* Performed thorough evaluation of the dataset and identified missing values and key patterns.
* Started with basic preprocessing:
* Applied label encoding to categorical variables.
* Used Gradient Boosting Regressor as the baseline model.
* Imputed missing categorical values with "Unknown" and numeric values with group-wise median.
* To improve RMSE and R²:
* Designed logical rules for imputing missing values in Outlet\_Size and Item\_Weight.
* This significantly improved model performance.
* Switched to **CatBoost Regressor** due to high number of categorical features.
* Tuned hyperparameters for better performance.
* CatBoost showed better RMSE than Gradient Boosting.
* Tried log transformation on target variable (highly right-skewed).
* Did not help; worsened performance.
* Created engineered features to test further improvement.
* Even though Gradient Boosting had lower RMSE in training, it gave higher submission errors.
* Checked feature importance of features
* **Final Model:** CatBoost Regressor

Checked feature importance and selected 4 most important features:

["Item\_MRP","Outlet\_Type", "Outlet\_Identifier", "Outlet\_Location\_Type"]

Also used entire data available for training, without train\_test\_split:

**Final RMSE: 1145.3081062456**

**Rank: 171**