David Brehm

DSC630

Course Project: Milestone 5

08/12/21

**Executive Summary**

This analysis aimed to predict national grocery store sales numbers based on a variety of economic indicators. The economic indicators used were from various US government entities such as the US Bureau of Economic Analysis or the US Census Bureau. Many of these indicators are compiled in one location by the Federal Reserve Bank of St. Louis, which is where the data used here was obtained from.

These indicators were fed into a polynomial regression model in an attempt to predict grocery store sales numbers. The model performed very well from the starting point of this analysis in 2003 through the months prior to the pandemic hitting the United States. After this however, the model performs much worse at forecasting the pandemic months. The fluctuations in our economic indicators might be too large to effective model, or we might need to incorporate additional indicators to try and improve the model.

**Introduction**

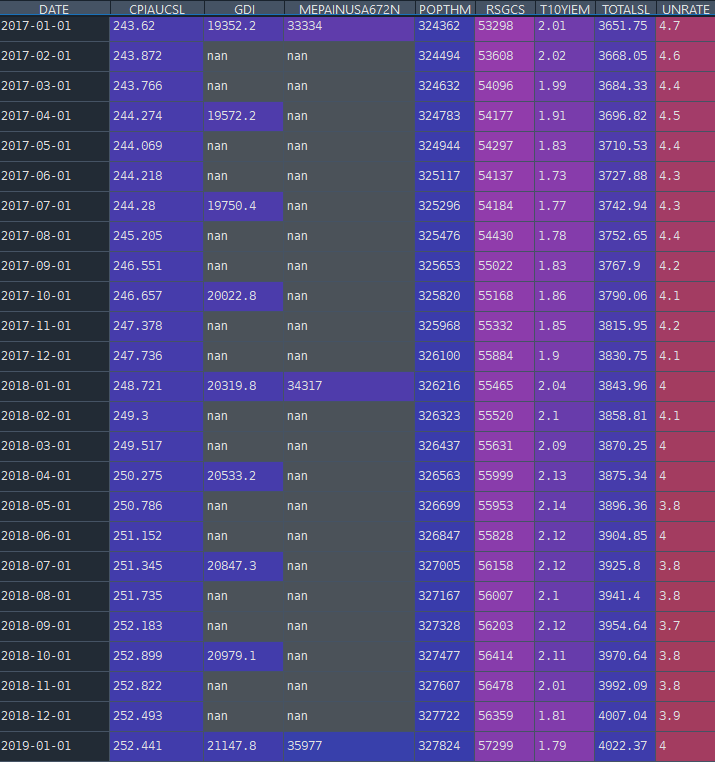
The recent pandemic has had an effect on many parts of our country, with the economy being one of the clearer ones. The shutdowns caused large spikes in many of our economic indicators that get tracked. One of these indicators that would have been very visible on a daily basis is grocery store sales. It was a prominent story at the beginning of the pandemic with how certain foods or materials were constantly out of stock. Some of this was likely due to production being shut down, but a lot of it was also sales increasing. With these fluctuations, I thought it would be an interesting problem to investigate which of the other main economic indicators track with the sales and if they could be used to predict future sales.

**Methods**

The data gathered was in the format of individual indicators. Attempting to keep these in the same date format made it simpler to combine. From the sources and structure of the data basically being just the date and one variable column, there wasn’t much cleaning required. With all of the date formats being YYYY-MM-DD, it was very easy to merge as well. After compiling all of the different data sources, features could be eliminated through the process of training the model through Recursive Feature Elimination. Once the model was trained, grocery store sales numbers could be predicted.

The model selected for this analysis was polynomial regression. This is a method of regression analysis where the relationship between the independent variables and dependent variable is modelled with an nth degree polynomial. The features are transformed to fit this selected polynomial degree. For example, if you have three features [a, b, c] and are attempting a polynomial with a degree of two, each row of data then becomes [1, a, b, c, a^2, b^2, c^2, ab, bc, ca].

