ORACLE

Sales Forecasting System Development

Modeling results and Evaluation

Minseok Oh
Principal Consultant

Sales Forecasting System Development Project

- Solution : ADW - Oracle ML, OAC
- Machine Learning Model – Prescription Drug Sales Forecast
- By analyzing data (EDA, correlation analysis, etc.) using internal/external data, the variables that affect sales of prescription drugs are identified, and the sales prediction model is implemented through machine learning algorithms.
- Visualize the forecast results using OAC, load it into SAP and use it for sales activities

Achieved 90% forecast result accuracy for Prescription Drug Sales Forecast

Key outcomes

- Minimize Human judgement by using machine learning to predict Prescription Drug Sales trend
- Improved data management efficiency and reduced human error through integrated data platform
- Identify sales trends of prescription drugs. Through this, it is utilized for preemptive response through sales activities.

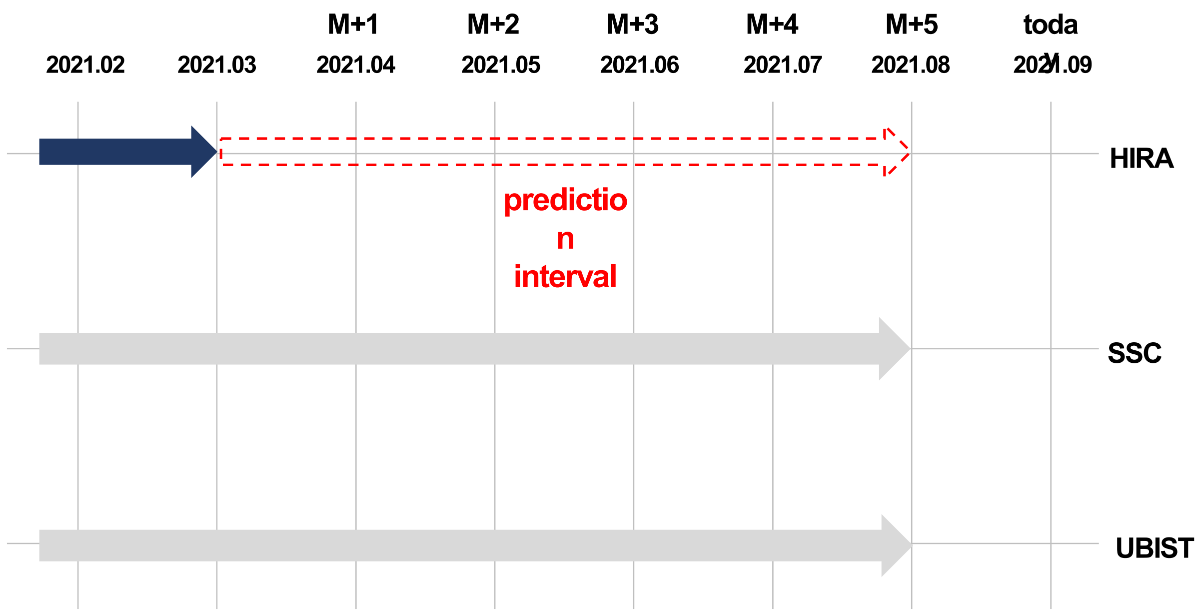
Prediction Result

- Using Regression Algorithm to predict Prescription Drug Sales for the coming 5 Months
- Achieved 90% forecast result accuracy - Through knowledge transfer, administrators continue to improve model performance after the project.

1.1 Model development goal and step-by-step task

☑ To predict Company’s sales performance by product unit route and person in charge, we develop a machine learning model using HIRA data as reference data, and aim to derive HIRA data prediction values from the previous month using data from 5 to 6 months ago. The model development steps are as follows, and the modeling and evaluation processes are repeated.

Model development goal

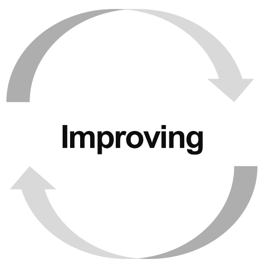


“ Company's business performance forecast by product unit route and person in charge ”

- Forecasting business performance by 250 cities and counties
- Forecasting business performance by 750 cities, counties and districts
- Forecasting sales performance by city, county, type, and product (PH5 level)

Step-by-step TASK

Data preparation	Modelling	Evaluation
Creating a foundation data set	Choosing a modeling technique	Model validation
Split foundation data set	Model test plan design	Model performance evaluation
	Create model	Evaluation of the modeling process
	Model training	Model applicability evaluation



1.2 Considerations and development direction before model development

☑ HIRA, UBIST, and SSC data before modeling and by identifying the characteristics Based on this, we developed a machine learning model .

Considerations before modeling

HIRA data

- Previous month forecast based on HIRA data from 5 to 6 months ago
- Data is unstable due to missing claims from nursing homes
  - In particular, in the case of Uijeongbu and hospitals, frequent omission of claims occurs (about 3-7% per month)
  - In the case of medical, frequent billing omissions in certain regions
  - Due to the discrepancy between the company's data accumulation period (2013.01~2021.04) and the other company's data accumulation period (2020.01~2012), it is difficult to build a model that integrates the company and other companies
  - Third-party data accumulation times are short, and stable patterns are suspected for use in predictive models, and performance verification is limited when models are integrated with third-party companies

UBIST Data

- Data accumulation period (2016.01 to 2020.08)
- Difference between HIRA data and UBIST data occurs depending on region

- ex (based on HIRA) UBIST prediction value accuracy: Gwangju Nam-gu (98%) > Gangnam-gu, Seoul (82%)

SSC Data

A large amount of shipments occur in a specific month and region, causing deviation from HIRA data - with irregular forwarding patterns

Model development direction

- Utilize time series algorithms with stable 5-6 month prediction

- Analysis of outlier patterns to learn the model after outlier processing

- Create separate models for our company and third-party companies

- If the company/third party data storage period becomes the same in the future, it can be implemented as an integrated model

- UBIST and SSC's specific areas where HIRA performance amount and error occur less are used as training data

2.1 Exponential Smoothing – Model Creation and performance evaluation method

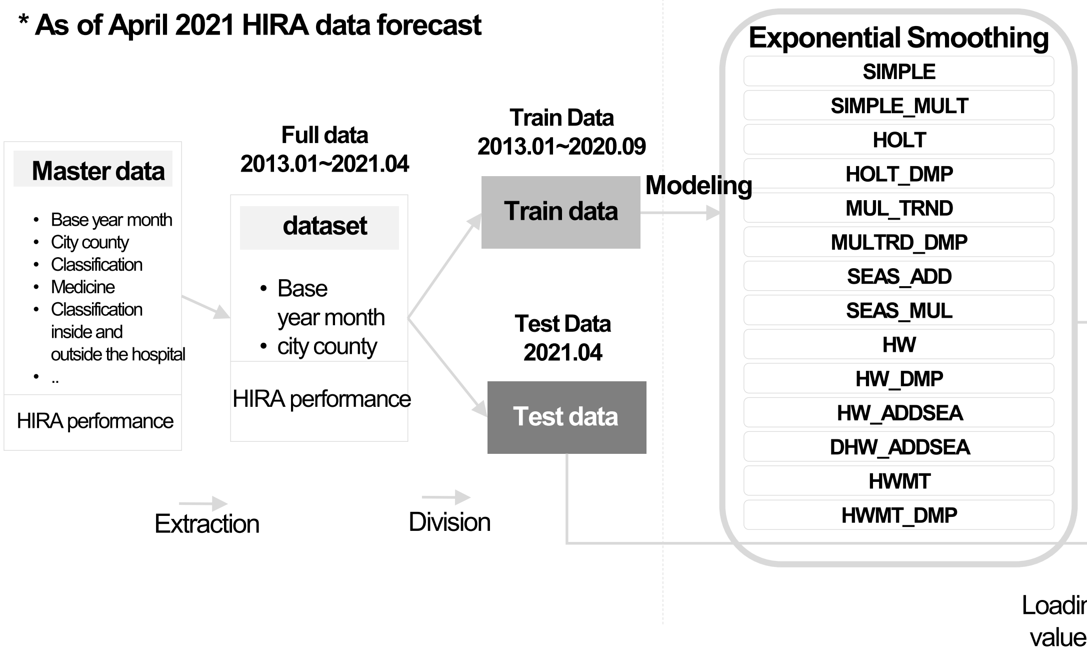
☑ To predict the amount of HIRA sales performance after 5 months, 14 models of Exponential Smoothing, a time series algorithm, are used to generate models and evaluate the performance of each model.

