

와인 가격에 영향을 미치는 요인에 대한 회귀분석

2019. 06. 21

강병우

목차

1	프로젝트 개요	3
	프로젝트 주제	4
	주제 선정 이유	4
	데이터에 대한 설명	5
2	프로젝트 수행 내용	8
	누락 데이터 처리	9
	EDA & Feature engineering	11
3	프로젝트 결과	23
	Baseline : Linear regression(OLS)	25
	Final: Gradient boosting regressor	31
4	한계 및 보완 방향	32

01

프로젝트 개요

1. 프로젝트 개요

1 프로젝트 주제

[WineEnthusiast](#)이라는 사이트에서 2017년에 수집된 데이터를 통해 가격에 영향을 미치는 요인과 그 정도를 분석하고 와인의 가격을 예측하는 회귀분석 프로젝트

2 주제 선정 이유

1. 개인적 흥미
 - 평소에 관심이 많은 커피와 비슷한 성격의 와인에 대한 흥미로운 데이터를 발견
2. Not Kaggle Competition
 - 이미 문제에 대한 정의가 되어있는 Kaggle Competition 데이터가 아닌, 내가 관심있는 주제를 바탕으로 하여 스스로 문제를 정의하고 결과를 분석해보는 프로젝트
 - 현업의 데이터처럼 누락 데이터가 상당부분 존재하는 데이터를 다루면서 전처리 과정과 feature engineering 작업에 대해 고민해보는 과정

1. 프로젝트 개요

3 데이터에 대한 설명

- Target(종속 변수)

- price: 와인 한 병의 가격

- Features(독립 변수)

- points: WineEnthusiast가 와인을 80 ~ 100 점 척도로 평가 한 점수
- country: 와인이 생산된 국가
- province: 와인이 생산된 지역(주)
- region_1, region_2 : 와인이 생산된 province의 구역
- designation: 포도가 생산된 winery의 포도원(Vineyard)
- winery: 와인이 생산된 양조장
- variety: 와인을 만드는데 사용된 포도의 품종
- title: 와인의 이름, 라벨
- description: 와인에 대한 시음 평가 텍스트
- taster_name: 와인을 시음하고 리뷰한 테이스터의 이름
- taster_twitter_handle: 위 테이스터의 twitter id

■ Excel로 확인해 본 데이터의 모습

A	B	C	D	E	F	G	H	I	J	K	L	M	N	
	country	description	points	price	province	region_1	region_2	taster_name	taster_twitter_title	variety	winery			
0	Italy	Aromas include tropical fruit, broom, brimstone and dried herb. The palate is tart and mineral.	87		Sicily & Sardinia	Etna		Kerin O'Sullivan @kerinokeefe	Nicosia 2013	White Blend	Nicosia			
1	Portugal	This is ripe and fruity, a wine that is smooth while still structured. Firm tannins add to the texture.	87	15	Douro			Roger Voss @vossroger	Quinta dos Azeiteiros	Portuguese Red	Quinta dos Avidagos			
2	US	Tart and snappy, the flavors of lime flesh and rind dominate. Some green tea notes are present.	87	14	Oregon	Willamette Valley	Willamette Valley	Paul Gregutt @paulgwine	Rainstorm 2011	Pinot Gris	Rainstorm			
3	US	Pineapple rind, lemon pith and orange blossom start off the aromas. The palate is bright and citrusy.	87	13	Michigan	Lake Michigan Shore		Alexander Peartree St. Julian	2011 Riesling	St. Julian				
4	US	Much like the regular bottling from 2012, this comes across as rather round and soft.	87	65	Oregon	Willamette Valley	Willamette Valley	Paul Gregutt @paulgwine	Sweet Cheeks	Pinot Noir	Sweet Cheeks			
5	Spain	Blackberry and raspberry aromas show a typical Navarran whiff of green leaf and earth.	87	15	Northern Spain	Navarra		Michael Schafer @wineschach	Tandem 2011 Tempranillo	Tempranillo	Tandem			
6	Italy	Here's a bright, informal red that opens with aromas of candied berry, white pepper and earth.	87	16	Sicily & Sardinia	Vittoria		Kerin O'Sullivan @kerinokeefe	Terre di Giurfo Frappato	Frappato	Terre di Giurfo			
7	France	This dry and restrained wine offers spice in profusion. Balanced with acidity, it has a long finish.	87	24	Alsace	Alsace		Roger Voss @vossroger	Trimbach 2011 Gewürztraminer	Gewürztraminer	Trimbach			
8	Germany	Savory dried thyme notes accent sunnier flavors of preserved peach in the nose.	87	12	Rheinhessen			Anna Lee C. Iijima	Heinz Eifel 2011	Gewürztraminer	Heinz Eifel			
9	France	This has great depth of flavor with its fresh apple and pear fruits and touch of oak.	87	27	Alsace	Alsace		Roger Voss @vossroger	Jean-Baptiste Pinot Gris	Pinot Gris	Jean-Baptiste Adam			
10	US	Soft, supple plum envelopes an oaky structure in this Cabernet, supported by fine tannins.	87	19	California	Napa Valley	Napa	Virginie Boon @vboone	Kirkland Signature	Cabernet Sauvignon	Kirkland Signature			
11	France	This is a dry wine, very spicy, with a tight, taut texture and strongly mineral-toned.	87	30	Alsace	Alsace		Roger Voss @vossroger	Leon Beyer 2011	Gewürztraminer	Leon Beyer			
12	US	Slightly reduced, this wine offers a chalky, tannic backbone to an otherwise soft and supple palate.	87	34	California	Alexander Valley	Sonoma	Virginie Boon @vboone	Louis M. Martini	Cabernet Sauvignon	Louis M. Martini			
13	Italy	This is dominated by oak and oak-driven aromas that include roasted coffee and dark chocolate.	87		Sicily & Sardinia	Etna		Kerin O'Sullivan @kerinokeefe	Masseria Setteporte Nerello Mascalese	Nerello Mascalese	Masseria Setteporte			
14	US	Building on 150 years and six generations of winemaking tradition, the wine is full-bodied and complex.	87	12	California	Central Coast	Central Coast	Matt Kettmar @mattkettmar	Mirassou 2011	Chardonnay	Mirassou			
15	Germany	Zesty orange peels and apple notes abound in this sprightly, mineral-toned wine.	87	24	Mosel			Anna Lee C. Iijima	Richard Burch Riesling	Riesling	Richard Burch			
16	Argentina	Baked plum, molasses, balsamic vinegar and cheesy oak aromas feed into the palate.	87	30	Other	Cafayate		Michael Schafer @wineschach	Felix Lavaque Malbec	Malbec	Felix Lavaque			
17	Argentina	Raw black-cherry aromas are direct and simple but good. This has a juicy, ripe texture.	87	13	Mendoza	Pro Mendoza		Michael Schafer @wineschach	Gaucho Andino Malbec	Malbec	Gaucho Andino			
18	Spain	Desiccated blackberry, leather, charred wood and mint aromas carry the wine through.	87	28	Northern Spain	Ribera del Duero		Michael Schafer @wineschach	Pradorey 2011	Tempranillo	Pradorey			
19	US	Red fruit aromas pervade on the nose, with cigar box and menthol notes.	87	32	Virginia	Virginia		Alexander Peartree	Quique Vremont Meritage	Meritage	Quique Vremont			
20	US	Ripe aromas of dark berries mingle with ample notes of black pepper, tobacco and earth.	87	23	Virginia	Virginia		Alexander Peartree	Quique Vremont Red Blend	Red Blend	Quique Vremont			
21	US	A sleek mix of tart berry, stem and herb, along with a hint of oak and charred wood.	87	20	Oregon	Oregon	Oregon Other	Paul Gregutt @paulgwine	Acrobat 2013	Pinot Noir	Acrobat			
22	Italy	Delicate aromas recall white flower and citrus. The palate offers passion fruit and lime.	87	19	Sicily & Sardinia	Sicilia		Kerin O'Sullivan @kerinokeefe	Baglio di Pianetto White Blend	White Blend	Baglio di Pianetto			
23	US	This wine from the Geneseo district offers aromas of sour plums and just-baked bread.	87	22	California	Paso Robles	Central Coast	Matt Kettmar @mattkettmar	Bianchi 2011	Merlot	Bianchi			
24	Italy	Aromas of prune, blackcurrant, toast and oak carry through to the extractive finish.	87	35	Sicily & Sardinia	Sicilia		Kerin O'Sullivan @kerinokeefe	Canicattì 2011 Nero d'Avola	Nero d'Avola	Canicattì			
25	US	Oak and earth intermingle around robust aromas of wet forest floor in the nose.	87	69	California	Sonoma Coast	Sonoma	Virginie Boon @vboone	Castello di Amorosa Pinot Noir	Pinot Noir	Castello di Amorosa			
26	Italy	Pretty aromas of yellow flower and stone fruit lead the nose. The bright palate is balanced by a touch of oak.	87	13	Sicily & Sardinia	Terre Siciliane		Kerin O'Sullivan @kerinokeefe	Stemmari 2011	White Blend	Stemmari			
27	Italy	Aromas recall ripe dark berry, toast and a whiff of cake spice. The soft, integrated tannins add to the complexity.	87	10	Sicily & Sardinia	Terre Siciliane		Kerin O'Sullivan @kerinokeefe	Stemmari 2011 Nero d'Avola	Nero d'Avola	Stemmari			
28	Italy	Aromas suggest mature berry, scorched earth, animal, toast and anise. The palate is balanced and expressive.	87	17	Sicily & Sardinia	Cerasuolo di Vittoria		Kerin O'Sullivan @kerinokeefe	Terre di Giurfo Red Blend	Red Blend	Terre di Giurfo			
29	US	Clarksburg is becoming a haven for Chenin Blanc in California. This bottling is no exception.	86	16	California	Clarksburg	Central Valley	Virginie Boon @vboone	Clarksburg 2011 Chenin Blanc	Chenin Blanc	Clarksburg Wine Company			
30	France	Red cherry fruit comes laced with light tannins, giving this bright wine airiness and elegance.	86		Beaujolais	Beaujolais-Villages		Roger Voss @vossroger	Domaine de la Madone Gamay	Gamay	Domaine de la Madone			
31	Italy	Merlot and Nero d'Avola form the base for this easy red wine that would be perfect with a meal.	86		Sicily & Sardinia	Sicilia				Duca di Salaparuta Red Blend	Red Blend	Duca di Salaparuta		

1. 프로젝트 개요

3 데이터에 대한 설명

■ 데이터 크기

총 119,988개의 row, 13개의 column을 지니고 있는 데이터 셋입니다.

