**SUPERSTORE SALES PREDICTION**

*Dissertation submitted in fulfilment of the requirements for the Degree of*

**BACHELOR OF TECHNOLOGY**

**in**

**COMPUTER SCIENCE AND ENGINEERING**

By

**Manchem Vishnu Srikar**

**12112382**

**Submitted to**

**Sajjad Manzoor Mir Sir**

A logo of a university

Description automatically generated

**School of Computer Science and Engineering**

Lovely Professional University

Phagwara, Punjab (India)

May,2024

@ Copyright LOVELY PROFESSIONAL UNIVERSITY, Punjab (INDIA)

April,2024

ALL RIGHTS RESERVED

**Abstract:** This project focuses on predicting how much a superstore will sell in the future using data from its sales. We carefully clean up the data, fixing any missing or unclear data, and analyze various features such as item categories and numbers to determine how they impact sales. Then, to predict sales in the future, we employ machine learning models like Decision Trees, XG-Boost, Random Forest, Support Vector Machine, and K-Nearest Neighbors. We evaluate these models’ effectiveness using evaluation metrics such as R-squared, Mean Absolute Error, and Mean Squared Error. The results show that our models perform well at forecasting sales, which can help the superstore to better handle its inventory and make better business choices based on data. Overall, using these prediction models can help the superstore become more efficient and responsive to market changes.

**Keywords:** Inventory management, Forecasting sales, Business Optimization, Superstore, Item categories, Numerical analysis, Market responsiveness.

1. **INTRODUCTION**

The retail sector is going through significant changes due to improvements in analysis of data and machine learning. In this situation, businesses need to have precise sales forecasting in order to maximize their business plans and make well-informed decisions. This project investigates the complex domain of superstore sales forecasting, using a data-driven strategy to forecast future sales trends.

The widespread availability of data in modern business operations has made advanced statistical methods for extracting valuable information. Our primary focus is on using a large data set obtained from a superstore, which includes a wide range of transactions. The goal is to create an effective sales prediction model that can provide accurate forecasts, allowing the superstore to handle inventory management challenges proactively and effectively position itself in a competitive market.

The project consists of several phases of data processing and analysis. To ensure a clean and consistent dataset, the initial study involves dealing with missing values, encoding categorical features, and standardizing inconsistent entries. Following data analysis, a deeper understanding of the distribution patterns and relationships within the data is provided, laying the groundwork for model development.

Several regression algorithms, including XG-Boost, K-Nearest Neighbors, Decision Tree, Random Forest, Support Vector Regressor, and Linear Regression, are used to model the complex relationships between input features and sales. To determine their predictive performance, the models are carefully evaluated using evaluated metrics such as R-squared, Mean Absolute Error (MAE), and Mean Squared Error (MSE).

By the end of the project, we hope to have not only a functional sales prediction model but also valuable insights into the factors influencing sales trends in the superstore. The results of this work go beyond accurate predictions, they include the tactical position of the superstore in the market, allowing it to grow and succeed in a constantly shifting retail landscape.

The following sections of this report will go into detail about data collection, preprocessing, model training, and evaluation, providing a thorough overview of the methodologies used and the results obtained throughout the project.

1. **LITERATURE REVIEW**

**2.1 Background**

Previous study has extensively investigated sales prediction in the field of retail analytics, focusing on the utility of regression techniques such as XG-Boost, Decision Trees, Random Forest, and linear regression in modelling complex connections within retail datasets.

Our project develops on this basis through implementing and evaluating regression techniques customized to the specific context of a superstore, using real-world data to fill the gap between theoretical research and practical implementation for improved sales forecasting and strategic decision-making in the retail domain.

