EAI Endorsed Transactions
on Internet of Things

Research Article EAI.EU

Big Mart Sales Prediction using Machine Learning

Koh Ya Wen[1,*], Minnu Helen Joseph[2] and V. Sivakumar[3]

[1]Asia Pacific University of Innovation and Technology, Kuala Lumpur, Malaysia
[2]Asia Pacific University of Innovation and Technology, Kuala Lumpur, Malaysia
[3]Asia Pacific University of Innovation and Technology, Kuala Lumpur, Malaysia

Abstract

INTRODUCTION: Sales prediction, also known as revenue forecasting or sales forecasting, refers to the process of accurately and timely estimating future revenue for manufacturers, distributors, and retailers, providing them with valuable insights. Sales prediction plays a crucial role in various industries, particularly in sectors such as retail, automotive leasing, real estate transactions, and other conventional businesses.
OBJECTIVES: This paper focuses on developing a sales prediction model for Big Mart, a supermarket chain, using machine learning algorithms. The developed model aims to provide Big Mart with accurate sales forecasts, enabling better decision-making, improved profitability, and enhanced customer service.
METHODS: The study utilises the CRISP-DM methodology and explores various machine learning algorithms, including Linear Regression, Decision Tree, Random Forest, XGBoost, Stacked Ensemble Model, and K-Nearest Neighbours (KNN). The dataset used for model development is sourced from Kaggle and includes information about products, stores, and sales. Pre-processing techniques are applied to handle missing data and feature engineering.
RESULTS: The XGBoost Regression Model Tuned with RandomizedSearchCV outperforms the existing models with an RMSE of 1018.82 and an R^2 of 0.6181.
CONCLUSION: This research contributes to the field of sales forecasting in the retail industry and provides insights for businesses looking to enhance their revenue prediction capabilities.

Keywords: Big Mart, sales prediction, machine learning, prediction model, regression, linear regression, decision tree, random forest, XGBoost regression, K-nearest neighbours

Received on 02 11 2024, accepted on 30 05 2024, published on 28 06 2024

doi: 10.4108/eetiot.6453

1. Introduction

In the modern era, a large number of shopping centres, including supermarkets and retail shops, keep records of information pertaining to the sale of merchandise and items with various dependent or independent characteristics, qualities, and customer information, as well as asset-related information. In order to draw in more customers over a short period, each market attempts to provide customised and time-limited offers. As a result, using machine learning algorithms, the acquired data can also be utilised to forecast future sales [12]. Without an accurate sales prediction model, an organisation is confronted with a number of issues. Sales prediction, provides manufacturers, distributors, and retailers with valuable insights. Short-term forecasts aid in production planning and inventory management, while long-term predictions support corporate growth. Sales forecasting is crucial in sectors with products of limited shelf life to minimize revenue loss in shortage and surplus situations [10].

Predicting sales is essential for businesses engaged in retailing, shipping, manufacturing, marketing, and

[*]Corresponding author. Email: tp056587@mail.apu.emy.my

wholesale. It enables businesses to allocate resources effectively, predict sales revenue, and design strategies that are better for the future of the company [13]. One of the most significant assets a supermarket like Big Mart may have is that it may possess customer data acquired through interactions with other supermarkets. Within these data lie substantial patterns and variables that can be examined through a machine learning algorithm, enabling the precise prediction of sales with exceptional accuracy [10].

In order to make precise predictions about forthcoming events, a machine learning model is trained using data to identify recurring patterns. An effective forecasting model has the potential to greatly boost supermarket revenue, making it an asset for the company by increasing profits and offering valuable insights into customer service [1]. Therefore, it is crucial for Big Mart to have an accurate sales prediction system in order to increase the sales performance of the organisation and provide better customer service to the customers.

2. Literature Review

2.1. Sales Prediction

In today's highly competitive and ever-changing consumer landscape, the objective of every business is to generate profit [10]. This is the result of increased product sales and a high turnover rate. A precise and timely prediction of future revenue, also known as revenue forecasting or sales forecasting, can provide manufacturers, distributors, and retailers with significant insight. Long-term predictions can deal with corporate growth and decision-making, but short-term forecasts mostly aid in production planning and inventory management. Due to the limited shelf life of many of the commodities, which results in a loss of revenue in both shortage and surplus circumstances, sales forecasting is particularly crucial in the sectors. The ability of a manager to predict sales trends, determine optimal inventory replenishment times, and coordinate workforce scheduling is a crucial factor affecting a company's capacity to enhance its sales [10]. Excess orders lead to product shortages, whereas insufficient orders lead to missed opportunities. Consequently, competition within the food industry experiences frequent fluctuations driven by factors like pricing, advertising, and increasing customer demand [4].

