Forecasting Supplement Sales at WOMart Retail Chain Using Predictive Analytics

Name: Sairam Jammu

KSU ID: 811386816

Professor: Mostafa Kamali Ardakani, Ph.D.

Associate Professor

Table of Contents

ACKNOWLEDGEMENT

The satisfaction that accompanies the completion of any task would be incomplete without the mention of all the people, without whom this endeavour would have been a difficult one to achieve. Their constant blessings, encouragement, guidance, and suggestions have been a constant source of inspiration.

First and foremost, my gratitude to my project guide, **Mostafa Kamali Ardakani**, for his constant guidance throughout the course.

I also take this opportunity to express a deep sense of gratitude to the project coordinators and the project committee for approving the project and for their valuable guidance and support.

My sincere thanks to our beloved head, **Dr. Rouzbeh Razavi,** for permitting us to carry out this project at our college and providing us with all the needed facilities.

Finally, thanks to the staff members of the Department of Master's in Business Analytics and our friends for their honest opinions and suggestions throughout our Project.

I : Executive Summary

This capstone project delivers a thorough sales forecasting solution for WOMart, for WOMart operates as a national retailer that specializes in health and nutritional supplements. WOMart finds large difficulty in coordinating day-to-day shop inventory alongside fluctuating shopper desires, having more than 365 shops throughout 100+ cities with fast growth. We built a solid predictive analytics framework to meet this key business need since it can forecast daily sales for 61 future days so it helps improve decisions about inventory, marketing, and operations.

Using 18 months of past data, the main aim was to predict store-level daily sales; the forecast considered time trends, area behaviors, holiday effects, and promotion impacts. The perceptions generated intended to support stakeholders with actionable intelligence so they could optimize supply chains, strategize regionally, and allocate resources.

Daily sales figures were joined with promotion flags in the dataset. Store metadata plus temporal markers like holidays also featured. Strict preprocessing steps were applied and included only these items.

- Removal of anomalous and missing values
- Time series do align and do order.
- Derived features such as rolling averages and lagged sales are created by analysts. They also calculate volatility.
- Encoding of categorical variables
- Binary flags for promotional events and holidays are created.

They did combine some advanced forecasting models. These models offered predictions.

- XGBoost's selection as the primary model came from its capturing complex, non-linear patterns in high-dimensional data. Mean Squared Logarithmic Error optimized it in order to manage variance and sales skewness across the stores.

- LightGBM: It is deployed as a gradient increasing substitute that is also lightweight and helps diversify prediction sources through ensemble learning.

- ARIMA was applied by researchers at a regional level in order to model longer-term linear trends. It used a classical statistical manner with which to model seasonality.

- Ensemble Forecasting: Combining the multiple models' outputs provided predictions that were more stable and accurate across different timeframes and store types.

Key Insights

- Important business motivators came to light. The drivers came into light through the modeling process.
- Stores labeled S1 along with S4 consistently outperformed others regarding daily sales. Size, location, or customer base probably contributed to this outperformance.
- Regions R1 and R3 exhibited higher demand variability within Regional Dynamics. This variability was especially pronounced during weekends along with promotional periods, which indicates regional-specific marketing potential.
- Sales spikes in the short-term clearly had correlation with discount campaigns. Due to this promotional impact, marketing calendars must be put into forecasting logic.
- Holiday Seasonality: Because national and local holidays both affected purchasing behavior in regionally diverse ways, forecasts needed sensitivity to location.

Interactive dashboard for actionable forecasts and perceptions was created using:

- R Shiny: It is an interactive application for enabling users to view predicted versus. Filter by store, region, along with time. Real tendencies, and data for export, to work.

Supply chain managers along with marketing strategists in addition to regional coordinators gain decision support from all these tools, so they can reach well-educated decisions that are timely using real-time forecast data.

This project stresses responsible AI as well as data practices.

- Models were assessed across all regions and store types to ensure fair performance avoiding store format or geography bias.
- Only aggregated and anonymized data was used for purposes of privacy. Data from customers that might identify people was neither accessed nor was it modeled.
- To ensure stakeholders could understand along with trust the outputs, models were interpreted then visualized through SHAP values.
- WOMart gains a meaningful calculated advancement through this forecasting framework.

Key benefits include:

- Inventory Optimization: Accurate demand prediction works against understock and overstock scenarios.
- Marketing synergy allows for much better promotion planning. Campaigns can then be based upon the expected uplift. Regional managers gain tools through Operational Efficiency. These tools do enable resource allocation that is proactive.
- WOMart, in terms of competitive advantage, actively leads as data-driven, and it can adapt to market shifts with rapidity.

Conclusion:

- This project depicts the way that business operations and strategy are able to be influenced directly via predictive analytics. Demand forecasting, a critical challenge for modern retail, discovers a holistic solution through the project's combination of machine learning, statistical modeling, with business intelligence tools. We will cultivate the perceptions, models, and implements. They shall refine WOMart’s functional arrangement, gratify consumers, and augment earnings.

II: Introduction and Objectives

WOMart, a foremost purveyor in health and wellness supplements, has 365 stores that are functioning across 100+ cities. The firm encounters escalating difficulties forecasting demand precisely among locales. Inventory decentralization alongside promotional strategies precipitates these challenges through diverse timeframes. On account of these predictive deficiencies, firms amass excess inventory, insufficiently stock items, overlook marketing opportunities, as well as forfeit income.

