Microenterprise Density Prediction System

Final Project Presentation

Daniel Felipe Gómez Miranda   Julian David Cabrera Barragán   Andrés Julián Vargas Medina   Geraldine Alejandra Vargas Moreno

Computer Engineering Program
Universidad Distrital Francisco José de Caldas

December 12, 2025

Project Overview

Objective:

- Forecast microenterprise density across U.S. counties
- Handle socioeconomic complexity and uncertainty
- Build a robust, scalable predictive system

Data Source:

- GoDaddy Microbusiness Density Forecasting
- 3M+ county-level observations
- Temporal + socioeconomic variables

Project Evolution

Workshop 1: Systemic Analysis
Workshop 2: Architecture Design
Workshop 3: Quality & Risk Management
Workshop 4: Simulation & Validation

Key Challenge

Managing chaos, nonlinearity, and external shocks in socioeconomic systems

Workshop 1: Systemic Analysis

System Characteristics Identified:

- **Multicausality:** Income, unemployment, demographics
- **Nonlinearity:** Complex variable interactions
- **Sensitivity:** Small changes $\rightarrow$ large effects
- **Chaos:** External shocks (crises, pandemics)

Key Insight:

"The system behaves as a dynamic socioeconomic ecosystem requiring adaptive modeling and continuous monitoring"

Complexity Factors

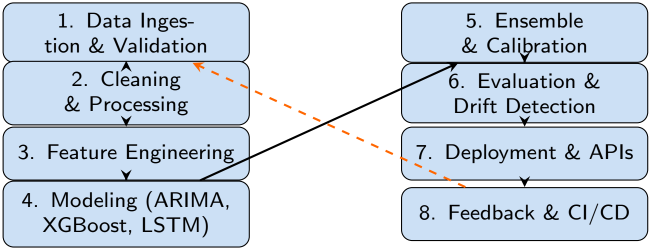
- High dimensionality
- Temporal dependencies
- Data noise & gaps
- Unpredictable events

Solution Approach

Robust preprocessing +
Hybrid modeling +
Feedback loops

Workshop 2: System Architecture

Eight-Layer Modular Architecture



Standards: ISO 9000, CMMI Level 3, Six Sigma

Workshop 3: Quality Assurance & Risk Management

Quality Controls:

- Schema validation
- Anomaly detection
- Drift monitoring
- Traceability documentation
- Automated testing

Risk Analysis:

Risk	Mitigation
Data integrity	Schema validation
Model drift	Auto-retraining
Security	Role-based access
Coordination	Clear roles & tools

Project Management

Methodology: Agile-Scrum
Tools: GitHub, Trello, MLflow
Roles: Manager, Analyst, Developer, Tester

Result

Robust, traceable, production-ready system

Workshop 4: Simulation Framework

Dual Simulation Approach

Data-Driven Simulation

Purpose: Baseline evaluation

Approach:

- Real historical data
- Random Forest model
- Stable conditions

Results:

- RMSE = 0.0727
- No drift detected (p ¿ 0.999)
- High predictive accuracy

Event-Based Simulation

Purpose: Stress testing

Perturbations:

- Income shocks (-15%)
- Unemployment spikes
- Noise injection

Observations:

- System sensitivity confirmed
- Recovery mechanisms activated
- Ensemble stabilization effective

Both approaches validate architectural robustness

Machine Learning Results

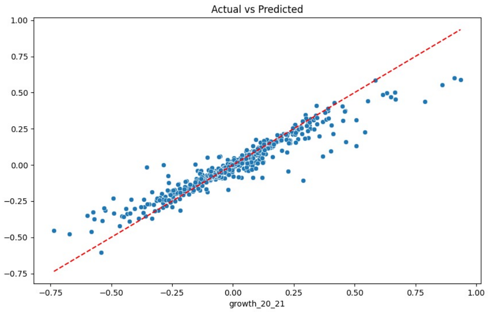
Random Forest Model:

- 100 estimators
- Median imputation
- Standard scaling
- Temporal features

Performance Metrics:

Metric	Value
RMSE	0.0727
MAE	0.0589
Drift p-value	0.9999

Actual vs Predicted



Interpretation

Model captures temporal trends with high fidelity and maintains stability across training cycles

