CDS 513 PREDICTIVE BUSINESS ANALYTICS

Presentation Title:

Predictive Analytics for Digital Commerce to improve Brand-level on Electronics and Buying Preference on Boutique Category

Presentation Date:

15 june 2020

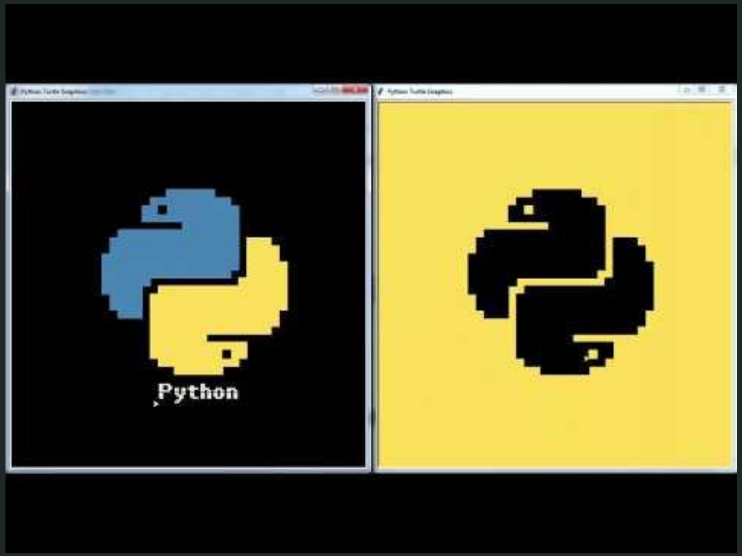
Group No. 7

1. CHEE SAI WAI
2. LIEW TIAN CHIN
3. SOO YIN YI

INTRODUCTION

Expanding in the segment drives economic growth to online retails companies producing consumer electronics and consumer products.

Businesses used complex software everything for demand planning can limit the insights into the appropriateness and effectiveness of different forecasting methods.




```
[ ]: df_raw2

[ ]: # Making each Invoice has an identical list
     trans = []
     for i in range(0, 28816):
         trans.append([str(df_raw2.values[i,j]) for j in range(0, 4632)])

     # Conveting it into an numpy array
     trans = np.array(trans)

     # checking the shape of the array
```

Market Basket Analysis (Extraction, Transformation, Load)

BESIDES RAPIDMINER...


```
Date =
VAR MinYear = 2010
VAR MaxYear = 2011
RETURN
ADDCOLUMNS (
    FILTER (
        CALENDARAUTO( ),
        AND ( YEAR ( [Date] ) >= MinYear, YEAR ( [Date] ) <= MaxYear )
    ),
    "Calendar Year", YEAR ( [Date] ),
    "Month Name", FORMAT ( [Date], "mmmm" ),
    "Month Number", MONTH ( [Date] ),
    "Weekday", FORMAT ( [Date], "dddd" ),
    "Week Number", WEEKNUM([Date]),
    "Weekday number", WEEKDAY( [Date] ),
    "Quarter", "Q" & TRUNC ( ( MONTH ( [Date] ) - 1 ) / 3 ) + 1
)
```

Time Series Forecasting (Extraction, Transformation, Load)

PROBLEM BACKGROUND

Number of goods and services that consumers will probably buy in the future

Retail demand likely to add volatility in local consumer preferences, due to their partial and imprecise models being used.

Optimizing product markdown by improving in-store merchandizing effectiveness.

Predicting the demand for the products of a particular brand or firm could see.

Retailers want to trend merchandizing effectiveness by item-based recommendation in store

- Business opportunity in online boutique products with supplement of existing customer’s transaction data.

BUSINESS DECISION/ PROBLEM STATEMENT

Retailers want to predict 6-month by brand and product sourcing in store.

- Fluctuation in near future by brand, week number, quarters, product category between Q3 to Q4 from year 2010 to year 2011.

MOTIVATION

Recommender system

- Cross-sell opportunity that is not easily noticed for the business owner.

MBA

- Seller aims to figure out the marketing strategies or the product which seems to be purchased together to improve the profit of the business.

Short –term forecasting

- The value of the data is to adequately forecast the size the trend of the change of point

Long–term forecasting

- The value of the data is to motivate the business by giving tactical information

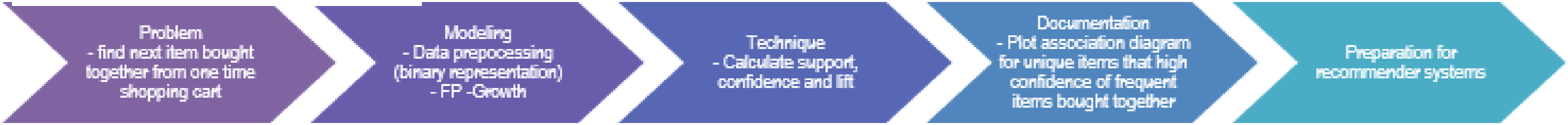
SCOPE and LIMITATION

Scope and Limitation

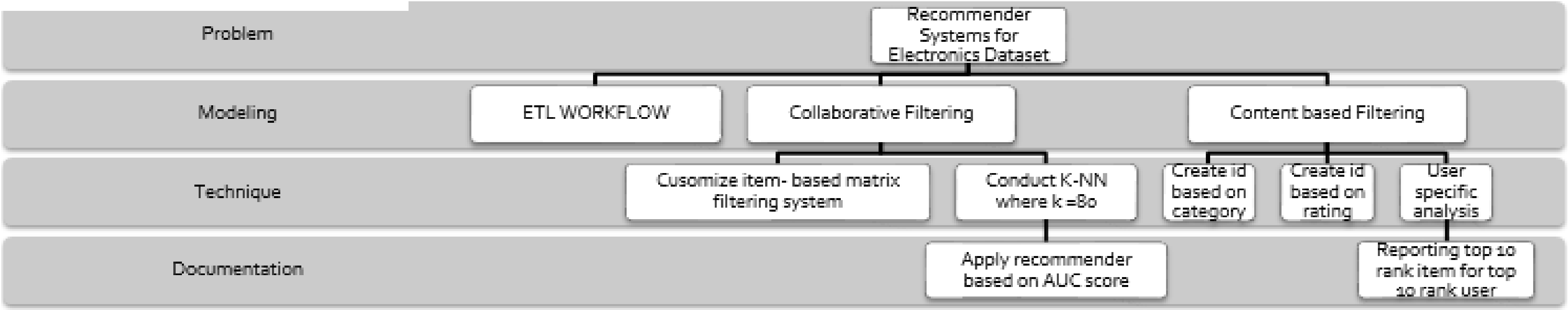
- New item is introduced into the system, this item may not have the same huge amount of history data as compare the previous item
- Sparsity of data and lead to demotivate the performance of the recommendation system.
- Noise from the times going, the time series data

