Microenterprise Density Prediction System: Systemic Design, Architecture, and Chaos Management

Daniel Felipe Gómez Miranda
systems engineering
Universidad Distrital Francisco José de Caldas
Email: dfgomezm@udistrital.edu.co

Julian David Cabrera Barragan
systems engineering
Universidad Distrital Francisco José de Caldas
Email: jdcabrerab@udistrital.edu.co

Andrés Julián Vargas Medina
systems engineering
Universidad Distrital Francisco José de Caldas
Email: ajvargasm@udistrital.edu.co

Geraldine Alejandra Vargas Moreno
systems engineering
Universidad Distrital Francisco José de Caldas
Email: geavargasm@udistrital.edu.co

***Abstract*—This paper extends the systemic analysis of the microenterprise density prediction system initiated in Workshop 1 by presenting a complete technical design. It defines measurable requirements, proposes a modular architecture, and establishes adaptive mechanisms for managing sensitivity and chaos in socioeconomic forecasting. The proposed framework integrates traditional statistical models and modern machine learning approaches to ensure stability, scalability, and robustness against uncertainty.**

***Index Terms*—Microenterprise density, Systems engineering, Chaos management, Sensitivity, Machine learning, Forecasting.**

I. Introduction

Building on the foundations of Workshop 1, this paper focuses on transforming the systemic understanding of the GoDaddy Microbusiness Density Forecasting problem into a practical and resilient system design. The work identifies measurable system requirements, develops a multi-layer architecture, and outlines mechanisms for handling data noise, model sensitivity, and chaotic behavior in socioeconomic data systems.

II. Review of Findings from Workshop No. 1

The initial analysis revealed critical constraints such as incomplete and noisy data, missing explanatory variables, and the influence of uncontrollable external events (e.g., crises or pandemics). The data exhibits high dimensionality, multicausality, and nonlinear behavior, increasing the model's complexity. Small variations in predictors (income, credit) can drastically alter results, demonstrating sensitivity to initial conditions. Thus, future models must be:

- Robust to noise and missing data.
- Regularized to prevent overfitting.
- Continuously retrained to adapt to new data patterns.

III. System Requirements Definition

Based on the findings, measurable technical and user-centered requirements were established (Table I).

TABLE I
System Requirements and Measurement Criteria

Category	Requirement	Criterion
Performance	Prediction error $< 10\%$ of true density	IoU ≥ 0.9
Reliability	Stability under $\pm 2\%$ data variation	Sensitivity $\leq 5\%$
Scalability	Handle multiple regions and time periods	Linear runtime
Adaptability	Automatic retraining per update	Periodic retraining
Robustness	Manage noise and missing data	95% success rate

User-centered goals include intuitive interfaces, interpretability of model results, and secure data access for socioeconomic information.

IV. High-Level System Architecture

The proposed architecture consists of six functional layers, ensuring modularity, scalability, and resilience (Fig. 1).

- **Data Ingestion:** Collects time series and socioeconomic variables from sources like Kaggle and institutional databases.
- **Processing and Cleaning:** Performs interpolation, normalization, and outlier detection to ensure data quality.
- **Feature Engineering:** Creates derived variables (lags, moving averages, and external indicators).
- **Modeling and Prediction:** Combines ARIMA, ETS, Prophet, XGBoost, Random Forest, and LSTM for hybrid modeling.

- **Evaluation and Calibration:** Evaluates results using RMSE, MAE, and IoU metrics; applies calibration for fairness across regions.
- **Deployment and Feedback:** Publishes results through dashboards or APIs with continuous monitoring and retraining loops.

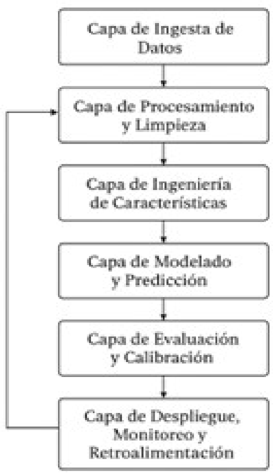

Fig. 1. High-level architecture of the proposed prediction system.

A. Systems Engineering Principles

The architecture follows principles of modularity, scalability, and traceability. Containerization (Docker), workflow automation (Airflow), and version control (Git) ensure reproducibility and maintainability.

V. Sensitivity and Chaos Management

Socioeconomic forecasting involves chaotic dynamics caused by external shocks and sensitivity to predictors. Identified instability sources and mitigation strategies are presented in Table II.

TABLE II
Instability Sources and Mitigation Strategies

Source	Strategy
Abrupt socioeconomic changes	Regularization (L2, dropout); variable selection via correlation.
Missing or noisy data	Robust interpolation, anomaly detection (Isolation Forest, LOF).
Model instability	Ensemble learning for smoothed predictions.
External disturbances	Scenario simulation for crisis forecasting.
Feedback drift	Adaptive retraining triggered by statistical drift.
Prediction uncertainty	Confidence intervals via bootstrapping.

A. Monitoring and Adaptive Feedback

A continuous monitoring subsystem includes:

- Dashboards for MAE, RMSE, and drift detection.
- Error logging and automatic alert triggers.
- Sensitivity indicators inspired by Lyapunov exponents.

An adaptive feedback cycle ensures stability:

$$\text{Data Update} \rightarrow \text{Sensitivity Evaluation} \rightarrow \text{Model Adjustment} \rightarrow \text{Deployme}$$

VI. Technical Stack and Implementation

Python serves as the main development environment:

1) **Data preparation:** using `pandas` and `numpy`.
2) **Visualization:** via `matplotlib` and `seaborn`.
3) **Modeling:** using `LightGBM`, `XGBoost`, and statistical baselines.
4) **Optimization:** with `Optuna` for hyperparameter tuning.
5) **Tracking:** managed by `MLflow`.
6) **Deployment:** using `Docker` and visualization through `Streamlit` or `Power BI`.

VII. Conclusion

This work extends the findings from Workshop 1 by presenting a full architectural framework for predicting microenterprise density. The proposed design addresses complexity, sensitivity, and chaotic factors using hybrid modeling and adaptive feedback. The result is a resilient, scalable system capable of learning from its environment and maintaining stability in dynamic socioeconomic contexts.

References

References

[1] C. E. Shannon, “A Mathematical Theory of Communication,” *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[2] Kaggle, “GoDaddy Microbusiness Density Forecasting Competition,” 2023. [Online]. Available: https://www.kaggle.com/competitions/godaddy-microbusiness-density-forecasting

[3] T. Akiba *et al.*, “Optuna: A Next-generation Hyperparameter Optimization Framework,” *KDD*, 2019.

[4] S. Taylor and B. Letham, “Forecasting at Scale,” *PeerJ Preprints*, 2017.