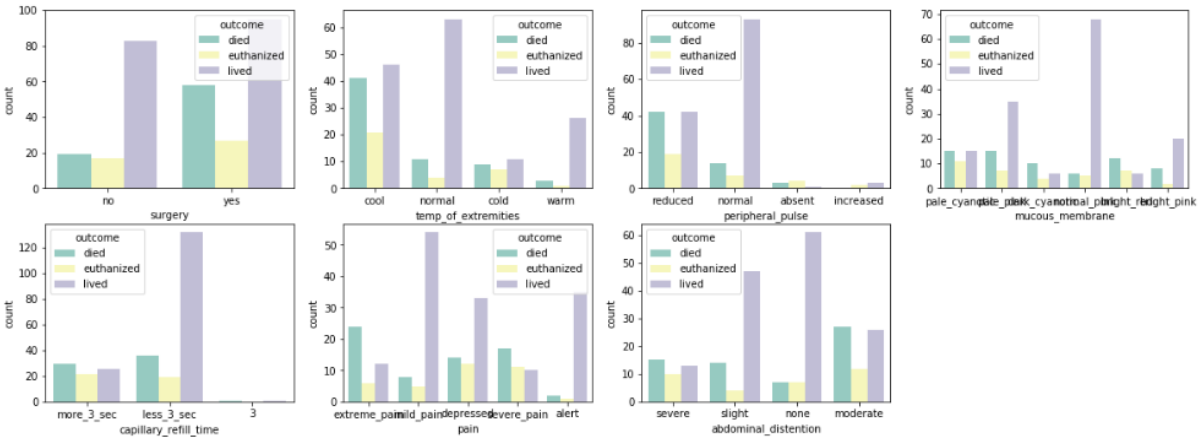
범주형 변수의 통계량은 데이터의 카테고리 비율, R-squared, F(p-value)를 구하여 보았다. 범주형 변수의 통계량 표는 아래와 같다.

Variables	Value	R^2	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 -	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

<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 그려보았다.

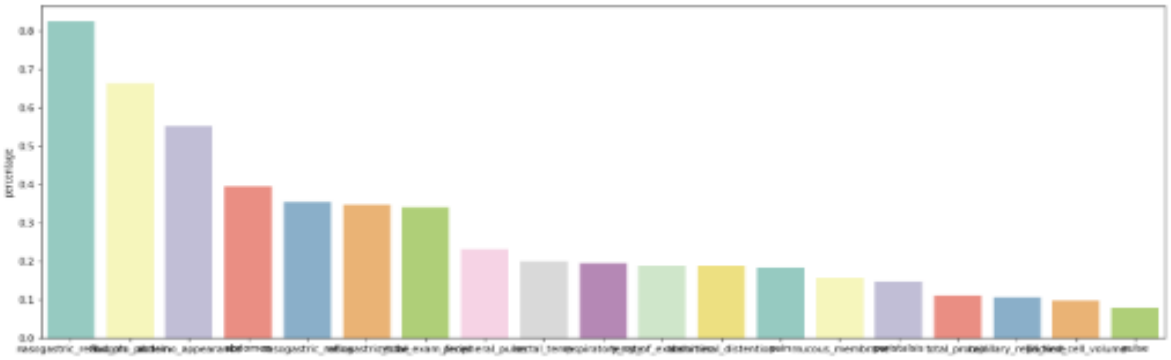


<figure 5> categoric feature count plot

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 언급했듯 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. 우선, 앞에서와는 다르게 가장 단순한 방법을 사용해 보았다. Missing value 비율이 50%이상인 변수는 제외하고 연속형 변수는 평균값을, 범주형 변수는 최빈값으로 설정하고 분석을 계속해보기로 한다. 추후 여러 방법을 적용하여 본 데이터를 가장 잘 설명할 수 있는 방법을 심층적으로 탐구해볼 것이다.



<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 만든다음 train data set으로 학습시키고 test data set으로 Precision을 비교해보기로 하였다. 또, Outcome의 결과가 3개인 점을 고려하여 로지스틱 회귀모형 보다 OVS(one vs rest)로지스틱 회귀모형을 만들어 보았다.

