National College of Ireland

Sales Forecasting for Small Medium Enterprises Using Machine Learning

MSc Research Project
Programme Name

Shailesh Subhashchand Yadav
Student ID: X21222801

School of Computing
National College of Ireland

Supervisor: Dr Syed Muslim Jameel

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Shailesh Subhashchand Yadav

Student ID: x21222801

Programme: Msc Data Analytics **Year:** 2022-2023

Module: MSc Research Project

Supervisor: Dr Syed Muslim Jameel

Submission Due Date: 18th September 2023

Project Title: Sales Forcasting for Small Medium Enterprises using Machine Learning

Word Count: 11189 **Page Count 30**

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Shailesh Subhashchand Yadav

Date: 18th September 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Sales Forecasting for Small Medium Enterprises Using Machine Learning

Shailesh Subhashchand Yadav
X21222801

Abstract

Sales prediction is a critical aspect of business operations, enabling Small and Medium Enterprises (SMEs) for making informed decisions, reducing risks, and capitalizing on opportunities in a fast-paced corporate environment. Traditional methods often lack accuracy and fail to account for the desired results using the sales data. This paper consists of two different datasets coming from different business domain which are our small enterprise data comes from superstore business domain whereas our medium enterprises comes from a pharma industry. It also presents a comprehensive approach to sales prediction for SMEs using machine learning techniques. It also showcases that machine learning algorithms are better at predicting the future sales when compared to traditional forecasting methods. Furthermore, the study investigates the impact of external factors on forecast accuracy, such as seasonal trends, marketing campaigns and location. Integrating these aspects improves model performance, allowing SMEs to handle market swings more effectively. Comparing multiple machine learning algorithms for small enterprises random forest gives best results with r2 score of 0. 61% while XGBoost came out to be the best performer with r2 score of 0.98 for medium enterprises.

1 Introduction

Small and medium-sized businesses (SMEs) are constantly challenged to improve their sales processes and increase profitability in today's fiercely competitive business environment. In this research we are considering two business sectors which are pharmacy store and a superstore where the market is too competitive on a daily basis. Making wise business decisions, making the most of resources, and guaranteeing sustainable growth now all depend on one's capacity to predict sales with accuracy. However, complex patterns and quickly changing market trends are frequently missed by manual sales forecasting systems, which limits their usefulness. Fortunately, the emergence of cutting-edge technologies has created new avenues for SMEs to greatly improve their capacity for sales forecast. Machine learning (ML) has emerged as a disruptive force among these ground-breaking technologies, altering how companies forecast and evaluate their sales success. With the use of machine learning algorithms, SMEs can use enormous volumes of historical sales data, consumer behaviour trends, and market indicators to produce data-driven predictions with previously unheard-of levels of accuracy. SMEs may get a competitive edge by leveraging the potential of ML by streamlining their overall sales processes, boosting marketing tactics, and optimizing inventory management.

This paper explores the use of machine learning techniques in small and medium businesses(pharmacy and superstore) sales forecasting. We will go into the fundamental ideas

behind machine learning and discuss how they apply to the process of drawing out useful information from challenging sales datasets. Additionally, we will discuss various ML techniques frequently applied to sales forecasting, contrasting their advantages and disadvantages to find the best models for particular business scenarios. Furthermore, we will examine the problems that SMEs may have when implementing ML-based sales prediction systems, such as data quality concerns, implementation costs, and the need for trained data scientists or analysts. Nonetheless, the potential benefits of precise sales forecasting far outweigh these hurdles, as organizations stand to capitalize on opportunities, avoid risks, and improve resource allocation.

Finally, the combination of small and medium-sized businesses with machine learning-based sales forecast tools heralds a new era in company management. The capacity to generate accurate forecasts, capitalize on market trends, and improve sales methods opens up a whole new universe of opportunities for long-term success and profitability. As we begin on this journey into sales prediction using machine learning, our goal is to encourage and equip SMEs to embrace this technology-driven future, driving their businesses to new heights.

1.1 Motivation and Project Background

Small and medium-sized firms (SMEs) confront the ongoing challenge of attaining sustainable development while keeping a competitive advantage in today's fiercely competitive and quickly changing business world. Accurate sales forecasting is a critical component of SMEs' decision-making, allowing them to manage inventories, allocate resources efficiently, and create successful marketing strategies. Traditional forecasting methodologies, on the other hand, frequently fail to capture nuanced market dynamics and sophisticated consumer behaviour patterns. The purpose for this project is to address the crucial need for SMEs to adopt data-driven insights and harness cutting-edge technology for sales forecasting. Machine learning (ML) has emerged as a disruptive force in predictive analytics, enabling SMEs to produce accurate and informed sales projections by evaluating vast amounts of historical sales data, customer interactions, and industry trends. The potential benefits of applying ML-driven sales forecast are enormous, ranging from reducing operational expenses and inventory waste to increasing revenue and customer happiness.

This project aims to bridge the gap between machine learning's promise and its practical use in the context of small and medium-sized businesses. This research attempts to discover the most successful models for a pharmacy shop and a superstore, as well as their specialized sales forecasting demands, by undertaking an in-depth investigation of the newest breakthroughs in ML algorithms.

1.2 Research Question

How can machine learning techniques be effectively leveraged to enhance sales prediction for small and medium enterprises?

The successful use of machine learning algorithms for predicting accurate sales forecasting in small and medium-sized businesses is very important in current competitive market, so this question aims to solve that challenge. It also aims to investigate how sophisticated methodologies might improve decision-making, capitalize on market trends, and improve sales tactics. The research question stated can be answered by executing the following research objective below.

1.3 Research Objective

The main objective of this study is to analyse and assess the potential of machine learning approaches to revolutionize sales forecasting for small and medium-sized firms (SMEs) such as a pharmacy store and a superstore. The study attempts to discover the best effective models for accurate sales forecasting in these two business scenarios by conducting an in-depth investigation of several ML algorithms. The paper tries to highlight the given objective below:-
-Using accessible information, create and execute reliable machine learning models to forecast future sales for small and medium-sized businesses.
-Compare multiple machine learning algorithms to identify the most effective way for predicting sales for SMEs, providing the highest degree of accuracy.
-Comparison of past models and currently developed model.

1.4 Our Contribution in the Scientific research

This paper contributes to the practical implementation and assessment of sales forecasting models for small and medium enterprises (SMEs) using real-world datasets available on Kaggle. I used two different datasets from Kaggle: one dataset represents the medium enterprise which is a pharmacy shop and the other dataset represents the small enterprises which in our case is a superstore. I hoped to improve the accuracy of sales forecasts and uncover significant insights for these SMEs by doing detailed studies and utilizing innovative approaches.
I used machine learning methods to build a robust sales forecasting model for the pharmacy shop dataset. During the research, I came across a condition that there are lot of null values in the dataset so I replaced them with 0 and figured that the existence of a large number of zero values does has an effect on the model's performance. To figure out the research, I created two different models: one model which includes all the zero values and another that excludes all the zero's. Through extensive evaluation, I established that the excessive presence of zero values did definitely affect the model's performance, underlining the need of proper data preprocessing.
In the instance of the superstore, I found overfitting concerns while developing the first model. To overcome this obstacle, I used cross-validation techniques, which allowed me to refine the model's performance and produce accurate sales predictions. This stage emphasizes the need of model development and validation, especially when dealing with complicated data.
Overall, my research helps to the knowledge of sales forecasting for SMEs by demonstrating practical ways for dealing with data problems, improving model accuracy, and providing actionable insights. I have demonstrated the effectiveness of these methodologies in improving sales predictions for both medium and small businesses using real-world datasets and advanced techniques, highlighting their potential for guiding informed decision-making and enhancing business performance.

1.5 Research Structure

The paper's structure is divided into the following sections: We have literature review of related past research papers in Section 2 of this study. Section 3 of this paper specifically discusses our data mining technique. Section 4 discusses the implementation of the research project. Section 5 gave a review of the various techniques, with the top ones graded according to their RMSE, MSE, R2 and MAE . Finally, in Section 6, we describe our findings and suggestions for further study.

2 Related Work

In this section, we will look at numerous studies that are helpful in the process of obtaining domain knowledge by comprehending and mastering a range of research methodologies supplied by researchers.

2.1 Sales Forecasting Approaches for Small and Medium Enterprises (SMEs)

The research (Purvika Bajaj1, 2020) displays various merits in its application of Machine Learning algorithms for sales forecasting. It performs a detailed analysis utilizing multiple models such as Linear Regression, K-Neighbours Regressor, XGBoost Regressor, and Random Forest Regressor, demonstrating the adaptability of diverse techniques. The data visualization component gives significant insights into attribute relationships, which aids data comprehension. Impact study of various qualities on sales improves decision-making. The study extensively outlines data pretreatment processes, assuring data quality. Multiple evaluation criteria, such as RMSE and Variance Score, bring rigor to performance assessment. However, several flaws are visible. Reusability of code is lacked due to limitation on particular dataset information. An explanation of algorithm limitations and biases would also help to enhance the paper (Purvika Bajaj1, 2020). A comparison of model strengths and shortcomings might improve comprehension. A comparison with other forecasting methodologies would be helpful in study's findings. Clarity would be improved if the abstract and conclusion were more concise. Finally, considering interpretability and actionable insights would be beneficial. Overall, the work adds to our understanding of sales forecast using Machine Learning, but fixing these flaws would increase its effect.

This paper's (Pavlyshenko, 2019) forecasting methodologies for sales predictive analytics incorporate a variety of machine-learning algorithms and techniques. The authors investigate several time series models, such as Holt-Winters, ARIMA, SARIMA, SARIMAX, GARCH, and others that are often used for sales forecasting. They also highlight the use of machine-learning algorithms, notably tree-based approaches such as Random Forest and Gradient Boosting Machine, which may discover complicated patterns in sales dynamics. Traditional time series methods are regarded less suited for sales prediction than regression-based alternatives. The literature (Pavlyshenko, 2019) study includes prior research on time series forecasting, multi-step forward forecasting, ensemble-based approaches, and merging multiple forecasting techniques. It also emphasizes the limits of traditional time series approaches for sales forecasting and the advantages of regression-based methodologies.

