

An Experimental Analysis of Machine Learning Model in the Context of Sales Forecasting: -A Walmart Dataset as A Case Study

Disha Wankhede¹, Abhishek Athanikar², Yash Borude³, Raviraj Chougule⁴ and Om Diware⁵

¹⁻⁵Department of Computer Engineering, Vishwakarma Institute of Information Technology

Email: disha.wankhede@viit.ac.in, {athanikarabhishek, yash.borude777, diwareom, atharvavalsange}@gmail.com

Abstract— Software programmers may get gradually precise at expecting consequences without being clearly coded using machine learning techniques. Machine learning is based on the idea that models and algorithms may collect input data then utilize statistical investigation to determine an output, while updating results as information become obtainable, as underlying principle. For example, models may be adapted and trained to meet management expectations so that correct measures are followed to reach a certain goal. Wall Mart, a one-stop shopping mall, has been used in this system to estimate sales of various products and to study impact that various variables have on sales of products. Predictive models may be built using different features of a Wall Mart dataset and methods used to construct them, and these findings can be used to make better business choices. This Predictive Algorithm is used in many fields over the market world for their profitability. Recently, Lot of companies are going to used this Machine Learning for better AI and new technologies in market. This report Includes combination of machine learning algorithm and Features Engineering to predict Walmart sales. The research topic of this paper primarily solves the problems of Walmart sales forecasting that is provided by the Kaggle competition.

Index Terms— Wall mart, Machine Learning, sales, predict, Feature engineering

I. INTRODUCTION

Recording information related to sales of things or products with their many dependent or independent aspects is an important phase in today's contemporary world of large shopping centers, such as major malls and marts. Dataset is a combination of element object and integral, float datatype. All constructed using numerous dependent & independent factors. After then, data is cleaned up to provide more precise forecasts and fresh, fascinating findings that add to our accepting of task's information. Machine learning methods like random forests as well as simple or multiple linear regression models may then be used to expect upcoming trades using this figures.

Based on the studies mentioned above, it is shown that there is a demand for making relevant forecasts in various fields. Moreover, machine learning methods are highly used in predicting models. The research topic of this paper primarily solves the problems of Walmart sales forecasting that is provided by the Kaggle competition. The rest of this paper is arranged as below: first, the methodology used in this paper will be introduced. In this section, feature engineering and machine learning method are detailly represented. Next, the procedures for the experiment

are shown, including the comparison of different models with the same metric. Finally, the conclusion and future work are represented.

II. LITERATURE SURVEY

In Past Years, there is lot of use of data innovation in large scale .The information where the conventional approach is no more compelling. Author [1]Chu, C. W., et al. are really worked for their paper which is study of retail sales forecasting published in 2003.They described research paper in their own way .Literature survey of there paper shows importance of machine learning algorithm used in that paper. [2] Author Zhao, J., et al . made paper on novel electricity sales forecasting method based on clustering regression and time-series analysis in 2015.This paper has more regression quality with different techniques. [4] Author D Suma, et.al. has incredible work on Data Mining based Prediction of Demand in Indian Market for Refurbished Electronics which is published in 2020. [6] Author Giuseppe Nunnari, et al. on Forecasting Monthly Sales Retail Time Series: A Case Study published in 2017.We have taken Weekly sales Prediction table from this paper. [7] Author Zone-Ching Lin,et.al. made paper on Multiple Linear Regression Analysis of the Overlay Accuracy Model Zone published in 1999. [8]Author O. Ajao Isaac, et.al. worked on Polynomial Regression Model of Making Cost Prediction In Mixed Cost Analysis and published in 2012. [9] Author C. Saunders, et al. worked on Ridge Regression Learning Algorithm in Dual Variables published in 1998. [11]Author Cheriyan,et.al. worked on Intelligent Sales Prediction Using Machine Learning Techniques and published in 2018. [12]Author Fawcett, et al. made paper on Combining Data Mining and Machine Learning for Effective User Profiling in 1996 .[16] Author Panjwani,et.al. really done their project hardly on Sales Prediction System Using Machine Learning in 2020.[15],[20]. Author Kulkarni,et.al. made project based on Random forest classifiers published in 2013. [16][17][18] Author had done the work in healthcare system using ML DL Likewise We created our paper similar to these other authors' papers. We have taken reference of these Papers. We are going to explain algorithm that used in our model. Some algorithm need to use libraries and packages for data cleaning and visualization. An algorithmic approach is used to collect, analyze and validate data. After data has been analysed, it may be utilized to introduce new products or conduct more market research. This paper mainly uses four steps to do with the feature, preprocessing and cleaning for data, data normalization and visualization, statistical features and heatmap, and feature selection. First, do some appropriate processing for the data. Then, extra the time characteristics from the timestamp, for example, the day of year and month. The second step is normalizing the data. By limiting the pre-processed data to a specific range, such as [0,1] or [-1,1], normalization aims to eliminate the undesired effects brought on by odd sample data. Next, statistical features in a certain period, like mean, minimum, maximum, standard deviation, etc. Meanwhile, visualizing the correlation between features by drawing a heatmap. The heatmap represents the correlation coefficient among those features, the numerical value of the correlation of the coefficient will be between -1 to + 1 [5]. The strength of the linear correlation between the two variables depends on how close the absolute value of the coefficient is to 1; conversely, the weaker the linear correlation between the two variables depends on how close the absolute value of the coefficient is to 0. ELI5 is a Python library that aids in explaining and debugging machine learning classifier predictions. This is classifier algorithm used in our study to debug and predict/explaining decision tress concept.

