2022 International Conference on Emerging Trends in Computing and Engineering Applications (ETCEA) | 978-1-6654-7709-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/ETCEA57049.2022.10009688

Design and Implementation of Business Intelligence Framework for a Global Online Retail Business

Razan Al-Omoush
King Hussein School of Computing Sciences
Princess Sumaya University for Technology
Amman, Jordan
razanomoush9@gmail.com

Salam Fraihat
Artificial Intelligence Research Center
College of Engineering and Information Technology
Ajman University, UAE
s.fraihat@ajman.ac.ae

Ghazi Al-Naymat
Artificial Intelligence Research Center
College of Engineering and Information Technology
Ajman University, UAE
g.alnaymat@ajman.ac.ae

Mohammed Awad
Department of Computer Science and Engineering
American University of Ras Al Khaimah, UAE
mohammed.awad@aurak.ac.ae

***Abstract*— Due to the intense competition in today's online retail environment, companies seek to enhance their strategies by adopting effective analytical techniques and infrastructure, allowing them to quickly analyze critical information that supports decision-making. A Business intelligence (BI) framework can promptly fulfill such needs by processing massive amounts of collected data from multiple sources and representing them in a way companies can utilize in their strategic decisions. This research paper presents a detailed design and implementation of a BI framework for the online retail business industry. It includes requirement analysis, data modeling, BI framework design, and the implementation of descriptive and predictive analytic tools to provide insights and decision support for retail businesses. Moreover, the paper details the implementation of various machine learning algorithms used in sales predictive analytics, such as Linear Regression, Lasso Regression, XGBoost, Random Forest, and LSTM. Interactive charts are provided to assist decision-makers in carrying informed decisions.**

Keywords*— Business Intelligence Architecture, Descriptive Analysis, Predictive Analysis, Machine Learning, Online Retail, Sales Prediction, Dashboards.*

I. INTRODUCTION

Online retail is one of the most diverse sectors of the "vertical industry" [1], as it is one of the industries with the most significant number of companies and employees worldwide. The tremendous growth of online shopping has led to a highly competitive business environment. So, an immediate response to market changes has become so crucial that collecting, storing, and analyzing the data continuously have shown high importance in achieving a competitive advantage in this field [2].

Massive amounts of data are gathered from operational systems and external environments. For example, data collected from Point-of-sale (PoS), customer relationship management, logistic management systems, and e-commerce initiatives can assist inventory management, sales forecasting, marketing, and customer understanding, ultimately satisfying organizations' expectations and needs [3]. However, this data needs to be stored and analyzed to produce dynamic analytics reports as closely as possible [4]. Many retail firms have a lot of data. Still, they do not know how to extract the most usable data to translate it into a helpful business insight that helps develop competitive strategies [1].

BI technology can transform the raw data into meaningful information by providing exploratory and explanatory analytical reports, such as static reports or dynamic dashboards [5]. Furthermore, it plays a critical role in generating up-to-date information for operational and strategic decision-making about future sales, supply chain, and customer care to enable the companies to be more efficient in decision-making [6]. Thus, we can say that the primary goal of using BI is to provide better insight into daily business operations [1].

This project aims to implement a business intelligence framework for companies operating in the online retail business, using an online retail company's data from a public resource to provide the ability to analyze critical business aspects quickly. The ability to make future forecasts and access key information immediately from the collected data utilizing Business intelligence; the framework implementation can be expressed in five aspects: identifying business requirements, designing framework architecture, building a data warehouse schema, applying descriptive analysis and predictive analytics -by using machine learning algorithms-, and finally, representing the results in interactive dashboards.

The paper is structured as follows: Section II highlights the background and related work. The problem is described in Section III. Section IV shows the predictive analysis, Section V demonstrates the experiments and results, and section VI presents the BI application dashboard. Finally, Section VII concludes the paper.