FRAMEWORK

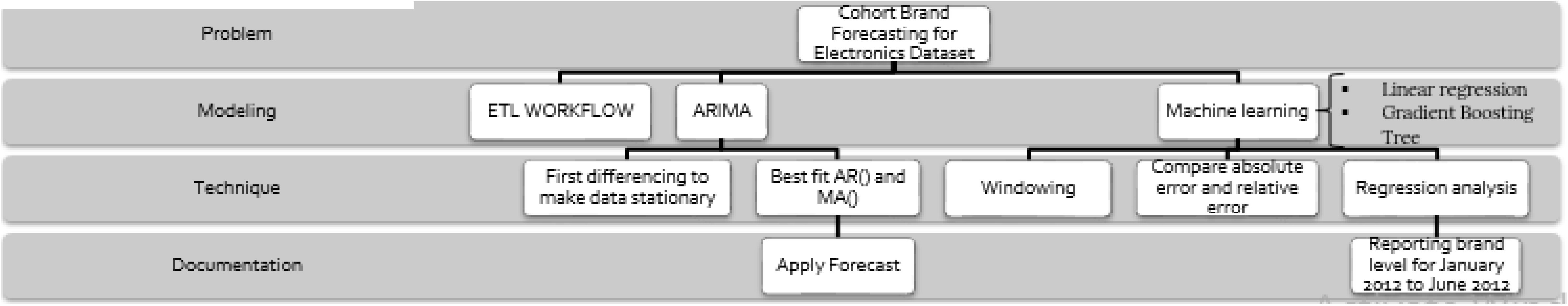
Market Basket Analysis Framework



Recommender Systems Framework



Time Series Forecasting Framework

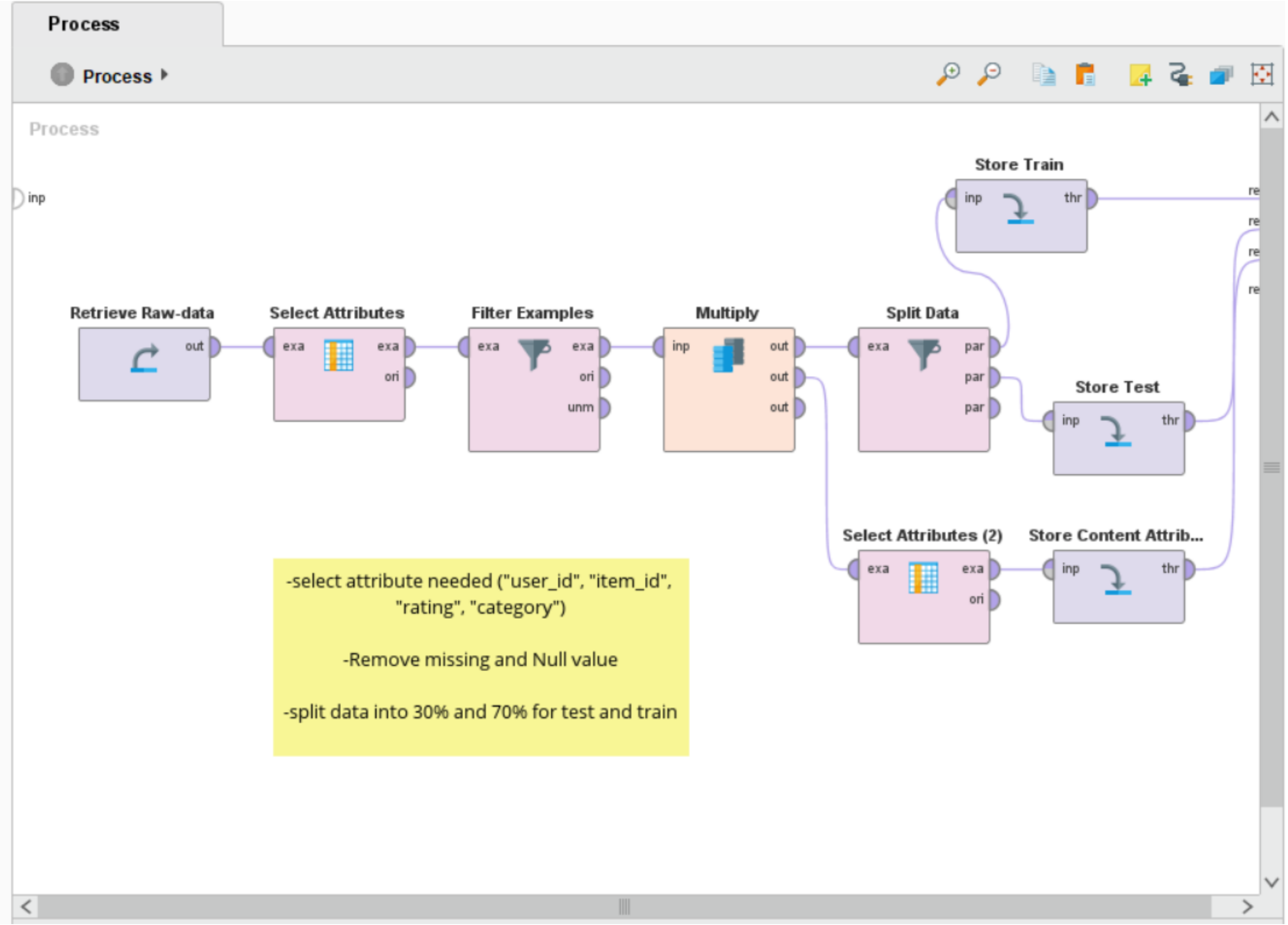


MARKET BASKET ANALYSIS

APPROACHES

Data Preprocessing

- Filter required Attribute “user_id”, “item_id” and “rating”
- Remove row with null value
- Split the data according to ratio 3:7 (Test:Train)

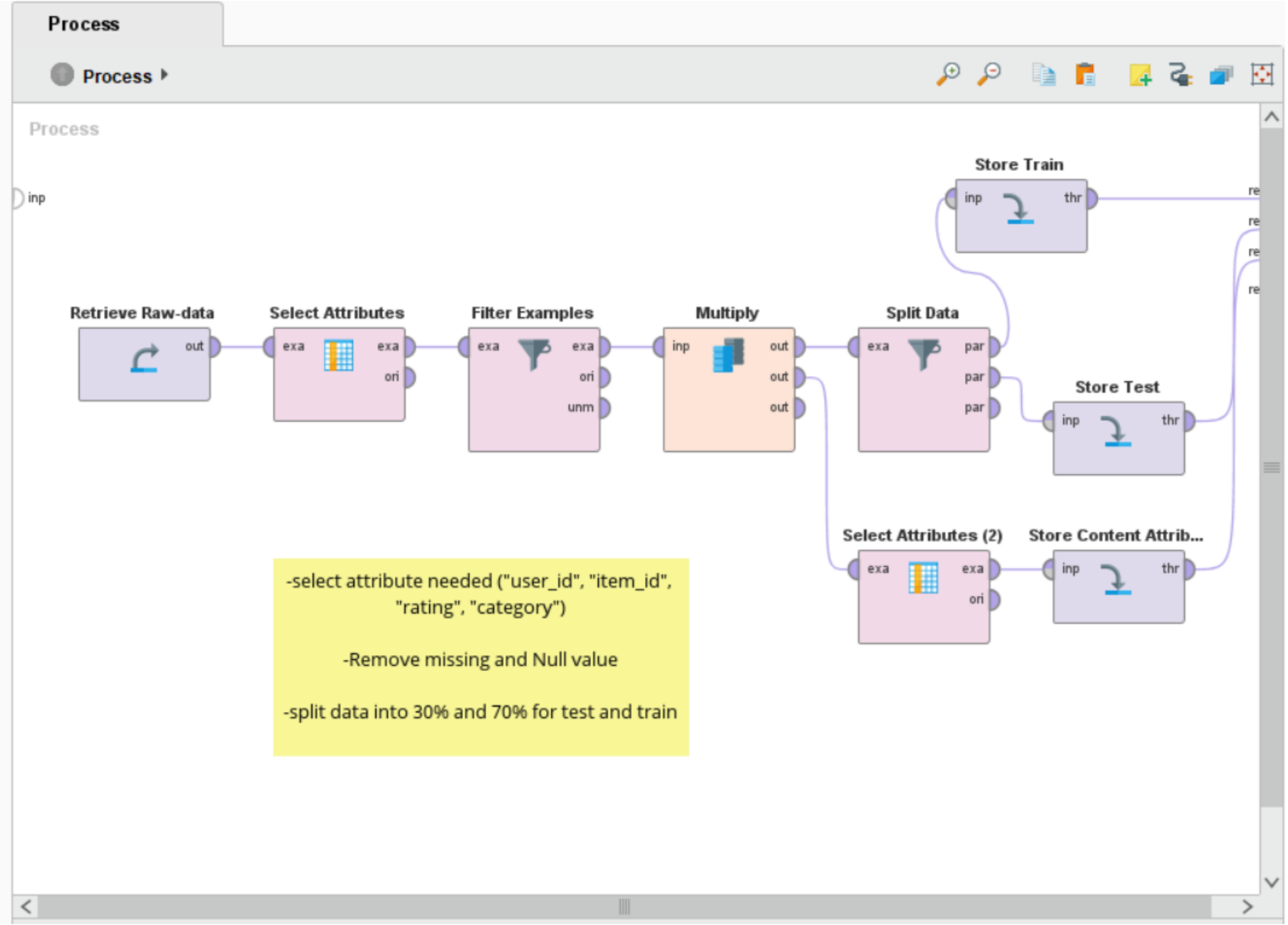


RECOMMENDER SYSTEMS

APPROACHES

Data Preprocessing

- Filter required Attribute “user_id”, “item_id” and “rating”
- Remove row with null value
- Split the data according to ratio 3:7 (Test:Train)



COLLABORATIVE FILTERING

Row No.	user_id	item_id	rank
1	767776	9557	1
2	767776	10	2
3	767776	9	3
4	767776	7	4
5	767776	8	5
6	767776	6	6
7	767776	4	7
8	767776	2	8
9	767776	3	9
10	767776	9556	10
11	753294	9558	1
12	753294	10	2
13	753294	9	3
14	753294	7	4
15	753294	8	5
16	753294	6	6
17	753294	4	7
18	753294	2	8

ExampleSet (3,728,680 examples, 0 special attributes, 3 regular attributes)