■ 데이터 특성

수치형 데이터 (continuous variable)	카테고리 데이터 (categorical variable)	텍스트 데이터 (text data)
price, points	country, province, region_1, region_2, winery, variety, taster_name, taster_twitter_handle	designation, title, description

4 모델 선택

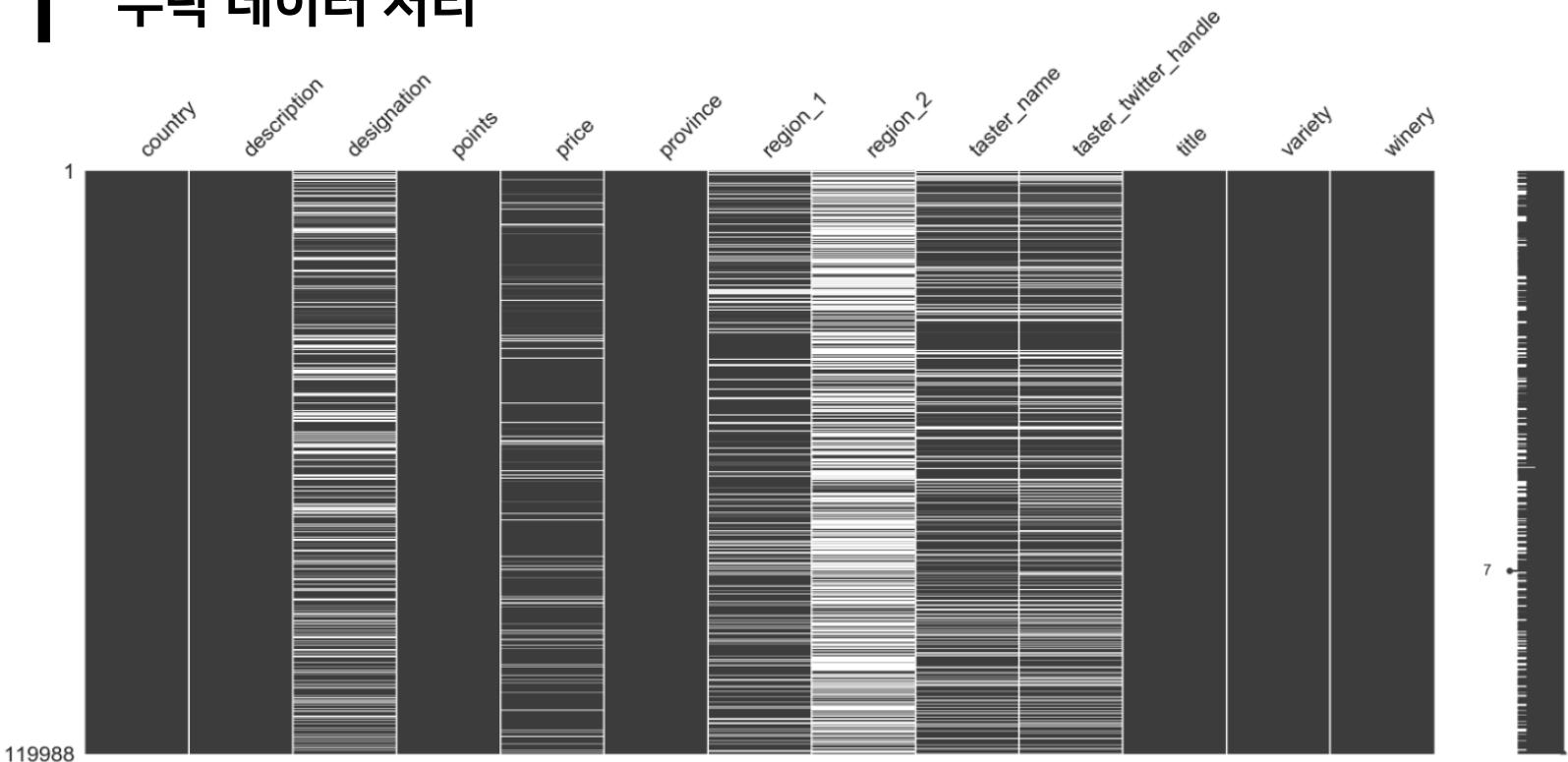
Baseline: Linear regression model(Statsmodels package)
Final: Gradient Boosting regressor(LightGBM)

02

프로젝트 수행 내용

2. 프로젝트 수행 내용

1 누락 데이터 처리



	Nan count	Nan rate
region_2	73219.0	61.02
designation	34545.0	28.79
taster_twitter_handle	29446.0	24.54
taster_name	24917.0	20.77
region_1	19560.0	16.30
price	8395.0	7.00
province	59.0	0.05
country	59.0	0.05
variety	1.0	0.00
winery	0.0	0.00
title	0.0	0.00
points	0.0	0.00
description	0.0	0.00

누락된 값이 전혀 없는 데이터(row)는 119988개의 데이터 중 20493개로, 전체 데이터의 20%도 되지 않습니다. 이 데이터만으로 프로젝트를 진행하는 것은 정보가 지나치게 손실되는 문제가 있으므로 적절한 누락 데이터 처리가 필요합니다.

2. 프로젝트 수행 내용

1 누락 데이터 처리

■ region_2

누락된 비율이 50%를 넘으며 대부분 region_1과 중복되는 값이 많아 feature에서 제외

■ price

정답이 없는 데이터를 학습시킬 수는 없으므로 price가 누락된 데이터들은 drop (전체 데이터의 7% 비율)

■ taster_twitter_handle

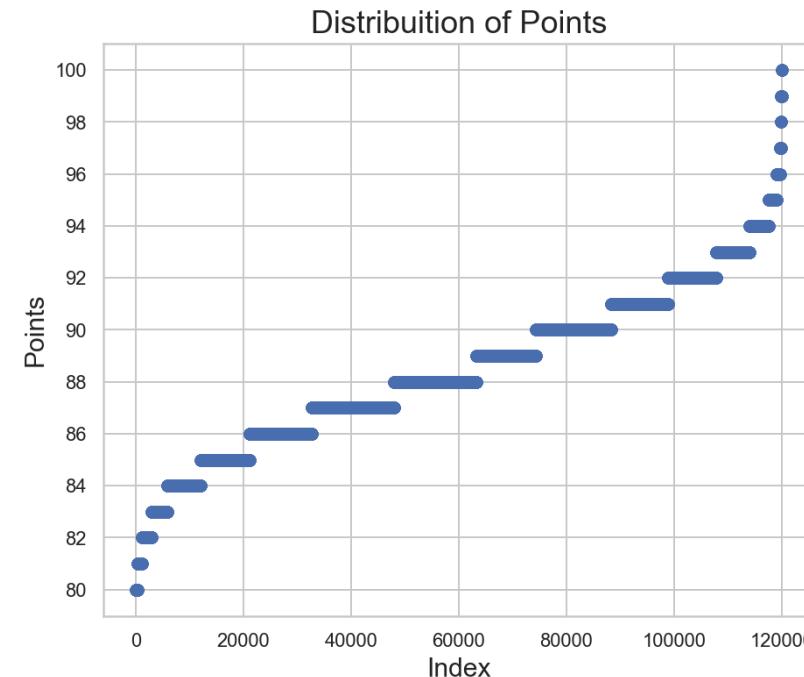
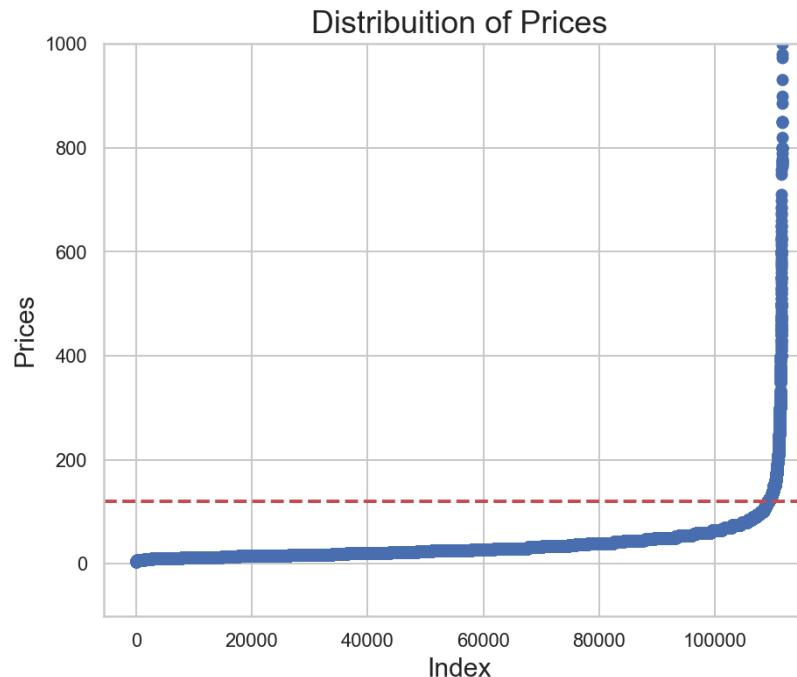
Twitter id 텍스트가 와인 가격에 영향력을 미치는 것은 아니므로 아이디가 존재하는지를 체크하는 binary feature로 가공

■ Etc.

