# Sales Analysis of a Trading Company for Data Driven Decision Making

Thejan Rupasinghe, Adeesha Jayasooriya, Ranula Liyadipita
Department of Computer Science & Engineering
University of Moratuwa
Email: thejan.20@cse.mrt.ac.lk, adeesha.20@cse.mrt.ac.lk, ranula.20@cse.mrt.ac.lk

***Abstract*—Trading companies work as middleman businesses between manufacturers (vendors) and customers, making the use out of their warehousing and transportation infrastructure. They are heavily dependent on their customer and supplier base, as any other company, and especially on the strength of their logistic system. So careful decisions needed to be made regarding customer lifetime value and, warehouse and transportation management to increase the revenue without overwhelming the infrastructure.**

**The complexity of this business model often forces decision-makers to make decisions based on subjective mental models which reflects their experience. However, researches have shown that companies perform better when they apply data-driven decision-making. This study analyses the sales data of a medical laboratory item trading company using the Descriptive, Diagnostic and Predictive Analysis to overcome the challenges faced by the company through data-driven decision making.**

***Index Terms*—sales analysis, demand prediction, sales prediction, data visualization**

## I. Introduction

Trading companies, also known as Merchandising companies, are businesses that work with different kinds of products which are sold for consumer, business or government purposes. In contemporary times, trading companies buy a specialized range of products, maintain a stock, broker them, and coordinate the delivery of products to customers [1].

Mainly two kinds of trading businesses can be defined; wholesalers and retailers. Wholesalers sell and deliver merchandise to other businesses and large end customers at a reduced rate, while Retailers sell inventory to the general public. In turn, a wholesaler can become a retailer by making the goods available to the public. Despite all these differences, today's trading company mainly refers to B2B traders, highly specialized in one goods category, working in a large geographical area with a strong logistic organization. Further following characteristics can be found in a typical trading company.

- Selling a variety of products in a specialized domain, accounting for sales through a cash register or point-of-sale system.
- Receiving payments from customers for merchandise purchased.
- Buying products from vendors in a large quantities.
- Managing inventory, such as products placed in a warehouse.
- Having back stock or excess inventory in a warehouse or stockroom.
- Earning revenue and profits from the merchandise sold.

As per the definitions and characteristics suggest, these trading businesses are heavily dependent on their customer base and the strength of the logistic system they have. So the following can be recognized as the most common challenges for these sorts of businesses.

- Minimizing customer churn.
- Improving customer lead generation.
- Keeping the warehouse storage at the optimum level.
- Maximizing the value of a customer.
- Deciding the right price for the right customer.

These issues should be addressed in a proper manner to make a trading business efficient, effective and successful.

This study focuses on a trading business of medical laboratory items to analyze the possible approaches to overcome these challenges using data science methodologies. In this domain, some of these become even more challenging, such as managing the warehouse; because many perishable items are involved. So the analysis of sales data in the company, using Descriptive, Diagnostic, and Predictive methods will assist the Data-Driven Decision Making (D3M) in approaching the above-mentioned challenges.

The outcomes of this study aim to maximize customer lifetime value and optimize the warehouse of the business. The analysis done on sales will reveal the customer buying patterns and, the predictions on sales will give an idea about future revenue and warehouse storage requirements.

The next subsection provides a detailed overview of the data set used in this research. Section II discusses the previous works which are related to our study. Section III describes the methodology followed in this study while the outcomes of the study are discussed in Section IV. Finally, we conclude our discussion in Section V.

### *A. Data Set*

This study uses sales data collected from a medical laboratory item trading company for the financial year of 2017-2018. The data set consists of 17938 of records about transactions, done for the government customers, giving details about 15 attributes. TABLE I explains each attribute following the common pattern for building a data dictionary. Data Type abbreviations used in the table are as follows.

