SALES PREDICTION USING MACHINE LEARNING ALGORITHMS

AMRITA SCHOOL OF ENGINEERING CHENNAI

Bachelor of Technology

RUPESH NAIDU.M

Department of Computer Science Engineering

maradanarupeshnaidu@gmail.com

ABSTRACT:

Every aspect of life is changing due to machine learning, which is also playing a significant role in real-world situations. Every sector has benefited from the innovative uses of machine learning, including those in educations, health issues, commerce, entertainment, and transportation. Because they are conducted without consideration of customers' buying habits, the traditional approach to sales and marketing goals no longer assist businesses in keeping up with in the pace of a competitive market.

As a result of improvements in machine learning, significant changes can be observed in the field of sales and marketing. Thanks to these developments, it is now possible to readily determine a number of important factors, including target market and forecasting sales for the upcoming years. informing the sales team's plans for a growth in their business. The purpose of this study is to suggest a factor for forecasting Mart Company upcoming sales while taking into account their past sales. Machine learning models like Linear Regression, K-Neighbors Regressor, XGBoost Regressor, and Random Forest Regressor.

Key parameters: XGBoost Regressor, Random Forest Regressor, Linear Regressor, K Nearest Neighbour Regressor.

INTRODUCTION:

Sales forecasting is always an important area to focus on. To maintain the effectiveness of the marketing organizations, all vendors now need to forecast in an effective and optimal manner. Manually performing this work could result in grave mistakes that would result in bad management of the organization, and most significantly, it would take time, which is not desired in today's time-constrained world. The business sectors, who are actually expected to generate enough goods in the right amounts to satisfy demand, are a significant component of the global economy. The main objective of business sectors is to target the market audience. It is crucial that the business has been successful in achieving its goal by using a system. In order to make predictions, it is necessary to analyse the data from a variety of sources, including market trends, customer behaviour, and other elements. The companies would benefit from this analysis by having better financial resource management. The forecasting method can be used for a variety of things, such as estimating future demand for the product or service and estimating how much of the product will be sold in a specific time frame. Here, machine learning has a lot of potential for use. In the

In our paper, we designed machine learning algorithms considering data collected from previous sales in a grocery store. The main objective is to predict the sales pattern and quantity of products to be sold based on some key features.

DATA PARAMETERS:

1.Item weight

2.Item fat content

3.Item visibility

4.Item type

5.Item MRP

6.Outlet establishment year

7.Outlet size

8.Outlet location

Analysis and survey of the collected data was also done to get a complete overview of the data. The analysis would help business organizations to make probabilistic decisions at every important stage of marketing strategy

LITERATURE SURVEY:

Identify new product issues, diagnose the root cause of manufacturing issues, and profile existing customers with more accurate and specific metrics. This huge collection of data values is either related or not related at all, So clustering is essential. Otherwise much of the backed data is useless to the user. Increase sales seasonally by updating inventory, discount offers, and store layouts based on knowledge discovered in your data. Authors analyzed sales data using clustering algorithms such as K-Means and EM. This revealed many interesting patterns that could help improve sales revenue and increase sales volume. Our research shows that segmentation methods such as K-Means and EM algorithms are better suited for analyzing sales data compared to density-based methods such as DBSCAN and OPTICS and hierarchical methods such as his COBWEB. Confirmed. Authors presented a new data clustering method for data mining in large databases. Simulation results show that the proposed new clustering method performs better. Moreover, our method produces much smaller errors than both the FSOM+K-Means approach and GKA in all investigated cases. Actual quantities and seasonal factors are very important for some product lines that are close to sales forecast results. Evaluating the forecast results suggested various campaigns and marketing techniques to market the company's products. By evaluating the predicted results, various campaigns and marketing techniques were suggested to market the company's products. This study analyzes data collected from retail companies to make predictions about the company's future business strategies. The effects of a large series of events such as weather, holidays, etc. can actually change the state of various departments. Therefore, these effects are also analyzed and their impact on sales is taken into account. In this theory, various individual algorithmic machine learning techniques are used to generate good and optimal results, which are further analyzed for the prediction task. Using ensemble techniques, 4 algorithms, etc.

