PREDICTIVE ANALYSIS FOR SALES FORECASTING

DPU

DR. D. Y. PATIL VIDYAPEETH, PUNE

(DEEMED TO BE UNIVERSITY)

CENTRE FOR ONLINE LEARNING

AKSHAY CHAVAN

PRN: 231105755

BATCH - 2023-2025

Problem Definition

Forecasting sales accurately remains one of the most critical and challenging aspects of retail planning. Businesses often rely on traditional techniques that fail to capture seasonality, regional variations, and shifts in customer behavior. This leads to inventory mismatches, lost revenue opportunities, and overstock issues.

The growing availability of retail data presents an opportunity to replace intuition-led decisions with predictive analytics. However, many organizations struggle with deploying scalable, explainable, and accurate forecasting solutions tailored to complex, multi-category retail environments.

Hypothesis

It is hypothesized that predictive analytics models—especially those using historical sales patterns, lag features, and rolling averages—can significantly improve sales forecasting accuracy compared to conventional methods.

Additionally, the hypothesis assumes that by segmenting data across time, region, and product category, and applying models like ARIMA and XGBoost, we can uncover underlying patterns and build forecasts that adapt to seasonality and demand fluctuations.

Objective of the Study

The primary goal of this project is to analyze historical sales data and develop a forecasting framework using machine learning and time-series modeling. Specifically, the objectives include identifying demand trends, engineering predictive features, training forecasting models, and evaluating their performance using appropriate error metrics.

Another key objective is to simulate a real-world business context where such forecasting insights guide inventory planning, category budgeting, and strategic resource allocation.

Data Collection and Sampling Method

The dataset used is the “Superstore Sales Dataset,” which comprises four years of transactional retail data across multiple U.S. regions, product categories, and customer segments. The data includes order date, product details, region, sales amount, and other attributes necessary for time-based analysis.

Sampling was done chronologically, respecting time-series integrity. A train-test split was applied to evaluate future performance and prevent data leakage. Data was grouped monthly to align with real-world forecasting needs.

Data Cleaning and Preprocessing

The raw data underwent several cleaning operations, including duplicate removal, parsing of date fields, and normalization of column headers. Missing values were handled, and outliers were retained cautiously due to their relevance in peak-period forecasting.

Feature engineering played a vital role: new features like monthly time stamps, 1-month lag values, 3-month rolling averages, and category-wise aggregations were created. These were essential for training context-aware predictive models.

Data Analysis

The exploratory data analysis phase revealed important sales trends and cyclicality. Category-wise monthly sales plots showed Technology as highly volatile, while Office Supplies remained stable. A rolling average graph helped in identifying long-term demand behavior.

Heatmaps and box plots further highlighted the regional sales dominance of the West and East. These visual insights helped shape the model features and pointed toward opportunities for segmentation and targeted forecasting.

Interpretation of the Data Analysis

The sales patterns observed were consistent with typical retail behavior: higher sales in Q4, especially in November and December, and lower volumes in summer months. The West region consistently outperformed others, likely due to concentrated urban retail activity.

The volatility in Technology sales highlighted the importance of dynamic models. The consistent, low-ticket sales in Office Supplies suggested a different strategy—perhaps bulk forecasting at category level rather than SKU level.

Achievement of Project

The project successfully established a comprehensive forecasting pipeline—from data preprocessing to visual trend analysis and model-readiness. It demonstrated the importance of time-aware data handling and laid a strong foundation for further model development.

Beyond technical outcomes, the project provided business-aligned insights that would help a retail firm streamline its decision-making, forecast demand more accurately, and allocate resources efficiently.

Suggestion and Future Enhancement

1.Future enhancements include implementing advanced models like SARIMA, Prophet, and XGBoost with hyperparameter tuning. Real-time data ingestion pipelines and forecast dashboards can be added to scale the solution for business use.

2.There is also scope for integrating profit and discount data, enabling ROI-based forecasting. With sufficient granularity, models can be extended to product-line or store-level forecasting. These steps would bring predictive analytics even closer to real-world application and strategic impact.

Thank you