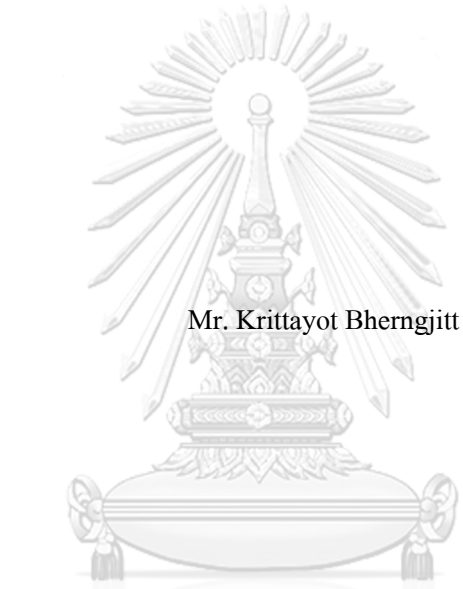
CUSTOMER DRINKS PURCHASING BEHAVIOR DURING COVID-19 PANDEMIC ANALYSIS

Mr. Krittayot Bherngjitt



A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science in Computer Science

Department of Computer Engineering

FACULTY OF ENGINEERING

Chulalongkorn University

Academic Year 2022

การวิเคราะห์พฤติกรรมการซื้อเครื่องดื่มของลูกค้าในช่วงโควิด-19ระบาด

นายกฤตยชญ์ พึ่งจิตต์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2565

Thesis Title	CUSTOMER DRINKS PURCHASING BEHAVIOR DURING COVID-19 PANDEMIC ANALYSIS
By	Mr. Krittayot Bherngjitt
Field of Study	Computer Science
Thesis Advisor	Assistant Professor SUKREE SINTHUPINYO, Ph.D.

Accepted by the FACULTY OF ENGINEERING, Chulalongkorn University in Partial Fulfillment of the Requirement for the Master of Science

..... Dean of the FACULTY OF ENGINEERING

()

THESIS COMMITTEE

..... Chairman

(Assistant Professor NATTEE NIPARNAN, Ph.D.)

..... Thesis Advisor

(Assistant Professor SUKREE SINTHUPINYO, Ph.D.)

..... Examiner

(Assistant Professor NUTTAPONG CHENTANEZ, Ph.D.)

..... External Examiner

(Assistant Professor Denduang Pradubsuwun, Ph.D.)

กฤตยชญ์ พึ่งจิตต์ : การวิเคราะห์พฤติกรรมการซื้อเครื่องดื่มของลูกค้าในช่วงโควิด-19 ระบาด. (CUSTOMER DRINKS PURCHASING BEHAVIOR DURING COVID-19 PANDEMIC ANALYSIS) อ.ที่ปรึกษาหลัก : ผศ. ดร.สุกรี สินธุภิญโญ

การระบาดของโรคโควิด-19นั้นทำให้วิธีการใช้ชีวิตของทุกคนเปลี่ยนไปเช่นการปิดเมืองทำให้ทุกคนต้องอยู่ที่บ้านทั้งวันเป็นเวลาหลายเดือน ทำให้พฤติกรรมการซื้อของของลูกค้าเปลี่ยนไปเช่นการสั่งซื้อแบบส่งของมีการเพิ่มขึ้นอย่างมาก งานวิจัยนี้จะวิเคราะห์พฤติกรรมการซื้อเครื่องดื่มของของลูกค้าโดยการนำข้อมูลการขายเครื่องดื่มจากบริษัทเครื่องดื่มขนาดใหญ่ในช่วงปลายปี2019ถึงปี2021มาทำการจัดกลุ่มข้อมูลและแบ่งลูกค้าออกเป็นกลุ่มตามการซื้อสินค้าแต่ละประเภทและนำข้อมูลการซื้อสินค้าของลูกค้ามาสร้างโมเดลที่จะทำการทำนายการพฤติกรรมการสั่งซื้อของลูกค้า สำหรับการจัดกลุ่มข้อมูลนั้นเราจะใช้วิธีK-meansโดยเราจะหาค่าKด้วยวิธีelbow methodและเราจะแบ่งข้อมูลเป็นการขายรายเดือนก่อนที่จะจัดกลุ่มลูกค้าแต่ละเดือน จากนั้นเราจะนำค่ากลางของแต่ละกลุ่มมาจัดกลุ่มอีกทีเพื่อหากลุ่มที่ครอบคลุมข้อมูลทั้งหมด เมื่อเราได้กลุ่มทั้งหมดแล้วเราก็จะเริ่มสร้างโมเดลที่จะใช้ทำนายการซื้อ สุดท้ายแล้วเราก็จะประเมินการทำนายของโมเดลโดยการใช้ Relative Root Mean Square Error เทียบกับวิธีอื่นๆ เช่น Random Forest Regression และ XGBoost โดยเราต้องการดูว่าการทำนายโดยการใช้ข้อมูลจากการจัดกลุ่มจะมีผลลัพธ์ที่ดีกว่าการใช้ข้อมูลการขายปกติหรือไม่ จากการเปรียบเทียบเราได้พบว่าการทำนายการขายโดยใช้ข้อมูลจากการจัดกลุ่มนั้นมีความแม่นยำใกล้เคียงกับการทำนายโดยใช้ข้อมูลขายปกติแต่ใช้เวลาในการฝึกโมเดลและทำนายน้อยกว่ามาก

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่อนิสิต .....

ปีการศึกษา 2565 ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

6270008021 : MAJOR COMPUTER SCIENCE

KEYWORD: K-means clustering, predictive modeling, Machine Learning

Krittayot Bherngjitt : CUSTOMER DRINKS PURCHASING BEHAVIOR DURING COVID-19 PANDEMIC ANALYSIS. Advisor: Asst. Prof. SUKREE SINTHUPINYO, Ph.D.

The COVID-19 pandemic has caused many changes to the lifestyle of people all over the world. The lockdown forced people to stay at home for many months. This has led to the changes in purchasing behavior as well such as the increase in delivery order. This research, which has received sales data of drinks during the pandemic from a large beverage company, seeks to analyze the changes in customer behavior during the pandemic by using machine learning to perform clustering and observe the changes in the purchases of each product type. We will use clustering to group customers based on their purchase behavior and create prediction models that can predict what customers will order based on the purchase history of the group of customers in the data. We will be using K-means clustering with elbow method for finding K. We will split the data into monthly sales and perform clustering on each month, and then we will perform clustering again with the data from each cluster to find global clusters that allow us to compare the clusters directly. We will then use the result to create 3 types of prediction models, namely LSTM, Random Forest Regression and XGBoost. Finally, we compare the result from the models trained by global cluster training data to the ones from the models trained by the customer's sales training data to see if global cluster training data can compete with using sales training data. We found that the models trained by global cluster customer training data performed similarly to the ones trained by sales training data but took much less time to train and run.

Field of Study:	Computer Science	Student's Signature .....
Academic Year:	2022	Advisor's Signature .....

ACKNOWLEDGEMENTS

I would like to thank Asst. Prof. Sukree Sinthupinyo, Ph.D. for his help in guiding me through the research. His guidance helped me complete this thesis.

I would also like to thank the thesis committee Asst. Prof. Nattee Niparnan, Ph.D., Assistant Professor Nuttapong Chentanez, Ph.D., Assistant Professor Denduang Pradubsuwun, Ph.D. for their comments and questions during the exam.

Krittayot Bherngjitt

TABLE OF CONTENTS

Page

..... iii

ABSTRACT (THAI) iii

..... iv

ABSTRACT (ENGLISH) iv

ACKNOWLEDGEMENTS v

TABLE OF CONTENTS vi

LIST OF TABLES ix

LIST OF FIGURES xii

1 Introduction 1

1.1 Background 1

1.2 Objective 2

1.3 Scope 2

1.4 Expected Outcomes 2

2 Related Theories 3

2.1 Clustering 3

2.1.1 K-means Clustering 3

2.1.2 Elbow method 4

2.2 Predictive Modeling 5

2.2.1 LSTM 5

2.2.2 XGBoost 7

2.2.3 Random Forest Regression 8

3 Related Works....10
3.1 Analyzing customer buying behavior....10
3.2 Data mining techniques: A source for consumer behavior analysis....10
3.3 Decision Tree Based Targeting Model of Customer Interaction with Business Page....10
3.4 Forecasting Sales in the Supply Chain Based on the LSTM Network: The Case of Furniture Industry....11
3.5 Sales forecasting using multivariate long short term memory network models....11
4 Methodology....12
4.1 Data gathering & preprocessing....14
4.1.1 Data gathering....14
4.1.2 Data Preprocessing....15
4.1.2.1 Separating the data by month....16
4.1.2.2 Data Aggregation....16
4.2 Machine learning....17
4.2.1 K-means clustering....18
4.2.1.1 Elbow Method....18
4.2.1.2 Clustering Result....19
4.2.2.3 Finding the global clusters....25
4.3 Prediction Model....29
4.3.1 Training the model....29
4.3.2 Evaluation....30
4.3.3 Building model for global cluster....31
5 Result and Analysis....32
6 Summary....46

REFERENCES 48

VITA 50

LIST OF TABLES

Page

Table 1: An example of sales data from year 2020.....15

Table 2: Examples of sales data being separated by month.....16

Table 3: An example of sales data being grouped by product types.....17

Table 4: Centroid of each product type of each cluster in a month 19

Table 5: The monthly cluster of customers in year 2020.....22

Table 6: the centroid of sales value for each cluster.....23

Table 7: Showing how monthly clusters change to global clusters 26

Table 8: Customers with monthly clusters (yellow) and global clusters (blue).....26

Table 9: Centroid of each product type in each global cluster.....27

Table 10: Comparison of RRMSE between the LSTM model with each type of training data for each customer for colored liquor products.....33

Table 11: Comparison of total RRMSE between the LSTM model with each type of training data for each customer for colored liquor products. 34

Table 12: Comparison of RRMSE between the LSTM model with each type of training data for each customer for white liquor products.....34

Table 13: Comparison of total RRMSE between the LSTM model with each type of training data for each customer for white liquor products. 34

Table 14: Comparison of RRMSE between the LSTM model with each type of training data for each customer for beer products. 35

Table 15: Comparison of total RRMSE between the LSTM model with each type of training data for each customer for beer products. 36

Table 16: Comparison of RRMSE between the LSTM model with each type of training data for each customer for Oishi products 37

Table 17: Comparison of total RRMSE between the LSTM model with each type of training data for each customer for Oishi products. 37

Table 18: The performance of LSTM model with each type or training data. 38

Table 19: Comparison of RRMSE between the XGBoost model with each type of training data for each customer for colored liquor products 38

Table 20: Comparison of total RRMSE between the XGBoost model with each type of training data for each customer for colored liquor products 38

Table 21: Comparison of RRMSE between the XGBoost model with each type of training data for each customer for white liquor products. 39

Table 22: Comparison of total RRMSE between the XGBoost model with each type of training data for each customer for white liquor products. 39

Table 23: Comparison of RRMSE between the XGBoost model with each type of training data for each customer for beer products.....40

Table 24: Comparison of total RRMSE between the XGBoost model with each type of training data for each customer for beer products. 40

