Jurnal Teknik Informatika (JUTIF)
Vol. 5, No. 2, April 2024, pp. 367-376

DOI: https://doi.org/10.52436/1.jutif.2024.5.2.1487
p-ISSN: 2723-3863
e-ISSN: 2723-3871

OPTIMIZING RAW MATERIAL INVENTORY MANAGEMENT OF MSME PRODUCT USING EXTREME GRADIENT BOOSTING (XGBOOST) REGRESSOR ALGORITHM: A SALES PREDICTION APPROACH

Muhammad Khusni Fikri[1], Farrikh Al Zami[*2], Ika Novita Dewi[3], Abu Salam[4], Ifan Rizqa[5], Mila Sartika[6], Diana Aqmala[7]

[1,2,3]Information System, Faculty of Computer Science, Universitas Dian Nuswantoro, Indonesia
[4,5]Information Engineering, Faculty of Computer Science, Universitas Dian Nuswantoro, Indonesia
[6,7]Faculty of Economics and Business, Universitas Dian Nuswantoro, Indonesia
Email: [1]fikrifizy@gmail.com, [2]alzami@dsn.dinus.ac.id

(Article received: October 24, 2023; Revision: November 17, 2023; published: April 04, 2024)

Abstract

Micro, Small and Medium Enterprises or MSMEs have a very important role for the survival of the economic sector in Indonesia. However, as the development of MSMEs, followed by a series of problems that arise. One of them is the problem of sales, business people have difficulty in determining the number of product sales in the future so that there is often an accumulation of raw materials or unsold products. This study aims to help MSMEs optimize raw material management by predicting product sales using the XGBoost Regressor Algorithm. Recently, the algorithm is very famous in the competition because of its reliability and no one has applied it to predict MSME product sales. Based on several other studies, this algorithm is accurate in predicting a value, such as predicting stock prices and the number of accidents in Bali, Indonesia. This research uses historical product sales data and weather data consisting of air temperature and relative humidity in Semarang Indonesia to train and evaluate the performance of the model. The prediction model was performed with predetermined variables and resulted in MAE 3.0752730568649156, MSE 38.25842541629838, and RMSE 6.185339555456788. In the end, it is concluded that the model built with XGBoost Regressor has a low error rate so that it can accurately predict the sales of an MSME product and optimize the management of raw materials for related products.

Keywords: *data science, MSMEs, predictive analysis, sales prediction, xgboost regressor.*

1. INTRODUCTION

Micro, Small, and Medium Enterprises or commonly called MSMEs are an important sector in Indonesia's economic survival [1]. MSMEs are businesses with relatively small operations, resources, and number of workers [2]. MSMEs have contributed more in contributing to employment and economic growth. Based on information provided by the Ministry of Cooperatives and SMEs, MSMEs contribute around 61% to the country's Gross Domestic Product (GDP) and provide jobs to nearly 97% of the overall labor force in Indonesia [3]. MSMEs can be found in various fields, such as the food and beverage industry, the fashion world, handicrafts, and various other services [4]. One example of MSME products in the food sector is Soto Betawi by Rumah Makan Kota Semarang, which will be the object of this study.

The progress of MSME development in Indonesia is in line with the constraints faced by business actors. The constraints vary depending on the sector and location. For MSMEs that want to grow their business, they need to be able to increase sales and expand their market share. This can be a challenging action, especially for MSMEs that have limitations in terms of managerial knowledge and skills [5]. One of the main things that constrains the entire MSME sector is the sales side, where business people often experience losses with falling profits due to the accumulation of unsold goods. This problem is related to the planning and management of raw materials for products to be sold [6]. Based on these problems, business people need to determine the right stock of raw materials and products so that there are no shortages and excesses that will have an impact on sales. Therefore, to achieve maximum profit, business people need to optimize the inventory of raw materials or products to be sold [7].

With the advancement of science and technology, these problems can be solved with less time and practice. For problems regarding the planning of raw materials and products sold, this can be solved by one branch of the data science discipline, namely Data Science [8]. Data Science is a discipline related to collecting, analyzing, and processing data to produce useful information [9]. Data Science components consist of descriptive, predictive, and prescriptive analysis [10]. Each component has its own role, but in this case the most instrumental in determining the amount of raw materials or products to be sold later with the help of predictive analysis.

Predictive analysis is an analytical technique used to predict a future value based on historical data [11]. The predictive analysis method used is the Extreme Gradient Boosting (XGBoost) Regressor Algorithm. XGBoost combines the advantages of gradient boosting and random forests, making it a powerful algorithm for predictive modeling [12]. The method was chosen because the algorithm produces good models in terms of regression and classification based on several studies [13], [14]. Moreover, XGBoost has proven its ability to provide stable and accurate predictions compared to other tree-based ensemble algorithms [13]. Currently, XGBoost has been implemented in various sectors, including finance and retail. It has been proven to be accurate in predicting BCA Bank's stock price and sales volume in the retail industry [15], [16]. In this study, the XGBoost algorithm is expected to accurately predict the sales of Soto Betawi MSME products so that business people can be helped in preparing and optimizing raw materials or products to achieve optimal sales which will have an impact on maximum profits.

