ESTIMATING STOCK KEEPING UNIT USING ML

Feature Selection Report

Overview:

Feature selection plays a crucial role in the performance of a machine learning model. In this project, we began with 9 raw features and expanded them through feature engineering to a total of 17 predictive features. The features were selected based on their relevance to sales forecasting, ability to capture time dependencies, and their statistical contribution to prediction accuracy.

Final Selected Features:

1. **Price and Promotion Variables**:
   a) `total_price`: Reflects the actual price paid by customers
   b) `base_price`: Original price, useful for understanding discounts
   c) `is_featured_sku`: Promotion indicator
   d) `is_display_sku`: Visual promotion indicator
2. **Lag Features (Time-Shifted Variables)**:
   a) `day_1` to `day_7`: Sales data from the previous 1 to 7 days
3. **Aggregate Statistical Features**:
   a) `rolling_mean_3`: Average of the last 3 days of sales

b) `expanding_mean`: Expanding window average of sales

4. **Interaction Features**:

a) `lag1_lag2_interaction`: Multiplicative interaction between `day_1` and `day_2`

b) `lag1_plus_lag2`: Additive interaction between `day_1` and `day_2`

5. **Encoded Identifiers**:

a) `store_encoded`: Mean target encoding of `store_id`

b) `sku_encoded`: Mean target encoding of `sku_id`

Feature Generation Rationale:

1) **Lag Features**: Capture recent patterns and trends in customer demand
2) **Rolling/Expanding Means**: Smooth out volatility and highlight consistent trends
3) **Interaction Terms**: Introduce complexity by modeling relationships between recent sales
4) **Target Encodings**: Convert categorical IDs to numerically meaningful values without one-hot encoding

Feature Evaluation Methodology:

- Correlation analysis
- Feature importance from Random Forest and XGBoost models
- Manual domain knowledge inspection

Result:

The combination of engineered, interaction, and encoded features resulted in high model accuracy. The selected features allowed models to learn from temporal patterns, price shifts, and store/product-level behavior without overfitting.