A Review on Sales Forecasting using Linear Regression and K Nearest Neighbor

Neha Sehgal[1], Deepika Garg[2]
[1]*Perusing M-Tech, Department of CSE, AITM at Palwal, Haryana, India*
[2]*HOD, Department of CSE, AITM at Palwal, Haryana, India*

***Abstract*—**Sales forecasting is a typical undertaking performed by deals associations giving precise conjectures enabling associations to make educated business decisions. Therefore, organizations infer gauges dependent on authentic deals, current economic situations, and gut intuition – anyway gut nature can acquaint human inclination driving with mistaken figures, therefore, under this scheme we tend device and earmarks of being very difficult to acquire ongoing writing on versatile forecasting procedures in sales (especially FMCG - fast moving consumer goods - Sector). We were searching for a review that could give understanding in the improvement of determining procedures or forecasting system, yet in addition, displayed as of effective and more accurate systems that could without much of a stretch be placed practically in forecasting scenarios for efficient results. In any case, the majority of the articles we experienced couldn't be isolated into one of the classifications. Possibly they were simply centred on giving a unique scientific confirmation of an answer for a specific issue, or it was stories to depict an effectively actualized procedure as an answer for a quite certain down to earth issue. Obviously, it isn't abnormal that little can be found in such a little piece of the absolute region of determining strategies, and that it appears that the vast majority of the experienced methods are accessible for a long time in the area or context to sales forecasting. Therefore, this proposed methodology using machine learning techniques as an amalgamated solution of Linear Regression and KNN we ensure the more effective and accurate framework will be established with elaboration for sales forecasting and predictions.

***Keywords*—** *Machine Learning, Sales Forecasting, Linear Regression, K-Nearest Neighbor, Fast Moving Consumer Goods.*

I. INTRODUCTION

For quite a long time humanity has tried to figure a wide range of speculating out of a variety of interests and potential in the field of forecasting. The contemplation and want to speculation procedures can be connected to practically. Subsequently a variety of procedure we put our brain at for determined results and figures. In the past, this concerned and is not a normal standard phenomenon but nowadays while using machine learning techniques one can conclude or niche the predictions, for example, anticipating the orbit of a planet, shortest path to cover from one location to another location even earth is on rotation and revolution. These days foresee or forecasting techniques are utilized as an approach to control and improve business targets. This can be an impressively all the more requesting procedure, particularly if the basic procedures are not acting as per some stationary example. At whatever point working with genuine information such as stationery or non-stationary conduct will be a common and standard circumstance. In most use cases the information will be hard to watch, and determinations one can make will be founded on fractional data. Luckily with time, forecasting strategies have advanced, which should empower us to make expectations notwithstanding when information changes after some time. The field that attempts to make such modern expectations is initiated adaptive forecasting framework. With this scheme, we touch base at the principal goal of machine learning models, which is to give an understanding/review in the sales forecasting models and can play in the field of Sales Management. To pick up this knowledge we isolate the fundamental goal in the accompanying stages:

1. Review of selected mathematical and machine learning forecasting procedures.
2. Experiments with these procedures on real-life datasets.

A. *Linear Regression:*

Calculations that build up a model dependent on conditions or scientific tasks on the qualities taken by the information ascribes to create a consistent incentive to speak to the yield are called of regression algorithms. The contribution to these calculations can take both persistent and discrete qualities relying upon the calculation, while the yield is constant esteem. Therefore, under this scheme, we inculcate the predictive analysis using linear regression which beholds and performs standard least squares regression to recognize straight relations in the preparation information. This calculation gives the best outcomes when there is some straight reliance on the information. It requires the information credits and target class to be numeric and it doesn't permit missing qualities esteems. The calculation ascertains a regression condition to anticipate the yield (x) for a lot of info qualities or attributes a1,a2,.....,ak. The condition to shape the yield is communicated as a straight mix of information properties with each characteristic related to its particular weight w0,w1,....,wk, where w1 is the weight of a1 and a0 is always taken as the constant 1. An equation takes the form. For an exemplary model on weather forecasting the equation learned would take the form : temp_Ct = w0 + wAt-2 temp_At-2 + wAt-1 temp_At-1 + wAt temp_At + wBt-2 temp_Bt-2 + wBt-1 temp_Bt-1 + wBt temp_Bt + wCt-2 temp_Ct-2 + wCt-1 temp_Ct-1 , where temp_Ct is value

