Stocks on the Menu:
Analyzing the Economic Impact of the Processed Food Industry

Alex Gao, Edmand Yu, Joel Kim, Zefeng Xu
August 4 2024

Contents

Abstract

This study explores the potential of incorporating non-financial data into ML models to enhance prediction of stock/ETF prices within the processed food industry. We analyzed the relationship between processed meat consumption, economic indicators (e.g., household income, unemployment rate, macroeconomic indicators), and stock market performance. Utilizing both financial and non-financial data, we developed models to forecast monthly stock price changes for companies within the processed food industry over the past decade. We identified strong linear relationships between processed meat consumption and income inequality (Gini index), as well as moderate correlations with major ETFs, highlighting the complex economic dynamics within the industry. The high performance of our models and the significant relationship discovered between the variables suggest that a more in-depth analysis of the connections could further enhance model accuracy. However, the study faced limitations due to data quality and availability, requiring estimations for processed meat consumption. Despite these challenges, the research provides valuable insights and sets the groundwork for future studies on the economic influences of the processed food sector.

Keywords: Machine Learning, stock prices, ETF, prediction

1 Introduction

With the emergence of advanced technologies, especially AI and ML techniques, have warranted a great opportunity to utilize available data to analyze various problems in the financial sector. One key component of the financial sector is the stock market. The stocks of different companies are being traded almost every day around the world at stock exchanges. Given the fact that one of the major reasons to trade stocks is to make money, it is thus exceedingly beneficial if traders are able to forecast the behaviors of the stock markets to take reasonable actions. Numerous methods and models have been proposed to predict the behavior of stock markets, especially the stock prices, using available data and machine learning techniques [1, 2, 3]. However, the existing literature still lacks research into stock prices prediction specifically targeted for the processed food industry [4], in which a series of restaurants, as well as some manufacturing and transporting companies are involved. In this study, we utilized not only the normal stock data (e.g., past stock prices, traded volumes, etc.), but also data of the food processing industry, as well as data from other indirectly related sources such as household income, unemployment rates, and macroeconomic indicators, to study and try to predict the stock prices of selected companies in the processed food industry for the past 10 years.

2 Non-Technical Summary

The major questions of this study were how to integrate various data (i.e., not only financial data) into ML models to predict stock/ETF prices and how these non-financial variables, or indicators, relate to the stock market. We have built two models, one for stocks and the other for ETFs, to study this major question.

The stock-price-predicting model was designed to predict whether a given stock's price will increase, stay the same, or decrease in a month (i.e., classification with 3 classes) given various factors. The ETF-price-predicting model was designed to predict the net change in ETF price in a month.

2.1 Major Question

2.1.1 Stock Price Prediction

Table 1 summarizes the best classification accuracy (i.e., after tuning of parameters) of the stock-price-predicting models employed for selected stocks, as well as the baseline accuracy computed using the target labels.

Table 1. Classification accuracy of various models

Company Abbv.	Baseline Accuracy	LR Accuracy	RF Accuracy	KNN Accuracy	SVM Accuracy	MLP Accuracy
MCD	0.5084	0.7083	0.9167	0.6667	0.9167	0.8333
HSY	0.5139	0.8333	0.9583	0.7917	0.7083	0.8333
COKE	0.5015	0.8750	0.8333	0.7500	0.7500	0.7917
CAG	0.4943	0.6667	0.9167	0.6667	0.6667	0.6667
SAP	0.5062	0.8333	1.0000	0.7083	0.8750	0.8333

* LR = Logistic Regression; RF = Random Forest; KNN = K Nearest Neighbor; SVM = Support Vector Machine; MLP = Multilayer Perceptron

As can be seen from the accuracies, Random Forest seems to be always working better than or at least as well as other models in our study. However, since the baseline accuracy was all close to 0.5 and all models perform better than that, it suggests that the models were able to learn meaningful patterns from the data and can successfully predict the change in stock prices to some extent.

2.1.2 ETF Price Prediction

Figure 1 shows the predicted ETF prices versus the actual ones. It can be observed that most of the points fall along the line with small deviations, which indicates that the although far from perfect, the model was able to learn from the features and make predictions of the ETF future prices. Table 2 summarizes the performance of Fast-KNN using different combinations of hyperparameters. The Information Coefficient (IC) were around 0.6 for all models, which indicates that the predicted values and the actual values have reasonably strong linear relationships. Thus, the Fast-KNN model was indeed able to learn from the features and make future price predictions.

