Abstract:

Abstract

Predictive sales analysis based on previous data is crucial for organizations to make educated decisions and remain competitive. Machine learning is a powerful technology that can automate this process, producing more accurate and informed forecasts. Machine learning has revolutionized many sectors, including sales and marketing. Machine learning algorithms can forecast consumer behaviour and sales trends by analyzing data and discovering patterns, hidden patterns, and linkages. The purpose of this research is to propose an understanding of the usage of machine learning algorithms for predicting future sales of Big Mart enterprises based on past year sales. Utilizing machine learning methods like Linear Regression and Gradient Boost, a thorough study of sales forecasting is undertaken. The performance of the Linear Regression and Gradient Boosting methods was evaluated using metrics such as mean absolute error, mean squared error, R2 score, and Accuracy. The study's findings can help businesses make better informed decisions about resource allocation, production planning, and marketing methods. Overall, machine learning is a powerful tool for forecasting sales and can assist organizations in staying ahead of the curve in a fast changing industry.

……..

Introduction:-

- Sales Prediction using Machine Learning is a popular application of Artificial Intelligence in business analytics.
- The goal is to forecast future sales based on historical data and other relevant factors. Accurate sales forecasts can help companies optimize resources, reduce waste, and increase profits.
- Regression, time series analysis, neural networks, and decision trees are some of the machine learning techniques used for sales prediction.
- Selecting which attributes to include in the model can be tough. Machine learning models require massive amounts of high-quality data to be adequately trained.
- A statistical method for predicting, in this case, sales, is linear regression. It models the relationship between a dependent variable (sales) and one or more independent variables.
- Itcan identify the relationship between historical sales data and various factors that may influence sales. Gradient
- Boosting is a powerful approach in machine learning for sales prediction. It combines multiple weak learning models into a strong predictive model.
- It can identify complex interactions between different factors. It is highly accurate and can handle non-linear relationships and missing data.

Problem Statement:-

- The challenges of estimating sales based on past data and machine learning.The goal is to create a model that can forecast future sales based on a range of factors such as previous sales data, market trends, economic indicators, and customer demographics. It is vital to analyze historical sales data, seek for connections and patterns, and employ machine learning algorithms when projecting future sales. A reliable, accurate, and scalable sales prediction system is being developed to help firms maximize their sales strategy and figure out how much each item sells in a certain location.

Literature Survey:-

A. Profit Prediction using ARIMA, SARIMA and LSTM Models in Time Series Forecasting: A Comparison - Uppala Meena Sirisha, Manjula C. Belavagi,Girija Attigeri(2022, IEEE)

This study explains how to predict and perform profit analysis of sales using Time series models in Machine Learning. Time series models like ARIMA, SARIMA and LSTM models were used. A comparative analysis of 3 models was done.

B. PromotionLens: Inspecting Promotion Strategies of Online E-commerce via Visual Analytics - Chenyang Zhang, Xiyuan Wang, Tianyu Zhang, Xiaojuan Ma,QuanLi(2022, IEEE)

This study discusses about multivariant time-series forecasting models and well-designed visualizations to demonstrate the impact of sales and promotional factors.

C. Predictive Analysis for Big Mart Sales using Machine Learning Algorithms - Ranjitha P, Spandana M(2021, IEEE)

This study explains about sales prediction in BigMart Stores to update inventory management. For the purpose of predicting the sales of BigMart stores, a predictive model was created utilising the Xgboost, Linear regression, Polynomial regression, and Ridge regression approaches.

D. Machine Learning Model for Sales Forecasting by using XGBoost - Xie Dairu, Zhang Shilong(2021, IEEE)

This paper explains how eXtreme Gradient Boosting (XGBoost) is used for efficient forecasting of the future sales amount in modern retail corporations.

E. Sales Prediction based on Machine Learning - Zixuan Huo(2021, IEEE)

This study discusses about sales prediction for e-commerce companies using 2 linear models, 3 machine learning models, and 2 deep learning models and determining which model can predict sales accurately.

