Hao Tran
Apr 2025

Food CPI Model

Overall Food Consumer Price Index and Predictions from Different Food Category Consumer Price Indexes using Linear Regression, Stepwise Regression, and XGBoost Regression

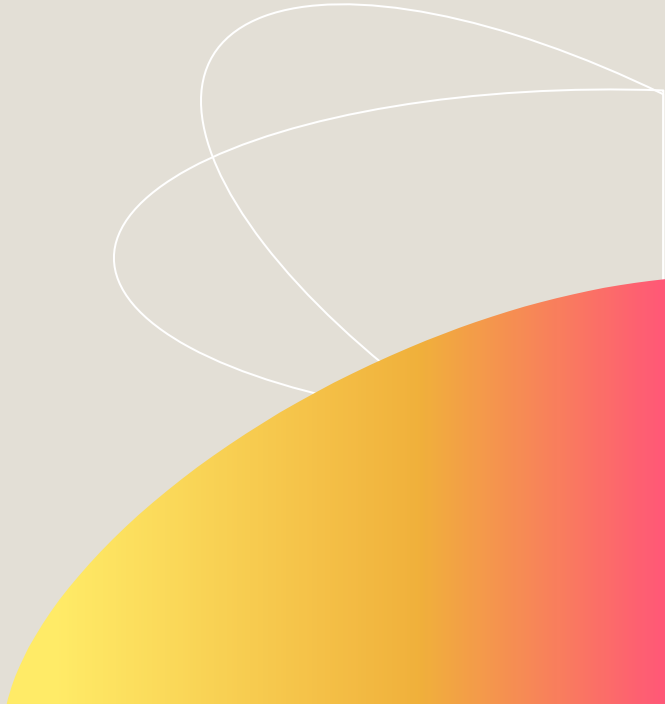
Project Goal

The objective of this project is to develop predictive models using **Linear Regression**, **Stepwise Regression**, and **XGBoost Regression** to estimate the contribution of various food category Consumer Price Indexes (CPI) to the overall Food CPI.

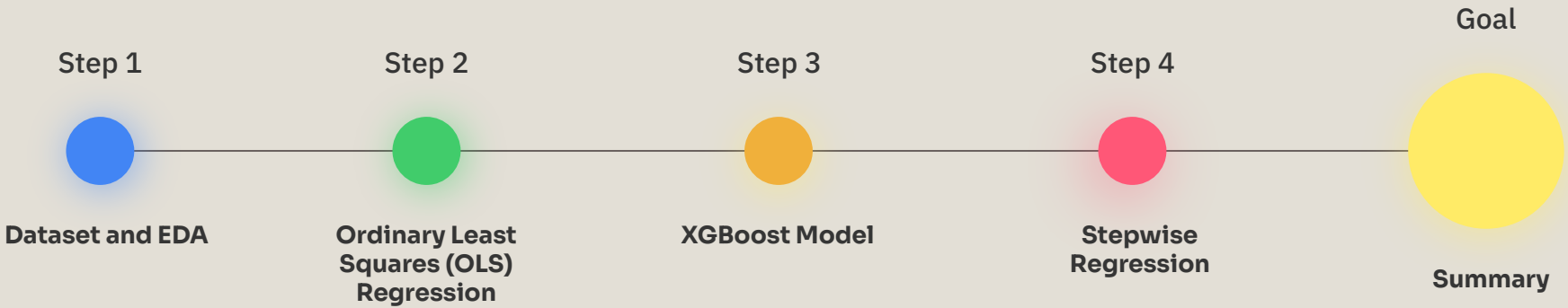
This type of model helps answer key questions such as:

- *Which food categories most influence the overall Food CPI?*
- *How much does each category contribute to fluctuations in food prices?*
- *Can we forecast future Food CPI trends based on category-level changes?*

GitHub Repo:
https://github.com/donxiya/Machine-Learning-Food-CPI



Road Map



Year/Quarter/Month

Dataset

Dataset Description

The All-Items Consumer Price Index (CPI) measures the average change over time in the prices paid by urban consumers for a representative basket of goods and services. The CPI for Food specifically tracks the changes in retail food prices.

Dataset Metadata

The dataset consists of 3 features: (Year, Category, Percent Change), with a total of 1,100 rows. The data represents the annual CPI percent change for various food categories from 1974 to 2023.

Source

U.S. Department of Agriculture, Economic Research Service. (2024). Annual percent changes in selected Consumer Price Indexes, 1974–2023. Retrieved from https://www.ers.usda.gov/data-products/food-price-outlook

	ɔnsumer Price Index item"	Year	Percent change
293	Beef and veal	2016	-6.3
294	Beef and veal	2017	-1.2
295	Beef and veal	2018	1.4
296	Beef and veal	2019	1.6
297	Beef and veal	2020	9.6
298	Beef and veal	2021	9.3
299	Beef and veal	2022	5.3
300	Beef and veal	2023	3.6
301	Pork	1974	-0.5
302	Pork	1975	22.4
303	Pork	1976	1.3

Cleaning Data

	Potential challenge	
1	Category of each food cpi is more meaningful when introduced as columns	Dataset pivot, and introduce multiple new variables.
2	Out of scope categories	I excluded the "Food at Home" and "Food Away from Home" categories from the dataset, as these categories were not relevant to the project’s goal.
3	Handling Outliers in CPI Data are retained	The CPI change percentages reflect significant economic events, such as inflation spikes or deflation periods.

