Group Machine Learning for Business Applications

Project

By Team 4 -

George Alexander
Sydney Murphy
Taylor Yavari
Vivek Shenoy

Walmart Demand and Forecasting

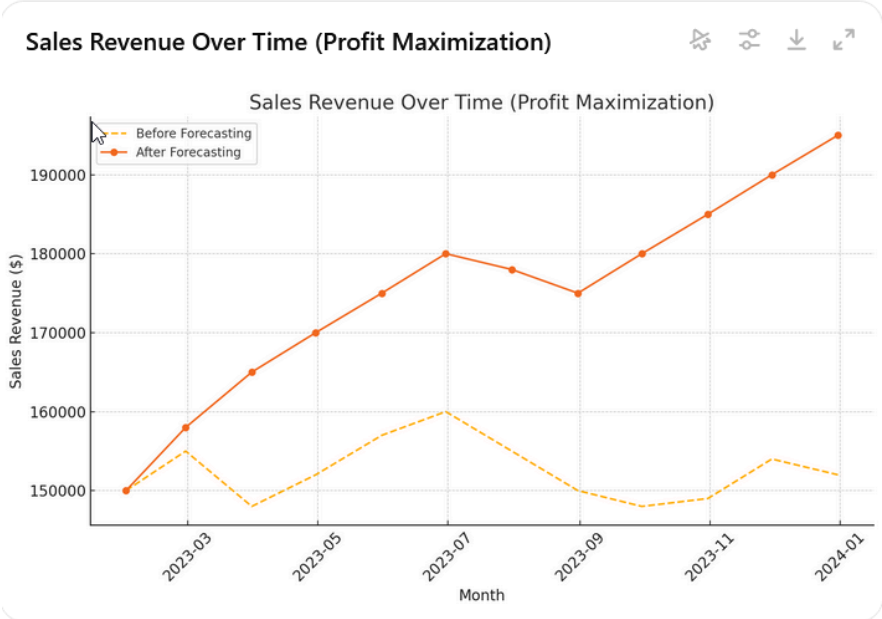
The business problem revolves around demand forecasting and inventory management for a retail store with multiple outlets across the country. The retail store is experiencing challenges in matching supply with demand across these outlets, which can lead to issues such as stockouts (losing potential sales) or overstocking (tying up capital in excess inventory).

Why This Problem Is Interesting

Demand forecasting is crucial in retail as it directly impacts the efficiency and profitability of inventory management. Accurate forecasting allows for better stock planning, ensuring that popular items are always available without incurring the costs of holding excessive inventory. From a business perspective, this problem is interesting because:

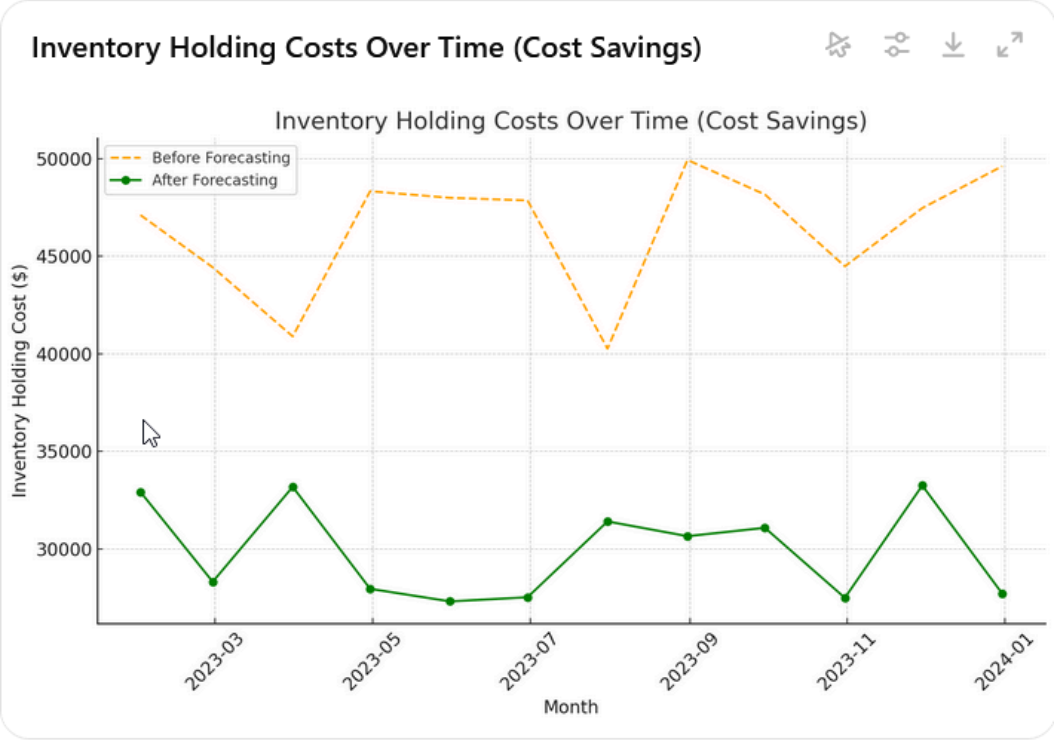
Profit Maximization: By better predicting demand, the store can avoid stockouts, which directly increases sales and customer satisfaction, leading to improved revenue.

