# Modified Prophet+Optuna Prediction Method for Sales Estimations

Kohei Arai[1], Ikuya Fujikawa[2], Yusuke Nakagawa[3], Tatsuya Momozaki[4], Sayuri Ogawa[5]
Information Science Department, Saga University, Saga City, Japan[1]
Success Institute Chain: SIC Co., Ltd, Hakata-ku, Fukuoka City, Fukuoka, Japan[2, 3, 4, 5]

***Abstract*—A prediction method for estimation of sales based on Prophet with a consideration of nonlinear events and conditions by a modified Optuna is proposed. Linear prediction does not work for a long-term sales prediction because purchasing actions are based on essentially nonlinear customers' behavior. One of nonlinear prediction methods is the well-known Prophet. It, however, is still difficult to adjust the nonlinear parameters in the Prophet. To adjust the parameters, the Optuna is widely used. It, however, is not good enough for parameter tuning by the Optuna. Therefore, the Optuna is modified with a short-term moving mean and standard deviation of the sales for final prediction. More than that, specific event such as typhoon event is to be considered in the sales prediction. Through experiments with a real sales data, it is found the sensitivity of the parameters the upper window, lower window, event dates, etc. for the final sales and the effect of the Optuna is 11.73%. Also, it is found that the effect of the consideration of Covid-19 is about 2.4% meanwhile the effect of the proposed modified Optuna is around 3 % improvement of the prediction accuracy (from 80 % to 83 %).**

***Keywords—Prediction method; nonlinearity; prophet; optuna; typhoon event; modified optuna; mean and standard deviation adjustment***

## I. Introduction

Periodicity, event effects, long-term trends, and outliers are not limited to this data, but are common features of general time series data. When creating a model for future prediction, it is necessary to incorporate these features into the model well. Prophet models each of the four features and combines them to predict future values. Such a model is called a Generalized Additive Model.

Prophet is a library for time series analysis developed by Facebook's Core Data Science team in 2017[1]. Libraries are provided in both Python and R. In addition, this Prophet is embedded as a template in AutoML services such as AWS, Azure, and DataRobot for flexible modeling in future forecasting tasks[2].

There are five advantages of Prophet:

*1) It can be made a model without knowledge of statistics:* Simply specify the data and perform the training to complete the model.

*2) Easy to incorporate domain knowledge:* It can be easily put in the domain knowledge that the data analyst has.

*3) No feature engineering required:* Prophet training uses minimally preprocessed data. There is no need to remove trend components or convert to a moving average series.

*4) There is no problem even if there are missing values:* Even if there is a defect in the training data, no error will occur, and training will be performed normally. Therefore, it is not necessary to fill in the missing values in advance.

*5) Easy to interpret prediction results:* Prophet is a model that adds four terms. Each term represents a trend, periodicity, event effect, and error, and after prediction, the components can be extracted for each term and the obtained prediction results can be considered.

On the other hand, Optuna is a Bayesian optimization package created for optimizing hyperparameters of machine learning models[3]. It performs optimization using TPE, which is a new method among Bayesian optimizations. It can be easily used in a single process, or it can be learned in parallel on many machines. When performing parallel processing, this is achieved by creating an Optuna file on the database and referencing it from multiple machines, so it is wonderful that all machines that can access the DB can participate in learning[4].

The proposed nonlinear prediction method is based on Prophet with Optuna for parameter tuning. It is not easy to optimize the parameters in Prophet and is not ensure the best fit parameters for Prophet. In this paper, therefore, some programmatic method for the parameter optimization is proposed and effectiveness of the proposed method is validated with a nonlinear sales data.

The biggest challenge of this research work is to predict one year term of sales (annual amount of sales). Although there are many prediction methods which allow to predict one day after the current time, there is no such method which allows forecast ahead for the following 365 days with an acceptable prediction accuracy. Therefore, nonlinearity, seasonal effect, event effect, the other influencing factors have to be considered.

One of nonlinear prediction methods is the well-known Prophet. It, however, is still difficult to adjust the nonlinear parameters in the Prophet. To adjust the parameters, the Optuna is widely used. It, however, is not good enough for

[1] https://facebook.github.io/prophet/
[2] https://peerj.com/preprints/3190/
[3] https://optuna.readthedocs.io/en/stable/tutorial/first.html
[4] https://optuna.readthedocs.io/en/stable/reference/index.html

parameter tuning by the Optuna. Therefore, the Optuna is modified with a short-term moving mean and standard deviation of the sales for final prediction.

