International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)
doi : https://doi.org/10.32628/IJSRSET

Predictive Analysis for Big Mart Sales Using ML Algorithms

Soubiya Hussain*[1], Dr. G. Kalaimani[2]

*[1] M.Tech Student CS, Department of Computer Science Engineering, Shadan Women's College of Engineering & Technology, Telangana, India

*[2] Professor, Department of Computer Science Engineering, Shadan Women's College of Engineering & Technology, Telangana, India

Article Info
Volume 9, Issue 5
Page Number : 282-289

Publication Issue :
September-October-2022

Article History
Accepted : 10 Oct 2022
Published: 30 Oct 2022

ABSTRACT

Big Marts, which are distribution centers for supermarket chains, now keep tabs on sales volume and revenue numbers for each product to anticipate domestic consumption and adjust inventory control. Examining the data warehouse's server database often reveals inconsistencies and overarching patterns. Companies like Big Mart can use the data with a variety of machine learning techniques to predict future product sales. Many different machine learning algorithms, including Linear Regression, Ridge Regression, Lasso Regression, Decision Tree Regression, Random Forest Regression, Support Vector Regressor, Adaboost Regressor, and XGBoost Regression, have been employed in this project to forecast Big Mart product sales. We find that XGBoost Regression performs the best in predicting sales volume among the listed algorithms. To this end, we have developed a model with XGBoost Regression and optimized it for maximum precision. This model is available through a flask application; users simply log in, specify the product's parameters, and receive sales forecasts.

Keywords: Linear Regression, Polynomial Regression, Ridge Regression, Xgboost Regression

I. INTRODUCTION

The fierce and increasingly aggressive competition between specialty shops and megastores is largely attributable to the rise of foreign retailers and online shopping. The company's management of inventory, shipping, and operational activities must be able to attract a large number of customers in a short amount of time and predict the amount of revenue for every product. To triumph over less expensive methods of prediction, the modern machine learning approach provides methods for estimating rather than forecasting sales trends for any kind of business. An ever-improving business plan, which is aided by more precise forecasting, is also of great value.

There has been a lot of work done up to this point that was legitimately meant for the field of transaction forecasting. The extensive research conducted on supermarket alliances is briefly summarised in this section. Regression, the Auto-Regressive Integrated Moving-Average (ARIMA), and the Auto-Regressive Moving-Average (ARMA) are just a few of the additional Measurable techniques that have been used to develop some deal prediction standards. As an

antidote to the stress of daily food deal anticipation, A. S. Weigend et al. suggested a hybrid approach involving the Auto-Regressive Integrated Moving Average (ARIMA) and the occasional quantum relapse strategy. Predicting future transactions is a challenging problem, affected by both internal and external factors.

II. PROBLEM STATEMENT

In recent years, people's ability to spend money has risen dramatically, both in traditional retail settings and online. Throughout the year, supermarkets frequently release a slew of deals to celebrate holidays like the New Year, Christmas, and others. Since it's a given that business will be brisk, top brass must make accurate projections about product demand and keep stock levels under control. Supermarkets typically stock up on products and sell them all before the end of the period. If it doesn't, the market risks incurring enormous losses because its predictions were off. Current methods cannot reliably extrapolate future product sales from historical data. Managers are solely responsible for doing this by carefully analyzing historical data and making best-guess projections of future sales activity in light of a variety of factors. This task can only be completed manually, as no automated system is currently capable of doing it.

III. OBJECTIVE

This study's primary objective is to foretell the sales volume of products carried by a supermarket. To do this, we have been using the Big Mart sales data set from Kaggle. The R2 score is compared after the dataset has been processed, analyzed, and fed to multiple regression algorithms. Stock booking and inventory upkeep are not part of this project's remit.

