Sales Forecast: Sporting Goods

- Adila Aghazada
- Bengi Koseoglu
- Chowdhury, Abdullah Al Murad
- Khizer Naushad
- Na Gong
- Qian Xia
- Sanjita Suresh

GROUP 4

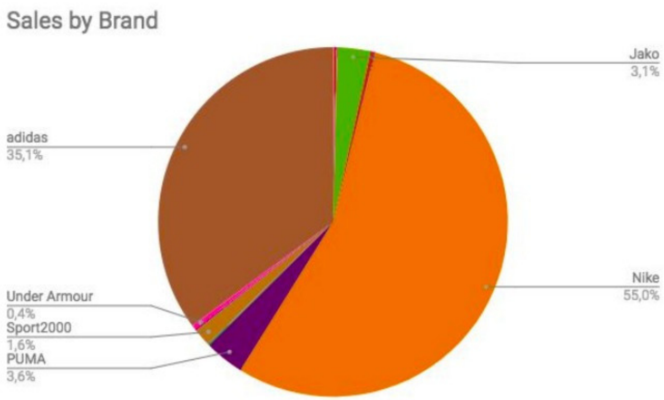
Agenda

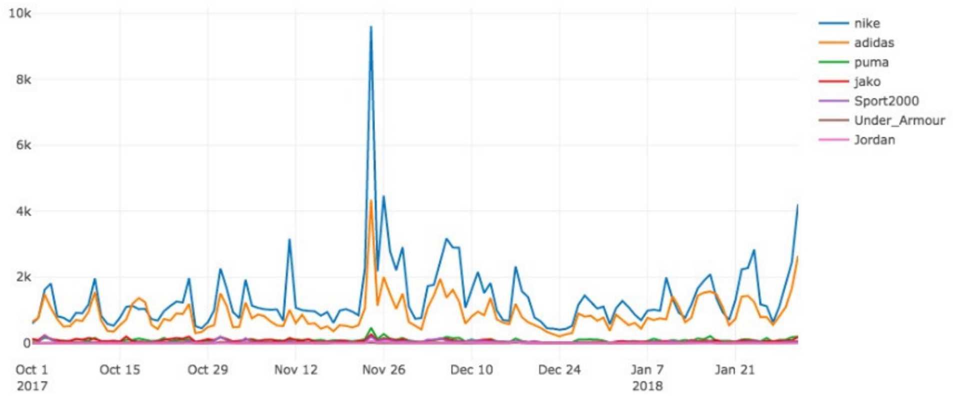
1. Data Exploring
2. Feature Engineering
3. Clustering
4. Modeling

