SALES PREDICTION USING MACHINE LEARNING ALGORITHM

Priyadharshini G1, Shreenidhi G L2, Dr. RakeshKumar3

12 Student, Dept of Computer Science and Engineering, Rajalakshmi Engineering College, Chennai, TamilNadu, India.

3 Assistant Professor, Dept of Computer Science and Engineering, Rajalakshmi Engineering College, Chennai, TamilNadu, India.

Abstract - The traditional way of finding the sales and marketing goals no longer help the companies to increase the pace of the sales of the companies among the competitive market. As these approaches do not have the insights to customers' purchasing patterns. The main issue faced by retailers is variations in sales of a product. In order to address this challenge, we aim to forecast sales by analyzing historical sales data across various stores. Thanks to these advancements, crucial aspects like consumer purchasing trends, target demographics, and forecasting sales for the upcoming years can now be easily discerned, aiding the sales team in crafting strategies to enhance business growth. This paper aims to introduce an application for predicting future sales of stores, leveraging insights from previous years' sales data. A thorough examination of sales prediction is conducted using Machine Learning models including Linear Regression, K-Neighbors Regressor, XGBoost Regressor, and Random Forest Regressor. Key predictive parameters encompass item weight, item fat content, item visibility, item type, item MRP, outlet establishment year, outlet size, and outlet location type.

Key Words: Label Encoder, Standard ScalerLinear Regression, K-Neighbors Regressor, XGBoost Regressor, Random Forest Regressor

1. INTRODUCTION

Sales predicting has always been a very important area to consider. The sales of a product plays a major role in predicting a company's sales. These future predictions help the company to increase their profit and they use different strategies to increase the sales[1]. Basically prediction involves various factors like customer taste, pattern of purchasing, surroundings of the shop. A good accurate prediction focus on these factors deeply. Shortly, prediction can be done by studying the previous years of sales.[2]. Every business strives for profitability, which entails not just maximizing stock sales but also managing inventory efficiently to avoid excess stock. It's essential for retailers to maintain optimal stock levels based on demand and address any issues or inefficiencies that may hinder sales. Thus, the study aims to address this challenge by predicting store sales[3]. Therefore, this paper divides the sales prediction into different phases, including Data preprocessing. In the first section, we import and observe the dataset and handle the missing values of the data by using statistics, then we use feature engineering to explore more data and are ready to pass in models. Splitting the dataset into train and test data with some ratio. We can use Scikit-learn to divide the dataset into test and train in the desired ratio.This helps to overcome the overfitting and underfitting issue. These analyses would help the business organizations to make a decision and implement different strategies at each important stage of business.

2. LITERATURE REVIEW

[4] ’Walmart's Sales Data Analysis - A Big Data Analytics Perspective’

This study involves data gathering from a retail store and forecasting the future sales of the store. Effect of various events such as the climatic conditions, holidays etc. can actually modify the state of different departments so this paper also studies these effects and attempts to examine its influence on sales.

[2] ’Applying machine learning algorithms in sales prediction’

This thesis employs various machine learning algorithms to enhance and optimize the results which is used further to predict the sales. There are four algorithms utilized along with ensemble technique. Feature selection has also been implemented.

[3] ’Sales Prediction System Using Machine Learning’

This paper aims to achieve the accurate predictions of future sales and demand of a company by implementing various machine learning techniques such as clustering models and sales predictions measures.

[4] ’Intelligent Sales Prediction Using Machine Learning Techniques’

This paper aims to make decisions from the experimental data and the insights from the data visualization. In this data mining techniques were used. To show the maximum accuracy, Gradient Boosting is implemented.

[5] ’Sales prediction and item recommendations using machine learning techniques’

This paper aims to predict the future sales along with the product recommendation system. In retail shops, the product recommendation system creates a great impact in sales. For designing the sales of each individual customer demographic has been used.

[6] ‘With the help of regression technique for constructing an intelligent sales prediction system’

This paper involves the deep neural network techniques which is used to know about the sales strategy. Additionally some optimization algorithms like genetic algorithms are used to enhance the results.

[7] ’Bayesian learning for sales rate prediction for thousands of retailers’

This paper aims to determine the correlation between dataset attributes for this heat map. This heat map, a feature of the data visualization library called seaborn which displays the color coded matrix.

