IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 4, April 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

International Journal of Innovative Reseh in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 4, April 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1004126|

Sales forecasting Using Machine Learning

Bharati Gawale, Anuja Bangal, Priyanka Gawale, Aishwarya Suryawanshi, Prof. N.B.Pokale

Dept. of computer Engineering, TSSM'S BSCOER, Narhe Pune, Maharashtra, India

ABSTRACT: supermarket run-centres , Big Marts keep track of each individual item's sales data in order to anticipate potential consumer demand and update inventory management. Anomalies and general trends are often discovered by mining the data warehouse's data store. For retailers like Big Mart, the resulting data can be used to forecast future sales volume using various machine learning techniques like big mart. A predictive model was developed using Linear regression, Random forest regression for forecasting the sales of a business such as Big -Mart, and it was discovered that the model outperforms existing models. The aim of this paper is to analyze the sales of a big superstore, and predict their future sales for helping them to increase their profits and make their brand even better and competitive as per the market trends by generating customer satisfaction as well. The technique used for prediction of sales is the Linear Regression Algorithm, which is a famous algorithm in the field of Machine Learning.

KEYWORDS: Random Forest Regressor, Testing and Training. Data Visualization, Forecasting, Machine Learning, Regression, Linear Regression. Sales Prediction, Data Analysis.

I. INTRODUCTION

Grocery services in many major cities are getting more and more popular. Here, we are talking about the Big Mart. There are 40 such Big Mart available over here. Although day to day the numbers are being increasing and spreading. Big Mart is one of the leading stores that sales even every kind of product needs to be in day-to-day life. Here, the various categories of products have been selected and being calculated from the 10 stores and takes as consideration we are going to predict the sales it has done at given time period.

For this purpose, we are going to use the machine learning technique for the further processes that consists of different algorithms that will help for the prediction of sales.We live in a world full of data. Data surrounds us everywhere. Right from handling monthly budgets, storing information on mobile phones, buying items from stores, all of it is stored in the form of data. In our everyday lives, we have to deal with a lot of data. This data could be as small as handling your monthly budgets to big ones like the data of a Multinational Company (often referred to as big data). These big shopping complexes have a lot of data to work upon. Handling inventory, maintaining purchase from manufacturers, handling inventory costs, handling supplies data, handling with their sales, profits and quantity data, and many more. This is a tremendous task to work upon such a big dataset. Our ultimate task is generating profits and customer satisfaction and to maintain brand name. A lot of work has to be done on the dataset for its analysis and prediction**.** This whole work is done so as to check the current position of sales and find out the future expected sales so that if any decline or anomalies is found could be worked upon by doing proper market research and evaluating the trends of the market so that customer base could be increased. Also, within the store, what techniques (like putting discounts or updating inventory) could be applied so that our target customers increase their purchase and become a satisfied and a happy customer. All this is very important for any business to survive in this cut-throat competition and undoubtedly data science is very much required to fulfill this purpose. In this paper, we will describe the methodology to deal with such data along with predicting sales of the superstore for next years from the available tools like machine learning. A brief description about the processes involved in fulfilling our objective is discussed further. Also, the tools used for various processes would be even discussed. The current machine learning algorithm is very advanced and provides methods for predicting or forecasting sales any kind of organization, extremely beneficial to overcome low – priced used for prediction. Always better prediction is helpful, both in developing and improving marketing strategies for the marketplace, which is also particularly helpful

II. PROBLEM STATEMENT

To identify and determine the sales of products of each category from Big Mart stores.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 4, April 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1004126|

III. OBJECTIVES

This paper emphasis two objectives. They are the following: 1. To predict the sales done by the stores.2. To help increasing their sales.3. To determine the short-term and long-term performance.4. To take various business decisions.

Data Collection and transformation- Data is collected from various sources and organized in a single file. Next, the data cleansing process is applied. Data cleansing is the process of detecting and correcting inaccurate or obsolete records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. Data cleansing may be performed interactively with data wrangling tools, or as batch processing through scripting. After cleansing, a data set should be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage, or by different data dictionary definitions of similar entities in different stores. Also, various noises within the data is also removed. Along with that, the data is categorized properly and all blank spaces or irrelevant information within data is also removed so that analytics could be performed easily on the data.

IV. LITERATURE SURVEY

“Intelligent Sales Prediction Using Machine Learning Techniques” The fashion store dataset for three consecutive years of sales data is used. The sales prediction is done for upcoming 3 years from 2015 to 2017. Exploratory analysis stages involved in the data mining model include data understanding, preparation, modelling, evaluation and deployment. The forecast is composed of a smoothed averaged adjusted for a linear trend. Then the forecast is also adjusted for seasonality. Machine learning algorithms such as Generalized Linear Model (GLM), Decision Tree (DT) and Gradient Boost Tree (GBT) are used in prediction of future sales. Based on the performance, Gradient Boost Algorithm is provides 98% overall accuracy and the second stands Decision Tree Algorithms with nearly 71% overall accuracy and followed by Generalized Linear Model with 64% accuracy. Finally, it can be compared based on the empirical evaluation of the three chosen algorithm the best fit for the model is Gradient Boosted Tree which provides the maximum accuracy of prediction across all the algorithms.