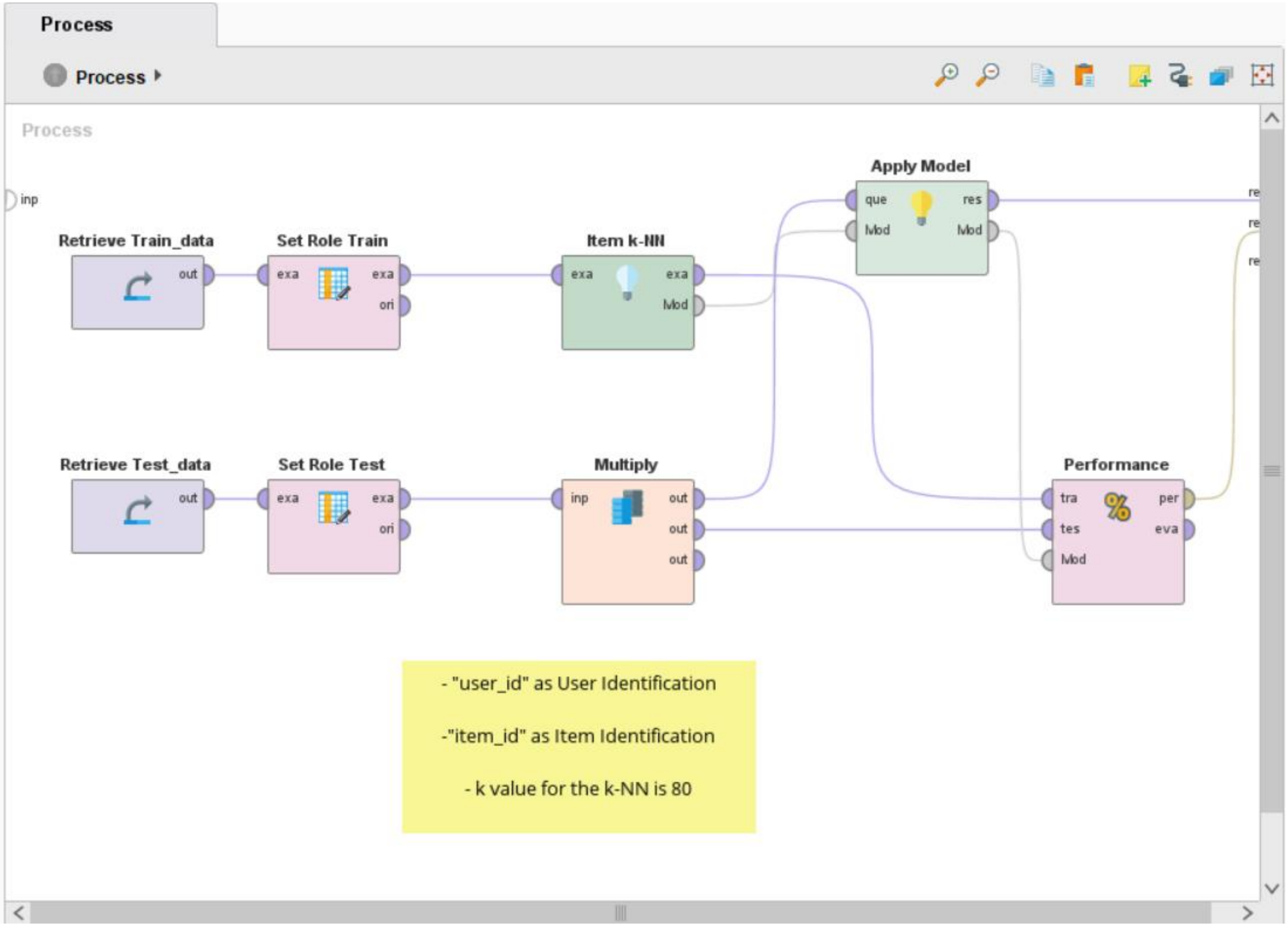
APPROACHES

COLLABORATIVE FILTERING

- “user_id” as user Identification
- “item_id” as item Identification
- K value of 80 for item k-NN

Parameter	Value
AUC	0.244
prec@5	0.000
prec@10	0.000
prec@15	0.000
NDCG	0.082
MAP	0.001

COLLABORATIVE FILTERING

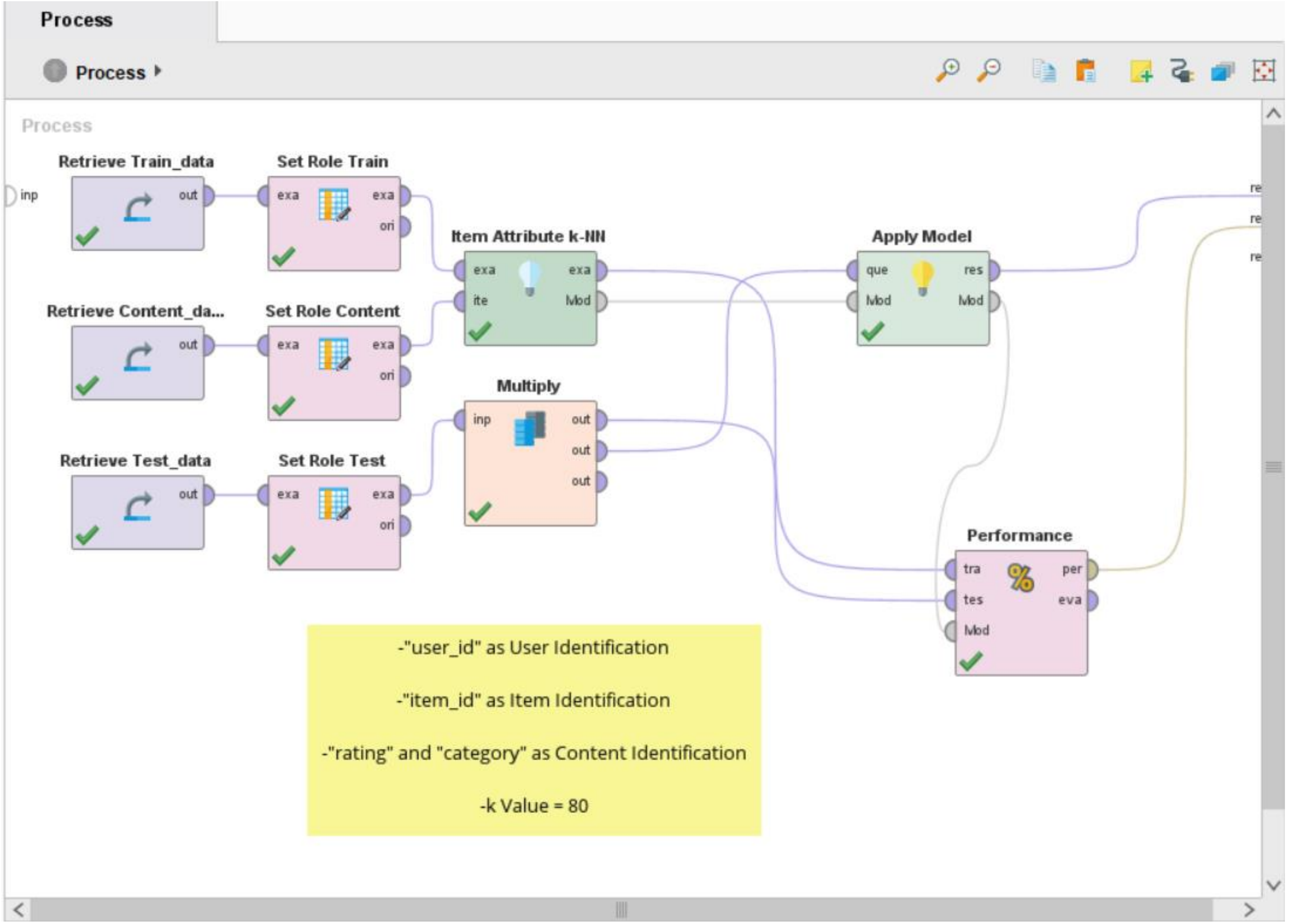


APPROACHES

CONTENT-BASED RECOMMENDATION

- “user_id” as user Identification
- “item_id” as item Identification
- Two content-based recommendation will be conduct
- “rating” and “category” as content identification
- K value of 80 for item k-NN

CONTENT BASED RECOMMENDATION



CONTENT-BASED RECOMMENDATION

- “rating” as content identification
- Parameter : 1, 2, 3, 4, 5

Parameter	Value
AUC	0.244
prec@5	0.000
prec@10	0.000
prec@15	0.000
NDCG	0.082
MAP	0.001

CONTENT-BASED RECOMMENDATION

- “rating” as content identification

Row No.	user_id	item_id	rank
1	767776	9558	1
2	767776	299	2
3	767776	2415	3
4	767776	2129	4
5	767776	1162	5
6	767776	4858	6
7	767776	6965	7
8	767776	8071	8
9	767776	54	9
10	767776	4161	10
11	767776	53	11
12	767776	52	12
13	767776	51	13
14	767776	2016	14
15	767776	5584	15
16	767776	50	16
17	767776	1124	17
18	767776	1784	18

ExampleSet (37,286,800 examples, 0 special attributes, 3 regular attributes)

