Vietnam Retail Diet Diversity

Luke Bravo | Alex McGraw | Jocelyne Walker

Modeling Food Retail Success

We are interested in opening a **food retail outlet** and want to analyze customers, products, and competition to maximize potential store viability, using **shop longevity** as a proxy for success.

Answering these questions can help us understand what factors lead to **stores' success** and identify areas with unmet demand.

Methodology

Cleaning & Analysis

Models & Feature Selection

Findings & Limitations

image: Flaticon.com

Data

Survey results of **563 food retailers**
Hanoi, Vietnam
Collected between January 2017 - January 2018

Over **200 feature columns** in three main categories:

Shop characteristics

- **Years in business**
- Location
- Type of shop, from supermarkets to specialty stores to street stalls

Description of food sold

- Ultra-processed food percentage
- Ready-to-eat food
- Availability of whole/processed foods for each of 19 diet subgroups

Safety claims

- Shop level signage
- Product-level signage
- Formal and informal certifications at each food subgroup level

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/ZWBUEK

Data transformation

Remove identifying characteristics

Factored shop name, exact address, survey timestamp will overfit classifiers.

Derive values to simplify analysis

For example, reduced 16 different certification characteristics to a single binary dimension.

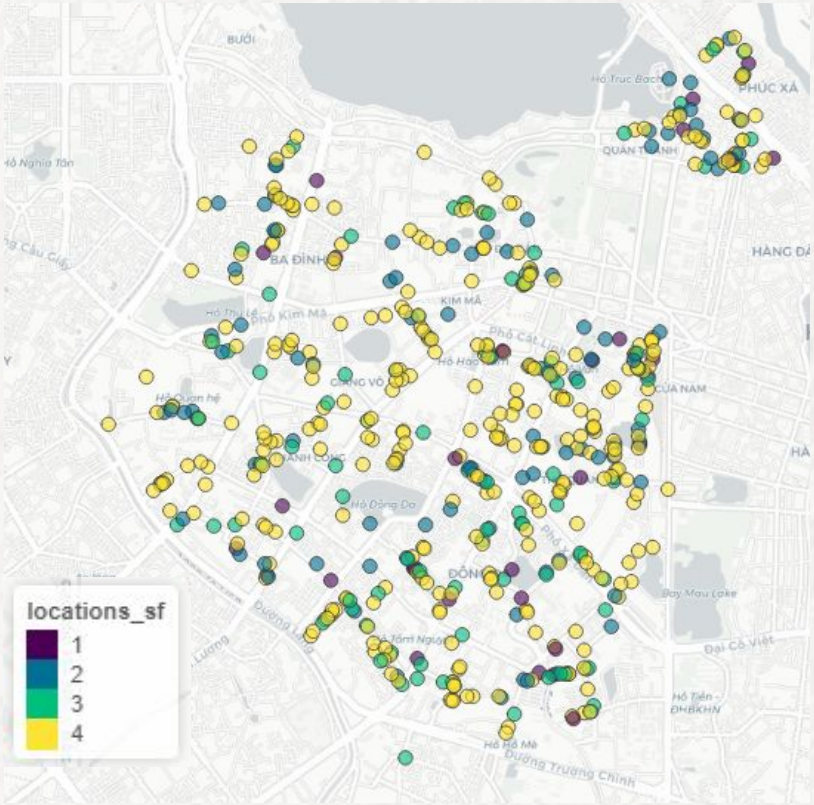
Cleaning data

Translating NAs in raw data to null cells, mapping location data, and dropping rows with null predictor values.

Principal components analysis

Understand the costs and benefits of simplifying predictor variables into components.