Based on study, for now there are still few studies that apply the XGBoost Regressor Algorithm to MSME data. In addition, there is no study that applies the XGBoost Regressor Algorithm to predict the sales of an MSME product, especially Soto Betawi products.

Based on the problems previously described, it is necessary to create a model to predict the sales of MSME products focusing on Soto Betawi using the XGBoost Regressor Algorithm to assist business people in optimizing the stock of raw materials and Soto Betawi products themselves. Where this has a direct impact on the level of sales success and the increase in profits earned by business people.

2. RESEARCH METHODOLOGY

2.1. Research Steps

The research stages used cover all aspects of the steps in implementing data science in a case. Problem understanding is obtained from discussions and literature studies on the problem of selling a product. In this study, the supporting weather data of average temperature and average humidity of Semarang City is used for the features used in the regression process. Starting with the understanding of the problem, followed by the problem solving process with data science steps including data collection from various sources, understanding the content and characteristics of the data, and data pre-processing. Data pre-processing includes feature reengineering, data cleaning, and merging sales data with Semarang City weather.

An overview of the research stages used can be seen through the flowchart below.

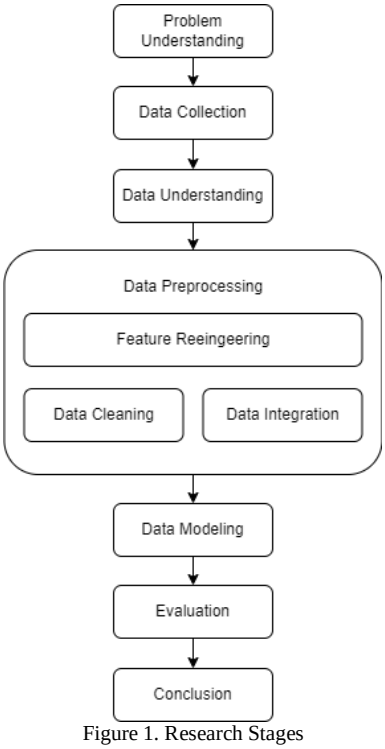

Figure 1. Research Stages

Afterwards, the process continued with data modeling using XGBoost Regressor, evaluating model performance using MSE (Mean Squared Error), MAE (Mean Absolute Error), and RMSE (Root Mean Squared Error), and drawing conclusions from the entire research activity.

2.2. Data Science

Data Science is a scientific discipline related to the collection, processing, and analysis of data to produce useful information [11]. Currently, data science has been used in various sectors both from the government, education, and business sectors, especially MSMEs [17]. There are several types of data analysis, but broadly speaking, they can be divided into three parts. Among them are descriptive, predictive, and prescriptive analysis [11]. The following is the explanation.

a) Descriptive Analysis is an analysis technique that describes the current condition of data. This analysis is used to determine the condition of sales and other things about MSMEs in a certain time span.
b) Predictive Analysis is an analytical technique used to predict a future value. This analysis is used to estimate the amount of demand for MSME goods so that the stock of goods provided can meet demand and nothing is wasted.
c) Prescriptive Analysis is an analytical technique used to provide recommendations or suggestions for certain actions based on data. This analysis is used in determining the optimal price and distribution of goods so as to achieve maximum profit.

2.3. Multiple Linear Regression

Multiple Linear Regression is a statistical analysis method used to determine the relationship between one dependent variable and more than one independent variable.

The purpose of this analysis is to determine the direction of the relationship between variables, whether each independent variable is positively or negatively related, and to predict the value of the independent variable to increase or decrease [18][19].

$$Y = a + b_1X_1 + b_2X_2 + \cdots + b_nX_n \quad (1)$$

In forecasting the value of Y, it is necessary to know all the variables that will be used. The equation above is a multiple linear regression equation used in forecasting. Here the following explanation equation above (1),

Y = Dependent variable or value that is predicted
X = Independent variable
a = Constant (the value of Y if X1, X2, ..., Xn = 0)
b = Regression coefficient or increase value or decrease.

2.4. eXtreme Gradient Boosting (XGBoost)

eXtreme Gradient Boosting or XGBoost is an ensemble learning algorithm used for machine learning, especially for data regression and classification problems [12]. XGBoost was chosen because it has been shown to produce stable and accurate predictions compared to other tree-based ensemble algorithms [13].

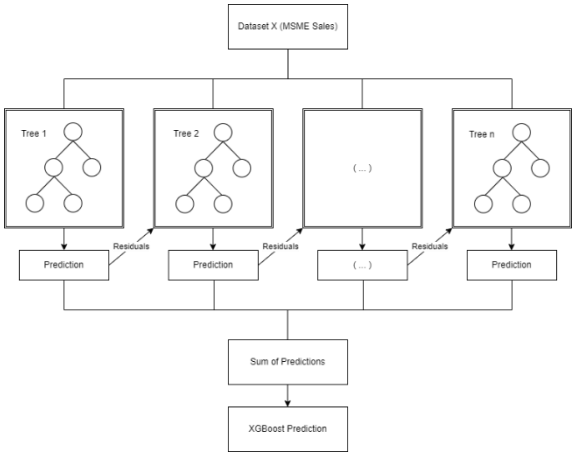

Figure 2. XGBoost Architecture

It is well known in competitive events for its speed, scalability, and performance in handling data sets [12], [20]. XGBoost combines the strengths of gradient boosting and random forests, making it a powerful algorithm for predictive modeling [12]. It uses the concept of gradient boosting by repeatedly training a weak prediction model like a decision tree, then combining the previous predictions to produce a final prediction model [13], [12]. In addition, this algorithm has been proven to produce stable and accurate predictions compared to other tree-based ensemble algorithms.