CONTENT-BASED RECOMMENDATION

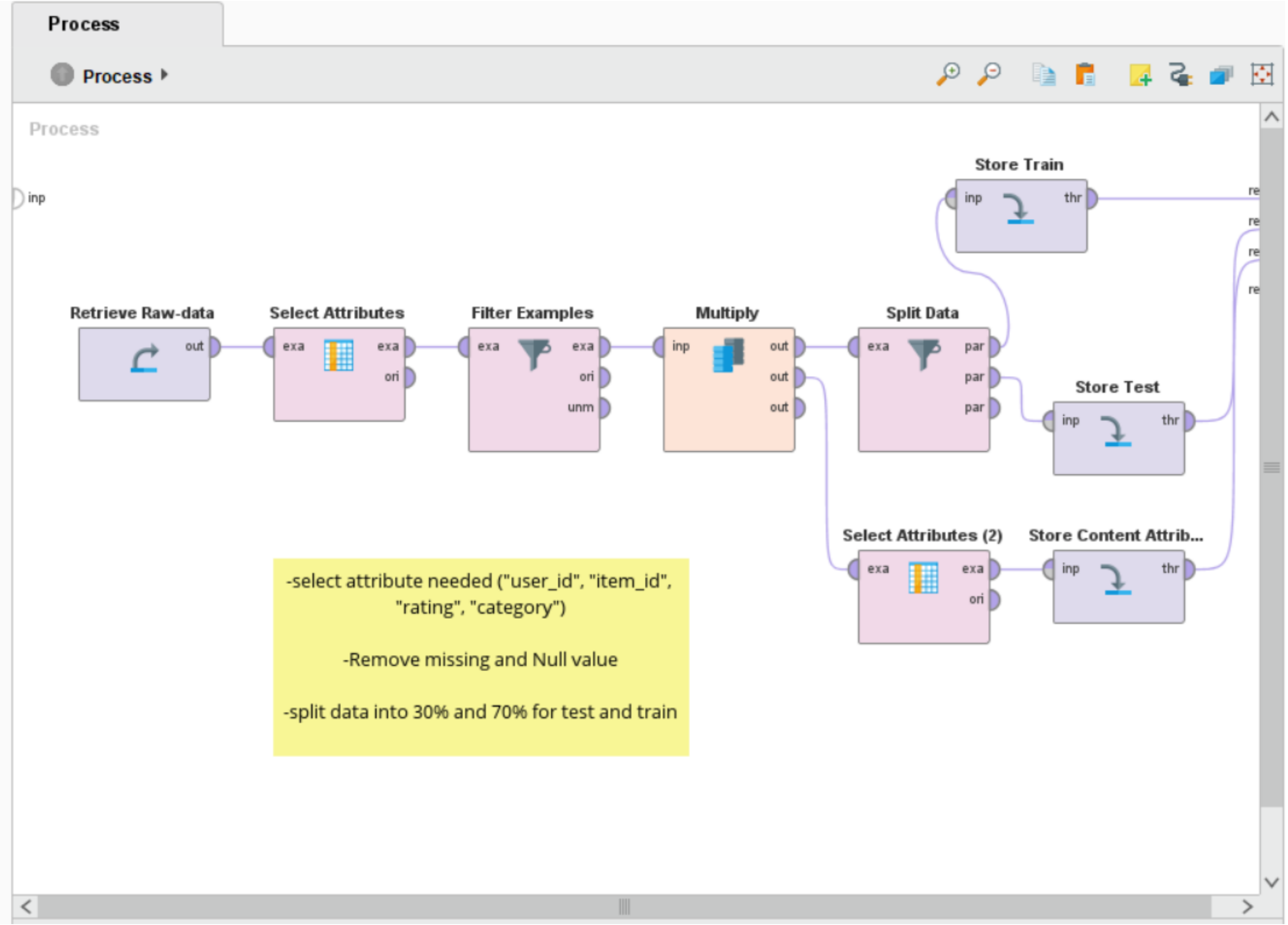
- “category” as content identification

Parameter	Value
AUC	0.232
prec@5	0.000
prec@10	0.000
prec@15	0.000
NDCG	0.080
MAP	0.000

APPROACHES

Data Preprocessing

- Filter required Attribute “user_id”, “item_id” and “rating”
- Remove row with null value
- Split the data according to ratio 3:7 (Test:Train)



COLLOBORATIVE FILTERING

Row No.	user_id	item_id	rank
1	767776	9557	1
2	767776	10	2
3	767776	9	3
4	767776	7	4
5	767776	8	5
6	767776	6	6
7	767776	4	7
8	767776	2	8
9	767776	3	9
10	767776	9556	10
11	753294	9558	1
12	753294	10	2
13	753294	9	3
14	753294	7	4
15	753294	8	5
16	753294	6	6
17	753294	4	7
18	753294	2	8

ExampleSet (3,728,680 examples, 0 special attributes, 3 regular attributes)

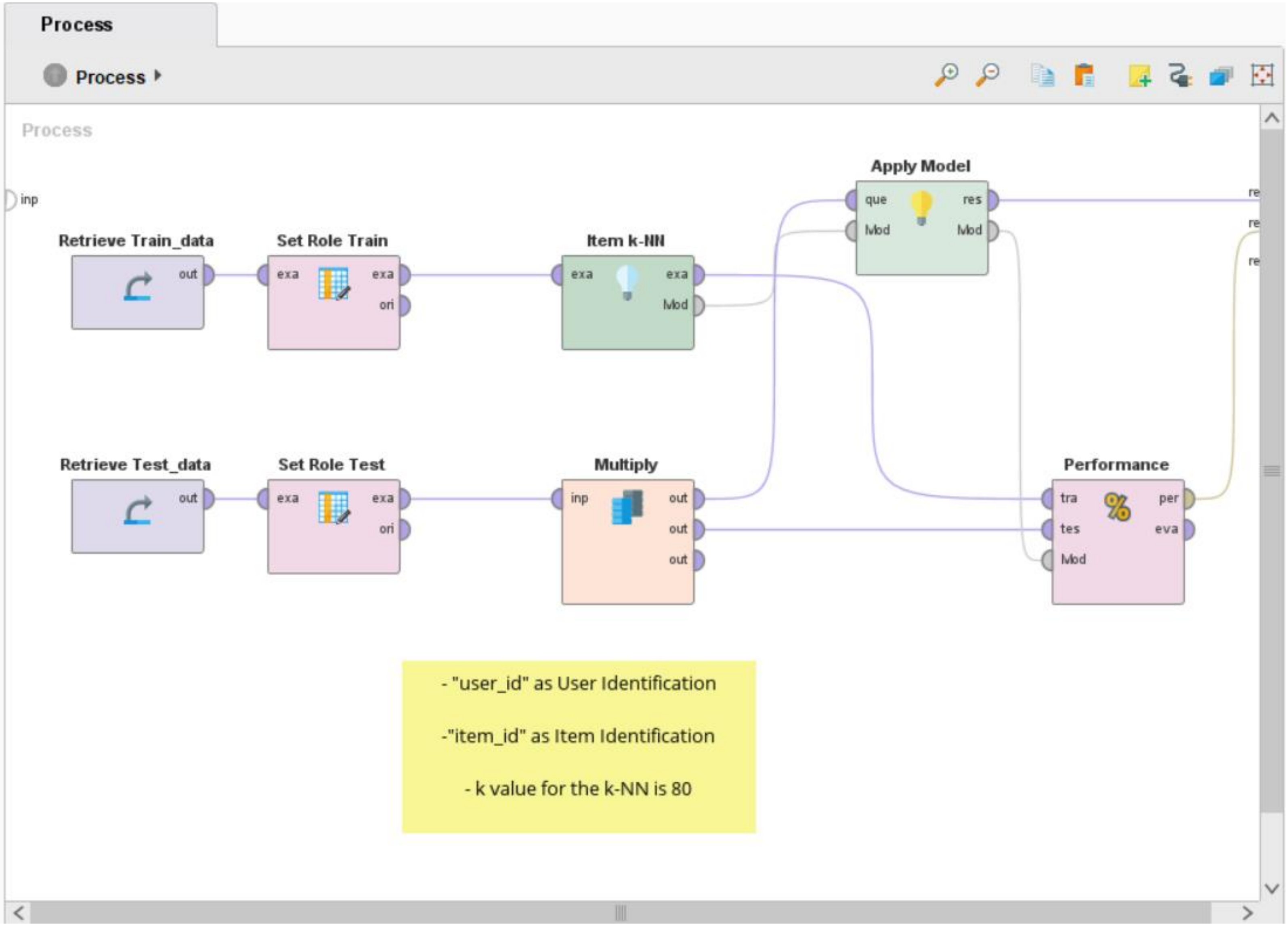
APPROACHES

COLLABORATIVE FILTERING

- “user_id” as user Identification
- “item_id” as item Identification
- K value of 80 for item k-NN

Parameter	Value
AUC	0.244
prec@5	0.000
prec@10	0.000
prec@15	0.000
NDCG	0.082
MAP	0.001