Moreover, Vengatesan et al. [14] claimed that sales forecasting plays a significant role in numerous domains, such as monetary forecasting, electric power selection, resource estimation etc. Sales prediction is essential for separate associations, particularly vehicle contracts, land transactions, and other customary endeavours. Beheshti-kashi et al. [3] pointed out that the accuracy of sales forecasting can play a pivotal role in the prosperity and effectiveness of enterprises. Incorrect predictions can lead to either stock shortages or surplus inventory, leading to financial losses for the organisations. In sectors like retail, electronic markets, and the fashion industry, where customer-oriented businesses are prevalent, precise forecasts hold paramount importance. Accurate forecasting poses several obstacles for businesses. For instance, they must establish their production schedules before specific information regarding future demand is available [3]. There are numerous methods for sales forecasting, with corporations historically focusing on statistical models like linear regression and time series, feature engineering, and random forest models to anticipate future sales and demand. Patterns, seasonality, irregularity, and cyclicity are the most crucial factors to analyse [4].

Tom et al. [13] stated that sales forecasting is essential for businesses involved in retailing, shipping, manufacturing, advertising, and distribution. It enables businesses to allocate resources effectively, estimate total sales, and design strategies that are better for the company's future [13]. The data accumulated through customers' interactions across various supermarkets represents a supermarket's largest asset. Hidden within this data are substantial patterns and variables that can be effectively analysed using machine learning algorithms, leading to highly accurate sales predictions [7]. To precisely predict an upcoming event, a machine learning model undergoes training using data to grasp patterns essential for forecasting future occurrences. An effective forecasting model holds the potential to substantially boost supermarket revenue and is typically highly valued by the company, as it enhances profitability and provides insights into customer service [10].

2.2. Machine Learning

Machine learning (ML) is a subfield of Artificial Intelligence that focuses on training computer programs to improve their performance on specific tasks by learning from data and experiences. It has gained significant attention in the digital realm and has been applied in various domains such as healthcare, virtual assistants, robotics, natural language processing, data mining, and more. Machine learning algorithms make data-driven decisions and predictions based on the patterns and insights they learn from large datasets. However, in complex scenarios, noisy gradients can cause variations in the error margin rather than a consistent reduction [11]. According to Sav et al. [12], machine Learning is a collection of techniques that enable programs to make increasingly accurate predictions without explicit programming. The basic objective of machine learning is to construct algorithms and models competent of gathering data and predicting output using mathematical analysis while evaluating findings. AdaBoost Regressor, XGBoost Regressor, Linear Regression, and Random Forest Regressor are instances of Machine Learning approaches. Niu [9] also noted that machine learning techniques are widely applied to effectively tackle challenges in a variety of industries. Researchers have recently focused a great deal of emphasis on the implementation of predictive models based on machine learning to forecast sales volume, which aids businesses in developing more accurate sales forecasts and plans.

Batta [2] stated machine learning enables computers to manage data with greater efficiency. With the plethora of available datasets, ML is becoming increasingly popular with numerous businesses using it to retrieve pertinent data, its primary objective being to acquire knowledge from data: many research studies have examined how robots learn without being explicitly programmed [2]. Machine Learning employs data mining methods for the retrieval of information from massive databases. Both ML and Data Mining examine data from beginning to end in order to uncover hidden patterns within datasets and have been applied in a variety of industries, from computer networking, the tourism and travel industry, financial forecasts, the telecommunications industry to power demand forecasting[8].

3. Materials and Methods

3.1. Materials

The Big Mart Sales Prediction Web Application is developed in Python using the JupyterLab IDE. JupyterLab was chosen as the integrated development environment (IDE) for its unified interface, combining notebook, text editor, console and directory views. Its advanced functionality and popularity among data scientists make it an ideal choice. JupyterLab supports various file types, facilitating tasks such as inspecting raw sales data and enabling work on project reports. It also allows code execution from text files and easy replication of cells between notebooks. The compatibility of JupyterLab with Python, its extensive features, keyboard shortcuts, and free availability contribute to its selection as the preferred IDE.