"Sales Prediction System Using Machine Learning"

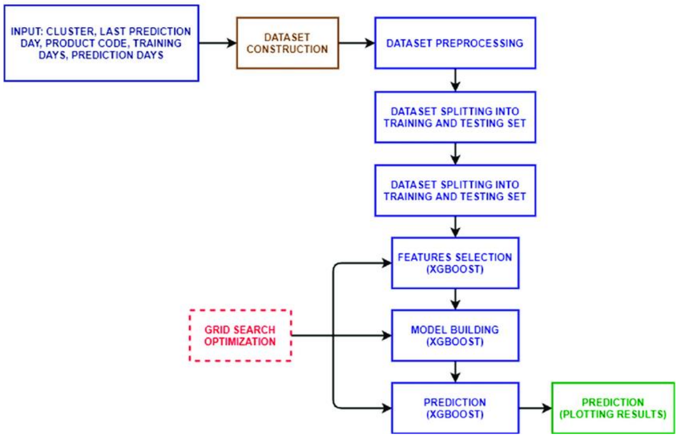
The main purpose of this study is to get accurate results for future forecasting. By using methods such as clustering and metric modeling to forecast sales, the goal of this article is to obtain accurate results to predict future sales or demand for a business. The potential of the algorithms is evaluated and used appropriately in further studies. This study examines how to make judgments based on experimental data and insights gained through data visualization.

Data mining techniques were used. The Gradient Boost algorithm has proven the highest level of accuracy in predicting upcoming trades. Retail sales forecasting and product recommendations based on customer demographics at the store level.

This article describes a product recommendation system and a sales forecasting system that was used to the benefit of a set of stores. Consumer demographic information was used to accurately design individual sales. "Building an intelligent sales forecasting system using artificial neural networks and GA" In the study, deep neural network techniques were used to predict their electronic components sales strategy.

This study shows how to use automated prototyping to identify suspicious behavior. To come up with this matching prototype, several machine learning methods have been leveraged. Here, differences in cell phone owners' behavior are detected using a combination of constructive inductive and data mining techniques.

FLOW DIAGRAM:


INPUT: CLUSTER, LAST PREDICTION DAY, PRODUCT CODE, TRAINING DAYS, PREDICTION DAYS
DATASET CONSTRUCTION
DATASET PREPROCESSING
DATASET SPLITTING INTO TRAINING AND TESTING SET
DATASET SPLITTING INTO TRAINING AND TESTING SET
FEATURES SELECTION (XGBOOST)
GRID SEARCH OPTIMIZATION
MODEL BUILDING (XGBOOST)
PREDICTION (XGBOOST)
PREDICTION (PLOTTING RESULTS)

DATA VISUALIZATION:

Heat map for finding the correlation between the dataset attributes:

The correlation between the target variable and the other attributes is shown using a heat map, an element of the data visualization library called Seaborn a color-coded matrix from the data visualization package Seaborn.

Lower is the target variable's dependence on the corresponding attribute, the more intense the color of an attribute is in relation to the target variable. The target variable has been seen to be Item_Outlet_Sales is least dependent on Item_Visibility and most dependent on Item_MRP. Thus, higher the MRP of an item, lower will be the Item_Outlet_Sales.

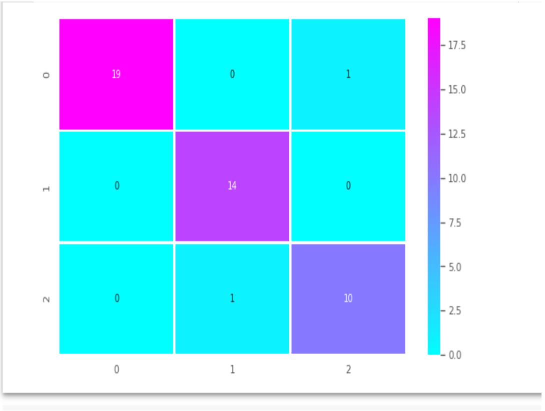


FIG-1: **Heatmap** for correlation between attributes

COUNT PLOTS:

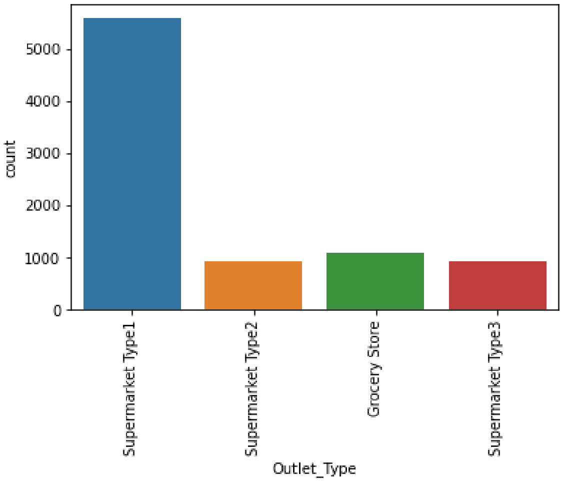


FIG 2: Count of each outlet type.

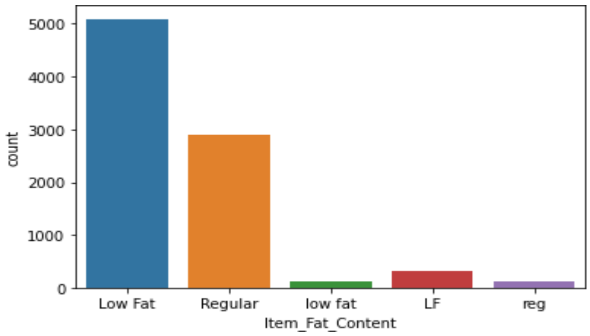


FIG-3: No. of item with each type of fat content

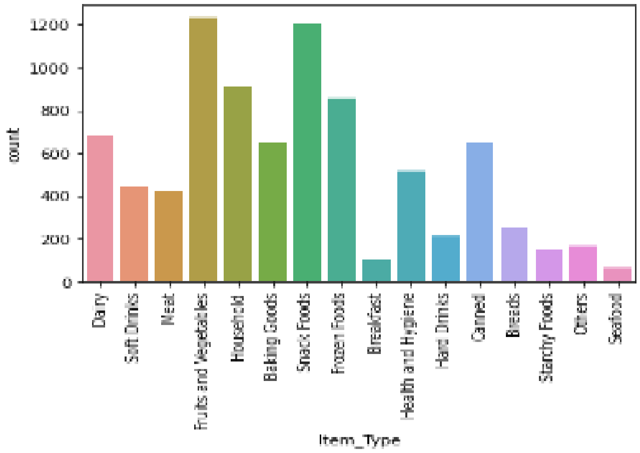


FIG-4: No.of items with each type

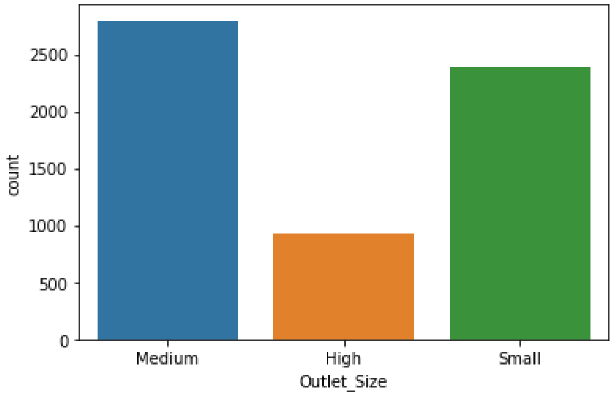


FIG-5: No.of outlets of different sizes

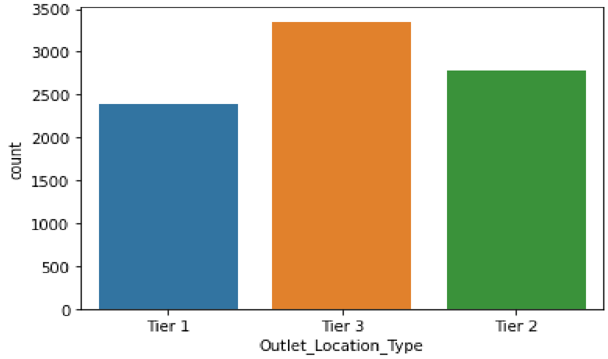


FIG-6: No.of outlets belonging to different categories of location type.

