Sales Prediction for a Pharmaceutical Distribution Company

[1]Smriti Keny, [2]Sagarika Nair, [3]Silka Nandi and [4]Deepak Khachane
[1,2,3,4]New Horizon Institute of Technology and Management, Thane,
University of Mumbai, India

Article Info

Article history:

Received Oct 30, 2020
Revised Jan 6, 2021
Accepted Jan 12, 2021

Keywords:

PDCs,SVR,
Inventory management,
Sales prediction,
Data Analysis,
Data Visualization.

ABSTRACT

The study aims to find an appropriate model to extract insights from the sales of a Pharmaceutical Distribution Company (PDC) and make it available in an interactive and readable manner for the company. In PDCs, it is highly important to obtain a good approximation of the medicine needs, due to the short shelf life of many medicines and the need to control stock levels. The presented method is a combination of analysis and interactive visualization tools along with prediction. In this paper, we explore the use of Support Vector Regression algorithm for the sales prediction of individual products. The proposed model helps to present the sales data in a better way such that understanding the trends and seasonality becomes easier for the PDCs. The dataset has information of hourly, daily, weekly and monthly sales of the drugs and hence the end results also give us a likely classified understanding of the sales. The study of the results obtained, suggest that the proposed model may be considered appropriate for product sales prediction.

Corresponding Author:

Smriti Keny
New Horizon Institute of Technology and Management, Thane,
University of Mumbai, India
Email: smritikeny172@nhitm.ac.in

1. INTRODUCTION

Sales prediction is an important part of business process management. Despite the difficulties and execution of forecasting processes across different businesses, the intended purpose of sales prediction stays the same: obtaining almost accurate estimation of a product or service, given the past data and the current state of the environment (e.g., political, social, economic), to plan and organize businesses accordingly. It plays a critical role in logistics and supply chain management. The purpose of our project is to examine pharmaceutical sales data and draw results from the data in an understandable and helpful format. According to the problem of having many new items with short number of historical records, and having great diversity of medicines, common prediction methods are mainly inappropriate or not that much effective for PDCs. The basic objective of this research is to give an exceptional and accurate sales prediction method to help PDC, to predict product sale which should help to tune inventory management policies in order to prevent costs of excessive inventory and prevent losing customers due to drug shortage. Considering the situation, PDCs are compelled to meet their customer needs by delivering right number of medicines at the right time. Scarcity and excess of goods can lead to loss of customers or income for PDCs. The prediction and analysis of their data, will help them with future planning of sales and thereby reduce excess inventory costs and increase profits for PDCs by keeping the customer satisfied. Precise sales analysis enables companies to make informed business decisions and perform short-term predictions. The purpose of our project is to predict, analyze and visualize sales data for a PDC using SVR algorithm and several python tools and libraries.

2. RESEARCH METHOD

The basic objective of the researches was to offer a novel and precise sales prediction to forecast product sale and tuning inventory management policies an aid to manage promotions may increase the frequency of prescribing and price growth may decrease it. When it comes to forecasting sales in a Pharmaceutical business it is a complex task Researchers have extensively worked and examined alternative methods to find out the most efficient sales prediction methodology. They have used different methods and algorithms to do the same. Linear Regression models were widely used in prediction problems. Linear models, such as autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA), are successful and famous linear methodologies, but their prediction ability is restricted by their supposition of a linear behavior [1]. It was also observed that the sales record available to work with was of very recent past. The forecast of sales was made based on the analysis of a historical data 24 months and for a 3-month time horizon [2]. Most of the methods used consisted of non- linear and stationary time series forecasting. To deal with the problem of non-linearity, Artificial Neural Networks (ANN) were introduced to regression problems.

The last few decades have seen the superior performance of Artificial Neural Networks (ANN) in classification and regression problems and have received focused of attention in the time series forecasting methods [3]. But ANN had certain limitations too, such as duration of development and amount of data required was quite high. It is necessary to build an accurate model for forecasting sales of pharmaceuticals using one of the methods of machine learning, taking into account the constant updating of medicines and the lack of sufficient data on past sales of preparations of each kind [4]. Results from Anusha et al., (2014) indicate that among Indian pharmaceutical retailers, guestimates were the most common forecasting technique applied and such results which were usually unreliable often resulting in unavailability of life saving commodities, which could be fatal or lead to the loss of customer loyalty and eventual profit reduction [5]. The need for a non- linear and efficient model for sales prediction was evident.