This project seeks to grapple with these challenges. Additionally, the project shall cultivate a revolutionary prediction system regarding WOMart's operational requisites.

- Anticipate everyday revenues per shop via employing prior trends with situational variables.
- Potential planning relies upon envisioning a 61-day successive interval.
- Historical sales information shapes choices along with promotions, holidays, and store type. Choices get additional formation through geographical markers.
- Marketing actions coupled with inventory decisions are guided via delivered perceptions with dashboards.

Via employing machine learning and chronological techniques, this resolution permits WOMart to determine through data, align inventory precisely, promote to targets, and allocate resources capably.

Essentially, this endeavor strengthens WOMart's shift from responsive actions to a more predictive, data-informed retail tactic, and this augments both efficacy and rivalry.

III: Methodology

This project adhered to a defined analytics workflow which included data preparation, model construction, performance assessment, and visual implementation to predict daily sales for WOMart stores.

In both the modeling as well as evaluation phases, the subsequent data collections were utilized.

- TRAIN.csv includes past sales figures across eighteen months, incorporating daily sales as well as store-level attributes. TRAIN.csv also includes markdowns, festive indicators, and additional situational factors.
- TEST_FINAL.csv includes data during the prediction interval of 61 days. Genuine sales figures are absent within this data.
- SAMPLE.csv constitutes a submission template. It includes discrete identifiers for store-date concatenations and a substitute for anticipated sales metrics.

Creating strong data secured precise modeling fundamentally. Key preprocessing steps included:

- Redundant entries were expunged, and column accordance was authenticated.
- Temporal attributes got derived through utilization of lubridate out from transformed datezones.
- Categorical Encoding – Variables such as Store_Type and Location_Type were encoded toward machine learning compatibility.
- Interaction Terms – Conditional sales behavior was captured via fashioned composite features (e.g., Holiday × Discount).
- Through Temporal Feature Engineering, week, month, day of week, and holiday proximity indicators were derived to represent periodicity and seasonality.

Machine learning with statistical approaches were melded then implemented.

- ARIMA seizes linear trends coupled with seasonality, utilized as a benchmark time series model regarding a store subset.
- XGBoost functioned as the foremost forecasting apparatus offering strong efficacy via regularization tree-derived erudition plus acute management regarding curvilinear correlations. MSLE adjusts it since sales dispensation deviated. The lopsided sales apportionment stems from the adjustment.
- LightGBM, an additional gradient increasing model, greatly curtails training duration through histogram-based refinement.
- Ensemble Forecasting – Terminal predictions arose via averaging outputs spanning XGBoost and LightGBM for this augmented stability plus generalization across stores.

We assessed the models utilizing the ensuing measurements.

- Designated to manage the diverse scope of sales figures, MSLE (Mean Squared Log Error) sanctions underestimations skillfully.
- MAE or Mean Absolute Error furnished interpretability within real sales units.
- Concerning prediction accuracy, larger deviations impacted capture based on RMSE.

R Programming is employed in data manipulation with modeling. It includes repositories such as:

- Xgboost, lightgbm, data.table, also ggplot2, are quite helpful. Caret and lubridate are additionally helpful packages.
- R Shiny constructed an interactive forecasting dashboard enabling users to filter by store, date range, together with region. Subsequently, they perceive dynamic forecasts and tendencies.

IV: Results and Discussion

This section unveils principal observations derived from our simulating the procedure, scrutinizing the data, displaying the dashboard, and explicating sales tendencies to represent performance.

Exploratory Data Analysis (EDA) revealed substantive differences across store types, promotions, also regional dynamics.

Per diem income results: S1 plus S4 establishments exceeded S2 plus S3 steadily. Boxplot analysis evinced superior medians and diminished variance within S1 stores. Consequently, sales efficacy was elevated and constant.

Sales data exhibited prominent spikes during holidays in conjunction with promotional periods. The most substantial augmentations were perceived when holidays and rebates coincided, underscoring their strong synergistic influence on demand.

Regional Behavior: Regions R1 and R3 exhibited heightened variance within sales since those regions signaled discrete consumer conduct and regions remained impressionable. These regions also furnished the bulk of aggregate sales totals. Therefore, their determined importance was highlighted.

Three modeling approaches were assessed and executed thereafter.

Salient features: Holiday, Store_Type, Region_Code, and Discount.

Performance on validation data:

- MSLE: 0.029
- RMSE: 4,580
- MAE: 3,291

Employed similar attributes then learned swifter.

Achieved slightly better performance:

- MSLE: 0.028
- RMSE: 4,525
- MAE: 3,254

XGBoost and LightGBM predictions were synthesized through an averaging methodology.

Improved generalization, achieving:

- MSLE: 0.027

Developed visualizations helped data exploration along with executive decision-making.