- CN - Categorical Nominal

- CO - Categorical Ordinal
- MD - Metric Discrete
- MC - Metric Continuous

TABLE I: Attribute Description

| Attribute Name | Data Type | Data Subtype | Description | Examples |
|---|---|---|---|---|
| Financial Year | MD | STRING | The financial year of the transactions happened | 2017-2018 |
| Description | CN | STRING | Description of the product sold | ALBUMIN REFRACTOMETER |
| Quantity | MD | NUMBER | Sold Quantity in Unit of Measure of the product | 10.5, 7000 |
| Unit Price | MD | NUMBER | Unit Price in Rupees | 58579.56, 6.47 |
| Total Price | MD | NUMBER | Total Price of the sale in Rupees | 615085.38, 45290 |
| Product Type | CN | STRING | Type of the product | CON, INS and SRV |
| Customer Name | CN | STRING | Name of the customer who bought the item/s | SRILANKA AIR FORCE |
| Date | MD | DATE | Date of the sale | 12/11/2017 |
| Unit of Measure | CN | STRING | Unit of measure for the Sold Quantity | PCS, PACK and Box |
| Supplier | CN | STRING | Supplier Name from where the distributor bought the item | 3M LANKA PVT LTD |
| CATNo | CN | ID | Identification number for each catalog of items | R-230, S4-175 |
| Product Group | CN | STRING | The group which the product belongs to | MICROBIOLOGY, IMMUNOLOGY |
| Region | CN | STRING | A Regional division of customers according to their physical locations | MED-REGION03, MED-REGION06 |
| Sales Person Code | CN | ID | Identification code of the person who did the sale | MED_SUD, MED_DK |
| CusType | CN | STRING | Type of the customer of the sale | GVT |

*Financial Year* attribute is 2017-2018 for all collected records in the data set as all the sales are done in the same financial year. *Description* field mainly has the product name. If it is a liquid product like a chemical solution, this description carries some other details such as concentration and volume per bottle. *Quantity* field contains numbers given in the unit of measure for the product. It can have fractional values when the measuring unit is Pieces (PCS). *Unit Price* is given in Sri Lankan Rupees (LKR) with cents separated by a period. *Total Price* in LKR is the multiplication of *Quantity* and *Unit Price*. Three short forms have been used in the *Product Type* attribute.

- CON - Chemical Constituents - Eg: CALCIUM8x20ML
- INS - Instrument - Eg: BIOSAFETYCABINET
- SRV - Service - Eg: INSTALLATIONCHARGERS

Any service that the company is providing to the customer, other than the transportation, is recorded also as a product sale but in the *Product Type* of SRV. Most of the time, *Customer Name* also carries the geographical area name of the customer. Examples; TEACHINGHOSPITAL-JAFFNA, DISTRICTGENERALHOSPITAL-MATALE. *Date* field gives the sale's date in MM/DD/YYYY format. Three types of *Unit of Measures* can be found in these sales records.

- PCS - Pieces
- PACK - Packets
- Box - Boxes

In *Supplier* attribute, 'VARIOUSPLACES' is used as a placeholder string to convey that there is no specific single supplier (vendor) for this product. Products are categorized according to the laboratory science they are used, in the attribute *Product Group*. Some examples are; BIOCHEMISTRY, IMMUNOLOGY, HISTOPATHOLOGY, HEMATOLOGY, and MICROBIOLOGY. Classification of *Regions* is done by the company and has named them from MED-REGION00 to MED-REGION11. Here in *CusType* column, only GVT can be found as all the sales related to the collected records are done for government customers.

## II. Related Work

Various researches have been carried out in sales analysis and prediction. The book; *Sales Management: Analysis and Decision Making* [2], gives an in-depth detailed understanding about the sales management and analysis. It also represents how this analysis can be used to manage a company's functions in the most effective way. A whole chapter is reserved in this book for Sales Forecasting, which brings out types and uses of forecasting, top-down and bottom-up approaches and finally using a regressive analysis to forecast future sales.

The patent research of *Method for performing retail sales analysis* [3] invents a novel system to provide a user with substantial flexibility in requesting and generating analysis projects on transaction and/or consumer data that is stored in one or more databases. It proposes to store these data as spreadsheet based interactive reports, which are easy to manipulate for further analysis and presentations. Furthermore, the study claims that the insights taken using the invention will lead to better decisions on new product launches, sampling, merchandising, assortment, distribution, and other sales and marketing priorities.