II. BACKGROUND

BI represents a broad area of applications and technologies for gathering, storing, analyzing, reporting, and enabling rapid data access to enhance business process quality [2]. BI is defined by the Data Warehousing Institute "as the processes, technologies, and tools needed to turn data into information, and information into knowledge and knowledge into plans that drive profitable business action. BI encompasses data warehousing, business analytics, and knowledge management" [7]. A BI dashboard offers an ideal data visualization platform to facilitate intuitive decision-making [8, 9].

The typical BI environment mainly consists of two perspectives, the data warehousing perspective, and the BI application perspective, as shown in Fig. 1. BI is essential in implementing the extraction-transformation-load (ETL) operations from the data warehousing perspective. The data is collected from multiple sources in different formats (structured and unstructured). Then, the data is transformed and integrated to be loaded into the data warehouse [10, 11]. From a BI application perspective, BI plays a vital role in two main aspects: descriptive and predictive analysis, which the orgnization can use to support decision-making as follows:

- In Assessment Systems such as employee tracking systems, BI helps build Key Performance Indicators (KPIs) metrics and create a standard for evaluating and tracking employee progress, thus, assisting business managers in analyzing business goals and priorities [11].
- In the Retail business, BI helps build a framework for generating analytical reports and interactive dashboards that assist in analyzing the performance of all business aspects, thus, allowing the business managers to make the right strategic decision at the right time [2, 11].
- BI can provide support for predictive analysis, in E-commerce and Retail businesses, by using prediction models and data mining techniques, for example, in sales growth forecasting [11].

III. Business Intelligence for a Global Online Retail Business

This section presents the problem, requirement analysis, and system architecture.

A. Problem Statement

Global-Mart is an online superstore serving customers from five market segments around the world: North America, Latin America, Africa, Asia, and Europe, offering over 13K products from three categories, including office supplies, furniture, and technology [12]. The company has been historical, for over four years, of Sales, products, customers, and regions. All data must be summarized, analyzed, reported, and interpreted quickly to stay on trend and expand its market share [2]. The company aims to create comparative reports yearly to follow up on sales and product preferences according to different regions. Moreover, it wants to make a decision on next year's investments in market segments and to determine whether there is a need to stop serving a specific segment or not. Besides, there is a need to generate weekly sales forecasting reports, which allow the company to track its sales performance across the five market segments.

Building a BI framework will quickly meet the company's needs, as it provides the ability to analyze critical business aspects rapidly and make future predictions in a short time by accessing essential information immediately from data located in the Data warehouse.

B. Requirement Analysis

One of the most critical steps in building the business intelligence framework lifecycle is to conduct the requirements analysis for the proposed solution. It is essential to evolve a deep understanding of the operational business processes and objectives by identifying business requirements, system functional and non-functional requirements, and Data requirements. The following section presents a detailed description of each requirement type [13, 14].

1) Business Requirement

Business requirements analysis aims to define a set of high-level company needs and objectives that need to be addressed by the proposed solution [13]. The business requirements are described through the following key points:

- Create a multi-dimensional view of the historical sales data to help decision-makers make better business decisions that would increase revenue.
- Make a targeted strategy, provide a customer's insight, and understand their purchasing behaviors and product preferences.
- Provide insight into geographic trends to understand sales performance and create strategic plans.
- Provide an interactive dashboard that visualizes the data shape to make it easy and intuitive to develop and understand the overall business performance and help decision-makers make reliable and quick strategic decisions.
- Provide an accurate predictive model that helps estimate the company's future sales.

2) Data Requirement

The ETL process, including data collection, preparation, and loading, plays an essential role in the quality of the resulting BI system. The correct implementation helps the system achieve business needs. Data requirements, including data sources, ETL, and data analytics, are illustrated in detail in the following section.

