Breadiction

Predicting Demand for Bakery Products

Erin Choi, Vicky Lin

Background

Problem

Predicting demand for food (bakery products) to reduce waste

Climate Change Connection

Food waste = 6% of global GHG emissions [1]

Quantifying demand prevents overproduction

Targeting food waste at the retailer and company stage of the supply chain

Previous ML Approaches

Regression: Predicting demand based on recipes [2]

CNN: Time-series analysis to predict hourly sales [3]

Methods: Data Preprocessing

Grupo Bimbo Inventory Demand (Kaggle) - data in Spanish [4]

Product

Product ID	Product name

Product ID	Product Type	Product Weight	Product Brand

Client

Client ID	Client name

Client ID	Client Type

State

Location ID	Town	State

Location ID	State

Data Preprocessing Cont.

Dataset Features After Feature Engineering

- State (33)
- Client Type (13)
- Product Type (11)
- Product Brand (55)
- Product Weight (grams)
- Product Price: sales in Pesos ÷ sales in unit quantity

Data Downsizing

- Downsize data from 75M to 8M rows - computation limits
  - Client: Drop unidentified clients
  - Product: Drop products with no brand, price, or weight
  - Sample 250k transactions from each state (33 states)

Scaling & Encoding

- Robust Scaler
- One-Hot Encoding → 107 features

Methods: Regression Models

80 train / 20 test split, retraining and retesting models with new hyperparameters
Regression problem, target = demand (sales minus returns)

Linear Regression

Lasso Regression

alpha

Ridge Regression

alpha

XGBoost

max_depth
learning_rate

Bagging

Decision tree: max_depth
n_estimators

AdaBoost

Decision Tree: max_depth, criterion
min_sample_split, n_estimators

Results: Metrics

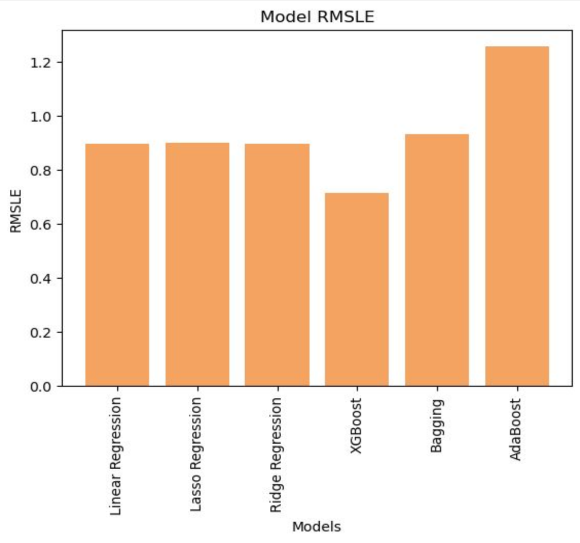
Root Mean Squared Log Error, Mean/Median Absolute Error

	Model	RMSLE	MAE	MedAE
0	Linear Regression	0.896453	6.322293	3.556854
1	Lasso Regression	0.899771	6.353492	3.661482
2	Ridge Regression	0.896465	6.322351	3.556877
3	XGBoost	0.716399	4.884369	2.474169
4	Bagging	0.931823	6.441883	4.831724
5	AdaBoost	1.257163	10.046502	8.493673

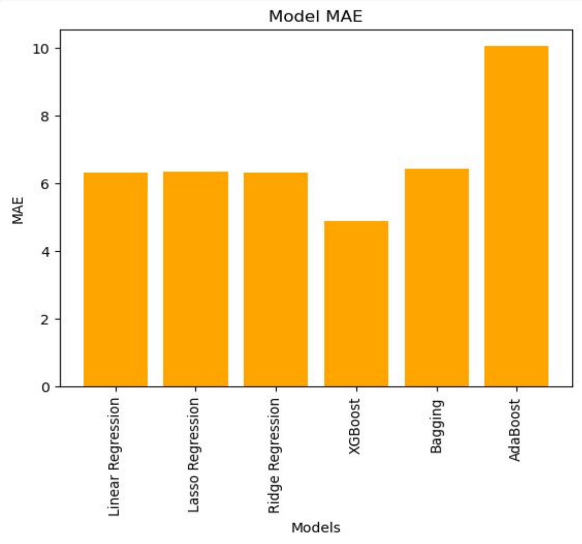
XGBoost outperformed all other models in all metrics

