PREDICTIVE ANALYSIS FOR BIG MART SALE PREDICTION USING MACHINE LEARNING ALGORITHMS.

**Saurabh Kale1, Shruti Gawande2, Shubham Saste3, Soham Jadhav4,Prof. Pradip Shinde**Department of Information Technology Dr. D. Y. Patil School of Engineering, Pune, Maharashtra, India.

*Abstract:* Technology and data analytics are driving the retail industry's rapid evolution. Within this framework, sales forecasting is essential for maximizing profitability, streamlining inventory control, and improving customer satisfaction. This study uses cutting-edge machine learning techniques to forecast sales for Big Mart, a prominent retail chain.The main goal of this study is to create a reliable and accurate prediction model to help Big Mart make data-driven decisions. Predictive models are created using a variety of machine learning algorithms, including regression, time series analysis, and ensemble approaches, and a complete dataset that contains historical sales data, product features, shop details, and external events like holidays and promotions. Data preprocessing methods are used in the study to address missing values, outliers, and feature.

**Keywords recommendation** - Video Surveillance, Anomaly detection, Machine learning, Convolutional neural networks, Image processing

# 1. Introduction:

Ordinary competitiveness among various buying centres as and as massive marts is becoming higher extreme, violent simply because of the quick improvement of worldwide shops additionally online buying. each marketplace seeks to offer customized and limit time deals to attract many clients counting on period of time, so that every item's quantity of income maybe anticipated for the company's inventory manage, transportation and logistical offerings. The on temporary device mastering algorithm is very superior and affords strategies for predicting or forecasting sales any kind of organisation, extraordinarily useful to conquer low – priced used for prediction. continually higher prediction is helpful, both in developing and improving marketing techniques for the market, which is likewise especially beneficial. Before submitting your final paper, check that the format conforms to this template. Specifically, check the appearance of the title and author block, the appearance of section headings, document margins, column width, column spacing and other features. To address these challenges, Big Mart has turned to data science and machine learning. This research endeavors to develop a robust sales prediction model customized to the specific requirements of Big Mart. The primary objective is to leverage historical sales data, product attributes, store-related information, and external factors such as holidays and promotions to create a predictive model that not only forecasts sales but also provides valuable insights into the factors driving sales performance.

**2. Literature Survey:** We have to gone through the different types of research paper and have been reported in the literature ,However few relevant and significant works are reviewed here.

**Kadam,et.al [1]:**Have suggested when the prediction for the sales for bigmart was done using the algorithm like random forest and LR for prediction analysis it gave lesser accuracy.So to overcome this problem we can use another algorithm which is XG boost algorithm which not only gives better accuracy but also is more efficient.

**Makridakis, et.al [2]:** Haave suggested predicting methods and applications containing Data Lack and short life cycles. So some data like historical data, consumer-focused markets face uncertain needs, which can be an accurate predictor of outcome.

as the concept which is very complex and less efficient concluded that we should use much simpler algorithm for the prediction.

# 3. Proposed System:

The Proposed Predictive Sales Analytics System for Big Mart aims to leverage historical sales data to forecast future sales trends accurately. By harnessing advanced data analytics and machine learning algorithms, the system will provide valuable insights and predictions to optimise inventory management, enhance marketing strategies, and improve overall sales performance.The objective of this proposed system is to predict the future sales from given data of the previous year's using Decision Tree Regression . Another objective is to conclude the best model which is more efficient and gives fast and accurate result by using Decision Tree Regression. To find out key factors that can increase their sales and what changes could be made to the product or store's characteristics. Experts also shown that a smartsalesforecasting program isrequired to manage vast volumes of data for business organizations.We are predicting the accuracy for Decision Tree Regression. Our predictions help big marts to refine their methodologies and strategies which in turn helps them to increase their profit. The results predicted will be very useful for the executives of the company to know about their sales and profits. This will also give them the idea for their new locations or Centre‟s of Bigmar.

Key Component:

**3.1Data Collection and Integration:** Gather comprehensive historical sales data, including product information, sales volumes, prices, promotions, and seasonal trends. Integrate data from various sources within the Big Mart network, including different store locations, product categories, and time periods.Below mentioned are important Key Components require under the Data Collection and Integration.