TABLE 1. DATASET DESCRIPTION

Variable Name	Description
Order ID	Unique Order ID (25,690 orders)
Order Date	The date on which the order was placed (1-01-2014:31-12-2017)
Customer ID	Unique Customer ID
Customer Name	Unique Customer Name (796 customers)
Segment	Customer segment (Consumer, Corporate, Home Office)
City	City of ordered product, (3650 city)
State	State of ordered product, (1091 state)
Postal Code	Postal code of the city.
Country	Country of ordered product (162 country)
Region	Region of ordered product (23 Region)
Market	Topographical market segment where the customer belongs to (North America, Latin America, Africa, Asia, and Europe)
Product ID	Unique identifier of product
Product Name	The name of product purchased
Subcategory	Subcategories of the product (17 subcategory)
Category	3 Categories (Furniture, Home Offices, or Technology)
Sales	Transaction final amount
Quantity	Amount of product being ordered
Profit	Profit amount produced on that transaction
Shipping cost	Total cost of shipping for that transaction
Order Priority	The priority of the product (Critical, medium, high, or low)
Ship Date	Date of the order shipment
Ship Mode	Shipping type (frits class , Same Daye, standard class, or second class)

Data collection: The data used in this project were collected from the data. World- public data sets provider -as a CSV file, the dataset belongs to an online retail company named as Global-Mart, contains data about historical sales, customers, products, and locations (Region, country, city,

and state), distributed as 23 features across 51,290 data record. Data fields are briefly described in Table 1.

Data Cleansing and Preparation: In this phase, the following steps were performed to ensure the correctness and quality of the data:

- The most relevant data from the previously specified business requirements were extracted. All the features provided in the dataset were used except (Order Priority, Ship Date, and Ship Mode).
- Handling missing and noisy data was performed to ensure data quality. One feature (Postal code) was removed due to a lot of nulls around 41K; on the other hand, a lot of punctuation was found in some categorical features such as (City, Country, and State) -which will influence the analysis process - this was treated as well.
- All data have been transformed into a suitable format, such as temporal data (Dates), to enable accurate data analysis.

Data analytics: Data analytics requirement was provided by using the BI package in Microsoft Power BI to perform the analytic process.

C. *Proposed BI Architecture Framework*

A successful BI solution should follow an evolving architecture based on multi-stages (layers) to build the system with an enduring business value to accommodate the growth and renovation of evolving business needs over time [13]. Typically, the BI system's architecture consists of four main layers: product architecture, technical architecture, data architecture, and information architecture [15].

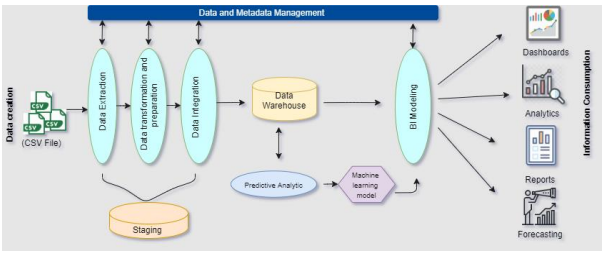

Fig 1. Information and Data Architecture

As shown in Fig. 1, the architecture describes the basic operations and processes that manage the data transformation from its sources (raw data) into valuable information [13, 15]. The following points illustrate these processes in Fig. 1:

- The proposed BI system is fed from a CSV file from a public data source.
- Data extraction, preparation, and integration were performed as staging operations. The most relevant data were extracted, cleaned, and transformed into a suitable format, and then data integration and modeling operations were applied to prepare data to be loaded into the data warehouse.
- At the data warehouse, a descriptive and predictive analysis was performed, in which machine learning algorithms were used for the predictive analysis.
- Finally, BI analytics are supported within interactive views using interactive dashboards and reports to represent the result as helpful information.

As indicated in Fig. 2, the architecture shows the basic technologies and how they fit together. In addition to a description of the products used to implement them [13, 15]. Fig. 2 illustrates the technologies and products used in the proposed BI system through the following points:

- ETL operations, including data extraction and transformation, were performed by Pandas, Numpy, and Datetime libraries in a Python environment using the Google Colab service.
- The integration and building of the data warehouse schema were implemented using Desktop Power BI.
- Predictive analysis was performed using machine learning algorithms within the python programming language using the Google Colab service. Various libraries were used.
- The used descriptive analytics illustrated through an interactive dashboard, implemented using Desktop Power BI tool functionalities.

