IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 13, No. 2, June 2024, pp. 1632~1639
ISSN: 2252-8938, DOI: 10.11591/ijai.v13.i2.pp1632-1639

An boosting business intelligent to customer lifetime value with robust M-estimation

Marischa Elveny[1], Rahmad B. Y. Syah[2], Mahyuddin K. M. Nasution[1]
[1]Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia
[2]Excellent Centre of Innovations and New Science, Faculty of Engineering, Universitas Medan Area, Medan, Indonesia

Article Info

Article history:

Received Feb 18, 2023
Revised May 22, 2023
Accepted Jun 30, 2023

Keywords:

Business intelligent
Customer lifetime value
M-estimation
Optimizing
Robust

ABSTRACT

When a business concentrates too much on acquiring new clients rather than retaining old ones, mistakes are sometimes made. Each customer has a different value. Customer lifetime value (CLV) is a metric used to assess long-term customer value. Customer value is a key concern in any commercial endeavor. When there are variations in customer behavior, CLV forecasts the value of total customer income when the data distribution is not normal, and outliers are present. Robust M-estimation, a maximum likelihood type estimator, is used in this study to enhance CLV data. Through the minimization of the regression parameter from the residual value, robust M-estimation eliminates data outliers in customer metric data. With an accuracy of 94.15%, R-square is used to gauge model performance. This research shows that CLV optimization can be used as a marketing and sales strategy by companies.

Corresponding Author:

Marischa Elveny
Faculty of Computer Science and Information Technology, Universitas Sumatera Utara
Padang Bulan 20155 USU, Medan, Indonesia
Email: marischaelveny@usu.ac.id

1. INTRODUCTION

Due to the intricacy of the competition in today's market, innovations must be created to boost client happiness [1]. Businesses frequently make the error of putting too much emphasis on acquiring new clients rather than providing for existing ones. In order to ensure economic value, modern power systems must comprehend real-time, comprehensive, and quick operating system operations through analysis and trend prediction [2].

Companies need measuring equipment to evaluate the effectiveness of business initiatives and how the company is positioned with customers [3]. Customer lifetime value (CLV) is a metric for the total revenue an organization anticipates receiving from a single customer. This metric displays a customer's value to a company based on each purchase they have made since the initial one [4]. When generating data to monitor business growth and make important sustainable business decisions, CLV measurement takes into account consumer purchasing behavior and patterns [5].

According to benefit segmentation, this study tries to evaluate clients from each segmentation. The fundamental components of CLV estimates include customer lifetime value analysis. For the company's marketing strategy to be effective, each segment has specific requirements that must be met by the appropriate marketing plans and initiatives [6], [7].

Real-life optimization issues involve uncertain data. Data can be incorrect or random by nature. Lack of understanding of model parameters, such as uncertain demand in model supply, or mistakes in the implementation of solutions in real-world situations, are causes of measurement/estimation data errors [8].

Outliers are caused by the extensive complexity of client data. By examining the CLV data for outliers, optimization is carried out [9]. With robust regression, when the final model incorporates estimates, outlier data can be handled. The capacity of the estimator to endure disruptions is what is meant by the crucial estimation performance index robust [10]. The article suggests a robust M-estimate approach to optimize CLV based on sampling in order to solve this issue by improving the overall state estimation accuracy by correctly changing the robustness [11].

2. METHODS

Data is provided by e-metrics merchants. The research is conducted in stages, beginning with an analysis of business needs by looking at objects, market needs, customer needs, and business goals. The data is also processed by calculating the CLV [12]. The information is separated into training and testing data. Training and testing data is shared between 70% and 30%. Where the training data is utilized to develop and train a regression prediction model using robust M-estimation. This method will evaluate the training model's performance at work and generate predictions in order to construct a robust regression model against outliers. Meanwhile, data testing is used to assess the capacity of the previously developed robust regression model to predict data outside of the training data. Robust M-estimation from outlier data on CLV is used for optimization. Figure 1 depicts the steps for resolving the problem.

