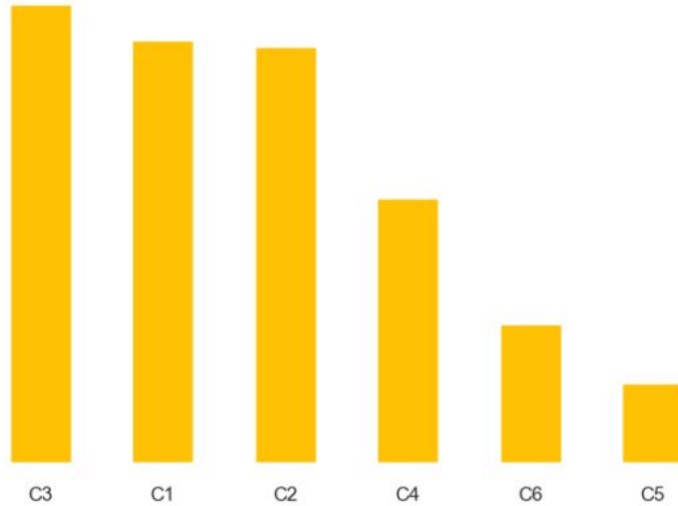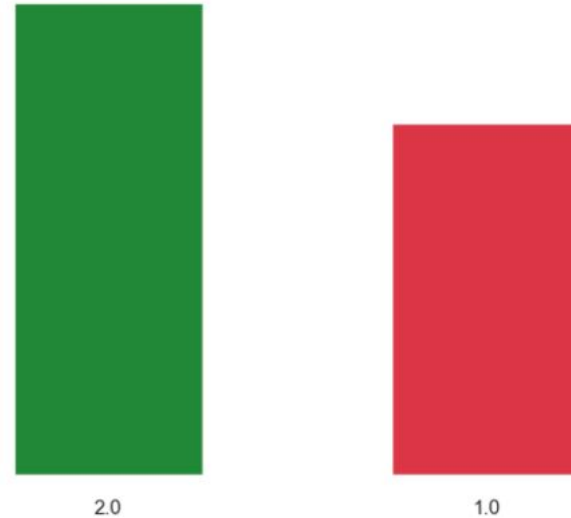# B2W Challenge
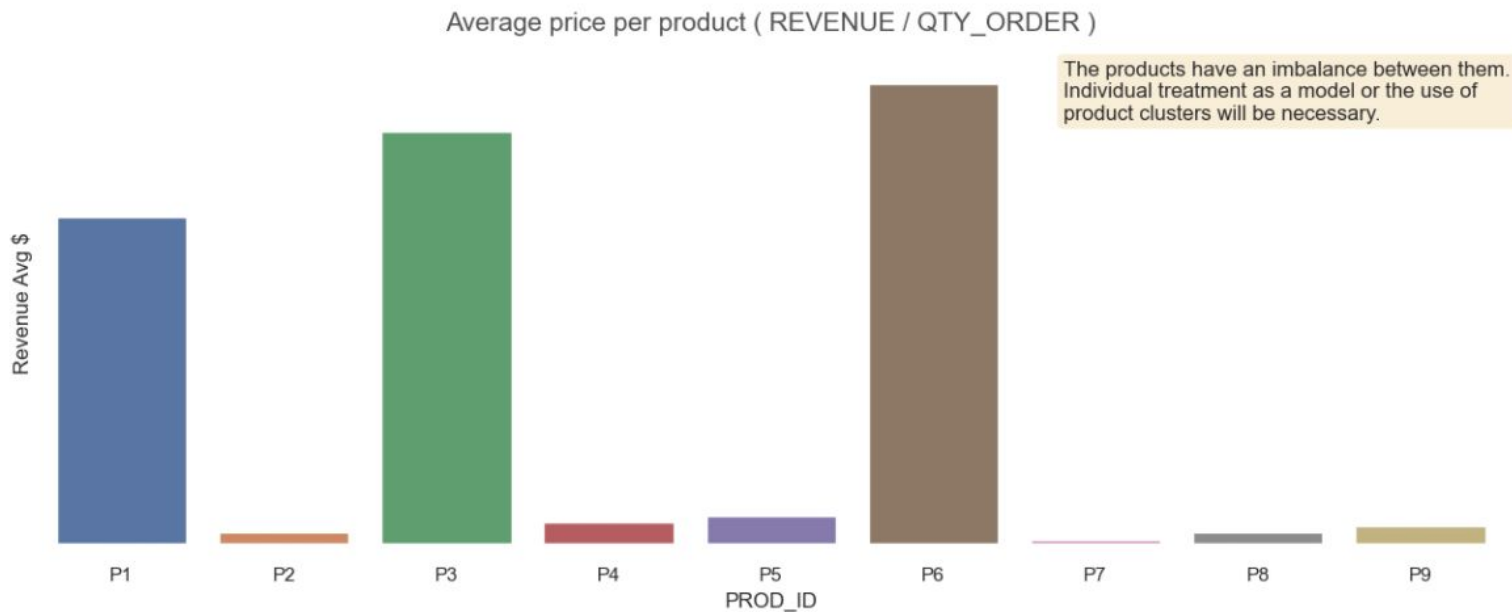
Model forecasting

