Monthly Sales Forecasting for a Retail Supermarket Using ARIMA, Prophet, and XGBoost

Soniya Malviya
Aryan Soni
Kalash Thakur

***Abstract*—This report presents a time-series forecasting study on a supermarket retail dataset containing 9,994 transactions from Tamil Nadu, India, with sales recorded across product categories, cities, regions, and dates between 2015 and 2018. The data were preprocessed by cleaning date formats, deriving temporal features (year, month, date), and computing variables such as discount amount. Exploratory analysis highlighted strong seasonality, with November and 2018 exhibiting the highest sales and profits, and February and 2015 showing the lowest. Monthly sales were modeled using ARIMA, Prophet, and XGBoost on a training period up to 2017, with 2018 held out for testing. Model performance was evaluated using RMSE, MAE, MAPE, and R^2, and the Prophet model achieved the lowest RMSE of approximately 64,625, a MAPE near 10%, and $R^2 \approx 0.86$, outperforming ARIMA and XGBoost. These results indicate that Prophet is well suited for capturing seasonality and trend in monthly retail sales, providing accurate forecasts for planning and inventory decisions.**

***Index Terms*—Time-series forecasting, retail analytics, ARIMA, Prophet, XGBoost, sales prediction.**

I. Introduction

Retail supermarkets generate large volumes of transactional data that support sales forecasting, inventory management, and strategic planning. Accurate monthly sales forecasts help prevent stockouts, reduce excess inventory, and support promotion planning. This work analyzes a "Supermart Grocery Sales – Retail Analytics" dataset from Tamil Nadu and develops statistical and machine learning models to forecast monthly sales.

The objectives of this study are:

- To perform exploratory data analysis (EDA) on category-, city-, region-, and time-wise sales and profit.
- To build ARIMA, Prophet, and XGBoost forecasting models using aggregated monthly sales.
- To compare models using multiple error metrics and recommend the best forecasting approach.

II. Dataset and Preprocessing

The dataset contains 9,994 transaction records with features such as Customer Name, Category, Sub Category, City, Order Date, Region, Sales, Discount, Profit, and State. The `Order ID` column was removed as it had no analytical value. No missing values were found among the relevant variables.

Order dates appeared in mixed formats (e.g., "8/27/2016", "06-11-2016") and were standardized using mixed-format parsing. From the processed dates, Year, Month, and Day-of-month fields were derived. A new variable, *Discount_Amt*, was computed using Sales and Discount percentage. The North region was excluded due to having only one record.

III. Exploratory Data Analysis

Category-wise analysis showed that Snacks accounted for the largest number of items sold, while Oil & Masala had the lowest volume. Subcategory-wise, Health Drinks and Soft Drinks recorded the highest sales, whereas Rice showed the lowest totals.

City-wise, Kanyakumari achieved the highest total sales, while Trichy was among the lowest. West was the strongest-performing region, followed by East, Central, and South. Year-wise, 2018 produced the highest sales and profits, while 2015 was the weakest. November was the most profitable month, whereas February consistently produced the lowest sales.

IV. Methodology

Transaction-level data were aggregated into monthly total sales using month-end resampling. The period from 2015 to 2017 served as the training set, and the 12 months of 2018 served as the test set.

Three models were implemented:

A. ARIMA

An ARIMA(p, d, q) model with parameters (5,1,2) was fitted to the monthly training series.

B. Prophet

The monthly series was reformatted into Prophet's (ds, y) structure and trained with yearly seasonality enabled. Forecasts were generated for the 12 months of 2018.

C. XGBoost

A gradient-boosted regression model using lag features (lag1, lag2) was trained to predict monthly sales. Regularization techniques such as reduced tree depth, subsampling, and L1/L2 penalties were used to avoid overfitting.

Model performance was evaluated using RMSE, MAE, MAPE, and R^2.

V. Results and Discussion

A. Descriptive Patterns

Category- and subcategory-level pivot tables showed that Eggs, Meat & Fish and beverage products generated high revenue. Snacks produced the highest total profit. November showed the highest profit, while February showed the lowest. About 30% of total revenue was allocated to discounts.

B. Forecasting Results

Table I summarizes the model performance on the 2018 test set.

TABLE I
MODEL PERFORMANCE COMPARISON

Model	RMSE	MAE	MAPE	R^2
ARIMA	174,247	156,415	46%	< 0
Prophet	64,625	47,993	10%	0.86
XGBoost	192,000	133,000	–	< 0

Prophet significantly outperformed both ARIMA and XGBoost across all accuracy metrics. Prophet's forecast captured seasonal peaks and troughs in 2018 more accurately.

VI. CONCLUSION

This study examined sales patterns in a supermarket retail dataset from Tamil Nadu and demonstrated strong seasonality, with peaks in November and troughs in February. Among the forecasting models evaluated, Prophet achieved the best performance in RMSE, MAPE, and R^2, making it well suited for monthly retail sales forecasting. Future work may integrate external variables such as holidays or promotions or explore deep learning models like LSTM networks.

REFERENCES

[1] "IEEE General Format," Purdue OWL, 2019. [Online]. Available: https://owl.purdue.edu

[2] "IEEE Paper Format: Template & Guidelines," Scribbr, 2023. [Online]. Available: https://www.scribbr.com/ieee/ieee-paper-format

[3] Google Colab Notebook: Supermart Grocery Sales Analysis. [Online]. Available: https://colab.research.google.com/drive/1TGXS_DqtkFJIxa-37Xpy63Ee9k78wQ3f

[4] A. Author, "Forecasting Sales Using Various Time-Series and Machine Learning Methods," M.Sc. thesis, Tilburg Univ., 2019. [Online]. Available: http://arno.uvt.nl/show.cgi?fid=160134