ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 6, June 2019

Sales Forecasting using Linear Regression and K-Nearest Neighbour

Neha Sehgal[1], Deepika Garg[2]

P.G. Student, Department of CSE, AITM at Palwal, Haryana, India[1]

HOD, Department of CSE, AITM at Palwal, Haryana, India[2]

ABSTRACT: Forecasting sales is a typical undertaking performed by associations. This generally includes physically concentrated procedures utilizing calculations that require a contribution from different degrees of an association. The proposed methodology presents inclination and is commonly not precise particularly during an underlying couple of long stretches of a sales period. Truth be told, that is the point at which an exact estimate that, where the sale can be increased or decreased or the specific product should be stopped selling in specific city or country even may have the most advantage, all things considered, there's little incentive in giving a precise conjecture in the most recent days sales based on past performance. In spite of the fact that the way toward gauging will, in general, be mind-boggling it is clear to decide its exactness. One basically needs to hold up until the finish of a determining period and afterward contrast estimates and actual. We are sure about the precision of our models and are welcoming deals pioneers to our Man versus Machine Forecasting Duel - allow us daily with your information and we'll give a calculation based fair-minded gauge or weight and measure. The procedure is straightforward and enables you to rapidly observe what AI using Linear Regression and KNN can accomplish for desired results in the proposed scheme.

KEYWORDS: Machine Learning, Linear Regression, K-Nearest Neighbour, Confusion Matrix, Sales Forecasting

I. INTRODUCTION

Under the proposed plan we will initially accumulate the information from International Data Repositories, from there on utilizing pre-handling systems the commotion will be sped up and the item social model will be made to characterize the social or non-social structures vide mapping among the information. Along these lines, utilizing Linear Regression the regressed sum of squares and the residual sum of squares will be assessed utilizing the slope and intercept. Along these lines shaping the disarray confusion matrix credited loads and measures. Thusly, utilizing KNN the distance using Euclidean function will be designated diverse ascribe gatherings to assess the closest neighbour utilizing different capacities as per previous sales of products based on categories, thereinafter the exhibition assessment utilizing Accuracy, Sensitivity, and Specificity with being forced for a powerful and exact conjecture on deals.

The following paragraphs show the growing importance of sales forecasting using the proposed scheme based on machine learning model composition produced using Linear Regression and K-NN Classification with Distance Calculation using Euclidean Method :-

1. Promotion of new business: using the above forecasting scheme is of utmost importance in setting up a new business. It is not an easy task to start a new business as it is full of uncertainties and risks. With the help of forecasting the promoter can find out whether he can succeed in the new business; whether he can face the existing competition; what is the possibility of creating demand for the proposed product etc. After discovering the business opportunity, he will see the possibilities of assembling men, money, materials etc. The success of a business unit depends upon as to how sound is the forecasting? Proper forecasting will help to minimise the role of luck or chance in determining business success or failure. A successful promoter is also the prophet of economic conditions.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 6, June 2019

2. Estimation of financial requirement :The above scheme produce or illustrate importance of forecasting can't be ignored in estimating the financial requirements of a concern in FMCG domain. Efficient utilisation of capital is a delicate issue before the management. No business can survive without adequate capital. But adequacy of either fixed or working capital depends entirely on sound financial forecasting. Financial estimates can be calculated in the light of probable sales and cost thereof. How much capital is needed for expansion, development etc., will depend upon accurate forecasting.

3. Smooth and continuous working of a concern: Above Forecasting illustrate the earnings mechanism which ensures smooth and continuous working of an enterprise, particularly to newly established ones. By forecasting, these concerns can estimate their expected profits or losses. The object of a forecast is to reduce in black and white the details of working of a concern.

4. Correctness of management decisions: The above scheme produces correctness of management decisions to a great extent depends upon accurate forecasting. As Meivin, T. Copeland says, "Administration is essentially a decision making process and authority has responsibility for making decisions and for ascertaining that the decisions made are carried out.In business, whether the enterprise is large or small, changes in conditions occur; shifts in personnel take place, unforeseen contingencies arise. Moreover, just to get the wheels started and to keep them turning, decisions must be made. This shows that the decision making process continues throughout the life of the concern. Forecasting plays an important role in various fields of the concern. As in the case of production planning, management has to decide what to produce and with what resources. Thus forecasting is considered as the indispensable component of business, because it helps management to take correct decisions.