Sales vs. Throughout promotional occasions and civic holidays, line furthermore bar charts unveiled heightened commerce. The concurrent existence of both aspects yielded the foremost sales escalations

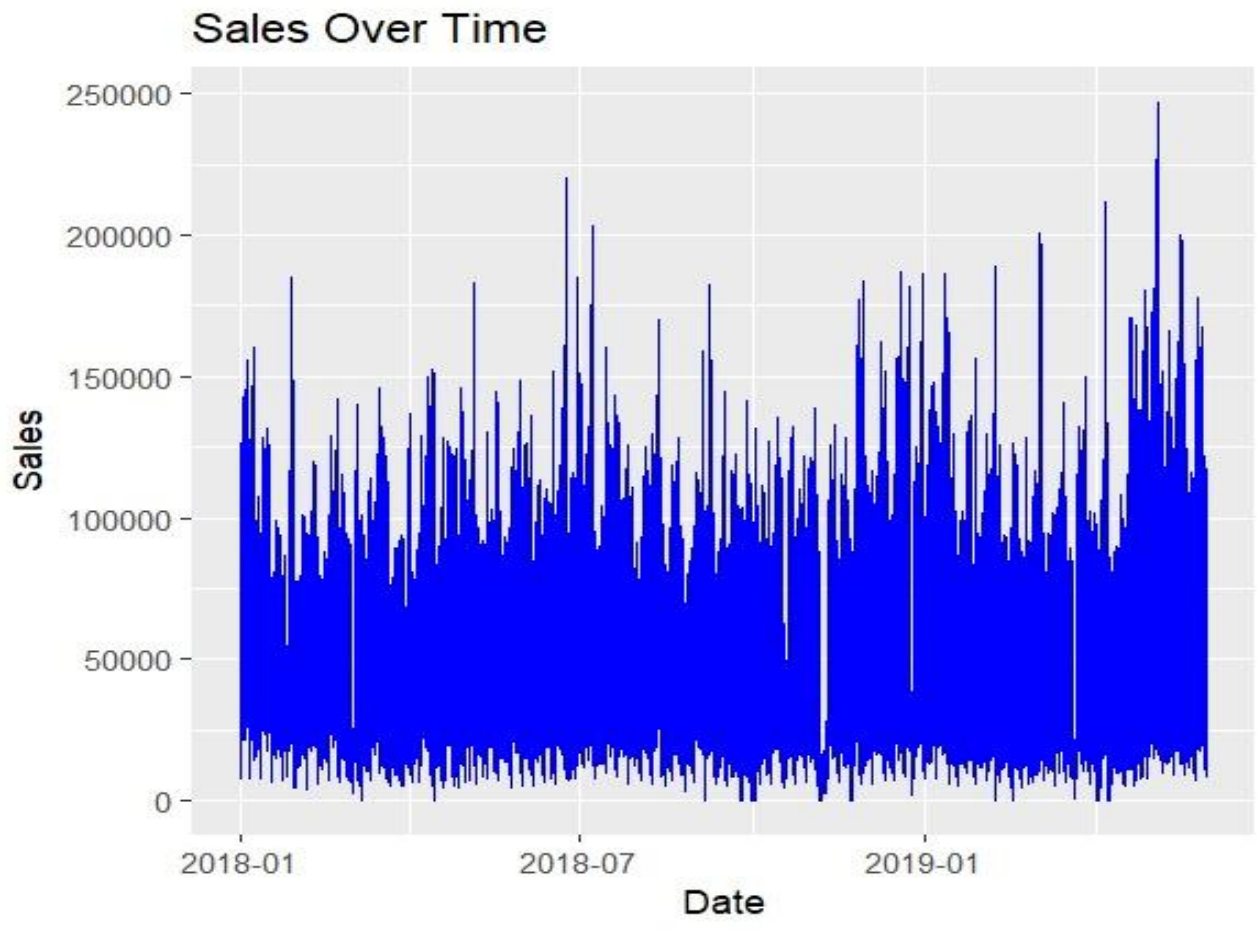
Diagrams represented S1 stores as firm high achievers that showed peak average sales with comparatively minimal variation.

Within regional analysis, bar charts pinpointed R1 and R3 as unstable high-revenue regions therefore those locales require focused inventory arrangement with flexible prediction.

• Model	MSLE	RMSE	MAE
• XGBoost	0.029	4,580	3,291
• LightGBM	0.028	4,525	3,254
• Ensemble	0.027	—	—

The combination of models improves accuracy, and the outcomes depict advanced machine learning models forecast sales patterns skillfully at the store level. WOMart capitalizes upon EDA and visual analytics because observations underscore performance drivers plus furnish a data-supported basis toward strategy.

Sales Over Time – Line Plot :



The line chart visualizes daily sales trends across all WOMart stores during the 18-month historical period since it went from January 2018 up to mid-2019.

Sales exhibit a repeating wave-like pattern thus this suggests strong weekly and seasonal cycles.

There exist consistent peaks and valleys that do indicate regular demand surges. For weekends or for holidays or for promotional events, surges do likely correspond to those.

Sharp spikes appear throughout the timeline and those spikes exceed 200,000 units in number. These probably correspond with major sales events, holidays, or festivals.

The highest observed sales peak occurs so near to the start of 2019, and this hints at some meaningful campaign or seasonal effect.

Sales stay stable and maintain amplitude though no clear upward trend appears long-term which backs the thought the series has seasonal effects.