IV. Existing system:

Sales are guaranteed to skyrocket around holidays like New Year's and Christmas, so it's crucial that management accurately predicts these spikes in demand so that they can keep inventory levels stable. Supermarkets typically stock up on products and sell them all before the end of the period. If it doesn't, the market risks incurring enormous losses because its predictions were off. Current methods cannot reliably extrapolate future product sales from historical data. Managers are the only ones responsible for doing so, and their efforts are needed to meticulously examine historical data, factor in multiple events, and estimate sales volume.

Problems with the Current System

- May be inaccurate, increasing the risk of financial setbacks.
- This requires human intervention.

V. A PROPOSED SYSTEM

It is proposed that multiple machine learning algorithms, such as Linear Regression, Ridge Regression, Lasso Regression, Decision Tree Regression, Random Forest Regression, Support Vector Regressor, Adaboost Regression, and XGBoost Regression, be used to forecast Big Mart product sales, with the best-performing algorithm being used to build a model to forecast sales volume. Ideally, this model would be hosted on a flask application, where users could log in, input product details, and view sales projections in real-time.

Advantages of Proposed System

- Computerized process
- Accurate
- Simple in design and implementation.

Machine Learning

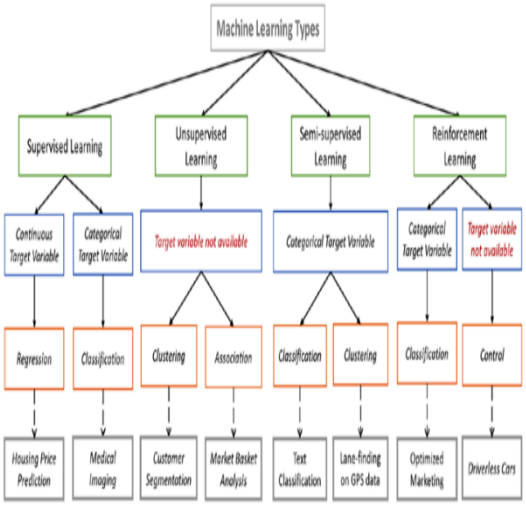
There are four types of machine learning. They all differ in their approach, the data type they take as input and give output.
The four types are:

1.Supervised Learning: it is a learning in which we train the models using the data which is well labelled that we already know the answers. Basically, it is a task of learning a function that represents an input to an output based on example input output pairs. It infers a function from labeled training data comprising of a set of training examples.

2.Unsupervised Learning: it is used on some data in which we don't know the output allowing the algorithm to act on that data without any guidance. It basically groups the data based on the similarities and pattern without any prior knowledge of data. Unlike Supervised Learning, no training will be given to the machine. Therefore, machine is limited to find the hidden structure in an unlabeled data by itself.

3.Semi-Supervised Learning: in the learning, it is assumed that the data is the combination of both labeled as well as unlabeled. Its major aim to extract the data from the unlabeled data that could enable learning a discriminative model with better execution.

4.Reinforcement Learning: it basically means improving the performance of the machine using trial and error experience. There is unkind teacher is given, where if a machine predicts wrong then penalty is given and if machine predicts correct then some reward is given.



PYTHON:

The Python 3.8 programming language is being used to create the project's applications. If you're looking for a powerful programming language that doesn't require you to memorize a bunch of syntaxes, look no further than Python. Python is an easy programming language to learn because its syntax is so similar to regular English. It can run sophisticated mathematical computations, incorporate machine learning algorithms, and support object-oriented programming. Its dynamic typing and data structures allow for swift application development. Python is widely regarded as a readable programming language because it uses a larger percentage of English words than any other language. The exponential growth of available data. Since Python's interpreter performs syntax analysis at runtime, the language doesn't need a compiler to fix syntax errors and compile programs. It can run on any OS and comes with a tonne of prebuilt libraries and packages for a wide range of tasks. It's a simple language that's straightforward to study, use, keep up with, and improve. You can access the vast majority of popular databases on the market today with just a few lines of Python code. Using GI programming can help you create applications for both Windows and the web. This language is unique in that it allows developers to choose between functional, structural, and object-oriented programming. Python code is easily portable to other languages, and it even has garbage collection built in!