Overall, stores are disproportionately older



Store Age	Frequency
< 1 year	33
1-2 years	92
3-5 years	93
>5 years	342

Newer stores are concentrated towards the **east** and **north**

-

60% of stores are >5 years old

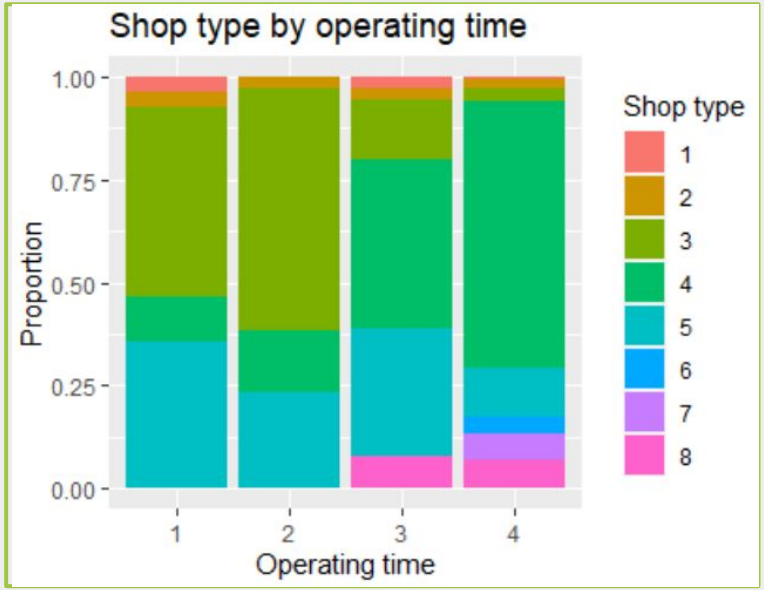
1

Geography plays a factor in **store age** and can be accounted for at the district or lat/long level.

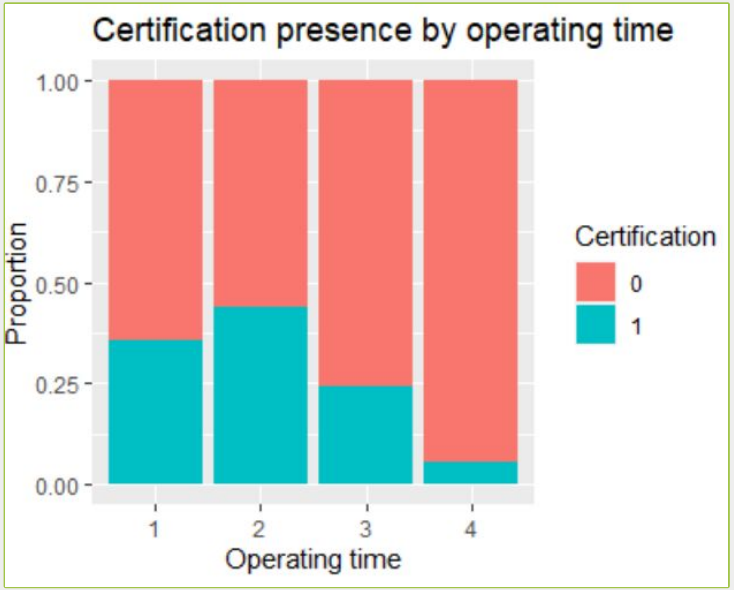
2

Binary models may be better to adjust **distribution** of store age.

Store characteristics varying by operating time

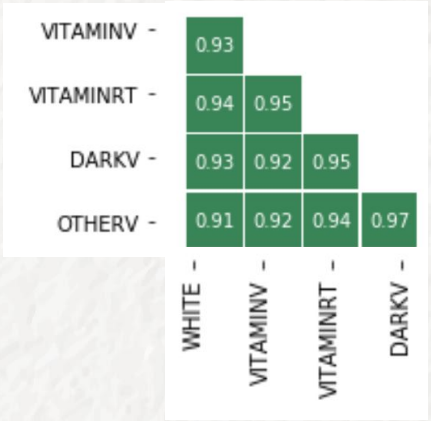


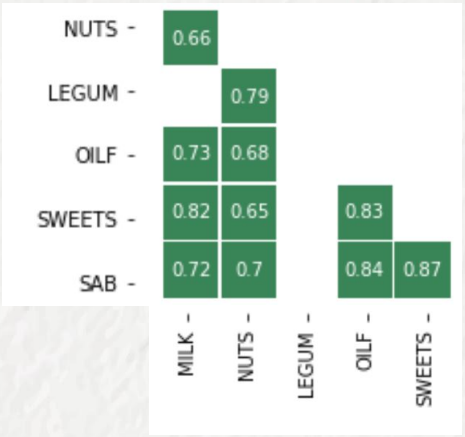
Stores <2 years old are most often **convenience stores and minimarts**, while stores >3 years old are most often **mom-and-pop shops.**



Over 25% of newer stores have posted certifications, while **less than 5%** of stores **greater than 5 years old** post certifications.

