İlker Poyraz, et. al. International Journal of Engineering Research and Applications
www.ijera.com
ISSN: 2248-9622, Vol. 10, Issue 7, (Series-III) July 2020, pp. 51-54

RESEARCH ARTICLE OPEN ACCESS

Drug Demand Forecasting for Pharmacies with Machine Learning Algorithms

İlker Poyraz[1], Ahmet Gürhanlı[2]
[1](Department of Computer Engineering, Istanbul Aydın University, Turkey
[2](Department of Computer Engineering, Istanbul Aydın University, Turkey

ABSTRACT
Today, with the developing technology, accurate and reliable demand forecasts play an important role in increasing the productivity of businesses. With reliable predictions made for the future, businesses; They have the opportunity to effectively plan sales and marketing activities, price policies, factors related to environmental changes. As in every sector, the issue of demand forecasts in pharmacies has an important place in the health sector. Drugs that play a key role in improving health behaviors and balances deteriorated in human life are increasingly purchased and consumed today. The aim of this study is to estimate the sales amounts in the next periods by using the data for drug purchase in a pharmacy. Thanks to these estimates, storage status of drugs can be taken under control over the drugs consumed in the pharmacy. In this study, a pharmacy in Turkey, 5 years of pharmaceutical sales data are arranged between the months of December 2019 and January 2015 and the Weka program with time series on a weekly basis to do the forecasting work by applying machine learning algorithms linearregresyo's, gaussianprocess, m5rules, multilayerperceptro's, smoreg, M5P, randomforest is used By comparing the average absolute percentage error (MAPE) of these algorithms, it has been tried to find the most successful prediction model.
Keywords: Machine Learning, Forecasting, Time Series, Drug Sales Forecasting

Date of Submission: 23-06-2020 Date of Acceptance: 11-07-2020

I. INTRODUCTION

In the past years, the flow of information in the pharmaceutical industry was relatively simple and technology implementation was limited. However, as we move into a more integrated world where technology has become an integral part of business life, the process of transferring information has become more complex. With today's technology, it is increasingly being used to help companies in the pharmaceutical industry manage their stocks and develop new products and services.

Data mining is the research and analysis of big data to discover meaningful patterns and rules. Time series analysis is one of the statistical analysis techniques frequently used in data mining. Time series are frequency series, and the frequencies of the series can take values varying annually, quarterly, monthly, weekly and daily [1]. Time series analysis is based on the behavior of historical data over time. The purpose of this is to make predictions about the future by looking at the historical data. It is used to make predictions about the future by examining the changes of the past data rather than the cause-effect relationship such as time series analysis and regression analysis. In the time series analysis, by looking at whether there is a certain trend by examining the past sales of an enterprise, making demand forecasts about the future [2]. Possible future events can be determined by time series analysis. Forecasting can also be carried out with time series analysis. It is especially important on financial data and strategic management decision making stages [3].

Linear Regression Algorithm

Defining the linear relationship between two variables as a line equation, When one of the values is known, it provides estimation about the other. In order to make the correct estimation among the data, it is necessary to create the best truth for the data. While creating the best line, the region closest to all points should be preferred. Linear

Since we will create a line in regression, we work on two variables, one dependent and one independent variable [4].

Gaussian Process Algorithm

Gauss processes are a statistical classification-prediction algorithm. In learning Gaussian processes algorithm, applies Bayesian Gaussian processes for nonlinear regression.

Nonparametric Bayesian Gauss approach; A smooth and continuous function f, which provides a high correlation between the outputs according to the proximity of the inputs, places a direct distribution on the space of nonlinear functions, and associates the data with the function outputs, using all possible probability sequences of the inputs. It provides supervised learning through final inference on functions before Gauss processes [5].

M5Rules Algorithm

The algorithm uses sperate and conquer technique. Creates a regression model or regression tree using the decision list for numerical prediction. In the regression tree, the leaves of the tree are the average of the samples reaching the leaf. It works better than a single linear regression equation, but model trees have been developed because the tree and its dimensions are large and difficult to interpret. In model trees, the regression equation is combined with regression trees. The leaves of the model trees are linearly expressed regression equations instead of the only estimated value [5].

Multilayer Perceptron Algoritması

While perceptrons are suitable for linear separable problems, the Multilayer perceptron (multilayer sensor) is also suitable for non-linear classifications. Multi-Layer Sensors (MLP) emerged as a result of efforts to solve the XOR Problem. Many problems in real life are nonlinear. MLP networks are the most frequently used ANN model in the solution of nonlinear problems. MLP The most popular learning method for web networks Back Propagation (back propagation) method. Back Propagation was first proposed by Werbos in 1974, the currently used version was developed in 1986 by Rumelhart, Hinton and Williams [6].

SMOreg Algorithm

This algorithm performs regression by applying it on Support Vector Machines. The SMOreg algorithm basically uses methods known as support vector machines. DVMs are a method of machine learning based on statistical learning theory, developed by Vladimir Vapnik and Alexey Chervonenkis in the late 1960s. The basic logic of DVM is to determine the best separator plane for data structures that can be separated linearly. DVMs can also classify new data that were not observed during training without any problems. This shows the ability of the support vector machines to generalize [7].