In the next section, related research works are reviewed followed by the proposed method. Then, the fact that a linear prediction method does not work for nonlinear time series of data is shown. After that, the validation of effectiveness of the proposed method is described together with effectiveness of the Optuna. Finally, conclusion and some discussions are described followed by future research work.

## II. Related Research Work

There are the following related research works on prediction,

Probabilistic cellular automata-based approach for prediction of hot mudflow disaster area and volume is proposed [1]. New approach of prediction of Sidoarjo hot mudflow disaster area based on probabilistic cellular automata is also proposed [2]. On the other hand, GIS based 2D cellular automata approach for prediction of forest fire spreading is proposed [3].

Cell based GIS as cellular automata for disaster spreading prediction and required data systems is investigated [4] together with hot mudflow prediction area model and simulation based cellular automata for LUSI and plume at Sidoarjo East Jawa [5].

Comparative study between eigen space and real space-based image prediction methods by means of autoregressive model is conducted [6] together with comparative study on image prediction methods between the proposed morphing utilized method and Kalman Filtering method [7].

Prediction method for time series of imagery data in eigen space is proposed [8]. Meanwhile, image prediction method with non-linear control lines derived from Kriging method with extracted feature points based on morphing is proposed [9]. On the other hand, cell-based GIS as cellular automata for disaster spreading predictions and required data systems is proposed [10].

Prediction method of El Nino Southern Oscillation event by means of wavelet-based data compression with appropriate support length of base function is proposed [11]. On the other hand, Question Answering for collaborative learning with answer quality prediction is created [12].

Wildlife damage estimated and prediction method using blog and tweet information is proposed [13]. Prediction method for large diatom appearance with meteorological data and MODIS derived turbidity as well as chlorophyll-a in Ariake Bay area in Japan is proposed and validated [14].

Method for thermal pain level prediction with eye motion using SVM is proposed [15] together with prediction method for large diatom appearance with meteorological data and MODIS derived turbidity and chlorophyll-a in Ariake bay area in Japan [16].

Smartphone image based agricultural product quality and harvest amount prediction method is proposed [17]. Meanwhile, data retrieval method based on physical meaning and its application for prediction of linear precipitation zone with remote sensing satellite data and open data is also proposed [18].

Recursive Least Square: RLS method-based time series data prediction for many missing data is proposed [19]. Furthermore, prediction of isoflavone content in beans with Sentinel-2 optical sensor data by means of regressive analysis is proposed and conducted [20].

## III. Proposed Prediction Method based on Prophet

A library often used for time series analysis using AI, especially for future prediction with the following features,

1) There is periodicity, weekly and yearly periodicity
2) There is an event effect
3) There is a long-term trend
4) There are outliers (noise)

Prophet is developed by Facebook in 2017.

The model formula of Prophet is as follows,

$$y=g(t)+s(t)+h(t)+\varepsilon_t \quad (1)$$

where *y(t)*: variable for prediction, *g(t)*: trend, *s(t)*: periodic, *h(t)*: event effect, $\varepsilon_t$: normal distribution of noise.

Basically, it can be used as the same as scikit-learn. Model instance creation can be done with fit flow then creation is also done with a data frame for prediction. After that, a prediction is made with the predict method.

Nonlinear for trend term: Linear by default Specify the upper limit of prediction (cap) to make it nonlinear. Fig. 1 shows the illustrative view of determination of upper limit of prediction.

