Sales Forecasting for Retail Businesses

A Project Based Learning Report Submitted in partial fulfilment of the requirements for the award of the degree

of

Bachelor of Technology

in The Department of CSE

BIG DATA ANALYTICS - 22DSB3303A

Submitted by

2210030420: K. Lahari Reddy
2210030403: K. Manaswija
2210030249: S. Neha Reddy
2210030399: Shreya Singh

Under the guidance of

SHAHIN FATIMA

Department of Electronics and Communication Engineering

Koneru Lakshmaiah Education Foundation, Aziz Nagar

Aziz Nagar – 500075

FEB - 2025

1. Introduction

The methodology for this sales forecasting project is structured to ensure **accuracy, efficiency, and scalability**. The development process follows a systematic approach, incorporating:

- **Data Collection**
- **Preprocessing**
- **Exploratory Data Analysis (EDA)**
- **Model Selection**
- **Implementation & Evaluation**
- **Deployment and Maintenance**

The goal is to build a predictive model to forecast sales of products at various stores, helping decision-makers identify key factors that influence sales. **By leveraging historical data and advanced analytical techniques, businesses can optimize inventory, pricing strategies, and marketing efforts.**

2. Hypotheses

The hypotheses explore different levels that could impact sales:

- **Store-Level**: Location, size, foot traffic, promotional strategies, store format.
- **Product-Level**: Category, price, demand elasticity, packaging, availability.
- **Customer-Level**: Buying behavior, preferences, seasonal demand, loyalty programs.
- **Macro-Level**: Economic conditions, market trends, competitor strategies, social trends.

3. Data Collection

3.1 Sources of Data

- **Datasets**: Train, Test.
- **Features**: 11 independent variables and 1 target variable (*Item_Outlet_Sales*) in the train dataset.
- **Dimensions**:
    - Train dataset: 8523 rows, 12 columns.
    - Test dataset: 5681 rows, 11 columns.
    - Additional external data sources such as macroeconomic indicators, weather data, and holiday schedules.

3.2 Exploratory Data Analysis (EDA)

- **Univariate Analysis**: Histograms and bar plots for individual feature distributions.
- **Bivariate Analysis**: Scatter plots, violin plots, and correlation heatmaps to understand relationships.
- **Time Series Analysis**: Identifying seasonality and long-term trends.
- **Outlier Detection**: Box plots and Z-score methods to identify anomalies.
- **Key Insights**: Patterns such as right-skewed *Item_Visibility* and *Item_MRP*, seasonal variations in sales, and the impact of store formats.

```
df_train descriptive statistics:
       Item_Weight  Item_Visibility     Item_MRP  Outlet_Establishment_Year  \
count  7060.000000      8523.000000  8523.000000                8523.000000
mean     12.857645         0.066132   140.992782                1997.831867
std       4.643456         0.051598    62.275067                   8.371760
min       4.555000         0.000000    31.290000                1985.000000
25%       8.773750         0.026989    93.826500                1987.000000
50%      12.600000         0.053931   143.012800                1999.000000
75%      16.850000         0.094585   185.643700                2004.000000
max      21.350000         0.328391   266.888400                2009.000000

       Item_Outlet_Sales
count        8523.000000
mean         2181.288914
std          1706.499616
min            33.290000
25%           834.247400
50%          1794.331000
75%          3101.296400
max         13086.964800

df_test descriptive statistics:
       Item_Weight  Item_Visibility     Item_MRP  Outlet_Establishment_Year
count  4705.000000      5681.000000  5681.000000                5681.000000
mean     12.695633         0.065684   141.023273                1997.828903
std       4.664849         0.051252    61.809091                   8.372256
min       4.555000         0.000000    31.990000                1985.000000
25%       8.645000         0.027047    94.412000                1987.000000
50%      12.500000         0.054154   141.415400                1999.000000
75%      16.700000         0.093463   186.026600                2004.000000
max      21.350000         0.323637   266.588400                2009.000000
```

4. Data Preprocessing

4.1 Handling Missing Values

- Imputation techniques (mean, median, predictive filling, KNN imputation).
- Removal of redundant and inconsistent records.