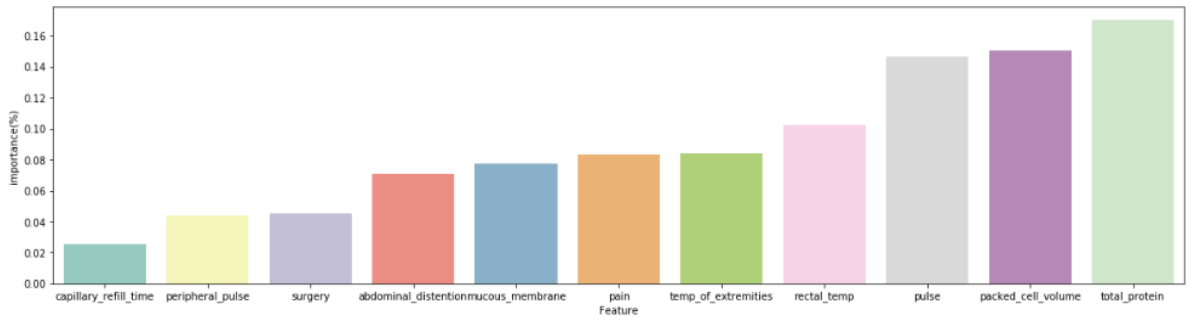
OVR Logistic Regression Model

OVS회귀모형은 python sklearn 패키지의 OneVsRestClassifier과 LogisticRegression함수를 이용하여 만들었다.
변수를 선택하는 과정에서 F(p-value)값을 고려해줬으므로 11개의 변수 그대로 사용하기로 하였고, test dataset에 대한 결과로 아래의 표를 얻었다. Precision이 075%로 만족할 만한 수준은 아니지만 조금 더 복잡한 모델을 사용하면 점수가 높아질 것이라는 생각이 든다.

	precision	recall	f1-score	support
1	0.83	0.93	0.88	43
2	0.60	0.67	0.63	9
3	0.50	0.12	0.20	8
avg / total	0.75	0.78	0.75	60

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.



<figure 7> feature importance plot

변수를 선택할 때 예상 했던 것처럼 total_protein, packed_cell_volume, pulse, rectal_temp변수가 생존에 미치는 영향이 크다는 것을 확인 할 수 있다. test data set에 대한 결과로는 precision 84%가 나왔다. 로지스틱 회귀모형보다 9%나 높게 나왔지만 한번 더 욕심 내어 precision값을 높일수 있는 방법을 생각해보자.

	precision	recall	f1-score	support
1	0.87	0.95	0.91	43
2	0.55	0.67	0.60	9
3	1.00	0.25	0.40	8
avg / total	0.84	0.82	0.80	60

Softmax Regression Model

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.

	precision	recall	f1-score	support
1	0.98	0.95	0.96	43
2	0.75	1.00	0.86	9
3	0.83	0.62	0.71	8
avg / total	0.92	0.92	0.92	60

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.

Test data prediction

이제 UCI의 'Horse Colic Data Set'에 포함되어 있는 test data를 random forest 모델과 softmax regression모델을 통해 적용시켜 보자
먼저 random forest model에서는 앞서 test해본 결과와 비슷한 83%가 나왔다. 다행이 random forest model에서는 overfitting문제가 발생하지 않은 듯 하다.

	precision	recall	f1-score	support
1	0.86	0.91	0.89	47
2	0.62	0.67	0.64	12
3	1.00	0.50	0.67	8
avg / total	0.83	0.82	0.82	67

다음으로 overfitting이 의심되는 softmax regression model에 적용시켜보았다. 우려했던 대로 precision 81%로 앞서 test해본 값과 차이가 심한 것을 확인할 수 있다.

	precision	recall	f1-score	support
1	0.95	0.74	0.83	47
2	0.56	0.83	0.67	12
3	0.42	0.62	0.50	8
avg / total	0.81	0.75	0.76	67

최종적으로 precision이 가장 높은 random forest model을 이용하여 말의 생존을 예측하면 될 것이다. Overfitting과 missing value이슈에 있어 아쉬운 부분이 있지만 컨넥트원에서 더 자세히 배운다면 좋은 분석 모델이 나올 것 같은 자신이 든다.