ANALYSIS AND DESIGN OF A MODULAR

PREDICTIVE SYSTEM FOR THE "SWEET REGRESSION COMPETITION"

Samuel Aljure Bernal - 20202020111
Carlos Alberto Barriga Gámez - 20222020179
David Santiago Aldana Gonzalez - 20222020158
Juan Diego Álvarez Cristancho - 20221020076

The Problem

Challenges of Traditional Methods

Chocolate sales in regional markets exhibit considerable demand variability, driven by complex and nonlinear dynamics that conventional methods fail to accurately model.

Traditional linear forecasting techniques are insufficient, unable to capture the intricate behaviors and heterogeneous data distributions that characterize this market.

Key Factors of Variability

Demographic Conditions
Population characteristics and specific consumption profiles of each region.

Climatic Patterns
Significant impact of weather and seasonal conditions on purchasing decisions.

Advertising Exposure
Effectiveness and reach of marketing campaigns and their influence on local demand.

Market Maturity
Stage of development and length of time the brand has been present in each geographic region.

Hybrid System Architecture

Data Engineering

Handles the ingestion and robust processing of heterogeneous data to ensure the quality and consistency of input information.

Python

Model Execution

This module houses the training and execution logic for the forecasting algorithms, optimizing prediction accuracy.

R

Deployment

This component handles the delivery of results and forecasts to end users through clear and accessible interfaces.

API / Dashboard

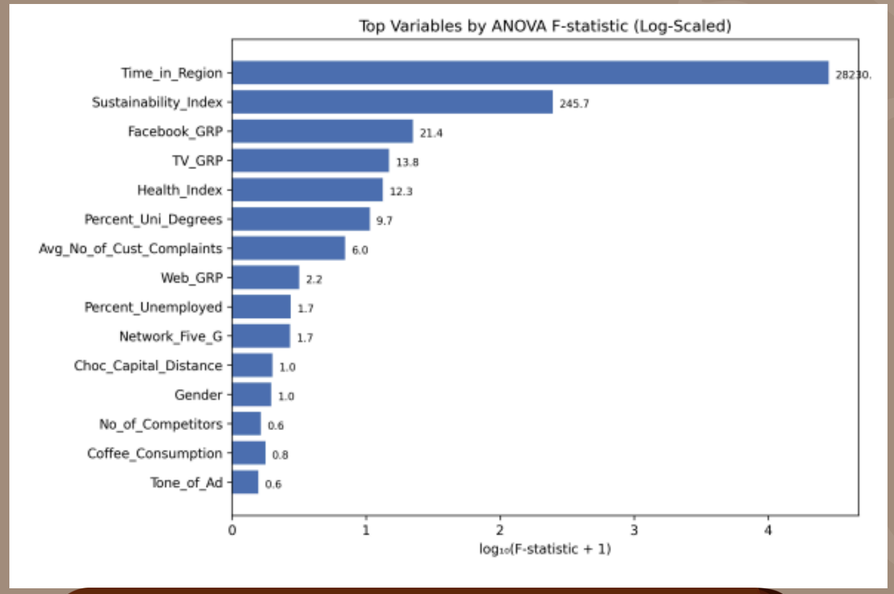
This integration of Python for processing and R for modeling ensures modularity, ease of debugging, and compliance with quality standards.

Data Processing in Python

Python is essential for processing heterogeneous datasets, enabling robust preprocessing and ensuring data quality for advanced analytics and reliable sales forecasting.

Transformation and Exploration
Detailed statistical exploration of heterogeneous data, encompassing correlation matrices and distributional analysis to guide modeling.

Data Diagnostics
Accurate identification of multicollinearity among predictors and non-normal distributions to ensure model validity.



Exploratory Statistical Analysis

Statistical exploration uses methods such as ANOVA and correlation matrices to assess the contribution of predictors, systematically identifying the variables that most influence chocolate sales.

Impact on the Model

Analyzing these variables is crucial for guiding modeling decisions and feature engineering strategies, thereby improving forecast accuracy.

Analysis of Variance (ANOVA)

Use F-statistics to rank the importance of predictors and their actual impact on sales variability.

Correlation Matrices

Evaluate the strength and direction of the linear relationship between predictor variables and sales volume.

Key Predictor: Time in Region

A fundamental variable that demonstrates a significant explanatory influence on sales dynamics.

Key Predictor: Facebook GRP

Identified as a critical predictor that directly impacts business performance and marketing strategies.