Agenda

1. Data Exploring
2. Feature Engineering
3. Clustering
4. Modeling

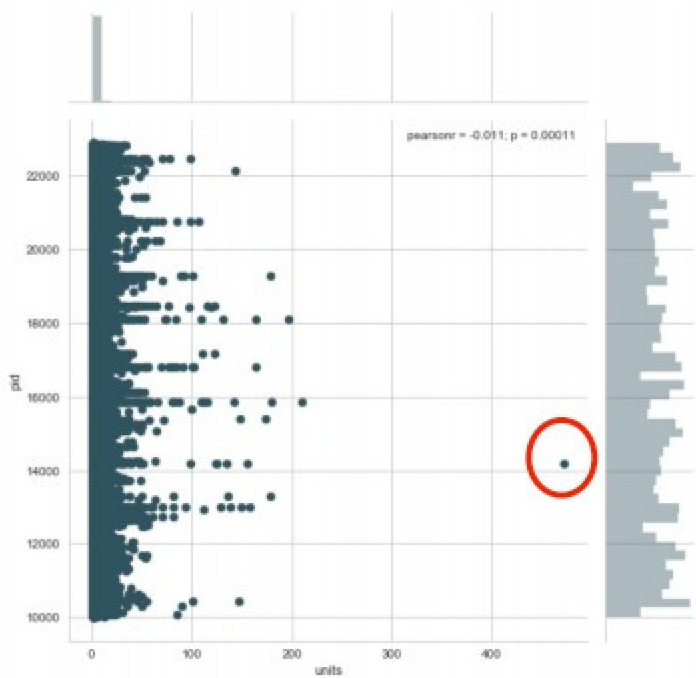
Data Exploration

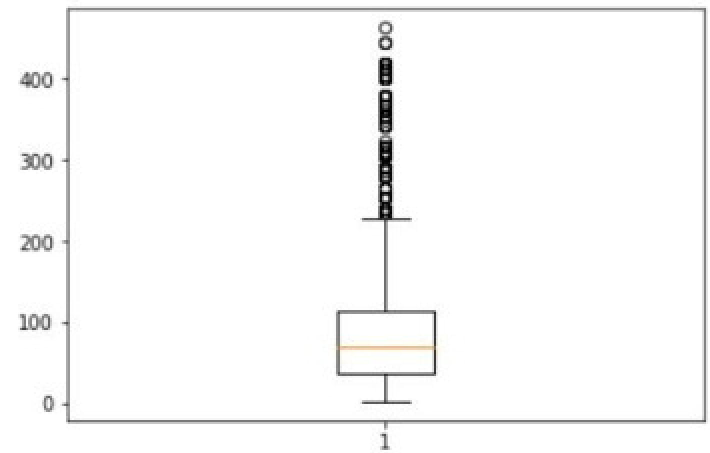




Xia; Bengi; Gong

Outlier Detection





Recommended Retail Price

Xia; Bengi; Gong

Agenda

1. Data Exploring
2. Feature Engineering
3. Clustering
4. Modeling

Feature Engineering

- Weather Data (from team 6)
- 11 Team Sports (from team 5)
- Holidays
- Weekends
- Size: new_size, size_classification
- New_product
- Price_daily_change
- Sum_unit_previous_month

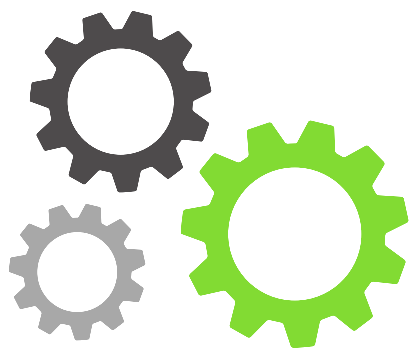
Dummy Features

Categorical At	Dummy Attribute
color	schwarz, blau, gruen, weis, braun, lila, grau, ...
brand	Nike, PUMA, adidas, Jako, ...
mainCategory	maincat_1, maincat_9, maincat_15
category	cat_2, cat_7, cat_16, cat_33, ...
subCategory	subcat_0, subcat_3, subcat_5, subcat_6,
size	size_S, size_M, size_L, size_XL, ...
...	...

(>300 features)

