XYZ Company Consulting Project

Levi Watson and Troy Rodriguez

STAT 480 – Spring 2023

# Introduction

The goal of this study is to explore the relationship between main economic factors and XYZ Company's revenue. To help make sense of the recent economic uncertainty as well as to plan for the future, knowing what economic factors will affect XYZ Company as well as the industry as a whole is a valuable asset. The following questions are the ones addressed by this study:

1. Do certain economic indicators correlate closer to XYZ Company, Total Retail, Non-Store Retail, and Electronic and Appliance Store sales than others? Which lags of which variables are best to watch?
2. How did the behavior of the economic indicators during the COVID-19 recession differ from their behavior during prior recessions?

In this report we will utilize a few methods to model and explain our findings. We will use Recurrent Neural Networks for modelling, ARIMAX for variable intuition, and Functional Principal Component Analysis on the variables to explore similarities and differences in the variables.

# Dataset

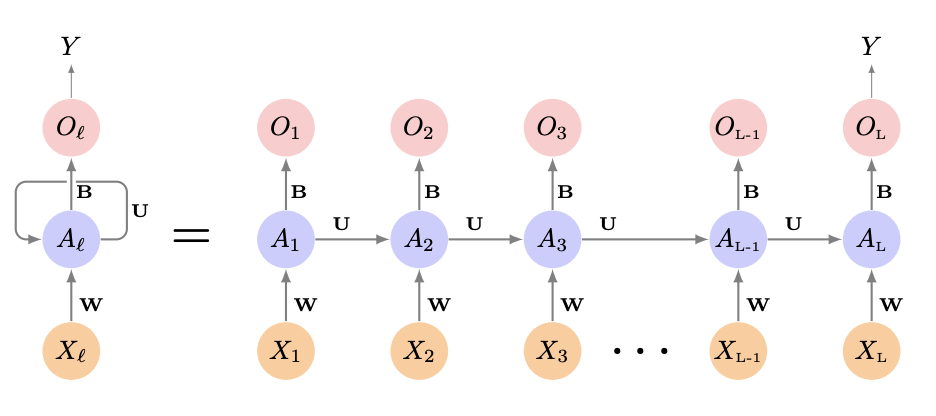
The data used in this experiment was gathered by XYZ Company utilizing many sources; the Bureau of Economic Analysis, the Bureau of Labor Statistics, the U.S. Census Bureau, the U.S. Employment and Training Administration, the National Association of Realtors, and various others calculated by Moody’s Analytics. Most of the response variables were captured by the U.S. Census Bureau, while XYZ Company’s revenue was captured by XYZ Company’s internal finance departments.

The data was stored in an .xlsx format, where all data (excluding XYZ Company’s Revenue) was stored as monthly time-series data, each attribute being recorded starting at different dates. Additionally, XYZ Company’s revenue was given in a quarterly format, so we extrapolated beyond that in proportion to industry performance to get estimated monthly revenue figures for XYZ Company. Information on recessions in the US was also included. A recession is defined as “the period between a peak of economic activity and its subsequent trough, or lowest point” with date ranges for each recession determined by the National Bureau of Economic Research.

# Data Analysis

Before models/methods were chosen, preliminary exploratory analysis was done on the data. This involved looking at each of the variables over time, histograms of values, correlation plots, among other methods. After that, models were chosen to analyze the data and answer the questions.

## RNN

The first part of this analysis was to use the economic factors to train a Recurrent Neural Network (RNN) predicting XYZ Company’s revenue. A Recurrent Neural Network is a series of nodes (each having their own input and output) that each feed into the following node in the network. 

In the above graph, X represents the inputs, A represents the nodes, and O represents the outputs. This specific kind of neural network is designed to look at not just the current values given to it, but the ones that come before. Each node feeds into the following one, meaning each node’s computation is dependent on the previous one’s, making it particularly useful in time-series and natural language processing applications.

The biggest shortcoming of this model is also its strength: its reliance on mathematical equations that are quite complicated and hard to interpret. This makes it particularly suited for finding complex relationships. But it also makes it difficult to fine-tune or explain, considering the inner workings of the model are mainly unintelligible.

Before feeding the data into the model, appropriate variable transformations were made, and the missing values were estimated for XYZ Company’s monthly revenue. The missing value estimation did not account for the seasonal spike in December 2012, so an average of December 2011 and 2013’s revenue was inserted at that point.

Training the model involves finding a middle ground between having an effective and a simple model. Setting the number of times the weights are recalculated (called epochs), the number of layers, the number of nodes per layer, which variables are used, and how much randomness is inserted in the model to prevent overfitting all were experimented with among other things until a sufficient model was trained.

The RNN that was trained provides a prediction for a year out for XYZ Company’s revenue, and it is also capable of displaying the ‘most important’ variables and lags based on its computations. This RNN was trained on lags 13-24 of all the economic factors originally given. Training on earlier lags allows this model to predict farther forward. After analyzing the important variables, a few redundant variables were removed to improve model performance. The final model decided on had 3 RNN layers.

Similar models were also used to predict the other response variables (industry revenue figures). The outcomes of these models were compared.

## ARIMAX

The main component of this analysis is to explain the intuition that is further explored by the ideas presented by the RNN. The intuition could be explained by a prescriptive regression model with ARIMA errors(ARIMAX). The regression portion would provide intuition, and the ARIMA errors will prevent violation of the white noise assumption.

The assumptions of an ARIMA model primarily involves checking the model with residual analysis. We need to ensure our model residuals have approximately normality, approximately constant variance, and no auto-correlation in the residuals of the model. These all can be assessed using the checkresiduals() function from the forecast library.

