



ESTIMATING STOCK KEEPING UNIT USING ML

Initial Model Training, Validation and Evaluation

Objective:

To build and compare regression models that can predict the number of units sold for a given SKU using the engineered features derived from historical data.

Training Data:

- Features (X): 17 columns including lag, statistical, price, and encoded fields
- Target (y): `units_sold`
- Train-Test Split: 80% training, 20% testing

Models Used:

- Random Forest Regressor (RF)
- XGBoost Regressor (XGB)

Model 1: Random Forest Regressor

- Ensemble of decision trees with bootstrap aggregation
- Captures non-linear patterns well
- Handles skewed distributions without preprocessing

Evaluation Metrics:

- R^2 Score: 0.8169
- Mean Absolute Error (MAE): 13.63

Model 2: XGBoost Regressor

- Gradient boosting technique for structured data
- Faster training and better generalization
- Built-in regularization and missing value handling

Evaluation Metrics:

- R^2 Score: 0.8267
- Mean Absolute Error (MAE): 13.61


Observations:

- XGBoost outperformed Random Forest slightly in both R^2 and MAE.
- Both models handled high variance in `units_sold` efficiently.
- Predictions remained stable despite outliers, thanks to ensemble learning.


Validation Strategy:

- Used hold-out validation (20% test set)
- Consistent scoring across multiple seeds and runs

- **Random Forest Performance (Initial):**

 Random Forest Performance:
R² Score: 0.8168695391415389
MAE: 13.628034350473376

- **XGBoost Performance (Initial):**

 XGBoost Performance:
R² Score: 0.8267875909805298
MAE: 13.616927146911621

Conclusion:

XGBoost was selected for further tuning due to its better performance. Both models demonstrated strong predictive power, justifying the quality of the engineered features and data pipeline.