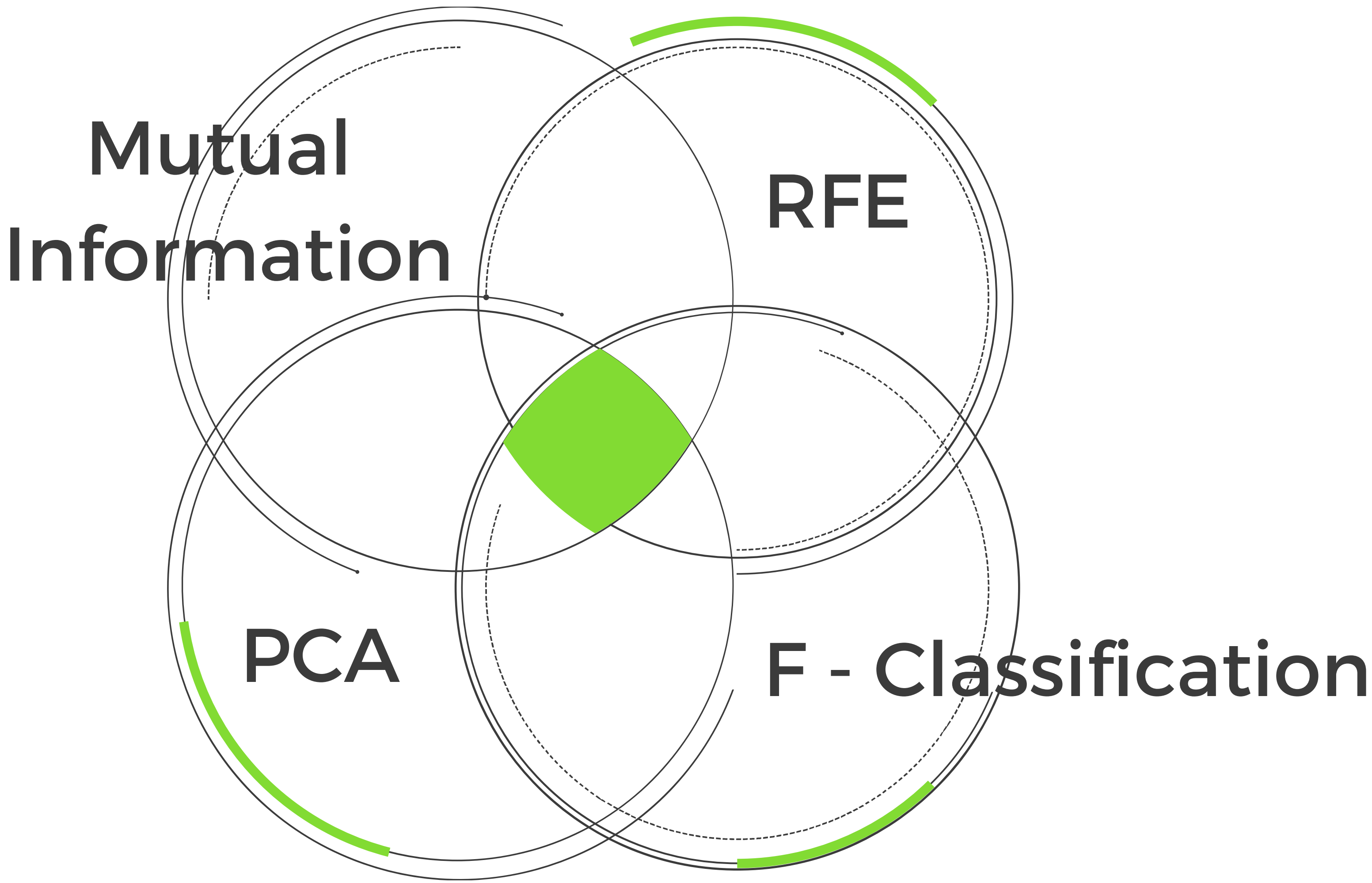
Long Format

Product	Sales Date	Units
10063 L	2017-10-04	12
10063 L	2017-10-11	1
10063 L	2017-11-20	3
10063 L	2017-11-30	4



Product	Sales Date	Units
10063 L	2017-10-01	0
10063 L	2017-10-02	0
10063 L	2017-10-03	0
10063 L	2017-10-04	12
10063 L	2017-10-05	0
10063 L	2017-10-06	0
10063 L	2017-10-07	0
10063 L	2017-10-08	0
10063 L	2017-10-09	0
10063 L	2017-10-10	0
10063 L	2017-10-11	1
10063 L	2017-10-12	0
10063 L	2017-10-13	0
10063 L	2017-10-14	0