Currently, XGBoost has been implemented in various sectors, be it finance, health, and so on. However, there is still no specific implementation of XGBoost on MSME sales data. For example, in this study, the algorithm is implemented to predict the number of sales of MSME products based on existing historical data.

The selection of parameters or features in the XGBoost algorithm is the same as selecting variables in linear regression or multiple linear regression algorithms as described in the previous section. The features used can be values in the dataset or outside the dataset. It is necessary to select features that have a high correlation value with the value to be predicted so that the model created has an accurate performance in predicting a value.

2.5. Model Performance Evaluation Metrics

In evaluating model performance, the Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) metrics are used. Some of these metrics are metrics that are suitable in various data sets, easy to understand, and most commonly used in evaluating the performance of a regression model, in this case the XGBoost Regressor [21].

The following is an explanation and mathematical form of these metrics.

a) Mean Absolute Error (MAE) is the average of the absolute difference between the actual value in the dataset (y_i) and the predicted value produced by the model $(\hat{y}_i)$. MAE is more interpretable than MSE [22], as it directly calculates the average error in a set of predictions as mention in equation (2) without looking at the direction of the resulting error values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (2)$$

b) Mean Squared Error (MSE) is the average of the squared difference between the actual value in the dataset (y_i) and the predicted value produced by the model $(\hat{y}_i)$. Mathematically, it is almost the same as MAE as mentioned in equation (3), except that the difference is squared.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (3)$$

c) Root Mean Squared Error (RMSE) is the root of the average of the squared difference between the actual values in the dataset (y_i) with the predicted value produced by the model $(\hat{y}_i)$ as mention in equation (4) . The purpose of this metric is to measure the standard deviation of the error.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (4)$$

3. RESULTS AND DISCUSSION

3.1. Problem Understanding

Problems are obtained from some literature and complaints directly from business people about the difficulty of determining the amount of stock of raw materials or products to be sold. From some literature, it is explained that it is difficult to determine which physical products such as property and FMCG are in demand and purchased in the future [23], [24].

As with the problems that will be raised in this study, Semarang City Restaurant needs to make stock planning of raw materials and products with the minimum possible error to increase sales which have an impact on preventing losses and increasing profits generated. This study will focus on the Soto Betawi product in the Semarang City Restaurant. This is because it is the product with the highest sales for six consecutive months. Therefore, assuming the highest profit-contributing potential, the product needs to be optimized by building a model that can predict future demand and sales so that the company's success can be achieved or improved.

In the problem solving process using the XGboost Regressor Algorithm, an independent variable of Semarang City weather which includes Average Air Temperature and Average Humidity is used to determine the number of sales. Some literature reveals that weather has a significant impact on customer buying interest in retail businesses [25], [26]. In this case, we can make a foundation to try to apply it to restaurant sales. The purpose of this study is to predict the number of sales based on the visit mode (dine in, online, and take away) using weather data for Semarang City.

3.2. Data Collection

The MSME product sales data used was obtained from one of the relevant business actors who donated the data for research purposes. The data contains sales of food and beverage products in Semarang City restaurants for six months from January 2023 to June 2023.

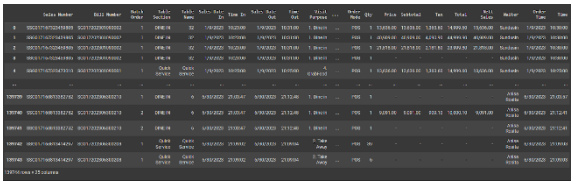

Figure 3. Sales Data View

This data contains ± 139 thousand records with ± 25 features divided by months. Before being received, this data was processed by a data engineer from the research team using the Elasticsearch Logstash Kibana (ELK Stack) tool so that it was ready to be processed in CSV format.

	Date	Temperature_Avg (°C)	Relative Humidity_Avg (%)
0	2023-01-09	28	84
1	2023-01-10	27,9	85
2	2023-01-11	28,3	85
3	2023-01-12	28,7	80
4	2023-01-13	27,4	86
...	...	...	...
168	2023-06-26	27,3	77
169	2023-06-27	28,6	79
170	2023-06-28	28,3	78
171	2023-06-29	28,9	80
172	2023-06-30	28,9	77

173 rows × 3 columns

Figure 4. Weather Data View

In addition, weather data for Semarang City was obtained from the Indonesian Meteorology, Climatology and Geophysics Agency (BMKG) website. Weather data includes Average Temperature (°C) and Average Humidity (%) per day for six months from January 2023 to June 2023. Because sales data and weather data are different sources, it is necessary to combine the two data in the Data Integration section.

