Asian Institute of Technology
School of Engineering and Technology

AT82.04 Business Intelligence and Analytics

Instructor:
Dr.Vatcharaporn Esichaikul

Project Report on
Prediction of monthly sales and customer segmentation based on real-time data

Submitted by:
Group7
Kunlanit Korsamphan (st121422)
Shin Thant (st121493)
Phway Thant Thant Soe Lin (st121494)
Vineela Mukkamala (st122050)

Submitted Date:
29th April 2021

ACKNOWLEDGEMENT

It gives us immense pleasure to express gratitude to our respected instructor Dr.Vatcharaporn Esichaikul, for her guidance, encouragement and enduring support during the course of this endeavor. Her technical knowledge has been a valuable asset, but more importantly madam has helped to keep us focused and pointed in the right direction throughout this course.

The completion of this project would not be possible, had there not been immense support from our faculty members. We owe a lot to them.

Our sincere thanks to our colleagues at our classroom for mutual learning and for the help they extended to us in different ways while we were undertaking the painful part of collecting information and designing models in course of preparing this assignment.

We also thank our parents for their blessings which made us stand despite many obstacles and successfully complete the project.

It has been an honor taking this course this semester and we appreciate all the knowledge that you best bestowed upon us.

Table of Contents

1. Background

Our project is proposed to develop a prediction model for monthly sales and customer segmentation using real-time data. The dataset is obtained from the actual business which is located in Myitkyina, Myanmar. This business is the distribution of personal care and home care products of a specific brand and it is the only distribution in Myitkyina. The distributor has seven salesmen and each salesman sells the products to retailers. As the business grows eventually, the distributor faces some difficulties in handling the data. Moreover, it is necessary to analyze the sales to unlock commercially relevant insights, increasing revenue and profitability, and improving brand perception. Customer segmentation is also important to perform because it will allow the distributor to better understand customers, products they are buying and keep these customers engaged with the business. Predicting the monthly sales of the business helps in business planning, budgeting, and making better business decisions. On the other hand, customer segmentation can also help the distributor to make better marketing strategies, gain competition over rival companies, and get a better knowledge of customers' needs and wants.

2. Objectives

The core objectives of the proposed system are as follows:

- To create a platform with real-time analytical dashboards for a business in Myanmar.
- To analyze the customers purchasing behavior from monthly sales reports based on the frequency rate of their purchase.
- To predict the next month's sales value from past sales reports.
- To make customer segmentation using RFM analysis

3. Scope

We are going to create models for predicting the monthly sales and segmenting customers based on their purchase history. And the model with the best performance will be selected to use for our system.

3.1. Dataset

Our dataset is a real-time dataset (in CSV) that includes data about the B2B system. B2B is a system in which the distributor distributes the products to salesmen and sales persons are supposed to sell them to the retailers or customers. Sample attributes of the monthly sales reports used for the project are ID and name for respective products, customers, and salesmen and also product category and sales price. Some attributes and values are listed in table 1.

Table 1. Attributes of the sales report dataset

Attribute	Description	Example Value
DOC_DATE3	Date of sale	2/12/2020
DOC_NO3	Document no.	CM-200002034820
DSR_NAME2	Salesman Name	Ko Thet Swe Aung
POPCODE3	Customer code	K100121721720105234

POP_NAME3	Customer name	Ko Maung Lin
POP_ADDRESS2	Customer address	NO(96)/Myot Ma Myot MaMyoe Ma
OUTLET_TYPE2	Outlet Type	SS RETAILER
SUB_OUTLET_TYPE2	Sub outlet type	GT BEAUTY RETAILER
PERFECT_TYPE2	Member Type	SILVER
SKU	Product code	67610231
SKU_DESC3	Product name	E-LAN POWDER ULTRA 100X45G
CATEGORY_DESC	Product category	FABRIC CLEANING
SALE_PRICE	Price for each package	28000
SELL_FACTOR2	Number of items per package	8
Textbox179	Amount of sold packages	15
QTY3	Amount of sold pieces	0
TOT_SALE_PCS	Change unit of pieces to packages (Textbox179 + (QTY3/SELL_FACTOR2))	15
SALE_VALUE1	SALE_PRICE * TOT_SALE_PCS	420000

We are considering data from June 2018 - December 2020, in order to do descriptive analytics and predictive analytics.

3.2. Model

This project is proposed to work on predictive analytics on monthly sales and segmentation analytics on customer segmentation. For the predictive model, LSTM is used to predict the sales of each month according to the data from June 2018- December 2020. For the customer segmentation purpose, we used RFM clustering using k means clustering.

4. Platform and Tools Used

4.1 Python

Python is an interpreted high-level programming language which would help in interpreting many programming paradigms, including object-oriented, imperative, functional and procedural and different libraries. We use python to develop machine

learning models and visualize some data in order to design the dashboard.

4.2 Jupyter Notebook

Jupyter Notebook, an open-source web application that allows us to create and share documents that contains live code, visualizations and narrative text was one of the tools used to build this system. Jupyter Notebook is used for data cleaning and transformation, statistical modeling, machine learning, visualization, simulation, etc.

4.3 Python Libraries

Python libraries were used for prediction and data visualization. We use seaborn, matplotlib for visualization and sklearn for implementing machine learning algorithms. We also use pandas and flask to create a backend system for our website.

4.4 ReactJS

React is an open-source, front end, JavaScript library for building user interfaces or UI components. We use ReactJS to create the front end of our website together with a fusion chart library for visualization.

4.5 MongoDB

MongoDB is a source-available cross-platform document-oriented database program. Classified as a NoSQL database program, MongoDB uses JSON-like documents with optional schemas. We use MongoDB as our project database. The backend system will pull data from the database, aggregate the data as the dashboard needs and visualize them.

