A glass of ML

Wine pricing interpretation and prediction

Chechkov Eugene

Main task

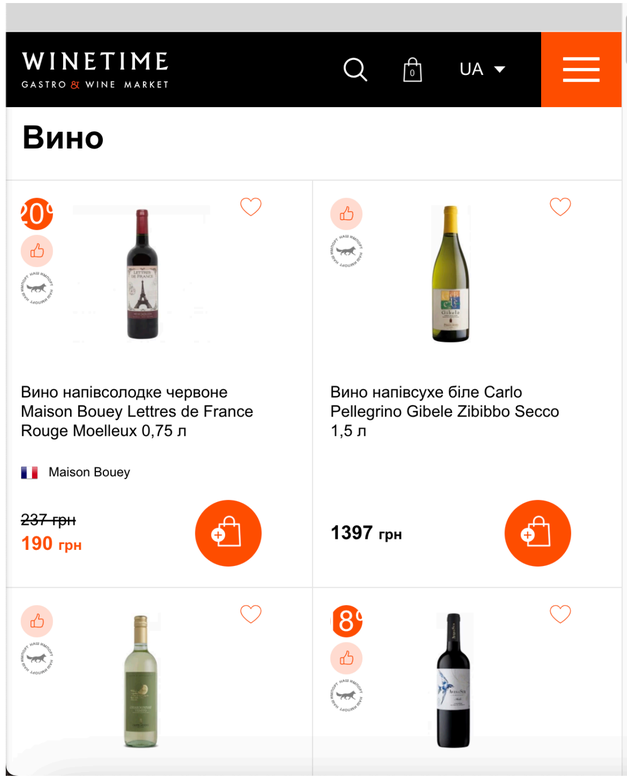
Can we understand by the data how wine prices are being formed?

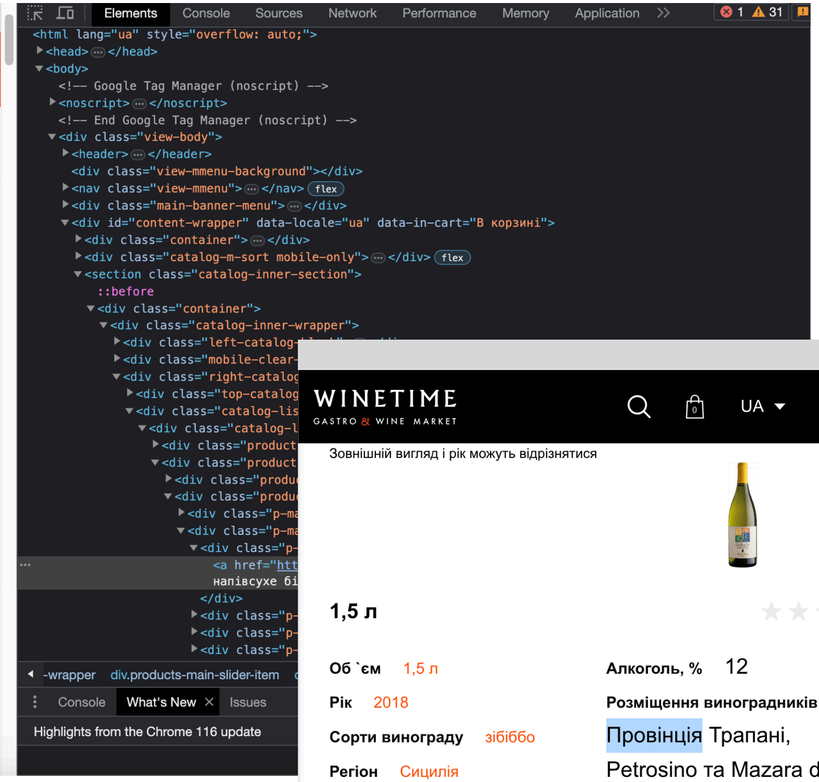
- Which features have most impact on wine price
- Which features have least impact on wine price
- Predict wine prices and whether the accuracy will be sufficient

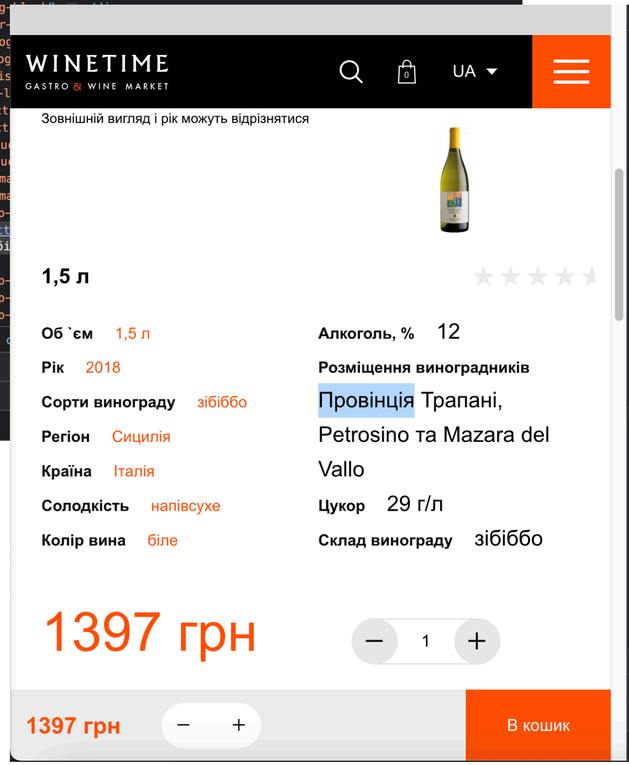
Who could be interested

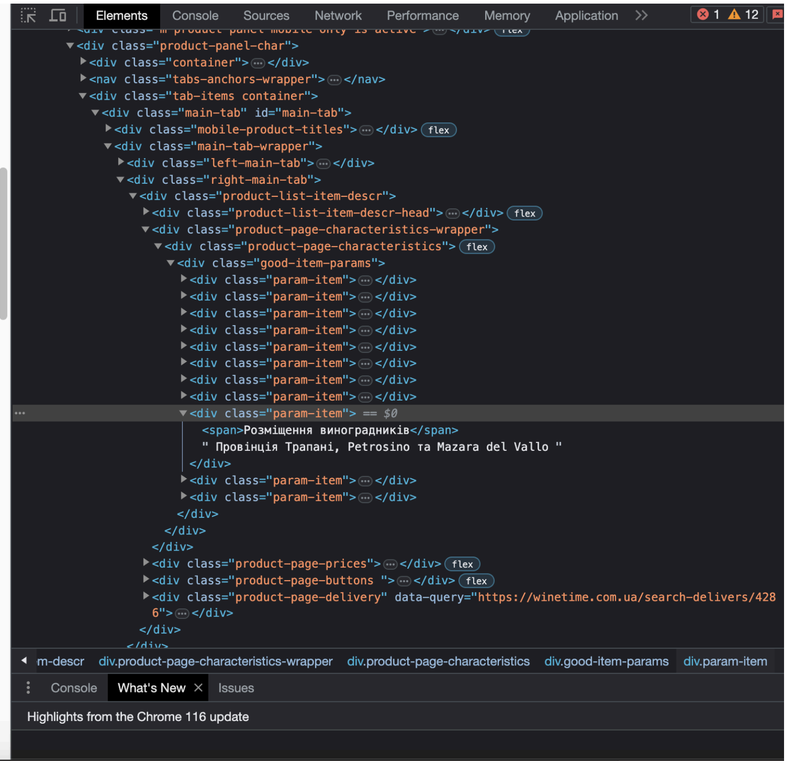
- Retailers
- Wine producers

Let’s parse a retailer









WINETIME
GASTRO & WINE MARKET

What do we have?

Quick look on data

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 4865 entries, 0 to 4864
Data columns (total 31 columns):
 #    Column                 Non-Null Count  Dtype
---   ------                 --------------  -----
 0    Title                  4865 non-null   object
 1    Producer               4304 non-null   object
 2    Vendor_code            4865 non-null   object
 3    Price                  4065 non-null   float64
 4    Temperature            3911 non-null   object
 5    Grape                  4523 non-null   object
 6    Technology             2590 non-null   object
 7    Volume                 4827 non-null   float64
 8    Year                   3736 non-null   float64
 9    Brand                  4304 non-null   object
 10   Region                 3486 non-null   object
 11   Country                4641 non-null   object
 12   Sweetness              4762 non-null   object
 13   Type                   4826 non-null   object
 14   Color                  4824 non-null   object