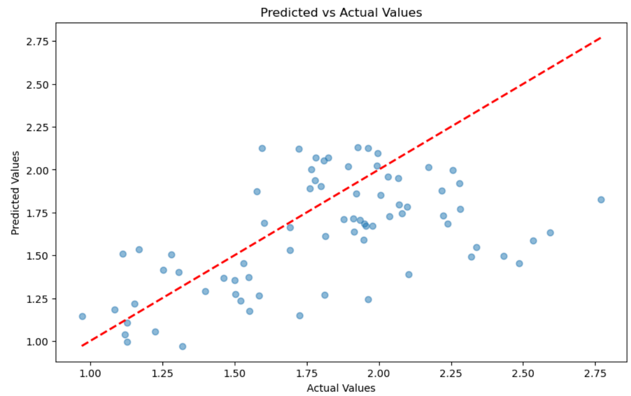

Figure 1. Predicted versus actual ETF prices

Table 2. Performance of different hyperparameters of KAN

Layers	Activation	LR	Grid_Size	Best_IC
5	GELU	1e-3	6	0.6356
5	RELU	1e-3	6	0.6102
5	tanh	1e-3	8	0.6211
4	GELU	1e-3	6	0.5823
6	GELU	1e-3	6	0.6032

* IC = Information Coefficient between actual values and predicted values.

2.2 Sub-questions

As a byproduct of the major problem, we have also studied the correlation between some selected variables. These include processed meat consumption, income inequality (measured using Gini index), buying power (measured using Consumer Price Index), broader economic indicators and unemployment rates. Four sub-questions were then asked between them:

(1) What is the relationship between processed meat consumption and income inequality?
(2) What is the relationship between processed meat consumption and buying power?
(3) What is the relationship between processed meat consumption and unemployment rate?
(4) What is the relationship between processed meat production and consumption and broader economic indicators like ETFs?

Regarding these sub-questions, we are giving the answers to each of them below.

(1) Processed meat consumption (in million pounds) and Gini index (representing income inequality) were computed/estimated for each month between 2014 and 2022, which is the year range of interest

in this study. Since both household and family income were provided, two versions of the Gini index (one for households, the other for family) were calculated. The Pearson Correlation Coefficient was computed between each Gini index version and the meat consumption statistics. The results showed that both correlations were very high and with significant p-values: 0.989 and 0.991 between processed meat consumption and household Gini index and family Gini index respectively. These high coefficients would mean that there is a strong linear relationship between processed meat consumption and both Gini indices. Such relationship was confirmed by plotting the variables in a scatter plot, as shown in Figure 2.

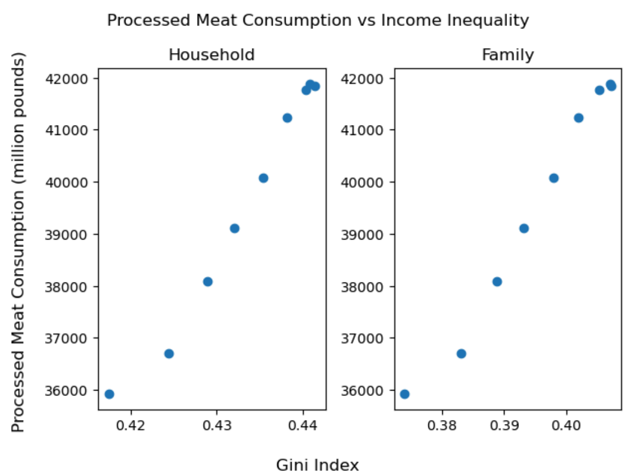

Figure 2. Processed meat consumption versus two versions of Gini index.

(2) Buying power was quantified by Consumer Price Index (CPI), which represents average change over time in the prices paid by urban consumers for a market basket of consumer goods and services. In this study, since the consumer goods are processed meat, we used the CPI specifically for meat and poultry. Both processed meat consumption and CPI data were obtained for each month between 2014 and 2022 inclusive. The Pearson Correlation Coefficient showed that there is no significant correlation between these two variables (coefficient=0.535, p-value=0.138). A scatter plot of these two variables confirmed this conclusion, as shown in Figure 3. Note that the baseline CPI (i.e., CPI=100) was in September 1984.

(3) Similarly, unemployment rate data was acquired for each month between 2014 and 2022 inclusive. Pearson Correlation Coefficient between unemployment rate and processed meat consumption showed

that there is no significant correlation between them: (coefficient=-0.107, p-value=0.784). A scatter plot of these variables confirmed this conclusion, as shown in Figure 4.