3.3. Data Understanding

Dataset understanding is done after the existing data can be accessed properly. In the previous process, data can be accessed in several tools such as Spreadsheet and Python on Google Colab. In the data understanding process, feature identification, feature selection, and Exploratory Data Analysis (EDA) can be carried out on the features to be used.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 139744 entries, 0 to 139743
Data columns (total 25 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   Sales Number         139744 non-null  object
 1   Bill Number          139744 non-null  object
 2   Batch Order          139744 non-null  int64
 3   Table Section        139744 non-null  object
 4   Table Name           139744 non-null  object
 5   Sales Date In        139744 non-null  object
 6   Time In              139744 non-null  object
 7   Sales Date Out       139744 non-null  object
 8   Time Out             139744 non-null  object
 9   Visit Purpose        139744 non-null  object
 10  Payment Method       139744 non-null  object
 11  Menu Category        139744 non-null  object
 12  Menu Category Detail 139744 non-null  object
 13  Menu                 139744 non-null  object
 14  Menu Code            139744 non-null  int64
 15  Order Mode           139744 non-null  object
 16  Qty                  139744 non-null  int64
 17  Price                139744 non-null  object
 18  Subtotal             139744 non-null  object
 19  Tax                  139744 non-null  object
 20  Total                139744 non-null  object
 21  Nett Sales           139744 non-null  object
 22  Waiter               139744 non-null  object
 23  Order Time           139744 non-null  object
 24  Time                 139744 non-null  object
dtypes: int64(3), object(22)
memory usage: 26.7+ MB
```

Figure 5. Sales Data Information

First, the feature identification process, in the picture below it can be seen that there are many features available in the dataset. Starting from sales number, sales date out, visit purpose, payment method, menu, and so on. However, this study will not use all existing features and select some features that can be used to predict sales. The target feature to be predicted is Qty (number of products sold) based on Visit Purpose (dine in, online, and take away). Therefore, the features that will be used in building the model include Sales Date Out (transaction date), Visit Purpose (visit mode), and Qty (number of products sold).

	Date	Mode	Payment Method	Sales Quantity	Price
0	1/9/2023	Online	Online	1	43.636
1	1/9/2023	Dine in	EDC	4	40.909
2	1/9/2023	Dine in	Cash	1	40.909
3	1/9/2023	Dine in	QRIS	1	40.909
4	1/9/2023	Dine in	Cash	1	40.909
...	...	...	...	...	...
3033	6/30/2023	Dine in	Cash	1	40.909
3034	6/30/2023	Dine in	EDC	2	40.909
3035	6/30/2023	Dine in	EDC	1	40.909
3036	6/30/2023	Dine in	EDC	2	40.909
3037	6/30/2023	Dine in	EDC	1	40.909

3038 rows × 5 columns

Figure 6. Sales Data (Feature Selected)

The table above is a form of data that has been sorted by features and content. In terms of features, the table above already contains features that have been planned previously. In terms of content, the table above only contains transaction data related to Soto Betawi. The process of sorting features and content is done with Spreadsheet so that it cannot be documented clearly.

The two figures below show the monthly and daily sales trends. From the first image, we can conclude that monthly sales trended upward from the beginning of the year until it peaked in April with the most sales.

Figure 7. Monthly Sales Trend

Figure 8. Daily Sales Trend

However, it decreased in the following month and was offset by an increase in sales in June. As for the second figure, daily sales tend to fluctuate and are difficult to read. But on average, daily sales tend to rise in the middle of the month from the 15th to the 20th of each month.

3.4. Data Preprocessing

Before entering the modeling stage, the data must pass the final stage of data pre-processing which includes feature reengineering, data integration, and data cleansing.

3.4.1. Feature Reengineering

Actually, we have done the feature selection before, the results can be seen in Figure 6. However, it is necessary to create a new feature from the Mode record so that the sales amount of each mode can be seen and predicted clearly.

Therefore, feature reengineering is carried out to create a new feature of Mode so that the number of sales from each purchase mode can be seen, be it dine in, online, or take away. The process can be done with the python code listed in the image below.

```
df_sum = data.groupby([pd.Grouper(key='Date'),
                       pd.Grouper(key='Mode')])['Sales Quantity'].sum().reset_index()

data_sum_daily = pd.DataFrame(df_sum)

