BASF — Grower Segmentation Analysis

Devam Parekh, Tristan Dietiker, Jeffrey Zhou, Grant Weingaertner, Nakshatra Bobbili, Shaurya Beriwala, Amit Jadhav

PURDUE UNIVERSITY | The Data Mine

BASF We create chemistry

INTRODUCTION

Understanding grower personas is critical for designing effective agricultural marketing strategies and improving customer engagement. BASF categorizes its customers into five distinct persona segments, labeled A-E, with each segment representing different purchasing behaviors, product preferences, and decision-making patterns.

This project explores whether transactional purchase data alone can be used to predict customer persona segments using machine learning. To address this, we developed and evaluated multiple classification approaches, including a two-phase stacking ensemble, and assessed several feature engineering strategies across five experiments.

Project Goals

- Predict customer personas from transactional purchase data using ML
- Engineer informative features from brand and product purchasing behavior
- Build and evaluate a stacking ensemble combining Logistic Regression, Random Forest, XGBoost, and LightGBM across multiple feature engineering strategies

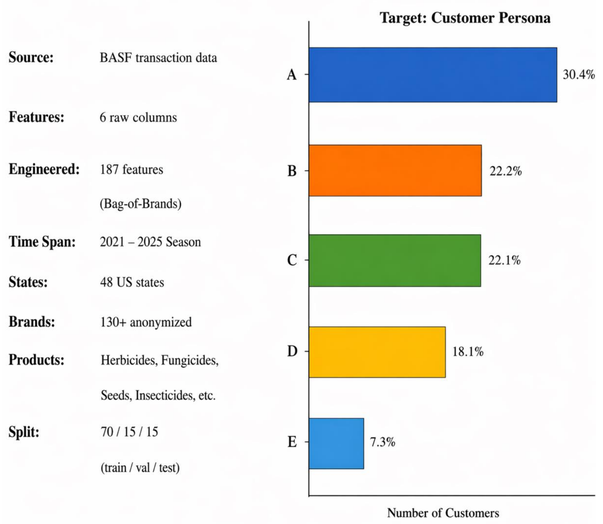
Motivation / Impact

- Enable data-driven customer segmentation for targeted marketing
- Identify the most discriminative purchasing behavior signals
- Determine if transactional data alone carries sufficient signal for persona prediction
- Provide actionable insights for BASF's agricultural customer engagement strategy

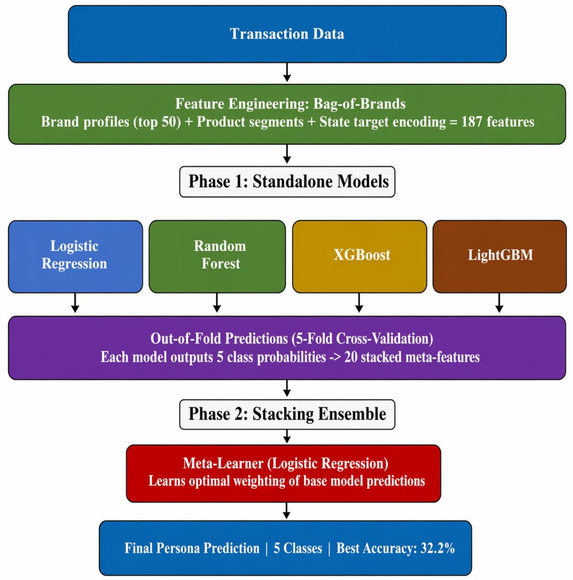
METHODOLOGY

DATASET OVERVIEW

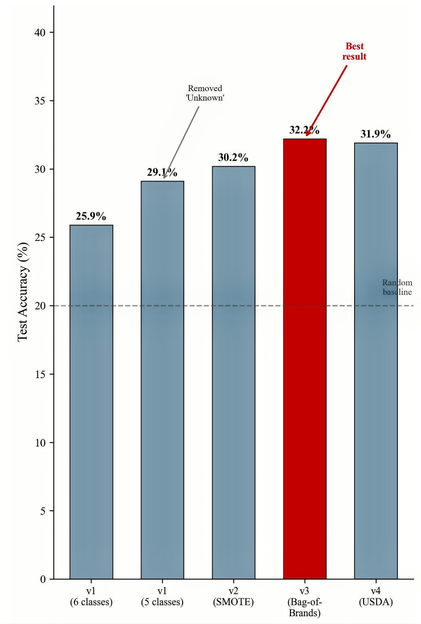
Source:	BASF transaction data
Features:	6 raw columns
Engineered:	187 features (Bag-of-Brands)
Time Span:	2021 – 2025 Season
States:	48 US states
Brands:	130+ anonymized
Products:	Herbicides, Fungicides, Seeds, Insecticides, etc.
Split:	70 / 15 / 15 (train / val / test)



