# **BigMart Sales Prediction – One Pager**

## **Objective**

Predict **Item Outlet Sales** using product and outlet-level features to help retailers understand drivers of sales and optimize inventory, pricing, and marketing strategies.**Data Understanding**

Dataset contains:

* **Item features**: Identifier, Weight, Fat Content, Visibility, Type, MRP.
* **Outlet features**: Identifier, Establishment Year, Size, Location Type, Outlet Type.
* **Target**: Item\_Outlet\_Sales.

Challenges observed:

* Missing values (Item\_Weight, Outlet\_Size).
* Inconsistent categories (Item\_Fat\_Content).
* Skewed variables (Item\_Visibility).

## **Approach & Thought Process**

1. **Preprocessing**
   * Imputed missing values (mean for weight, mode for outlet size based on outlet type).
   * Standardized categorical labels in Item\_Fat\_Content
   * Transformed Item\_Visibility into log and treated -inf values with 25 %ile
2. **Feature Engineering**
   * Buckets features for MRP, Visibility and Weight
   * Broad categories (Food, Drinks, Non-Consumables), target encoding for Item\_Type and Outlet age from their Establishment years
3. **Encoding Strategy**
   * Applied Label Encoding (IDs) + One-Hot Encoding (Other Categorical features )
4. **Modeling**
   * Baseline: Linear Regression, Ridge/Lasso.
   * Advanced: Random Forest, XGBoost, CatBoost.
   * Final: **CatBoost Regressor** chosen for its ability to handle categorical data efficiently.
5. **Experimentation**
   * Evaluated models using **RMSE** and **R²**.
   * Applied **Hyperopt** for CatBoost hyperparameter tuning (depth, learning rate, l2\_leaf\_reg, bagging temperature).
   * Derived features from MRP and Outlet Age significantly reduced RMSE.

## **Results & Learnings**

* CatBoost outperformed linear and tree-based baselines with **lowest RMSE**.
* Feature engineering (esp. MRP transformations & Outlet features) had the **biggest impact**.
* Target encoding for high-cardinality variables improved stability.