BigMart Sales Prediction - Approach Note

Achievement: Rank #250 | RMSE: 1146.10 | Top 5%

Strategic Philosophy

Core Approach: Retail domain expertise + systematic optimization beats algorithmic complexity.

Hypothesis: Clean, business-logical features would outperform complex models on poorly understood data.

Phase 1: Domain-Driven Discovery

Critical Insights:

- Data quality issues: 17% missing weights, 28% missing outlet sizes, 526 zero-visibility items
- Business violations: Inconsistent labels, non-edible items with fat content
- **Key Pattern**: Outlet hierarchy - Supermarket Type3 (₹3,694) > Type1 (₹2,316) > Grocery (₹340)
- **Breakthrough Hypothesis**: Store assortment diversity = primary sales driver

Phase 2: Strategic Feature Engineering

Created 16 Business-Driven Features:

- **Temporal**: Outlet_Age, maturity groups (customer loyalty patterns)
- **Economic**: Price_per_Weight, MRP_Category, competitive positioning
- **Business Logic**: Food vs Non-Food, Perishability (inventory patterns)
- **Market Structure**: **Outlet_Item_Diversity** (assortment breadth), penetration metrics
- **Interactions**: Category-channel synergies

Rule: Every feature required business rationale + performance improvement.

Phase 3: Algorithm Selection

Systematic Testing: Linear models failed (Ridge: 1276 RMSE). **CatBoost won** (1147 RMSE) due to categorical handling + overfitting protection.

Baseline: 1151.0 RMSE with conservative parameters.

Phase 4: Learning Through Failures

Strategic Failures:

- Multi-Algorithm Ensemble (LightGBM+CatBoost+GB): +2.1 RMSE → Weaker algorithms dilute performance
- Mixed Ensemble (CatBoost+Ridge+Lasso): +2.2 RMSE → Components need performance parity
- Target Encoding: +5.0 RMSE → Data leakage despite precautions
- Aggressive Tuning: +4.4 RMSE → Conservative changes outperform

Key Learning: Algorithm diversity ≠ improvement. Quality over quantity.

Phase 5: Breakthrough - Seed Ensemble

Strategic Pivot: Model diversity through random seeds, not algorithms.

Discovery: Seeds 46 (1147.38) + 48 (1147.84) with 50.8%/49.2% weights → **1146.85 RMSE**

Why This Worked: Same algorithm, different patterns = optimal diversity without performance sacrifice.

Phase 6: Micro-Optimization

Conservative Refinement: iterations +5, learning_rate -0.005, depth -1 → **1146.10 RMSE**

The Game-Changer

Feature Importance Results:

1. **Outlet_Item_Diversity: 47.11%** - Store assortment breadth dominates
2. Item_MRP: 17.41% - Price foundation
3. Price_Rank_in_Category: 8.86% - Competitive positioning
4. Temporal features: 12.06% - Store lifecycle

Breakthrough: Single engineered feature = nearly half of predictive power!

Performance Journey

```
Baseline (1151.0) → Features (1148.5) → Failures (1149-1154) → Recovery (1148.87) → Ensemble (1146.85) → Final (1146.10)
```

Results: 4.9 RMSE improvement, Rank ~800 → #250 (550 positions up)

Success Factors

What Worked:

- Domain expertise driving feature engineering

- Systematic methodology with validated improvements
- Conservative optimization over aggressive tuning
- Homogeneous high-quality ensembles
- Learning from systematic failures

What Failed:

- Algorithm diversity ensembles
- Aggressive parameter changes
- Statistical feature selection over domain knowledge

Key Insights

Strategic Differentiators:

1. **Outlet_Item_Diversity Discovery**: Store assortment as hidden competitive advantage
2. **Failure-Driven Learning**: Setbacks guided optimal approach
3. **Business Logic First**: Intuitive features outperformed statistical artifacts
4. **Conservative Excellence**: Small systematic improvements compounded

Core Learning: Deep domain understanding + disciplined ML execution beats algorithmic sophistication. Retail assortment diversity dominated prediction, proving business insight trumps complexity.

Validation: 5-fold CV confirmed 1146.23 (+/- 2.84) stable performance.

Conclusion

Methodology Success: Systematic retail expertise integration achieved top 5% through discovering that store assortment breadth is the primary sales driver - a business insight that became 47% of predictive power.

Achievement: Rank #250 via domain mastery + methodical optimization, not algorithmic complexity.