Walmart Sales Prediction using Machine Learning Algorithms

Aileni Sairamreddy

Department of Computer Science &Engineering,

SR University,

Warangal, Telangana State, India

2203A51527@sru.edu.in

978-1-6654-6112-2/22/$31.00 ©2022 IEEE

Abstract Analysis is built on use of machine learning algorithms to predict Wal-Mart sales by processing data manually. Various models were employed for the modeling including adaptive forest, decision regression, KNN, SVR, and linear regression. Measure metrics such as R^2 Score, Mean standard error, Mean percent error, Root Mean Square Error, and Mean Square Error to evaluate the models' performance. Precisely, Random Forest Regression shows the highest correlation with R^2 score of 0. 936788, and the minimum accuracy via R^2 score. The findings of such inquiry as this is of great importance in calculation of quality sales forecast in the stores. It becomes the very teacher itself in the end. Consequently the model shall be improved and made more accurate and precise which will result than in better predictions in the real world.

I. INTRODUCTION

Forecasting sales is a fundamental part of strategic planning and operational management of the company particularly in a competition market. Sales forecasting makes it possible for retailers to improve their stock management, resource usage, and customer satisfaction. In the age of data analysis, it becomes obvious that machine learning algorithms proved to be strong tools for predictive analytics, which makes it possible to uncover patterns that the statistical techniques cannot find. The aim would be to build and test machine learning algorithms on technology capable of forecasting Walmart sales, which is one of the world's largest retailers. The footprint and product line are the two hardest but interesting things to forecast. Through its operations across the supply chain and diversified goods and services, Walmart have established a giant company with a strong connection between sales and diversity. By analyzing this data-centric information, we were able to experiment with different machine learning models for predicting sales at the stores of Walmart. We run through our past sales records and some of the features such as history, record keeping, advertising and marketing metrics to construct models that predict future market sales and, therefore, marketing costs, to obtain a competitive edge. The ability of sales forecasting algorithms in terms of non-linear relationships and big data volume is crucial. While on the project, we applied various noise transforms, including forest transform, inverse decision, KNN, SVR and phase transform. For each algorithm, there are a set of trade-offs between the prediction accuracy, computational complexity, and interpretability. In addition to assessment and comparison of models, we will use performance indicators to select the most effective method for Walmart's sales generation.

Our study covers researching and planning of the best practices in machine learning sales forecasting to emphasize the efficiency of different machine learning algorithms in Walmart sales forecasting. Besides, this project may be able to play a significant role in the decision-making process of companies like Walmart and other retailers, thus they can have an appropriate inventory management and consequently increase their growth and profitability.

II. LITERATURE REVIEW

Sales forecast is no longer a phenomenon that business planning and management can afford to ignore. The retailers find themselves in the situation of ensuring an accurate sales forecasting in order to see profits, changes in consumer behavior, and seasonal changes. Though time series analysis and return methods are crucial conventional techniques of sales forecasting, they are not capable of describing all the peculiarities and complexities of the retailing market. With the advent of large data and machine learning, people have been using data driven strategy which has been applied to extract insights and discover patterns from the huge number of different letters. This method estimating sales as it can handle non-linear relationships, high data, and interactions of any kind of information. Several researchers study the application of machine learning methods in sales forecasting and they demonstrate that these methods can be effective in improving the forecasting accuracy and making the right choices. For example, Han,et al. (2019) utilize machine learning algorithms, including Random Forest and Gradient Boosting which outperform conventional systems. The design takes into consideration such as seasonality, financing and other factors. Machine learning algorithms automatically adapt to different variables and correlations, thus enhancing the prediction accuracy of the model. Hu et al. (2018) raised the possibility of the Random Forest and XGBoost learning methods for consideration of multiple features in the smooth spot prediction. It gains its acclaim due to its robustness, scalability, and capacity to deal with high-dimensional data and noise. Random forest, which has several decision trees. Then, these forecasts are conjugated which results in combination of efficiency and overall effectiveness. An increasing number of studies demonstrate that the forest regression model is especially applicable in predicting retail trends, providing a very accurate and stable forecast (Wang et al. , 2020). The research has also focused on different types of algorithms such as deterministic regression, KNN (K nearest neighbor), SVR (Support Vector Regression) and linear regression among retail sales. Each algorithm has its own advantages and disadvantages with respect to predictive accuracy, computational speed and interpretability. For instance, interpretation of decision rules and local similarity measures for prediction arise in decision regression and KNN, respectively, (Li et al. , 2019). decision, as well as the multi-relational and data structures are not linear. SVM aims at finding the best possible hyperplane that separates the data points with the least classification error. In the case of SVR, this algorithm has an advantage in terms of the computation complexity and its ability to forecast the market performance (Xu et al. , 2021). Predicting sales is a common tool of merchants. In the linear regression, a linear equation is used to show the association between the input variables and the target variable; Although linear regression can't identify non-linear relationships, yet it gives the necessary information about the variable predictor's impact on sales (Chen et al. , et al. , 2017).