FLASK:

Flask is a framework for creating small, fast, and simple web apps. It does this by making available Python's library system. The WSGI Development Environment Flask is a web application framework based on the Andy jinja 2 template engine. Flask is a microframework due to its lightweight nature. The Web Server Gateway Interface (WSGI) toolkit is a standardized protocol and interfaces for implementing common web objects like requests, responses, and helper functions. Python now uses this standard as its baseline for building web applications. It is the universally accepted interface specification for establishing communication between a web server and

a web application. Jinja 2 is a popular Python templating engine. Using this method, templates can be linked to various data sources to generate dynamic web content. Flask's architecture makes it easy to create basic web applications. Unfortunately, it lacks basic features like database support and built-in form validation abstraction layers. However, the flask allows for the incorporation of all these features. This flask is a lightweight web framework that can be easily customized.

MYSQL:

Our project makes use of MySQL, an open-source RDBMS, to establish database connections and store information. To store, retrieve, and modify information in a database, My SQL relies heavily on structured query language. There are rows and columns in the tables that hold the data in the database. To facilitate a client-server architecture, MYSQL was developed. The user, or client, connects to the MYSQL client to perform tasks on the database, such as creating tables, modifying data, and retrieving records. The MYSQL client initiates an operation, which is sent to the server, and the server either completes the operation successfully or returns an error message.

Regression Algorithm

BigMart's data scientists have compiled sales information from 10 stores in different cities for 1559 products in 2013. The method was developed by Barun Waldron and Sanjeev trivedi[1]. Finding out if certain characteristics of products and/or stores significantly affect sales is the primary goal. To accomplish this, they developed a predictive model to ascertain the sales of each product in a specific store, thereby enabling BigMart to increase sales through the discovery of optimal product organization within stores. The method they used was called Linear Regression. While this makes the methods they use more refined and easier to analyze, it comes at the cost of some accuracy.

For businesses in the retail, logistics, manufacturing, marketing, and wholesale sectors, accurate sales forecasting is crucial. This paves the way for better resource management in the future of businesses. The authors Conley, T G, and Gleason, D W[2] estimated sales revenue to help better plan for the company's future expansion. In their paper, they present a two-tiered method for predicting product sales at a store that outperforms the most common single-model predictive learning algorithms. In this method, 2013 sales data from Big Mart is used as the basis for analysis. Correct predictions rely heavily on data exploration, data transformation, and feature engineering. The outcome proved that a two-level statistical approach outperformed a single model approach due to the latter's lack of information and the former's ability to better predict outcomes. Decision trees and linear regression were the methods used. It has the benefit of being extremely trustworthy. On the downside, it's less precise.

It can be challenging to analyze large datasets and derive useful insights from them. Consequently, a robust and efficient data mining tool is required for mining complex data sets to extract the information and make better decisions in the future. For this purpose, the potent and efficient xgboost is employed. The efficiency gained from using the decision tree regressor's built-in packages (ggplot2, VIM, etc.) is considerable. An essential part of statistics is the process of transforming raw data into knowledge and understanding it, and Alishahi et al.[3] used the open-source data analysis environment and programming language R to accomplish this. It enables users to perform several crucial operations for the efficient processing and analysis of big data. R includes many pre-built statistical modeling and machine learning algorithms that can be used to develop data products and conduct reproducible research. They implemented the R linear regression, RF, DT, and R algorithms. Handle such a large dataset with missing values and regularities. But the precision is better now than it was before.