The literature (Elabbasy, 2014) covers the issue faced by pharmaceutical distribution companies (PDCs) in effectively anticipating sales and maintaining inventory levels in order to avoid excessive inventory expenditures and medicine shortages. The suggested system combines network analysis tools and forecasting methodologies for time series. Because each medicine has limited historical sales data, an exploratory network-based analysis is performed to identify clique sets and group members, and their sales data is utilized to forecast. To construct reliable sales prediction models, three forecasting methodologies are used: ARIMA, neural networks, and a hybrid neural network approach. (Elabbasy, 2014) Because of the

limited shelf-life of products and the crucial necessity of product quality for human health, the study underlines the need of exact sales forecasting in the pharmaceutical business. Traditional approaches such as ARIMA are frequently inadequate for dealing with the complexities of pharmaceutical sales forecasting, necessitating the adoption of sophisticated data mining techniques such as neural networks.The suggested hybrid strategy, which incorporates individual medication sales data as well as the records of their group members, outperforms classic ARIMA and single neural network methods. Using real sales data from a renowned PDC in Iran, the study process includes data gathering, exploratory analysis, graph-based analysis, model construction, and assessment.

The study's (Huo, 2021) goal is to examine alternative prediction models for e-commerce sales forecasting using an actual data set from Walmart. The study's core findings include the rapid expansion of China's e-commerce sector, the significance of precise sales forecasting for firms to remain competitive, and the incorporation of calendar and pricing information to improve model performance. The researcher (Huo, 2021) looks at two linear models, three machine learning models, and two deep learning models. Surprisingly, the results demonstrate that basic linear regression models outperform complicated machine learning and deep learning models. The inclusion of date and price information, on the other hand, considerably improves the forecast accuracy for all models. The literature's strengths include using real Walmart data, doing a thorough assessment of prediction algorithms, and emphasizing the significance of calendar and pricing information. However, shortcomings include single-store data, low generalizability, inadequate examination of the influence of external factors, and a lack of model interpretability for business choices. Some research has looked into how customer sentiment analysis and promotional activities affect sales forecasts. For example, (Shih Y S, 2019)] takes into account short-term sales forecasting based on customer sentiment from e-commerce platforms, whereas (Zhuang Q, 2019) adds promotion activities into the GA-BP algorithm model for sales prediction.

The review discusses the influential M-competitions, which began in 1982, as well as recent research ((Casper Solheim Bojer, 2021) (Tao Hong, 2019) (Lloyd, 2014); (Spyros Makridakis, 2022)). While these competitions have offered useful insights into forecasting methodologies, they have not revealed a clear overall advantage of one strategy over another. In many empirical comparisons, ensemble strategies that combine diverse methodologies work well. The review analyzes articles in the field of food demand forecasting, which is analogous to horticulture sales in terms of perishable and seasonal items. (Arunraj, 2014) demonstrated that SARIMAX with holiday and promotional characteristics beat its univariate counterpart SARIMA in forecasting banana sales in a retail shop. (Liu, 2017)tested multiple ML approaches for predicting food sales in a Japanese supermarket chain, with LSTM outperforming Random Forest. A comparison study on bakery sales projections revealed that ML systems outperformed humans (Huber, 2020). Overall, the study of literature emphasizes the difficulty of projecting demand in a variety of fields, such as horticulture sales, food, and tourism. While certain approaches have demonstrated potential in specific settings, there is no one-size-fits-all answer, and ensembles of diverse methods are often beneficial.

A data analytics-based technique to anticipating sales figures in supermarkets and shopping malls is discussed in the (K. VENGATESAN1, 2020) study. Because consumer needs vary regularly, stock management is a priority. To improve overall profitability, the proposed

approach uses 12 months of historical sales data to generate reports that highlight the best months for sales, calculate monthly earnings from various products, identify cities with the highest sales, determine the best time for advertisements, and identify frequently purchased items. The importance of sales forecasting is emphasized in a variety of sectors, including financial forecasting, electric power forecasting, and asset forecasting. (W. HUANG, 2015) Sales forecasting is essential for successful company planning and decision-making, particularly in areas such as vehicle sales, real estate, and conventional enterprises. For historical sales data, traditional approaches such as regression or autoregressive-moving-average (ARMA) models are frequently utilized. Data mining techniques, on the other hand, are increasingly being used to uncover probable patterns and trends in sales data for more accurate forecasts.

(Florian Haselbeck, 2022) looks on sales forecasting for perishable horticulture items. Demand forecasting is critical owing to limited shelf life and external effects such as seasonality and weather. While both classical forecasting (Exponential Smoothing, ARIMA) and machine learning (ML) approaches (Linear Regression, ANN, LSTM, etc.) are utilized for time series forecasting, their relative effectiveness is unclear. The paper tries to close this gap by actually comparing nine ML and three traditional forecasting methods. The dataset depicts horticulture sales issues, and the study investigates multivariate techniques with external elements. Computational efficiency and the capacity to record rapid changes, such as the pandemic's impact, are taken into account. The review covers the study's aims, methodologies, and importance in improving horticulture sales forecast systems. ML algorithms (e.g., XGBoost) routinely beat traditional approaches for horticulture sales forecasts, especially when external factors like as weather are considered. ML models handled rapid demand shifts efficiently during the epidemic. However, there are limitations such as complexity, data reliance, and the possibility of overfitting. The interpretability of certain ML models is limited, and the resource-intensive training of these models raises difficulties.

The difficulty in estimating future demand stems from defining what demand is, as it is influenced by factors such as order times, preferences, and prior sales. Previous sales data is a great place to start since it includes more than simply quantity sold, but also information about changes in customer behavior, including price. (Rosa María Cantón Croda, 2018) significe that it investigates the resilience of ANNs for sales forecasting, showing the possibility for good forecasts despite small sample sizes. The research demonstrates the viability of using ANNs to predict sales trends by analyzing sales data from a chemical firm, showing their independence from data volume and their capacity to offset the constraints
caused by restricted data availability.

	Previous Research Paper Analysis			
Author	Special Features	Results	Benefits	Drawbacks
(Purvika Bajaj1, 2020)	It employs feature significance analysis (e.g., heatmaps and count plots) to determine the important features that have a major impact on sales.	93.53% accuracy is achieved	Detailed Data Pre-processing	There is no mention of external validation or cross-validation.
(Elabbasy, 2014)	Uses Hybrid Strategy for model building	90.568 % accuracy is achieved	Overcame the traditional approach and method	If additional factors had been addressed instead of focusing just on one, greater results may have been obtained.
(Florian Haselbeck, 2022)	Implemented Machine Learning models in Horticulture sales prediction outperforming traditional forecasting technique	Has achieved a RMSE value of 571.11	Clearly explained the difference between all models built	Data Pre-processing is not clearly mentioned

Table 1: Summary of Previous Research Paper Machine Learning Model

2.2 Challenges in Sales Forecasting in Small Medium Enterprises (SMEs)

Small and medium-sized enterprises (SMEs) have distinct problems in sales forecasting due to their limited resources and unpredictable business situations. This article examines the current literature that focuses on these issues, offering insights into the obstacles that SMEs experience when seeking to effectively anticipate their future sales. (K. VENGATESAN1, 2020) highlights and discusses the primary issues that SMEs encounter when anticipating demand. It identifies data scarcity, insufficient experience, and restricted financial resources as major impediments. The strength of this study rests in its identification of major difficulties, which assists SMEs in understanding their pain areas. However, the report does not provide concrete answers or techniques to overcome these difficulties, limiting its practical relevance. (Shih Y S, 2019) digs into the critical problems that SMEs face when it comes to demand forecasting. The difficulties mentioned include many parts of the forecasting process. These problems include data scarcity, in which SMEs may lack previous data for reliable projections, making advanced quantitative methodologies difficult to implement. Another key concern raised is a scarcity of forecasting skills and resources. SMEs, unlike bigger organizations, may lack the financial resources or skilled employees to establish and maintain a comprehensive forecasting system. Furthermore, the article emphasizes the dynamic character of SMEs' business environments, with factors such as seasonality, market trends, and external shocks influencing demand patterns. The study's weakness is that it focuses on identifying challenges without providing concrete techniques or solutions for overcoming these hurdles.

(Lloyd, 2014) provides a broader view by doing a meta-analysis of predicting accuracy inside SMEs. The report seeks to collect insights into the overall forecasting performance of SMEs

by evaluating several research across multiple industries and approaches. The data indicate that forecasting accuracy varies, with some SMEs reaching high accuracy and others struggling. However, the research lacks a full assessment of the causes that contribute to these variances. The problem here is that the meta-analysis presents a generic perspective without diving into the details that would underlie the variations in predicting effectiveness. Furthermore, the report does not give advice on alternative tactics that SMEs may use to enhance their forecasting accuracy.

(Rosa María Cantón Croda, 2018) highlights a variety of issues faced when using neural networks for sales forecasting in constrained data scenarios. The main barrier is the dataset's small size, which only includes a 12-month time period. This restriction makes it difficult to capture nuanced sales trends and linkages, reducing the predictability of projections. Overfitting becomes a major risk since the model may mistakenly learn noise rather than important patterns. Due to the shortage of data, the choice of neural network design and activation functions becomes even more important. Balancing model complexity with available dataset becomes a sensitive undertaking, as too detailed models may limit generalization. Parameter tuning, particularly the learning rate, emerges as a delicate task, requiring careful calibration for convergence and correct predictions. Given the narrow breadth of the dataset, effective model validation is difficult. To reliably assess the model's performance on unseen data, rigorous validation procedures are required. Furthermore, because of the intrinsic model complexity, understanding the neural network's results is difficult. The objective of this paper (W. HUANG, 2015) is to address these problems by bringing insight on the complexities of using neural networks for sales prediction in restricted data circumstances. By overcoming these challenges, the study contributes to the discussion of the usefulness and limits of such approaches in actual forecasting scenarios.

2.3 Conclusion and Justification

Exploration of machine learning approaches for improving sales forecast in small and medium-sized firms (SMEs) gives significant insights into the complexities and constraints of this undertaking. The research articles mentioned give a complete grasp of the application of different machine learning techniques, such as linear regression, ensemble approaches, neural networks, and deep learning, in sales forecasting. The studies highlight the need of precise sales forecasting for effective decision-making and inventory management in a variety of industries, including medicines, e-commerce, and horticulture. The reviewed research papers address the study topic by providing in-depth studies of machine learning algorithms' usefulness in predicting sales for SMEs. They give empirical studies that use real-world data from diverse domains to validate the applicability of machine learning methods. These studies emphasize the issues that SMEs confront, such as insufficient historical data, data unavailability, a lack of forecasting abilities, and unpredictable business settings. Furthermore, the study demonstrates the advantages of using machine learning over traditional methodologies, such as improved prediction accuracy and the ability to detect complex patterns in sales dynamics. While earlier research in the subject of sales forecasting and machine learning has been undertaken, the study issue remains important owing to the ever-changing technological landscape, the unique constraints encountered by SMEs, and the continual search for more accurate and adaptive forecasting approaches. The argument emphasizes the necessity for a concentrated investigation into the implementation of machine learning techniques customized to the demands of SMEs, ultimately leading to enhanced sales forecast and business outcomes.