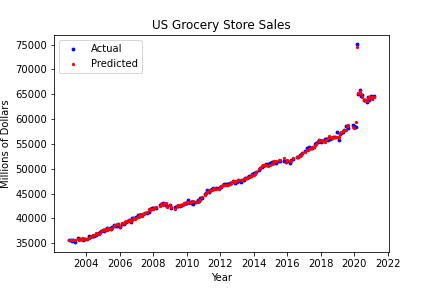
A concern for this model’s accuracy was the data size. Even if more features are added, the corresponding sales numbers which we were trying to predict were set. As a result of the grocery store sales data being monthly points from 1992, there was only around 350 data points. Through trying to add more features to the model, the data did not align well. There was monthly, quarterly, and annual data. This introduced missing values, not even accounting for the differences in seasonally adjusted vs non-seasonally adjusted data. Below is an example of how the data looks after its initial merge.



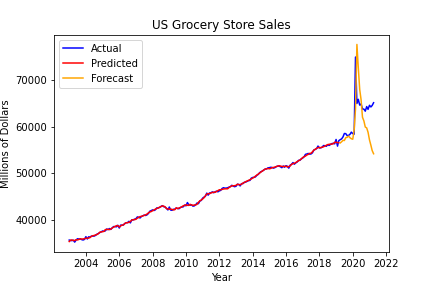
**Results**

The polynomial regression was first trained and tested on the entire dataset with a variety of degrees. The most accurate result was a polynomial with a degree of two. Below is a table of the R2 values for the model with various degrees and a plot of the 2nd degree polynomial predicted grocery store sales versus the actual store sales.

|  |  |
| --- | --- |
| **Polynomial Degrees** | **R2** |
| 2 | 0.9946 |
| 3 | 0.9832 |
| 4 | 0.8116 |
| 5 | 0.9333 |
| 6 | 0.9570 |



As we can see, the results through this process were relatively accurate with an R2 of 0.9946. The next step was attempting to see how this model predicted grocery store sales from past months. In order to do this, the training and testing datasets were split by month instead of a ratio of random selections. The forecast range selected here was 25 months, so slightly over the past two years. Attempting to forecast this way had poor results. The R2 value in this case was -0.81605. This indicates that the model now performs worse than if we were to just use the average. Below is a plot of actual, predicted, and forecasted grocery store sales numbers.



**Discussion and Conclusion**

Using a polynomial regression model to predict national grocery store sales had mixed results. Before the pandemic, the selected features here did a very good job at predicting grocery store sales. By narrowing the focus around the pandemic however, the features here performed much worse.

There are a few routes that could be explored from here.

1. Explore other features. Look into seeing if different or additional economic indicators improve performance.
2. There appeared to be a lag between indicators which would require some deeper analysis. For example, there was a large spike in grocery store sales and unemployment rate last year, but the spike in unemployment rate occurred a month after the increase in grocery store sales.
3. The large changes in multiple economic indicators as a result of the pandemic might simply be too difficult to model.

**References**

* Gross Domestic Income
  + Seasonally adjusted, Quarterly data since 1947.
  + Source: <https://fred.stlouisfed.org/series/GDI>
* Consumer Price Index for Urban Consumers
  + Seasonally adjusted, Monthly data since 1947.
  + Source: <https://fred.stlouisfed.org/series/CPIAUCSL>
* Grocery Store Sales
  + Seasonally adjusted, Monthly data since 1992.
  + Source: <https://fred.stlouisfed.org/series/RSGCS>
* Unemployment Rate
  + Seasonally adjusted, Monthly data since 1948.
  + Source: <https://fred.stlouisfed.org/series/UNRATE>
* Total Consumer Credit
  + Seasonally adjusted, Monthly data since 1943.
  + Source: <https://fred.stlouisfed.org/series/TOTALSL>
* Population
  + Monthly data since 1959.
  + Source: <https://fred.stlouisfed.org/series/POPTHM>
* Real Median Personal Income
  + Annual data since 1974.
  + Source: <https://fred.stlouisfed.org/series/MEPAINUSA672N>
* 10-Year Breakeven Inflation Rate
  + Monthly data since 2003.
  + Source: <https://fred.stlouisfed.org/series/T10YIEM>