좋은 와인을 고르기 위한 데이터 분석

1. 김영희 (19991212)
2. dsc2018@navercorp.com
3. 김영희(19991212)/이철수(19980101)/박동수(20000914)

"클로드 모네의 '산보 파라솔을 든 여인' 입니다. '에세조'에는 프레시한 아로마, 그리고 서양의 허브. 상쾌하고 우아한, 그래요, 금발의 귀부인이 산들바람을 맞으며 서 있는 모습."

- 신의 물방울 中, 학산문화사

만화 '신의 물방울'에서 주인공이 와인을 한 모금 마시고 그 맛을 표현한 대사다. 와인에 조예가 깊고 미각이 뛰어나다면, 한 모금의 시음으로도 와인을 평가할 수 있겠지만 와인 가게의 수백 종의 와인에 대하여 그런 경험과 지식을 가지는 것은 어려운 일이다. 와인을 평가하기 위한 기준으로 포도원이나 양조장, 빈티지 등 다양한 기준들이 있지만, 일단은 와인의 화학적 성질 분석만으로 시음을 대체할 수 있는 품질 평가가 가능할 수 있을지, 신의 물방울에서 소개될 만큼 훌륭한 와인을 구분할 수 있을지 궁금하여 분석을 시작하게 되었다.

와인의 화학적 성질만으로 좋은 와인을 구분할 수 있는 방법을 찾아보려고 한다. 개인적으로 화이트와인은 산미가 적당하고 너무 달지 않으며 너무 가볍지 않은 와인을 선호하는데 과연 나의 취향이 품질과 연관이 있을지 확인하여 보겠다.

분석에는 UCI에서 제공하는 'Wine Quality Data Set' 의 화이트 와인 데이터를 이용하였다.

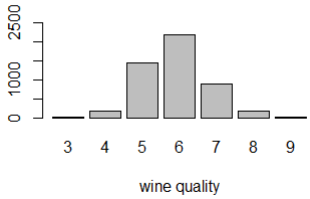
- 데이터 출처 : UC Irvine Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/wine+quality)
- 포르투갈의 특정 지역에서 생산된 화이트 와인의 품질 점수와 11가지의 화학적 성질 데이터
- 4,898 종의 와인 데이터

변수 이름과 설명

1. fixed acidity (tartaric acid) : 타르타르산은 비 휘발성 산으로 쉽게 증발하지 않음
2. volatile acidity (acetic acid) : 아세트산은 휘발성 산으로 와인의 아세트산 함량이 너무 높으면 불쾌한 식초 맛이 날 수 있음
3. citric acid : 소량의 구연산은 와인에 신선함과 풍미를 더할 수 있음
4. residual sugar : 발효가 끝난 후 잔류하는 설탕의 양으로 1g/liter 이하인 와인은 드물고, 45g/liter 이상의 와인은 달콤하다고 여겨짐
5. chlorides (sodium chloride) : 소금의 양
6. free sulfur dioxide : 미생물 성장과 포도주의 산화를 방지함
7. total sulfur dioxide : 와인에서 저농도의 SO2는 잘 느껴지지 않지만, Free sulfur dioxide의 농도가 50ppm 이상이면 맛과 향이 분명하게 느껴짐
8. density : 와인의 농도는 알코올과 설탕 함량에 따라 결정됨
9. pH : 와인이 산성 혹은 염기성인지 나타내며, 대부분의 와인은 pH 3~4 사이임
10. sulphates (potassium sulphate) : 이산화황과 같은 와인 첨가제로 항균 및 항산화제 역할을 함
11. alcohol : 와인의 알코올 함량
12. quality : 0에서 10점 사이의 품질 점수

몇 점 이상이면 좋은 와인이라고 할 수 있을까?

품질 점수(quality)는 0에서 10 사이의 값으로 8점 이상이면 좋은 와인이라고 할 수 있다. 점수 분포를 보면, 5~7인 와인의 비중이 높고, 품질이 아주 좋거나(8~9) 아주 낮은(3~4) 와인은 적은 비중을 차지한다. 데이터를 수집한 지역의 와인은 아주 훌륭한 와인은 드물지만 아주 품질이 떨어지지도 않는, 대부분의 와인이 평균 이상의 품질을 보이는 무난한 지역인 것으로 예상된다.



quality	3	4	5	6	7	8	9
와인 개수	20	163	1,457	2,198	880	175	5
(%)	0.4%	3.3%	29.7%	44.9%	18.0%	3.6%	0.1%

좋은 와인과 나쁜 와인은 화학적인 성질이 다를까?