[8] ’Combining Data Mining and Machine Learning for Effective User Profiling’ This

research involves various automated prototypes to detect suspicious activity. Different machine learning algorithms are employed to develop this prototype. In this data mining and constructive induction techniques to recognize the pattern and nature of customer shopping trends.

3.METHODOLOGY

Predicting mart sales is an exciting challenge; integration of data science enhances sales and business.

This proposed research method has been divided into following steps: i) hypothesis making, ii) data exploration, iii) data cleaning, and iv) feature engineering. To build the model, we first examined the dataset hypothetically, features extraction in order to understand the data. Next, finding the relationship between attributes for a meaningful conclusion. Handling missing values is one of the most important steps and for that we need to clean the data before proceeding to feature engineering.

3.1. The problem assertion

We have also gathered the data of 8524 items from various stores in order to predict the outcome i.e the predicted price of a product in various stores and try to analyze the difference in sales. The most critical factor is to find out attributes which cause more impact in the sale of a product. Let's discuss the various hypotheses that depend on sales factors .

3.2. Mart hypothesis

− As per general hypothesis, store sales are influenced by the store location into urban or rural areas.

− Population: Here, it can be seen that more population = more sale.

− The map can also be a determining factor on the sales since the larger the store would be the bigger the store quality print would be.

− Opponents consider how many hurdles are there in the market to succeed successfully.

− Promotions are also an important part of sales as well.

− Location matters. It should satisfy the sale if a store is located in an area with high population density.

− Customer satisfaction.

3.3. Product hypothesis

− Consumers always require products that are of good quality within certain brands.

− It also prints the excellent quality of the mind of the customer during the packaging of the product.

− Whether or not the stores sell the related product. that it should be related to daily use, it means that stores should have the product.

− Another reason for consumers to patronize Discount Promotions.

− Product price.

Here are some postulates which we thought should be taken into account; this is all not enough.

There are also many imponderables that are related to sales, but we write what we find in a search. These hypotheses are for a better understanding of the overall design and to be able to isolate the potential confounding variables that may influence the outcomes of the study. Therefore, the description of the used dataset is described in the next section.

3.4. Data exploration

In this step, we begin the implementation process where we compute and compare the attributes of the data set and check for null values or graphs of the analyzed data.

3.5. Data cleaning

Outliers and missing valuables affect any data set greatly thus correcting for outliers and dealing with missing values are important. Data cleaning is how we enhance our dataset, to do it. We have two basic attributes in our datasets and have missing variables that we have to complete. The used codes produce a summary of missing values for the before and after case. For instance, from Table II, product weight has a total of 1463 missing values before the mean and all zeroes after implementing it. Output tells us that we can see there are no more missing values. Further, another attribute. The code in the following cell has a missing value, Outlet_size, by using mode in aggfunc**.** So, there are two critical characteristics here and two unknown values that should be added. These codes also help in calculating how many missing values are present initially and how many are remaining at the end of the process. For instance, for the product weight, the number of missing entries is 2410, prior to applying the implemented method, and before the zeroes method is executed. It is also evidence of the output that there is no more information missing in the analysis. In addition, another attribute that has a missing value is Outlet size that employs mode in aggfunc.

After this step, the feature engineering process is ready on our dataset. The topic of the feature engineering step is continued in the next section.

3.6. Feature engineering

In this stage, feature engineering assists in the appreciation of the data that can be used for more effective analysis. Here in this write up, it will be appropriate to develop some new variables from the original information given in the Research Proposal Framework . If we look into our data set for analysis, a number of shades has to be addressed. For instance, do you remember the time when we discussed the concatenation of two Supermarkets ? Indeed we almost tend to believe that both have more or less the same sale. Well, then, check and see if the item is true or false. The output compares the supermarket sales and from there you can clearly see the variation between the two hence this proposal to aggregate the data of markets will not benefit the idea. Data cleaning and data wrangling have been done, and it was time for us to create our earliest model. Predicting sales is a very crucial aspect, which is why this model is very useful. Our study incorporates six decision tree algorithms, including Random Forest, Linear regression, and many more. Therefore, let us first start evaluating the model with the initial basic model present in the program. The baseline is independent of the criteria for the forecasting model; the average value of all the sales is used in making the sales forecasts. This simplifies it since we can predict the probability of mart sales on this day by taking an average of all mart sales using the baseline model. On the other hand, linear regression is a valuable and fundamental model of predictive analysis, and the library, sclikit-learn, provides many artificial intelligence models, one of which is the modes of regression.