Optimized OLS Model

OLS Regression Results

Dep. Variable:	All_food	R-squared:	0.956
Model:	OLS	Adj. R-squared:	0.952
Method:	Least Squares	F-statistic:	232.6
Date:	Mon, 28 Apr 2025	Prob (F-statistic):	1.60e-28
Time:	10:34:34	Log-Likelihood:	-35.603
No. Observations:	48	AIC:	81.21
Df Residuals:	43	BIC:	90.56
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.5448	0.129	4.239	0.000	0.286	0.804
Meats	0.1861	0.020	9.091	0.000	0.145	0.227
Fruits_and_vegetables	0.1437	0.032	4.505	0.000	0.079	0.208
Nonalcoholic_beverages	0.0609	0.011	5.527	0.000	0.039	0.083
Other_foods	0.4798	0.036	13.192	0.000	0.406	0.553

Omnibus:	0.063	Durbin-Watson:	1.582
Prob(Omnibus):	0.969	Jarque-Bera (JB):	0.025
Skew:	-0.016	Prob(JB):	0.987
Kurtosis:	2.892	Cond. No.	16.7

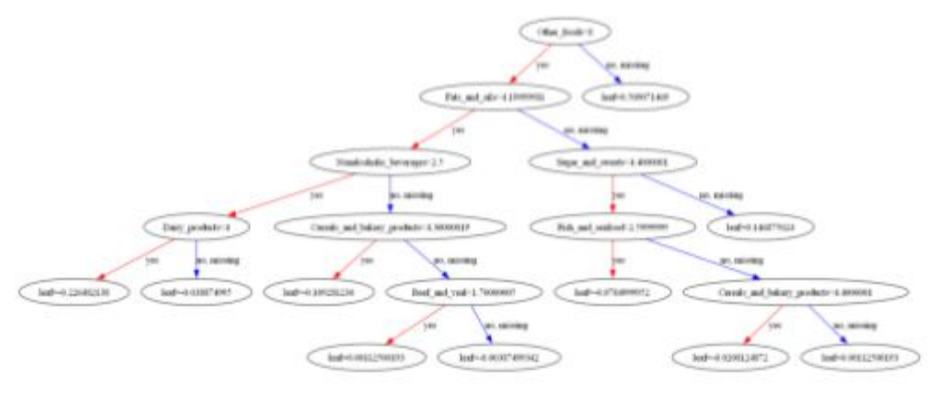
Steps

- Include interaction between common combination of food types
- Iteratively remove predictors with high p-values and confidence intervals covering zero.
- Refit the model after each removal to reassess which variables remain significant.
- Continue the process until all remaining predictors are statistically significant at the 95% confidence level.

Result

- R-squared: 0.956
- All factors with minimal P-value

Hyperparameter Tuned XGBoost Model



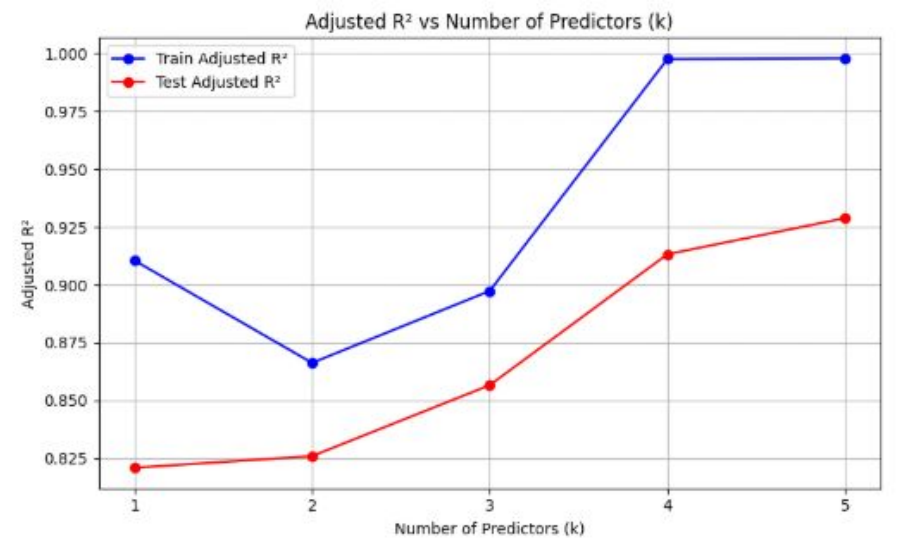
Steps

- XGBoost (Extreme Gradient Boosting) is a powerful machine learning algorithm based on the gradient boosting framework.
- To further enhance the performance of the XGBoost model, we can perform hyperparameter tuning.
- Weight, gain, and coverage.

Result

- Test MSE: 0.2021
- Test MAE: 0.3380
- Test RMSE: 0.2021
- Test R^2: 0.9628

Stepwise Regression



Steps

- Forward Selection, where we start with no variables and add them one by one based on statistical significance.
- Selected the top k = 5 features to balance model simplicity and predictive performance.

Result

- Adjusted R^2 = 0.9971

Conclusion

This project utilized OLS regression, XGBoost, and Stepwise regression to model the All_food CPI using food category data. The models identified significant predictors and achieved strong performance.

Through iterative modeling, we built a robust and interpretable model for forecasting food-related CPI changes. The final models are promising and can be enhanced with more data and advanced techniques.