먼저 좋은 와인과 나쁜 와인의 화학적 성질이 어떻게 다른지를 알아보기 위해서 일부 표본의 데이터를 비교하여 보았다. 품질 점수가 높은 좋은 와인 5개와 품질 점수가 낮은 나쁜 와인 5개의 화학적 성질 데이터는 다음과 같다.

	fixed .acidity	volatile .acidity	citric .acid	residual .sugar	chlorides	free. sulfur. dioxide	total. sulfur. dioxide	density	pH	sulphates	alcohol
품질 상위 와인 (quality=9)	9.1	0.27	0.45	10.6	0.035	28	124	0.997	3.2	0.46	10.4
	6.6	0.36	0.29	1.6	0.021	24	85	0.98965	3.41	0.61	12.4
	7.4	0.24	0.36	2	0.031	27	139	0.99055	3.28	0.48	12.5
	6.9	0.36	0.34	4.2	0.018	57	119	0.9898	3.28	0.36	12.7
	7.1	0.26	0.49	2.2	0.032	31	113	0.9903	3.37	0.42	12.9
품질 하위 와인 (quality=3)	6.8	0.26	0.34	15.1	0.06	42	162	0.99705	3.24	0.52	10.5
	7.1	0.49	0.22	2	0.047	146.5	307.5	0.9924	3.24	0.37	11
	6.1	0.2	0.34	9.5	0.041	38	201	0.995	3.14	0.44	10.1
	5.8	0.24	0.44	3.5	0.029	5	109	0.9913	3.53	0.43	11.7
	6.7	0.25	0.26	1.55	0.041	118.5	216	0.9949	3.55	0.63	9.4

화학적 성질 중 알코올 함량(alcohol)이 가장 먼저 눈에 띄었다. 좋은 와인은 알코올 함량이 12 정도인 반면 나쁜 와인은 10 안팎인 것을 알 수 있다. chlorides도 특징적인데, 나쁜 와인이 소금 함량이 높고 좋은 와인은 상대적으로 낮은 것을 확인할 수 있다. free.sulfur.dioxide는 품질 상위 와인에서는 20~60 정도의 분포인 반면, 하위에서는 100 이상인 경우를 볼 수 있다. free.sulfur.dioxide의 함량이 50ppm 이상이면 이산화황 맛이 느껴진다고 하니 그 이유 때문인 듯 하며, 품질에 큰 영향을 미칠 것으로 예상된다. 10개의 표본 만을 살펴보고 판단을 내리는 것은 성급하다고 할 수 있으므로 전체의 분포를 살펴보는 것이 안전할 것 같다.

와인의 품질과 화학적 성질의 관계를 자세히 살펴보기 위해, 좋은 와인과 나쁜 와인을 아래와 같이 엄밀히 정의하고 두 그룹의 화학적 성질의 분포를 살펴보았다.

- 좋은 와인 : 품질 점수가 8~9 사이인 180개의 와인
- 나쁜 와인 : 품질 점수가 3~4 사이인 183개의 와인

분포를 살펴보기 위해서 두 집단의 히스토그램을 서로 다른 색으로 겹쳐 그리기로 했다. R의 ggplot 함수를 이용하여 간단히 그릴 수 있으며 fixed.acidity, alcohol, citric.acid의 분포를 그리면 다음과 같다.

두 와인 그룹의 fixed.acidity 분포를 보면 상당 부분 겹쳐져 있는 것을 볼 수 있고, alcohol은 두 그룹의 분포가 확연히 구분되어 있는 것을 알 수 있다. Alcohol은 fixed.acidity에 비하여 좋은 와인과 나쁜 와인의 구분 기준으로 더 적합하다는 생각이 든다

또 눈에 띄는 점은 citric.acid인데, 좋은 와인은 0.25~0.4 사이에 집중되어있고 나쁜 와인은 상대적으로 넓게 퍼져있는 것을 볼 수 있다. 이렇게 집중도의 차이도 분포의 중요한 특성이고 좋은 와인을 찾기 위한 훌륭한 힌트가 될 수 있으나 꽤 높은 수준의 통계적 지식을 필요로 할 것 같다. 매우 도전적인 과제이므로 다음 번에 다루도록 하겠다.

좋은 와인과 나쁜 와인의 구분 기준은 어떤 것이 좋을까?

11개의 화학적 성질을 나타내는 변수는 좋은 와인과 나쁜 와인의 분포가 조금씩은 차이가 있는 것을 알 수 있다. 어떤 변수는 그 차이가 크고 어떤 변수는 작다. 먼저 분포의 대표값으로 평균을 이용하여 좋은 와인과 나쁜 와인 집단 차이를 변수 별로 비교하여 보았다.