5. Methodology

5.1 Data Cleaning

The initial datasource was transaction data from a business in Myanmar. The time period of the data is June 2018 to December 2020. First, the business process is needed to understand to determine the importance of each field in the transaction. The customers can order several products in one order, therefore there were many records with duplicate transaction numbers. Customers can order the product with the whole package or purchase only some pieces in the package. The price per package and price per piece is needed to be calculated based on the transaction data in order to create the Product table.

The values of product category, product brand, customer outlet type, customer sub outlet type and member type are constants. They are fixed value and are not expected to change but they were needed to clean up. To clean up these values, first we changed these values to be all lower case. Then, we replace the symbol and space with underscore. Next, we need to ensure that every value does not contain misspelling. For Member Type value, we convert the original text value into ordinal value in order to rank it later and it was easy for comparison as well. The table belows shows the changes on these constants for data cleaning.

Table 2: Data Cleaning on Constant

Original Value	Cleaned up Value
Product Category	
FABRIC CLEANING	fabric_cleaning
FABRIC SENSATIONS CATEGORY	fabric_sensations
HAIR CARE	hair_care
HOME & HYGIENE	home_and_hygiene
ORAL CARE	oral_care
SKIN CLEANSING	skin_cleansing
FABRIC ENHANCERS	fabric_enhancers
Product Brand	
3D	3d
DOMI	domi
E.CO	e_co
E-LAN	e_lan
FAMILY CARE	family_care
FINO	fino
MISSS	misss
O2	o2
ROSE	rose
WIN	win
MR CARE	mr_care
BON (HC&BPC)	bon

Outlet Type	
SS WHOLESELLER	wholesaler
SS RETAILER	retailer
DISTRIBUTOR	distributor
Sub Outlet Type	
FAMILY GROCERY RETAILER	family_grocery_retailer
FAMILY GROCERY WHOLE SELLER	family_grocery_wholesaler
GT BEAUTY RETAILER	gt_beauty_retailer
GT BEAUTY WHOLESALER	gt_beauty_wholesaler
MOM & POPS RETAILER	mom_and_pops_retailer
MOM & POPS WHOLESALER	mom_and_pops_wholesaler
MS FOOD RETAILER	ms_food_retailer
MS FOOD WHOLESALER	ms_food_wholesaler
MS HPC RETAILER	ms_hpc_retailer
MS HPC WHOLESELER	ms_hpc_wholesaler
CANVASSER	canvasser
Member Type	
BRONZE	1
DIAMOND	5
GOLD	3
PLATINUM	4
SILVER	2
NA	0

After cleaning up the values of all transactions and calculating the sale price of the product, the data will be extracted to generate customer data, product data, and salesman data. Transactions that contain product quantity as 0 or total sale price as 0 will be ignored.

Generating customer data from transactions, we tried to determine the registered date of the customers by defining the first date of transaction as registered date. We can monitor

the growth rate of customers later with this piece of data. Also, the changes in member type will be stored as logs as well. With this data, we hoped we could extract some useful information from it, but there were not many changes in member type.

After the data extraction, the number of customer records, product records and salesman records are displayed in the following table.

Table 3: Total records of each table

Table	Total Records
Transaction	159,682
Customer	894
Product	309
Salesman	7

5.2 Descriptive Analysis

The visualizations are made with tableau and python (using matplotlib and seaborn libraries) for analysis and understanding data. This helps with dashboard design on the website.

5.2.1 Area graph for sales amount with respect to years.

This area graph depicts the sales amount in different years 2019 and 2020. We can see that there is a low peak in 2020 during the months february to may due to lockdown.

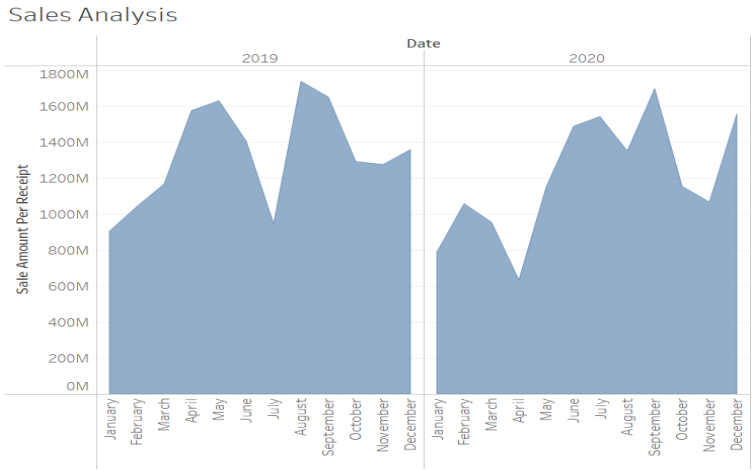

Figure 1: Area graph for sales amount with respect to years

5.2.2 Analysis of salesman performance

This graph shows the sales made by each sales man and discount given by him for the customers. We can see that sales man with id SO3 has made more sales comparative to others.