A) Source Identification: Identify diverse sources of data within Big Mart's ecosystem, including transactional databases, sales records, inventory logs, customer information, promotional activities, and external data like economic indicators or weather patterns.Collaborate with different departments and branches to ensure comprehensive data collection across various store locations, product categories, and time frames.

B)Data Gathering Mechanisms: Implement automated data collection mechanisms to streamline the extraction process. This may involve APIs, data connectors, or ETL (Extract, Transform, Load) processes to pull data from different databases and sources. Ensure data collected is in structured formats compatible with analysis and processing.

C) Data Quality Assurance: Perform rigorous data cleaning and validation processes to ensure data quality and integrity. This includes handling missing values, correcting errors, removing duplicates, and standardizing data formats. Validate collected data against predefined business rules to ensure accuracy and reliability.

**3.2Data Preprocessing and Cleaning:**Perform thorough data cleaning, handling missing values, and addressing outliers to ensure data quality and accuracy. Normalize and standardize datasets to prepare them for analysis. We are applied below some operations using python Data Processing Libraries.

A) Missing Values Handling: Identify and handle missing values in the dataset appropriately. Employ techniques such as imputation (mean, median, mode), interpolation, or using predictive models to fill in missing data. Assessthe nature of missingness (whether it'srandom orsystematic) to determine the most suitable approach for handling missing values.

B) Outlier Detection and Treatment: Identify outliers or anomalies in the data that may skew analysis or model performance. Use statistical methods, visualisation techniques, or machine learning algorithms to detect outliers. Apply techniques like trimming, winsorizing, or transforming skewed distributions to mitigate the impact of outliers on the analysis without losing critical information.

C) Data Normalisation and Standardisation:Normalize or standardize numerical features to bring them to a common scale, preventing dominance of certain variables in the analysis due to differing magnitudes. Techniques like Min-Max scaling, Z-score normalization, or robust scaling can be employed to standardize numerical data.

D) Handling Categorical Data:Encode categorical variables into numerical formats suitable for analysis. Use techniques like one-hot encoding, label encoding, or embedding methods to represent categorical data as numerical features. Ensure proper handling of categorical variables with a large number of unique categories to avoid issues like the curse of dimensionality.

**3.3 Feature Engineering:** Extract relevant features such as time-based patterns, seasonal variations, product attributes, and store-specific characteristics. Create new variables or features that could enhance the predictive power of the model.

A) Time-based Features: Extract temporal features such as day of the week, month, season, or holiday indicators from timestamps. These features can capture seasonal variations, weekday/weekend patterns, or holiday effects on sales. Create lag features representing past sales data (e.g., sales in the previous week/month) that can capture trends and seasonality.

B) Aggregated Features: Aggregate sales data over different time intervals (daily, weekly, monthly) or across various store locations, product categories, or customer segments. This can provide insights into overall trends and performance.Compute statistical aggregates like mean, median, standard deviation, or percentiles for sales volumes, prices, or promotional activities over specific time windows.

C)Product Attributes and Categories: Extract information from product attributes such as brand, weight, size, or category. Convert categorical product features into numerical representations using encoding techniques. Create indicators for new products, popular brands, or specific categories that may influence sales patterns.

D) Market and External Factors: Incorporate external data sources such as economic indicators (e.g., inflation rates, GDP), weather conditions, or local events that might impact sales. Generate features like consumer sentiment indices, competitor activities, or advertising expenditures that correlate with sales fluctuations.

**3.4 Predictive Modeling:** Employ machine learning algorithms such as regression, time series analysis, or ensemble methods to build predictive models. Train the models using historicalsales data, validating and fine-tuning them to achieve optimal accuracy and performance. Data Splitting: Split the available data into training, validation, and test sets. The training set is used to train the model, the validation set helps tune hyperparameters, and the test set evaluates the final model's performance. Model Selection: Choose appropriate predictive modeling techniques based on the nature of the problem. Common approaches for sales prediction include: Regression Models: Linear Regression, Polynomial Regression, Ridge Regression, Lasso Regression, etc., to predict continuous sales values. Time Series Models: ARIMA (AutoRegressive Integrated Moving Average), SARIMA (Seasonal ARIMA), Prophet, etc., for handling temporal data and seasonality. Machine Learning Models: Random Forests, Gradient Boosting Machines (GBM), XGBoost, Neural Networks, etc., for more complex relationships and feature interactions. Feature Selection and Engineering: Identify and use the most relevant features from the dataset. Techniques like Recursive Feature Elimination, feature importance from tree-based models, or domain knowledge can aid in feature selection. Utilize the engineered features (as discussed earlier) that capture the most meaningful information for predicting sales. Model Training: Train the selected models using the training dataset while tuning hyperparameters to optimize model performance. Techniques like cross-validation help in finding the best hyperparameters.