“MachineLearning Models for Sales Forecasting” “Rosemann Store Sales” dataset is used to predict the future sales. The calculations were conducted in the Python environment using the main packages pandas, sklearn, numpy, keras, matplotlib, seaborn. Analysis is done using Jupyter notebook. Regression algorithm captures the patternsin the whole set of stores or products.The analysis includes the attributes such as meansales value of historical data, state and school holiday flags, distance from store to competitor‟s store, store assortment type are considered in prediction. Various machine learning models such as Random Forest, Neural network, Lasso regularization, Arima model and ExtraTree model are used to analyze the data. The models in the first level (ExtraTree, Lasso, Neural Network) have non-zero coefficients for their results.

The solution is based on three level models. On the first level, many models were based on thel linear regression machine learning algorithm. In the second level, models from Python scikit-learn package, Extra tree model, linear model as well as Neural network model. The results from the second level were summed with weights on the third level .The use of regression approaches for sales forecasting can give better results compared to time series methods. ExtraTree method provides more stacking weights for regressors compared to other approaches. "Walmart’s Sales Data Analysis- A Big Data Analytics Perspective". The Walmart has 45 stores in geographically diverse locations, each of the store having 99 departments. The dataset contains the weekly sales and the factors affecting sales such as (Temperature, fuel price, unemployment rate, holiday) for each store locations for 3 years. Apache data science platforms, libraries, and tools are used in this work. Tools like Hadoop Distributed File Systems (HDFS), Hadoop MapReduce framework and Apache Spark along with Scala, Java and Python high-level programming environments are used to analyse and visualize the data. Machine learning library is employed with a simple regression model to predict future sales. The results are predicted from data analysis and based on the predicted results the Retailers need to plan and evaluate according to the market driving factors which are, and not limited to, the temperature, fuel prices holidays, human resources, geographical location and many more. Effective and efficient supply chain, inventory, human resource management is needed to avoid losing competitive edge in the market, especially planning sales at different locations

“Forecast of Sales of Walmart Store Using Big Data Applications” The forecasting process uses Walmart sales data. The different types of stores such as convenient store, department store, luxury store, super market, shopping malls etc. helps in determining the business models and strategy of operations. The process involves the stages such as determine dependent and independent variables, develop forecast procedures, select forecast analysis method, gather and analyze data, present assumptions about data, make and finalize forecast, evaluate results. The strategy includes the collection of huge data of sales and then it is transferred on HDFS (Hadoop distributed file system) and map reduced is performed on the data sets. The Holt winters algorithm is used to predict the sales .The seasonality, trend and randomness is observed in the algorithm .The algorithm is used for train data sets and then the sales prediction. The final results represents that the numerical representation of the forecasted sales and the accuracy of sales predicted is measured by 80% low confidence sales, 80% high confidence sales and 95% low confidence sales and 95% high confidence sales , error factor can be found between the predicted sales

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 4, April 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1004126|

and the observed sales data, i.e. to find the error factor of month June in both the predicted sales and the observed sales data then the difference between predicted sales and the observed sales data is obtained, if the difference between them is very low or negligible and thus the sales prediction will be more accurate.

V. PROPOSED SYSTEM

In proposed system, various analysis method is used in order to predict data from the sales record of daily sales of 1,115 stores. The project makes use of the complete dataset and make predictions for the upcoming 6 weeks. Various algorithms such as Linear Regression, Time series analysis, Benchmark method (Seasonal Naive method), Exponential smoothing method, Arima model method are used to forecast the data, so that predictions can be compared across all the results, to determine the most accurate result. The forecasted data of each algorithm are compared and the final efficient resultset is identified for the prediction of sales.Time series data is used by the algorithm, so that data for the upcoming 3 years can be predicted with less variation across time and data. A time series can be broken down to its components to systematically understand, analyze, model and forecast it. Variance and effect of seasonality is monitored, so that it should not increase over time, which will increase chance to predict most accurate results.

VI. ALGORITHM

A. Linear Regression

• Build a fragmented plot.1) a linear or non-linear pattern of data and 2) a variance (outliers). Consider a transformation if the marking isn't linear. If this is the case, outsiders, it can suggest only eliminating them if there is a non-statistical justification. • Link the data to the least squares line and confirm the model assumptions using the residual plot (for the constant standard deviation assumption) and the normal probability plot (for the normal probability assumption) A transformation might be necessary if the assumptions made do not appear to be met.

• If required, convert the data to theleast square using the transformed data, construct a regression line. • If a change has been completed, return to the previous process 1. If not, continue to phase 5. • When a "good-fit" classic is defined, write the least-square regression line equation. Consist of normal estimation, estimation, and Rsquared errors.

Linear regression formulas look like this:

$$Y=o1x1+ o2x2+\ldots\ldots\ OnXn$$

R-Square: Defines the difference in X (depending variable) explains the total variance in Y (dependent variable) (independent variable).