그 외의 카테고리 변수에서 누락된 값들은 'unknown'으로 채운 다음 프로젝트를 진행

2. 프로젝트 수행 내용

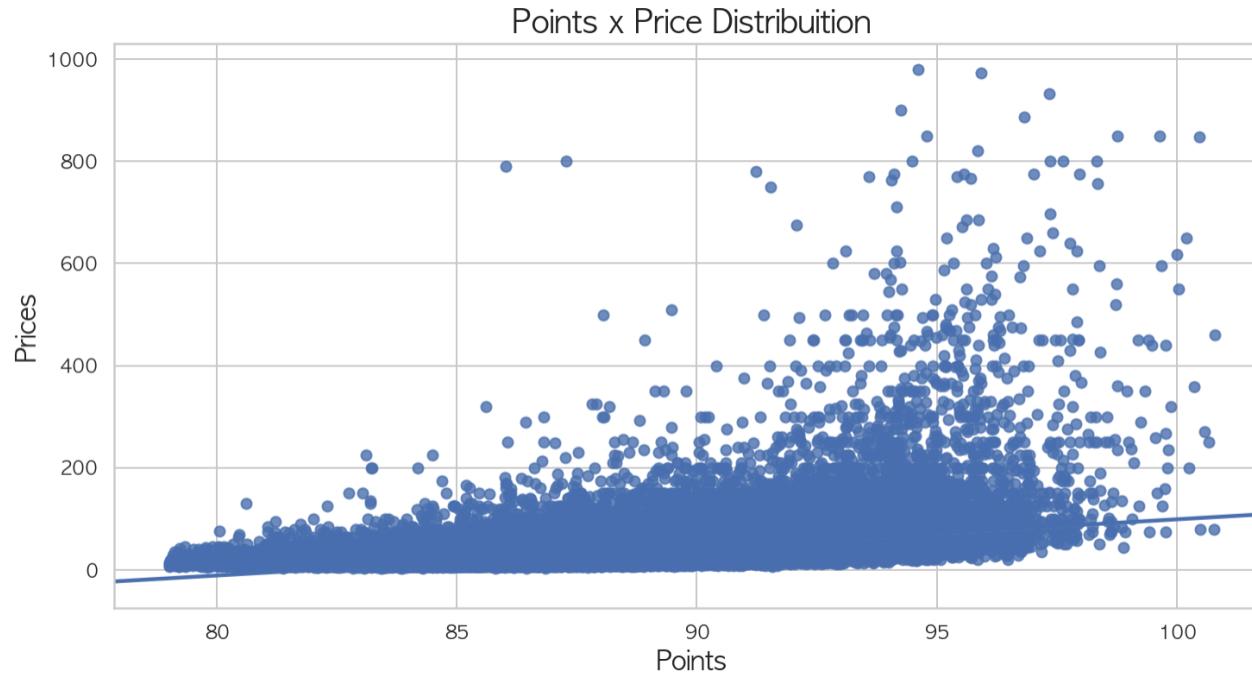
1 EDA: price, points



120 이하의 price 데이터가 전체 데이터의 98%를 차지하고 있으며, 소수의 고가 와인 데이터가 존재합니다.
원활한 모델링을 위해서, z score 2 이상에 해당하는 소수의 고가 와인 데이터를 아웃라이어로 규정하여 제외했습니다.

2. 프로젝트 수행 내용

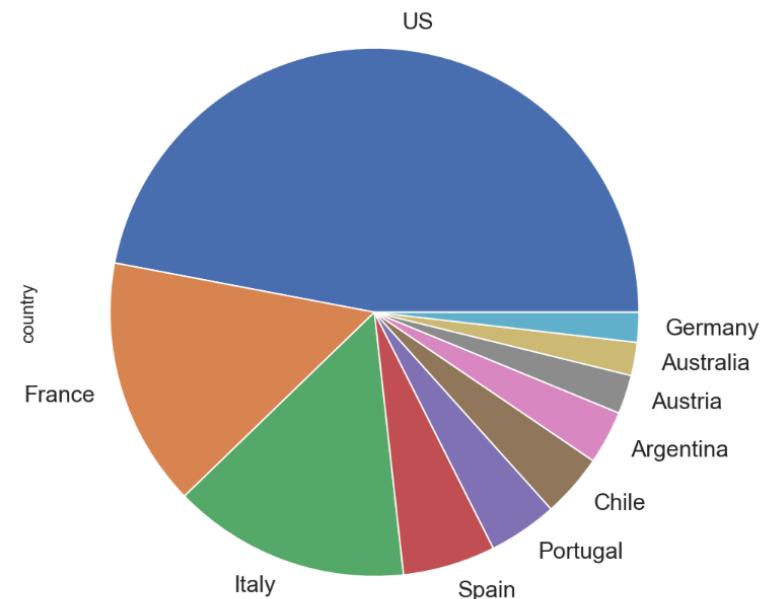
1 EDA: price, points



Points는 price와 0.42의 양적 상관관계를 지니고 있습니다. 후술할 선형 회귀분석 모델 결과에서도 가장 높은 회귀계수로써 와인 가격에 가장 큰 영향력을 미치는 feature로 분석되었습니다.

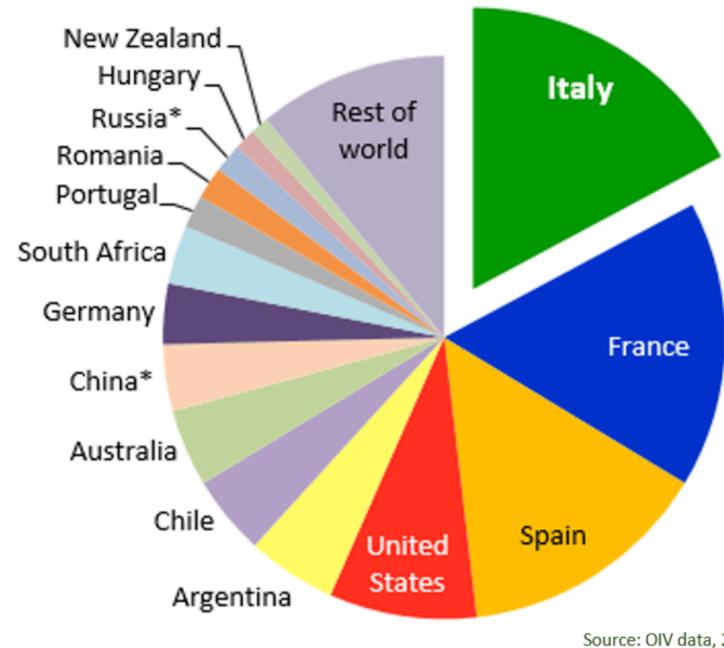
2. 프로젝트 수행 내용

1 EDA: country



〈데이터셋의 각 국가 와인 생산량〉

World Wine Production, 2018

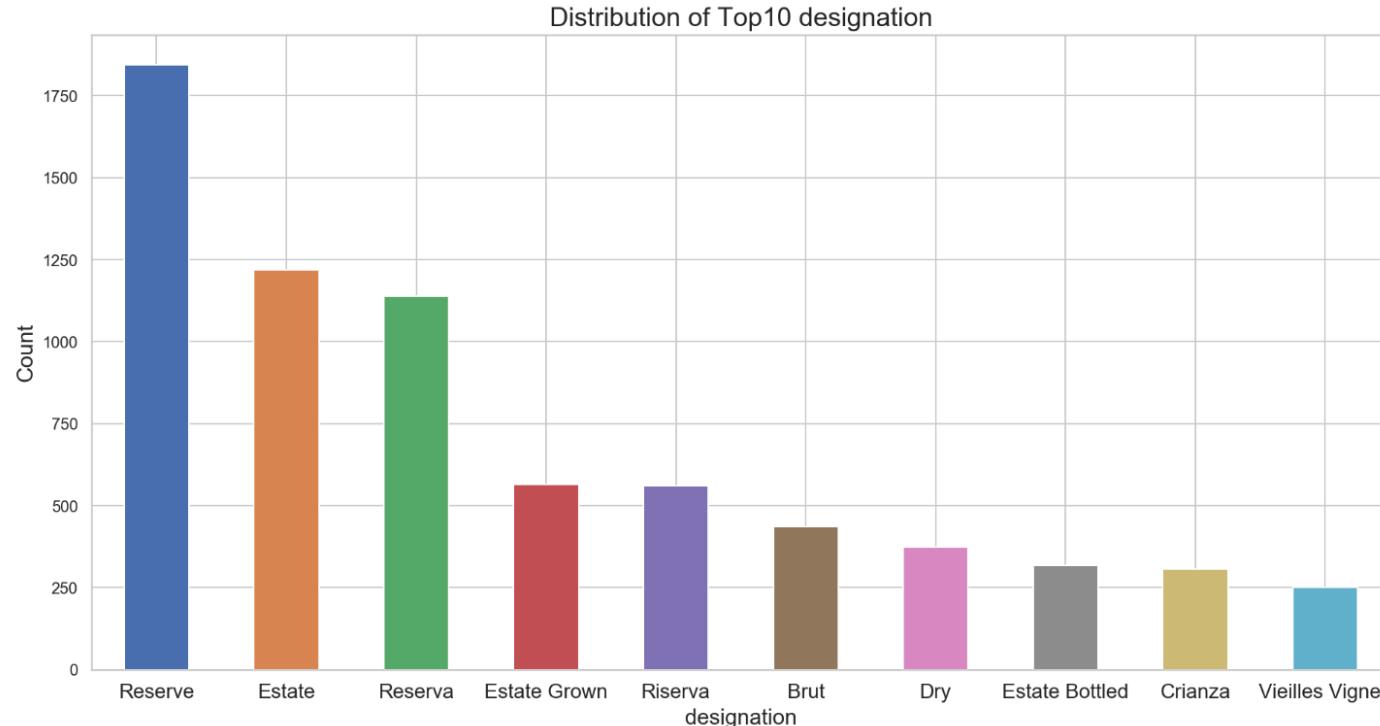


〈실제 세계의 각 국가 와인 생산량〉

프로젝트 데이터는 실제 세계와 다르게 미국에서 생산된 와인의 비율이 상당히 높습니다. 따라서 프로젝트의 데이터는 실제 세계의 와인 생산 국가 분포와 다소 거리감이 있는, 미국에 편향된 데이터라 할 수 있습니다.