Data extraction and segmentation	Create and train models	Model performance evaluation
• Create master data based on HIRA data and process data according to predictive purpose • Purpose of prediction: Forecasting business performance by city, county, city, county, and type, city, county, and product • Splitting Train Data Set, Test Data Set	• Machine learning is performed using data from 5 months prior to the forecast year as training data • Generating models and deriving predicted values	• The value of the predicted year and month derived from the algorithm is compared with the HIRA data to measure the prediction error and evaluate it using performance indicators • Utilization Performance Indicators MAPE * MAPE: Absolute value (actual value-predicted value)/actual value*100

Model creation and performance evaluation

Example

*** As of April 2021 HIRA data forecast**



*** MAPE : ABS(actual value - predicted value) / actual value* 100**

ARA N.M.	Predicted value by model			2021.02 Actual value	SIMPLE model MAPE	HOLT model MAPE	…
	SIMPLE	HOLT	…				
Gangneung–si	91,266,830	91,350,488	…	87472623	4.34	6.4	…
Goseong–gun	986,865	986,723	…	1192832	17.27	7.90	…
Donghae–si	16,557,016	16,695,591	…	14455591	14.54	4.14	…
Samcheok–si	18,918,257	18,918,763	…	20595334	8.14	15.2	…
…	…	…	…	…	10.24	13.4	…

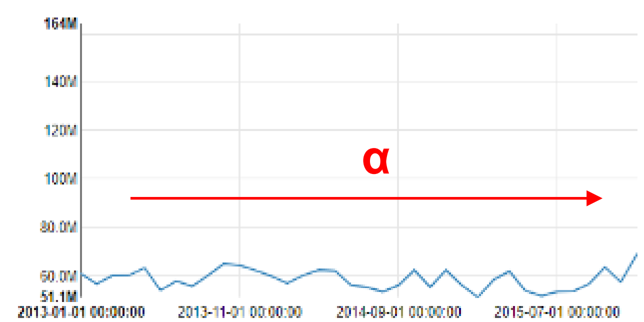
Performance evaluation by model

[Reference] Exponential Smoothing - Algorithm Explanation

✔ We generate models using Exponential Smoothing, a time series algorithm that can be predicted for a long time. (5-6 month data utilization), Exponential Smoothing is an algorithm that predicts future values by adding large weights to recent data and decreasing weights exponentially as they go back to the past

Exponential Smoothing Overview

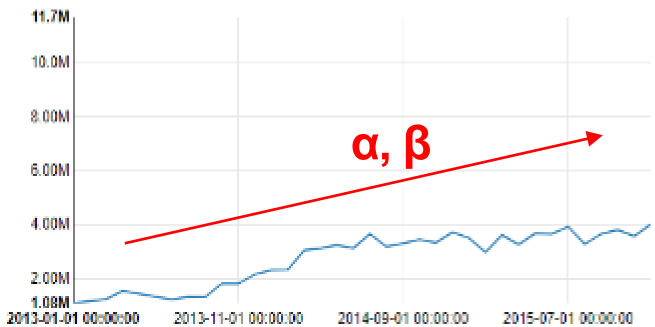
• **Facilitates time series data to stabilize graphs and capture patterns to increase predictive power**



Simple exponential smoothing model

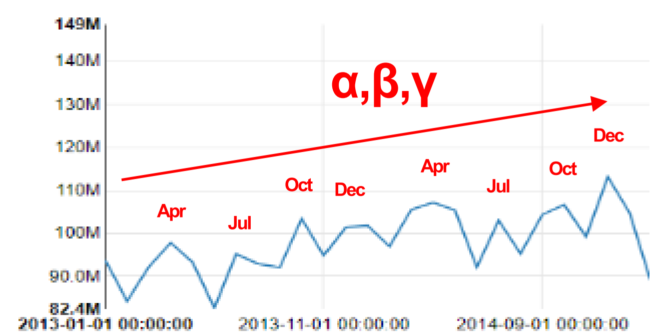
- Model suitable for predicting time series without trend or seasonality
- Prediction model: $F_{t+1} = \alpha y_t + \alpha(1-\alpha)y_{t-1} + \alpha(1-\alpha)^2 y_{t-2} + \dots$

Year	Actual value	Weight (α= 0.8)	Forecasts
2021			**11.04**
2020	11	0.8	**8.8** = 0.8*11
2019	12	0.16 = 0.8 * (1−0.8)	**1.92** = 0.16*12
2018	10	0.032 = 0.8 * (1−0.8) * (1−0.8)	**0.32** = 0.032*10



Double exponential smoothing model

- A model that reflects trends by expanding the simple exponential smoothing model
- The model is created using the smoothing factor α and the trend slope value β created
  - Prediction model (additional): $F_{t+k} = L_t + kt$
  - $L_t = \alpha y_t + (1-\alpha)(L_{t-1} + T_{t-1})$
  - $T_t = \beta(L_t - L_{t-1}) + (1-\beta) T_{t-1}$



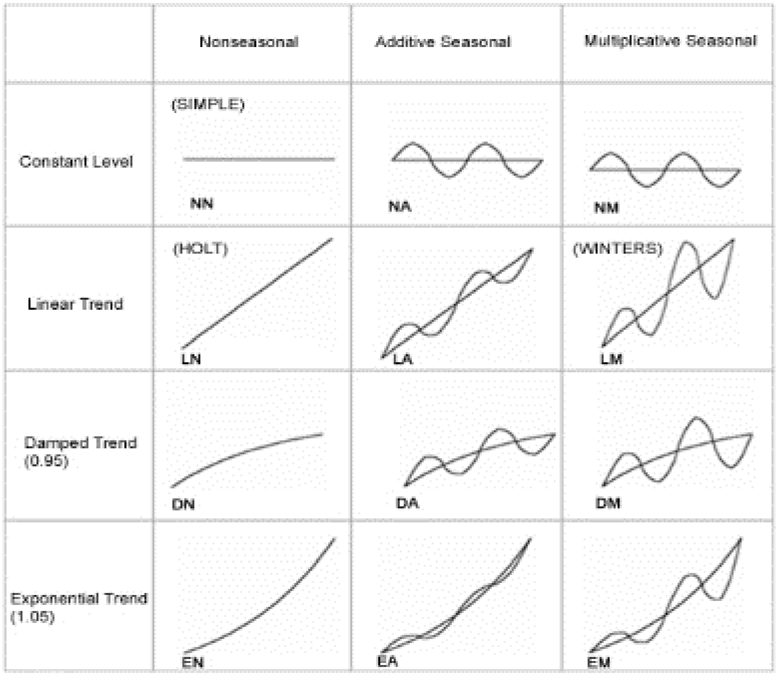
Triple exponential smoothing model

- A model that reflects trends and seasonality by extending the simple exponential smoothing model
- Create a model with the smoothed coefficient α and the seasonal parameter γ in the trend slope β
  - Predictive model (additive): $F_{t+k} = (L_t + kT_t) S_{t+k-m}$
  - $L_t = \alpha y_t / S_{t-m} + (1-\alpha)(L_{t-1} + T_{t-1})$
  - $T_t = \beta(L_t - L_{t-1}) + (1-\beta) T_{t-1}$
  - $S_t = \gamma(y_t / L_t) + (1-\gamma) S_{t-m}$

[Reference] Exponential Smoothing - Model

☑ The OML Exponential Smoothing algorithm provides 14 models depending on the time series type, and the accuracy of the prediction can be improved by utilizing suitable models.

Classification of models by time series type



- **N: constant model (Constant)**
- **A : Additive model**
- **Ad: Damped Additive**
- **M : Multiplicative model**
- **Md : Damped Multiplicative model**

OML – Model provided by Exponential Smoothing algorithm

Division	Model name	Explanation
NN	EXSM_SIMPLE	Simple exponential smoothing model
	EXSM_SIMPLE_MULT	Simple exponential smoothing model with multiplicative error
LN	EXSM_HOLT	Holt linear exponential smoothing model
DN	EXSM_HOLT_DMP	Holt linear exponential smoothing model with decaying trend
EN	EXSM_MUL_TRND	Exponential smoothing model with multiplicative tendency
DN	EXSM_MULTRD_DMP	Exponential smoothing model with multiplicative decay tendency
NA	EXSM_SEAS_ADD	Exponential smoothing model with additive seasonality but no trend
NM	EXSM_SEAS_MUL	Exponential smoothing model with multiplicative seasonality but no trend
LM	EXSM_HW	Holt–Winters triple exponential smoothing model Additive trend , multiplicative seasonality applied
DM, EA	EXSM_HW_DMP	Holt–Winters multiplicative exponential smoothing model
LA	EXSM_HW_ADDSEA	Holt–Winters Additive Index Smoothing Model, Apply Additive Trend, Additive Seasonality
DA	EXSM_DHW_ADDSEA	Holt–Winters additive exponential smoothing model with
EM	EXSM_HWMT	Holt–Winters Multiplicative Exponential Smoothing Model with
DM	EXSM_HWMT_DMP	Holt–Winters multiplicative exponential smoothing model

2.2 Performance Evaluation by Exponential Smoothing Model - Area

☑ The '14 Model Mixing Model' is a model that uses the median value of the values derived from the 14 models as a prediction value and is more accurate than when each model is used alone, so the '14 Model Mixing Model' is adopted for predicting sales.

