**Big Mart Sales Forecasting: Approach & Findings**

**1. Item Data Analysis & Preprocessing**

* **Handling Missing Values:**
  + Item\_Weight and Outlet\_Size had missing values, which were imputed using median values.
  + Certain items (FDN52, FDK57, FDE52, FDQ60) lacked weight information, so their values were filled using the median weight of their respective Item\_Type.
* **Category Standardization:**
  + The Item\_Fat\_Content column had multiple representations of the same categories, which were standardized as follows:
    - Low Fat → Low Fat, low fat, LF
    - Regular → Regular, reg
* **Feature Engineering:**
  + Converted Outlet\_Establishment\_Year into **Outlet Age**.
  + Created new features such as:
    - Total Unique Items per Outlet
    - Average Item MRP per Outlet Type
    - MRP per Visibility
    - Interaction Terms like MRP × Outlet Size

**2. Categorization Using Clustering**

* Tested various clustering techniques, including **K-Means, Hierarchical Clustering, DBSCAN, and Gaussian Mixture Models (GMM)**.
* The best-performing algorithm was **K-Means with 4 clusters**, chosen based on the **highest Silhouette Score**.
* Used the generated cluster labels as a **new feature for training regression models**.

**3. Feature Encoding & Dimensionality Reduction**

* Applied **One-Hot Encoding** to all categorical variables.
* Performed **Principal Component Analysis (PCA)** to reduce dimensionality and retain only the most important components.

**4. Model Training & Feature Selection**

* Trained multiple regression models, using **Recursive Feature Elimination (RFE)** for feature selection. The models included:
  + **Linear Models:** Ridge Regression, Lasso
  + **Tree-Based Models:** Random Forest, XGBoost, LightGBM, CatBoost
  + **Neural Networks:** MLP Regressor
  + **Boosting Models:** Gradient Boosting, AdaBoost, HistGradientBoosting
  + **Polynomial Regression:** Degree 2 and 3
* Selected the **top three models** based on the lowest **Root Mean Squared Error (RMSE)**.

**5. Stacked Ensemble Modeling & Experimentation**

* Developed a **Stacked Ensemble Model**, combining:
  + **CatBoost**
  + **Gradient Boosting Regressor (GBR)**
  + **Quadratic Regression**
* Conducted multiple experiments:
  + Tested different K-Means cluster combinations.
  + Applied PCA after feature selection.
  + Compared **Stacked Ensemble Models vs. Weighted Ensemble Models**.
  + Performed hyperparameter tuning using **Random Search, Optuna, and Bayesian Optimization**.

**6. Handling Skewness in Data**

* Transformed skewed features such as **Item Visibility, Item MRP, and Item Sales** to normalize distributions.

**7. Key Takeaways: What Worked Best?**

* **Data Cleaning**: Handling missing values and standardizing categories.
* **Using Clustering as a Predictor**: K-Means clustering improved performance.
* **One-Hot Encoding**: Beneficial for categorical features.
* **Feature Selection via RFE**: Helped refine the feature set.
* **Stacked Ensemble Model**: Combining **CatBoost, GBR, and Quadratic Regression** yielded the best results, especially with **Random Search hyperparameter tuning**.