International Journal of Latest Trends in Engineering and Technology
Special Issue SACAIM 2017, pp. 095-098
e-ISSN:2278-621X

MACHINE LEARNING APPROACH FOR THE PREDICTIVE ANALYSIS OF SALESIN GROCERY STORE

Kavya Ramesh[1], Jithin Jose Mathew[2] & Hemalatha N[3]

Abstract- The main aim and intention of this study is to understand the pattern of sales in grocery stores. Machine learning is an advancing field with its implementation of various aspects of business environments. Here we focus on the implementation of Machine Learning algorithms like Linear Regression in small scale business like a store with day to day grocery movement for predicting the sales for the upcoming month. Analytical techniques like Linear Regression, Generalized Linear Model, and Artificial Neural Networks were tested in order to come up with accurate result and to identify the most suitable technique with best predicting results. The Linear Regression with feature selection M5 prime, min tolerance 0.05, ridge 1.0E-8 forecasted the result with highest accuracy with just 2.307 +/- root mean square error. With real time training and testing of data, a clear pattern could be obtained which helps in the analyzing the performance and predicting the sustainability of the business. Understanding the sales pattern will also help predict the product movement which in turn can be utilized to bring in the product with accurate quantity thereby minimizing the wastage.
Keywords- Machine Learning, Predictive analysis, Linear Regression.

1. INTRODUCTION

In earlier years, data processing techniques like machine learning were developed in the field of information science, which are being applied in various practical areas today[1]. These data mining techniques, are an area of interest, and researchers continue to discover ideas and opinions which can be useful to businesses from the large volumes of customer information data accumulated by businesses using these techniques.Machine learning is effective for data with strong insufficiency. Due to the strong competition in the present scenario, most retailers are in an effort to increase profits and reduce the cost[1].

A Predictive analysis on retail stores is an efficient way to achieve the aforementioned goals and lead to improve the customers' satisfaction, reduce destruction of products, increase sales revenue and make production plan efficiently. The purpose of this research is to construct a sales prediction model for retail stores as a way to introduce the machine learning approach in the field of marketing so that retailers can use it for marketing strategies. The research constructs a sales prediction model from the point-of-sale (POS) data of a supermarket and proposes a method for using the advanced data analysis of machine learning for sales analysis [1].An effective method can help the decision makers in this case, retailerscalculate the production costs.

2. PROPOSED MODEL

2.1 Linear Regression

Simple Linear Regression analysis is the simplest form of Regression where the data is fit into a line $y = \beta 0 + \beta 1x$, where x is the dependent variable and y is the dependent or response variable. In a probabilistic model for linearly related data, paired data points are (x1, y1),(x2, y2), . . . ,(xn, yn), where we assume that as a function of xi , each yi is generated by using some true underlying line $y = \beta 0 + \beta 1x$ that we evaluate at xi , and then adding some Gaussian noise. Therefore $yi = \beta 0 + \beta 1xi + \varepsilon i$, where εi is the noise in the data which shows the fact that data cannot fit the model perfectly. This is due to the fact that real world data in any research analysis never form a straight line.

2.2 Generalized Linear Models (GLMs)

GML usually refers to conventional linear regression model which include Multiple linear regression, ANOVA, ANCOVA [3].
Three components of GLM:

2.2.1 Random Component –
Probability distribution of the responsevariable (Y)

2.2.2 Systematic Component –
Specifies the explanatory variables (X_1, X_2, ... X_k)

2.2.3 Link Function, η or g(μ) –
Link between random and systematic components

[1] II M.Sc ST, AIMIT, Beeri, Mangaluru, India
[2] II M.Sc BI, AIMIT, Beeri, Mangaluru, India
[3] Department of BioInformatics, AIMIT, Beeri, Mangaluru, India

3. METHODOLOGY

The analysis was carried out on a set of items available in Supermarket like Dairy, Grocery, beauty products etc. Data for a series of months starting from January to June was trained in the training model and the test sample was applied to forecast the outcome of July. Regression analysis techniques like Linear Regression, Generalized linear regression, were carried out in order to obtain the best outcome.

The entire work was carried out using the RapidMinerdata science software platform tools [2]. RapidMiner is one of the world's most widespread and most used open source data mining solutions whichis used in various applications disciplines like physics, mechanical engineering, medicine, chemistry, linguistics and social sciences. A typical goal of the RapidMiner is, based on a series of observations for which a certain target value is known, to make forecasts for observations where this target value is not known.It is licensed under the GNU Affero General Public License version 3 and is currently available in version 7.6[2]. With this academic background, RapidMiner continues to not only address business clients, but also universities and researchers from the most diverse disciplines. This includes computer scientists, statisticians and mathematicians on the one hand, who are interested in the techniques of data mining, machine learning and statistical methods.

3.1 Dataset-

The dataset consists of columns for items, id, and label, perishable or non-perishable. The supermarket items were given taken as family and the month "july" was taken as the label set. Set of values representing the sales of the items for the given respective month (January to July) was used to train the model to produce the best result.