* MAPE : ABS(actual value - predicted value) / actual value* 100

Model name	Forecast year / month MAPE									Adoption
	September 2020	October 2020	November 2020	December 2020	January 2021	February 2021	March 2021	April 2021	Average	
EXSM_SIMPLE	14.82	14.29	13.98	12.81	8.87	13.66	13.23	10.23	12.74	
EXSM_SIMPLE_MULT	15.68	15.47	15.14	13.99	9.79	14.55	13.80	11.09	13.69	
EXSM_HOLT	15.89	17.44	18.79	16.41	11.83	18.40	12.22	10.41	15.17	
EXSM_HOLT_DMP	14.27	14.54	13.92	12.21	8.66	14.69	12.42	9.61	12.54	
EXSM_MUL_TRND	14.48	17.32	17.44	14.42	11.55	19.58	11.61	9.90	14.54	
EXSM_MULTRD_DMP	14.55	15.77	14.98	12.23	9.61	15.62	12.72	10.21	13.21	
EXSM_SEAS_ADD	15.58	14.52	13.94	12.87	9.57	10.81	11.99	10.42	12.46	
EXSM_SEAS_MUL	16.89	17.14	16.07	14.42	10.98	13.30	13.20	11.00	14.12	
EXSM_HW	17.71	15.65	16.81	13.61	13.12	15.50	13.55	11.88	14.73	
EXSM_HW_DMP	17.13	15.21	13.18	13.17	11.54	13.58	13.23	11.21	13.53	
EXSM_HW_ADDSEA	16.84	17.29	19.12	14.64	11.79	16.91	11.10	9.73	14.68	
EXSM_DHW_ADDSEA	14.97	14.42	14.28	12.20	9.45	11.87	11.13	9.77	12.26	
EXSM_HWMT	17.77	18.14	34.58	14.86	14.20	17.88	12.31	11.52	17.66	
EXSM_HWMT_DMP	17.98	17.58	18.32	13.63	12.32	14.98	12.68	11.10	14.82	
14 model mixed model	14.6	14.37	13.34	11.49	9.31	13.10	11.24	9.08	12.07	Y

[Back Data] Prediction results of cities, counties and districts

☑ This is the result of predicting monthly sales of cities, counties and districts after adopting the '14 model mixed model' of Exponential Smoothing.

[City, County, District] Monthly forecast accuracy (%)

Division	2020.09	2020.10	2020.11	2020.12	2021.01	2021.02	2021.03	2021.04
Prediction accuracy	86.97%	89.29%	89.61%	89.12%	91.46%	87.65%	88.76%	90.92%

The number and specific gravity of cities, counties and districts by prediction accuracy section

division	20.04 –> 20.09 prediction		20.05 –> 20.10 prediction		20.06 –> 20.11 prediction		20.07 –> 20.12 prediction		20.08 –> 21.01 prediction		20.09 –> 21.02 prediction		20.10 –> 21.03 forecast		20.11 –> 21.04 prediction	
	city county (number of regions)	importance	city county (number of regions)	importance	city county (number of regions)	importance	city county (number of regions)	importance	city county (number of regions)	importance	city county (number of regions)	importance	city county (number of regions)	importance	Number	Importance
95% or more	43	17%	75	30%	76	30%	79	32%	110	44%	76	30%	69	28%	106	42%
Less than 95% More than 90%	71	28%	74	30%	86	34%	70	28%	66	26%	63	25%	75	30%	68	27%
Less than 90% More than 85%	61	24%	43	17%	35	14%	45	18%	40	16%	37	15%	42	17%	36	14%
Less than 85% More than 80%	32	13%	19	8%	22	9%	23	9%	12	5%	26	10%	32	13%	18	7%
less than 80%	43	17%	39	16%	31	12%	33	13%	22	9%	48	19%	32	13%	22	9%
Sum	250	100%	250	100%	250	100%	250	100%	250	100%	250	100%	250	100%	250	100%

Evaluation by Exponential Smoothing Model – City, County , District and Category

- Although the performance of other models (MUL_TRND, MULTRND_DMP, HW, HW_DMP, HWMT, and HWMT_DMP) seems to be superior to the performance of 14 models mixed, they use a stable model for prediction when missing data in Train Data.

*** MAPE : ABS(actual value - predicted value) / actual value* 100**

model name	Forecast year month MAPE									Adoption
	September 2020	October 2020	November 2020	December 2020	January 2021	February 2021	March 2021	April 2021	Average	
EXSM_SIMPLE	22.47	94.36	156.32	49.32	29.23	27.98	18.09	29.77	53.44	
EXSM_SIMPLE_MULT	35.55	97.91	147.27	51.45	35.77	34.88	22.31	34.59	57.47	
EXSM_HOLT	22.38	94.57	156.55	48.23	33.56	32.44	15.59	27.72	53.88	
EXSM_HOLT_DMP	21.67	94.17	156.26	48.21	29.71	29.44	16.29	27.98	52.97	
EXSM_MUL_TRND	29.39	30.07	34.61	30.75	21.43	27.54	15.65	29.25	27.34	
EXSM_MULTRD_DMP	27.34	28.84	34.32	28.87	20.75	26.02	16.60	30.26	26.63	
EXSM_SEAS_ADD	23.77	88.05	160.31	52.83	30.07	20.78	17.75	31.09	53.08	
EXSM_SEAS_MUL	36.25	35.14	43.6	35.50	27.46	29.41	25.03	25.60	32.25	
EXSM_HW	39.75	38.21	45.17	37.22	30.99	36.43	24.13	24.77	34.58	
EXSM_HW_DMP	38.07	34.02	41.33	34.35	29.40	31.05	24.54	24.69	32.18	
EXSM_HW_ADDSEA	23.18	92.57	162.74	54.58	37.09	28.84	15.01	29.50	55.44	
EXSM_DHW_ADDSEA	23.05	86.38	160.43	52.16	31.44	23.68	16.14	29.43	52.84	
EXSM_HWMT	37.27	36.28	45.29	38.39	31.04	36.63	23.03	25.15	34.14	
EXSM_HWMT_DMP	36.86	34.59	41.54	35.63	27.45	31.03	22.91	24.42	31.80	
14 model mixed model	25.88	56.77	88.21	36.25	23.57	22.01	15.66	28.67	37.13	Y

[Back Data] Prediction results

☑ This is the result of predicting sales by type after adopting Exponential Smoothing's '14 Model Mixed Model'.

[Type - definition] Monthly prediction accuracy (%)

Division	2020.09	2020.10	2020.11	2020.12	2021.01	2021.02	2021.03	2021.04
Prediction accuracy	74.12%	43.23%	11.79%	63.75%	76.43%	77.99%	84.34%	71.33%

[Categories - definitions] Number and proportion of cities, counties, and districts by prediction accuracy section

Division	20.04 -> 20.09 prediction		20.05 -> 20.10 prediction		20.06 -> 20.11 prediction		20.07 -> 20.12 prediction		20.08 -> 21.01 prediction		20.09 -> 21.02 prediction		20.10 -> 21.03 forecast		20.11 -> 21.04 prediction	
	city county (number of regions)	importance	city county (number of regions)	importance	city county (number of regions)	importance	city county (number of regions)	importance	city county (number of regions)	importance	city county (number of regions)	importance	city county (number of regions)	importance	city county (number of regions)	importance
95% or more	34	20%	32	19%	37	22%	34	20%	45	27%	55	33%	22	13%	39	23%
Less than 95% More than 90%	25	15%	24	14%	39	23%	37	22%	40	24%	28	17%	40	24%	43	25%
Less than 90% More than 85%	23	14%	33	20%	26	15%	25	15%	27	16%	20	12%	35	21%	29	17%
Less than 85% More than 80%	18	11%	22	13%	14	8%	18	11%	22	13%	17	10%	28	17%	19	11%
less than 80%	69	41%	58	34%	53	31%	55	33%	35	21%	49	29%	44	26%	39	23%
Sum	169	100%	169	100%	169	100%	169	100%	169	100%	169	100%	169	100%	169	100%

2.3.2 Performance evaluation by Exponential Smoothing model (Hospital)

- When 14 models are mixed to derive predictions, the '14 model mixed model' is adopted because each model is more accurate than when used alone.