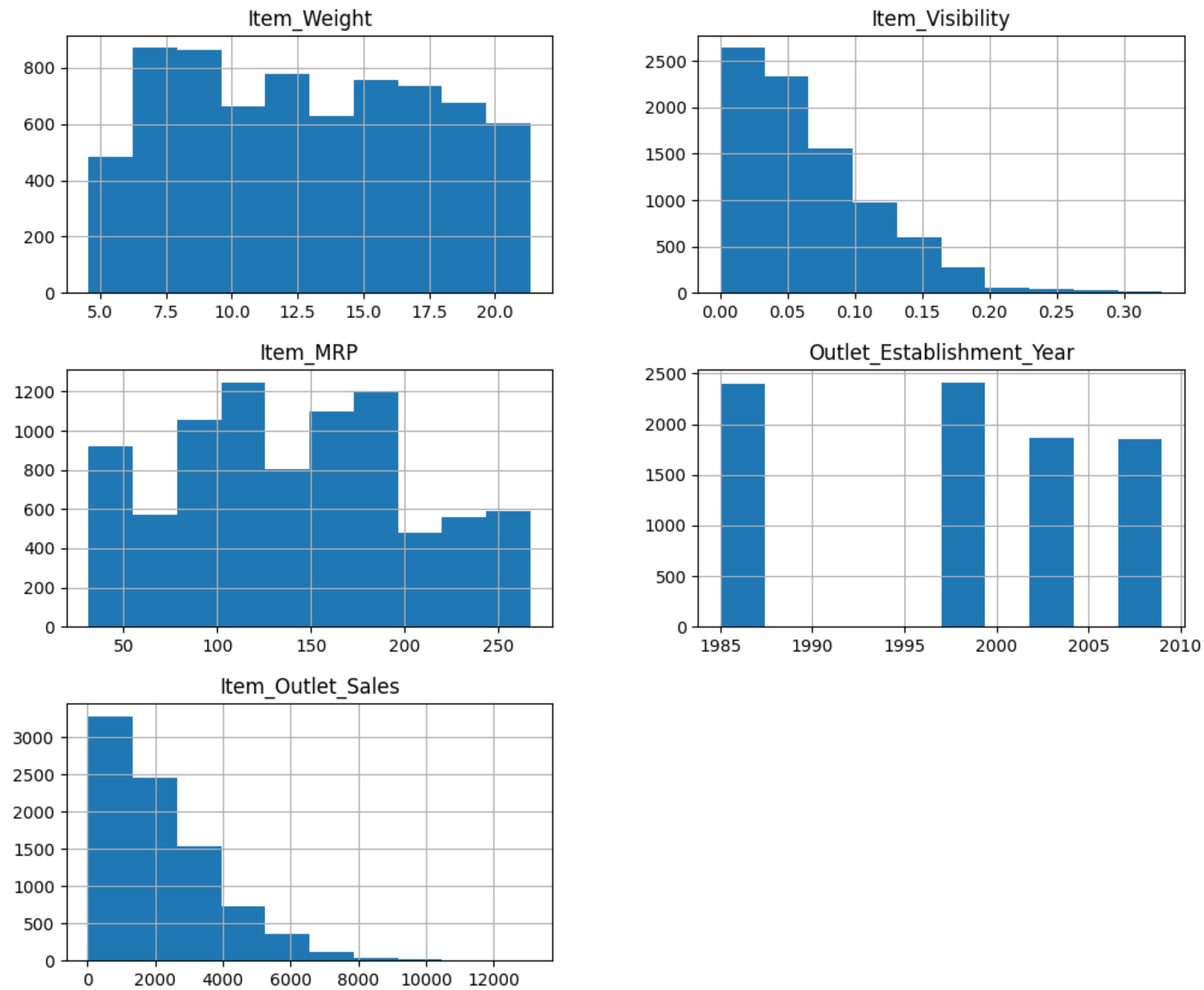
```
df_train missing values:
 Item_Identifier                0
Item_Weight                  1463
Item_Fat_Content                0
Item_Visibility                 0
Item_Type                       0
Item_MRP                        0
Outlet_Identifier               0
Outlet_Establishment_Year       0
Outlet_Size                  2410
Outlet_Location_Type            0
Outlet_Type                     0
Item_Outlet_Sales               0
dtype: int64

df_test missing values:
 Item_Identifier                0
Item_Weight                   976
Item_Fat_Content                0
Item_Visibility                 0
Item_Type                       0
Item_MRP                        0
Outlet_Identifier               0
Outlet_Establishment_Year       0
Outlet_Size                  1606
Outlet_Location_Type            0
Outlet_Type                     0
dtype: int64
```