First, let's explore the relationship between item weight and count through a bar plot as depicted below in figure 3.1. Figure 3.2 depicts there is maximum visibility of items when density is around 11 Figure 3.3 contains the bar diagram which defines the relations between item mrp and density. This figure spotlights item outlet sales and density as presented below in Figure 3.4. As presented in figure 3.5, low fat content has the highest number of counts among all the possible contents. Figure 3.6 presents the item outlet sales and the density of outlets for each item in the stores. Table 4 presents the various items that were sold in marts, categorized by their count. The following different types of marts have been illustrated in figure 3.8.

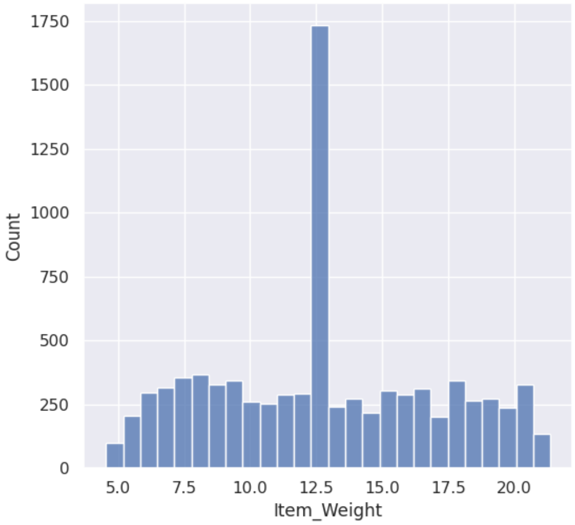


Figure 3.1

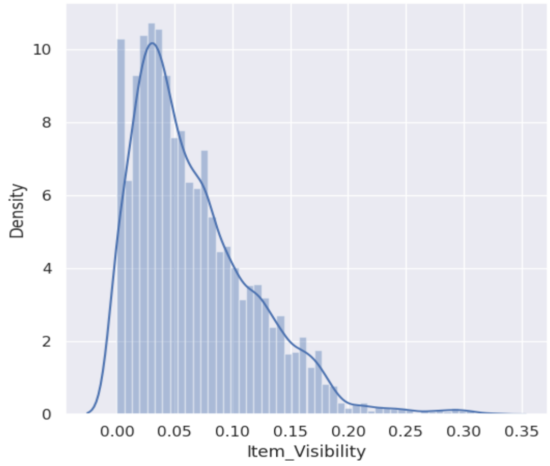