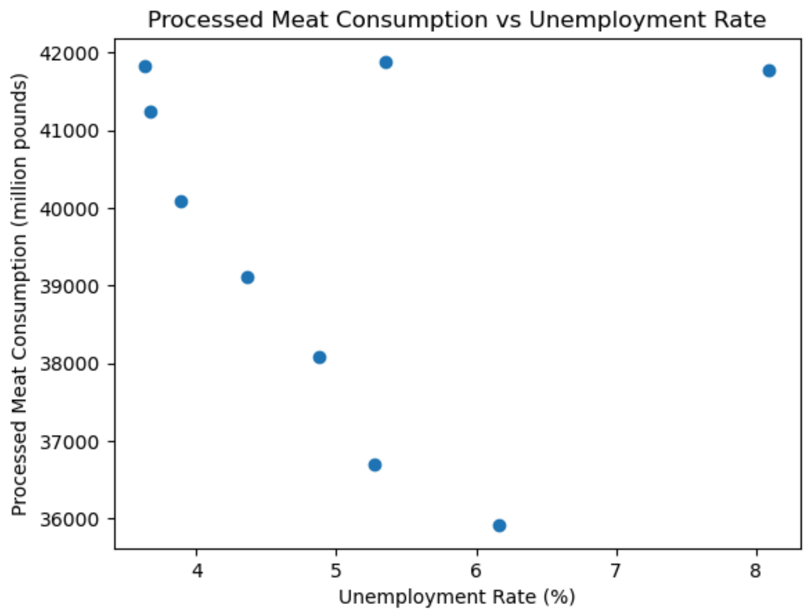

Figure 3. Processed meat consumption versus Consumer Price Index for meat and poultry.

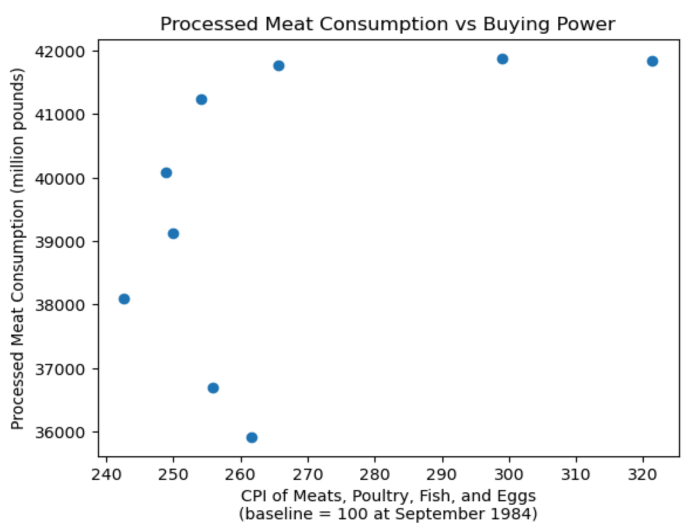

Figure 4. Processed meat consumption versus unemployment rate.

(4) The relationship between processed meat consumption and broader economic indicators like ETFs appeared to be complex and moderately strong, based on our analysis.

- Correlation with major ETFs:

From the correlation heatmap (Figure 5), we can see that processed meat consumption has a moderate positive correlation with major ETFs:

  - 0.72 correlation with DIA_mean (SPDR Dow Jones Industrial Average ETF)
  - 0.72 correlation with SPY_mean (SPDR S&P 500 ETF Trust)
  - 0.52 correlation with VOO_mean (Vanguard S&P 500 ETF)

This suggests that as processed meat consumption increases, there's a tendency for these major market indices to also increase. In addition, the correlation being stronger with DIA and SPY compared to VOO may indicate that processed meat consumption might have a stronger relationship with some market segments than others.