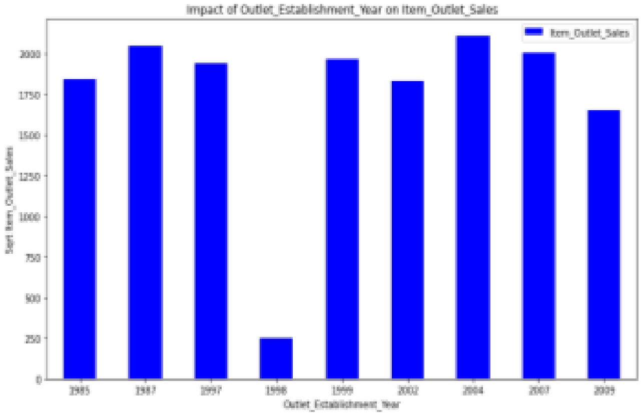


FIG-7: Impact of Outlet Establishment Year on Outlet Sales.

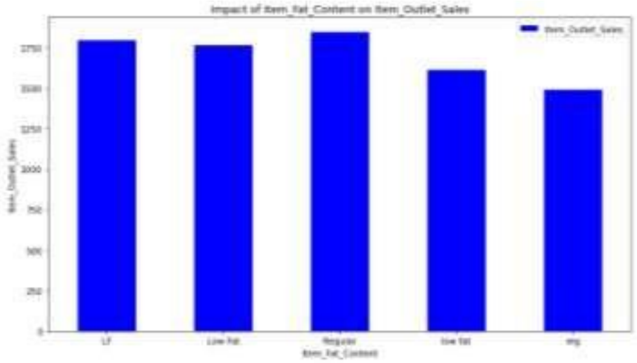


FIG-8: Impact of Item Fat Content on Outlet Sales.

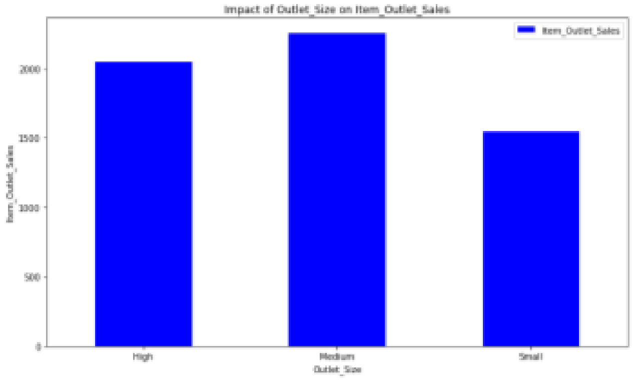


FIG-9: Impact of Outlet Size on Outlet Sales.

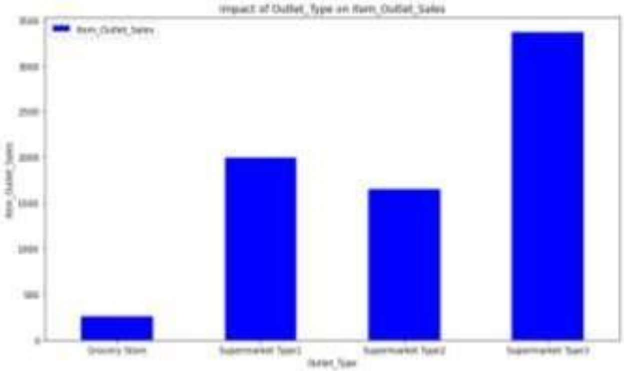


FIG-10: Impact of Outlet Type on Outlet Sales.