RapidMiner makes it possible and easy to implement new analysis methods and approaches and compare them with others. It can be used as such a tool, since it provides a wide range of methods from simple statistical evaluations such as correlation analysis to regression, classification and clustering procedures as well as dimension reduction and parameter optimization[2].

In this paper it is shown how RapidMiner can be optimally used for these tasks. Here, RapidMiner is used to predict the analysis of grocery sales. Data is analyzed to predict the sales in grocery stores so that the retailer manages the wastage of items and gain profit in business.

3.2 Linear Regression:

The raw data of the grocery sales of the past few months was split into two with the ratio of 0.7 and 0.3 for training and testing respectively. Since the Linear Regression cannot handle Polynomial data, the data conversion from polynomial to numeral was carried out using the appropriate operator in RapidMiner. The Linear Regression model was applied to the dataset, validation was carried out using the test data with the outcome listed in the result section. The performance of the model was calculated using the Performance operator.

3.3 Generalized Linear model—

Under the Generalized Linear Model different parameters were changed. The analysis were carried out using methods like Gaussian, poisson, tweedie. The poisson method provided the best result in comparison with the other operators with the Generalized linear Models.

3.3.1 Gaussian method-

Gaussian method is a probabilistic method used in both classification and regression. It is a stochastic process which consists of random values that is linked with every point of range so that each random variable has a normal distribution. Moreover, each finite set of random values had a multivariate normal distribution. This process is important in statistical process because of the properties that is inherited from normal. For instance, if a random process is considered as a Gaussian method, the distributed sharing of various derived quantities can be obtained explicitly[4].

3.3.2 Poisson Distribution-

Poisson distribution is a discrete probability distribution used in regression that describes the probability of a finite set of events occurring in a fixed time intervals even if events occur in a known constant rate and independent of time from last event. The Poisson distribution can be applied to systems with a large number of possible events, each of which is rare.[5]

3.3.3 Tweedie Distribution

Tweedie Distribution is one type of exponential distribution. It can have a cluster of data items at zero, which is particularly useful for modeling claims in the insurance industry, in medical/genomic testing, or anywhere else there is a mixture of zeros and non-negative data points.

4. EXPERIMENT AND RESULT

Linear Regression-

root_mean_squared_error: 3.207 +/- 0.000

squared_error: 10.285 +/- 23.020

Generalized Linear Regression-
Gaussian : root_mean_squared_error: 11.750 +/- 0.000
Poisson : root_mean_squared_error: 8.236 +/- 0.000
Tweedie : root_mean_squared_error: 11.749 +/- 0.000

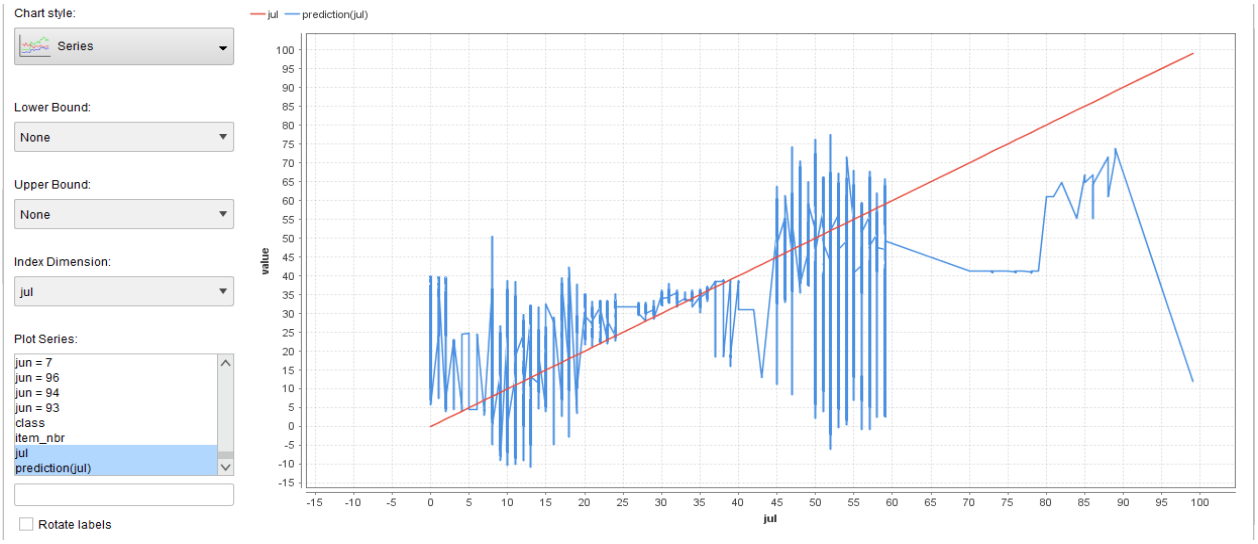

Figure 1. Shows the plot of prediction of the july month versus the linear line using the Linear Regression method, The straight line in red represents the accurate values of the month july, and the blue line indicates the deviation in the prediction of Generalized Regression.