F. Sales Forecast of Manufacturing Companies using Machine Learning navigating the Pandemic like COVID-19 – Prabhat Sharma, Shreyansh Khater,Vasudha Vashisht(2021, IEEE)

This study discusses about prediction of turnover of the automobile industry in the era of COVID-19 using various machine learning models.

G. Prediction Analysis Sales for Corporate Services Telecommunications Company using Gradient Boost Algorithm - Oryza Wisesa, Andi Adriansyah ,Osamah Ibrahim Khalaf(2020, IEEE)

This study discusses about analysis ofthe reliability of B2B sales using Gradient Boost algorithm. It discusses how Telecommunication Company increase its market growth.

H. A Hybrid Machine Learning model for Sales Prediction - Jingru Wang(2020, IEEE)

This study explains about the LightGBM framework and the XGBoost framework which is used to build a sales prediction model. Weights are assigned based on the prediction results of these two models and model integration is performed later.

I. A Stock Price Prediction method based on Deep Learning technology - Xuan Ji,Jiachen Wang, Zhijun Yan(2020, IEEE)

This study proposes a prediction model based on deep learning technology called Long Short Term Memory model(LSTM) which uses stock financial index variables as its inputs to predict sales of stocks.

J. Forecasting Promotional Sales within the Neighbourhood – Carlos Aguilar, Sergio Romero, José Luis Rojo-Álvarez (2019, IEEE)

This study discusses about machine learning methods for automatic prediction of promotional sales in real-market applications. It explains about fully automated weighted k-nearest neighbors which were used in sales prediction

Existing System With Drawbacks:-
Arima

- ARIMA (Auto Regressive Integrated Moving Average) is a time series algorithm used in machine learning. ARIMA assumes that the input time series is stationary, meaning that its statistical properties (mean, variance, etc.) remain constant over time.

- ARIMA is a simple model that does not account for complex inter-variable connections or the impacts of exogenous variables. As a result, it may be unsuitable for modelling complex data. ARIMA can be an effective time series forecasting model, but its accuracy may be restricted by issues such as non-stationarity, model parameter misspecification, outliers, limited model complexity, and lack of interpretability.

Sarima:

- SARIMA (Seasonal Auto Regressive Integrated Moving Average) is a time series algorithm used in machine learning. The differencing order, the order of the moving average and autoregressive components, and the seasonality period are among the hyperparameters that must be chosen for SARIMA models.

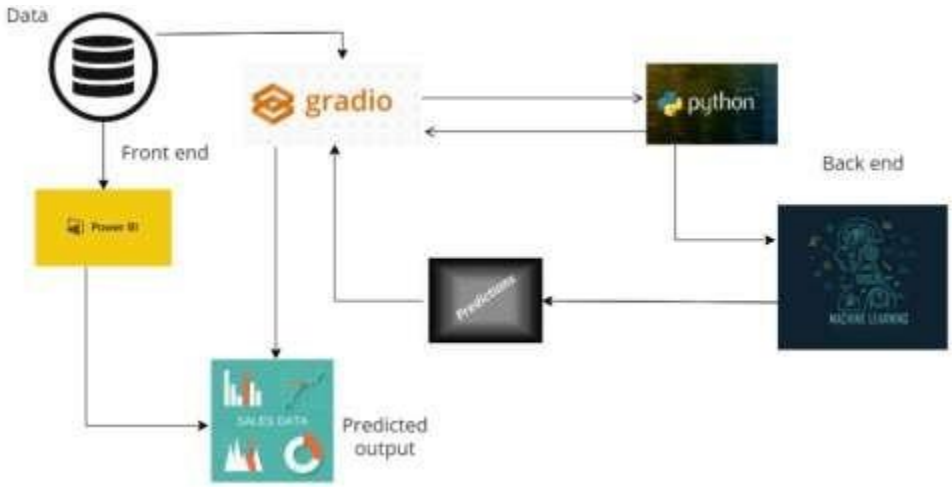
- The accuracy of the model depends on selecting the appropriate hyperparameters, which can be time-consuming and challenging. Time series predictions made by SARIMA models are susceptible to overfitting, which occurs when the model fits the training data too closely and fails to generalize to new, previously unseen data. This can lead to poor forecasting accuracy.