data_transpose = data_sum_daily.pivot(index='Date', columns= 'Mode', values='Sales Quantity')
```

Figure 9. Feature Reengineering Python Code

In the first step, the data in Figure 9 groups the total amount of the Sales Quantity column based on the Date and Mode columns and converts it into a dataframe called data_sum_daily. After that, a pivot table is made on the data_sum_daily dataframe with the parameters of the Date index, Mode column, and Sales Quantity value.

The Figure 10 below is a table of results from feature reengineering. Data contains the Date and all mode (Dine in, Online, and Take Away). The data in the table will be combined with the weather data of Semarang City that has been prepared previously to enter the data modeling stage.

Date	Mode Dine in	Online	Take Away
2023-01-09	14.0	1.0	NaN
2023-01-10	18.0	2.0	NaN
2023-01-11	16.0	NaN	NaN
2023-01-12	12.0	NaN	NaN
2023-01-13	19.0	NaN	NaN
...	...	...	...
2023-06-26	20.0	NaN	NaN
2023-06-27	17.0	NaN	NaN
2023-06-28	23.0	NaN	NaN
2023-06-29	26.0	NaN	NaN
2023-06-30	30.0	2.0	NaN

171 rows × 3 columns

Figure 10. Sales Data (Feature Reengineered).

3.4.2. Data Integration

Data integration is done by combining sales data that has been created in the previous stage with weather data for Semarang City. The weather data to be combined adjusts the sales data based on the date of each dataset. Both data start from January 9, 2023 to June 30, 2023. The command below is the command to merge data in Pyhton.

```
data_final = data_transpose.merge(weather, on='Date', how='inner')
```

Figure 11. Data Integration Python Code

The meaning of the above command is that the previously prepared sales data (data_transpose) is combined with the Semarang City weather data (weather) based on the date (Date) by melting. The operation produces a new dataset named data_final.

	Date	Dine in	Online	Take Away	Temperature_Avg (°C)	Relative Humidity_Avg (%)
0	2023-01-09	14.0	1.0	NaN	28.0	84
1	2023-01-10	18.0	2.0	NaN	27.9	85
2	2023-01-11	16.0	NaN	NaN	28.3	85
3	2023-01-12	12.0	NaN	NaN	28.7	80
4	2023-01-13	19.0	NaN	NaN	27.4	86
...	...	...	...	...	...	...
166	2023-06-26	20.0	NaN	NaN	27.3	77
167	2023-06-27	17.0	NaN	NaN	28.6	79
168	2023-06-28	23.0	NaN	NaN	28.3	78
169	2023-06-29	26.0	NaN	NaN	28.9	80
170	2023-06-30	30.0	2.0	NaN	28.9	77

171 rows × 6 columns

Figure 12. Data Integrated (Sales & Weather)

Figure 12 is the result of the data integration process between sales data and weather data for Semarang City. The data still looks like it has a lot of empty values so it still needs to be cleaned in the next stage.

3.4.3. Data Cleansing

In the last stage, the merged data needs to be cleaned by filling in the value 0 in some empty records. This is important because regression calculations are numeric, a numeric value is needed in order to produce a predicted value. This can be done with the following command.

```
data_final.fillna(0, inplace=True)
```

Figure 13. Fill Missing Value Python

The meaning of the above command is that all records that are detected empty (NaN) from the previously merged data (data_final) are filled automatically with the value 0 on the record directly.

	Date	Dine in	Online	Take Away	Temperature_Avg (°C)	Relative Humidity_Avg (%)
0	2023-01-09	14.0	1.0	0.0	28.0	84
1	2023-01-10	18.0	2.0	0.0	27.9	85
2	2023-01-11	16.0	0.0	0.0	28.3	85
3	2023-01-12	12.0	0.0	0.0	28.7	80
4	2023-01-13	19.0	0.0	0.0	27.4	86
...	...	...	...	...	...	...
166	2023-06-26	20.0	0.0	0.0	27.3	77
167	2023-06-27	17.0	0.0	0.0	28.6	79
168	2023-06-28	23.0	0.0	0.0	28.3	78
169	2023-06-29	26.0	0.0	0.0	28.9	80
170	2023-06-30	30.0	2.0	0.0	28.9	77

171 rows × 6 columns

Figure 14. Final Data

Figure 14 above is the result of the data cleaning operation by filling in blank values. After this, the data is ready to be entered into the data modeling process.

3.5. Data Modeling

3.5.1. Descriptive Analysis

Before going into predictive analysis, descriptive analysis needs to be done to see a clear picture of the data. Here, it is done to see the correlation between tables to ensure that the features we will use (air temperature and humidity features) actually affect the target values (dine in, online, and take away sales).

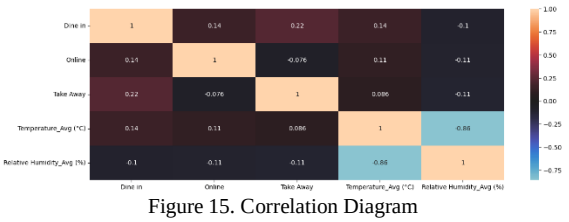

Figure 15. Correlation Diagram

From the correlation diagram above, it can be concluded that the air temperature feature (Temperature_Avg) has a positive correlation with the sales feature (Dine in, Online, and Take Away), although the correlation is relatively weak. On the other hand, the air humidity feature (Relative Humidity_Avg) has a negative correlation with the sales features (Dine in, Online, and Take Away) and

is classified as less strong. However, both features are still used in the model to predict sales because the selected model belongs to a regression algorithm that can still operate with positive-negative and strong-weak features.

3.5.2. Predictive Analysis

As previously described, the process of predicting sales value (Dine in, Online, and Take Away) will be done by inputting the average air temperature (Temperature_Avg) and average humidity (Relative Humidity_Avg) features. In the first step, the XGBoost library will be imported, this library will be used to develop a model to predict sales value.

In addition, in this process, split mode will also be used, which divides the dataset into two parts (training dataset and test dataset). The training dataset is set at 80%, which is used to train the XGBoost model in performing regression operations. While the test dataset is set in a portion of 20%, where this dataset is used to test the model that has been done the training process.