The study [4] explains a critical analysis on predicting sales using various machine learning models. Predicting sales being a classic problem engaging time-series analysis, researches like [5]–[7] presents how ARIMA (Auto Regressive Integrated Moving Average) and Neural Network based models can be used in this scenario. Here in this study we try to use Facebook's™ Prophet [8] open-source library to predict sales.

## III. Methodology

Three data science analysis techniques; Descriptive, Diagnostic and Predictive Analysis, are applied on sales data to assist in deriving business insights.

### A. Descriptive Analysis

Each of the item that is sold by the company belongs to a specific product group. Overall there are fifteen different product groups.

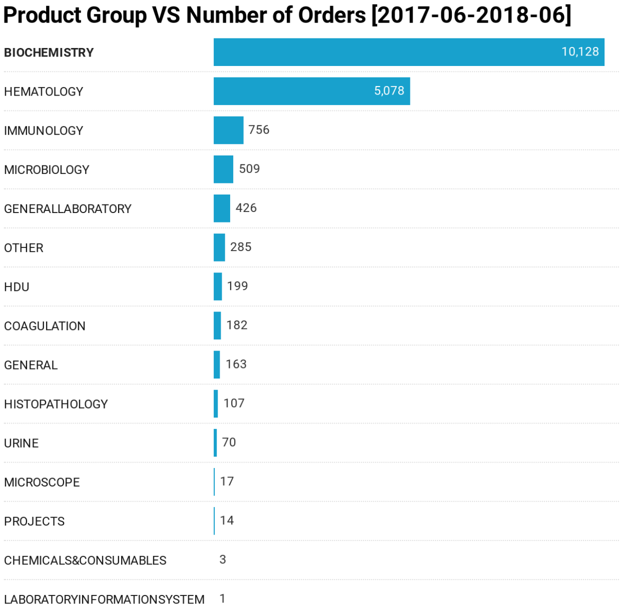

Fig. 1: Variation of Number of orders with regard to product group

It is evident from the Fig. 1 that Biochemistry product group had the highest demand for the 2017-2018 financial year. More than 50% of the orders were under that product group. Major supplier for the company during the financial year was the Chinese bio medical company named Shenzhen Mindray Bio-Medical Electronics Co.Ltd. According to Fig. 2, 46% of the orders were completed through this supplier.

This yields an interesting insight with regard to the company strategies. That is, the company is heavily dependent on a single supplier and going forward into the business it will drastically affect them if that supplier to go through any unforeseen tough business years. Hence for them it would be better to have multiple suppliers with more of an even breakdown.

Fig. 3 illustrates the highest sold catalog numbers for the latter 3 months of the financial year. It shows an abnormal demand for specific three items during the last quarter. That might provide a better insight into the aspect of stock management in the company for the upcoming years.

Customers can be prioritized according the average order quantity. Top twenty customers according to their respective