The development took place on the developer’s personal computer, running Windows 10 and equipped with an Intel® Core™ i7-8750H CPU @ 2.20GHz 2.20 GHz and 8gb RAM.

3.2. Methodology

The development of the Big Mart Sales Prediction project adheres to the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology, a widely recognised and adopted approach in the field of data science. CRISP-DM is versatile and iterative, allowing developers to handle data science problems effectively. Its iterative nature enables flexibility and the ability to address issues and improve models through experiment, configuration changes and testing. Alterations can be made and earlier stages revisited to uncover hidden problems and ensure project success. The six stages of CRISP-DM (business understanding, data understanding, data preparation, modelling, evaluation, and deployment) help to lower the risk of failure and improve quality and productivity. Overall, CRISP-DM was considered the most appropriate approach for developing the Big Mart sales prediction web application.

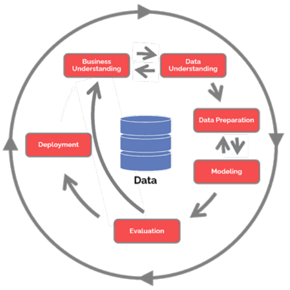

Figure 1: Stages of CRISP-DM Methodology [6]

The Big Mart Sales Prediction web application dataset was sourced from Kaggle [5], a renowned platform for data science and machine learning. This dataset is available in CSV format and comprises 8,523 rows, representing different products sold at multiple Big Mart locations. Each row in the dataset corresponds to a distinct product-store combination and includes 12 attributes providing detailed information.

Table 1: Metadata of Big Mart Sales Data

Attributes	Metadata
Item_Identifier	Unique product ID
Item_Weight	Weight of the product
Item_Fat_Content	Fat contact of the product (Low Fat / Regular)
Item_Visibility	Percentage of total display area for the product
Item_Type	Category of the product (e.g., Dairy, Breads, Meat)
Item_MRP	Maximum Retail Price (MRP) of the product
Outlet_Identifier	Unique store ID
Outlet_Establishment_Year	Year of the store was established
Outlet_Size	Size of the outlet (Small / Medium / High)
Outlet_Location_Type	Type of city where the outlet is located
Outlet_Type	Type of store (e.g., Grocery Store, Supermarket)
Item_Outlet_Sales	Sales of the product in the particular store

The researcher conducted several preprocessing steps to prepare the datasets for model training. Firstly, the inconsistent naming of item fat content categories was addressed, such as "Low Fat, Regular, LF, reg, and low fat," by renaming to "Low Fat" and "Regular" to ensure uniformity. To handle missing values, the missing values were imputed in the Item Weight column by using the mean weight for the corresponding Item Identifier. Similarly, missing values were imputed in the Outlet Size column with the mean weight for the corresponding Outlet Type. Additionally, 0 values in the Item Visibility column were replaced with the mean item visibility for each unique item identifier.

In order to derive meaningful features, a new variable was created named "Outlet Operation Years" by calculating the difference between the outlet establishment year and 2013. This variable provides insights into how long each outlet has been in operation. The "Item Category" variable was also designated, grouping items based on the first two letters of the item identifier, with categories such as 'FD' representing food, 'DR' representing drinks, and 'NC' representing non-consumables. Categorical variables, including the Outlet Identifier, underwent label encoding, assigning a numerical label to each unique category, facilitating further analysis. Additionally, specific categorical variables, including item fat content, outlet size, outlet location type, outlet type, item category and outlet, were one-hot encoded. This conversion allows categorical variables to be represented as binary values, making them suitable for modeling purposes.

To streamline the dataset and remove unnecessary variables, the developer removed the Outlet Establishment Year, Item Identifier, Item Type, and Outlet Identifier columns, as they were not utilized in the modeling process. Finally, the dataset was split into training and testing sets, with a test size parameter of 0.20, ensuring that model performance could be evaluated effectively.

After conducting an extensive literature review, the sales prediction models were developed following the completion of data preprocessing. The machine learning models employed for this task encompassed a range of techniques, including Linear Regression, Decision Tree, Random Forest, XGBoost, Stacked Ensemble Model, and K-Nearest Neighbors (KNN). To optimize their performance, some of the models underwent a tuning process. Evaluation of these models was carried out by analyzing their RMSE (Root Mean Square Error) and R-square scores. These metrics provide valuable insights into the accuracy and fit of the models. By assessing the RMSE, which measures the average prediction error, and the R-square, which indicates the proportion of variance explained by the models, their effectiveness in capturing and predicting sales patterns could be determined.