Classifiers Algorithms

The Big Mart Company uses Sales Prediction to forecast sales of their various products across their many retail locations and cities. The number of goods was a topic covered in a paper by Aghion et al.[4]. The proliferation of retail locations is making accurate manual forecasting difficult. Sellers' ability to anticipate the correct level of product demand is crucial in terms of physical location, time, and money. Depending on storage costs and availability, sellers may need to move their products quickly. Xgboost's - Linear Regression was used in this approach. When compared to alternative methods, their accuracy is very high. Some people may have a skewed perception of the drawback. Many authors have researched sales forecasting and sales forecasting analysis, which can be summarised as follows: This paper also investigates and describes in detail the automated process of knowledge acquisition using computational and statistical methods. Machine learning is the process by which a machine acquires knowledge from its experiences, based on data, using statistical or computational methods. In [6], we are introduced to several machine learning methods, each with a specific industry in mind. The most popular data mining strategy was identified by Pat Langley and Herbert A. Everyone today relies heavily on data analysis to improve their decision-making, and this is especially true in the modern era. Tackling big data analysis and gleaning useful insights from it is a challenge. So, a robust and efficient data mining tool is required for the mining of complex datasets to extract the information and make better decisions in the future. In this case, we will use R, a powerful and free data mining tool developed by the community. R comes with a plethora of pre-installed packages. That efficiency comes from sources like Fikes, Richard E et al.[7ggplot2,]'s VIM, etc. R is a free and open-source software framework and programming language for statistical computing and analysis. Data analysis, an essential part of statistics, is the transformation of raw data into useful information. It enables users to perform several crucial operations for the efficient processing and analysis of big data. R includes many pre-built statistical modeling and machine learning algorithms that can be used to develop data products and conduct reproducible research. Linear Regression, Decision Trees, K-Means, and Naive Bayes can all be used with this approach. Various algorithms, such as random forest and eclat', are available to us, and we plan to use them in our future projects. We also use means to cluster the dataset into categories, and we use a naive Bayes classifier to determine the fat content of individual items. The primary goal of this paper is to demonstrate strategies for dealing with a massive dataset that frequently contains missing values. In the realm of business, the Rule Induction

When compared to other data mining methods, the RI method is superior. While in [8] we learn how to forecast sales for a pharmaceutical distributor. This paper is concerned with two additional issues: I the stock state should not experience out-of-stock, and (ii) it avoids customer dissatisfaction by predicting the sales that manage the stock level of medicines. In [3], the authors discuss methods for dealing with sales fluctuations in the footwear industry over time. Also discussed in this paper is the use of neural networks to predict weekly retail sales, which can help reduce the amount of guesswork involved in making near-term sales projections. There is a proposal for a sales forecasting model in retail that makes use of both linear and non-linear [1] analysis. The fashion industry sales forecast was done by Beheshti-Kashi and Samaneh. Forecasting sales at the mega mart is done using a two-tiered statistical approach [7]. In their proposal, Xia and Wong distinguished between classical (based on mathematical and statistical models) and modern heuristic approaches, among which they included exponential smoothing, regression, Auto Regressive Integrated Moving Average (ARIMA), and Generalized Auto-Regressive Conditionally Heteroskedastic (GARCH) methods. In most cases, the

asymmetrical nature of real-world sales data is too complex for the linear models commonly used. Several obstacles make accurate forecasting difficult, including a lack of historical data, consumer-oriented markets that face uncertain demands, and short life cycles of prediction methods. According to the author's interpretation [9], the question of "what is data analysis, and how we can do it efficiently?" is answered. This paper suggests using R for data analysis due to its many advantages, including its extensive built-in packages, its ease of use concerning implementing various machine learning algorithms, etc. Given that R is both a statistical language and a programming language, it can be used to create more accurate models for making predictions and to improve visualization. The survey's authors concluded, then, that data analysis with R is more productive. Assesses the strengths and weaknesses of three widely used data mining tools (Rapid Miner, Weka, and R) through a detailed comparison and analysis.

This paper provides an explanation of how to use R, Weka, and Rapidminer for time series analysis and structural health monitoring.