5. Success in business: using the above scheme, accurate forecasting of sales helps to procure necessary raw materials on the basis of which many business activities are undertaken. The accurate sales forecasting becomes the basis for several other budgets. In the absence of accurate sales forecasting, it is difficult to decide as to how much production should be done. Thus, to a great extent, the budgets of other departments depend upon the compilations based on the sales forecasts and the accuracy of these budgets also depends upon correctness of sales forecasting. Thus, the success of a business unit depends on the accurate forecasting by the various departments.

6. Plan Formulation: The above scheme illustrates the importance of correct forecasting is apparent from the Key role it plays in planning. It should not go unaccounted that forecasting is an essential element in planning since planning premises include some forecasts. There are forecast data of a factual nature having enormous implication on sound premises. Undoubtedly, forecasting is a prelude to planning and indeed it is the foundation on which planning takes place. Infact, planning under all circumstances and in all occasions involve a good deal of forecasting, i.e. appraising the future in the light of existing conditions and environment. Forecasting and planning are closely related. Adequate planning, no matter whether it is overall or sectoral, short-term or long term, largely depends on forecasting.

7. Cooperation and co-ordination: using above scheme forecasting will incorporate the proper co-ordination of all departmental heads in a company. Thus, by bringing participation of all concerned in the process of forecasting, team spirit and co-ordination is automatically encouraged. According to Henry Fayol, "The act of forecasting is of great benefit to all who take part in the process, and is the best means of ensuring adaptability to changing circumstances. The collaboration of all concerned leads to a united front, an understanding of the reasons for decisions and a broadened outlook.

8. Complete Control: the above scheme will provide the information which helps in the achievement of effective control. The managers become aware of their weaknesses during forecasting and through implementing better effective control they can overcome these weaknesses.

II. RELATED WORK

Gelper, Sarah & Fried, Roland & Croux, Christophe [1] Outliers affect the forecasting methods in two ways. First, the smoothed values are affected since the update equations involve current and past values of the series including the outliers. The second effect of outliers is on the selection of the parameters used in the recursive updating scheme. These

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 6, June 2019

parameters regulate the degree of smoothing and are chosen to minimize the sum of squared forecast errors. Forecasting techniques are often used as much for their explanatory power as for their predictive power for practical implementation. Understanding of trend, level and seasonal behavior of the data provided would result in a better forecast.

Koehler, Anne & D. Snyder, Ralph & Ord, Keith & Beaumont, Adrian [2] there are two variations to winter's method of forecasting that differ in the nature of the seasonal component. The additive method is preferred when the seasonal variations are roughly constant through the series, while the multiplicative method is preferred when the seasonal variations are changing proportional to the level of the series. With the additive method, the seasonal component is expressed in absolute terms in the scale of the observed series, and in the level equation the series is seasonally adjusted by subtracting the seasonal component. Within each year the seasonal component will add up to approximately zero. With the multiplicative method, the seasonal component is expressed in relative terms (percentages) and the series is seasonally adjusted by dividing through by the seasonal component.

A.A. Syntetos, J.E. Boylan [3] studied the causes of the unexpected performance of the Croston Method and developed first step towards it. Certain limitations are identified in Croston's approach and a correction in his derivation of the expected estimate of demand per time period is presented. In addition, a modification to his method that gives approximately unbiased demand per period estimates is also introduced.

Teunter, Ruud & Sani, Babangida [4] presented a predictive error-forecasting model which compares and evaluates forecasting methods based on their factor levels when faced with intermittent demand. The modified procedure uses a new method for estimating the mean demand and a smoothing method for estimating the variance of the forecasted demand rate. It is found that the smoothed variance estimate is based on an invalid measure of forecast accuracy. Moreover it is shown that the method of forecasting mean demand produces biased forecasts.

III. PROPOSED ALGORITHM

A. *Linear Regression:*

LR may be a basic and ordinarily used kind of prognosticative analysis. The plan of regression is to look at 2 things will a collection of predictor variables do an honest job in predicting associate degree outcome (dependent) variable. That variable especially is important predictors of the result variable and in what approach do they–indicated by the magnitude and sign of the beta estimates–impact the result variable. These regression estimates are required to justify the link between one variable and one or a lot of freelance variables. The only sort of the regression of y on x with one dependent and one variable is outlined by the formula y = c + b*x, wherever y = calculable variable score, c = constant, b = parametric statistic, and x = score on the variable.