4. Results and Discussion

Table 2: Model Comparisons Table

Model	**RMSE**	**R-Square**
Linear Regression (without tuning)	1066.99	0.5811
Decision Tree (without tuning)	1469.97	0.2050
Random Forest (without tuning)	1077.68	0.5727
XGBoost (without tuning)	1104.28	0.5513
Stacked Ensemble Model of Random Forest, Gradient Boosting, and XGBoost Regressor Tuned with Random Search	1035.35	0.6056
XGBoost Regression Model Tuned with Random Search	1018.82	0.6181
K-Nearest Neighbours (KNN) Base Model	1154.95	0.5092
K-Nearest Neighbours (KNN) Tuned with Grid Search	1073.83	0.5757

The table provided in this context compares different machine learning models for sales prediction. Each row in the table represents a different model, and the columns show the model's Root Mean Squared Error (RMSE) and R-Square (R^2) values. These metrics are used to evaluate the performance of the models. RMSE measures the differences between the predicted values and the actual values. Lower RMSE values indicate that the model's predictions are closer to the actual values, indicating better accuracy. On the other hand, R^2 is a measure of how accurately the model's predictions align with the real data, with higher R^2 values indicating a better fit.

Observing the table, Linear Regression (without tuning) provides a reasonable baseline performance with an RMSE of 1066.99 and an R^2 of 0.5811. However, the Decision Tree (without tuning) performs less effectively than the Linear Regression model with an RMSE of 1469.97 and an R^2 of 0.2050. The Random Forest (without tuning) and XGBoost (without tuning) models have slightly worse performance than the Linear Regression model with an RMSE of 1077.68 and 1104.28, respectively, and R^2 of 0.5727 and 0.5513. The Stacked Ensemble Model, which combines Random Forest, Gradient Boosting, and XGBoost Regressor Tuned with Random Search, performs better than the previous models with an RMSE of 1035.35 and an R^2 of 0.6056. The XGBoost Regression Model Tuned with Random Search has the best performance, with an RMSE of 1018.82 and an R^2 of 0.6181. The K-Nearest Neighbours (KNN) Base Model has an RMSE of 1154.95 and an R^2 of 0.5092, which is worse than the Linear Regression model. The K-Nearest Neighbours (KNN) Tuned with Grid Search has an RMSE of 1073.83 and an R^2 of 0.5757, which is slightly better than the Random Forest model but worse than the Linear Regression model.

In summary, the XGBoost Regression Model Tuned with Random Search provided the best performance among the models presented in the table with the lowest RMSE and the highest R² value. Hence, this model is recommended for predicting sales based on its performance metrics.

5. Conclusions

The primary objective of this project is to utilise machine learning algorithms to assist companies in accurately predicting product sales, optimising resource allocation, and managing cash flow. The report of the investigation contains essential components such as objectives, deliverables, intended consumers, and a problem statement. A comprehensive literature review was conducted, focusing on sales prediction, machine learning algorithms, and existing systems in the market. Multiple research platforms were employed to obtain pertinent sources.

In the course of evaluating programming languages, IDEs, libraries, and operating systems, Python, JupyterLab, and libraries such as NumPy, Pandas, Matplotlib, and Streamlit were chosen. The chosen system development methodology is CRISP-DM, and all six phases are described in detail. The paper emphasises the importance of accurate sales predicting for Big Mart and discusses inventory instability, rising costs, and inadequate cash management. The developer overcame obstacles such as insufficient experience and time constraints through exhaustive research and effective time management. Moreover, it is anticipated that the practical execution of the undertaking will benefit from the gained insights. Overall, the developer has acquired extensive knowledge and practical skills in machine learning and sales prediction, empowering them to tackle future projects more effectively. Throughout the model implementation process, hands-on experience and the fine-tuning of algorithms enhanced technical knowledge and highlighted the practical applications of machine learning and sales forecasting. This experience, combined with a solid theoretical foundation, enabled the developer to overcome obstacles and achieve success in the development of effective, efficient, and scalable Big Mart Sales Prediction Web Application.

Acknowledgements.