*** MAPE : ABS(actual value - predicted value) / actual value* 100**

model name	Forecast year month MAPE									Adoption
	September 2020	October 2020	November 2020	December 2020	January 2021	February \2021	March 2021	April 2021	Average	
EXSM_SIMPLE	38.88	62.73	40.73	38.24	37.57	50.13	38.36	33.05	42.46	
EXSM_SIMPLE_MULT	47.37	58.23	47.55	40.85	44.09	59.67	44.43	37.31	47.44	
EXSM_HOLT	56.09	87.43	60.23	56.21	41.38	57.23	62.43	55.41	59.55	
EXSM_HOLT_DMP	41.47	65.96	43.02	40.09	38.79	52.06	41.84	36.72	44.99	
EXSM_MUL_TRND	72.91	86.2	58.48	49.84	51.27	70.07	45.82	45.85	60.06	
EXSM_MULTRD_DMP	45.95	64.49	43.18	39.19	41.83	59.75	41.08	40.93	47.05	
EXSM_SEAS_ADD	43.45	63.37	41.2	40.25	37.34	48.47	41.21	34.48	43.72	
EXSM_SEAS_MUL	48.62	60.21	50.09	46.79	43.31	60.25	48.22	40.32	49.73	
EXSM_HW	49.74	62.42	54.57	46.74	50.68	66.22	55.28	53.35	54.88	
EXSM_HW_DMP	45.04	58.04	48.52	42.00	45.05	60.76	50.07	47.33	49.60	
EXSM_HW_ADDSEA	53.24	95.39	62.53	57.88	39.93	55.18	63.61	54.29	60.26	
EXSM_DHW_ADDSEA	48.38	67.93	43.99	40.90	38.79	51.10	43.69	37.14	46.49	
EXSM_HWMT	55.86	81.72	70.03	53.74	52.64	67.76	55.18	53.70	61.33	
EXSM_HWMT_DMP	51.04	71.22	56.38	44.58	44.02	60.82	47.49	47.38	52.87	
14 model mixed model	38.22	57.29	40.69	36.45	37.60	51.25	40.16	35.07	42.09	Y

[Back Data] Prediction results of city, county, and district (Hospital)

☑ This is the result of predicting sales by type (hospital) after adopting Exponential Smoothing's '14 model mixed model'.

[Type - Hospital] Monthly Prediction Accuracy (%)

Division	2020.09	2020.10	2020.11	2020.12	2021.01	2021.02	2021.03	2021.04
Prediction accuracy	57.34	36.34	56.49	60.8	61.72	48.62	59.84	64.93

[Type - hospital] Number and proportion of cities, counties, and districts by prediction accuracy section

Division	20.04 -> 20.09 prediction		20.05 -> 20.10 prediction		20.06 -> 20.11 prediction		20.07 -> 20.12 prediction		20.08 -> 21.01 prediction		20.09 -> 21.02 prediction		20.10 -> 21.03 forecast		20.11 -> 21.04 prediction	
	city county (number of regions)	importance	city county (number of regions)	importance	city county (number of regions)	importance	city county (number of regions)	importance	city county (number of regions)	importance	city county (number of regions)	importance	city county (number of regions)	importance	city county (number of regions)	importance
95% or more	25	10%	37	15%	48	20%	49	20%	45	19%	42	17%	43	18%	35	14%
Less than 95% More than 90%	19	8%	37	15%	35	14%	33	14%	38	16%	37	15%	38	16%	32	13%
Less than 90% More than 85%	29	12%	33	14%	27	11%	33	14%	31	13%	28	12%	28	12%	46	19%
Less than 85% More than 80%	34	14%	27	11%	28	12%	22	9%	20	8%	17	7%	26	11%	23	10%
less than 80%	135	56%	108	45%	104	43%	105	43%	108	45%	118	49%	107	44%	106	44%
Sum	242	100%	242	100%	242	100%	242	100%	242	100%	242	100%	242	100%	242	100%

2.3.3 Performance Evaluation by Exponential Smoothing Model (Medical)

☑ When 14 models are mixed to derive predictions, the '14 model mixed model' is adopted because each model is more accurate than when used alone.

*** MAPE : ABS(actual value - predicted value) / actual value* 100**

Model name	Forecast year month MAPE									Adoption
	September 2020	October 2020	November 2020	December 2020	January 2021	February 2021	March 2021	April 2021	Average	
EXSM_SIMPLE	12.77	10.49	10.58	12.54	32.76	19.84	11.34	9.56	14.99	
EXSM_SIMPLE_MULT	13.52	12.36	10.99	12.85	39.22	21.45	11.73	10.47	16.57	
EXSM_HOLT	17.3	23.29	12.03	12.30	11.08	36.60	11.73	9.96	16.79	
EXSM_HOLT_DMP	12.5	10.92	10.84	12.41	27.78	19.30	11.37	9.59	14.34	
EXSM_MUL_TRND	12.55	14.3	22.98	14.58	22.83	22.02	10.62	11.55	16.43	
EXSM_MULTRD_DMP	12.7	12.28	14.79	13.01	18.69	17.21	11.29	10.45	13.80	
EXSM_SEAS_ADD	13.43	11.22	10.5	11.54	28.46	13.03	10.30	9.67	13.52	
EXSM_SEAS_MUL	14.54	14.06	11.75	12.47	39.11	27.33	11.46	10.67	17.67	
EXSM_HW	17.97	16.71	13.28	12.61	14.38	16.64	11.68	10.99	14.28	
EXSM_HW_DMP	17.53	15.65	12.51	12.76	12.87	14.81	11.99	10.53	13.58	
EXSM_HW_ADDSEA	18.52	24.53	11.86	11.10	11.76	39.06	10.75	8.85	17.05	
EXSM_DHW_ADDSEA	13.12	11.61	10.83	11.24	28.00	14.73	10.21	9.75	13.69	
EXSM_HWMT	16.9	17.4	36.18	16.21	25.82	18.84	11.87	12.63	19.48	
EXSM_HWMT_DMP	16.92	16.34	16.01	13.96	22.54	15.86	11.77	11.57	15.62	
14 model mixed model	12.75	10.9	10.83	11.27	10.80	14.73	10.07	9.46	11.35	

[Back Data] Prediction results

☑ This is the result of measuring sales by type (medical) after adopting Exponential Smoothing's '14 Model Mixed Model'.

[Type - definition] Monthly prediction accuracy (%)

Division	2020.09	2020.10	2020.11	2020.12	2021.01	2021.02	2021.03	2021.04
Prediction accuracy	87.25	89.1	89.21	88.77	81.65	85.27	89.93	90.58

[Categories - definitions] Number and proportion of cities, counties, and districts by prediction accuracy section

division	20.04 -> 20.09 prediction		20.05 -> 20.10 prediction		20.06 -> 20.11 prediction		20.07 -> 20.12 prediction		20.08 -> 21.01 prediction		20.09 -> 21.02 prediction		20.10 -> 21.03 forecast		20.11 -> 21.04 prediction	
	city county (number of regions)	importance	city county (number of regions)	importance	city county (number of regions)	importance	city county (number of regions)	importance	city county (number of regions)	importance	city county (number of regions)	importance	city county (number of regions)	importance	city county (number of regions)	importance
95% or more	64	26%	117	47%	103	41%	80	32%	108	43%	48	19%	88	35%	121	48%
Less than 95% More than 90%	62	25%	59	24%	74	30%	70	28%	70	28%	65	26%	76	30%	70	28%
Less than 90% More than 85%	52	21%	33	13%	30	12%	56	22%	30	12%	60	24%	37	15%	30	12%
Less than 85% More than 80%	39	16%	14	6%	15	6%	11	4%	18	7%	33	13%	21	8%	11	4%
less than 80%	33	13%	27	11%	28	11%	33	13%	24	10%	44	18%	28	11%	18	7%
Sum	250	100%	250	100%	250	100%	250	100%	250	100%	250	100%	250	100%	250	100%

Performance evaluation by Exponential Smoothing – city , county , type (uijeong), product

- X, Y, and Z, which have a high proportion of sales, were tested at PH5 level for the product, and the SEAS_ADD model with the highest accuracy was adopted as a result of prediction by product in the definition.

* MAPE : ABS(actual value - predicted value) / actual value* 100

Model name	Forecast year month MAPE					Adoption
	January 2021	February 2021	March 2021	April 2021	Average	
EXSM_SIMPLE	46.93	43.70	35.88	31.93	39.61	
EXSM_SIMPLE_MULT	58.46	39.01	47.41	42.85	46.93	
EXSM_HOLT	51.17	46.11	37.44	35.76	42.62	
EXSM_HOLT_DMP	47.36	39.22	34.60	30.86	38.01	
EXSM_MUL_TRND	100.90	78.05	81.14	76.26	84.09	
EXSM_MULTRD_DMP	95.38	78.63	82.22	77.09	83.33	
EXSM_SEAS_ADD	52.26	29.99	32.82	33.12	37.05	Y
EXSM_SEAS_MUL	71.26	51.26	55.97	53.79	58.07	
EXSM_HW	114.73	78.87	80.15	78.98	88.18	
EXSM_HW_DMP	110.77	78.26	80.85	80.25	87.53	
EXSM_HW_ADDSEA	57.46	48.60	33.24	36.66	43.99	
EXSM_DHW_ADDSEA	53.60	40.70	30.31	34.21	39.71	
EXSM_HWMT	114.80	79.47	81.11	80.41	88.95	
EXSM_HWMT_DMP	110.35	79.13	81.69	80.88	88.01	
3 model mixed model	42.13	41.21	46.30	43.89	43.38	

Performance evaluation by Exponential Smoothing– city , county, type (hospital), product

- ☑ X, Y, and Z, which have a high proportion of sales, were tested at PH5 level for the product, and the hospital adopted the SIMPLE model with the highest accuracy as a result of prediction by product.

