BigMart Sales Prediction: Project Development Approach

Project Development Timeline:

- Exploratory Data Analysis: Identified right-skewed sales distribution, missing value patterns, and performance differences across outlet types and item categories
- Missing Value Strategy: Evolved from simple mean imputation to hierarchical approach using item-outlet combinations with fallbacks to item type averages
- Initial Feature Engineering: Created competition-based features, cross-product relationships, cannibalization metrics, and outlet-level characteristics driven by retail domain knowledge
- Compared target encoding (with cross-validation) against label encoding, choosing target encoding since it gave marginally better results.
- Evaluated log transformation versus Box-Cox transformation, selecting Box-Cox for automatic parameter optimisation
- Built weighted ensemble of Random Forest, XGBoost, and LightGBM with performance-based weight allocation
- Residual analysis revealed systematic underperformance on low-sales items, leading to implementation of weighted RMSE loss function
- Concentrated on Random Forest after consistent superior performance, removing other ensemble components for optimisation efficiency
- Implemented Bayesian optimisation through Optuna for systematic parameter space exploration
- Developed three segmentation strategies (outlet type, individual outlets, price bins) with performance-weighted combination
- Applied two-stage feat selection process combining importance scoring and recursive elimination, customized per segment
- Tested equal-width versus quantile-based binning strategies for price segment models
- Achieved substantial improvements through segmented approach with weighted RMSE addressing low-sales prediction concerns, picked the from a selection of individual segment models or their weighted ensemble.

Problem Understanding & Initial Analysis

My exploration of the BigMart dataset revealed a heavily right-skewed sales distribution with most items showing low sales and occasional high performers. Missing values in item weights and outlet sizes required strategic treatment beyond simple imputation. The data showed clear patterns where supermarkets consistently outperformed grocery stores, and certain item categories like fruits and vegetables exhibited higher sales variability. These insights shaped my decision to eventually pursue segmented modelling rather than a single global approach.

Data Preprocessing

I started with basic mean imputation but realised this approach ignored crucial business relationships. For item weights, I developed a hierarchical strategy that first attempted imputation using item-outlet combinations, then fell back to item type averages when specific combinations weren't available. This preserved natural product relationships while handling edge cases. Outlet size imputation leveraged business rules derived from analysing relationships between outlet types and locations, ensuring imputations reflected real retail patterns.

Feature Engineering

Feature creation process was driven by retail domain knowledge. I built competition-based features including price rankings within categories and cheaper alternative counts to capture product positioning effects. Understanding shopping behaviour led to creating cross-product relationship features for complementary items like dairy and breakfast products. Cannibalization features quantified competitive pressure through price gap calculations and substitution intensity metrics. Outlet-level features captured market penetration and assortment diversity, recognising different customer demographics across store formats.

Modelling Journey and Performance Insights

I initially implemented an ensemble of Random Forest, XGBoost, and LightGBM with performance-weighted averaging. Random Forest consistently outperformed the others, leading me to focus solely on this algorithm while investing effort in other optimisation areas. Detailed residual analysis revealed models systematically underperformed on low-sales items while achieving reasonable accuracy on high-sales products. This prompted me to implement weighted RMSE loss function that emphasised prediction accuracy proportional to sales magnitude, better aligning model objectives with business impact.

Segmented Modelling

Error analysis across different business dimensions showed that a global model struggled with diverse patterns across outlet types, price segments, and individual stores. This insight drove the development of three segmentation strategies: outlet type grouping for format-specific patterns, individual outlet models for location effects, and price bin segmentation for budget versus premium dynamics. Rather than simple averaging, I implemented performance-weighted combination where superior segment types received higher influence in final predictions.

Technical Optimisation Experiments

I replaced manual hyper parameter tuning with Bayesian optimisation through Optuna for systematic parameter space exploration. Feature selection combined Random Forest importance scoring with recursive feature elimination, applied separately to each segment since different business contexts benefited from different feature sets. This two-stage approach removed the bottom 20% of features by importance, then refined selections to optimal subsets per segment.

Encoding and Transformation Experiments

Categorical variable treatment significantly impacted performance and pipeline stability. I initially used target encoding for high cardinality variables with cross validation to prevent overfitting. Experimenting with label encoding for all categorical variables simplified the pipeline while maintaining Random Forest's ability to learn categorical patterns, ultimately providing better robustness. For target transformation, I tested both log and Box-Cox approaches, selecting Box-Cox for its automatic parameter optimisation despite marginal complexity increases.

Results and Key Learnings

The segmented approach delivered substantial improvements over baseline single models. Weighted RMSE successfully addressed low-sales prediction concerns, while outlet-type segmentation provided the most consistent cross-validation improvements. Feature selection reduced dimensionality by 30% while maintaining accuracy, demonstrating genuine value rather than noise. The progression from simple ensembles to sophisticated segmented modeling reflected my deepening understanding of both the domain and technical constraints, ultimately achieving a solution that balanced predictive performance with business alignment and operational feasibility.