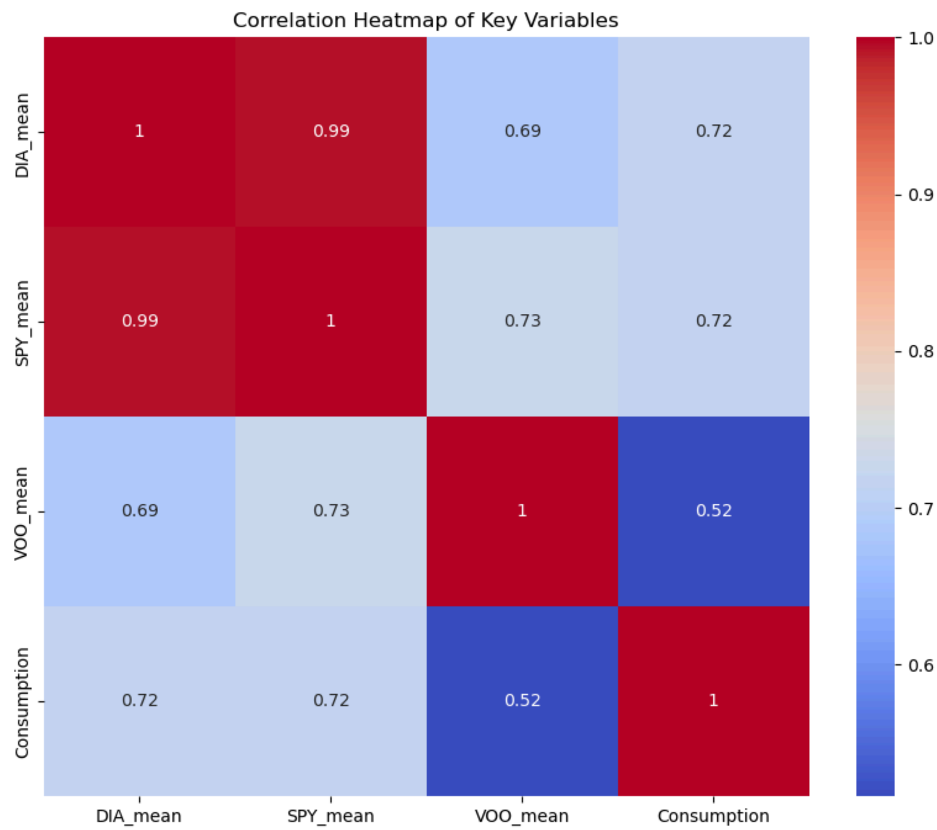

Figure 5. Correlation heatmap of key variables

Apart from the major question and the sub-questions, we have also tested for our expectation that the price trends of Exchange-Traded Funds (ETFs) will follow more or less exactly as the trend of the unemployment rate over time. We experimented with four major ETFs: SPDR Dow Jones Industrial Average ETF Trust, Fidelity Nasdaq Composite Index ETF, SPDR S&P 500 ETF Trust, and Vanguard

500 Index Fund. We extracted the price data for these four ETFs, as well as the unemployment rate data, from January 1, 2014, until now. Figure 6 shows the general trend of the unemployment rate and the price trend for the four ETFs on the same axis for comparison. As can be seen from the figure, when the unemployment rate was increasing significantly around year 2020 to 2021, the prices of the four ETFs were all decreasing. Also, when the unemployment rate increases, the prices decrease and when the unemployment rate decreases, the prices increase. This confirms with our expectation that the prices of ETFs are generally negatively related to unemployment rate. Note especially that the time period between 2020 and 2021 was exactly when the COVID-19 pandemic situation around the world was the most severe.

