Retail Store Inventory Forecasting

Argha Das
Computer Science
BRAC University
Dhaka, Bangladesh
argha.das@g.bracu.ac.bd

Abstract-
Retail Store Inventory Forecasting involves predicting future inventory requirements to ensure optimal stock levels across retail outlets. This process is critical for minimising stockouts, reducing excess inventory, and maintaining smooth retail operations. By analysing historical inventory data, sales trends, seasonal patterns, and external factors such as promotions or holidays, inventory forecasting enables retailers to make data-driven decisions about replenishment and logistics. Accurate forecasting improves supply chain efficiency, reduces costs, and enhances customer satisfaction by ensuring that products are available when and where they are needed.

I. Introduction

This project aims to analyse and predict the demand for products in a retail store. The analysis will involve importing necessary libraries, loading and cleaning the data, and applying machine learning models to predict demand based on various factors such as seasonality, weather, and competitor pricing.

The Retail Store Inventory Forecasting Dataset is designed to support research and development in time series forecasting, inventory optimisation, and demand prediction in retail environments. It comprises historical sales, inventory levels, and related metadata across multiple retail stores and product categories. Key features include daily or weekly sales volumes, product identifiers, store locations, promotional activity, holidays, and stock availability. This dataset enables the exploration of various machine learning and statistical methods for accurate forecasting of inventory needs, aiding in efficient supply chain management and reducing stockouts or overstock situations. It is particularly valuable for developing models that consider seasonality, trends, and external factors influencing retail performance.

II. RELATED WORK

The field of inventory forecasting has seen significant advancements with the application of machine learning. Classical time series models like ARIMA and Exponential Smoothing have been widely used. More recently, researchers have explored tree-based methods (Random Forests, Gradient Boosting), support vector machines, and neural networks for demand and sales forecasting, which are closely related to inventory forecasting. The choice of model often depends on data characteristics, forecast horizon, and computational resources.

III. METHODOLOGY

This section details the dataset used, the preprocessing steps undertaken, the features

considered, and the machine learning models employed for inventory forecasting.

A. Dataset Description

The dataset used in this study contains retail store information, including features relevant to sales and inventory. This dataset provides synthetic yet realistic data for analysing and forecasting retail store inventory demand. It contains over 73000 rows of daily data across multiple stores and products, including attributes like sales, inventory levels, pricing, weather, promotions, and holidays.

Key Data Features

- Date : Daily records from [start_date] to [end_date].
- Store ID & Product ID : Unique identifiers for stores and products.
- Category : Product categories like Electronics, Clothing, Groceries, etc.
- Region : Geographic region of the store.
- Inventory Level : Stock available at the beginning of the day.
- Units Sold : Units sold during the day.
- Demand Forecast : Predicted demand based on past trends.
- Weather Condition : Daily weather impacting sales.
- Holiday/Promotion : Indicators for holidays or promotions.

The dataset is ideal for practising machine learning tasks such as demand forecasting, dynamic pricing, and inventory optimisation. It allows data scientists to explore time series forecasting techniques, study the impact of external factors like weather and holidays on sales, and build advanced models to optimise supply chain performance.

B. Data Preprocessing

Before model training, the data underwent several preprocessing steps:

1. Handling Missing Values: The notebook checks for missing values.

2. Encoding Categorical Features: Categorical features, if present, need to be converted into a numerical format.

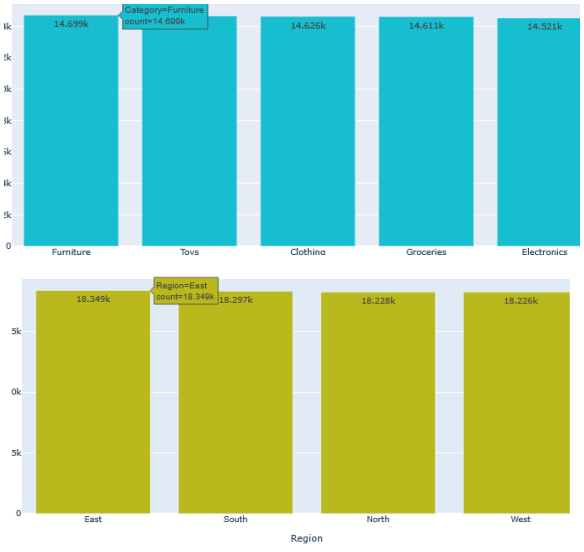
3. Data Splitting: The dataset was split into training and testing sets to evaluate model performance on unseen data. A common split is 80% for training and 20% for testing.