 15   Soil                   2316 non-null   object
 16   Serve_with             3860 non-null   object
 17   Classification         2731 non-null   object
 18   Vineyards_placement    2398 non-null   object
 19   Endurance              2685 non-null   object
 20   Grapes_composition     4523 non-null   object
 21   Sweet                  4110 non-null   float64
 22   Alcohol                4649 non-null   float64
 23   Harvest                2171 non-null   object
 24   Additional_color       2982 non-null   object
 25   Aroma                  4333 non-null   object
 26   Taste                  4351 non-null   object
 27   Interesting            279 non-null    object
 28   Style                  71 non-null     object
 29   Potential              401 non-null    object
 30   Degustations           4 non-null      object
dtypes: float64(5), object(26)
memory usage: 1.2+ MB
```

Preprocessing. Preprocessing? Preprocessing!

Data example

	Title	Producer	Vendor_code	Price	Temperature	Grape	Technology	Volume	Year	Brand	...	Sweet	Alcohol	Harvest
0	Вино сухе біле Dominio de Punctum Viento Alise...	Dominio de Punctum	15055373	439 грн	8°C	віоньє	Вино виробляють за принципами «живого» господа...	0,75 л	2022.0	Dominio de Punctum	...	1 г/л	13.0	Виноград збирають вночі.
1	Вино напівсолодке червоне Castelnuovo Vino Ros...	Castelnuovo	10369400	217 грн	14-15°C	корвіна, рондінелла	NaN	0,75 л	NaN	Castelnuovo	...	35 г/л	11.0	NaN
2	Вино сухе червоне Vintae El Picaro 0,75 л	Vintae	15426282	557 грн	16-18°C	тінта де торо	Біодинамічний спосіб ведення господарства. Фер...	0,75 л	2022.0	Vintae	...	2.4 г/л	14.5	Добірний урожай з 90-100-річних лоз, зібраний ...
