Mind over Data:
Retail Challenge

Group Y:

- João Alves Henriques;
- João Paulo César
- Pedro Sancho
- Vilmar Adriano Bussolaro

NOVA IMS Information Management School

Context

Mind over Data trusted us to help them on our first real world problem and challenged us to do produce solutions for the following challenges:

- Data/Feature Engineering
- Quarterly analysis of each Point-of-Sale characteristics
- Point-of-Sales Clustering
- Units Product forecast

Our Data

Our original data was a csv file with 27GB containing:

- 9 columns
- 21 Families of Products
- 178 Categories of Products
- 1 535 Brands Products
- 8 660 SKUs
- 410 Point-of-Sales
- Sales data from almost 3 years

Data Preprocessing

Steps performed:

- Regular expression (*Regex*) preprocessing
- Feature *downcasting*
- Data trimming
- Data aggregation
- Feature Engineering

NOVA IMS Information Management School

Quarterly analysis

Solution implemented using:

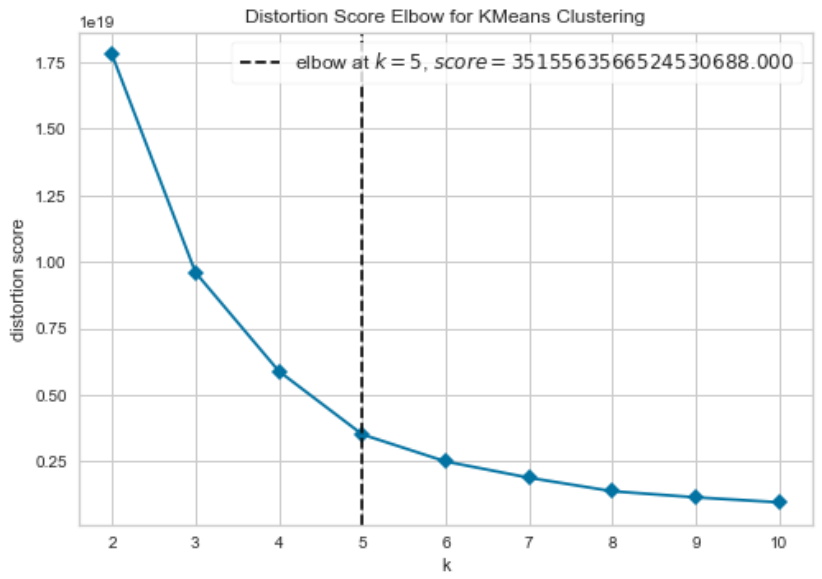
Clustering

K-Prototypes:

- POS_ID as guide (*groupby*)
- Metric Features: Total_Value, Total_Units
- Categorical Features: ProductFamily_ID, ProductName_ID, ProductBrand_ID

