International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 06 Issue 08 August 2023
DOI: 10.47191/ijcsrr/V6-i8-35, Impact Factor: 6.789
IJCSRR @ 2023

# E-Commerce Product Demand Modelling Using Machine Learning Algorithm Case Study of Rice Trading Products in PT XYZ

**Vinsensia Fresian Meiliana, S.Si[1], Taufik Faturohman, S.T, MBA, Ph.D.[2]**
[1,2] School of Business and Management ITB

**ABSTRACT:** E-commerce XYZ is an Indonesian commerce company that have 3 types of products in its B2C business line: trading, consignment, and marketplace. From January 2021 until October 2022, the company's trading rice category product sales generated a negative profit. Even though for the last several years e-commerce has been focused on growth instead of profitability, the current economic environment is forcing e-commerce companies to focus on profitability as well. For trading products, maximum profit can be achieved in two ways: selling products with a very high margin but with less quantities or selling in large quantities but with a sub-optimal margin. Hence, the company needs to find a demand function model that can be used to generate maximum profit. To find the best model, the researcher first created a baseline model by using median for every product group which is already grouped based on their Unit of Measurement. Next, to find the best model, the researcher will create a demand function using 4 other models. It is found that Gradient Boosted is the best algorithm to model the demand function. Although this model successfully models a demand function for a product category in e-commerce, business context still needs to be added before this model can be implemented in real life as well as finding other features that might affect the demand function.

**KEYWORDS:** Demand Function, E-Commerce, Gradient Boosted, Machine Learning, Rice Product.

## 1 INTRODUCTION

Just as other e-commerce company in Indonesia, PT XYZ also started to focus on generating profit instead of only focusing on growth. PT XYZ has a certain profit target for each of its sales categories or brand, but sometimes the target is not achieved. One of the probable reasons is that there is a gap between the profit target and the actual profit difference between COGS and sales price. COGS (Cost of Goods Sold) amount is decided from the agreement between the brand manager and the supplier. Sales price amount is also decided by the brand manager by analyzing several factors such as competitiveness, product aging, and the product historical performance. Currently, PT XYZ doesn't have any benchmark in terms of the price needed to be applied for each product group, and this might cause the margin generated to be not optimal and the profit target cannot be achieved. For example, during 1 January 2021 – 31 October 2022, rice category trading products record a negative profit up to hundreds of million. Hence, this paper will try to create the best demand model for each product. This model will then be assessed by the campaign team and brand managers to see whether this model is relevant and can be used for future pricing decision making. There are several scope and limitations for this research:

1. This research just focused on PT XYZ case.
2. The model will be used for trading products only.
3. The product category is limited to rice products only.
4. The products that will be included are the products with complete data (It should have sales quantity and price).
5. The outlier products will be excluded.
6. The data that will be used is historical weekly data of sales, price, and COGS from 1 January 2021 until 31 October 2022.
7. Assumption: price is the only factor that affects demand function.
8. The product will be group based on Unit of Measurement.

## 2. LITERATURE REVIEW

### 2.1 E-Commerce

E-commerce is a business in which information technology is used to increase sales, business efficiency and provide a basis for new products and services (Išoraitė and Miniotienė, 2018). E-commerce is formed from two words: that is generated by margin from

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 06 Issue 08 August 2023
DOI: 10.47191/ijcsrr/V6-i8-35, Impact Factor: 6.789
IJCSRR @ 2023

selling trading products. The margin from trading products is acquired from the electronic and commerce, which means utilizing electronic media to carry commerce activity (Whinston et al,1997). There are three types of products that are sold in ecommerce: marketplace, consignment, and trading. A trading product is a product that is stored, sold, and fulfilled by the company. For trading products, the profit is generated from margin or difference between sales price and Cost of Goods Sold (COGS). The maximum margin can be achieved by setting a higher price, but doing so high might cause a drop in sales, and results in lesser profit in general.

### 2.2 Machine Learning

Machine Learning is one of the computer science fields that researches algorithms and methods to automate solutions for complicated problems that are challenging to program using traditional programming techniques. There are 3 types of problems that can be solved by using machine learning: Classification is classifying something into several groups/ categories/ classes. Clustering is organizing a big collection of data points into a few clusters by clustering them so that each cluster contains points with similar characteristics. Prediction is creating models based on historical data and using them for future projection. (Rebala et al., 2019). There are several models that are frequently used by machine learning researchers.