B. *K-Nearest Neighbor:*

K nearest neighbour 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 neighbour, with the case being assigned to the class most common amongst its K nearest neighbour measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbour.**Distance Function of KNN:** **T**he **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,

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 6, June 2019

$d_q(X,Y) = \sqrt[q]{\sum_{i=1}^{n-1} w_i |x_i - y_i|^q}$. If q = 2, this gives the Euclidean function. If q = 1, it gives the **Manhattan** distance, which is $d_1(X,Y) = \sum_{i=1}^{n-1} |x_i - y_i|$. If q = ∝, it gives the **max** function $d_\infty(X,Y) = \max_{i=1}^{n-1} |x_i - y_i|$.

IV. PSEUDO CODE

The Algorithm's pseudo-code

Step 1

Linear regression is the most popular regression model. In this model, we wish to predict response to n data points $(x_1, y_1),(x_2, y_2),.....,(x_n, y_n)$ by a regression model given by $y = a_0 + a_1 x$ where a_0 and a_1 are the constants of the regression model.A measure of goodness of fit, that is, how well $a_0 + a_1 x$ predicts the response variable y is the magnitude of the residual ε_i at each of the n data points. $E_i = y_i - (a_0 + a_1 x_i)$

Step 2

Consider k as the desired number of nearest neighbour and S:=p1,...,pn be the set of training samples in the form p1=(xi,ci), where xi is the d-dimensional feature vector of the point pi and ci is the class that pi belongs to.
For each p'=(x',c')

- Compute the distance d(x′,xi) between p′ and all pi belonging to S
- Sort all points pi according to the key d(x′,xi)
- Select the first k points from the sorted list, those are the k closest training samples to p′
- Assign a class to p′ based on majority vote: c′=argmaxy∑(xi,ci) belonging to S, I(y=ci)

End

Step 3

Calculating Euclidean Distance
Where as the initial distance is dist=0
for d=1 to N // d = dimension
dist=dist+(x1[d]-x2[d])^2
next
return sqrt(dist)

V. SIMULATION RESULTS

We are concerned with whether the relationship pattern between two values of variables can be described as a straight line, which is the simplest and most commonly used form. Remember from geometry class that a line is described by the formula: Y = a + bX (in geometry we said Y = mx + b where m was slope and b was y-int) Where Y is the dependent variable, measured in units of the dependent variable, X is the independent variable, measured in units of the independent variable, and a and b are constants defining the nature of the relationship between the variables X and Y. The a or Y-intercept (aka Yint) is the value of Y when X = 0. The b is the slope of the line and is known as the regression coefficient and is the change in Y associated with a one-unit change in X.

The greater the slope or regression coefficient, the more influence the independent variable has on the dependent variable, and the more change in Y associated with a change in X. The regression coefficient is typically more important than the intercept from a policy researcher perspective as we are usually interested in the effect of one variable on another. Coming back to the equation, we also have a term to capture the error in our estimating equation, denoted ε or e. Also known as the residual, it reflects the unexplained variation in Y, and its magnitude reflects the

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 6, June 2019

goodness of fit of the regression line. The smaller the error, the closer the points are to our line. So our general equation describing a line is: Y = a + bX + e Remember, b is the regression coefficient and is interpreted as the change in Y associated with a one-unit change in X.

Category	UnitPrice Slope	UnitPrice Intercept	UnitsinStock Slope	UnitsinStock Intercept	UnitsOnOrder Slope	UnitsOnOrder Intercept
Beverages	-0.15461307939296315	52.45540924444525	-0.5476753553886536	63.345876971854786	-0.016057481513603537	5.609849766652068
Condiments	0.28987665813358154	35.56471957179428	0.018120814485784526	22.067728921308937	-1.3322564306150664	44.89183059772663
Confections	0.1254741978329276	26.535376874831236	0.11029740886258516	21.571169244541704	-0.35244670260136757	22.713712883604252
Dairy Products	-0.010006805033997158	39.58749550862674	-7.191467134995194E-4	28.52826246584053	-0.6287668455084486	32.06447147145773
Grains/Cereals	-0.777492058690062	59.744214188473755	-0.11885092001502065	25.37229762351805	1.4733020723037362	-16.9772241070078
Meat/Poultry	-0.3041443606385235	43.925823103551195	-0.3144591002557801	62.314291923700615	0.0	0.0
Produce	0.2625147288958665	11.502398225640801	0.5627306273062731	20.94538745387454	-0.37601757550187503	16.17168891899569
Seafood	-0.5849731730667772	70.51537431862029	-0.07070698884426957	24.54713326498608	-0.20319614884135592	14.202604348411345

Figure 1: Confusion Matrix Result based on Linear Regression

Consequently, the resulted structures are achieved on the basis of Linear Regression and K-NN along-with Euclidean Distance Function thereafter, features and variant factors will drawn with results based on categories, products, cities and countries which will proposed various escalating and illustrated stages for sales forecasting. Subsequently, we believe scheme proposed produces betters and effective results then the existing results produced by various machine learning techniques. Thus we can proceed with the new scheme as the defacto model for the sales forecasting or prediction or in future as ready reference model for FMCG sector or domain.