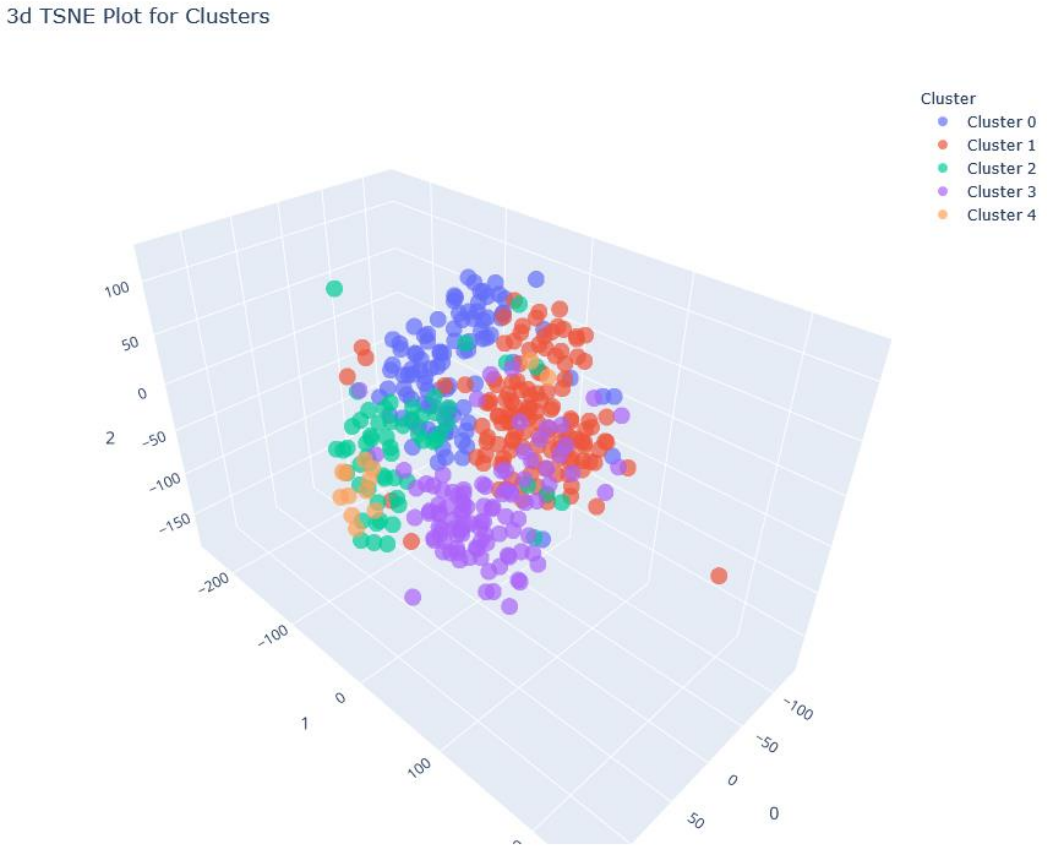
Silhouette	Calinski-Harabasz	Davies-Bouldin
0.54	1356.65	0.50

POS_ID	Total_Value	Total_Units	ProductFamily_ID	ProductBrand_ID	ProductName_ID
1	978670349.77	609151	21	1472	1228
2	635534919.91	390484	12	133	253
3	1048120296.56	602120	21	1472	253
4	1261300284.63	797480	21	1472	253
5	668177864.75	423022	21	1472	253
6	628293081.42	391811	21	133	253
7	438050582.83	292544	21	133	253
8	1196780448.41	776041	21	1472	253
9	752055977.18	456590	21	1472	1228
10	425764712.91	261983	21	133	253