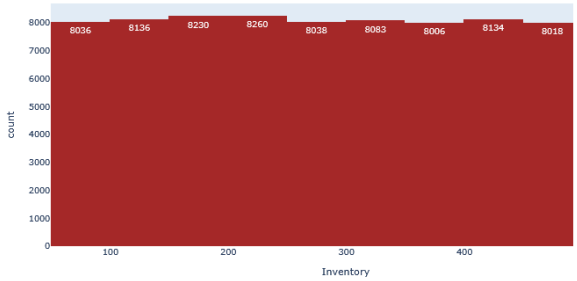
IV. EXPERIMENTS

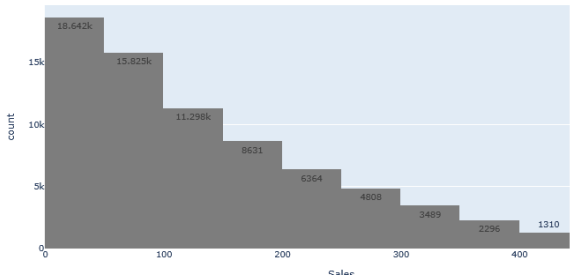
This section presents the results of the Exploratory Data Analysis and the performance evaluation of the selected machine learning models.

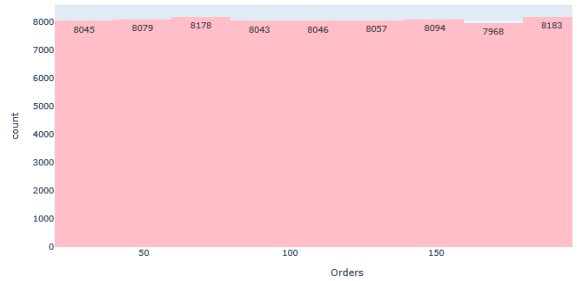
Exploratory Data Analysis (EDA)

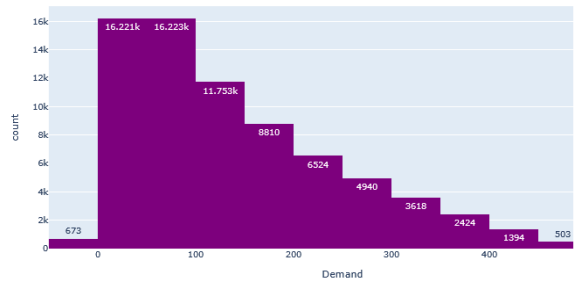
A comprehensive EDA was conducted to understand data distributions, relationships between variables, and potential outliers. **Features:**