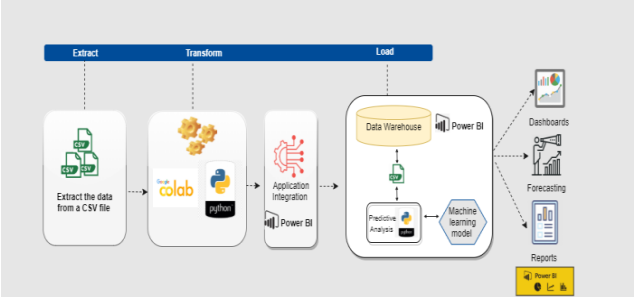

Fig 2. Technical and Products Architecture

IV. PREDICTIVE ANALYSIS

Predictive analytics is a proactive strategy that allows business stakeholders to use historical data to predict expected sales growth due to customer behavior changes. It can also help predict market trends, customer churn, and other scenarios; this can assist business stockholders in staying ahead of the curve, competing effectively, and gaining significant market share [10].

Various Machine Learning algorithms, including Linear Regression, Lasso Regression, Random Forests, XGBoost, and deep learning algorithms such as LSTM, can predict sales growth [16-19]. From a Global-Mart perspective, the company needs to build a robust sales prediction model for each market segment (Africa, Europe, North America, Asia, and Latin America) to predict the following week's sales. Fig. 3 represents the proposed solution for building the five prediction models. The solution was implemented with a Python environment and associated libraries, including Pandas, Numpy, Statsmodels, Seaborn, Matplotlib, Sklearn, and Keras. Data processing and building prediction model pipelines are described in the following sections.

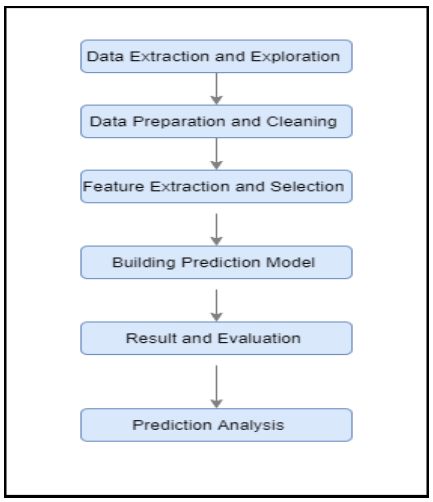

Fig 3. Prediction Model Building Pipeline.

A. Data Extraction and Exploration

The company needs to build prediction models for each market segment. Thus, the required data fields (Sales, Market segment, and Order date) were extracted from the data warehouse as a CSV file to build the models. The data represents daily sales over four years from (1/1/2014 - 1/1/2017), consisting of 51,290 transactions distributed over 1429 days. Fig. 4 illustrates the total sales distribution, where most data range from 3,000 to 12,000, with sales values skewed to the right. In subsequent operations, the distribution will be normalized to stationary form.

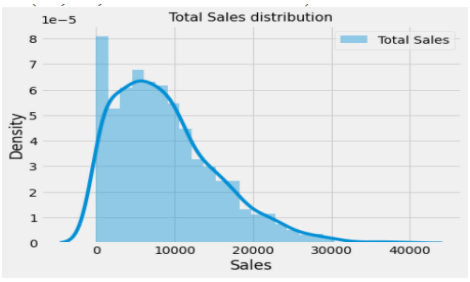

Fig 4. Total Sales Distribution.

B. Data Preparation and Cleaning

One of the most important preprocessing steps in any time series problem is to make the data stationary. Data preparation steps were done by applying a logarithmic transformation to each market segment to convert data into a stationary structure. Furthermore, to ensure the stationary structure, a standard statistical test was used, known as the Augmented Dickey-Fuller (ADF) test, where data is stationary if the p-value is less than 0.05 and the test statistic is negative [20, 21].

C. Data Preparation and Cleaning

The only features used from the dataset are the date and sales data, as additional features were derived from sales and date features to build the model. The derived features from the date are (year, month, day, day of the week, week, and week of the year). On the other hand, features derived from sales are known as lags and window statistics features. Lags features are the values of previous time steps, such as (the previous day or week). In contrast, the window statistics features are summary statistics values over a fixed window of previous time steps, such as (the mean or the max value of the last 3-days) [20, 21].