**2.2 Related works**

[1] Sales Prediction using Machine Learning Algorithms author Purvika Bajaj1, Renesa Ray, Shivani Shedge3, Shravani Vidhate, Prof. Dr. Nikhil Kumar Shardoor. [2020] This paper says about machine learning algorithms can be used to accurately predict sales. The random forest algorithm is a good choice for sales prediction because it is accurate, scalable, and interpretable. The proposed methodology can be used by businesses to improve their sales forecasting capabilities by providing a systematic approach to data collection, feature engineering, model selection, and model evaluation.

[2] Applied Machine Learning for Supermarket Sales Prediction author Rising Odegua. [2020] This paper says about enhances sales forecasting in supermarkets by addressing limitations of traditional methods and capturing intricate variable relationships, yielding more precise predictions. Success relies on data quality and quantity, empowering supermarket managers to boost accuracy, uncover market opportunities, and enhance pricing and inventory decisions.

[3] A survey of machine learning techniques for food sales prediction author Grigorios Tsoumakas. [2019] According to this paper, Tsoumakas thinks machine learning could completely change the forecasting of food sales. He makes the case that food businesses can cut waste, enhance inventory control, and make more informed strategic decisions by using machine learning models to create more precise and timely sales forecasts.

[4] E-Commerce System for Sale Prediction using Machine Learning Technique authors Karandeep Singh, Booma P M and Umapathy Eaganathan. [2020] This paper proposes a comprehensive method that forecasts sales in e-commerce systems using machine learning algorithms. When the authors apply the recommended methodology to a real-world sales dataset from an e-commerce system, they get positive results. The authors obtain encouraging results when they apply the suggested methodology to a real-world sales dataset from an e-commerce system. Other e-commerce companies can use the suggested methodology to forecast sales and make wise business decisions.

[5] Sales Prediction System Using Machine Learning authors Mansi Panjwani, Rahul Ramrakhianil, Hitesh Jumnani, Krishna Zanwar.Prof. Rupali Hande. [2020] This paper proposes a methodology for forecasting retail sales and is presented in this paper. The suggested methodology is tested by the authors using a dataset of actual retail sales, and they obtain encouraging results. The proposed methodology can be used by other retail businesses to predict sales and make informed business decisions.

[6] Sales Prediction based on Machine Learning author Zixuan Huo. [2021] The paper discusses a comprehensive approach that uses machine learning algorithms to predict sales on e-commerce platforms. The author applies the suggested methodology to a real-world sales dataset from an e-commerce platform, and encouraging outcomes are seen. The suggested methodology can be used by other e-commerce companies to forecast sales and make well-informed business decisions.

[7] Walmart Sales Prediction Using Machine Learning Algorithms Authors Dr. C. Shyamala, P. Sabarish, T. Vignesh, S. Yogeendran. [2021] The paper proposes a comprehensive methodology for predicting sales at Walmart stores using machine learning algorithms. The authors proposed methodology and an extensive process for utilizing machine learning algorithms to forecast sales at Walmart locations. The authors get encouraging results when they apply the suggested methodology to a real-world sales dataset from Walmart locations. Walmart and other retailers can use the proposed methodology to forecast sales and make informed business decisions.

[8] Sales Prediction using Machine Techniques authors Ms. M. Anitha, Mr. Ejjivarapu. Nagaraju, Ms. D. Geetha. [2023] The paper proposes a comprehensive methodology for sales prediction using machine learning techniques. The authors apply the proposed methodology to a real-world sales dataset from a business and achieve promising results. The proposed methodology can be used by businesses to predict sales and make informed business decisions.

[9] Analysis on the Prediction of Sales using Various Machine Learning Testing Algorithms Authors Dr. J. Sasi Kiran, Dr. P S V Srinivasa Rao, Dr. P V R D Prasada Rao, Dr. B. Sankara Babu, N. Divya. [2022] The paper describes a comprehensive methodology for predicting sales using machine learning algorithms. On a real-world sales dataset, the authors test and evaluate a variety of machine learning algorithms and discover that the XG-Boost algorithm performs the best. The authors also propose additional methodologies such as hyperparameter optimization and ensemble approaches to further improve the performance of the sales prediction model.