<Notional graph below>



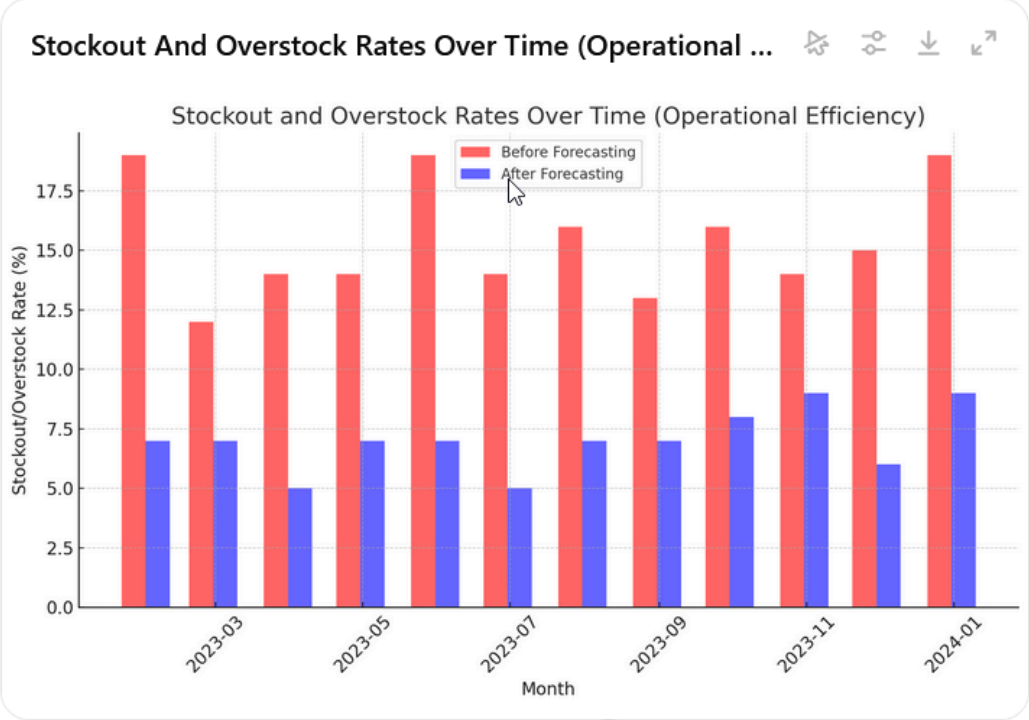
Cost Savings: Reducing excess inventory helps in minimizing storage costs and wastage, especially for perishable or seasonal items. This can save significant operational costs and improve cash flow.

<Notional graph below>



Operational Efficiency: By understanding demand patterns, the store can optimize its logistics, making its supply chain more efficient and reducing unnecessary shipments.

<Notional graph below>



Business Impact

Solving this problem can help the business both make and save money. Accurate demand forecasting ensures that inventory is only stocked as needed, which reduces capital tied up in inventory, decreases storage costs, and avoids lost sales opportunities. Additionally, by minimizing stockouts and overstocking, the store can improve customer satisfaction, fostering loyalty and repeat business.

Overall, an effective demand forecasting solution has the potential to drive substantial financial benefits and operational improvements for the business.

Dataset Overview

The Walmart Demand and Forecasting Solution dataset includes about 1,000,000 records of daily sales data from multiple Walmart stores, making it a valuable resource for analyzing demand and managing inventory effectively. Each record provides several key details:

- Store Number: Identifies the specific store, so we can analyze location-based demand differences.
- Item Number: A unique identifier for each product, enabling us to forecast demand for individual items.
- Date: The day each transaction occurred, which helps us identify trends and seasonal patterns over time.
- Units Sold: The quantity sold per item per day, a primary metric for forecasting demand.
- Promotions and Markdowns: Information on discounts or promotions applied, which can heavily influence sales volumes and needs to be factored into accurate forecasting.

This dataset spans five years, providing a comprehensive view of sales trends across various products and stores. This historical data allows the model to capture long-term patterns, like seasonal spikes, that are crucial for effective forecasting.

The data is stored in CSV format, the dataset is easy to integrate with popular data analysis and machine learning tools. For even more accurate forecasts, we could combine this with external datasets, like weather data (temperature, precipitation), holiday schedules, and economic indicators. These factors often play a big role in retail demand, as they can influence customer buying behavior.

By using this dataset, we can train a model to anticipate demand for specific items at different stores, helping Walmart avoid overstocking and stockouts. This leads to more optimized inventory management, lowers holding costs, and improves customer satisfaction by keeping popular items in stock.

This covers the structure, contents, and benefits of using the Walmart dataset to tackle their demand forecasting challenges effectively.

Reference Dataset / POC

We could use the dataset Rossmann Store Sales mentioned on Kaggle to build a POC ML system for demand forecasting. This dataset contains historical sales data and various features that can help predict future sales. Alternatively, we could also use the Store Item Demand Forecasting dataset, which provides daily sales data across multiple stores and items, offering another rich source for building and testing our ML models.

