# Approach note — Big Mart Sales Prediction

**Summary:**

- **Objective** was to get minimum RMSE score by using **cleaned** and **meaningful** features, testing with **multiple models** and **hyper parameter tuning**.

- **Models tried**: CatBoost, XGBoost, RandomForest.

RandomForest was selected as best model, therefore hyper-parameter tuning was only applied on RF in interest of compute time and code complexity.

**Data cleaning & missing value treatment**

- **Inspected** types, distributions and missing counts.

- **Item Visibility:** “Visibility Was zero” flag was added where Item visibility was 0. Imputed by median at Item level, then global median fallback.

- **Item Weight**: Imputed by median at Item level, then global median fallback.

- **Outlet Size**: recorded missing indicator “Outlet Size Missing”. Imputed by mode within (Outlet\_Type, Outlet\_Location\_Type), then global mode.

- **Standardized Item\_Fat\_Content variants:** (e.g., "lf" → "Low Fat", "reg" → "Regular")

**Feature engineering**

- **Extracted Item\_Category** as first two chars of Item\_Identifier (Food/Drinks/Non consumables).

- **Calculated Age of Outlet** as 2013 − Outlet\_Establishment\_Year.

- **store\_count:** added the number of outlets the item was sold in as a proxy for popularity.

- **Price\_per\_gram** = Item\_MRP / Item\_Weight

- **Kept indicators for imputation events:** (Visibility\_WasZero, Outlet\_Size\_Missing) to capture information loss patterns.

**Evaluation**

- For each model: OOF predictions collected across folds, fold RMSEs printed, and overall CV RMSE computed from OOF.

- Best model chosen by lowest OOF RMSE.

**Hyperparameter tuning**

- Selected RandomForest for tuning (compute-efficient and chosen as best).

- RandomizedSearchCV with custom param distribution (n\_estimators, max\_depth, min\_samples\_split/leaf, max\_features, bootstrap).

- Refit and saved best\_estimator\_.

**Key observations & experiments**

- Handling zero visibilities improved signal when combined with indicator.

- Per-item medians for weight/visibility preserve item-specific patterns vs. global imputation.

- CatBoost can handle categories directly; however, RF performed competitively for this dataset and compute budget.

- Randomized search (limited iterations) provided modest gains without exhaustive compute.

**Next steps / improvements**

- Expand tuning for all models (XGBoost/LightGBM/CatBoost).

- Add more models into the mix.

- Model explainability to validate feature importances and refine features.