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*—**
TODO: Talk about,

- **trading companies buying and selling inventory**
- **problems in above areas**
- **specify medical trading company usecase**
- **how data science can help in D3M**
- **prdiction**

This study analyses the sales data of a medical lab item trading company using the methodologies in data science to overcome the challenges faced by the company through data-driven decision making.

A complexity of business dynamics often forces decision-makers to make decisions based on subjective mental models, reflecting their experience. However, research has shown that companies perform better when they apply data-driven decision-making. This creates an incentive to introduce intelligent, data-based decision models, which are comprehensive and support the interactive evaluation of decision options necessary for the business environment.

***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. 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 Con (????) - 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.

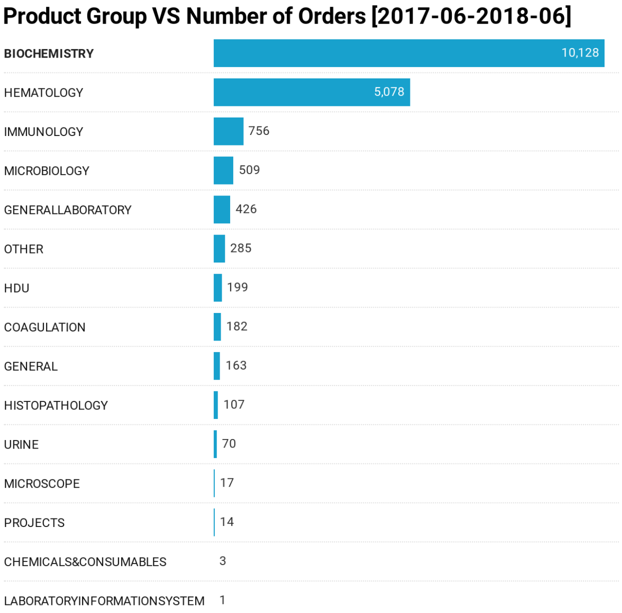

Fig. 1: Variation of Number of orders with regard to product group

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.

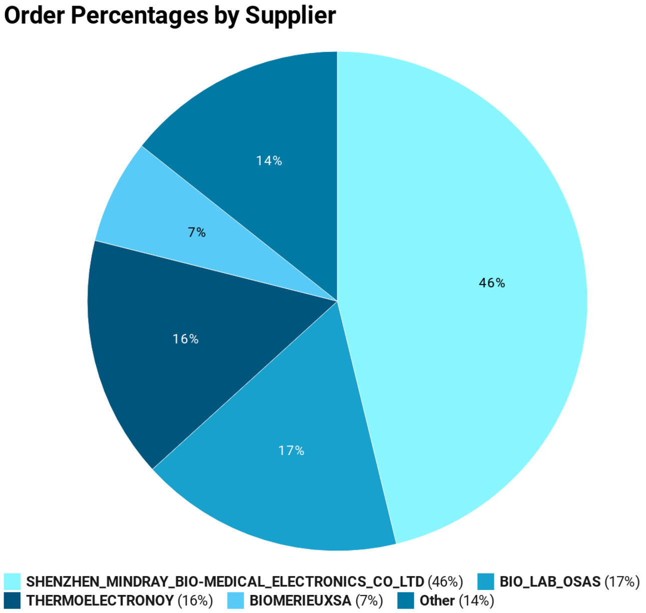

Fig. 2: Variation of Number of orders with regard to supplier

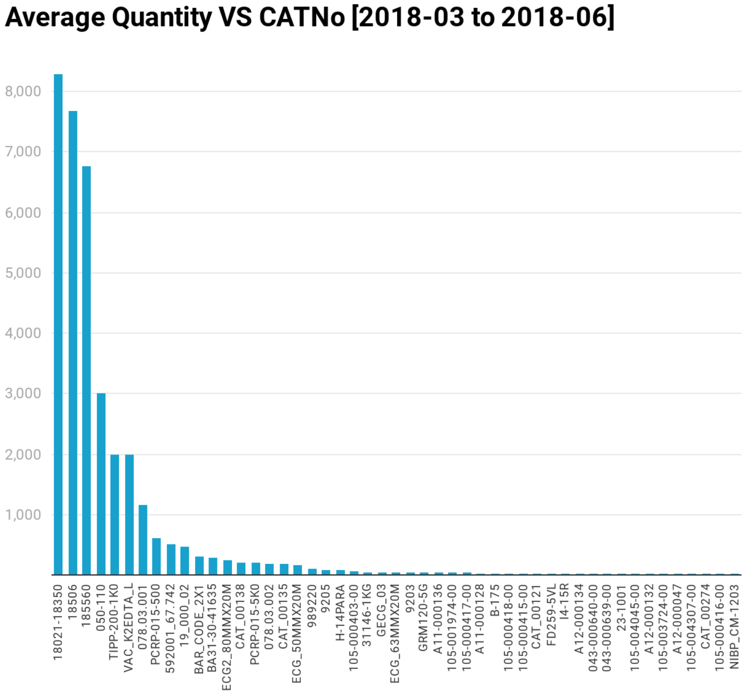

Fig. 3: Average quantity for a category

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

Customers of the company are distributed across the country. Fig. 5. highlights that highest number of customers are

