Prediction Analysis Sales for Corporate Services Telecommunications Company using Gradient Boost Algorithm

Oryza Wisesa
Department of Electrical Engineering,
Universitas Mercu Buana,
Jakarta, Indonesia
oryza49@gmail.com

Andi Adriansyah
Department of Electrical Engineering,
Universitas Mercu Buana,
Jakarta, Indonesia
andi@mercubuana.ac.id

Osamah Ibrahim Khalaf
College of Information Engineering,
Al-Nahrain University
Bagdad, Iraq
usama.ibrahem@coie.nahrain.edu.iq

***Abstract*—Sales prediction analysis requires smart data mining techniques with accurate prediction models and high reliability. Essentially, most market segments rely on the know-how base and the demand trend forecast for analysis of Business To Business (B2B) sales data. Data are provided by sales on how Telecommunication Company should manage its sales team, its products and also its budgeting flows. Precise estimates make it possible for Telecommunication Company to survive the market war and increase its market growth. In this research, the study and analysis of comprehensible predictive models use machine learning techniques to improve future sales predictions. Traditional forecasting systems are difficult to deal with big data and the accuracy of sales forecasting. In this paper, a brief analysis of the reliability of B2B sales using machine learning techniques. The latter part of this research explains a range of sales prediction strategies and interventions. Based on the performance assessment, a best-adapted predictive model for the B2B sales trend forecast is suggested. Projection, estimation and analysis findings are summarized in terms of reliability and consistency of efficient prediction and forecasting techniques. The results of this analysis are expected to generate reliable, accurate and effective forecasting data, a valuable resource for sales predictions. Research has shown that Gradient Boost Algorithm shows good accuracy in forecasting and future B2B sales prediction with MSE = 24,743,000,000.00, and MAPE: 0.18.**

***Keywords*—gradient boosted trees, prediction, reliability, sales forecasting, business to business (B2B), telecommunication**

I. INTRODUCTION

Sales forecast analysis includes smart data mining techniques with accurate forecasting models and high reliability. Sales forecasts provide details on how an organization can handle its sales force, its goods and even its budget flows. Accurate forecasts enable organizations to raise the highest potential amount of sales in line with market growth. Data mining is very successful in the processing of extensive volume data into useful cost information and revenue forecasts. Based on the background above, this research will focus on comparing forecast analysis and Business to Business (B2B) sales reliability using machine learning techniques.

The problem in this study is how to determine the type of prediction analysis that is accurate and accurate by comparing four machine learning techniques in B2B sales with data for 2016-2017 and 2018. The results of this analysis are expected to generate reliable, accurate and useful forecasting data, a valuable resource for sales predictions.

II. LITERATURE REVIEW

In [1] multiple prediction techniques and methods are studied. Comparative machine tuning analysis and various clustering algorithms are tested on sales data [2]. In decision-taking, as discussed in [3], data classification is significant. The methods of clustering are useful for evaluating distribution patterns and clustering algorithms by distance measurements. The data from a broad collection of data can be converted into a responsive format using useful techniques for data mines and can be done by supervised and unattended learning [4]. An accurate sales forecast methodology can be used to make successful business decisions. The terminology and algorithms are discussed in [5]. The research is conducted using various data mining technologies to forecast sales of the B2B drug. The goal was to forecast sales of all sorts of goods so that we could get to know the problem we learned before. Several studies have emerged from this problem to help predict and predict sales. The research attempted to develop an analysis system using machine-learning predictive methods.

Throughout 2015, D'Arcy, Gallagher and Madden [6] researched many companies using predictive techniques to determine whether or not revenue incentives will be received. Three key differences were identified in alternative methods to estimate sales of chance: (1) qualitative evaluations augmented by quantitative data that explain opportunity; (2) Replacing the mass factor with the classification of the increased Naïve Bayes Tree (TAN), where dependencies between variables are collected and probabilistic outcomes are generated where thresholds exist.; (3) Historical data are used for TAN studying, while the existing Quantstamp (QSP) is fixed. For an exactness of 90.6 percent from the approach taken, profits will either be gained or lost, the exactness of the current approach is improved significantly by around 75.6 percent.