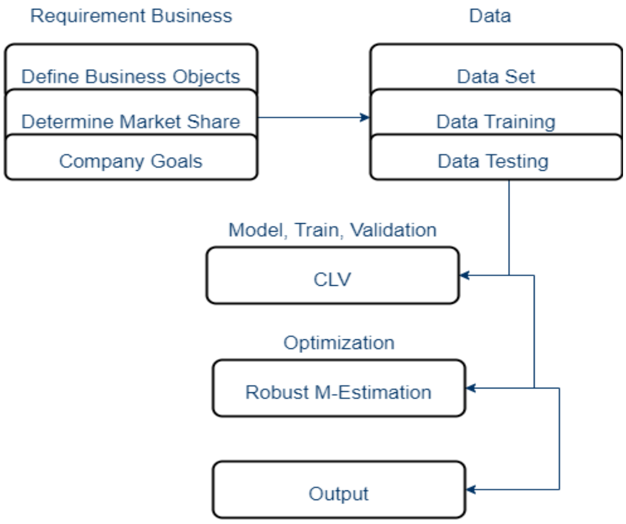


Figure 1. State of the art

2.1. Customers lifetime value

CLV considers retention time, or the period of customer interaction with the business [13]. Customers who make repeat purchases of products over time have a higher value than customers who make one-time purchases [14]. In (1) and (2), the basic formula for calculating CLV is given [15].

$$CLV_{a,b} = \sum_{a=0}^{B} \frac{profit_{a,b}}{1+x)^{b}} \quad (1)$$

$$CLV_{a=1} = \sum_{a=1}^{B} \frac{Revenue_{a,b}}{(1+x)^{a}} - \sum_{a=1}^{B} \frac{Cost_{a,b}}{(1+x)^{a}} \quad (2)$$

A customer's average purchase value represents the average amount of money a customer spends on each purchase, which can be calculated by dividing total revenue by the number of purchases. The number of repeat purchases can be calculated by dividing the total number of purchases by the number of unique customers, and the average customer age is measured by analyzing historical customer data to determine the duration of a customer relationship. These findings define potential customer value as the profit expected from specific customers when using additional services, as shown in (3) [16].

$$Potential\ Value = \sum_{a-1}^{x} probability_{pj} \times profit_{pj} \quad (3)$$

2.2. Robust M- estimation

Outliers are data that deviate from the norm. Outlier was detected using the difference in fitted value (DFFITS') which can be seen in the stages below [17],

- Create a regression model: begin by fitting the linear regression regression model to the data.
- Calculate DFFITS: after fitting the model, compute the DFFITS statistic for each observation. DFFITS tracks changes in anticipated value.
- Establish a DFFITS statistical criterion to identify outliers. Observations with DFFITS values that surpass this limit are considered outliers.
- Outlier detection: compare the DFFITS value for each observation to a threshold. Mark the applicable observation as a potential outlier if the DFFITS value exceeds the threshold.

In (4) and (5) show the formula for determining DFFITS.

$$DFFITS_i = a_i\sqrt{\frac{b_{ii}}{1-b_{ii}}} \tag{4}$$

$$b_{ii} = r_i'(R'R)^{-1}r_i \tag{5}$$

The dependent variable's value is determined by patterns or events in the independent variable; values range from negative to positive numbers. As a result, this process generates a model with target values of 0 and 1. The graphical method is used in this study to determine whether the model violates any of its classical assumptions [18]. In order to determine the relationship between the dependent variables, a regression analysis was performed, which included a systematic calculation of the data [19]. The regression analysis results are the regression models with the highest accuracy when used to explain data trends. It is evident in (6) [20].

$$R_a = \beta_c + \beta_a\beta_a + \beta_p X_p + e \tag{6}$$

R_a = Dependent variable
$\beta_c\ dan\ \beta_a$ = Parameter estimation function
X_p = Independent variable
e = Error value

By minimizing the regression parameter from the residual value, this study employs M-estimation robust regression (Maximum likelihood type estimator) to robustly address outlier data on merchant customer e-metric data. The steps for measuring robust M-estimation are,
robust M-estimation measures,

- Mean absolute deviation (MAD) calculation is the average absolute deviation in a data center.
- The value is compared to the initial residual value to determine the new residual value.
- Add the initial estimated value to the new residual value to get the new estimated value.
- Rerun the regression analysis with the new estimated value as the dependent variable.
- Repeat until the iterations have reached a point of convergence.