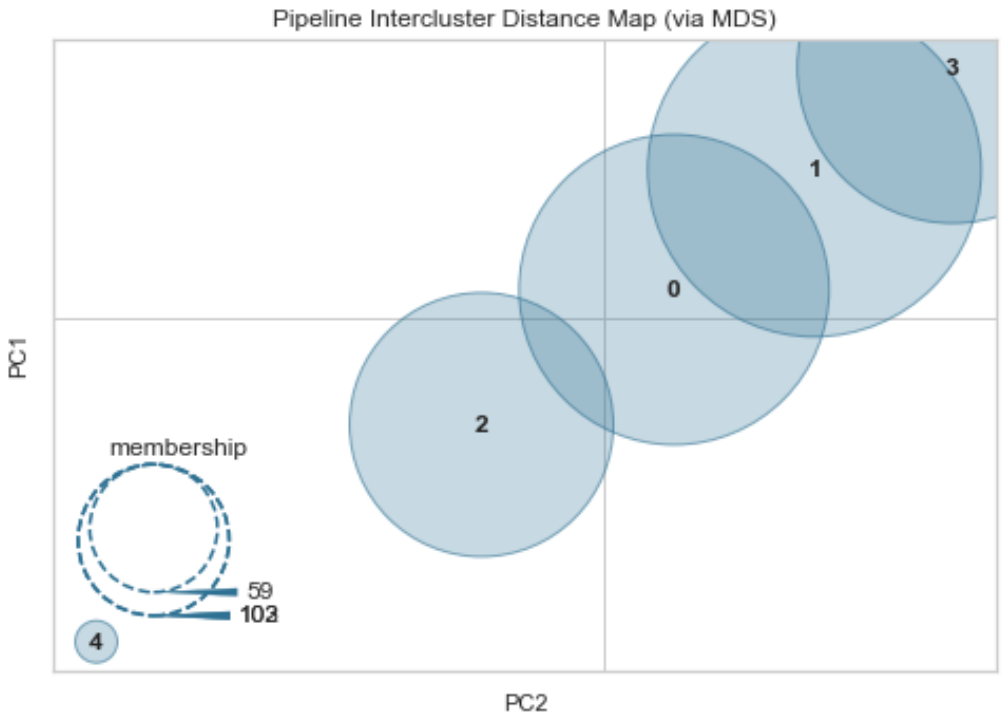


NOVA IMS Information Management School

Clustering



Clustering



Cluster 3: lowest value, less purchases

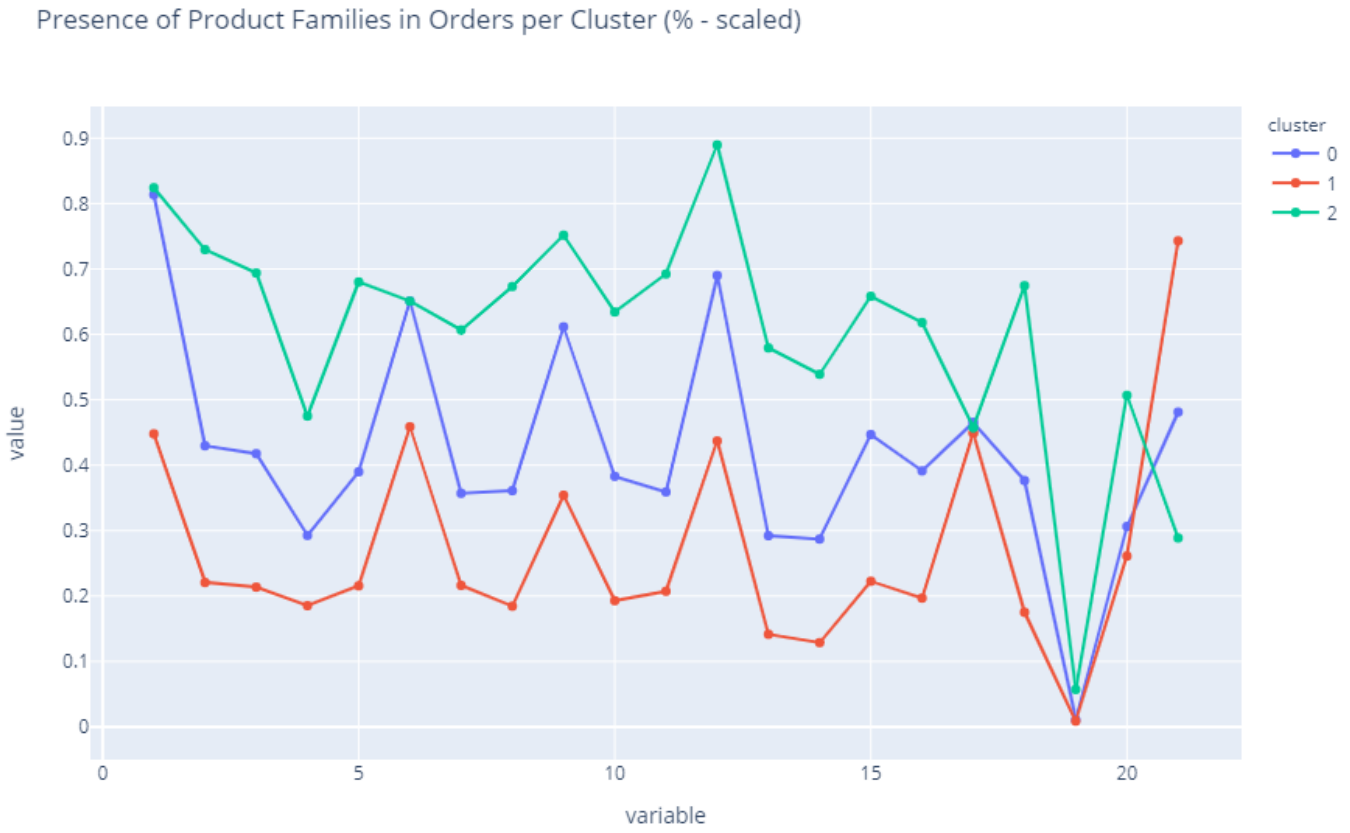
Cluster	Total_Value	Total_Units	ProductFamily_ID	ProductBrand_ID	ProductName_ID
Cluster 0	926312759.01	576697.31	21	133	253
Cluster 1	650379848.28	404556.80	21	133	253
Cluster 2	1258010167.01	775852.29	21	133	253
Cluster 3	404022574.45	250576.04	21	133	253
Cluster 4	1841490763.73	1118103.08	21	1472	253