3	Вино сухе біле Dominio de Punctum Viento Alise...	Dominio de Punctum	15055373	439 грн	8°C	віоньє	Вино виробляють за принципами «живого» господа...	0,75 л	2022.0	Dominio de Punctum	...	1 г/л	13.0	Виноград збирають вночі.

Producer	Vendor_code	Price	Temperature	Grape	Technology	Volume	Year	Brand	...	Sweet	Alcohol	Harvest	Additional_color
Dominio de Punctum	15055373	439 грн	8°C	віоньє	Вино виробляють за принципами «живого» господа...	0,75 л	2022.0	Dominio de Punctum	...	1 г/л	13.0	Виноград збирають вночі.	Яскравий золотистий зі відблисками лайму.
Castelnuovo	10369400	217 грн	14-15°C	корвіна, рондінелла	NaN	0,75 л	NaN	Castelnuovo	...	35 г/л	11.0	NaN	Багатий рубіновий.
Vintae	15426282	557 грн	16-18°C	тінта де торо	Біодинамічний спосіб ведення господарства. Фер...	0,75 л	2022.0	Vintae	...	2.4 г/л	14.5	Добірний урожай з 90-100-річних лоз, зібраний ...	Насичений, глибокий рубіновий.
Dominio de Punctum	15055373	439 грн	8°C	віоньє	Вино виробляють за принципами «живого» господа...	0,75 л	2022.0	Dominio de Punctum	...	1 г/л	13.0	Виноград збирають вночі.	Яскравий золотистий зі відблисками лайму.