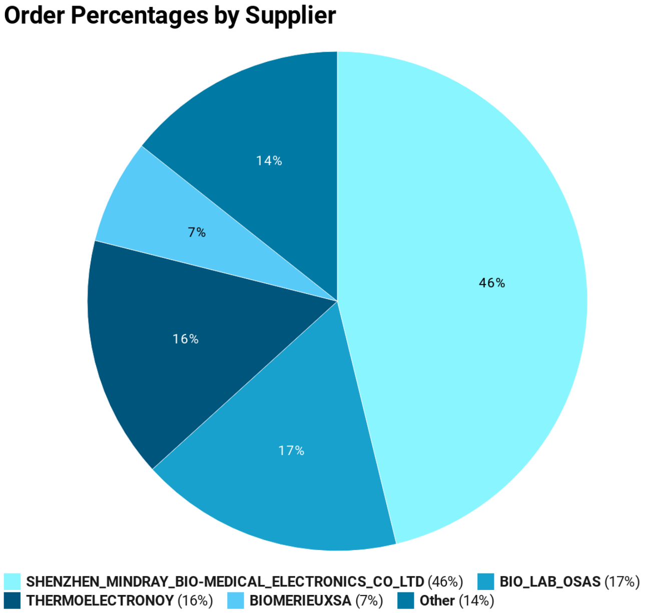

Fig. 2: Variation of Number of orders with regard to supplier

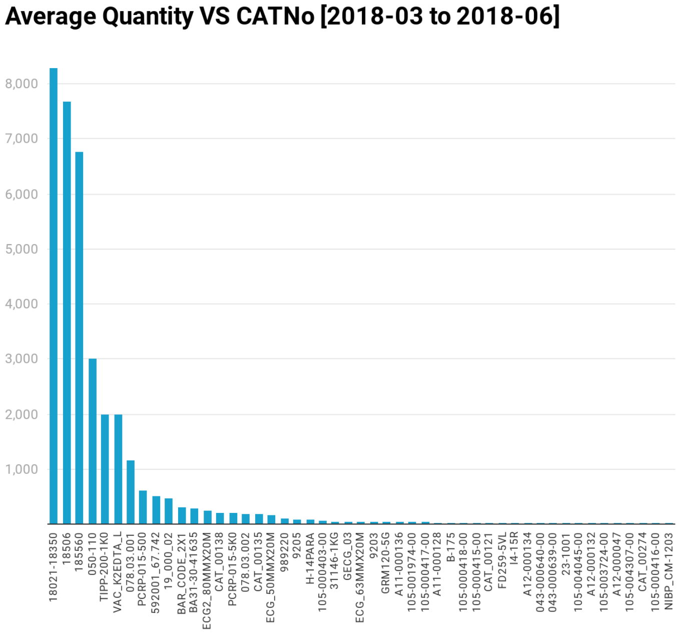

Fig. 3: Average quantity for a category

order quantities are shown in Fig. 4. Sri Lanka Navy Headquarters is the top most client of the business and seven of the top twenty are generally hospitals across the country.

Customers of the company are distributed across the country. Fig. 5 highlights that highest number of customers are based in Colombo district. Kurunagala, Kandy and Galle happen to have a high density of customers compared to the rest of the districts. Northern regions happen to have the lowest customer density among rest.

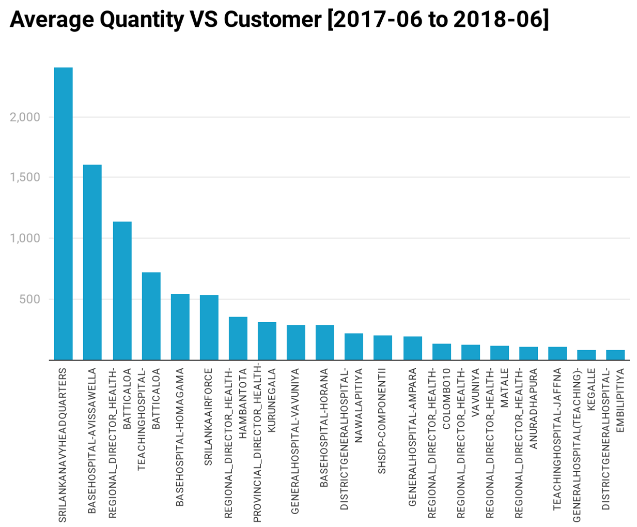

Fig. 4: Variation of Average Quantity with regard to the customer for the financial year 2017-2018

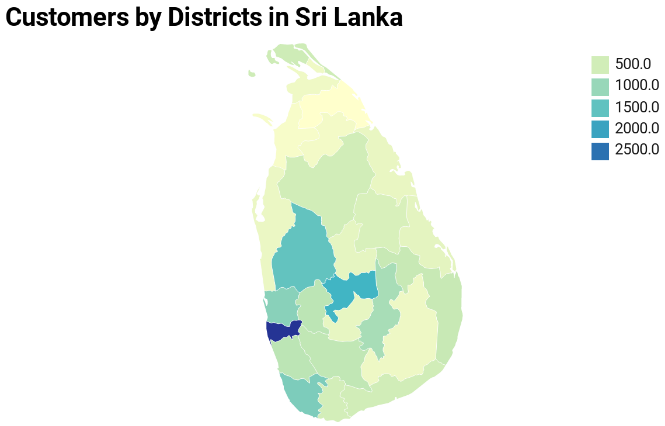

Fig. 5: Customers classified in to districts according to their geographical location

For management purposes company has divided the country into eleven different regions. Fig 6 shows how the number of orders varied on the region basis.