3 Research Methodology

3.1 Introduction

The research is divided into phases that follow the KDD process approach (Knowledge Discovery in Databases is utilized for various compelling reasons, the most important of which is to harness the abundance of information buried inside enormous databases and translate it into usable knowledge and it is well suited for detection and prediction challenges. The technique is made up of several procedures that are divided into six separate groups. Dataset selection, pre-processing, exploratory data analysis (EDA), transformation, model building and evaluation are all part of the process. The flow of the KDD approach is depicted in the design specification below Fig 1.

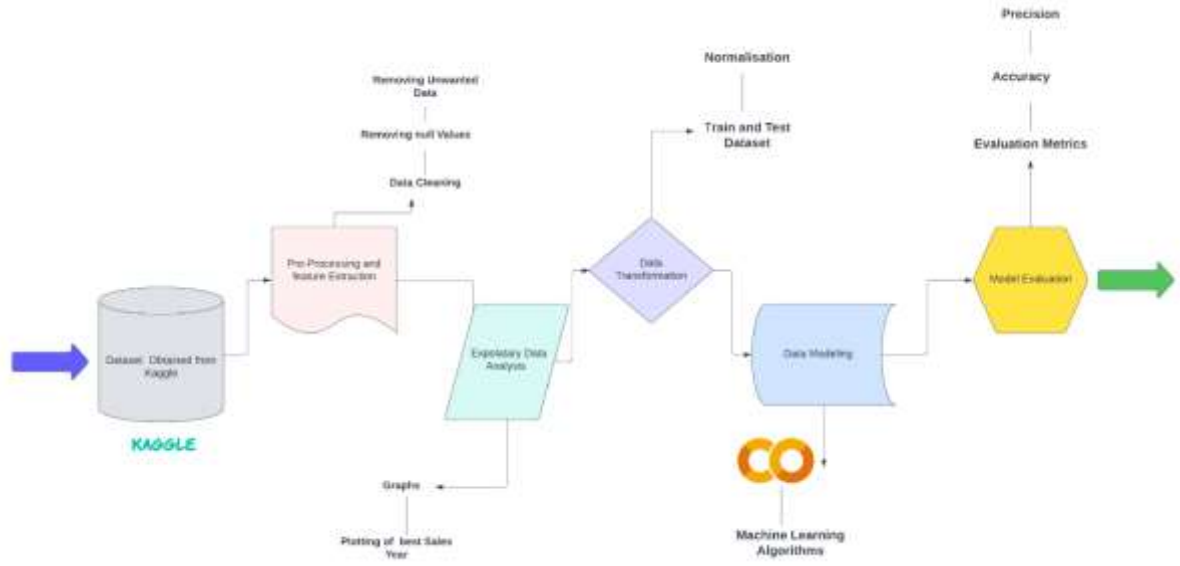


Fig 1:- Modified KDD Methodology

The modified KDD methodology comprises of 6 steps which starts from obtaining dataset from Kaggle and ends at model evaluation. This modified methodology is specifically designed for this research.

3.2 Data Selection

KDD process begins with selecting suitable dataset for the research. At this level, it is critical to understand and build the requisite project-plan expertise. This will help with data selection since a good understanding of the research domain will make it easier to interpret the dataset's columns and finalize a suitable dataset. The dataset for Medium Enterprise which is a pharmacy store was selected from Kaggle that has 1017210 rows and 9 columns and it is divided into two folders one consists of historical sales record while the other has store datasets for which the sale prediction are to be performed. Another dataset which is for the small enterprises of superstore is also taken from Kaggle it comprises of 8524 rows and 12 columns. These data contains 2 files train and test. The datasets that are selected are optimal for performing the required research and it fulfils are the requirements.

3.3 Data preprocessing required for the Research

Preprocessing the data is the second stage in the KDD process. This process is performed to get the best possible results using machine learning models. The data must be cleansed and processed before applying the models so that the computers can better interpret it. Data preparation includes determining if the data is category or numerical, identifying missing values, and identifying outliers. The data will be cleansed when these factors have been determined. The sales column contained 172817 rows with 0 sales. The medium enterprise dataset had 6 null values in the sales_df and in store_df it had 354 in CompetitionOpenSinceMonth,354 in CompetitionOpenSinceYear, 544 in Promo2SinceWeek ,544 in Promo2SinceYear and 544 in PromoInterval respectively so all this null values are cleaned. In the small enterprises dataset the train set has 8523 rows and 12 columns, whereas the test set has 5681 rows and 11 columns. Columns in the train set contain both dependent and independent variables, but columns in the test set contains only independent variables. There are 1463 null values in tem_Weight and 2410 null values in Outlet_Size which will be cleaned and used for further work.

3.4 Selection of Algorithms for the research study

We used relevant literature to discover successful strategies for selecting algorithms for sales prediction in Small and Medium Enterprises (SMEs). Based on their success in similar applications and compliance with the KDD technique, the following algorithms were considered:

Linear Regression: Linear regression is a simple yet effective method that develops a linear relationship between the dependent variable (sales) and the independent variables (predictors). It's especially beneficial for investigating the impact of specific variables on sales and their linear relationships.

Decision Trees: Decision trees are simple and easy-to-understand models for segmenting data based on feature values. They can capture nonlinear correlations and interactions between predictors, making them suited for SMEs with complicated sales patterns. Ensemble approaches, such as Random Forests, can enhance accuracy even more by integrating numerous decision trees.

XGBoost : XGBoost is an enhanced gradient boosting technique that extends the classic gradient boosting framework. It uses regularization techniques to reduce overfitting and handles missing data well, making it a strong competitor for reliable sales prediction.

Random Forest: Random Forest: Random Forest is another ensemble learning technique that creates numerous decision trees and combines their predictions to improve accuracy and decrease overfitting. Each tree is trained on a random part of the data and a random collection of characteristics, resulting in varied trees that collectively generate a robust prediction. (contributors, 2023)

Apart from these there are other several algorithms that can be useful in the research such as lasso and ridge regression or Adaboost as well. The algorithms are chosen based on a past relevant paper and balance of prediction accuracy, model interpretability, computational

efficiency, and fit for the special characteristics of SME sales data. Multiple scenarios are created to test the performance of each algorithm, and hyperparameters were tweaked to maximize their performance inside the KDD framework.

3.5 Evaluation

To evaluate the classification models that were created, a range of assessment measures such as MAE, RMSE, MSE and R2 score are employed. This helps determine the chance that the predicted classes will match with the real classes.

3.5.1 Mean Absolute Error (MAE):

Mean Absolute Error (MAE): The average absolute difference between forecast and actual sales numbers is measured by MAE. It gives a simple grasp of the model's prediction errors in the original data units.

- $MAE = (1/n)\ \Sigma(i=1 \text{ to } n)\ |y\_i - \hat{y}\_i|$

Fig 2 :- MAE Formula (DeepChecks, 2023)

where:
n is the number of observations in the dataset.
y_i is the true value.
ŷ_i is the predicted value.

3.5.2 Root Mean Square Error (RMSE):

Root Mean Square Error (RMSE): The square root of the average squared differences between expected and actual sales is calculated. It penalizes greater mistakes more severely, providing an indication of the total amount of forecast variances. (Frost, Statistics By Jim, 2023)

$$RSME = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{N - P}}$$

Fig 3: RMSE Formula (Frost, 2023)

Where
yi is the true value for the ith observation.
i denotes the anticipated value for the ith observation.
N denotes the number of observations.
P is the total number of parameter estimations, including the constant.

3.5.3 Mean squared error (MSE):

Mean squared error (MSE) quantifies the amount of error in statistical models. It computes the average squared difference between observed and expected values. When a model has no errors, the MSE is 0. As model inaccuracy grows, so does the value. The mean squared error is also known as the mean squared deviation (MSD). (Frost, Mean Squared Error, 2023)

$$MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n}$$

Fig 4 : MSE Formula (Frost, Mean Squared Error, 2023)

Where

yi is the ith observed value.
i is the projected value
n = the number of observations.

3.5.4 R-squared (R2) or coefficient of determination:

The coefficient of determination, or R 2, is a metric that indicates the quality of fit of a model. It is a statistical measure of how well the regression line approximates the real data in the context of regression. It is so critical when a statistical model is employed to forecast future events or to evaluate hypotheses. The one seen here is extensively used. (Grant-Walker, 2023)

$$R^2 = 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}},$$
$$= 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}.$$

Fig 5 :- R2 Formula (Grant-Walker, 2023)

The entire sum of squares equals the square of the data's distance from the mean. Because it is a percentage, it will accept values between 0 and 1.

3.6 Conclusion

The KDD approach was employed in the study, with minor modifications. This approach is applied to the flow of the two-tier project design process. The information comes from the Kaggle repository.

4 Design Specification

The design flow depicted in Figure 6 clearly depicts the flow of this research. To begin, a dataset from the Kaggle data set is retrieved and preprocessed using data cleaning, data transformation, and feature selection. Following that, several subcategories are created to train the dataset. The results are then examined using metrics like R2 score, RMSE, MSE, and MAE. Finally, the findings are presented in tabular form so that they may be easily understood.