Architecture:-

- The data in form of user inputs is given in frontend(Gradio) and user's input request is sent via web service to backend to fetch necessary sales predicted data.

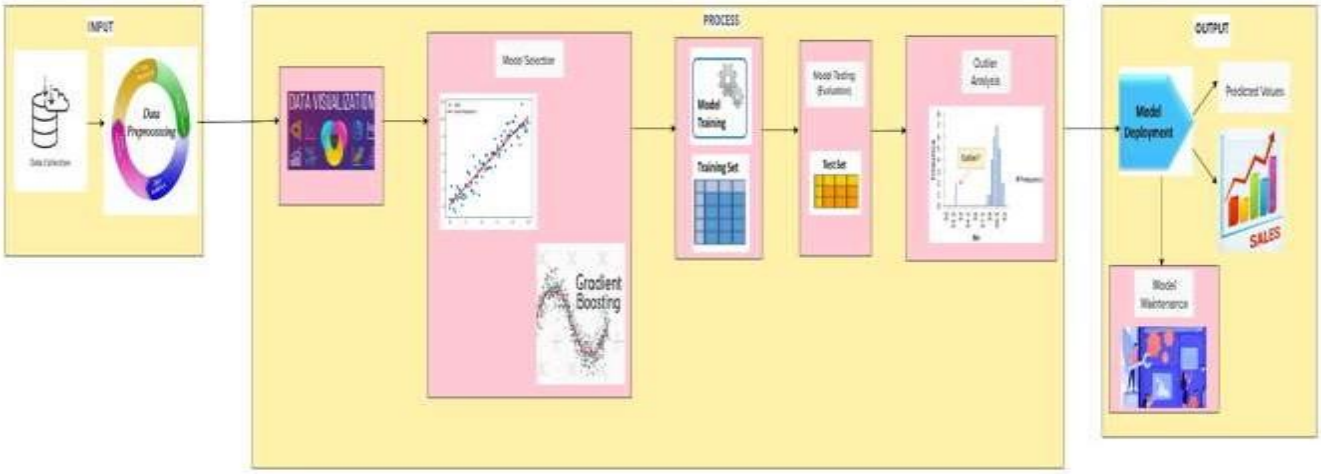
- The Sales prediction is done using Machine Learnings models using Python and the predicted data is sent as a response back to user via frontend.

- The user perceives the sales predicted data. The user loads the dataset to the Data source tool(Power BI) to get the predicted output in form of graphs.



Project Architecture.

The initial step is to collect and pre-process data. In the second stage, the data is visualized and selection of model is done. Later on Model training, testing and outlier analysis is done. In the third stage, the deployment of model is done with its results of output graphs and sales predicted values. Later on model maintenance is done.



System Architecture.

Proposed Solution:-

Linear Regression

• A supervised machine learning technique called linear regression models a dependent variable (commonly referred to as the target variable) and one or more independent variables (also known as predictors or features). It presupposes that the dependent and independent variables have a linear relationship.

• The linear regression procedure identifies the line that best matches the connection between the variables by minimizing the sum of the squared errors between the predicted and actual values.

• The target variable can be predicted for new data points using the coefficients of the line that best fits the data, which are calculated by the algorithm.

• Both simple and multiple regression analyses can be performed using linear regression. While multiple linear

regression involves two or more independent variables, simple linear regression simply requires one.

• The widely used algorithm of linear regression is utilized for a variety of tasks, including forecasting stock prices, sales volumes, and home price indices. It is a straightforward but effective method for creating predictions since it explains the link between the variables clearly.

