Hybrid Data-Driven Forecasting System for Regional Chocolate Sales: An Engineering Approach for Operational Decision Support

Samuel Aljure Bernal, Carlos Alberto Barriga Gámez, David Santiago Aldana González, Juan Diego Alvarez Cristancho
Department of Computer Engineering
Universidad Distrital Francisco José de Caldas
Emails: {saljureb, cabarrigag, daldanag, jalvarezc}@udistrital.edu.co

***Abstract*—Accurate regional sales forecasting is essential for chocolate manufacturers that must plan production, inventory, logistics, and marketing in markets shaped by demographic and environmental variability. This work presents a complete predictive system that integrates Python-based data processing with R-based modeling to generate reliable forecasts using robust preprocessing, statistical exploration, and ensemble learning. The implemented system achieves substantial improvements in prediction accuracy, identifies critical instability drivers, and demonstrates operational reliability through software quality tests and robustness experiments.**

***Index Terms*—Sales Forecasting, Machine Learning, Hybrid Architecture, Python-R Integration, Ensemble Learning, Predictive Modeling**

I. Introduction

Forecasting sales in regional markets is a fundamental capability for companies in the consumer goods industry. For a chocolate manufacturer, demand variability arises from interactions among demographic conditions, advertising exposure, weather patterns, and the temporal maturity of each market. These factors create complex dynamics that challenge traditional linear prediction techniques and require modeling strategies capable of capturing non-linear behavior and heterogeneous data distributions. Developing an accurate forecasting system is therefore not only a statistical task but also an engineering requirement that influences production scheduling, procurement, supply-chain coordination, and marketing investment.

Previous work in quality-managed analytical systems emphasizes the importance of process consistency and controlled workflows. ISO 9001 highlights principles for structured, auditable processes and organizational reliability [1]. Similarly, the Capability Maturity Model Integration (CMMI) framework argues that analytical and software processes benefit from clearly defined stages, standardized interfaces, and managed lifecycles [2]. In addition, methodologies derived from Six Sigma promote defect prevention and encourage early identification of data quality issues that can propagate through analytical pipelines [3]. These perspectives converge on the need for forecasting systems that operate within reproducible, traceable, and verifiable workflows rather than ad-hoc or exploratory environments.

The dataset provided by the company consists of structured records describing demographic indicators, weather conditions, advertising variables across digital and traditional channels, and variables reflecting time elapsed since the company's entry into each region. Initial exploratory analysis reveals significant heterogeneity: multicollinearity among marketing features, non-normal distributions, categorical variables without ordinal meaning, and a strong temporal component that influences long-term sales behavior. These findings guide the architectural decision to combine flexible Python-based data preparation with specialized R-based modeling tools.

The core motivation of this work is to design, build, and validate a predictive system that transforms these heterogeneous sources into a reliable forecasting asset. Unlike isolated scripts, the system is engineered as an integrated pipeline with modular stages for data ingestion, transformation, modeling, robustness evaluation, and service-oriented deployment. The following sections present the design decisions, experiments, and validation strategies that demonstrate the system's operational viability.

II. Methods and Materials

A. System Architecture Overview

As a solution for automated and accurate chocolate sales prediction, an artificial intelligence system was developed integrating Python and R. The system is designed to facilitate the analysis of historical data and the generation of future sales forecasts through a straightforward and visual interface, without requiring advanced programming knowledge. Its architecture is structured into three specialized layers:

Data Processing Layer: This layer ingests historical sales files and prepares them for analysis. Cleaning routines convert non-numerical data into machine-readable formats and extract relevant patterns. The layer produces graphical summaries, including heatmaps of correlations and feature importance charts, to help users understand the drivers of sales without requiring technical expertise.

Training Layer: This layer implements advanced machine learning algorithms. Multiple models, including Random Forest and XGBoost, are trained and compared using historical

data. The system automatically selects and saves the best-performing model for subsequent predictions.

Presentation Layer: A user-friendly web interface allows uploading of new data (e.g., product or date information) and generates instant predictions displayed on-screen or as downloadable files. Visual elements enhance usability, allowing non-technical users to interpret predictions and make informed decisions regarding inventory or marketing strategies. The system automates the entire workflow with a single command, guiding the user at each step.

Figure 1 illustrates the three-layer architecture and workflow.