```
import xgboost as xgb
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(data_final[['Temperature_Avg (°C)','Relative Humidity_Avg (%)']],
                                                    data_final[['Dine in','Online','Take Away']], test_size=0.2, random_state=42)


model = xgb.XGBRegressor(
    max_depth=3,
    n_estimators=100,
    learning_rate=0.1,
    objective='reg:squarederror'
)

model.fit(X_train, y_train)
```

Figure 16. Data Modeling XGBoost Python

In this model, it is set for the maximum depth of the decision tree in the XGBoost ensemble is 3 branches and the number of trees estimated is 100 trees. In addition, the learning rate is the learning rate that controls the extent to which the model learns from the data at each iteration. This value is 0.1, which means the model will update the estimation with 10% of the remaining error at each iteration. Once the model is trained, it is time to test the model.

```
predictions = model.predict(X_test)
```

Figure 17. Predict Command Python

```
[ 3.20230484e+01,  4.34270650e-01,  7.10781753e-01],
[ 1.32934980e+01,  1.95599735e-01,  1.54000089e-01],
[ 2.52434578e+01,  4.87159669e-01,  1.02233678e-01],
[ 1.28537140e+01,  2.58969992e-01,  8.50283921e-01],
[ 3.01722622e+01,  5.79376459e-01,  7.55799711e-01],
[ 2.02107487e+01,  1.95599735e-01,  1.54000089e-01],
[ 1.75685253e+01,  8.31402838e-01,  7.01312184e-01],
[ 2.60853863e+01,  2.69575715e-01,  1.22399487e-01],
[ 1.34740772e+01,  3.89836609e-01,  2.33483449e-01],
[ 1.53508749e+01,  2.12604284e-01,  4.33490813e-01],
[ 2.20806732e+01,  3.76040787e-01,  4.62009341e-01],
[ 2.95404797e+01,  5.09912252e-01,  4.18293446e-01],
[ 8.82058811e+00,  8.38408470e-02,  4.98142727e-02],
[ 1.85413666e+01,  6.39134586e-01,  7.12331057e-01],
```

Figure 18. Predicted Value

After testing, the prediction results appear according to the number of target features and the portion of the test dataset that has been set before. This shows that the trained model can predict the value well.

3.6. Evaluation

As previously described, the model performance evaluation is done using Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). Data Modeling shows that the XGBoost Regressor model can predict the number of sales. Therefore, it needs to be evaluated with some of the above metrics with the following commands.

```
from sklearn.metrics import mean_squared_error
from sklearn.metrics import mean_absolute_error

mse = mean_squared_error(y_test, predictions)
mae = mean_absolute_error(y_test, predictions)
rmse = np.sqrt(mean_squared_error(y_test, predictions))
```

Figure 19. Metrics Command

Before that, it is necessary to import the metrics in the sklearn.metrics module for mean_squared_error and mean_absolute_error before starting to perform the operation. However, since RMSE is the root result of MSE, the RMSE operation only uses the numpy math operation with np.sqrt.