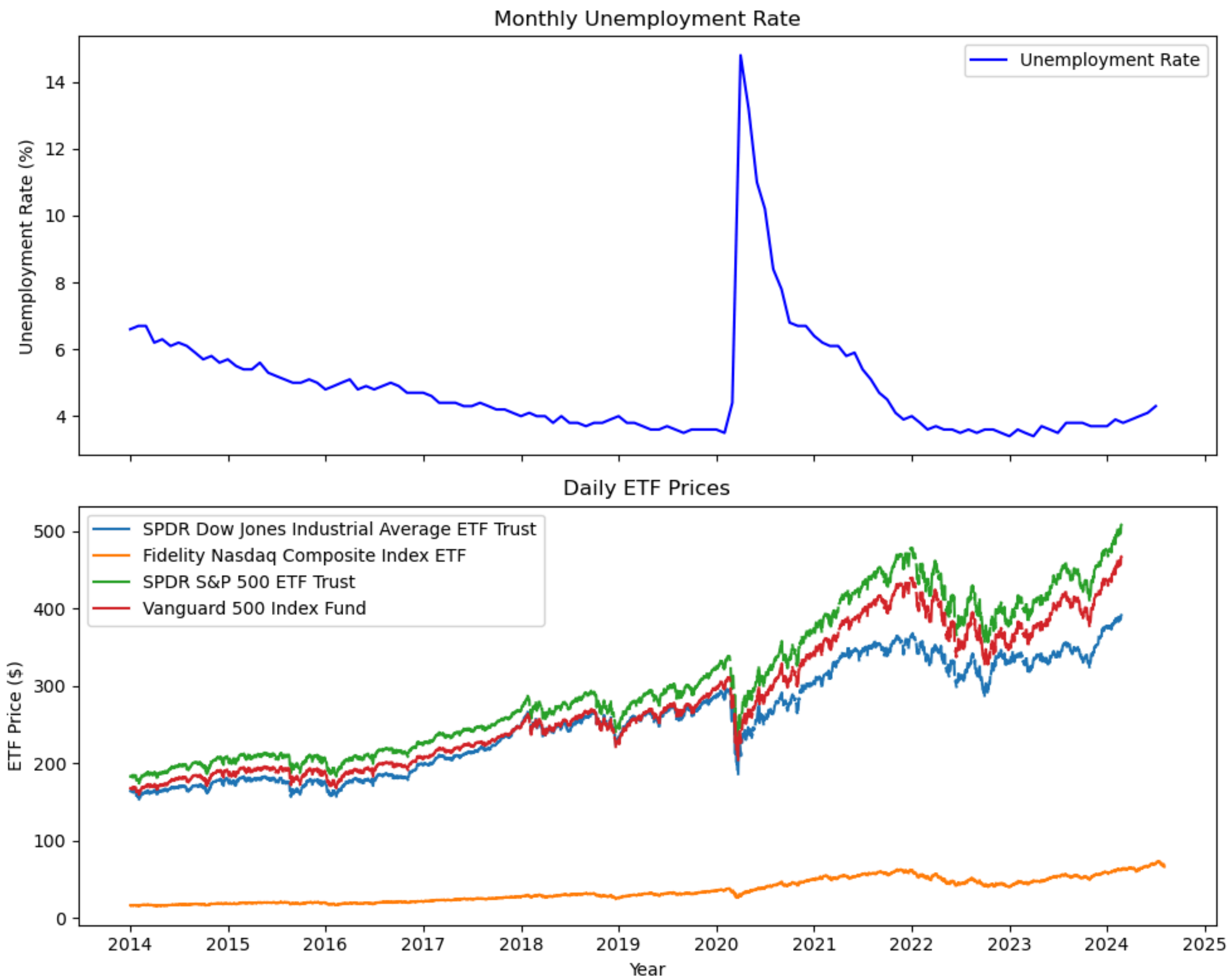

Figure 6. Trend of the unemployment rate and four major ETFs from 2014 to now

3 Technical Summary

In this technical summary, we will first discuss in depth the datasets used in this study and our approaches to clean and preprocess them to better fit our objectives. In addition, any problems we encountered during this preprocessing process will be outlined and explanations of our strategies to solve them will be provided. After that, we will discuss the models we chose to implement our primary objective to utilize machine learning techniques to predict monthly stock/ETF price changes. Lastly, in the special considerations section, we will discuss any assumptions we made in this study and provide possible justifications.

3.1 Datasets

To generate the influencing factors (i.e., independent variables) of our ML models, we utilized both the provided datasets and some public datasets from external sources.

For the provided datasets, we utilized the following parts/sections.

1. Household and family income sections of the "American Community Survey - Selected Economic Characteristics" dataset
2. The entire "Meat Production" dataset
3. The entire "Cold Storage" dataset
4. Part of the "Nutrition Physical Activity and Obesity Data" dataset
5. The entire "Stocks and ETFs" dataset
6. The entire "Commodities" dataset

For external datasets, we utilized the following:

1. A Consumer Price Index (CPI) dataset
    i. We obtained this dataset from the U.S. Bureau of Labor Statistics' official website. We specifically chose to download the dataset of CPI for Food and Beverages in the U.S. because our problem setting is related to food.
2. An "Unemployment Rate" dataset
    i. We obtained this dataset from Federal Reserve Economic Data official website.
3. A "High school graduates" dataset
    i. We obtained this dataset from the National Center for Education Statistics official website. This dataset contains the number of high school graduates by state from 1980-81 through 2031-32.
4. An "Adult population by state and year" dataset
    i. We obtained this dataset from the KFF - The independent source for health policy research, polling, and news official website. Note that this source does not provide the data for year 2020 and the year 2020 data was obtained from the "American Community Survey - ACS Demographic and Housing Estimates" dataset at United States Census Bureau's official website.

3.2 Preprocessing

In this section, we will discuss in detail how we preprocessed each dataset. All the preprocessing done were not only towards the purpose of simply cleaning the datasets, but also towards the computation of key variables of interest in our models. In general, we were interested in predicting the change in stock prices over time using processed food consumption, multiple economic indicators, health conditions of people, price of commodities, and past stock data. The economic indicators were determined to include Consumer Price Index (CPI), unemployment rate, and income inequality measures.