assigned to the output attribute, and each term on the right hand side is the product of the values of the input attributes and the weight associated with each input. The exactness of anticipating the yield by this calculation can be estimated as the supreme distinction between the genuine yield watched and the anticipated yield as acquired from the regression condition, which is likewise the mistake. The loads must be picked in, for example, way that they limit the mistake. To show signs of improvement exactness higher loads must be allocated to those properties that impact the outcome the most. A lot of test cases is utilized to update and evaluate the weights using a confusion matrix. At the begging, the loads can be appointed arbitrary qualities or all set to a steady, (for example, 0). For the primary case in the preparation information, the anticipated yield is gotten as $w_0 + w_1 a_1^{(1)} + \ldots\ldots\ldots + w_k a_k^{(1)} = \sum_{j=0}^{k} w_j a_j^{(1)}$ where the superscript for qualities gives the example position in the preparation information. After the anticipated yields for all examples are acquired, the loads are reassigned in order to limit the aggregate of squared contrasts between the real and anticipated result. Therefore the point of the load refresh process is to limit $\sum_{i=1}^{n}\left(x^{(i)} - \sum_{j=0}^{k} w_j a_j^{(i)} \right)$ which is the sum of the squared differences between the observed output for the i^{th} training instance ($x^{(i)}$) and the predicted outcome for that training instance obtained from the linear regression equation. Regression is the endeavor to clarify the variety in a needy variable utilizing the variety in free factors. Regression is along these lines a clarification of causation. In the event that the autonomous variable(s) adequately clarify the variety in the needy variable, the model can be utilized for expectation. The below figure depicts the linear regression scenario using residual sum of square and regressed sum of square using confusion matrix.

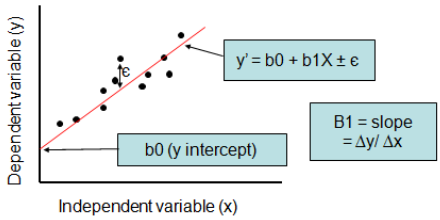

Figure 1: Linear Regression Model depicts that utput of a regression is a function that predicts the dependent variable based upon values of the independent variables. Thus linear regression fits a straight line to the data.

B. K Nearest Neighbors – Classification

K nearest neighbors is a indispensable calculation that stores every accessible case and groups new cases dependent on a similitude measure (e.g., remove capacities). KNN has been utilized in measurable estimation and example acknowledgment as of now at the start of the 1970s as a non-parametric system. Therefore, the case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor.

Distance Function of KNN: The **Euclidean** similarity function is defined as $d_2(X,Y) = \sqrt{\sum_{i=1}^{n-1}(x_i - y_i)^2}$. It can be generalized to the **Minkowski** similarity function, $d_q(X,Y) = \sqrt[q]{\sum_{i=1}^{n-1} w_i |x_i - y_i|^q}$. If q = 2, this gives the Euclidean function. If q = 1, it gives the **Manhattan** distance, which is is $d_1(X,Y) = \sum_{i=1}^{n-1} |x_i - y_i|$. If q = ∝, it gives the **max** function $d_\infty(X,Y) = \max_{i=1}^{n-1} |x_i - y_i|$.

Consider the following data concerning credit default. Age and Loan are two numerical variables (predictors) and Default is the target.

