Supermarket Sales Analysis and EDA : Final Report

1. Introduction & Problem Statement

Every day, supermarkets produce enormous volumes of sales data that document seasonal patterns, product demand, and consumer preferences. Using exploratory data analysis (EDA) to analyze this data reveals important insights that inform business choices. Supermarkets, however, frequently struggle to forecast sales patterns, effectively manage inventory, and pinpoint the main variables affecting profits. Inaccurate demand forecasting, overstocking, and stockouts can all have a detrimental effect on company performance if they are not properly analyzed.

This study aims to analyze supermarket sales data in order to find important trends, spot irregularities, and determine how various attributes—like product category, price, and store type—relate to one another. We hope to offer practical insights that supermarkets can use to improve marketing tactics, streamline operations, and increase overall profitability by utilizing EDA techniques.

2. Dataset Overview

5,681 records with 11 attributes pertaining to retail product sales and outlet characteristics make up the dataset. Product-specific information like `ProductID`, `Weight`, `FatContent`, `ProductVisibility`, `ProductType`, and `MRP` are included, as are outlet-related characteristics like `OutletID`, `EstablishmentYear`, `OutletSize`, `LocationType`, and `OutletType`. Interestingly, there are 2,582 missing values in the dataset, especially in columns like `Weight` and `OutletSize`. These may need to be preprocessed in order to be analyzed accurately. Pricing strategies and inventory management may benefit from the dataset's apparent utility for examining sales trends, product performance, and outlet efficacy.

3. Technology Stack

Programming Language : Python

Libraries & Tools: Pandas, NumPy, Matplotlib, Seaborn, Scikit (Classification Model - Random Forest)

4. ML Model Implementation & Evaluation

Four machine learning models—K-Nearest Neighbors (KNN), Decision Tree, Random Forest, and Gradient Boosting—are used in the provided code to classify supermarket sales data. First, the dataset is preprocessed by encoding categorical variables and splitting it into training and test sets. Each model is trained and evaluated using classification reports, confusion matrices, and accuracy scores. Random Forest and Gradient Boosting usually provide higher accuracy because of ensemble learning, but KNN and Decision Tree offer solutions that are simpler to comprehend. The analysis helps identify the best model for product category prediction by using a range of features.

5. Results & Insights

The analysis shows that Random Forest and Gradient Boosting are the most accurate models for supermarket sales classification. MRP, Product Type, and Outlet Type were the most crucial factors in predicting sales. However, inconsistent FatContent labeling and missing values in Weight and OutletSize necessitated preprocessing. ProductVisibility and MRP outliers may have had an impact on the model's performance. Overall, simplifying feature engineering and data preprocessing can further improve accuracy. For more precise forecasts, future studies might look into deep learning models .

6. Challenges & Future Improvements

The dataset has issues that can affect model accuracy, including missing values in Weight and OutletSize, inconsistent labeling in FatContent, and imbalanced product categories. Furthermore, weak feature correlations decrease effectiveness, and outliers in ProductVisibility and MRP may skew predictions. Future research can concentrate on imputation for handling missing values, categorical encoding refinement, hyperparameter tuning, and deep learning model exploration to enhance performance. Supermarket sales forecasting can be made even more reliable by implementing real-time predictions using cloud solutions and balancing data with SMOTE.

7. Conclusion & Learnings

The study discovered that Random Forest and Gradient Boosting were the most effective models for categorizing supermarket sales, with MRP, Product Type, and Outlet Type acting as significant predictors. Data preprocessing, which included handling outliers, inconsistent labels, and missing values, had a significant effect on model performance. Ensemble models outperformed simpler classifiers, indicating the importance of combining multiple learners. Key takeaways include the necessity of feature selection, data cleaning, and outlier detection to improve accuracy. Future developments in real-time prediction, deep learning, and cloud deployment can further enhance supermarket sales forecasting.

8. References

Kuila, A. (n.d.). *Big Mart Sales Dataset*. Kaggle. Retrieved from https://www.kaggle.com/datasets/akashdeepkuila/big-mart-sales/data