As shown in (7) [21]–[24].

$$\beta_s = \sum_{i=x}^{x} p\,(e_i) \tag{7}$$

β_s = Regression coefficient
p = Parameter
e_i = Residual

3. RESULTS AND DISCUSSION

3.1. Understanding of business to customers

Consumer behavior is the study of how people, groups and organizations select, buy, use and dispose of goods, services, ideas or experiences to meet needs and wants. Consumer behavior can change over time as consumers' needs and desires become more diverse and keep up with the times. Can be seen in Figure 2. Customers' perspectives on the decision-making lifecycle differ from how companies design acquisition-centric initiatives; efforts are made to select analytical techniques and align lifecycle programs with how customers make decisions [25].

Figure 1. Customer side view

3.2. Data preparing customers lifetime value

This stage consists of four steps: data reduction, feature selection, data transformation, and data normalization. The date of the last purchase, calculating customer purchases, total income from customers for one year, and calculating product items purchased by customers are all features [26], [27]. The final feature is based on the opinion of an expert. Tables 1-3 shows process data.

3.3. Optimization robust M-estimation

The size of the DFFITS value determines detection optimization. DFFITS is used to calculate the fit value of an i-th observation on the regression model. In (4) and (5) are utilized with observational values of 11, 13, 17, 23, 35 based on Tables 2 and 3. Table 4 shows the magnitude of the DFFITS value.

Table 1. Data preparing

Preliminary Data	Process Data
Type Data	Number Customer
Type Transaction	Frequency
Type Customers	Behaviour
Total Spending	Calculate Costs
Total Income	Count Purchase

Table 2. Customer segmentation based on CLV calculations

Data	Number Customer	Frequency	Behaviour	Calculate Costs	Count Purchase
A1	234	36.897	67.383	2.97	5637.83
A2	124	56.862	11.537	1.67	2511.12
A3	112	12.638	324.77	34.67	3638.26
A4	342	63.738	52.729	22.43	5373.89
A5	135	66.353	112.73	112.64	3441.63

Table 3. Estimating CLV value

Data	Percent of customers	Number Customer	Frequency	Behaviour	Calculate Costs	Count Purchase	CLV Value
A1	17	234	36.897	67.383	2.97	5637.83	0.6788
A2	15	124	56.862	11.537	1.67	2511.12	0.7899
A3	10	112	12.638	324.77	34.67	3638.26	0.3245
A4	19	342	63.738	52.729	22.43	5373.89	0.5678
A5	16	135	66.353	112.73	112.64	3441.63	0.2134

Table 4. Data observation DFFITS

Observation	$DFFITS_i$
11	2.146
13	1.764
17	1.567
23	0.637
35	1.685

Robust visualizes outliers and results after applying robust regression based on outlier tests [28]. Can be seen in Figure 3, where the outlier test is used to identify and evaluate the presence of outliers in the data set. Outliers are indicated by data points that deviate significantly from most of the data and have a disproportionate impact on analysis and modeling. Figure 3(a) focuses on the linearity test plot and shows that there are violations of the classical assumptions, outliers, and the regression model does not fit the data, it is shown that the data points are displayed far from and not parallel to the lines and diagonal lines, there are non-constant patterns in plots. Then, after M-estimation is applied, Figure 3(b) it can be seen that outliers are not detected, so the model is declared Robust to the outlier data and Fit to the data. In this case, a robust M-estimation system will make comparisons with a minimum of 80 transactions. Starting with the input values based on (6) and (7).

– R_1=100000 and R_2=100
β_0=- 1.7848598e-01; β_1=6.486588e-07; dan β_2 = -6.486373e-07.
y = -1.7848598e-01+ (6.486588e-07*100000) + (-6.486373e-07*100)
y = -1.7848598e-01+ 0.006486588 + -0.06486373
y = -1.67E-01

– R_1=200000 and R_2=200
y = -1.7848598e-01+ (6.486588e-07*200000) + (-6.486373e-07*200)
y = -1.7848598e-01+ 0.065547 + -0.1283746
y = -2.26E-01

– R_1=300000 and R_2 =300
y = -1.7848598e-01+ (6.486588e-07*300000) + (-6.486373e-07*300)
y = -1.7848598e-01+ 0.05437263 + -0.2847758
y = -4.14E-01