#### 2.2.1 Machine Learning Model

1. Linear Regression is an approach to model a linear relationship between variables (Witten et al. ,2017).
   Linear Regression can be express by

$y = ax + b$

   with b as constant and a as coefficient
2. Random Forest Regressor is a model that consists of decision trees where each tree has their own configuration. Random forest algorithm will generate output by averaging the result of each of its trees. This will result in reduction of overfitting of the trees while keep the prediction capability of the tree. (Witten et al., 2017)
3. SVR or Super Vector Regression is based on linear regression but instead of fitting a line to the data point, this algorithm creates support vectors to create an area around the line that can fit the data points and hence minimize the error. (Witten et al., 2017)
4. Gradient Boosting Regressor is an ensemble of decision trees but instead of creating the trees randomly this algorithm will create new trees based on the learning of previous tree. (Witten et al., 2017)

#### 2.2.2 Data Splitting

The researcher must evaluate the model to see how good it predicts the data but are unable to do so using the data that was used to create the model as it can recall the entire training set and accurately forecast each point. So, to evaluate the model's performance, the researcher divided the collected data into training data and test data. Training data is the data that is inputted to the model to learn. After getting the initial result, the researcher tests the model against the test data. (Mueller and Guido, 2016).

#### 2.2.3 Evaluation

After getting the result from the initial model, the researcher is evaluating the model by looking at the error of the model compared to the test data. There are two commonly used calculations to measure the error which are RMSE or Root Mean Squared Error and MAE or Mean
Absolute Error.

$$root\ mean - squared\ errod \sqrt{\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{n}}$$

$$mean\ absolute\ error \frac{|p_1 - a_1| + \cdots + |p_n - a_n|}{n}$$

*With: p: prediction a: actual n: number of data point*

The main difference between these two calculations is that while MAE will value all errors evenly, RMSE values bigger error more, as the error will be squared. (Witten et al. ,2017).

## 3. RESEARCH METHODOLOGY

### 3.1 Research Design

In order to know the root cause of the problem this research will be started by identifying the problem/business. The next step after knowing the issue is to create a model to solve this problem, by getting the required data for the model which will be collected, such

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 06 Issue 08 August 2023
DOI: 10.47191/ijcsrr/V6-i8-35, Impact Factor: 6.789
IJCSRR @ 2023

as historical data of sales quantity and price of the product. To create a good prediction, the training set should be of good quality. One of the methods to have a good training set is to group the data with similar property, or in this case the researcher groups the data based on UOM or Unit of Measurement. The decision to group data based on unit of measurement is based on an assumption that products with similar unit tend to have similar price range. The next step is to choose a model that can best describe the data. The steps are as follows: First, split the data into two data sets; training set and test set. The training data set will be fed to the model candidates. Next, compare the accuracy of the models with a baseline model. The purpose of comparing the model to a baseline model is to prove that the machine learning model will have a better accuracy of predicting the output in comparison of using generic statistics, in this case median. After comparing the baseline model, compare the model with each other to decide which model has the best prediction by choosing the model with the least errors.

### 3.2 Data Collection Methods

The historical data will be collected by generating the data from database using BigQuery. First, generate the list of products with filter as follows; trading, rice category products that have orders from Jan 2021 – Oct 2022. Then, get the average selling price, and selling quantity of each product on a weekly basis. Next, group the product based on their unit of measurement (UOM). Since there is no available data regarding the UOM and size of the product, the researcher also needs to create an algorithm to identify the size amount and the UOM based on the product name. The data then will be standardized to the same unit after getting the size amount and the UOM, where in this case is in Kilogram.

### 3.3 Data Analysis Methods

The model will be created using Python to search the best demand model function that can be applied for each product or UOM.

#### 3.3.1 Demand Model Function

To create a demand function, the researcher uses historical data which consists of average selling price and selling quantity per UOM on a weekly basis. This data will be split into training and test set with a ratio of 3:1. Where the data should always be consistent in each iteration. There are two main processes in machine learning: learning and predictions.

