IJNRD.ORG ISSN : 2456-4184
International Research Journal
IJNRD
Research Through Innovation
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Sales Prediction Using Machine Learning Techniques

Hitesh S.M[1], Yukthi A[2], Prof.Ramya B.N[3]

[1]Student,Department of AI & ML,Jyothy Institute of Technology,Bengaluru,Karnataka,India
[2] Student,Department of AI & ML,Jyothy Institute of Technology,Bengaluru,Karnataka,India
[3] Professor,,Department of AI & ML,Jyothy Institute of Technology,Bengaluru,Karnataka,India

Abstract - *An Intelligent Decision Analytical System necessitates the fusion of decision analysis and predictive methodologies. Within business frameworks, reliance on a knowledge base and the ability to predict sales trends holds paramount importance. The precision of sales forecasts profoundly influences business outcomes. Leveraging data mining techniques proves highly effective in unveiling concealed insights within vast datasets, thereby amplifying the accuracy and efficiency of forecasting. This study deeply examines and analyzes transparent predictive models aimed at refining future sales predictions. Conventional forecasting systems struggle with handling extensive data, often compromising the accuracy of sales forecasts. However, these challenges can be surmounted by employing diverse data mining techniques. The paper provides a succinct analysis of sales data and forecast methodologies, elaborating on various techniques and metrics crucial for accurate sales predictions. Through comprehensive performance evaluations, a well-suited predictive model is recommended for forecasting sales trends. The findings are encapsulated, emphasizing the reliability and precision of the adopted techniques for prediction and forecasting. The research identifies the Gradient Boost Algorithm as the optimal model, demonstrating superior accuracy in forecasting future sales trends*

Key Words: Data mining techniques, Machine Learning Algorithms, Prediction, Reliability, Sales forecasting

1.INTRODUCTION

This research aims primarily to develop a dependable mechanism for predicting sales trends using data mining techniques, ultimately optimizing revenue generation. Present-day businesses grapple with vast data repositories, a volume projected to exponentially expand. Consequently, imperative measures must accommodate transaction speed and anticipate the burgeoning data volume alongside evolving customer behaviors. Notably, the E-commerce sector seeks novel data mining methods and intelligent sales trend prediction models characterized by heightened accuracy and reliability.

Sales forecasting stands as a crucial element in guiding workforce management, cash flow, and resource allocation within companies. It forms the bedrock for enterprise planning and decision-making, empowering organizations to strategize effectively. Precise forecasts not only foster market growth but also elevate revenue generation. Leveraging data mining techniques to distil vast data into actionable insights is fundamental for cost prediction and sales forecasts, forming the cornerstone of sound budgeting

At the organizational level, accurate sales forecasts serve as pivotal inputs across various functional areas such as operations, marketing, sales, production, and finance. For businesses seeking investment capital, predictive sales data plays a pivotal role in efficiently leveraging internal resources. This study approaches the challenge with a new perspective, focusing on selecting the most suitable approach for highly precise sales forecasting.

The initial dataset analysed in this research contained a substantial volume of entries. However, the final dataset used for analysis was considerably smaller, achieved by eliminating non-usable data, redundant entries, and irrelevant sales data, ensuring a more refined analysis.

The research delves into data mining techniques and predictive methods in Section I, followed by a review of existing literature on sales forecasts in Section II. Section III outlines the objectives of sales prediction, while Section IV covers predictive analytics and methodologies concerning sales pricing, maintaining the essence while refining the content.

2.LITERATURE REVIEW

"Numerous sales prediction methods, totaling over 200, have emerged, categorized broadly into subjective and objective approaches (cheriyan, 2018). Subjective methods heavily rely on expert experiences, utilizing techniques such as the Delphi method (Linstone & Turof, 1975), brainstorming (Tremblay, Grosskopf, & Yang, 2010), and subjective probability (Hogarth, 1975). These techniques integrate expert opinions, offering flexibility but bearing a strong subjective element.

In contrast, objective prediction methods leverage raw data, employing mathematical and statistical models (sakib, 2019). This category includes regression analysis (such as simple and multivariate regression) and time series analysis (like moving average, exponential smoothing, seasonal trends, autoregressive-moving-average, and generalized autoregressive conditional heteroscedastic models) based on actual sales data.

