Wine, Rating & Price.

| Numan SAHNOU
| Matthieu ECCHER

Summary

Data exploratory

| Interest of the project

| Data wrangling / cleaning

| Exploratory data analysis (EDA)

Machine learning

| Purpose

| Features selection

| Test set split for prediction

| Model selection

Data exploratory

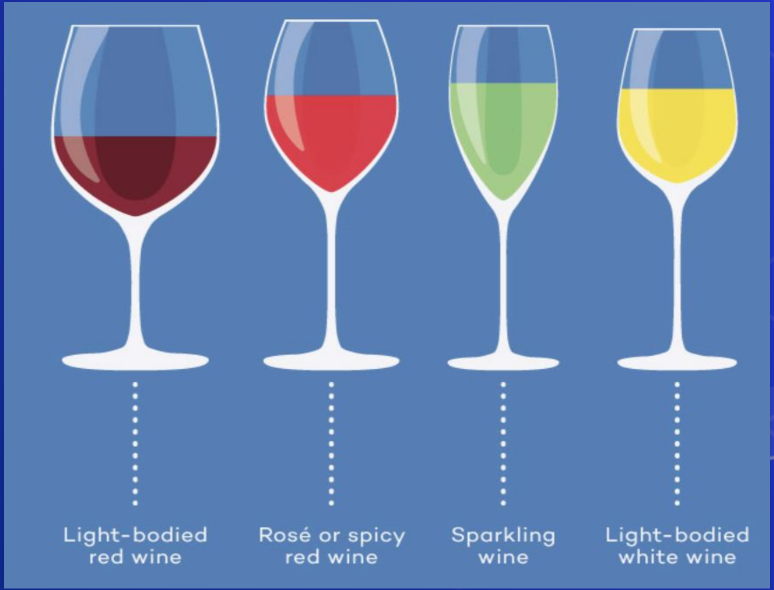
| Interest of the project

| Data wrangling / cleaning

| Exploratory data analysis (EDA)

Interest of the project

Predict wine rating from features



- 40 000
- 3 € to 3000€

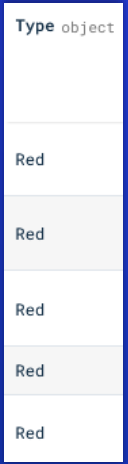
Interest of the project

The Vivino dataset

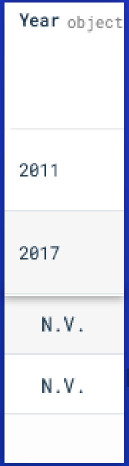
Features

| Name of the bottle
| Country of provenance
| Region in the country
| Winery in the region
| Rating of the bottle 0 to 5 step 0.1
| Number of Ratings : the number of people which give a grade to the bottle
| Price of 1 bottle
| Year of production
| Type of the wine (Red, White, Rosé, Sparkling)

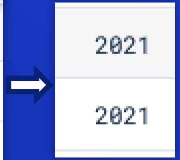
Data wrangling / cleaning



Type object
Red
Red
Red
Red
Red



Year object
2011
2017
N.V.
N.V.