Table 25: Comparison of RRMSE between the XGBoost model with each type of training data for each customer for Oishi products. 41

Table 26: Comparison of RMSE between the LSTM model with Random Forest Regression and XGBoost for Oishi products using global cluster centroid as training data.....41

Table 27: The performance of XGBoost model with each type or training data. 41

Table 28: Comparison of RRMSE between the Random Forest model with each type of training data for each customer for colored liquor products. 42

Table 29: Comparison of total RRMSE between the Random Forest model with each type of training data for each customer for colored liquor products 42

Table 30: Comparison of RRMSE between the Random Forest model with each type of training data for each customer for white liquor products. 43

Table 31: Comparison of total RRMSE between the Random Forest model with each type of training data for each customer for white liquor products. 43

Table 32: Comparison of RRMSE between the Random Forest model with each type of training data for each customer for beer products. 44

Table 33: Comparison of total RRMSE between the Random Forest model with each type of training data for each customer for beer products. 44

Table 34: Comparison of RRMSE between the Random Forest model with each type of training data for each customer for Oishi products. 45

Table 35: Comparison of total RRMSE between the Random Forest model with each type of training data for each customer for Oishi products. 45

Table 36: The performance of Random Forest model with each type or training data. 45

Table 37: The comparison of time taken to train and run the model for each kind of models and training data 46

LIST OF FIGURES

Page

Figure 1: An example of data pre-K-means clustering 3

Figure 2: An example of data post K-means clustering 4

Figure 3: An example graph for elbow method 4

Figure 4 : The process of predictive modeling. 5

Figure 5: A single LSTM cell. 6

Figure 6: Overview of XGBoost 8

Figure 7: Overview structure of Random Forest regression 9

Figure 8: The process of the research 13

Figure 9: An example of using Elbow method to find the number of clusters, with the number of clusters in X axis and inertia in Y axis. 18

Figure 10: The result of a clustering of January 2020 sales data as a graph with product groups in X axis and centroid of sales value in Y axis. 20

Figure 11: The result of a clustering of February 2020 sales data as a graph with product groups in X axis and centroid of sales value in Y axis. 21

Figure 12: The result of a clustering of March 2020 sales data as a graph with product groups in X axis and centroid of sales value in Y axis. 22

Figure 13 The number of clusters in each month 24

1 Introduction

1.1 Background

The COVID19 pandemic has resulted in a drastic shift in the lifestyles of everyone. In order to prevent large scale infection, people use multiple methods to avoid contact with one another such as social distancing and working from home. This has led to changes in customers' purchasing behavior like more food delivery. This research is interested in the purchasing behavior of drinks because they have been heavily affected by the pandemic. The lockdown that happened due to the COVID19 pandemic forced restaurants to close early and to seat fewer customers. This leads to more people preferring to order food and drinks online. Since we don't know the customers' purchasing behavior from the raw data, we should use clustering to group customers based on purchasing behavior.

Clustering is a method to analyze the sales data is to group customers together with other customers with similar purchasing behaviors. A way to use post clustering sales data to benefit the company is to create prediction models which can forecast the sales of products. There are multiple kinds of prediction model such as decision trees, regression, neural networks.

This research seeks to see how the customer behaviors during a lockdown affect a company which sells products to restaurants and supermarkets as well as small stores. This research has been provided the drink sales data of a certain sales team from a certain big beverage company. The company also has many brands under its umbrella which allows this research to see if certain brands perform better or worse relative to each other during the lockdown. The research will use clustering to split the customer into multiple clusters based on the number of each type of product they purchase. The result of the clustering will be used to create prediction models to predict the sales of products for each customer. To identify the model which best fit the set of sales data and to see how using the data from clustering as training data for the model can compare to simply using sales data as training data, this research will compare the result from 3 types of prediction model, and

the result from different sets of training data, between the global cluster centroids and the sales data.

1.2 Objective

1. To study the changes in customer purchasing behavior as time went on during the pandemic.
2. To create prediction models that can predict the sales of each product type for each customer.
3. To compare the performance of 3 types of prediction models and the performance of the models with different training data types.

1.3 Scope

This research uses sales data from October 2019 to December 2021 and the. This data allows us to compare the sales of different months and years in order to see the shift in purchasing behaviors and trends caused by living under the pandemic and the lockdown. This research will be comparing the results from 3 types of prediction models which are LSTM, XGBoost, and Random Forest.

1.4 Expected Outcomes

The expected outcomes for this research are as follow

1. Show how similar each customer's purchasing behavior is compared to other customers in the same month and group them together.
2. Show what kind of prediction model works the best for the drink sales data from the company and show how each type of prediction model performs with different types of training data

2 Related Theories

The backgrounds of this research consist of methods commonly used in data mining and tools used for programming. To create a prediction model, we will be using Google Colab to run python code and perform clustering

2.1 Clustering

Clustering is a method of categorizing data into groups which help in data mining. There are multiple methods that can be used for clustering such as k-means clustering, PAM, CLARA. Picking suitable clustering methods for the data type is essential to get good result from data mining. This research will be using k-means clustering. [1]

2.1.1 K-means Clustering

K-means clustering is a method of clustering that splits the data into K clusters. K-means clustering works by identifying K number of centroids and assign each data row to the closest one, creating K clusters. In order to find the optimal value of K, we will be using the elbow method. [1]

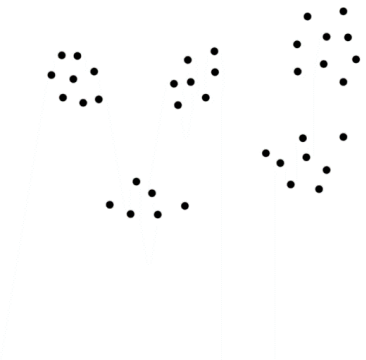


Figure 1: An example of data pre-K-means clustering

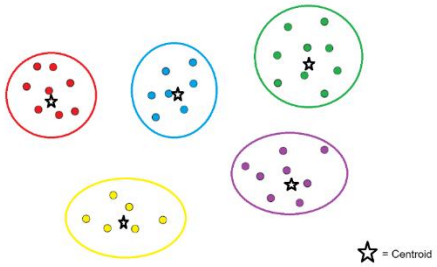


Figure 2: An example of data post K-means clustering

Figure 1 and 2 show how data sets would look like before and after k-means clustering. Figure 2 also shows the centroid of each cluster as stars.

2.1.2 Elbow method

Elbow method is a way to find the best value of K in K-means clustering. It works by plotting the Within-Cluster Sum of Square (WCSS), the sum of the distance between the object and the centroid, against the value of K. We will then select the value of K that is at the value of k where the value of WCSS began changing less compared to the value of k, the elbow of the graph. [2]

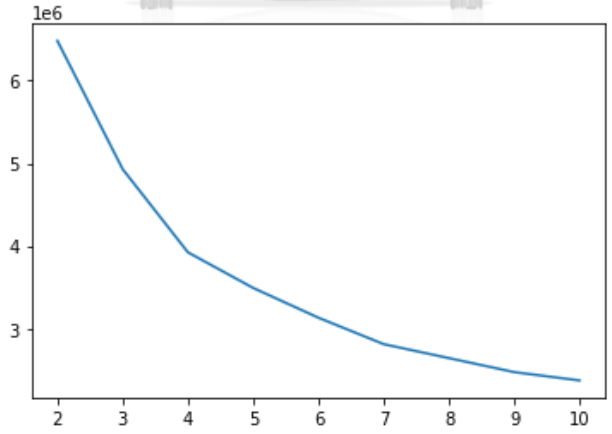


Figure 3: An example graph for elbow method

We can see from figure 3 that the point where WCSS was changing rapidly with k until the point where k = 4, we can call k = 4 is the elbow of the graph, so we pick k=4 for our k-means clustering

2.2 Predictive Modeling

Predictive modeling is a technique of using machine learning and data mining to predict the future by using data input. The model predicts the most likely outcome based on the data given as input. The model will be using is the clustering model, which predicts the future by assigning data into groups called clusters. The model will predict the future by putting the new customer into the most similar cluster.

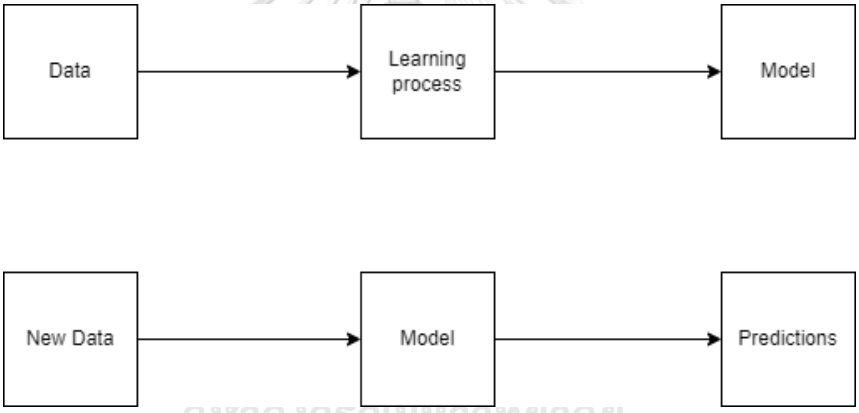


Figure 4 : The process of predictive modeling.

Figure 4 shows how prediction model works, we use the first set of data, training data, to train the model and use the second set, test data, to get prediction results from the model.

2.2.1 LSTM

The first type of prediction model we will use is the Long Short Term Memory (LSTM). LSTM is a type of recurrent neural network (RNN) that is separated into multiple layers and can send information to the next states which makes it suitable for time series data. LSTM combines hidden states which contain short term memory with cell states which contain long

term memory. It also employs function gates to determine if the information will be dropped or sent to the next layer. [3, 4]

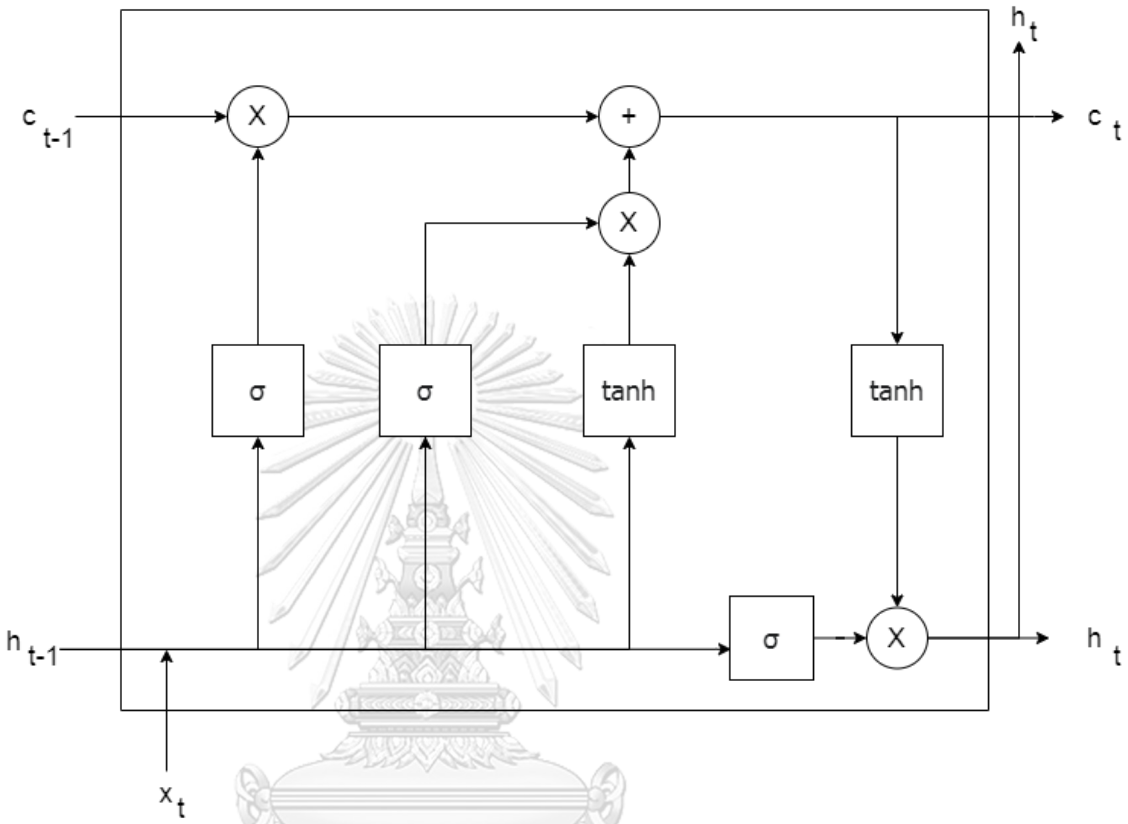


Figure 5: A single LSTM cell.