**3.5 Model Evaluation and Selection:** Evaluate the performance of multiple models using appropriate metrics (e.g., RMSE, MAE, R-squared) to select the best-performing model. Ensure the chosen model can handle various scenarios and adapt to changing sales patterns.

**3.6 Prediction and Visualization:** Deploy the selected model to generate future sales forecasts based on new or unseen data. Visualise predictions and trends using interactive dashboards and visual representations to facilitate easy interpretation by stakeholders. Exploratory Data Analysis (EDA): Before modeling, visualizations help in understanding the underlying patterns and distributions within the data. Histograms, box plots, scatter plots, and correlation matrices visualize relationships between variables. Time series plots can reveal seasonality, trends, and irregular patterns in sales data, aiding in feature selection and engineering. Model Evaluation and Validation: Visualizations assist in evaluating model performance. For regression models, scatter plots of predicted versus actual values provide an intuitive way to assess accuracy. ROC curves, precision-recall curves, and confusion matrices are visual tools for evaluating classification models used in sales prediction. Feature Importance: Visualizations like bar plots or heatmaps of feature importance derived from models (like decision trees, random forests, or gradient boosting) showcase which features contribute most to the predictions. This helps in refining feature engineering efforts. Prediction Intervals and Uncertainty: Visualizing prediction intervals (ranges where predictions are likely to fall) helps understand the uncertainty associated with forecasts. This can be visualized using shaded areas or confidence bands in time series plots. Dashboard and Reporting: Interactive dashboards or visual reports provide a comprehensive overview of predictions and insights derived from models. They allow stakeholders to explore predictions across different product categories, time periods, or store locations interactively. Forecasting Trends and Anomalies: Visualizations help in showcasing forecasted trends against historical data. They can also highlight anomalies or deviations from expected sales patterns, prompting further investigation or strategy adjustments.

**3.7 Integration with Decision Support Systems:** Integrate the predictive sales analytics system with Big Mart's decision-making processes. Provide actionable insights and recommendations to stakeholders, aiding in inventory planning, marketing campaigns, and strategic decision-making.

**3.8 Continuous Improvement and Maintenance:** Implement a feedback loop to continuously refine and improve the predictive models. Regularly update the system with new data and adapt models to evolving sales patterns and market dynamics.

**4. Methodology:** Data Collection :Gather historical and real-time data on water quality parameters. Include measurements such as temperature, pH, dissolved oxygen, chemical concentrations, etc. Collect relevant environmental and meteorological data. Data Preprocessing: Clean data by addressing missing values, outliers, and inconsistencies. Perform data normalization, scaling, and feature engineering. Data Splitting: Divide the dataset into training, validation, and test sets. Training: For model training. Validation: For hyper parameter tuning and model evaluation. Test: To assess model generalization. Feature Selection: Identify and select relevant features to improve model performance. Consider correlation analysis and feature importance techniques. Model Selection: Choose appropriate machine learning algorithms (e.g., regression, decision trees, neural networks) for prediction. Model Training : Train the selected model using the training dataset. Model learns to map input features to target variables (e.g., pollutant concentration). Model Deployment: Deploy the model for real-time or periodic predictions. Deployment platforms can include cloud servers or IoT devices.

**4.1 Hardware Requirements:** System : Pentium i3 Processor , Hard Disk : 500GB , Ram : 4GB.

**4.2 Software Requirements: Programming language :** Python 3.10, Operating System : MacOs,Windows, Web Framework : Flask Stream-light , Python Libraries for Data Analysis : Pandas , Numpy , Dtale , Seaborn , Matplotlib , Klib, Python Libraries for Model Building: Sk Learn,Tensor Flow.