Features Selection



Bengi; Gong

Features Selection

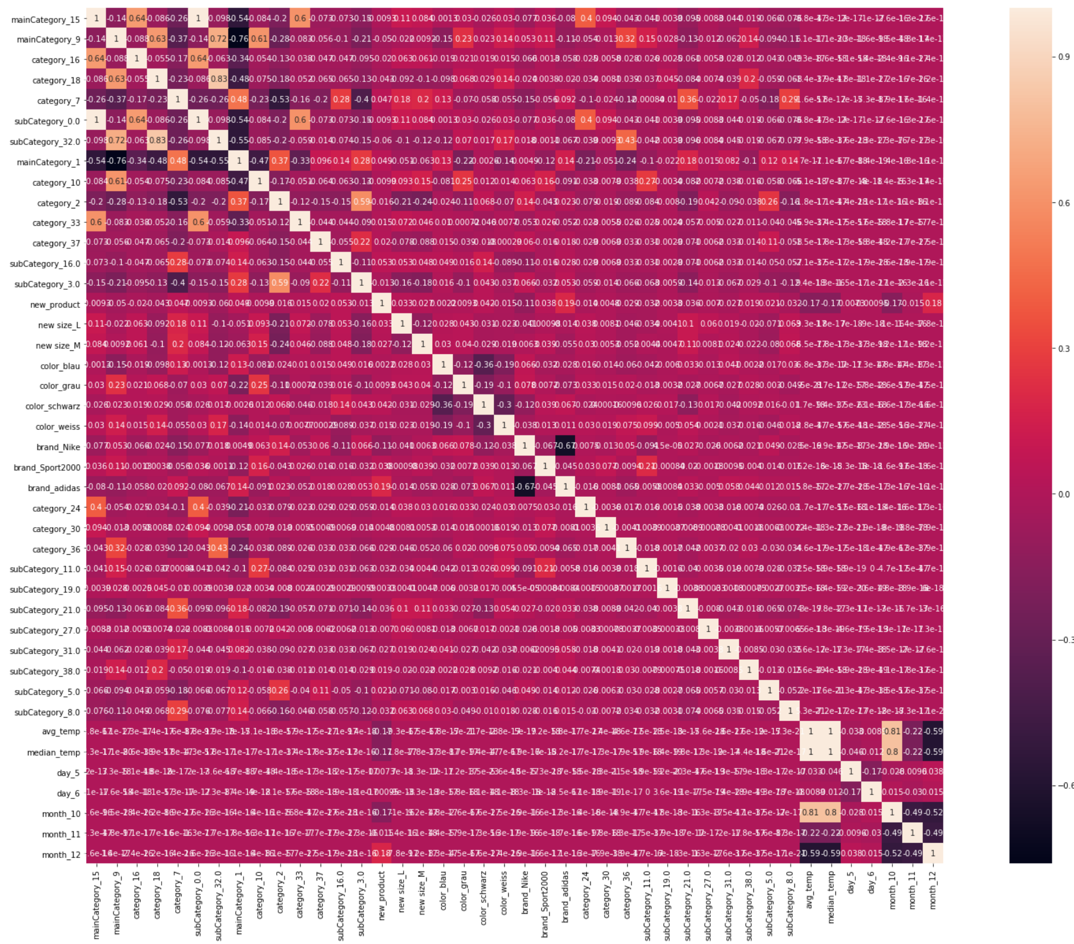
3

- mainCategory_15
- mainCategory_9
- Category_16
- Category_18
- Category_7
- subCategory_0
- subCategory_32

2

- mainCategory_1
- Category_10
- Category_2
- Category_33
- Category_37
- subCategory_16
- subCategory_3
- new_product

Correlation Matrix



Bengi; Gong

Agenda

1. Data Exploring
2. Feature Engineering
3. Clustering
4. Modeling

Clustering Approaches

1 BRAND

- Nike
- Adidas
- PUMA
- Jako
- Others brands

2 CATEGORY

- mainCategory_1
- mainCategory_9
- mainCategory_15

3 SALES

- low sales
- average sales
- good sales
- top sales
- new in Jan

4 K-Means Clustering

- Use the selected features with Oct-Dec data:

```
Index(['sales_day', 'sales_weekday', 'sales_month', 'rrp', 'new size_L',
       'new size_M', 'new size_S', 'new size_XL', 'color_blau', 'color_rot',
       'color_schwarz', 'brand_Nike', 'brand_adidas', 'category_10',
       'category_16', 'category_7', 'subCategory_13.0', 'subCategory_14.0',
       'subCategory_16.0', 'subCategory_22.0', 'new size_41', 'new size_42',
       'new size_43', 'new size_44'],
```

- Decide the optimal cluster number:
  - calinski_harabaz_score = 7
  - elbow analysis = 12