In 2018, the research solution for estimating house levels with the Regressive Technology was conducted by Leopoldovich, Viktorovich and Viktor Aleksandrovich [7], which aims to forecast housing prices using the values of attributes of the structure, the house area etc. This research employs classical algorithms for machine learning and explains the original process. Wu, Patil, Saravana Gunaseelan, Ching-she, Pratik Patil [8] conducted research was carried out on Black Friday (discount days) to develop a precise and efficient algorithm for analyzing past customer expenses and future customers' output with the same features. Within this analysis, various machine learning approaches are

978-1-7281-7450-1/20/$31.00 ©2020 IEEE

applied, such as regression and neural newer networks, for the creation of predictive models. Such methods are implemented using different algorithms and several tools for the best estimate. For this study, there were seven separate computer studies algorithms. Subsequently, work on the definition of sales data, and the revenue estimate was carried out in 2018, by Cheriyan, Ibrahim, Mohanan, Treesa [9]. The next part of the research will explain various techniques and sizes for sales predictions. Based on performance evaluations, the most effective predictive model for forecasting trends in sales is suggested. The results of the predictive and predictive results are summarized by the reliability and accuracy of effective techniques. Studies showed that the Gradient Boost Algorithm is the most suitable model, offering maximum precision in projections and future sales predictions.

In 2019, Ullah, Raza, Malik, Imran, Islam, Kim [10] researched this research by proposing a churn-prediction model using the classification and clustering methods for identifying clients in the churning telecoms industry and providing factors behind churning customers. Feature selection is performed using the correlation attribute filter and gain information. The first suggested model classifies consumer data using the Random Forestry (RF) algorithm with 88.63 percent correctly categorized RF algorithm. CRM is an essential activity to prevent the creation of efficient retention. Once classified, the proposed model segments are categorized into groups that combine with cosine similarities to offer group-based retention services to customers. In addition the churn element, essential in determining the root cause of churn is defined by this study. CRM will improve efficiency, suggest effective incentives to churn customer groups based on the same patterns of activity by understanding significant churning factors from consumer data and over-intensifying marketing strategies for the client. Metrics such as accuracy, precision, recall, f-size and Receiving Operating Characteristics (ROC) will be evaluated in the proposed churn prediction model. The results show that the RF Algorithm and Client Profiles with K-means clustering are used to boost the churn-prediction model. In addition, the churn customer churn factor is also given by rules created using the selected classification algorithm attribute.

III. Method

The aim of this research is to evaluate and analyze the use of machine learning techniques for sales prediction to know which machine learning techniques are more reliable for forecasting B2B Sales.

A. Data Collection and Preparation

For the three consecutive years of sales data, the data used for this analysis is based on the B2B revenue. In order to forecast B2B revenue, historical record revenue for three years was obtained between 2016, 2017 and 2018. The database covers a category, region, item type and opportunity ID, quarter, product name, product sub-component, service product (MIDI) and sales revenue. Originally the data contained a large number of entries. Still, we have eliminated several non-usable data, redundant and irrelevant for the last selected data collection, which is the quarter, year and service items, and quantity [17].

B. Exploratory Analysis

After data preprocessing, an exploratory study was carried out [18] to understand the essence of our results better. The experimental study consists of the six necessary data mining measures, as illustrated in Fig. 1.

Data mining processes include data awareness, planning, simulation, assessment and implementation. Table 1 shows an analysis of the sales data collected.

Fig. 2 shows the results of sales generated during the years 2016, 2017 and 2018, respectively, for quarters 1, 2, 3 and 4. Between MIDI and NON-MIDI goods, the sales rise is not necessarily the same.

C. Outlier Detection

All data preprocessing and this process carry out model optimization. More detection can be used to use the model or as a starting point for further optimization and helpful in presenting generic, model-independent information. The main focus is on data quality, in particular, the quality of each attribute of the data. We also suggest the elimination of less critical data attributes. Data: after transformation of the dataset for modelling. Correlations: a matrix is showing the relationship between the characteristics of the revenues shown in Table 2 with a positive correlation.