x_t denotes Input for current layer

h_{t-1} denotes Output from last layer

h_t denotes Output from current layer

c_{t-1} denotes Memory from last layer

c_t denotes Memory from current layer

The model works by using the forget gate to drop out irrelevant information, using sigmoid function to determine if the input is worth keeping or not. The input with results equal to 1 are kept and those with result equal to 0 are dropped. The input gate then decides which input will be added to the cell state by using the hidden state. This will result in a new cell state and the process is repeated until the last layer. In order to determine the result of the LSTM model, the output gate put the current input and the output from last layer into a sigmoid function gate and tanh function gate then perform multiplication and addition with the memory from last layer to make the decision. [3, 4]

2.2.2 XGBoost

We will be using an XGBoost model as the second type of model. XGBoost is a type of decision tree that can use the result from a tree to improve the performance of the next tree, this is called boosting. XGBoost provides the ability to boost multiple trees at the same time unlike normal gradient boosted decision trees. XGBoost can be used to build a prediction model by combining the weighted results from each tree. [5, 6]

The general equation of XGBoost model can be shown as follow:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \quad (1)$$

The equation (1) represents the general equation of the XGBoost model Where K is the number of trees, $f_k(x_i)$ is the prediction of the k^{th} tree and $\hat{y}_i$ is the predicted value [5, 6]

The objective function of the model can be defined as follow:

$$\varphi(\emptyset) = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (2)$$

The equation (2) represents the objective function of the XGBoost model Where y_i *is the actual value,* $L(y_i, \hat{y}_i)$ *is the loss function which is* the difference between the predicted value and the actual value and Ω *is the regularization term which controls the complexity of the model.* [5, 6]

The overview of the XGBoost model can be seen in the figure below. Each tree uses the prediction from the previous tree to improve its own prediction and the final result is the weighted sum of the prediction sum from all trees.

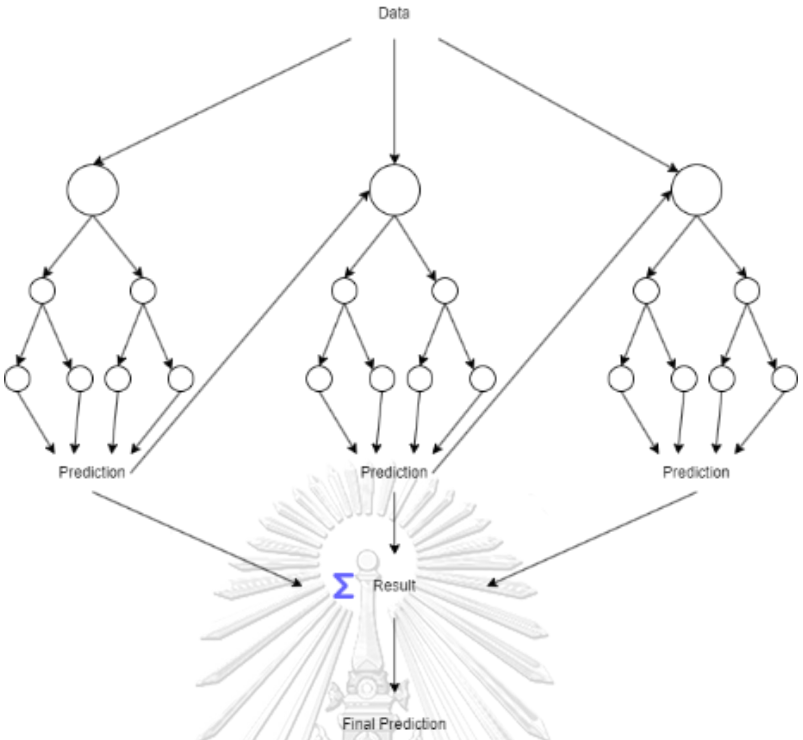


Figure 6: Overview of XGBoost

2.2.3 Random Forest Regression

The last type of prediction model we will be using is the random forest regression model. Random Forest Regression is a type of decision tree that operates by running multiple decision trees with random data from the training dataset and random feature per decision split. It can be used to make a prediction model by having each tree makes its own prediction and use the average result as the prediction for the model. [7, 8]

The overview of the Random Forest Regression model can be seen in the figure below. The trees are separated from each other and the only interaction between them is the average of their predictions for the final result.

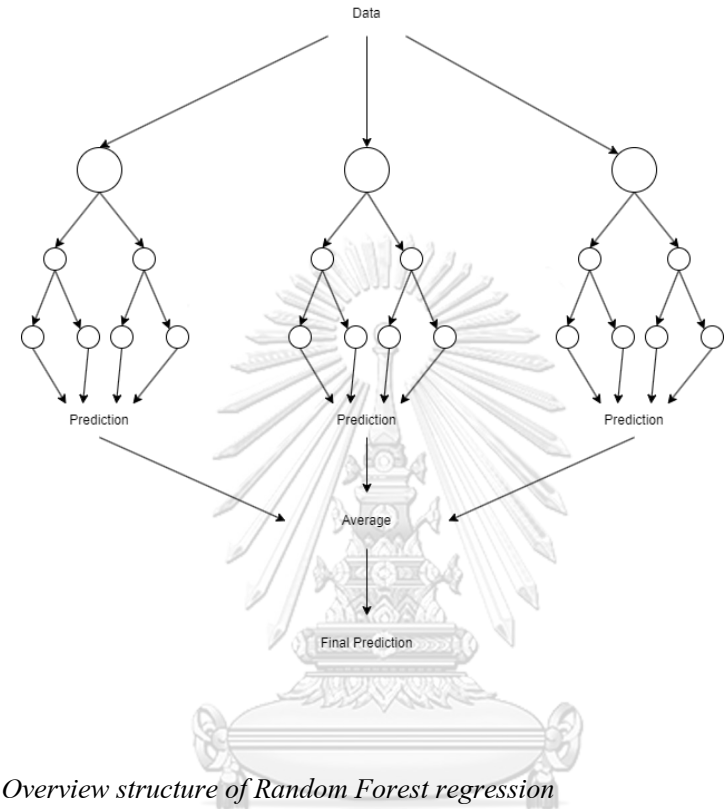


Figure 7: Overview structure of Random Forest regression

3 Related Works

There are many researches about analyzing customer purchasing behavior with data mining techniques in order to improve the performance of the companies. A few of the researches are as follow.

3.1 Analyzing customer buying behavior

Analyzing customer buying behavior by Tanya Nayyar [9] is about using multiple data mining methods to analyze the buying behavior of customers in order to retain existing customers and expand the customer base of mid-west tools manufacturing company in the USA. The research used multiple methods such as logistic regression, decision trees, support vector machine, naïve bayes and random forest to perform the analysis of data given by the company.

3.2 Data mining techniques: A source for consumer behavior analysis

This research [10] is about how consumers may be influenced by their environments when looking to buy a product. It used various data mining techniques such as association rules to help understand consumers' feeling toward purchase differs for similar products, how they make their decision, and how their decisions are influenced by various external factors such as marketing. Finally, it used the knowledge it gains to tell the management on how to improve their marketing so that they can reach the customer more effectively. The research concludes that data mining is useful for business to know about their customers' buying habit and trends, which can then be used to update their services to satisfy the customers.

3.3 Decision Tree Based Targeting Model of Customer Interaction with Business Page

This research [11] is about using decision trees to create a model of customer interaction with business page on Facebook so that the companies can improve their advertisement and target the customers more effectively. This research used recursive separation and regression tree to construct the model and R language for

programming. It looked at the age, sex, the number of actions on the business page, and the number of times the advertisement was successful in reaching a specific goal. This research concludes that decision trees can simplify data flows and that the company can take the result of the research into account when making an advertisement so that it would reach more people.

3.4 Forecasting Sales in the Supply Chain Based on the LSTM Network: The Case of Furniture Industry

This research [4] is about using LSTM to forecast the sales of furniture with historical sales data as input. The research uses data from January 2017 to March 2019 to build the LSTM network, with the data from 2017 to 2018 as training set and the last 2 months of 2018 being the validation set and the data from 2019 as test set. The research used Keras to build LSTM network with 2 layers, with the second being the output layer. It also used min max scaling to normalize the data. The research concludes that the LSTM network can recognize long term relationships.

3.5 Sales forecasting using multivariate long short term memory network models

This research [3] is about improving the ability of LSTM model to forecast sales by using peephole connections. It used sales data of 1,115 Rossmann stores in Germany.This research compared the result of the improved model with basic LSTM model, XGBoost, Random Forest regression by using Root Mean Square Error and Mean Absolute Error to evaluate the performance of the model and found that the normal LSTM had the least accurate predictions among the prediction methods tested for half of the test customers and the most accurate predictions for the other half, but the improved LSTM using peephole connection had the most accurate predictions. It concludes that the improved LSTM model performed 20% better than the initial LSTM model.

4 Methodology

In order to create a prediction model, we will need to prepare training data. To get the training data, we will need to perform data gathering and preprocessing to group the sales data by product type. Then we can use machine learning to perform clustering to separate the sales data into multiple clusters with similar sales. We will then repeat the process for every month so we can get the cluster changes of the whole period.

After getting every monthly cluster from the data, we will perform another clustering using each cluster as a node to find the global clusters of data. The new clusters will be used to find how similar the clusters in different months are to clusters in other months. Once we have gotten the global cluster, we will create the prediction model using 3 types of models, which are LSTM, XGBoost, and Random Forest Regression. We will use the sales data from the last month as test data and use Relative Root mean Square Error to evaluate the performance of the models. We will compare the result of the prediction of the models with global cluster as training data to the model with sales data as training data. We will also compare the results of the 3 types of models to see which one performed the best for our data set.

The figure below shows the process of the research.