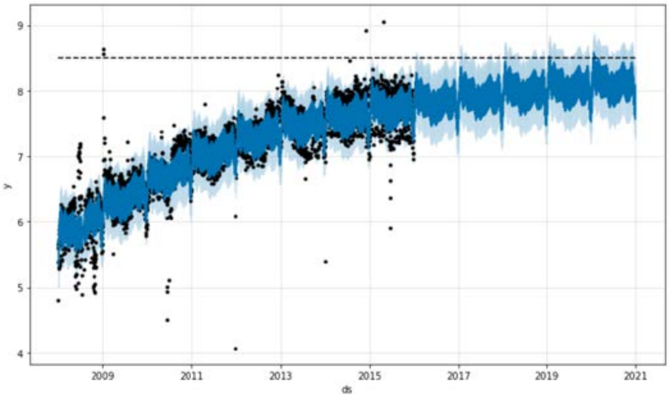
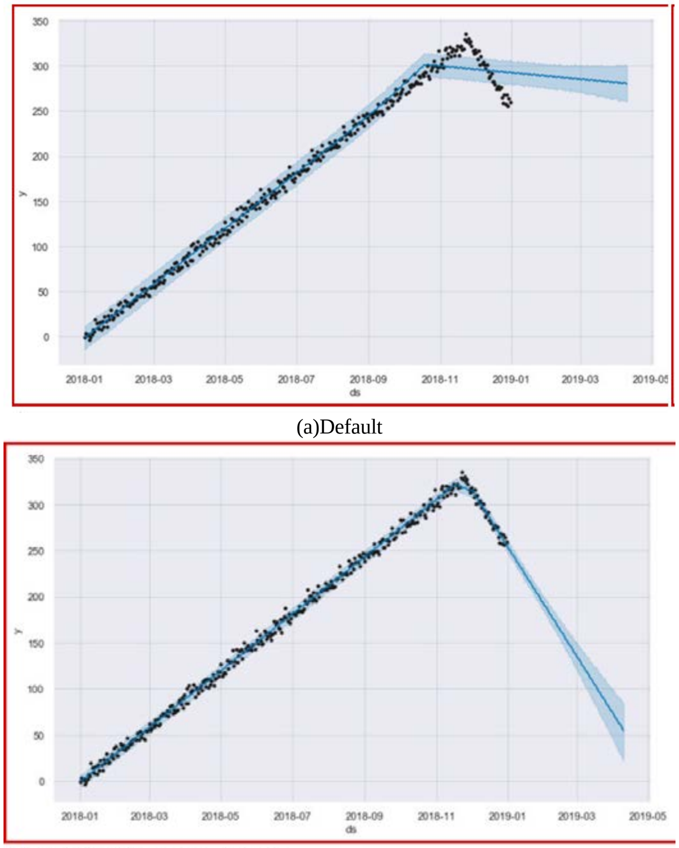


Fig. 1. Illustrative View of Determination of Upper Limit of Prediction.

*1) Changepoint-range:* The default setting does not reflect the most recent change point as shown in Fig. 2 (a). The data used by Prophet to estimate the trend change point is 80% of the total by default. Resolved by setting "changepoint-range = 1". It tends to be predicted that the latest data will be overwhelmed as shown in Fig. 2 (b).



(a)Default

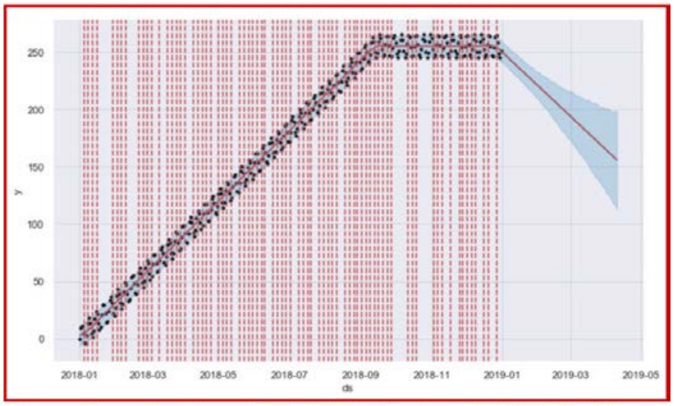
「changepoint_range=1」としたときの予測

(b) Changepoint-range = 1

Fig. 2. Changepoint-Range Setting.