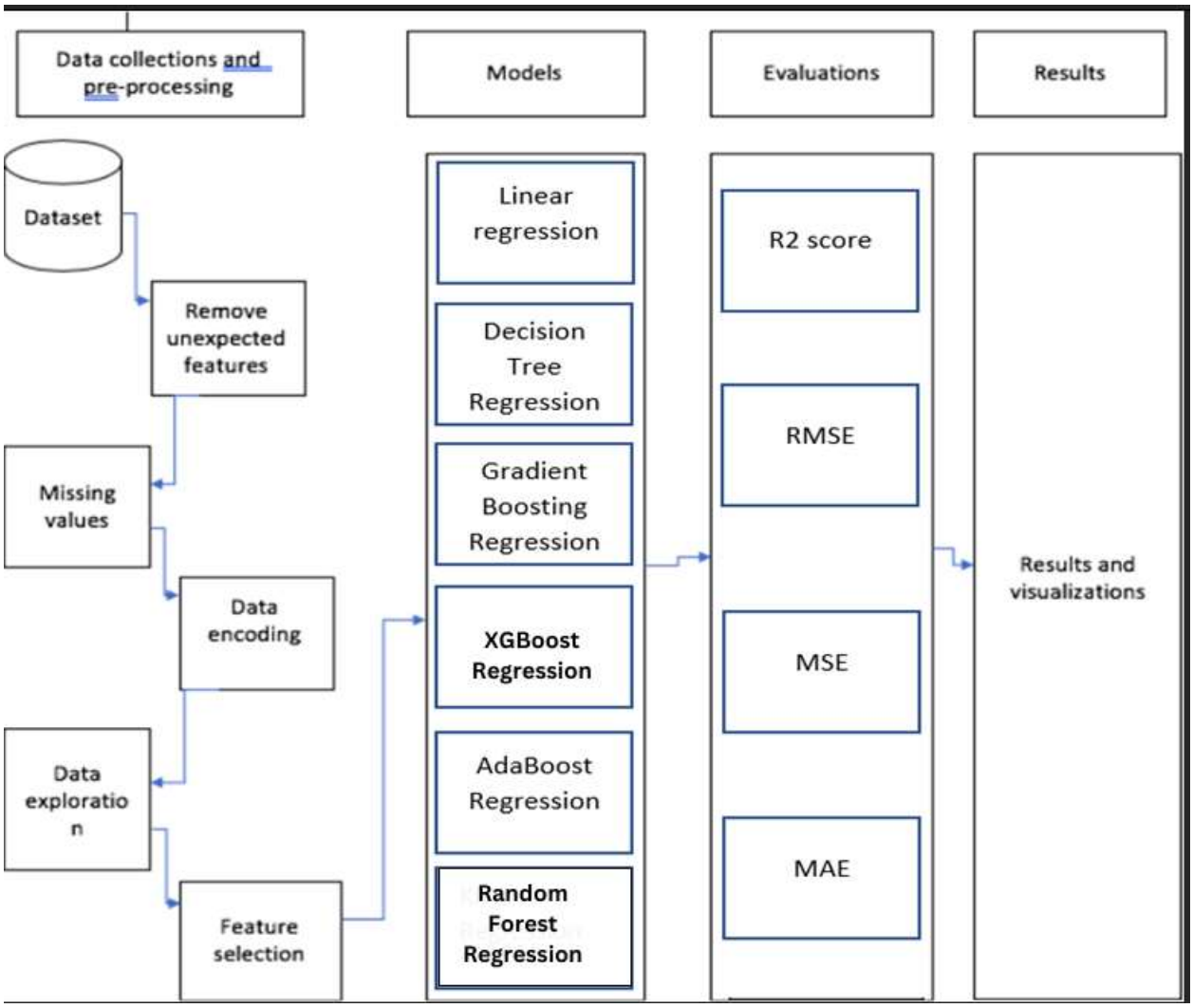


Fig 6 :- Design Flow of the Research

There are no additions of any new machine learning model as whatever models are required is already described in detail in the methodology section. So this is the final design flow that our research will follow.

5 Data Pre-processing, Implementation, Evaluation and Results of Implemented Machine learning models to predict the sales for Small and Medium Enterprises (SMEs)

5.1 Introduction

This component of the research addresses the practical use of machine learning models for sales prediction in Small and Medium Enterprises (SMEs). We present the data pre-processing procedure, which involves cleaning and transforming raw data for analysis. The metrics used to measure model performance are highlighted in the assessment approach. Finally, we provide the obtained findings, providing insights into the accuracy and efficacy of the models used. This section highlights how our research may be used in practice by providing SMEs with helpful tools for forecasting sales trends and optimizing decision-making.

5.2 System Environment setup used for the research

Setting up the environment is an important step in adopting machine learning models for sales prediction in Small and Medium Enterprises (SMEs). The research was done using a 64-bit system having windows operating system and 16 GB of RAM. Python was used for model building because of it's simplicity, powerful libraries, and vibrant ecosystem.. Google Colab is used to run the code on browser. Colaboratory (or "Colab" for short) is a Google Research product. (Google, 2023) Colab is ideal for machine learning, data analysis, and teaching since it allows anybody to create and execute arbitrary Python code using the internet.
During the coding process, the most recent Python release (version 3.10.12) was used.

5.3 Data loading and Pre-processing

This process begins by uploading both the dataset that is downloaded from Kaggle and to start the data processing in Google Colab, we began by uploading both the Kaggle-sourced datasets to Google Drive once it is uploaded Mount your Drive within Colab to access uploaded data. On the left-hand side the datasets uploaded will be visible and can be used for the pre-processing. Using Pandas we load the datasets as they're in CSV format.
Once loaded we preprocess both the dataset of small and medium enterprises. The medium enterprises dataset is huge and sales_df contains 1017208 rows and 9 columns and store_df contains 1116 rows and 10 columns whereas the small enterprises contains 8523 rows and 12 columns, and the test set has 5681 rows and 11 columns. We have transformed the categorical data into numerical as per the requirement and drop few columns which were of no use in the research such as Item_Identifier','Outlet_Identifier','Outlet_Establishment_Year these columns are being dropped from the small enterprise dataset as they don't provide any valuable insights for the research. Once the data pre-processing is finished, we will have clean and usable data to assist us achieve better outcomes and achieve our goal.

5.4 Exploratory Data Analysis (EDA)

After the data preprocessing stage the next step is exploratory data analysis, or EDA, it is utilized to thoroughly explore the data. This analysis entails representing the data in a diagram to make it easier to understand. EDA is used to analyse the data and generate a summary of the most important findings. Finding trends or patterns with its aid is possible thanks to statistical insights and visualizations. The programs seaborn and pyplot, both of which are included in the matplotlib collection, are required to plot the visualizations. In this research we have performed several EDA to understand the dataset in a better way. Such as finding the best business year or month or day of a week. We also checked the sales and competition in month as well as figured out whether holidays affect in business. We performed EDA for the small enterprise as well where we checked the types of items available in the superstore and in which location what type of business would be best.

Fig 7 :- Best Month of Sales of Medium Enterprises

Here in fig 7 we have plotted the graph between the sales and CompetitionOpenSinceMonth which shows us about the sales which is done based on the competitions organised each month and we can see that at the end of year which is in December during Christmas time in winters the sales is highest followed by July as 2nd highest whereas in January, May and October we see that the sales are lowest .

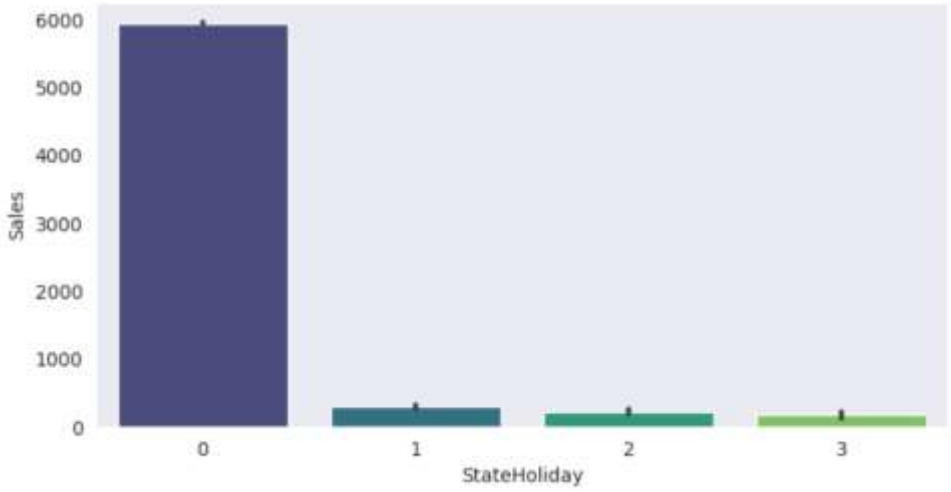


Fig 8:- Sales on Public Holiday is highest of medium dataset

Here in fig 8. We have done the visulation to check the sales during holidays which can help understand the sales based on holidays in a better way so in the fig 8 , 0 denotes Public holidays, 1 = Easter holiday, 2 = Christmas Holidays and 3 = None and we can see that on public holidays which is represented by 0 the sales is highest compared to other holidays.

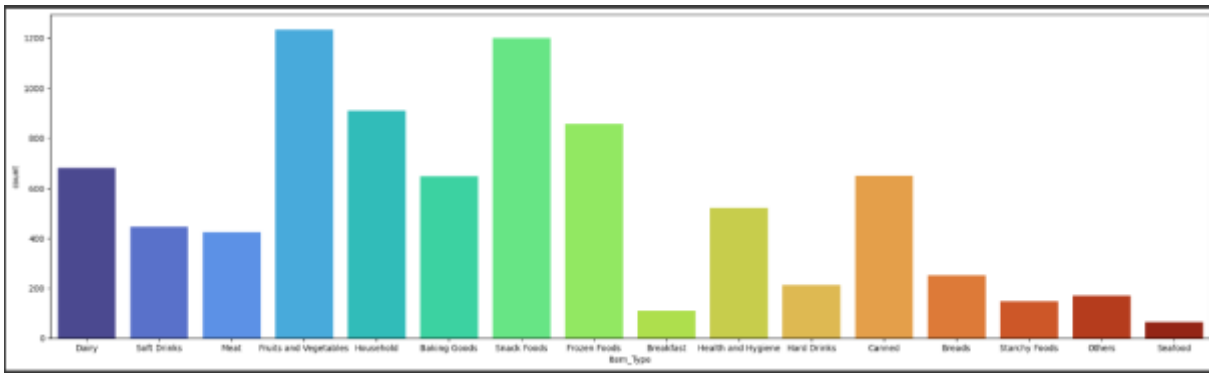


Fig 9: Types of Items Available in the Small Enterprise Dataset

Here in fig 9. We tried to understand what category of item has best sales figure in the supermarket dataset and after visulation we see that fruits and vegetables are highest in numbers followed by snack food , household and frozen food respecitively whereas seafood and breakfast are having lowest count follwed by starchy foods and hard drinks.