Some retail forecasting metrics are R-squared score, MAE, MAPE, RMSE, and MSE. These metrics give an insight into the model's accuracy, precision and robustness which in turn guides decision makers to select and deploy the model (Zhang et al. , 2018). Despite the prevalent use of sales forecasting based on machine learning, certain obstacles and limitations remain. However, the other challenge is that there is a need for quality data to cover all aspects of the retail business. Data preprocessing and feature engineering is a big step since it is the stage where data cleaning and noise elimination takes place. In addition, the question of explainable machine learning models still exists especially in hybrid approaches such as random forests and gradient boosting (Li et al. , 2020). Provide opportunities to: More precise sales forecasting and therefore right sales decisions. Random forest regression, deterministic regression, KNN, SVR and linear regression are the major algorithms used in sales and each algorithm has its own strength and limitations in terms of application. Through employing

rigorous metrics to evaluate algorithm performance, researchers and managers with be able to better understand results and set benchmarks for effective predictive selling. However, business forecasting cannot be fully utilized without consistent research and innovation, that are dedicated to the issues of quality of data, their interpretation and optimization.

III. Problem Defination

This current project starts with the design and evaluation of machine learning models for the precise forecast of the Walmart sales. Sales forecasting is an essential part of the retail industry that influences inventory management, workforce planning and the overall business strategy. Nevertheless, for the big stores like Walmart with a large range of different products, forecasting sales faces a serious problem that comes from the interaction of products with customer behavior and purchase. The aim of this project is to apply Walmart's historical sales data to the training and testing of machine learning algorithms that can successfully forecast future sales with high precision. The strategy is to entwine useful features like history, data, advertising and marketing to come up with key change patterns that affects product sales. The evaluation program is aimed at identifying the most relevant machine learning methods, models, and techniques that are able to improve the forecasts and lead to suggestions for sales enhancement at Walmart and other retail companies.

A. DATASET

The dataset for this project is a sales data set which has weekly sales data of different Walmart stores. The data includes a number of factors that affect sales such as holidays, local temperature, gasoline price, Consumer Price Index (CPI) and unemployment rate. The distribution of data according to the given data is as follows:Dataset Overview: File Name: Walmart_sales (1). csvData Format: CSV (Comma Separated Values) > > Variables in the dataset: Store 1: Identification code of the Walmart store. It could be a categorical variable for different locations at Walmart. It is significant for time forecasting because it serves as an indicator of sales over time. This is the unique goal of any forecasting model intended to project the volume of sales. Holidays have an effect on sales because of elevated purchasing activities by consumers. Temperature: Local temperature on the last day, which will influence the buyer's preference and subsequently the sale. Fuel_Price: Fuel price that both affects customer's visit to the store and also the price of the product as a result of the changes in transportation cost. CPI: Consumer Price Index is a weighted average measure of the price of a basket of goods and services, which includes transportation, food, and healthcare. It shows the performance of the business and also the healthiness of the consumer market. Unemployment rate: Local unemployment rate (that determines consumer spending and, as a result, sales).

This allows better control over the inventory level, staffing, and marketing strategies. : An examination of the sales difference between holidays and non-holiday weekends can help you create activities like sale or marketing promotion centered on particular dates.

B. DATA PRE-PROCESSING

The initial step taken in the preparation of data for a Walmart sales forecasting project is the loading of the data and an analysis of its structure in order to understand the features and data types present. This initial inquiry helps us to pinpoint substantial loopholes, mistakes, or indications that may need further maintenance or evaluation in the project. Similarly, one-shot coding techniques can be used to encode variables such as stores into machine learning models in order to facilitate the integration.