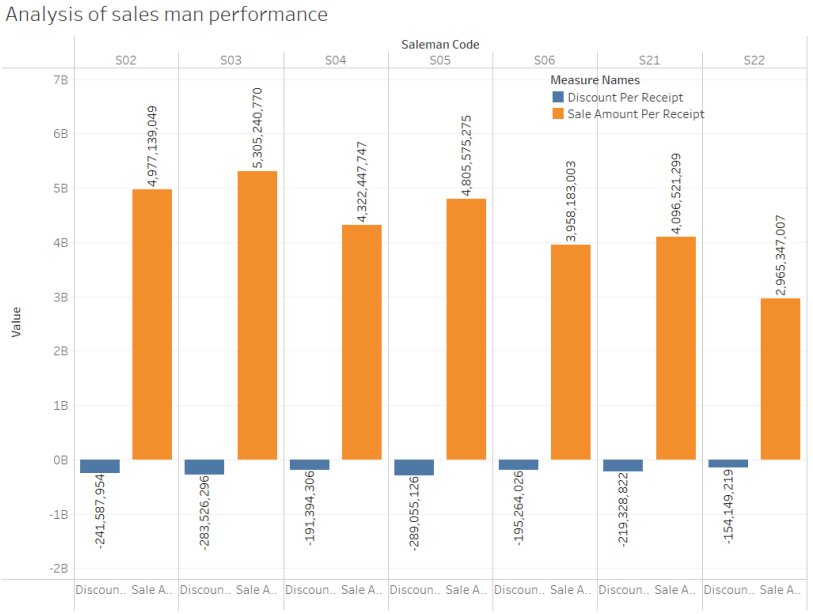

Figure 2: Bar graph for sales amount and discount for sales man code

5.2.3 Sales amount per month vs Discount per month

This graph shows the relationship between the sales amount and the discount for the years 2020 and 2019.

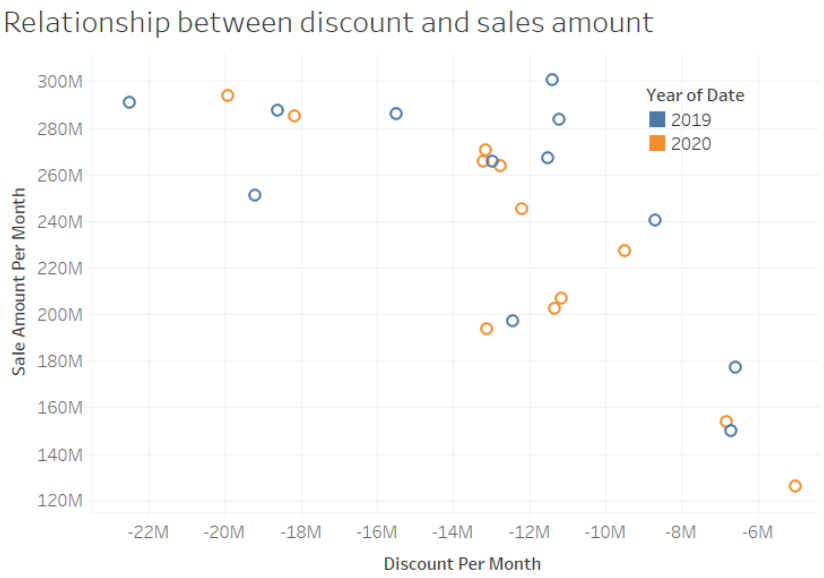

Figure 3: sales amount per month vs Discount per month

5.2.4 Sales by outlet type

This bar graph shows us the sales amount made by different outlets in the years 2019 and 2020.

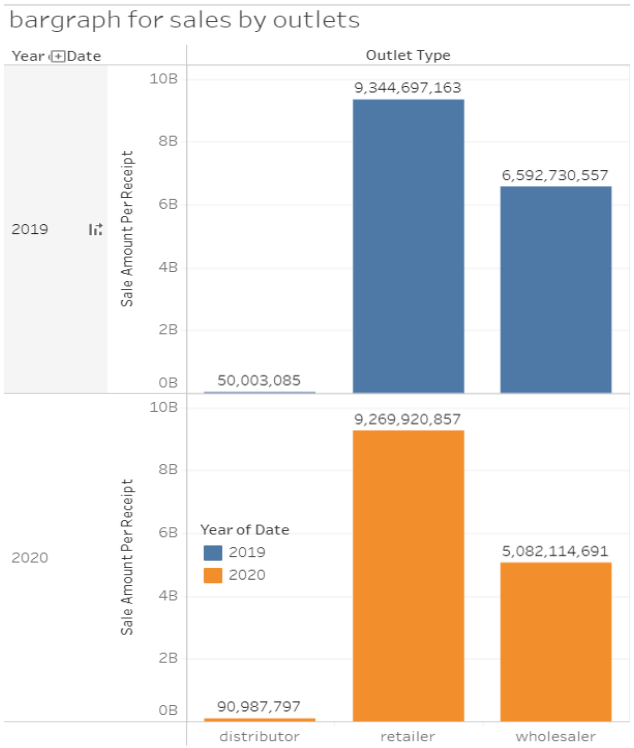

Figure 4: Bar Graph for sales by outlet type

5.2.5 Bar graph for sales amount per product category

This bar graph depicts the sales amount made by each product category for the years 2019 and 2020.

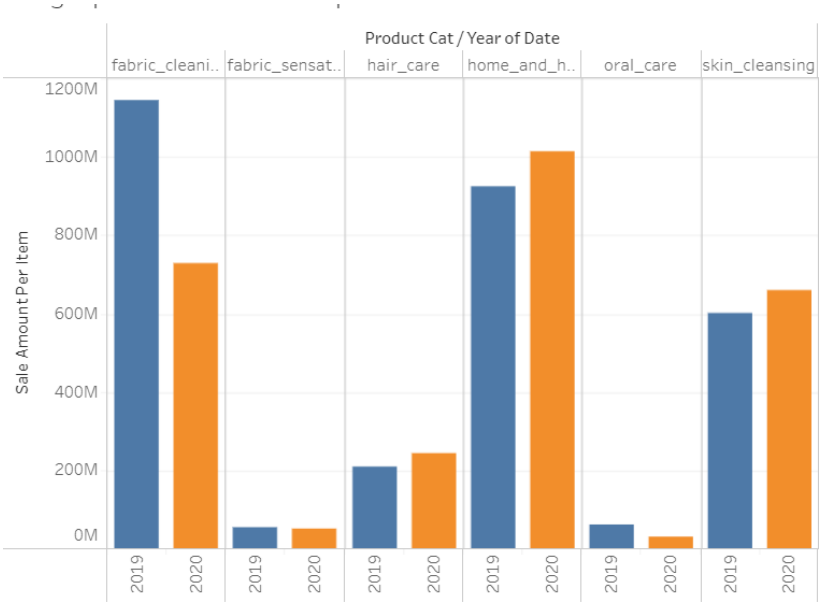

Figure 5: Bar graph for sales amount for product category

5.2.6 Grading of each product depending on the sales amount

The below packed bubbles show which products are sold more and which product category products are sold more. We can see that the home and hygiene products category as well as fabric cleaning category products contribute more to the sales amount.