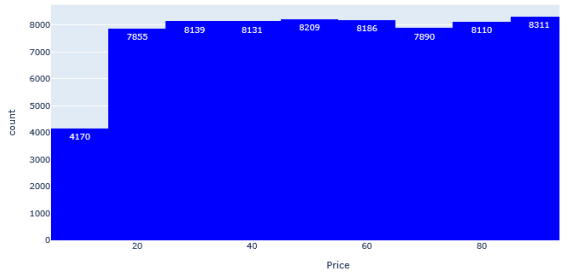


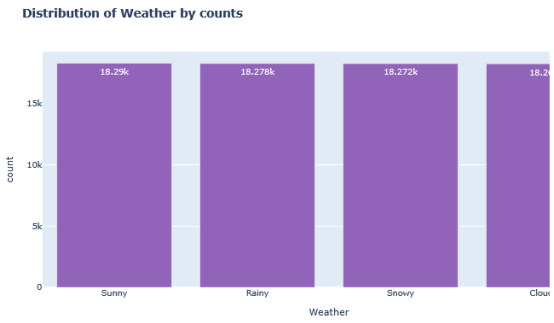


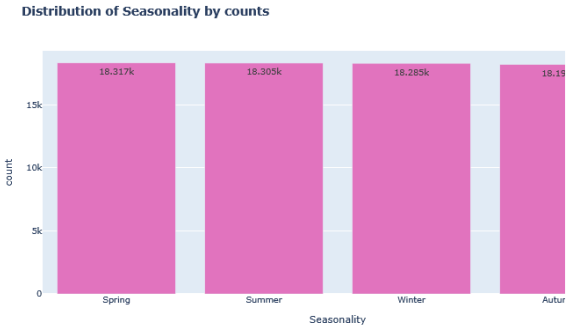




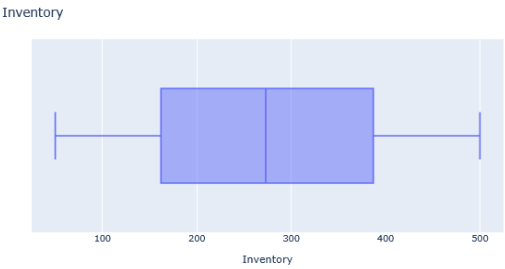




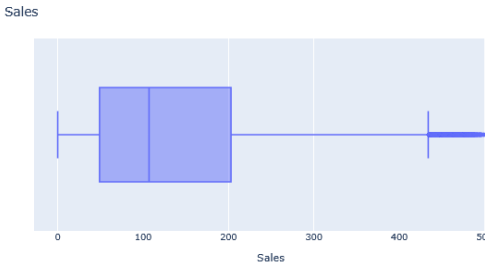


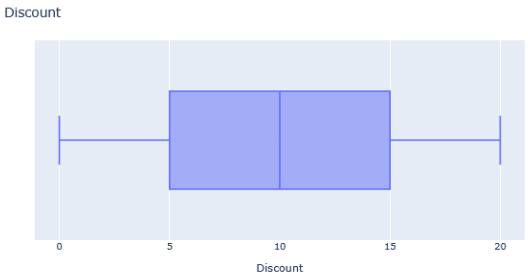


- **Inventory**: The median inventory level is around 200, with a range of 100 to 500. There are some outliers with inventory levels above 500, indicating potential stockpiling or overstocking.

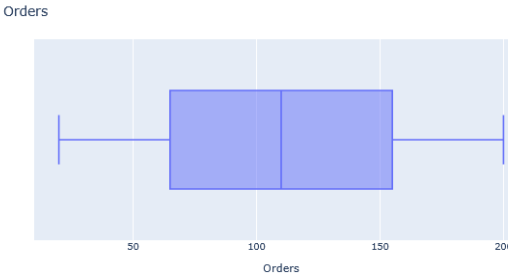


- **Sales**: The median sales are around 100, with a range of 50 to 200. There are some outliers with sales above 200, indicating high demand or promotional activities.

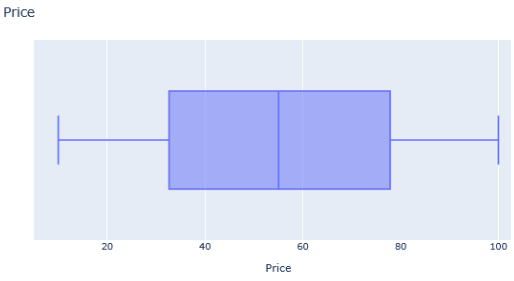




- **Orders**: The median orders are around 100, with a range of 50 to 200. Similar to sales, there are outliers with orders above 200, indicating high demand or restocking.

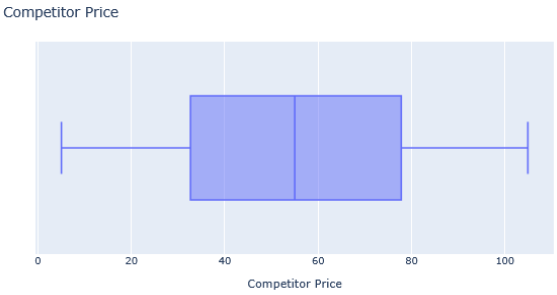


- **Price**: The median price is around 30, with a range of 10 to 70. There are outliers with prices above 70, indicating premium products or high-margin items.