Figure 3.2

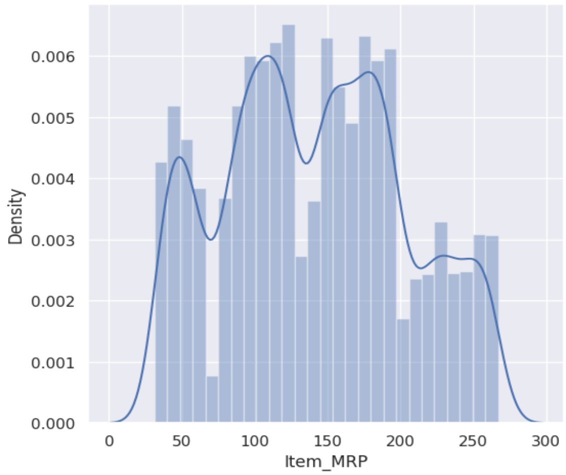


Figure 3.3

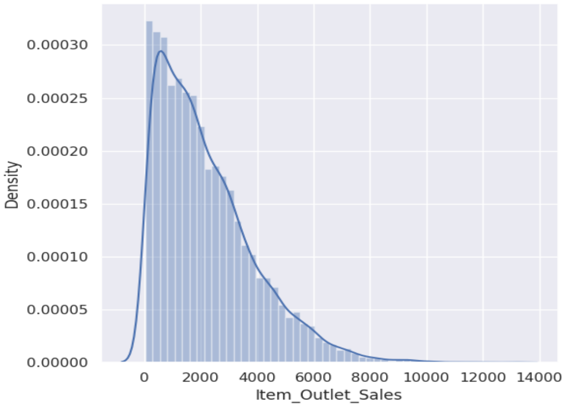


Figure 3.4

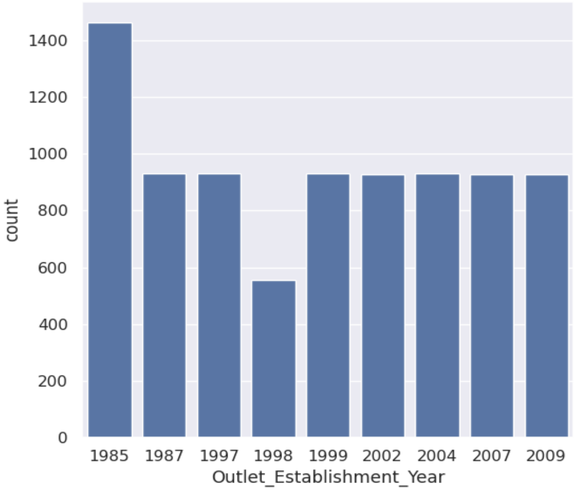


Figure 3.5

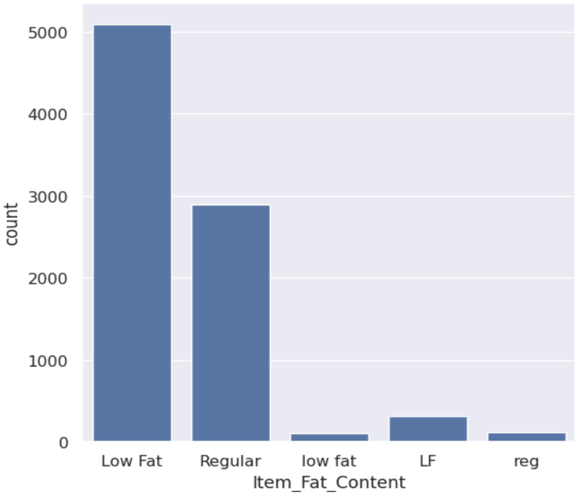


Figure 3.6

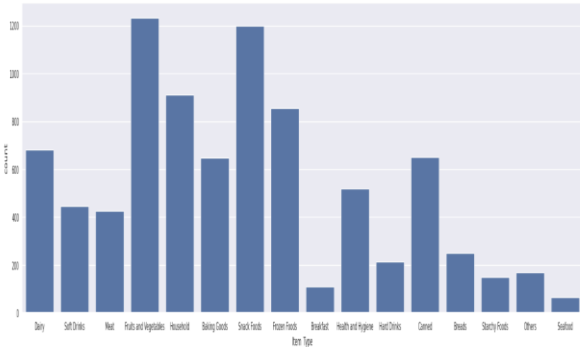


Figure 3.7

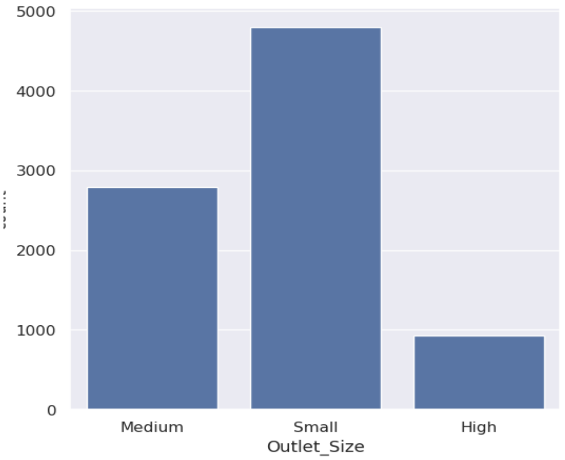


Figure 3.8

5.IMPLEMENTATION

5.ALGORITHMS USED

5.1 Linear Regression

Linear regression is one of the supervised learning algorithms. It is based on the statistical method which is used for prediction. This algorithm makes predictions based on the continuous variables such as weather, salary, price etc. It shows the linear relationship between dependent variable and independent variable. Let the dependent variable is (y) and the independent variable is (x). This algorithm shows the relationship in terms of slope.