*** MAPE : ABS(actual value - predicted value) / actual value* 100**

Model name	Forecast year month MAPE					Adoption
	January 2021	February 2021	March 2021	April 2021	Average	
EXSM_SIMPLE	85.89	80.01	54.97	57.28	69.54	Y
EXSM_SIMPLE_MULT	124.71	84.22	51.29	59.72	79.99	
EXSM_HOLT	149.44	132.56	85.39	77.59	111.25	
EXSM_HOLT_DMP	104.32	111.84	62.57	68.37	86.78	
EXSM_MUL_TRND	140.61	118.54	102.37	80.83	110.59	
EXSM_MULTRD_DMP	140.40	109.38	85.57	81.12	104.12	
EXSM_SEAS_ADD	99.91	81.36	62.45	60.22	75.99	
EXSM_SEAS_MUL	121.90	103.55	94.97	72.20	98.16	
EXSM_HW	215.45	118.69	83.91	86.80	126.21	
EXSM_HW_DMP	162.85	120.53	82.91	85.73	113.01	
EXSM_HW_ADDSEA	150.80	117.38	91.16	76.67	109.00	
EXSM_DHW_ADDSEA	118.67	107.33	65.80	70.10	90.48	
EXSM_HWMT	219.05	121.12	85.46	87.19	128.21	
EXSM_HWMT_DMP	163.58	121.23	83.23	86.03	113.52	
3 model mixed model	82.53	76.70	61.47	63.61	71.08	

Performance evaluation by Exponential Smoothing – city , county , type (medical), product

- X, Y, and Z, which have a high proportion of sales, were tested at PH5 Level for the product, and the SIMPLE model, which showed the highest accuracy, was adopted as a result of medical product prediction.

*** MAPE : ABS(actual value - predicted value) / actual value* 100**

Model name	Forecast year month MAPE					Adoption
	January 2021	February 2021	March 2021	April 2021	Average	
EXSM_SIMPLE	18.40	15.90	16.90	14.95	16.54	Y
EXSM_SIMPLE_MULT	18.93	15.80	18.88	15.49	17.28	
EXSM_HOLT	24.93	31.37	18.47	19.64	23.60	
EXSM_HOLT_DMP	20.08	24.59	15.56	15.75	19.00	
EXSM_MUL_TRND	36.05	48.86	28.12	29.72	35.69	
EXSM_MULTRD_DMP	22.71	25.52	18.77	17.81	21.20	
EXSM_SEAS_ADD	19.43	16.47	20.89	15.13	17.98	
EXSM_SEAS_MUL	20.58	18.01	22.99	16.64	19.56	
EXSM_HW	25.11	26.86	20.29	26.90	24.79	
EXSM_HW_DMP	21.29	24.13	20.36	21.62	21.85	
EXSM_HW_ADDSEA	26.15	30.89	18.85	20.14	24.01	
EXSM_DHW_ADDSEA	20.22	22.76	17.00	16.36	19.09	
EXSM_HWMT	47.59	45.19	29.00	40.99	40.69	
EXSM_HWMT_DMP	24.69	27.42	21.57	22.57	24.06	
3 model mixed model	18.07	15.57	17.05	16.04	16.68	

[Reference] How to derive the median value

The median value is centered when the total number of observations is arranged in order of magnitude, and when the number of observations (N) in the model is odd, when the number of observations in the model is even, the mean of the N/2nd and (N+1)/2nd observations is derived as median values.

1. The number of models used : 14 (even number)

Example

ARA N.M.	prediction result by model													
	SIMPLE	SIMPLE MULT	HOLT	HOLT DMP	MUL TRND	MULTRD DMP	SEAS ADD	SEAS MUL	HW	HW DMP	HW ADDSEA	DHW ADDSEA	HWMT	HWMT DMP
Gangneung-si, Gangwon-do	91,266,830	91,350,488	92,667,060	91,332,524	91,984,319	91,279,352	92,375,164	92,332,187	93,918,341	93,462,558	93,648,642	92,415,316	92,831,453	93,334,705
Goseong-gun, Gangwon-do	986,865	986,723	1,053,177	1,043,154	1,031,862	1,034,643	1,000,648	1,006,709	1,070,274	1,044,629	1,058,830	1,050,759	1,051,049	1,054,008
Donghae-si, Gangwon-do	16,557,016	16,695,591	16,468,692	16,613,296	16,799,061	16,637,990	17,389,333	16,710,613	16,915,161	16,698,251	17,268,077	17,365,607	17,180,173	16,752,607
Samcheok-si, Gangwon-do	18,918,257	18,918,763	19,014,030	18,919,889	19,068,626	18,921,120	20,073,993	20,047,872	20,556,862	19,983,210	20,120,182	20,017,731	20,244,423	19,984,314
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

Deriving the median value of 14 models by region

1 . Number of Models Utilized : 3 (number of holes)

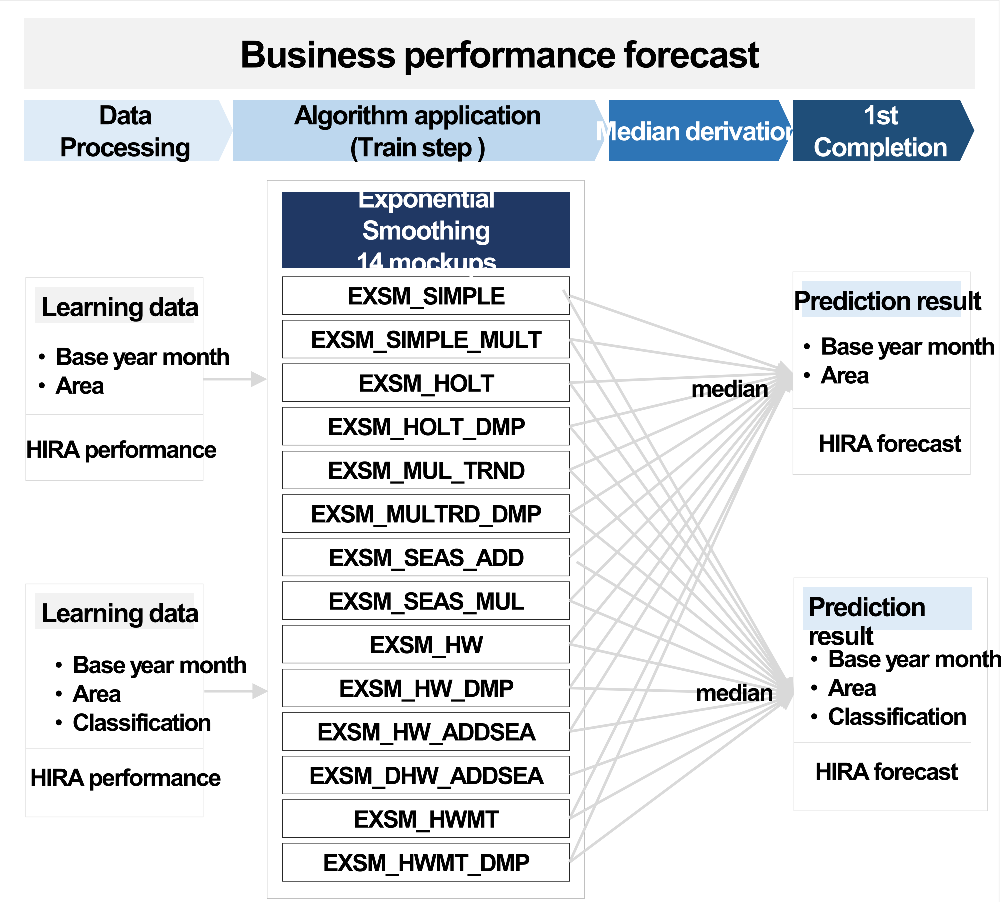
Example

ARA N.M.	prediction result by model		
	SEAS ADD	HW ADDSEA	DHW ADDSEA
Gangneung-si, Gangwon-do	92,375,164	93,648,642	92,415,316
Goseong-gun, Gangwon-do	1,000,648	1,058,830	1,050,759
Donghae-si, Gangwon-do	17,389,333	17,268,077	17,365,607
Samcheok-si, Gangwon-do	20,073,993	20,120,182	20,017,731
…	…	…	…

