**Big Mart Sales Prediction Approach – Akshay**

**1. Exploratory Data Analysis (EDA):**   
I started with automated EDA using pandas profiling and Sweetviz, quickly spotting missing values and inconsistent categories (like standardizing ‘LF’ to ‘Low Fat’). The target (Item\_Outlet\_Sales) showed strong right skewness, confirmed non-normal by KS and Anderson-Darling tests. Correlation analysis (Pearson, Spearman, and Kendall) and scatter/residual plots showed that the data is non-linear and Heteroscedastic—linear models would likely underperform.

**2. Feature Engineering & Selection:**   
**Missing values -** Item\_Weight values were imputed using known weights per Item\_Identifier or, if needed, group medians based on visibility and MRP bins. Outlet Size was not important and hence imputation was not needed.

**Feature engineering** focused on non-linear binning (Visibility/MRP/Weight), adding temporal features (Outlet Age, Decade), and aggregating sales means by key groups to capture business patterns. Feature importance with ExtraTrees and RandomForest, and Correlation highlighted that Item\_MRP, Outlet\_Type, Item\_Visibility, Outlet\_Identifier, and Item\_Identifier as the most predictive input features. After iterative validation, these formed the core feature set (removing created features, Item Weight etc.). Surprisingly Item Identifier didn’t have much impact on predictions, but they helped reduce data wrangling later and in interpretation. Standard scaling is not needed when using Tree based models, however removing it showed reduced accuracy and higher RMSE on the Training and Test dataset.

**3. Outlier Handling:**   
Boxplots and IQR confirmed many outliers in sales. Experiments with outlier removal or capping actually hurt test performance, so outliers were retained—this proved crucial, as retail data often includes rare but important events.

**4. Model Benchmarking & Tuning:**   
Tried a range of models; tree-based (Gradient Boosting, LightGBM, CatBoost) vastly outperformed linear models. Hyperparameter optimization (Optuna, scikit-optimize) and KFold CV focused on LightGBM, using the Tweedie objective for skewed data. LightGBM delivered the lowest RMSE (≈1133) on the test set. Random seed used was 7 to deliver consistent and reproducible results.

**5. Final Model & Interpretation:**   
The final pipeline included robust preprocessing and tuned LightGBM. Though GBM showed better and lower RMSE during training, LightGBM was consistent in Training and on the Test dataset. SHAP values confirmed price (Item MRP), outlet type, and identifier as key drivers, aligning with retail intuition and boosting model trust. SHAP can be used to remove features in the future for better accuracy.

**Conclusion:**   
Retaining outliers, robust feature engineering and selection, and **LightGBM with Tweedie** loss and fine-tuned parameters were instrumental to success. LightGBM’s speed, handling of skew, and interpretability made it the clear winner in the end achieving a rank of 15 and an RMSE of 1133.28 on the Test dataset.