KNN Distance Evaluation using Euclidean Function

Sno	Country	City	Category	Least-DM	Max-DM
1	Venezuela	San Cristóbal	Seafood	6.675859079133145	38.38023209876539
2	Venezuela	San Cristóbal	Produce	19.9989166373248	317.1800000000003
3	Venezuela	San Cristóbal	Meat/Poultry	4.949747468305833	19.360000000000014
4	Venezuela	San Cristóbal	Grains/Cereals	9.787321731028703	71.84375
5	Venezuela	San Cristóbal	Dairy Products	13.386280289908765	187.26543209876542
6	Venezuela	San Cristóbal	Confections	12.429074449317081	168.15216875
7	Venezuela	San Cristóbal	Condiments	6.399316369738252	10.400624999999934
8	Venezuela	San Cristóbal	Beverages	5.862721211178305	34.923553719008275
9	Venezuela	San Cristóbal	Total Sale	13.565463996279059	211.03786498765436
10	Venezuela	I. de Margarita	Seafood	17.669788583099358	273.19374999999985
11	Venezuela	I. de Margarita	Produce	0.0	0.0
12	Venezuela	I. de Margarita	Meat/Poultry	0.0	0.0
13	Venezuela	I. de Margarita	Grains/Cereals	7.424621202458749	27.5625
14	Venezuela	I. de Margarita	Dairy Products	12.650296439214378	106.68666666666665
15	Venezuela	I. de Margarita	Confections	3.979330275868602	14.075617283950654
16	Venezuela	I. de Margarita	Condiments	7.534421012924617	27.13388888888892
17	Venezuela	I. de Margarita	Beverages	2.2499206335208615	1.984375

Figure 2 Results Evaluated using K-NN Euclidean Method based on Distance depicting Least Distance and Maximum Distance inculcating Country, City and Category

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 6, June 2019

VI. CONCLUSION AND FUTURE WORK

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.

Future Scope : The Above proposed scheme can be incorporated with mammoth repositories based on Big-Data using Hadoop or Spark like technologies which will act on extremely large datasets based on MapR and Parallel execution and can deliver likewise accurate and efficient results thereinafter.

REFERENCES

1. Gelper, Sarah & Fried, Roland & Croux, Christophe. (2010). Robust Forecasting with Exponential and Holt-Winters Smoothing. Journal of Forecasting. 29. 285-300. 10.2139/ssrn.1089403.
2. Koehler, Anne & D. Snyder, Ralph & Ord, Keith & Beaumont, Adrian. (2012). A study of outliers in the exponential smoothing approach to forecasting. International Journal of Forecasting, 28, 477-484. International Journal of Forecasting - INT J FORECASTING. 28. 10.1016/j.ijforecast.2011.05.001.
3. Teunter, Ruud & Sani, Babangida, 2009. "On the bias of Croston's forecasting method," European Journal of Operational Research, Elsevier, vol. 194(1), pages 177-183, April.
4. Nikolopoulos, Konstantinos & A. Syntetos, Aris & Boylan, John & Petropoulos, Fotios & Assimakopoulos, Vassilis. (2011). An Aggregate-Disaggregate Intermittent Demand Approach (ADIDA) to Forecasting: An Empirical Proposition and Analysis. JORS. 62. 544-554. 10.1057/jors.2010.32.
5. Lindsey, Matthew & Pavur, Robert. (2016). A Comparative Evaluation of Intermittent Demand Forecasting with Updated Smoothing Constants. 10.1108/S1477-407020160000011004.
6. Anderson E & Anderson M 2000, Are your decision today creating your future competitors? Avoid the outsourcing trap, The Systems Thinker, Pegasus Communication, Waltham, MA.
7. Ballot, M 1986, Decision-making models in production & operations management, Florida, Robert E. Krieger Publishing Company, Inc.
8. Ballou, RH 2004, Business logistics/ supply chian management (ed.) 5, New Jersey, Pearson Education, Inc.
9. Christopher, M 2005, Logistics and supply chain management creating value-adding networks (ed.) 3, Harlow, Pearson Education Limited.
10. Fawcett, SE, Ellram, LM, & Ogden, JA 2007, Supply Chain Management: From vision to implementation, New Jursey, Pearson Prentice Hall.