Aroma	Taste	Interesting	Style	Potential	Degustations
Насичений із нотами запашних квітів.	Збалансований з нотами магнолії і персикового ...	NaN	NaN	NaN	NaN
Освіжаючий букет з нотками полуниці та смородини.	Вишуканий, солодкий смак з привабливим освіжаю...	NaN	NaN	NaN	NaN
Інтенсивний соковитий аромат темних ягід (чорн...	Теплий, соковитий смак, в якому домінують стиг...	На етикетках вин Matsu зображені фотографії ви...	NaN	NaN	NaN
Насичений із нотами запашних квітів.	Збалансований з нотами магнолії і персикового ...	NaN	NaN	NaN	NaN

Feature engineering

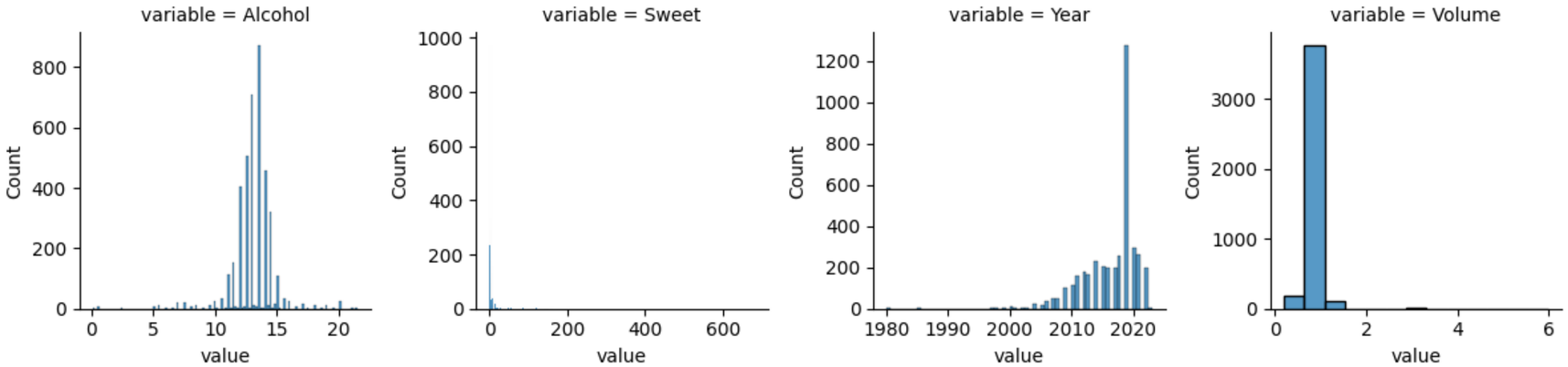
Classification

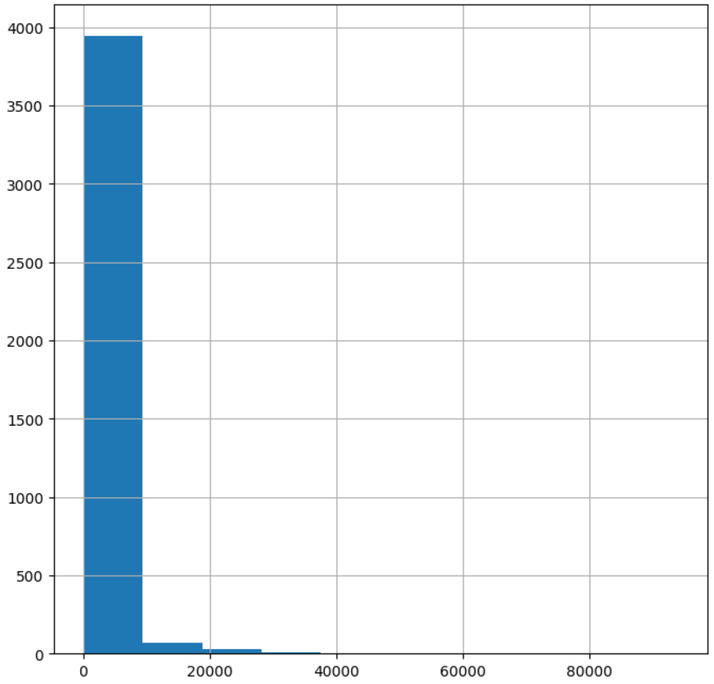
- Data for a couple of countries
- Countries classifications are different
- Spelling is different
- Result - binned into 3 categories : Regional wines, table wines, best wines

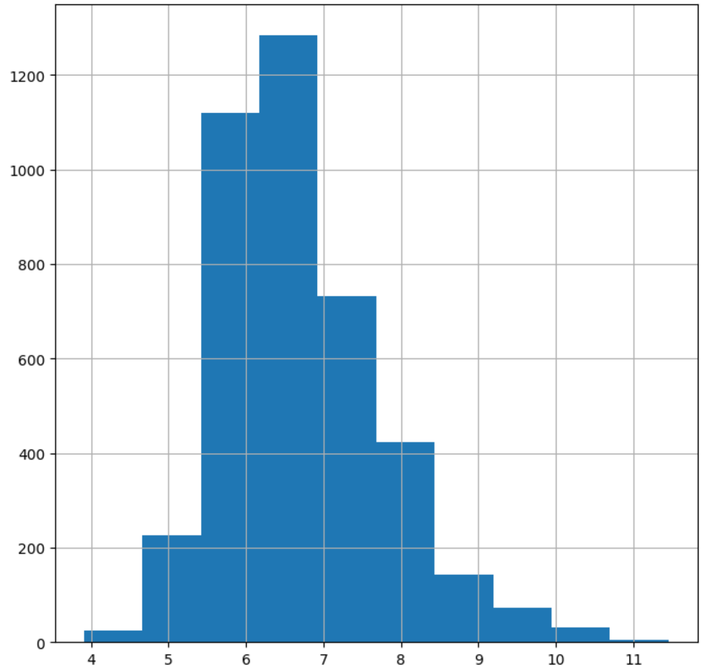
```
df['Classification'].unique()
```

```
array(['Indicazione Geografica Tipica (IGT)', nan, 'Toro DO', 'Kurant',
       'Denominazione di Origine Controllata (DOC)', 'Reserve',