– R_1=500000 and R_2 =500
y = -1.7848598e-01+ (6.486588e-07*500000) + (-6.486373e-07*500)
y = -1.7848598e-01+ 0.0636889 + -0.63763823
y = -5.46E-01

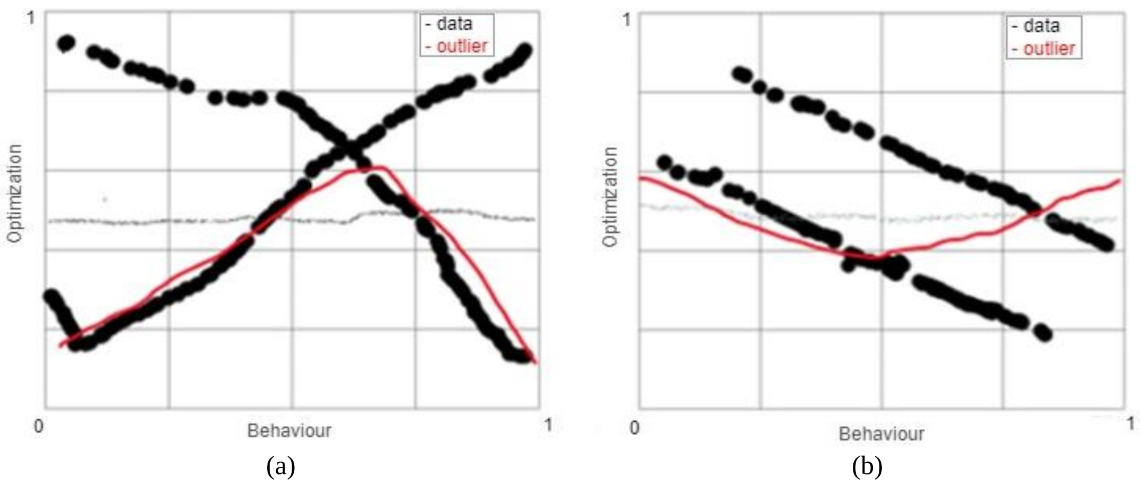


Figure 3. Outliers test, (a) Linearity test and (b) Outlier detection

3.4. Validation

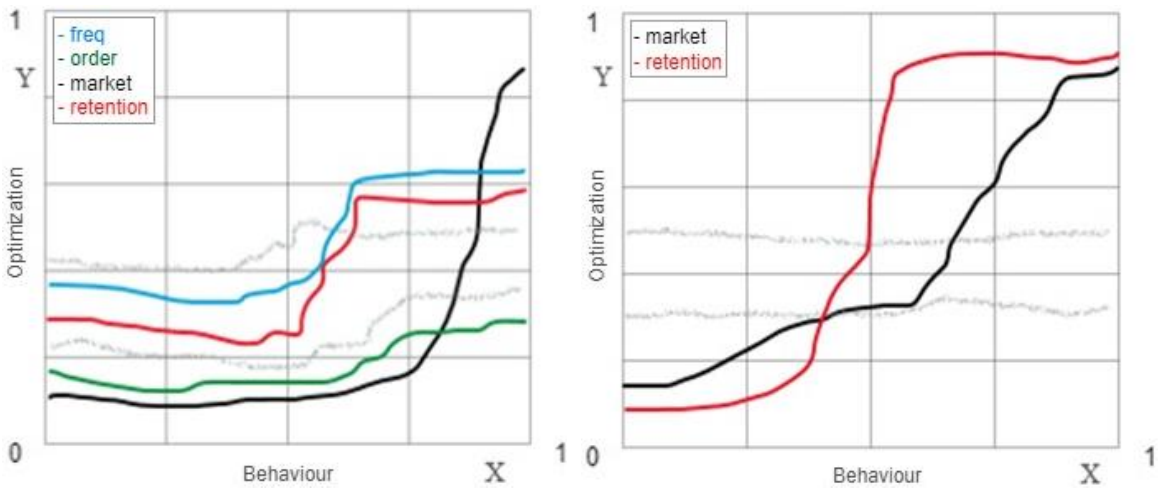
To construct an appropriate model, robust regression functions must be used to determine the relationship between the dependent variable and the independent variable without ignoring outliers. As a result, the value of the dependent variable (Y) is determined by patterns or events in the independent variable (X), allowing its development to range from negative to positive numbers. Validation employs R-Squared, which indicates how much the independent variable (exogenous) influences the dependent variable (endogenous) by measuring the influence of specific latent independent factors on the latent dependent variable. R-Squared has a value between 0 and 1, with the closer to one, the better [29].