- K-means clustering: k = 7

5 Time Series Clustering

Dynamic Time Warping

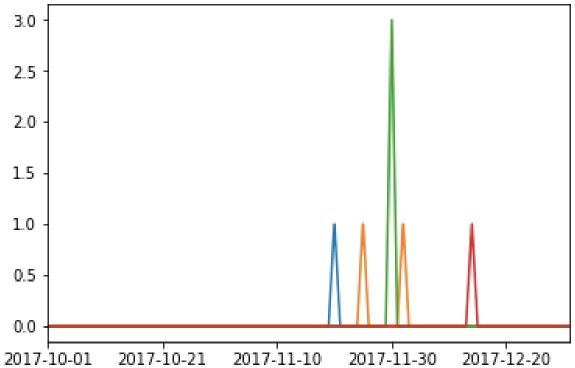
+

Hierarchical Clustering

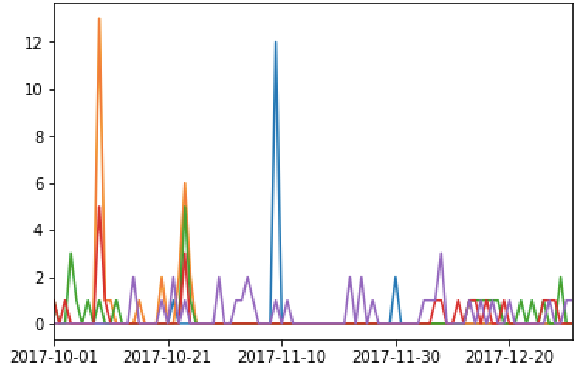
+

7 Clusters

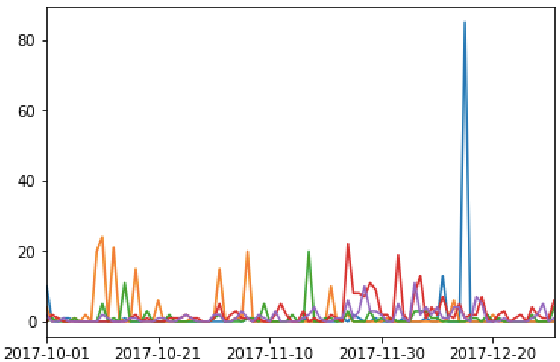
Bengi



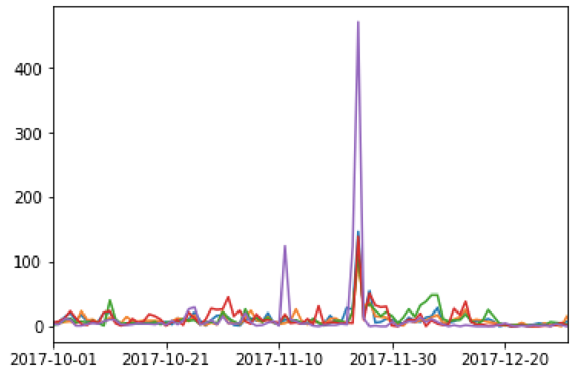
Cluster 1



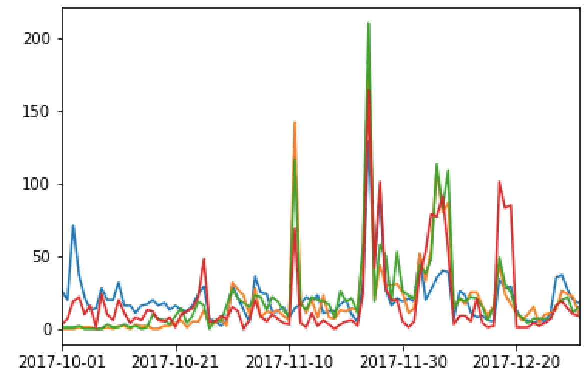
Cluster 2



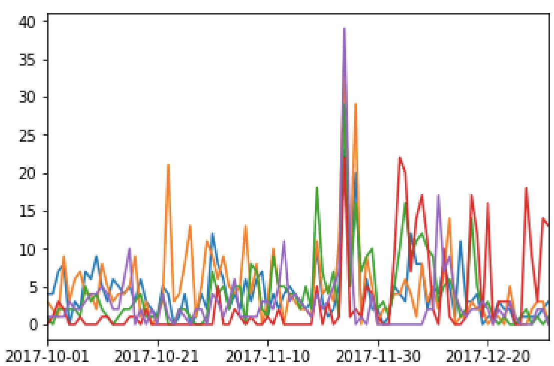
Cluster 3



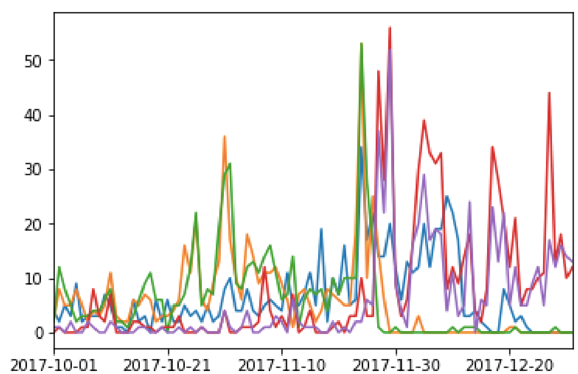
Cluster 4



Cluster 7



Cluster 5



Cluster 6

Clustering Comparison

Clustering	Performance
K-Means	280
Category	276.7
Brand	276.1
Time Series	**261**

* GBoostRegression, w/o parameter tuning, test_0, assigning 31st of January

Agenda

1. Data Exploring
2. Feature Engineering
3. Clustering
4. Modeling

Baseline

- January 15th
- Error function:
- Baseline = 261.02