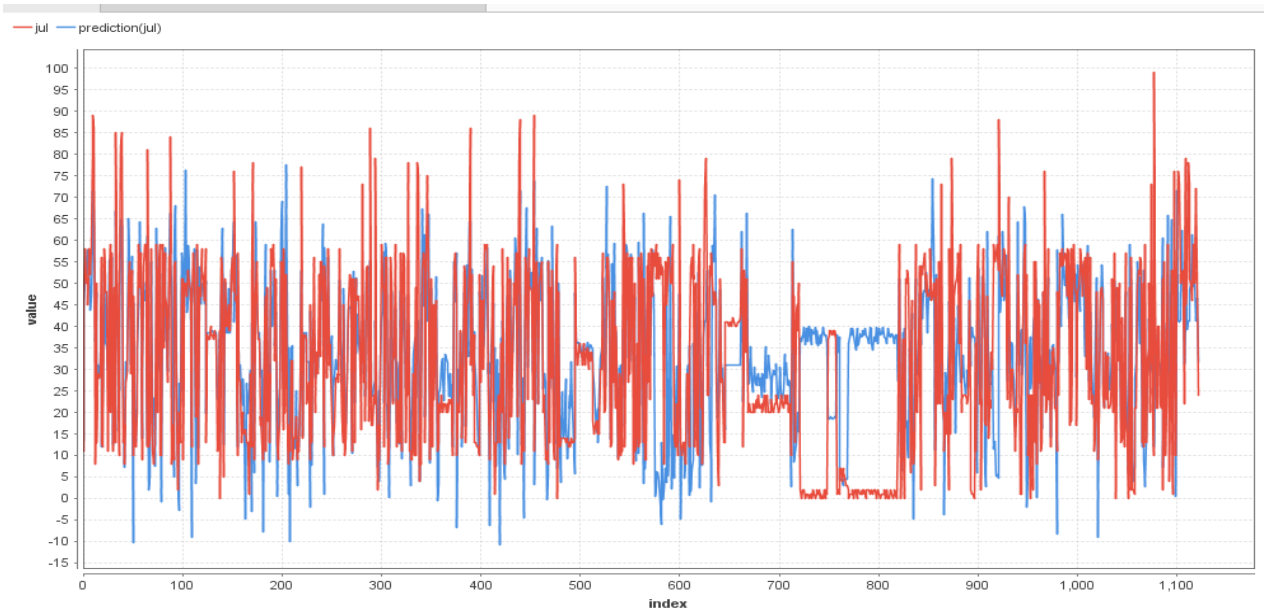

Figure 2 represents the comparative plot of Sales and predictions on the july month implementing the linear regression method.

The statistical plot in red indicates the actual data of the month of July where the blue indicates the prediction plot obtained used using Linear regression. The resulting data shows that the Linear Regression was able to make a accurate prediction when the blue and red indicating the prediction and actual data are in alignment with each other at most of the regions. The prediction values range of -15 to 100 which is plotted in the Y axis.
The root mean squared error show a good prediction with the sale of the particular item in the store with a deviation scale of 3.207 +/- 0.000.

Among the different algorithms applied on the data to predict the grocery sales, The Linear Regression provides the best result with the minimal root mean squared error. This shows that the sales could be predicted with the scale difference of 3.207 from the past data. The dataset was trained with the data of past 6 months to predict the 7^{th} month sales and the Linear Regression proved to be the best model with the most accuracy.

5. CONCLUSION

Machine learning enables programming applications to end up plainly more exact in foreseeing results without being expressly customized [6]. The procedures engaged with machine learning are like that of data mining and predictive modeling. In the real world, ML techniques give a way to identify trends, forecast behavior, and make fact-based recommendations.Machine Learning provides various algorithms based on which datasets are analyzed. Here, data is predicted for the grocery sales store for predicting the sales in third month. Various algorithms were implemented and it was found that among the different algorithms applied on the given dataset, Linear Regression showed the best result which shows that a data could be analyzed to predict of the sales of the upcoming moth with the help of previous months.

6. REFERENCES

[1] Kaneko, Yuta, and Katsutoshi Yada. "A Deep Learning Approach for the Prediction of Retail Store Sales." Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference on. IEEE, 2016.

[2] Land, Sebastian, and Simon Fischer. “RapidMiner 5 :RapidMiner in academic use.” RapidMiner 5, vol. 5, 27 Aug.2012, www.rapid-i.com

[3] 6.1-Introduction to Generalized Linear Models.” 6.1 - Introduction to Generalized Linear Models | STAT 504, onlinecourses.science.psu.edu/stat504/node/216.

[4] GmbH, RapidMiner. “Gaussian Process (RapidMiner Studio Core).” Gaussian Process - RapidMiner Documentation, docs.rapidminer.com/studio/operators/modeling/predictive/functions/gaussian_process.html.

[5] “Poisson distribution.” Wikipedia, Wikimedia Foundation, 3 Nov. 2017, en.wikipedia.org/wiki/Poisson_distribution.

[6] Statistics for Research Projects. www.mit.edu/~6.s085/notes/lecture3.pdf.

[7] L'Heureux, Alexandra, et al. "Machine Learning with Big Data: Challenges and Approaches." IEEE Access ,2017.