SMART COOKIES:
A PREDICTIVE APPROACH TO GIRL SCOUT COOKIE SALES

Purdue University — Mitchell E. Daniels, Jr. School of Business

Rishita Korapati | Gabriel Morales | Ramya Chowdary Polineni

rkorapat@purdue.edu | moral136@purdue.edu | rpolinen@purdue.edu

ABSTRACT | BUSINESS PROBLEM

Every year, **thousands of Girl Scouts** rely on cookie sales for fundraising. The current forecast method—based solely on the **previous year's numbers**—explains only **70% of sales variability.** This limitation often leads to **missed revenue opportunities** or **excess stock**. By leveraging **advanced predictive models**, we can **bridge this gap** and **unlock new opportunities** for **growth, efficiency, and increased fundraising success.**

KEY BENEFITS

Increased Inventory Management & Cost Savings → Enhanced Marketing Strategies → Increased Troop Engagement → Data Driven Decision Making

Research Questions

1. Can **machine learning models** effectively integrate historical sales data, troop participation rates to **improve the accuracy of cookie sales forecasts** beyond traditional methods?
2. How can insights from predictive models be used to **optimize inventory management and marketing strategies** for Girl Scout troops, ensuring that each troop meets demand without costly surplus or shortages?

ANALYTICAL PROBLEM

Analytical Context: The context involves analyzing **historical sales data and external factors** to improve forecasting accuracy by troop and cookie type.

Success Metrics: Success is measured by the **reduction of the forecast error (RMSE) from 12 to 9** and increasing alignment between projected demand and final orders.

Challenges: Challenges include **high variance in sales across troops and regions**, impact from **external factors like weather and local events, and ensuring model reliability for troop leaders**.

Solution Focus: The solution focuses on **implementing machine learning models** to enhance **forecasting accuracy** and optimize inventory **management.**

Fig 1. **Cookie Funnel** AI image generated by DALLE)

DATA DICTIONARY

Key Variables

Date: Sales transaction date.
Number of Cases Sold: Total boxes sold.
Cookie Type: Different cookie varieties.

Troop ID: Identifies the troop responsible for sales.
Number of Girls: Number of girl scouts participating in sales.
Period: Specific sales time window.

Data Insights

Total Rows: 68,966 **Total Columns:** 6	**Dataset** covers multiple sales periods with different troop participation rates
Outlier Treatment: Removed extreme cases after validation.	**Missing Values:** Minimal missing values, handled through imputation.

Troop Sales Data → Historical Sales Trends → Sales Period

→ Cookie Type Sales Performance → Future Demand Predictions → Number of Girls Participating → Sales Performance

PROJECT METHODOLOGY

1. Data Understanding
- Entity relationship mapping of troop sales.
- Understanding categorical and numerical variables.
- Initial data exploration and cleaning.

2. Exploratory Data Analysis (EDA)
- Segmentation of cookie sales by troop.
- Time-series decomposition to identify patterns.
- Graphical analysis of seasonality and trends.

3. Data Preprocessing
- Remove irrelevant columns & missing values.
- Scale input features using Standard Scaler.
- Apply data transformation & feature engineering.

4. Modeling & Validation
- Group data by troop & cookie type.
- Split into training (periods 1-4) & testing (period 5).
- Test models: Ridge, Random Forest, Polynomial, XG Boost, Linear Regression.
- Validate predictions using RMSE.

5. Predictive Modeling
- Dynamic selection of the best model based on validation results like MAE, MSE, RMSE, MAPE, R^2.

LANGUAGE WE USED:
Python

6. Reporting & Insights
- Website and demo for sales trends forecast and business use.
- Prediction of next sales cycle quantities.
- Recommendations report for troop-level forecasting strategies.

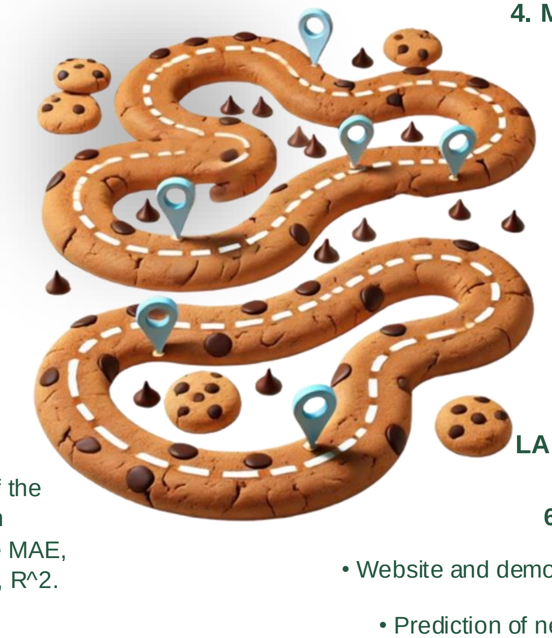


Fig 2. **SEMMA Roadmap** AI image generated by DALLE)

MODEL SELECTION

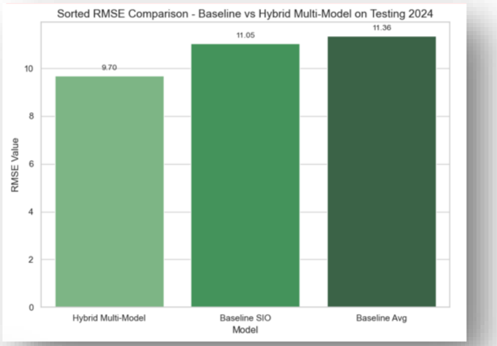
ASSUMPTIONS:
- The model assumes **past sales patterns** are predictive of **future sales performance.**
- **No major disruptions** in cookie availability, troop operations, or supply chains are expected.