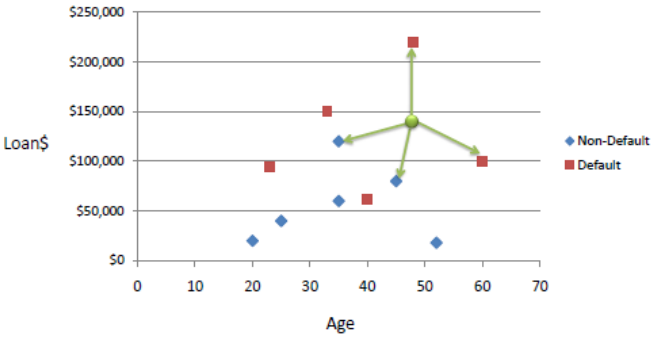

Figure 2: The above figure depicts use of the training set to classify an unknown case (Age=48 and Loan=$142,000) using Euclidean distance. If K=1 then the nearest neighbor is the last case in the training set with Default=Y.

II. RELATED STUDY

William Perrizo, QinDing,Anne Denton [1] depicts that, lazy classifiers store all of the training samples and do not build a classifier until a new sample needs to be classified. It differs from eager classifiers, such as decision tree induction, which build a general model (such as a decision tree) before receiving new samples. K-nearest neighbor (KNN) classification is a typical lazy classifier. Given a set of training data, a k-nearest neighbor classifier predicts the class value for an unknown tuple X by searching the training set for the k nearest neighbors to X and then assigning to X the most common class among its k nearest neighbors. Lazy classifiers are faster at training time than eager classifiers, but slower at predicating time since all computation is delayed to that time. In this paper, we introduce approaches to efficient construction of lazy classifiers, using a data structure, Peano

Count Tree (P-tree)[1*]. P-tree is a lossless and compressed representation of the original data that records the count information to facilitate efficient data mining. With P-tree structure, we introduced two classifiers, P-tree based k-nearest neighbor classifier (PKNN), and Podium Incremental Neighbor Evaluator (PINE). Performance analysis shows that our algorithms outperform classical KNN methods.

Juan Rincon-Patino, Emmanuel Lasso and Juan Carlos Corrales [2] depicts that, persea americana, commonly known as avocado, is becoming increasingly important in global agriculture. There are dozens of avocado varieties, but more than 85% of the avocados harvested and sold in the world are of the Hass one. Furthermore, information on the market of agricultural products is valuable for decision-making; this has made researchers try to determine the behavior of the avocado market, based on data that might affect it one way or another. In this paper, a machine learning approach for estimating the number of units sold monthly and the total sales of Hass avocados in several cities in the United States, using weather data and historical sales records, is presented. For that purpose, four algorithms were evaluated: Linear Regression, Multilayer Perceptron, Support Vector Machine for Regression and Multivariate Regression Prediction Model. The last two showed the best accuracy, with a correlation coefficient of 0.995 and 0.996, and a relative absolute Error of 7.971 and 7.812, respectively. Using the Multivariate Regression Prediction Model, an application that allows avocado producers and sellers to plan sales through the estimation of the profits in dollars and the number of avocados that could be sold in the United States was created.

Robert Fildes, Shaohui Ma and Stephan Kolassa [3] depict that, forecasting problems faced by large retailers, from the strategic to the operational, from the store to the competing channels of distribution as sales are aggregated over products to brands to categories and to the company overall. Aggregated forecasting that supports strategic decisions is discussed on three levels: the aggregate retail sales in a market, in a chain, and in a store. Product-level forecasts usually relate to operational decisions where the hierarchy of sales data across time, product and the supply chain is examined. Various characteristics and the influential factors which affect product level retail sales are discussed. The data-rich environment at lower product hierarchies makes data pooling an often appropriate strategy to improve forecasts, but success depends on the data characteristics and common factors influencing sales and potential demand. Marketing mix and promotions pose an important challenge, both to the researcher and the practicing forecaster. Online review information to adds further complexity so that forecasters potentially face a dimensionality problem of too many variables and too little data. The paper goes on to examine evidence on the alternative methods used to forecast product sales and their comparative forecasting accuracy. Many of the complex methods proposed have provided very little evidence to convince as to their value, which poses further research questions. In contrast, some ambitious econometric methods have been shown to outperform all the simpler alternatives including those used in practice. New product forecasting methods are examined separately where limited evidence is available as to how effective the various approaches are. The paper concludes with some evidence describing company forecasting practice, offering conclusions as to the research gaps but also the barriers to improved practice.

