Big Mart Sales Prediction: A Comprehensive ML Pipeline

1. Data Understanding Through Comprehensive EDA

Exploratory Data Analysis revealed critical patterns that shaped our modeling strategy:

- `Item_MRP` **is the dominant predictor**: Strong correlation (0.568) with sales, showing clear tiered pricing effects.
- **Outlet heterogeneity is extreme**: Sales vary dramatically across outlets (OUT027: 3694 avg vs OUT010: 339 avg), making outlet features crucial.

Our core insight: **Target variable is highly skewed (skewness: 1.177), while missing values and outlet variations demand intelligent preprocessing.**

2. Data Preprocessing Pipeline

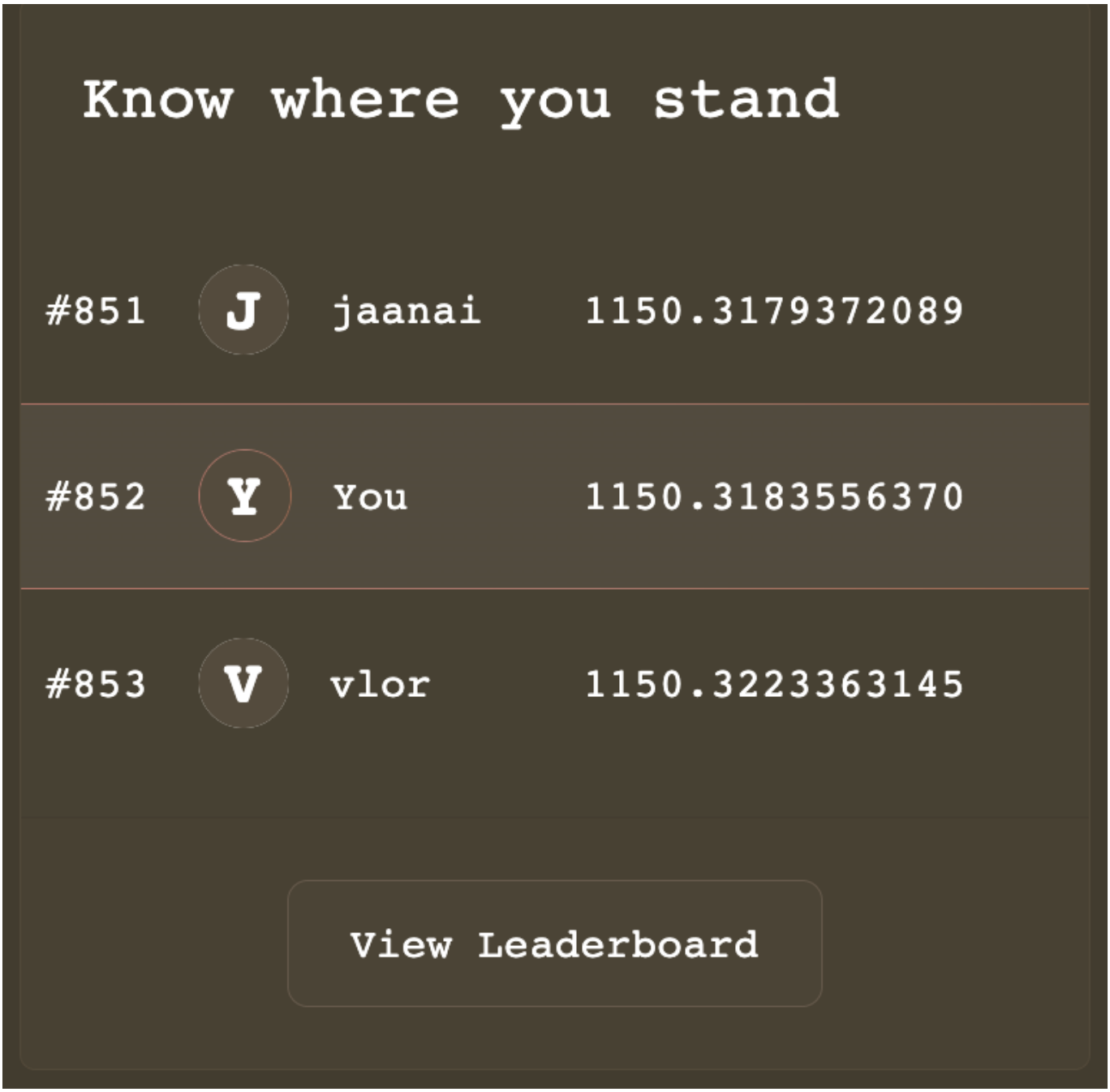
I developed a comprehensive `DataProcessor` class with intelligent feature handling:

- **Smart Missing Value Imputation**:
  - `Item_Weight`: Group-based imputation using `Item_Identifier` patterns.
  - `Outlet_Size`: Mode imputation based on `Outlet_Type` and `Location_Type` combinations.
  - `Item_Visibility`: Zero values replaced with `Item_Type`-specific medians.
- **Advanced Feature Engineering**:
  - **Target Encoding for Outlets**: Converted high-cardinality `Outlet_Identifier` into mean-encoded performance scores.
  - **Temporal Features**: Calculated `Outlet_Age` from establishment year.
  - **Categorical Binning**: Created `MRP_Category` to capture pricing tier effects.

3. Ensemble Modeling Pipeline

- **Three-Model Weighted Ensemble**: Optimized combination for robust predictions:
  1. **RandomForest** (40% weight): Captures non-linear interactions and feature importance.
  2. **XGBoost** (40% weight): Gradient boosting for complex pattern recognition.
  3. **Ridge Regression** (20% weight): Provides stable linear baseline and regularization.
- **Hyperparameter Optimization**: GridSearchCV across all models with 5-fold cross-validation.

This systematic pipeline, from comprehensive EDA through sophisticated preprocessing to tuned ensemble modeling, demonstrates a production-ready approach to sales prediction with robust evaluation metrics and scalable architecture.


Know where you stand
#851 J jaanai 1150.3179372089
#852 Y You 1150.3183556370
#853 V vlor 1150.3223363145
View Leaderboard

just a note: scores in leaderboard and upload is slightly different because they are screenshotts of same output submitted twice