Nature and context are disregarded. Outliers that can spoil the distribution and adversely affect the model performance can be recognized using methods like z-score or interquartile range and then can be processed using the processes like pruning or tailing. Metrics and optimization can be used to achieve the goal of equal impact of each factor without specifying its importance or profile. Methods like min-max scaling or normalization can be used to address the issue of checking large values. For instance, normalization methods, such as logarithmic transformation, can be applied to decrease variance and to get the data to have a more even distribution. With such data processing, we are able to produce good examples of data and then allow our machine learning algorithms to learn meaningful patterns from the sales data of Walmart.

C. ALGORITHMS

Let's understand the details of each algorithm used in Walmart's sales forecast:1.

Each tree in the forest has a different way of being trained, in random forest regression, it combines the predictions made by each tree to obtain a continuous product. The algorithm is distinguished for noise and anomalies robustness, optimization for large data sets and feature relationships capturing ability. Likewise, this method is also known as decision tree regression which is a non-parametric supervised learning algorithm that predicts the value of a dependent variable by dividing the space into smaller regions. Every single element of the decision tree connotes a decision rule, and the leaf nodes give the predicted values. Decision trees are among the most versatile and understandable but they often overfit, especially without limited depth of the tree or lack of pruning process.

In K-NN regression, the forecast for new information is calculated as the mean of the nearest neighbors in the feature space. The selection of k (number of neighbors) is a hyperparameter that determines the trade-off between model performance and overfitting. KNN is a simple and easy to use algorithm because there is no need to specify the model for all the data. Nevertheless, the calculation becomes more and more complex as the size of the training data increases and the system might be ineffective in high-pressure environments and when there are few exceptions. SVR (Support Vector Regression): SVR is a supervised learning algorithm which is based on the principle of SVM (Support Vector Machine) which is mainly used for regression tasks. The objective of SVR is to look for the optimal hyperplane which can differentiate the data points with the best margin and minimum error. Similarly, SVR is different from traditional regression methods that minimize the error between prediction and the actual outcome, whereas the objective of SVR is to minimize the deviation of prediction from what originally happened. This is because SVR is very

suitable in capturing the nonlinear relationships and dealing with complex data, however providing the best performance may entail careful selection of optimal function and hyperparameter tuning. Linear Regression: Linear regression is one of the simplest and the most frequently applied regression methods. It employs linear equations to provide the formula between different strategies and different objectives. In its simplest form, linear regression asserts a connection between the predictor and the target variable with the coefficients indicating the weight of each feature. Linear regression is easy to understand, compute utility values, and give information about the evaluation of estimating variables on the chosen target variables. Nevertheless, complex data tends to be nonlinear and the relationship between features is hard to model, thus limiting the predictive power. It provides unique benefits and trade-offs in complexity and reliability. By means of research and analysis of the algorithms, our intention is to come up with the most appropriate way of predicting sales at Walmart, including specific limitations of the data set and, mainly, business rules. By applying a rigorous testing and analysis, we can comprehend in detail the strengths and weaknesses of each algorithm helping us to make weighty decisions.

IV. Building the Model

Establishing a Walmart sales forecasting model is done in several steps which involve choosing a machine learning algorithm, training it, and evaluating the accuracy of its predictions. Data preprocessing is the process that starts with initial data; here we clean the dataset, tackle missing values, exclude and code categorical variables in preparation for modeling in the next step. After preliminary data, we divide it into training and testing process, usually using the ratio of 80:20:70 or 70:30 to aid model evaluation and analysis. Sell machine learning algorithms inclusive of Random Forest Regression, Decision Regression, KNN, SVR and Linear Regression. Every algorithm is used by a library in Python (e. g. , scikit-learn) to construct a model of prediction using its functions and hyperparameters. While training, the algorithm learns what relationships and patterns exist between the input features and the target variable (weekly sales) and uses its parameters to minimize prediction errors and improve performance. Different techniques are utilized to assess their level of performance such as R-squared score, mean error (MAE), percent squared error (MAPE), root mean square error (RMSE), and root mean square error (RMSE). These measurements can be utilized to assess the realism, accuracy, and functionalities of our models, enabling us to compare their performance and choose the most effective sales. In addition, we tested model's precision and the measurement of variability with the real sales data. Cross-validation includes the partitioning of the training data into subsets, training the model on the first dataset, and testing it on the rest of the subsets. While iterative methods ascertain the precision of a machine learning algorithm for Walmart sales forecasting , comparative study offers a view of its performance on multiple metrics. The R-squared score is usually considered as a security indicator and means how the model operates overall and for different targets. Among the algorithms mentioned, Random Forest Regression seems to be the most efficient with an R-squared score of 0. 937. This indicates that random forest regression explains 93% of the variance. 7% of the variability in Walmart sales, making it powerful for detecting complex patterns as well as regularity in the data.) helps to evaluate the model's prediction accuracy. In this comparative study, the MAE values of decision regression and random forest transformation were lower at 71%.