Khalid Alkhatib, Hassan Najadat, Ismail Hmeidi and Mohammed K. Ali Shatnawi [4] depicts that, Stock prices prediction is interesting and challenging research topic. Developed countries' economies are measured according to their power economy. Currently, stock markets are considered to be an illustrious trading field because in many cases it gives easy profits with low risk rate of return. Stock market with its huge and dynamic information sources is considered as a suitable environment for data mining and business researchers. therefore, We applied k-nearest neighbor algorithm and non-linear regression approach in order to predict stock prices for a sample of six major companies listed on the Jordanian stock exchange to assist investors, management, decision makers, and users in making correct and informed investments decisions. According to the results, the kNN algorithm is robust with small error ratio; consequently the results were rational and also reasonable. In addition, depending on the actual stock prices data; the prediction results were close and almost parallel to actual stock prices.

III. PROPOSED METHODOLOGY

Under the proposed scheme we will first gather the data from International Data Repositories, thereafter using pre-processing techniques the noise will be expedited and the object-relational model will be created to define the relational or non-relational structures vide schema among the data. Subsequently, using linear regression the residual sum of squares and the regressed sum of squares will be evaluated using the slope and intersect. Thus forming the confusion matrix originating attributed weights and measures. Consequently, using KNN the K-weight will be allocated different attribute groups to evaluated the nearest neighbour using various function i.e. Euclidean, Minkowski or Manhattan thereinafter, the performance evaluation using Accuracy, Sensitivity and Specificity with being imposed for an effective and accurate forecast on sales and sales management. Below diagram depicts the workflow of proposed scheme with nitty-gritty.

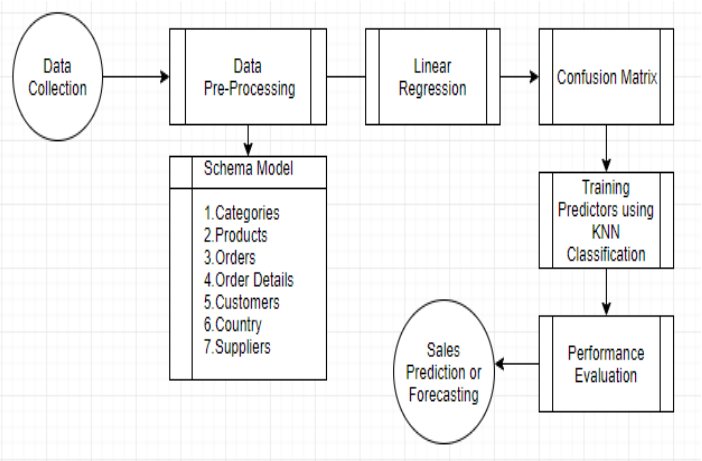

Figure 3: Workflow Diagram of Proposed Scheme using Linear Regression and KNN Classifier

IV. CONCLUSION

An information-driven for sales forecasting model expands both verifiable information to measure patterns and regularity just as the present pipeline of chances to figure deals for the following a year. This model runs consequently and presents a month to month gauge while continually learning on new information that ends up accessible approach using machine learning techniques under the unsupervised learning model amalgamated using linear regression and KNN for accurate, effective, potentially niche as per the consumption or demand of market specially in FMCG (fast moving consumer goods) sector. However, under this review, we affirm to incorporate the above-mentioned scheme in future scenarios

ACKNOWLEDGMENT

I take this opportunity to express my profound gratitude and deep regards to my guide Assistant Prof. Ms.Deepika Garg for her exemplary guidance, monitoring and constant encouragement throughout the scheme/scenario of this research. She motivated and inspired me through her knowledge base and skill set in the entire duration of work, without which this paper could not have seen the light of the day. I convey my regards to all the faculty members of the Department of Computer Science and Engineering, AITM for their valuable guidance and advices for developments at appropriate time to time.

REFERENCES

[1] William Perrizo ,QinDing,Lazy Classifiers Using P-trees, Anne Denton, Conference: Proceedings of the 15th International Conference on Computer Applications in Industry and Engineering, November 7-9, 2002, Clarion Hotel Bay View, San Diego, California, USA.