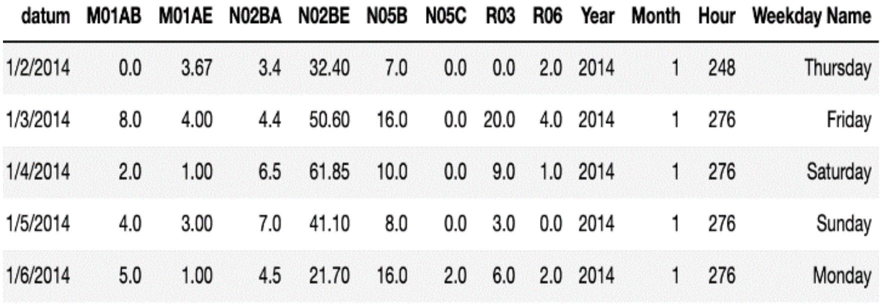
3. PROPOSED METHODOLOGY

The application of supply chain analytics to the analysis of complex data sets provides supply chain managers with the ability to respond to relevant problems in a timely manner by providing accurate business insights [5]. The overall procedure consists of Data collection, Treatment and Analysis, Modelling, Iterating and Deployment. The dataset is available on Kaggle. To approach the objective of the research, relevant data has to be extracted and tested for missing values and outliers. When the data is ready, algorithms are to be tested, to find the one that fits well such that, satisfactory results of the predicted values with more accuracy and less errors are obtained. The Analysis and Visualization of data can be done for the data to be readable. Once the model is prepared after enough training and testing, the results have to be presented in an interactive manner, for the users to be able to understand and study the data. For live data, continuous monitoring of the model is necessary to obtain updated and precise values

4. DATASET

The data set consists of hourly, daily, weekly and daily sales data of the medicines. It was built from the initial dataset which consisted of 600000 transactional data collected in 6 years (period 2014-2019), indicating date and time of sale, pharmaceutical drug brand name and sold quantity, exported from Point-of-Sale system in the individual pharmacy. Selected group of drugs from the dataset (57 drugs) is classified to the following Anatomical Therapeutic Chemical (ATC) Classification System categories:

Table 1. weekly Dataset



datum	M01AB	M01AE	N02BA	N02BE	N05B	N05C	R03	R06	Year	Month	Hour	Weekday Name
1/2/2014	0.0	3.67	3.4	32.40	7.0	0.0	0.0	2.0	2014	1	248	Thursday
1/3/2014	8.0	4.00	4.4	50.60	16.0	0.0	20.0	4.0	2014	1	276	Friday
1/4/2014	2.0	1.00	6.5	61.85	10.0	0.0	9.0	1.0	2014	1	276	Saturday
1/5/2014	4.0	3.00	7.0	41.10	8.0	0.0	3.0	0.0	2014	1	276	Sunday
1/6/2014	5.0	1.00	4.5	21.70	16.0	2.0	6.0	2.0	2014	1	276	Monday

M01AB - Anti-inflammatory and antirheumatic products, non-steroids, Acetic acid derivatives and related substances

- M01AE - Anti-inflammatory and antirheumatic products, non-steroids, Propionic acid derivative.
- N02BA - Other analgesics and antipyretics, Salicylic acid and derivatives
- N02BE/B - Other analgesics and antipyretics, Pyrazolones and Anilides
- N05B - Psycholeptics drugs, Anxiolytic drugs
- N05C - Psycholeptics drugs, Hypnotics and sedatives drugs
- R03 - Drugs for obstructive airway diseases
- R06 - Antihistamines for systemic use.

Sales data are resampled to the hourly, daily, weekly and monthly periods. Data is already pre-processed, where processing included outlier detection and treatment and missing data imputation.