COLLABORATIVE FILTERING

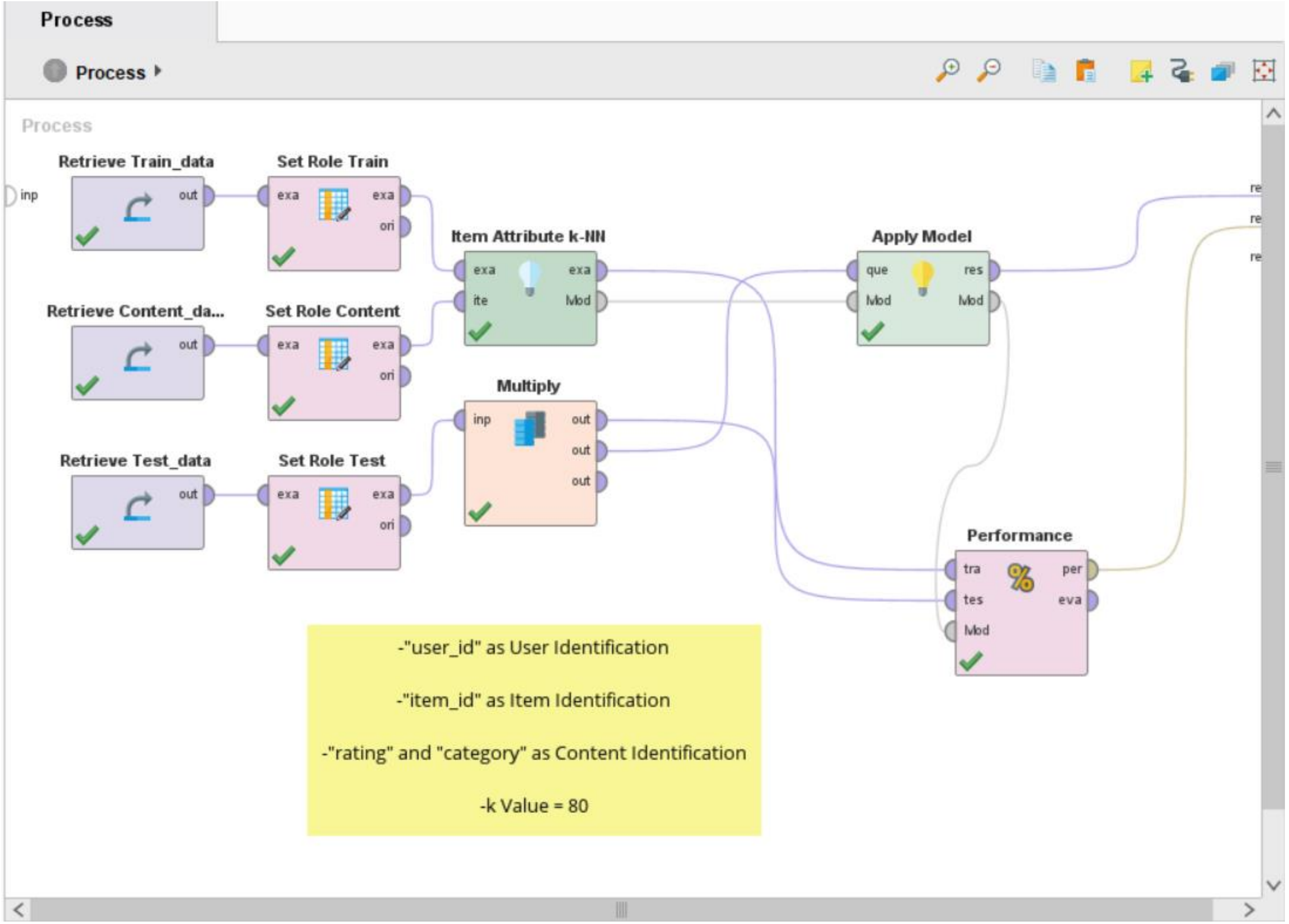


APPROACHES

CONTENT-BASED RECOMMENDATION

- “user_id” as user Identification
- “item_id” as item Identification
- Two content-based recommendation will be conduct
- “rating” and “category” as content identification
- K value of 80 for item k-NN

CONTENT BASED RECOMMENDATION



CONTENT-BASED RECOMMENDATION

- “rating” as content identification
- Parameter : 1, 2, 3, 4, 5

Parameter	Value
AUC	0.244
prec@5	0.000
prec@10	0.000
prec@15	0.000
NDCG	0.082
MAP	0.001

CONTENT-BASED RECOMMENDATION

- "rating" as content identification

Row No.	user_id	item_id	rank
1	767776	9558	1
2	767776	299	2
3	767776	2415	3
4	767776	2129	4
5	767776	1162	5
6	767776	4858	6
7	767776	6965	7
8	767776	8071	8
9	767776	54	9
10	767776	4161	10
11	767776	53	11
12	767776	52	12
13	767776	51	13
14	767776	2016	14
15	767776	5584	15
16	767776	50	16
17	767776	1124	17
18	767776	1784	18

ExampleSet (37,286,800 examples, 0 special attributes, 3 regular attributes)

CONTENT-BASED RECOMMENDATION

- “category” as content identification

Parameter	Value
AUC	0.232
prec@5	0.000
prec@10	0.000
prec@15	0.000
NDCG	0.080
MAP	0.000

CONTENT-BASED RECOMMENDATION

- “category” as content identification

Row No.	user_id	item_id	rank
1	767776	9558	1
2	767776	299	2
3	767776	2415	3
4	767776	2129	4
5	767776	1162	5
6	767776	4858	6
7	767776	6965	7
8	767776	8071	8
9	767776	54	9
10	767776	4161	10
11	767776	53	11
12	767776	52	12
13	767776	51	13
14	767776	2016	14
15	767776	5584	15
16	767776	50	16
17	767776	1124	17
18	767776	1784	18

Recommendation System Findings

- Outcome to recommend potential product to customer met, but the performance of the model is not satisfying.
- Content-based recommendation is aim to recommend potential product by user-specific classification where collaborative filtering is used when it is new user and doesn't have enough data for him
- User with id 76776 have the highest ranking in all recommendation because he have the most transaction history
- Solution : Increase the size of data set and reduce item type.

TIME SERIES FORECASTING

PREPARED BY P-COM-0145-19

ETL WORKFLOW

DATASET
ELECTRONICS

- Gather data updated data with an additional variable
- Pre-process the data
- Explore the data and decide on ...
- Run ...
- Analyze...and visualize results

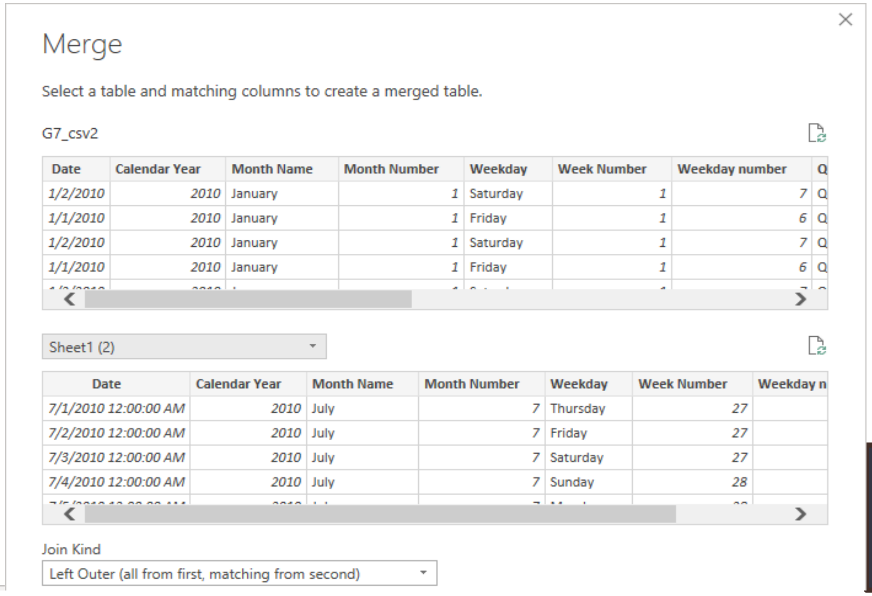
NEXT SLIDE»
«RETURN PREVIOUS SLIDE

Transform Original Data and store Extracted Data

Experiment : 1) Attributes timestamp we use is only two years (2010/2011).
2) Missing Values in 'Brand' is 23K.