ERYC MASSELLI

Relating competitors

Occurrence of lower prices from competitors
C3 is the most relevant competitor in the scenario

Payment methods of the most relevant competitors

C3   C1   C2   C4   C6   C5

2.0   1.0

Understanding the difference between products

Average price per product (REVENUE / QTY_ORDER)

The products have an imbalance between them. Individual treatment as a model or the use of product clusters will be necessary.
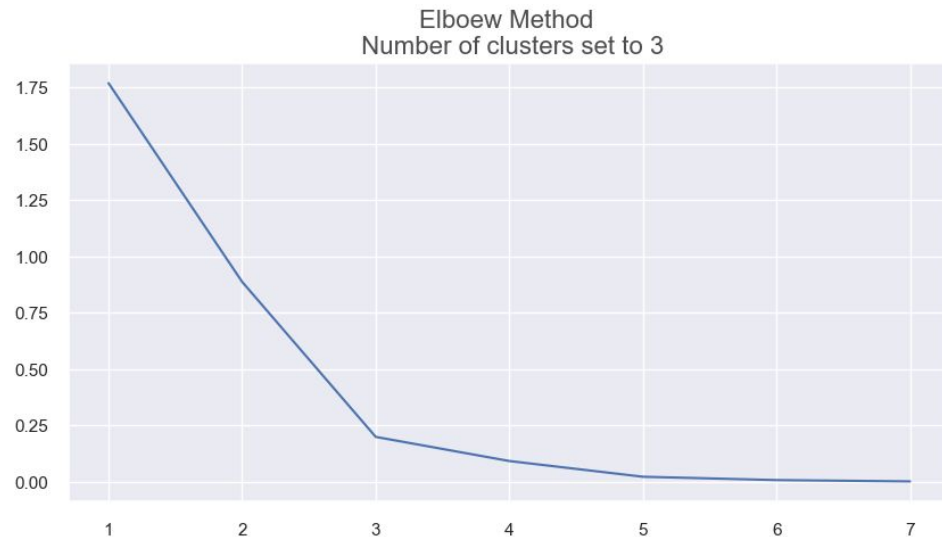
Transformation to supervised model

