Spring 2022

Finding the Best Predictors for Foot Traffic in US Seafood Restaurants

Isabel Paige Beaulieu
University of New Hampshire

Finding the Best Predictors for Foot Traffic in US Seafood Restaurants

Isabel Beaulieu

The University Of New Hampshire

College of Engineering & Physical Sciences

Honors Thesis

Faculty Advisors:

Professor Easton White

Assistant Professor Biological Sciences

Professor Philip Ramsey

Principal Lecturer Mathematics & Statistics

Contents

1 Abstract

COVID-19 caused state and nation-wide lockdowns, which altered human foot traffic, especially in restaurants. The seafood sector in particular suffered greatly as there was an increase in illegal fishing, it is made up of perishable goods, it is seasonal in some places, and imports and exports were slowed. Foot traffic data is useful for business owners to have to know how much to order, how many employees to schedule, etc. One issue is that the data is very expensive, hard to get, and not available until months after it is recorded. Our goal is to not only find covariates that can accurately predict foot traffic in United States seafood restaurants, but find ones that are free, easily accessible, and available in real time.

We built multiple models in R-Studio using cross validation and compared using mean squared error. We found that the google search trend of 'seafood takeout' was most strongly correlated to foot traffic, correlation coefficient is .602, and appeared in every model. The model with the lowest mean squared error is a generalized linear model (GLM) with google search trends, unemployment rate, and average covid cases per week as the predictors. These predictors were selected using backwards selection by akaike information criterion (AIC). We are still awaiting new foot traffic data to test our model on, which shows how long it takes to receive the data. Next steps would be to look at if more models can be made that focus on individual states as well as finding other covariates that could be used.

2 Literature Review

2.1 *COVID-19 and US Fisheries*

COVID-19 negatively impacted a lot of industries in the United States, the seafood sector especially. COVID-19 is known as a 'black swan' event which changes the way people think, buy, and consume food. The pandemic has caused labor problems, shutdown of factories, and food shortages (James, Witten, Hastie and Tibshirani, 2021). In the seafood sector, fishing seasons were shortened and workers often are working in tight working conditions, raising the risk of spreading the disease (White et al., 2020). The United States has wild lobster, farmed salmon, and geoduck that is exported to China. Due to imports and exports being affected, this greatly affected the supply chain in the seafood sector (Love et al., 2021). One study sums it up nicely saying that:

> Finally, the consequences of the pandemic in fisheries have been also exposed by the Food and Agriculture Organization, among which are the restrictions imposed on fishing activities (management, production, and supply of fisheries products), worse working conditions (concerns and difficulties to work in the safe and confidential environment, disruption of at-sea surveys affecting stock assessments), and the decrease in sales due to the closure of the hospitality industry (Ruiz-Salmón et al., 2021).

Seafood is one of the most traded food commodities both regionally and globally. For many people, seafood plays a big role in food and nutrition security. One issue with policies in place regarding pandemic aid is that women and migrant workers might not benefit from this aid (Love et al., 2021). Another downfall for seafood is that it is perishable. Processed and

non-perishable food companies have benefited the most from COVID-19. There was also a three to four fold increase in sales in foods known to boost the immune system (James, Witten, Hastie and Tibshirani, 2021). The people that depend on seafood are not getting an adequate amount of aid and the seafood industry is not getting enough funding when it is struggling greatly.

2.2 *Google Trend Data*

Google search trends have been used to predict various trends. For example, a study in 2011 shows that web search data is a useful predictor in tourism rates in Hong Kong (Gawlik, Kabaria and Kaur, 2011). When looking at the financial industry, it is shown that search terms related to finance patterns may indicate 'early warning signs' of changes in the stock market (Preis, Moat and Stanley, 2013). One goal of this paper is to see if the Google search term 'seafood takeout' will be a good predictor for foot traffic in seafood restuarants.

Google trend data is available in real time and is a weekly index of the volume of queries someone enters into Google. A query index is based on a share, which is:

> the total query volume for the search term in question within a particular geographic region divided by the total number of queries in that region during the time period being examined. The maximum query share in the time period specified is normalised to be 100, and the query share at the initial date being examined is normalised to be zero.(CHOI and VARIAN, 2012)

It is noted that for the seafood industry in particular, important data is not available for months or years later, so the real time feature of google search data is extremely beneficial for businesses (White et al., 2020). Varian points out that google trends are helpful in 'predicting the present' and it is uncertain if they can predict the future (CHOI and VARIAN, 2012). This

paper focuses on how accurate this data is in predicting future foot traffic.

2.3 *Lockdowns and Foot Traffic*

Because of COVID-19, there were country and nationwide lockdowns put in place to help slow the spread of the disease. Foot traffic in a lot of places and in certain industries changed drastically. In the entertainment industry, many public events were postponed or canceled and major companies such as Netflix, Amazon, and Apple dropped out of events, premieres, and panels (Moon, 2020). Because of this, financial constraints were felt by studios, theaters, and filmmakers (Moon, 2020). The less money people have, the less likely they are to get take out or go out to eat. Lockdowns also caused many restaurants to be take out only. This is negative for the seafood industry, as 65% of spending on seafood in the United States is in restaurants (White et al., 2020). COVID-19 caused a shift in food sourcing, and citizens favored eating at home compared to in a restaurant. There was also an increase in online search terms for 'seafood recipes' and 'seafood delivery' compared to the previous four years (Love et al., 2021).