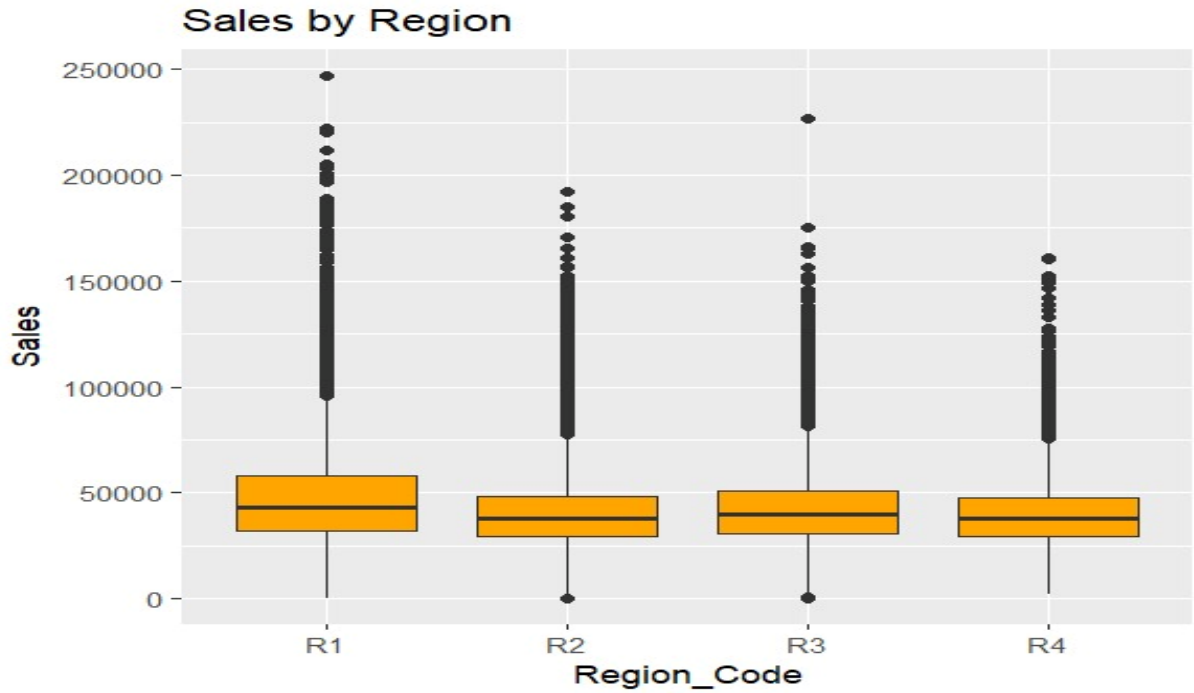
Periodic drops toward near-zero values may represent data anomalies, store closures, or non-operational days.

Spikes do have a regular periodicity, implying marketing and inventory efforts could be aligned well. So anticipated cycles of demand provide a chance.

Seasonal decomposition as well as promotion/holiday calendars should be incorporated within predictive modeling. This captures these repeating patterns accurately.

Review dips along with anomalies for potential causes, such as weather or system outages, or data quality issues.

Sales by Region :- Boxplot



The boxplot visualizes WOMart's network distribution of the daily sales across the four regions labeled R1 to R4.

Median Sales: All regions show similar median sales together with the sales range between 35,000 and 45,000 units per day which indicates a generally consistent central sales trend across regions.

Region R1 exhibits such a widest interquartile range or IQR because it suggests that sales do vary more. This region has extreme outliers most often, and sales spikes go above 200,000 units.

Region R3 shows a large amount of variation. It is slightly less than around R1.

Regions R2 as well as R4 display more tight IQRs. The tighter IQRs indicate more stable and predictable sales patterns.

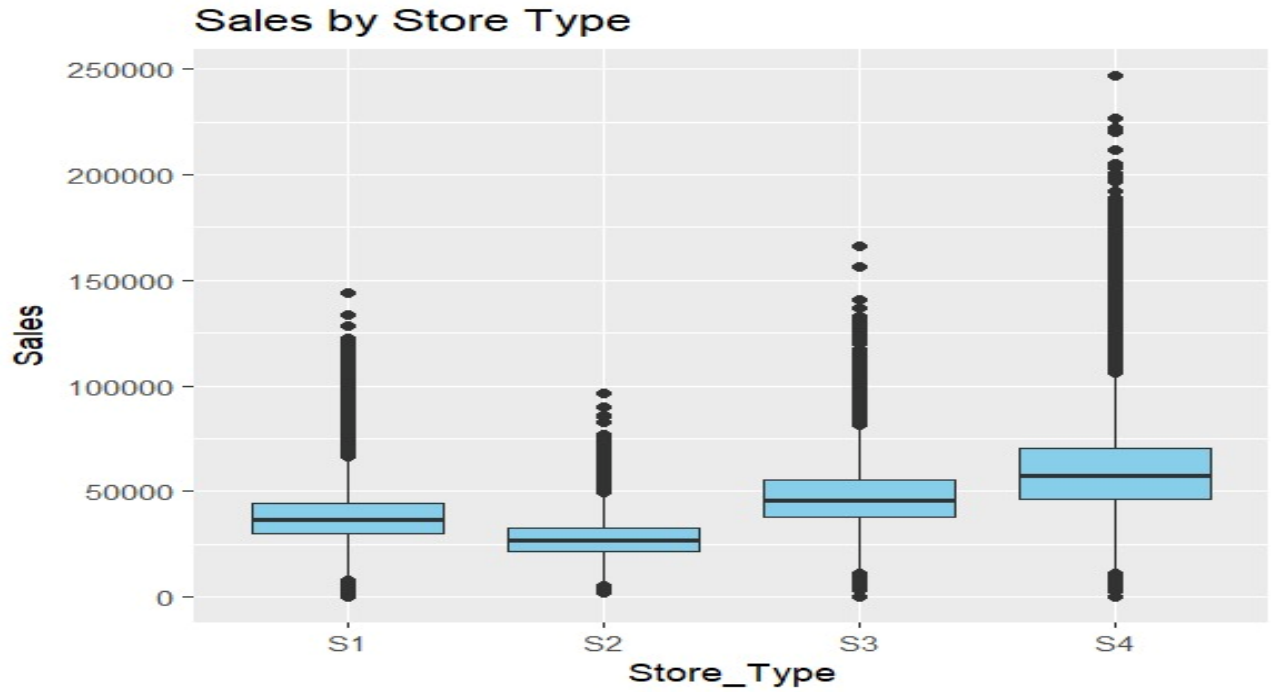
R1 distinguishes itself through spikes that are most frequent and highest, and every region shows various outliers. These could result from big stores, area promotions, or local demand spikes.