Categories of correlated predictor values





Very strong correlation between **vitamin A rich vegetables, roots, and dark leafy greens.**

Strong correlation between **oils/fats**, **spices**, other **base cooking ingredients**, and **sweets.**

Considering PCA

Using **Principal Components Analysis**

Benefits	Tradeoffs
• Reduces data dimensionality • Maximizes explained feature variance • Can reduce overfitting and noise	• Assumes linear relationship • Features lose their original meaning • Not as useful for categorical features

Our models only use **direct variables** from dataset and easily **derived** variables as PCA:

1. **Does not improve classification accuracy**
2. **Reduces interpretability of our results**

Modeling Store Age

Classification Methods

Continuous Linear Model

Proxy the four categories as a continuous predictor variable

Discrete Choice Model

Multinomial logistic regression to predict 1 of 4 categories

Binary Logistic Regression

Simple 0/1 model to identify stores greater than or less than five years old

Ordered Choice Model

Adds ordered context to the multinomial logistic regression

Continuous Model

Baseline Accuracy

60%

Lasso Model

Alpha = 0.005

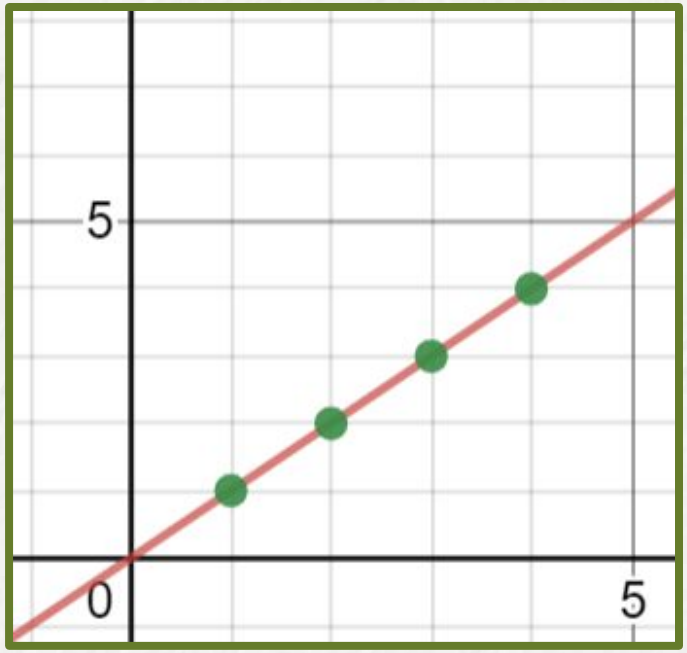
5 Selected Features

R^2: **18.82%**

Accuracy: **20.54%**

Selected Features

Shop Type | Dark Vegetables | Vitamin Fruits
Other Fruits | Certifications



Plotting discrete categories on a continuous line graph.