B. Random Forest Regressor –

A Random Forest is an outfit method that can perform both the regression and classification tasks by using the multiple decision trees and Bootstrap Aggregation technique, generally known as bagging [7], [14]. The fundamental thought of this is to combine multiple decision trees in deciding the final outputs instead of depending on any individual decision tree. The features of Extreme gradient boosting are listed below: i). Sparse Aware – The missed data values are automatic handled in XGBoost. ii). It can support parallelism in construction of tree. iii). It has continuous training and that leads to fitted model that will be able to more boost with new dataset. All the models receive parameters as inputs and these are separated into training and test sets. Here, test set will be used for predicting the sales. The Random forest (RF) model is an additive model that predicts the sales by combining decisions from a sequence of base models. Finally, the equation of RF model is given as Eq. (1):

$$P(x)=f0(x)+f1(x)+f3(x)+\ldots\ldots\ldots\ldots\ldots\quad (1)$$

Different types of models have different advantages. The random forest model is the best at handling tabular data with categorical features, or numerical features with least than several categories. In contrast to linear models, random forests can catch non-linear collaboration between the features and the target. Trees run in parallel in Random Forest. No interaction is present between the trees while building it. The Random forest algorithm is given as below: Step1: Select the random samples from the data set.

Step2: Construct the decision trees for all samples to obtain the prediction results. Step3: Voting should be done for all the predicted results. Step4: Select the majority voted prediction result as the final prediction result. The data set is divided into testing and training sets by the 80% and 20% ratios respectively. The standard scalar methods are adopted to normalize the each field value. The maximum leaf node 900 with 20 estimators are utilized.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 4, April 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1004126|

VII. CONCLUSIONS

We have proposed a system for prediction of sales of big mart which consists of input from different stores and collecting those data and after using all those machine learning algorithms prediction is done. The motive of this system is to provide the sales which would be further beneficial for increasing the sales of the stores also to setup a comparison between the past and future results. Here we have used the machine learning algorithms basically the two of them those were linear regression algorithm and random forest algorithm. Machine Learning (ML) can be used for the various tasks. This research work presents the use of ML algorithm for the prediction of the amount that a customer is likely to spend on next "Black Friday" sale. It has been performed that the exploratory data analysis is used to find interesting trends from the dataset .This research work suggests that when the user tries to predict the product that the customer is more likely to purchase, according to the customer's gender, age and occupation. Experiments states that our method can produce more accurate prediction when compared to the techniques like decision trees, ridge regression etc. A comparison of various methods are summarized. Also, we have concluded that our model with lowest RMSE perform better than exiting models.

REFERENCES

[1] Yukta Kaneko, Katutoshi Yada,"A Deep Learning Approach for the Prediction of Retail Store Sales., Institute of Electrical and Electronics Engineers, Electronic ISSN: 2375-9259, 12-15 Dec. 2016. https://ieeexplore.ieee.org/document/7836713

[2] Muhhamad Adnana Khan, Shazia Shaqib, Tahir Alyash, Anees Ur Rehman, Yousaf Saeed, Asim Zeb,"Effective Demand Forecasting Model Using Business Intelligence Empowered With Machine Learning. Institute of Electrical and Electronics Engineers, Electronic ISSN: 2169-3536, 19th June, 2020. https://ieeexplore.ieee.org/document/9121220

[3] James Le,"THE TOP 10 MACHINE LEARNING ALGORITHMS EVERY BEGINNER SHOULD KNOW, September, 2020 https://builtin.com/data-science/tour-top-10-algorithms-machine-learning-newbies

[4] Reena Shaw,"The 10 Best Machine Learning Algorithms for Data Science Beginners. , June, 2019. https://www.dataquest.io/blog/top-10-machine-learning-algorithms-for-beginners/

[5] Beheshti-Kashi, S., Karimi, H.R., Thoben, K.D., Lutjen, M., Teucke, M.: " A survey on retail sales forecasting and prediction in fashion markets, " Systems Science & Control Engineering 3(1), 154, 161(2015)

[6] Smith, Oliver, and Thomas Raymen. " Shopping with violence: Black Friday sales in the British context. " Journal of Consumer Culture 17.3 (2017): 677-694.

[7] Majumder, Goutam. " ANALYSIS AND PREDICTION OF CONSUMER BEHAVIOUR ON BLACK FRIDAY SALES. " Journal of the Gujarat Research Society 21.10s (2019): 235-242.

[8] Challagulla, Venkata Udaya B., et al. " Empirical assessment of machine learning based software defect prediction techniques. " International Journal on Artificial Intelligence Tools 17.02 (2008): 389-400.

[9] Chu, C.W., Zhang, G.P.: " A comparative study of linear and nonlinear models for aggregate retail sales forecasting, " International Journal of production economics 86(3), 217{231(2003) }

[10] Makridakis, S., Wheelwright, S.C., Hyndman, R.J.: " Forecasting methods and applications, " John wiley & sons(2008)

[11] Correia, Alvaro, Robert Peharz, and Cassio P. de Campos. " Joints in Random Forests. " Advances in Neural Information Processing Systems 33 (2020).

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details