```
Mean Squared Error: 38.25842541629838
Mean Absolute Error: 3.0752730568649156
Root Mean Squared Error: 6.185339555456788
```

Figure 20. Metrics Results

The results of the three metric operations can be seen in Figure 20 above, or more clearly in Table 1 below.

Table 1. Metrics Results

Metrics	Rate
Mean Squared Error (MSE)	38.25842541629838
Mean Absolute Error (MAE)	3.0752730568649156
Root Mean Squared Error (RMSE)	6.185339555456788

Based on the results of the metric calculation, the model built using the XGBoost Regressor Algorithm can predict the number of product sales well. The metric calculation mentions a small prediction error rate. The smaller the value of the three metrics, the better the resulting model [21]. Therefore, the model can be used to predict the number of sales in the future.

With this model, it is hoped that it can assist businesses in improving production cost efficiency in preparing optimal product raw materials according to the number of predicted sales for a certain period. This can certainly reduce the risk of losses due to unsold raw materials or products and increase the optimal profit of product sales.

4. DISCUSSION

Predicting sales is very important to be done by business people for the MSME scale, not only for the

large business scale. This is because MSMEs are businesses at the development level towards large. In this development stage, it is necessary to make resource efficiency by optimizing product raw materials through sales or demand predictions. In the end, business people can predict the amount of demand or sales in the future. With this, businesses can increase profits and reduce the risk of loss.

In previous study, many have made predictive models for MSME product sales. However, it has not been found that uses XGBoost Regressor to do so, mostly using other models such as Linear Regression, K-Means, and so on. Although models other than XGBoost can produce models well, XGBoost is very popular and excels in competitions. Therefore, it is necessary to try the algorithm in the business sector, namely the sale of MSME products.

Applied to the sales data of Soto Betawi MSMEs, which has hundreds of rows and several features, XGBoost Regressor proved to be able to make a good prediction model as well. With MAE, MSE, and RMSE scores that are relatively small, proving that the model has a minimal error rate. This also proves that the model built is good and can be used to predict sales and optimize raw material management of related products.

5. CONCLUSION

This study uses the XGBoost Regressor Algorithm to predict the number of sales of Soto Betawi MSME products in Semarang City Restaurants. The product was chosen because of the high number of sales among other products for 6 consecutive months. The data used is Soto Betawi sales data and weather consisting of Average Temperature and Average Air Humidity from January 09, 2023 to June 30, 2023. This study uses Average Temperature and Average Humidity to predict the amount of demand or sales of related products in the future.

The model scores MAE 3.0752730568649156, MSE 38.25842541629838, and RMSE 6.185339555456788. From these scores, it shows that the model error rate in predicting a value is small. This proves that the performance of the XGBoost Regressor model is successful in conducting predictive analysis.

Thus, with good metric evaluation results, it can be concluded that: 1) the XGBoost Regressor model can be used to predict the number of sales and optimize raw material management for future Soto Betawi MSME products in Semarang City Restaurants. 2) weather attributes such as daily temperature and humidity can be used together with daily sales to predict next day sales.

For future research, it is highly recommended to build models using additional features such as changes in raw material prices, other weather aspects, inflation rates, and so on to predict the number of sales of related products. With these additional features, it is expected to increase the accuracy of the XGBoost model in conducting predictive analysis.

ACKNOWLEDGEMENT

We sincerely express our gratitude to the Directorate General of Higher Education, Research, and Technology, Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia for funding a portion of this project through the Kedaireka Program. This work has also received support from Dian Nuswantoro University (UDINUS) through the Center of Excellence in Science and Technology, UDINUS, and the Association of Food and Beverage Souvenir Entrepreneurs in Central Java. This is related to the grant contract document entitled "Implementation of Supply Chain Management System for MSMEs in the Food and Beverage Souvenir Sector, A Case Study in Central Java" with contract number 4501/F.9.02/UDN-01/IV/2023.

REFERENCES

[1] Universitas Pembangunan Panca Budi, E. Erwansyah, R. Saragih, and T. O. H. Purba, “Pendampingan Pemilihan Merek, Pentingnya Merek dan Nilai Yang Dihasilkan Melalui Merek" Bagi Pelaku Umkm Di Desa Lau Bakeri, Kecamatan Kutalimbaru, Kabupaten Deli Serdang, Sumatera Utara,” *methabdi*, vol. 2, no. 1, pp. 26–31, Jun. 2022, doi: 10.46880/methabdi.Vol2No1.pp26-31.