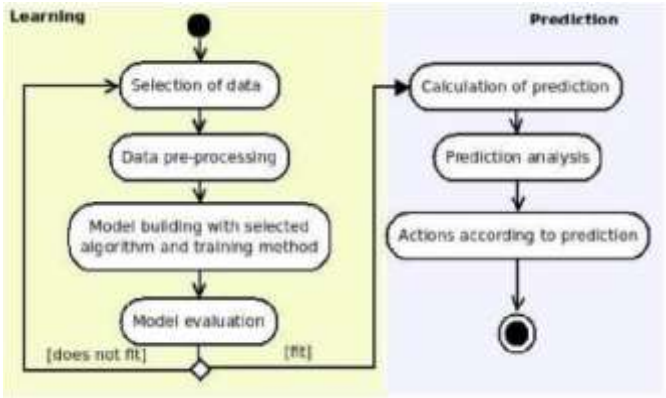


**Figure 1:** Machine Learning Process: Prediction
**Source:** Vitolina, 2015

1. Training or learning process.

Analyzing the issue domain and choosing the data are the first steps in the machine learning process. After thorough investigation of the problem context, several data selection approaches are allowed. The goal of data pre-processing is to produce a sample database that contains the model's training and test data. The training algorithm will then receive training data and a prediction model is being built based on the data. This phase is referred to as the creation of the knowledge base in literature. It should be noted that no machine learning algorithm or approach is obviously superior to another; nonetheless, for more accurate results, each of the machine learning methods should be evaluated with the test data set. The evaluation of model accuracy is a crucial stage. Making sure the prediction model can predict with other data as correctly as it does with the training data set is the aim of the evaluation stage.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 06 Issue 08 August 2023
DOI: 10.47191/ijcsrr/V6-i8-35, Impact Factor: 6.789
IJCSRR @ 2023

2. Second prediction process.

After the model is trained using the training set, the model is then tested using the test data set. There are 4 models that will be compared; linear regression, random forest regression, support vector regression, and gradient boosted regression based on their Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). These 4 models will also be compared to a baseline model, in this case a median model. The reason the researcher is using median instead of mean is due to there are some product's price outliers. Median is used to represent central tendency while shielding it from the impact of outliers or extreme value. (Walpole et al, 2016). It is usually caused by premium rice which has higher price that might skew the mean to the right side. The model that will be taken is the one that has the least error. If the baseline has better error than the other models, it means median can predict demand better than the other models and the researcher will need to find another model for the demand function.

## 4. EXPECTED FINDINGS / SOLUTIONS

### 4.1 Analysis

The analysis started by getting the average selling price and the total quantity sold per product per week between January 2021 until October 2022. The researcher also categorizes each product by its unit of measurement (UOM) to group products with similar price. The data is limited only to trading products from rice product category. The reason the researcher choose rice product in the model is because of its role as the main commodities in Indonesia, rice price is regulated by the government, where for every type of rice & unit of measurement, the price will range around the same price. Hence, it will be easier to model compared to non-commodities products such as fashion or beauty products for example. There are 184 products data from 91 weeks that will be split randomly with ratio 3:1 to training and test data set.

**1 Data Description**

**Table 1:** Descriptive Statistics of the Data

| | UOM | Weighted Average Price | Quantity |
|---|---|---|---|
| **count** | 6,737 | 6,737 | 6,737 |
| **mean** | 5.09 | 65,891.57 | 175.60 |
| **std** | 2.72 | 31,853.55 | 566.20 |
| **min** | 0.48 | 10,953.38 | 1.00 |
| **25%** | 5.00 | 54,000.00 | 4.00 |
| **50%** | 5.00 | 56,463.64 | 19.00 |
| **75%** | 5.00 | 62,500.00 | 95.00 |
| **max** | 20.00 | 259,900.00 | 9,795.00 |

The data contains 6,737 observation points. The minimum UOM observable is 480 grams and maximum at 20 kg. The mean of the weighted average price is Rp 65,891.57, while the median price is at Rp 56,463.64. The fourth quartile of the data is at Rp 62,500, not far from the average price. This means the weighted average price distribution is centered around Rp 60,000. Which is the price point for premium price of 5 kg. This fact is also strengthened by the fact that the mean, median, first quartile, and third quartile of UOM data is at 5 kg. The lowest price of the product is Rp. 10,953,38 and the highest price of the product is Rp. 259,900. The mean of the quantity sold is 175 units, but the median is only 19 units. It implies that the quantity sold distribution is skewed to the right side. This skew is likely caused by upper outliers in quantity sold, where the max quantity sold is at 9,795 units, far higher than the median or mean quantity sold.

**2 Demand Model Construction**

After feeding the training set to each of the models and compare it between each other and to the baseline model. Based on MAE, the only model that has better error compared to the baseline is SVR. While based on RMSE all the other models have better models than baseline with models with the least error in order are Gradient Boosted, Linear Regression, Random Forest, and SVR. In this case, the researcher will use the result from RMSE error, since RMSE will penalize high number error better than MAE. Hence the

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 06 Issue 08 August 2023
DOI: 10.47191/ijcsrr/V6-i8-35, Impact Factor: 6.789
IJCSRR @ 2023

lesser the error, the lesser high number error that might exist in the data. So, based on the RMSE error, the best model to predict the quantity is Gradient Boosted Regression.