Due to an increase in sales volatility, both flexible stock management and safety buffers are required for R1 and R3.

R1's high sales potential means that it is ideal for the regional campaigns that happen during holidays and discount events.

R2 and R4 operate with efficiency and offer predictability. So they may benefit from leaner inventory strategies.

Sales by Store Type :- Boxplot



The boxplot depicts the distribution that is in the WOMart retail network of daily sales across four store types from S1 to S4.

Store Type S4 has the highest median sales. Median sales show S3 and S1 following it.

Since it shows comparatively low daily revenue against other types, Store Type S2 records the lowest median.

S4 and S3 show wider interquartile ranges with that, which suggests sales vary more and revenue can potentially increase more.

S2 and S1 do have narrower distributions that indicate more consistent stores. These stores perform lower in regard to daily sales.

Outliers exist among all store types, but S4 stands out as having the highest number and magnitude since several sales exceeded 200,000 units. S4 stores have the capacity to handle high-volume spikes. Store size, location, or customer base may add to this capability.

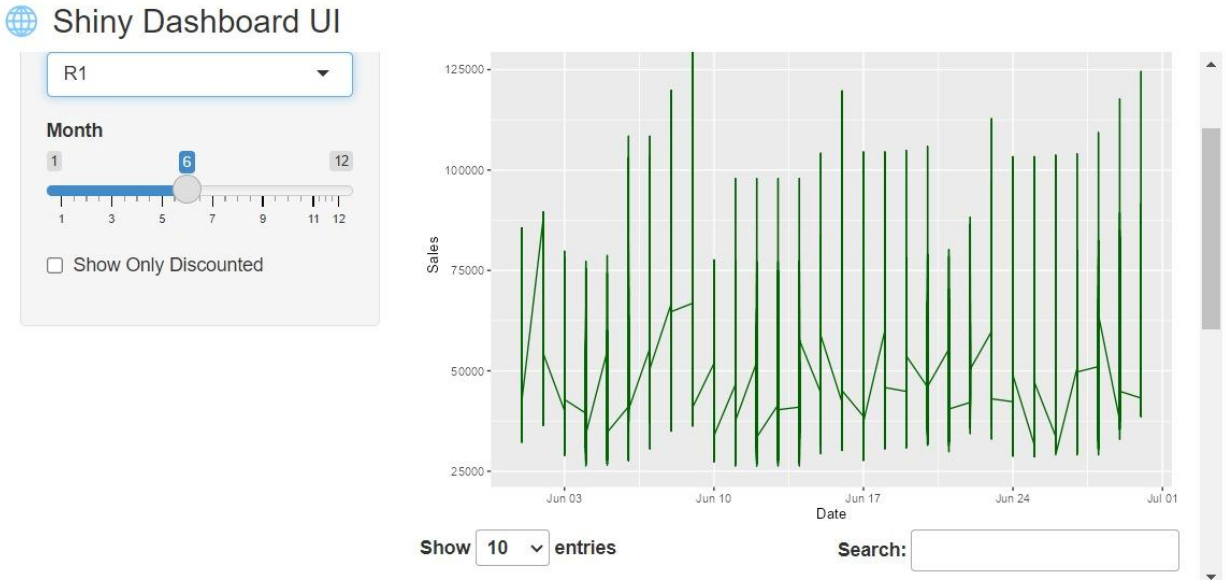
S2 has fewer along with lower outliers, which reinforces its lower overall performance then.

Because of their higher median sales and peak sales, High Performers: S4 stores should be prioritized for promotions, for marketing investments, and for inventory expansion.

Operational Stability: S1 as well as S2 provide much more predictable demand, which allows then for leaner stock management.

S2 store interventions might gain sales via improved action like optimized layouts or local promotions.