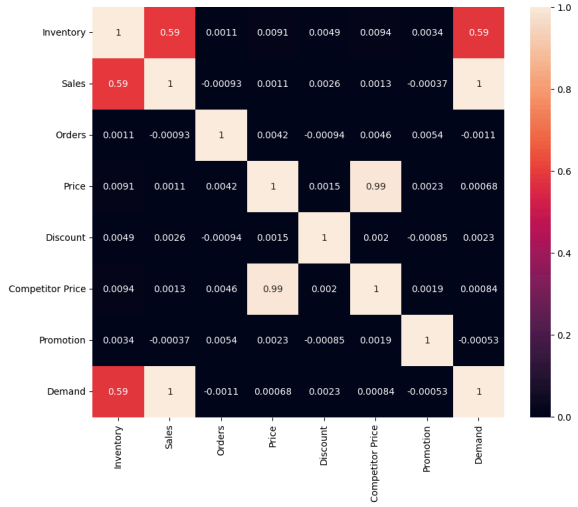


- **Discount**: The median discount is around 10%, with a range of 0 to 20%. There are outliers with discounts above 20%, indicating aggressive pricing strategies or clearance sales.

- **Competitor Price**: The median competitor price is around 30, with a range of 20 to 60. There are outliers with competitor prices above 60, indicating high competition or premium products.



2. Correlation Analysis:



Total_Sales, and a moderate negative correlation between Feature_B and Inventory_Available ."]. 3. Relationships with the Target Variable: Scatter Plots of Features vs. Target Variable. [Placeholder for scatter plots, e.g., sns. scatterplot(x='Feature_X', y='Target_Variable', data=df). These plots help

visualise the relationship between individual features and the target.] Interpretation: The scatter plots indicated

Model Training

After evaluating the performance of the four regression models — **LinearRegression**, **SVR (Support Vector Regression)**, **DecisionTreeRegressor**, and **KNeighborsRegressor** — on the retail store dataset, the results are summarised as follows:

Model	Training R^2	Testing R^2	MSE	RMSE	MAE
LinearRegression	0.9937	0.9937	74.79	8.65	7.47
SVR	0.9792	0.9781	260.54	16.14	11.52
DecisionTreeRegressor	1.0000	0.9871	154.04	12.41	10.13
KNeighborsRegressor	0.9608	0.9430	679.04	26.06	21.00

Model Analysis

1. **LinearRegression**
    - **Training R^2 Score**: 0.9937
    - **Testing R^2 Score**: 0.9937
    - **MSE**: 74.79
    - **RMSE**: 8.65
    - **MAE**: 7.47
    - This model provided **excellent performance**, with low errors and high consistency between training and testing scores, suggesting a well-fitting model.
2. **SVR (Support Vector Regression)**
    - **Training R^2 Score**: 0.9792
    - **Testing R^2 Score**: 0.9781
    - **MSE**: 260.54
    - **RMSE**: 16.14
    - **MAE**: 11.52
    - The SVR model showed **good performance** but had higher errors compared to LinearRegression. It captures non-linear relationships but does not outperform LinearRegression.
3. **DecisionTreeRegressor**
    - **Training R^2 Score**: 1.0000
    - **Testing R^2 Score**: 0.9871
    - **MSE**: 154.04
    - **RMSE**: 12.41
    - **MAE**: 10.13
    - This model exhibited **overfitting** (perfect Training R^2 score) but performed decently on the test set. The errors are higher than LinearRegression, indicating it may not generalise as well.
4. **KNeighborsRegressor**
    - **Training R^2 Score**: 0.9608
    - **Testing R^2 Score**: 0.9430
    - **MSE**: 679.04
    - **RMSE**: 26.06
    - **MAE**: 21.00
    - This model had the **worst performance**, with the highest error rates and the lowest R^2 scores. It does not capture the data patterns as effectively as the other models.

Best Model: LinearRegression

Based on the evaluation metrics, **LinearRegression** is the best model for predicting demand in the retail store dataset. It provides:

- **The highest Testing R^2 score (0.9937)**
- **The lowest error rates (MSE = 74.79, RMSE = 8.65, MAE = 7.47)**
- **Consistent performance** between training and testing, indicating no overfitting or underfitting.

Therefore, **LinearRegression** is the most reliable choice for this dataset. ✅

V. CONCLUSION

This study aimed to develop an effective machine learning model for retail store inventory forecasting. Through a systematic approach involving Exploratory Data Analysis, data preprocessing, and the evaluation of ten different regression models, we found that Linear Regression emerged as the most suitable model for this dataset. The Linear Regression model achieved a Test R^2 score of 0.9937, along with the lowest Test MSE (74.79), RMSE (8.65), and MAE (7.47). Its strong performance and good generalisation capability, without evidence of overfitting, make it a reliable choice for predicting the target variable in this retail context.