$$E = \sqrt{\sum_i |d_i - \hat{d}_i|}$$

Test 0	260,77
Test 1	260,88
Test 2	261,50
Test 3	260,99
Test 4	260,92
Average	**261,02**
Standard Dev.	0,26

Failure Cases

pid	size	oct_openin	sales_oct	nov_oepning	sales_nov	dec_opening	sales_dec	jan_openin	sales_ja	feb_openi
10000	XL (158-170)	2	0	2	1	1	0	1	0	
10001	L	5	0	5	1	4	1	3	2	
10003	3 (35-38)	16	1	15	14	1	0	1	0	
10003	4 (39-42)	4	0	4	3	1	0	1	0	
10003	5 (43-46)	12	7	5	0	5	3	2	1	
10006	XL	2	0	2	0	2	1	1	0	
10008	XL	18	0	18	2	16	0	16	4	1
10013	L	2	0	2	0	2	0	2	1	
10013	M	5	0	5	0	5	1	4	3	
10013	S	2	0	2	1	1	0	1	0	
10015	L	7	1	6	1	5	0	5	0	
10015	S	2	0	2	0	2	1	1	0	
10017	L	5	1	4	3	1	0	1	0	
10020	XL	5	0	5	3	2	0	2	1	

Stock Calculation

Xia; Bengi; Gong: Murad

Failure Cases

- Feature: total_sales_current_month
- Cluster: sales binning
- Model: GBoost Regression
- Error = 203.6