Cluster 4: highest value, more purchases, top Brand 1472

Clustering

- **Additional Solution (Feature Engineering + K-means)**:

  Family Preference percentage for POS_ID



Predictions

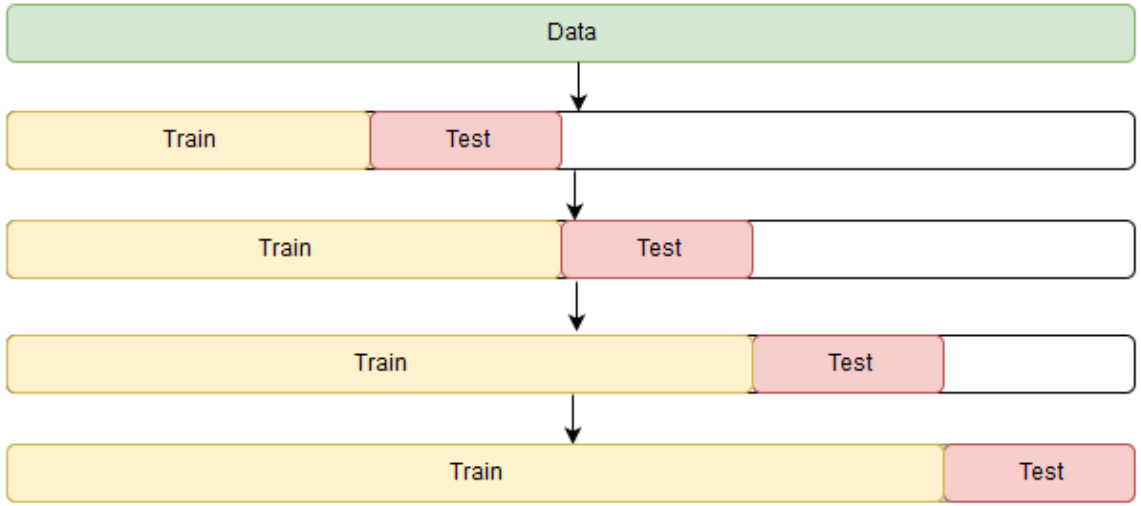
Final preprocessing to trim the number of time series to predict (pair of POS-Product):