Moreover, three lag features were extracted, along with the mean, maximum, and minimum of the three previous days. The number of lags and the window size were chosen based on an experiment applied using the SelectKBest method from the sklearn library [22, 23] lags\window size were tested, while the size of 3 for both obtained the best result. This step was applied to all five market segments to build a stand-alone model for each.

D. Building Prediction Model

Five different machine learning algorithms were used to build an accurate sales prediction model. To test which one can perform efficiently on our data, the five algorithms are (linear regression and lasso regression from the statistical algorithms) [24, 25], XGBoost and RF from ensemble learning algorithms [26, 27], and LSTM from deep learning algorithms. These algorithms are chosen because they are widely used in sales prediction tasks [16, 17,18].

V. Experiments and Results

TABLE 2: BEST ACCURACIES OF ALL MACHINE LEARNING MODELS

Market	Measure Algorithms	MAPE (%)	RMSE (%)	MAE
Asia	***XGBoost***	**2.390**	**25.19**	**0.1660**
	LSTM	2.456	28.16	0.1893
	Random Forest	2.651	29.55	0.1990
	Lasso Regression	7.848	68.03	0.4680
	Linear Regression	6.488	65.81	0.4562
Europe	***XGBoost***	**2.538**	**32.26**	**0.1986**
	LSTM	3.085	33.27	0.2386
	Random Forest	3.428	37.20	0.2610
	Lasso Regression	6.171	68.03	0.4679
	Linear Regression	5.996	68.54	0.4535
Latin America	***XGBoost***	**3.205**	**29.25**	**0.2063**
	LSTM	3.684	34.96	0.2452
	Random Forest	3.645	32.56	0.2289
	Lasso Regression	7.844	65.31	0.4463
	Linear Regression	7.812	71.74	0.4794
North America	***XGBoost***	**3.470**	**31.10**	**0.2237**
	LSTM	4.408	41.61	0.2959
	Random Forest	4.843	39.59	0.2907
	Lasso Regression	8.690	75.99	0.5355
	Linear Regression	6.488	65.81	0.4562
Africa	***XGBoost***	**5.144**	**44.99**	**0.2798**
	LSTM	5.472	46.68	0.3086
	Random Forest	6.721	54.9	0.3428
	Lasso Regression	11.47	73.08	0.5298
	Linear Regression	13.04	96.10	0.6458

The model-building process followed the following steps: firstly, the dataset was split into 90% for training and 10% for testing (nearly the last five months). Secondly, all the models were trained on the training set, then evaluated on the test set, and finally, the algorithm that achieved the best

result was used to build the final prediction model. To assess the performance of the above models, the accuracy measures of Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percent Error (MAPE) are calculated as listed in Table 2.

As shown in Table 2, the XGBoost generates the best result compared to the other machine learning algorithms. Parameter tuning was performed on lasso regression, XGBoost, RF, and LSTM, to select the best parameters to build our models, the tuning task was done by using the Grid search method in the sklearn library except for LSTM, which is done by manual experiment. Table 3 illustrates the parameters selected for each model .

TABLE 3: MODELS PARAMETER TUNING

Method	Parameter Search space	Best parameter	
XGBoost	learning_rate= [0.01,0.1] max_depth= [3,4,5,11,13,14] n_estimators= (100, 150,200)	**Europe:**	learning_rate=0.1 max_depth=4 n_estimators= 150
		N-America	
		L-America:	
		Africa:	
		Asia:	learning_rate=0.1 max_depth=5 n_estimators= 150
LSTM	Unit = [20,40,50,60] Epoch= [100,200,250,500] Dense = [1] for all	**Europe:**	Unit =50 Epoch=200
		N-America	
		L-America:	
		Africa:	
		Asia:	
RF	max_depth= [3,4,5,11,13,14] n_estimators= (100, 150,200)"	**Europe:**	max_depth= 14 n_estimators=100
		N-America	
		L-America:	max_depth= 13 n_estimators=100
		Africa:	max_depth= 11 n_estimators=200
		Asia:	max_depth= 11 n_estimators=150
Lasso	alpha = (0.0, 0.1, 0.01)	**Europe:**	alpha = 0.1
		N-America	alpha=0.0
		L-America:	alpha=0.01
		Africa:	alpha=0.01
		Asia:	alpha = 0.0