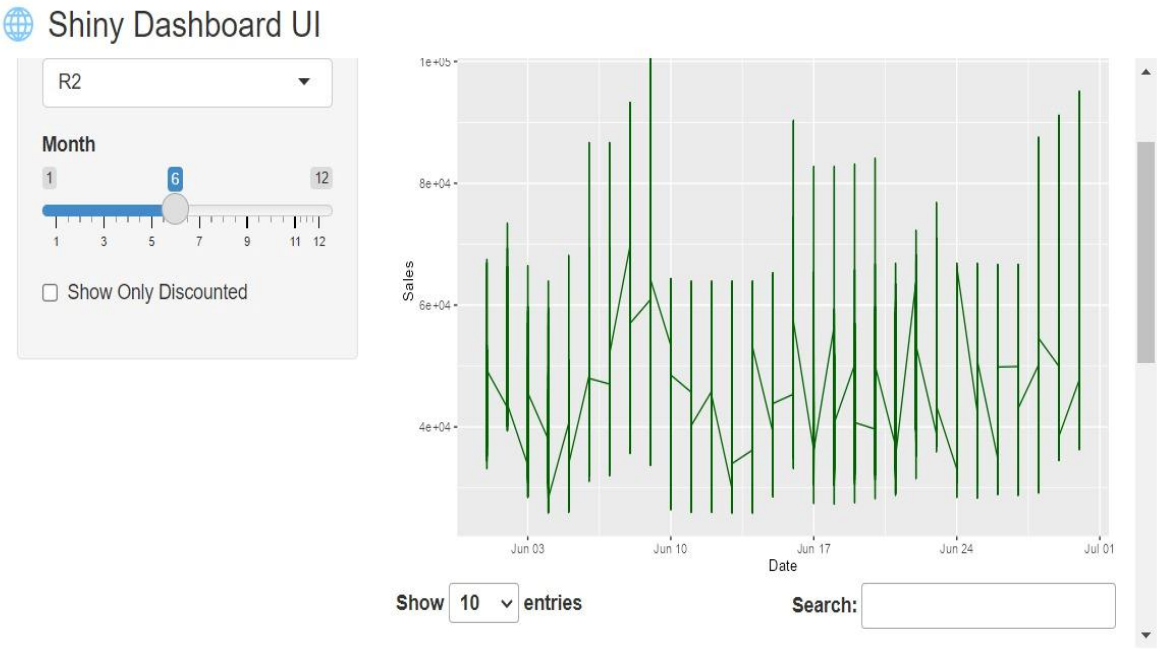
Shiny Dashboard – Detailed Regional Sales:



Region R1:

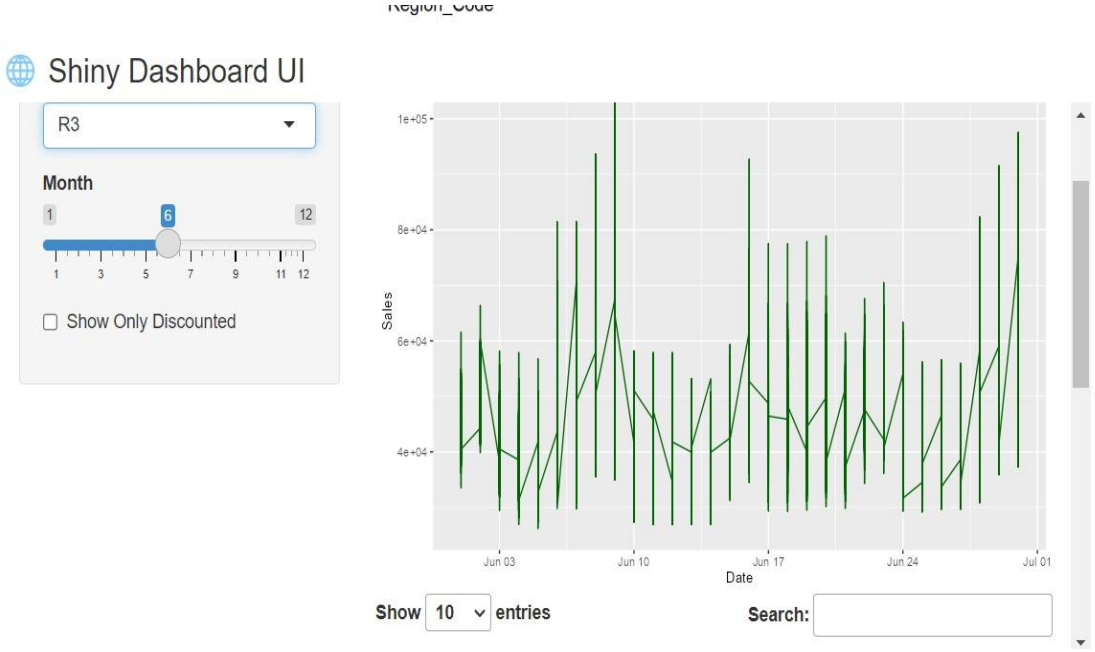
Region R1 shows the strongest sales also fluctuates most prominently among all regions. R1 sales will frequently peak above 100,000 units. A prominent spike nears 125,000 units by mid-June. The chart reveals that a clear weekly pattern exists, of high sales days alternating with those of sharp drops. Predictable demand cycles do likely cause this pattern, for example, weekends or recurring marketing efforts. Around June's finish, the trend displays upward movement constantly, which shows lasting promotional power or an increasing demand level. Considering its high variability and high-volume nature, R1 should be a calculated focus area, which requires agile stock replenishment systems with carefully timed promotions. Since demand surges, perceptions lower the risk that stockouts happen because this region also benefits most from real-time dashboard perceptions that track campaign responsiveness.