Mathematically it can be derived as:

y=mx+c

Where

y= Dependent variable

x=Independent variable

m=Linear regression coefficient

c=Intercept

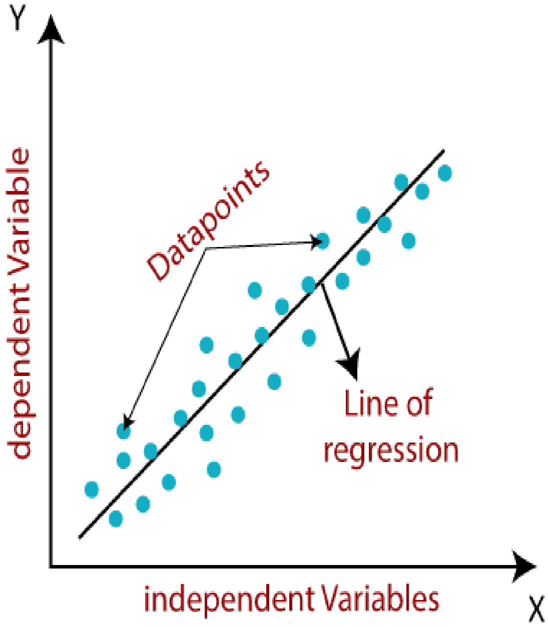


Fig 5.1.1 Linear Regression graph

5.2 K-Neighbors Regressor

K Nearest Neighbour is one of the easiest and widely used supervised learning algorithms. It can be used for regression as well as classification. The idea behind KNN algorithm is to find and group the data points which all have more similarities. Also called Lazy Learner algorithm. The working of KNN is first to select the number of neighbors, simply say number of clusters to be formed and calculate euclidean distance of K neighbors. The formula for calculating the euclidean distance is as follows:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

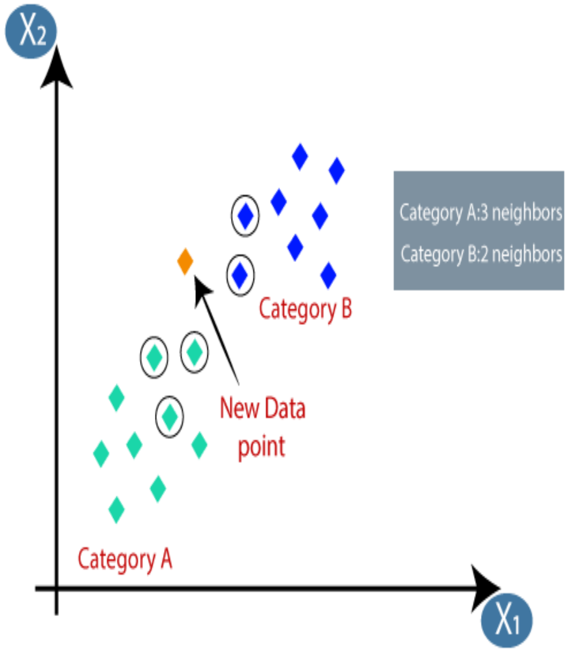


Fig 5.2.1 K-Neighbours illustration

Data points are forming a group based on minimum euclidean distance.

5.3 XGBoost Regressor

XGBoost means Extreme Gradient Boosting. It comes under Boosting which is a popular ensemble learning algorithm to classify. In this algorithm, a decision tree is used to classify the data points. In XGBoost, a decision tree is trained and a dataset is passed as input to that decision tree that generates some predictions. These predictions consist of some amount of error, These errors are residual errors. In the next set of iteration again the combination of these residual errors and data set will pass as input to the next set of decision trees. In XGBoost various optimization and regularization techniques are used to enhance the accuracy, to reduce overfitting and underfitting.