2. 프로젝트 수행 내용

1 EDA: designation



designation은 winery에서 와인을 제조하는 포도의 포도원(vineyard: 포도밭)을 의미합니다. 그러나 designation에서 높은 빈도로 등장하는 값들은 특정한 포도원 지명이 아닌 포도의 유통 및 관리 방식을 의미하며, designation의 값들은 이러한 개념들과 실제의 포도원 이름이 혼재되어 있는 특징을 가지고 있습니다. Ex. 'Estate Riverbend Vineyard'

2. 프로젝트 수행 내용

1 Feature engineering: designation

```
def remove_special_numbers(sentence):

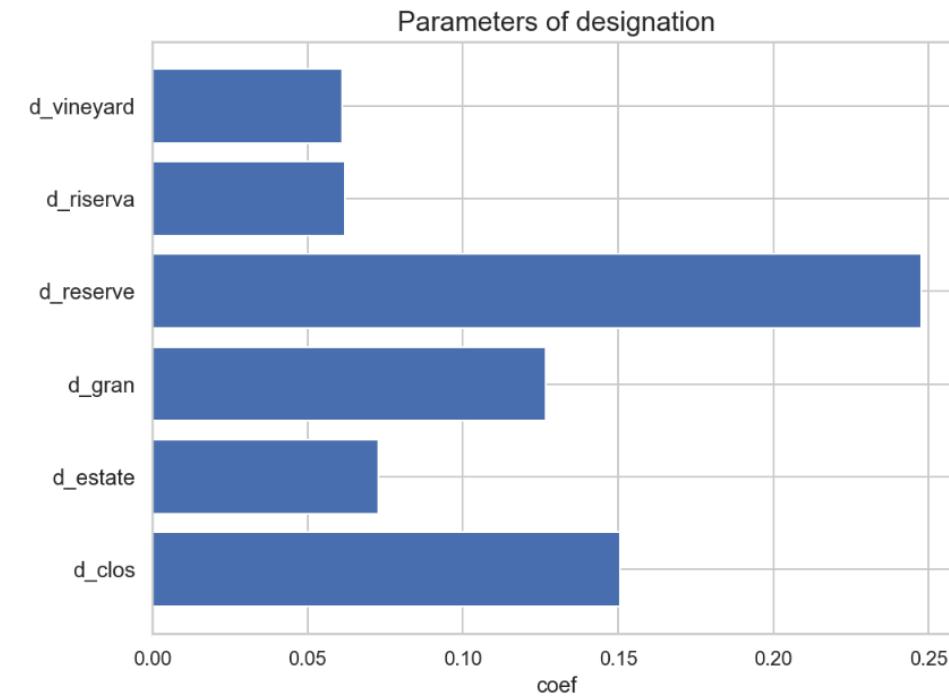
    text = ''.join([word for word in sentence if not word.isdigit()])
    stripped = re.sub("[^\w\s]", ' ', text)
    stripped = re.sub("_", ' ', stripped)
    stripped = re.sub(r"\b[a-zA-Z]\b", "", stripped)
    stripped = re.sub("\s+", " ", stripped)

    return stripped.strip()

onehot_train_df['designation'] = train_df['designation'].apply(lambda x: x.lower())
onehot_train_df['designation'] = train_df['designation'].apply(remove_special_numbers)

vect = CountVectorizer(stop_words=['winery', 'wine'],
                      ngram_range=(1, 2),
                      max_features=100) # max_features로 빈도가 높은 값을 추출
X = vect.fit_transform(onehot_train_df['designation'])
print("Number of extracted designation features: ", X.shape[1])

Number of extracted designation features: 100
```



〈후술할 L1 regularization 작업으로 선택된 designation 요소들〉

따라서 designation은 카테고리 데이터보다는 텍스트 데이터로 간주하여 feature engineering 작업을 진행하는 것이 적절하다고 판단하여 각 텍스트 값들을 BoW 방식으로 tokenizing했습니다. 앞서 확인했던 높은 빈도로 등장 포도의 유통 및 관리 방식을 의미하는 개념 위주로 정보들이 추출될 수 있도록 max_features를 적용했습니다.

2. 프로젝트 수행 내용

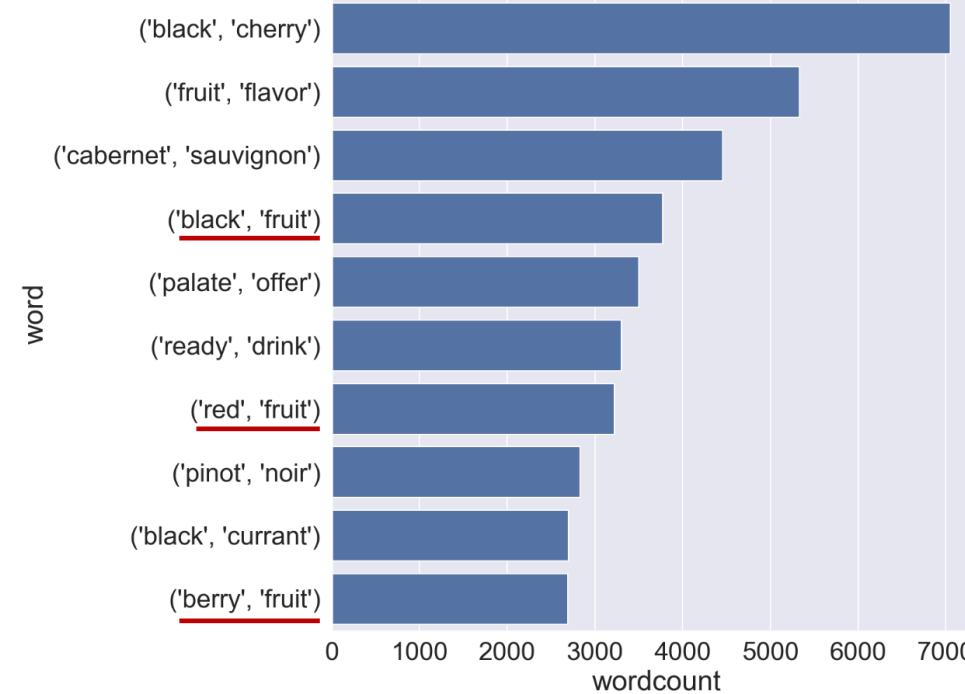
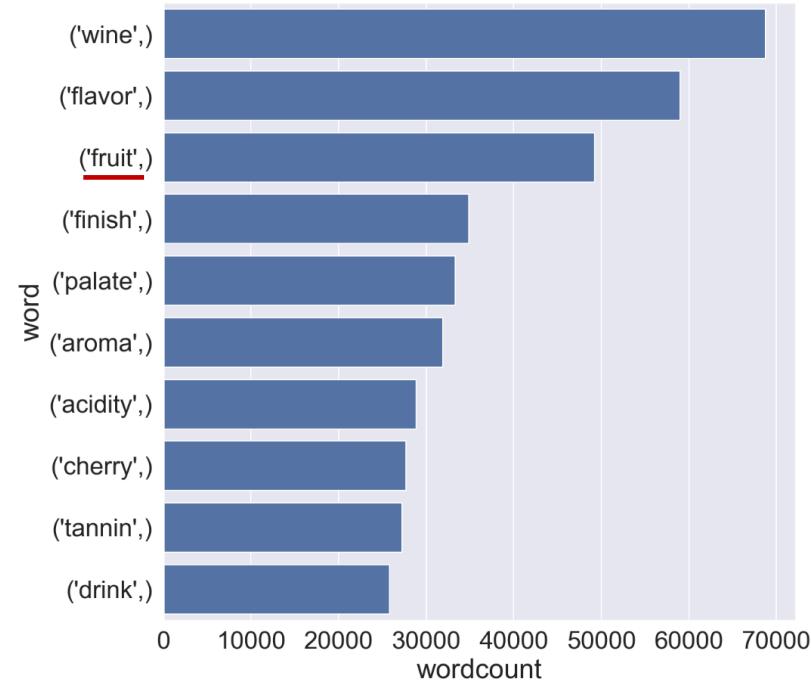
3 Feature engineering: Text data: title -> vintage

title
Nicosia 2013 Vulkà Bianco (Etna)
Quinta dos Avidagos 2011 Avidagos Red (Douro)
Rainstorm 2013 Pinot Gris (Willamette Valley)
St. Julian 2013 Reserve Late Harvest Riesling (Lake Michigan Shore)
Sweet Cheeks 2012 Vintner's Reserve Wild Child Block Pinot Noir (Willamette Valley)
Tandem 2011 Ars In Vitro Tempranillo-Merlot (Navarra)
Terre di Giurfo 2013 Belsito Frappato (Vittoria)
Trimbach 2012 Gewürztraminer (Alsace)
Heinz Eifel 2013 Shine Gewürztraminer (Rheinhessen)
Jean-Baptiste Adam 2012 Les Natures Pinot Gris (Alsace)
Kirkland Signature 2011 Mountain Cuvée Cabernet Sauvignon (Napa Valley)

title은 와인의 이름으로 포도의 품종, 생산지, winery, 빈티지(vintage: 와인을 만들기 위해 포도를 수확한 해) 등의 정보가 복합적으로 담겨있습니다. title에서는 vintage에 대한 정보만을 추출하여 새로운 feature로 가공했습니다.