	좋은 와인 집단 평균	나쁜 와인 집단 평균	평균의 차이
fixed.acidity	6.678	7.181	-0.503
volatile.acidity	0.278	0.376	-0.098
citric.acid	0.328	0.308	0.020
residual.sugar	5.628	4.821	0.807
chlorides	0.038	0.051	-0.013
free.sulfur.dioxide	36.628	26.634	9.994
total.sulfur.dioxide	125.883	130.232	-4.349
density	0.992	0.994	-0.002
pH	3.221	3.183	0.038
sulphates	0.486	0.476	0.010
alcohol	11.651	10.174	1.478

평균의 차이가 가장 큰 변수는 free.sulfur.dioxide이고 가장 작은 변수는 density이다. 하지만 각 변수의 측정 단위가 서로 다르기 때문에 평균의 차이가 크다고 좋은 판단 기준이라고 예단할 수는 없다.

특정 변수가 두 집단을 얼마나 잘 나누는 지에 대한 좀 더 정교한 기준이 필요하다. 모든 변수는 측정 단위가 다르니 수치의 차이보다는 부등호의 방향에 의미를 두는 방법을 생각해 보았다. 우리가 원하는 것은 좋은 와인과 나쁜 와인을 구분하는 것이니 특정 변수의 수치가 모든 좋은 와인과 나쁜 와인의 쌍에 대해서 일관된 부등호를 가지고 있다면 매우 좋은 구분 기준이라고 할 수 있을 것이다. 좋은 와인과 나쁜 와인의 쌍에 대해서 부등호의 방향을 세는 방식을 고려해 보았다. 180개의 좋은 와인과 183개의 나쁜 와인으로 32,940개의 (좋은 와인, 나쁜 와인) 쌍을 만들 수 있다. 특정 변수가 좋은 구분 기준은 일관된 부등호를 가질 것이고 나쁜 구분 기준은 약 반반의 부등호를 가지고 있을 것이다.

	best<worst 인 pair	best>=worst 인 pair		
alcohol	6,347	26,593	19%	81%
density	24,367	8,573	74%	26%
free.sulfur.dioxide	8,714	24,226	26%	74%
chlorides	23,870	9,070	72%	28%
volatile.acidity	22,632	10,308	69%	31%
residual.sugar	13,505	19,435	41%	59%
fixed.acidity	19,434	13,506	59%	41%
pH	13,922	19,018	42%	58%
citric.acid	14,510	18,430	44%	56%
total.sulfur.dioxide	16,344	16,596	50%	50%
sulphates	16,410	16,530	50%	50%

두 그룹에 대해서 가장 부등호가 일관된 변수는 alcohol이었다. 전체 와인 쌍 중에 좋은 와인의 알코올 함량이 나쁜 와인의 수치보다 적은 쌍의 비율은 19%에 불과했으며, 이는 임의로 좋은 와인과 나쁜 와인 쌍을 알코올 함량을 기준으로 판단하면 81% 정도는 맞출 수 있다는 것을 의미한다. 다음으로 density가 좋은 와인과 나쁜 와인 쌍에 대해서 일관된 부등호를 가지고 있는데, 74%정도의 정확성을 가지고 있다. density 분포의 평균은 -0.002의 미미한 차이를 보였으나 실제 와인을 구분하는 데는 매우 유용한 변수라는 것을 알 수 있다.

와인 판별 모형

지금까지의 결과를 바탕으로 와인의 품질을 평가하기 위한 가장 좋은 하나의 성질 만을 고른다면 알코올 함량이 가장 좋은 구분 기준으로 생각된다. 그렇다면, 하나의 변수가 아니라 여러 화학 성질을 종합적으로 고려한 구분 기준을 만들면 어떻게 될까? 여러 변수의 선형 결합으로 더 좋은 변수를 만들 수 있지 않을까? 통계학 교과서에 나와있는 로지스틱 회귀모형을 이용하여 판별식을 만들어 보았다.

Logistic Regression Model

로지스틱 회귀모형은 R의 glm 함수를 이용하여 만들 수 있다. Stepwise로 변수 선택을 하였으며 아래와 같은 결과가 나왔다.

