# 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**  *Insight: Model was severely overfitting.* |
| **Data Analysis** | (Pre-V2) Performed multicollinearity checks (Pearson, Spearman) and SHAP analysis. | *Insight: Found Outlet\_Type & Outlet\_Location\_Type were collinear. SHAP showed Outlet\_Type had negligible importance.* |
| **V2** | **Addressed Overfitting:**  • Reduced model complexity (max\_depth, num\_leaves ↓).  • Added early\_stopping with a validation set. | **1162** (Change: **-363**)  *Result: Confirmed overfitting was the primary issue.* |
| **V3** | **Data Enhancement:**  • Added missing value imputation (for Item\_Weight, Outlet\_Size).  • Added new features (outlet age, etc.). | **1162** (Change: 0)  *Result: No score change, but validated model stability.* |
| **V4** | **Fine-Tuning:**  • Lowered learning\_rate for more precise steps.  • Further reduced max\_depth & num\_leaves. | **1157** (Change: -5)  *Result: Gained from patient, iterative refinement.* |
| **V5** | **Objective Optimization:**  • Changed objective from 'regression' to **'tweedie'**.  • *Insight: Tweedie is better suited for non-negative, skewed sales data.* | **1152.88** (Change: -4.12)  *Result: Final model, secured by matching the objective to the data's nature.* |

### 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.