Gradient Boosting

• A machine learning approach called gradient boosting is applied to classification and regression issues. Gradient Boosting is known for its ability to handle complex datasets with high-dimensional features and non-linear relationships.
• It is also robust to outliers and missing values, as it uses the median to handle missing values and can handle outliers through the use of robust loss functions.
• It is a type of ensemble learning algorithm that combines multiple weak learners to create a stronger predictor
• Gradient Boosting has three main components:

Loss Function- The purpose of the loss function is to calculate how well the model predicts given the available data. We can determine the discrepancy between the expected and observed weights with the use of the loss function.

Weak Learner - A weak learner is one that classifies the data but does so poorly, perhaps no better than random guessing. In other words, it has a high error rate. These are typically decision trees which are built sequentially, each subsequent tree building on the errors of the previous one which can be improved later.

Additive Model - This is the iterative and sequential approach of adding weak learners one step at a time. After each iteration, it needs to be closer to the final model. The loss function's value should decrease with each repetition.

• The algorithm works by iteratively training decision trees to predict the errors of the previous trees. In each iteration, the algorithm calculates the negative gradient of the loss function, which measures the difference between the predicted and actual values. The negative gradient is then used to update the weights of the samples that were misclassified by the previous tree. The updated weights are used to train the next decision tree, which focuses on correcting the errors of the previous tree. This process is repeated until the specified number of trees is reached. The final prediction is the weighted sum of the predictions of all the trees in the ensemble.

Methodology:-

Data Collection

• Data gathering is the initial stage of the machine learning life cycle. As machines initially learn from the data that is given to them so it is necessary to collect reliable data so that machine learning model can find the correct patterns.
• The quality and quantity of data collected impacts the accuracy of predictions. It is crucial to identify reliable data sources to collect accurate data. Outdated or incorrect data can lead to inaccurate predictions.

Data Pre-Processing:

• It is the step that requires the most time and labour. In this step one explores, pre-processes, conditions, and transforms data prior to modelling and analysis.
• This step can be further divided into 2 processes:

Data Exploration:

It is used to understand the nature, characteristics, format, and quality of data.

Data Wrangling:

Cleaning and transforming raw data into a usable format is known as Data Wrangling. It also involves putting together all the data and randomizing it.

Data cleaning is required to address the quality issues. Cleaning the data is done to remove unwanted data, missing values, rows and columns, duplicate values, data type conversion, etc.

Data Visualization

• Exploratory Data Analysis (EDA) is performed via Data Visualization. When working with vast amounts of data, a visual representation makes interpreting the data much simpler.

• Data patterns and trends can be found via visualization, as can the relationships between different variables and classes. Data Visualization also helps for faster decision making through the visual representation.

Model Selection

• It is crucial to select a machine learning model that is appropriate for the task at hand since it governs how a machine learning algorithm applied to the data collected produces its output. Determining whether the model works best with numerical or categorical data is also necessary.
• The best machine learning model generates predictions using a machine learning algorithm using input data.

Model Training

• Training is the stage of machine learning that is most important. To detect patterns and create predictions, we feed the prepared data to the machine learning model during training.
• The model then learns from the data so that it can accomplish the task set. Over time, the model becomes more accurate in predicting.
• The model is being trained to perform better in order to solve problems more effectively.
• Datasets are used to train the model using machine learning algorithms. It involves splitting the cleaned data into a training set. The set that the model learns from is called the training set.

Model Evaluation (Model Testing)

• After training the model, we have to check to see how the model performs. This is accomplished by evaluating the model's performance using test data that has never been seen before.
• When applied to test data, the model's accuracy and speed can be determined. It involves splitting the cleaned data into a testing set. A testing set is used to check the accuracy of the model after training.

Outlier Analysis

• The Outlier analysis stage of machine learning (ML) and data analysis is critical. Outliers are data points that deviate noticeably from the other data points in a dataset and can skew the analysis's final conclusions.
• Measurement flaws, data gathering problems, and other types of data noise can all lead to outliers.
• Outliers must be properly identified and handled because they can significantly affect the precision and dependability of machine learning models.
• Outlier analysis can reduce overfitting and increase the precision of predictive models.
• Visual inspection, clustering, and density-based methods are just a few statistical and machine learning techniques that can be used to identify outliers.

