Mark Kaplan, Qianwen Liu, Avi Skidelsky, Nosson Weissman

DAV 6150 - Data Science

Professor James Topor

Summer 2022

DAV 6150 Final Project Proposal 1st Draft

Introduction

As of recently, Americans spend close to 9% of their disposable income on food[1]. Additionally, according to data from the Bureau of Labor Statistics, the US has been facing extreme inflation on food prices in 2022[2]. With all the inflation, food pricing has been on our minds recently. With this in mind, we managed to procure food pricing data from kosher supermarkets in NYC which we will focus on in this research project.

Research Questions

In completing this project, we hope to gain a better understanding of food pricing and share that with the reader. We want to find the driving factors of item pricing and understand our categorical variables and which items are mostly likely to be on sale. We also aim to build a model which can predict item-price using four or five explanatory variables.

Data To Be Used

Supermarket data being used was scraped from websites of three different supermarkets.

Each observation contains:

- store - categorical nominal
- item-id - identifier[3]
- name –
- brand – categorical
- price – numerical continuous
- weight – composed of numerical and potential categorical data
- sale-price – numerical continuous
- category - categorical nominal
- main-category - categorical nominal
- date – date (does not necessarily fit the typical typology)

[1] USDA ERS - Food Prices and Spending

[2] Food prices up 10.8 percent for year ended April 2022; largest 12-month increase since November 1980 : The Economics Daily: U.S. Bureau of Labor Statistics (bls.gov)

[3] Item-id may only really be used as an id when considered with the actual store, so it is essentially one part of a composite id

- sub-category - categorical nominal

Below is an example of what some of our observations may look like:

store	item-id	name	brand	price	weight	sale-price	regular-price	category	date	main-category	sub-category
moishas	10724857	Korns Sesame Rings 2Pk	Korn's Bakery	2.49	Â | 8 Oz			Bread & Bakery;Bread & Challah	7/5/2022	Bread & Bakery	Bread & Challah
pomegranate	967448	Lender Bagels Plain 6 Pk.	Lender's	2.99	Â | 12 Oz			Bread & Bakery;Bread & Challah	7/5/2022	Bread & Bakery	Bread & Challah
moishas	1729317	Toufayan Bagels, Cinn Raisin	Toufayan Bakeries	2.79	Â | 20 Oz			Bread & Bakery;Bread & Challah	7/5/2022	Bread & Bakery	Bread & Challah
glattmart	1725992	Toufayan Bagels, Classic	Toufayan Bakeries	2.79	Â | 20 Oz			Bread & Bakery;Bread & Challah	7/5/2022	Bread & Bakery	Bread & Challah
glattmart	1725993	Toufayan Bagels, Everything	Toufayan Bakeries	2.79	Â | 20 Oz			Bread & Bakery;Bread & Challah	7/5/2022	Bread & Bakery	Bread & Challah

Project Outline

This project can be broken down into smaller parts:

1. Project Proposal
2. Data Gathering
3. Data Cleansing / Preparation
4. Data Analysis
5. Feature reduction – if necessary
6. ML
    a. Three ML models
    b. Ensemble model
7. Conclusion
8. Video Presentation

Approach

We explained where the data is coming from earlier.
While some data cleansing will likely be necessary, if we end up with too many categories in some of our fields, we may also need to aggregate some of our categorical data.

In our data analysis we will want to see how prices varies by brand category, store etc. We will want to know which stores offer the most competitive prices for each category and we may want to consider pricing variance by category.
As we are working with a dataset that has few explanatory variables, feature reduction may not be necessary, but that is to be determined. We may use some feature reduction methods just to double check.
As far as ML algorithms, with this amount of categorical data, we will likely want to consider random forest; KNN and logistic
Below is an overview of what each one of our team members will be contributing:

- Project Proposal – Nosson and team
- Data Gathering – Nosson
- Data Cleansing / Preparation – Nosson, Avi and team
- Data Analysis – Avi, Qianwen and team
- Feature Reduction – Qianwen, Avi and team
- ML – Mark, Avi and team
- Conclusion - Mark, Qianwen, Avi, Nosson
- Video Presentation – Mark, Qianwen, Avi, Nosson