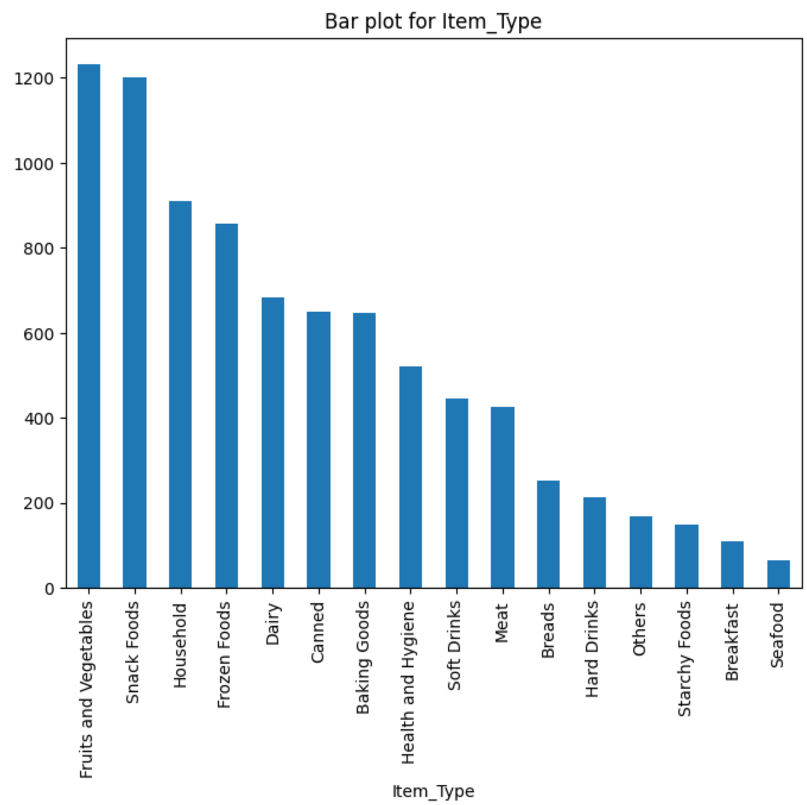
4.2 Data Normalization

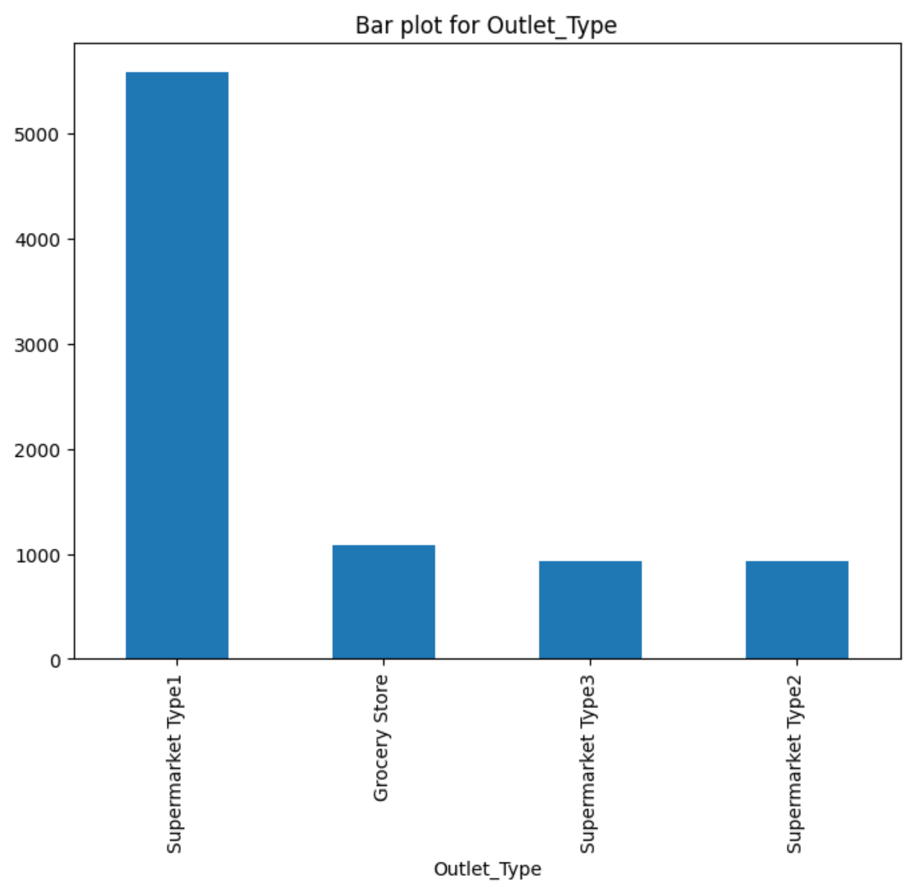
- Scaling numerical variables for model efficiency (Min-Max scaling, Standardization).
- One-hot encoding and label encoding for categorical variables.