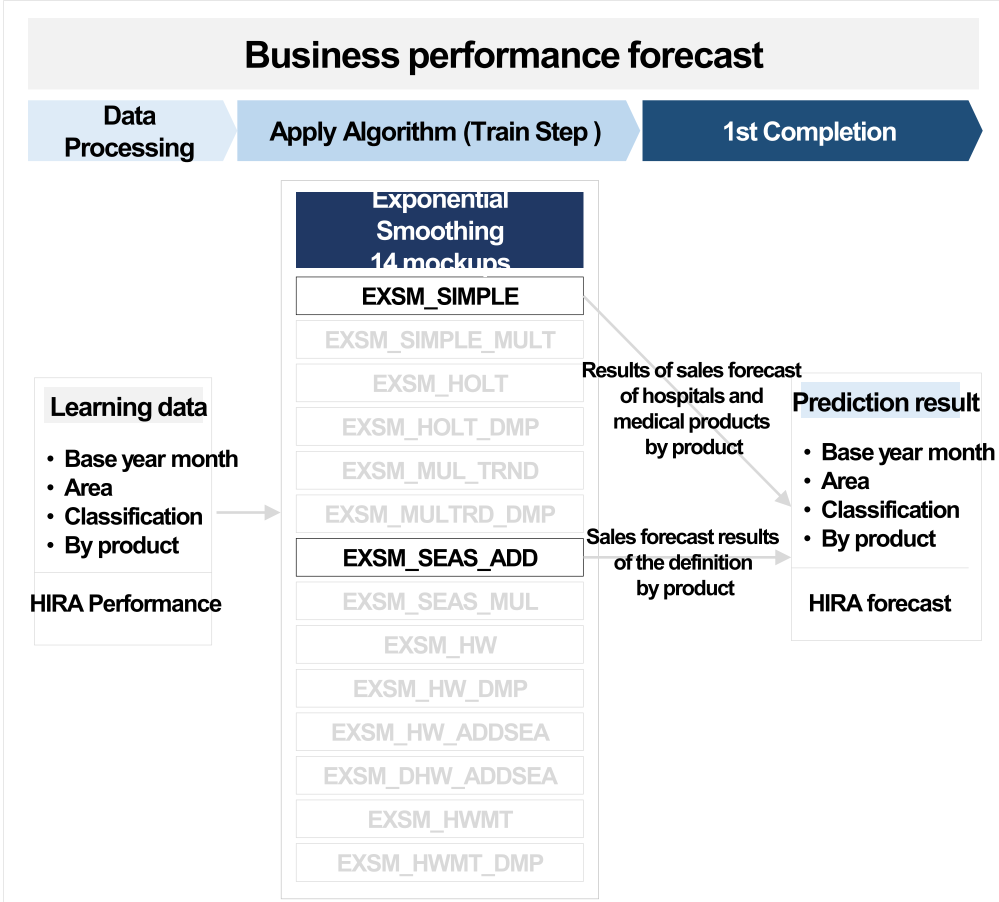
Deriving the median value of 3 models by region

2.2 Exponential Smoothing – How to apply the model

As a result of the model test, the accuracy of the predictions derived by mixing the models is better than that derived by adopting individual models. In the case of predictions by city, county, and district, the median value of the values derived from 14 models, and sales by city, county, district, and product show high accuracy when using the SEAS_ADD and SIMPLE models, so the model is used as a prediction value.



• When 14 models are mixed to derive predicted values, the performance (accuracy) is better than when each model is used alone → Use 14 model mixed models for predictions by city, county, and type



• When the SEAS_ADD model is used for product definition and the simple model is used for hospitals and medical, the performance is excellent

By Regression Algorithm model creation and model performance evaluation method

The Regression algorithm needs sufficient explanatory variables to explain the response variable (HIRA sales amount), but it is currently difficult to obtain meaningful data for prediction both internally and externally, so it generates a model using UBIST and SSC as explanatory variables and evaluates the performance of each algorithm.

Data extraction and segmentation	Create and train models	Model performance evaluation
• Integrate predictions derived from HIRA, UBIST, SSC, and primary algorithms to generate master data and process data according to the prediction purpose (city, city, county, and type) • Splitting Train Data Set, Test Data Set	• Run machine learning using data from 5 months prior to the forecast year as train data • Create Linear Regression, SVR, Neural Network model	• The prediction error is measured by comparing the value of the prediction year and month derived from the algorithm with HIRA data, and the prediction result of each algorithm is evaluated using performance indicators • Utilization Performance Indicators → MAPE

Model creation and performance evaluation

Example

*** As of April 2021 HIRA data forecast**

Master data

- Base year month
- Area
- Classification
- Cedicine
- Classification

HIRA performance
UBIST Amount
SSC amount

Full data
2013.01~2021.04

Dataset

- Base year month
- Area
- UBIST Amount
- SSC amount

HIRA performance

Division

Train Data
2013.01~2020.09

Train data

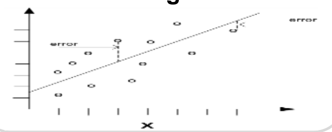
Test Data
2021.04

Test data

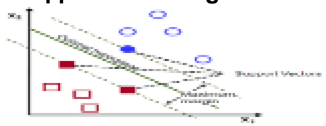
Create

Extraction

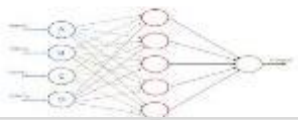
Linear Regression



Support Vector Regression



Neural Network



*** MAPE : ABS(actual value - predicted value) / actual value* 100**

ARA N.M.	by model predicted value			2021.02 actual value	Linear MAPE	SVR MAPE	NN MAPE
	Linear Regression	Support Regression	Neural Network				
Gangneung-si	91,266,830	91,350,488	90,345,313	87472623	4.34	6.4	6.7
Goseong-gun	986,865	986,723	990,345	1192832	17.27	7.90	12.3
Donghae-si	16,557,016	16,695,591	16,960,302	14455591	14.54	4.14	12.5
Samcheok-si	18,918,257	18,918,763	19,984,403	20595334	8.14	15.2	2.6
…	…	…	…	…	10.24	13.4	7.05

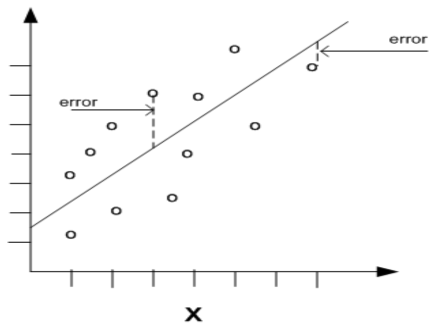
Performance evaluation by algorithm

[Reference] Regression Algorithm Description

☑ The regression algorithms provided by OML include Linear Regression, Support Vector Regression, and Neural Network, and are techniques to model the correlation between dependent and one or more independent variables, which are used to predict numerical data.

Linear Regression

Explanation



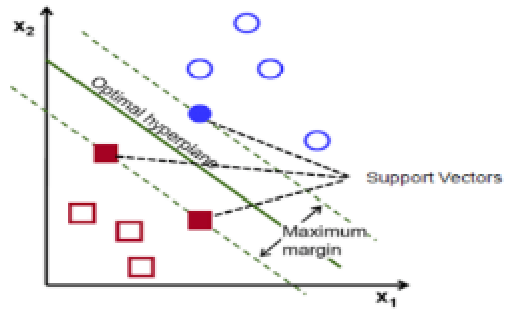
- To assume a linear relationship between an explanatory variable and a response variable and predict the results

Advantages

- The most common method for modeling numerical data
- Provides an estimate of the strength and magnitude of the relationship between features and results

Support Vector Regression (SVR)

Explanation



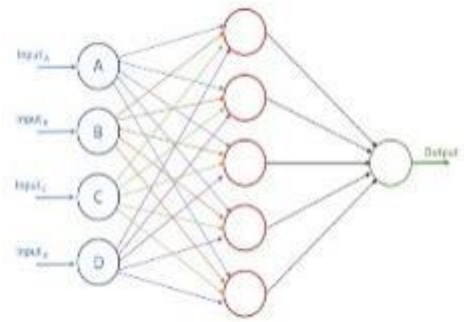
- Model to define baseline for Decision Boundary

Advantages

- Noise is rarely affected, overfitting is not easy, and accuracy is high

Neural Network

Explanation



- A model that imitates the human brain structure, in which multiple hidden layers are interconnected to predict the optimal output value corresponding to the input

Advantages

- Suitable for classification or numerical prediction problems
- High accuracy

3.2 Regression model performance evaluation ①

☑ We separately constructed datasets to train UBIST and SSC by region, and adopted a method to supplement the predictions of primary algorithms in specific regions by generating and applying models.

[City/gun/gu prediction results - MAPE]
Explanatory variables : year/month , region , 1st algorithm predicted value , UBIST amount , SSC amount

Year month	Linear Regression	Support Vector Regression	Neural Network	Exponential Smoothing
Feb 2021	32.86	280.45	35.77	13.10
Jan 2021	52.73	471.36	169.37	9.31

Explanatory variables : year and month , region , 1st algorithm predicted value , UBIST amount

Year month	Linear Regression	Support Vector Regression	Neural Networks	Exponential Smoothing
Feb 2021	42.32	286.15	32.69	13.10
Jan 2021	46.73	416.83	93.09	9.31

Explanatory variables : year and month , region , 1st algorithm predicted value , SSC amount

Year month	Linear Regression	Support Vector Regression	Neural Networks	Exponential Smoothing
Feb 2021	46.82	473.83	29.17	13.10
Jan 2021	27.22	28.245	21.97	9.31

3.2 Regression model performance evaluation ② - City, county and district

☑ We generate a dataset of predictions and SSC data derived from the primary algorithm as explanatory variables, and compared with the Regression model, the predictions of the Neural Network model improve over the accuracy of the primary predictions.