REGION 4 is the top region with respect to the number of orders. Top 3 cities of the REGION04 as per the number of orders are as follows,

- PERADENIYA - 830 orders
- KEGALLE - 698 orders
- KANDY - 662 orders

### B. Diagnostic Analysis

To elaborate more on the Average order quantity of customers we did an analysis considering the last month of the financial year. Fig . 7 shows an abnormal average for Sri Lanka Navy Headquarters. Furthermore, when we analyze the Fig. 4 we can see that which is shown in 7 is an abnormality and in-

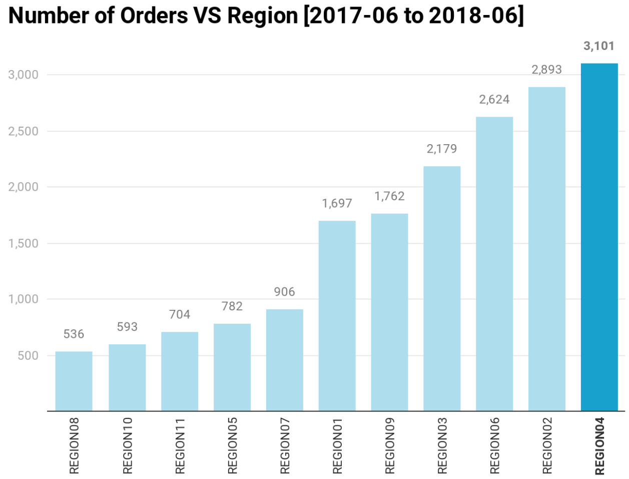

Fig. 6: Variation of Number of orders with regard to region

fact for the month June, average quantity is almost as twice to the overall average quantity. Reason for the spike in average order quantity was the lack of ordering during the months April-May.

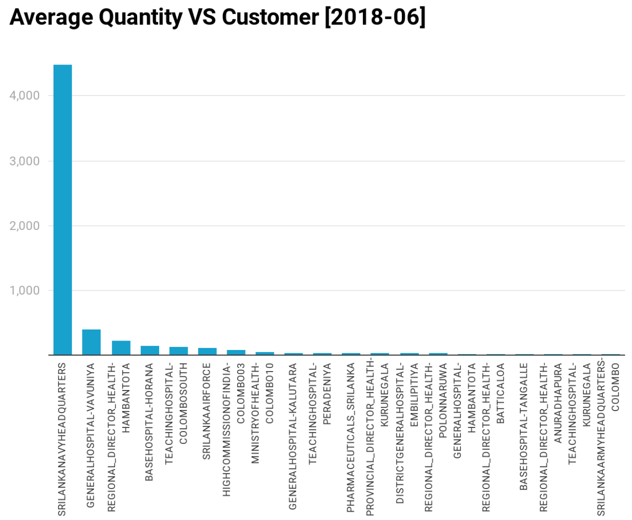

Fig. 7: Variation of Average Quantity with regard to the customer for the month of June 2017

REGION 8 is the region with least number of orders. We analyzed the trend in Fig. 8 for average order quantity in REGION 8. It shows in the beginning of the financial year there was a higher order quantity and as the year progresses average lessened. Further when we looked into the number of cities that represents REGION 08 it was shown that only five cities ( HOMAGAMA, ANGODA, COLOMBO10, AVISSAWELLA, MULLERIYAWA) are included for that region. It is the region with least number of cities.