4.3 Data Aggregation

- Grouping data by time intervals (*daily, weekly, monthly*) to extract meaningful patterns.

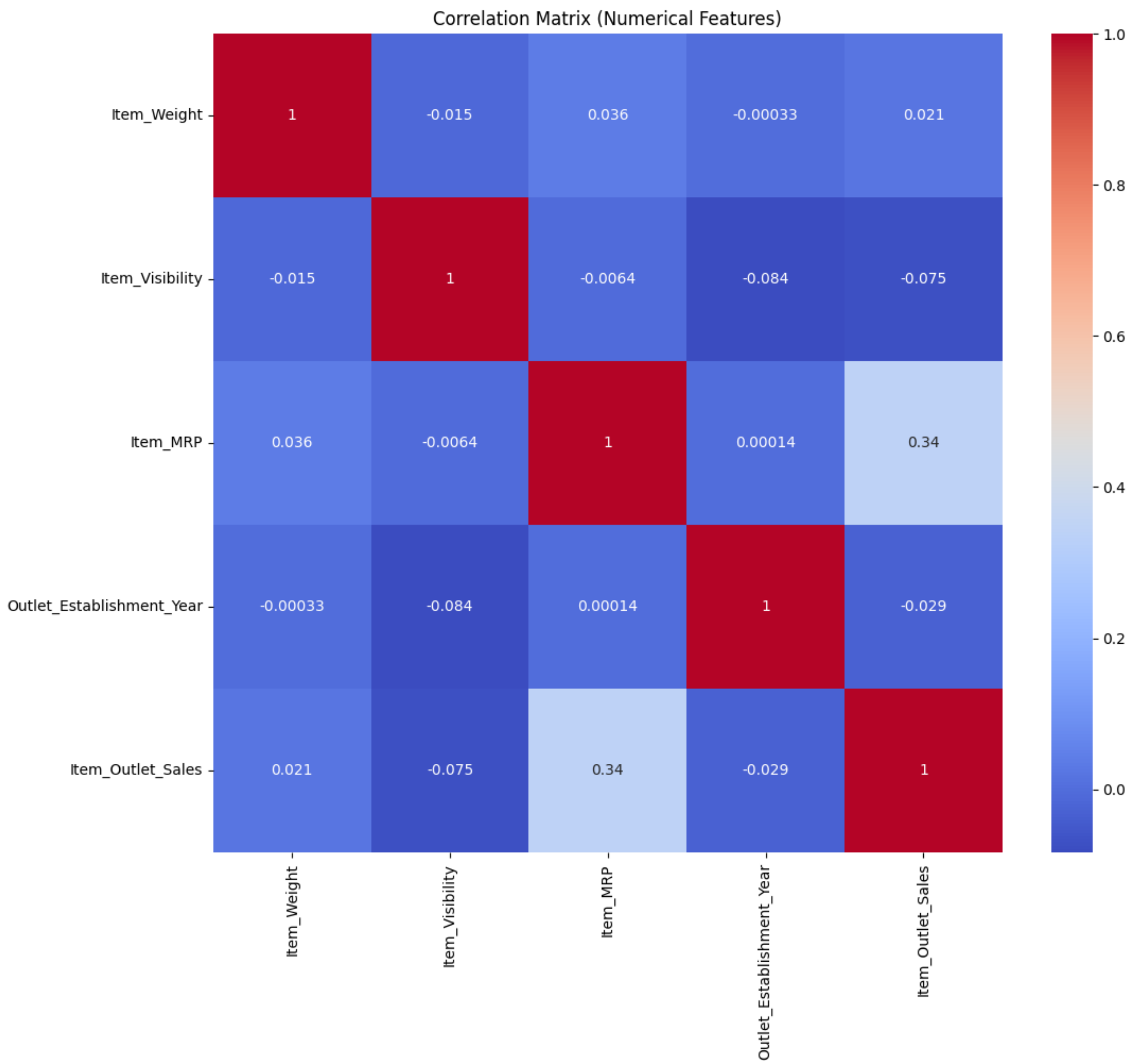



Bar plot for Item_Type
0
200
400
600
800
1000
1200
Fruits and Vegetables
Snack Foods
Household
Frozen Foods
Dairy
Canned
Baking Goods
Health and Hygiene
Soft Drinks
Meat
Breads
Hard Drinks
Others
Starchy Foods
Breakfast
Seafood
Item_Type


Bar plot for Outlet_Type
0
1000
2000
3000
4000
5000
Supermarket Type1
Grocery Store
Supermarket Type3
Supermarket Type2
Outlet_Type

4.4 Feature Engineering

- Created new features like *Item_Type_new, Item_category, Outlet_Years, price_per_unit_wt, Item_MRP_clusters, Discount_Percentage, Weekend_Sales_Boost*.
- Used **Principal Component Analysis (PCA)** to reduce dimensionality.
- Created lag variables to incorporate past sales trends.



5. Model Selection and Development

5.1 Statistical Methods

- **ARIMA** (AutoRegressive Integrated Moving Average) – Time series forecasting.

- **Exponential Smoothing** – Capturing trends and seasonality.
- **Holt-Winters Method** – Seasonal adjustments.

5.2 Machine Learning Models

- **Linear Regression** – Establishing relationships between features.
- **Decision Trees & Random Forests** – Handling nonlinear interactions.
- **XGBoost & LightGBM** – High-performance boosting algorithms.
- **Support Vector Regression (SVR)** – Handling outlier sensitivity.

5.3 Deep Learning Models

- **Recurrent Neural Networks (RNN)** – Sequential data dependencies.
- **Long Short-Term Memory (LSTM) Networks** – Handling long-term dependencies.
- **Transformer Models** – Enhanced forecasting using attention mechanisms.

6. Model Training and Evaluation

6.1 Training Process

- Splitting data into training, validation, and test sets.