$$R^2 = 1 - \frac{S\,Regression}{S\,Total} \tag{8}$$

$$R^2 = 1 - \frac{\sum(x_i - x_i)^2}{\sum(x_i - x)^2} \tag{9}$$

$$= 1 - \frac{8}{83468340} = 1 - 0.0584472 = 0.9415528 \times 100\% = 94.15\%$$

Because the R-Squared value is close to one 1, the model is considered good. In contrast, if the R-Squared value approaches zero 0, the model becomes weaker [30]. Based on validation results, Figure 4 depicts the validation of the robust M-estimation optimization model, where the results are comparable 1. Shown in Figure 4(a), purchase frequency (blue): How often customers make purchases tends to have a higher CLV compared to those who make purchases occasionally. Average order value (green): The average amount a customer spends per transaction tends to provide more value. Retention rate (red): The ability to retain a customer over a period of time, which can reduce churn and promote repeat purchases, tends to have a higher CLV. Market conditions and competitive landscape (black): External factors such as market conditions and competitive landscape can influence customer behavior, and consequently, changes in economic conditions, industry trends, or the entry of new competitors can impact customer loyalty and spending patterns. Figure 4(b), retention rate (red) Market conditions and the competitive landscape (black) have a greater influence.



(a) (b)

Figure 4. Validation models, (a) Behavior category and (b) Impact

4. CONCLUSION

Customers lifetime value depicts the customer's worth to the company based on their first purchase. CLV measurement takes into account purchasing behavior and patterns and is dependent on the company's ability to conduct the analysis. CLV was optimized using robust M-estimation, and the results were 94.15% with an R-Squared close to 1, indicating that the model was good. CLV can assist businesses in implementing strategies to make customers last longer based on the results obtained.

ACKNOWLEDMENTS

Thanks to Universitas Sumatera Utara for providing research funding for the PKUU equity project program scheme with contract number 7092.1/UN5.1.R/PPM/2023.

REFERENCES

[1] A. O. Dandis, M. B. Al Haj Eid, R. Robin, and N. Wierdak, "An empirical investigation of the factors affecting customer lifetime value," *International Journal of Quality and Reliability Management*, vol. 39, no. 4, pp. 910–935, Jun. 2022, doi: 10.1108/IJQRM-12-2020-0412.

[2] A. Mosaddegh, A. Albadvi, M. M. Sepehri, and B. Teimourpour, "Dynamics of customer segments: A predictor of customer lifetime value," *Expert Systems with Applications*, vol. 172, p. 114606, Jun. 2021, doi: 10.1016/j.eswa.2021.114606.

[3] W. D. Dahana, Y. Miwa, and M. Morisada, "Linking lifestyle to customer lifetime value: An exploratory study in an online fashion retail market," *Journal of Business Research*, vol. 99, pp. 319–331, Jun. 2019, doi: 10.1016/j.jbusres.2019.02.049.

[4] P. Jasek, L. Vrana, L. Sperkova, Z. Smutny, and M. Kobulsky, "Modeling and application of customer lifetime value in online retail," *Informatics*, vol. 5, no. 1, p. 2, Jan. 2018, doi: 10.3390/informatics5010002.

[5] A. O. Dandis and M. B. Al Haj Eid, "Customer lifetime value: investigating the factors affecting attitudinal and behavioural brand loyalty," *TQM Journal*, vol. 34, no. 3, pp. 476–493, Jun. 2022, doi: 10.1108/TQM-12-2020-0311.

[6] H. Castéran, L. Meyer-Waarden, and W. Reinartz, "Modeling customer lifetime value, retention, and churn," in *Handbook of Market Research*, Springer International Publishing, 2022, pp. 1001–1033.

[7] S. Monalisa, P. Nadya, and R. Novita, "Analysis for customer lifetime value categorization with RFM model," *Procedia Computer Science*, vol. 161, pp. 834–840, 2019, doi: 10.1016/j.procs.2019.11.190.