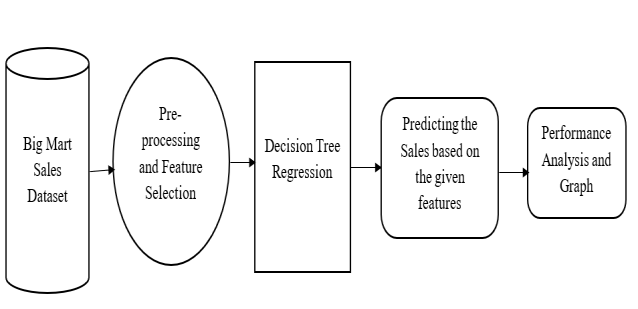
**5 Algorithm: Linear Regression:** Build a fragmented plot.1) a linear or non-linear pattern of data and 2) a variance (outliers). Consider a transformation if the marking isn't linear. If this is the case, outsiders, it can suggest only eliminating them if there is a non-statistical justification. Link the data to the least squares line and confirm the model assumptions using the residual plot (for the constant standard deviation assumption) and the normal probability plot (for the normal probability assumption) A transformation might be necessary if the assumptions made do not appear to be met. If required, convert the data to theleast square using the transformed data, construct a regression line. If a change has been completed, return to the previous process 1. If not, continue to phase 5. When a "good-fit" classic is defined, write the least-square regression line equation. Consist of normal estimation, estimation, and Rsquared errors. Linear regression formulas look like this: Y=o1x1+ o2x2+……… onxn R-Square: Defines the difference in X (depending variable) explains the total variance in Y (dependent variable) (independent variable).

# 5.1 Polynomial Regression Algorithm: Polynomial Regression is a relapse calculation that modules the relationship here among dependent(y) and the autonomous variable(x) in light of the fact that as most extreme limit polynomial. The condition for polynomial relapse is given beneath: y= b0+b1x1+ b2x1 2+ b2x1 3+...... bnx1 n. It isregularly alluded to asthe exceptional instance of various straight relapse in ML. Since we apply some polynomial terms to the numerous straight relapse condition to change it to polynomial relapse adjustment to improve accuracy. The informational collection utilized for preparing in polynomial relapse is of a non-straight nature. It uses a linear regression model to fit complex and non-linear functions and datasets. C. Ridge Regression Ridge regression is a model tuning tool used to evaluate any data that suffers from multicollinearity. This method performs the L2 regularization procedure. When multicollinearity issues arise, the least squares are unbiased and the variances are high, resulting in the expected values being far removed from the actual values. The cost function for ridge regression: Min(||Y – X(theta)||^2 + λ||theta||^2) D. XGBoost Regression “Extreme Gradient Boosting” is same but much more effective to the gradient boosting system. It has both a linear model solver and a tree algorithm. Which permits “xgboost” in any event multiple times quicker than current slope boosting executions. It underpins various target capacities, including relapse, order and rating. As "xgboost" is extremely high in prescient force however generally delayed with organization, it is appropriate for some rivalries. It likewise has extra usefulness for cross-approval and finding significant factors.

**5.2 XGBoost Algorithm:** What is XGBoost in Machine Learning? XGBoost is designed for speed, ease of use, and performance on large datasets. It does not require optimisation of the parameters or tuning, which means that it can be used. XGBoost is a widespread implementation of gradient boosting. Let’s discuss some features of XGBoost that make it so attractive: XGBoost offers regularization, which allows you to control overfitting by introducing L1/L2 penalties on the weights and biases of each tree. This feature is not available in many other implementations of gradient boosting. Another feature of XGBoost is its ability to handle sparse data sets using the weighted quantile sketch algorithm. This algorithm allows us to deal with non-zero entries in the feature matrix while retaining the same computational complexity as other algorithms like stochastic gradient descent. XGBoost also has a block structure for parallel learning. It makes it easy to scale up on multicore machines or clusters. It also uses cache awareness, which helps reduce memory usage when training models with large datasets. XGBoost Algorithm 14 | P a g e 4) Finally, XGBoost offers out-of-core computing capabilities using disk-based data structures instead of in-memory ones during the computation phase.