- Hyperparameter tuning using **Grid Search & Bayesian Optimization**.

6.2 Evaluation Metrics

- **Mean Absolute Error (MAE)** – Measures prediction accuracy.
- **Root Mean Squared Error (RMSE)** – Penalizes large errors.
- **Mean Absolute Percentage Error (MAPE)** – Percentage-based accuracy.
- **R-squared Score** – Model fit assessment.

Key Findings:

- Identified *Item_MRP, price_per_unit_wt, Outlet_Years*, and *Item_MRP_Clusters* as key features.
- **XGBoost** outperformed other models with the best RMSE score.

7. Model Deployment and Maintenance

7.1 Deployment

- **Dockerized application** for easy deployment.
- **CI/CD pipelines** for automated updates.
- **Cloud-Based API** using Flask or FastAPI.

7.2 Maintenance

- Continuous data updates and model retraining.
- Real-time monitoring and anomaly detection.
- Model versioning for performance tracking.

8. Model Ensembling

- **Objective**: Improve overall prediction accuracy.
- **Techniques**: Stacking, bagging, boosting, blending.

Additional Considerations

- **Explainability**: Using SHAP values for feature importance analysis.
- **Hyperparameter tuning**: Optimized configurations using **Optuna**.
- **Ethical Considerations**: Ensuring fair and unbiased predictions.

9. Conclusion

By combining **statistical, machine learning, and deep learning techniques**, this project ensures **robust and accurate sales forecasting**. The developed application empowers retailers with **data-driven insights**, leading to:

- Improved **inventory management**.
- **Cost reduction**.
- Increased **profitability**.
- **Better demand planning**.

10. Future Scope

- Integration with **real-time data pipelines**.
- Incorporation of **external economic indicators**.
- **Reinforcement learning** for adaptive forecasting.