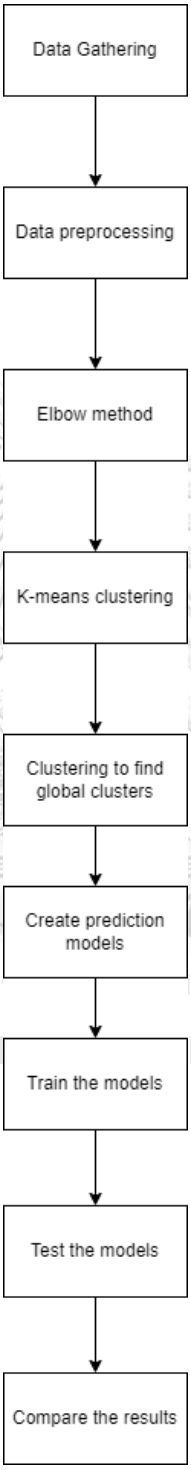


Figure 8: The process of the research

We can see from figure 8 that the research performs clustering twice, the first time to find the clusters in each month, and the second time to find the global clusters. We then create the prediction models with the global cluster data as training data and another set of models with sales data as training data. Finally, we compare the performance between the multiple types of prediction model and the different types of training data.

4.1 Data gathering & preprocessing

We have gotten sales and product data from a large beverage company. This includes the product name, quantity sold, payment amount and sale order in which the products are sold. As the data is confidential, we need to use the auto generated number in the product and customer table instead of their name to differentiate the products and customers. This research will be performing clustering of data to group them into separate clusters for each month, then perform another clustering using the cluster themselves to find the clusters for the whole period which will be used for creating the prediction model later.

4.1.1 Data gathering

We got the sales data of customers in a certain sales team by querying the data from the company's database. We got the sales data of customers in a certain sales team from October 2019 to December 2021. Due to the data being confidential, we have to use product id and customer id in place of their name. We are interested in the customer and the sales quantity of each product. The sample data can be seen in table 1 below.

SaleOrderId	ProductId	CustomerId	OnDate	Quantity	Amount	CustomerCatId	Product Type
45288221	2864	1957178	23-12-2019 15:20:57	1	589	507	Beer
45287697	3002	1957177	23-12-2019 11:55:13	3	0	507	Oishi
45287697	3181	1957177	23-12-2019 11:55:13	6	77.52	507	Oishi
44164603	3254	1947585	20-11-2019 18:35:15	2	1250	504	Beer
45186193	1901	1945583	20-12-2019 13:32:03	6	316.5	504	White liquor
43062475	2864	1418515	19-10-2019 13:05:44	1	589	507	Beer

Table 1: An example of sales data from year 2020

Table 1 shows the sales data before the conversion into monthly sales per product type per customer. We can see the date of the purchase and the quantity of the items purchased by the customer in each order, as well as the product type and the total amount of payment. The customer names and product names are replaced by id number to keep the data confidential.

The sales data contain 12 product types in total

- Colored liquor
- White liquor
- Beer
- Oishi
- Water
- Est
- Group1
- GF&N1
- Soda
- RTD
- 100Plus
- Other

From the company's database, we have 12 total product types as shown in table 2. For our prediction models, we choose to use only the first 4 types of products due to the low sales number of other product types.

4.1.2 Data Preprocessing

Before we can begin clustering and data mining, we must prepare the data first. We perform data preprocessing in multiple steps as detailed below.

4.1.2.1 Separating the data by month

First, we must separate the yearly sales data into monthly sales data. We do this so we can look at the changes in sales data over the months as the lockdown continues. As shown in Table 3

SaleOrderId	ProductId	CustomerId	OnDate	Quantity	Amount	CustomerCatId	ProductGroupName
57669879	3511	2076392	2020-12-28 11:59:09.000	1	304	506	Beer
57300809	3511	2076192	2020-12-17 17:15:45.000	3	912	506	Beer
57300809	3511	2076192	2020-12-17 17:15:45.000	3	5472	506	Beer
57760905	3511	75987	2020-12-30 10:06:55.000	1	304	507	Beer
57760905	3511	75987	2020-12-30 10:06:55.000	1	304	502	Beer
57760905	3511	75987	2020-12-30 10:06:55.000	1	304	507	Beer
57760905	3511	75987	2020-12-30 10:06:55.000	1	304	507	Beer
57286193	3511	68954	2020-12-17 15:31:45.000	1	304	507	Beer
57286193	3511	68954	2020-12-17 15:31:45.000	1	304	507	Beer

Table 2: Examples of sales data being separated by month

4.1.2.2 Data Aggregation

We perform data aggregation by pivot the data so that the raw sales data are turned into sales categorized by product types such as water, alcoholic drinks, etc. We repeat this process for each month of sales data. The aggregated data can be seen in the table below.

CustomerId	Colored liquor	White liquor	Beer	Oishi	Water	Est	Group1	GF&N1	Soda	RTD	100Plus	Other
67552	132	56	12	0	0	0	0	0	0	0	0	0
67573	32	28	0	0	0	0	0	0	0	0	0	0
67577	111	90	24	72	0	0	0	0	0	0	0	0
67598	75	39	33	84	0	27	0	0	0	0	0	0
67604	32	232	8	0	0	0	0	0	0	8	0	0
67608	39	12	51	24	150	15	0	0	18	0	0	0
67620	40	64	4	0	0	0	0	0	0	0	0	84
67633	111	126	0	0	0	0	0	0	0	0	0	0
67654	252	260	300	216	276	4	0	0	44	0	0	0
67686	75	72	24	0	0	0	0	0	0	0	0	0
67701	52	144	72	16	72	0	0	0	0	0	0	84
67708	0	3	0	0	0	0	0	0	0	0	0	0
67726	212	92	508	16	0	0	0	0	24	12	0	92

Table 3: An example of sales data being grouped by product types.

Table 3 shows the monthly sales data after it has been grouped by product types and customers. The number refers to the quantity of the item bought by that customer in each product type.

4.2 Machine learning

From the data, we can see that the data is separated by customers and product types. We can use K-means clustering to group customers with similar purchasing behavior together. We use machine learning to perform analysis of the data due to its ability to handle large amount of data and the ability to make prediction based on training. This research will be using machine learning to perform K-means clustering of the sales data and create a prediction model that can predict the behavior of new customers. This research will be using the data from October 2019 to December 2021 in order to detect the purchasing behavior of customers and group them into clusters. We will be able to identify the product types that are in demand during the pandemic by looking at the cluster with high number of customers. The

models created by this research will be able to predict the sales of certain product types for each customer, which can help the company make decision related to production.

The steps followed by this research to use machine learning to perform the clustering of data are as follow.

4.2.1 K-means clustering

4.2.1.1 Elbow Method

We use K-means clustering to perform clustering of our sales data. We begin by using the elbow method to determine the suitable K for the clustering. We can do that by plotting the number of cluster (K) against inertia.

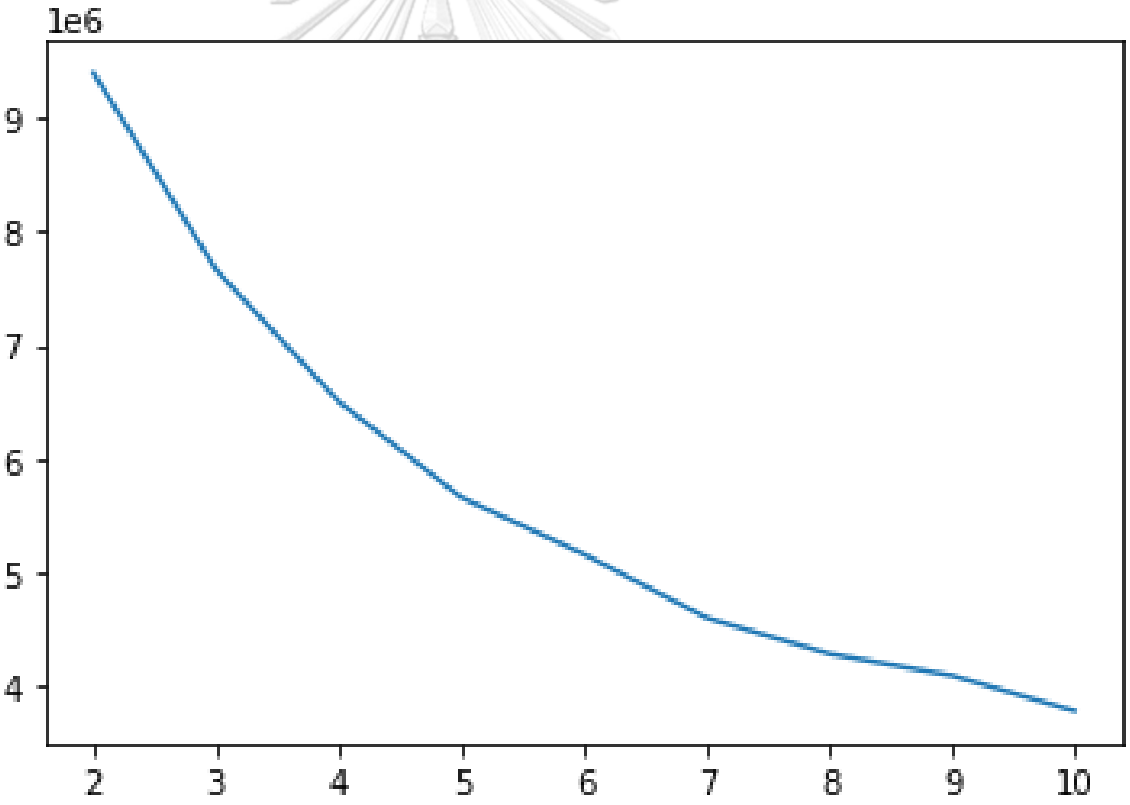


Figure 9: An example of using Elbow method to find the number of clusters, with the number of clusters in X axis and inertia in Y axis.

From figure 9, we can see the elbow of the graph, the point where WCSS began changing less, is at when K = 5. Thus; we can conclude that the number of clusters of this data set should be 5.

4.2.1.2 Clustering Result

After we have found the suitable number of clusters, we will get the centroid of sales per product type in each cluster for the month of January 2021 as shown in the table below.

Cluster	Colored liquor	White liquor	Beer	Oishi	Water	Est	Group 1	GF&N 1	Soda	RTD	100Plus	Other
Jan-2021-0	72.19	190.18	16.77	16.45	5.20	1.91	0.00	0.00	0.59	0.00	0.00	2.41
Jan-2021-1	24.95	38.98	9.96	6.70	4.90	0.39	0.00	0.00	0.65	0.00	0.00	0.14
Jan-2021-2	223.47	426.23	47.07	52.47	52.83	3.13	0.00	0.00	4.50	1.47	0.00	4.03
Jan-2021-3	134.73	124.73	260.60	46.07	54.00	0.53	0.00	0.00	7.87	0.00	0.00	8.80
Jan-2021-4	52.14	84.69	26.14	151.40	14.46	3.88	0.00	0.00	1.51	0.06	0.00	0.61

Table 4: Centroid of each product type of each cluster in a month

We can visualize the centroid of each cluster using bar chart to show how different clusters have significantly different sales for each product type.