       "Appellation d'Origine Controlee (AOC)", 'Vin de Table (VDT)',
       'Denominazione di Origine Controllata e Garantita (DOCG)',
       'Denominacion de Origen Calificada', 'Denominacion de Origen (DO)',
       'D.O. Carinena', 'KOP', 'Denominazione Di Origine Controllata',
       'Denomination of Origin', 'DOP', 'Appellation Bordeaux Contrelee',
       'витримане', 'Denominacion de Origen (D.O.)', 'Vin de Pays (VdP)',
       'DOCa Rioja', 'Indicazione Geografica Protette', 'DOC',
       'Appellation Alsace Controlee', 'Denominacao de Origine Protegida',
       'Qualitatswein',
       'Vino a Denominazione di Origine Controllata e Garantita (DOCG)',
       'DOC Douro', 'Denominazione Di Origine Controllata E Garantina',
       'D.O.C.G', 'IGP', 'Indication Geographique Protegee (IGP)',
       "Appellation d'Origine Protegee", 'Barolo D.O.C.G',
       'Appellation Medoc Controlee',
       'Appellation Saint-Emilion Grand Cru Controlle',
       'Indication geographique protegee (I.G.P.)',
       "Appellation d'Origine Protegee (AOP)", "Vin de Pays d'Oc",