- Various possibilities for parameterization and diversity of models

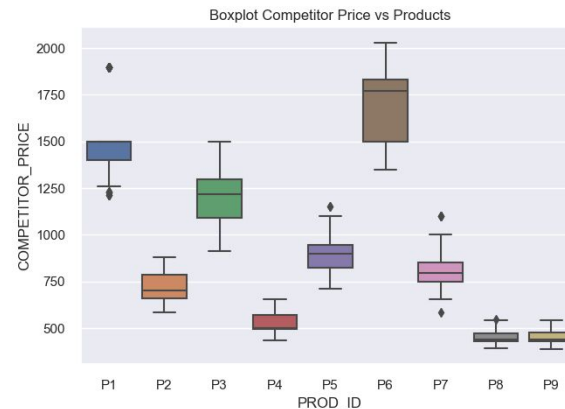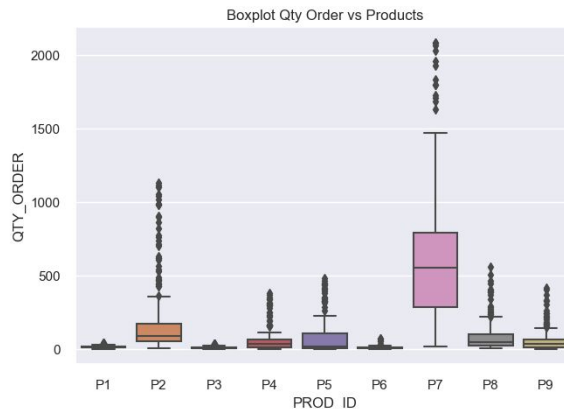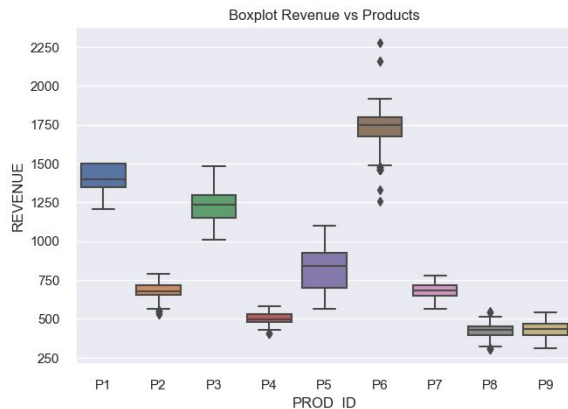- Used shift to shift the data, thus creating temporal features (t-2, t-1, t+1, t+2)

Feature correlation
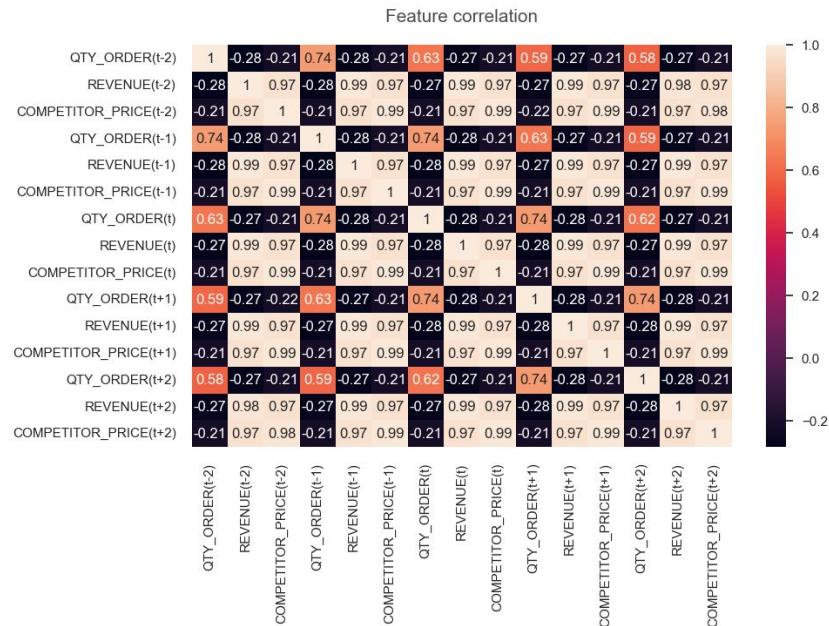
K-means for product clustering