2. 프로젝트 수행 내용

3 EDA: description -> unigram, bigram



description에서는 (형용사+명사) 등의 두 단어로 이루어진 텍스트가 하나의 단어보다 와인의 맛과 풍미를 보다 효과적으로 표현하는 경향이 있습니다. 따라서 unigram+bigram으로 CountVectorizing 한 데이터를 stacked auto-encoder로 차원을 축소하며 sparse한 데이터 구조를 연속적인 수치형 데이터로 바꾸어, 보다 선형 회귀분석에 적합한 형태로 만들었습니다.

2. 프로젝트 수행 내용

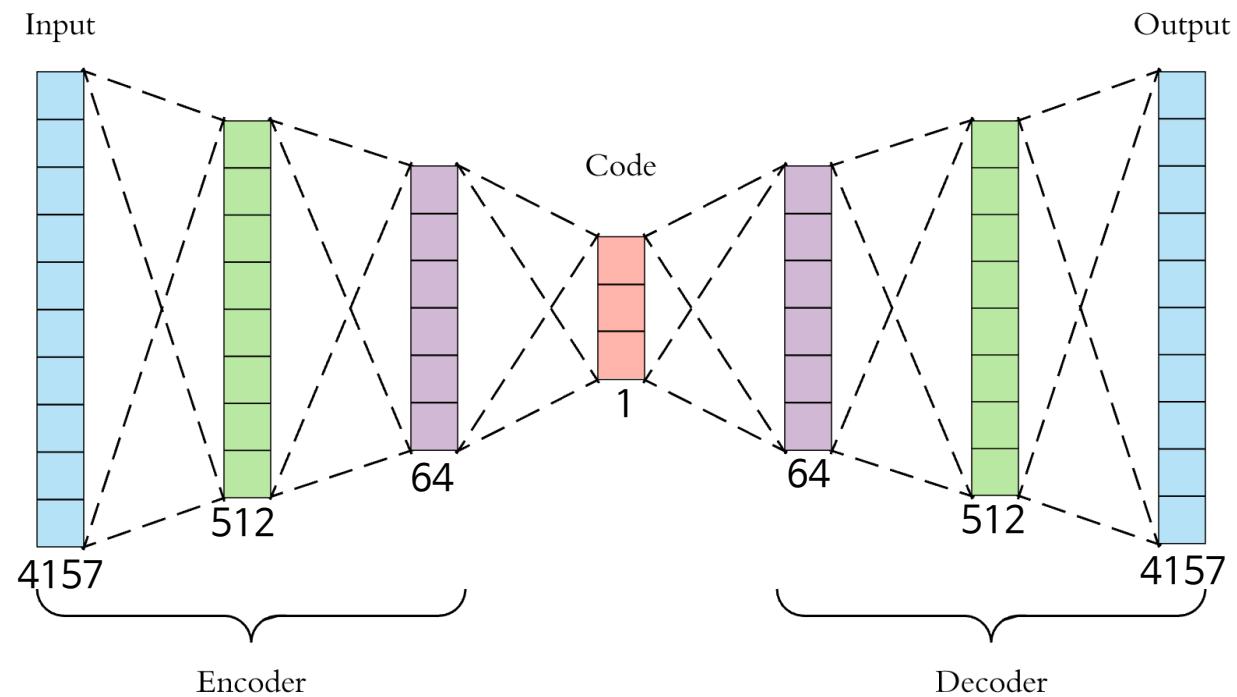
3 Feature engineering: description -> stacked auto-encoder

ReviewAutoencoder_state_dict.pth

```
class Autoencoder(nn.Module):
    def __init__(self):
        super(Autoencoder, self).__init__()

        self.encoder = nn.Sequential(
            nn.Linear(4157, 512),
            nn.ReLU(),
            nn.Linear(512, 64),
            nn.ReLU(),
            nn.Linear(64, 1)
        )
        self.decoder = nn.Sequential(
            nn.Linear(1, 64),
            nn.ReLU(),
            nn.Linear(64, 512),
            nn.ReLU(),
            nn.Linear(512, 4157)
        )

    def forward(self, x):
        encoded = self.encoder(x)
        decoded = self.decoder(encoded)
        return encoded, decoded
```



2. 프로젝트 수행 내용

2 Feature engineering: Categorical variable – issue of encoding

■ One hot encoding

Number of categories	
Categorical features	
country	42
province	422
region_1	1204
region_2	17
designation	35776
winery	15855
variety	697
taster_name	19

```
train_df.info()
```

