

Big Mart Sales Prediction Challenge: One-Page Approach Summary

Author: Ishaan Jain

Challenge: Big Mart Sales Prediction (Analytics Vidya)

Final Rank: 1373

Final Model RMSE: 1152.88

1. Objective

To build a regression model to accurately predict product sales across various store outlets (Evaluation Metric: RMSE). This document summarizes the iterative approach, which reduced the RMSE from **1525** to **1152.88**.

2. Iterative Model Development

The final model was the result of a five-version process centered on a LightGBM regressor. The diagnostic and tuning process is summarized below.

Version	Key Actions & Insights	Result & RMSE Score
V1	Baseline: LightGBM with default parameters. No feature engineering, imputation, or validation.	1525 <i>Insight: Model was severely overfitting.</i>
Data Analysis	(Pre-V2) Performed multicollinearity checks (Pearson, Spearman) and SHAP analysis.	<i>Insight: Found Outlet_Type & Outlet_Location_Type were collinear. SHAP showed Outlet_Type had negligible importance.</i>
V2	Addressed Overfitting: <ul style="list-style-type: none">Reduced model complexity (<code>max_depth, num_leaves</code> ↓).Added <code>early_stopping</code> with a validation set.	1162 (Change: -363) <i>Result: Confirmed overfitting was the primary issue.</i>
V3	Data Enhancement: <ul style="list-style-type: none">Added missing value imputation (for <code>Item_Weight, Outlet_Size</code>).	1162 (Change: 0) <i>Result: No score change, but validated model stability.</i>

	<ul style="list-style-type: none"> Added new features (outlet age, etc.). 	
V4	<p>Fine-Tuning:</p> <ul style="list-style-type: none"> Lowered learning_rate for more precise steps. Further reduced max_depth & num_leaves. 	1157 (Change: -5) <i>Result: Gained from patient, iterative refinement.</i>
V5	<p>Objective Optimization:</p> <ul style="list-style-type: none"> Changed objective from 'regression' to 'tweedie'. <i>Insight: Tweedie is better suited for non-negative, skewed sales data.</i> 	1152.88 (Change: -4.12) <i>Result: Final model, secured by matching the objective to the data's nature.</i>

3. Key Strengths of the Approach

- Systematic & Iterative:** Followed a logical flow (Baseline → Diagnose → Fix → Enhance → Tune).
- Data-Driven Diagnosis:** Used SHAP and correlation matrices to guide decisions, not guesswork.
- Targeted Problem Solving:** Solved the biggest problem (overfitting) first, which yielded the largest gain.
- Understanding the "Why":** Aligned the model's objective (tweedie) with the statistical properties of the target variable (sales).

4. Conclusion

The final score was achieved through a structured process of identifying and solving the most significant problem first (overfitting), followed by methodical data enhancement and parameter tuning.