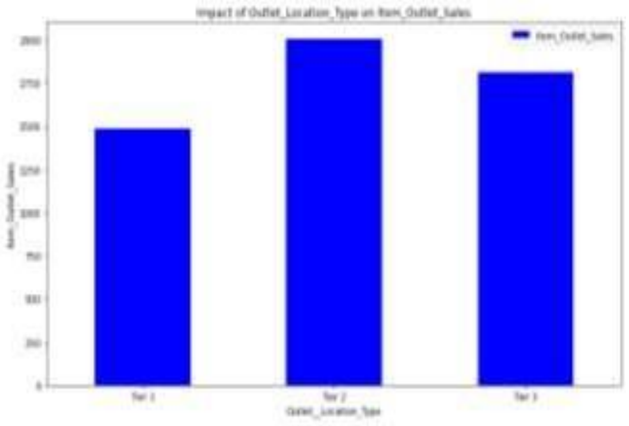


FIG-11: Impact of Outlet Location Type on Outlet Sales.

DATA PREPROCESSING:

Data must be created before being utilized in machine learning algorithms because it cannot be used in its original format due to the method of acquisition. This method is employed to resolve issues that the knowledge extractor is still learning about. We refer to this as preparatory work. Finding out what information the car requires in order to decide whether or not to use it is the aim of preprocessing. Clean, properly formatted data is need for preprocessing.

The following tasks are included in data preprocessing:

Importing: In our project we have collected the data from a grocery store to predict the future sales.

The dataset can be categorized into two forms which are test and train.

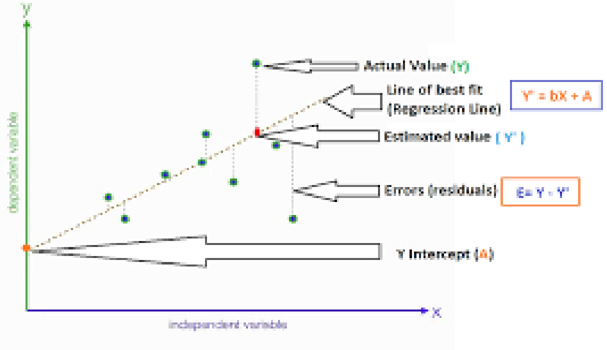
To predict the outcomes of a test we use the test data which is the major part of the process.

FOLLOWING ALGORITHMS ARE USED:

1. **LINEAR REGRESSION**
2. **XGBOOST REGRESSOR**
3. **RANDOM FOREST**

LINEAR REGRESSION:

Linear regression is the most popular and widely used algorithm in machine learning algorithms. It is used to find or create or establish a linear relationship between a target or dependent variable and a response or independent variable. A linear regression model is based on the following equation:



$$y^{\wedge} = \theta0 + \theta1x1 + \theta2x2 + \theta3x3 + \ldots\ldots + \theta nxn$$

where, y^ is the target variable, θ0 is the intercept, x1,x2,x3,….,xn are the independent variables and θ1, θ2, θ3, ……, θn are their respective coefficients.

The main goal of this algorithm is to find the best fit path with the target variable and the independent variables of the data. It is obtained by finding the most optimal values for all θ. In the most appropriate way, we mean that the predicted value should be very close to the true value and have minimal error.

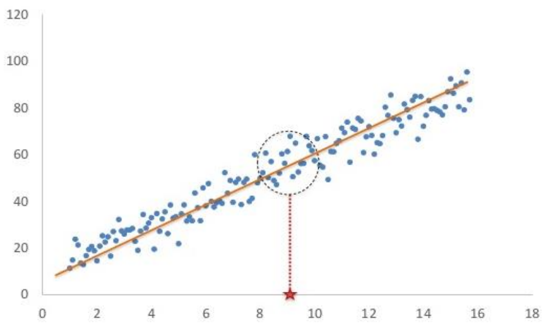
Error is the distance between the data points and the fit regression line and can usually be calculated using the following equation:

Error=y – y',
where, y is the true value and y' is the value guess.