5. COMPARISON OF DIFFERENT ALGORITHM

Comparison of different algorithms, commonly used for prediction purposes compared based on different factors, such as Linear Regression, Polynomial Regression and SVR will help get a clear idea of which one is to be selected. Linear regression is a method of modelling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables) denoted X in statistics. By applying a linear model to a linear dataset, a decent result can be obtained the same, like in Simple Linear Regression, but when the same model is applied to a non-linear dataset without any modifications, a dramatic result is obtained. As a result of the increased loss function, the error rate will be high, and accuracy will be reduced. In such cases, where data points are organized non-linearly, the Polynomial Regression model is required. Support Vector Regression.

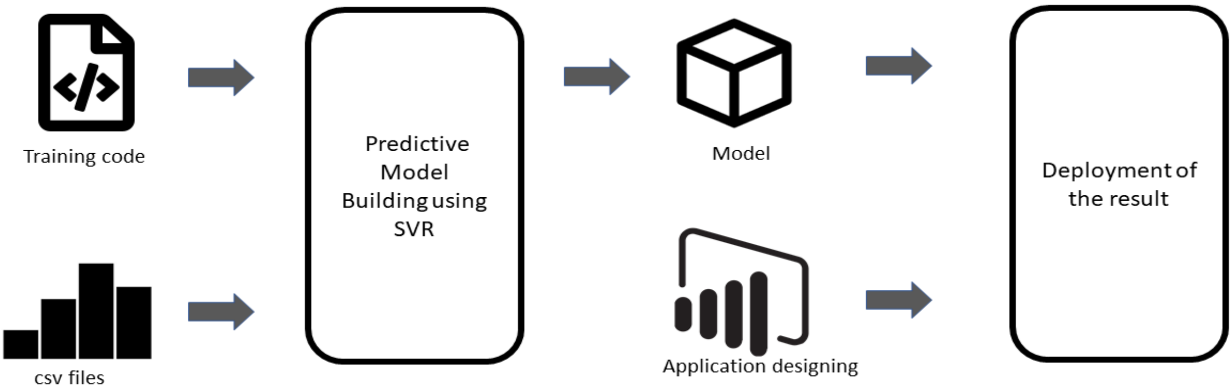


Figure 1. Architecture

(SVR) is a supervised machine learning algorithm that is mostly used to divide data into categories. It is derived from the popular classification algorithm SVM. SVM, unlike most algorithms, employs a hyperplane as a decision boundary between the different groups. SVM can be used to create several separating hyperplanes, splitting the data into segments with only one form of data in each segment.

After comparing Linear Regression, Polynomial Regression and SVR prediction functions along with the calculated accuracy and error values the effectiveness of the model can be obtained.

5.1. SELECTION OF MODEL

The most important thing to ensure that your model is adequately tested, is to not train it on the entire dataset. 70 percent of the data is used for preparation, and 30 percent is used for testing in a traditional train/test split. In order to avoid over fitting to the training set, it is critical to evaluate the model with new data. However, it is often helpful to analyze the model while constructing it in order to find the best parameters, but this cannot be done with the test set because you might end up choosing the parameters that work best on the test data but not necessarily the parameters that generalize best. Instead, generate a third subset of data known as the validation set to test the model while it is still being built and tuned. 60 percent of the data will be used for preparation, 20% for validation, and 20% for testing in a traditional train/test/validation split. After getting the accuracy for all the three models, SVR gave the highest accuracy for prediction.

6. VISUALIZING DATA

6.1. DATA VISUALIZATION

Data visualization makes it easy to see and understand trends, outliers, and patterns in data by using visual elements including charts, graphs, and maps Businesses produce and collect vast quantities of data, and it is up to marketing and sales departments to bring it to good use. Data visualization is the method of transforming a vast amount of data into a digital representation that highlights its meaning

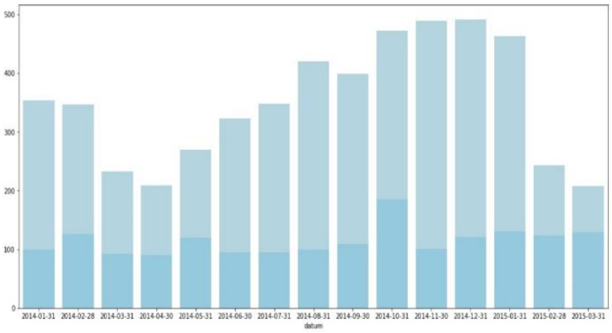


Figure 2. Bar plot

6.2. DATA ANALYSIS

Cleaning, transforming, and modelling data to discover useful knowledge for business decision-making is known as data analysis. Data analysis' aim is to derive valuable knowledge from data to make decisions based on that information.