Firstly, we needed lagged predictor variables to use in the model. To help determine interesting lags, it was necessary to pre-whiten the predictor variables so that we could better predict the response variable at present time with lag of both predictor and response variables. Doing this helped determine which lags were significant, despite their seasonal patterns, and other confounding effects. This process involved determining an approximate and suitable time series model so that the ACF (Auto-Correlation Function) and PACF (Partial Auto-Correlation Function) appear as white noise. We limited this to only autoregressive models with a max order of three and at most one difference for simplicity and the large number of variables. With this AR transformation, we filter the response variable using the coefficients of the AR transformation. For example, an ARIMA(2,1,0) model could be written as:

Which we then apply to the response variable and expand:

With this, we would observe the ccf plots of the ARIMA predictor’s residuals against the filtered y values and create a list of significant lags for predicting the response.

We also wanted more predictors with lags to test during model fitting, so we also included other methods to pick out specific lags of predictors alongside our manual method. We used the prewhiten() function from the TSA package on R which accepts a predictor, response, and predictor ARIMA model. When implemented the predictor ARIMA models were fitted with the auto.arima() function, which limits the reliability of the results. We also, to some extent, picked and chose lags based on the CCF of the predictors against the response. With this we can fit the model by comparing coefficient significance, coefficient sign meaning, whether the residuals are white noise or not and if there were any patterns left in the residuals of the model plotted against each predictor.

When it came to fitting the model, there were multiple attempts with different approaches. The process that resulted in our current best model was formed using a recursive algorithm which can generally be summed up as:

1. Initialize model fitting with few predictors to a SARIMA(0,0,0)(0,0,0,12) model.
2. Add predictors incrementally until the residuals are uncorrelated.
3. Adjust SARIMA order.
4. Delete Insignificant predictors incrementally while making sure the residuals are uncorrelated with any lags.
5. Add lagged predictors incrementally until the residuals are uncorrelated.
6. Repeated 4-5 until no further improvements are found.

Periodically throughout this process, cross-referencing the correlations between all the predictors as well as checking the model residuals against the predictors assisted in giving further direction. We would also change from the method argument in the stats::arima() function between Conditional Sums of Squares, and Maximum Likelihood to see which fit better.

With this model, we’ll be able to say that the included predictors as a set, at the specified lags, can provide intuition on already determined revenue numbers. This model was not built with the intention of prediction. The RNN will work better for that as it can include many lags of all the predictors,

## Functional PCA

The last component of this analysis is clustering done on a Functional PCA on economic factors during recessions. This variation of PCA focuses on the slopes and trends in variables to determine similarities and differences between the different recessions. The number of months in a recession was limited to 10 in order to capture the whole curve of the COVID-19 Recession.

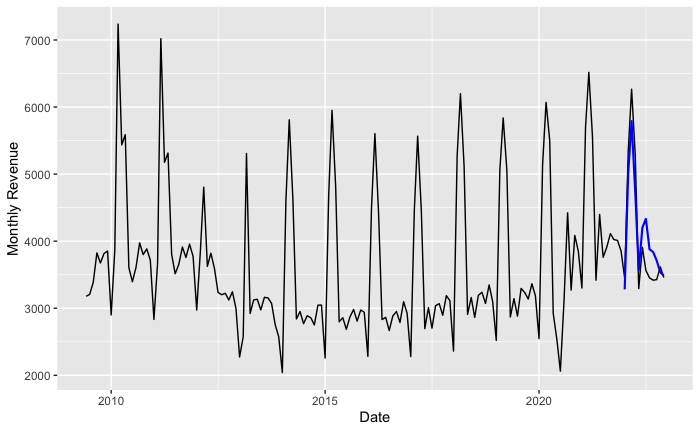
# Results

## RNN

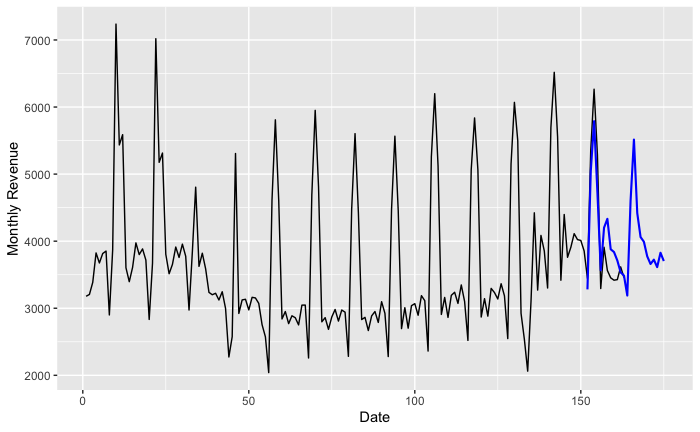
The recurrent neural network took the following as inputs (with transformations):

* GDP (YoY)
* PCE (YoY)
* Hourly Earnings
* Hourly Earnings (YoY)
* CPI
* CPI (YoY)
* Available Credit
* Disposable Income (YoY)
* CCI
* Home Sales (Seasonally Adjusted)
* New Home Sales (Seasonally Adjusted)
* Plans to Buy an Automobile
* S&P 500 (log)
* Unemployment Rate (seasonally adjusted, log10),
* Unemployment Claims (seasonally adjusted, log)
* Debt (log10)
* Disposable Income (log)
* Savings Rate (log)
* Presence of Recession
* Month of the year

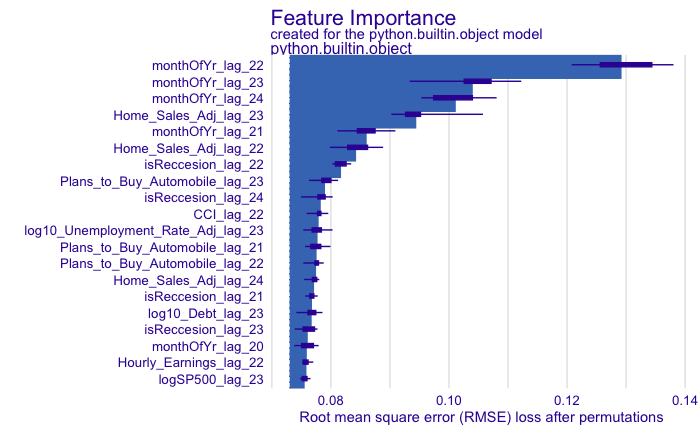
The final model contained 3 simple RNN layers. After training the model, the prediction for the last year of present data was given, with a mean squared error of 0.00578. (Plots by R’s ‘ggplot2’)