Results: Metrics

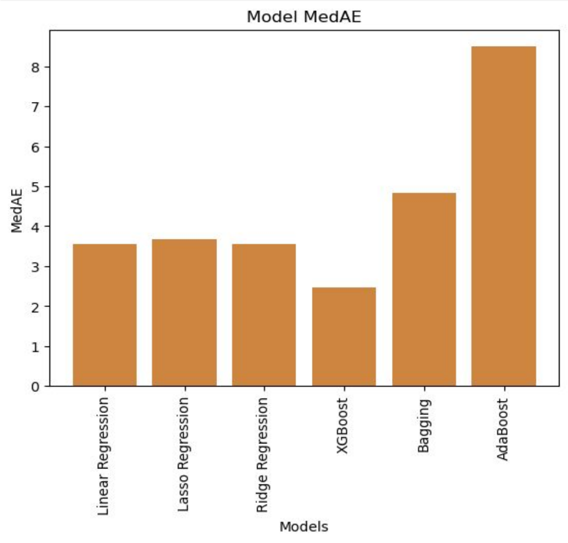
RMSLE



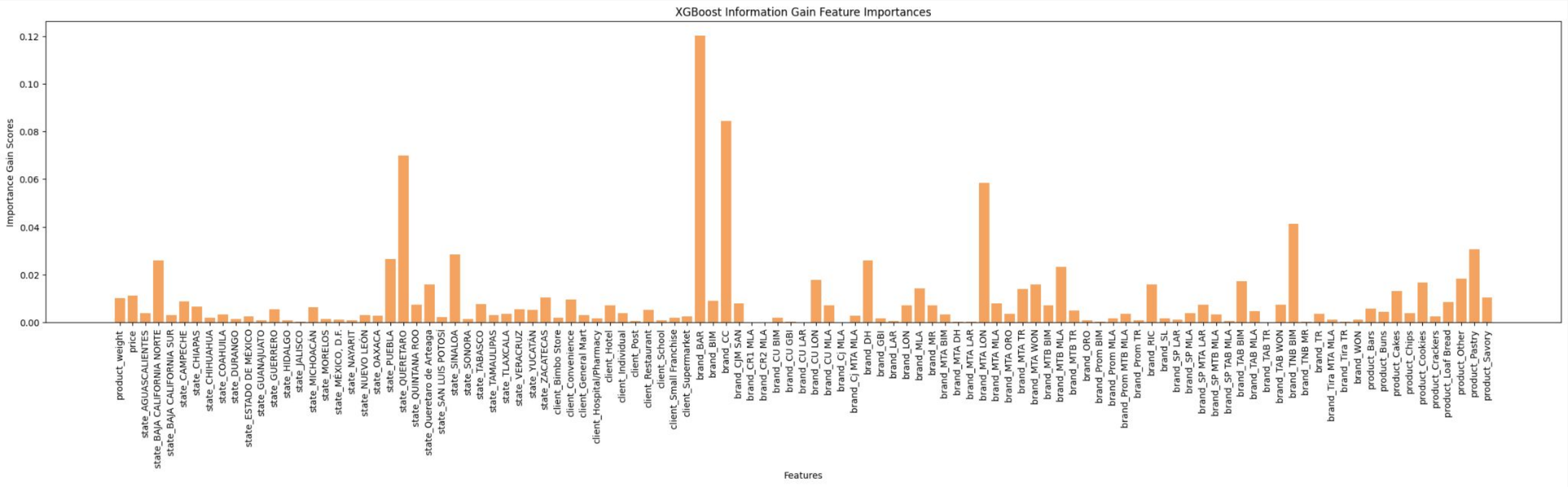
MAE



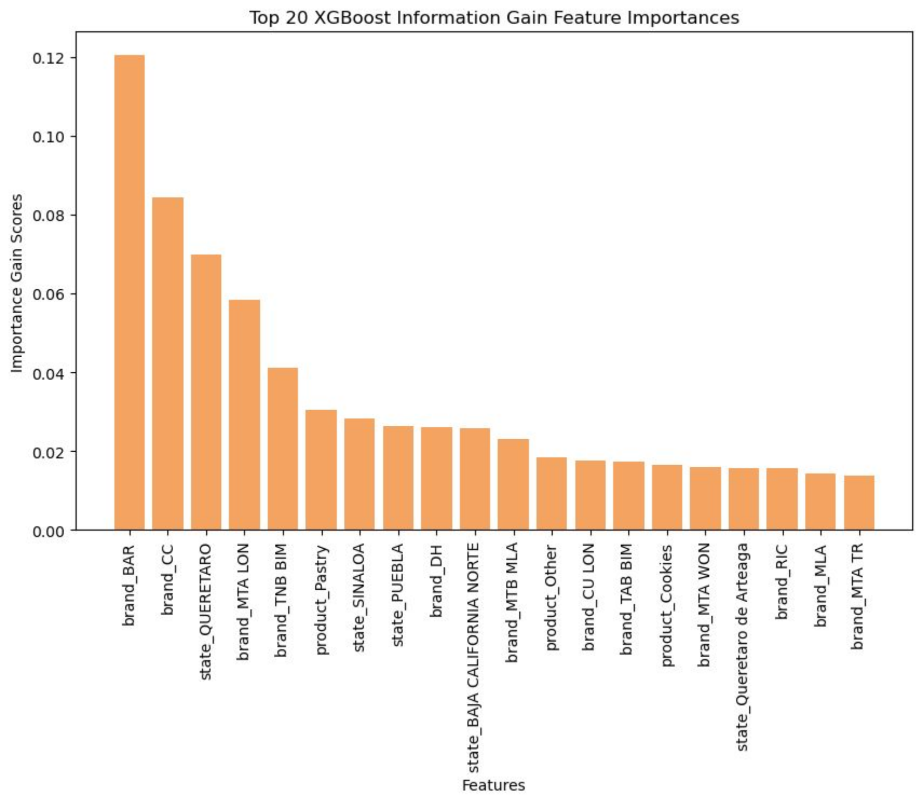
MedAE



Results: XGBoost Feature Importances



Results: XGBoost Feature Importances



	Feature	Importance
0	brand_BAR	0.120298
1	brand_CC	0.084348
2	state_QUERETARO	0.069850
3	brand_MTA LON	0.058350
4	brand_TNB BIM	0.041204
5	product_Pastry	0.030590
6	state_SINALOA	0.028307
7	state_PUEBLA	0.026467
8	brand_DH	0.025971
9	state_BAJA CALIFORNIA NORTE	0.025872
10	brand_MTB MLA	0.023099
11	product_Other	0.018366
12	brand_CU LON	0.017755
13	brand_TAB BIM	0.017302
14	product_Cookies	0.016527
15	brand_MTA WON	0.015913
16	state_Queretaro de Arteaga	0.015825
17	brand_RIC	0.015726
18	brand_MLA	0.014243
19	brand_MTA TR	0.013777

Discussion

Why XGBoost?