III. DATASET AND PROPOSED METHODOLOGY

The dataset for our study can getting from Kaggle website hosted by Walmart, this dataset contains large entries which is unstructured data, this can be found here: <https://www.kaggle.com/datasets/aslanahmedov/walmart-sales-forecast>

Along with department information for the specific stores, it includes details about 45 Walmart locations. The examination. The csv data file is exclusively utilized for model value prediction. The dataset does not contain a target variable. Rather, 'train' is the training dataset. For this study, the CSV file is divided into training and validation datasets. This study's primary objective is to forecast the department's weekly sales for the entire year. The training and testing datasets have previously been separated from the dataset; the testing dataset is the same as the training dataset except for the weekly sales data. Weekly sales data for the stores and departments from 2010-02-05 to 2012-11-01 is included in the training dataset. There is one column i.e. Is holiday which shows target variable. In total, there are 4,21,570 rows in the training dataset and 1,15,064 rows in the testing dataset.

There is another dataset called 'stores.csv' contains information about the type and size. One more big challenge of this study is to check whether there is an increase in the weekly store sales because of changes in temperature, fuel prices, holidays, markdowns, unemployment rate, and fluctuations in consumer price indexes, The file 'features.csv' contains all detailed information about these factors .

TABLE I. DATASET

Date	Weekly_Sales	Size	Temperature	Fuel_Price	CPI	Unemployment	Total_MarkDown
2010-02-05	2492.450	151315	42.31	2.572	211.096358	8.106	0.0
2010-02-05	668.48	125833	38.01	2.572	214.655459	6.415	0.0
2010-02-05	693.87	125833	38.01	2.572	214.655459	6.415	0.0
2010-02-05	8654.60	155078	34.14	2.572	214.655459	6.299	0.0
2010-02-05	1112.356	125833	38.01	2.572	214.655459	6.415	0.0

TABLE II. STORE DATASET

Index	Store	Type	Size
0	1	A	151315
1	2	A	202307
2	3	B	37392
3	4	A	205863
4	5	B	34875

A summary of the features dataset is displayed in the image below. (Figure 1)