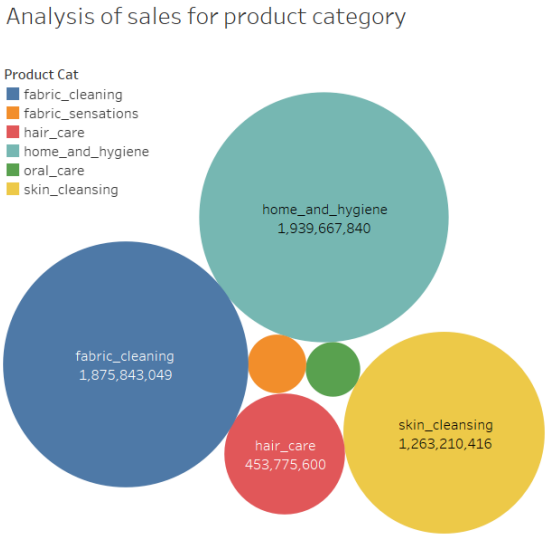

Figure 6: visualizing with respect to product category and sales

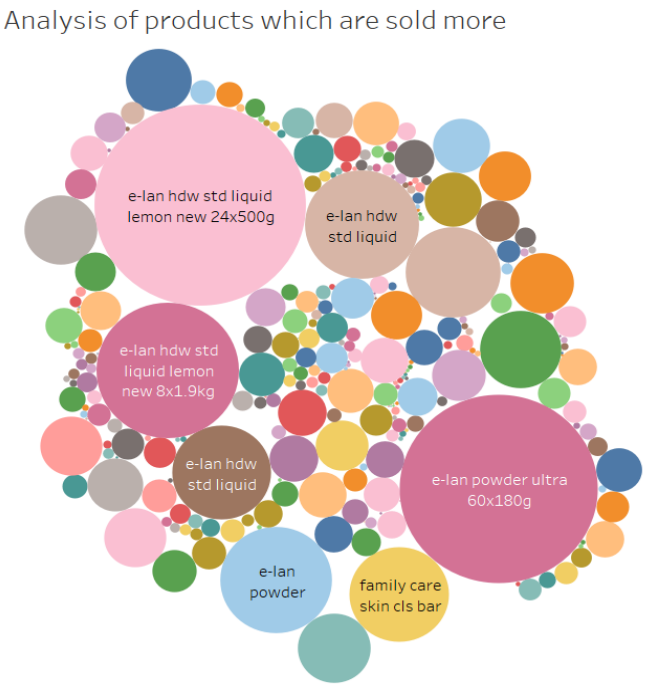

Figure 7: visualizing products with respect to their sales amount.

5.2.7 Analysis of sales amount by month

This tree graph shows us the month in which more amount has been received.

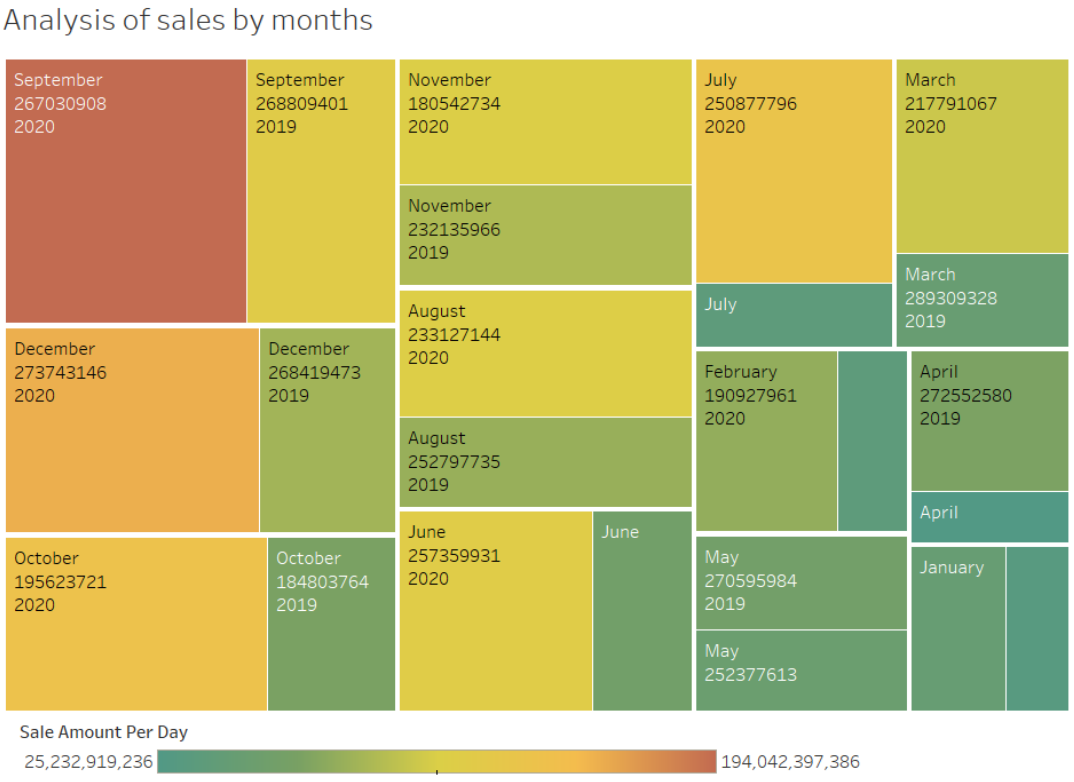

Figure 8: Tree Map for sales amount per month

5.3 Sales Prediction

Prediction of sales over time is a very important analysis on a retail system. There are many methods in the literature to achieve time series forecasting including Long Short-term Memory (LSTM), Autoregressive Integrated Moving Average (ARIMA), Seasonal Autoregressive Integrated Moving-Average (SARIMA), Vector Autoregression (VAR), and so on. In our system, we focus on the LSTM method, which is a quite popular Deep Learning model.