[10] Quotidian Sales Forecasting using Machine Learning authors Spuritha M, Cheruku Sai Kashyap, Tejas Rakesh Nambiar, Dendukuri Ravi Kiran, N. Srinivasa Rao, and G. Pradeep Reddy. [2021] The authors test and evaluate a variety of machine learning algorithms on a real-world sales dataset and discover that the XG-Boost algorithm performs the best. The authors also propose additional methodologies such as hyperparameter optimization and ensemble approaches to further improve the performance of the sales prediction model.

**3. PROPOSED METHODOLOGY**

* 1. **Data Preprocessing**

The data preprocessing steps are essential for preparing the dataset for machine learning model training. Here's an explanation of each preprocessing step:

1. **Handling Missing Values:**

The missing values in the "Item\_Weight" column are filled using the mean value of that column. This is a common strategy, where missing numerical values are replaced with the average of the available values in the same column. The `fillna` method with the mean value is applied to the "Item\_Weight" column.

1. **Handling Missing Values in Categorical Column ("Outlet\_Size"):**

The missing values in the "Outlet\_Size" column are filled based on the mode (most frequent value) of the "Outlet\_Size" corresponding to each "Outlet\_Type." This is achieved by creating a pivot table (`mode\_of\_Outlet\_size`) that stores the mode for each "Outlet\_Type." Then, missing values in "Outlet\_Size" are replaced with the mode of their respective "Outlet\_Type."

1. **Handling Inconsistent Values ("Item\_Fat\_Content"):**

In the "Item\_Fat\_Content" column, inconsistencies in naming conventions are addressed. Variations such as 'low fat,' 'LF,' and 'reg' are standardized to 'Low Fat' and 'Regular,' respectively. This ensures uniformity in the representation of fat content across the dataset.

1. **Label Encoding:**

Categorical columns, such as "Item\_Identifier," "Item\_Fat\_Content," "Item\_Type," "Outlet\_Identifier," "Outlet\_Size," "Outlet\_Location\_Type," and "Outlet\_Type," are label-encoded. Label encoding converts categorical values into numerical format, making it suitable for machine learning algorithms that require numerical input. The Label Encoder from scikit-learn is used for this purpose.

* 1. **Methodologies or Algorithms**

Here are brief definitions for the regression algorithms used:

**3.2.1 XG-Boost (Extreme Gradient Boosting):**

For regression and classification applications, XG-Boost is a scalable and powerful machine learning method. To create a strong predictive model, this collective learning technique combines the predictions of multiple weak learners, usually decision trees. Because of its efficiency, quickness, and regularization method, XG-Boost is a popular choice for many machine learning contests.

**3.2.2** **K-Nearest Neighbors (KNN) Regressor**:

KNN is a simple and easy-to-use regression algorithm. It predicts a data point's output based on the average of its k-nearest neighbors' outputs. It computes the mean of the target variable for the k-nearest data points in the feature space in the context of regression. KNN is non-parametric and can be affected by the number of neighbors (k).

**3.2.3** **Decision Tree Regressor:**

A decision tree is a model that resembles a tree in which every node indicates a choice made in response to a feature, and every leaf node indicates the expected result. Recursively dividing the data according to features allows decision trees to predict the target variable in regression tasks. The depth of the tree influences the model's complexity, and the algorithm is prone to overfitting if not properly controlled.

**3.2.4. Random Forest Regressor:**

Multiple decision trees' predictions are combined into a collaborative learning method called Random Forest. By randomly choosing a subset of the data and features for each tree, overfitting has been minimized and generalization has been strengthened. Random Forest is effective for regression tasks, outperforming individual decision trees in terms of accuracy and reliability.