Elboew Method
Number of clusters set to 3

	PROD_ID	QTY_ORDER	REVENUE	LABEL
0	P1	3597.143168	351287.831255	0
1	P2	49132.265996	194956.364284	1
2	P3	2470.633106	308111.805211	0
3	P4	12431.904156	99455.123366	1
4	P5	14940.066835	149408.145137	1
5	P6	2723.437701	491197.768680	0
6	P7	177069.942954	194498.385509	2
7	P8	18049.193527	88309.741122	1
8	P9	11923.209047	91042.659545	1

Outlier handling

Boxplot Revenue vs Products

Boxplot Qty Order vs Products

Boxplot Competitor Price vs Products
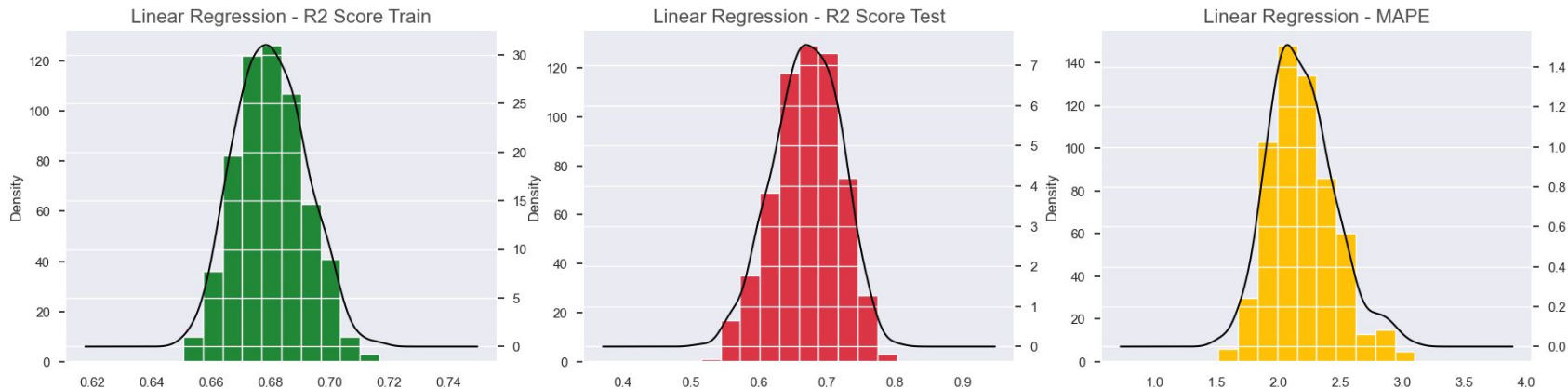
Outliers were filtered based on a standard deviation above 3 in their product dataset. These values (REVENUE) have been replaced by your average.
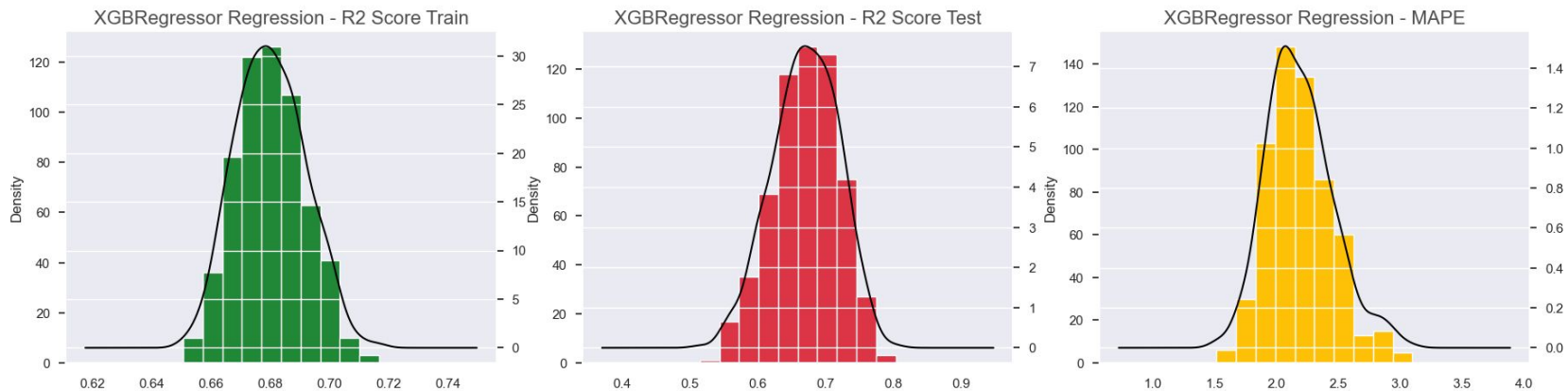
For COMPETITOR_PRICE the outliers were replaced by the same corresponding REVENUE value.
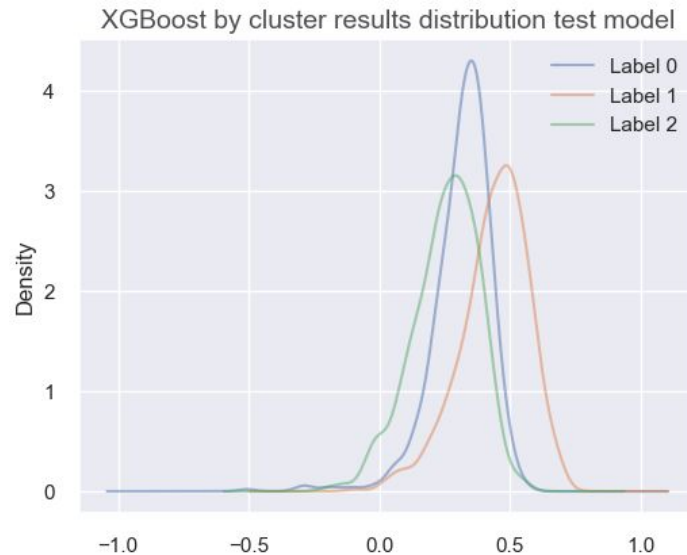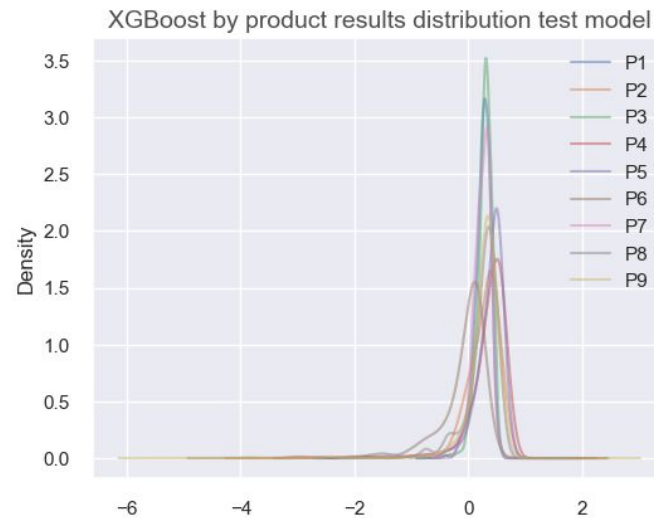
Linear Regression