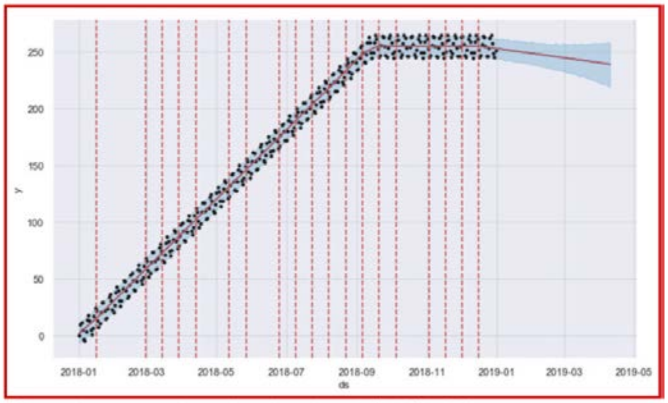
*2) Changepoint-prior-scale:* It represents the variance of the Laplace distribution, which is the prior distribution of the trend term, and the larger it is, the easier it is to detect the change point as shown in Fig. 3 (a).,

*3) n-changepoints:* It represents the number of change point candidates to be detected, and the larger the number, the easier it is to detect more change points as shown in Fig. 3 (b).



「changepoint_range=1, n_changepoints=100, changepoint_prior_scale=10」としたときの予測

(a) Changepoint-Prior Scale.



「changepoint_range=1, changepoint_prior_scale=10」としたときの予測

(b) n-changepoints

Fig. 3. Another Parameter Setting.

Trend term g(t) is represented as equation (2) and can be determined as follows,

$$g(t) = \frac{C}{1+\exp(-k(t-m))} \tag{2}$$

where *C*: Upper limit, *k:* Growing ratio, *m*: Offset

This is the base logistic curve which is shown in Fig. 4 Phenomenon with a flow of less at the beginning then more in the middle, less again after that.

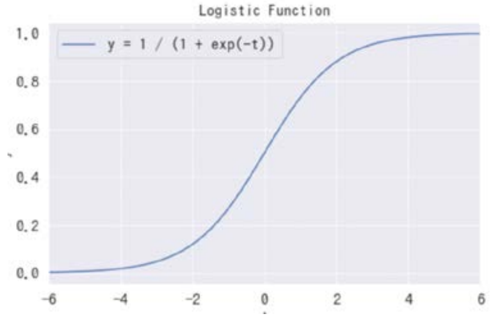


Fig. 4. Logistic Function.

Upper limit, growing ratio and offset are determined as follows,

Since *m* is an expression that directly subtracts the value of *t* as shown in Fig. 5 (c). The curve simply moves from side to side.

There are seasonal fluctuations. There is periodicity. It can be expressed like signal processing.

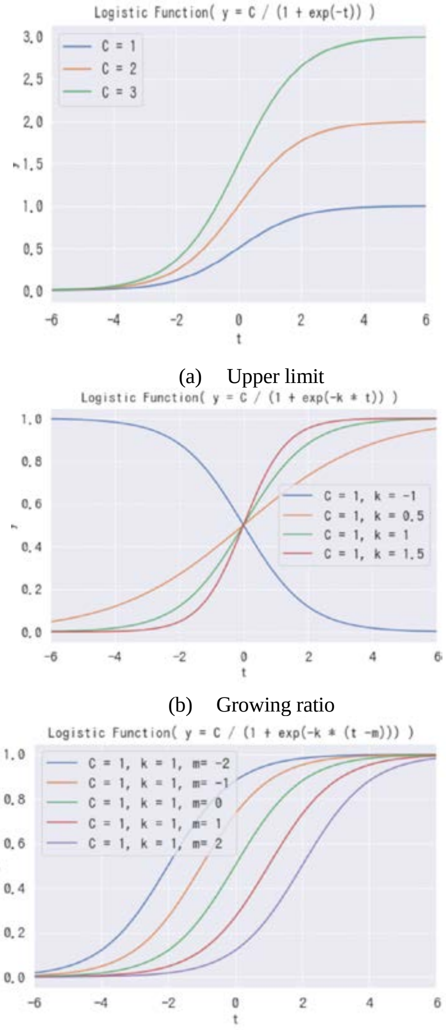
$$s(t) = \sum_{n=1}^{N} (a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \sin(\frac{2\pi nt}{P})) \tag{3}$$

Fit with N = 10 for a yearly cycle and N = 3 for a weekly cycle.

$$\beta = (a_1, b_1, \ldots, a_N, b_N)^T \tag{4}$$

$$X(t) = [\cos\left(\frac{2\pi(1)t}{365.25}\right), \sin\left(\frac{2\pi(1)t}{365.25}\right), \ldots, cos\left(\frac{2\pi(10)t}{365.25}\right), sin\left(\frac{2\pi(10)t}{365.25}\right)] \tag{5}$$

$$s(t) = X(t)\beta \tag{6}$$



(a) Upper limit

(b) Growing ratio

(c) Offset

Fig. 5. Determination of Upper Limit, Growing Ratio and Offset Determinations.