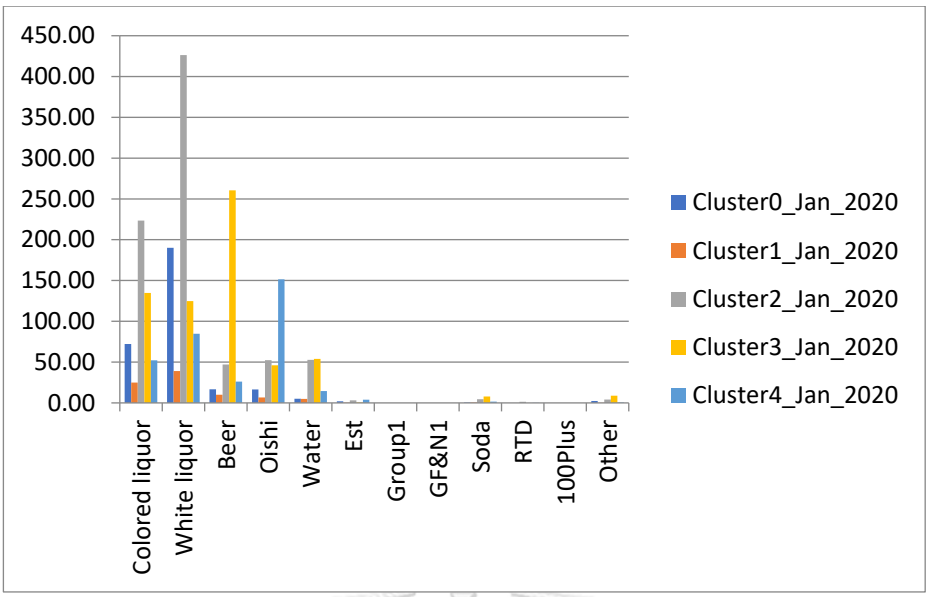


Figure 10: The result of a clustering of January 2020 sales data as a graph with product groups in X axis and centroid of sales value in Y axis.

- Cluster 0 has the second highest sales for white liquor and the third highest for colored liquor but low sales for other product types
- Cluster 1 has low sales for every product types.
- Cluster 2 has the highest sales for both colored and white liquor products.
- Cluster 3 has the highest sales for beer and second highest for colored liquor and third highest for white liquor.
- Cluster 4 has the highest sales for Oishi products.

The chart below shows that there are similar clusters in each month.

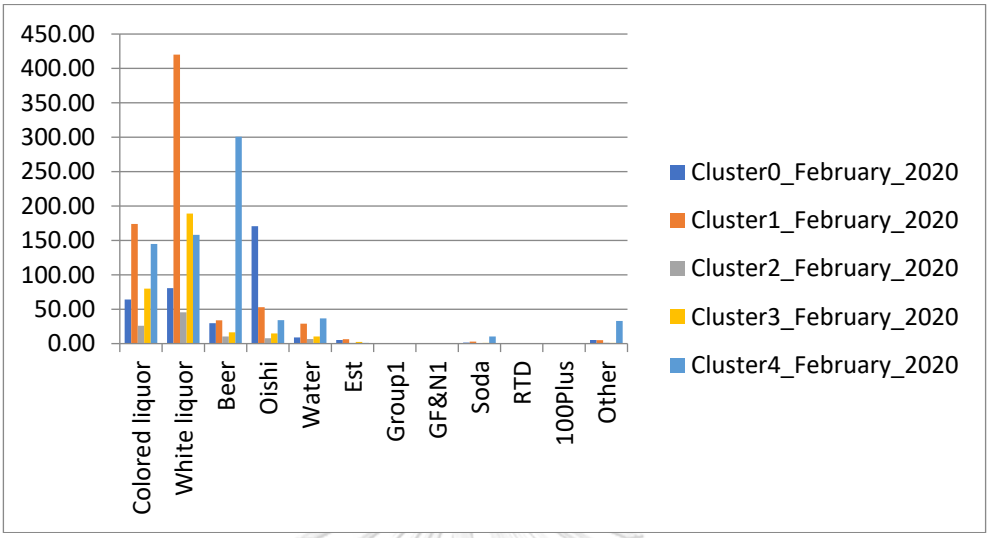


Figure 11: The result of a clustering of February 2020 sales data as a graph with product groups in X axis and centroid of sales value in Y axis.

We can see from figure 11 that there are some clusters similar to the clusters from January 2020.

- Cluster 0 has the highest Oishi sales similar to cluster 4 from January 2020
- Cluster 1 has the highest white liquor and colored liquor sales similar to cluster 2 from January 2020
- Cluster 4 has the highest beer sales and second highest colored liquor sales similar to cluster 3 from January 2020

The chart below shows that the similarity between clusters continue more than 2 months.

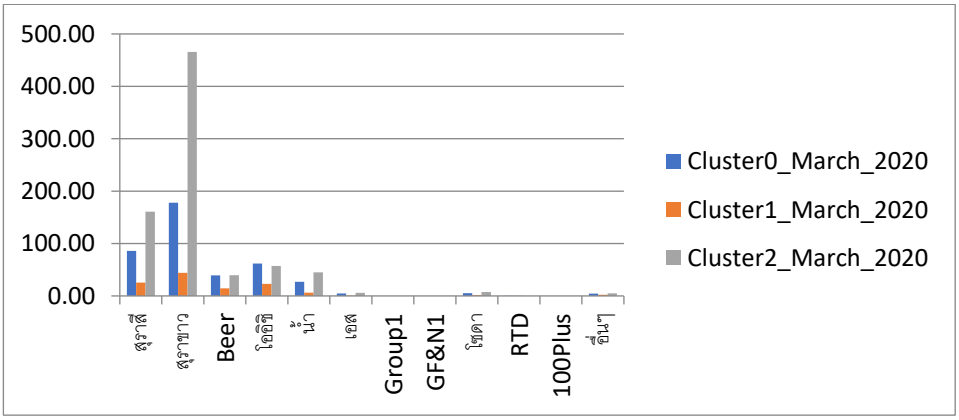


Figure 12: The result of a clustering of March 2020 sales data as a graph with product groups in X axis and centroid of sales value in Y axis.

We can see from figure 12 that cluster 2 has the highest amount of both color and white liquor products bought, this makes it looks similar to cluster 2 from January 2022 and cluster 1 from February 2022.

Figure 10, 11, and 12 shows us the sales of each product type of each cluster during the first three months of 2020.We can see that January and February have similar looking graphs but March has a very different looking graph and clusters. We repeat the previous step with other months until December 2021 so we can get the clusters of each month. We record the resulting cluster for each customer in each month and the sales of each product type in each cluster of each month. As shown in the tables below.

The table below shows the customers changing cluster in each month.

CustomerId	Cluster_Jan_2020	Cluster_Feb_2020	Cluster_Mar_2020	Cluster_Apr_2020	Cluster_May_2020
67552	1	0	0	0	0
67577	4	3	0	3	2
67598	4	0	1	3	2
67608	4	2	1	1	2
67633	0	3	0	0	0
67654	3	4	0	1	1
67686	1	2	1	3	0

Table 5: The monthly cluster of customers in year 2020.

From the raw cluster number obtained from K-means clustering, we can't know how similar or different the clusters are to each other so we decided to compare the centroid of different clusters to see their similarity. We can see clusters in different month have centroids that are close to each other in the table below.

Period	Colored liquor	White liquor	Beer	Oishi	Water	Est	Group1	GF&N1	Soda	RTD	100Plus	Other
Cluster0_January_2020	72.19	190.18	16.77	16.45	5.20	1.91	0.00	0.00	0.59	0.00	0.00	2.41
Cluster1_January_2020	24.95	38.98	9.96	6.70	4.90	0.39	0.00	0.00	0.65	0.00	0.00	0.14
Cluster2_January_2020	223.47	426.23	47.07	52.47	52.83	3.13	0.00	0.00	4.50	1.47	0.00	4.03
Cluster3_January_2020	134.73	124.73	260.60	46.07	54.00	0.53	0.00	0.00	7.87	0.00	0.00	8.80
Cluster4_January_2020	52.14	84.69	26.14	151.40	14.46	3.88	0.00	0.00	1.51	0.06	0.00	0.61
Cluster0_February_2020	64.04	80.65	29.60	170.75	8.88	5.19	0.00	0.00	1.62	0.00	0.00	5.29
Cluster1_February_2020	174.03	420.00	33.95	53.00	29.08	6.38	0.00	0.00	2.92	0.00	0.00	5.05
Cluster2_February_2020	25.98	45.53	10.41	7.71	6.51	1.08	0.00	0.00	0.60	0.05	0.00	1.30
Cluster3_February_2020	79.84	189.05	16.48	14.78	10.42	2.50	0.00	0.00	0.94	0.00	0.00	0.95
Cluster4_February_2020	144.93	158.07	300.71	34.07	36.64	1.14	0.00	0.00	10.29	0.29	0.00	33.00

Table 6: the centroid of sales value for each cluster.

Table 6 shows the centroid of each product type in each monthly cluster. From it we can see that

- Cluster 2 of January 2020 is similar to cluster 1 of February 2020, having high white liquor and colored liquor sales.
- Cluster 3 of January is similar to cluster 4 of February 2020, having high beer sales and the second highest colored liquor sales in their respective months.

The number of clusters for each month can be seen the chart below

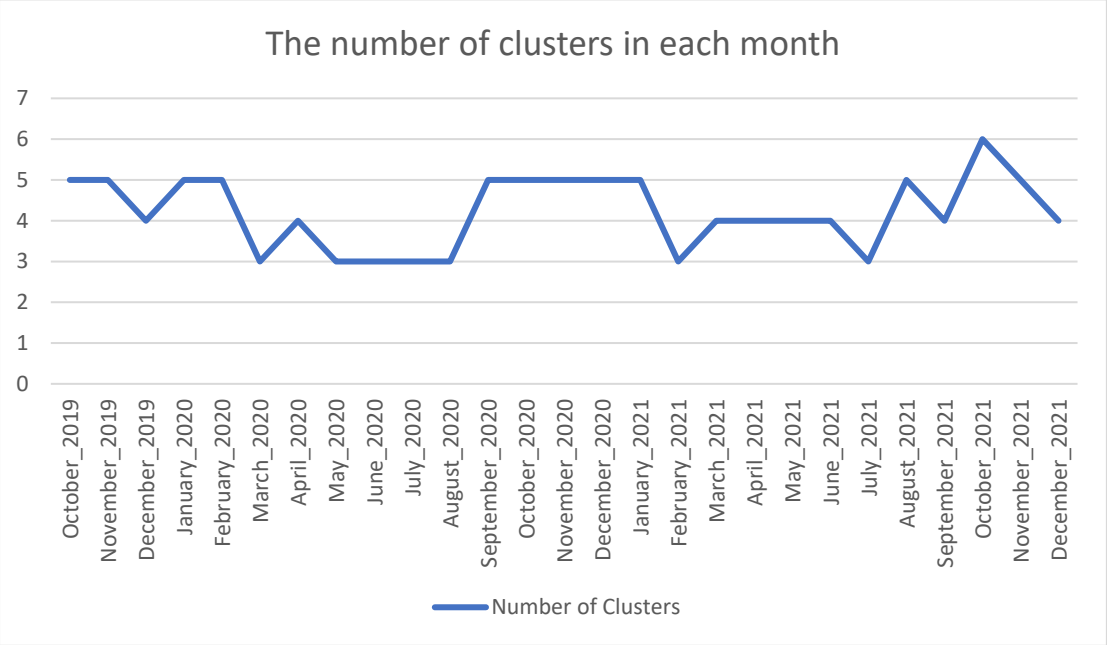


Figure 13 The number of clusters in each month

4.2.2 Cluster Analysis during the pandemic

From the chart, we found that most months have between 4-5 clusters for both the pre-pandemic and post-pandemic periods, with some exceptions. The month with the highest number of clusters is October 2021 with 6 clusters. The period with the lowest number of clusters is between May 2020 to August 2020, where every month has only 3 clusters in that period. The only month with 3 clusters outside of that period are March 2020, February 2021, and July 2021.