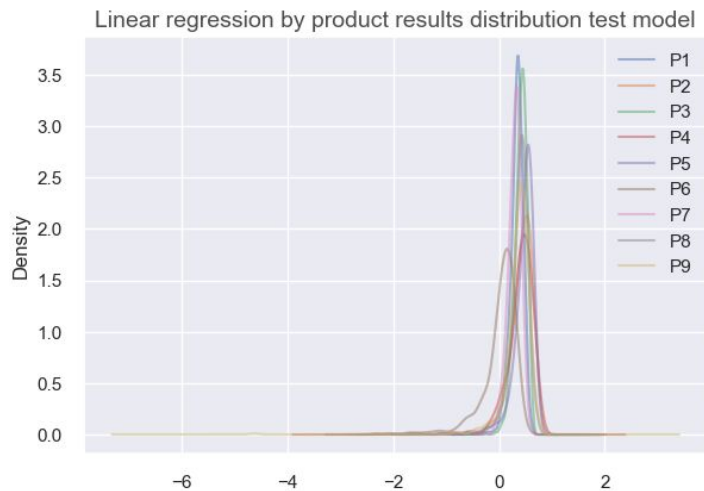Linear Regression - R2 Score Train

Linear Regression - R2 Score Test

Linear Regression - MAPE

XGBoost

Clustering Regression Model

Linear regression by cluster results distribution test model

XGBoost by cluster results distribution test model

Product Regression Model

Linear regression by product results distribution test model
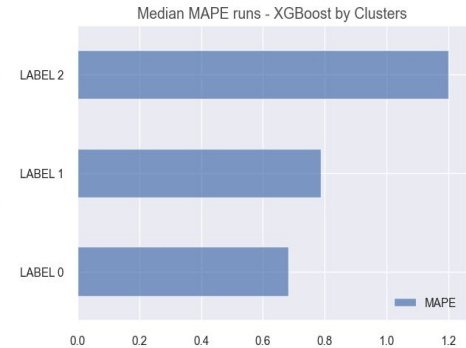

XGBoost by product results distribution test model
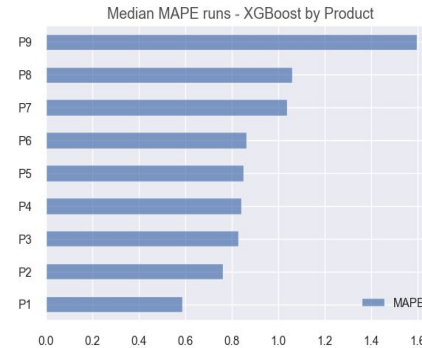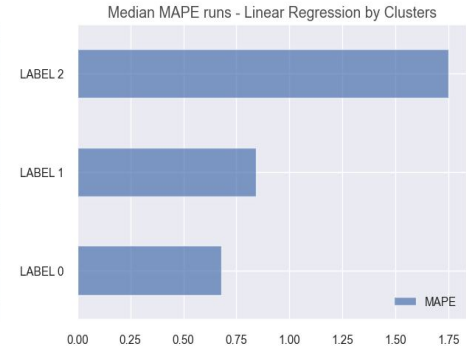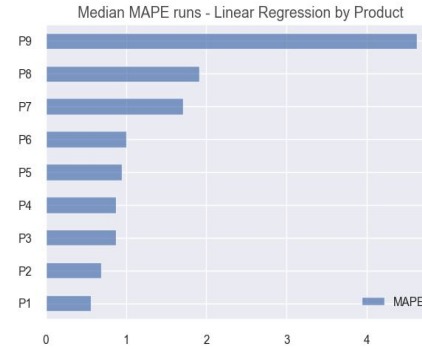
Results and Next Steps

- Despite the significant relevance of the results in the execution of the models with all data available in R2 score, there is a problem in the prediction by products analyzed in the MAPE method.

- Both algorithms presented greater prediction difficulties for the products P1, P2 and P3, even resulting in a negative R2 for P1 (worse than a horizontal line).

- Clustering in turn alleviated discrepancies in results when compared to one-to-one products.

Median MAPE runs - Linear Regression by Product

Median MAPE runs - Linear Regression by Clusters

Median MAPE runs - XGBoost by Product

Median MAPE runs - XGBoost by Clusters

Thank You!

erycmasselli@gmail.com