Event effect is defined by incorporate sudden event effects into the model *h(t).* Prophet is designed so that the analyst can create a list of event calendars and incorporate them into the model. The coefficient parameter for each event *i* is $\kappa_i$, and the vector is represented b*y κ.*

$$D_i = (\dots, 1975/12/25, 1976/12/25, \dots, 2020/12/25, \dots) \quad (7)$$

$$Z(t) = [1(t \in D_1), \dots, 1(t \in D_L)] \quad (8)$$

$$h(t) = Z(t)\kappa, \kappa \sim Normal(0, \nu^2) \quad (9)$$

Probability Density Function: PDF is defined as follows,

If you assume the distribution for each parameter, you can treat it as a state space model. In fact, in Prophet, this model formula is described in Stan and optimized by the L-BFGS method etc.

$$\beta \sim Normal(0, \sigma^2) \quad (10)$$

## IV. Experiments

### *A. Example of Sales Prediction and Sensitivity of the Parameters*

Through an adjustment of the Prophet parameters to forecast Mega-Donki Hair Salon (One of the Hair Salons in concern) sales, and then prediction of the sales. Fig. 6 shows the prediction result. In this case, the sales data of 2015 to 2020 is used for training data and also the sales data of 2021 is used for validation data.

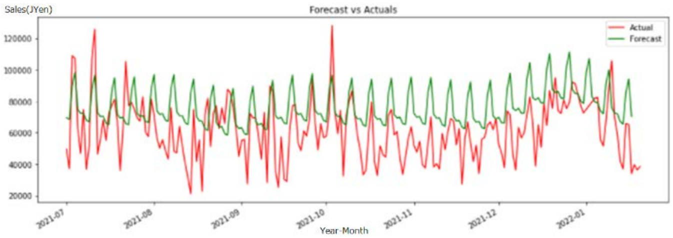


Fig. 6. Prediction of Mega-Donki Hair Salon Sales in 2021.

There is a systematic error. Also, prediction error is not so small. Therefore, some parameter adjustments are required to improve the prediction accuracy.

Optuna is a software framework for automating hyperparameter optimization. The author adjusted the parameters of Prophet using Optuna and tried to forecast the sales of the Mega-Donki Hair Salon. Parameters that can be tuned such as seasonal prior distribution, degree of influence, range of use of data used for detection of change points, influence of trends, etc. After the adjustment by Optuna, prediction result is improved as shown in Fig. 7. The total prediction error is reduced from 38.09 to 26.36. Namely, 11.73 % of improvement is confirmed.

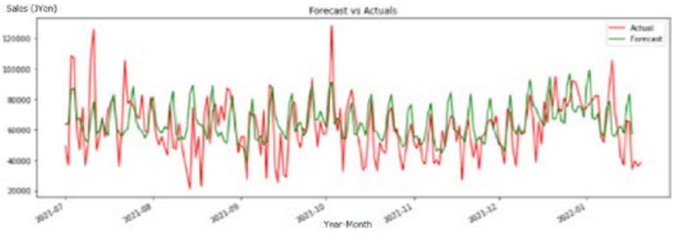


Fig. 7. Prediction of Mega-Donki Hair Salon Sales after the Parameter Adjustment with Optuna.

Tuned parameters are as follows,

*1) Changepoint-range:* Percentage of what range of data is used to detect the trend change point. The default is 0.8, which uses the first 80% of the data to detect the trend change point. According to the formula [0.8, 0.95] Range is reasonable

*2) n-changepoints:* A parameter that represents the number of candidates change points to detect. The larger the parameter, the easier it is to detect more change points. The default is 25.