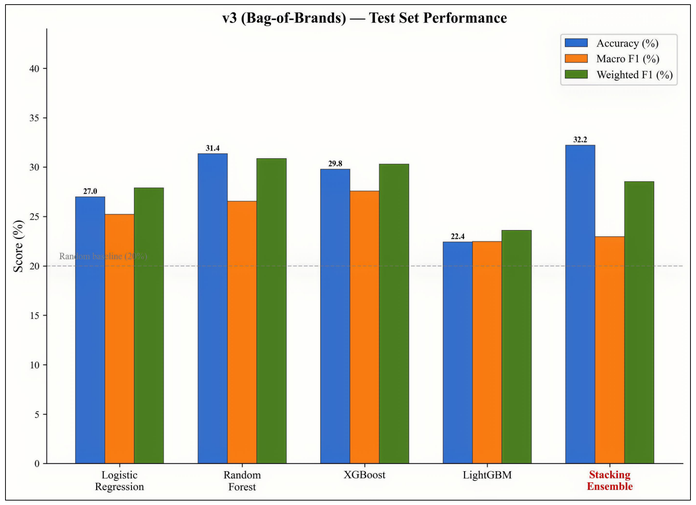
STACKING ENSEMBLE ARCHITECTURE



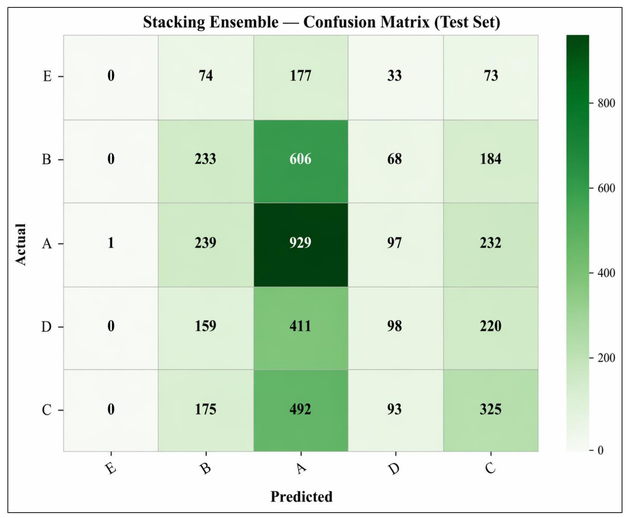
EXPERIMENT PROGRESSION



RESULTS



Best Model: Stacking Ensemble (32.2% accuracy)
Random Forest strongest standalone (31.4%)
Bag-of-Brands features most effective strategy



Ensemble predicts Persona A best (62% recall)
Persona E nearly impossible to predict (0% recall)
Model overwhelmingly predicts majority classes (Persona A, Persona B)

CONCLUSIONS

- Stacking Ensemble achieved best test accuracy of 32.2% using Bag-of-Brands features
- Brand identity (which brands a customer buys) provides the strongest signal
- Spending patterns are nearly identical across all 5 persona types
- No single feature exceeds 1.5% importance — signal is spread thinly across 187 features
- Personas are likely defined by psychographic/survey data not present in transaction records

FUTURE GOALS

- Enrich dataset with survey, digital engagement, and agronomic profile data
- Explore reducing to 3 super-classes for higher prediction accuracy
- Implement probability-based soft targeting for marketing campaigns

ACKNOWNLEDGEMENTS

We would like to sincerely thank Melissa Barona, our BASF mentor, for her guidance throughout this project; Saksham Singh, Teaching Assistant and Project Lead at The Data Mine, for his support and leadership throughout the semester; and Cai Shun Chen, Corporate Partners Technical Specialist at The Data Mine, for his valuable guidance and assistance throughout the project.