3.2.1 "American Community Survey - Selected Economic Characteristics"

The objective of using this dataset would be to generate income inequality measures. Thus, we first selected the data with category "INCOME AND BENEFITS." The returned output showed that income data for both households and families are available from 2010 to 2022 inclusive. To measure income inequality, we chose to compute the Gini Index for each year. Thus, we aggregated the counts across different U.S. states to be the counts of households/families at a specific income level.

Given a list of income levels called X and a list of number of households/families in each income level called P, Gini Index can be calculated using the following algorithm.

- Suppose there are n entries in both lists.
- Sort income levels X and corresponding number of households/families P by income levels X in ascending order.
- Calculate Total Income $TI = \sum_{i=1}^{n} X_i * P_i$
- Calculate Cumulative Income Shares
    - The Cumulative Income Share up to income level i is S_i
    - $S_i = \frac{\sum_{j=1}^{i} X_j * P_j}{TI}$

      where $\sum_{j=1}^{i} X_j * P_j$ is cumulative income up to income level i
- *Gini Index* $G = 1 - \sum_{i=1}^{n} (S_{i-1} + S_i) * P_i$

3.2.2 "Meat Production" & "Cold Storage"

The objective of using these two datasets was to obtain the processed food consumption values. We estimated the consumption of processed meat using the following formula, leveraging available data on production and cold storage of meat products:

$$\begin{aligned} Consumption &= (Percentage\ of\ Processed\ Meat) * \\ &\quad (1 - Food\ Waste\ Percentage) * \end{aligned}$$

$$(Production - \Delta Storage)$$

The term (1 - Food Waste Percentage) * (Production - ΔStorage) represents the amount of food actually consumed. Given that 30-40% of food is typically wasted [5], we estimated the food waste percentage at 35%.

For the percentage of processed meat, a study from 2019 indicated that, on average, Americans consume 284 grams per week of unprocessed meat and 187 grams per week of processed meat [6]. This means processed meat makes up approximately 187/471 of total meat consumption. Additionally, for the sake of our study, we assumed this ratio remains constant over time, as the same study revealed that there have been no changes in the amount of processed meat consumed by US adults over the last 18 years [6].

This information initially led us to define consumption as:

$$Consumption = \frac{187}{471} * (1 - 0.35) * (Production - Storage)$$

However, this approach led to unrealistic negative consumption values on certain dates where storage exceeded production. The problem arose because stored meat from previous dates could be carried over into the present date. To address this, we refined our formula to account for changes in storage rather than absolute storage values:

$$Consumption = \frac{187}{471} * (1 - 0.35) * (Production - \Delta Storage)$$

This refined formula is more accurate as it accounts for changes in stored meat products. If there is no change in storage, consumption equals production (adjusted for waste). A positive change in storage indicates that some of the current production is being stored rather than consumed, while a negative change signifies that stored meat has been consumed.

Therefore, our final formula for estimating the consumption of processed meat is:

$$Consumption = \frac{2431}{9420} * (Production - \Delta Storage)$$

Note that the unit for consumption is millions of pounds and we ignored the “Frozen Eggs” category in the storage data because it is not present in the production data.

3.2.3 “Nutrition Physical Activity and Obesity Data”

This dataset was used to retrieve data about people’s health conditions, especially obesity status and whether people engage in physical activity. Thus, we have picked a total of five different questions to consider as variables in our predictive models. The five questions are: “Percent of students in grades 9-12 who have obesity,” “Percent of students in grades 9-12 who participate in daily physical education,” “Percent of students in grades 9-12 who drank regular soda/pop at least one time per day,” “Percent of adults who engage in no leisure-time physical activity,” and “Percent of adults aged 18 years and older who have obesity.”

To transform these percentages into actual numbers, which would be more informative, we used the "High school graduates" and the "Adult population by state and year" datasets. The process for the mapping of the two adult-related questions was simple: finding the corresponding number of adults in the state for the year suffices. However, since students in grades 9-12 does not only include one type of people, a different approach was needed.

Since U.S. high schools students typically graduate in 4 years (after their 12th grade), we decided to use the following strategy to estimate the number of 9-12 grade students in each year in each state:
If the percentage is for year n

- Grade 9 students will graduate in 4 years, so find the number of graduates in year n+4 to represent the number of students in Grade 9 in year n.
- Grade 10 students will graduate in 3 years, so find the number of graduates in year n+3 to represent the number of students in Grade 10 in year n.
- Grade 11 students will graduate in 2 years, so find the number of graduates in year n+2 to represent the number of students in Grade 11 in year n.
- Grade 12 students will graduate in 1 year, so find the number of graduates in year n+1 to represent the number of students in Grade 12 in year n.
- Sum number of students in each grade to get the total number of high school students in year n.