2021
2021

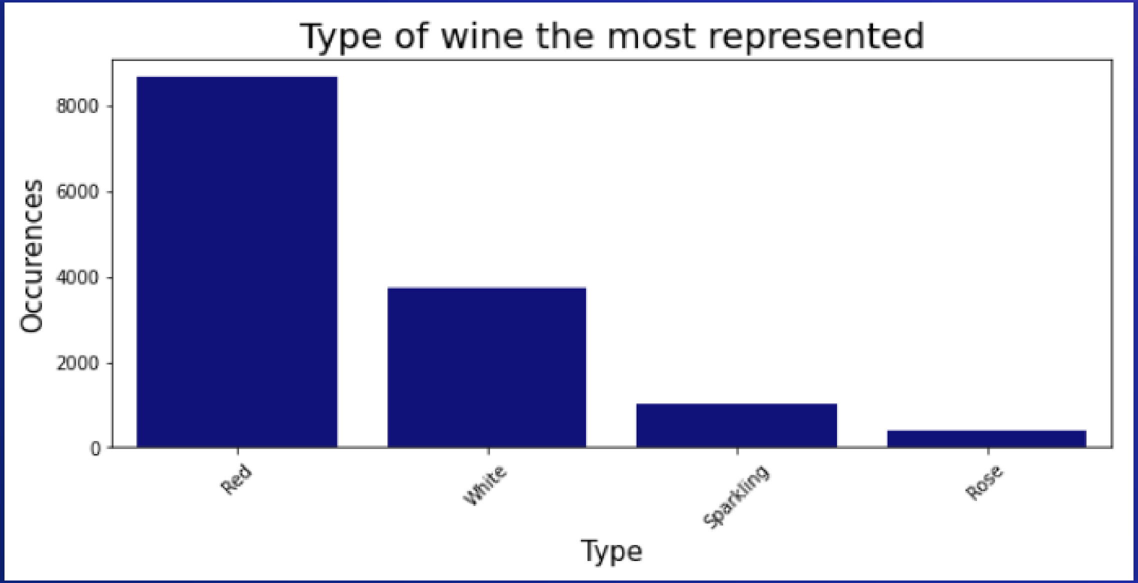
Column created ✓

Encode variables ✓

Features

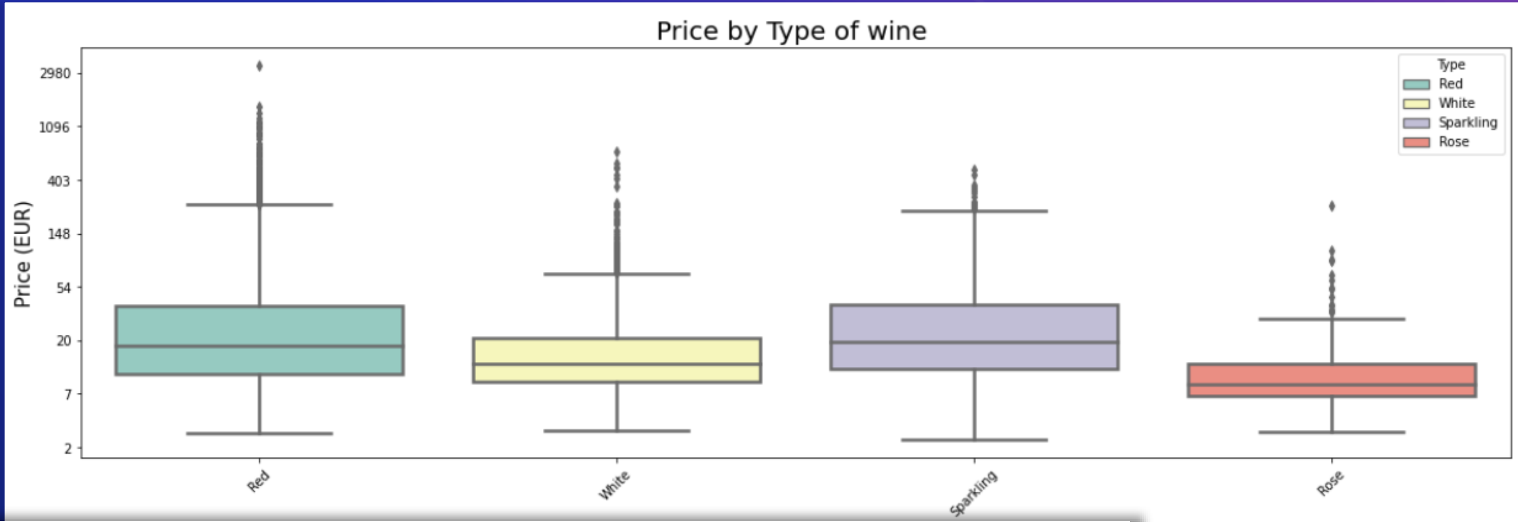
| Name
| Country
| Region
| Winery
| Rating
| NumberOfRatings
| Year
| Type