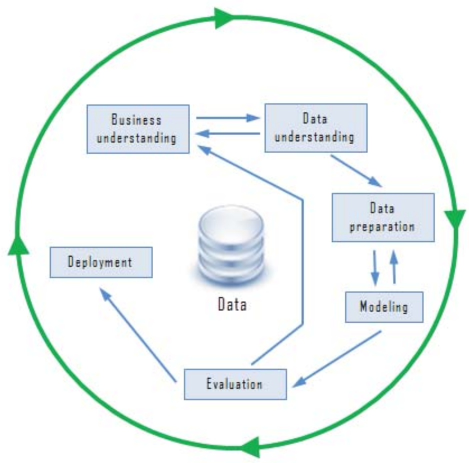

Fig. 1. Data Mining Process

TABLE I. Yearly Sales Data By Product

Sales	Year of Data					
Quarter	**Y2016**		**Y2017**		**Y2018**	
Product	**MIDI**	**NON-MIDI**	**MIDI**	**NON-MIDI**	**MIDI**	**NON-MIDI**
Q1	976 M	228 M	648,047 M	122,800 M	1,937,887 M	850,250 M
Q2	1,031 M	4,292 M	572,748 M	138,784 M	2,112,247 M	764,069 M
Q3	13,875 M	248 M	1,344,548 M	154,728 M	2,099,508 M	1,554,222 M
Q4	71,547 M	4,191 M	1,642,016 M	227,790 M	1,629,096 M	2,342,905 M

Fig. 2. Yearly Sales Graph by Product

D. Forecasting and Trends

The product grouping MIDI and NON-MID also create the pattern I for MIDI and NON-MIDI for forecasting and the future prediction for revenue for Q1 of 2016 to Q4 of 2021. Fig. 3 and Fig. 4 shows further. The trend shows the quarter's total sales revenue. The blue colour shows real sales produced and the orange colour, which shows a small rise in revenues for the current quarter.

Often by consolidating sales and quantities sold in cluster revenues and cluster quantities for each quarter. The result shows a trend line showing that revenue from MIDI product is more, but the quantity is less in Fig. 5 and Fig. 6.

To optimize the forecast, whether moving average, exponential smoothing or another form of a forecast, we need to calculate and evaluate MAD, MSE, RMSE, and MAPE. They are used to determine the precise prediction of the future volume by the trend. Table 3 shows the results.

E. Prediction

The forecast addresses events in the future. Without human interference, enhance computer intelligence by using machine-learning algorithms. Machine Learning (ML), as described in Ethem Alpaydin [19], is intended to refine the results by sampling data and previous experience.

TABLE II. CORRELATION MATRIX

	Quarter	Product	Quantity	Revenue
Quarter	1	0.183809187	0.000935759	0.008941
Product	0.183809187	1	0.006986816	0.020283
Quantity	0.000935759	0.006986816	1	0.050925
Revenue	0.008941057	0.020282845	0.050925088	1