Since we have only two years data, prediction for overall sales may not give a good result. Since our data is obtained from an actual retail system, the sales trend is affected by the pandemic (Covid-19). For these reasons, we decided to predict sales over specific products. We focus on three best selling products under the 'O2' product brand. We analyze individual product trends to predict each product's sales respectively.

5.3.1 Sales Prediction using Long Short-term Memory (LSTM)

Time Series Sales Prediction is performed using LSTM with PyTorch in Python. Advanced deep learning models such as Long Short Term Memory Networks (LSTM), are capable of capturing patterns in the time series data, and therefore can be used to make predictions regarding the future trend of the data. LSTM is one of the most widely used algorithms to solve sequence problems. We will be using the PyTorch library, which is one of the most commonly used Python libraries for deep learning to perform time series analysis using LSTM.

a) Dataset Preparation

We extract monthly sales records for each specific product from our core dataset. The extracted dataset has two columns: date, and sales_amount_per_item. The sales_amount_per_item column contains the total amount of sales per item in a specified month.

b) Data Preprocessing

Splitting of train and test data: Since LSTM is a classification model, we split our dataset into 80 percent for training and 20 percent for testing. The data from the first 20 months will be used to train our LSTM model, whereas the model performance will be evaluated using the values from the last 4 months. In contrast, we predicted the amount of sales in the last 4 months based on the first 20 months.

Normalization: It is very important to normalize the data for time series

predictions. We performed min/max scaling on the dataset which normalizes the data within a range of -1 and 1.

Converting training data into sequences and corresponding labels: The final preprocessing step is to convert our training data into sequences and corresponding labels. Since we have only two years data, we set the input sequence length for training to 4. Next, we defined a function that accepts the raw input data and will return a list of pair data. In each pair, the first element will contain a list of 4 items corresponding to the amount of sales for the specific product in 4 months, the second tuple element will contain one item i.e. the amount of sales in 4+1st month.

c) Training and Testing the LSTM Model

We have pre-processed the data, now is the time to train our model.We use input_size (number of features) as 1. Though our sequence length is 4, for each month we have only 1 value. We define one hidden layer of 10 neurons. Since we want to predict the amount of sales for 1 month in the future, the output size will be 1. We will use the cross entropy loss. For the optimizer function, we will use the adam optimizer. We will train our model for 150 epochs. Since we normalized the dataset for training, the predicted values are also normalized. We need to convert the normalized predicted values into actual predicted values. We can do so by passing the normalized values to the inverse_transform method of the min/max scaler object that we used to normalize our dataset. We illustrated the predicted sales for the last four months of 2020 (orange line) for three top products of the 'O2' brand in figure 9. The last four months are affected by pandemic so the trend is almost dropping to zero. So, the prediction also gives the wrong result. We can be claimed from this that the data we used is not a good one to do the sales prediction because of the unstable trend during the pandemic.

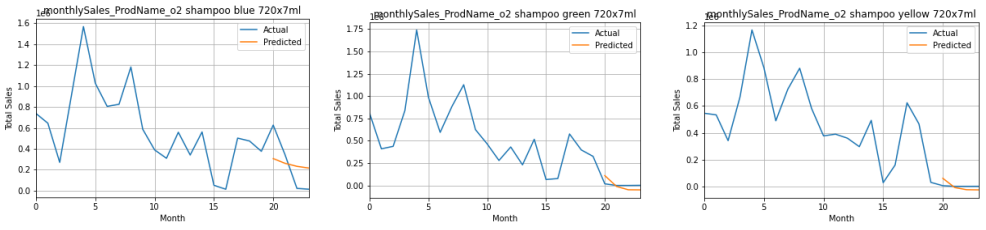

Figure 9: Illustration of the trend of sales (blue line) together with predicted sales result for last four months of 2020 (orange line) for three top products of 'O2' brand

5.4 Customer Segmentation

5.4.1 RFM Clustering of Customers using K-Means

RFM (Recency, Frequency, Monetary) analysis is a proven marketing model for behavior based customer segmentation. RFM analysis is applied to present data at aggregate level and is used to segment customers into homogenous groups. Three main variables resulting from the analysis are R-recency, F-frequency, and M-monetary. These three values are important as F and M indicate the value of customers, and R indicate customers' engagement and satisfaction. The values are easy to obtain from the transaction history and are grouped by customer ID. Recency shows the length of duration since last purchase. Frequency is the number of orders over the period and monetary is the sum of the total amount spent over the period. By using RFM, it will help to find who are the most valuable customers, who are churned, who are potential customers and many others customer groups according to their RFM values. R is calculated by finding the difference between analysis date and latest purchase date. The number of invoices for each customer ID is counted to get F and sum the purchasing amount over the period of time for each customer ID.

After all the RFM variables have been calculated, elbow method is used to get the optimum number of K-means clusters. K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. K- means clustering modeling is built using the optimal number of k from the elbow method.