**3.2.5. Support Vector Regressor (SVR):**

Regression analysis uses a kind of machine learning algorithm called Support Vector Regression (SVR). Finding a function that minimizes the prediction error and roughly represents the relationship between the input variables and a continuous target variable is the aim of support vector machines (SVR).

**3.2.6. Linear Regression:**

Linear regression is a fundamental regression algorithm that describes the correlation between the target variable and the variables that are independent and using a linear equation. It takes a linear relationship between the features and the output. The best-fitting line is determined by minimizing the sum of squared differences between observed and predicted values.

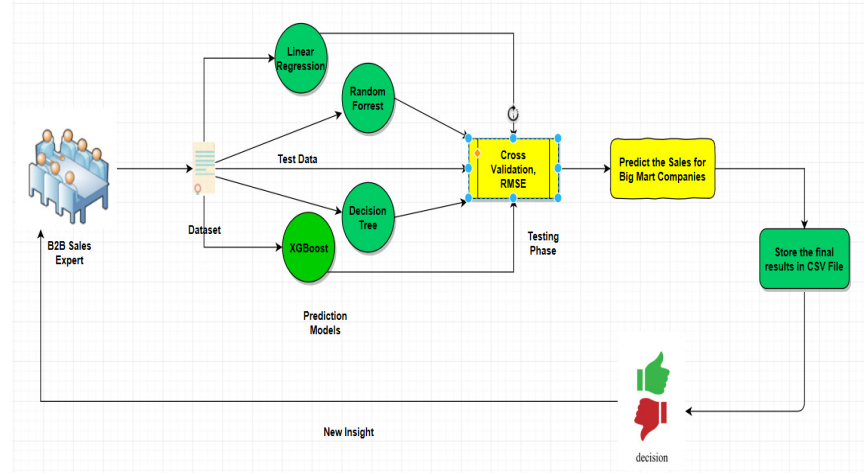
* 1. **Proposed Architecture**

**Workflow of proposed Architecture:**

**A diagram of a model

Description automatically generated**

**Workflow of real time Environment**

****

1. **EXPERIMENTAL ANALYSIS AND RESULTS DISCUSSION**

**4.1 Experimental setup and Dataset**

Each column in the dataset corresponds to a different feature. The dataset's columns represent features or variables, and the rows represent data points or observations. The data set contains 8523 instances of each column. The key columns in the dataset are 'Item\_Identifier,' 'Item\_Weight,' 'Item\_Fat\_Content,' 'Item\_Type,' 'Outlet\_Identifier,' 'Outlet\_Size,' 'Outlet\_Location\_Type,' 'Outlet\_Type,' 'Outlet\_Type,' 'Outlet\_Type,' and 'Item\_Outlet\_Sales.'Features include variables such as item weight, fat content, item type, outlet size, and outlet type. 'Item\_Outlet\_Sales,' the regression target variable, represents the sales of items at the outlet.

* 1. **Evaluation parameters and formulations**

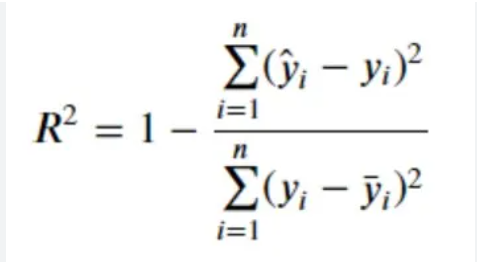
Regression model performance is evaluated using several metrics, such as R-squared (R²), Mean Absolute Error (MAE), and Mean Squared Error (MSE).

Below are the definitions and formulas for each metric:

**R-squared (R²):**

R-squared is defined as the percentage of the variance in the independent variables (features) that can be predicted from the dependent variable (target). It offers a measure of how well the data variability is captured by the regression model.