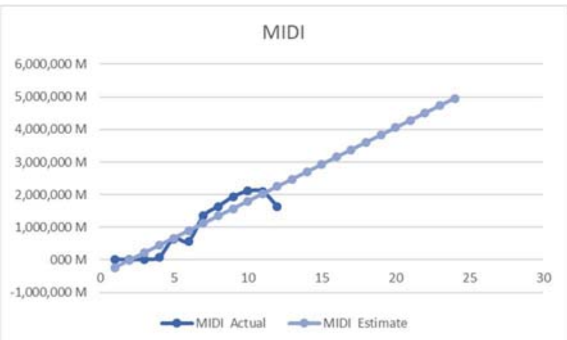

Fig. 3. Forecast for five years MIDI Product

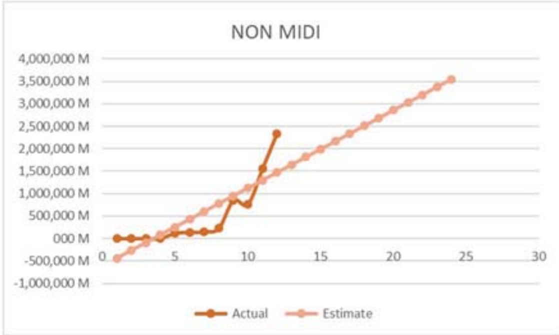

Fig. 4. Forecast for five years NON-MIDI Product

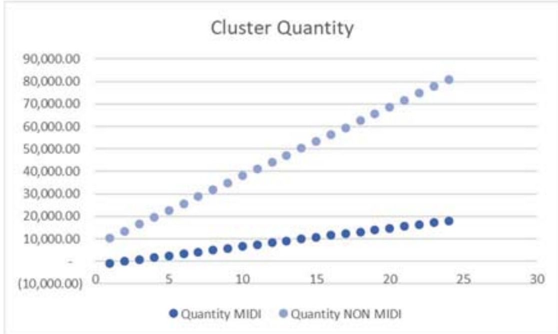

Fig. 5. Trend Analysis using Clusters Quantity

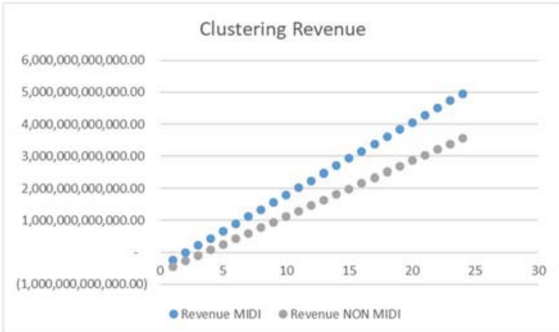

Fig. 6. Trend Analysis using Clusters Revenue

TABLE III. FORECAST TABLE

Period	Actual	Forecast	Error	Absolute Value of Error	Square of Error	Absolute Values of Errors Divided by Actual Values.
t	**At**	**Ft**	**At -Ft**	**| At -Ft|**	**(At -Ft) ^2**	**|(At -Ft)/At|**
Q1 2016	1203M	-676249M	677452M	677452M	458941129576390B	562.978203
Q2 2016	5324M	-276960M	282284M	282284M	79684242583633B	53.02570361
Q3 2016	14124M	122328M	-108204M	108204M	11708119865150B	7.661197668
Q4 2016	75737M	521616M	-445879M	445879M	198807652897953B	5.887167771
Q1 2017	770847M	920904M	-150057M	150057M	22516972605367B	0.194664406
Q2 2017	711532M	1320192M	-608660M	608660M	370467072062137B	0.855421756
Q3 2017	1499276M	1719480M	-220204M	220204M	48489887965007B	0.146873673
Q4 2017	1869806M	2118769M	-248962M	248962M	61982243485751B	0.133148738
Q1 2018	2788137M	2518057M	270080M	270080M	72943140839484B	0.096867522
Q2 2018	2876316M	2917345M	-41029M	41029M	1683346612449B	0.014264289
Q3 2018	3653731M	3316633M	337098M	337098M	113634981190092B	0.092261278
Q4 2018	3972002M	3715921M	256081M	256081M	65577312028869B	0.064471439
Total	**18238035M**	**18238035M**	**0.003051758**	**3645989M**	**1506436101712280B**	**631.1502452**

MAD	303,832,391,369.09
MSE	125,536,341,809,357,000,000,000.00
RMSE	354,311,080,562.49
MAPE	52.60

All the disciplines that be subject to machine learning techniques. In order to solve a number of problems, machine learning uses statistics. They are watched, unattended and half-controlled. Within this study only discuss Gradient Boost Tree (GBT), that could be used for predicting. Fig. 7 shows a system architecture of the process.

In this risk, we have implemented Gradient Boosted Trees algorithms on the training dataset and testing the output models. For the prediction, the best algorithm is selected based on the accuracy of the output.

1) Gradient Boosted Trees

Gradient boosting is a regression and classification method for machine learning. This method could combine a large number of decision-making bodies with a final prediction model [15]. This model builds on the idea that, by using the boosting method, a combination of weak students will generate a strong student. The weak learner is the decision tree [16], a strong additional training method which is required for the inclusion of a new weak learner.

The new tree applied to the model is F(x) complete after round T-1 and H(x).

$$F0 = 0 \tag{1}$$

$$ft(x) = Ft - 1(x) + h(x) \tag{2}$$

The new function attempts to correct the model errors generated in preceding rounds. The new feature(x) must therefore be able to predict the Ft-1(x) residual. Fig. 8 shows gradient boosted tree for this research.

2) Forecast Estimation, Evaluation & Transformation

We must calculate and analyze the MAD, MSE, RMSE and MAPE to optimize the forecast, whether it be average moving, exponential smoothing or a different type of prediction. These are used to assess how correctly the future volume is supposed to be estimated.