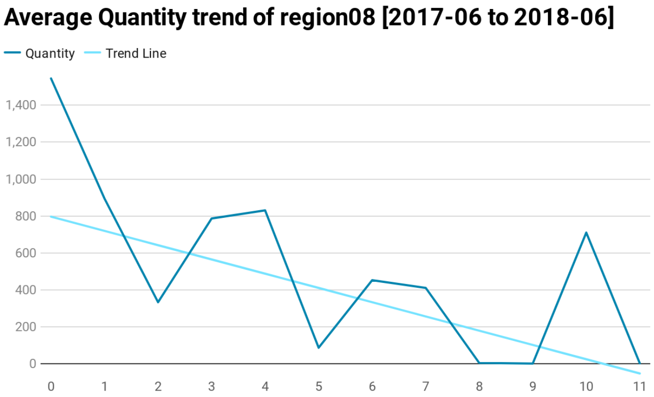

Fig. 8: Average quantity trend in MED-REGION08

### C. Predictive Analysis

Considering our problem is to identify the demand for lab equipment and chemicals we choose to predict the *sales quantity* field. Since the data set was really clean on its raw form, we did not have to do any complex data preprocessing before feeding into the learning models. However, we have a problem with the wrong data in the quantity field that need to be preprocessed before feeding to models. Besides, we used famous time-series machine learning models like Prophet to predict the demand for lab instruments and chemicals.

*1) Preprocessing:* Since data is in raw form and no missing values not much preprocessing is required. However, in the quantity field, some minus values should not be there. Accordingly, we decided to remove those records from the data set.

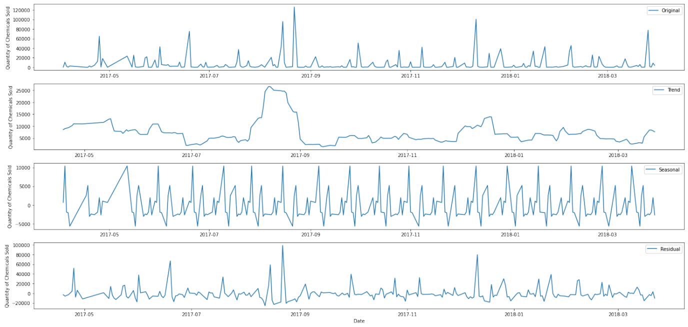

Fig. 9: Time series charts of quantity of sold chemicals

*2) Time series Analyze:* At the end of the initial preprocessing, we started to analyze the quantity field against time for chemicals and lab equipment separately. Figures 9 and 10 contain the original data, the trend of original data with a window size of 14, seasonal data with window size 14 and residuals data which is a result of reducing seasonal and trend data from original data. Charts are generated from seasonal decompose of *scikit-learn library* [9].

There is no clear upper trend or down trend in sold lab equipment and chemicals. And can see some spikes in the residual charts which may cause of the unknown fact. In our data set, there is no clear evidence to predict these residual changes. Therefore our prediction results of chemicals have high error values as you can see in Tables II and III. Therefore we decided to predict for each product separately.

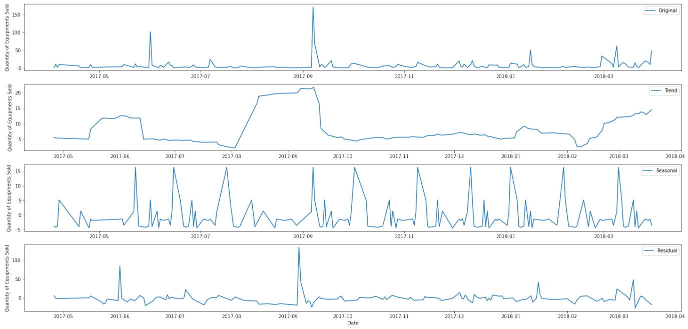

Fig. 10: Time series charts of quantity of sold lab equipments

## IV. Results

Here we used four machine learning models to predict the quantity of each product. Linear Regression, SVM, Random Forest and Prophet. Predicted results evaluated using cross-validation of 10 folds in Linear regression, SVM and Random forest. In prophet cross-validation calculation initial set to 180 days, period 30 days and horizon 60 days. Here we used RMSE(Root Mean Squared Error) as our error function.

TABLE II: RMSE of models for Chemicals

| Model | RMSE |
|---|---|
| prophet | 15509.70 |
| Linear Regression | 16688.59 |
| SVM | 17944.47 |
| Random Forest | 19349.49 |

TABLE III: RMSE of models for Lab Equipments

| Model | RMSE |
|---|---|
| prophet | 18.57 |
| Linear Regression | 18.51 |
| SVM | 19.01 |
| Random Forest | 22.07 |

*1) Model Selection:* Even though Linear regression shows the best results with lab instruments it does not show good results with Chemicals. The Prophet model is showing better results with both chemicals and lab instruments. Therefore we selected Prophet to predict each product separately.