7. BUILDING THE MODEL

The model building phase consists of training and testing the clean data, using SVR and then deploying the final results on an interactive dashboard. With the help of the supervised learning algorithm SVR, and the popular scikit-learn Python library, the prediction model can be trained, using the labeled input data. The model should then be integrated into a simple web application using dash or similar framework. The prediction of sales of the medicines in grams, for the next few month should be obtained. The visualization of the results can be done using the python library Matplotlib. The Visualized data helps get a clear representation of the trends the customers followed over the years. For analysis, the use of numpy library is suggested. The user could find the highest and lowest selling medicines at different time, using the Analyzed data. If studied the sales of the drugs with the presented data, stock management would be much easier for the PDCs.

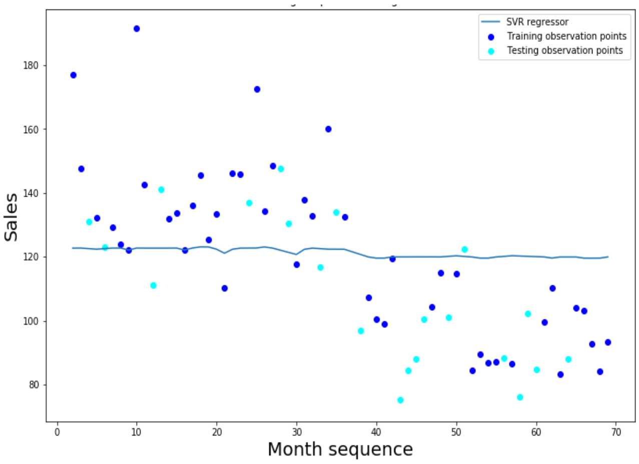
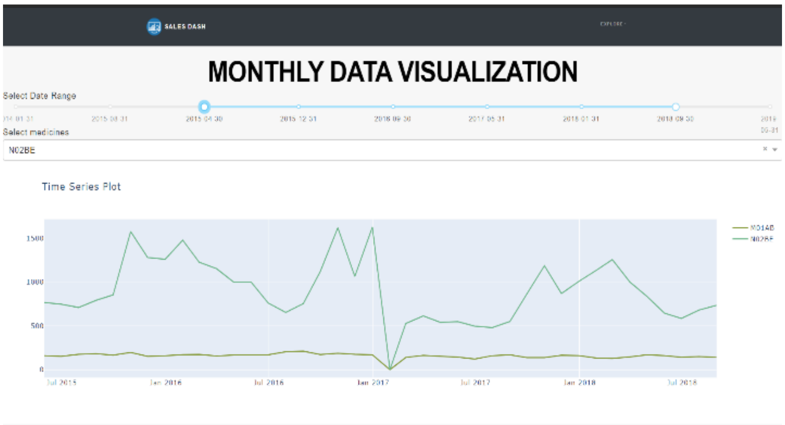
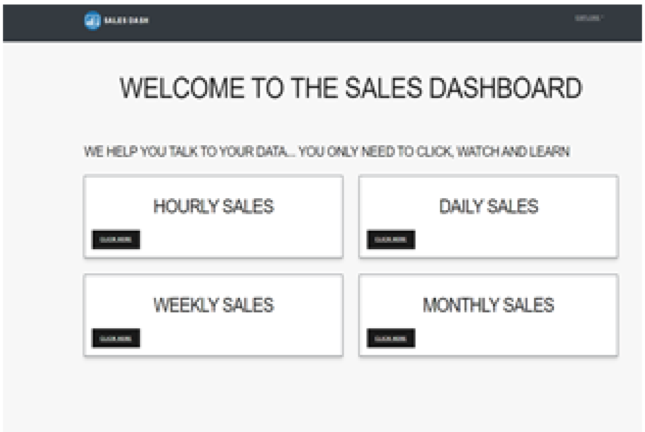


Figure 3. Prediction using support vector regression (SVR)

8. FINAL RESULTS

The results were deployed on a website using Dash. The interactive website allows the users to view the analysis, visualization and the prediction of the data required for inventory management.