Solution : Power query allows to restructure general times in per week frequency and rename timestamp within two years by left outer join later. Given a sequence of missing value more than 30% Rapidminer remove the next value in 'Brand' header.

```
Date =
VAR MinYear = 2010
VAR MaxYear = 2011
RETURN
ADDCOLUMNS (
    FILTER (
        CALENDARAUTO( ),
        AND ( YEAR ( [Date] ) >= MinYear, YEAR ( [Date] ) <= MaxYear )
    ),
    "Calendar Year", YEAR ( [Date] ),
    "Month Name", FORMAT ( [Date], "mmmm" ),
    "Month Number", MONTH ( [Date] ),
    "Weekday", FORMAT ( [Date], "dddd" ),
    "Week Number", WEEKNUM([Date]),
    "Weekday number", WEEKDAY( [Date] ),
    "Quarter", "Q" & TRUNC ( ( MONTH ( [Date] ) - 1 ) / 3 ) + 1
)
```



Date	Calendar Year	Month Name	Month Number	Weekday	Week Number	Weekday number	Q
1/2/2010	2010	January	1	Saturday	1	7	Q
1/1/2010	2010	January	1	Friday	1	6	Q
1/2/2010	2010	January	1	Saturday	1	7	Q
1/1/2010	2010	January	1	Friday	1	6	Q

Date	Calendar Year	Month Name	Month Number	Weekday	Week Number	Weekday n
7/1/2010 12:00:00 AM	2010	July	7	Thursday	27	
7/2/2010 12:00:00 AM	2010	July	7	Friday	27	
7/3/2010 12:00:00 AM	2010	July	7	Saturday	27	
7/4/2010 12:00:00 AM	2010	July	7	Sunday	28	

NEXT SLIDE»
«RETURN PREVIOUS SLIDE

Figure III-1: Load extracted data

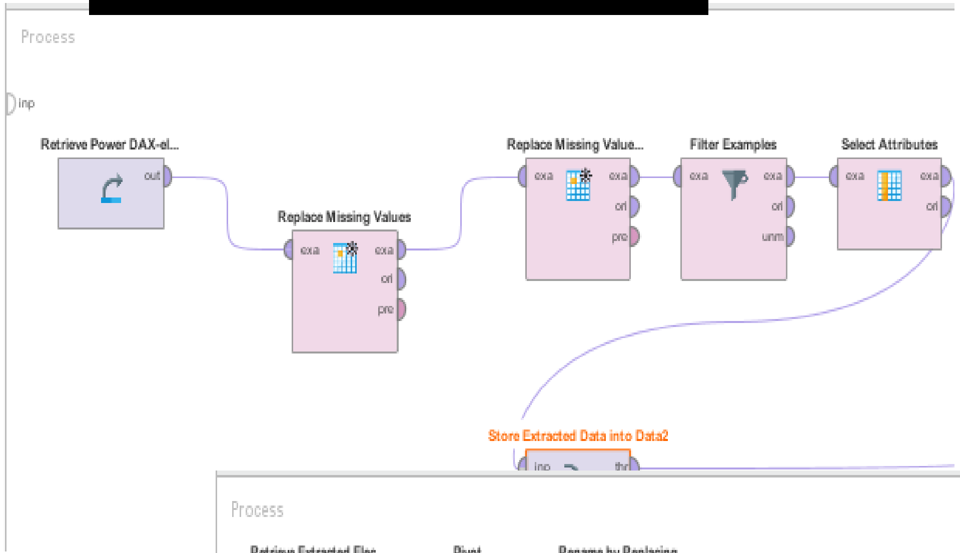


Figure III-3: Normalize data

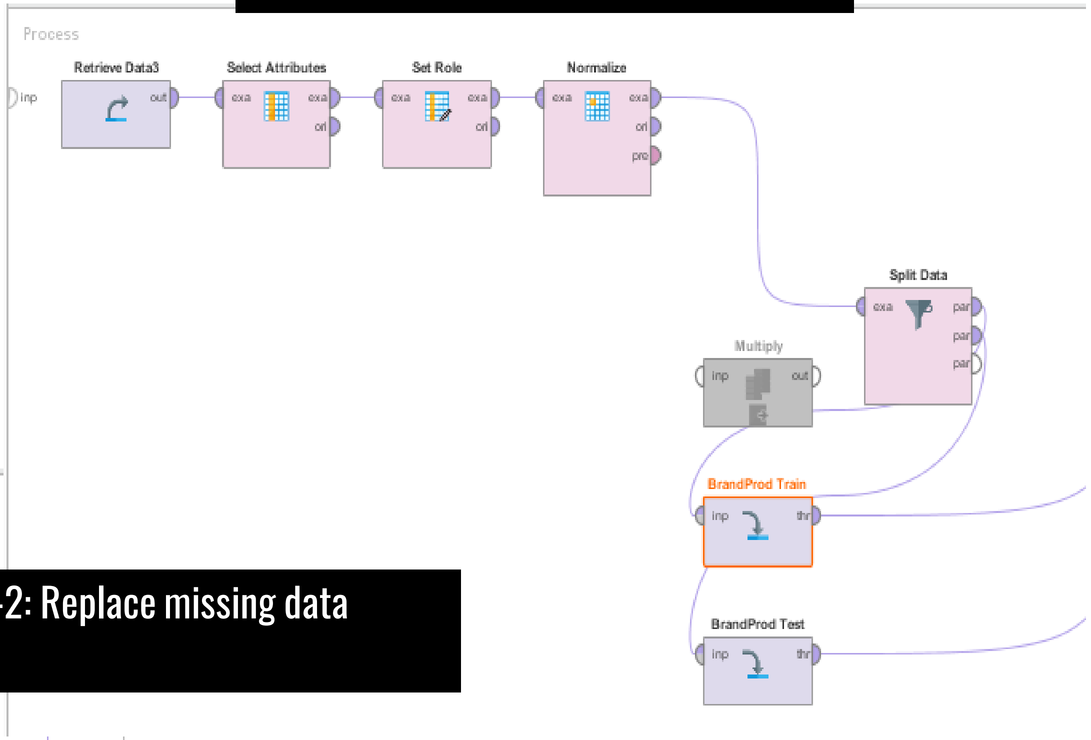
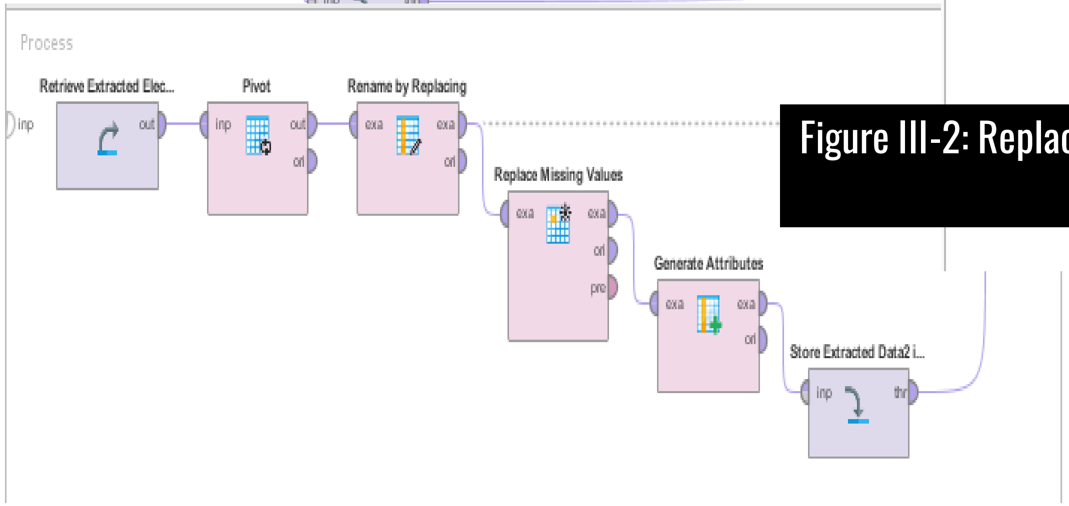
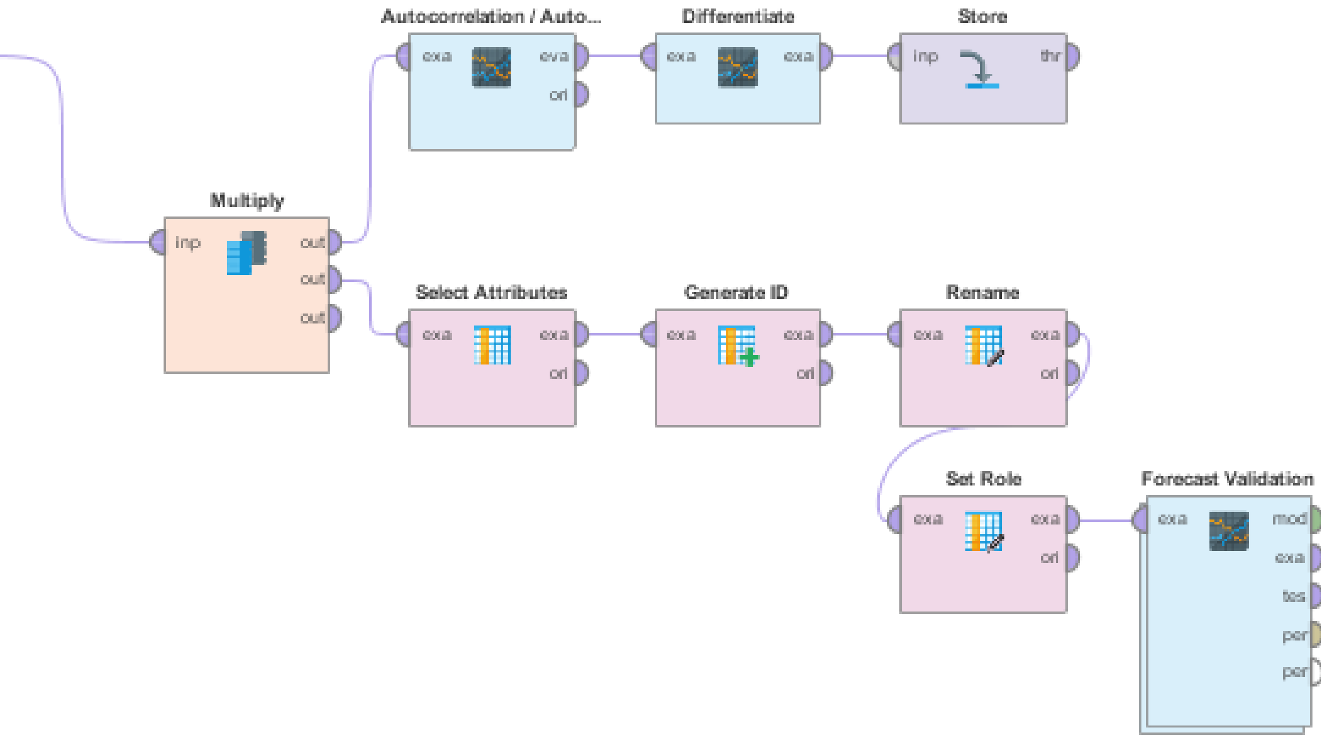
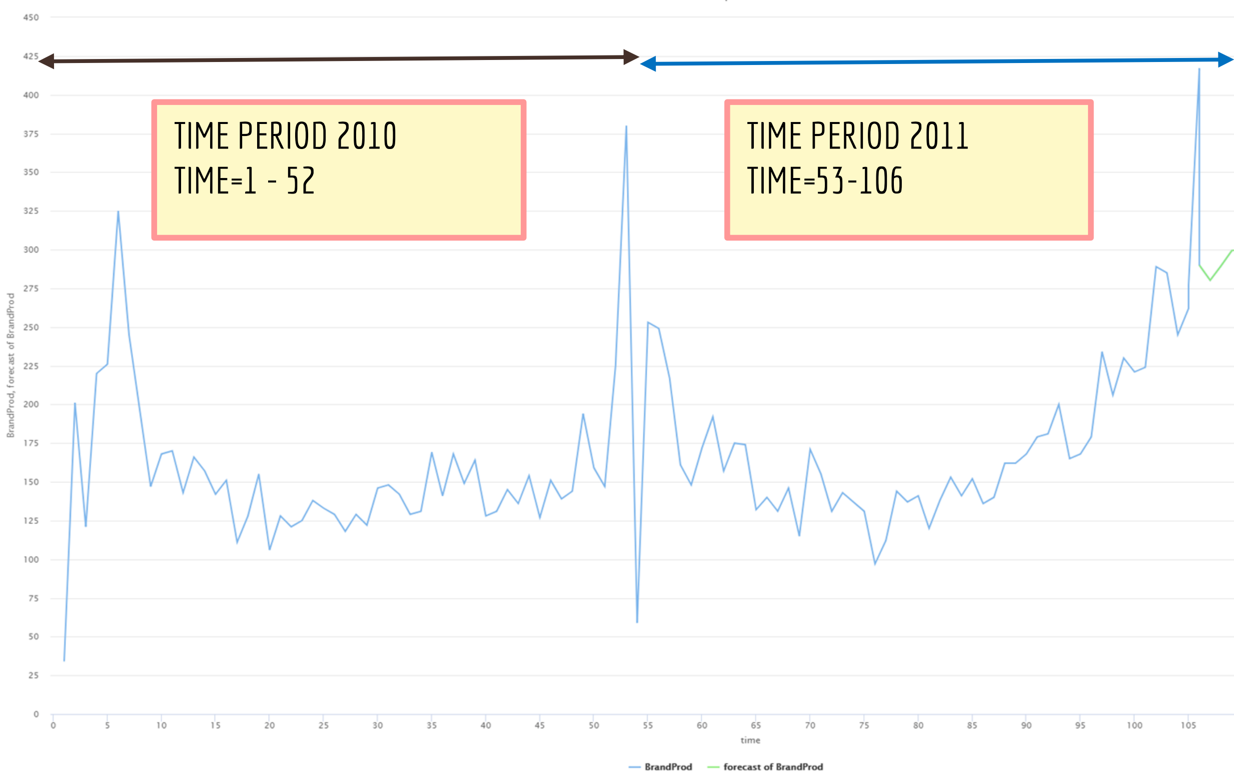