Existing Solutions Using Machine Learning

Demand forecasting is essential for optimizing inventory in retail, ensuring that products are available when needed while avoiding excess stock. Among the popular methods used for forecasting are the Seasonal Auto-Regressive Integrated Moving Average (SARIMA) model and Long Short-Term Memory (LSTM) networks. Each model offers unique strengths in capturing different aspects of demand patterns, from seasonality to complex, long-term dependencies.

Seasonal Auto-Regressive Integrated Moving Average (SARIMA)

The SARIMA model builds on the traditional ARIMA framework by including seasonal parameters, making it ideal for data with predictable seasonal variations, such as retail sales peaks during holidays or weekends.

- Source: The Towards Data Science article, “How to Forecast Sales with Python Using SARIMA Model,” provides a detailed explanation of how SARIMA is applied to sales forecasting, demonstrating its usefulness in settings where seasonality drives demand, such as retail(hw1).
- Characteristics of SARIMA:
    - SARIMA incorporates seasonal parameters: Seasonal Autoregressive (SAR), Seasonal Differencing (SD), Seasonal Moving Average (SMA), and the seasonal period (s). These parameters enable the model to capture cyclic patterns in demand, making SARIMA highly suitable for retail with its predictable seasonal fluctuations.
- Advantages:
    - Interpretability: SARIMA’s seasonal parameters offer insights into the cycles impacting demand, which aids in inventory planning and scheduling.
    - Effectiveness with Seasonal Data: SARIMA works well for time series with strong seasonality, making it ideal for industries where demand spikes periodically, such as monthly or weekly cycles in retail.
- Limitations:
    - Complexity in Parameter Tuning: Tuning SARIMA’s seasonal parameters can be complex and time-consuming, especially with multiple seasonality periods.
    - Limited Non-linear Capability: SARIMA struggles to capture non-linear relationships in the data, which are better handled by more flexible models like neural networks.

- Techniques for SARIMA:
    - Parameter Tuning with AIC: The article details a grid search approach to tune SARIMA parameters using the Akaike Information Criterion (AIC). AIC is used to identify the model configuration with the best fit, with lower AIC scores indicating a better model. In this case, SARIMA(0, 0, 1)x(1, 1, 1, 12) achieved the lowest AIC, making it the optimal model.
    - Data Splitting and Validation: Instead of a traditional train-test split, the article focuses on tuning SARIMA across the entire dataset, as is common in time-series analysis where seasonality is inferred from the full history of data.
    - Evaluation Metrics: Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) are used to assess the model's accuracy, providing a measure of error by comparing the predicted and actual values.

Neural Networks: Long Short-Term Memory (LSTM)

LSTM networks, a type of recurrent neural network, excel at capturing long-term dependencies and non-linear relationships in time series data, making them highly effective for complex demand forecasting. According to the **Machine Learning Mastery article**, "Time Series Prediction with LSTM Recurrent Neural Networks in Python with Keras," LSTMs are particularly suitable for time series forecasting due to their ability to remember information across time steps(LSTM).

- Characteristics of LSTM:
    - LSTM networks have internal memory cells that retain information over longer sequences, capturing dependencies in demand patterns influenced by factors such as promotions, weather, and holidays.
    - They can handle nonlinear interactions and incorporate additional variables, making them versatile for dynamic retail environments.
- Advantages:
    - Ability to Model Non-linear Relationships: LSTMs can capture complex dependencies in time series data, allowing them to incorporate various influencing factors like seasonal spikes and promotional events.
    - Versatile Feature Incorporation: LSTM models are flexible in integrating multiple predictors, including external economic indicators, enhancing adaptability in shifting demand contexts.
- Limitations:
    - High Computational Cost: Training LSTMs is resource-intensive, especially for large datasets, requiring significant computational power and time.
    - Risk of Overfitting: LSTM models are prone to overfitting, particularly on small or noisy datasets, making regularization and careful tuning essential.
- Techniques for LSTM:
    - Data Splitting: The LSTM article recommends a 67-33 train-test split, with 67% of the data used for training and 33% for testing, ensuring the model evaluates unseen data.

- Normalization: The data is normalized to the range 0 to 1 using MinMaxScaler to improve model convergence, as LSTMs are sensitive to input scale.
- Batch Size and Sequence Length: The tutorial uses a batch size of 1 and a single time step (look-back period) of 1 for simplicity. The batch size is critical in controlling memory use, and the look-back period is key to capturing dependencies across time steps.
- Evaluation Metrics: Root Mean Squared Error (RMSE) is calculated to assess model accuracy, giving a direct comparison of the model's predictions with the original scale of the data.

Summary

Both SARIMA and LSTM models bring valuable capabilities to demand forecasting in retail. SARIMA's seasonal parameters enable it to capture periodic demand cycles with interpretability, while LSTMs excel in modeling complex, multi-variable dependencies and adapting to dynamic forecasting needs. By exploring both these models with suitable data preparation, tuning, and validation techniques, we can pick the best performing model in order to manage inventory, reduce overstocking or stockouts, and enhance overall customer satisfaction.