[2] Juan Rincon-Patino, Emmanuel Lasso and Juan Carlos Corrales, Estimating Avocado Sales Using Machine Learning Algorithms and Weather Data ,Grupo de ingeniería Telemática, Universidad del Cauca, Campus Tulcán, Popayán 190002, Colombia; 29 September 2018.

[3] Robert Fildes, Shaohui Ma and Stephan Kolassa, Retail forecasting research and practice, The Department of Management Science Lancaster University Management School Lancaster LA1 4YX UK April,2018

[4] Khalid Alkhatib, Hassan Najadat, Ismail Hmeidi and Mohammed K. Ali Shatnawi Stock Price Prediction Using K-Nearest Neighbor (kNN) Algorithm, International Journal of Business, Humanities and Technology Vol. 3 No. 3; March 2013

[5] Saravanan Thirumuruganathan, A Detailed Introduction to K-Nearest Neighbor (KNN) Algorithm, https://saravananthirumuruganathan.wordpress.com/2010/05/17/a-detailedintroduction-to-k-nearest-neighbor-knn-algorithm/ , [last accessed: 06-18-2015].

[6] Khalid Alkhatib, Hassan Najadat, Ismail Hmeidi, Mohammed K. Ali Shatnawi, Stock Price Prediction Using K-Nearest Neighbor (k-NN) Algorithm, International Journal of Business, Humanities and Technology (2013) 3 No. 3, 32 - 44

[7] Fazli Wahid1 and DoHyeun Kim A Prediction Approach for Demand Analysis of Energy Consumption Using K-Nearest Neighbor in Residential Buildings, International Journal of Smart Home Vol. 10, No. 2, (2016), pp. 97-108 http://dx.doi.org/10.14257/ijsh.2016.10.2.10

[8] D. TREN, "EU energy and transport in figuresâ€"Statistical pocketbook 2009, Directorate-General for Energy and Transport", Office for Official Publications of the European Communities, Luxemburg, (2009).

[9] Su Yang, Shixiong Shi, Xiaobing Hu, "Discovering Spatial temporal weighted model on Map- Reduce for short term traffic low forecasting ", IEEE international conference on transportation, pp. 364 – 367, 2015.

[10] Donghai Yu, Yang Liu, and Xiaohui Yu, "A Data Grouping CNN Algorithm for Short-Term Trac flow Forecasting",2016.

[11] YuqiWang ,Jiannong Cao , Wengen Li and Tao Gu ,"Mining Traffic Congestion Correlation between Road Segments on GPS Trajectories" , IEEE, pp. 1 – 8, 2016.

[12] Hao-Fan Yang, Tharam S. Dillon, Life Fellow, IEEE, and Yi-Ping, "Optimized Structure of the Traffic flow Forecasting Model With a Deep Learning Approach" IEEE transactions on neural network, pp. 1 – 11, 2016.

[13] JinyoungAhn, EunjeongKo, Eun Yi Kim "Highway Traffic Flow Prediction using Support Vector Regression and Bayesian Classifier", IEEE, pp. 239 – 244, 2016.

[14] ZhongshengHou, Senior Member, IEEE, and XingyiLi,"Repeatability and Similarity of Freeway Traffic Flow and Long-Term Prediction Under Big Data", IEEE transactions on intelligent transportation system pp.1786 – 1796, 2016.

[15] JiwanLee ,Bonghee Hong , Kyungmin Lee and Yang-Ja Jang , "A Prediction Model of Traffic Congestion Using Weather Data" ,IEEE International Conference on Data Science and Data Intensive Systems, pp. 81 – 88, 2015.

[16] Zhiyuan Ma and GuangchunLuo, "Short Term Traffic Flow Prediction Based on On-line Sequential Extreme Learning Machine", 8th

[17] International Conference on Advanced Computational Intelligence, pp. 143 – 149, 2016.

[18] KalliSrinivasaNageswara Prasad SeelamRamakrishn, "An Efficient Traffic Forecasting System Based on Spatial Data and Decision Trees", Department of Computer Science, Sri Venkateswara University,