The author provides a concise explanation of data analysis and provides tips for getting the most out of the process. The author of this paper suggests using Random Forest for data analysis due to its many useful features, including its robust data exploration capabilities, its built-in packages, and its ease of implementation of several machine learning algorithms. Because of its dual nature as a programming language and a statistical language, R facilitates accurate model prediction and enhanced visualization. Researchers concluded that the boost significantly improved the efficiency of data analysis after conducting their survey. Examines the capabilities of three well-known data mining mechanisms (Rapid Miner, Weka, and R) and how they can be applied to the problem of monitoring the condition of structures. This paper explains how to use R, Weka, and Rapid miner for time series analysis for structural health monitoring, including how to visualize data, apply filters and statistical models, and more. The authors of this paper provide an in-depth analysis of the research done to date on the subject of estimating product sales and profits. The implementation of the proposed system is described in the following chapter. In this paper, we take a look at the challenge of trying to predict traffic on an online store. To solve this problem, we proposed a stacked generalization strategy that uses regressors at multiple levels. Separately from the overall model, we have also evaluated the performance of individual classifiers. Experiments have shown that our method is at least as effective at predicting demand as individual classifiers, and often performs even better while requiring significantly less data for training (only 20 percent of the dataset). We believe that by using more data, our method will be able to make much more accurate predictions. Because the proposed model is statistically indistinguishable from the random forest, it can be used to predict demand with a minimum of information. The results of this effort will be incorporated into a later effort to solve the problem of optimal pricing. It was suggested that long-term annual growth factors could be used to predict the demand for electricity in the future. Based on the weather data, a forecasting method using multiple linear regressions for bicycle rental demand were proposed.

VI. PROPOSED MODULAR IMPLEMENTATION

An outline of the project's proposed modular structure is provided below.

It has two separate parts:

1. Admin
2. User

Admin Module:

Activities such as these fall under the purview of the system administrator.

Step 1: Transferring the Dataset

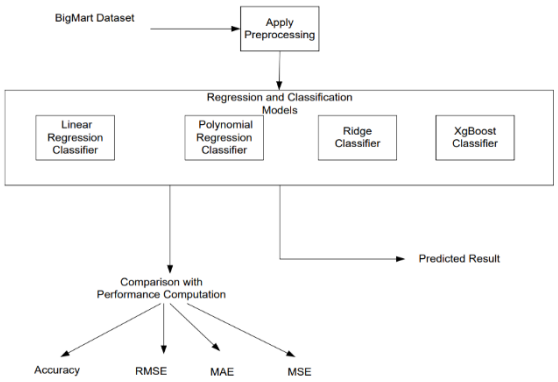
Step 2: Analysis of the data set

Step 3: Pre-processing of Data

Step 4: Dividing the data into a test and a training set
Step 5: Teaching the model to use a number of different regression technique.
Step 6: Analyze how well the algorithms work with the provided data set.
Step 7: The XGBoost regressor algorithm is used to build the model in step seven.

Architecture diagram:

An architecture diagram is a visual representation of a system's concepts, principles, elements, and components. It is an abstract representation of the system's structure and behaviour. It's possible to define different levels of abstraction, such as a purely conceptual one, in which only the ideas behind the system are shown. Logical abstractions provide further insight into the workings of the concepts by illustrating their underlying principles and constituent parts. A system's design can be seen at the physical abstraction level, the lowest level of abstraction. Any of the aforementioned three abstraction levels may be shown in the Architecture diagram as appropriate.



VII. Conclusion and Future Work

Conclusion:

We built and deployed a machine learning model to predict sales of different products in a superstore as part of this project. To that end, we've decided to use the publically available Big Mart sales data set from kaggle.com. We have prepared the data for analysis by performing standard data preparation steps such as sorting it into training and test sets and then feeding it to a variety of regression algorithms, including Linear regression, support vector regression, ridge regression, lasso regression, decision trees regression, random forest regression, AdaBoost regression, and xgboost regression. We've seen that conversation and can attest that it's spot-on. Since then, we've adjusted the xG boost regressor so that it's approximately 98% accurate. To date, we have not found a machine-learning model that can compete with this one on this dataset. We hope to incorporate ARIMA time series analysis into this project down the road.