VI. Business Intelligence Application

The final component of the BI solution is the BI Application, where an interactive dashboard is created to represent all specified business requirements. The dashboard designed for Global Mart's BI system contains all previous charts in one interactive view. The dashboard has filters to control overall data by years and quarters, along with filters to select the basic business performance metrics, including Sales, Profit, and Product Quantity, and filters to break down the data based on specific market segments or product categories.

The BI dashboard consists of three screens, shown in the following figures. Fig. 5 shows the first screenshot that provides an executive analysis of a company's overall performance. Fig. 6 screenshot provides customer and product analysis.

VII. Conclusion

This paper illustrated a general design and implementation of a BI framework for an online retail business. The implementation process was conducted through detailed lifecycle phases to create a BI application with valuable insights and analytics. Descriptive and predictive analytics were implemented to analyze the critical business aspects to support the decision-making process. This paper detailed the sales predictive analysis implementation process to forecast seven business days (one week). This was achieved by implementing five stand-alone prediction models using five machine learning algorithms, including Linear regression, Lasso regression, and XGBoost.

The XGBoost algorithm achieved the best result with a minimal RMSE of 25.19%. Both descriptive and predictive analytics results are presented in an interactive dashboard consisting of three informative, interactive, and insightful screens that the managers can utilize to support decision-making. Finally, using the BI framework in the retail business will provide high competitive advantages by generating comparative reports, quicker access to critical information, and help decision-makers to make accurate decisions.

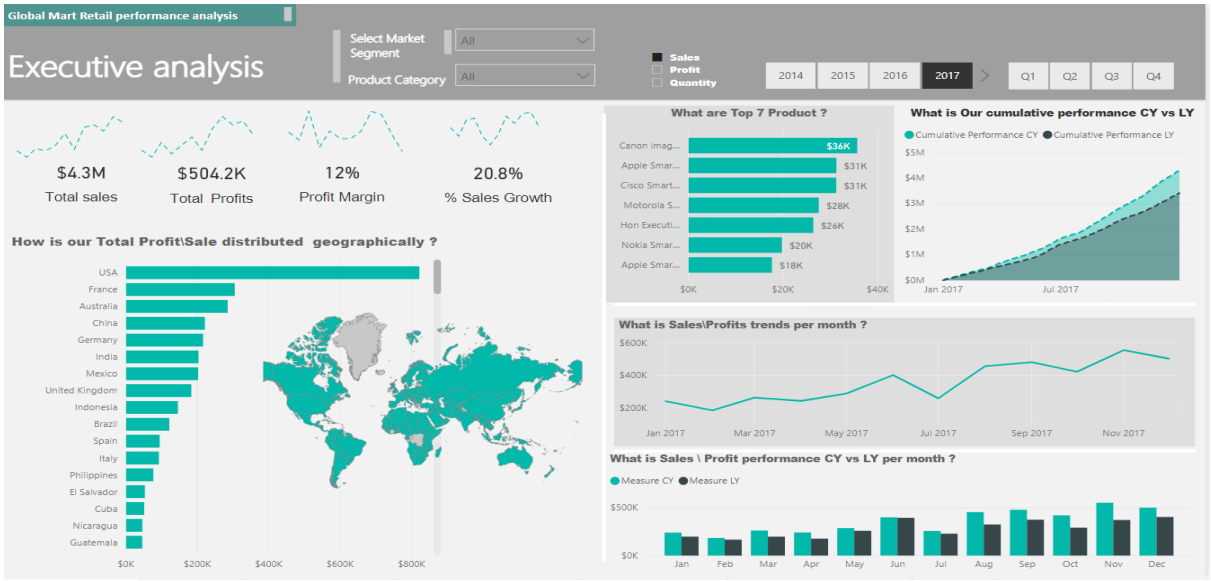

Fig 5. Dashboard Overall Company's Performance.