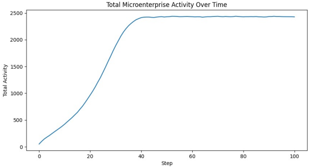
Configuration:

- Grid: 50×50 cells
- Steps: 100 iterations
- Gaussian noise: $\sigma = 0.05$
- Growth threshold: 0.6
- Decay probability: 0.02

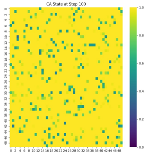
Emergent Phases:

1. Growth (steps 0-40)
2. Stabilization (steps 40-60)
3. Steady-state (steps 60+)

Activity Over Time



Grid State at Step 100



Model Comparison & Integration

Approach	Characteristics
Random Forest	High accuracy, smooth predictions, robust under stable conditions, low drift
Cellular Automata	Nonlinear dynamics, sensitive to perturbations, emergent behaviors, stress-testing
Historical Baseline	Ground truth patterns, low volatility, reference trajectory

Integrated Interpretation

- **ML model:** Optimal for forecasting and operational deployment
- **CA model:** Reveals chaotic tendencies and resilience capacity
- **Together:** Comprehensive understanding of system dynamics

Dataset Summary

Data Source:

- GoDaddy Kaggle Competition
- County-level U.S. data
- Full temporal coverage
- 3M+ observations

Key Variables:

- `microbusiness_density`
- `population`
- `median_income`
- `unemployment_rate`

Preprocessing Pipeline:

1. Schema validation
2. Missing value treatment
3. Outlier detection (IQR)
4. Normalization (MinMax)
5. Feature engineering (lags, windows)
6. Version control & metadata

Final Dataset Quality

- 100% completeness (target variable)
- 95%+ completeness (features)
- Uniform scaling

Key Results & Achievements

- ✓ **Systemic understanding** of socioeconomic complexity achieved
- ✓ **Eight-layer architecture** designed following ISO 9000, CMMI, Six Sigma
- ✓ **Quality assurance** framework with risk mitigation strategies
- ✓ **Dual simulation approach:** data-driven + event-based validation
- ✓ **ML model accuracy:** RMSE = 0.0727, no drift detected
- ✓ **CA emergent behavior:** growth, stabilization, steady-state phases
- ✓ **Robust project management:** Agile-Scrum with clear roles
- ✓ **Complete documentation:** traceability and reproducibility ensured

Result: Production-ready forecasting system capable of handling uncertainty, adapting to perturbations, and maintaining long-term stability

Lessons Learned

Technical Insights:

- Chaos management requires ensemble methods
- Drift detection is critical for long-term stability
- Feature engineering drives model performance
- Redundancy improves fault tolerance

Methodological Insights:

- Iterative refinement essential
- Clear documentation prevents errors
- Version control enables reproducibility

Project Management:

- Agile approach provided flexibility
- Role definition reduced conflicts
- Weekly sprints maintained momentum
- Early feedback improved quality

Interdisciplinary Integration:

- Systems thinking + ML = holistic solution
- Simulations validate theoretical insights
- Engineering standards ensure rigor

Future Work

Technical Enhancements:

- Cloud deployment (AWS/Azure)
- Real-time data streaming
- Deep learning models (LSTM, Transformer)
- SHAP-based interpretability
- Multi-region comparative analysis

System Expansion:

- Interactive dashboards for policymakers
- Mobile application for field analysis
- Integration with external APIs
- Enhanced CA rules with richer variables
- Automated report generation

Long-term Vision

Transform the system into a comprehensive decision support platform for economic development and microenterprise policy planning

Project Summary

The Microenterprise Density Prediction System demonstrates how **systems engineering methodologies** can effectively model **complex socioeconomic phenomena**.

Core Contributions

- Integrated systemic analysis, robust architecture, and validated simulations
- Combined ML predictions with CA emergent behavior modeling
- Established quality assurance and risk management framework
- Delivered scalable, adaptive, production-ready forecasting platform

The system is ready for deployment and continuous improvement

Thank You!

Questions?

Project Repository:
`https://github.com/Sukedas/Systems-Analysis-and-Design-Project`

Universidad Distrital Francisco José de Caldas
Computer Engineering Program