*3) Changepoint-prior-scale:* Parameters that control the flexibility of the trend. If it is too small, the trend will be inadequate, and if it is too large, the trend will be overfitted. The default is 0.05, and the formula says that the range [0.001, 0.5] is reasonable.

*4) Seasonality-prior-scale:* Parameters that control seasonal flexibility. The default is 10, and the formula says that the range [0.01, 10] is reasonable.

Add-seasonality (period, Fourier-order): Not only the specified year / week / day periodicity, but also the model of any cycle can be set by the user. For each periodicity, the unit of periodicity (period), the Fourier series that is the basis of the seasonal component (Fourier-order), and the degree of influence of seasonality (prior-scale) are set. In order to be able to adjust the Fourier series and the degree of influence of each seasonality, all the prescribed seasonality is set to False, and weekly, monthly, yearly, and quarterly periodic fluctuations are added.

### *B. Sales Prediction Results with Parameter Setting*

Data preprocessing (missing value completion) is needed. Then, consideration of event effect (entrance ceremony, graduation ceremony, pension payment date) is necessary. There is a date with zero sales. Until now, it was excluded and calculated. Therefore, complement the interpolation by the average value is needed. The average value is the day of zero. Complemented with the average value of the day of the week. With lower-window and upper-window, the range can be extended the range to which the event effect is applied to the days around the event day. Also, if Christmas is set as an event and lower window is set to -1, the event effect can be applied until Christmas Eve.

The date of the pension payment can be considered. After the 15th of even-numbered months. Also, one week defined as (7 days). Improvement of the prediction accuracy (MAPE: Mean Absolute Prediction Error) for the specific two shops: Konoha Mall Hashimoto Hair Salon (This Hair Salon is another Hair Salon in concern and has low prediction accuracy).

*1)* Before considering the event: MAPE= 37.8

*2)* After considering the entrance ceremony, graduation ceremony, and pension: MAPE=36.4

For the Hakata Station South Hair Salon (Another Hair Salon in concern) case,

*1)* Before considering the event: MAPE=25.5

*2)* After considering the entrance ceremony, graduation ceremony, and pension: MAPE=24.7

*3)* After considering the entrance ceremony and pension, MAPE=24.8

Sales are declining near the graduation ceremony due to learning from training data. The sales are increasing near the entrance ceremony. Also, sales increase (upside) can be dealt with relatively, but sales decrease (downside) cannot be dealt with (red ... actual green ... forecast) Hakata-eki-minami Hair Salon (another Hair Salon in concern) data as shown in Fig. 8. There are some strange prediction errors marked with blue ellipsoids. For instance, sales are gotten down on August 25 due to the typhoon #15 is hit over these areas as shown in Fig. 9. These events can be considered in the prediction by Prophet.

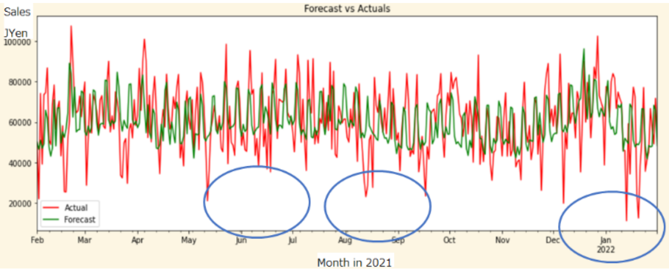


Fig. 8. Sales Prediction Result for Hakata-eki-minami Hair Salon.