*2) Results by Product:* The table IV shows some products showing lower RMSE while some products showing higher RMSE. For instance, MAGNESIUM8*10ML showing an error of 0.07 while SCAL3MLX10 showing an error of 6.50.

TABLE IV: RMSE of Prophet by Products

| Product | RMSE |
|---|---|
| MAGNESIUM8*10ML | 0.07 |
| HDL/LDLCALIBRATOR2X2ML | 1.10 |
| GAMMAGT10X10ML | 3.07 |
| CK-NAC(IFCC)20X3ML | 2.93 |
| SCAL3MLX10 | 6.50 |
| CALCIUM8x20ML | 5.66 |

## V. Conclusion and Future Work

Analysis we performed on the medical sales data set yielded interesting set of insights. Data set contained sales information for the financial year 2017-2018. Having data for a single year was a limitation for us to provide deeper insights into the how sales behaved over the years. Nevertheless, in Descriptive Analysis it was evident that company's sales are majorly focused on Western and Central provinces. Region breakdown was not directly correlating to geographical setting of the cities. Nonetheless, it showed that company can potentially grow their business into areas such as Kalutara, Ratnapura as these districts are surrounded by higher customer density districts and they already have the logistics and infrastructure through these higher customer density districts. Another major highlight of the analysis was company being heavily dependent on a sole supplier ( 50%) and that might lead to unfavourable consequences. Even though the company provides many items it was interesting to see only a few items have a higher demand during the year. That might provide a useful insight when it comes to stock management.

Diagnostic analysis showed that regions breakdown needs to be re looked at and SL Navy HQ is the major customer of the company. It is important that it's requirements/orders are satisfied on time as the priority customer. The predictive analysis was rather challenging with the accuracy of the models. Because of predicting residuals is difficult. Even though we were able to give accurate results for some products while some are still having high error margins.

As to the future work it would be fruitful to perform an analysis over the years to identify demands and provide much effective predictions. To achieve effective predictions this application can be further extended to multivariate time series analysis. For instance, here we cannot predict residuals accurately since it does not depend on time. But if we are able to find some features which correlated with quantity predicting residuals will be more accurate. That may lead us to less error margin of predictions. Furthermore, time series machine learning models like ARIMA (Auto Regressive Integrated Moving Average) and other Neural Network-based models can be used for this sales prediction.

## References

[1] "What is trading company? definition and meaning - businessdictionary.com," http://www.businessdictionary.com/definition/trading-company.html, (Accessed on 04/10/2020).

[2] T. N. Ingram, R. W. LaForge, C. H. Schwepker, and M. R. Williams, *Sales management: Analysis and decision making*. Routledge, 2015.

[3] P. Springfield, E. Blake, and D. Stern, "Method for performing retail sales analysis," Jul. 3 2012, uS Patent 8,214,246.

[4] M. Bohanec, M. K. Borštnar, and M. Robnik-Šikonja, "Explaining machine learning models in sales predictions," *Expert Systems with Applications*, vol. 71, pp. 416–428, 2017.

[5] H. Omar, V. H. Hoang, and D.-R. Liu, "A hybrid neural network model for sales forecasting based on arima and search popularity of article titles," *Computational intelligence and neuroscience*, vol. 2016, 2016.

[6] S. P. Shakti, M. K. Hassan, Y. Zhenning, R. D. Caytiles, and I. N. C. SN, "Annual automobile sales prediction using arima model," *Int. J. Hybrid Inf. Technol*, vol. 10, pp. 13–22, 2017.

[7] A. R. Abdel-Khalik and K. M. El-Sheshai, "Sales revenues: Time-series properties and predictions," *Journal of Forecasting*, vol. 2, no. 4, pp. 351–362, 1983.

[8] "Prophet — prophet is a forecasting procedure implemented in r and python. it is fast and provides completely automated forecasts that can be tuned by hand by data scientists and analysts." https://facebook.github.io/prophet/, (Accessed on 04/10/2020).

[9] "scikit-learn: machine learning in python — scikit-learn 0.22.2 documentation," https://scikit-learn.org/stable/, (Accessed on 04/10/2020).