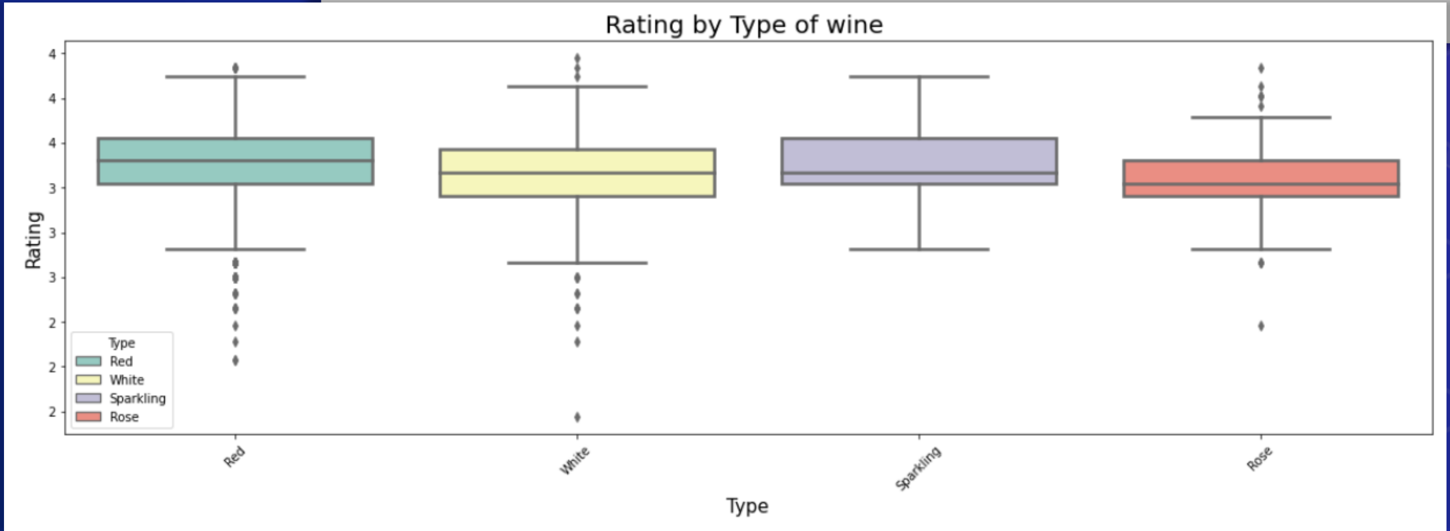
Exploratory data analysis



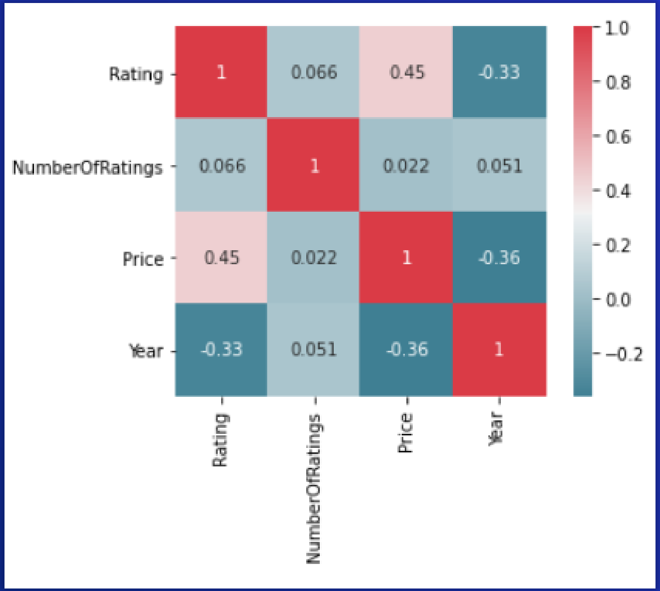
13 834 rows

Boxplot

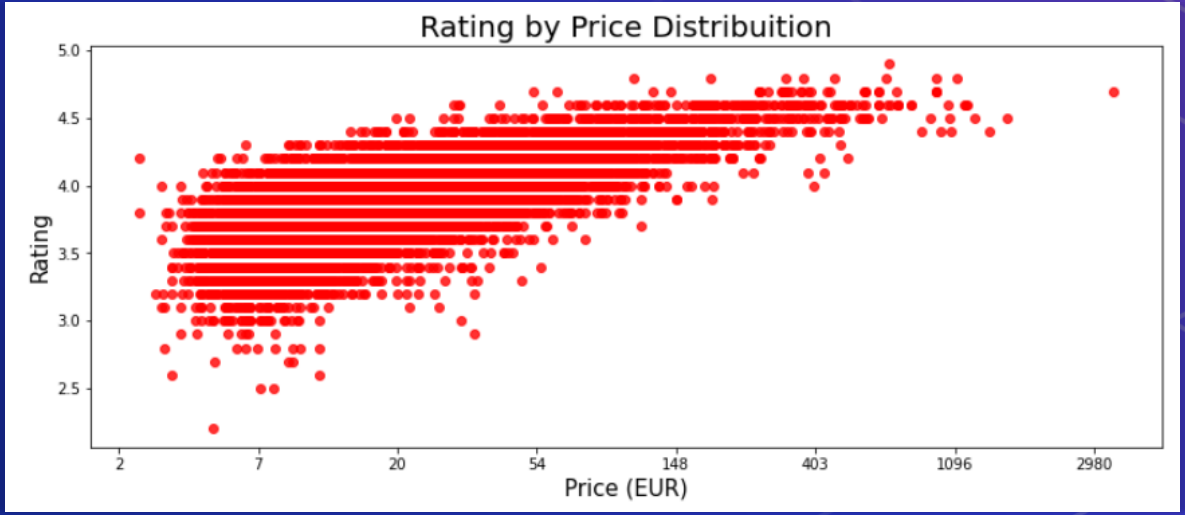

Price by Type of wine
Price (EUR)
2980
1096
403
148
54
20
7
2
Red
White
Sparkling
Rose
Type
Red
White
Sparkling
Rose


Rating by Type of wine
Rating
4
4
4
3
3
3
2
2
2
Type
Red
White
Sparkling
Rose
Red
White
Sparkling
Rose
Type