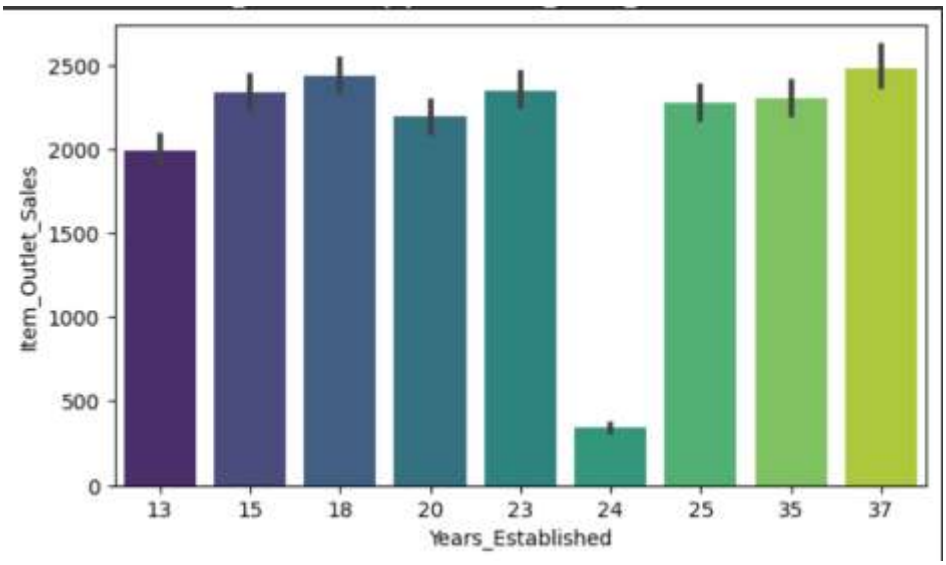


Fig 10 : No. of years of business established for the small enterprise dataset

In Fig 10. We tried to understand that whether years of establishment of shops affects numbers of sales or not ,so after visualization we see that shops that had been in the market for most time which is for 37 years is having better sales compared to other shops but at the same time we also see that a shop which is established 24 years ago has lowest sales , lowest even compared to shop that has been in market just for 13 years so we can say that years of establishment has nothing to do with sales until and unless the items are marketed based on client interest and demands.

5.5 Correlation Matrix

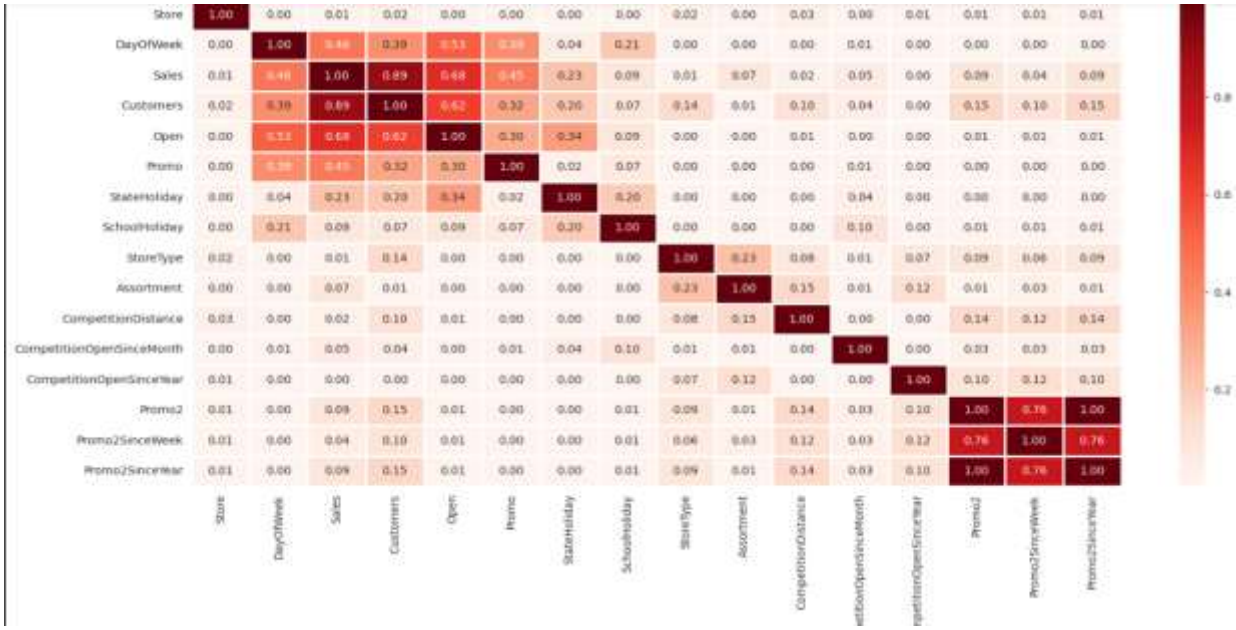


Fig 11 :- Correlation Matrix of Medium Enterprise (Pharmacy Store)

In fig 11, The color of the cell reflects the direction and degree of the connection: a warm color (such as Red) indicates a positive correlation, while a cold color (such as Orange) suggests a negative correlation.The color intensity shows the strength of the association.

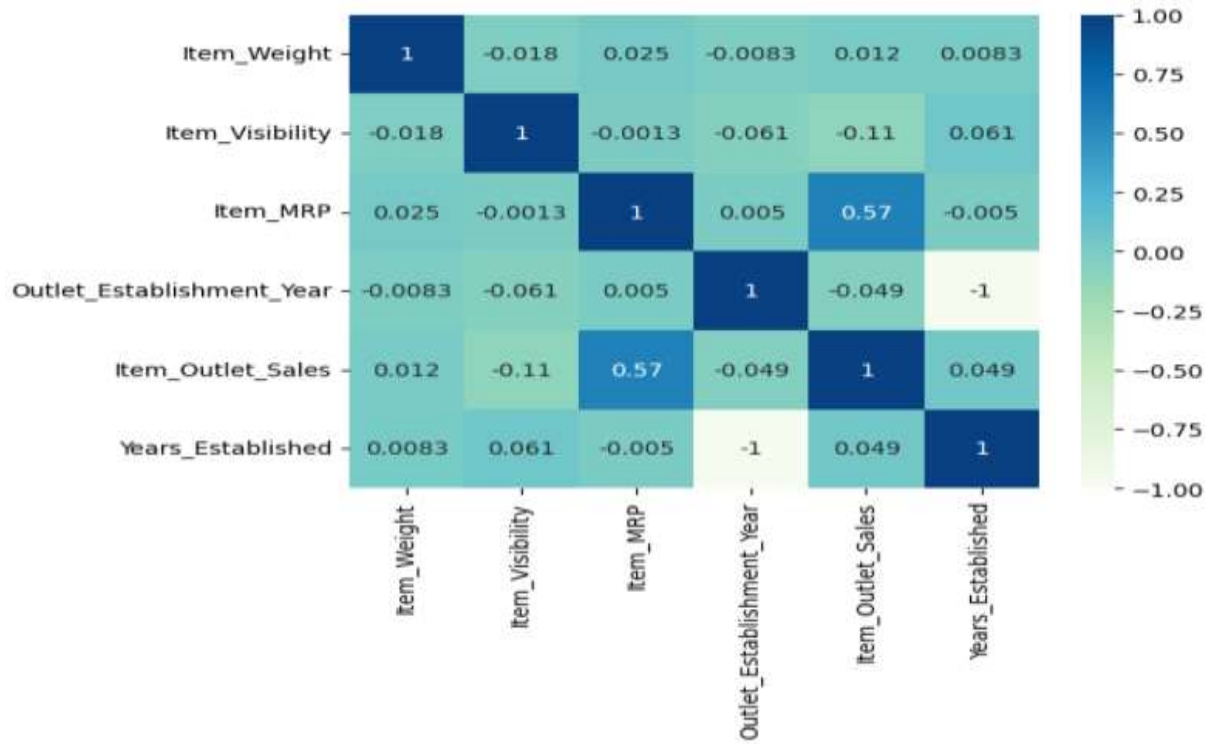


Fig 12: Correlation Matrix of Small Enterprises (Superstore)

In fig 12 , We can in see that Item_Outlet_Sales is significantly associated with Item_MRP, which means that when Item_MRP rises, so does Item_Outlet_Sales.

5.6 Feature Selection

After the completion of the EDA process the datasets of small as well as medium enterprise are merged with their respective common columns, such as in case of medium enterprises the sales_df and store_df are merged on the store column as that was the common column in both the dataset. Whereas in the small dataset the train and test datasets are not merged as both the dataset have same columns only the train dataset has the Item_Outlet_Sales columns which is our dependent variables for which the prediction is being performed. The final dataset is tested for correlation again, the sets sales _df and store_df is put to the final data frame (final1). To decrease mistakes in model development, square root transformation is used on the small enterprise's dataset. Variable skewness is unfavourable for predictive modelling. Some machine learning algorithms presuppose normally distributed data, and a skewed variable can be modified by taking its log, square root, or cube root to get its distribution as near to normal as feasible.
This data is divided into train and test sets, and machine learning models are evaluated using metrics. To give a solution for the research topic, modeling was done individually for each dataset because they are from two distinct businesses and answer two different research questions.

5.7 Evaluation Metrics

In Small and Medium Enterprises (SMEs), evaluating the effectiveness of predictive models for sales forecasting is based on well-defined measures that assess the models. RMSE, MSE, MAE and R2 are the various metrics that helps us decide the accuracy and value of the model built. All the models are assessed with the required metrics and are explained in detail below.

5.8. Experiment 1 :- Implementation, Evaluation and Results on Small Enterprises Dataset

All machine learning models utilized in this study were developed in Python using the sklearn module, which allowed all machine learning models to be imported. The small enterprise dataset consists of two datasets one is train and other is test. These datasets have all the columns similar except of Item_Outlet_Sales which is our dependent variable and rest all the variable are independent. The dataset are combined and then splitted into 80:20 ration for training and testing datasets.

5.8.1 Experiment 1 :- Implementation of Linear regression on Small Enterprise Dataset

In Python, the "sklearn" package is used to build linear regression. LinearRegression is the function that implements it ().We import the libraries required to work with data and develop models. We also create a unique scoring mechanism to assess how well our model predicts. Following that, we divide our data into training and testing halves.
We employ a Linear Regression model to learn from the data and generate predictions. This model attempts to discover patterns in the data in order to forecast values. We assess the model with our own scoring algorithm. This informs us how close the model's predictions are to the actual values. We generate two graphics to help you comprehend the model's predictions. The first is a display of prediction inaccuracy. It assists us in determining where the model's predictions diverge from the actual results. A residuals plot is the second visualization. It displays the amount by which the model's predictions differ from the actual values across the dataset.

