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 Alvarez Cristancho - 20221020076

INTRODUCTION

The Sweet Regression challenge aimed to build a regression system predicting chocolate sales, with results delivered in R. The analysis showed that relative evaluation and submission limits increase sensitivity when data noise or randomness adds instability. A hybrid system using Python for exploration and R for final modeling was proposed.

GOAL

Design a modular, reproducible architecture to generate sales predictions with minimal error (MAE), using feedback and control mechanisms to stabilize performance and ensure model generalization.

PROPOSED SOLUTION

A Hybrid Modular Predictive System (Python/R) was developed with independent modules for data ingestion, preprocessing, modeling, and validation.

Python handled exploratory analysis (correlations, ANOVA, etc.) and R was used for final model comparison with methods such as Ridge, Lasso, Random Forest, Gradient Boosting, and XGBoost.

The Validation Unit stops training when the MAE no longer improves.

EXPERIMENTS

The experimental process focused on exploratory statistical analysis using Python. This phase included validating the consistency of the data schema, coding categorical variables (e.g., *Tone_of_Ad*, *Weather*), and performing Pearson correlation and variance (ANOVA) analyses. The goal was to quantify the influence of predictors on sales and identify the most relevant characteristics, laying the groundwork for the future modeling phase in R.

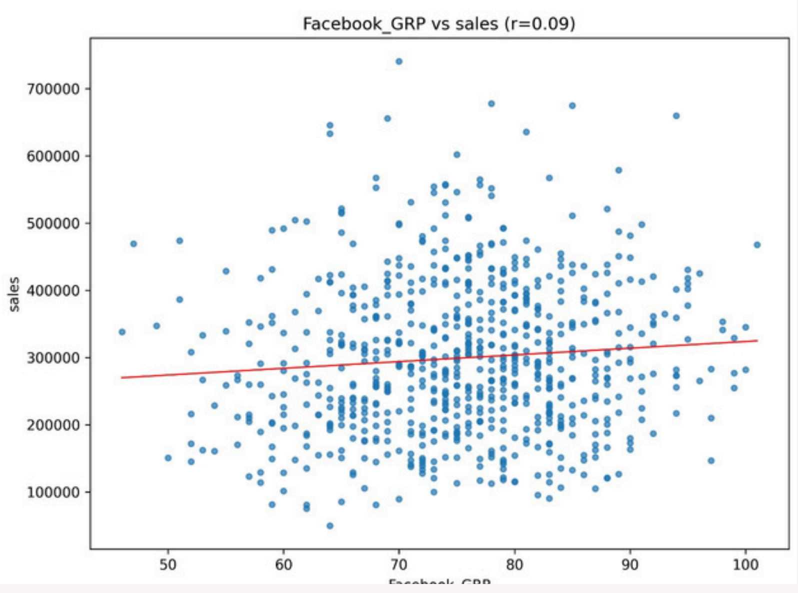
RESULTS

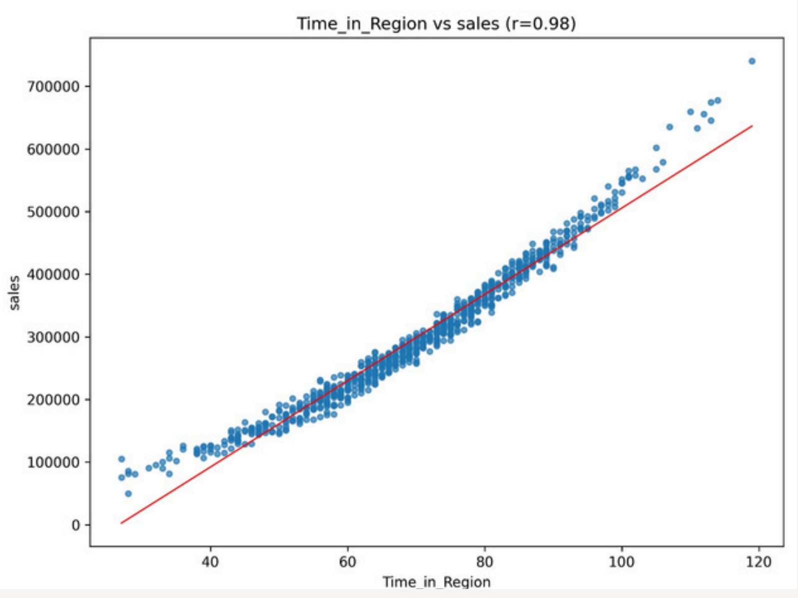
1. Key Correlations
   - Time_in_Region showed a strong positive correlation with sales (r = 0.98) — the dominant predictor.
   - Facebook_GRP, Choc_Capital_Distance, Gender, and Sustainability_Index showed weaker but consistent trends, suggesting potential nonlinear effects.
2. ANOVA Insights

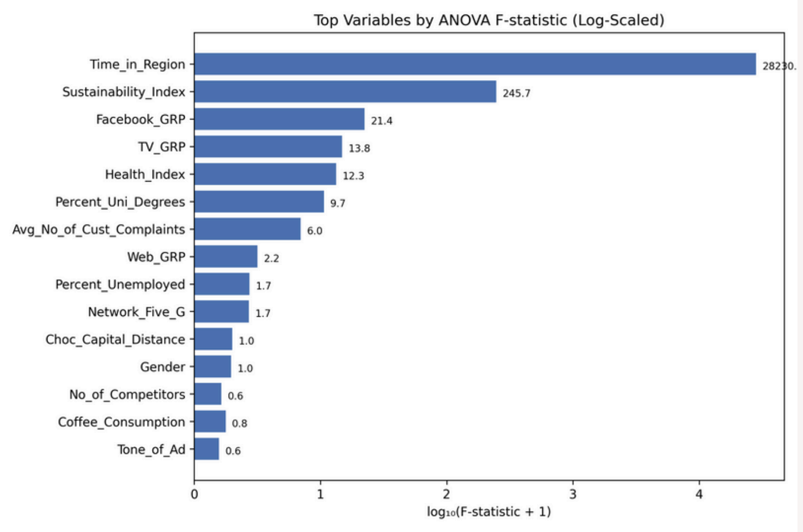
   The ANOVA F-statistic highlighted the most influential predictors:
   - Extreme Impact: Time_in_Region (F = 2823.0)
   - High Impact: Sustainability_Index (245.7), Facebook_GRP (21.4), TV_GRP (13.8)
   - Low Impact: Tone_of_Ad, Coffee_Consumption, No_of_Competitors (F < 10)
3. Interpretation

   Results confirm that geographic exposure (Time_in_Region) drives sales, while marketing variables play a secondary but complementary role.

   The contrast between linear and nonlinear behaviors supports the need for ensemble and hybrid models to capture complex relationships.







CONCLUSIONS

The statistical analysis confirms Time_in_Region is the dominant predictor. However, it also shows that key marketing variables (like Facebook_GRP) have a weak linear influence , even though the ANOVA analysis proves they are statistically relevant.

The hybrid modular design enables testing of multiple regression approaches, improving stability and addressing data sensitivity.

BIBLIOGRAPHY

Asuero, A. G., Sayago, A., & González, A. G. (2006). The Correlation Coefficient: An Overview. Critical Reviews in Analytical Chemistry, 36(1), 41–59. https://doi.org/10.1080/10408340500526766

Terrádez, M., & Juan, Á. A. (s.f.). Análisis de la varianza (ANOVA). Universitat Oberta de Catalunya.