An Eye for CPI : A Time Series Analysis & Modeling

Victor Thai, Stephanie Chavez, Michael Chen, Zed Siyed

Introduction

The Bureau of Labor Statistics holds a vast amount of past CPI data that presents advantageous and practical insight on the trends of the US economy. This project uses such CPI data collected from 2012-2021 to forecast future CPI's using models surrounding linear machine learning, autoregression, and multivariate time series forecasting. In this presentation, we will discuss the importance of CPI as a factor in distinguishing the trends of the US economy, critiques for the predictive models that we tested in order to forecast future CPI data, and our analysis on the most accurate model we generated.

What is CPI?

CPI, also known as consumer price index, is a measure of the average change of a basket of goods or services over a time period. It is most commonly used as a gauge for trends in inflation. In recent years, CPI has been fluctuating against typical norms, namely through the COVID pandemic. Through our modeling, we hope to gain a better insight on the trends of CPI, what factors may influence its behaviors, and to predict upcoming price indexes.

Dataset / EDA

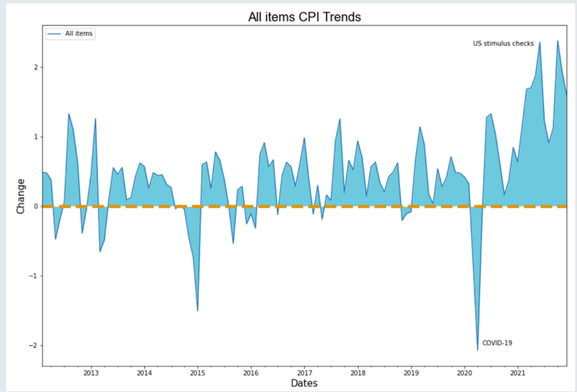
The dataset that we were given to work with was taken from the Bureau of Labor Statistics in a rather messy format. Our data was first retrieved in XLSX files with significant whitespace and labels out of order.

- total of 315 expenditure categories for the years of 2012-2020 with categories consisting of of food types, housing items, and even transportation expenditures.
- CPI not simply influenced by its past data, but rather through other gauges as well such as gold prices and oil prices.
- located datasets for US gold prices and US crude oil prices from the Federal Reserve of Economic Data (FRED) API.

Consumer Price Index for All Urban Consumers (CPI-U): U.S. city average, by expenditure category, December 2021
[1982-84=100, unless otherwise noted]

inconsistent spacing

Indent Level	Expenditure category	Relative importance Nov. 2021	Dec. 2020	an. 2021	Feb. 2021	Mar. 2021
0	All items	100.000	260.474	261.582	263.014	26
1	Food	13.990	270.023	270.938	271.363	27
2	Food at home	7.722	251.253	252.107	252.716	25
3	Cereals and bakery products	0.978	283.735	282.911	284.095	28
4	Cereals and cereal products	0.299	232.794	232.504	231.699	23
5	Flour and prepared flour mixes	0.041	232.911	239.720	238.561	23
5	Breakfast cereal[4]	0.139	227.887	226.376	223.913	22
5	Rice, pasta, cornmeal	0.118	244.594	243.333	244.664	24



Models

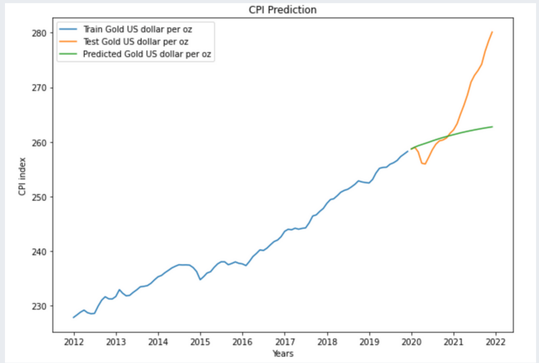
VARIMA Time Series -

The VARIMA model is a multivariate forecasting algorithm that is used when two or more time series data influence each other. In our model, we used the features of past CPI data, US Crude Oil prices, and US Gold prices.