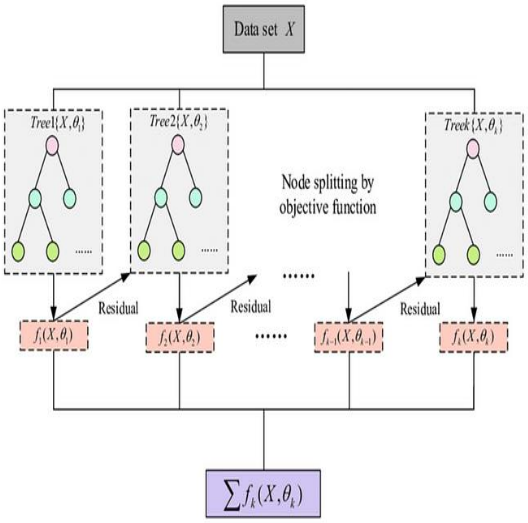


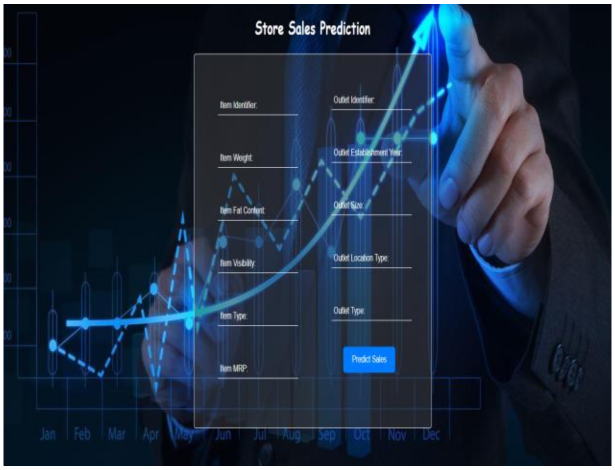
Fig 5.3.1 XG Boost illustration

8.RESULTS

With the help of Machine Learning techniques including Linear Regression, K Nearest Neighbors algorithm, XGBoost algorithm, it is possible to predict future selling rates of various outlets in the shop. Some factors Variance Score, Training and Testing accuracies as to the precision of results is included as follows below for the three algorithms.

Table 8.1: Comparitive analysis of algorithms

ALGORITHM USED	TRAIN DATA ACCURACY	TEST DATA ACCURACY



Store Sales Prediction

Item Identifier:

Item Weight:

Item Fat Content:

Item Visibility:

Item Type:

Item MRP:

Outlet Identifier:

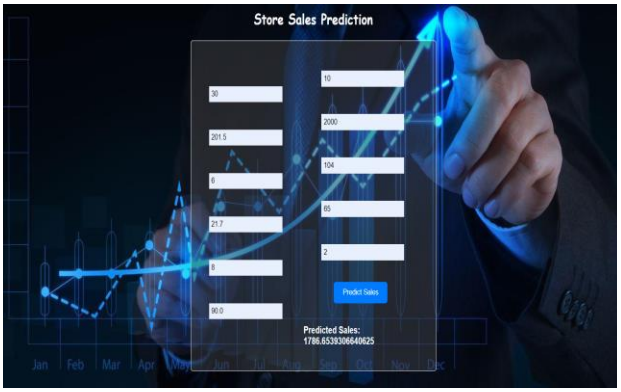
Outlet Establishment Year:

Outlet Size:

Outlet Location Type:

Outlet Type:

Fig 8.1 Mart Sales Prediction



Store Sales Prediction

30

201.5

6

21.7

8

90.0

10

2000

104

65

2

Predicted Sales:
1786.6539306640625

Fig 8.2 Predicted Value

9. CONCLUSION

When traditional methods failed to support the revenue generation of the business organizations, adopting Machine Learning approaches turned out to be helpful for planning business strategies taking into account the consumers' purchasing behaviors. The sales forecasting with reference to various factors like previous year sales, new product sales, etc. , thus providing actual strategies required for raising sales and setting their foot firm in the competitive world.

10.REFERENCES

[1] L. Bornmann and H. D. Daniel, welches in:We know that h index is a measure of the performance of a journal or a researcher over a given time frame. Jasmin, K. (2007) The impact of personalized email campaigns on e-mail list rental and co-registration. Journal of the American Society for Information Science & Technology, 58 (9), 1381-1385.

