**Big-Mart Sales Prediction: Approach and Methodology**

**Problem statement**

The Big Mart sales prediction challenge required developing a machine learning solution to predict item sales across various outlets. My approach combined thorough exploratory data analysis, data preprocessing, and a comprehensive model evaluation to achieve high predictive accuracy.

**Data Understanding**

I began by examining the dataset structure, which contained 8523 training samples and 5681 test samples with features describing items and outlets. Initial analysis revealed missing values in 'Item\_Weight' (17.2%) and 'Outlet\_Size' (28.3%), which required careful imputation. The target variable 'Item\_Outlet\_Sales' showed a right-skewed distribution with a mean of approximately 2181 and values ranging from 33 to 13086.

**Exploratory Data Analysis**

My exploratory analysis help me to find several key insights:

Supermarket Type 3 outlets had the highest average sales

A strong positive correlation existed between Item\_MRP and sales

Certain item types (Starchy Foods, Seafood, Fruits and Vegetables) consistently outperformed others

Outlet establishment year showed interesting patterns with newer and very old outlets performing differently

Zero visibility values appeared suspicious and required investigation

I visualized these relationships through comprehensive plots including histograms, bar charts, scatter plots, and correlation heatmaps to guide feature engineering decisions.

**Data Preprocessing Strategy**

My preprocessing pipeline addressed several challenges:

1. Missing Values: Implemented a multi-tiered approach for 'Item\_Weight' using item-level, category-level, and global means. For 'Outlet\_Size', I used context-aware imputation based on outlet type and location.

2. Standardization: Unified inconsistent categorical values (e.g., 'low fat', 'Low Fat', 'LF' → 'Low Fat')

3. Feature Engineering: Created 15+ new features including:

- Time-based features (outlet age, age categories)

- Item categorization (from identifier patterns)

- Price-weight relationships

- Visibility adjustments for zero values

- Market positioning based on price and fat content

- Interaction features combining outlet and item characteristics

4. Encoding: Applied label encoding to all categorical variables after standardization

**Modeling Approach**

I implemented a diverse ensemble of models to capture different patterns in the data:

- Tree-based models: XGBoost, HistGradientBoosting, RandomForest, AdaBoost

- Linear models: LinearRegression, Ridge, HuberRegressor, TheilSenRegressor

- Other approaches: KNN, Neural Network (MLP)

Also worked on many trials and error to get best hyper parameters for the models

Each model was evaluated using a consistent validation framework with metrics including RMSE, R², and MAE. MLP consistently outperformed other models, likely due to its ability to handle non-linear relationships and interactions between features.

Feature Importance Analysis

Feature importance analysis from the best model revealed that store-level statistics, MRP, and outlet type were the strongest predictors. This aligned with business intuition that pricing strategy and store characteristics heavily influence sales performance.

**Final Results**

The final solution achieved strong predictive performance with RMSE(1145.09) significantly below baseline models. The comprehensive approach combining domain-specific feature engineering with advanced modelling techniques proved effective for this retail sales prediction challenge.