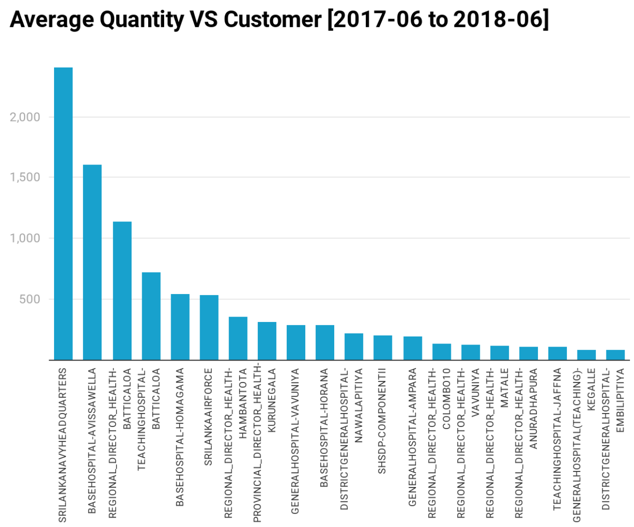

Fig. 4: Variation of Average Quantity with regard to the customer for the financial year 2017-2018

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.

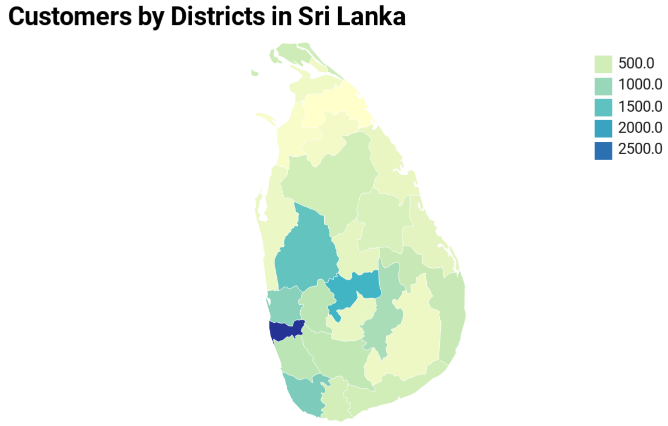

Fig. 5: Customers classified in to districts according to their geographical location

For management purposes company as 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

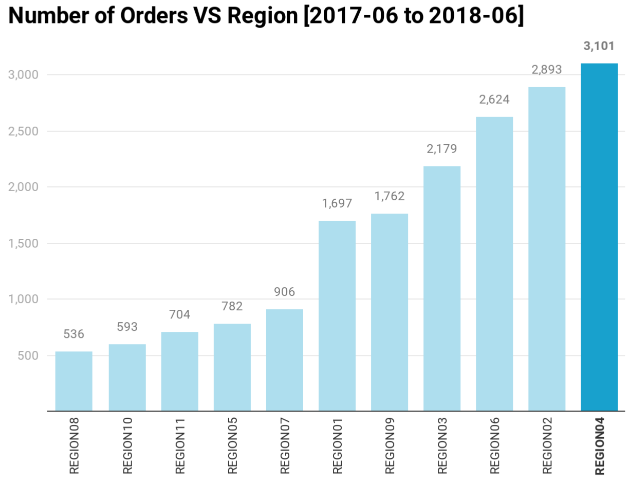

Fig. 6: Variation of Number of orders with regard to region

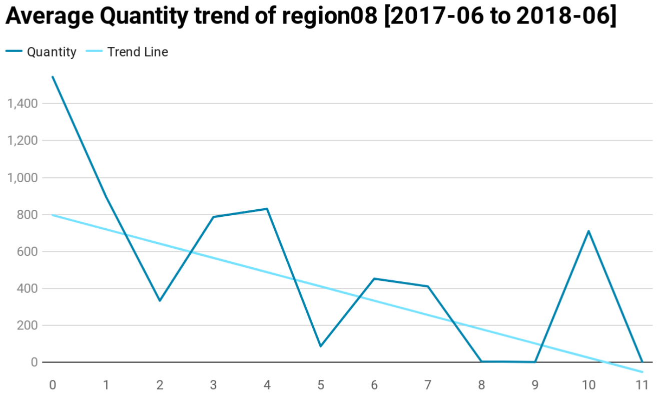

Fig. 7: Average quantity trend in MED-REGION08

B. *Diagnostic Analysis*

C. *Predictive Analysis*

Since our problem is to identify demand of lab equipment and chemicals we choose to predict quantity of demand of lab equipment's and chemicals. Since the data set was really clean on its raw form, we did not have to do any complex data reprocessing before feeding in to the learning models. although we has a problem with wrong data in the quantity field. Moreover we used time series machine learning models to predict the demand of lab equipment's and chemicals.