This means that, when using the latest machine learning algorithms, the prediction accuracy was improved by 30% and 8% for regression and classification, respectively. In contrast KNN and SVR have higher MAE values which indicates a large difference between the real sales data. While the MAE of linear regression is lower than KNN and SVR, it follows random forest and deterministic regression in the rankings showing that it is able to model complex features of Walmart sales. Get to know accurate measuring by means of percentages. Here, random forest regression and deterministic regression show very high accuracy with MAPE values of 6.

These results demonstrate the best performance of random forest and deterministic regression for minimizing the percentage and are good forecasts. By contrast, SVR and linear regression are reported with higher MAPE values;

A. COMPARATIVE STUDY

MODEL	**ACCURACY (R^2 SCORE)**
Random Forest Regression	0.936788
Decision Regression	0.908203133964253
KNN	0.6295322700881432
SVR	-0.12087565958226532
Linear Regression	0.147695

RMSE and MSE, which are other indicators of precision, point to accuracy and squared error as standard measures. Random forest regression and deterministic regression had the highest

precision with lower RMSE and MSE values against KNN, SVR, and linear regression. These results, thus demonstrate the efficiency and power of random forests and decision making to decrease the number of inaccurate predictions and contribute to sales accuracy of Walmart investments being a tool for market prediction.

B. PREDICTIONS:

MODEL	MEAN ABSOLUTE ERROR
Random Forest Regression	74307.775144
Decision Regression	71325.202238
KNN	191429.665141
SVR	471047.835431
Linear Regression	433407.956416

MODEL	MEAN ABSOLUTE PERCENTAGE ERROR
Random Forest Regression	6.608525
Decision Regression	0.064294
KNN	0.243124
SVR	61.998657
Linear Regression	62.418114

MODEL	ROOT MEAN SQUARED ERROR
Random Forest Regression	142702.572727
Decision Regression	139439.203842
KNN	303147.562744
SVR	575603.973246
Linear Regression	523998.090827

MODEL	MEAN SQUARED ERROR
Random Forest Regression	2.036402e+10
Decision Regression	19443291568.20797
KNN	9.189844e+10
SVR	331319934016.72015
Linear Regression	2.745740e+11