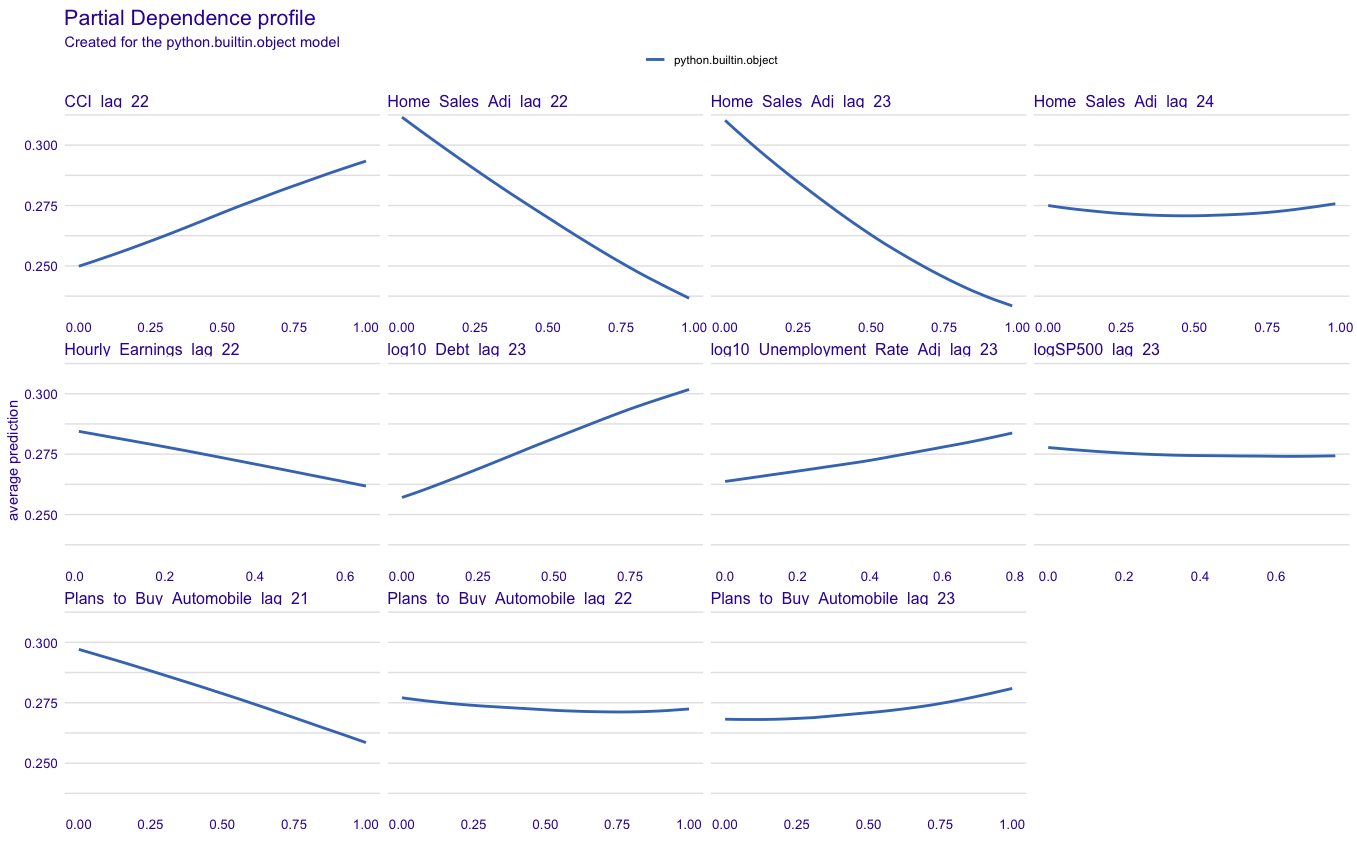
Because the model was trained on lags 13-24, this allows us to predict 1 year out with this model. The graph below shows the prediction for XYZ Company revenue:



In terms of important features, these are the variables that are the most important (e.g. most harmful to the model’s performance when randomized). (Plot by R’s ‘DALEX’)