[Prediction results - MAPE]

- Explanatory variables: year month, region, prediction value of primary algorithm, UBIST amount
- Areas subject to algorithm application: areas where the UBIST amount is lower than the predicted value derived through the primary algorithm for 3 months as of the MAX year of HIRA data

Year month	Linear Regression	Support Vector Regression	Neural Networks	Exponential Smoothing	number of regions
Apr 2021	12.60	49.40	8.80	8.18	8
Mar 2021	7.47	9.42	8.51	9.47	7
Feb 2021	25.95	125.87	12.79	8.33	10
Jan 2021	6.28	17.48	8.10	5.42	13
Average	13.08	50.54	9.55	7.85	

- Explanatory variables: year month, region, prediction value of primary algorithm, SSC amount
- Areas subject to algorithm application: areas where the SSC amount is lower than the predicted value derived through the primary algorithm for 3 months as of the MAX year of HIRA data

Year month	Linear Regression	Support Vector Regression	Neural Network	Exponential Smoothing	number of regions
Apr 2021	7.95	43.29	8.73	10.92	19
Mar 2021	7.99	33.64	7.78	9.64	28
Feb 2021	10.54	34.00	11.51	13.46	34
Jan 2021	12.39	57.82	10.28	10.95	28
Average	9.72	42.18	9.67	11.24	

3.2 Regression model performance evaluation ② - City , county and district (regulation)

☑ The predictions derived from the primary algorithm and SSC data are generated as explanatory variables, and the predictions of the Neural Network model are improved over the accuracy of the primary predictions when the model is compared.

[Prediction results - MAPE]

- Explanatory variables: year month, region, prediction value of primary algorithm, UBIST amount
- Areas subject to algorithm application: areas where the UBIST amount is smaller than the predicted value derived through the primary algorithm for 3 months based on the MAX year of HIRA data

Year month	Linear Regression	Neural Network	Exponential Smoothing	number of regions
Apr 2021	7.76	5.54	4.49	5
Mar 2021	11.11	11.64	13.39	7
Feb 2021	9.69	11.28	11.56	10
Jan 2021	6.73	7.58	4.71	12
Average	8.82	9.01	8.53	

- Explanatory variables: year month, region, prediction value of primary algorithm, SSC amount

- Areas subject to algorithm application: areas where the SSC amount is smaller than the predicted value derived through the primary algorithm for 3 months as of the MAX year of HIRA data

Year month	Linear Regression	Neural Networks	Exponential Smoothing	number of regions
Apr 2021	10.86	11.28	14.85	8
Mar 2021	10.58	10.33	13.17	6
Feb 2021	10.39	9.12	14.62	8
Jan 2021	15.66	15.66	17.25	9
Average	11.87	11.59	14.97	

3.2 Regression model performance evaluation ② - City , county and district (hospital)

☑ The predictions derived from the primary algorithm and the UBIST data are generated as explanatory variables, and the predictions of the Linear Regression model are improved over the accuracy of the primary predictions when compared with the model.

[Hospital prediction - MAPE]

- Explanatory variables: year month, region, prediction value of primary algorithm, UBIST amount

→ Areas subject to algorithm application: areas where the UBIST amount is smaller than the predicted value derived through the primary algorithm for 3 months based on the MAX year of HIRA data

Year month	Linear Regression	Neural Network	Exponential Smoothing	number of regions
Apr 2021	27.65	27.06	26.02	4
Mar 2021	33.16	32.00	46.16	3
Feb 2021	10.32	15.59	35.83	3
Jan 2021	20.43	23.83	25.84	7
Average	22.89	24.62	33.46	

- Explanatory variables: year month, region, prediction value of primary algorithm, SSC amount

- Areas subject to algorithm application: areas where the SSC amount is smaller than the predicted value derived through the primary algorithm for 3 months as of the MAX year of HIRA data

Year month	Linear Regression	Neural Networks	Exponential Smoothing	number of regions
Apr 2021	33.85	24.19	27.81	16
Mar 2021	42.32	32.36	39.12	17
Feb 2021	51.25	44.45	40.30	13
Jan 2021	28.61	28.61	14.93	17
Average	39.01	32.4	30.54	

3.2 Regression model performance evaluation ② - City , county and district (medical)

☑ The predictions derived from the primary algorithm and SSC data are generated as explanatory variables, and the predictions of the Neural Network model are improved over the accuracy of the primary predictions when the model is compared.

[Predicted result - MAPE]

- Explanatory variables: year month, region, prediction value of primary algorithm, UBIST amount

- Areas subject to algorithm application: areas where the UBIST amount is smaller than the predicted value derived through the primary algorithm for 3 months based on the MAX year of HIRA data

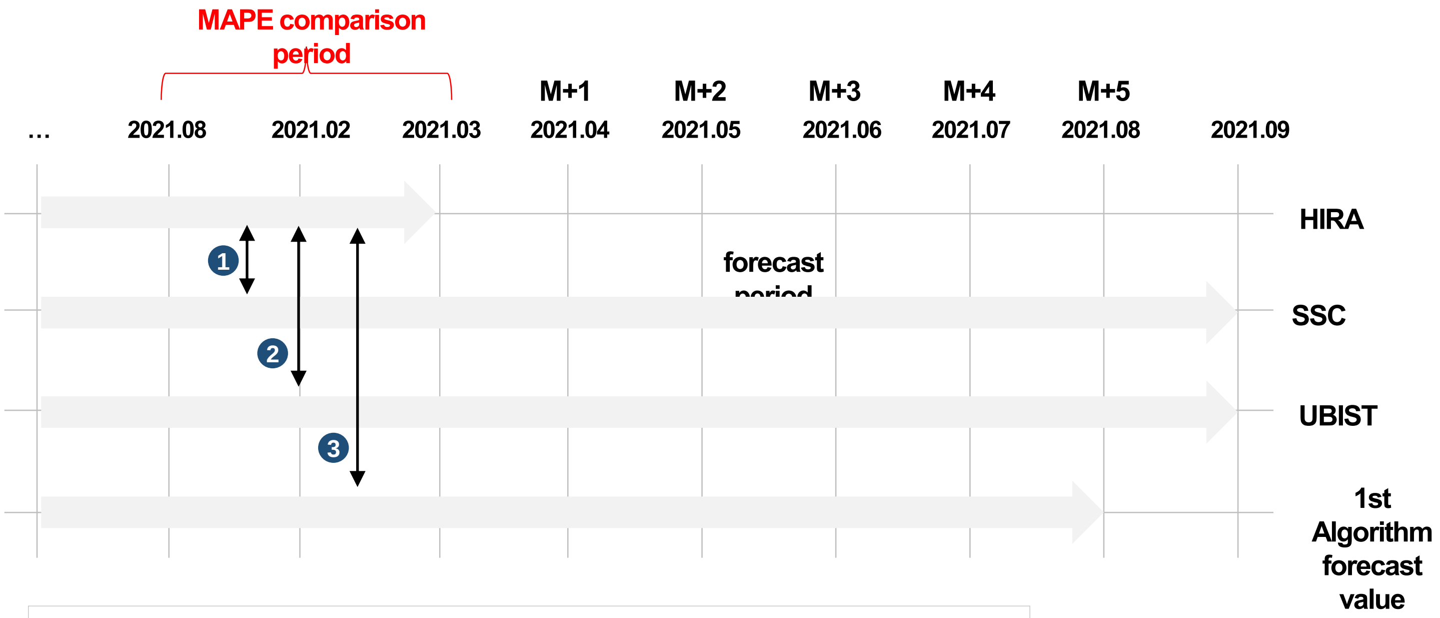
Year month	Linear Regression	Neural Networks	Exponential Smoothing	number of regions
Apr 2021	5.17	6.66	5.18	41
Mar 2021	13.07	12.62	6.25	35
Feb 2021	13.66	10.90	9.32	37
Jan 2021	10.10	9.19	9.39	37
Average	10.50	9.84	7.53	

Explanatory variables: year month, region, prediction value of primary algorithm, SSC amount
Areas subject to algorithm application: areas where the SSC amount is smaller than the predicted value derived through the primary algorithm for 3 months as of the MAX year of HIRA data

Year month	Linear Regression	Neural Network	Exponential Smoothing	number of regions
Apr 2021	6.83	5.63	6.35	103
Mar 2021	8.95	6.10	7.08	98
Feb 2021	6.79	8.81	10.35	105
Jan 2021	6.24	6.24	7.00	90
Average	7.20	6.69	7.69	

[Reference] 2nd Algorithm Application Area

☑ By comparing the accuracy of the three-month HIRA amount, SSC, UIBST, and 1st algorithm prediction based on the MAX year of HIRA data, the area where the SSC and UBIST amount approximates the HIRA amount is reflected as the 2nd algorithm application area.