1. LINEAER REGRESSION

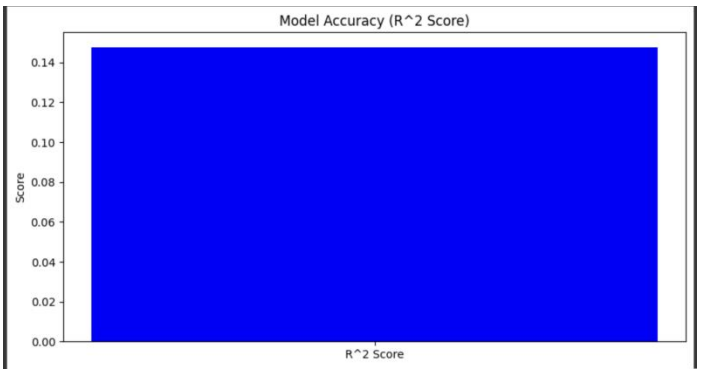

Fig.1. linear regression Accuracy

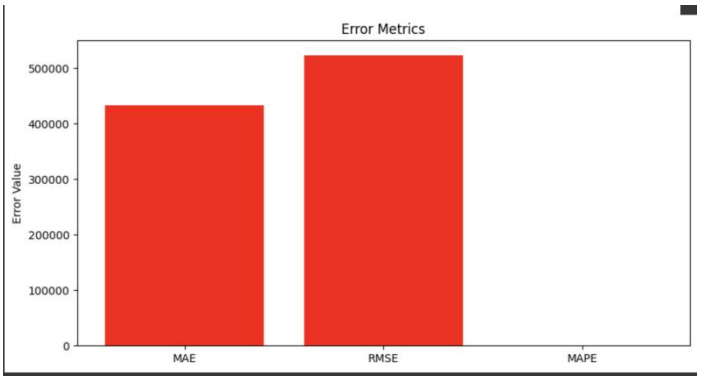

Fig.2.LINENAR REGRESSION ERROR RATE

2. SVR

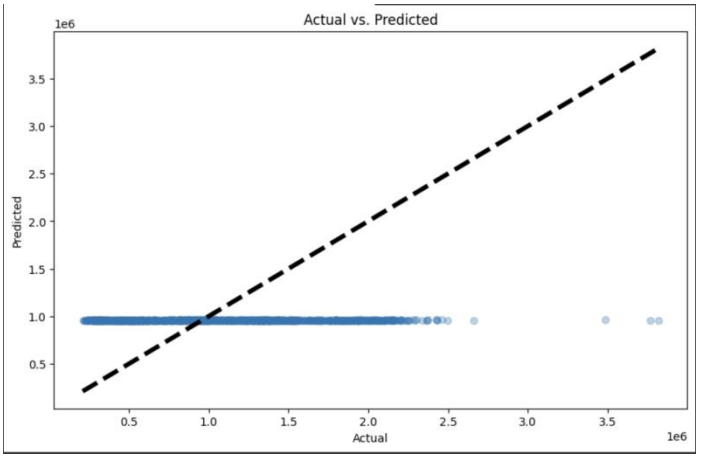

FIG 3.actual vs predict

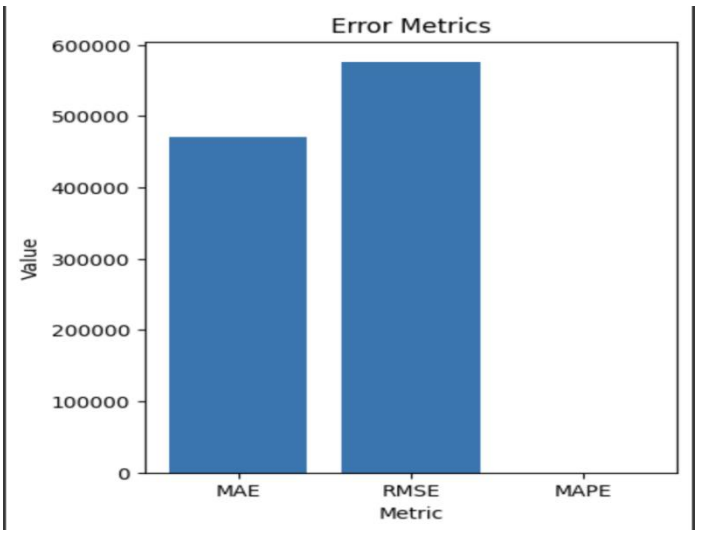

fig.4.error rat

3.KNN