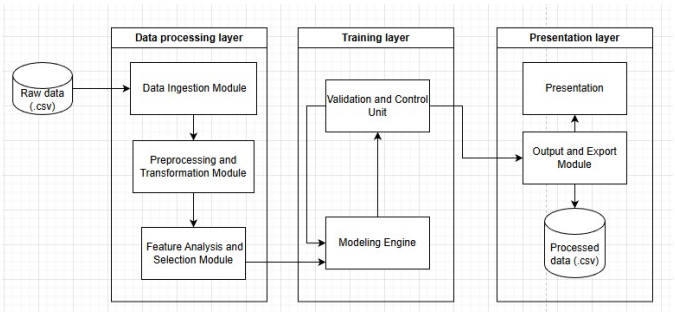

Fig. 1. Feature importance analysis using ANOVA F-statistics. Critical drivers such as *Time in Region* and Facebook GRP are highlighted.

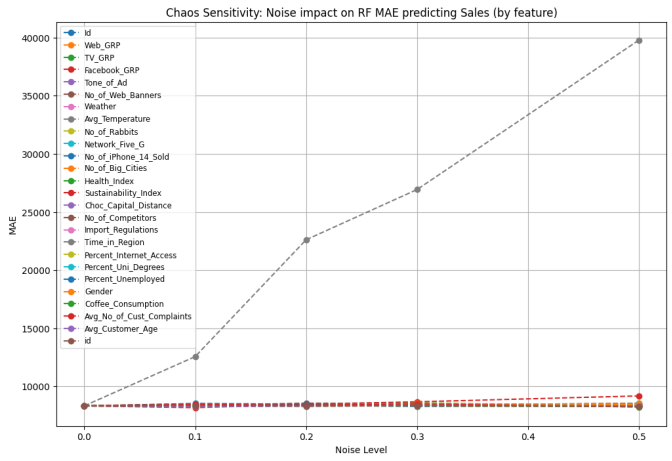

Fig. 2. Correlation matrix of the main predictors, illustrating multicollinearity among socioeconomic and marketing variables.

B. Modeling Engine

The Modeling Layer, implemented in R, evaluates Linear Regression, Random Forest, GBM, and XGBoost. A Stacking Ensemble is trained as the meta-model, combining the base learners to leverage complementary strengths. Cross-validation using MAE ensures generalization. The serialized model is then integrated into the deployment pipeline.

C. Robustness Evaluation

a) Gaussian-Noise Sensitivity Analysis:: Controlled Gaussian perturbations (10%, 20%, 30%, 50%) are injected into numerical features. Most variables, such as marketing and environmental features, produce linear changes in MAE. *Time in Region* demonstrates exponential sensitivity, making it a structural instability driver.

b) Cellular Automata Simulation:: A 50×50 grid models spatial market dynamics. Positive-feedback (contagion) and negative-feedback (cooling) rules simulate local interactions and saturation. Emergent clusters of high sales intensity are observed, confirming non-linear spatial dynamics in market adoption.

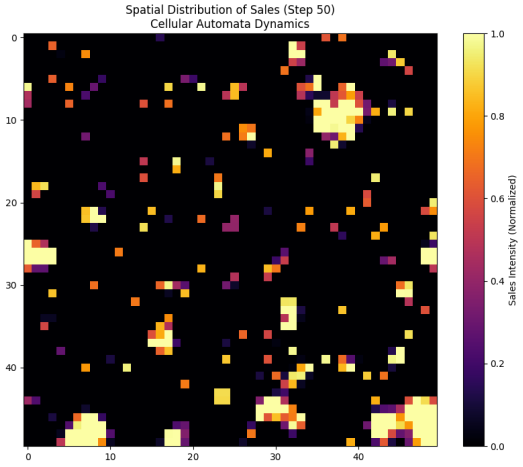

Fig. 3. Spatial distribution of sales at t=50. High-intensity clusters form naturally due to local interactions.

D. Presentation and Deployment

The system exposes a FastAPI interface. Inputs are validated, preprocessed, and passed to the R engine. Outputs are delivered in JSON or CSV format. Modular design supports integration into dashboards or ERP systems.

E. Assumptions and Limitations

The system assumes historical data represent future scenarios and that temporal variables remain consistent. Cross-language execution introduces latency. Ensemble complexity improves accuracy but reduces interpretability.

III. Results and Discussion

A. Predictive Performance

Table I summarizes model performance. Linear Regression performs poorly due to non-linear effects. Random Forest and XGBoost reduce error, while the Stacking Ensemble achieves the lowest MAE, confirming the advantage of combining base models.