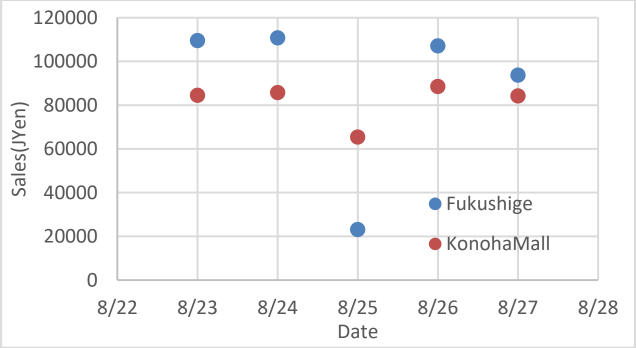
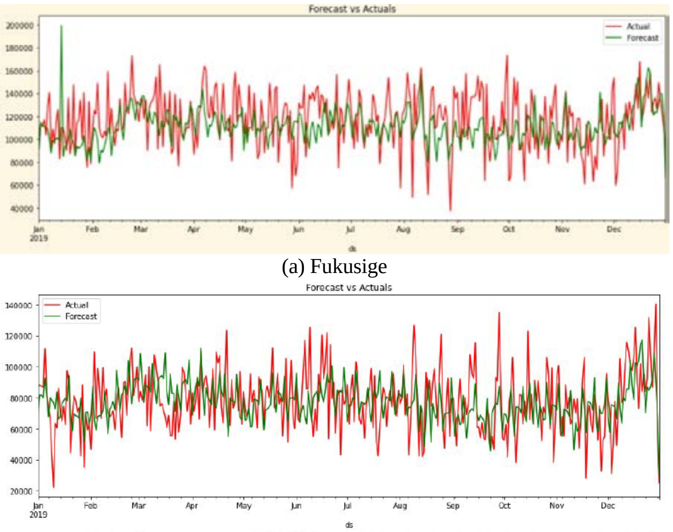


Fig. 9. Sales of the Fukushige and the Konoha-Mall Hair Salons during from Aug.22 to Aug. 28.

### *C. Influence Due to the Covid-19*

The sales of the hair salon have been changed due to the Covid-19. To investigate the influence of Covid-19, the sales of the Fukushige hair salon have been predicted for one year of 2019 utilizing the nine years sales data, 2010 to 2018. The actual and the predicted sales with Prophet and the proposed modified Optuna are shown in Fig. 10(a). As the result, it is found that the MAPE is improved from 18.09 to 16.61. Also, as shown in Fig. 10(b), it is found that the actual and the predicted sales of Shingu hair salon is improved from 23.6 to 21.1.



(a) Fukusige

(b) Shingu

Fig. 10. Influence of Covid-19 on the Sales Prediction.

## V. Conclusion

Through the experiments, it is found the followings,

*1)* Although Prophet is basically linear prediction method, it does work because it can consider trend, seasonal changes, event effect.

*2)* Upper window, lower window, event dates, etc. need to be entered from specialized knowledge and experience.

*3)* The proposed Optuna parameter tuning shows 11.73% of improvement in mean prediction error for the specific Hair Salon sales in comparison to the Prophet prediction without Optuna.

*4)* When the events of typhoons, heavy rain, pension payment date, entrance ceremonial date, etc. are considered in the proposed Optuna parameter tuning, then the sale prediction error is reduced.

*5)* The effect of the considering the entrance ceremony, graduation ceremony, and pension payment days is less than 2%.

*6)* Influence of Covid-19 on the sales prediction is clarified. If the influence is considered in the sales prediction processes, MAPE is improved from 8.2 to 10.6 %.

## VI. Future Reaesrch Work

Further investigations are required for improvement of prediction accuracy by considering the other influencing factors such as coupon, special campaign, etc. to the sales. Weather forecast data, geolocation, population, environmental factors are other candidates of the influencing factors.

## Acknowledgment

The authors would like to thank to the cooperative staff of the Success Institute Chain: SIC Holdings Co., Ltd, for their effort to conduct this research work. Also, the authors would like to thank to Professor Dr. Hiroshi Okumura and Professor Dr. Osamu Fukuda for their valuable discussions.

## References

[1] Achmad Basuki, Tri Harsono and Kohei Arai, Probabilistic cellular automata based approach for prediction of hot mudflow disaster area and volume, Journal of EMITTER1, 1, 11-20, 2010.

[2] Kohei Arai, Achmad Basuki, New Approach of Prediction of Sidoarjo Hot Mudflow Disaster Area Based on Probabilistic Cellular Automata, Geoinformatica - An International Journal (GIIJ), 1, 1, 1-11, 2011.

[3] Kohei Arai, Achmad Basuki, GIS based 2D cellular automata approach for prediction of forest fire spreading, International Journal of Research and Reviews on Computer Science, 2, 6, 1305-1312, 2011.

[4] Kohei Arai, Cell based GIS as Cellular Automata for disaster spreading prediction and required data systems, CODATA Data Science Journal, 137-141, 2012.