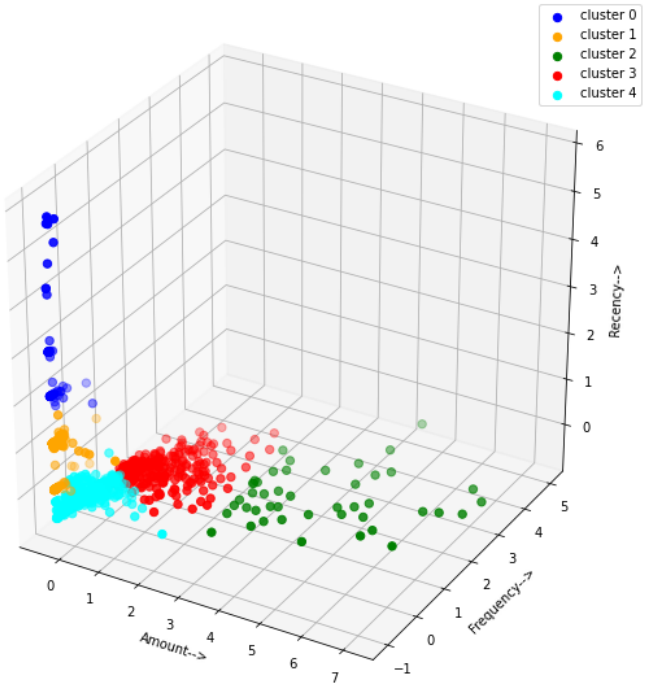

Figure 10: K-means Clustering Results

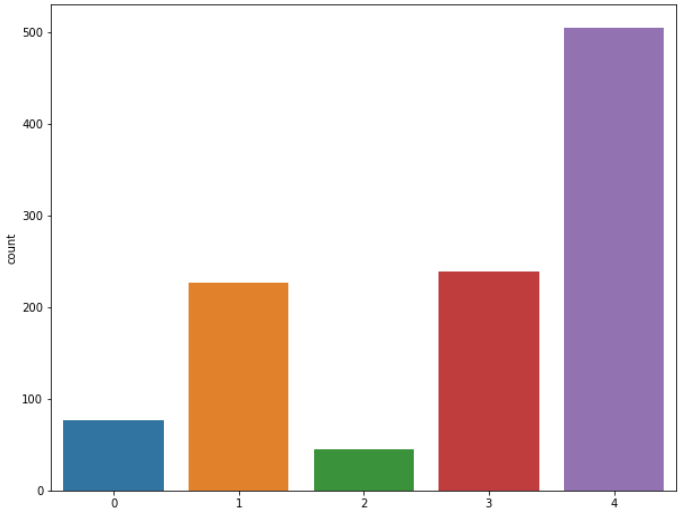

Figure 11: Customer Count in each Cluster

	Recency			Frequency			Amount			
	mean	min	max	mean	min	max	mean	min	max	count
Cluster_Id										
0	391.894737	276	724	11.644737	1	142	1.568783e+05	38.00	4686682.00	76
1	154.097345	76	250	13.743363	4	158	2.297433e+05	156.00	10541644.00	226
2	4.000000	0	17	413.333333	123	858	3.073049e+07	18906102.91	51988149.35	45
3	5.112971	0	52	334.527197	150	630	6.273947e+06	585174.00	18080877.90	239
4	11.633663	0	87	120.314851	8	253	9.213003e+05	390.00	15737440.92	505

Figure 12: Statistics summary for RFM variables of each clusters

From the statistics summary, cluster 2 is the most valuable group of customers with highest mean F, lowest mean R and they also purchased the most. Cluster 3 has the second lowest mean R and second largest mean M (amount) and F. Therefore, it can be recognized as the second most valuable customer group. Cluster 0 is the worst group with lowest F and M and highest R that means they are inactive and almost lost them.

5.4.2 RFM Clustering of Customers using Quintiles RFM Scoring Method

Different businesses may use different methods of rfm formulas for ranking the RFM values on the scale. In this project, quintiles are used to calculate the RFM score because using quintiles is more flexible as the ranges will adapt to the data and would work across different industries. Quantiles are similar to percentile but it divides the data into five equal parts. At this point, the values of Recency, Frequency and Monetary parameters are divided into five equal parts, so each quintile contains 20%. Each

customer will get a note between 1 and 5 for each parameter. Higher Recency score means the customers have bought recently, higher Frequency score means the customers have bought many times and higher Monetary scores means the customers bought in a large amount. In contrast, lower R, F and M scores represent worse results. The RFM scores give $5^3 = 125$ segments which is not easy to identify all the segments . Therefore, ten segments are identified based on the RFM scores. The description of each segment is as shown in Table 2.

Table 4: Description of the customer segments

Segment	Description	Recency Score Range	Frequency and Monetary Range
Champions	Bought recently, buy often and spend the most	4-5	4-5
Loyal Customers	Buy on a regular basis. Responsive to promotions.	3-4	4-5
Potential Loyalist	Recent customers with average frequency.	4-5	2-3
New Customers	Bought most recently, but not often.	4-5	0-1
Promising	Recent shoppers, but haven't spent much.	3-4	0-1
Customers Needing Attention	Above average recency, frequency and monetary values. May not have bought very recently though.	2-3	2-3
About to Sleep	Below average recency and frequency. Will lose them if not reactivated.	2-3	1-2
At Risk	Purchased often but a long time ago. Need to bring them back!	1-2	3-4

Can’t Lose Them	Used to purchase frequently but haven’t returned for a long time.	1-2	4-5
Inactive	Last purchase was long back and a low number of orders. May be lost.	1-2	1-2

By this way, customers are assigned to their related segments based on their RFM score range as shown in Figure 13.

	customer_code	Amount	Frequency	Recency	R	F	M	RFM Score	Segment
0	15388969F200610610620807897	156.00	4	76	2	1	1	211	inactive
1	15388969F200610610620807898	4405953.98	160	5	4	4	5	445	loyal customers
2	15388969F200610610620807899	1610801.00	119	5	4	3	4	434	potential loyalists
3	15388969F200610610620807900	1415948.00	177	5	4	4	4	444	loyal customers
4	15388969F200610610620807901	156.00	4	76	2	1	1	211	inactive

Figure 13: Example of RFM score and Customer Segments

Some data visualizations are created to get a better idea of our customers portfolio. First, the distribution of R, F and M are visualized using bar charts to compare quantities of R, F and M. The distribution of segments are also visualized to see the number of customers and percentage of each segment.

