DATA MINING AND TEXT MINING

BIP PROJECT 2020

10529039 Cappelletti Andrea, 10532096 Maglione Sandro

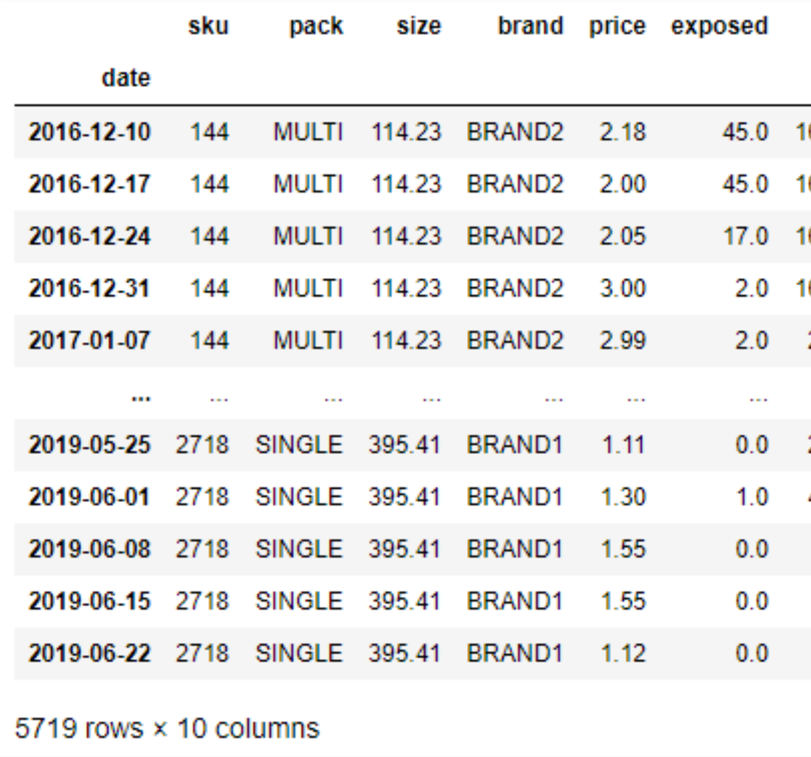
POLITECNICO
MILANO 1863

UIC

1) PREPROCESSING

❑ Check for missing weeks

❑ Fill missing values with data from next week



	sku	pack	size	brand	price	exposed
date						
2016-12-10	144	MULTI	114.23	BRAND2	2.18	45.0
2016-12-17	144	MULTI	114.23	BRAND2	2.00	45.0
2016-12-24	144	MULTI	114.23	BRAND2	2.05	17.0
2016-12-31	144	MULTI	114.23	BRAND2	3.00	2.0
2017-01-07	144	MULTI	114.23	BRAND2	2.99	2.0
...	...	...	...	...	...	...
2019-05-25	2718	SINGLE	395.41	BRAND1	1.11	0.0
2019-06-01	2718	SINGLE	395.41	BRAND1	1.30	1.0
2019-06-08	2718	SINGLE	395.41	BRAND1	1.55	0.0
2019-06-15	2718	SINGLE	395.41	BRAND1	1.55	0.0
2019-06-22	2718	SINGLE	395.41	BRAND1	1.12	0.0

5719 rows × 10 columns

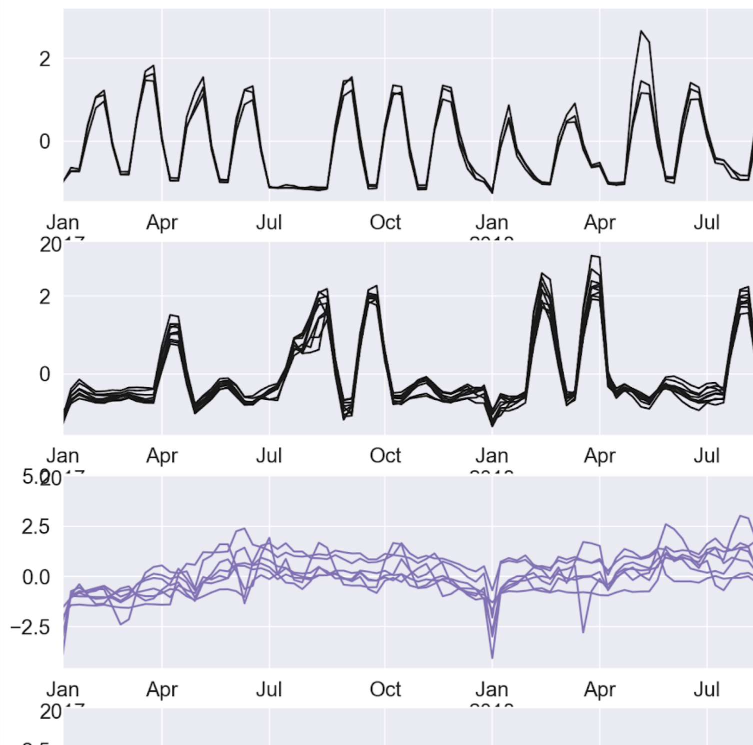
2) DATA ANALYSIS

❑ **Hierarchical Clustering**

Products in scope similar time series (2 clusters)