Since the lockdown started in April 2020, we can see that there were 5 clusters for most months before the lockdown. The number of clusters went down to 3 and 4 during the start of the lockdown. The number of clusters stabilized at 3 for the first few months of the lockdown, between May 2020 to August 2020. The number of clusters went up to 5 during September 2020 to January 2021. This is similar with the number from 2019 but the number of clusters went down in February instead of March. In the middle of 2021, the number of clusters stabilized at 4

clusters from March to June instead of 3 clusters from May to August like in 2020. This is then followed by the number of clusters changing every month for the rest of the year.

There are 2 notable clusters that disappeared during the early period of the pandemic, the first one is the cluster with very high white liquor sales number. The sales of white liquor declined to around 100-200 and would only begin to break 300 in August 2020. The second one is the cluster with high beer sales, the sales of beer would be below 100 until September 2020.

4.2.2.3 Finding the global clusters

We can't compare the clusters from each month directly, so we propose the term global cluster, global cluster would group clusters with similar centroids together to make comparison between clusters in different month easier.

Now that we have the monthly clusters and their centroid, we can compare the cluster to see which clusters are similar to clusters from other months. We can see from table 7 that some clusters are similar to clusters from another month such as cluster 2 of January 2020 and cluster 1 of February 2020. We will create global clusters so we can group similar clusters together and observe the changes in customer throughout the period. We perform another K-means clustering using the centroid of each cluster from figure.8 in place of raw sales data.

Monthly Cluster	Global Clusters
Cluster0_October_2019	2
Cluster1_October_2019	1
Cluster2_October_2019	0
Cluster3_October_2019	4
Cluster4_October_2019	2
Cluster0_November_2019	1
Cluster1_November_2019	2
Cluster2_November_2019	4
Cluster3_November_2019	0
Cluster4_November_2019	2
Cluster0_December_2019	2
Cluster1_December_2019	2
Cluster2_December_2019	1
Cluster3_December_2019	4

Table 7: Showing how monthly clusters change to global clusters

From table 7, we can see how each monthly cluster fits into global clusters. We can also see that there are months where multiple monthly cluster fits into a single global cluster such as October 2019 where its cluster 0 and cluster 4 are both in global cluster 2.

The resulting global clusters show that the customers can be divided into 5 clusters. When we look at the global clusters of customers during the pre-pandemic time, we can see that there are no customers in global cluster 3. When looking at the customers in global clusters 3, we found that the first member of that cluster only appears during January 2021. Therefore, we can conclude that global cluster 3 is the result of customer behavior changing due to the pandemic and the lockdown. This means that before the pandemic and during the first year of the pandemic, the customers can be split into 4 clusters.

4.2.3 Benefits of using global cluster

We can see in the yellow column, representing monthly clusters, that for some customers. We don't know if the purchasing behavior changes or not, so we can use global cluster to compare them instead. We can see how the changes in cluster for customers as time passes in the tables below.

CustomerId	Cluster_Oct_2019	Global Cluster	Cluster_Nov_2019	Global Cluster	Cluster_Dec_2019	Global Cluster	Cluster_Jan_2020	Global Cluster
67552	0	2	1	2	0	2	1	2
67577	2	0	3	0	1	2	4	0
67598	2	0	3	0	0	2	4	0
67608	2	0	1	2	0	2	4	0
67633	4	2	4	2	1	2	0	2
67654	3	4	2	4	3	4	3	4
67686	0	2	1	2	0	2	1	2

Table 8: Customers with monthly clusters (yellow) and global clusters (blue).

Table 8 shows how each customer fits into monthly cluster and how the monthly clusters fit into global clusters We can see how some customers like 67552 is consistently in global cluster 2 while some customers like 67598 switches between global cluster 0 and 2 depending on the month.

4.2.3 Global Cluster characteristic

We get the centroid of the 5 global clusters as seen in the table below

Global Cluster	Colored liquor	White liquor	Beer	Oishi	Water	Est	Group1	GF&N1	Soda	RTD	100Plus	Other
Cluster-0	98.49	117.82	51.60	249.71	30.66	6.89	0.00	0.00	6.14	0.11	0.00	10.26
Cluster-1	195.31	495.15	64.83	116.90	55.14	7.76	0.00	0.00	7.32	0.27	0.00	15.23
Cluster-2	57.33	119.42	19.49	25.80	14.36	2.42	0.00	0.00	2.65	0.07	0.00	4.45
Cluster-3	197.86	1019.48	121.15	251.06	345.77	9.58	0.00	0.00	3.48	1.13	0.00	11.75
Cluster-4	154.50	159.91	302.97	61.29	62.47	4.83	0.00	0.00	12.74	1.32	0.00	25.86

Table 9: Centroid of each product type in each global cluster

From table 9, we can see that each global cluster centroid has distinct sales amount for each product type as follow

The highest selling product type in cluster-0 is Oishi with around 250 unit sold while colored liquor and white liquor are at around 100 unit sold. There are around 50 units of beer, 30 units of water, 6 units of Est and soda and 10 units of other sold. This cluster has the highest amount of Oishi sold so we can call it the Oishi cluster. There are 25 monthly clusters that fit into this global cluster.

Cluster-1 has very high amount of white liquor sold at almost 500 units and it also has the highest amount of colored liquor sold at around 200 units, nearly double the amount sold in cluster-0, while the number of beers is slightly higher than cluster-0 at around 65 units sold. The number of waters is also almost double of cluster-0 at 55 units but the number of oishi sold is around half of cluster-0 at 117 units. This cluster can be called the liquor cluster due to the high amount of liquor sales. There are 20 monthly clusters that fit into this global cluster.

Cluster-2 has around the same amount of white liquor sold as cluster-0 as 119 units but the amount of colored liquor is almost half at 57 units, it also has very low number of sales for other product groups such as beer at 19 units, oishi at 26 units, and water at 14 units. This cluster can be called the low-sales cluster due to the low number of sales for every product types. There are 54 monthly clusters that fit into this global cluster, making it the most common global cluster.

Cluster-3 has the highest amount of white liquor sold at 1019 units, double of cluster-1, and the highest amount of colored liquor sold at 198 units, shared with cluster-1. It also has the highest amount of water sold at 346 units and around 121 units of beer sold which is almost twice the sale of beer in cluster-1. This cluster has high number of sales in almost every category so it can be called the high-sales cluster. There are 4 monthly clusters that fit into this global cluster, making it the rarest global cluster.

The last cluster, cluster-4, has the highest number of sales for beer at 303 units and the second highest amount of water, colored liquor and white liquor sold at 62 units, 155 units and 160 units respectively. This cluster also has the highest amount of water sold at 25 units and soda at 13 units sold. This cluster can be called the beer cluster due to the high amount of beer sales. There are 11 monthly clusters that fit into this global cluster, making it the second rarest global cluster.

From the table, we can also see that the sales amount for Group1, GF&N1, 100Plus and RTD are either 0 or very close to 0. This means that the customers in this sales team do not buy these products at all. Finally, the type other, soda and Est also sold very poorly in this sales team but their sales number are above 0. This result leads to us only using the first 4 product types for the prediction models.

4.3 Prediction Model

We will create models to predict the sales of products by each customer. The model will use the sales data from customers to predict the number of products the customer will order based on product type. The model will be using 5 months of sales data to predict the sales for the next month.

In this research, we will compare the performance of 3 kinds of models with global cluster training data and sales training data. We will compare the performance of the models trained by the data from the customer's past purchase and the models trained by using other customers from the same global cluster. We will compare the accuracy of the prediction and the time taken to train and run the model. If models with global cluster training data can perform at a similar level to the ones with sales training data, we can save time by preparing models that have been trained with global cluster centroid of the customer and use them while getting results comparable to the models that have to be trained again by the sales data.

4.3.1 Training the model

We will use global cluster data as training data for the prediction model. First, we need to split the data into training data and test data. We use the data from December 2020, the last month, as the test data and the rest of the data as training data for the model. Each row of the data will contain the 5 months of sales as the input(X) and the sixth month as the output(Y). We will be using 2 types of training data, the global cluster training data and sales training data.

Global cluster training data is using the sales data from customers in the same global cluster as the test customer to train the model. In case the customer is in multiple global clusters, we use the cluster they are in for the majority of the times during the last 6 months as

training data. The sales training data means using the past sales of the test customer as training data for the model.

4.3.2 Evaluation

We will be using the data of the last month, December 2021 as the test data for evaluation of the prediction models. We will be evaluating the models by using Root Mean Square Error (RMSE). It can be defined as below

- $\text{RMSE} = \sqrt{\frac{1\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{n}}$
- y_j is the actual value
- $\hat{y}_j$ is the predicted value
- n is the total amount of data

The lower the RMSE means the lower difference between the predicted value and actual value, which means the better performance of the model. However, the difference between sales number may cause comparison of different customers to be difficult. So, we will use Relative RMSE (RRMSE) to compare the performance of different predictions. Relative RMSE can be defined as

- $\text{RRMSE} = \sqrt{\frac{\mathbf{1}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{n(y_j)^2}}$
- y_j is the actual value
- $\hat{y}_j$ is the predicted value
- n is the total amount of data

We will compare the result from models using global cluster as training data to the ones with sales data as training data and compare all 3 model types including LSTM, XGBoost, Random Forest Regression model. We will use the aforementioned models as benchmarks. We will compare the RRMSE of each model type and analyze the reason for the performance.

4.3.3 Building model for global cluster

We will compare the performance of the 3 different kinds of models which are LSTM, XGBoost and Random Forest Regression to see which one works well with global cluster training data. The models that work well with global cluster training data are the ones that has similar performance than the models trained using sales training data or in some cases, better performance. We will also look at the model training time when using sales training data compared to global cluster training data.

5 Result and Analysis

The clustering of monthly clusters resulted in 5 global clusters. We can see that the pre-lockdown data only has 4 clusters compared to the 5 clusters from the post-lockdown data, with global cluster 3 only appearing in 2021, around a year after the lockdown began. We can also see from the pre-lockdown data during the year 2019, the customers in global cluster 0 during October are very likely to move to global cluster 2 by the end of the year while the customers in other global clusters are more likely to stay in the same cluster. The changes in global clusters as time went on can be observed as followed.