First of all, I would like to express my sincere appreciation to Ms Minnu Helen Joseph, who provided the author with tremendous assistance and encouragement during the entire process of accomplishing the paper. Without her assistance and suggestions for improvement, finishing the paper would have been a tough and challenging task. Her commitment, participation and support aided the developer in accomplishing everything necessary, facilitating comprehensive research and the submission of the investigation report. Ms Minnu was very competent and skilled in machine learning and analytics and always capable of solving the concerns and inquiries of the researcher. The author genuinely acknowledges the exertion and insights shared by Ms Minnu, as well as Assoc. Prof. Dr. V. Sivakumar. Appreciation is also given to both the supervisor and second marker for providing valuable guidance and feedback throughout the journey and up to project completion.

In addition, the author is incredibly grateful to her family members for their unconditional assistance, care, and encouragement throughout this challenging time. On top of that, thanks are given to her course mates and friends for helping her uncover solutions to the issues encountered while conducting her investigation.

The author would also like to express gratitude to the Asia Pacific University of Technology and Innovation (APU) for providing the golden opportunity to carry out this project. This project would not have succeeded without the dedication and contributions of all involved parties. This investigation has assisted the developer in exploring more educated paths relating to the topic, which will undoubtedly be helpful within the future.

References

[1] Bajaj, P., Ray, R., Shedge, S., Vidhate, S., & Shardoor, N. (2020). Sales prediction using machine learning. International Research Journal of Engineering and Technology (IRJET), 7(6), 3619–3625. https://doi.org/10.1063/5.0078390

[2] Batta, M. (2018). Machine Learning Algorithms - A Review. International Journal of Science and Research (IJSR), 18(8), 381–386. https://doi.org/10.21275/ART20203995

[3] Beheshti-kashi, S., Karimi, H. R., Thoben, K., Lütjen, M., & Teucke, M. (2015). A survey on retail sales forecasting and prediction in fashion markets. Systems Science & Control Engineering: An Open Access Journal, 6, 37–41. https://doi.org/10.1080/21642583.2014.999389

[4] Boyapati, S. N., & Mummidi, R. (2020). Predicting sales using Machine Learning Techniques. May. https://www.diva-portal.org/smash/get/diva2:1455353/FULLTEXT02

[5] Brij. (2017). BigMart Sales Data. https://www.kaggle.com/datasets/brijbhushannanda1979/bigmart-sales-data

[6] Hotz, N. (2022). What is CRISP DM? - Data Science Process Alliance. Data Science Process Alliance. https://www.datascience-pm.com/crisp-dm-2/

[7] Malik, N., & Singh, K. (2020). Sales Prediction Model for Big Mart. Parichay: Maharaja Surajmal Institute Journal of Applied Research, 3(1), 22–32. https://www.researchgate.net/publication/344099746_SALES_PREDICTION_MODEL_FOR_BIG_MART

[8] Nagar, R., & Singh, Y. (2019). A literature survey on Machine Learning Algorithms. Journal of Emerging Technologies and Innovative Research (JETIR), 6(4), 471–474. https://doi.org/10.22214/ijraset.2021.37969

[9] Niu, Y. (2020). Walmart Sales Forecasting using XGBoost algorithm and Feature engineering. Proceedings - 2020 International Conference on Big Data and Artificial Intelligence and Software Engineering, ICBASE 2020,

458–461. https://doi.org/10.1109/ICBASE51474.2020.00103

[10] Odegua, R. (2020). Applied Machine Learning for Supermarket Sales Prediction. https://www.researchgate.net/publication/338681895

[11] Ray, S. (2019). A Quick Review of Machine Learning Algorithms. International Conference on Machine Learning, Big Data, Cloud and Parallel Computing, 35–39. https://ieeexplore.ieee.org/abstract/document/8862451

[12] Sav, R., Shinde, P., & Gaikwad, S. (2021). Big Mart Sales Prediction Using Machine Learning. International Journal of Creative Research Thoughts (IJCRT), 9(6), 674–678. https://ijcrt.org/papers/IJCRT2106802.pdf

[13] Tom, M., Raju, N., Isaac, A., James, J., & R, R. S. (2021). Supermarket Sales Prediction Using Regression. International Journal of Advanced Trends in Computer Science and Engineering, 10(2), 1153–1157. https://doi.org/10.30534/ijatcse/2021/951022021

[14] Vengatesan, K., Visuvanathan, E., Kumar, A., Yuvaraj, S., & Tanesh, P. S. (2020). An approach of sales prediction system of customers using data analytics techniques. Advances in Mathematics: Scientific Journal, 9(7), 5049–5056. https://doi.org/10.37418/amsj.9.7.70