M5P Algorithm

M5P is a method derived from the decision trees method. It is used for processes such as classification and regression by forming many decision trees and combining their results. It is a useful and simple method that is used mostly in decision-making needs. The final result is expressed in decision tree algorithms and the way the rule is created is clear. Also, the advantage of other methods is that decision trees are more understandable [8].

Random Forest Algorithm

Random Forest is one of the popular machine learning models since it gives good results even without hyper parameter estimation, and is applicable to both regression and classification problems. overfitting decreases [9]. The Random Forest algorithm can be used in both classification and regression problems, such as the decision tree. Working logic creates more than one decision tree. The average in these decision trees when it will produce a result value is taken and the result is produced [10].

II. PURPOSE

This study was prepared by using 5-year pharmacy sales data from January 2015 to December 2019. By applying machine learning techniques to the data set, it was aimed to estimate the data for drug purchase in a pharmacy and the subsequent sales quantities. It is aimed to use the results of the sales forecast according to the algorithm that gives the best results by using 7 machine learning methods to evaluate the sales data related to the previous years.

III. SCOPE

Sustainability of pharmacy services is of great importance in order to maintain the continuity of health services without interruption in human life. It is important to ensure the continuity and sustainability and to manage the demand-related amount in stock management in a rational and systematic manner without sacrificing service quality. When the machine learning methods used in this study were integrated into pharmacies, the importance of inventory control and order management over drugs consumed by purchasing in the pharmacy was emphasized thanks to drug-based estimations, and a prediction model was proposed for inventory and order management.

IV. METHOD

Since we will make the weekly estimation of 5-year Pharmaceutical sales data between 01.01.2015 - End date 31.12.2019, we obtained from the pharmacy, and the weekly sales data was created to be the sum of the sales data, so that the Weka Arff file format is compatible with the structure. While

choosing the best 10 drugs, we took the top 10 drugs in the 5-year sales data set and analyzed the weekly time series based on the total 5-week weekly sales. The study was carried out on a laptop with 16 GB RAM, Intel Core (TM) i7-8550 CPU 1.80GHZ Processor. 3/4 (75%) of the 261 Weekly data based on the sales data prepared was used for training and the remaining 1/4 (25%) was used for the test. We make the .arff file structure, which we have created based on the sales data we will estimate in Wekada, by using the parametric values to take into account in the Forecasting field in the Configuration parameters in the Wekada Forecast Tab. While using these, we reach the Min and Max Lag value by increasing the max lag value by evaluating the weekly values that we previously determined as periods over 52 weeks in a year. The delay generation panel allows the user to control and change how delayed variables are generated. Delayed variables are the main mechanism by which the relationship between past and present values of a series (Weekly) can be captured by suggestion learning algorithms. As a result of the studies on Machine Learning algorithms in Weka application, an estimation was made for 3 weeks and weka was tried to catch the minimum error rate with the given parameters. In this study, it is aimed to increase the accuracy rate by using different algorithms to catch the minimum error rate.

V. RESULTS

When we analyze the time series analysis results of a random drug among the 10 drugs that were estimated, the 1-week sales amount we received from the test data set had the best error rate between the actual and estimated sales quantity results of the drug named "DEVIT-3 BULB" that occurred in the forecast results. When we remember the working principle of the 5Rules algorithm, we can express that it creates a regression model or regression tree by using the decision list for the numerically predicted prediction.

Table 1: Actual and Predictions DEVIT-3 AMPUL

Algorithms	Actual	Predicted	Error
Linear Regression	30	32.9234	2.9234
Gaussian Processes	30	31.7304	1.7304
SMOreg	30	22.8297	-7.1703
M5P	30	32.9234	2.9234
Random Forest	30	24.8966	-5.1034
Multilayer Perceptron	30	20.6701	-9.3299
M5Rules	30	29.4488	-0.5512

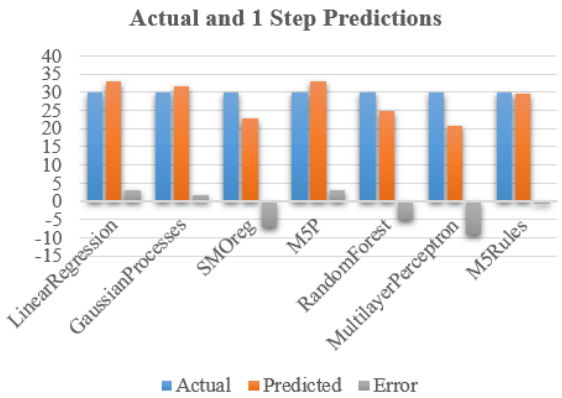

Figure 1: Algorithm Predictions Comparison

When we look at other algorithms, it is seen in Table 1 that Linear Regression and GaussianProcesses algorithms produce close estimation results. Linear Regression algorithm is defined as a straight line equation between two variables, and when one of the values of the variable is known, the Gauss Process Algorithm is a statistical classification. These are algorithms that operate on prediction as a prediction algorithm.