**5.3 Benefits and Attributes**: XGBoost is a highly portable library on OS X, Windows, and Linux platforms. It's also used in production by organisations across various verticals, including finance and retail. XGBoost is open source, so it's free to use, and it has a large and growing community of data scientists actively contributing to its development. The library was built from the ground up to be efficient, flexible, and portable. You can use XGBoost for classification, regression, ranking, and even user-defined prediction challenges! You can also use this library with other tools like H2O or ScikitLearn if you want to get more out of your model-building process.

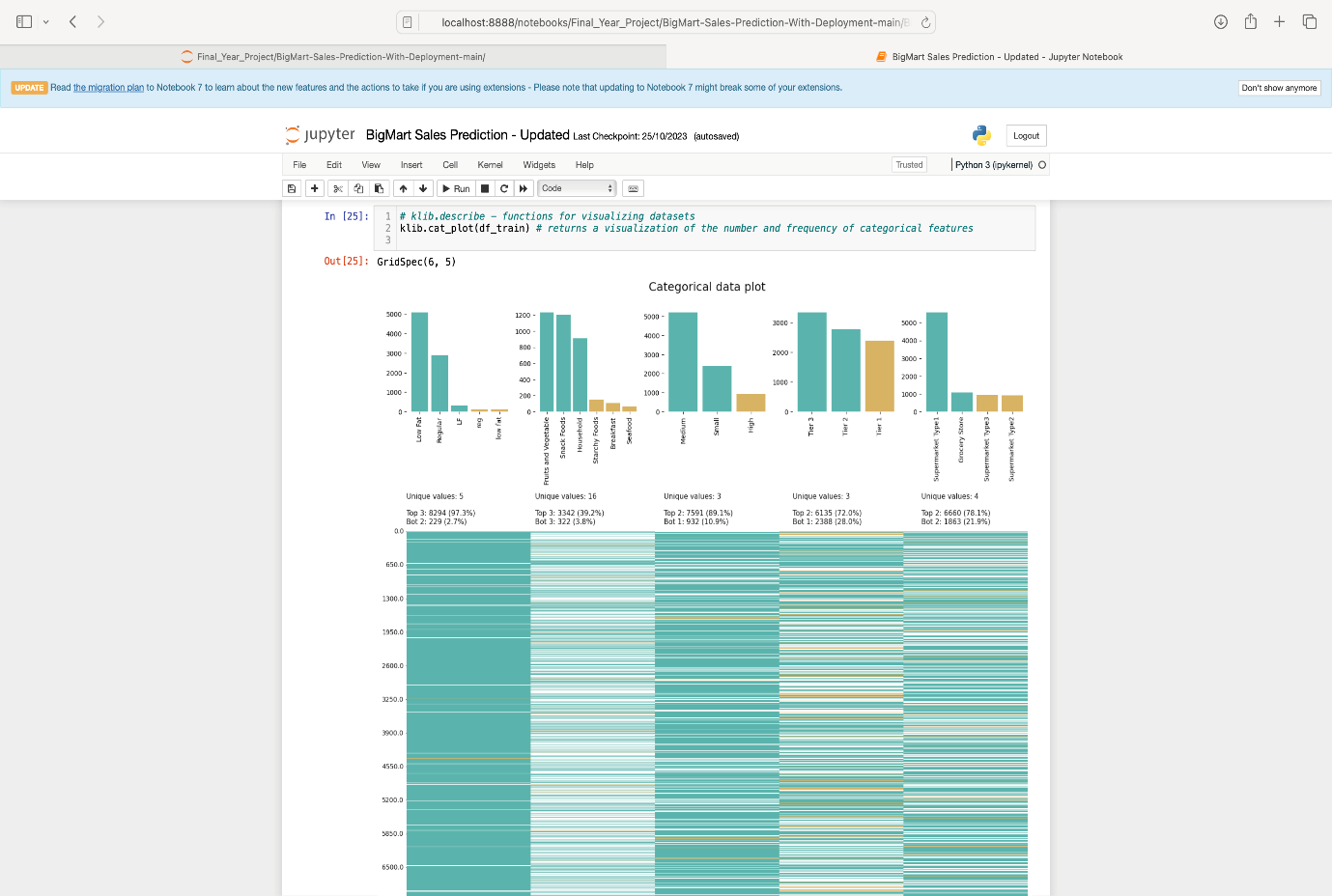
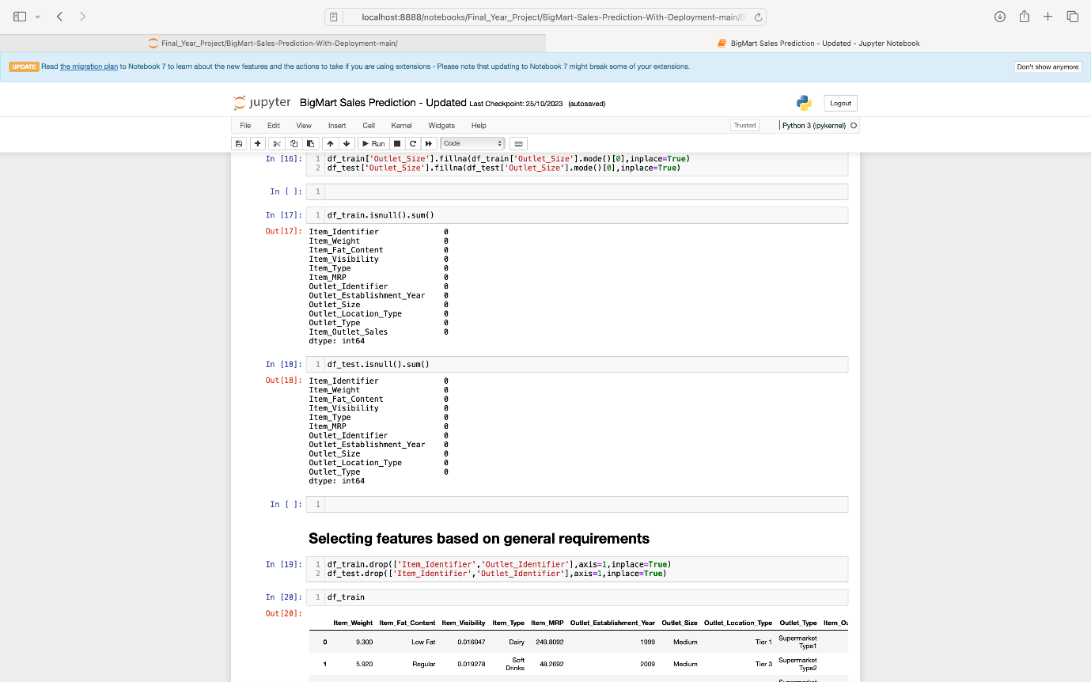
**6. Schematic:** Data Collection: Gather historical sales data including date, product, location, promotions, etc. Collect additional data like store size, competition, demographics, etc., if available. 1. Data Preprocessing: Handle missing values, outliers, and inconsistencies in the data. Convert categorical variables into numerical representations (one-hot encoding, label encoding, etc.). Normalize or scale numerical features. 2. Feature Engineering: Create new features that might impact sales (e.g., average sales per month, seasonal trends, promotional impact). Extract relevant information from date/time fields (day of the week, month, year). 3. Split Data: Divide the dataset into training and testing sets to train and evaluate the model. 4. Model Selection: Choose a suitable algorithm for regression (e.g., Linear Regression, Random Forest, Gradient Boosting, Neural Networks). Train multiple models and evaluate their performance using metrics like RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), etc. 5. Hyperparameter Tuning: Optimize the selected model's hyperparameters to improve performance. 6. Validation: Validate the model's performance using the testing dataset to ensure it generalizes well to new data.