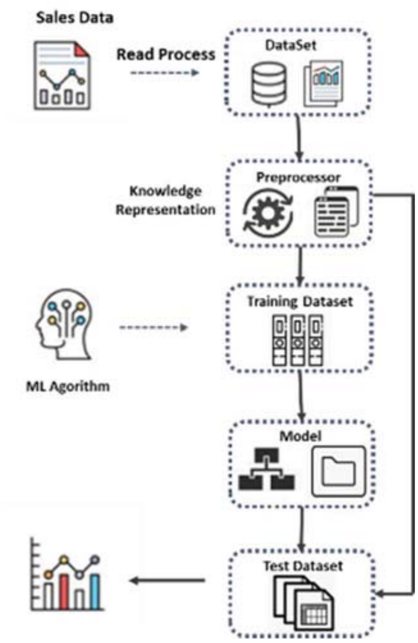

Fig. 7. System Architecture

Fig. 8. Gradient Boosted Tree Model

a) Mean Absolute Percentage Error (MAPE)

Expresses precision in a proportion of the mistake. As this figure is a measure, it is simpler than other numbers to comprehend.

The Equation is:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \quad (3)$$

Where At is the current value and Ft is the expected value. The gap between At and Ft is again split into At. At each predicted point in time, the absolute value in this equation is calculated and divided by n. It is a percentage error by multiplying by 100 percent.

b) Mean Absolute Deviation (MAD)

The precision of the data is represented in the same units that help to conceptualize the amount of error. Outliers have less effect on MAD than on MSD. The equation is that of:

$$MAD = \frac{\sum_{t=1}^{n}|A_t - F_t|}{n} \quad (4)$$

c) Mean Squared Error (MSE)

A precision estimation of the time series values that are widely used. The impact of outliers on MSE is more significant than on MAD. The same is true:

$$MSE = \frac{\sum_{t=1}^{n}(A_t - F_t)^2}{n} \quad (5)$$

d) Root Mean Squared Error (RMSE)

A root-mean-square error (RMSE) is a common measure of the differences between values, predicted by the model or an estimator, and observed values (sample or population values). The RMSE represents the square root of the second sample moment of the variations in the quadratic mean of expected values and observed values. These deviations are called residue when calculations take place through the data sample used to estimate and when measured out-of-sample errors are named. The RMSE adds the magnitude of the error in forecasts to a single predictive power factor for different periods. RMSE is an accuracy metric for comparing predictive errors of various models in a given dataset, not between data sets since they are scale-driven.

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(A_t - F_t)^2}{n}} \quad (6)$$

IV. Results and Discussion

The prediction performance concentrates primarily on precision, accuracy in each class and uncertainty matrix, which indicates how many predictions Root Mean Square Error, Mean Square Error, Absolute error and the average error are measured in the production in Table 4. This test helps to see if the forecast is incorrect on average.

V. Conclusion

The result is seeks in the form of comparison between four machine learning approaches used in B2B sales in 2016–2017 and in 2018 data to determine the type of prediction analysis that is reliable and accurate. We evaluate MSE = 122,883,547,626 and822,000,000,000,000 MAPE = 111.64 before we do the research using a gradient boosted tree. The only way to compare mean squared error (MSS), mean absolute percent error (MAPE), since MSS is based on the scale, is to compare accuracy across time series with the same scales. Specific errors squared. MAPE is often chosen since the variation in percentage between the actual data is the only variation.

In GBT, we have introduced Gradient Tree in this study on the algorithm of the decision tree. At the beginning to the end of the MSE and MAPE data analysis, the results of the graph have good results than the other form, MSE equal to 24,743,000,000.00, and MAPE equal to 0.18.

TABLE IV. Prediction Value

	Manual	**GBT**
MAD	289,297,865,468.19	880,640,000,000,000,000,000.00
MSE	122,883,547,626,822,000,000,000.00	24,743,000,000.00
RMSE	350,547,496,962.71	157,299.08
MAPE	111.64	0.18

REFERENCES

[1] S. H. Sastry and M. S. P. Babu, "Analysis & Prediction of Sales Data in SAP-ERP System using Clustering Algorithms," International Journal of Computational Science and Information Technology (IJCSITY), vol. 1, no. 4, November 2013. DOI: 10.5121/ijcsity.2013.1407

[2] V. Shrivastava and N. Arya, "A study of various clustering algorithms on retail sales data," International Journal of Computer Computation Network, vol. 1, no. 2, 2012.