[8] M. De Marco, P. Fantozzi, C. Fornaro, L. Laura, and A. Miloso, "Cognitive analytics management of the customer lifetime value: an artificial neural network approach," *Journal of Enterprise Information Management*, vol. 34, no. 2, pp. 679–696, Feb. 2021, doi: 10.1108/JEIM-01-2020-0029.

[9] D. Q. F. de Menezes, D. M. Prata, A. R. Secchi, and J. C. Pinto, "A review on robust M-estimators for regression analysis," *Computers and Chemical Engineering*, vol. 147, p. 107254, Apr. 2021, doi: 10.1016/j.compchemeng.2021.107254.

[10] C. Yang, W. Shi, and W. Chen, "Robust M–M unscented Kalman filtering for GPS/IMU navigation," *Journal of Geodesy*, vol. 93, no. 8, pp. 1093–1104, Jan. 2019, doi: 10.1007/s00190-018-01227-5.

[11] P. S. Dąbrowski and M. Hubert Zienkiewicz, "Impact of cross-section centers estimation on the accuracy of the point cloud spatial expansion using robust M-estimation and Monte Carlo simulation," *Measurement: Journal of the International Measurement Confederation*, vol. 189, p. 110436, Feb. 2022, doi: 10.1016/j.measurement.2021.110436.

[12] N. Donthu, S. Kumar, and D. Pattnaik, "Forty-five years of Journal of Business Research: A bibliometric analysis," *Journal of Business Research*, vol. 109, pp. 1–14, 2020, doi: 10.1016/j.jbusres.2019.10.039.

[13] P. P. Chen, A. Guitart, A. F. Del Río, and A. Periáñez, "Customer lifetime value in video games using deep learning and parametric models," in *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, Dec. 2019, pp. 2134–2140, doi: 10.1109/BigData.2018.8622151.

[14] L. L. (Luke) Chiang and C. S. Yang, "Does country-of-origin brand personality generate retail customer lifetime value? A Big Data analytics approach," *Technological Forecasting and Social Change*, vol. 130, pp. 177–187, May 2018, doi: 10.1016/j.techfore.2017.06.034.

[15] Y. Sun, H. Liu, and Y. Gao, "Research on customer lifetime value based on machine learning algorithms and customer relationship management analysis model," *Heliyon*, vol. 9, no. 2, p. e13384, Feb. 2023, doi: 10.1016/j.heliyon.2023.e13384.

[16] B. Hajipour and M. Esfahani, "Delta model application for developing customer lifetime value," *Marketing Intelligence and Planning*, vol. 37, no. 3, pp. 298–309, May 2019, doi: 10.1108/MIP-06-2018-0190.

[17] K. Rajabi Khamseh, A. Nikbakht Shahbazi, H. Fathian, and N. Zohrabi, "Determination of critical aquifer subsidence points by DFFITS-COOK distance method using Importance sampling reliability analysis," *Ain Shams Engineering Journal*, vol. 14, no. 4, p. 101915, Apr. 2023, doi: 10.1016/j.asej.2022.101915.

[18] J. H. Sullivan, M. Warkentin, and L. Wallace, "So many ways for assessing outliers: What really works and does it matter?," *Journal of Business Research*, vol. 132, pp. 530–543, Aug. 2021, doi: 10.1016/j.jbusres.2021.03.066.

[19] S. Wang, L. Zhu, J. Y. H. Fuh, H. Zhang, and W. Yan, "Multi-physics modeling and Gaussian process regression analysis of cladding track geometry for direct energy deposition," *Optics and Lasers in Engineering*, vol. 127, p. 105950, Apr. 2020, doi: 10.1016/j.optlaseng.2019.105950.

[20] V. W. Y. Tam, A. Butera, K. N. Le, L. C. F. D. Silva, and A. C. J. Evangelista, "A prediction model for compressive strength of CO2 concrete using regression analysis and artificial neural networks," *Construction and Building Materials*, vol. 324, p. 126689, Mar. 2022, doi: 10.1016/j.conbuildmat.2022.126689.

[21] D. Barrow, N. Kourentzes, R. Sandberg, and J. Niklewski, "Automatic robust estimation for exponential smoothing: Perspectives from statistics and machine learning," *Expert Systems with Applications*, vol. 160, p. 113637, Dec. 2020, doi: 10.1016/j.eswa.2020.113637.