❑ **Scope**

BRAND2 and BRAND4

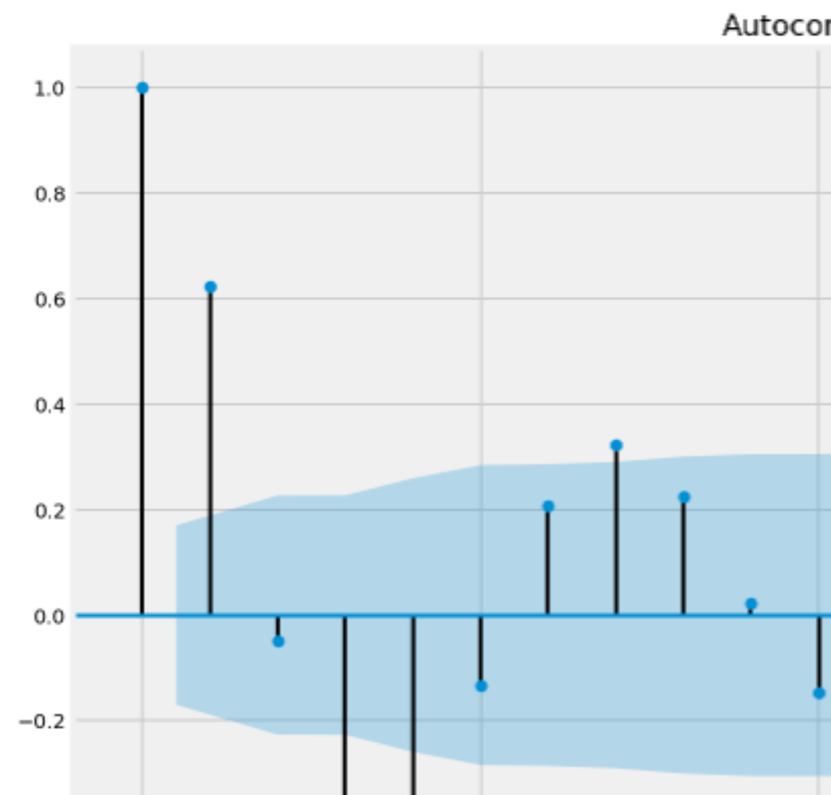


3) MODEL

- ❑ Build new DataFrame
  Past data to predict current week

- ❑ Autocorrelation

- ❑ Features selection



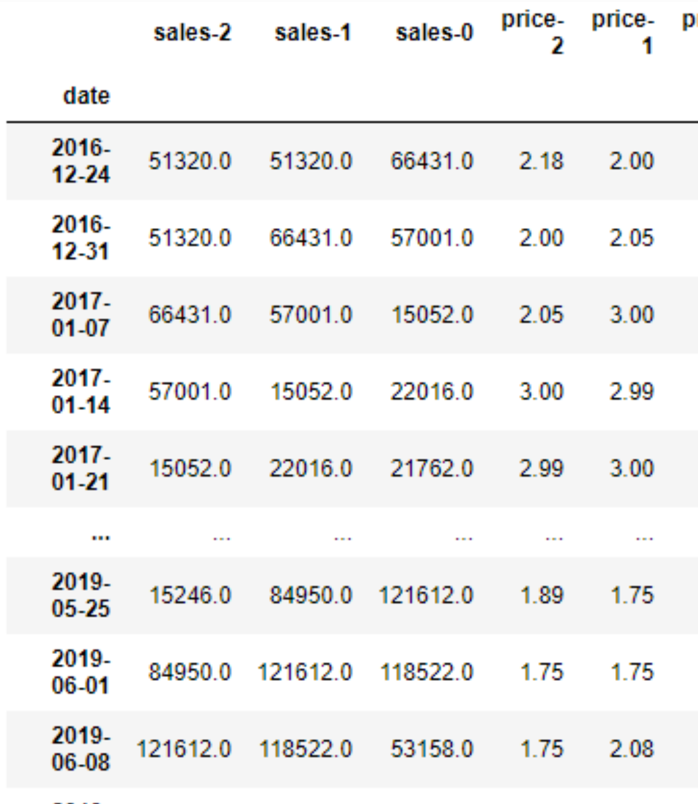
4) CONCLUSION

10-Fold Cross Validation

Random Forest: MAPE = 11.26

Lasso: MAPE = 17.90

AdaBoost: MAPE = 15.08



	sales-2	sales-1	sales-0	price-2	price-1	p
date						
2016-12-24	51320.0	51320.0	66431.0	2.18	2.00	
2016-12-31	51320.0	66431.0	57001.0	2.00	2.05	
2017-01-07	66431.0	57001.0	15052.0	2.05	3.00	
2017-01-14	57001.0	15052.0	22016.0	3.00	2.99	
2017-01-21	15052.0	22016.0	21762.0	2.99	3.00	
...	...	...	...	...	...	
2019-05-25	15246.0	84950.0	121612.0	1.89	1.75	
2019-06-01	84950.0	121612.0	118522.0	1.75	1.75	
2019-06-08	121612.0	118522.0	53158.0	1.75	2.08	