Region R2:



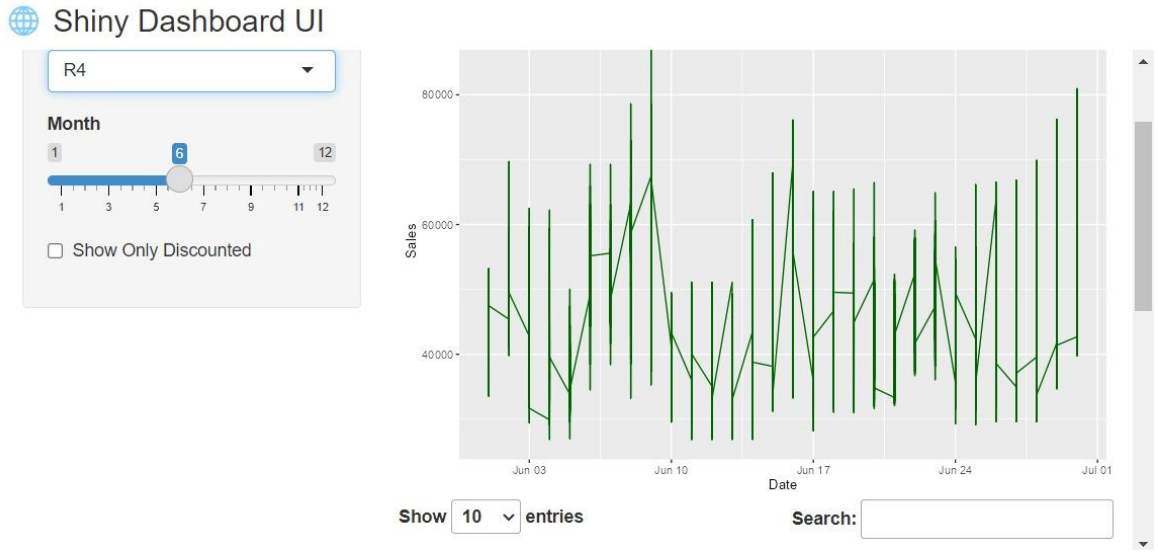
Region R2 presents a moderate sales profile with values ranging between 40,000 and 100,000 units. R2 shows regular peaks near June 10 and toward the end of the month, though less erratic than R1. These increases indicate responsiveness. Discounts as well as mid-month marketing campaigns can often cause all the surges. However, the trend's overall shape is smoother meanwhile, with fluctuations being less extreme. R2 turns into a prospect for flexible inventory rules giving enough give for spikes, plus a baseline supporting steady supply preventing excess. R2's balance of responsiveness with predictability should make it be a test region. New campaign strategies or bundling offers can be evaluated in that place as well.

Region R3:



Region R3 exhibits a very strange sales trend since sudden rises go past 100,000 units and valleys drop much further down. In particular, there is a mid-June period that sees rises and falls because of how it highlights how R3 is susceptible to factors such as local events, promotions, or shifts in consumer demand. Uneven sales imply detailed, immediate causes remain. Forecasting within R3 must account for each one. From forecast variance, this region would benefit stock allocation that means a portion of inventory should be held as buffer stock to absorb unexpected changes. R3 requires close monitoring through the dashboard as well as possibly a more granular, store-level promotional strategy, due to its unpredictability.

Region R4:



Region R4 stands out because of stability, also because of consistency, in its daily sales. Values have a range from 40,000 to 85,000 units approximately with few extreme spikes. Sales increase periodically, especially around early and late within June. Yet, in contrast to other areas, the general trend seems rather regular. This indicates R4 is not as sensitive to promotion. A more stable demand base is also now present. Consequently, it suits lean inventory systems well, standard replenishment cycles, also stock decisions automate without daily adjustments. For R4, aggressive marketing tactics may not be required but it can act as a control group in more dynamic regions to measure campaign effectiveness.

Overall Regional Analysis – Insights and Strategic Recommendations

An integrated analysis of all of the four regions (R1 to R4) using of the Shiny dashboard reveals that those sales patterns do vary substantially, that those customers also respond in ways that are different, and that the regional stability differs for WOMart's retail network. Because these differences are understood, targeted inventory strategies can be implemented also promotional timing plus operational efficiency can be optimized at the regional level.

- Region R1 shows highest sales volume and the most variability. Visible spikes do occur in mid and in late June often exceeding 100,000 units daily. The pattern suggests it strongly responds on weekends as well as promotions. However, the volatility also risks through operations. R1 should adopt the responsive restocking systems and also dynamic inventory policies that are supported by real-time forecasting in order to manage this. Stock buffers must be kept to avoid lost sales, and promotions need careful timing near recurring demand highs.