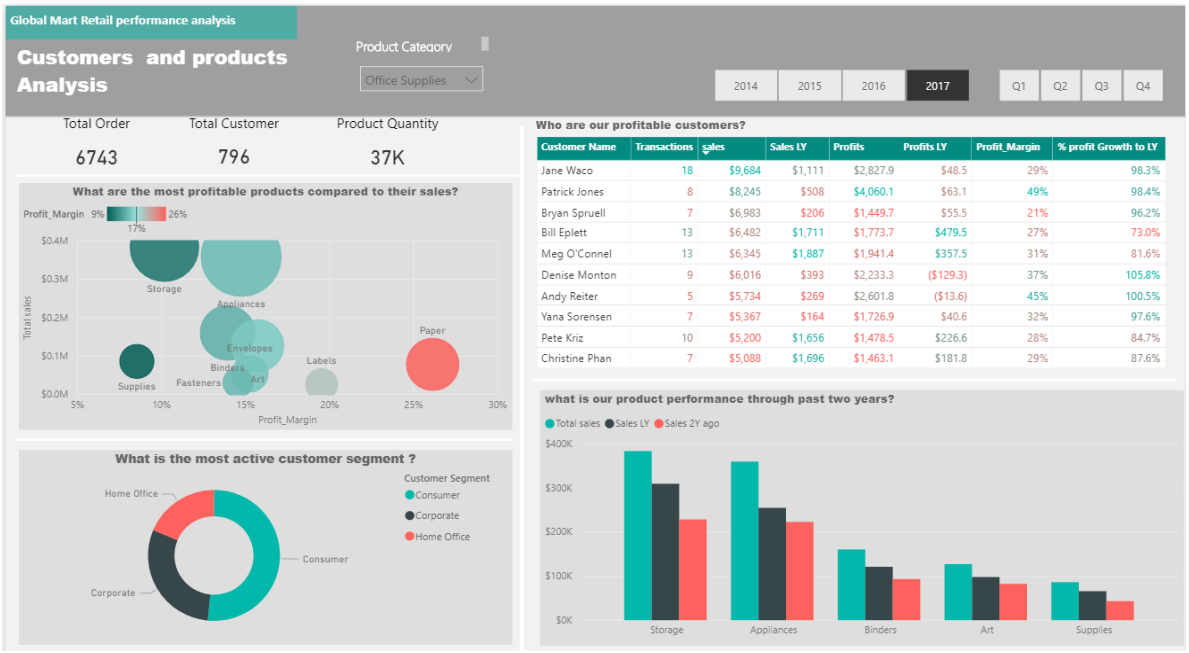

Fig 6. Customer and Products Performance.

REFERENCES

[1] S. Chandramana, "Retail analytics: driving success in retail industry with business analytics," *Res J Soc Sci Manag*, vol *7*, 2017, pp. 2251-1571.

[2] I. D. Kocakoç and S. Erdem, "Business intelligence applications in retail business: OLAP, data mining & reporting services," *Journal of Information & Knowledge Management*, vol *9*, no. 02, 2010, pp. 171-181.

[3] D. D., Phan and D. R. Vogel, "A model of customer relationship management and business intelligence systems for catalogue and online retailers," *Information & management*, vol. *47, no.* 2, 2010, pp. 69-77.

[4] A. Sato and R. Huang, "From data to knowledge: A cognitive approach to retail business intelligence," In *2015 IEEE International Conference on Data Science and Data Intensive Systems, 2015,* pp. 210-217.

[5] B. AlArmouty and S. Ki, "Data analytics and business intelligence framework for stock market trading," In 2019 2nd International Conference on new Trends in Computing Sciences (ICTCS), 2019, pp. 1-6.