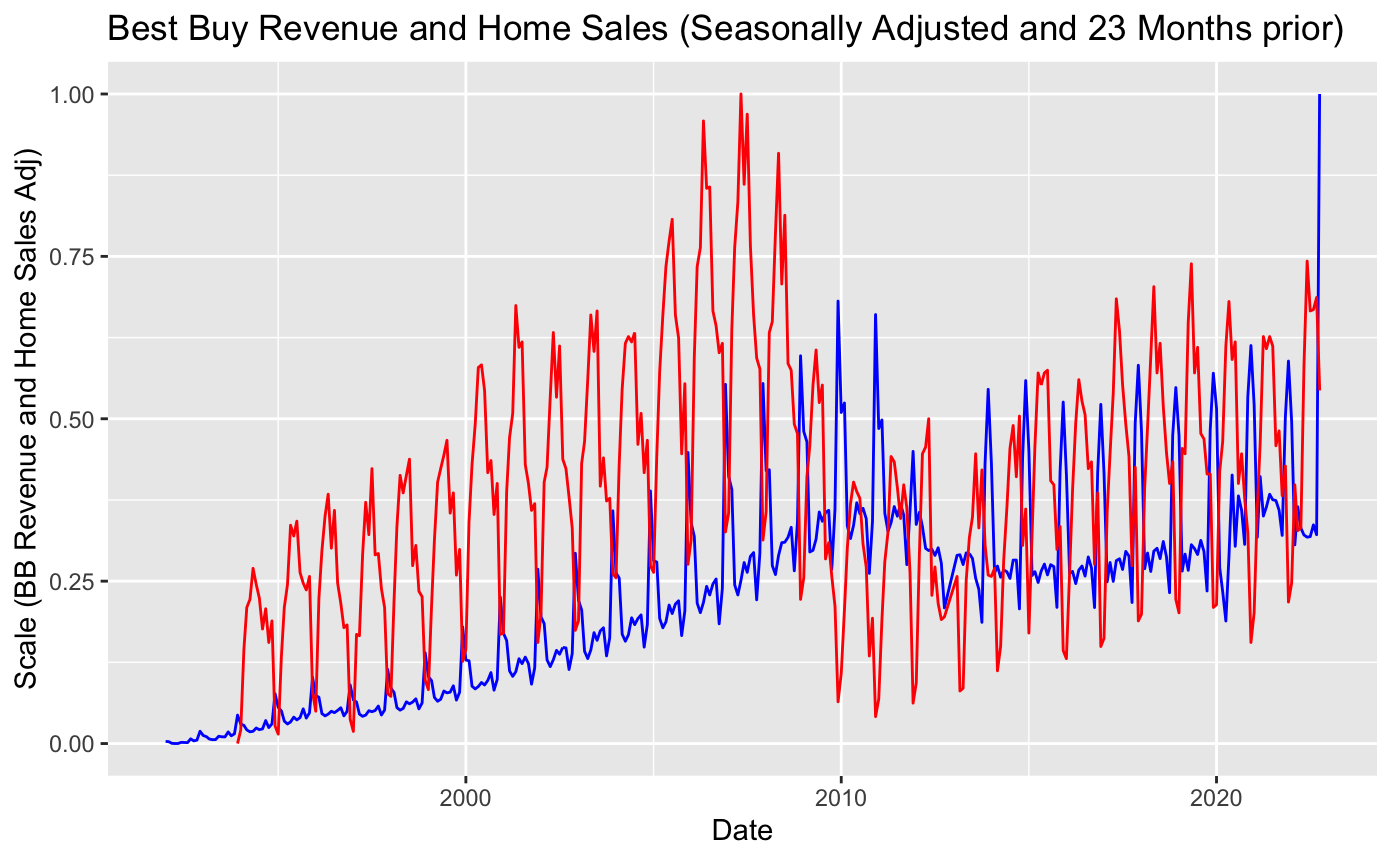


The model based a large amount of its predictions off of month of the year. This suggests that a large amount of XYZ Company’s revenue can be explained by seasonal fluctuations. Apart from that, the largest factors according to the model are Home Sales, the presence of a recession, plans to buy an automobile, CCI, Unemployment Rate, Debt, Hourly Earnings, and the S&P 500 Stock Index. The following plot shows how each of these variables influence the average predictions: (Plot by R’s ‘ingredients’)



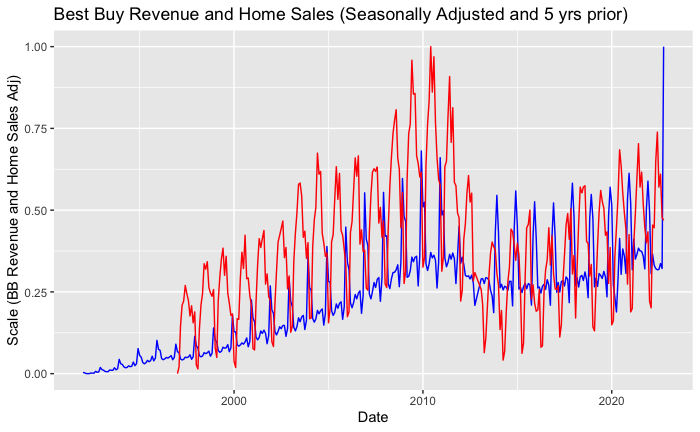
Following are some closer looks at the most important variables according to the model:

* Home Sales Seasonally Adjusted, Lag 23:



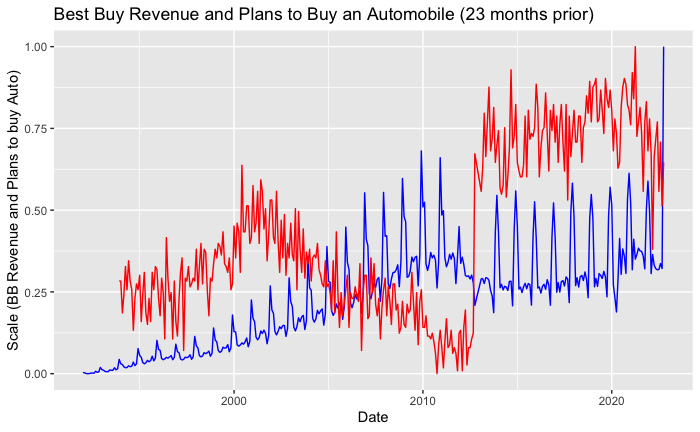
At 23 lags back, we can see that the spikes in XYZ Company Revenue and the downward spikes in Home Sales line up, especially during the training set (2010-2021). It’s not apparent if there is a causation or not between home sales and XYZ Company’s Revenue, but it’s clear that the seasonal fluctuations line up at this lag.

Additionally, it appears that Home Sales (Seasonally adjusted) follows the same trend as XYZ Company’s revenue, but 5 years before. The model says 23 years is the most important predictor, but it was only given 13-24 months as lags. Does XYZ Company’s revenue actually follow home sales 5 years prior? Chances are, the change in reporting month is the cause for XYZ Company’s revenue to change trend in 2013, and the recession in 2008 caused Home Sales to dip, and those just happen to line up. Below is the graph showing XYZ Company’s revenue and the Home Sales for 5 years prior.