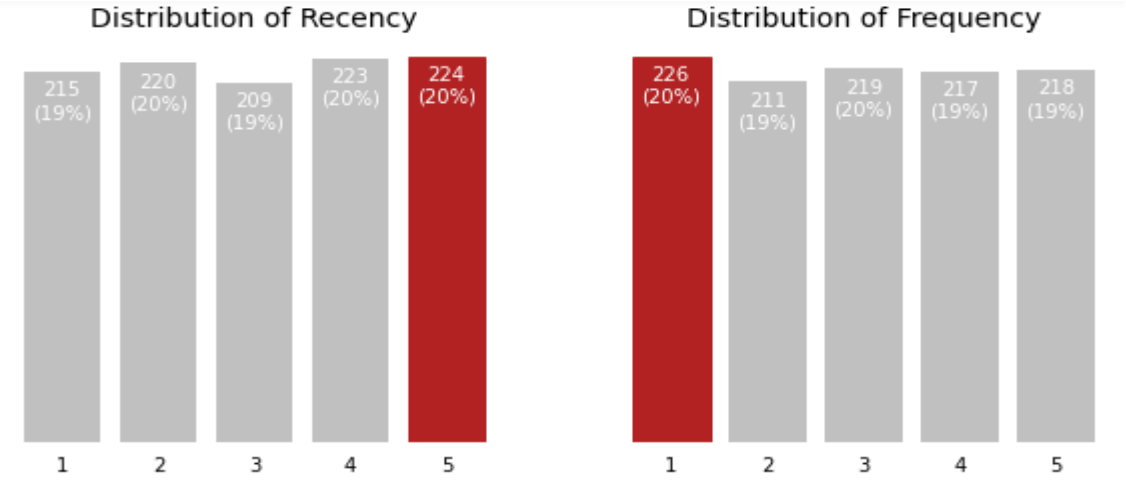

Figure 14: Distribution of Recency and Frequency

It can be seen that the distributions of customers in terms of recency and frequency are

almost evenly distributed.