Model Deployment (Making Predictions)

• Deploying the model in a real-world system is a crucial stage of the machine learning life cycle.
• Some of the important factors that should be tested and evaluated before deploying a model are robustness, compatibility and scalability.
• Making Predictions: In the end, we can make precise predictions using the model on unobserved data.

Model Maintenance

Machine learning models require ongoing maintenance to ensure that they continue to perform well over time.
Some common tasks involved in maintaining machine learning models:
• Data drift correction: Making updates to the model to account for changes in the distribution of the input data.
• Regular retraining: Retraining the model on updated data to reflect changes in the real-world distribution of the input data.
• Model interpretation: Interpreting the model's predictions and understanding its decision-making process, to ensure that the model is making predictions that are aligned with the intended application.
• Parameter Tuning: After model development, training and evaluation it is necessary to improve its accuracy. The model's parameters are tuned to achieve this.

At a particular value of parameter, the accuracy will be the maximum. Parameter tuning refers to finding these values and maintaining it without getting decreased to maintain accuracy.

Significance:-

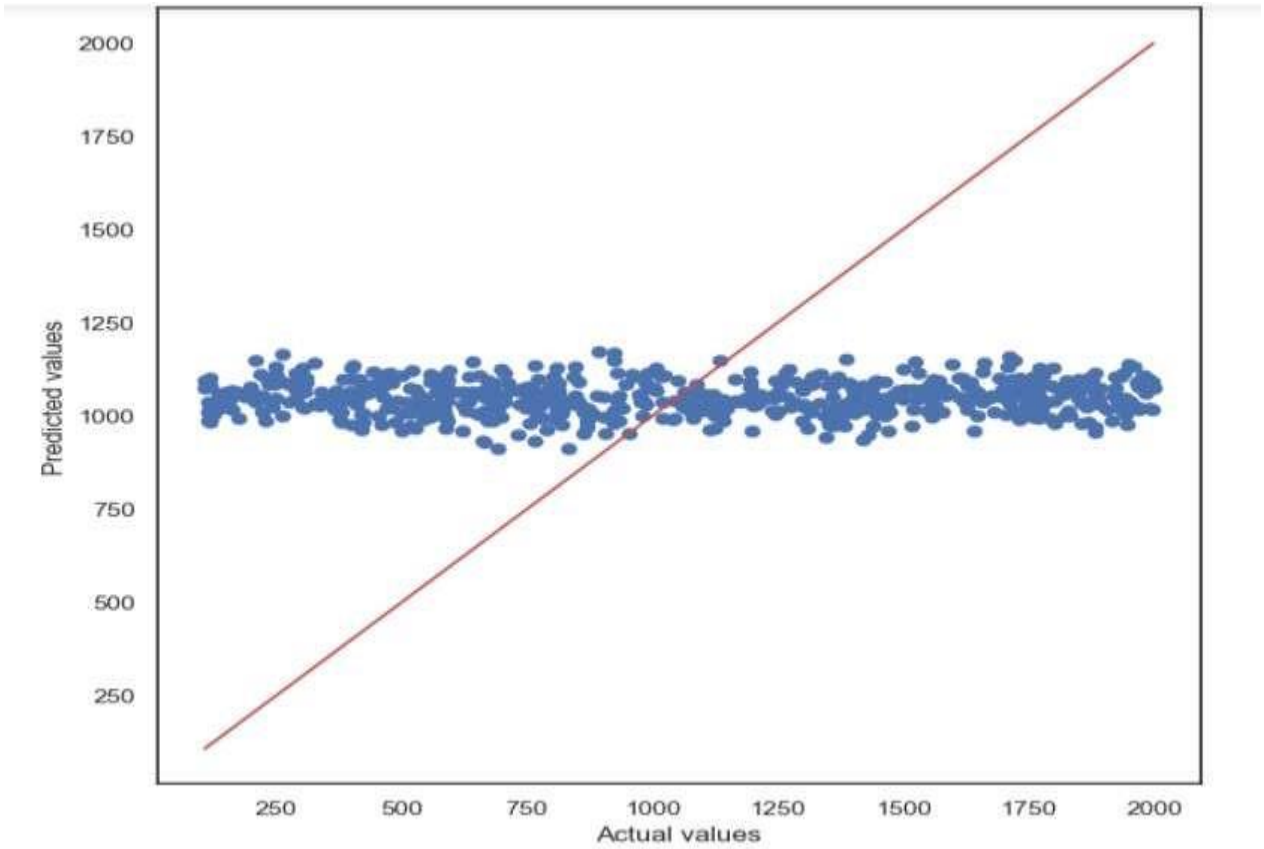
Sales prediction using machine learning techniques such as gradient boosting and linear regression can provide businesses with valuable insights and improve their decision-making processes. Some of the key significant features of sales prediction using these techniques include:

1. **Accurate Sales Forecasting:** Machine learning algorithms can analyze large amounts of historical data and identify patterns and trends that can be used to predict future sales accurately. This can help organisations reach more informed decisions regarding pricing, inventory management, and manufacturing.
2. **Improved Planning:** With accurate sales forecasts, businesses can plan their resources more effectively, such as their workforce, inventory, and marketing efforts. They can minimize costs and optimize their operations thanks to this.
3. **Better Marketing Strategies:** By analyzing historical data on customer behavior and preferences, machine learning algorithms can help businesses to identify the most effective marketing strategies for different customer segments. This can improve customer engagement and increase sales.
4. **Reduced Risk:** Sales prediction can help businesses to identify potential risks and opportunities in advance. Businesses can move promptly to decrease risk and adapt their plans if the algorithm forecasts a drop in sales for a specific product, for example.
5. **Improved Customer Experience:** By predicting customer behavior, businesses can tailor their products and services to meet their needs and preferences. Sales and profitability may increase when customer loyalty and satisfaction rise.

Results:-

ML

Linear Regression



Comparison of Actual vs Predicted sales.

From this graph it is inferred that the Sales predicted values for all stores is in range of Rs 1000 to 1250 per day based on Linear Regrssion algorithm.It shows rough analysis of sales