Results and Evaluation of Linear Regression on Small Enterprises Dataset

In fig 13 we see the graph between residuals and predicted value , residuals help us analyse the difference between actual value and predicted value, we see that the actual value are closely related with the predicted value as Initially on default we got r2 value of 0.482 which is very less so we did a cross validation and build the model again and predicted the model on test data and we got r2 value of 0.49

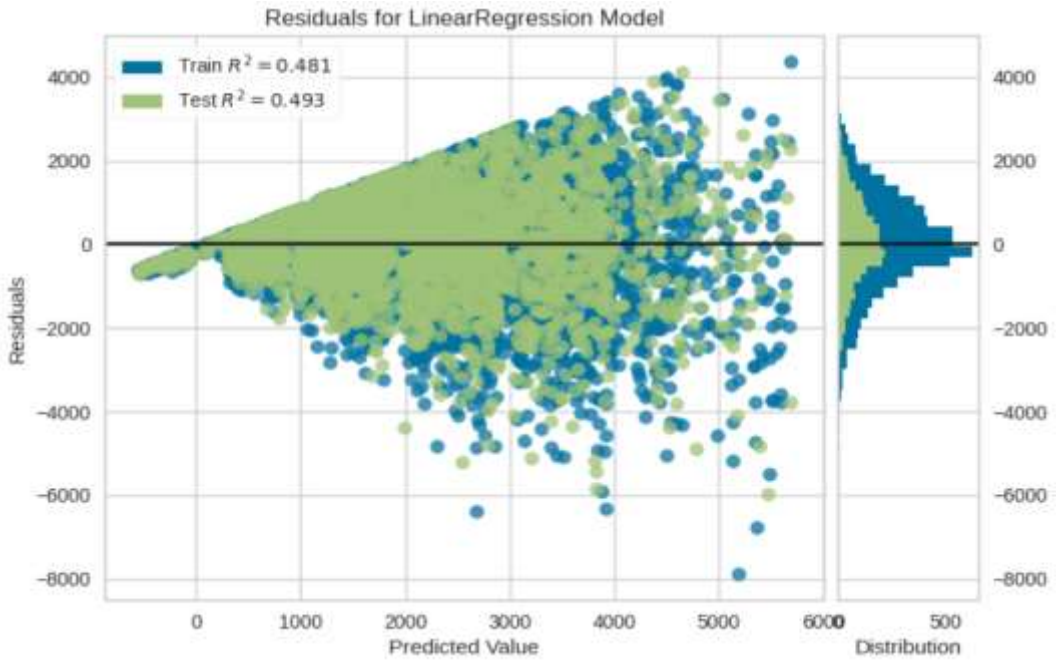


Fig 13 Predicted value using Linear Regression for Small Enterprises

5.8.2 Experiment 2 – Implementation, Evaluation and Results using Lasso Regression on Small Enterprises

Lasso Regression improves feature selection by identifying key drivers of sales and reducing overfitting. This technique leads to the development of strong and interpretable prediction models for the informed decision-making of SMEs. Lasso Regression is a strong feature selection and regularization approach that is very effective in predictive modelling for sales forecasting in Small and Medium Enterprises (SMEs).

Implementation of Lasso Regression on Small Enterprises Dataset

The "sklearn" Python library is used to create lasso regression. We're analysing and making predictions from a dataset using Lasso Regression, a sort of predictive modeling. First, we import the data manipulation and modeling tools and libraries that are required. After that, we divided our dataset into two parts: one for training the model and one for assessing its correctness. We use the Yellowbrick package to investigate various alpha values in order to determine the appropriate regularization strength for the Lasso Regression. This helps to strike a balance between model complexity and data fit. We train the Lasso model to predict outcomes after determining the appropriate alpha. Using cross-validation, we develop a custom scoring mechanism to evaluate the model's correctness. We use this function to assess the model's ability to anticipate fresh data. We create two types of graphs to show the results. The first is a prediction error plot, which shows where the model's predictions differ from the actual values.

The second type of figure is a residuals plot, which shows the disparities between expected and actual values across the dataset. These visualizations give useful information on how well the model predicts and where it might be improved. Overall, this technique assists us in developing and refining a Lasso Regression model that can generate accurate predictions based on our data while also recognizing its strengths and shortcomings.

Results and Evaluation of Lasso Regression on Small Enterprises Dataset

Lasso Regression is a feature selection and regularization technique. Using the normal regression r2 value we got it 0.481 and Manual alpha selection is enabled for fine-tuning the regularization strength to obtain optimal model performance. After Enabling ManualAlphaSelection we got a r2 value of 0.493. We see that in fig 13 and fig 14 there isn't much difference between both the models and they have similar score as well so we'll be moving ahead on building a new model using random forest regression hoping to get better results.

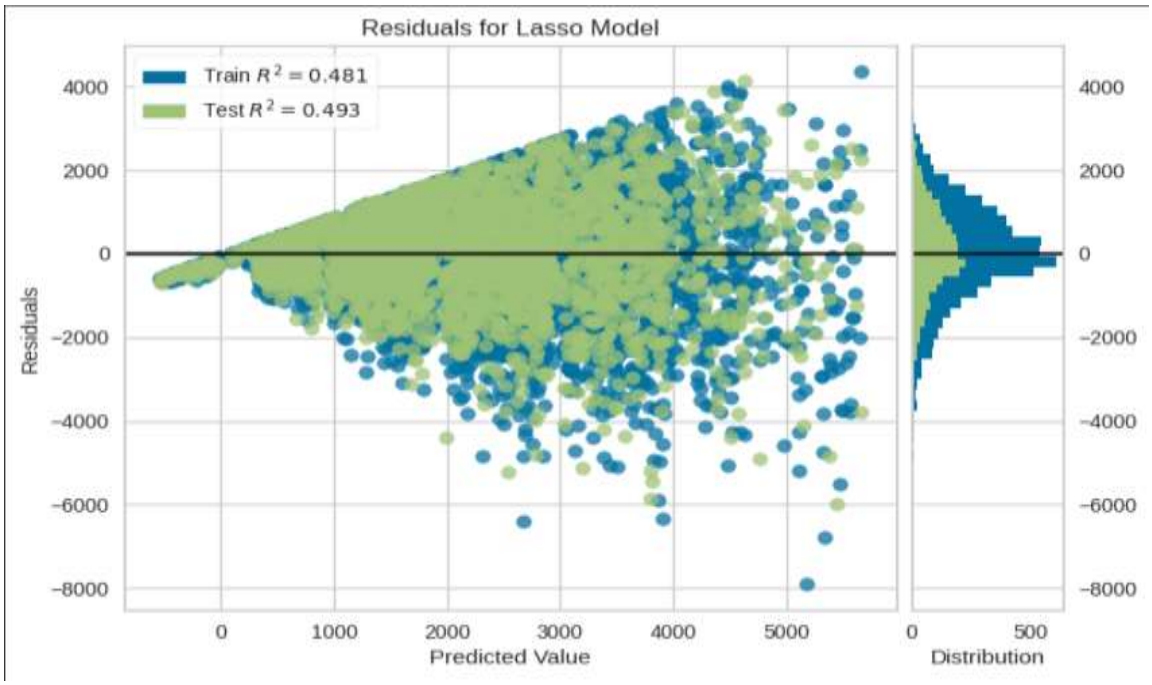


Fig 14: Predicted Value using Lasso Regression

5.8.3 Experiment 3:- Implementation, Evaluation and Results using Random Forest Regression on Small Enterprises

Random Forest is an ensemble learning approach well-known for its resilience and predictive performance in sales forecasting for SMEs. Random Forest is particularly good at capturing complicated connections and preventing overfitting. SMEs receive reliable insights for sales forecasting by applying and assessing this approach, allowing for informed decision-making and strategic planning.

Implementation of Random Forest on small enterprises Dataset

In Python, the "sklearn" package is used to build Random forest regression. Random Forest Regression is the function that implements it (). First, we divide the dataset into training and

testing subsets, which helps us determine how well the model works on unknown data. Using the visualization features of the Yellowbrick library, we investigate the effect of varying max_depth values—basically, the depth of decision trees—on the RandomForestRegressor's performance. We improve the model's forecast accuracy by determining the ideal max_depth. Following that, we train a RandomForestRegressor model with a given max_depth to balance complexity and generality. To assess its prediction ability, we use a proprietary scoring algorithm that computes and displays the average R-squared score and root mean squared error via cross-validation. These measurements show us how well the model's projections match actual values.

Results and Evaluations of Random Forest on small enterprises dataset.

Initially as the model is build using random forest regression we get r2 value of 0.59 and the reason is that the dataset is limited in number and does not have any balance so even after cleaning and using learning curve and cross validation the maximum we got r2 value of 0.61 and it gives us optimal result.

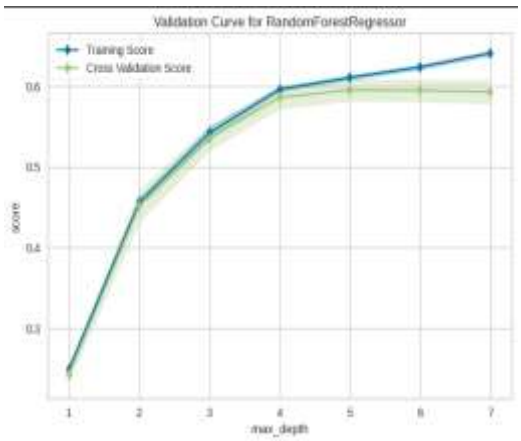


Fig 15 (a) :- Validation Curve

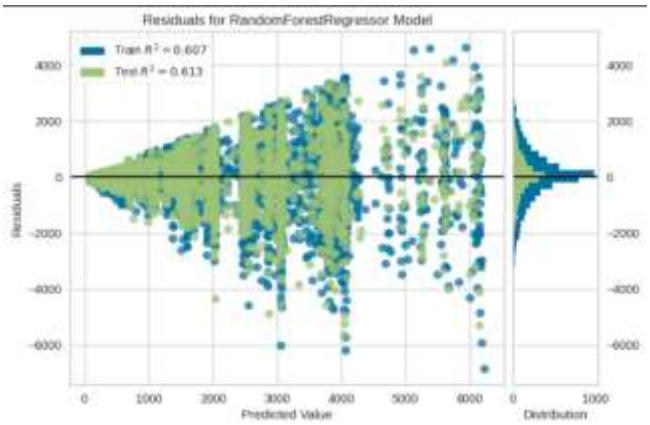


Fig 15(b)Predicted Value using Random Forest

In fig 15 (a) using validation curve we found that the r2 score decreases after depth 5 so we took max_depth of 5 to get the best results for our model which can be seen in fig 15(b) as it shows the difference between residuals and predicted values.