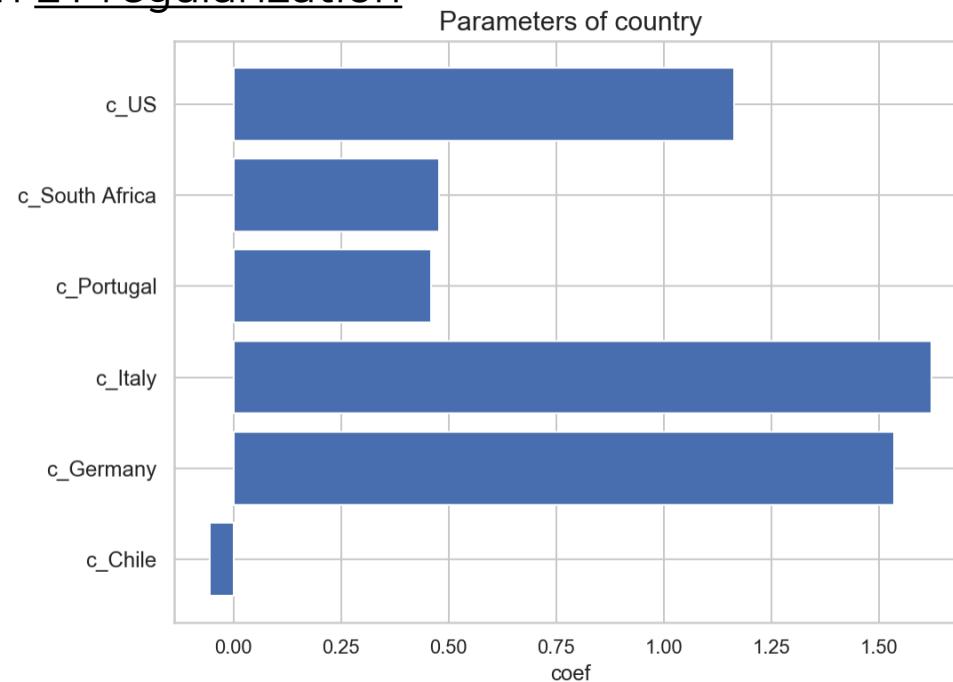
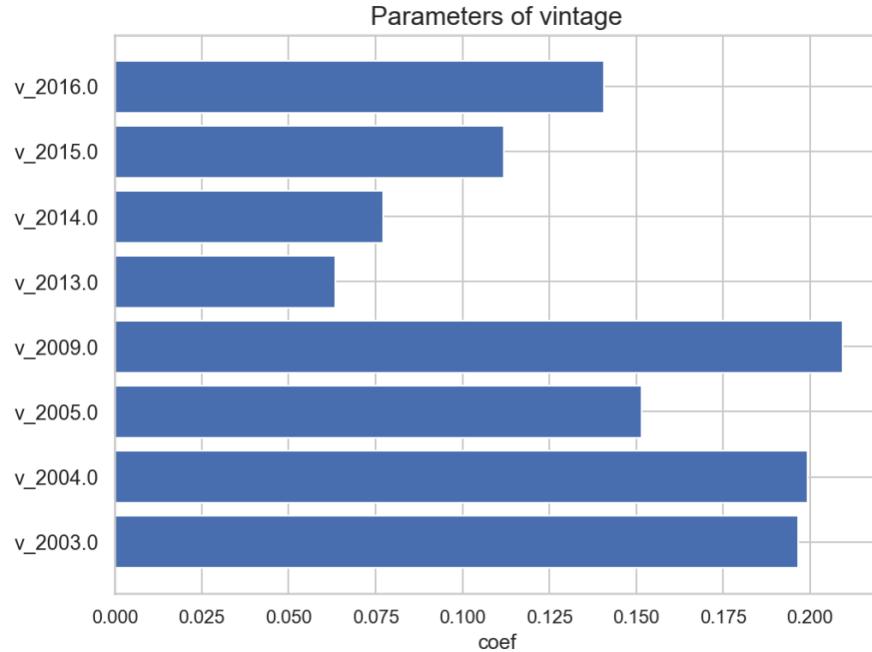
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 108984 entries, 0 to 111537
Columns: 40298 entries, points to age
dtypes: float64(3), uint8(40295)
memory usage: 4.1 GB
```

프로젝트 진행 초반부에 카테고리 변수의 수많은 카테고리를 one hot encoding한 결과, 데이터의 차원(columns)이 과도하게 높아지고 메모리 자원 낭비가 심해져 프로젝트 작업이 매우 비효율적으로 진행되는 문제가 발생했습니다.

2. 프로젝트 수행 내용

2 EDA: Categorical variable – issue of encoding

- One hot encoding -> Feature selection with L1 regularization



따라서 One hot encoding은 L1 regularization을 통해 가격에 실질적으로 영향을 미치는 세부 카테고리 변수가 어떤 것인지 확인하고 후술할 target encoding의 결과 향상 비교 용도로 활용하였습니다.

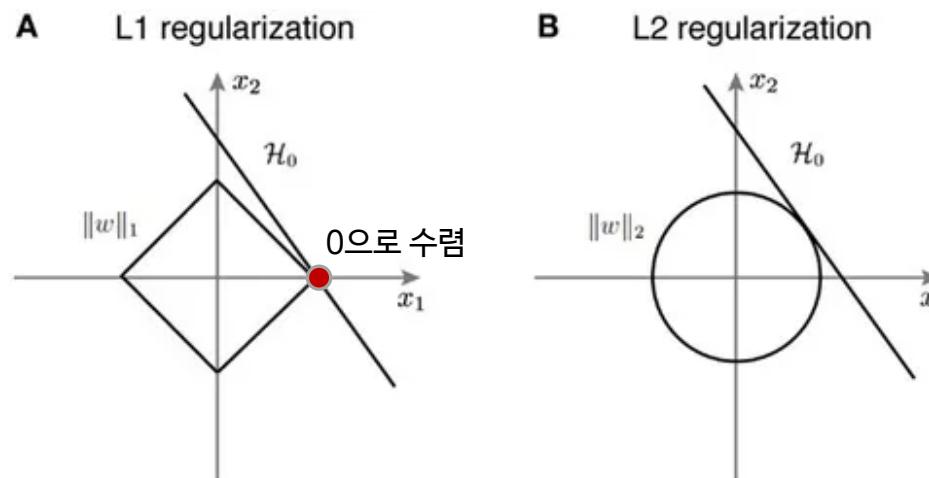
2. 프로젝트 수행 내용

2 EDA: Categorical variable – issue of encoding

■ Feature selection with L1 regularization

- L1 regularization은 가중치의 절대값의 합을 최소화하는 것을 추가적인 제약 조건으로 걸어서 입력된 독립변수 중 일부를 0으로 빠르게 수렴하도록 만들고 overfitting을 방지하는 효과가 있습니다.

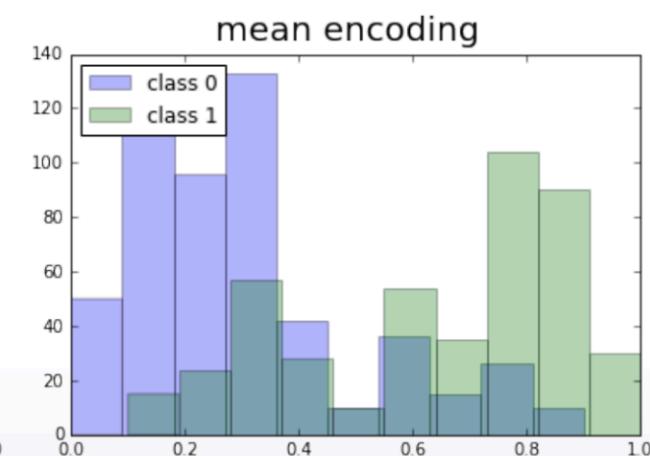
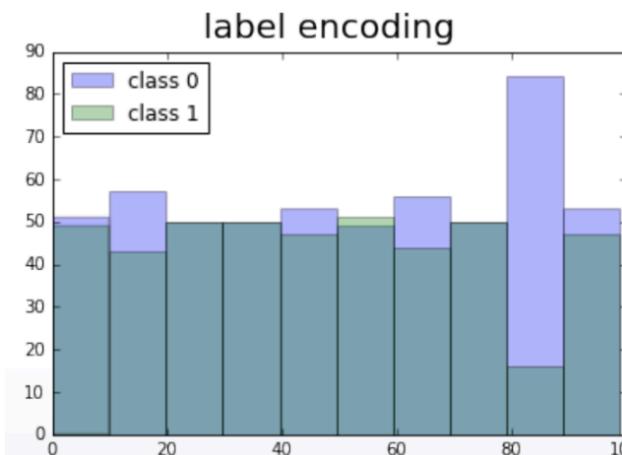
- 이 과정을 통해 성능을 저하하지 않으면서도 모델에 사용할 features의 개수를 최소화할 수 있는데, 주로 타겟에 대한 레버리지가 높으면서도 많은 데이터에서 등장한 독립변수가 선택됩니다.



2. 프로젝트 수행 내용

2 Feature engineering: Target encoding(=Mean encoding)

	feature	feature_label	feature_mean	target
0	Moscow	1	0.4	0
1	Moscow	1	0.4	1
2	Moscow	1	0.4	1
3	Moscow	1	0.4	0
4	Moscow	1	0.4	0
5	Tver	2	0.8	1
6	Tver	2	0.8	1
7	Tver	2	0.8	1
8	Tver	2	0.8	0
9	Klin	0	0.0	0
10	Klin	0	0.0	0
11	Tver	2	0.8	1



Target encoding은 categorical feature를 카테고리 단위로 groupby 하여 target에 대한 평균값으로 연산하는 인코딩 방식입니다. High cardinality categorical feature에 적합한 방식이며 One hot encoding과 label encoding과 비교했을 때 차원이 증가하는 문제를 방지하면서도 feature와 target의 상관관계를 반영할 수 있다는 장점이 있습니다.

03

프로젝트 결과

3. 프로젝트 결과

Metrics : RMSE(Root Mean Square Error)

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (Predicted_i - Actual_i)^2}{N}}$$

모델의 목표는 와인의 가격을 정확히 예측하는 것이므로 평균 제곱근 오차 RMSE score를 평가 척도 기준으로 합니다. 모델의 성능은 RMSE가 낮을수록 좋습니다.

3. 프로젝트 결과

1 Baseline : Linear regression(OLS): OHE with selected features(by L1 regularization)