[3] S. Rajagopal, "Customer data clustering using data mining technique," International Journal of Database Management System, vol. 3, no. 4, pp. 1-11, November 2011. DOI: 10.5121/ijdms.2011.3401

[4] A. Saxena et al., "A review of clustering techniques and developments," Neurocomputing, vol. 267, pp. 664-681, December 2017. DOI: 10.1016/j.neucom.2017.053

[5] N. Shah, M. Solanki, A. Tambe, and D. Dhangar, "Sales Prediction Using Effective Mining Techniques," International Journal of Computer Sciences and Information Technologies, vol. 6, no. 3, pp. 2287-2289, 2015

[6] B. D'Arcy, C. Gallagher, and M. G. Madden, "A Bayesian Classification Approach to Improving Performance for a Real-World Sales Forecasting Application," 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), Miami, US, 2015, pp. 475-480. DOI: 10.1109/ICMLA.2015.150

[7] P.V. Aleksandrovich, K.I. Leopoldovich, and P.A. Viktorovich, "Predicting Sales Prices of the Houses Using Regression Methods of Machine Learning," 2018 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC), Vladivostok, Rusia, 2018, pp. 1-5. DOI: 10.1109/RPC.2018.8482191

[8] W. Ching-She, P. Patil, and S. Gunaseelan, "Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data," 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, 2018, pp. 16-20. DOI: 10.1109/ICSESS.2018.8663760

[9] S. Cheriyan, S. Ibrahim, J. Mohanan, and S. Treesa, "Intelligent Sales Prediction Using Machine Learning Techniques," 2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE), Shouthend, UK, 2018, pp. 53-58. DOI: 10.1109/iCCECOME.2018.8659115

[10] I. Ullah, B, Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim, "A Churn Prediction Model using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector,", IEEE Access, vol. 7, pp. 60134-60149, May 2019. DOI: 10.1109/ACCESS.2019.2914999

[11] C. Lazăr, and M. Lazăr, "Using the Method of Decision Trees in the Forecasting Activity," Petroleum-Gas University of Ploiesti Bulletin, Technical Series, vol. 67, no. 1, pp. 41-48, 2015

[12] B. Flesch, R. Vatrapu, R. Mukkamala, and A. Hussain, "Social set visualizer: A set theoretical approach to big social data analytics of real-world events", 2015 IEEE International Conference on Big Data (Big Data), Santa Clara, CA, 2015, pp. 2418-2427. DOI: 10.1109/BigData.2015.7364036

[13] K. Asooja, G. Bordea, G. Vulcu, and P. Buitelaar, "Forecasting Emerging Trends from Scientific Literature". In LREC, pp. 417-420, 2016

[14] E. Alpaydin," Introduction to Machine Learning (Adaptive Computation and Machine Learning)", The MIT Press, 2004

[15] B. Stearns, F. Rangel, F. Rangel, F. de Faria, J. Oliveira, and A.A.D. Ramos, "Scholar Performance Prediction using Boosted Regression Trees Techniques," In European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), Belgium, 2017, vol. 25, pp. 329-334

[16] F. Zhou, Q. Zhang, D. Sornette and L. Jiang," Cascading logistics regression onto gradient boosted decision tree for forecasting and trading stock indices," Applied Soft Computing, vol. 84, November 2019. DOI: 10.1016/j.asoc.2019.105747

[17] T. D. Rey, C. Wells, and J. Kauhl, "Using data mining in forecasting problems." In SAS Global Forum 2013: Data Mining and Text Analytics, 2013, pp. 1-17

[18] W. Huang, Q. Zhang, W. Xu, H. Fu, M. Wang, and X. Liang, "A Novel Trigger Model for Sales Prediction with Data Mining Techniques," Data Science Journal, vol. 14, pp. 15, 2015. DOI: 10.5334/dsk-2015-015

[19] A. Holzinger, "Introduction to Machine Learning and Knowledge Extraction (MAKE)," Machine Learning and Knowledge Extraction, vol. 1, pp. 1-20, 2019. DOI: 10.3390/make1010001

[20] C. Preda and G. Saporta, "PLS approach for cluster wise linear regression on functional data," In Classification, Clustering, and Data Mining Applications, Springer, Berlin, Heidelberg, 2004