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 421570 entries, 0 to 421569
Data columns (total 17 columns):
#   Column              Non-Null Count  Dtype
---  -
0   Store                421570 non-null  int64
1   Dept                 421570 non-null  int64
2   Date                 421570 non-null  object
3   Weekly_Sales         421570 non-null  float64
4   IsHoliday_x          421570 non-null  bool
5   Type                 421570 non-null  object
6   Size                 421570 non-null  int64
7   Temperature          421570 non-null  float64
8   Fuel_Price           421570 non-null  float64
9   Markdown1            421570 non-null  float64
10  Markdown2            421570 non-null  float64
11  Markdown3            421570 non-null  float64
12  Markdown4            421570 non-null  float64
13  Markdown5            421570 non-null  float64
14  CPI                  421570 non-null  float64
15  Unemployment          421570 non-null  float64
16  IsHoliday_y          421570 non-null  bool
dtypes: bool(2), float64(10), int64(3), object(2)
memory usage: 52.3+ MB
```

Figure 1. Summary of Feature dataset

IV. MODELS AND ALGORITHM

Machine learning (ML) is a subset of artificial intelligence that allows computers to learn and make decisions without being explicitly programmed. In this study, several machine learning algorithms are applied to the prediction of sales. In this section, these methods are introduced.

a. Linear Regression

A simple and widely used kind of predictive analysis is linear regression. Regression's main goal is to look at two things: (1) How well can a group of predictor variables predict a dependent variable that is the outcome? (2)

Which specific variables, as shown by the size and sign of the beta estimates, are important predictors of the outcome variable, and how do they affect the outcome variable? There are two main types:

1) Simple regression

$$y=mx+b \quad \dots\dots\dots(1)$$

2) Multivariable regression

$$f(x,y,z)=w1x+w2y+w3z \quad \dots\dots\dots(2)$$

b. KNN Classification Model

For classification and regression applications, the K-Nearest Neighbour (KNN) algorithm is a well-liked machine learning method.

c. Random Forest Regressor

Random forest is a supervised learning algorithm. It has two variations – one is used for classification problems and other is used for regression problems.

d.. XGBoost Classification Model

It is essential that you understand your data in order to make better decisions, and XGBoost can help you with this. Decision trees based on gradient-boosting can be implemented using XGBoost. Machine-learning models optimized by data scientists and researchers worldwide have been optimized using this tool. In addition to being fast and easy to use, XGBoost can handle large datasets with ease. The parameters do not need to be optimized or tuned, so it can be used as soon as it's installed.

e. ARIMA (AutoRegressive Integrated Moving Average):

ARIMA is combines autoregression, differencing, and moving averages. It is particularly effective for univariate time series data where patterns are observed over time. ARIMA models are widely used in various domains, including finance, economics, and demand forecasting

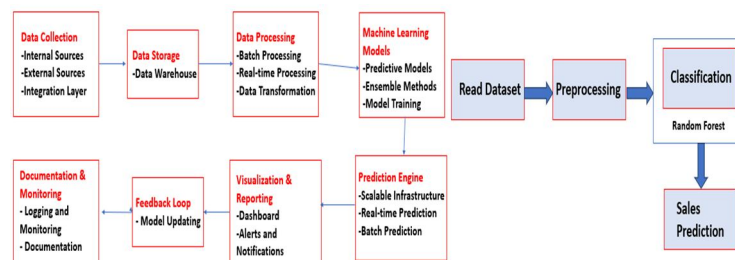


Figure 2 : System Architecture

Above architecture collects the data from the dataset. Performs preprocessing i.e. cleans the data and then applies classification. Having successfully completed preceding steps, dataset is now prepared to be used to develop suggested model. It is utilized as a prediction model for Wall Mart sales after model is built.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