       'Appellation Bordeaux Controlee', 'Barbaresco D.O.C.G',
       'Qualitatswein trocken', 'Appellation Cahors Controlee', 'IGT',
       'Appellation Medoc Contrelee',
       'Indication Geographique Protegee\xa0(IGP)',
       'Indicazione Geografica Tipica (IGT).', 'Eiswein',
       'Denominacao de Origem Controlada (DOC)',
       'Appellation Pomerol Contrelee', 'ординарне столове',
       'Denominazione di Origine Protetta (DOP)',
       'Vino ad Indicazione Geografica Protetta, Terre Siciliane',
```

Numerical features and target variable



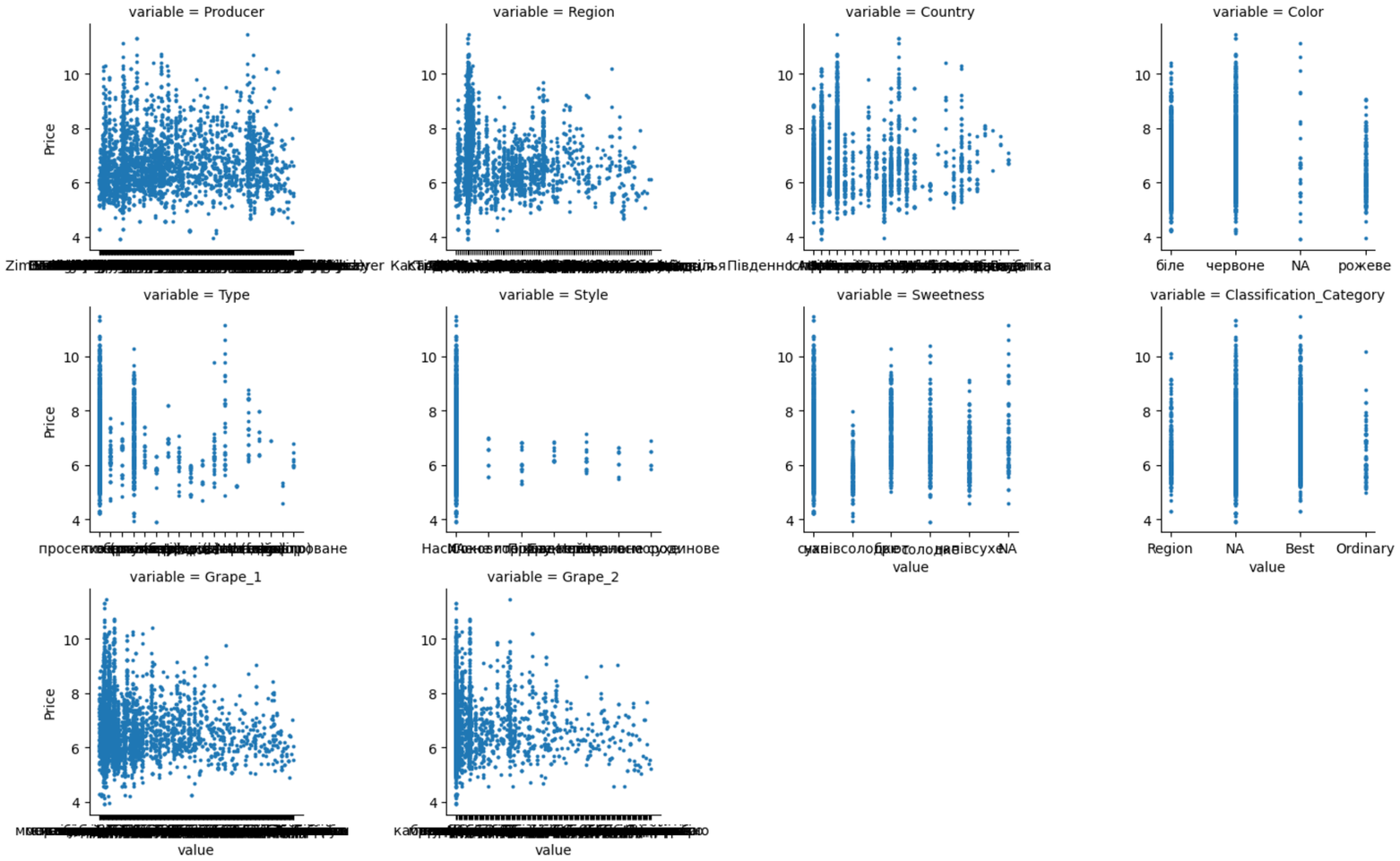




Categorical features

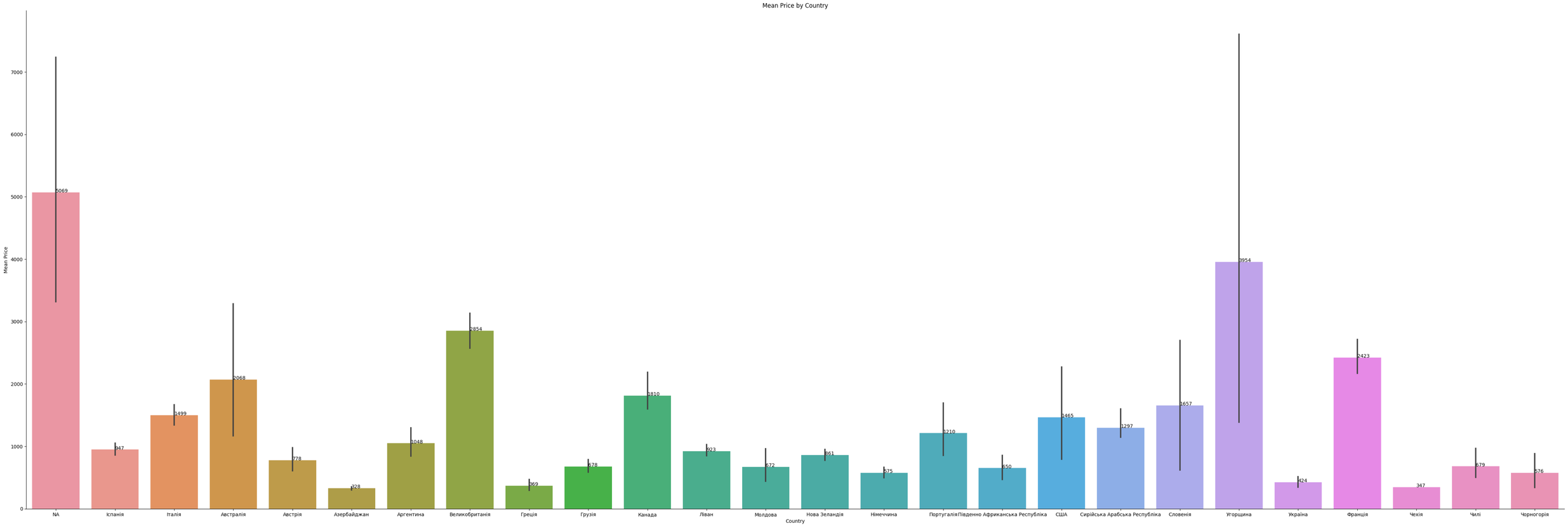
Main problem

- A lot… A LOT of unique categories. That could be a problem



Another problem

Some countries are absent



Let’s test some models

- Naive approach (just take average value
- Lasso/Elastic regression
- XGBoost

And the winner is definetely: dmlc XGBoost

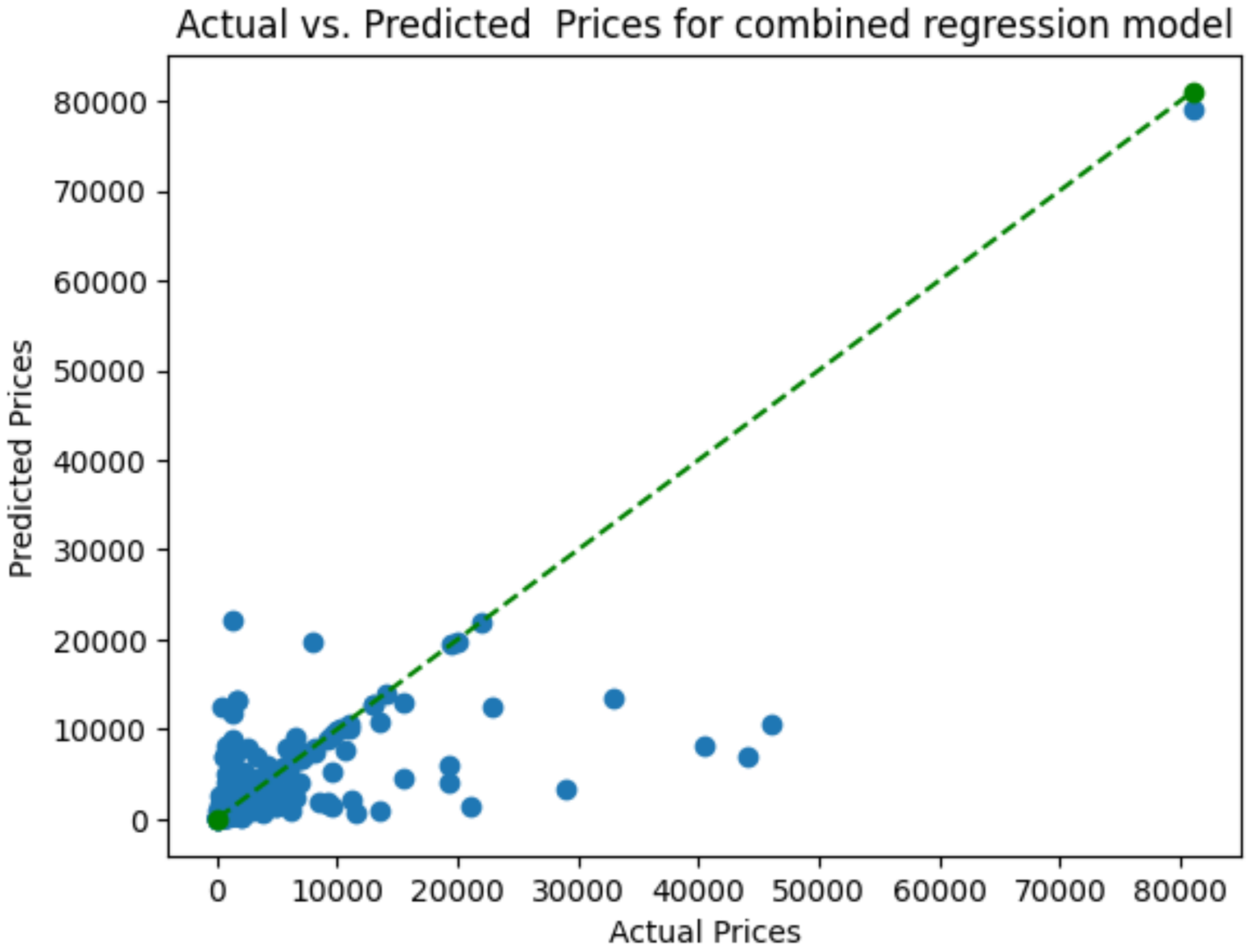
Results

	error_type	naive	lasso_cv	elastic_cv	xgb_regressor
0	MAE	1316.632166	1063.017128	1075.738863	663.959524
1	RMSE	4482.624008	4169.771960	4223.256121	2802.594270