Model Improvements

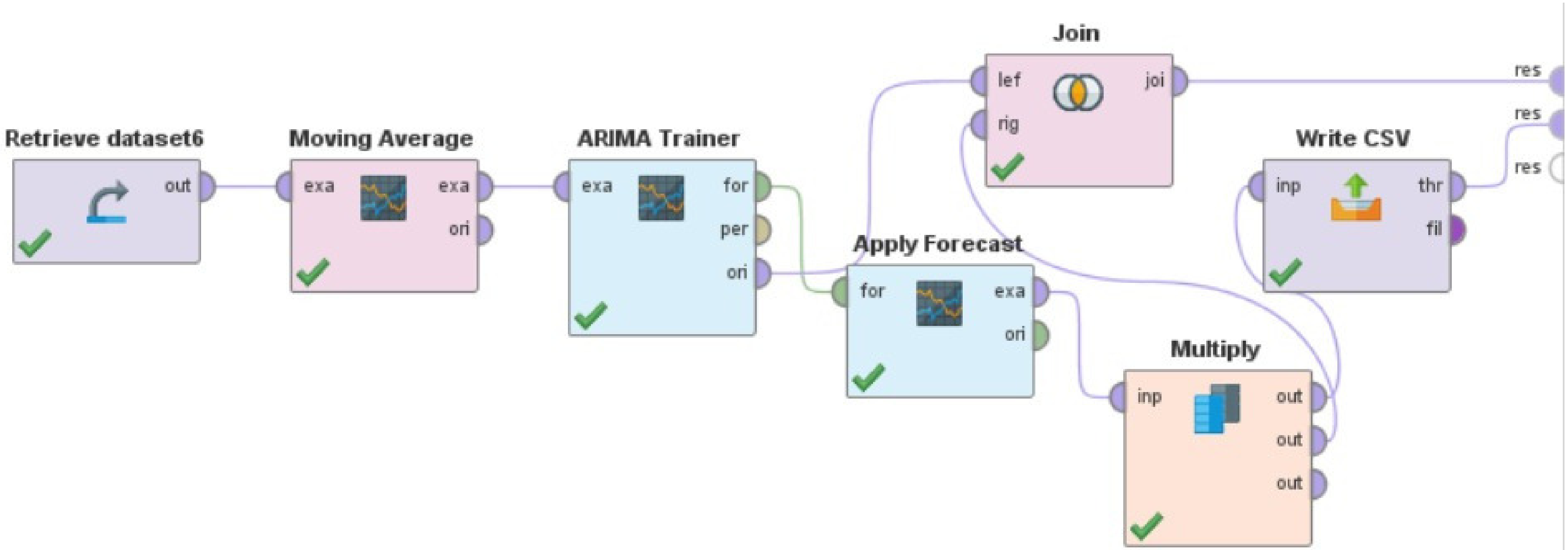
1. Handle Missing Values
2. Predicting decimals
3. January 31st -> January 22nd
4. Remove Black Friday
5. Based on Time Series Clustering
6. Regression Parameter Tuning
7. Cluster-based Model Combination
8. Ensemble - stacking

Xia; Bengi; Gong

Model Approaches

1. ARIMA
2. Windowing
3. Regression
4. Combine Model
5. Ensemble

ARIMA Model



Model	Average Performance	Standard Dev.
Arima all	262,034	0,50
Arima Time Clustering	**248,928**	0,34

Windowing Model

Model	Average Performance	Standard Dev.
Windowing Individually Lag(1)- Linear Reg	256,621	0,65
Windowing Individually Lag(30)- Linear Reg	264,042	4,16
Windowing Individually Lag(30)- ANN	259,428	0,22
Windowing Individually & Weather Data (Lag=1)- Regr	253,125	0,33
Windowing Individually & Weather Data (Lag=1)- ANN	268,706	0,31
Windowing Time Series Clustering (Lag=1)- Regr	254,947	0,35
Windowing Time Series Clustering (Lag=5)- Regr	256,922	0,36
Windowing Time Series Clustering (Lag=10)- Regr	254,988	0,34
Windowing Time Series Clustering (Lag=15)- Regr	254,822	0,30
Windowing Time Series Clustering (Lag=20)- Regr	252,602	0,30
Windowing Time Series Clustering (Lag=20)- ANN	**250,848**	**0,37**
Windowing Time Series Clustering (Lag=30)- Regr	266,374	0,38
Windowing Time Series Clustering (Lag=30)- Gradiant Boost	259,160	0,31
Windowing Time Series Clustering (Lag=30)- Random Forest	257,716	0,32

Regression Model

LRegression	308.6
GBRegression	285.7
DTRegression	293.5
RFRegression	293.9
WeekdayRegression	276