In the year 2020, we find that customers in global cluster 0 at the start of the year will move between global cluster 0 and 2 throughout the year, with most customers moving to global cluster 2 during April. Customers in global cluster 1 in January are very likely to stay in the same cluster until April, which they move to global cluster 2. Most of them move back to global cluster 1 in August and global cluster 0 in September before spreading between global cluster 1,2,4 in December. Most customer in global cluster 2 at the start of the year stay there throughout the whole year with minimal movements. Most of the customers in global cluster 4 at the start of the year move to global cluster 2 in March, before moving back to global cluster 4 in April and back to global cluster 2 in May. They then stay in global cluster 2 until September where they move to global cluster 0. They move back to global cluster 4 in October and stay there until the end of the year.

In the year 2021, The customers that start in global cluster 0 move to global cluster 2 in February. They then split into 2 groups in March, with one group staying in global cluster 2 and the other group moving back to global cluster 0. The customers maintain the same behavior until the end of the year, jumping between global cluster 0 and 2. Most of the customers that started the year in global cluster 1 will move to global cluster 2, with a tiny amount moving to global cluster 3. The customers that moved to global cluster 3 will move back to global cluster 1 in April while the customers that moved to global cluster 2 will move between every global cluster each month, most of them will move to global cluster 2 and 3 in March and global cluster 1 and 2 in April. Most of the customers will move back to global cluster 1 in June before moving to global cluster 2 in July and back to global cluster 1 in August. Most of the customers stay in global cluster 1 until November, where most customers move to global cluster 2. Most of the customers

stay in global cluster 2 until the end of the year, with around a third of the customers moving back to global cluster 1 in December. Most of the customer that started the year in global cluster 2 stay there for the whole year, with a few moving between global cluster 0 and 2. The customers that started the year in global cluster 3 stay there until March and move to global cluster 1 in April. They then stay in global cluster 1 until June and move between global cluster 0, 2, 3 until September where they move to global cluster 1 and stay there until the end of the year.

We test the performance of the prediction models by predicting the sales of a single type of product using 5 months of previous data as input. The model uses the sales of a specific product type in all months except the last one as training data to predict the sales of a that product type. We compare the performance of the model with global clusters centroids as training data to the ones with actual sales data of the customer as training data and the model trained with the data from customers in the same global cluster as the test customer. The comparison of the RRMSE can be found in the tables below

Customer	Sales Training RRMSE	Global Cluster Centroid RRMSE	Global Cluster All Customer RRMSE
71866	0.064	0.699	0.816
71872	0.384	0.222	0.664
71909	0.079	0.026	0.580
71910	0.117	0.034	0.730
71911	0.598	0.170	0.062
72010	0.044	0.195	0.270
72022	0.388	0.517	0.528

Table 10: Comparison of RRMSE between the LSTM model with each type of training data for each customer for colored liquor products.

Table 10 shows the result of using LSTM model to predict colored liquor product sales with each type of training data. We can see that the result varies from customer to customer, so we sum up the total error for each type of training data in the table below

Customer	Sales Training RRMSE	Global Cluster Centroid RRMSE	Global Cluster All Customer RRMSE
Total	167.8631	795.1886	195.1786

Table 11: Comparison of total RRMSE between the LSTM model with each type of training data for each customer for colored liquor products.

The result shows that for colored liquor products, the LSTM model performed the best with sales training data, with global cluster customer training data in the second place and global cluster centroid training data providing the worst result.

Next, we use the LSTM model to predict the sales of white liquor products with all 3 types of training data. The comparison of the RRMSE can be found in the tables below

Customer	Sales Training RRMSE	Global Cluster Centroid RRMSE	Global Cluster All Customer RRMSE
71866	0.338	0.013	0.465
71872	0.247	0.326	0.128
71909	0.000	0.000	0.000
71910	2.906	0.374	0.954
71911	0.000	0.000	0.000
72010	0.008	0.065	0.429
72022	0.181	0.162	0.484

Table 12: Comparison of RRMSE between the LSTM model with each type of training data for each customer for white liquor products.

Table 12 shows the result of using LSTM model to predict white liquor product sales with each type of training data. We can see that the result varies from customer to customer, so we sum up the total error for each type of training data in the table below

Customer	Sales Training RRMSE	Global Cluster Centroid RRMSE	Global Cluster All Customer RRMSE
Total	507.471	764.482	384.554

Table 13: Comparison of total RRMSE between the LSTM model with each type of training data for each customer for white liquor products.

The result shows that for white liquor products, the LSTM model performed the best with global cluster customer training data, with sales training data in the second place and global cluster centroid training data providing the worst result.

Next, we use the LSTM model to predict the sales of beer products with all 3 types of training data. The comparison of the RRMSE can be found in the tables below

Customer	Sales Training RRMSE	Global Cluster Centroid RRMSE	Global Cluster All Customer RRMSE
71866	0.362	1.150	0.338
71872	0.444	0.169	0.813
71909	0.000	0.000	0.000
71910	0.066	0.045	0.071
71911	0.000	0.000	0.000
72010	0.000	0.000	0.000
72022	0.000	0.000	0.000

Table 14: Comparison of RRMSE between the LSTM model with each type of training data for each customer for beer products.

Table 14 shows the result of using LSTM model to predict beer product sales with each type of training data. We can see that the result varies from customer to customer, so we sum up the total error for each type of training data in the table below

Customer	Sales Training RRMSE	Global Cluster Centroid RRMSE	Global Cluster All Customer RRMSE
Total	238.792	527.323	254.936

Global Cluster	Global Clusters Training Data LSTM	Self Training Data LSTM	Random Forest
2	0.245337	0.253370227	0.117083
2	8.361145	6.524695587	1.329
2	0.538996	0.681042804	0.1275
2	0.107213	0.032000542	0.098667
2	0.623632	0.084182247	0.089412
2	0.098577	0.201059977	0.06234
1	0.260564	0.897367896	0.274643
0	0.166221	0.672392874	0.271579

1	0.393691	0.105171428	0.018214
2	0.041175	0.043627636	0.082812
2	1.466319	1.116770903	1.185
2	0.112963	0.57988563	0.122727
2	0.725777	0.495100212	0.176429
2	0.227056	0.403122614	0.068095
2	0.457409	0.151580811	0.175
2	4.500155	2.989559937	1.2
2	0.18557	0.32465299	0.050278
1	0.282837	0.729153035	0.143529
2	0.377874	0.322909219	0.151486
2	0.346698	0.159477303	0.041728
	19.51921	16.76712387	5.785524

Table 15: Comparison of total RRMSE between the LSTM model with each type of training data for each customer for beer products.

The result shows that for beer products, the LSTM model performed the best with sales training data, with global cluster customer training data in the second place and global cluster centroid training data providing the worst result.

Next, we use the LSTM model to predict the sales of beer products with all 3 types of training data. The comparison of the RRMSE can be found in the tables below

Customer	Sales Training RRMSE	Global Cluster Centroid RRMSE	Global Cluster All Customer RRMSE
71866	0.362	1.150	0.338
71872	0.444	0.169	0.813

71909	0.000	0.000	0.000
71910	0.066	0.045	0.071
71911	0.000	0.000	0.000
72010	0.000	0.000	0.000
72022	0.000	0.000	0.000

Table 16: Comparison of RRMSE between the LSTM model with each type of training data for each customer for Oishi products

Table 16 shows the result of using LSTM model to predict Oishi product sales with each type of training data. We can see that the result varies from customer to customer, so we sum up the total error for each type of training data in the table below

Customer	Sales Training RRMSE	Global Cluster Centroid RRMSE	Global Cluster All Customer RRMSE
Total	242.816	475.278	185.264

Table 17: Comparison of total RRMSE between the LSTM model with each type of training data for each customer for Oishi products.

The result shows that for Oishi products, the LSTM model performed the best with global cluster customer training data, with sales training data in the second place and global cluster centroid training data providing the worst result.

We can summarize the test result of the LSTM model in the table below

Product Type	Sales Training RRMSE	Global Cluster Centroid RRMSE	Global Cluster Customer RRMSE
Colored Liquor	322.652	700.611	370.37
White Liquor	507.471	764.482	384.554
Beer	238.792	527.323	254.936
Oishi	242.816	475.278	185.264

Table 18: The performance of LSTM model with each type or training data.

We can see that for the LSTM model, the sales training data provided the best result for colored liquor and beer products while the global cluster customer training data provided the best results for the white liquor and Oishi products.

We then test the performance of the XGBoost model with all 3 training data types and the 4 product types. The result for colored liquor products can be seen in the table below

Customer	Sales Training RRMSE	Global Cluster Centroid RRMSE	Global Cluster All Customer RRMSE
71866	0.262	1.166	0.178
71872	0.271	0.200	0.161
71909	0.004	0.163	0.070
71910	0.001	0.086	0.017
71911	0.023	0.046	0.008
72010	0.006	0.329	0.101
72022	0.071	0.500	0.085

Table 19: Comparison of RRMSE between the XGBoost model with each type of training data for each customer for colored liquor products

Table 19 shows the result of using XGBoost model to predict colored liquor product sales with each type of training data. We can see that the result varies from customer to customer, so we sum up the total error for each type of training data in the table below

Customer	Sales Training RRMSE	Global Cluster Centroid RRMSE	Global Cluster All Customer RRMSE
Total	167.863	795.189	195.179

Table 20: Comparison of total RRMSE between the XGBoost model with each type of training data for each customer for colored liquor products

The result shows that for colored liquor products, the XGBoost model performed the best with sales training data, with global cluster customer training data in the second place and global cluster centroid training data providing the worst result.

Next, we use the XGBoost model to predict the sales of white liquor products with all 3 types of training data. The comparison of the RRMSE can be found in the tables below

Customer	Sales Training RRMSE	Global Cluster Centroid RRMSE	Global Cluster All Customer RRMSE
71866	71866	0.022	0.092
71872	71872	0.046	0.761
71909	71909	0.000	0.000
71910	71910	0.011	0.217
71911	71911	0.010	0.000
72010	72010	0.027	0.330
72022	72022	0.013	0.388

Table 21: Comparison of RRMSE between the XGBoost model with each type of training data for each customer for white liquor products.

Table 21 shows the result of using XGBoost model to predict white liquor product sales with each type of training data. We can see that the result varies from customer to customer, so we sum up the total error for each type of training data in the table below.

Customer	Sales Training RRMSE	Global Cluster Centroid RRMSE	Global Cluster All Customer RRMSE
Total	131.927	1776.584	153.260

Table 22: Comparison of total RRMSE between the XGBoost model with each type of training data for each customer for white liquor products.

The result shows that for white liquor products, the XGBoost model performed the best with sales training data, with global cluster customer training data in the second place and global cluster centroid training data providing the worst result.

Next, we use the XGBoost model to predict the sales of beer products with all 3 types of training data. The comparison of the RRMSE can be found in the tables below

Customer	Sales Training RRMSE	Global Cluster Centroid RRMSE	Global Cluster All Customer RRMSE
71866	0.242	3.059	1.048
71872	0.075	0.658	0.334

71909	0.000	0.000	0.000
71910	0.003	0.019	0.000
71911	0.000	0.000	0.000
72010	0.000	0.000	0.000
72022	0.000	0.000	0.000

Table 23: Comparison of RRMSE between the XGBoost model with each type of training data for each customer for beer products.

Table 23 shows the result of using XGBoost model to predict beer product sales with each type of training data. We can see that the result varies from customer to customer, so we sum up the total error for each type of training data in the table below.