Future Work:

As part of future enhancement, we can combined with effective data mining methods and properties, the traditional means could be seen to make a higher and positive effect on the overall development of corporation's tasks on the whole. One of the main highlights is more expressive regression outputs, which are more understandable bounded with some of accuracy. Moreover, the flexibility of the proposed approach can be increased with variants at a very appropriate stage of regression model-building. There is a further need of experiments for proper measurements of both accuracy and resource efficiency to assess and optimize correctly.

VIII. REFERENCES

[1]. Ching Wu Chu and Guoqiang Peter Zhang, A comparative study of linear and nonlinear models for aggregate retails sales forecasting, Int. Journal Production Economics, vol. 86, pp. 217231, 2003.

[2]. Wang, Haoxiang. "Sustainable development and management in consumer electronics using soft computation." Journal of Soft Computing Paradigm (JSCP) 1, no. 01 (2019): 56.- 2. Suma, V., and Shavige Malleshwara Hills.

[3]. Suma, V., and Shavige Malleshwara Hills. "Data Mining based Prediction of Demand in Indian Market for Refurbished Electronics." Journal of Soft Computing Paradigm (JSCP) 2, no. 02 (2020): 101110

[4]. Giuseppe Nunnari, Valeria Nunnari, Forecasting Monthly Sales Retail Time Series: A Case Study,

Proc. of IEEE Conf. on Business Informatics (CBI), July 2017.

[5]. https://halobi.com/blog/sales-forecasting-five-uses/. [Accessed: Oct. 3, 2018]

[6]. Zone-Ching Lin, Wen-Jang Wu, Multiple Linear Regression Analysis of the Overlay Accuracy Model Zone, IEEE Trans. on Semiconductor Manufacturing, vol. 12, no. 2, pp. 229 237, May1999.

[7]. O. Ajao Isaac, A. Abdullahi Adedeji, I. Raji Ismail, Polynomial Regression Model of Making Cost Prediction In Mixed Cost Analysis, Int. Journal on Mathematical Theory and Modeling, vol. 2, no. 2, pp. 14 23, 2012.

[8]. C. Saunders, A. Gammerman and V. Vovk, Ridge Regression Learning Algorithm in Dual Variables, Proc. of Int. Conf. on Machine Learning, pp. 515 521, July 1998.IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 56, NO. 7, JULY 2010 3561.

[9]. Robust Regression and Lasso. Huan Xu, Constantine Caramanis, Member, IEEE, and Shie Mannor, Senior Member, IEEE. 2015 International Conference on Industrial Informatics-Computing Technology, Intelligent Technology, Industrial Information Integration.An improved Adaboost algorithm based on uncertain functions.Shu Xinqing School of Automation Wuhan University of Technology.Wuhan, China Wang Pan School of the Automation Wuhan University of Technology Wuhan, China.

AUTHOR PROFILE

Soubiya Hussain completed B. Tech in Computer science and engineering from Shadan women's Engineering College. Her area of interest include machine learning and data science. At present she is pursuing M. Tech in Computer science and engineering.

Dr. G. KALAIMANI received the Ph.d degree in Information and Communication Engineering from Anna University, Chennai. She has 16 years of teaching experience. Her areas of interest include computer networks, database management systems, advanced algorithm, data structure, cloud computing. At present she is working as a professor in Department of Computer Science and Engineering at Shadan Women's College of Engineering and Technology, hyderabad. She has published 14 papers in International Journal, 2 papers in International Conferences.

Cite this article as :

Soubiya Hussain, Dr. G. Kalaimani, "Predictive Analysis for Big Mart Sales Using ML Algorithms", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 9 Issue 5, pp. 282-289, September-October 2022.
Journal URL : https://ijsrset.com/IJSRSET229544