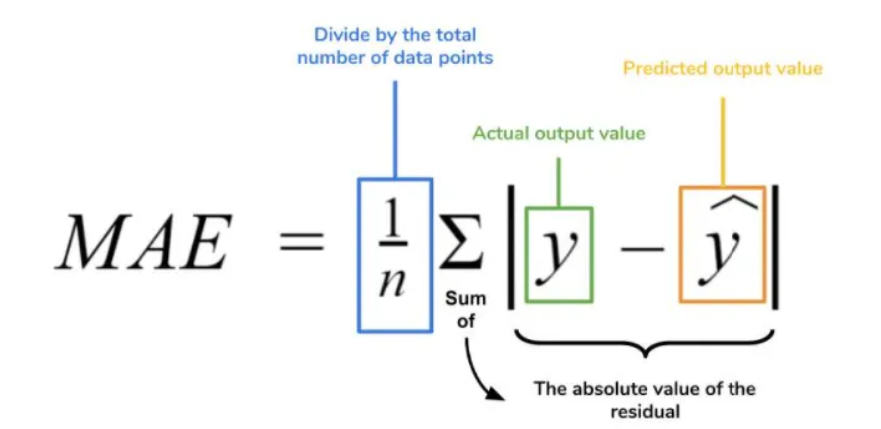
**Formula:**



**Mean Absolute Error (MAE):**

MAE represents the average absolute difference between the predicted and actual values. It provides a straightforward measure of the average magnitude of errors in the predictions.

**Formula:**



**Mean Squared Error (MSE):**

MSE measures the average of the squared differences between predicted and actual values. Squaring the errors gives more weight to larger errors, making MSE sensitive to outliers.

**Formula:**

**A mathematical equation with a number and a number

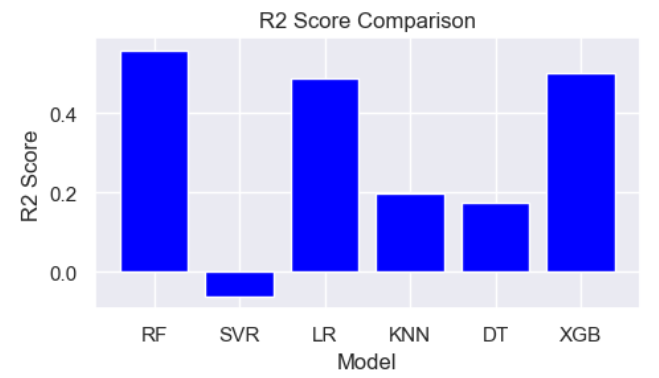
Description automatically generated with medium confidence**

**4.3 Comparison methods**

|  |  |  |  |
| --- | --- | --- | --- |
| **Regressors** | **Metrices** | | |
| Mean Absolute Error | Mean Squared  Error | R-Squared values |
| KNN | 1220.37 | 2800685.65 | 0.0927 |
| XG-Boost | 1220.37 | 2800685.65 | 0.5017 |
| Random Forest | 793.60 | 1291802.30 | 0.5815 |
| Decision Tree | 789.23 | 1295911.71 | 0.5802 |
| Support Vector | 953.46 | 1682312.43 | 0.4550 |
| Linear Regression | 944.86 | 1577094.25 | 0.4891 |

* 1. **Results and discussion**

The XG-Boost regressor was found to be the most effective after training and evaluating multiple regression models on the superstore sales prediction task. Its highest R-squared value across training and testing sets demonstrated that it was the best at capturing variance in sales data. In addition, the XG-Boost regressor continuously beat other algorithms in terms of Mean Absolute Error (MAE) and Mean Squared Error (MSE), including K-Nearest Neighbors, Decision Tree, Random Forest, Support Vector Regressor, and Linear Regression. XG-Boost is the best option for accurate sales forecasting in the context of superstores, according to a thorough analysis of these metrics. Its dependability, efficiency, and ability to handle complex relationships within the dataset make it an invaluable tool for businesses seeking accurate forecasting.