Figure III-2: Replace missing data



NEXT SLIDE»
«RETURN PREVIOUS SLIDE





PLOT TIME SERIES CHART

Experiment :The time plot about label 'BRANDPROD' shows some sudden changes, particularly the big drop in December 2010/2011. These changes are due to the beginning of seasonal holidays and the end of the holidays.

Solution : Trend analysis allows to identify general trends upward and downward. Given a sequence of events predict the next event(s).

NEXT SLIDE»
«RETURN PREVIOUS SLIDE

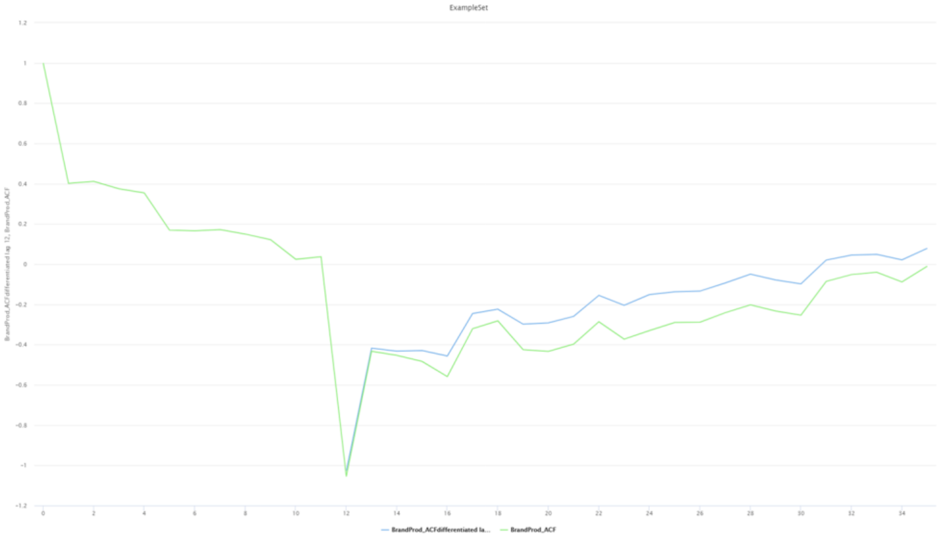
ARIMA MODEL PROCESS TO PLOT AND FORECAST

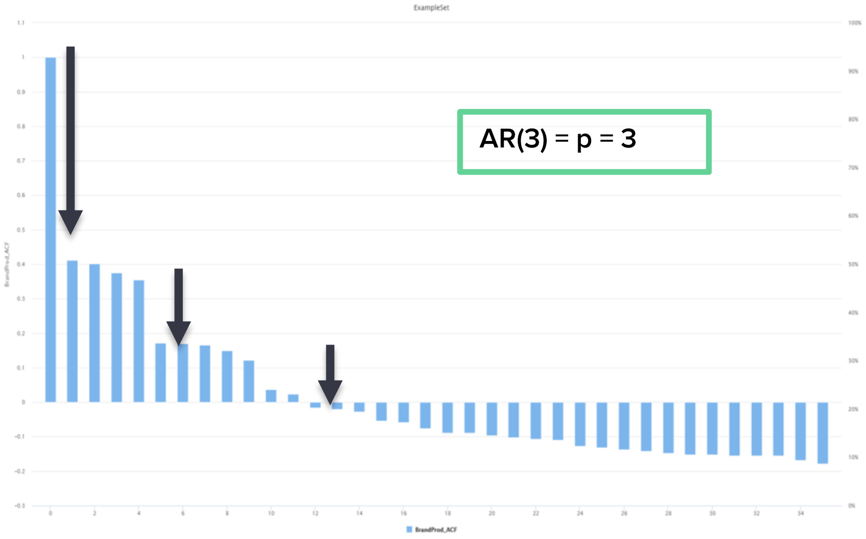
Experiment :

1) Difference data to make data stationary
2) Plot ACF that predicts p,d, q of the fitted model
3) Demand prediction remain any products sell quickly for say the least a six month period.

Solution :

1) The data are clearly non-stationary, as the series wanders up and down for long periods.
2) Create a plot is order of the autoregression (3), degree of differencing (0), and order of moving-average (0).
3) Short-term forecasting suggests making forecasts for a 24 units of time, such as outer join forecast and historical electronics data.