WE HELP YOU TALK TO YOUR DATA... YOU ONLY NEED TO CLICK, WATCH AND LEARN

9. FUTURE RESEARCH

Despite the results obtained have been satisfactory, it is considered that there is still room to introduce improvements in the technique of forecast. On the one hand, it would be interesting to see if, using a broader forecast horizon than that was used in the work, the results obtained could have a precision that could be considered even more acceptable. The inclusion of a broader forecast horizon (for example, adding more techniques for data visualization, analysis and generating report) in the sales forecast, the pharmaceutical distribution company may achieve a greater room for inventory management as well.

10. CONCLUSION

The proposed model shows how SVR can be used in sales forecasting. From the study and application carried out during the work, it is possible to conclude that the performance of the Web Application using SVR and other data visualizing/ analyzing methods was favourable in forecasting product sales for individual products, obtaining results that are closer to the real ones and more reliable for the company.

ACKNOWLEDGEMENTS

We express our gratitude to everyone who supported us throughout the course of research project report. We are thankful for the aspiring guidance, invaluably constructive criticism and friendly advice during the project work we are sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to the project. We are also thankful to our Principal, Dr. Prashant Deshmukh and management of NHITM for their support and encouragement. We are highly indebted to our Head of Department, Dr. Sanjay Sharma and guide Mr. Deepak Khachane for their guidance and constant supervision as well as for providing necessary information regarding the project and also for the support in completing the project. We are thankful to the review committee for their valuable suggestions and feedback. We also thank Laboratory staff for their valuable support. Last but not least, we sincerely thank those from teaching and non-teaching staff from NHITM who were somehow attached with our endeavor.

REFERENCES

[1] Khalil Zadeh, Neda &Sepehri, Mohammad Mehdi &Farvaresh, Hamid. (2014). “Intelligent Sales Prediction for Pharmaceutical Distribution Companies: A Data Mining Based Approach.” Mathematical Problems in Engineering. 2014. 1-15. 10.1155/2014/420310.

[2] Ribeiro, I. Seruca and N. Durão, "Sales prediction for a pharmaceutical distribution company: A data mining-based approach," 2016 11th Iberian Conference on Information Systems and Technologies (CISTI), Las Palmas, 2016, pp. 1-7, doi: 10.1109/CISTI.2016.7521397.

[3] Keerti Nilesh Mahajan et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, “Business Intelligent Smart Sales Prediction Analysis for Pharmaceutical Distribution and Proposed Generic Model”, Vol. 8 (3), 2017, 407-412

[4] A G Kravets, M A A,-Gunaid, V I Loshmanov, S SRasulov, L B Lempert, Model of medicines sales forecasting taking into account factors of influence. International Conference Information Technologies in Business and Industry 2018.

[5] C. I. Papanagnou and O. Matthews-Amune, "An Estimation Model for Hypertension Drug Demand in Retail Pharmacies with the Aid of Big Data Analytics," 2017 IEEE 19th Conference on Business Informatics

BIOGRAPHIES OF AUTHORS

	Smriti Keny is a final year student of Computer Engineering at New Horizon Institute of Technology and Management, Thane which is affiliated to the University of Mumbai.
	Sagarika Nair is a final year student of Computer Engineering at New Horizon Institute of Technology and Management, Thane which is affiliated to the University of Mumbai.
	Silka Nandi is a final year student of Computer Engineering at New Horizon Institute of Technology and Management, Thane which is affiliated to the University of Mumbai.

	Deepak Khachane is a professor at the Computer Department at the University of Mumbai affiliated College New Horizon Institute of Technology and Management. He also holds a Masters in Computer engineering from University of Mumbai along with 10 years of teaching experience