Modeling Strategy in R

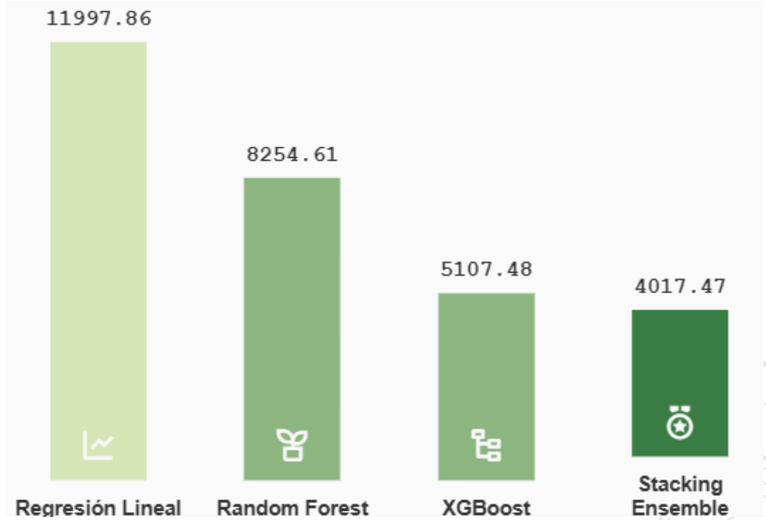
Model Comparison

The predictive core was implemented in R. Using 5-fold cross-validation, four models were evaluated: Linear Regression, Random Forest, GBM, and XGBoost. The results showed variable performance among the base models.

Ensemble Implementation

To significantly improve accuracy, a Stacking Ensemble was implemented. This advanced approach integrates the base models, using their predictions as input for a meta-learner that generates the final output.

Performance Comparison (Mean Absolute Error - MAE)



The Stacking Assembly achieves the lowest MAE with 4017.47, demonstrating a substantial improvement in predictive accuracy over individual models.

Model Robustness Assessment

Applied Methodology

Controlled perturbation tests were performed by injecting Gaussian noise into each feature of the model. The objective was to observe the response of the Mean Absolute Error (MAE) and quantify the sensitivity of the variables.

This approach allows for the simulation of fluctuations in the input data and the evaluation of how they impact the accuracy of sales predictions.

Highly Sensitive Variable

'Weather in the Region exhibited exponential sensitivity. Small fluctuations generated large deviations in the prediction, identifying it as a high-risk operational variable.

Highly Stability Variables

Advertising variables based on GRP (Gross Rating Point) maintained remarkable stability. Noise injection had a minimal impact on the prediction error.

Spatial Simulation with Cellular Automata

Market Dynamics Modeling

A Cellular Automata model is used to analyze how localized interactions and spillover effects influence regional chocolate demand.

Emergence of Clustering Patterns

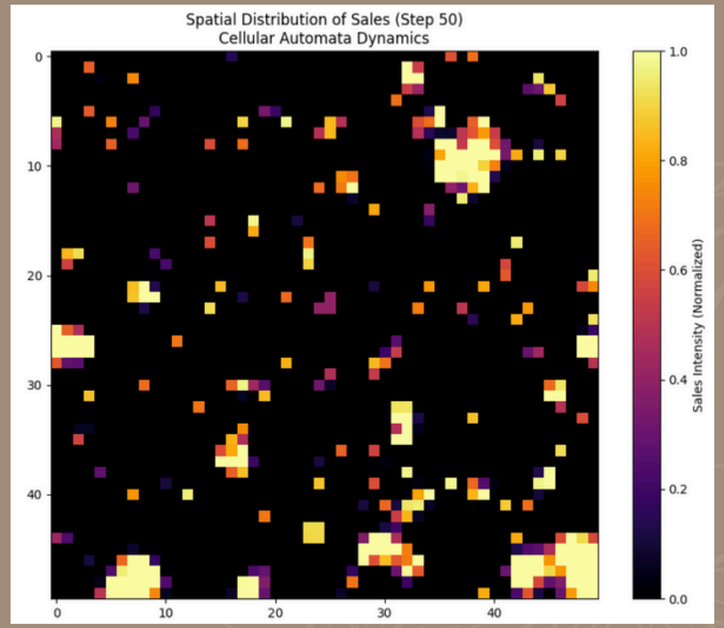
Iterative simulation reveals the emergence of sales clusters, demonstrating that market adoption tends to concentrate in specific geographic areas.

Complex System Dynamics

These results validate the hypothesis that chocolate demand operates as a complex adaptive system, shaped by spatial and temporal dependencies.

Strategic Perspectives

The simulation identifies nonlinear market behaviors, crucial for effective strategic planning and a deep understanding of market penetration.



Influence of Feature Engineering

Impact of Key Variables

Advertising and time-related maturity variables (e.g., Time in Region, GRP) significantly impact forecast accuracy, based on feature importance determined by ANOVA.

High Temporal Sensitivity

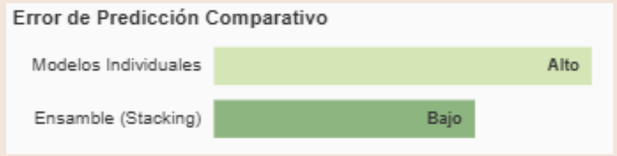
Temporal variables such as 'Time in Region' show high sensitivity; minor fluctuations can cause substantial deviations in the prediction.