1) Preprocessing: Since data is in raw form and no missing values not much preprocessing is required. But in the quantity field there is some minus values which should not be. Therefore we decided to remove those columns from the data set.

2) Time series Analyze: At the end of initial preprocessing we started with analyze the Quantity field against time for chemicals and lab equipment separately. Figure 9 and 10 contain the original data, trend of original data with window size of 14, seasonal data with window size 14 and residuals

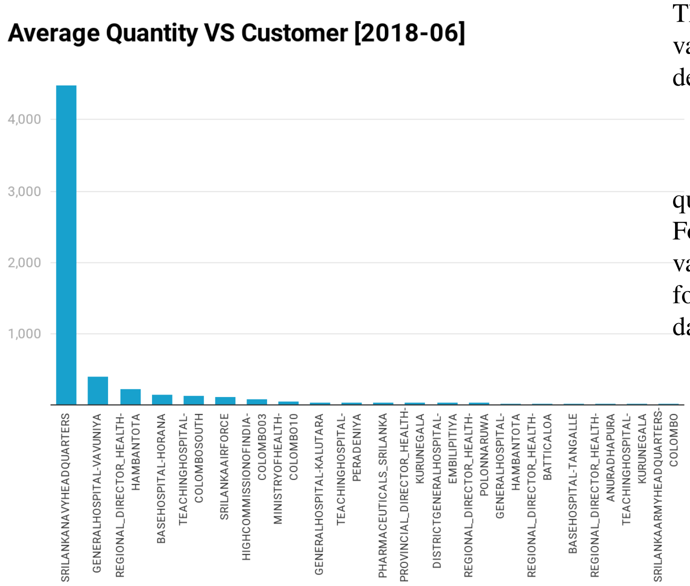

Fig. 8: Variation of Average Quantity with regard to the customer for the month of June 2017

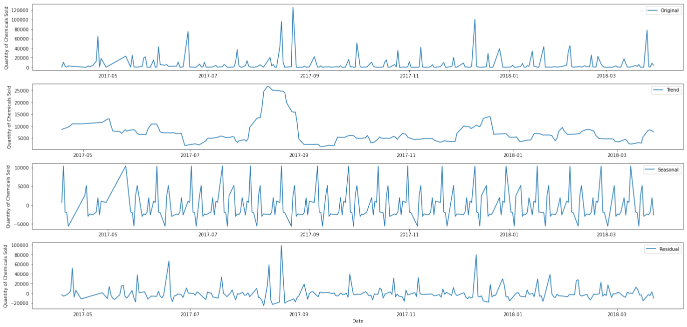

Fig. 9: Time series charts of quantity of sold chemicals

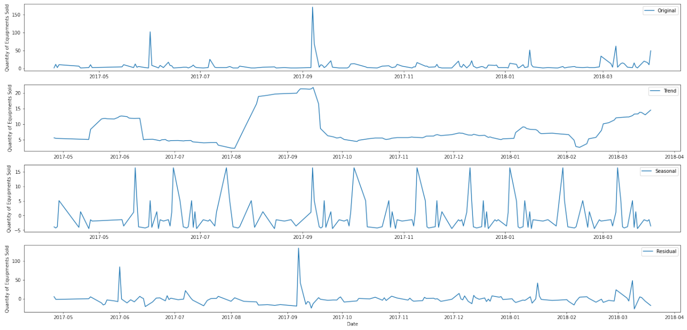

Fig. 10: Time series charts of quantity of sold lab equipments

data which is 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 in the residual charts which may cause of 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 the Table II and IV. Therefore we decided to predict for each product separately.

IV. Results

Here we used four machine learning models to predict the quantity of each products. 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.

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 best results with Lab equipments it not shows good results with Chemicals. Prophet is showing better result with both chemicals and Lab equipments. Therefore we select prophet to predict each product separately.

TABLE IV: RMSE of Prophet by Products

Product	RMSE
MAGNESIUM8*10ML	0.07
Linear Regression	1.10
SVM	19.01
Random Forest	22.07

V. Conclusion and Future Work

Developing a methodology to detect situations where multiple viewpoints are provided in regard to the same discussion topic within a court case transcript is the major research contribution of this study. This study has introduced novel approaches to detect deviations in the opinions provided by two sentences regarding the same topic. At the same time, existing methodologies to detect contradiction and change of perspectives have been evaluated within the study. Additionally, it has been empirically demonstrated the way in which the outcomes of the study can be used to facilitate the process of identifying relationships between sentences in court case transcripts. Evaluation of the performance of existing semantic similarity measures in relation to identifying verbs with similar meaning can be considered as another key research contribution of the study.

The proposed approach can also be used to facilitate several other information extraction tasks related to the legal domain such as identifying counter arguments to a particular argument, determining representatives related to the proposition party and the opposition party in a court case.

The accuracy of the approaches proposed in this study can be further improved by developing semantic similarity measures and sentiment annotators which can perform in the legal domain with an improved accuracy. Coming up with such mechanisms can be considered as the major future work.

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).