5.9 Experiment 1:- Implementation, Evaluation and Results on Medium Enterprises

All machine learning models utilized in this study were developed in Python using the sklearn module, which allowed all machine learning models to be imported. The dataset are divided into cleandata and cleandata1 in the ration of 80:20 . During the data preprocessing we came across a situation where this dataset has 172817 rows with 0 sales so we will be building two models one including 0 and other excluding them.

5.9.1 Experiment 1 – Implementation, Evaluation and Results of Linear Regression on Dataset excluding 0

Here in this model of linear regression we are not considering the rows which contains 0, so excluding the rows having 0 in them we initially build the linear regression model with

dependent variable *Sales* and independent variables 'Store', 'DayOfWeek', 'Customers', 'Open', 'Promo', 'StateHoliday', 'SchoolHoliday', 'StoreType', 'Assortment', 'CompetitionDistance', 'CompetitionOpenSinceMonth', 'CompetitionOpenSinceYear', 'Promo2', 'Promo2, 'PromoInterval_0', 'PromoInterval_Feb,May,Aug,Nov', 'PromoInterval_Jan,Apr,Jul,Oct', 'PromoInterval_Mar,Jun,Sept,Dec', 'PromoInterval_Mar,Jun,Sept,Dec'. The dataset trained and once the trained is complete the test set is then predicted and evaluated using various evaluation metrics like RMSE, MSE, MAE and R2.
The equation for a linear regression line is Y = a + bX,
where X is the explanatory variable and Y is the dependent variable. The line's slope is b, and the intercept (the value of y when x = 0) is a. (Grant-Walker, 2023)

Implementation of Linear Regression on Medium Enterprises Dataset
In Python, the "sklearn" package is used to build linear regression. LinearRegression is the function that implements it (). In linear regression implementation, I began by identifying the dependent variable as 'Sales' and picking pertinent independent factors from the dataset. These independent factors were used to forecast sales results. I separated the dataset into training and testing sets to make model training and evaluation easier. The training set allowed the model to learn patterns, while the testing set was used to evaluate performance. I evaluated the linear regression model's performance on the training data using the R-squared value, a statistic that reflects the model's capacity to capture data variation. In addition, I investigated the coefficients of independent variables and the intercept of the model to better understand their contributions to prediction. I generated predictions on the test dataset using the trained model and evaluated its accuracy using metrics like as Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). Finally, the R-squared value on the test data revealed how well the model's predictions matched the actual values. Overall, my implementation provided a thorough examination of the model's performance in explaining and forecasting sales depending on selected factors."

Results and Evaluation of Linear Regression
The model is tested on train data and it got a r2 value of 0.7635 whereas it gets a r2 value of 0.7634 when it is tested on test data. It gets MSE value of 2265580 and RMSE value of 1505

	actual	pred
0	5495	5324.460835
1	5472	4849.366393
2	7969	7238.309617
3	7384	6085.335663
4	13212	7828.978735
...	...	...
168874	16337	16398.001796
168875	9195	7062.893253
168876	2938	4098.084042
168877	10413	7772.916183
168878	5828	5368.302742

Fig 16: Actual vs Predicted Value using Linear Regression

In fig 16 , we have compared the actual values with the predicted values using linear regression and we see that the predicted value is mostly less compared to actual value which helps us understand that this model is not giving us the optimal results what we want to achieve so we'll be moving ahead to next model building using lasso regression

5.9.2 Experiment 2 :- Implementation, Evaluation and Results of Lasso Regression

Lasso Regression is a strong feature selection and regularization approach that is very effective in predictive modeling for sales forecasting in Small and Medium Enterprises (SMEs).

Implementation of Lasso Regression on Medium Enterprises Dataset without 0

The "sklearn" Python library is used to create lasso regression. Lasso Regression improves feature selection by identifying key drivers of sales and reducing overfitting. This technique leads to the development of strong and interpretable prediction models for the informed decision-making of SMEs. It is being used to predict outcomes such as sales. I split my data into training and testing sets, trained the model using training data, and then used it to predict test data. I also utilized cross-validation to optimize the model's parameters.

Results and Evaluations on Lasso Regression without 0

GridSearchCV, combined with cross-validation, allows us to systematically search for the ideal alpha value for Lasso Regression, delivering a well-tuned model with increased predictive capabilities for sales forecasting. We have used GridSearch as on default lasso regression we got a r2 value of 0.758 so to improve the r2 we implemented GridSearch combined with cross-validation and now we got a r2 value of 0.76.

	actual	pred
0	5203	6741.076637
1	8590	8148.519988
2	6465	6432.563829
3	7250	6330.804763
4	4339	6047.776042
...	...	...
168874	10096	7916.676674
168875	12400	8283.759472
168876	7499	6232.132109
168877	11606	7564.019185
168878	17389	6007.753939

Fig 17: Actual Vs Predicted value using lasso Regression

In fig 17 we have compared the actual values with the predicted values using lasso regression and we see that the predicted value is mostly less compared to actual value which helps us understand that there are still some growth opportunities such as counting 0 values which can give us better results.

5.9.3 Experiment 3 : Implementation, Evaluation and results on Medium Enterprise dataset using Decision Tree

The decision tree is hierarchical, like a computer engineering tree, with terminals connected by edges. A decision tree organizes information by asking questions at each level. The decision tree uses these trays or qualities to determine whether or not to ask the necessary questions at the appropriate phase to assess the extent to which the prediction/accuracy may be supplied to the individual. Decision trees give interpretable information about the factors that drive sales. By applying and evaluating this technique, SMEs may gain actionable insights to modify strategy and optimize decision-making.

Implementation of Decision tree on Medium Enterprises without 0

The Python package "sklearn" is used to generate the Decision tree. We use Python's scikit-learn module and essential tools to investigate Decision Tree Regression. Using train_test_split, we divide the data into training and testing sets. We train a decision tree with a predefined maximum depth using the DecisionTreeRegressor model. Predictions are produced for both training and testing data. We assess prediction accuracy and alignment with actual values by computing mean_squared_error and r2_score. We then present the RMSE and RMPSE measures, which normalize mistakes and measure performance in relation to typical sales. This technique demonstrates how Decision Tree Regression, driven by scikit-learn, enables us to successfully analyze, forecast, and interpret data trends.

Results and Evaluation of Decision tree on Medium Enterprises

The dataset stored into dataframe before adding it into decision tree it is stored into sales_mean_new dataframe where it is trained and predicted on the test data and We got the MSE value of 2006720 and RMSE value of 1416, RMPSE 0.203 and R2 : 0.79

5.9.4 Experiment 4:- Implementation, Evaluation and Results on Medium Enterprises including 0 using Linear Regression

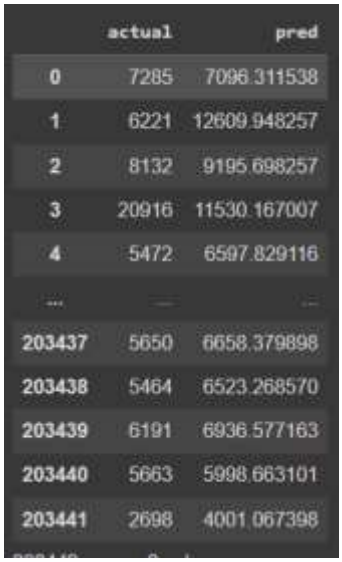
Here we are considering the whole dataset including 0 as well. So a new dataframe is created and it is subdivided into U_train and U_test data. Using Standard scalar the data is fitted into linear regression model. Linear Regression is a key approach for forecasting sales in SMEs based on linear connections between variables and goal sales values.

Implementation of Linear Regression on whole dataset of Medium Enterprises

Linear Regression reveals the direct effect of attributes on sales. Implementing and assessing this model provides SMEs with a comprehensive understanding of the elements that influence sales, allowing for informed decision-making and effective tactics. I've experimented with sales prediction using Python's scikit-learn module and pandas. I divided my data into training and testing sets by creating dependent and independent variables. I achieved consistency by using StandardScaler for feature scaling. I used machine learning to find data associations by using a Linear Regression model. I used measures like RMSE, RMPSE, and R2 to predict sales on the testing set. By visualizing this trip with a DataFrame, I was able to compare real and expected sales. I was able to handle data manipulation, modeling, and assessment with ease because to scikit-learn's capabilities, resulting in insights for informed decision-making.

Results and Evaluation of Linear Regression on Medium Enterprises Dataset

After considering the whole dataset including 0 as well we trained the model again and used standard scalar function to normalize the data. After considering the whole dataset we got a better results and the r2 value is 0.868 which is higher and better than compared to the linear regression model without considering 0 values. MSE value of 1944291, RMSE : 1394 RMPSE : 0.2415 and R2 : 0.8684



	actual	pred
0	7285	7096.311538
1	6221	12609.948257
2	8132	9195.698257
3	20916	11530.167007
4	5472	6597.829116
...	...	...
203437	5650	6658.379898
203438	5464	6523.268570
203439	6191	6936.577163
203440	5663	5998.663101
203441	2698	4001.067398

Fig 18:- Linear regression Actual vs Predicted value by taking whole dataset

In fig 18 , we have comapared the actual values with predicted value and we see that we have positive growth in the predicted values as they are higher than the actual values after considering 0 rows as well which helps us understand that adding 0 values helps us get more data and also increases the values and we got a good r2 score of 0.86 which is higher than the linear regression model without 0 which is shown in fig 16.

5.9.5 Experiment 5 :- Implementation, Evaluation and results of Decision Tree on Medium Enterprises Dataset

The decision tree is hierarchical, like a computer engineering tree, with terminals connected by edges. A decision tree organizes information by asking questions at each level. The decision tree uses these trays or qualities to determine whether or not to ask the necessary questions at the appropriate phase to assess the extent to which the prediction/accuracy may be supplied to the individual. Decision trees give interpretable information about the factors that drive sales. By applying and evaluating this technique, SMEs may gain actionable insights to modify strategy and optimize decision-making.

Implementation of Decision Tree using complete dataset including 0

The Python package "sklearn" is used to generate the Decision tree. We began the development of the decision tree using the entire dataset using scikit-learn and its DecisionTreeRegressor. I

used machine learning to uncover data trends by building a decision tree model with a depth of only 5. I forecasted sales for both the testing and training sets after training the model using standardized data. I assessed the model's accuracy by computing error measures such as MSE, RMSE, RMPSE, and R2. This initiative, made possible by scikit-learn's tools, has yielded useful insights into anticipating sales patterns and evaluating model performance, therefore facilitating informed decision-making.