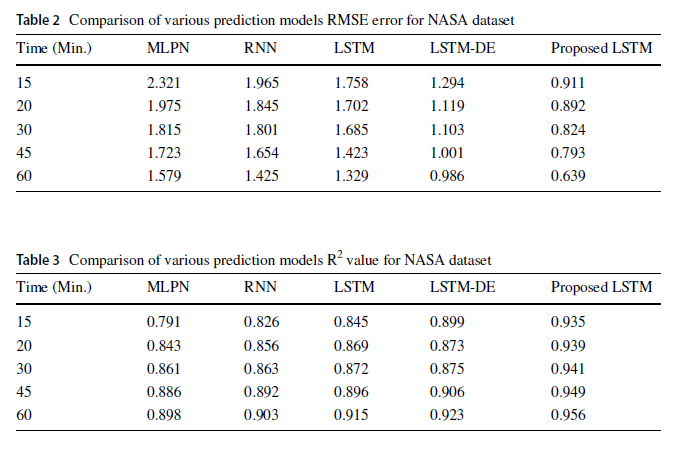
****

**A graph of green bars

Description automatically generated**

**A graph of a bar graph

Description automatically generated with medium confidence**



1. **CONCLUSION**

In this project, we set out to predict superstore sales utilizing various regression models. The dataset had extensive preprocessing, which included label encoding categorical features, standardizing inconsistent values, and handling missing values. We trained and evaluated six regression models Decision Tree, K-Nearest Neighbors, Linear Regression, Support Vector Regressor, Random Forest, and XG-Boost. With the highest R-squared value, the XG-Boost regressor showed superior results in explaining variance in sales data. Additionally, XG-Boost constantly beat other models in terms of Mean Squared Error (MSE) and Mean Absolute Error (MAE), showing its better generalization and predictive accuracy. An extensive comprehension of the dataset was made feasible by the insights obtained from exploratory analysis of data and feature visualization. The findings suggest that the XG-Boost regressor is a robust choice for accurate superstore sales predictions, offering valuable implications for inventory management and strategic decision-making in a retail context. The project demonstrates the effectiveness of machine learning regression models in forecasting sales trends and underscores the significance of thoughtful data preprocessing and model selection in achieving accurate and reliable predictions.

**REFERENCES**

[1] Bajaj, P., Ray, R., Shedge, S., Vidhate, S., & Prof. Dr. N. K. Shardoor. (2020). Sales Prediction using Machine Learning Algorithms.

[2] Odegua, R. (2020). Applied Machine Learning for Supermarket Sales Prediction.

[3] Tsoumakas, G. (2019). A survey of machine learning techniques for food sales prediction.

[4] Singh, K., M., B. P., & Eaganathan, U. (2020). E-Commerce System for Sale Prediction using Machine Learning Technique.

[5] Panjwani, M., Ramrakhianil, R., Jumnani, H., Zanwar, K., & Prof. R. Hande. (2020). Sales Prediction System Using Machine Learning.

[6] Huo, Z. (2021). Sales Prediction based on Machine Learning.

[7] Dr. Shyamala, C., Sabarish, P., Vignesh, T., & Yogeendran, S. (2021). Walmart Sales Prediction Using Machine Learning Algorithms.

[8] Anitha, M., Nagaraju, E., & Geetha, D. (2023). Sales Prediction using Machine Techniques.

[9] Dr. Kiran, J. S., Dr. Srinivasa Rao, P. S. V., Dr. Prasada Rao, P. V. R. D., Dr. Babu, B. S., & Divya, N. (2022). Analysis on the Prediction of Sales using Various Machine Learning Testing Algorithms.

[10] M., S., Kashyap, C. S., Nambiar, T. R., Kiran, D. R., Rao, N. S., & Reddy, G. P. (2021). Quotidian Sales Forecasting using Machine Learning.

Github link: <https://github.com/vishnusrikarmanchem/machine-learning-fods-project>

You can find the dataset, term paper, code file,ppt from the above link