Gradient Boosting

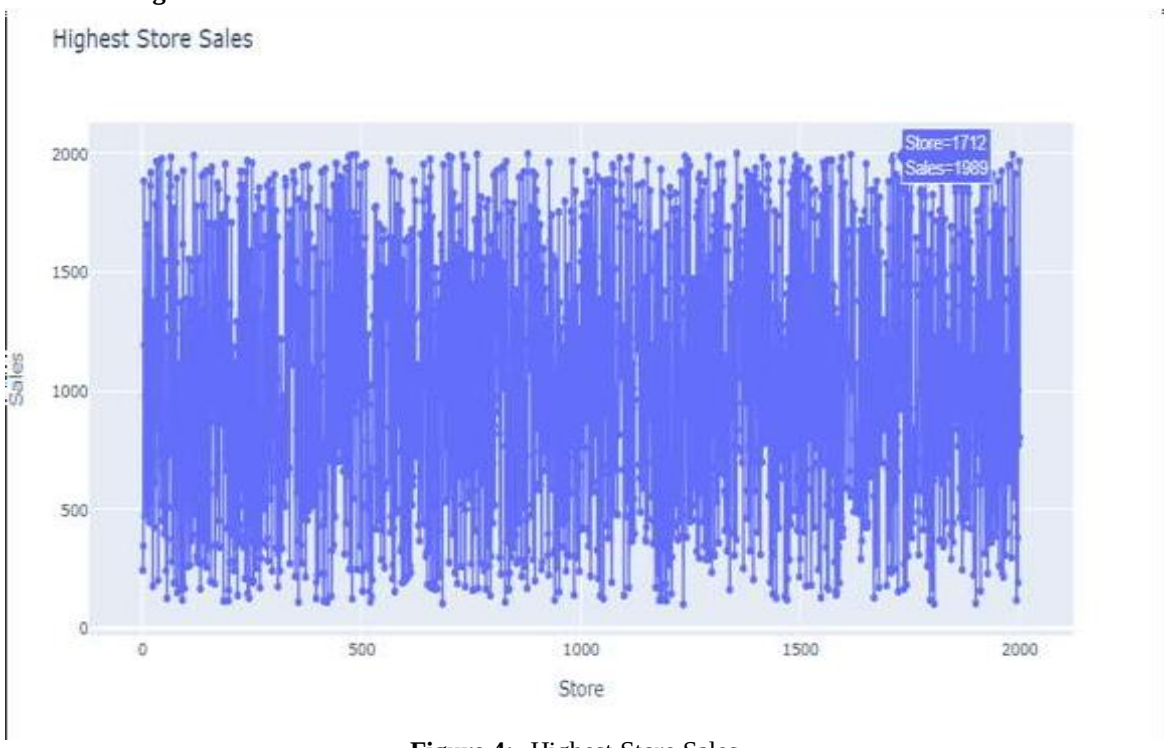
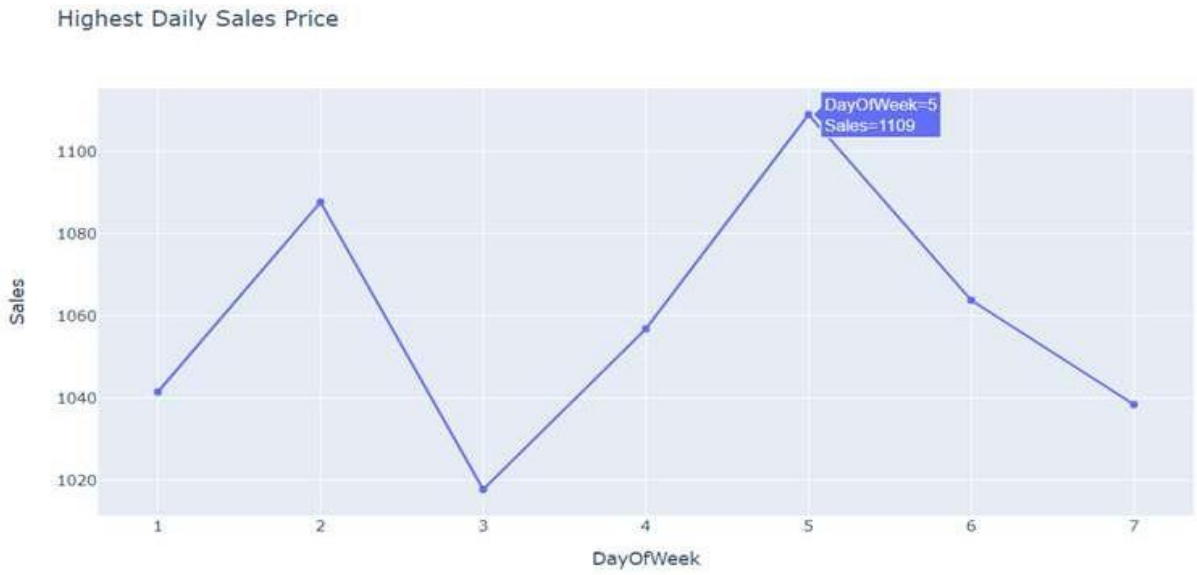


Figure 4:- Highest Store Sales.



Highest Daily Sales Price.

From these graphs it is inferred that on day five of the week, the maximum sales is occurred and its corresponding value for the specific store is described. Also Gradient boosting shows exact sales value of every store and its storetype accordingly.

Power BI

Estimation of Daily,Monthly Total Sales and its Profit value.

Estimation of Annual Sales and Total Store Sales and its Profit Percentage value.