1 MAPE = (HIRA Amount -SSC Amount)/(HIRA Amount)*100

2 MAPE = (HIRA Amount -UBIST Amount)/(HIRA Amount)*100

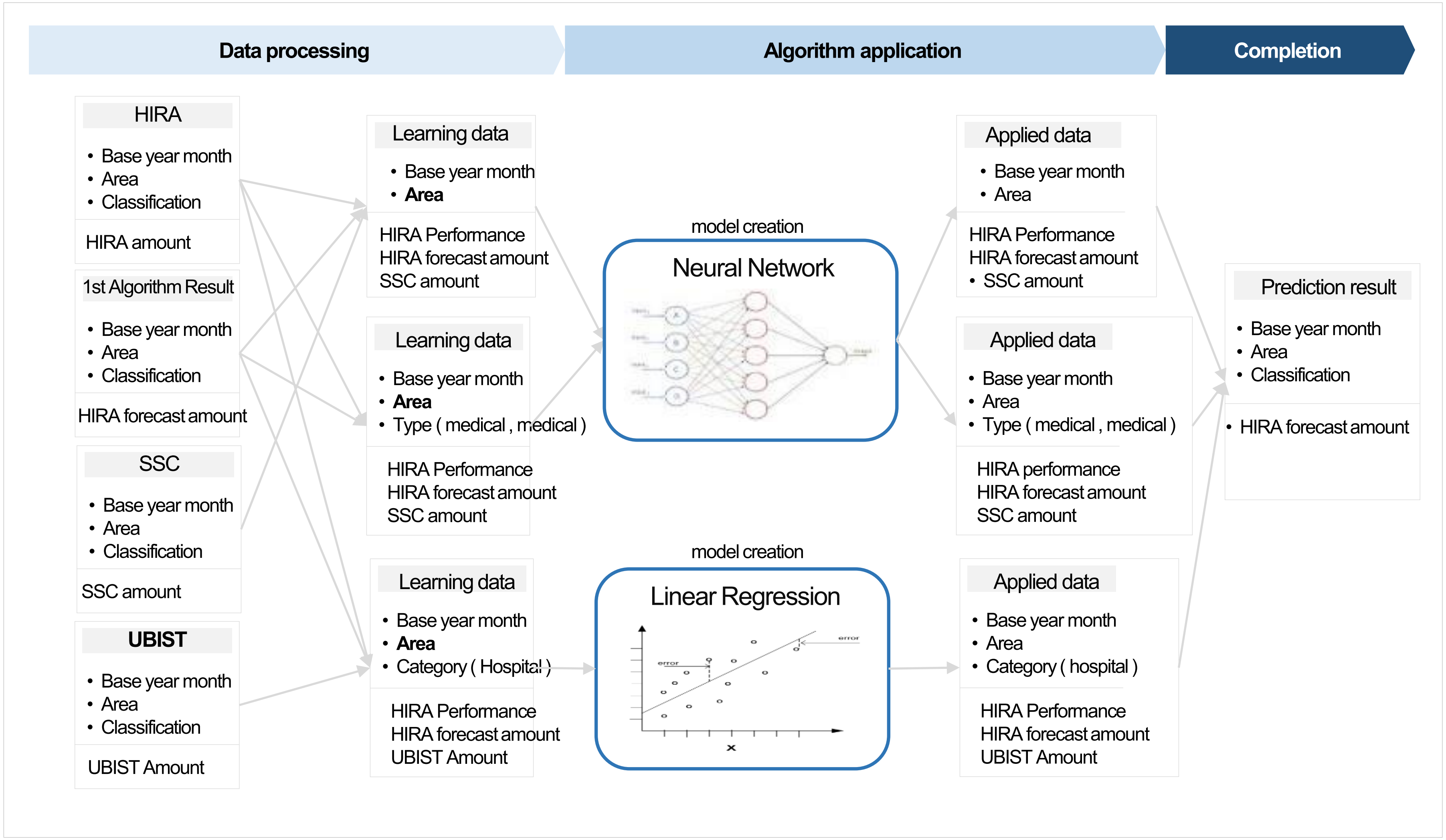
3 MAPE = (HIRA Amount - Primary Algorithm Prediction Amount)/(HIRA Amount)*100

Secondary algorithm application area

- Based on the MAX year of HIRA data, the accuracy of the HIRA amount for 3 months, SSC, UIBST, and 1st algorithm prediction amount is calculated, and if ①②'s MAPE is lower than ③, that is, the area where SSC and UBIST amount approximate to HIRA amount is reflected as the 2nd algorithm application area
- As a result, the application area of the secondary algorithm is changed every month

3.4 Regression – How to apply the model

- ☑ The HIRA amount is a response variable, the predictions derived through the primary algorithm, SSC, and UBIST are created as explanatory variables, and the model is learned to complement the predictions of the primary algorithm.



4.1 Machine Learning Process for Predicting Sales Sales

☑ Data processing and HIRA data-based time-series algorithms are used to derive the predicted values, and the final sales forecast reflecting the derived predicted values, UBIST, and SSC data in the regression algorithm.

Data Processing

- Based on the performance data (HIRA) of the Korea Appraisal Board 5 to 6 months before the forecast year, Train Data Set is created by city, county, type, and product

Apply first- order algorithm

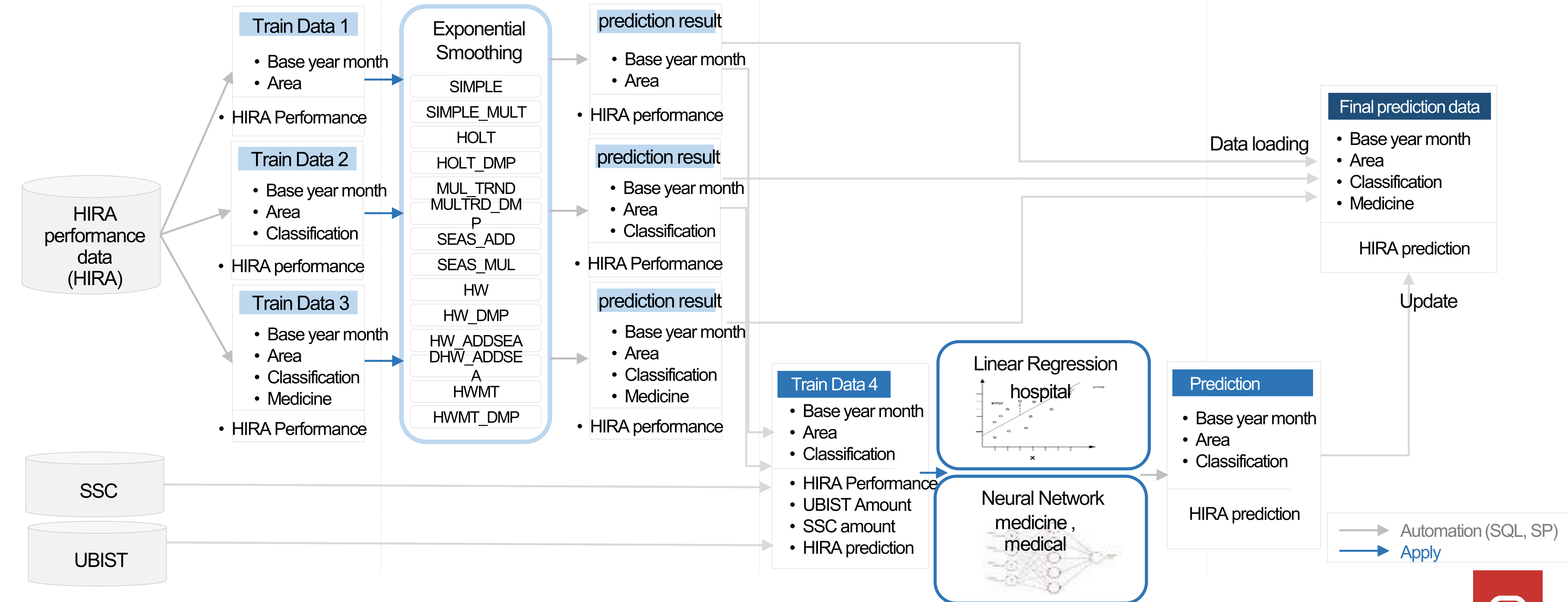
- Predictive values by city, county, type, and product are derived using the Exponential Smoothing
- Use HIRA data as learning data and improve prediction accuracy with 14 models provided by Exponential Smoohting

Apply quadratic algorithm

- Using Linear Regression, the predictions derived after applying the primary algorithm are supplemented
- The target of applying the secondary algorithm is city, county, and district, and type-definition, and the learning data uses SSC, predictions derived from the primary algorithm, and HIRA data

Derivation of final prediction data

- Load the final prediction by updating the predictions derived from the primary algorithm to the predictions derived from the secondary algorithm



End of Document