2	R2	-0.040911	0.099314	0.076060	0.593117
3	MAPE	93.010235	56.099967	56.217935	39.029760

Predicting

Main thoughts

- Not so good especially for the outliers by the price



Features interpretation

Main thoughts

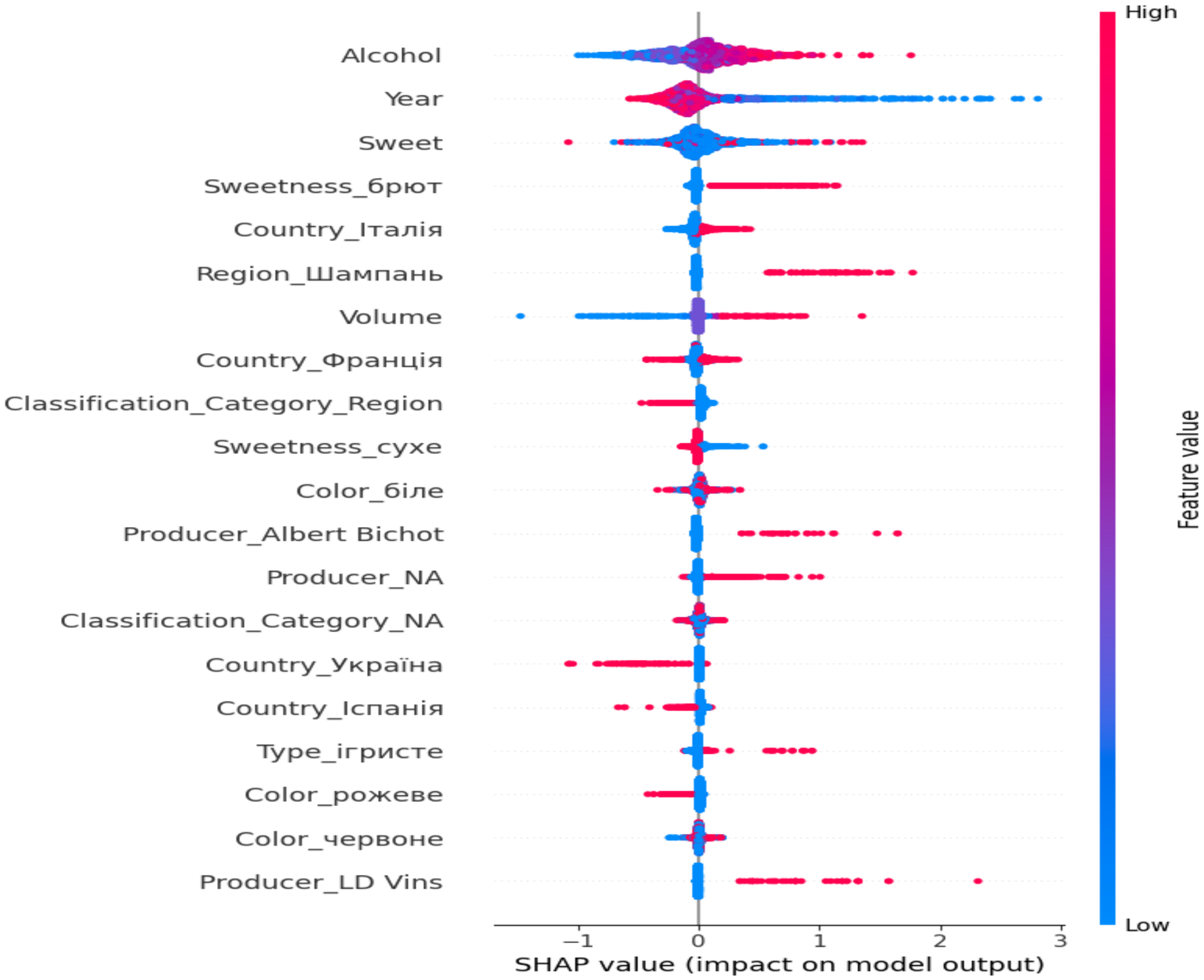
- Champaigns and brut wines features have the biggest impact
- Most of impact features related with country/region or producer or wine type/ sweetness
- Alcohol, color, category have no impact
- A lot of features with small impact

Weight	Feature
0.0888	Region_Шампань
0.0350	Sweetness_брют
0.0275	Country_Україна
0.0184	Producer_LD Vins
0.0146	Country_Угорщина
0.0145	Producer_E.Guigal
0.0124	Grape_2_ віоньє
0.0118	Producer_Albert Bichot
0.0110	Type_ігристе
0.0108	Sweetness_солодке
0.0101	Region_Тоскана
0.0099	Producer_Cristom Vineyards
0.0098	Producer_Georg Breuer Weingut
0.0092	Producer_Les Grands Chais
0.0090	Region_Брда
0.0088	Producer_Bodega Chacra
0.0086	Country_Південно Африканська Республіка
0.0081	Producer_Maison Louis Latour
0.0079	Country_Італія
0.0078	Color_червоне
	... 1064 more ...