Binary Logistic Regression

Baseline Accuracy

60%

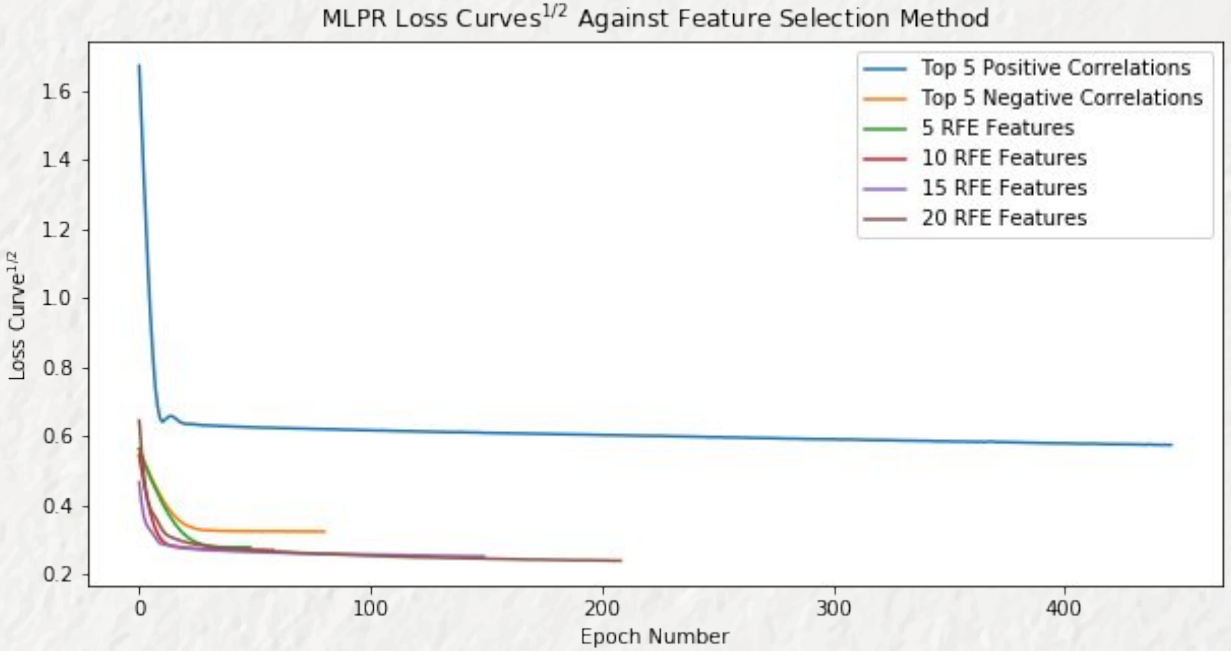
Best MLP Model

10 Features Selected

R^2 Score: **44.67%**

Accuracy: **83.93%**

Employing Multinomial Logistic Regression



Discrete Choice Model

Baseline Accuracy

60%

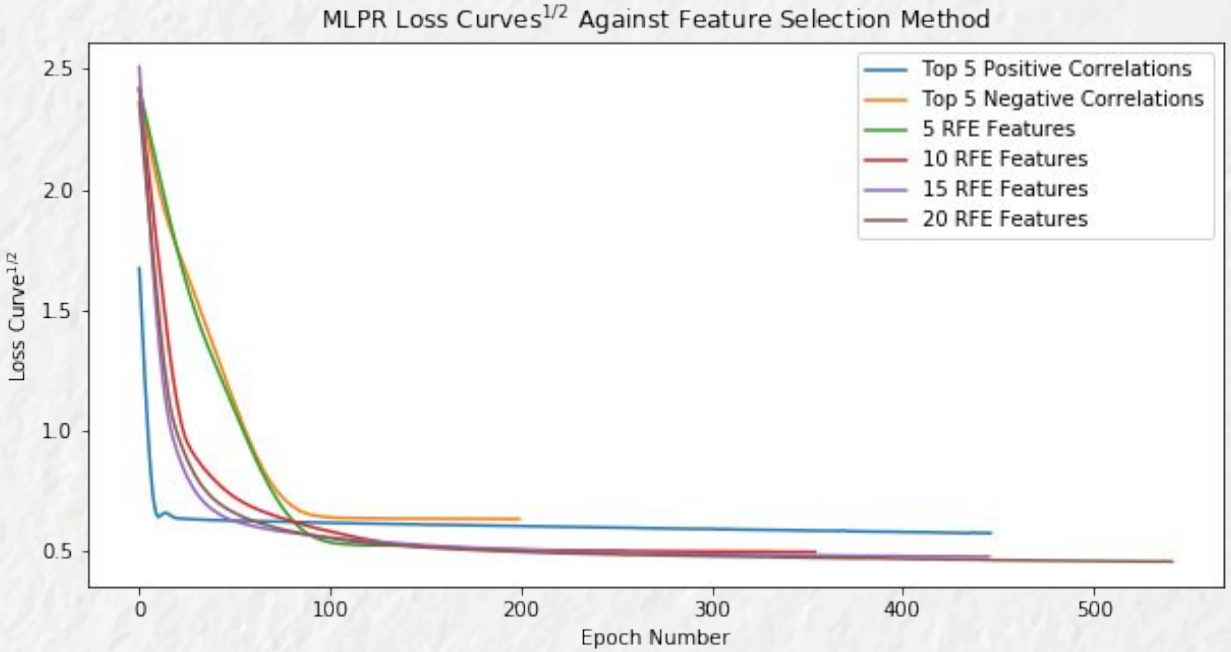
Best MLP Model

10 Features Selected

R^2 Score: **46.89%**

Accuracy: **72.3%**

Employing Multinomial Logistic Regression



Ordered Choice Model

Improving the Multinomial Logit Predictions by Adding Order to Categories

Baseline Accuracy

60%

Best MLP Model

R^2 Score: **37.33%**

Accuracy: **75%**

Six Features Selected

Shop Type
Other Fruits
Organ Meat
Insects
Milk & Milk Products
Oils & Fats

Classification Results

Unordered Multinomial Logistic Regression

Actual / Predicted	< 1 year	1-2 years	3-5 years	>5 years
< 1 year	0	0	0	0
1-2 years	2	11	1	0
3-5 years	1	5	5	2
>5 years	2	3	12	68

75.00% Accuracy

Ordered Multinomial Logistic Regression

Actual / Predicted	< 1 year	1-2 years	3-5 years	>5 years
< 1 year	0	0	0	0
1-2 years	3	16	6	2
3-5 years	0	0	0	0
>5 years	2	3	12	68

75.00% Accuracy

Features Selected and Odds Ratios

Coefficient	Odds ratio