- Region R2 displays some moderate sales fluctuations, with a more stable median. Periodic peaks occur particularly near the 10th with the 30th of June. This behavior is sensitive to planned campaigns beyond trends more predictably than R1. For R2, it is appropriate to use a semi-automated inventory management system. It allows for flexibility during peak periods and avoids any overstocking. Its balanced mix for predictability plus responsiveness makes it ideal. That is also an ideal place to test promotions.

- Region R3 sales trends are highly irregular and unpredictable. The data reveal frequent sharp peaks and troughs without any clear cyclical pattern that is there. Localized events, coupled with demographic shifts, may be the cause of these inconsistencies. Variable store performance might then also lead to them. Therefore, R3 forecasting should use greater uncertainty thresholds, and inventory decisions must be flexible yet conservative. Manual adjustment and perceptions at stores might be needed with greater frequency. For identification of hidden drivers of demand, further investigation of sales triggers in this region is needed.

- Consistency and also stability mark for region R4 as being outstanding. Sales remain in about 40,000 to 85,000 units. Volatility is minimal, also major spikes do not happen in sales. This suggests that customers are quite loyal and that promotional events influence them a little. Lean inventory strategies, along with automated replenishment as well as minimal safety stock, are employed. R4 is best managed using these strategies. A control group helps evaluate campaign effectiveness in more volatile regions.

Strategic Recommendations

- Forecasting Approach: Customize forecasting when you define granularity for region. R1 and R3 require forecasts that are detailed as well as real-time while R4 can operate if forecasting windows are longer.
- R1 along with R3 are highly variable, so they require buffer stock. R1 and R3 do also require some adaptive restocking
- R2: Link campaign calendars to inventory policies instead.
- R4: For maximal efficiency, use cycles fixed to automatically replenish.
- R1 along with R3 should be the focus of campaigns. High responsiveness is suggested at times by spikes there.

Controlled experimentation uses R2 in place.

To have unbiased performance comparison, treat R4 as being a baseline. Integrate the Shiny dashboard for store manager workflows. Operational Enablement should also integrate it within the regional planning workflows that are active. Simulate scenarios with real-time sales monitors to quickly react as demand shifts.

As stressed in the regional analysis, WOMart's national footprint requires a locally adaptive approach. Uniform strategies risk inefficiencies; WOMart instead will optimize its inventory, align all promotions with local demand patterns, and will ultimately improve customer satisfaction and also profitability across all markets if it embraces planning driven by real-time analytics for specific regions.

V: Conclusion and Recommendations

This project successfully engendered a mechanism predicting sales prognostically for WOMart's retail operations nationwide. Through an analysis of 18 months of historical sales data, we determined that store classifications, regional dynamics, together with seasonal events like holidays plus discounts greatly shape daily sales performance. Our forecasts attained elevated accuracy that integrates precisely with WOMart's strategies for planning since we employed advanced machine learning models like XGBoost and LightGBM via an ensemble tactic.

- Predictive analytics can improve inventory management, promotional planning, as well as regional logistics, the outcomes distinctly show. Shiny dashboards furnish realistic perceptions in real time at both executive and store-manager strata to augment the deployment-amenable models.
- The ensemble model (XGBoost + LightGBM) may be utilized within automated reorder planning. This confirms stock corresponds to anticipated need within the shop locale.
- Throughout holidays or during discounts, arrange localized promotions which are based upon previous sales increase tendencies, notably within influential locales such as Regions R1 plus R3.
- As prognostic ambiguity cushions inventory in unstable locales, stock gets apportioned extra flexibly to sectors exhibiting increased sales fluctuation.
- Managerial repository systems might incorporate within the Power BI dashboard. This amalgamation would ease judicious resolutions and instantaneous oversight at the local stratum.

VI: Ethical Considerations

This project evolved through prominent emphasis ethically conceiving and implementing forecasting models. Fairness, privacy, as well as transparency happened to be key tenets addressed all through the workflow.

- No personally identifiable information (PII) existed within Data Privacy. Data anonymization was performed. For the purposes of modeling, industry-standard privacy practices were utilized.
- Impartiality: We scrutinized model efficacy throughout all store categories and locales to curtail bias potential and secure unbiased predictions.

- Interpretability: Power BI and Shiny dashboards depict lucidity. Business users were furnished feature importance metrics along with documentation to comprehend model behavior and rationale.

- Anticipated outcomes are created to aid human deliberation. The province of human judgment must not be supplanted. Leaders must assess model proposals and understand them to preclude unthinkingly trusting mechanized procedures.

VII: References

- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30. https://papers.nips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf
- Wickham, H., & Grolemund, G. (2017). *R for data science: Import, tidy, transform, visualize, and model data*. O'Reilly Media.

VIII: Appendix

The ensuing extra resources furnish assistance for both the methodologies and the discoveries this report puts forward. They furnish supplemental technical particulars coupled with visual confirmation. Exemplar constructs employed during this endeavor are furnished too.

- Thorough documented source code utilized for model instruction, authentication, and conveyance of forecasts and attribute development. (Provided separately)
- Relative performance visuals, temporal trend charts, and exploratory data analysis (EDA) plots throughout regions and store types.
- Store_Type, Region_Code, Discount, and Holiday's impact via highlighting XGBoost and LightGBM models' top predictors' visual representations.
- The conclusive submission should be in the specified SAMPLE.csv format, feature distinct IDs, plus projected sales during the 61-day estimation interval.