In this part, the process of the experiment will be introduced. This section starts with an overall introduction to the dataset used in this paper, which includes descriptive statistics, then followed by pre-processing of the dataset. The third step is the selection of features. Finally, the machine learning models are applied to the dataset and compared with the same metric to select the best model. Firstly, We have taken large data to predict accuracy in machine learning.so we have imported required packages and libraries to run Python code. We are using different platform like VS Code, Jupyter Notebook ,etc. Now moving towards data preprocessing and cleaning the noisy data from given dataset.

Data Processing

Missing Value:- The markdown columns were given by Walmart to see the effect of markdowns on sales. In fact, there are many Nan values for markdowns, more than 250000 in each markdown column (i.e., there are many

missing values in markdowns). The solution for them is to fill the missing value with 0. It shows that there is no markdown for that week.

Encoding Categorical Data :- The type of Is holiday is bool and contains true and false which define into numerical values as 1 and 0. Meanwhile, replace the type of stores from A, B, and C with 1, 2, and 3.

Data Normalization :- This method restricts the unnormalized data to a specified upper and lower bound by linear operations. The function converts the original data to the range of [0, 1], and the final result is considered normalized data. The linear formula is shown below(3):

$$x' i, = xi, - \min(xi) \max(xi) - \min(xi) (nMax - nMin) + nMi \quad \dots\dots\dots(3)$$

After removing all null values, blank spaces, noisy data we have to merge datasets. In data preprocessing, we are going to merge three dataset together and making new dataset as our targeting field and split the date by year, month, week and add this new column to show properly.

add total markdown in single column to reduce column size. Outlier Detection and Abnormalities or describe the data column. calculate numeric data from given dataset. making holiday column as int data type 0/1.

Data Visualization:-

After Data Preprocessing we are going to visualize given data with different parameters. We have already imported required packages and libraries so we can move further for visualize the data by different methods like Matplotlib, seaborn, heatmap, plotly, etc. From that we can figure out average monthly sales, monthly sales for each year, average sale per store, average sale per department, effect of temperature on sale, Pie chart distribution, correlation matrix and many other difference we can easily getting by visualization technique. Figure 4 shows Correlation matrix

V. RESULT

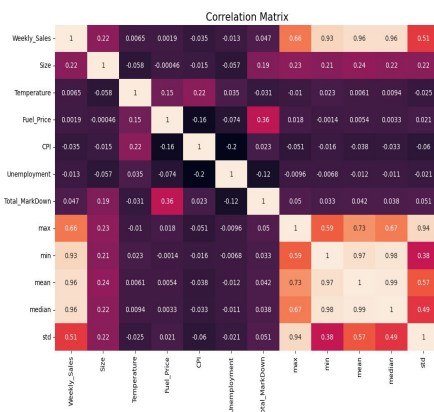


Figure 3. Correlation Matrix

TABLE III. MODEL ACCURACY

Algorithm	Accuracy
Linear Regression:-	92.28
Random Forest Regressor	97.89
KNN Classification Model:	95.62
XGBoost Classification	97.34
ARIMA Model:-	97.80

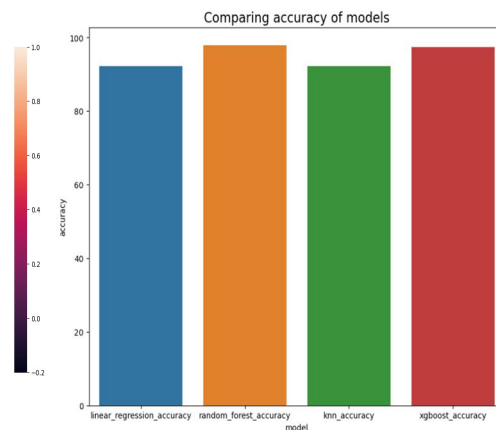


Figure 4. Accuracy Comparison between Model

TABLE IV. MODEL MEAN SQUARE ERROR RATE

Algorithm	MAE	MSE	RMSE	R ²
Linear Regression	0.0300	0.003483	0.0590	0.92283
Random Forest Regressor	0.01548	0.00094	0.0308	0.97896
KNN	0.02114	0.001975	0.0444	0.9562
XGBoost	0.0197	0.0011	0.0346	0.97346
ARIMA Model	0.07604	0.0091	0.0456	0.9129

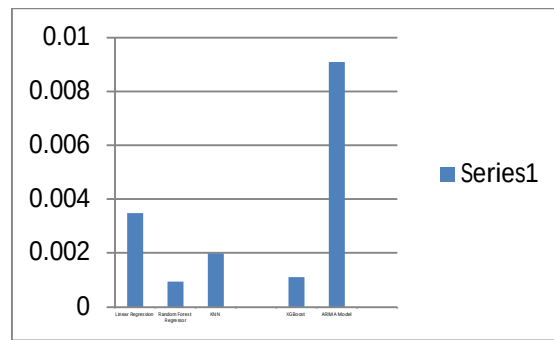


Figure 6. All Model MSE

VI. CONCLUSION

In this paper, the sales of Walmart retail outlets are examined using regression tools, including linear regression, random forest regressor. Using the dataset provided by Walmart in Kaggle. The Best Accuracy from 5 Model is "Random Forest Regressor" which is 97.89697340526645. The experiment We can Use this Model anywhere for large Dataset and apply easily. From this Project We can Consider that this Model is Feasible Model for us. The experiment shows that the linear and evaluation metrics regression provides higher performance and accuracy compared to other regression models. The variables involved in this experiment are very limited and should include more external factors such as the location of the supermarket, and more internal factors such as the sales of each product for future study. The author also believes that it would be beneficial to evaluate the performance of the same strategies in the same situation but with historical data. In circumstances when time series data is available, such as in the case of continuity products, it seems relevant to examine the potential of these models.