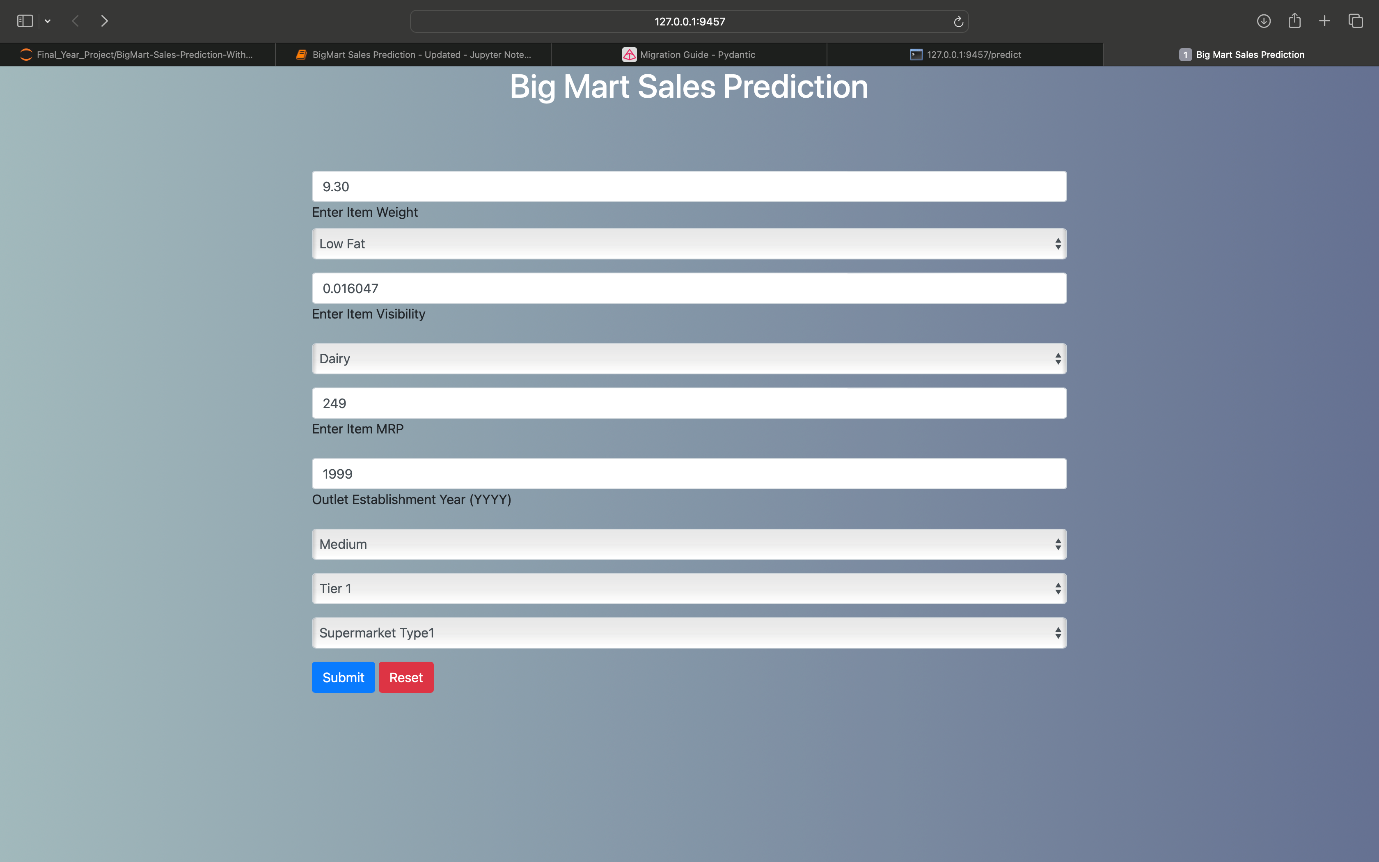
*Big mart sale Prediction*

**7.EDA and Result:**

Some EDA Operation on Dataset applying for Data Cleaning and Data Abstraction purpose. Graphical Representation using Seaborn and Pandas Library.Final model using Flask Library.



**7.2 Final Model:**



**7.1 Advantages:** Inventory Management: Accurate predictions help optimize inventory levels. Avoid overstocking or understocking by anticipating demand, reducing costs, and maximizing revenue. Resource Allocation: Efficiently allocate resources like manpower, marketing budgets, and stock across different locations or products based on predicted sales patterns. Marketing Strategies: Tailor marketing campaigns and promotions based on anticipated demand. Optimize timing, channels, and content to maximize impact and sales. Customer Satisfaction: Anticipating demand ensures products are available when customers need them, improving customer satisfaction and loyalty. Profit Maximization: Predictions aid in setting optimal pricing strategies, maximising profits while staying competitive in the market. Reduced Wastage: Minimize waste by aligning supply with demand, reducing the chances of unsold products reaching their expiration dates. Operational Efficiency: Streamline operations by aligning production schedules, staffing, and logistics with predicted sales volumes. Forecasting Trends: Analyzing historical sales data can reveal trends and patterns, enabling better long-term strategic planning and decision-making.