```
OLS Regression Results
=====
Dep. Variable:                      y   R-squared:                 0.971
Model:                             OLS   Adj. R-squared:            0.971
Method:                            Least Squares   F-statistic:             2.805e+04
Date:                            Thu, 20 Jun 2019   Prob (F-statistic):        0.00
Time:                            04:45:58   Log-Likelihood:          -55184.
No. Observations:                  65390   AIC:                     1.105e+05
Df Residuals:                      65311   BIC:                     1.112e+05
Df Model:                           79
Covariance Type:                nonrobust
Omnibus:                          1502.633   Durbin-Watson:           1.980
Prob(Omnibus):                   0.000   Jarque-Bera (JB):       2003.329
Skew:                            0.283   Prob(JB):                  0.00
Kurtosis:                         3.644   Cond. No.:                 77.7
=====
```

RMSE score of test set: 21.346859478400802

cv scores: [21.31918813 21.04491947 21.44173443 21.27667076 20.95184523]

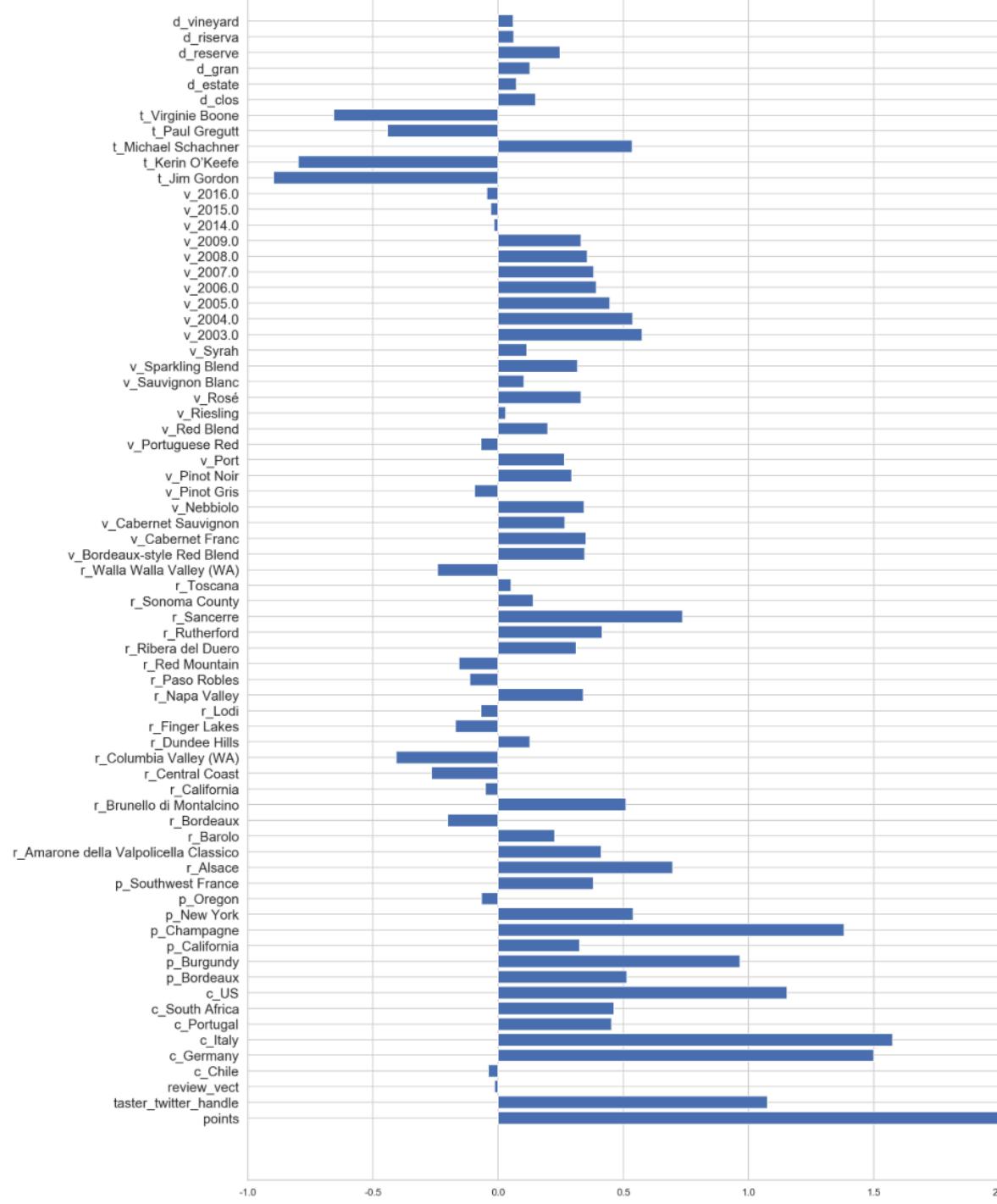
*단일계수 t 검정(H0: w = 0) 결과, 회귀계수에 대한 신뢰성이 떨어지는 features
'p_Port', 'p_Sicily & Sardinia', 'r_Mendoza', 'v_Pinot Grigio', 'v_Portuguese White', 'v_2013.0', 'd_cru', 'd_dry'

Performance:

- RMSE score: 21.35
- R-squared: 0.971

Cross validation(cv=5):

- 교차 검증 결과 test set의 RMSE 와 균등한 결과값들이 나왔으므로 overfitting 문제는 없습니다.



회귀계수 결과

*신뢰성 있는 회귀계수 분석을 위해서
단일계수 t 검정의 p-value가 5% 이상인 feature는 제외한 그래프입니다.

1. 가격에 대해 영향력이 가장 강한 요인은 points 입니다.
2. vintage는 2003년까지 시간을 거슬러 올라갈수록 가치가 상승합니다.
3. variety에서 백포도 품종(Sauvignon Blanc, Riesling, Pino Gris 등)에 비해서 적포도 품종(Cabernet Sauvignon, Syrah, Nebbiolo 등)이 가격에 대한 영향력이 더 강하고 가치가 높은 편입니다.
4. winery의 모든 값들이 feature selection 과정에서 탈락되었습니다
5. country 중 ‘미국, 남아프리카, 포르투갈, 이탈리아, 독일, 칠레’가 feature selection 결과로 선택되었습니다.

3. 프로젝트 결과

1 Baseline : Linear regression(OLS): Target encoding(=mean encoding)

OLS Regression Results						
Dep. Variable:	price	R-squared:	0.763			
Model:	OLS	Adj. R-squared:	0.763			
Method:	Least Squares	F-statistic:	1.912e+04			
Date:	Thu, 20 Jun 2019	Prob (F-statistic):	0.00			
Time:	04:46:31	Log-Likelihood:	-12173.			
No. Observations:	65390	AIC:	2.437e+04			
Df Residuals:	65378	BIC:	2.448e+04			
Df Model:	11					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Intercept	2.9434	0.004	700.043	0.000	2.935	2.952
points	0.7674	0.009	83.855	0.000	0.749	0.785
scale(country)	0.0035	0.002	1.655	0.098	-0.001	0.008
scale(designation)	0.1569	0.001	112.793	0.000	0.154	0.160
scale(province)	-0.0251	0.002	-13.805	0.000	-0.029	-0.022
scale(region_1)	0.0954	0.002	57.479	0.000	0.092	0.099
scale(taster_name)	-0.0153	0.002	-8.141	0.000	-0.019	-0.012
taster_twitter_handle	0.0072	0.003	2.463	0.014	0.001	0.013
scale(variety)	0.0550	0.001	37.434	0.000	0.052	0.058
scale(winery)	0.2577	0.002	150.787	0.000	0.254	0.261
scale(vintage)	0.0353	0.001	29.572	0.000	0.033	0.038
review_vect	0.0258	0.002	12.383	0.000	0.022	0.030
Omnibus:	3120.632	Durbin-Watson:	2.012			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5277.477			
Skew:	0.398	Prob(JB):	0.00			
Kurtosis:	4.142	Cond. No.	16.8			

Performance:

- RMSE score: 11.78
- R-squared: 0.763

앞의 One hot encoding 모델과 비교했을 때
R-squared는 약 0.21가 감소하지만, RMSE score가
9.57만큼 감소하는 의미 있는 성과가 있었습니다.

country의 경우 단일 계수 t검정에 대한 유의확률
(p-value)이 유의수준 5%를 넘기 때문에, 귀무가설
(H0: w = 0)을 채택하여 종속변수 price에 영향을 미
치지 않는 변수로 분석하였습니다.

3. 프로젝트 결과

1 Baseline : Linear regression(OLS): target encoding(=mean encoding)

	sum_sq	df	F	PR(>F)
points	597.532030	1.0	7031.664363	0.000000e+00
scale(country)	0.232651	1.0	2.737800	<u>9.800442e-02</u>
scale(designation)	1081.102725	1.0	12722.249380	0.000000e+00
scale(province)	16.193981	1.0	190.568256	2.748642e-43
scale(region_1)	280.753814	1.0	3303.867388	0.000000e+00
scale(taster_name)	5.631961	1.0	66.276047	<u>3.988293e-16</u>
taster_twitter_handle	0.515545	1.0	6.066848	<u>1.377681e-02</u>
scale(variety)	119.081683	1.0	1401.334803	1.787745e-303
scale(winery)	1932.111605	1.0	22736.790045	0.000000e+00
scale(vintage)	74.314374	1.0	874.520041	6.204393e-191
review_vect	13.030717	1.0	153.343454	3.529461e-35
Residual	5555.647577	65378.0	Nan	Nan

F 검정을 사용한 변수 중요도 분석

전체 모델과 각 변수를 하나씩 뺀 모델들의 성능을 비교하는 방식으로 독립 변수의 영향력을 측정합니다.

검정 결과 country, taster_name, taster_twitter_handle의 경우 중요도가 가장 낮은 변수로 분석되었습니다. 이 feature를 제외하여 모델링을 진행해도 성능의 큰 차이가 없을 것이라 가정하고 추가적으로 모델링을 진행해보았습니다.

3. 프로젝트 결과

1 Baseline : Linear regression(OLS): target encoding(=mean encoding)

OLS Regression Results						
Dep. Variable:	price	R-squared:	0.763			
Model:	OLS	Adj. R-squared:	0.762			
Method:	Least Squares	F-statistic:	2.624e+04			
Date:	Thu, 20 Jun 2019	Prob (F-statistic):	0.00			
Time:	04:46:35	Log-Likelihood:	-12220.			
No. Observations:	65390	AIC:	2.446e+04			
Df Residuals:	65381	BIC:	2.454e+04			
Df Model:	8					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Intercept	2.9528	0.004	755.368	0.000	2.945	2.960
points	0.7684	0.009	85.219	0.000	0.751	0.786
scale(province)	-0.0309	0.001	-20.850	0.000	-0.034	-0.028
scale(designation)	0.1569	0.001	112.790	0.000	0.154	0.160
scale(region_1)	0.0923	0.002	56.939	0.000	0.089	0.096
scale(variety)	0.0558	0.001	38.089	0.000	0.053	0.059
scale(winery)	0.2575	0.002	150.622	0.000	0.254	0.261
scale(vintage)	0.0351	0.001	29.699	0.000	0.033	0.037
scale(review_vect)	0.0157	0.001	12.487	0.000	0.013	0.018
Omnibus:	3193.235	Durbin-Watson:	2.012			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5418.680			
Skew:	0.404	Prob(JB):	0.00			
Kurtosis:	4.155	Cond. No.	15.0			
cv scores:	[11.54175747 11.40245485 11.60910512 11.86176653 11.65547524]					

Performance:

- RMSE score: 11.79
- R-squared: 0.763

All features Selected features

R_squared	0.762846	0.762508
RMSE	11.775978	11.797275

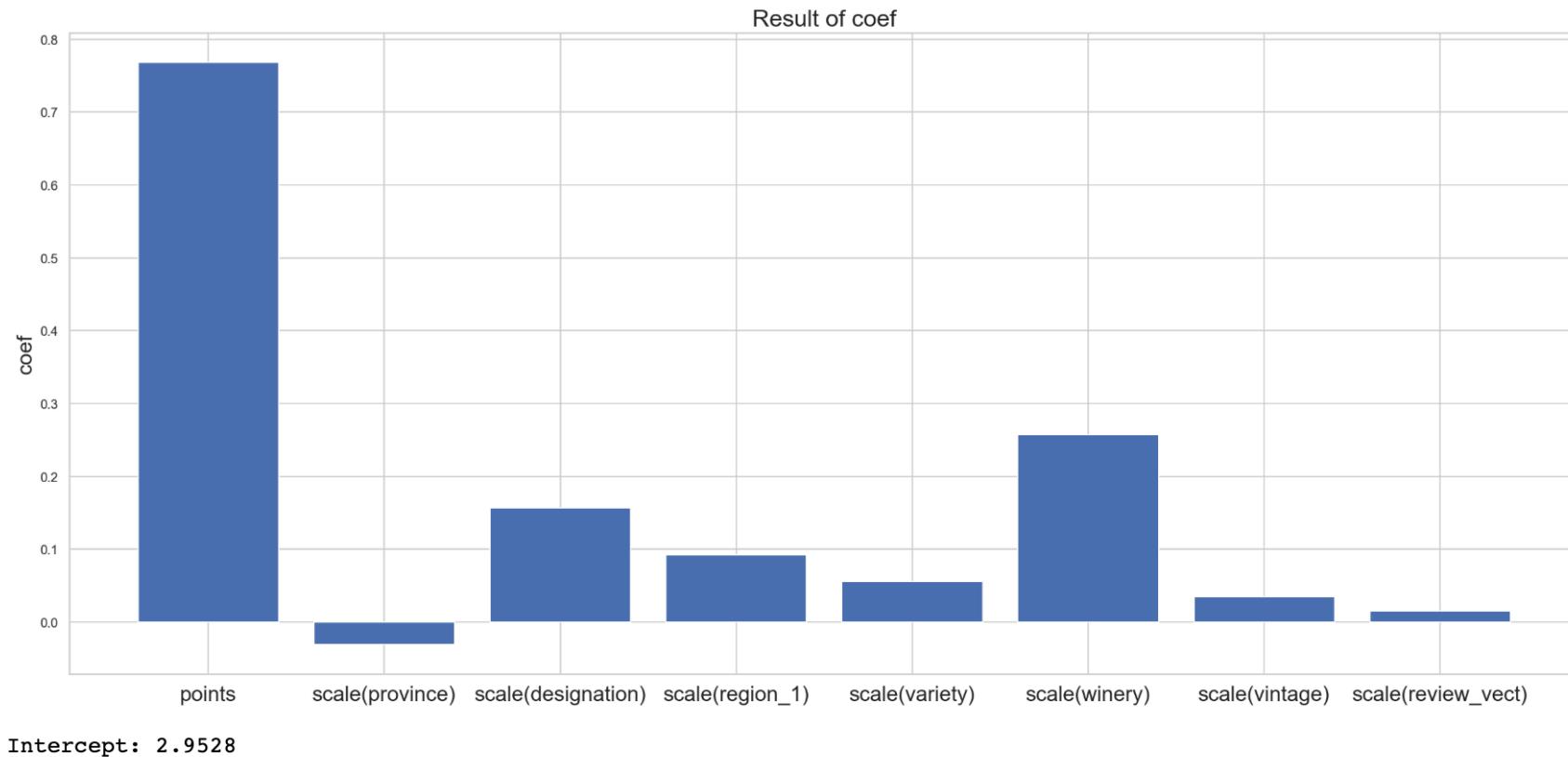
앞의 F 검정에서 중요도가 낮은 변수로 분석된 country, taster_name, taster_twitter_handle을 제외해도 성능의 차이가 존재하지 않는 것으로 확인되었습니다.

Cross validation(cv=5):

- 교차 검증 결과 역시 균등한 값으로 overfitting 문제는 없습니다.

3. 프로젝트 결과

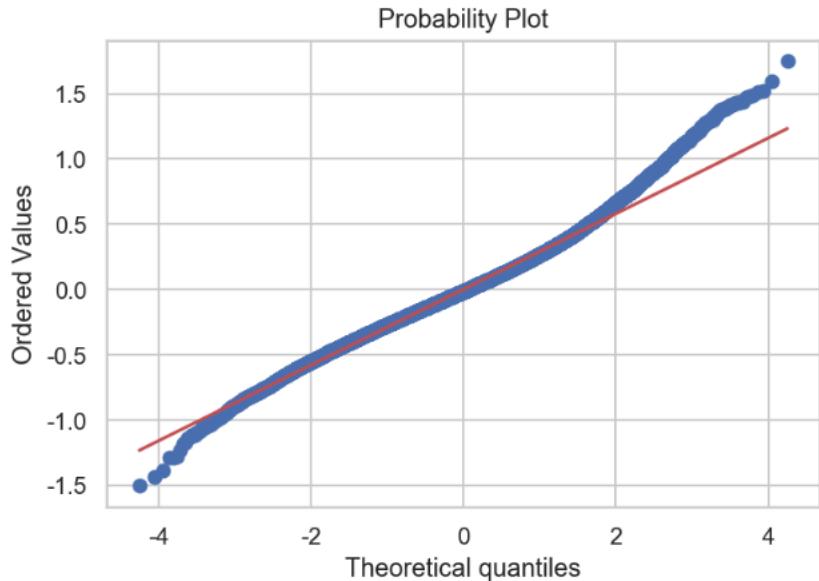
1 Baseline : Linear regression(OLS): target encoding(=mean encoding)



3. 프로젝트 결과

1 Baseline : Linear regression(OLS): target encoding(=mean encoding)

- 잔차 정규성 검사(Q-Q plot, D'Agostino's K-squared test):
 - 회귀분석에 사용된 데이터가 회귀분석에 사용된 모델 가정을 만족하고 있는지 확인합니다.



양쪽 꼬리가 명확하게 중심선에서 벗어나 있기 때문에 잔차에 정규성이 있다고 판단하기 어렵습니다.