REFERENCES

- [1] Chu, C. W., & Zhang, G. P. (2003). A comparative study of linear and nonlinear models for aggregate retail sales forecasting. *International Journal of production economics*, 86(3), 217-231.
- [2] Zhao, J., Tang, W., Fang, X., Wang, J., Liu, J., Ouyang, H., ... & Qiang, J. (2015, October). A novel electricity sales forecasting method based on clustering regression and time-series analysis. In *Proceedings of the 2015 International Conference on Artificial Intelligence and Software Engineering*.
- [3] Ching Wu Chu and Guoqiang Peter Zhang, "A comparative study of linear and nonlinear models for aggregate retail sales forecasting", *Int. Journal Production Economics*, vol. 86, pp. 217-231, 2003.
- [4] 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. "Data Mining based Prediction of D 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): 101- 110
- [5] Giuseppe Nunnari, Valeria Nunnari, "Forecasting Monthly Sales Retail Time Series: A Case Study", *Proc. of IEEE Conf. on Business Informatics (CBI)*, July 2017.
- [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, May 1999.
- [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] Baba, Norio, and Hidetsugu Suto. "Utilization of artificial neural networks and GAs for constructing an intelligent sales prediction system." In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, vol. 6, pp. 565-570. IEEE, 2000.
- [10] Cheriyan, Sunitha, Shaniba Ibrahim, Saju Mohanan, and Susan Treasa. "Intelligent Sales Prediction Using Machine Learning Techniques." In *2018 International Conference on Computing, Electronics & Communications Engineering (icCECE)*, pp. 53-58. IEEE, 2018.
- [11] Fawcett, Tom, and Foster J. Provost. "Combining Data Mining and Machine Learning for Effective User Profiling." In *KDD*, pp. 8-13. 1996.
- [12] Friedman, Jerome H. "Stochastic gradient boosting." *Computational Statistics & Data Analysis* 38.4 (2002): 367-378.
- [13] Giering, Michael. "Retail sales prediction and item recommendations using customer demographics at store level." *ACM SIGKDD Explorations Newsletter* 10, no. 2 (2008): 84-89.

- [14] Kulkarni, Vrushali Y., and Pradeep K. Sinha. "Random forest classifiers: a survey and future research directions." *Int J AdvComput* 36.1 (2013): 1144-53.
- [15] Panjwani, Mansi, Rahul Ramrakhiani, Hitesh Jumnani, Krishna Zanwar, and Rupali Hande. *Sales Prediction System Using Machine Learning*. No. 3243.EasyChair, 2020.
- [16] Mrs. Disha Sushant Wankhede, Dr. Selvarani Rangasamy, "REVIEW ON DEEP LEARNING APPROACH FOR BRAIN TUMOR GLIOMA ANALYSIS" *Journal of Information Technology in Industry*, VOL. 9 NO. 1 (2021) pp. 395 - 408 , DOI: <https://doi.org/10.17762/itii.v9i1.144>
- [17] Disha Sushant Wankhede, R. Selvarani, Dynamic architecture based deep learning approach for glioblastoma brain tumor survival prediction, *Neuroscience Informatics*, Volume 2, Issue 4, 2022, 100062, ISSN 2772-5286, <https://doi.org/10.1016/j.neuri.2022.100062>. (<https://www.sciencedirect.com/science/article/pii/S2772528622000243>)
- [19] Wankhede, D.S., Shelke, C.J. (2023). An Investigative Approach on the Prediction of Isocitrate Dehydrogenase (IDH1) Mutations and Co-deletion of 1p19q in Glioma Brain Tumors. *Intelligent Systems Design and Applications. ISDA 2022. Lecture Notes in Networks and Systems*, vol 715. Springer, Cham. https://doi.org/10.1007/978-3-031-35507-3_19
- [20] Sutar, S. ., Jose, K. ., Gaikwad, V. ., Mishra, V. ., Wankhede, D. ., & Karnik, M. . (2023). Enhancing Data Management: An Integrated Solution for Database Backup, Recovery, Conversion, and Encryption Capabilities. *International Journal of Intelligent Systems and Applications in Engineering*, 12(6s), 720–734. Retrieved from <https://ijisae.org/index.php/IJISAE/article/view/4011>