Demand Forecasting

Introduction:

This project analyzes sales and stock (inventory) patterns from a large e-commerce dataset to understand how demand varies across regions, product categories, and time periods. By combining historical sales behavior with available inventory levels, we can build a demand forecasting model that predicts future product demand and helps optimize stock planning. The goal is to use data-driven forecasting to help the business maintain the right inventory levels, avoid stockouts, reduce overstock, and better understand customer buying patterns.

overview:

Sales Data :

The columns the data set contains :

- **sales_date** – When each purchase was made
- **sep_region** – Region where the sale happened
- **category** – Product category (Electronics, etc.)
- **product** – Specific product sold
- **dealer_code / sales_agent_code** – Sales personnel details
- **payment_method & type_of_contract** – Debit, EMI, one-time, etc.
- **total_price** – Final amount customer paid
- **deposit_amount** – Advance paid (if any)
- **payment_frequency** – One-time / recurring
- **quantity** – You likely have quantity or can derive demand from count of rows

What This Tells Us

- Customer demand over time
- Popular categories and products

- Revenue-generating regions
- Payment behavior insights
- Seasonal or monthly patterns

This dataset is the core input for demand forecasting.

Stocks Data:

Columns:

benchmark_date – Date when stock was recorded
ec_region – Region of distribution center
category – Category grouping
product_name – Product identifier
ec_id – Warehouse/center ID
inventory_qty – Units available in stock

What This Tells Us

- Inventory availability at each time point
- Whether stock was sufficient during sales periods
- Gaps or mismatches between supply and demand
- Early signals of stockouts or overstock situations

This dataset supports forecasting by helping compare **actual sales vs. available stock**.

Model Experiments, Evaluation :

To build a robust and scalable **demand forecasting model**, multiple machine learning algorithms were trained and evaluated on the same engineered feature set derived from the sales and inventory datasets. The objective of this experiment was to determine which model delivers the best predictive accuracy while maintaining computational efficiency and generalization capability.

All models were evaluated using the following metrics:

- **RMSE (Root Mean Square Error):** Measures the average magnitude of error
- **MAE (Mean Absolute Error):** Measures the average absolute difference
- **WMAPE (Weighted Mean Absolute Percentage Error):** Measures relative % error
- **R^2 (Coefficient of Determination):** Measures how well the model explains variance

1. LightGBM Regressor

Performance

- **RMSE: 0.9573**
- **MAE: 0.3531**
- **WMAPE: 1.09%**

Interpretation

- The RMSE of **0.95** is the lowest among all tested models, meaning its predictions stay closest to the actual demand values.
- WMAPE of **1.09%** indicates extremely low percentage error.
- LightGBM uses a **leaf-wise growth strategy**, capturing deeper relationships in the data.

Strengths

- Handles large-scale tabular data (millions of rows) efficiently.
- Supports categorical features and high cardinality.
- Faster training and inference compared to XGBoost and SVM.
- Less overfitting due to regularization and boosting mechanisms.

Conclusion: LightGBM is the most accurate and scalable model in the experiment.

XGBoost Regressor

Performance

- **RMSE: 1.0774**
- **MAE: 0.4227**
- **WMAPE: 1.30%**

Interpretation

- XGBoost's RMSE is higher than LightGBM by ~0.12.
- It tends to overfit on high-dimensional datasets unless tuned extensively.
- Although accurate, its performance falls behind LightGBM.

Strengths

- Very strong gradient boosting model
- Good at capturing nonlinear interactions
- More stable than Random Forest

Conclusion: Great performance, but not as accurate or efficient as LightGBM.

Random Forest Regressor

Performance

- **RMSE: 2.3733**
- **MAE: 1.2565**
- **WMAPE: 3.86%**

Interpretation

- Errors are significantly higher than gradient boosting models.
- Tree ensembles like Random Forest do not learn sequential patterns in the data.
- Cannot capture complex feature interactions like boosting methods.

Strengths

- Stable baseline
- Less prone to overfitting
- Easy to train

Conclusion: Performs reliably but is far from optimal for forecasting in this dataset.

Support Vector Regressor (SVM)

Performance (Best Found)

- **RMSE: 12.0908**
- **WMAPE: 18.76%**

Interpretation

- The RMSE is drastically higher compared to tree-based models.
- SVM does not scale well to large datasets (2M+ rows).
- Training time is extremely slow.
- Unable to capture complex non-linear relationships in demand data.

Strengths

- Works well on small, simple datasets
- Good theoretical foundations

Conclusion: Not suitable for large-scale forecasting in this project.

Why LightGBM Was Selected as the Final Forecasting Model

Based on the experimental results, **LightGBM is the best-performing model for demand forecasting** for the following reasons:

Best Predictive Accuracy

- Lowest RMSE (0.9573)
- Lowest WMAPE (1.09%)
- Most consistent predictions across product categories and regions

Superior Handling of Large-scale Tabular Data

- Designed for millions of rows
- Efficient memory usage
- Handles missing values & categorical variables effectively

Fast Training & Inference

- Much faster than XGBoost and SVM
- Enables retraining and deployment at scale

Leaf-wise Tree Growth Captures Deep Relationships

- Discovers complex patterns in sales, seasonality, inventory, and pricing
- Performs better than Random Forest and linear models

Better Generalization

- Avoids overfitting through boosting and regularization
- Produces stable predictions across multiple test sets

After evaluating multiple models including Random Forest, XGBoost, Support Vector Regression, and a weighted ensemble model, LightGBM emerged as the most accurate and efficient model for demand forecasting. It achieved the lowest RMSE (0.9573) and WMAPE (1.09%) among all tested algorithms while also offering the best scalability for large datasets. The combination of high accuracy, fast training time, and strong generalization makes LightGBM the ideal choice for our final forecasting system.”