Capture of Interactions

Ensemble models effectively capture complex and nonlinear interactions among various features, an advantage over individual models.

Ensemble Advantage

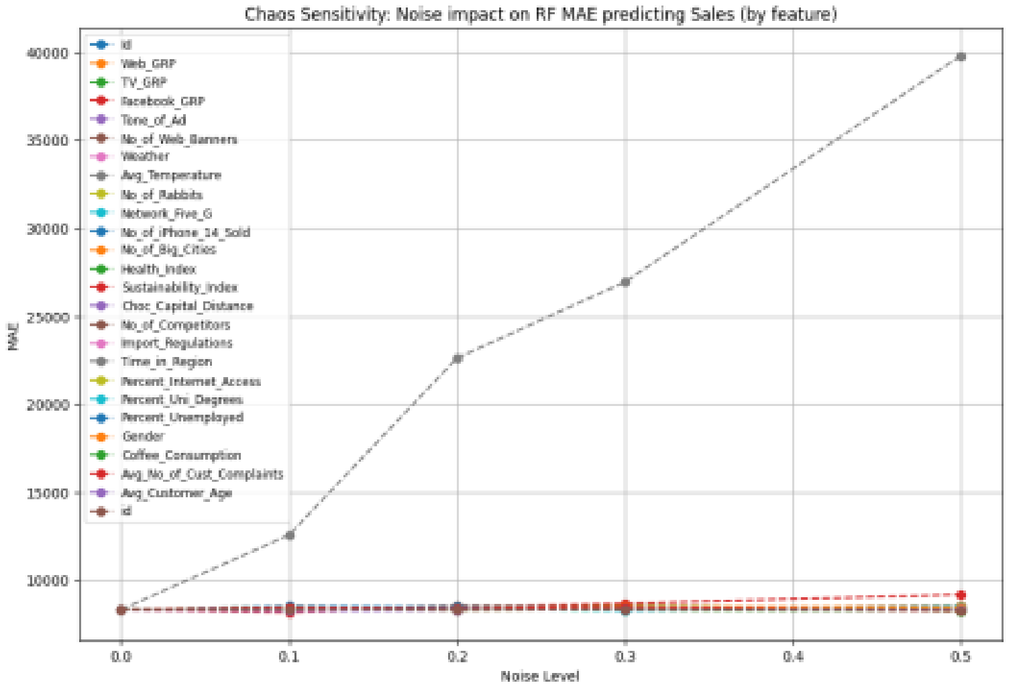
Stacking significantly reduced the prediction error compared to base models, demonstrating its superiority.



Sensitivity and Chaos Analysis

Key Findings

- Controlled perturbation experiments reveal the system's stability and breaking points.
- Injection of Gaussian noise demonstrates an exponential degradation of the Mean Absolute Error (MAE) in critical variables.
- The "Time in Region" variable exhibits chaotic sensitivity, with exponential increases in MAE in response to minor perturbations.
- These findings are crucial for prioritizing data quality control and managing dependencies in the forecasting system.



Software Quality Validation

System Testing

Unit Tests
Validated the integrity of data ingestion, preprocessing, and schemas.

Integration Tests
Ensured efficient communication between the Python and R components.

Acceptance Tests
Confirmed the API's responsiveness and functionality under concurrent requests.

Compliance with Standards

ISO 9001
Adherence to principles for structured, auditable processes and organizational reliability.

CMMI
Benefits throughout the software development lifecycle through standardized stages and interfaces.

Six Sigma
Promoting defect prevention and early detection of data quality issues.

Conclusions and Practical Applications

Advantages of the Hybrid System

It integrates Python for robust data processing and R for advanced statistical modeling.

It ensures modularity, ease of debugging, and compliance with the organization's quality practices.

It offers accurate predictions through robust preprocessing and advanced assembly learning (Stacking Ensemble).

It provides reliable deployment mechanisms for actionable forecasts.

Error Reduction and Operational Improvement

Substantial reduction in prediction errors, significantly outperforming individual models.

Improves operational decision-making for production, inventory, logistics, and marketing investment.

Identifies critical drivers of instability and nonlinear market behavior for strategic planning.

Supports reproducible, traceable, and verifiable workflows, ensuring operational reliability.

Future Work

Spatial Regression Integration

Analyze geographic patterns and market dependencies for more accurate and localized sales forecasts.

External Economic Indicators

Integrate macroeconomic factors to gain a comprehensive view of the market context and consumer trends.

Real-Time Monitoring

Implement systems for the proactive detection of data drift, ensuring the long-term adaptability and reliability of the model.

THANKS