Historically, conventional sales prediction methods introduced factors or time series for forecasting purposes. McElroy and Burmeister (1988) applied Arbitrage Pricing Theory in a multivariate regression model, while Lee and Fambro (1999) utilized the autoregressive-integrated-moving-average model for traffic volume forecasting. Huang and Shih (2003) forecasted short-term loans using ARMA, and Tay and Cao (2001) delved into time series forecasting. However, complexities in the relationship between influencing factors, time series data, and sales predictions often led to unsatisfactory results.

Consequently, recent focus has shifted toward intelligent models like artificial neural networks (ANN), support vector machines (SVM), and other cutting-edge approaches. Kuo and Xue (1998) proposed a sales prediction decision support system using fuzzy neural networks, while Hill, Marquez, and O'Connor (1994) reviewed artificial neural network models for forecasting and decision making. Cao (2003) combined SVM with time series for sales prediction, and Gao et al. (2014) advocated extreme learning machines. Yuan (2014) introduced an online user behavior-based data mining method for ecommerce sales prediction.

However, previous research primarily aimed at enhancing prediction accuracy via single model algorithm optimization or analyzing influencing factors. Limitations emerged in scenarios with zero sales volume, and most methods only forecasted for singular items rather than a broader product range.

To address these limitations, we devised a trigger model system instead of relying solely on a single algorithm. This system, grounded in data concerning sales-influencing factors, triggers one of the previously discussed prediction models. Consequently, it generates more accurate predictions and accommodates a significantly larger scale of sales prediction scenarios.

3. METHODOLOGY

The following picture Fig. 1 shows the sequence steps and the stages of the proposed prediction process. By utilizing these steps the big mart sales prediction model is built. In this flow diagram there are 5 major steps and each plays a significant role in building the model.

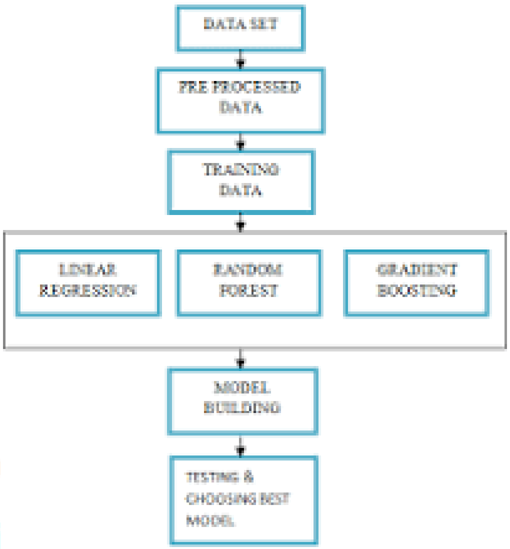


Fig. 1. Block diagram of Big mart sales Prediction

A. Data Gathering and Preparation

This study draws from a dataset sourced from an efashion store, covering three consecutive years of sales data. To forecast the efashion store's sales, historical sales records from 2015 to 2017 were compiled. The dataset includes diverse fields such as Category, City, Item Type and Description, Quantity, Quarter, Sales Revenue, Year, SKU Description, Week, and Year. Initially abundant, the dataset underwent refinement by eliminating unusable, redundant, and irrelevant entries, resulting in a significantly reduced final dataset [12].

B. Data Analysis

Stage B involves delving into Exploratory Data Analysis (EDA) and Preprocessing, pivotal for comprehending datasets and outlining their core attributes through visualizations. This phase enables a profound grasp of fundamental details essential for subsequent stages. The amalgamation of train and test data proves advantageous. Within EDA, a comprehensive Univariate and Bivariate analysis is conducted to formulate data hypotheses. During this process, observations might reveal synonymous categories such as "LF," "low fat," and "Low Fat," as well as correspondences like "reg" and "Regular," which need rectification due to repetition.

The scrutiny extends to unraveling relationships between bivariate features. Preprocessing emerges as the cornerstone of predictive analysis. The dataset may harbor undesired elements such as missing data or irregularities, demanding conversion into a structured format compatible with machine learning models. Statistical parameters—mean, median, mode, standard deviation, count of values, maximum values, etc.—are ascertained using the data.describe() function.