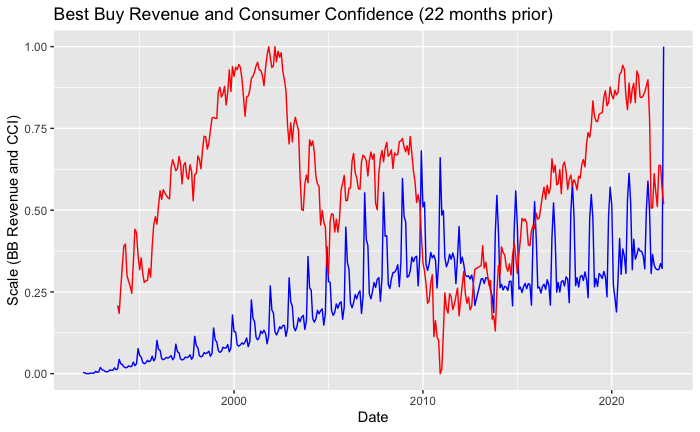
If Home Sales influences XYZ Company’s revenue from that distance, the case would probably be this: When homes are bought, give the owners a few years, and those people will be wanting to buy new appliances or TVs. This is a possible explanation for why the graph looks this way.

* Plans to Buy an Automobile, Lag 23:



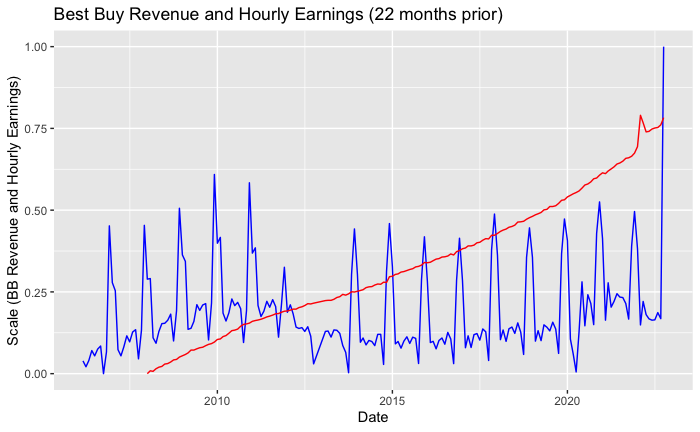
The large jump around 2011 in Plans to Buy an Automobile appears to be the result of an Auto industry boom in 2010-11, and we see XYZ Company’s revenue change trend because of the change in reporting month. After that point, we can see that XYZ Company’s Revenue follows a similar trend as Plans to Buy an Automobile. It appears that there isn’t enough data to see if these really follow the same trend.

* Consumer Confidence Index, Lag 22:



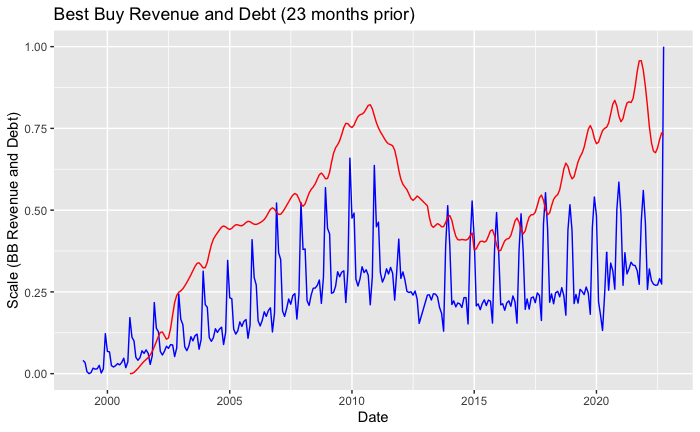
For CCI, the relationship isn’t so easy to pick out. We can see potentially that when there’s a lower CCI, XYZ Company’s Revenue decreases, but CCI is so volatile that it’s hard to see what’s going on.

* Hourly Earnings, Lag 22:



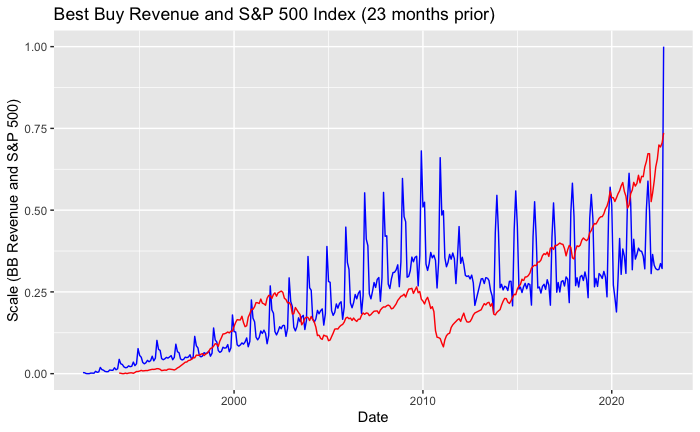
For this one it’s a little hard to interpret as well, considering Hourly Earnings just keeps increasing while XYZ Company’s revenue fluctuates. This one may just provide a sort of baseline for the prediction.

* Debt, Lag 23:



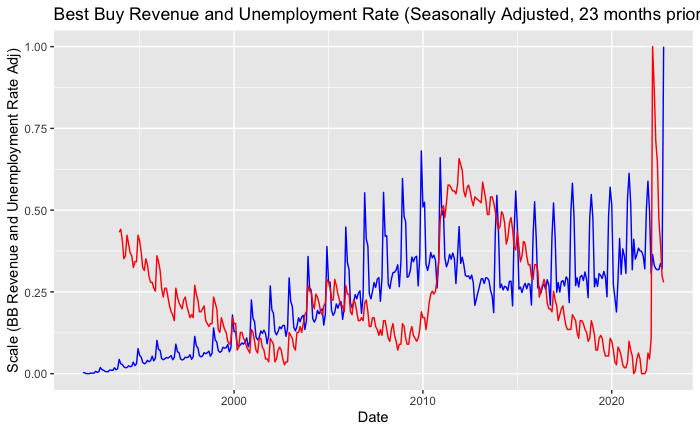
For Debt, this one’s a little more simple. These follow roughly the same trend.