Removing discontinued products

3-fold Cross-Validation on a rolling basis

25 algorithms tested

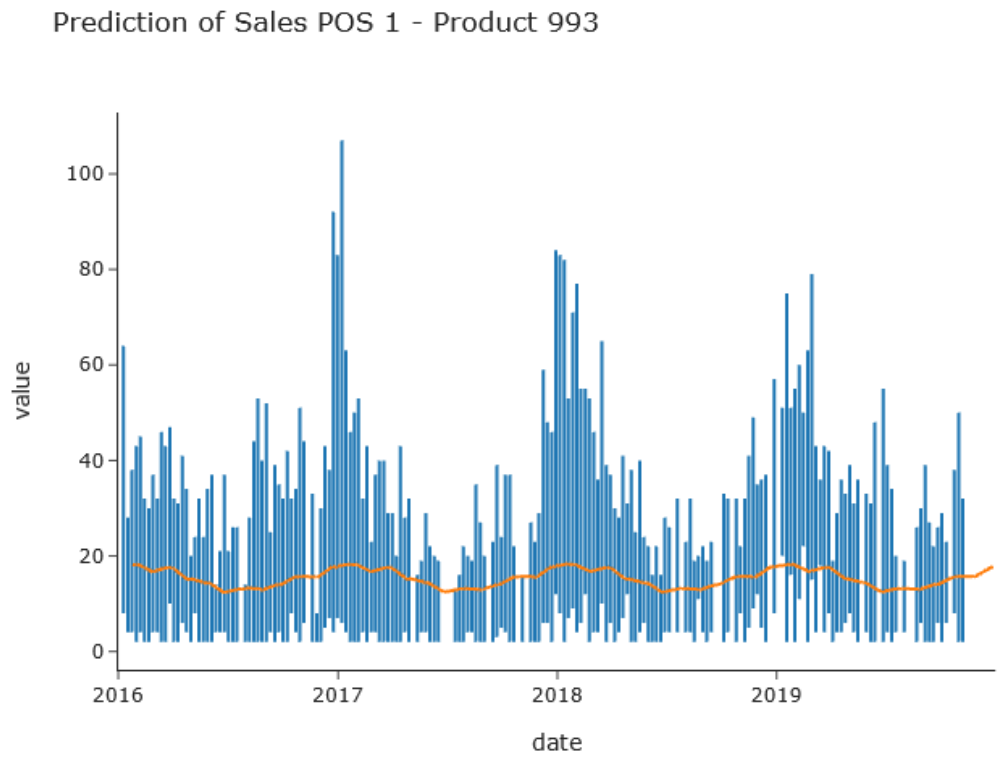
MAE used to select algo for each time series



Predictions

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)	time_series
huber	Huber Regressor	15.2322	455.6069	21.3306	-0.1644	1.0480	1.5959	0.0100	1_993
huber	Huber Regressor	6.7912	114.3427	10.6565	-0.0059	0.8108	1.1844	0.0100	1_356
knn	K Neighbors Regressor	10.8454	296.2128	16.3733	0.5624	0.4765	0.5365	0.0100	1_481
knn	K Neighbors Regressor	12.5603	425.1140	19.9768	0.5030	0.4194	0.3674	0.0067	1_1234
knn	K Neighbors Regressor	12.4270	402.6272	18.7425	0.4651	0.4087	0.3748	0.0100	1_1147

```
{'1_993': HuberRegressor(alpha=0.0001, epsilon=1.35, fit_intercept=True, max_iter=100,
                tol=1e-05, warm_start=False),
 '1_356': HuberRegressor(alpha=0.0001, epsilon=1.35, fit_intercept=True, max_iter=100,
                tol=1e-05, warm_start=False),
 '1_481': KNeighborsRegressor(algorithm='auto', leaf_size=30, metric='minkowski',
                      metric_params=None, n_jobs=-1, n_neighbors=5, p=2,
                      weights='uniform'),
 '1_1234': KNeighborsRegressor(algorithm='auto', leaf_size=30, metric='minkowski',
                      metric_params=None, n_jobs=-1, n_neighbors=5, p=2,
                      weights='uniform'),
 '1_1147': KNeighborsRegressor(algorithm='auto', leaf_size=30, metric='minkowski',
                      metric_params=None, n_jobs=-1, n_neighbors=5, p=2,
                      weights='uniform'),
```



Last remarks on VAR for COVID impact

Thank you!