Correlation and trend

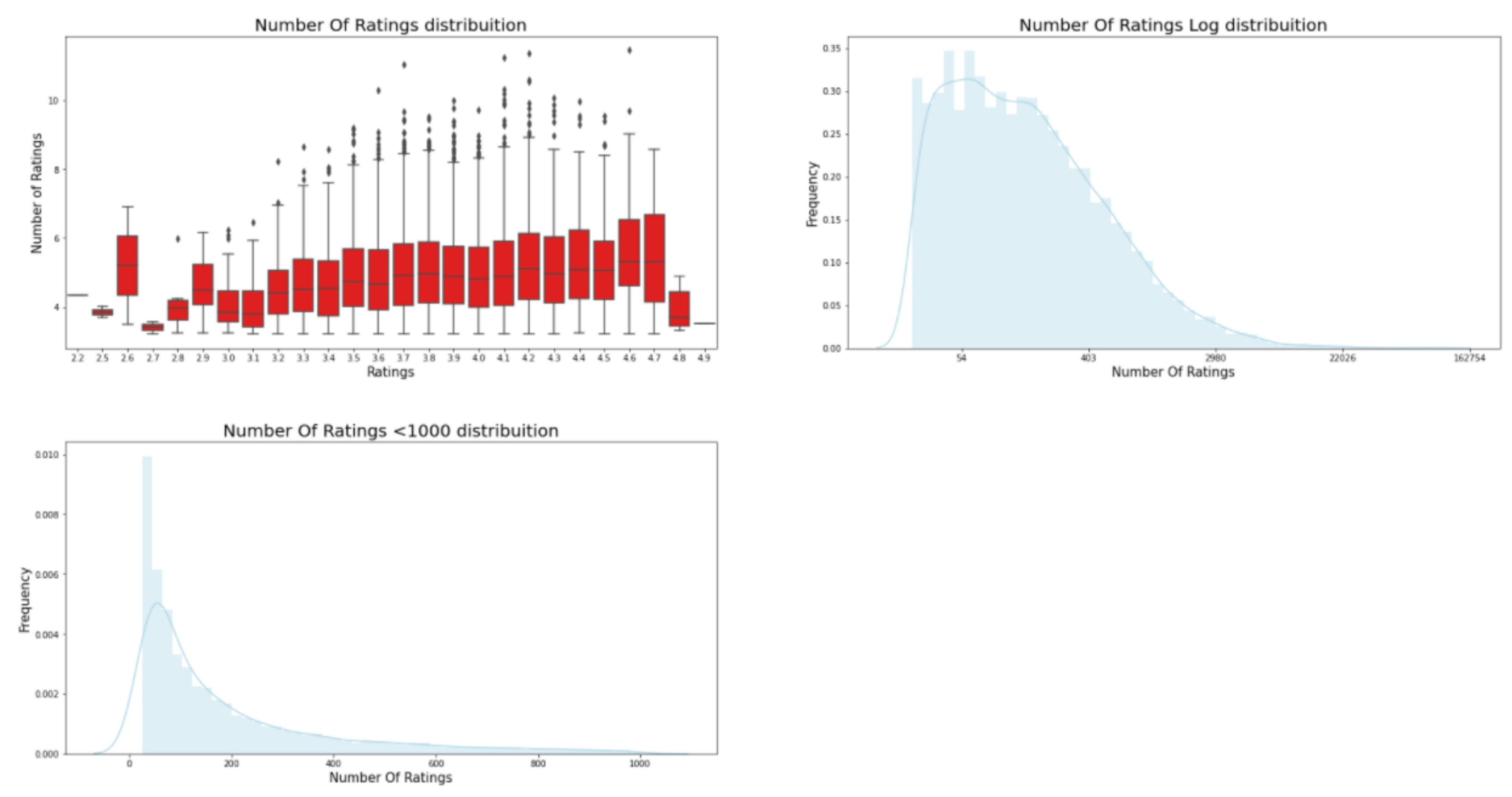


	Rating	NumberOfRatings	Price	Year
Rating	1	0.066	0.45	-0.33
NumberOfRatings	0.066	1	0.022	0.051
Price	0.45	0.022	1	-0.36
Year	-0.33	0.051	-0.36	1



Rating by Price Distribuition

Rating plot



Number Of Ratings distribuition

Number Of Ratings Log distribuition

Number Of Ratings <1000 distribuition

Visual Conclusion

“

| Number of Ratings has an exponential distribution

| For a large number of wines existing in Vivino, there is no Rating at all

| Problem for business ⚠️

Machine learning

| Purpose

| Features engineering

| Test set split for prediction

| Model selection

Purpose

1 Lack of Rating

2 Predict Rating

3 Inform the customer

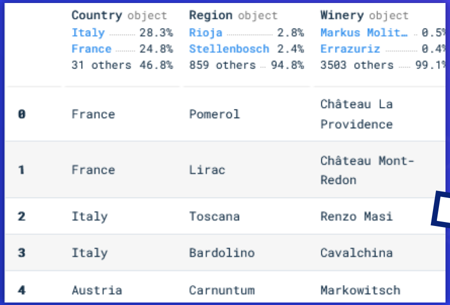