* S&P 500, Lag 23:



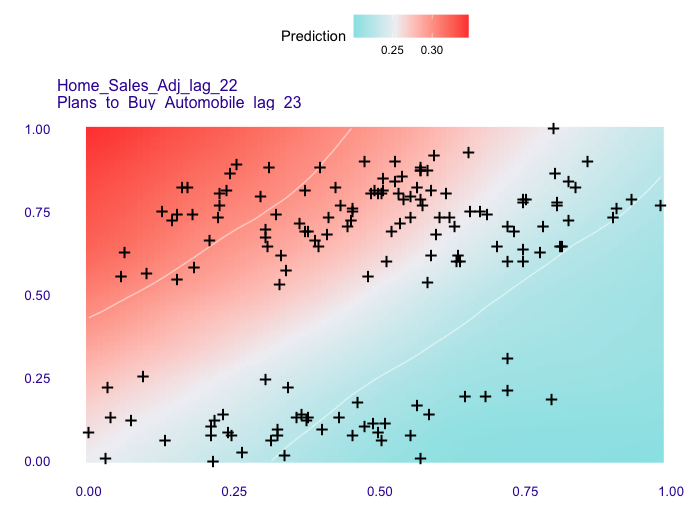
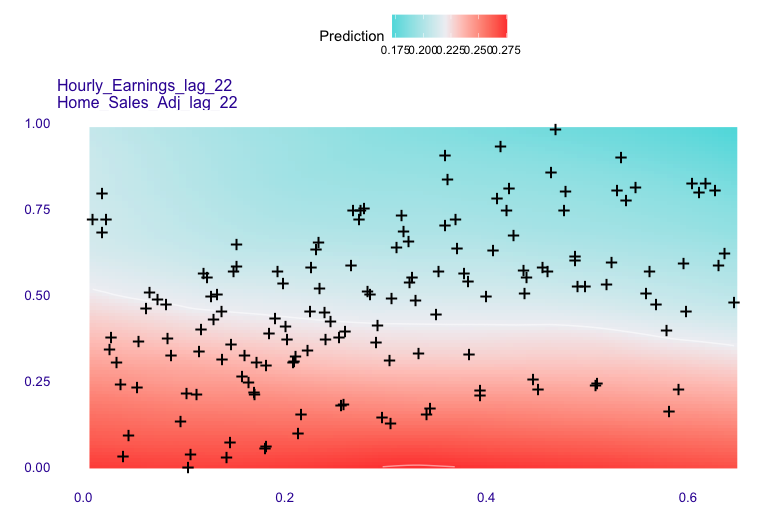
For the stock index, we can see a somewhat similar trend, as well as some seasonal line-ups at certain points in the graph.

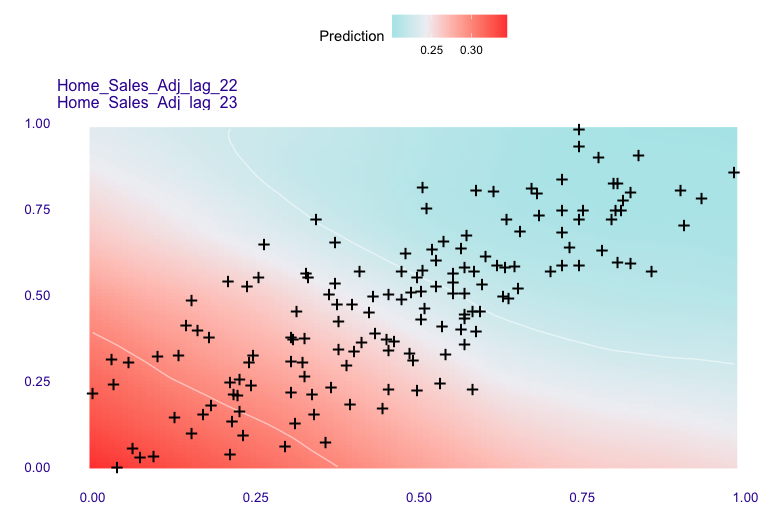
* Unemployment Rate Seasonally Adjusted, Lag 23:

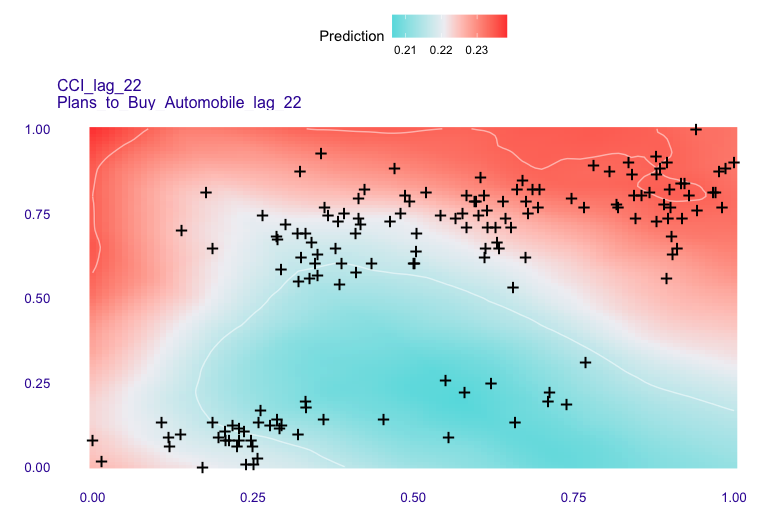


For unemployment rate, we can see XYZ Company’s revenue’s slope decreases when unemployment rate is increasing and the slope increases when unemployment rate decreases.