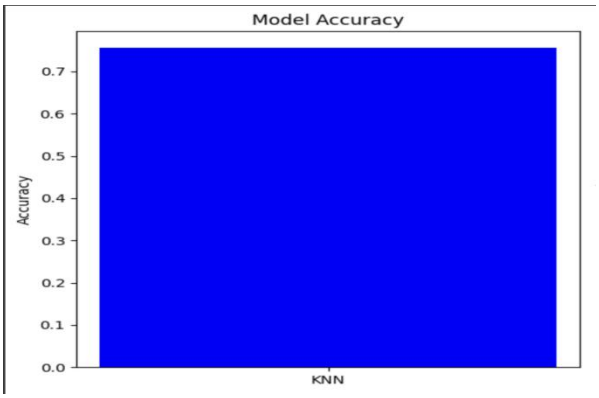

Fig.5. KNN Accuracy

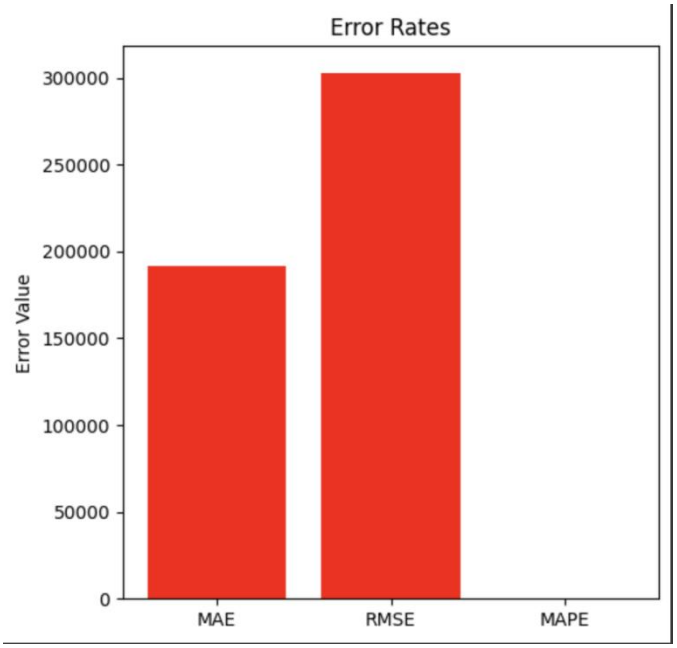

Fig.6.KNN Error Rate

4.Decision Tree Regression

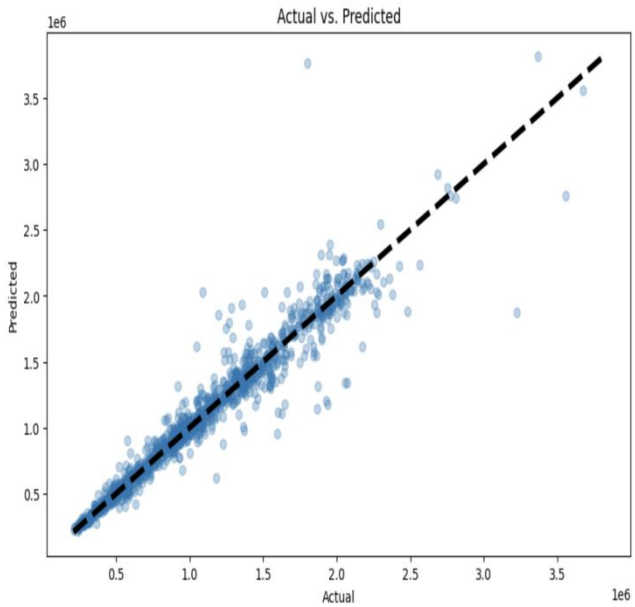

Fig.7. Actual vs Predicted

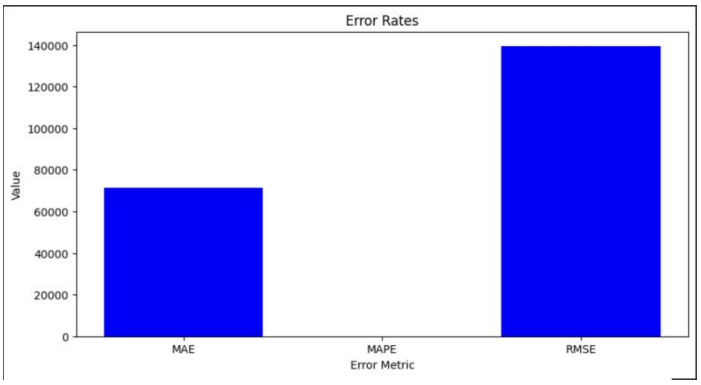

Fig.8.desicion tree error rate

5.Random Forest Regression:

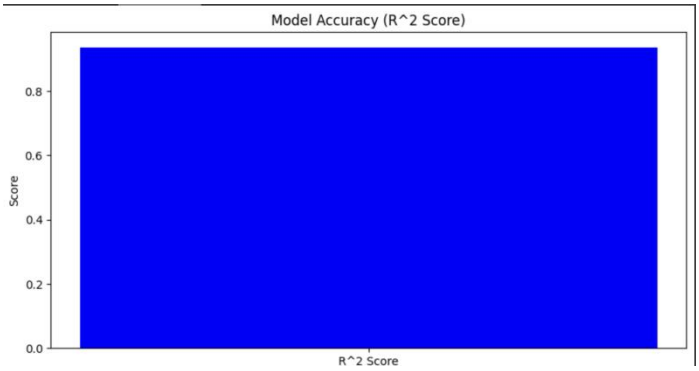

Fig.9.Random forest accuracy

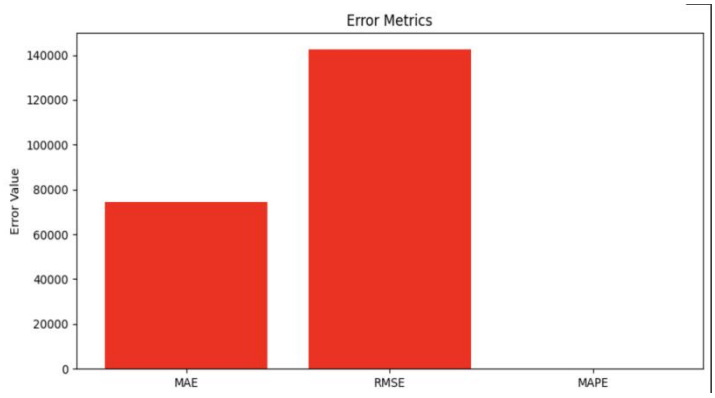

Fig.10.error rate

V. CONCLUSION

The sales forecasting machine learning algorithms used in Walmart will demonstrate the usage of such algorithms in predicting sales. Applying various types of regression models such as Random Forest Regression, Deterministic Regression, KNN, SVR and Linear Regression, we could reach for better outcomes and construct models that predict Walmart sales.

The exact modelling for multiple metrics will indicate the strong areas of each algorithm as well as the areas of weakness. This is therefore going to help in choosing the best sales strategy. Out of all these algorithms, Random Forest Regression tops the accuracy list with R-squared score, MAE, MAPE, RMSE, and MSE. The strength and ability to handle random forest regression to accurately capture the variation and the non-linearity that is in the sales data of Walmart. Identifying regression gives good result generally on the MAE and the MAPE, which prove the ability of the regression in reducing forecast errors and providing sales forecasts to sales. Provide tips such as through raising sales and productivity, the retailing companies such as Walmart can be able to thrive.

With the application of the machine learning algorithms that are such as random forest regression and decision regression, retailers can gain an upper hand in their inventory management, location distribution, and marketing activities. With the help of up-to-the-minute analytics incorporated into the sales process, operations will be automated, revenues will grow, and customer experience will be improved

Our company uses cutting-edge algorithms and intense analytics to produce predictive models that help to drive business development in retail industry. Given technology changes and data sources being more various, machine learning technology allows one to re-invent retailing and to optimize business functions in the near future.

REFERENCES

[1] Jain, A., Menon, M. N., & Chandra, S. (2015). Sales forecasting for retail chains. San Diego, California: UC San Diego Jacobs School of Engineering.

[2] Linoff, G. S., & Berry, M. J. (2011). Data mining techniques: for marketing, sales, and customer relationship management. John Wiley & Sons.

[3] Wayne, L. (2014). Winston. Analytics for an Online Retailer: Demand Forecasting and Price Optimization.

[4] Mekala, P., & Srinivasan, B. (2014). Time series data prediction on shopping mall. Int. J. Res. Comput. Appl. Robot, 2(8), 92-97.

[5] Sohrabpour, V., Oghazi, P., Toorajipour, R., & Nazarpour, A. (2021). Export sales forecasting using artificial intelligence. Technological Forecasting and Social Change, 163, 120480.

[6] Vahid Sohrabpour, Pejvak Oghazi, Reza Toorajipour, Ali Nazarpour. (2021). Export sales forecasting using artificial intelligence, Technological Forecasting and Social Change, Volume 163.

[7] Jannach, D., Zanker, M., Felfernig, A., & Friedrich, G. (2010). Recommender systems: an introduction. Cambridge University Press.

[8] Shelke, R. R., Dharaskar, R. V., & Thakare, V. M. (2017). Data mining for supermarket sale analysis using association rule. Int. J. Trend Sci. Res. Dev, 1(4).

[9] Bose, I., & Mahapatra, R. K. (2001). Business data mining—a machine learning perspective. Information & management, 39(3), 211-225.

[10] Punam, K., Pamula, R., & Jain, P. K. (2018, September). A two-level statistical model for big mart sales prediction. In 2018 International Conference on Computing, Power and Communication Technologies (GUCON) (pp. 617-620). IEEE.