[5] Kohei Arai, A.Basuki, T.Harsono, Hot mudflow prediction area model and simulation based cellular automata for LUSI and plume at Sidoarjo East Jawa, Journal of Computational Science (Elsevior) 3,3,150-158, 2012.

[6] Kohei Arai, Comparative Study between Eigen Space and Real Space Based Image Prediction Methods by Means of Autoregressive Model, International Journal of Research and Reviews in Computer Science (IJRRCS) Vol. 3, No. 6, 1869-1874, December 2012, ISSN: 2079-2557.

[7] Kohei Arai, Comparative Study on Image Prediction Methods between the Proposed Morphing Utilized Method and Kalman Filtering Method, International Journal of Research and Reviews in Computer Science (IJRRCS) Vol. 3, No. 6, 1875-1880, December 2012, ISSN: 2079-2557.

[8] Kohei Arai Prediction method for time series of imagery data in eigen space, International Journal of Advanced Research in Artificial Intelligence, 2, 1, 12-19, (2013).

[9] Kohei Arai Image prediction method with non-linear control lines derived from Kriging method with extracted feature points based on morphing, International Journal of Advanced Research in Artificial Intelligence, 2, 1, 20-24, (2013).

[10] Kohei Arai, Cell based GIS as cellular automata for disaster spreading predictions and required data systems, Advanced Publication, Data Science Journal, Vol.12, WDS 154-158, 2013.

[11] Kohei Arai, Prediction method of El Nino Southern Oscillation event by means of wavelet based data compression with appropriate support length of base function, International Journal of Advanced Research in Artificial Intelligence, 2, 8, 16-20, 2013.

[12] Kohei Arai, Anik Nur Handayani, Question Answering for collaborative learning with answer quality prediction, International Journal of Modern Education and Computer Science, 5, 5, 12-17, 2013.

[13] Kohei Arai, Shohei Fujise, Wildlife Damage Estimated and Prediction Using Blog and Tweet Information, International Journal of Advanced Research on Artificial Intelligence, 5, 4, 15-21, 2016.

[14] Kohei Arai, Prediction method for large diatom appearance with meteorological data and MODIS derived turbidity as well as chlorophyll-a in Ariake Bay area in Japan, International Journal of Advanced Computer Science and Applications IJACSA,8,3,39-44, 2017.

[15] Kohei Arai, Method for Thermal Pain Level Prediction with Eye Motion using SVM, International Journal of Advanced Computer Science and Applications IJACSA, 9, 4, 170-175, 2018.

[16] Kohei Arai, Prediction method for large diatom appearance with meteorological data and MODIS derived turbidity and chlorophyll-a in Ariake bay area in Japan, International Journal of Advanced Computer Science and Applications IJACSA, 10, 9, 39-44, 2019.

[17] Kohei Arai, Osamu Shigetomi, Yuko Miura, Satoshi Yatsuda, Smartphone image based agricultural product quality and harvest amount prediction method, International Journal of Advanced Computer Science and Applications IJACSA, 10, 9, 24-29, 2019.

[18] Kohei Arai, Data Retrieval Method based on Physical Meaning and its Application for Prediction of Linear Precipitation Zone with Remote Sensing Satellite Data and Open Data, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 10, 56-65, 2020.

[19] Kohei Arai, Kaname Seto, Recursive Least Square: RLS Method-Based Time Series Data Prediction for Many Missing Data, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 11, 66-72, 2020.

[20] Kohei Arai, Prediction of Isoflavone Content in beans with Sentinel-2 Optical Sensor Data by Means of Regressive Analysis, Proceedings of SAI Intelligent Systems Conference, IntelliSys 2021: Intelligent Systems and Applications pp 856-865, 2021.

## Authors' Profile

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is a Science Council of Japan Special Member since 2012. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Science Commission "A" of ICSU/COSPAR since 2008 then he is now award committee member of ICSU/COSPAR. He wrote 77 books and published 680 journal papers as well as 550 conference papers. He received 66 of awards including ICSU/COSPAR Vikram Sarabhai Medal in 2016, and Science award of Ministry of Mister of Education of Japan in 2015. He is now Editor-in-Chief of IJACSA and IJISA http://teagis.ip.is.saga-u.ac.jp/index.ht