Conclusion:-

Sales predictive analysis utilizing machine learning algorithms is a useful tool that can help firms get insights into their sales trends and anticipate future sales outcomes. Businesses may create reliable predictive models using machine learning techniques such as linear regression and gradient boosting. This allows them to make data-driven decisions and optimize their sales strategy. These prediction models' accuracy and performance can be enhanced by employing proper feature selection strategies and model evaluation procedures. Due to the increased accessibility of data and the rising popularity of machine learning, predictive sales analysis is swiftly evolving into an essential tool for companies wanting to remain competitive in today's market. Predictive sales analysis using machine learning algorithms has the potential to fundamentally alter how businesses make sales decisions while also improving their bottom line. Businesses can get a competitive edge and boost sales performance through the power of data and machine learning.

Merits And Demerits:-

Merits:

1. **Enhanced operational efficiency:** Machine learning models can help businesses identify sales patterns and optimize their inventory management, production, and supply chain processes. This can help businesses reduce waste, increase sales efficiency, and optimize their resources.
2. **Improved pricing strategies and increased revenue:** Machine learning models can help businesses analyze customer behavior, purchase history, and market trends to develop effective pricing strategies. By optimizing pricing strategies, businesses can increase sales revenue and profit margins.
3. **Enhanced marketing strategies and improved customer engagement:** Machine learning models can help businesses analyze customer data to identify the most effective marketing channels and messages. Marketing tactics that are more successful and targeted may result in a rise in sales.
4. **Competitive advantage and new opportunities**: By leveraging machine learning models to analyze sales data, businesses can identify new market opportunities, detect emerging trends, and stay ahead of the competition. This can provide companies a competitive edge and spur growth.
5. **Informed critical business decisions**: Machine learning models can provide businesses with real-time insights into sales performance, customer behavior, and market trends. This can assist companies in making decisions that will have a favourable effect on their sales results.
6. **Increased revenue and profit:** By leveraging machine learning models to optimize their sales processes, pricing strategies, and marketing efforts, businesses can increase sales revenue and profit margins. Machine learning models can also assist companies in finding chances for cost reduction, which boosts revenues.

Demerits:

1. **Reliance on historical data:** Machine learning models rely on historical data to make predictions. However, due to alterations in the market, alterations in consumer behaviour, or alterations in other factors, previous data might not always be trustworthy. This may result in incorrect sales projections.
2. **Limited human input and external factor considerations:** Machine learning models may not consider external factors such as seasonality, economic conditions, or customer sentiment. This can lead to inaccurate sales predictions and missed opportunities.
3. **High complexity and technical expertise requirements:** Machine learning models can be complex, requiring technical expertise to develop and maintain. This can be costly and time-consuming for businesses that lack the necessary expertise.
4. **Privacy and data security concerns:** Machine learning models require access to customer data, which raises privacy and data security concerns. Businesses must ensure that their data management practices correspond to privacy regulations and industry standards in order to protect their customers' data.
5. **Limited interpretability and difficult to understand predictions:** Machine learning models can produce complex and difficult to understand predictions, making it challenging for businesses to interpret and act on them. This can limit the practical value of machine learning for sales prediction.
6. **Potential bias from biased training data:** Machine learning models can be biased if the training data is biased. This can lead to inaccurate sales predictions and perpetuate systemic biases that affect customer behavior and business outcomes. Businesses must ensure that their data is representative and unbiased to avoid this issue.

Future Enhancements:-

The subject of sales prediction utilizing machine learning algorithms is continually evolving, and there are several prospects for future improvements to improve the accuracy and effectiveness of these models.

Among the probable future enhancements are:

1. **Adding new data sources:** Companies can improve their understanding of sales trends and the precision of their prediction models by including new data sources like social media, consumer behaviour, and macroeconomic data.
2. **Using more advanced machine learning algorithms:** While gradient boosting and linear regression are valuable machine learning approaches, deeper learning and neural networks can yield more accurate predictive models.
3. **Implementing real-time data processing:** Businesses can keep up with the most recent sales trends and respond rapidly to market developments thanks to real-time data processing.
4. **Using natural language processing:** By adding NLP, firms can analyze consumer feedback and sentiment to acquire insights into their sales strategy.
5. **Creating personalized sales suggestions:** Businesses can improve customer engagement and drive sales by utilizing predictive data to provide personalized sales recommendations for specific consumers.