Results and Evaluation of Decision Tree on Medium Enterprise Dataset including 0

The dataset stored into dataframe before adding it into decision tree it is stored into sales_mean_new dataframe where it is trained and predicted on the test data and We got the MSE value of 1938824 and RMSE value of 1392 , RMPSE 0. 241 and R2 : 0. 868
If we compare this decision tree model to the one which is built using the dataset that did not count 0 values, this decision tree has better results compared to the earlier model.

	actual	pred
0	7285	6405.437098
1	6221	10731.782531
2	8132	9096.412211
3	20916	11835.129880
4	5472	5476.684725
...	...	...
203437	5650	5476.684725
203438	5464	5476.684725
203439	6191	8169.463417
203440	5663	6405.437098
203441	2698	2906.987485

Fig 19: Decision tree Actual Vs Predicted Value on Medium Enterprise

In fig 19, we see the actual values compared to predicted values which we have got using decision tree and we see that in most of the cases the value has increased in the predicted outcome and we get a good r2 score of 0.86 using decision tree including the 0 values as well.

5.9.6 Experiment 6 :- Implementation, Evaluation and results of XGBoost on Medium Enterprises Dataset

XGBoost (Extreme Gradient Boosting) is a sophisticated ensemble learning algorithm that is extensively used for sales prediction in Small and Medium Enterprises (SMEs) because to its predictive accuracy and flexibility. XGBoost excels in capturing complex relationships in data. By implementing and evaluating this model, SMEs gain robust predictive insights, empowering data-driven decisions and strategic planning.

Implementation of XGBoost in Medium Enterprises Dataset

In Python, the "sklearn" package is used to build XGBoost . I've harnessed the power of gradient boosting for exact sales forecast using the xgboost package. I've used machine learning to identify difficult data patterns by using certain model settings such as 500 estimators and a

maximum depth of 8. After training the XGBoost model with standardized data, I projected sales for testing and measured its performance using measures like MSE, RMSE, RMPSE, and R2. This journey, aided by the xgboost toolset, has allowed me to delve into sophisticated boosting methodologies and obtain deeper insights into sales forecasting accuracy, further enhancing my decision-making process.

Results and Evaluations of XGBoost on Medium Enterprises Dataset

Upon considering the whole dataset and using XGBoost model we got a very good score where our RMSE value is 421 and as we know the lower the RMSE value is the better is the model build. So we got MSE value of 177443 and RMPSE value of 0. .072 and R2 value of 0.987. Upon comparing all the models previously built we get the best r2 score of 0.987 using XGBoost.

Results :-

No.	Model for Small Enterprise Dataset	Average RMSE	R2
1	Linear Regression	-1225.6	0.48
2	Lasso Regression	-1225.8	0.48
3	Random Forest Regression	-1082	0.61

Table 2: Comparison of Model Build for Small Enterprises

No.	Model for Medium Enterprise Dataset	RMSE	MSE	R2
1	Linear Regression (Excluding 0)	1505	2265580	0.76
2	Lasso Regression (Excluding 0)	1508	2267788	0.76
3	Decision Tree (Excluding 0)	1416	2006720	0.79
4	Linear Regression	1394	1944291	0.86
5	Decision Tree	1392	1938824	0.86
6	XGBoost	421	177443	0.98

Table 3 : Comparison of Model build for Medium Enterprises

5.10 Conclusions for Implementation, Results and Evaluation

All the machine learning models are implemented and their results are shown in the table 2 and 3 for Small and Medium Enterprises respectively. We can see that small enterprises results are very low compared to results achieved by medium enterprises. In small enterprises we got highest r2 value of 0.61 using random forest. The reason is that the small enterprises dataset is relatively small in size as compared to medium enterprises, so we could say that data size does affect in the prediction of the model on the other hand in medium enterprises we could easily see that XGBoost has out performed all other models and has a R2 value of 0.98. Rest all the results are mentioned in Table 2 & 3 and Table 1 has all the previous research papers, if compared to previous papers our XGBoost has highest accuracy among all previous papers.

6 Discussion & Conclusion

This study intended to answer the fundamental research question: "How can machine learning techniques be effectively leveraged to improve sales prediction for small and medium enterprises?" The research concluded that this endeavour of sales prediction for small and medium establishments (SMEs) has verified the importance of data-pushed insights in enhancing choice-making and operational performance. By leveraging diverse regression algorithms together with Linear Regression, Decision Trees, and XGBoost, we have correctly developed predictive fashions that capture the relationships inside the dataset, allowing us to forecast sales with good accuracy. Through the usage of key metrics which includes RMSE, RMPSE, and R2, we have quantified the predictive performance of these models, providing valuable benchmarks for evaluating the future forecasts. The study technique includes working with two unique datasets supplied from Kaggle, representing a pharmacy shop (medium enterprise) and a superstore (small enterprise). Machine learning models were built and rigorously tested using diligent preprocessing, addressing missing information, and picking relevant features. Notably, the study revealed the need of controlling excessive zero values in data, since their presence harmed predictive model accuracy. Furthermore, while the superstore dataset initially exhibited overfitting concerns, the use of cross-validation effectively resolved this issue, resulting in better model efficacy.

The findings of this study demonstrate its success in answering the research question and aims. The pharmacy store and superstore datasets both produced models with impressive accuracy in sales prediction. Medium Enterprise having large amount of data using that XGBoost has outperformed all other models and has r2 score of 0.98 , while for small enterprises even though having less amount of data random forest gave the best r2 score of 0.61 among all the models built. This study provides practical insights for firms aiming to improve inventory management, resource allocation, and promotional methods by demonstrating the usefulness of machine learning approaches in boosting sales forecasting for SMEs. However, the study also acknowledged its shortcomings. The study was limited to two datasets, which may restrict the findings' generalizability. Furthermore, improving the interpretability of the models and diving further into the effect of certain qualities on sales forecast accuracy are areas for future research.

Future Work

In terms of future work, there are several interesting pathways for expanding the effect of this research. Using datasets that have enough information and give important features that could enhance the prediction accuracy we could obtain higher results for small enterprises as well only if data would be more. Furthermore integrating emerging technologies like AI can further empower SMEs in accurate sales forecasting. Commercially, the findings of this study can be converted into practical applications, perhaps leading to the development of sales forecast systems designed specifically for SMEs. Such technologies might help organizations make better decisions, reduce costs, and optimize their operating procedures.

7 References

Arunraj, N. &. (2014). Time series sales forecasting to reduce food waste in retail industry. *Research Gate*.

Arunraj, N. S. (2015). A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting. *International Journal of Production Economics*, 321–335.

Casper Solheim Bojer, J. P. (2021). Kaggle forecasting competitions: An overlooked learning opportunity. *International Journal of Forecasting*, 587-603.

contributors, W. (2023). Random forest. *Wikipedia, The Free Encyclopedia*.

DeepChecks. (2023). Mean Absolute Error. *DeepChecks*, 1. Retrieved from https://deepchecks.com/glossary/mean-absolute-error/

Elabbasy, E. K. (2014). Intelligent Sales Prediction for Pharmaceutical Distribution Companies: A Data Mining Based Approach. *Hindawi Publishing Corporation*.

Florian Haselbeck, J. K. (2022). Machine Learning Outperforms Classical Forecasting on Horticultural Sales Prediction. *Machine Learning with Applications*.

Frost, J. (2023). Mean Squared Error. *Statistics By Jim Making statistics intuitive*, 1. Retrieved from https://statisticsbyjim.com/regression/mean-squared-error-mse/

Frost, J. (2023). Statistics By Jim. *Statistics By Jim Making statistics intuitive*, 1. Retrieved from https://statisticsbyjim.com/regression/root-mean-square-error-rmse/

Google. (2023). Machine Learning Crash Course . *Google Developers Course* .

Grant-Walker, A. (2023). Coefficient of Determination(R-squared). *Numeracy, Maths and Statistics - Academic Skill Set*, 1. Retrieved from https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/coefficient-of-determination-r-squared.html

Huber, J. &. (2020). Daily retail demand forecasting using machine learning with emphasis on calendric special days. *International Journal of Forecasting*, 1420-1438.

Huo, Z. (2021). Sales Prediction based on Machine Learning . *International Conference on E-Commerce and Internet Technology (ECIT)*.

K. VENGATESAN1, E. V. (2020). AN APPROACH OF SALES PREDICTION SYSTEM OF CUSTOMERS USING DATA ANALYTICS TECHNIQUES. *Advances in Mathematics: Scientific Journal 9 (2020)*, 5049-5056.

Liu, X. &. (2017). Food sales prediction with meteorological data — A case study of a Japanese chain supermarket. *Data Mining and Big Data:Second International Conference, DMBD 2017, Fukuoka, Japan*, 93-104.

Lloyd, J. R. (2014). GEFCom2012 hierarchical load forecasting: Gradient boosting machines and Gaussian processes. *International Journal of Forecasting*, 369-374.

Pavlyshenko, B. M. (2019). Machine-Learning Models for Sales Time Series Forecasting. *Data Stream Mining & Processing (DSMP)*.

Purvika Bajaj1, R. R. (2020). SALES PREDICTION USING MACHINE LEARNING ALGORITHMS . *International Research Journal of Engineering and Technology (IRJET)*, 1-7.

Rosa María Cantón Croda, D. E. (2018). Sales Prediction through Neural Networks for a Small Dataset. *Deanship of Graduate Studies in Engineering and Businesses*.

Shih Y S, L. M. (2019). A LSTM Approach for Sales Forecasting of Goods with Short-Term Demands in E-Commerce. *Asian Conference on Intelligent Information and Database Systems. Springer*, 244-256.

Spyros Makridakis, E. S. (2022). The M5 competition: Background, organization, and implementation. *International Journal of Forecasting*, 1325-1336.

Tao Hong, J. X. (2019). Global energy forecasting competition 2017: Hierarchical probabilistic load forecasting. *International Journal of Forecasting*, 1389-1399.

W. HUANG, Q. Z. (2015). A Novel Trigger Model for Sales Prediction with Data Mining Techniques. *Data Science Journal,*, 15.

Zhuang Q, Z. X. (2019). A Neural Network Model for China B2C E-Commerce Sales Forecast Based on Promotional Factors and Historical Data. *International Conference on Economic Management and Model Engineering (ICEMME). IEEE*, 307-312.