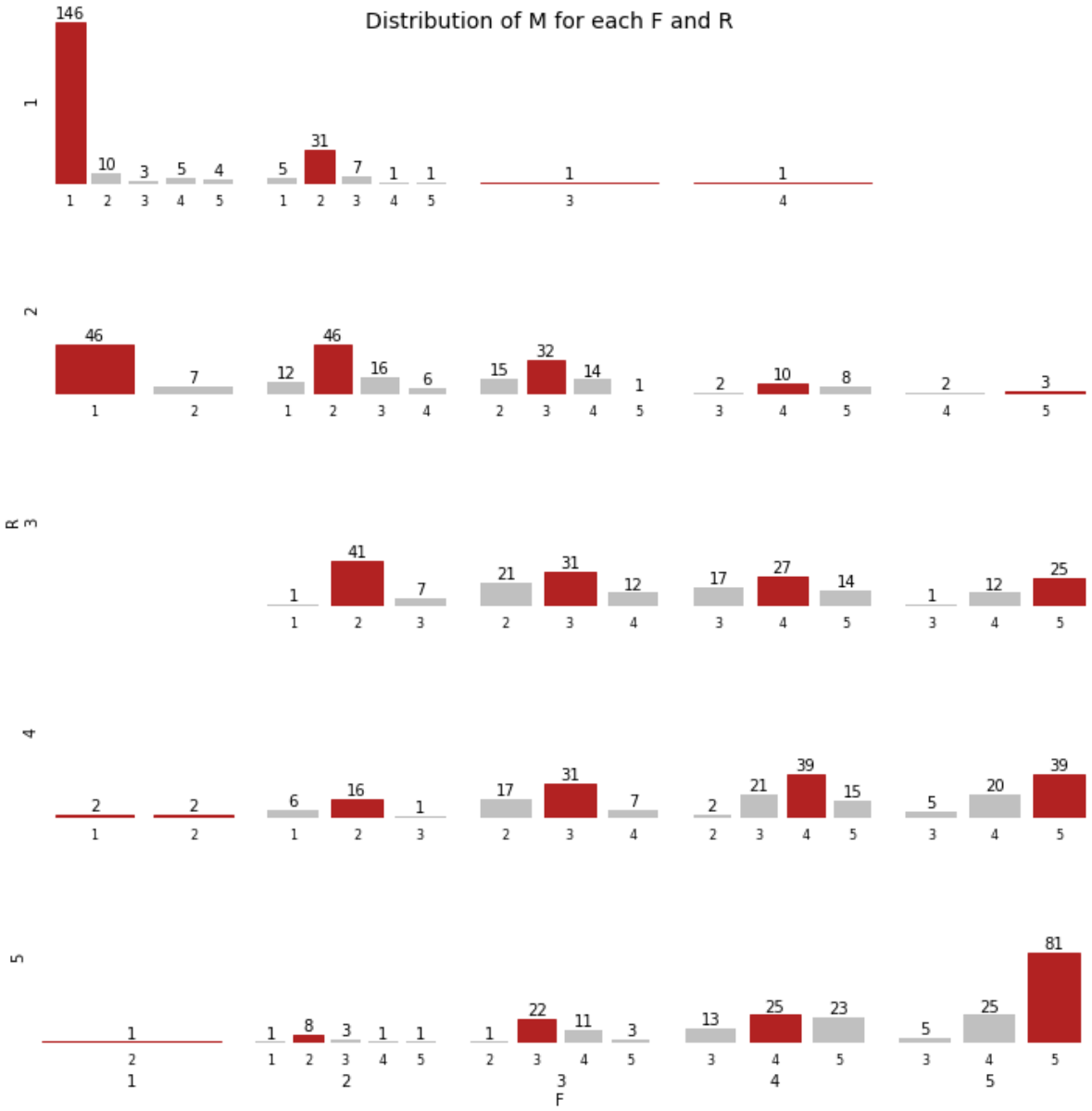

Figure 15: Distribution of Monetary for each F and R

When looking at the monetary value in Figure 15, it can be seen that the customers spending the most are those with the lowest R and F score which means that this business has very large sales to the customers but they are not frequent and recent customers. On the other hand, there also has customers who bought frequently and recently with the second largest amount of purchase.

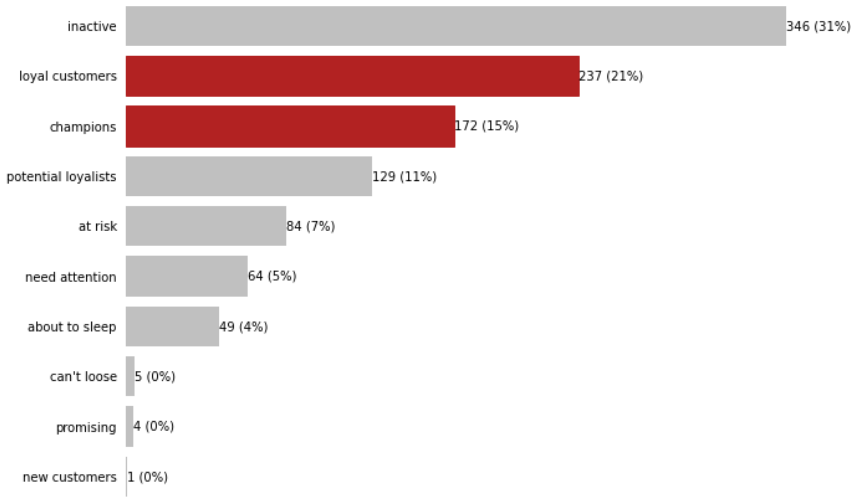

Figure 16: Customer Distribution in each Segment

	Recency			Frequency			Amount			
	mean	min	max	mean	min	max	mean	min	max	count
Segment										
about to sleep	10.040816	8	13	62.000000	13	84	1.399546e+05	13810.0	714129.00	49
at risk	32.261905	14	280	132.023810	86	270	2.082451e+06	97425.0	26707334.82	84
can't loose	15.800000	14	20	328.400000	291	363	1.593062e+07	2251724.0	41756024.82	5
champions	1.174419	0	2	329.476744	152	858	9.231342e+06	370648.0	51988149.35	172
inactive	189.488439	14	724	18.193642	1	83	2.098424e+05	38.0	10541644.00	346
loyal customers	7.206751	3	13	273.535865	150	700	5.669678e+06	131731.0	48266019.90	237
need attention	10.546875	8	13	112.281250	86	149	6.302348e+05	55348.0	3482371.00	64
new customers	2.000000	2	2	12.000000	12	12	3.029600e+04	30296.0	30296.00	1
potential loyalists	3.441860	0	7	96.860465	13	149	8.171531e+05	672.0	15737440.92	129
promising	3.000000	3	3	11.500000	10	12	3.043625e+04	390.0	60376.00	4

Figure 17: Statistics summary for RFM variables of each clusters

According to the customer distribution bar graph and statistics summary, there are 237 loyal customers and 172 champions, and 129 customers are also expected to become loyal customers soon. On the other hand, there are 346 inactive customers which is the biggest number out of all clusters, 84 are at risk and 64 need attention. Therefore, the business owner should figure out the reasons and customize plans to encourage the customer purchase. Moreover, the quality of services should be enhanced to avoid further losing.

6. Web Application

Web application was developed to display dashboards for business people. The main target for this web application are the business owner, marketing team and sales team. The

dashboard consists of a summary section, a sales section and a customer section. The information shown on the dashboards are from the descriptive analysis and predictive analysis.

This project aims to create the real-time web application in order to display the dashboard with real-time data. This will be beneficial for business as they can notice the anomaly in their business and execute the proper action in the right time. However,creating a real-time system required abundant technology resources and time to provide a proper user experience. The plan and idea to create the real-time web application has been laid, but not successfully executed.

The front end for the application was developed using ReactJS together with FusionChart for visualization. Users can choose the year they would like to see the information and drill down the charts to see everything in more detail.

The data will be pulled from the database, MongoDB, which was called through API. API was developed with python as we aimed to execute machine learning models in real-time as well. Web socket will be used to monitor and update the dashboard in real-time.

Figure 18: The user interface of the web application on Summary section

Figure 19: The user interface of the web application on Sales section

Figure 20: The user interface of the web application on Customer section

7. Conclusion

Business analytics is very important to reduce risk and make data-driven decisions that have more essential business insights. Moreover, it can give an overview and insight on how the business can become more efficient. Our project is expected to support dashboards for the distributor that have both descriptive and predictive analyses for sales and customers. The distributor can use these dashboards effectively to quickly gain insights into the sales and predict the monthly sales so that he can make the right decisions and plan according to the prediction. Moreover, the distributor can keep track of the customers' behavior and develop more targeted customer retention strategies by identifying the customer dashboard.

8. Reference

1) https://www.putler.com/rfm-analysis/
2) https://guillaume-martin.github.io/rfm-segmentation-with-python.html
3) https://stackabuse.com/time-series-prediction-using-lstm-with-pytorch-in-python/