NEXT SLIDE»
«RETURN PREVIOUS SLIDE

ARIMA MODEL PROCESS TO PLOT AND FORECAST

Experiment :

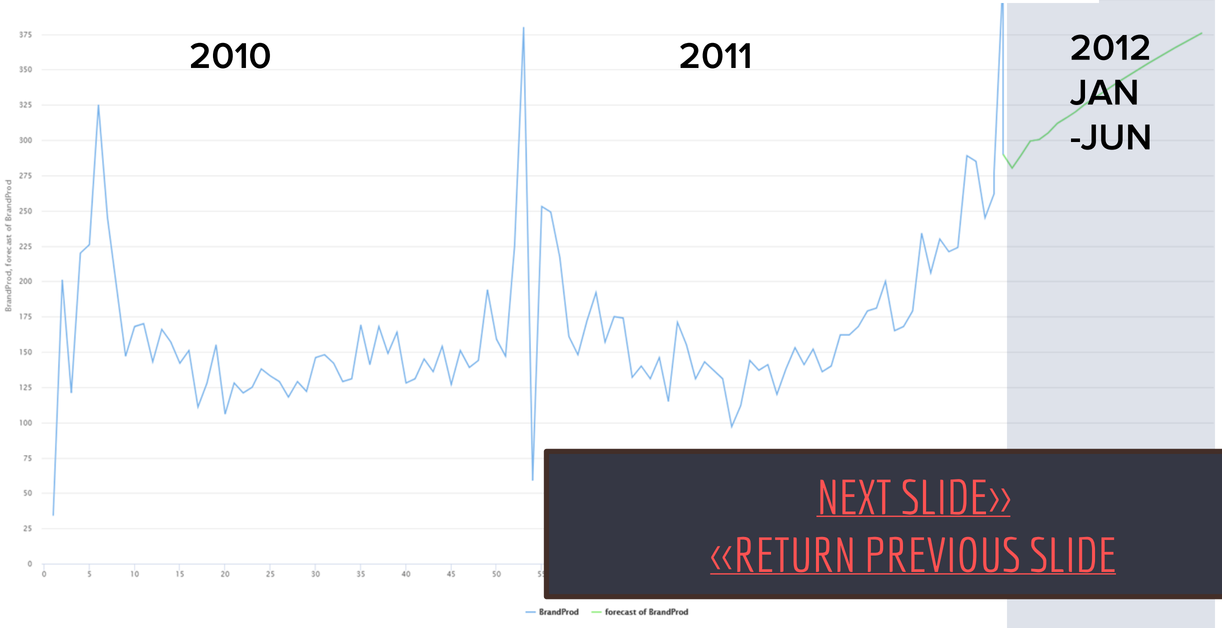
1) Difference data to make data stationary
2) Plot ACF that predicts p,d, q of the fitted model
3) Demand prediction remain any products sell quickly for say the least a six month period.

Solution :

1) The data are clearly non-stationary, as the series wanders up and down for long periods.

2) Create a plot is order of the autoregression (3), degree of differencing (0), and order of moving-average (0).

3) Short-term forecasting suggests making forecasts for a 24 units of time, such as outer join forecast and historical electronics data.



NEXT SLIDE»
«RETURN PREVIOUS SLIDE

Last time in ...	BrandProd +...	BrandProd - 6	BrandProd - 5	BrandProd - 4	BrandProd - 3	BrandProd - 2	BrandProd - 1	BrandProd - 0
7	196	34	201	121	220	226	325	245
8	147	201	121	220	226	325	245	196
9	168	121	220	226	325	245	196	147
10	170	220	226	325	245	196	147	168
11	143	226	325	245	196	147	168	170
12	166	325	245	196	147	168	170	143
13	157	245	196	147	168	170	143	166
14	142	196	147	168	170	143	166	157
15	151	147	168	170	143	166	157	142
16	111	168	170	143	166	157	142	151
17	128	170	143	166	157	142	151	111
18	155	143	166	157	142	151	111	128

Windowing results to perform machine learning model for Linear regression and GBT.

- COMPARE BETWEEN ALL THE MODELS AND PERFORM SOME ANALYSIS
- Experiment : Windowing generates set of attributes into cross sectional data Random sampling will make an each data of sample forecast to a fold. Determine how accurate the forecast and ARIMA (3,0,0).
- Solution : Linear regression and GBT classifiers USE windowing to forecast unit according to prediction and actual classes of brand electronics. Compare absolute error, relative error, and Root mean square error (RMSE).

NEXT SLIDE»
«RETURN PREVIOUS SLIDE

		Parameter	Actual brand produced, y	Slope of regression, h	intercept	$Prediction, \hat{Y}$
1	BrandProd2	M5 Prime	250 (time, t=106)	0.328	10.448	303.09
2	BrandProd3	M5 Prime		0.310	10.448	300.75
3	BrandProd2	Greedy		0.356	16.277	312.56
4	BrandProd3	Greedy		0.350	16.277	311.78

Forecast of BrandProd	Linear Regression		Gradient Boosting Tree			
	M5 PRIME	GREEDY	Trees =100; Learning rate=0.01	Trees =1000; Learning rate=0.01	Trees =100; Learning rate=0.001	Trees =1000; Learning rate=0.001
Absolute error (%)	27.75	27.205	26.085	30.430	33.444	26.090
Relative error (%)	18.51	18.23	16.82	19.43	20.45	16.82

Absolute and relative error and regression output results to illustrate a set of parameters for Linear regression.

$\hat{y}(prediction)$ - y (actual) = ht $+$ $intercept$

- COMPARE BETWEEN ALL THE MODELS AND PERFORM SOME ANALYSIS
- Experiment (LINEAR REGRESSION M5 ALGORITHM) : A store manager by Amazon is analyzing cohort brand demand forecasting data and its relationship to brand names and time. **Brand names rises as individual's time increases.**
- Solution : At the same time when June, 2012, the store manager will finally able to stock for say at least 6 months based on regression analysis and cohort brand forecasting with **304 and 301 stock** units followed by BrandProd2 and BrandProd3, respectively.

NEXT SLIDE»
«RETURN PREVIOUS SLIDE

		Parameter	Actual brand produced, y	Slope of regression, h	intercept	$Prediction, \hat{Y}$
1	BrandProd2	M5 Prime	250 (time, t=106)	0.328	10.448	303.09
2	BrandProd3	M5 Prime		0.310	10.448	300.75
3	BrandProd2	Greedy		0.356	16.277	312.56
4	BrandProd3	Greedy		0.350	16.277	311.78

Forecast of BrandProd	Linear Regression		Gradient Boosting Tree			
	M5 PRIME	GREEDY	Trees =100; Learning rate=0.01	Trees =1000; Learning rate=0.01	Trees =100; Learning rate=0.001	Trees =1000; Learning rate=0.001
Absolute error (%)	27.75	27.205	26.085	30.430	33.444	26.090
Relative error (%)	18.51	18.23	16.82	19.43	20.45	16.82

Absolute and relative error and regression output results to illustrate a set of parameters for Linear regression.

$\hat{y}(prediction)$ - y (actual) = ht $+ \, intercept$

- COMPARE BETWEEN ALL THE MODELS AND PERFORM SOME ANALYSIS
- Experiment (LINEAR REGRESSION GREEDY ALGORITHM) : A store manager by Amazon is analyzing cohort brand demand forecasting data and its relationship to brand names and time. **Brand names rises as individual's time increases.**
- Solution At the same time when June, 2012, the store manager will finally able to stock for say at least 6 months based on regression analysis and brand forecasting with **313 and 312 stock** units followed by BrandProd2 and BrandProd3

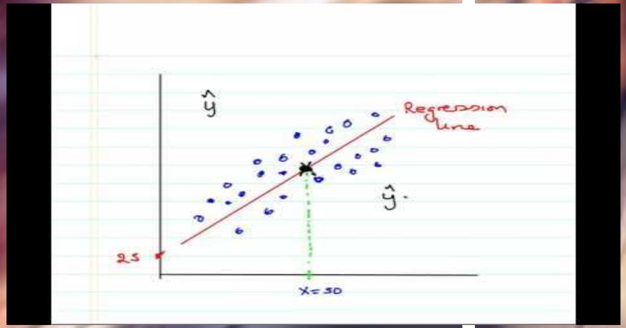
NEXT SLIDE»
«RETURN PREVIOUS SLIDE

Market basket Analysis

A table of item description based on “StockCode” is recorded in Appendix II. 85099B (JUMBO BAG RED RETROSPOT) and 85099C (JUMBO BAG BAROQUE BLACK WHITE) are two relatively strong attributes that associate with many other attributes. Both made up 31% of all the rules.

Recommender Systems

Significant weakness for both collaborative filtering and content-based recommendation in this dataset is the limited data analysis. Top 10 rank for three of these models which are collaborative filtering, rating based recommendation and category-based recommendation are all recommended around this specific user.



Time Series Forecasting

Regression analysis will also yield more informative results as it demonstrates the impact of more than one ARIMA (3,0,0) to a good forecast. Sign of the R-squared is 0.897, it means that low p-value and high confidence of regression. This version of Gradient Boosting Tree is also generally cheaper and quicker to implement than the Neural Network (NN) such as number of trees or learning rate

Conclusion and Future Works

NEXT SLIDE>>
<<RETURN PREVIOUS SLIDE

Sources

- WEB.
- http://www.real-statistics.com/time-series-analysis/arima-processes/comparing-arima-models/

- ACADEMIC PAPER.
- Chatfield, C. (2000). ***TIME-SERIES FORECASTING***.
  Web Services, A. (2020). ***Time Series Forecasting Principles with Amazon Forecast Technical Guide***.

NEXT SLIDE>>
<<RETURN PREVIOUS SLIDE

Q & A