Table 2: Mean Absolute Percentage Error Results DEVIT-3 AMPUL

Algorithms	1-Step-ahead	2-Step-ahead	3-Step-ahead
Linear Regression	32.2101	32.6641	32.9555
Gaussian Processes	31.6892	32.112	32.3981
SMOreg	33.5285	33.7108	33.7522
M5P	32.7653	33.2706	33.6094
Random Forest	40.3129	40.8241	41.1429
Multilayer Perceptron	42.6856	42.5711	42.388
M5Rules	32.1142	32.5082	32.7854

When we examine the Table 2 above according to the results of Mean Absolute Percentage error, the best estimate in the 1st week forecasts is the result of the Gaussian Processes algorithm, which is the closest result to zero as a result of 31.6892, for the first week.As a result, the best estimate in the 2nd week forecasts is the closest result to zero as a result of 32,112, the result produced by the Gaussian Processes algorithm is the best result for the 2nd week and the best estimate for the 3rd week forecasts is the Gaussian Processes

algorithm, which is the closest result to zero as a result of 32,391. We can express the result it produces as the best result for the third week. The Gaussian Processes algorithm is seen as the algorithm that gives the best results in the 3-week series.

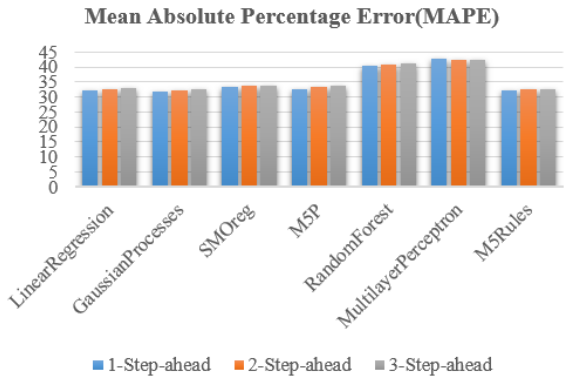

Figure 2: Algorithm MAPE Comparison

Mean absolute percentage error values are evaluated differently according to numerical ranges in different sources. We can express these evaluations as follows.

Mean absolute percentage error review

I. Models whose value range is below Mape <10% are referred to as "**very good**".
II. Models with a value range of 10%<Mape<20% are considered to be "**good**".
III. Models with a value range of 20% <Mape <50% are "**acceptable**"
IV. Finally, models above 50%<Mape are referred to as "**faulty and wrong**".

VI. CONCLUSION

The best result of the "GaussianProcesses" algorithm in the "Mean absolute percentage error" values in the 3-week estimation results of the "DEVIT-3 AMPUL" drug, which has the highest sales amount in the 5-year time period we obtained from the Pharmacy drug Sales data, on 7 algorithms. In the 3-week estimation, the "M5Rules" algorithm gives the closest values to the "GaussianProcesses" algorithm in the 1st, 2nd and 3rd week estimation. We have examined the "M5Rules" algorithm in Table 1 as the algorithm that yields the best results on the quantitative basis. It is seen that "M5Rules" algorithm produces successful results in time series as performance.

REFERENCES

[1]. Time Series Analysis Access Address: https://ekonometrice.blogspot.com/2015/04/zaman-serisi-bolum1.html?m=1 April 04, 2015
[2]. Dr. Aysel Çetindere Filiz Faculty of Economics and Administrative Sciences Access Address : https://avys.omu.edu.tr/storage/app/public/aysel.cetindere/131686/11.%20Hafta%20Talep%20Tahmin%20Y%C3%B6ntemleri.pdf
[3]. Şeker, Sadi Evren(2015). Time Series Analysis, MIS Encyclopedia, Volume 2, Number 4.
[4]. Ecem Bölük(2019) What are Regression Algorithms? Access Address: https://medium.com/womancoder/regresyon-algoritmalar%C4%B1-nedir-2-e66b12907b57
[5]. Esin Erguvan ETGİN (2017) Implementation of the efficiency and efficiency of data mining forecasting algorithms in Time Series on bist100 stocks
[6]. Dr. Gökçen UYSAL(2019) Hydrological Modeling with Artificial Neural Networks
[7]. Zeynep Behrin Güven Population Growth Prediction Application Using Time Series Mining
[8]. Evin GARİP (2017) Estimating CO2 Emissions in OECD Countries by Machine Learning
[9]. Hakkı Kaan Simsek(2018) Machine Learning Lessons 5a: Random Forest (Classification) Access Address: https://medium.com/data-science-tr/makine-%C3%B6%C4%9Frenmesi-dersleri-5-bagging-ve-random-forest-2f803cf21e07
[10]. Ekrem Hatipoglu(2018) Machine Learning — Classification — Decision Tree — Random Forest Access Address: https://medium.com/@ekrem.hatipoglu/machine-learning-classification-decision-tree-random-forest-part-12-8c9515d811b9