K-Means	280
Category	276.7
Brand	276.1
TimeSeries	261

GBoost & XGBoost + TS Cluster

Parameter Tuning

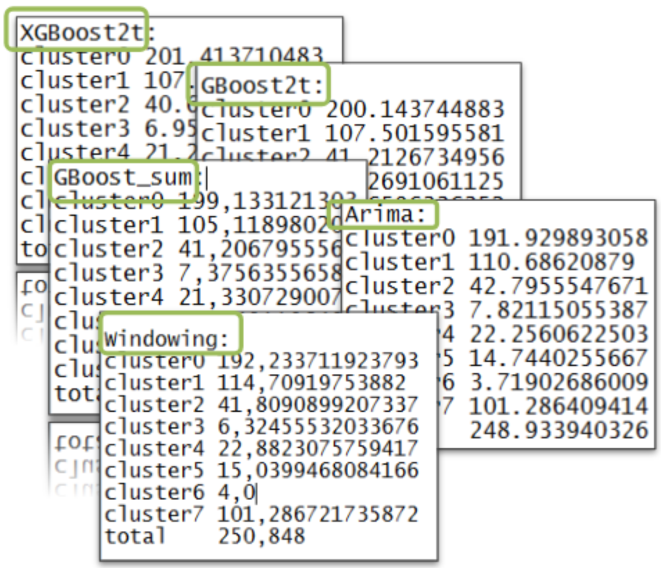
Cluster	Parameter - Gboost
2	{'learning_rate': 0.05, 'max_depth': 3, 'min_samples_split': 2, 'n_estimators': 500}
3	{'learning_rate': 0.05, 'max_depth': 3, 'min_samples_split': 10, 'n_estimators': 1000}
4	{'learning_rate': 0.01, 'max_depth': 10, 'min_samples_split': 5, 'n_estimators': 100}
5	{'learning_rate': 0.05, 'max_depth': 3, 'min_samples_split': 2, 'n_estimators': 100}
6	{'learning_rate': 0.01, 'max_depth': 15, 'min_samples_split': 10, 'n_estimators': 100}

5	{'colsample_bytree': 0.7, 'learning_rate': 0.03, 'max_depth': 6, 'min_child_weight': 7, 'n_estimators': 500, 'objective': 'reg:linear', 'silent': 1, 'subsample': 0.8}
6	{'colsample_bytree': 0.8, 'learning_rate': 0.03, 'max_depth': 6, 'min_child_weight': 7, 'n_estimators': 500, 'objective': 'reg:linear', 'silent': 1, 'subsample': 0.7
7	{'colsample_bytree': 0.8, 'learning_rate': 0.03, 'max_depth': 9, 'min_child_weight': 6, 'n_estimators': 500, 'objective': 'reg:linear', 'silent': 1, 'subsample': 0.7}

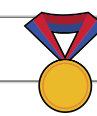
Model	Average Performance	Standard Dev.
GBoost Regression	253.6	0,59
GBoost Regression_sumunit	**251.7**	**0,45**
XGBoost Regression	254.4	0,66

Gong; Xia; Bengi

Combine Model



Model	Average Performance	Standard Dev.
Combine Model-1	248.7	0,33
Combine Model-2	247.0	0,37
Combine Model-3	246.8	0,34
Combine Model-4	**245.8**	**0,37**



cluster 1	ARIMA
cluster 2	GBoostRegression_sumUnits
cluster 3	XGBoost Regression
cluster 4	Windowing Time Series Clustering (Lag=20)- ANN
cluster 5	XGBoost Regression
cluster 6	GBoostRegression_sumUnits
cluster 7	Windowing Individually Lag(30)- Linear Reg
cluster 8	GBoost Regression

Model	Average Performance	Standard Dev.
Ensembling Stacking-1	**250.9**	**0,61**
Ensembling Stacking-2	253.8	0,59

Bengi; Gong; Xia

February Prediction

DATA MINING CUP
International Student Competition

pid|size|soldOutDate
15835|39 1/3|2018-02-22
15835|40|2018-02-03
15835|41 1/3|2018-02-10

Column name	Value range
pid	Natural number
size	String
soldOutDate	Format YYYY-MM-DD

pid|size|soldOutDate
10000|XL (158-170)|2018-02-18
10001|L|2018-02-18
10003|3 (35-38)|2018-02-11
10003|4 (39-42)|2018-02-18
10003|5 (43-46)|2018-02-18
10006|XL|2018-02-18
10008|XL|2018-02-22
10013|L|2018-02-22
10013|M|2018-02-18

12824 * 3

```
SoldOutDay_predict.dtypes
```

```
pid                     int64
size                   object
soldOutDate    datetime64[ns]
```

Bengi; Gong; Sanjita

MANY THANKS