**7.2 Disadvantages:**  Data Dependency: Predictive models heavily rely on historical data. Changes in consumer behavior, market trends, or unforeseen events (like pandemics, economic crises) can render historical data less relevant or accurate. Complexity of Factors: Sales prediction involves numerous variables (seasonality, promotions, competition, etc.), making it challenging to capture all influencing factors accurately in a model. Accuracy Challenges: Even sophisticated models might struggle to accurately predict sales due to the inherent unpredictability of human behavior and external factors. Overreliance on predictions can lead to significant errors. Costly Errors: Inaccurate predictions can lead to overstocking, understocking, or misallocation of resources, resulting in financial losses or missed opportunities. Model Maintenance: Models need continuous updating and fine-tuning to remain relevant. Failure to adapt to changing trends or market dynamics might lead to decreased accuracy over time. Ethical Concerns: Over-reliance on predictive models might lead to biases or discriminatory outcomes, especially if models are trained on biased historical data. Resource Intensive: Building and maintaining predictive models require skilled personnel, computing power, and ongoing data collection efforts, which can be resource-intensive. Security Risks: Gathering and storing large amounts of customer data for predictive analysis can pose security and privacy risks if not handled securely.

**7.3Application:** 1. Marketing and Promotions: Anticipating sales patterns enables better planning of marketing campaigns and promotions. Targeted promotions at specific times can maximize their effectiveness. 2. Supply Chain Optimization: Predictive models aid in streamlining the supply chain by forecasting demand for suppliers and distributors, reducing lead times and costs. 3. Financial Planning: Accurate sales predictions assist in budgeting, financial forecasting, and setting revenue targets, allowing for better financial planning and allocation of resources. 4. Staffing and Operations: Understanding sales patterns helps in scheduling staff effectively to meet customer demand, improving operational efficiency and customer service. 5. Product Placement and Assortment: Predictions guide decisions on product assortment and placement in stores, optimizing layout and shelf space to maximize sales. 6. Price Optimization: Sales predictions inform pricing strategies, helping to set competitive prices while maximizing profits based on anticipated demand. 7. Forecasting Trends: Analyzing sales data over time can reveal market trends and patterns, guiding long-term strategic decisions and new product introductions.

**8. Conclusion:** Predicting sales for big mart establishments is a pivotal tool that offers multifaceted benefits for operational efficiency, resource optimization, and strategic decision-making. By leveraging historical data, predictive models, and sophisticated algorithms, businesses can anticipate consumer behavior, market trends, and demand fluctuations, enabling them to: Optimize Inventory and Resources: Accurately forecast demand to maintain optimal inventory levels, preventing overstocking or understocking while efficiently allocating resources. Enhance Marketing and Promotions: Tailor marketing strategies and promotional activities based on anticipated sales patterns, maximizing their impact and effectiveness. Streamline Operations and Supply Chain: Improve operational efficiency by aligning staffing, production, and distribution with predicted sales, optimizing the supply chain. Improve Customer Satisfaction: Ensure products are available when and where customers need them, enhancing customer experience and loyalty. Drive Profitability and Growth: Facilitate informed decision-making regarding pricing strategies, product assortment, and expansion plans, ultimately maximizing profitability and supporting sustainable growth.

**9. Reference:** [1]Nikita Malik, Karan Singh MSI, Janakpuri, New Delhi. Sales Prediction Model for Big Mart. [2] Cerrada, M., & Aguilar, J. (2008). Reinforcement learning in system identification. In Reinforcement Learning. [3] Smola, A., & Vishwanathan, S. V. N. (2008). Introduction to machine learning. Cambridge University, UK, 32, 34 [4] Das, P.Chaudhury. Prediction of retail sales of footwear using feed forward and recurrent Neural Networks. [5] Kadam, H., Shevade, R., Ketkar, P., and Rajguru. A Forecast for Big Mart Sales Based on Random Forests and Multiple Linear Regression.