[2] Dikson Efrando Sidabutar, “Pengaruh Kualitas Sumber Daya Manusia Terhadap Pertumbuhan Usaha Mikro Kecil Menengah (Umkm) Di Kecamatan Rambah,” *cano*, vol. 11, no. 2, pp. 24–34, Dec. 2022, doi: 10.30606/cano.v11i2.1618.

[3] Febri Azka Dzikrullah and Vadilla Mutia Zahara, “Pengaruh Tingkat Jumlah Umkm, Impor, Dan Expor Terhadap Pertumbuhan Ekonomi Di Indonesia Pada Tahun 1997-2019,” *FLURALIS*, pp. 62–68, Jul. 2023, doi: 10.61252/fjeb.v2i2.93.

[4] E. Subiyantoro, A. R. Muslikh, M. Andarwati, G. Swalaganata, and F. Y. Pamuji, “Analisis Pemilihan Media Promosi UMKM untuk Meningkatkan Volume Penjualan Menggunakan Metode Analytical Hierarchy Process (AHP),” *JTMI*, vol. 8, no. 1, pp. 1–8, Jul. 2022, doi: 10.26905/jtmi.v8i1.6760.

[5] B. E. Rokhmah and I. Yahya, “Tantangan, Kendala, Dan Kesiapan Pemasaran Online Umkm Di Desa Nglebak, Kecamatan Tawangmangu, Kabupaten Sukoharjo,” *Final Mazawa*, vol. 1, no. 1, pp. 20–31, Jun. 2022, doi: 10.22515/finalmazawa.v1i1.2363.

[6] J. Jabbar, “Sistem Informasi Stok Barang

Menggunakan Metode Clustering Kmeans (Studi Kasus Rmd Store)," *infotech*, vol. 8, no. 1, pp. 70–75, Apr. 2022, doi: 10.31949/infotech.v8i1.2280.

[7] Harkim Simamora, Rejekia Vaizal Simanungkalit, Maya Andriani, and Bambang Sugiharto, "Socialization of UMKM Merchandise Inventory Management During the Covid 19 Pandemic in Desa Suka Makmur," *gandrung*, vol. 2, no. 2, pp. 266–275, Aug. 2021, doi: 10.36526/gandrung.v2i2.1383.

[8] R. Rios, C. Childs, S. Smith, N. Washburn, and K. Kurtis, "Advancing cement-based materials design through data science approaches," *RILEM* Tech Lett, vol. 6, pp. 140–149, Dec. 2021, doi: 10.21809/rilemtechlett.2021.147.

[9] P. Maslianko and Y. Sielskyi, "Data Science — definition and structural representation," *SRIT*, no. 1, pp. 61–78, Jul. 2021, doi: 10.20535/SRIT.2308-8893.2021.1.05.

[10] A. Asllani, Business analytics with management science models and methods. Upper Saddle River, New Jersey: Pearson Education, 2015.

[11] A. T. Sasongko, "Studi Literatur Konsep dan Implementasi Sains Data untuk Memaksimalkan Kinerja Industri Manufaktur," *JTEKSIS*, vol. 5, no. 2, pp. 90–94, Apr. 2023, doi: 10.47233/jteksis.v5i2.778.

[12] R. Natras, B. Soja, and M. Schmidt, "Ensemble Machine Learning of Random Forest, AdaBoost and XGBoost for Vertical Total Electron Content Forecasting," *Remote Sensing*, vol. 14, no. 15, p. 3547, Jul. 2022, doi: 10.3390/rs14153547.

[13] S. Demir and E. K. Sahin, "An investigation of feature selection methods for soil liquefaction prediction based on tree-based ensemble algorithms using AdaBoost, gradient boosting, and XGBoost," *Neural Comput & Applic*, vol. 35, no. 4, pp. 3173–3190, Feb. 2023, doi: 10.1007/s00521-022-07856-4.

[14] T. Kavzoglu and A. Teke, "Predictive Performances of Ensemble Machine Learning Algorithms in Landslide Susceptibility Mapping Using Random Forest, Extreme Gradient Boosting (XGBoost) and Natural Gradient Boosting (NGBoost)," *Arab J Sci Eng*, vol. 47, no. 6, pp. 7367–7385, Jun. 2022, doi: 10.1007/s13369-022-06560-8.

[15] B. Jange, "Prediksi Harga Saham Bank BCA Menggunakan XGBoost," *ARBITRASE*: *Journal of Economics and Accounting* vol. 3, no. 2, pp. 231-237, 2022, doi : 10.47065/arbitrase.v3i2.495.

[16] L. Zhang, W. Bian, W. Qu, L. Tuo, and Y. Wang, "Time series forecast of sales volume based on XGBoost," *J. Phys.: Conf. Ser.*, vol. 1873, no. 1, p. 012067, Apr. 2021, doi: 10.1088/1742-6596/1873/1/012067.

[17] S. Rautenbach, I. De Kock, and J. Grobler, "DATA SCIENCE FOR SMALL AND MEDIUM-SIZED ENTERPRISES: A STRUCTURED LITERATURE REVIEW," *SAJIE*, vol. 32, no. 3, 2022, doi: 10.7166/33-3-2797.

[18] A. Pamuji and H. S. Setiawan, "LINEAR REGRESSION FOR PREDICTION OF EXCESSIVE PERMISSIONS DATABASE ACCOUNT TRAFFIC", *J. Tek. Inform. (JUTIF)*, vol. 3, no.2, pp. 367–374, Apr. 2022, doi: 10.20884/1.jutif.2022.3.2.206.

[19] S. Adiguno, Y. Syahra, and M. Yetri, "Prediksi Peningkatan Omset Penjualan Menggunakan Metode Regresi Linier Berganda," *j. sist. inf. trig. dhar. JURSI TGD*, vol. 1, no. 4, p. 275, Jul. 2022, doi: 10.53513/jursi.v1i4.5331.

[20] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[21] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Computer Science*, vol. 7, p. e623, Jul. 2021, doi: 10.7717/peerj-cs.623.

[22] S. M. Robeson and C. J. Willmott, "Decomposition of the mean absolute error (MAE) into systematic and unsystematic components," *PLoS ONE*, vol. 18, no. 2, p. e0279774, Feb. 2023, doi: 10.1371/journal.pone.0279774.

[23] G. N. Ayuni and D. Fitrianah, "Penerapan Metode Regresi Linear Untuk Prediksi Penjualan Properti pada PT XYZ".

[24] A. Alfani W.P.R., F. Rozi, and F. Sukmana, "PREDIKSI PENJUALAN PRODUK UNILEVER MENGGUNAKAN METODE K-NEAREST NEIGHBOR," *jipi. jurnal. ilmiah. penelitian. dan. pembelajaran. informatika.*, vol. 6, no. 1, pp. 155–160, Jun. 2021, doi: 10.29100/jipi.v6i1.1910.

[25] A. Abdurrahim and A. Hartono, "Pengaruh Cuaca Terhadap Perilaku Belanja Konsumen Minimarket: Studi Pada Minimarket Indomaret," *Ecombis. Rev. J. Ilm. Eco. and*

Bussines., vol. 10, no. 2, Jul. 2022, doi: 10.37676/ekombis.v10i2.2649.

[26] N. Rose and L. Dolega, "It's the Weather: Quantifying the Impact of Weather on Retail Sales," *Appl. Spatial Analysis*, vol. 15, no. 1, pp. 189–214, Mar. 2022, doi: 10.1007/s12061-021-09397-0.