- Model is good for numerical & categorical training data with high dimensionality
- Execution speed is faster than other models when trained on large sample sizes
- Uses more accurate approximations to find the best tree model

Compared to Previous Work

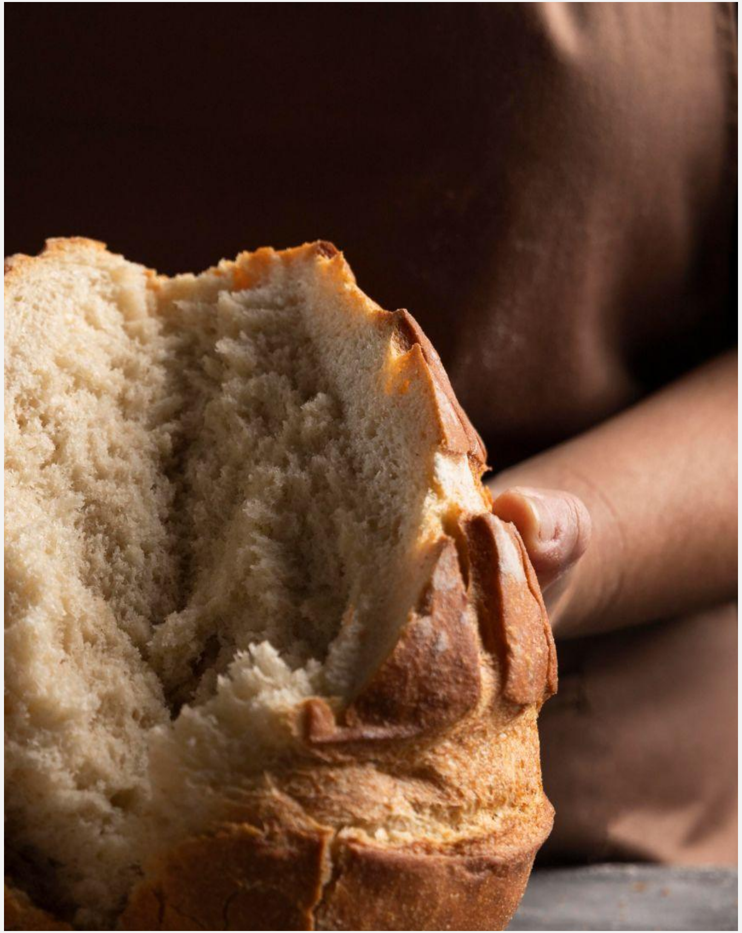
- Previous work's XGBoost model resulted in better performance (smaller errors)
- Need more features and data points to improve performance of models
- Need longer timespan for time-series analysis [3]

Discussion

Future Improvements

- With higher computing power:
    - Use the whole dataset
    - Build a Random Forest Model
    - Further hyperparameter tuning (GridSearch)

- Generalizability
    - Current model is specific to features in Mexico (brands, prices)
    - Requires similar features from other countries for generalization



Thank You!

Q&A

References

1. J. Poore and T. Nemecek, “Reducing food’s environmental impacts through producers and consumers,” *Science*, vol. 360, no. 6392, pp. 987–992, 2018. www.science.org/doi/10.1126/science.aaq0216.
2. A. Garre, M. C. Ruiz, and E. Hontoria, “Application of Machine Learning to support production planning of a food industry in the context of waste generation under uncertainty,” *Oper. Res. Perspect.*, vol. 7, no. 100147, 2020. doi.org/10.1016/j.orp.2020.100147.
3. N. Xue, I. Triguero, G. P. Figueredo, and D. Landa-Silva, “Evolving deep CNN-LSTMs for inventory time series prediction,” *2019 IEEE Congress on Evolutionary Computation (CEC)*, 2019. www.cs.nott.ac.uk/~pszjds/research/files/dls_cec2019.pdf.
4. A. Montoya, Grupo Bimbo, M. O’Connell, and Wendy Kan, “Grupo Bimbo Inventory Demand,” *Kaggle.com*, 2016. www.kaggle.com/competitions/grupo-bimbo-inventory-demand.