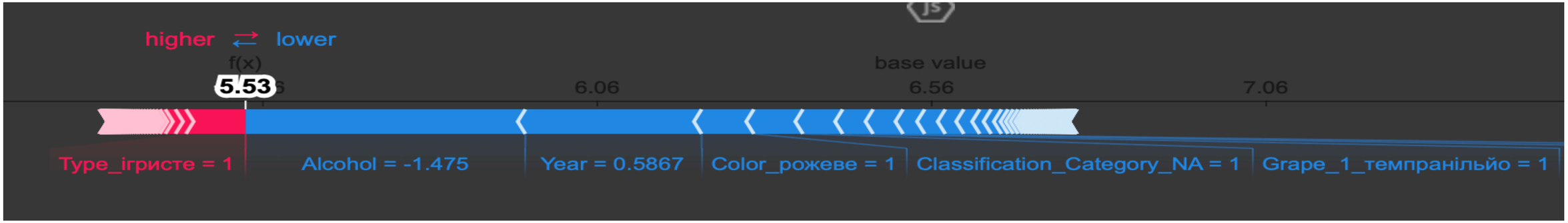
Predicting

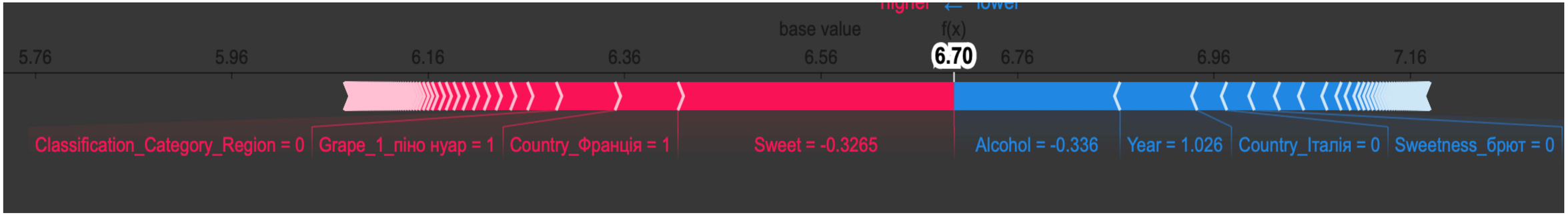
Main thoughts

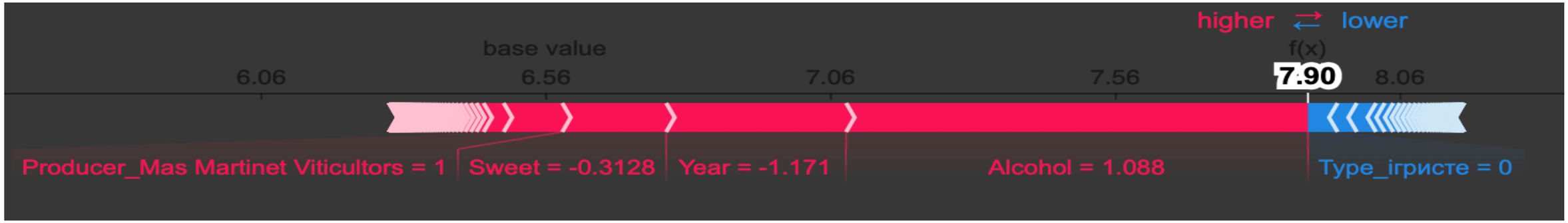
- SHAP Visualisation



SHAP


higher ⇄ lower
f(x)
5.53
base value
6.56
6.06
7.06
Type_ігристе = 1
Alcohol = -1.475
Year = 0.5867
Color_рожеве = 1
Classification_Category_NA = 1
Grape_1_темпранільйо = 1


base value
f(x)
6.70
5.76
5.96
6.16
6.36
6.56
6.76
6.96
7.16
Classification_Category_Region = 0
Grape_1_піно нуар = 1
Country_Франція = 1
Sweet = -0.3265
Alcohol = -0.336
Year = 1.026
Country_Італія = 0
Sweetness_брют = 0


higher ⇄ lower
base value
f(x)
7.90
6.06
6.56
7.06
7.56
8.06
Producer_Mas Martinet Viticultors = 1
Sweet = -0.3128
Year = -1.171
Alcohol = 1.088
Type_ігристе = 0

Eli5

Contribution?	Feature
+0.074	Sweetness_напівсолодке
+0.059	Country_Україна
+0.056	Country_Іспанія
	... 220 more positive ...
	... 227 more negative ...
-0.031	Region_Шампань
-0.054	Color_рожеве
-0.069	Color_біле
-0.095	Sweetness_брют
-0.101	Classification_Category_NA
-0.191	Year
-0.658	Alcohol

Contribution?	Feature
+0.340	Sweet
+0.052	Producer_Vignerons Catalans
+0.050	Country_Україна
	... 243 more positive ...
	... 194 more negative ...
-0.037	Color_червоне
-0.046	Region_Шампань
-0.053	Grape_1_шардоне
-0.056	Grape_2_NA
-0.072	Producer_Badet Clement
-0.117	Sweetness_брют
-0.182	Alcohol

Contribution?	Feature
+0.781	Year
+0.391	Alcohol
+0.077	Sweet
+0.076	Producer_Mas Martinet Viticultors
+0.068	Country_Іспанія
+0.066	Country_Чилі
	... 237 more positive ...
	... 199 more negative ...
-0.053	Volume
-0.057	Type_ігристе
-0.066	Grape_1_каберне совіньйон
-0.088	Producer_Maison Louis Latour

Next steps

What can be improved

- Manually fill unfilled data
- Use more data (Maybe use a couple of retailers)
- Try add textual features
- Add more features (pH levels, ratings e.t.c)

That’s all and thank you