Feature engineering

Features

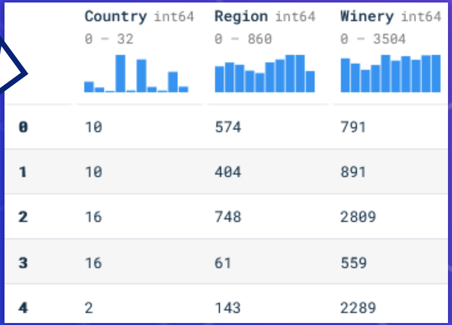
| ~~Name~~
| Country (String)
| Region (String)
| Winery (String)
| NumberOfRatings
| Price
| Year
| Type (Red, White, Sparkling, Rosé)

⇨ Label Encoder

⇨ One hot encoder



	Country object	Region object	Winery object
	Italy 28.3% France 24.8% 31 others 46.8%	Rioja 2.8% Stellenbosch 2.4% 859 others 94.8%	Markus Molit… 0.5 Errazuriz 0.4 3503 others 99.1
0	France	Pomerol	Château La Providence
1	France	Lirac	Château Mont-Redon
2	Italy	Toscana	Renzo Masi
3	Italy	Bardolino	Cavalchina
4	Austria	Carnuntum	Markowitsch



	Country int64	Region int64	Winery int64
	0 – 32	0 – 860	0 – 3504
0	10	574	791
1	10	404	891
2	16	748	2809
3	16	61	559
4	2	143	2289