**Table 2:** MAE and RMSE Table Result

| | mae | rmse |
|---|---|---|
| Linear Regression | 251.898463 | 657.986746 |
| Random Forest | 209.068257 | 660.751476 |
| SVR | 182.094231 | 686.234826 |
| Gradient Boosted | 224.454537 | 644.300509 |
| Baseline | 185.28368 | 687.017411 |

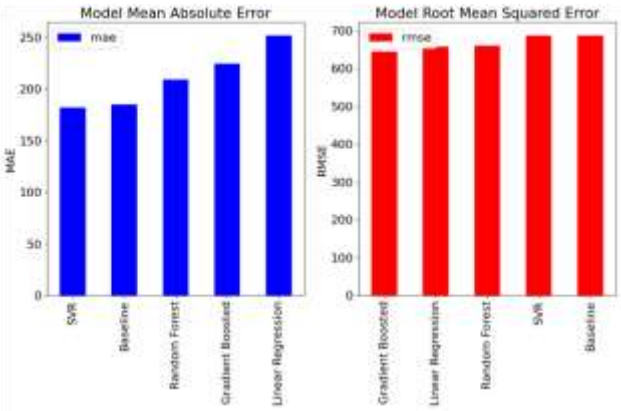


**Figure 2:** MAE and RMSE Plotting Result

## 5. CONCLUSION AND RECOMMENDATION

### 5.1 Conclusion

This paper has a goal to create the best model to maximize profit of e-commerce XYZ based on the historical data of trading rice product sales from January 2021 – October 2022. After testing several models to predict the total quantity of each product group based on price, it's decided that the best demand function model based on RMSE is Gradient Boosted.

### 5.2 Recommendation

This model still needs development in terms of features used and on the classification of the product. There are some features that might be useful to model sales quantity, such as seasonality, user traffic, brand exposure, and price competitiveness. The products can also be classified in one or more units, such as by brand or by product type. It is also encouraged to model other product categories, starting from other commodities such as cooking oil or milk products. The ideal model would be able to model all product categories within the e-commerce company.

## REFERENCES

1. Mussi, M., Genalti, G., Trovò,F., Nuara, A., Gatti, N., & Restelli, M. (2022). Pricing the Long Tail by Explainable Product Aggregation and Monotonic Bandits. Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, USA, 3623–3633. https://doi.org/10.1145/3534678.3539142.
2. Boer, A.V. (2015). Dynamic pricing and learning: Historical origins, current research, and new directions. Surveys in Operations Research and Management Science, 1876-7354. https://doi.org/10.1016/j.sorms.2015.03.001.

International Journal of Current Science Research and Review
ISSN: 2581-8341
Volume 06 Issue 08 August 2023
DOI: 10.47191/ijcsrr/V6-i8-35, Impact Factor: 6.789
IJCSRR @ 2023

3. Lu, C.J. & Kao, L. J. (2016). A clustering-based sales forecasting scheme by using extreme learning machine and ensembling linkage methods with applications to computer server. Engineering Applications of Artificial Intelligence, 09521976. http://dx.doi.org/10.1016/j.engappai.2016.06.015.
4. Dai, W., Chuang, Y. Y., Lu, C. J. (2015). A clustering-based sales forecasting scheme using support vector regression for computer server. 2nd International Materials, Industrial, and Manufacturing Engineering Conference, Indonesia, 2351-9789. https://doi.org/10.1016/j.promfg.2015.07.014
5. Khouja, M., & Robbins, S. S. (2005). Optimal pricing and quantity of products with two offerings. European Journal of Operational Research, 163(2), 530-544. https://doi.org/
6. Rebala, G., Ravi, A., Churiwala, S. (2019). Machine Learning Definition and Basics. In: An Introduction to Machine Learning. Springer, Cham. https://doi.org/10.1007/978-3030-15729-6_1.
7. Witten, I. H., Frank, E., Hall, M., & Pal, C. J. (2017). Data Mining Practical Machine Learning Tools and Techniques (4th ed.). Morgan Kaufmann.
8. Mueller, A. C., & Guido, S. (2016). Introduction to Machine Learning with Python: A Guide for Data Scientists. O'Reilly.
9. Išoraitė, Margarita, and Neringa Miniotienė. "Electronic Commerce: Theory and Practice." Integrated Journal of Business and Economics, vol. 2, no. 2, 2018, pp. 73-79, doi:10.33019/ijbe.v2i2.78.
10. Walpole, Ronald E., Myers, Raymond H., Myers, Sharon L., Ye, Keying. (2016). Probability & statistics for engineers & scientists (9th ed. Global ed.). Tokyo: Pearson.
11. Whinston A. B. Stahl D. O. & Choi S.-Y. (1997). The economics of electronic commerce. Macmillan Technical Pub.