[2] Bornmann L, & Daniel H. D. (2009). The state of h index research: In the light of the above discussions, it can be posited that while the h index offers a more accurate depiction on the productivity of its subject of study than the ISI, it may not be the perfect way of measuring the research performance of individuals or institutions. EMBO reports, 10(1), 2-6.

[3] Carpenter, M. P. , & Narin F. Predictors of freshman course difficulty: Cross sectional data. ISSN: 0005-9360, Information Sciences The adequacy of the science citation index as an indicator of international scientific activity. Information processing & management, 20 (4-5), 447-454.

[4] Dašić P, Moldovan L, Grama L Effect of Protein Kinase C on Fatty Acid Composition in Adipose Tissue of Diabetic Rats. Publication analysis of research papers emerging from Romanian and Serbian institutions in the SCI, SCI-E and SSCI citation indices. Procedia Technology, 19, 1075-1082.

[5] Egghe, L. (2006). G-index and complementary concepts in the organizational framework. Scientometrics, 69(1), 131-152. [6] Egghe, L. (2006). An improvement of the h-index: The g-index The h-index is a measure of an individual's academic and research performance, which is integral for academic promotion and tenure assessments worldwide, particularly in the United States of America. ISSI.

[6] Garfield, E. (2007). It returned back a list of the evolution of science citation indexes in the university of science technology. International microbiology, 10(1), 65.

[7] Hirsch, J. E. (2005). To tier or not to tier?: A synthesis and evaluation of a decade of research on citation practices by humanities scholars. Academicus: The International Scientific Journal , 6, 5 – 27. According to the given context, it can be defined as: "The term that has been used to create a measure of productivity of a scientist". PNAS. 102 (46): 19849 19852.

[8] Jacsó, P. (2008). CASH HAMMOND.It is evident that the availability of knowledge resources has increased with the rise of internet usage, as mentioned in an article published in the Online Information Review where it highlighted that knowledge resources are accessible by 32(4), 524–535.

[9] Malin, M. V. (Guest editor), American Notes: Archibald MacLeish's last Odyssey-Part I, The Kansas Quarterly, Vol. IX No. 4, Summer 1968 Science citation index: On this basis, a new concept in the development of indexing is presented. Library Trends, 16(3), 374-374.

[10] K. Punam, R. Pamula, and P. K. Jain, "A two-level statistical model for big mart sales prediction," in 2018 International Conference on Computing, Power and Communication Technologies (GUCON), Sep. 2018, pp. 617–620, doi: 10. 1109/GUCON. 2018. 8675060.

[11] P. Das and S. Chaudhury, "Prediction of retail sales of footwear using feedforward and recurrent neural networks," Neural Computing and Applications, vol. 16, no. 4–5, pp. 491–502, May 2007, doi: 10. The1007/s00521-006-0077-3.

[12] S. Beheshti-Kashi, H. R. Karimi, K. -D. Thoben, M. Lütjen, and M. Teucke, "A survey on retail sales forecasting and prediction in fashion markets," Systems Science & Control Engineering, vol. 3, no. 1, pp. 154–161, Jan. 2015, doi: 10. 1080/21642583. 2014. 999389.

[13] S. Asur and B. A. Huberman, "Predicting the future with social media," in 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Aug. 2010, pp. 492–499, doi: 10. 1109/WI-IAT. 2010. 63. [14] A. Bermingham and A. F. Smeaton, "On using Twitter to monitor political sentiment and predict election results," in Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011), 2011, pp. 2–10.

[15] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls : "Linking text sentiment to public opinion time series," in Proceedings of the Fourth International Conference on Weblogs and Social Media 2010, pp. 23–26. [16] E. Gilbert and K. Karahalios, "Widespread worry and the stock market," in Proceedings of the Fourth International Conference on

Weblogs and Social Media ICWSM 2010, 2010, pp. 23–26.

[17] R. K. Agrawal, F. Muchahary, and M. M. Tripathi, “Long term load forecasting with hourly predictions based on long-short-term memory networks,” in 2018 IEEE Texas Power and Energy Conference (TPEC), Feb. 2018, pp. 1–6, doi: Introduction: 10. Eigenvalues and Eigenvectors1109/TPEC. 2018. 8312088.