Pandas tools are employed to streamline data preprocessing, encompassing scrutiny of independent variables for null values in each column and their subsequent replacement with appropriate data types. This meticulous step rectifies repeated features, missing values, and extraneous columns, ensuring the dataset is primed for model training aimed at forecasting Outlet sales. The process involves addressing missing values by computing the mean and median for the respective features. Consequently, the dataset stands prepared to facilitate accurate model training, optimizing predictions for Outlet sales.

C. Model Building

After the cleaning processes of the data, now the dataset is ready to adapt for a model. Here the model is built using three algorithms.

-- Linear regression.
-- Random forest regression.
-- Gradient boosting.

To track the machine-learning system on wholesome basis we have used Scikit-Learn here. Algorithms for Predicting the dataset are discussed below.

1.Linear regression serves to establish a relationship between two variables by employing a linear equation fitted to observed data. This method aims to forecast the value of a dependent variable (y) based on a provided independent variable (x). Consequently, it unravels a linear connection between the input (x) and the output (y). The essence of linear regression lies in determining the optimal straight line that best encapsulates the given data points.

The formula for the linear regression equation is expressed as: $y=a+bx+\varepsilon$
Where:
y represents the predicted value.

x denotes the independent variable.

a signifies the Y-intercept of the line.

b represents the slope of the line.

ε accounts for the disparity between actual and predicted values.
This technique inherently acknowledges the presence of an irreducible error (ε), signifying the difference between actual and predicted values. Hence, complete reliance on predicted outcomes from the learning algorithm may be limited due to this inherent variability.

2.Random Forest Regression stands as a robust supervised learning algorithm adept at handling both classification and regression tasks. Operating on the principle of an ensemble method, it comprises an assembly of 'n' decision trees derived from different subsets of the dataset. The algorithm amalgamates these decision trees to harness their collective predictive power, enhancing the overall accuracy of predictions by aggregating their results, typically through averaging.

Its operation unfolds in two key phases:

Random Forest Creation: This initiates the amalgamation of 'N' decision trees, each trained on distinct subsets of the dataset. By harnessing multiple trees, the random forest maximizes the diversity of predictions.

Prediction Process: Once the forest of trees is established, the algorithm offers predictions by aggregating the outcomes from the various trees crafted in the initial phase. This ensemble approach culminates in a robust prediction by leveraging the collective insights derived from the multitude of decision trees.

A distinct advantage of Random Forest lies in its proficiency in handling extensive datasets characterized by numerous dimensions. Its capacity to navigate such complex data landscapes contributes significantly to its efficacy in prediction tasks.

3.Gradient Boosting Within gradient boosting, two primary types of base estimators are utilized: an average-type model and decision trees with full depth. The sequence of steps characterizing gradient boosting encompasses:

i) Creation of Average Model: An initial average model is crafted.

ii) Calculation of Residuals: Residuals are computed by contrasting actual values with predictions from the average model.

iii) Model Creation (RM1): A new model (RM1) is constructed, taking these residuals as the target.

iv) Prediction of New Residual Values: This model (RM1) predicts new residual values, subsequently leading to the calculation of updated predicted values.

v) Iteration with Residuals: The cycle continues with a recalculated set of residuals (Actual – Predicted), wherein a new model (RM2) is trained on these residuals as the target. This model then generates fresh predictions for the updated residuals.

This iterative process perpetuates, refining predictions by iteratively optimizing subsequent models based on the residuals derived from preceding iterations. This method enhances predictive accuracy by systematically refining the predictions through each iterative step.

4.RESULTS AND ANALYSIS

From the spectrum of algorithms employed, the selection of the most efficient model defines the subsequent output's accuracy. In essence, an algorithm showcasing a lower RMSE value tends to yield predictions of higher accuracy.

Among the trio of algorithms assessed, the Gradient Boosting algorithm emerges as the frontrunner, achieving the highest prediction accuracy of 0.69, representing the pinnacle among the selections. Moreover, its RMSE value, a mere 10.343, stands as the lowest among the algorithms scrutinized.