TABLE I
MODEL PERFORMANCE USING 5-FOLD CROSS-VALIDATION (MAE).

Model	MAE
Linear Regression	11997.86
Random Forest	8254.61
XGBoost	5107.48
Stacking Ensemble	**4017.47**

B. Feature Importance and Sensitivity

Figures 4 and 3 illustrate sensitivity and spatial behavior. Marketing variables show stable linear responses under noise. *Time in Region* produces exponential MAE increases, emphasizing the need for temporal data quality. Sensitivity experiments define operational boundaries for reliability.

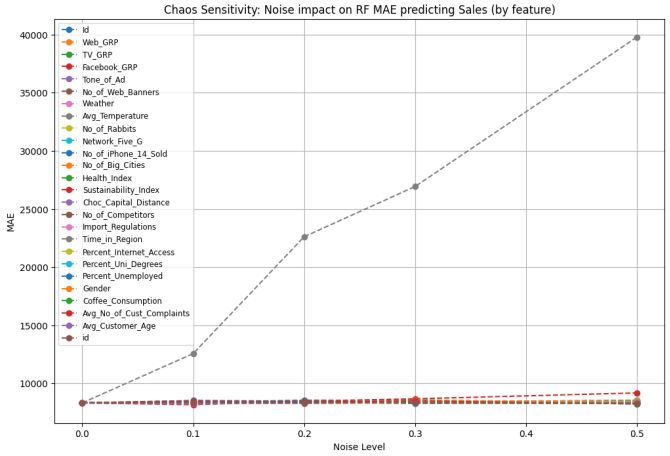

Fig. 4. MAE degradation under Gaussian noise. *Time in Region* exhibits exponential sensitivity compared to stable predictors.

C. Spatial Market Insights

Cellular Automata simulations reveal emergent clusters (Figure 3). High-intensity zones form surrounded by low-activity regions, demonstrating local interactions and saturation dynamics. This insight supports the inclusion of spatial-lag variables in future models.

D. Software Quality and Testing

Unit, integration, and acceptance tests ensure reproducibility and reliability. Unit tests verify schema integrity, preprocessing, and feature consistency. Integration tests confirm Python–R communication. Acceptance tests validate API responsiveness and functionality. Testing aligns with ISO 9001, CMMI, and Six Sigma principles [1]–[3].

The combination of predictive accuracy, sensitivity analysis, and software quality confirms that the system can reliably support operational planning in production, logistics, and marketing.

IV. CONCLUSIONS

This work presents a complete data-driven forecasting system designed to support decision-making in chocolate sales across multiple regions. Through a hybrid architecture integrating Python and R, the system delivers accurate predictions, robust preprocessing, and reliable deployment mechanisms. The Stacking Ensemble reduces prediction error substantially compared to individual models, demonstrating the advantage of combining complementary learning strategies. Robustness experiments identify critical sensitivity drivers, and spatial simulations reveal non-linear market behaviors relevant for strategic planning.

The system succeeds in transforming heterogeneous data into actionable forecasts while maintaining process structure aligned with established engineering and quality-management principles. Future work includes integrating spatial regression models, incorporating external economic indicators, and deploying real-time monitoring components to detect data drift and maintain long-term reliability.

ACKNOWLEDGMENTS

We would like to express our sincere gratitude to Professor Andrés Sierra for his guidance, mentorship, and technical insights, which were fundamental to the success of this work. We also wish to acknowledge our own dedication, resilience, and teamwork in bringing this project to fruition.

Finally, we wish everyone a Merry Christmas and a prosperous New Year 2026.

REFERENCES

[1] L. Fonseca and P. Domingues, "Iso 9001: 2015 edition-management, quality and value," *International Journal for Quality Research*, vol. 11, no. 1, pp. 149–158, 2017.

[2] F. S. Silva, F. S. F. Soares, A. L. Peres, I. M. de Azevedo, A. P. L. F. Vasconcelos, F. K. Kamei, and S. R. L. Meira, "Using cmmi together with agile software development: A systematic review," *Information and Software Technology*, vol. 58, pp. 20–43, 2015.

[3] Y. Z. Mehrjerdi, "Six-sigma: methodology, tools and its future," *Assembly Automation*, vol. 31, no. 1, pp. 79–88, 2011.