[22] G. Boente, M. Salibian-Barrera, and P. Vena, "Robust estimation for semi-functional linear regression models," *Computational Statistics and Data Analysis*, vol. 152, p. 107041, Dec. 2020, doi: 10.1016/j.csda.2020.107041.

[23] A. Khan *et al.*, "Robust extreme learning machine using new activation and loss functions based on m-estimation for regression and classification," *Scientific Programming*, vol. 2022, pp. 1–10, Oct. 2022, doi: 10.1155/2022/6446080.

[24] M. J. Tarokh and M. EsmaeiliGookeh, "Modeling patient's value using a stochastic approach: An empirical study in the medical industry," *Computer Methods and Programs in Biomedicine*, vol. 176, pp. 51–59, Jul. 2019, doi: 10.1016/j.cmpb.2019.04.021.

[25] P. Song and H. Zhao, "Robust diffusion affine projection M-estimate algorithm for distributed estimation over network," *IFAC-PapersOnLine*, vol. 52, no. 24, pp. 290–293, 2019, doi: 10.1016/j.ifacol.2019.12.423.

[26] N. S. Shaeeali, A. Mohamed, and S. Mutalib, "Customer reviews analytics on food delivery services in social media: A review," *IAES International Journal of Artificial Intelligence*, vol. 9, no. 4, pp. 691–699, Dec. 2020, doi: 10.11591/ijai.v9.i4.pp691-699.

[27] R. Syah, M. K. M. Nasution, E. B. Nababan, and S. Efendi, "Sensitivity of shortest distance search in the ant colony algorithm with varying normalized distance formulas," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 19, no. 4, pp. 1251–1259, Aug. 2021, doi: 10.12928/TELKOMNIKA.v19i4.18872.

[28] Al-Khowarizmi, R. Syah, M. K. M. Nasution, and M. Elveny, "Sensitivity of MAPE using detection rate for big data forecasting crude palm oil on k-nearest neighbor," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 3, pp. 2696–2703, Jun. 2021, doi: 10.11591/ijece.v11i3.pp2696-2703.

[29] K. Taira, D. McInnes, and L. Zhang, "How many data points and how large an R-squared value is essential for Arrhenius plots?," *Journal of Catalysis*, vol. 419, pp. 26–36, Mar. 2023, doi: 10.1016/j.jcat.2023.01.033.

[30] R. Di Mari, S. Ingrassia, and A. Punzo, "Local and overall deviance R-squared measures for mixtures of generalized linear models," *Journal of Classification*, Apr. 2023, doi: 10.1007/s00357-023-09432-4.

BIOGRAPHIES OF AUTHORS

Marischa Elveny received (bachelor's) degree in information technology at the universitas sumatera utara and earned his M.Kom. (master) degree in informatics engineering, university of Sumatera Utara. Doctorate (Ph.D.) at the University of Sumatera Utara. Currently working as a lecturer in the faculty of computer science and information technology at the Universitas Sumatera Utara. Research interests are Artificial Intelligence, Data Science and Computational Intelligence. Can be contacted via email: marischaelveny@usu.ac.id.

Rahmad B. Y. Syah is an Associate Professor in Universitas Medan Area and received head of Excellent Centre Innovations and New Science (PUIN) Universitas Medan Area. His research interests are modeling and computing, artificial intelligence, Data Science, Business Intelligent, Metaheuristics Hybrid Algorithm and Computational Intelligent. He is a member of the IEEE, Institute for System and Technologies of Information, Control, and Communication, IAENG (International Association of Engineers). He should be contacted at email: rahmadsyah@uma.ac.id.

Mahyuddin K. M. Nasution is a Professor from Universitas Sumatera Utara, Medan Indonesia. Worked as a Lecturer at the Universitas Sumatera Utara, is an Drs. Mathematics (USU Medan, 1992); MIT, Computers and Information Technology (UKM Malaysia, 2003); PhD in Information Science (Malaysian UKM). fields: Mathematics, computer science, Data Science, Artificial Intelligence, Modelling and Computation, New Science, and Information Technology. Can be contacted via email mahyuddin@usu.ac.id.