Street food stall (Shop type 8)	2.64
Oils & fats	2.41
Mom-and-pop store (Shop type 4)	1.45
Supermarket (Shop type 2)	1.31

Variables with odds ratios greater than 1 are associated with **older stores**

Variables with odds ratios less than 1 are associated with **younger stores**

Coefficient	Odds ratio
Convenience store/minimart (Shop type 3)	0.09
Insects	0.32
Other fruits	0.41
Specialty food shop (Shop type 5)	0.53
Milk	0.56
Organ meats	0.61

Note: Shop Type 6 and Shop Type 7 only occur in stores >5 years old.

Understanding Hanoi's diet landscape

Supply Side Trends

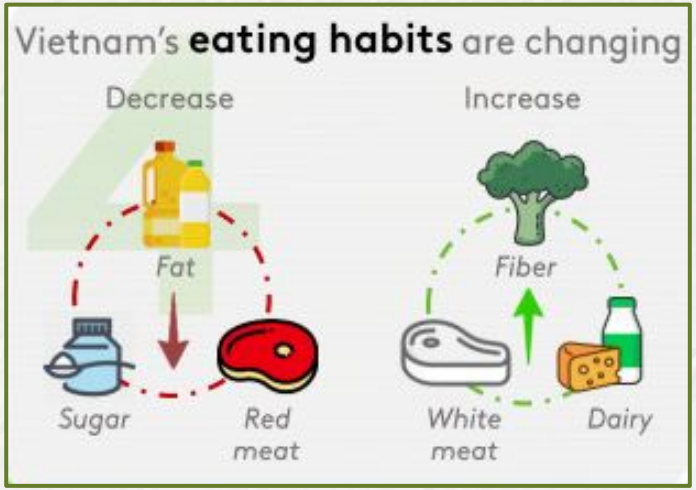
Newer **convenience** stores and minimarts quickly entering market

>40% Established **mom-and-pop** stores continue to dominate food retailing

Demand Side Trends

Increasing **milk** and **meat** demand reflected in newer stores' offerings

Wet **markets** and **fresh** food purchasing still important part of diet behavior



Source: Hanoi Times Diet Study

General Environment Trends

Influence of **globalization** and demographic change on diet trends

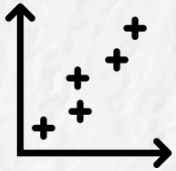
Linking **economic** status of country to retailer profitability and consumer health

