

Sales Forecasting Report: Weekly Product Demand Prediction for a Single Store

Objective

To develop a data-driven model that accurately forecasts weekly product sales quantities, supporting better inventory planning, pricing strategy, and promotional targeting.

Approach & Methodology

1. Data Preparation

- **Handle Missing Values:** No missing value found
- **Handle Outliers:** Using Interquartile Range(IQR), Winsorization to cap extreme value at a define percentile
- **Feature Extraction:** Extract “Month” and “Week” from “Sales_Week” given

2. Feature Engineering

- Revenue-based features: reflect business value, capture price impact
- Aggregated features (mean, median, sum, count) for price and sales
- Lag-based features & rolling averages for time series: to capture temporal dependencies

3. Development

- **Train-test split:** train on historical data
- **Regression-based Model:** XGBRegressor
- **Sample weights:** Add weights to `y_train` to reduce the imbalanced of low-volume samples
- **Gain-based Feature Importance:** Identified historical product performance as the strongest driver of future sales
- **Hyperparameter Tuning:** Used *RandomizedSearchCV* rather than *GridSearchCV*, allowed faster experimentation across a broader range of values

4. Model Evaluation & Visualization

- **Evaluation Metrics**
 - **Mean Absolute Error (MAE):** The average prediction error is only **3.75%** of **weekly average sales**, showing high practical accuracy
 - **Mean Absolute Percentage Error (MAPE):** Predictions are off by **40.89%**, which indicates rooms for improvement, possibly due to outliers or volatile sales weeks. Performs very well with **0.59% target grouping MAPE**, suggesting better aggregated forecasting accuracy
 - **R² score:** The model explain **89%** of the variance in weekly quantity sales, captures most patterns and trends in data, indicating reliable predictions
- **Error Distribution Plot (`y_test-y_pred`):** The model prediction errors are tightly clustered around zero, with minimal bias and few extreme deviations, indicating strong and reliable performances.

- Comparison using **Baseline (4-weeks moving average)**: XGBoost significantly outperformed this baseline in R^2 , MAPE and MAE.
- **Forecasting vs Actual Sales Quantity** over time chart: The model closely tracks actual sales across weeks, accurately capturing fluctuations and seasonal patterns. This indicates strong generalization ability and reliability for forward-looking sales quantity forecasting.



- **Compare metrics performance** with other Regression-based Model: **XGBoost** provides the best balance of accuracy, flexibility, and explainability.
 - Its ability to model complex sales behaviour, capture non-linearity
 - Robust to handle imbalances, missing values and outliers
 - Deliver actionable feature importance made it the suitable choice for sales prediction

Metrics	XGBoost	Random Forest	Linear Regression	Baseline (4-weeks MA)
MAE (% of avg sales weekly)	3.75%	17.37%	11.43%	16.70%
MAPE (Overall)	40.89%	44.85%	54.06%	57.86%
MAPE (Target Grouping)	0.59%	2.71%	1.78%	2.60%
R ² SCORE	0.89	0.86	0.84	0.79

Model Accuracy: XGBoost > RandomForest > Linear Regression

Key Findings

- **Weekly-level** features and aggregation help reduced error metrics significantly
- The model closely tracks actual sales across weeks, accurately capturing fluctuations and seasonal pattern
- **Limitations:**
 - **Overfitting Risk**
 - **Features Sensitivity:** Performances heavily depends on input features' quality, may not generalize well if external factors change (not include promotions or holidays in training)
 - **Computational Cost:** Require more resources for training or tuning