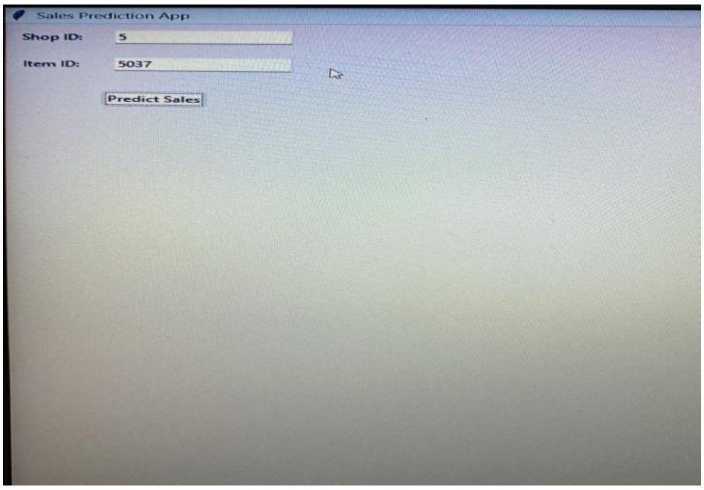
Where: Accuracy = Number of Correct Predictions/ Total Number of Predictions

RMS= sqrt(1-r2)SDy

The Gradient Boosting algorithm not only delivers the highest accuracy, as defined by correct predictions over the total, but also exhibits a minimal RMSE value, affirming its superior predictive capability within this study's context

ALGORITHM	ACCURACY	RMSE
Linear Regression	0.58	11.787
Random Forest Regression	0.649	10.821
Gradient boosting algorithm	0.69	10.343

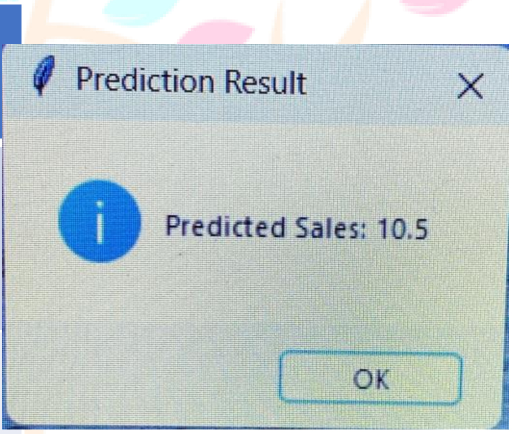
TABLE I
COMPARISON OF RESULTS OF THE ALGORITHM



Shop ID: 5

Item ID: 5037

Fig. 2 OUTPUT OF THE PROJECT



Predicted Sales: 10.5

Fig. 3 OUTPUT OF THE PROJECT

5. CONCLUSIONS

The evaluation of classification algorithms primarily revolves around key metrics such as Classification Accuracy, Accuracy per Class, and the Confusion Matrix, which illustrates the frequency of predictions for each class, allowing comparison against the instances of each class. Additionally, metrics like Root Mean Square Error, Mean Square Error, and Absolute Error are computed, culminating in an Error Rate displayed in Table III. This metric aids in identifying the average incorrectness of predictions.

The comparative analysis of the three algorithms, depicted in Table 1 and visualized in the accompanying figure, distinctly indicates the performance disparities. Notably, the Gradient Boost Algorithm showcased an impressive 98% overall accuracy, followed by the Decision Tree Algorithm achieving nearly 71% overall accuracy. The Generalized Linear Model trailed with a 64% accuracy rate. Ultimately, upon empirical evaluation, the Gradient Boosted Tree emerges as the most fitting model.

While classification accuracy rates can theoretically attain 100%, the empirical analysis of the GBT model, achieved approximately 98% accuracy. This corroborates its exceptional performance and reliability.

6. References

[1] cheriyan, S. (2018). sales prediction using ml techniques. *IEEE*, 10.

[2] fng, y. (2022). sales prediction analysis. *science gate*, 7.

[3] ibahim, s. (2018). intelligent techniques of ml in sales prediction. *semantic scholar*, 6.

[4] sakib. (2019). ML predictive analysis. *engrxiv*, 8.

[5] varshini, d. p. (2021). analysis of ml algorithms to predict sales. *ijsr*, 6.