Limitations, risks, and extra info

Survivorship Bias

Financial Information

Collinearity



Continuous Target

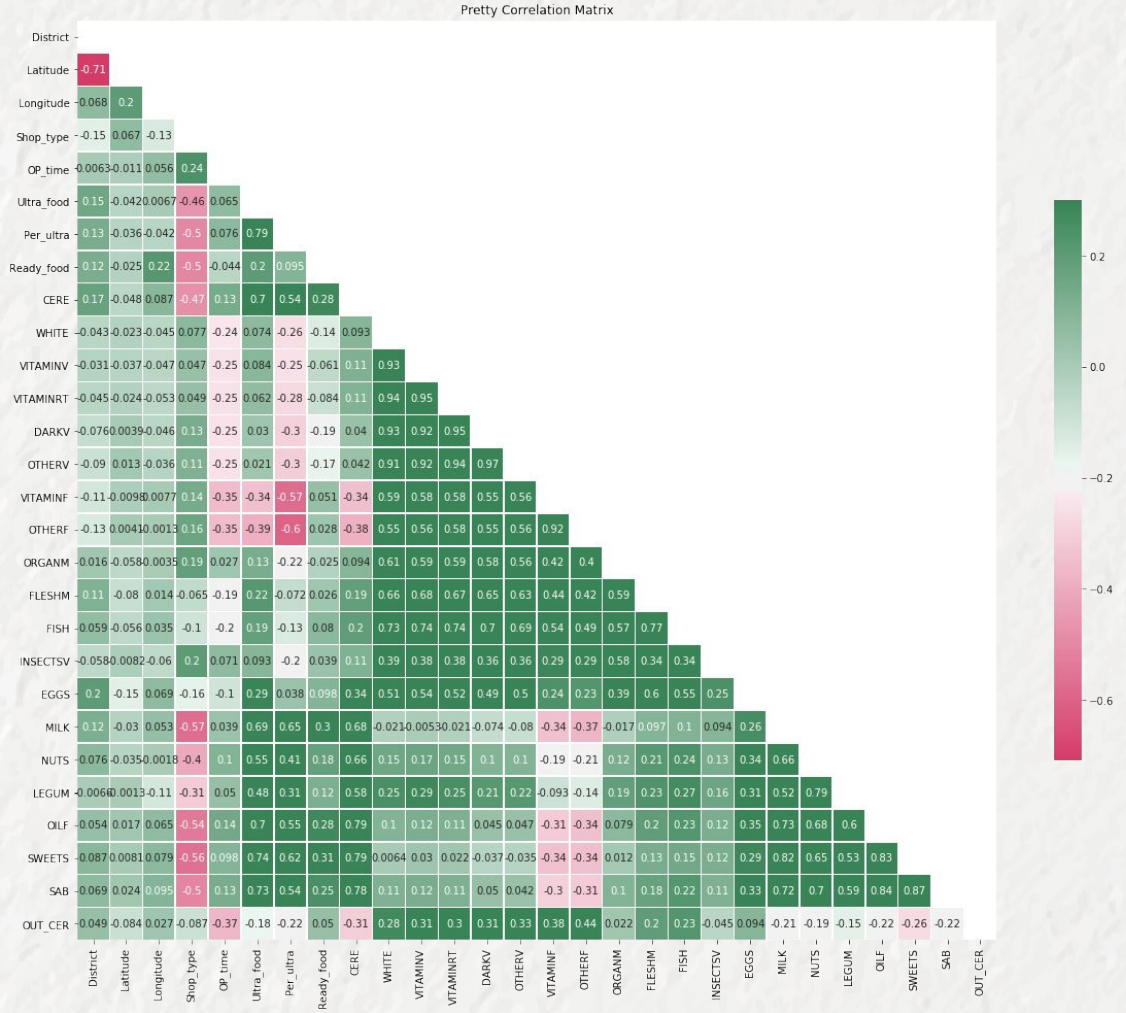
Data Models **Findings**

image: Flaticon.com

THANK YOU!

Appendix

Complete Correlation Matrix



Reference to raw data

Shop type

1 = Hypermarket

2 = Supermarket

3 = Convenience store/minimart

4 = Mom-and-pop store

5 = Specialty food shop

6 = Wet market

7 = Street market

8 = Street food stall

Operating time

1 = <1 year

2 = 1-2 years

3 = 3-5 years

4 = >5 years

District

1 = Ba Dinh

2 = Dong Da

The only slide that matters