Because of the number of variables, visualizing interactions between variables would result in an excessive number of plots. Some variables have no interaction (See first plot), some have fairly simple interactions (see plots 2 and 3), but some have some interesting trends and shapes (see plot 4). Top feature listed is the X axis, and the bottom feature listed is the Y axis. Color represents the prediction at that specific point: (Plots by R’s ‘ingredients’)







## ARIMAX

For the ARIMAX model, using the steps outlined in the Methods section, the model was regressed on the response variable ln(Est XYZ Company Monthly Revenue), which had the missing values imputed using the na\_kalman() from the imputeTS library.

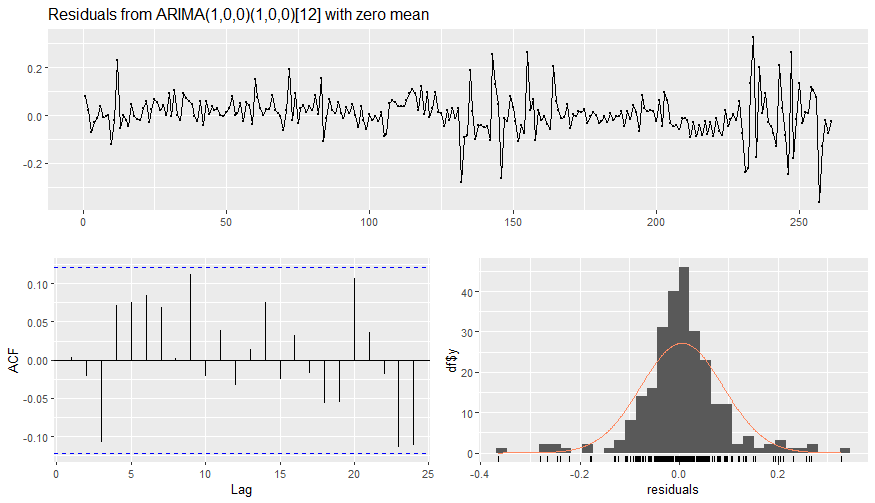
This model omits the use of Hourly Earnings and the included variants to try and capture more of the Revenue figures, as the earlier timestamp for Hourly Earning was 2007. In the end the time window was limited to Jan 2001 to October 2022 as the maximum lag used was -24, and the variables Debt and Available Credit initiated recording in Jan 1991.

Model fitting took into consideration each of the remaining supplied predictor variables with natural log transformations for Unemployment Claim variants. There was also no use of dummy variables for recessions, to see if the data could be explained with the predictors explicitly.

Throughout the fitting process, there was much difficulty prioritizing predictors, as there is such high collinearity in them, ex. GDP and PCE are highly correlated, despite measuring different things. Overall, this leads to a more fragile model, as there are many possibilities available. The result in our final regression model utilizes the following predictors and coefficients:



This is fitted with ARIMA(1,0,0)(1,0,0,12) errors using the Maximum Likelihood method. Performing residual analysis, a Ljung-Box test resulting in a P-value of 0.12 indicating the data is independently distributed. The model has the following residuals:



In the graph plotting the residuals over time, in general we don’t observe patterns in the data, the large spikes are attributed to the various recessions. The ACF plot shows no significant lags, and the distribution of the residuals appears fairly normal. This model performs well.

From this model, we can try to gain intuition on how XYZ Company Revenue will change over time based on the sign of the coefficients. In general, the coefficients for the predictors match Pearson’s r value for the predictor at lag 0 against Revenue.

So, in this case an increase in Seasonally Adjusted Home Sales from 17 months ago, Seasonally Adjusted Unemployment Claims from 11 months ago, and PCE YoY growth from 19 months ago all led to an increase in revenue for the current month. This makes sense as new homeowners may wait to purchase home furniture and entertainment systems until they’ve settled into their home and have adjusted to the new home’s cost. The growth in Unemployment Claims leading to increased revenue is a bit shocking, but considering how large the stimulus checks were during the COVID recession, this makes sense. For other recessions, unemployment claims could still lead to an increase in revenue for XYZ Company, as those on unemployment have more time that could be filled with entertainment found at XYZ Company. For PCE growth, this in general means consumers are spending more money, which naturally indicates the increase.

The only negative coefficient is Adjusted New Home Sales from 7 months prior. A rationale for this could be that new homes simply have appliances preinstalled, which delays a possible revenue source for XYZ Company.

The crossed terms are in general more difficult to interpret. An increase in both GDP YoY and Adjusted Unemployment Claims from 13 months prior indicates an increase in revenue, which makes sense as mentioned before in the rationale for the Unemployment Claims, and for PCE YoY as well as GDP YoY is highly correlated with PCE YoY. A decrease in GDP YoY with an increase in Adjusted Unemployment Claims might be the model’s attempt to stabilize revenue because of the astronomical increase in Unemployment claims and steep plummet of GDP during the Covid-19 Recession.

Chart, line chart

Description automatically generated

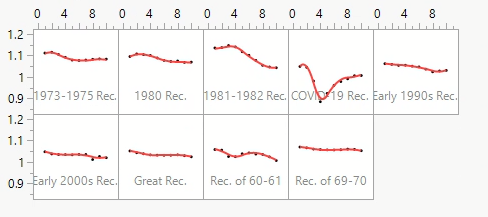
In this graph we see that GDP YoY and Adjusted Unemployment Rate tend to move in opposition, and when they do, there are irregularities in Revenue.

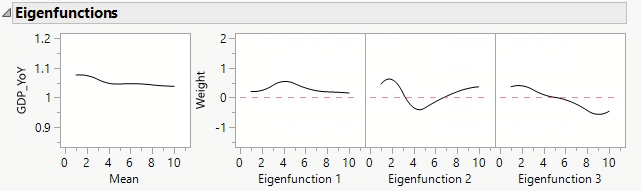
The last crossed term is Seasonally Adjusted CPI crossed with Debt from 14 months prior. An increase in both can be translated to “as goods and services get more expensive, and consumers have more debt, 14 months later revenue is expected to increase”. This could be interpreted as: it takes 14 months for consumers to stop trying to hold out from making purchases while prices are increasing due to inflation.