```
normaltest(lm_2_result.resid)
```

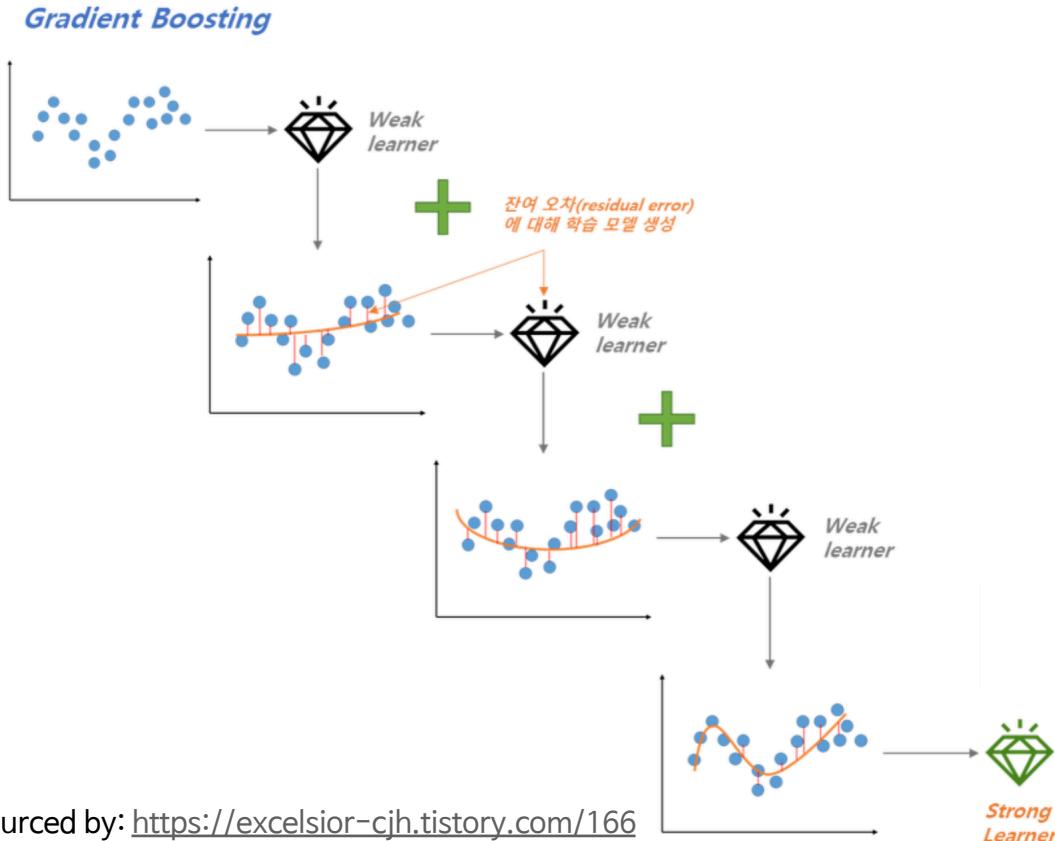
```
NormaltestResult(statistic=3193.2352819131256, pvalue=0.0)
```

검정 귀무가설(H0): 잔차가 정규성을 지닌다.

검정 결과, p-value < 0.01로 귀무가설이 기각되어 잔차가 정규성을 가지고 있지 않습니다. 따라서 회귀분석에 사용된 데이터가 회귀분석에 사용된 모델의 가정을 제대로 만족하고 있다고 할 수 없습니다.

3. 프로젝트 결과

2 Final: LightGBM Gradient boosting regressor



Sourced by: <https://excelsior-cjh.tistory.com/166>

1. 회귀분석 모델 진단 결과, 데이터가 모델의 가정을 제대로 만족하고 있지 않은 문제
2. 복수의 모델을 결합하는 원리로 단일 모델에 비해 모델의 성능을 더욱 끌어올릴 수 있는 Ensemble model의 강력함

위의 두 가지 사항을 고려하여 최종 모델을 LightGBM 패키지의 Gradient Boosting Regressor로 선정하였습니다.

Final performance:

- RMSE score of validation set: 10.40
- RMSE score of test set: 10.48
- 선형 회귀모델에 비해 약 1.3 정도의 수치적 향상이 있습니다.

validation set과 test set의 score에 차이가 없기 때문에 overfitting 문제는 없다고 진단하였습니다.

04

한계 및 보완 방향

4. 한계 및 보완 방향

1 적절하지 못한 누락 데이터 처리 방법

누락된 데이터 처리에 대한 근거가 부족하다. 일괄적으로 'unknown'으로 규정하여 처리하는 방식은 쉽게 납득하기 어렵다.

2 Feature 사이의 상호관계 반영 X

근본적으로 One hot encoding은 feature 사이의 관계가 전혀 반영되지 않는 문제가 있다. country, province, region_1 등의 지리적인 feature는 서로 포함관계가 있는 특징이 있지만, 프로젝트 과정에서 이러한 특성이 고려된 전처리가 이루어지지 않았다. 이러한 feature간의 상호관계가 무시되는 방법은 충분히 비판의 여지가 있다.

3 미흡한 feature 분석 과정과 시각화 기술

프로젝트 기한과 모델링 결과 향상에 쓸겨 Chi-square test, t-test 검정 등의 방법을 활용한 통계적인 feature 분석 과정이 미흡했다. 분석 과정에 꼭 필요한 시각화 스킬에 대해서도 스스로 한계를 많이 느꼈다.

감사합니다