K NEAREST NEIGHBOUR REGRESSOR:

A supervised learning strategy is used in the KNN method for regression. Based on the similarity with other available cases, it predicts the target. The distance measure is used to calculate the similarity,

KNN regression algorithm being the supervised learning method. It predicts the target based on similarity with other available cases. Similarity is calculated using a distance measure



With Eucidian distance being the most common approach.

$$\sqrt{\sum_{i=1}^{n}(yi - xi)\,\hat{}\,2}$$

XGBOOST REGRESSOR:

Extreme Gradient Boosting, or XGBoost, has been used to create a model with great computational speed and efficacy. An ensemble method that models the expected error of several decision trees to optimize previous predictions is used in formulas to create predictions. The development of this model also reports the importance of each feature's contributions to the prediction of the final construction performance score. This feature value shows the impact each attribute has on predicting school achievement in absolute terms. XGBoost supports parallelization by building decision trees in parallel. This algorithm can evaluate any huge and complex model, which makes distributed computing another important characteristic it possesses. Due to the vast and diverse datasets it analyses, it is an out-core computation. Resource utilization is handled fairly well by this computational model. Additional models should be implemented for each individual step to reduce the errors encountered.

XGBoost function at iteration t is:

L(t)= ∑i=1nL(y_outi, y_out1i(t-1) + ft(xi) + g(ft)

where, y_out = real value knowm from the training dataset, and the summation part could be said as

f(x + dx) where x= y_out1i(t-1)

We need to take the Taylor approximation. Let's take the simplest linear approximation of f(x) as:

f(x)= f(b) + f `(b)(x-b)   dx= ft(xi)

Where, f(x) is the loss function L, while b is the previous step (t-1) predicted value and dx is the new learner we need to add in step t.

Second order Taylor approximation is:

f(x)= f(b) + f`(b)(x-b) + 0.5f``(b)(x-b)^2

L(t)=∑i=ln[L(y_outi,y_out1(t-1))+hift(xi)+
0.5kift^2(xi)] + g(ft)

If we remove the constant parts, we have the following simplification objective which is minimized at step t.

L(t) = ∑i = ln[hift(xi) + 0.5kift^2(xi)] + g(ft)

RANDOM FOREST REGRESSOR:

A random forest is defined as a collection of decision trees that help provide correct output using a bagging mechanism. Bagging and boosting are the two most popular ensemble techniques aimed at dealing with higher variability and bias. In bagging, there are some basic learners. In other words, the basic model that takes different random samples of the data set from the training data set. For the Random Forest Regressor, the decision tree is the base learner and is trained on the collected data. Decision trees themselves are not exact learners. If implemented to the full depth, we often run the risk of overfitting where the training accuracy is high but the actual accuracy is low.

As a result, we use a bootstrapping technique to distribute samples from the main data file to each decision tree. This methodology uses row sampling with replacement and feature sampling. The end result is that each model is trained using all these data files, and each time we feed test data into one of the already used trained model, the predictions made by each model are combined. and the output looks like this: is the average of all results produced. Aggregation here refers to the process of combining various results. The hyperparameter that needs to be controlled in this algorithm is the number of decision trees, which should be considered when creating the random forest.

RESULTS

Machine learning algorithms such as linear regression Algorithm, K- Nearest Neighbors Algorithm, XGBoost, and Random forest algorithms were used to predict the sales of various mart outlets. Various parameters such as root mean squared error (RMSE), variance values, and training and testing accuracies that determine the accuracy of the results are tabulated for each of the four algorithms. The XGBOOST and Random Forest algorithms turned out to be the best algorithms with 96.57% accuracy.

CONCLUSION

The use of machine learning approaches has proven to be an important aspect for designing business strategies while considering consumer purchasing behavior, as traditional methods are not very helpful in increasing sales for business organizations. Forecasting sales in relation to a variety of factors, including previous year's sales, helps companies formulate appropriate strategies to increase sales and remain fearless in a highly competitive world.

REFERENCES

1. https://ieeexplore.ieee.org/document/8659115
2. https://thecleverprogrammer.com/2021/05/19/sales-prediction-with-machine-learning/
3. https://graphite-note.com/machine-learning-sales-forecasting
4. https://nebula.wsimg.com/9c57886ee8e628af1879d9497f1b379f?AccessKeyId=DFB1BA3CED7E7997D5B1&disposition=0&alloworigin=1