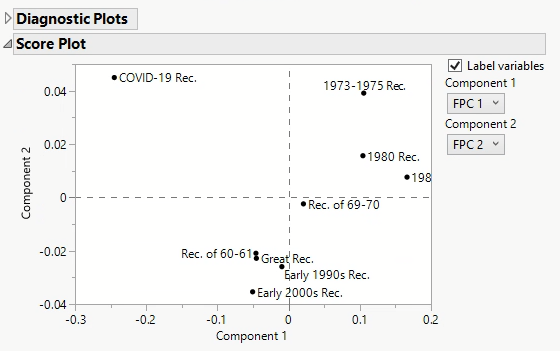
## Function PCA

The COVID-19 Recession was quite different than the other recessions, mainly in the fact that it was very short and very abrupt. We can see based on the first set of graphs for each of these variables this abruptness. The 2nd set of graphs under each variable are the eigenfunctions: the functions that are used to describe the current variable. The Score plots show how similar each recession is to the first two eigenfunctions for each variable.

Year-over-year GDP:

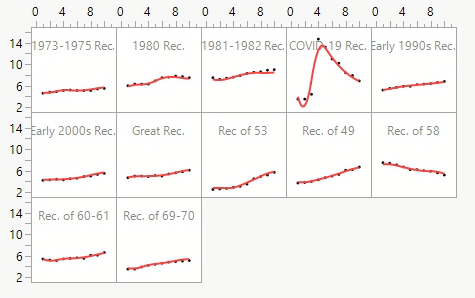


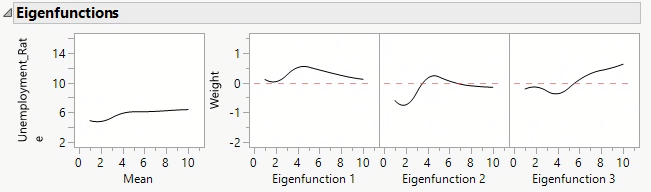


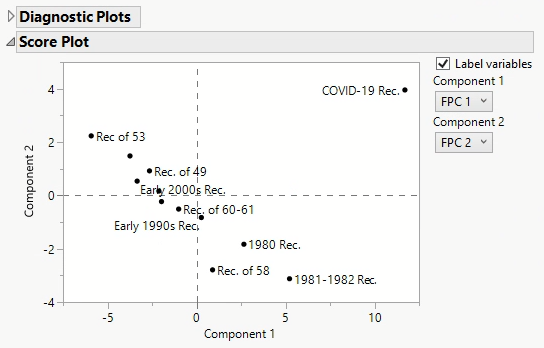


As we can see from this plot, the COVID-19 Recession follows the curve of Eigenfunction 2 quite a bit, but so did the first few months of the 1973-1975 Recession. What really sets the COVID-19 recession apart is that it inversely follows Eigenfunction 1 a lot more than the other recessions.

Unemployment Rate:

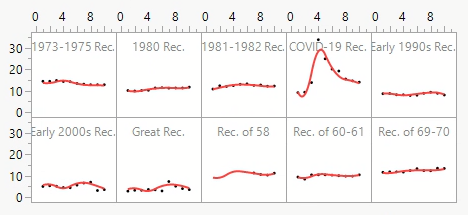


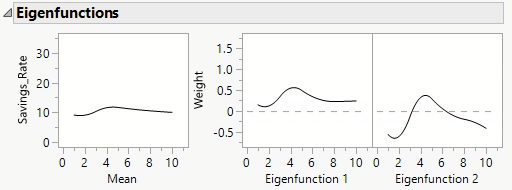


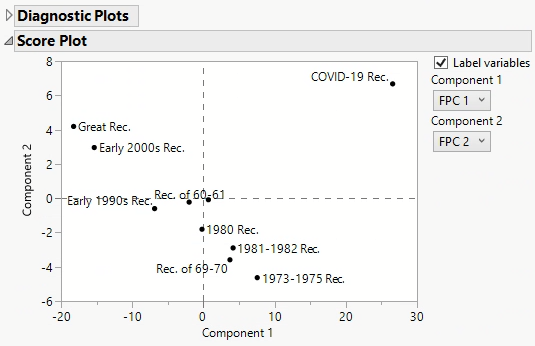


From these plots, we can see that most recessions had a very gradual climb in unemployment rate, but the COVID-19 recession dramatically spiked. The score plot shows that the COVID-19 recession follows the trends of Eigenfunction 1 and 2 both more than any of the other recessions.

Savings Rate:







From these plots we can see the savings rate during the COVID-19 Recession also spiked dramatically, where in other recessions it gradually grew or shrank.

Conclusion

From this report we’ve created a Recurrent Neural Network model that accepts inputs of the predictors and determines the best lags to predict a year out and then perform that prediction. We’ve also created a Regression model with ARIMA errors to help give intuition on which predictors and what lags are profound. Finally we performed Functional PCA to determine how the various recessions are different from each other, and determines that the COVID 19 recession is much different from many that came before.

The RNN provided a fairly accurate representation of XYZ Company’s revenue so far, and it drew attention to the concept that economic indicators and company revenue are all influenced heavily by the seasons. It did pick out home sales, plans to buy an automobile, consumer confidence, and the unemployment rate as things that heavily influenced its calculations.

The ARIMAX model accounted for the seasonality present in the data and leaves a few key indicators to watch: Home Sales, New Home Sales, Unemployment Claims, and PCE.

Both the RNN and the ARIMAX models agreed that Home Sales is the best metric to watch in order to predict XYZ Company’s revenue. Give homeowners a year or two to get acclimated to their new house, and they’ll be coming to XYZ Company for things like appliances or entertainment systems.