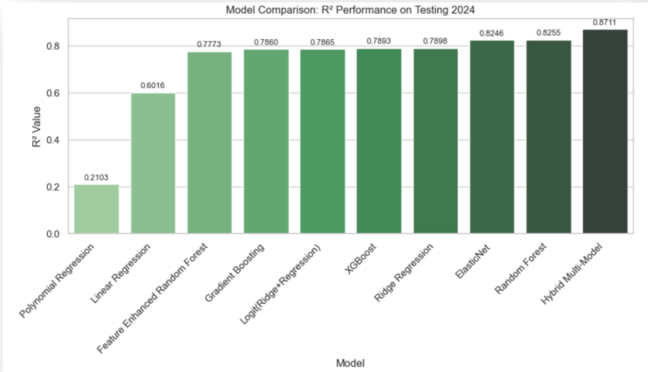
LIMITATION:
- The model may **underperform** for troops with **limited or erratic historical data**.

Fig 3. **Cookie Box** AI image generated by DALLE)

SPLITTING:
- The dataset is grouped by **troop ID** and **cookie type**. For each group, the data is **split by year** into **Training (2020–2023)** and **Testing (2024).**
- Cluster-based modeling is applied within **each location and cookie type**, allowing for more personalized predictions.

BASE MODEL APPROACHES:
- **SIO Model:** Uses last year's sales and troop participation to estimate.
- **Avg Model:** Averages past sales from 2021-2023 to predict 2024. However, these models had **higher RMSE values**, indicating **significant prediction errors**.





MODEL SELECTION:

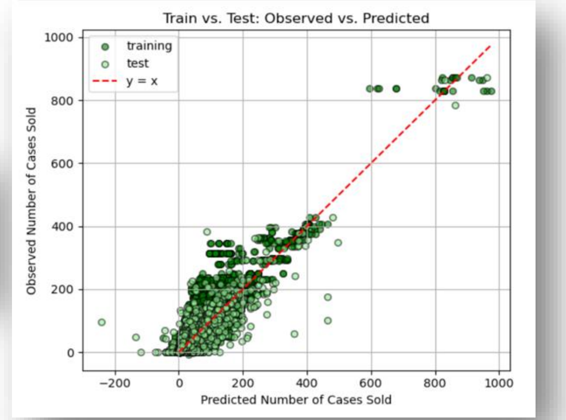
To improve accuracy, we built a **Hybrid Multi-Model system** that automatically selects the best among **Clustered Ridge Regression**, **Troop-Level Ridge Regression**, **Linear Regression**, **SIO & Average**, and **Location-Level Ridge Regression.** Each troop-cookie prediction uses the method with the **lowest error**, dynamically chosen based on past performance. We tested other models and found the Hybrid had the best performance.

VALIDATION

- **Confidence in Predictions**: Model predictions are validated using **cross-validation (CV)** within each training group to optimize regularization strength (λ), minimizing overfitting and ensuring stable performance across troops and cookie types.
- **Dynamic Error-Based Method Selection**: For each troop-cookie pair, the model dynamically selects the prediction method with the **lowest expected error (MSE)**, based on past performance. This approach ensures predictions are **customized and evidence-driven**.
- **Robustness Across Segments**: The use of **clustering + Ridge + fallback heuristics** allows the model to adapt to sparse, dense, and even noisy troop histories—leading to **consistent accuracy improvements over baseline models**.

FINAL MODEL RESULTS

Metric	Value
MAE	4.59
MSE	94.07
RMSE	9.7
R^2	0.8711



KEY FINDINGS:

Our **Hybrid Multi-Model approach**, which blends Ridge Regression (with CV), linear models, and PGA heuristics, achieved the **highest R^2** and **lowest error metrics**, highlighting its adaptability and accuracy. The model **dynamically selects the best prediction method** (from 6 options) for each troop-cookie pair, based on historical performance and data quality.

By **improving prediction accuracy** by **1.35 cases per troop per cookie type** compared to the SIO tool, our model provided a **significant advantage** in planning and inventory. Scaled across **1,401 troops**, **8 cookie types**, and **12 boxes per case**, this translates to a potential impact of **over 181,000 boxes**—equivalent to more than

$1,089,000 in value generated

This enables GS Indiana to **optimize inventory, reduce over-ordering, and boost revenue** for smarter, more profitable decisions.

MODEL LIFE CYCLE MANAGEMENT

Enhance Data Collection & Quality by maintaining detailed sales histories for more precise forecasting and better model generalization.

Incorporate External Factors like adding weather, marketing efforts, and regional trends to further refine forecasting accuracy.

Use this model for troop-level sales forecasting to minimize errors and maximize predictive accuracy.

Regularly retrain the model with new data, gather user feedback, and refine the model based on new insights and changing business needs.

ACKNOWLEDGEMENTS

We appreciate and extend our gratitude to **Professor Davi Moreira** for his invaluable guidance and support throughout this project. We also acknowledge **the Girl Scouts of Central Indiana** for providing the essential data that played a crucial role in the success of this research. Additionally, we would like to thank **the Krenicki Center** for its support and resources that contributed to the project's development.