Customer	Sales Training RRMSE	Global Cluster Centroid RRMSE	Global Cluster All Customer RRMSE
Total	142.653	1223.482	159.064

Table 24: Comparison of total RRMSE between the XGBoost model with each type of training data for each customer for beer products.

The result shows that for beer products, the XGBoost model performed the best with sales training data, with global cluster customer training data in the second place and global cluster centroid training data providing the worst result.

Next, we use the XGBoost model to predict the sales of Oishi products with all 3 types of training data. The comparison of the RRMSE can be found in the tables below

Customer	Sales Training RRMSE	Global Cluster Centroid RRMSE	Global Cluster All Customer RRMSE
71866	0.000	0.000	0.000
71872	0.185	5.429	0.076
71909	0.000	0.000	0.000

71910	0.006	0.420	0.037
71911	0.008	0.399	0.003
72010	0.000	0.000	0.000
72022	0.000	0.000	0.000

Table 25: Comparison of RRMSE between the XGBoost model with each type of training data for each customer for Oishi products.

Table 25 shows the result of using XGBoost model to predict Oishi product sales with each type of training data. We can see that the result varies from customer to customer, so we sum up the total error for each type of training data in the table below.

Customer	Sales Training RRMSE	Global Cluster Centroid RRMSE	Global Cluster All Customer RRMSE
Total	97.954	1013.283	79.148

Table 26: Comparison of RMSE between the LSTM model with Random Forest Regression and XGBoost for Oishi products using global cluster centroid as training data.

The result shows that for Oishi products, the XGBoost model performed the best with global cluster customer training data, with sales training data in the second place and global cluster centroid training data providing the worst result.

We can summarize the test result of the XGBoost model in the table below

Product Type	Sales Training RRMSE	Global Cluster Centroid RRMSE	Global Cluster Customer RRMSE
Colored Liquor	167.863	795.189	195.179
White Liquor	131.927	1776.584	153.26
Beer	142.653	1223.482	159.064
Oishi	97.954	1013.283	79.148

Table 27: The performance of XGBoost model with each type or training data.

We can see that for the XGBoost model, the global cluster customer training data provided the best result for only Oishi products while sales training data provided the best result for all other types of products.

We then test the performance of the Random Forest model with all 3 training data types and the 4 product types. The result for colored liquor products can be seen in the table below

Customer	Sales Training RRMSE	Global Cluster Centroid RRMSE	Global Cluster All Customer RRMSE
71866	0.469	1.035	0.119
71872	0.226	0.377	0.214
71909	0.025	0.263	0.007
71910	0.075	0.040	0.171
71911	0.273	0.060	0.185
72010	0.039	0.371	0.049
72022	0.056	0.803	0.232

Table 28: Comparison of RRMSE between the Random Forest model with each type of training data for each customer for colored liquor products.

Table 28 shows the result of using Random Forest model to predict colored liquor product sales with each type of training data. We can see that the result varies from customer to customer, so we sum up the total error for each type of training data in the table below.

Customer	Sales Training RRMSE	Global Cluster Centroid RRMSE	Global Cluster All Customer RRMSE
Total	153.944	1063.369	163.409

Table 29: Comparison of total RRMSE between the Random Forest model with each type of training data for each customer for colored liquor products

The result shows that for colored liquor products, the Random Forest model performed the best with sales training data, with global cluster customer training data in the second place and global cluster centroid training data providing the worst result.

Next, we use the Random Forest model to predict the sales of white liquor products with all 3 types of training data. The comparison of the RRMSE can be found in the tables below

Customer	Sales Training RRMSE	Global Cluster Centroid RRMSE	Global Cluster All Customer RRMSE
71866	0.178	0.000	0.144
71872	0.175	0.544	0.075
71909	0.000	0.000	0.000
71910	0.707	0.253	0.404
71911	0.170	0.000	0.000
72010	0.095	0.083	0.140
72022	0.062	0.259	0.180

Table 30: Comparison of RRMSE between the Random Forest model with each type of training data for each customer for white liquor products.

Table 30 shows the result of using Random Forest model to predict colored liquor product sales with each type of training data. We can see that the result varies from customer to customer, so we sum up the total error for each type of training data in the table below.

Customer	Sales Training RRMSE	Global Cluster Centroid RRMSE	Global Cluster All Customer RRMSE
Total	211.081	1159.224	153.389

Table 31: Comparison of total RRMSE between the Random Forest model with each type of training data for each customer for white liquor products.

The result shows that for white liquor products, the Random Forest model performed the best with global cluster customer training data, with sales training data in the second place and global cluster centroid training data providing the worst result.

Next, we use the Random Forest model to predict the sales of beer products with all 3 types of training data. The comparison of the RRMSE can be found in the tables below

Customer	Sales Training RRMSE	Global Cluster Centroid RRMSE	Global Cluster All Customer RRMSE
71866	0.245	2.305	0.193
71872	0.247	0.469	0.350

71909	0.000	0.000	0.000
71910	0.000	0.062	0.013
71911	0.000	0.000	0.000
72010	0.000	0.000	0.000
72022	0.000	0.000	0.000

Table 32: Comparison of RRMSE between the Random Forest model with each type of training data for each customer for beer products.

Table 32 shows the result of using Random Forest model to predict beer product sales with each type of training data. We can see that the result varies from customer to customer, so we sum up the total error for each type of training data in the table below.

Customer	Sales Training RRMSE	Global Cluster Centroid RRMSE	Global Cluster All Customer RRMSE
Total	113.242	946.070	107.879

Table 33: Comparison of total RRMSE between the Random Forest model with each type of training data for each customer for beer products.

The result shows that for beer products, the Random Forest model performed the best with global cluster customer training data, with sales training data in the second place and global cluster centroid training data providing the worst result.

Next, we use the Random Forest model to predict the sales of Oishi products with all 3 types of training data. The comparison of the RRMSE can be found in the tables below

Customer	Sales Training RRMSE	Global Cluster Centroid RRMSE	Global Cluster All Customer RRMSE
71866	0.000	0.000	0.000
71872	2.123	3.524	1.455
71909	0.000	0.000	0.000

71910	0.149	0.654	0.154
71911	0.210	0.590	0.130
72010	0.000	0.000	0.000
72022	0.000	0.000	0.000

Table 34: Comparison of RRMSE between the Random Forest model with each type of training data for each customer for Oishi products.

Table 34 shows the result of using Random Forest model to predict Oishi product sales with each type of training data. We can see that the result varies from customer to customer, so we sum up the total error for each type of training data in the table below.

Customer	Sales Training RRMSE	Global Cluster Centroid RRMSE	Global Cluster All Customer RRMSE
Total	118.950	821.423	94.453

Table 35: Comparison of total RRMSE between the Random Forest model with each type of training data for each customer for Oishi products.

The result shows that for Oishi products, the Random Forest model performed the best with global cluster customer training data, with sales training data in the second place and global cluster centroid training data providing the worst result.

We can summarize the test result of the Random Forest model in the table below

Product Type	Sales Training RRMSE	Global Cluster Centroid RRMSE	Global Cluster Customer RRMSE
Colored Liquor	153.944	1063.369	163.409
White Liquor	211.081	1159.224	153.389
Beer	113.242	946.07	107.879
Oishi	118.95	821.423	94.453

Table 36: The performance of Random Forest model with each type or training data.

We can see that for the Random Forest model, the sales training data provided the best result for only colored liquor products while global cluster customer training data provided the best result for all other types of products.

Next, we measure the time taken to train and run the model for each type of training data and prediction models. The result can be seen in the table below

Model	Sales training data	Global Cluster Centroid	Global Cluster Customer
LSTM	99 minutes 16 seconds	3 minutes 20 seconds	13 minutes 34 seconds
XGBoost	1 minute 5 seconds	20 seconds	30 seconds
Random Forest	1 minutes 20second	15 seconds	33 seconds

Table 37: The comparison of time taken to train and run the model for each kind of models and training data

From table 37, we can see that the sales training data took the longest time to train and run the mode, due to training the model once per customer. The global cluster centroid training data was the fastest one due to using the least amount of data and only needing to train once per global cluster. Finally, the global cluster customer training data took longer than the global cluster centroid training data due to amount of data used in each cluster.

6 Summary

The test result has shown that the global cluster customer training data performed the best for LSTM model with white liquor and Oishi products, XGBoost model with Oishi products and Random Forest model with white liquor, beer and Oishi products. The sales training data performed the best for LSTM model with colored liquor and beer products, XGBoost model with colored liquor, white liquor and beer products and Random Forest model with colored liquor

products. The global cluster centroid training data performed the worst for every kind of products and models.

We can see that out of the 12 tests,3 kinds of models and 4 types of products, we can see that the global cluster customer training data provided the best result for 6 of them. Sales training data worked the best for the other 6. We can also see from the time comparison that the global cluster centroid and global cluster customer training data took much less time to train and run the model than the sales training data. This is due to the model needing to train once for every customer when using sales training data while for the global cluster training data, we can train the model once and use it to predict the sales for many customers in the same global cluster. We conclude that the global cluster customer training data is the best out of the 3 types of training data due to performing similarly to sales training data while taking much less time to train and run the model.

REFERENCES

1. Li, Y. and H. Wu, *A Clustering Method Based on K-Means Algorithm.* Physics Procedia, 2012. **25**: p. 1104-1109.
2. Marutho, D., et al. *The Determination of Cluster Number at k-Mean Using Elbow Method and Purity Evaluation on Headline News*. in *2018 International Seminar on Application for Technology of Information and Communication*. 2018.
3. Helmini, S., et al., *Sales forecasting using multivariate long short term memory network models.* PeerJ PrePrints, 2019. **7**: p. e27712v1.
4. Pliszczuk, D., et al., *Forecasting Sales in the Supply Chain Based on the LSTM Network: The Case of Furniture Industry.* European Research Studies, 2021. **24**: p. 627-636.
5. Chen, T. and C. Guestrin, *XGBoost: A Scalable Tree Boosting System*, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, Association for Computing Machinery: San Francisco, California, USA. p. 785–794.
6. Ibrahem Ahmed Osman, A., et al., *Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia.* Ain Shams Engineering Journal, 2021. **12**(2): p. 1545-1556.
7. Ali, J., et al., *Random Forests and Decision Trees.* International Journal of Computer Science Issues(IJCSI), 2012. **9**.
8. Biau, G., *Analysis of a random forests model.* J. Mach. Learn. Res., 2012. **13**(null): p. 1063–1095.
9. Nayyar, T. and A. Townsend, *Analysing Customer Buying Behavior.* Dissertation, 2019.
10. Raorane, A. and R. Kulkarni, *Data mining techniques: A source for consumer behavior analysis.* arXiv preprint arXiv:1109.1202, 2011.
11. Lipyanina, H., et al. *Decision tree based targeting model of customer interaction with business page*. in *CMIS*. 2020.

VITA

NAME	Krittayot Bherngjitt
DATE OF BIRTH	20 September 1993
PLACE OF BIRTH	Thailand
INSTITUTIONS ATTENDED	Saint Gabriel's College Chulalongkorn University
HOME ADDRESS	242 Samsen 13 Samsen Road Vajira Hospital Dusit Bangkok 10300