- we must test for stationarity in order for modeling to work
- stationarity is when the mean and variance are not dependent on time (no clear trends over a time period).
- our data features all had trends of stationarity using the Dickey - Fuller test
- differenced the data to transform our data into a new format with no apparent trends
- among all our other models, the VARIMA model was the most optimal model as it offered a solution to our overfitting problem that we had with other models while also allowing us to use multiple parameters in our analysis.
- didn't have to sacrifice choosing one specific parameter column to use for CPI forecast

Process:

1. Make sure data is stationary
   a. Perform Dickey-Fuller Test
2. If data not stationary, perform a first-order differencing on data
3. Determine the value of p, d, q modeling parameters
4. Run model for all categories using same p, d, q parameters to determine lowest RMSE
5. Train and fit VARIMA model



p: the number of preceding lagged Y values between point and other Y values (autoregression)
d: the number of times our data is differenced
q: the number of preceding lagged values for the error term (moving average)

$$Y_{1,t} = \alpha_1 + \beta_{11,1}Y_{1,t-1} + \beta_{12,1}Y_{2,t-1} + \beta_{13,1}Y_{3,t-1} + \beta_{11,2}Y_{1,t-2} + \beta_{12,2}Y_{2,t-2} + \beta_{13,2}Y_{3,t-2} + \epsilon_{1,t}$$
$$Y_{2,t} = \alpha_2 + \beta_{21,1}Y_{1,t-1} + \beta_{22,1}Y_{2,t-1} + \beta_{23,1}Y_{3,t-1} + \beta_{21,2}Y_{1,t-2} + \beta_{22,2}Y_{2,t-2} + \beta_{23,2}Y_{3,t-2} + \epsilon_{2,t}$$
$$Y_{3,t} = \alpha_3 + \beta_{31,1}Y_{1,t-1} + \beta_{32,1}Y_{2,t-1} + \beta_{33,1}Y_{3,t-1} + \beta_{31,2}Y_{1,t-2} + \beta_{32,2}Y_{2,t-2} + \beta_{33,2}Y_{3,t-2} + \epsilon_{3,t}$$

Other Notable Models

SARIMA Time Series -

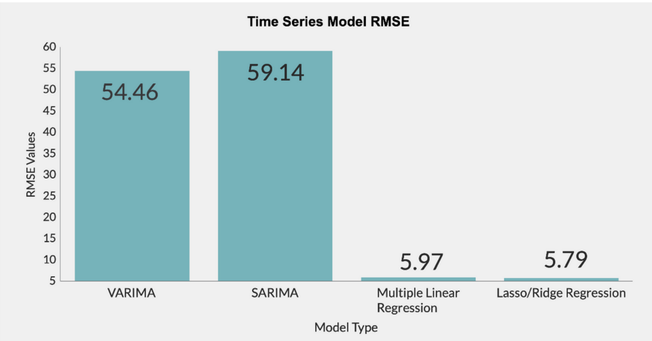
- similar to the VARIMA model that is auto regressive, handles seasonality, and a moving average.
- lacks the ability to intake more than one predictor variable.
- poor RMSE in comparison to VARIMA.

Multiple Linear Regression -

- similar to simple Linear Regression where a predictor variable, X, is used to predict an outcome variable, Y.
- The Multiple Linear Regression model took the major expenditure categories as predictor variables which generated overly accurate results yielding a RMSE of 5.97

Ridge / Lasso Regression -

- Ridge/Lasso Regression has the benefit of prioritizing the mean of the data points for a specific variable.
- given data has high collinearity, model selects the one best feature to be used for the whole model.
- Ridge uses L2 regularization, Lasso uses L1 regularization



So What?

CPI is a key indicator of how the economy is doing. Understanding the causes of CPI and inflation provide valuable information to consumers, and our government deciding fiscal and monetary policy. It also influences wages, retirement benefits, and tax brackets. Our model allows us to predict these changes and gain a greater understanding of what causes these changes. With so many Americans feeling the effects of inflation, predicting how long spikes will be and how severe they will be is crucial to understanding the United States' economic outlook.

Conclusion

This project provided exposure to a vast amount of forecast models, where we concluded that the VARIMA provided the most accurate and reliable predictions. However, there are some limitations that we did not handle. Given that we had more time, we would want to look more into the effects of inflation on CPI. Not only that, but since the majority of our testing data was centered around a pandemic, this influenced our predictions so we hope to further analyze data once the economy stabilizes from the current COVID-19 pandemic.

Acknowledgements

"Time-series Modeling for Consumer Price Index Forecasting using Comparison Analysis of AutoRegressive Integrated Moving Average and Artificial Neural Network". 2021. Accessed March 09, 2022.