[6] J. Ranjan, "Business intelligence: Concepts, components, techniques and benefits," *Journal of Theoretical and Applied Information Technology*, vol. *9, no.* 1, 2009, pp. 60-70.

[7] S. Williams and N. Williams, *The profit impact of business intelligence*, 2010, Elsevier.

[8] S. Fraihat, W. A. Salameh, A. Elhassan, B.A. Tahoun and M. Asasfeh, "Business Intelligence Framework Design and Implementation: A Real-estate Market Case Study," *ACM Journal of Data and Information Quality (JDIQ),* vol. 13, no. 2, 2021, pp.1-16.

[9] M. Awad, A. Al Redhaei and S. Fraihat, "Using Business Intelligence to Analyze Road Traffic Accidents," In *Central and Eastern European eDem and eGov Days (CEEeGov), September 22-23, 2022, Budapest, Hungary*. ACM, New York, NY, USA. https://doi.org/10.1145/3551504.355150

[10] H. Kumar, "Predictive Analytics Use Cases Retail Industry," Acuvate.Com. Accessed November 2021, from https://acuvate.com/blog/top-5-predictive-analytics-use-cases-retail-industry/

[11] M. Arora and D. Chakrabarti, |Application of Business Intelligence: A Case on Payroll Management," In *2013 International Symposium on Computational and Business Intelligence, 2013,* pp. 73-76. IEEE.

[12] A. R. Banjanagari and B. Vijaykumar, "Retail giant sales forecasting using machine learning," *Int. J. Recent Technol. Eng.*, vol. 8, no. 2 Special Issue 11, 2019, pp. 2408–2411.

[13] R. Sherman, "Business intelligence guidebook: From data integration to analytics" Newnes, 2014

[14] Y. Wang, S. Yu, and T. Xu, "A user requirement driven framework for collaborative design knowledge management," *Advanced Engineering Informatics*, vol 33, 2017, pp. 16-28.

[15] I. L. Ong, P. H. Siew, and S. F. Wong, "A five-layered business intelligence architecture," *Communications of the IBIMA*, 2011

[16] B. Lakshmanan, P.S.N Vivek Raja and V. Kalathiappan, "Sales. Demand Forecasting Using LSTM Network," Advances in Intelligent Systems and Computing, vol 1056, 2020, Springer, Singapore.

[17] A. Krishna, V. Akhilesh, A. Aich, and C. Hegde, "Sales-forecasting of Retail Stores using Machine Learning Techniques," In *2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS)*, 2018, pp. 160-166.

[18] B. Singh, P. Kumar, N. Sharma, and K. P. Sharma, "Sales Forecast for Amazon Sales with Time Series Modeling," In *2020 First International Conference on Power, Control and Computing Technologies (ICPC2T)*, 2020, pp. 38-43.

[19] B. M. Pavlyshenko, "Machine-learning models for sales time series forecasting," *Data*, vol 4, no. 1, 2019.

[20] J. Brownlee, "Deep learning for time series forecasting: Predict the future with MLPs, CNNs and LSTMs in Python," Machine Learning Mastery, 2018.

[21] J. Brownlee, "Introduction to time series forecasting with python: how to prepare data and develop models to predict the future," Machine Learning Mastery, 2017.

[22] G. Richards, W. Yeoh, A. Y. L. Chong and A. Popovič, "Business intelligence effectiveness and corporate performance management: an empirical analysis," *Journal of Computer Information Systems*, vol *59, no.* 2, 2019, pp. 188-196.

[23] P. Myers, "*Create date tables in Power BI Desktop - Power BI*," Microsoft Docs, Accessed November 2021 from https://docs.microsoft.com/en-us/power-bi/guidance/model-date-tables.

[24] *Lasso Regression*, Scikit-Learn.Org. Accessed November 2021 from https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html.

[25] *LinearRegression*, Scikit-Learn.Org, Accessed November 2021 from https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html.

[26] *XGBoost Regression*, Accessed October 2021, from https://xgboost.readthedocs.io/en/latest/python/.

[27] RandomForestClassifier, Scikit-Learn.Org, Accessed October 2021 from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.htm