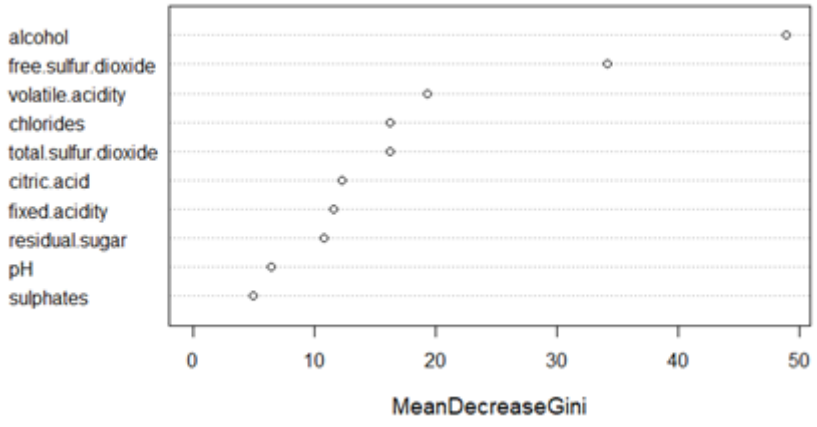
Logit(좋은 와인일 확률) = -24.96 + 1.64*alcohol - 10.48*volatile.acidity + 0.27*residual.sugar + 2.80*pH
+ 0.03*free.sulfur.dioxide - 0.01*total.sulfur.dioxide

	best<worst 인 pair	best>=worst 인 pair		
로지스틱 회귀모형	2,732	30,208	8%	92%

모형을 이용하여 좋은 와인과 나쁜 와인을 구분하여 보면 83.7%의 정확도를 보인다. 알코올 함량 한 가지만 고려하였을 때 보다 여러 변수들을 고려하니 더 좋은 성능을 보임을 알 수 있다.

Random Forest

그렇다면 이제 좀 더 욕심을 내어서 로지스틱 모형의 선형 결합보다 좀 더 복잡한 형태의 함수를 고려하면 어떻게 될까? 기계학습 교과서에 있는 Random Forest 모형을 적용하여 보았다. R의 randomForest 함수를 이용하였고, 150개의 의사결정나무모형이 만들어졌다. 랜덤포레스트모형은 그 결과를 직접적으로 표현하기 어려우니 각 변수의 영향력을 이용하여 모형에서의 간접적인 영향을 살펴보면 다음과 같다.



랜덤포레스트도 역시 alcohol이 가장 중요한 변수로 와인의 품질에 미치는 영향이 가장 크다고 판단하였다. Alcohol에 비해 중요도는 떨어지나 free.sulfur.dioxide, volatile.acidity 순으로 영향도가 높았고, sulphates와 pH는 품질에 미치는 영향이 가장 적었다.

	best<worst 인 pair	best>=worst 인 pair		
랜덤포레스트 모형	0	32,940	0%	100%

랜덤포레스트 모형을 이용하여 좋은 와인과 나쁜 와인을 구분하여 보면 100%의 정확도를 보인다. 고차원의 복잡한 모형을 사용했기 때문에 과적합 현상(*과적합 현상을 설명한 블로그 https://boolio.blog.me/220928734664*)이 발생했을 우려가 있다. 과적합 현상을 확인하기 위해서는 별도의 테스트용 표본을 활용하면 될 것 같으나 어차피 랜덤포레스트처럼 복잡한 모형으로 와인 가게에서 와인을 고르는 것은 무리라는 생각에 이쯤에서 분석을 멈추려고 한다.

이제 데이터 분석의 결과를 실제에 적용하여 보자.

내가 가지고 있는 와인들이다. 와인의 판단 기준 중에 가장 쉬우면서 판별력도 쓸만한 알코올 함량을 이용하여 좋은 와인을 판단하여 보니 가장 맛있는 와인으로 네 번째 와인이 뽑혔다. 과연 나의 모형이 추천해준 알코올 함량이 가장 높은 4번 와인은 정말 맛이 좋은 와인일까?

와인의 맛은 가격에 비례한다고 생각하는데 다행히도 가장 비싼 와인의 알코올 함량이 가장 높았다.

(왼쪽부터 순서)	와인 이름	알코올 함량	가격(추정)
1	Chateau Lafon-Rochet 2013	13%	130,000
2	Chateau La Tour Carnet 2014	13.5%	200,000
3	Pio Cesare Barolo 2012	14%	134,000
4	Opus One 2013	14.5%	600,000
5	Domaine Faiveley Mercurey	13%	100,000

와인들의 가격 차이가 상당하다. 가격대비 성능을 고려하면 어떤 와인이 가장 좋을지, 더 간단한 와인 선별 기준이 있는지 8월에 커넥트원에서 더 이야기를 나누어 보자