Since in this study, we did not compare anything between states, after the above mappings, we took the mean of all counts across all U.S. states in a year to get a national estimate in real numbers (i.e., not percentages) for each of the five questions for that year. Moreover, since our primary problem was to predict the change of stock/ETF prices for each month, we need to divide this national estimate into 12 months for each year. We employed a "random number division" algorithm, in which n random numbers are generated such that their sum equals to a specified target. In our case, we utilized this algorithm to divide each yearly national estimate into 12 different random values, each representing a month.

3.2.4 "Stocks and ETFs"

For the stock-price-predicting model, we were predicting the monthly changes in stock prices. This target label was created by first finding for each (year, month) combination, the earliest and latest dates on which the markets were open and stock prices were recorded. Then the change in stock price for this month was calculated as the latest price at market closure minus the earliest price at market closure. A label was assigned for each month: if the change is positive then label was 1; if the change is zero then label was 0; if the change is negative then label was -1.

For the ETF-price-predicting model, the mean "Open," "Close," "High," "Low" prices and traded "Volume" were calculated for each ETF monthly. The means were directly set to be the target variables in regression. The date range of this model was determined to be from November 1999 to February 2024 instead of the past 10 years in the stock-price-predicting model.

3.2.5 “Commodities”

For both models, we were using prices of all commodities as features in our ML models.

3.3 Modeling

3.3.1 Stock Price Prediction

For this model, integrating economic indicators, health question data, commodity prices, previous stock prices, and traded volume into one dataset resulted in 16 features being used to predict the change in stock prices over months for selected companies. The split of training and testing data followed the strategy of using previous data as training data to predict later data. In this case, the last two years' data were set as testing and the rest were used as training data. After that, z-score normalization was performed to make each feature follow a standard normal distribution.

We chose to predict the stock price changes for a selected subset of companies including McDonald's Corporation, Hershey Co, Coca-Cola Consolidated Inc., ConAgra Foods Inc, and SAP SE ADR because they are representative of each sector in the food-processing industry. Machine learning models including Logistic Regression, Random Forest, K Nearest Neighbor, Support Vector Machine, and Multilayer Perceptrons were deployed for each company to predict the change in stock prices over months. A baseline accuracy calculated as the sum of squared class fractions was also calculated to serve as a comparison. Machine learning models are only considered meaningful if they perform better than the baseline model. Otherwise, the machine learning models are not learning valuable insights from the given data.

3.3.2 ETF Price Prediction

For this model, we integrated the meat statistics (production and cold storage), commodity price data. After splitting the data into training and validation sets in a temporal fashion (training: before Jan 2022; validation: from Jan 2022 onwards), we normalized all numerical columns using z-score normalization. To predict the monthly mean ETF prices, we implemented Fast-KAN [7], a modified version of Kolmogorov-Arnold Networks (KANs) [8], which is a recently proposed novel neural network architecture inspired by the Kolmogorov-Arnold representation theorem, that incorporates several optimizations and modifications to enhance performance and training stability. Our implementation relied on PyTorch and the model was trained with learning rate 0.001, batch size of 10 and 100 epochs.

KANs differ fundamentally from traditional Multi-Layer Perceptrons (MLPs) in their approach to non-linearity:

1. Learnable Activation Functions: Instead of fixed activation functions on nodes, KANs employ learnable activation functions on edges.
2. Absence of Linear Weights: KANs replace every weight parameter with a univariate function, eliminating the need for linear weight matrices.
3. Simple Node Operations: Nodes in KANs perform simple summation of incoming signals without applying non-linearities.

Our Fast-KAN implementation includes the following key features and modifications:
1. Gaussian Approximation of Splines: Rather than using B-splines directly, we approximate the spline functions using Gaussian basis functions. This approach potentially simplifies optimization and improves computational efficiency.
2. Layer Normalization: We incorporate standard layer normalization between KAN layers. This helps to stabilize the learning process by normalizing the inputs to each layer, potentially addressing internal covariate shift.
3. GELU Activation: While traditional KANs don't use fixed activation functions, we introduce Gaussian Error Linear Unit (GELU) activations in our model. GELU has shown promise in extracting features more effectively than other common activations like ReLU or tanh.