Test set split for prediction

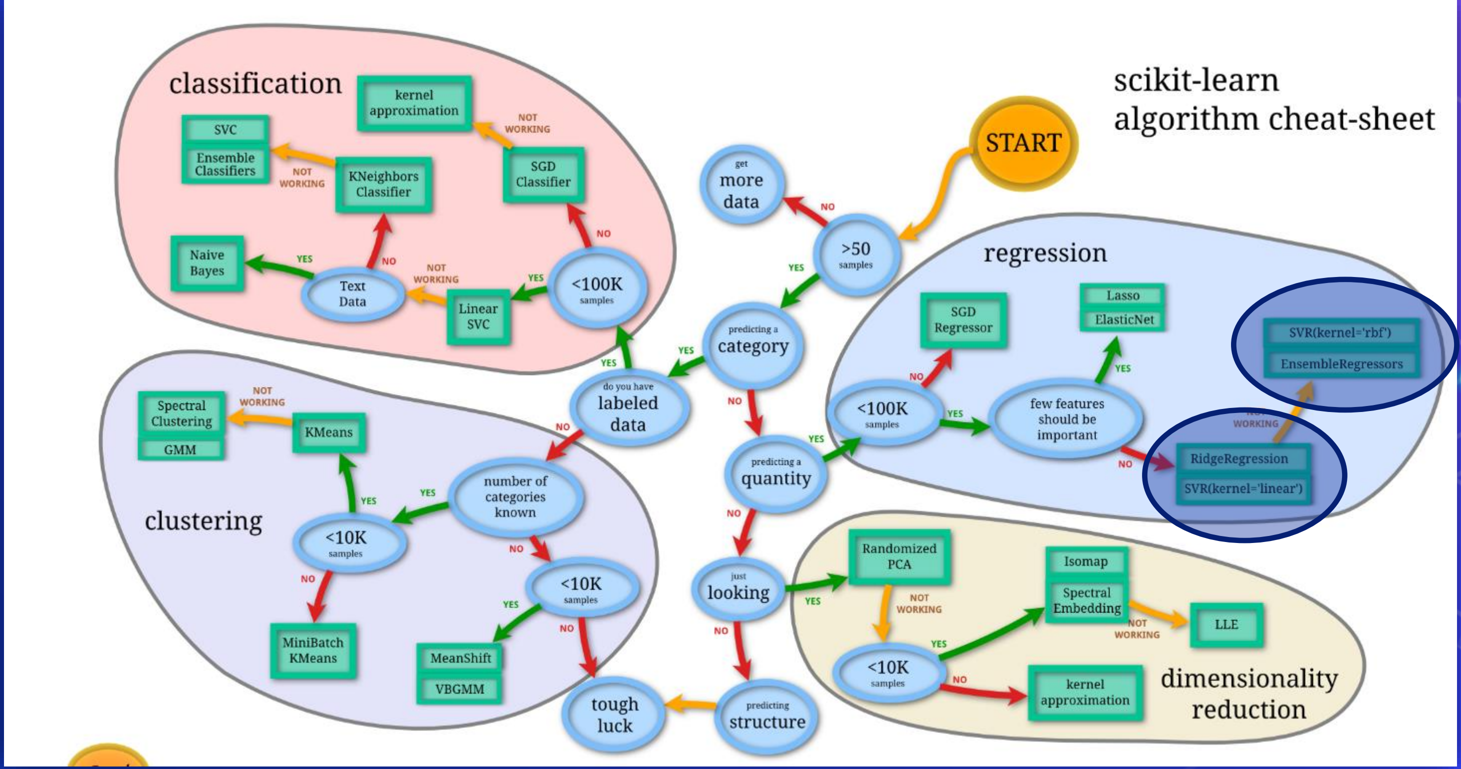
- Definition of 3 set of data based on the **NumberOfRatings**
- **We want to compare the performance between these sets**

Number Of Ratings LOW : < 40

Number Of Ratings MIDDLE : > 40 & < 850

Number Of Ratings HIGH : > 850

Model selection



Comparison between Ridge Regressor and Ensembling model

Ridge Regressor 😕

Accuracy (MAE)

Number Of Ratings LOW : 0,22

⚠ Number Of Ratings MIDDLE : 0,20

Number Of Ratings HIGH : 0,20

Random Forest 🙂

Accuracy (MAE)

Number Of Ratings LOW : 0,17

Number Of Ratings MIDDLE : 0,13

Number Of Ratings HIGH : 0,11

Thanks!

Any questions?