BigMart Sales Prediction

Problem Overview

The objective was to predict product level sales across multiple BigMart outlets using historical data containing both product attributes and outlet characteristics. The focus was not only on building an accurate model but also on understanding the key factors influencing sales.

Key Insights from EDA

- From the Item Identifier, three major product groups were identified, and a new feature **Item_Category** was created.
- Supermarket Type3 had the highest median sales, while Grocery Stores consistently underperformed, indicating that outlet format plays a significant role in revenue generation.
- Medium-sized outlets performed better than large outlets, showing that bigger size does not guarantee higher sales.
- Sales variation across **Outlet_Type** was stronger than across **Outlet_Size**, highlighting the importance of store structure over physical scale.
- Non-Consumables and Food categories performed better than Drinks. Snack Foods emerged as a reliable high performing category due to balanced sales volume and revenue.
- **Item_MRP showed a moderate positive correlation (0.57) with sales**, making pricing the strongest linear predictor, while other numerical features had limited influence.

Feature Engineering

- Based on these insights, I engineered key features including:
- Item_Category
- Item_Age (from establishment year)
- Visibility correction and normalized Visibility_MeanRatio
- MRP_Band for price segmentation
- Cleaning categorical inconsistencies
- These steps improved data quality and strengthened predictive signals.

Modeling & Results

Random Forest, XGBoost, and CatBoost were evaluated using consistent 5-fold cross-validation. After hyperparameter tuning, **CatBoost achieved the lowest CV RMSE (approx. 1075)** and demonstrated stable performance.

The final submission achieved a leaderboard RMSE of approx. 1145, securing Rank **227**.