Our KAN models were structured as follows:
1. Input Layer
2. Series of KAN layers, each followed by layer normalization
3. GELU activation applied after each normalized KAN layer
4. Output Layer

Our training process was as follows:
The model was trained using backpropagation, leveraging the differentiable nature of the Gaussian basis functions and GELU activations. The learnable parameters include the coefficients of the Gaussian basis functions for each edge and the parameters of the layer normalization layers.

4 Conclusions and Future Directions

This study presents a comprehensive analysis of the economic impact of the processed food industry, focusing on the integration of various data sources to predict stock prices using machine learning (ML) techniques. By incorporating both financial and non-financial variables, such as household income, unemployment rate, and macroeconomic indicators, we aimed to develop models that predict the stock prices of companies within the processed food industry. The findings highlight the potential and challenges of using diverse datasets and ML models in economic predictions, offering valuable insights and paving the way for future research.

4.1 Conclusion

The study explores the potential of using non-financial data, such as processed meat consumption, into ML models to improve stock/ETF price predictions within the processed food industry. Models, particularly those utilizing diverse data sources, showed high performance in predicting stock prices, with some algorithms outperforming others depending on the specific company and variables used. A strong linear relationship was observed between processed meat consumption and income inequality (Gini index), suggesting that as processed meat consumption increases, the income inequality increases. This relationship should be further explored, possibly by including other variables such as time in the analysis to gain a deeper understanding of the relationship. Such insights could allow for the simplification of the models by discarding one of the highly correlated variables. Furthermore, the study found a moderately strong relationship between processed meat consumption and broader economic indicators like ETFs, indicating a complex economic ecosystem with processed meat consumption being one of several influencing factors. Additionally, a negative correlation was identified between ETFs and the unemployment rate. Similarly, further analysis with a third variable could refine the models or adjust them to better account for these relationships, ultimately improving their accuracy.

4.2 Limitations

Despite the promising results, the study faced several limitations. The data quality and availability were limited, necessitating estimations for processed meat consumption due to insufficient data. Several assumptions, such as the constant ratio of processed to unprocessed meat consumption over time, may not hold true in all contexts and could affect the robustness of the results. Additionally, the models developed may not generalize well across different time periods or economic conditions due to the specific nature of the datasets and the period studied (2014-2022). Furthermore, our analysis did not account for additional variables such as changes in the US population, time. This omission could compromise the validity of the observed relationship between processed meat production/consumption and broader economic indicators like ETFs.

4.3 Future Directions

Future research should address these limitations by obtaining more precise and comprehensive datasets, particularly for processed meat consumption, to improve model accuracy and reliability. Implementing dynamic models that adapt to changing economic conditions and consumer behaviors over time would enhance predictive power and generalizability. Additionally, expanding the range of economic indicators and considering other sectors beyond processed food could provide a more holistic understanding of the economic impact and improve model performance. Finally, further research should focus on identifying causal relationships rather than mere correlations to better understand the underlying mechanisms driving the observed trends.

5 References

[1] Carson Kai-Sang Leung, Richard Kyle MacKinnon, and Yang Wang. 2014. A machine learning approach for stock price prediction. In Proceedings of the 18th International Database Engineering & Applications Symposium (IDEAS '14). Association for Computing Machinery, New York, NY, USA, 274–277. https://doi.org/10.1145/2628194.2628211

[2] Payal Soni et al 2022 J. Phys.: Conf. Ser. 2161 012065
DOI 10.1088/1742-6596/2161/1/012065

[3] Mehar Vijh, Deeksha Chandola, Vinay Anand Tikkiwal, Arun Kumar, Stock Closing Price Prediction using Machine Learning Techniques, Procedia Computer Science, Volume 167, 2020, Pages 599-606, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.03.326.

[4] Milosavljevic, Milos, Katarina Nedovic, and Zeljko Spasenic. "Predicting Stock Prices Using Deep Learning Algorithms: A Case of Food-Processing Industry." Athens Journal of Technology & Engineering: 29.

[5] USDA, “Food Waste FAQs”, *usda.gov*, https://www.usda.gov/foodwaste/faqs. Accessed 1 August 2024.

[6] Zeng L, Ruan M, Liu J et al. (2019) Trends in processed meat, unprocessed red meat, poultry, and fish consumption in the United States, 1999-2016. J Acad Nutr Diet 119, 1085.e12-1098.e12.

[7] Li, Ziyao. "Kolmogorov-Arnold Networks are Radial Basis Function Networks." arXiv preprint arXiv:2405.06721 (2024).

[8] Liu, Ziming, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y. Hou, and Max Tegmark. "KAN: Kolmogorov–Arnold Networks." arXiv preprint arXiv:2404.19756 (2024).