2.4 *Big Data*

The source we used to get foot traffic data for United States seafood restaurants was from SafeGraph. This company uses points of interest (POI) and tracks how many people visit each one by cell phone data. Each POI gets a Safegraph place ID so it is easy to look at data for certain places (Unique IDs for Store Locations | SafeGraph Places, 2022). How accurate is SafeGraph? There was a study done in Yellowstone National Park to assess the validity of the service in 2021. The results of this study found that there was no significant difference between the SafeGraph metrics and official park visitation statistics (Liang et al., 2021). SafeGraph

appears to be accurate and is a good source to use for this project.

There is a lot of benefit in using 'big data'. For example, seafood restaurant owners would definitely benefit if they were able to know what their expected foot traffic would be in six months. This would help them budget, know how much to order, and anticipate how many employees they will need. To be successful with big data though, companies need to pay attention to data flows and rely on data scientists to analyze the data (MIT Sloan Management Review, 2012). Unless the seafood restaurant is a chain, it is highly unlikely they have a team of data scientists to look at foot traffic data. Another boundary for restaurants in particular is the cost of big data. When looking at data transfer, the biggest problem is the economics, not necessarily the technology (Marx, 2013). The SafeGraph dataset we used was approximately one million rows of data. On their website, they say each record is 10 cents, which brings this dataset out to about $100,000. Most businesses cannot afford this and this is why we want to see if there are other variables that could predict foot traffic that are free and easily accessible.

3 Methods

3.1 *Variable Selection and Collection*

As found during the literature review, the goal of this research is to find covariates that are not only accurate in predicting foot traffic, but ones that are accessible and available almost instantly after being recorded. Because foot traffic depends on consumer behavior, we tried to pick variables that captured that. One covariate was the weekly hits that the google search trend 'seafood takeout' had in the United States. The google search trend 'seafood takeout' will be referred to as 'Hits' in this paper. To gain insight about the economy, the weekly unemployment rate was a variable as well as the closing price of the NASDAQ. Because we expect COVID-19 to impact foot traffic in seafood restaurants, we added a seven day rolling average of the daily share of the population receiving a COVID-19 vaccine dose to the model. All doses, including boosters, are counted. The seven day average of new COVID-19 cases was also added to the model. Another variable was a factor variable that indicated pre COVID-19 versus after. Lastly, an index measuring how strict the United States is with restrictions, mask and vaccine mandates, was added to the model. All of this data was collected from free, online sources and gathered weekly. We compiled these many datasets into one before fitting various models.

3.2 *Generalized Linear Model*

Four out of the five models that were fit were GLMs. This type of model was chosen because it is very flexible and all our response variables are continuous. All of the potential covariates are also continuous, with the exception of pre COVID-19 versus after which is a factor variable. In a

GLM, the response is from the exponential family. The distribution of the exponential family is:

$$f(y|\theta,\phi) = exp[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)]$$

The mean and variance of the exponential family are:

$$EY = \mu = b'(\theta)$$

$$varY = b''(\theta)a(\phi)$$

The way that a GLM estimates the parameters β is by using the maximum likelihood estimate (MLE). The log likelihood of a single observation is:

$$logL(\theta_i,\phi;y_i) = w_i[\frac{y_i\theta_i - b(\theta_i)}{\phi}] + c(y_i,\phi)$$

where

$$a_i(\phi) = \frac{\phi}{w_i}$$

(Faraway, 2006)

3.3 *Ridge Regression*

Very similar to least squares regression, coefficients in ridge regression are estimated by minimizing

$$RSS = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}(\beta_j x_i j))^2$$

where p is the number of predictors (James, Witten, Hastie and Tibshirani, 2021). Ridge regression helps address overfitting and we used this because there was some evidence of multicollinearity in the generalized linear models.

3.4 *Normalizing Foot Traffic Data*

The foot traffic data that was used as the response is from SafeGraph. This company collects data on every individual business and categorizes each business type. To collect the data, they partner with mobile applications that obtain opt-in consent from its users to collect anonymous location data (Unique IDs for Store Locations | SafeGraph Places, 2022). The data was presented as a row for each seafood restaurant in the United States on a single day with the total number of visitors for that day. To get an estimate for all seafood restaurants in a given day, using the tidyverse package in R, we took the sum of all of the visits per day at each restaurant and divided that number by the total number of phones seen that day. We then took the average per week to get the data in a weekly format.

3.5 *Data Partitioning*

To avoid over fitting the model, which is when the model corresponds too much to one data set, we separated the data into training and test sets. It is standard practice to make about 70-80% of the data the training data, and the remainder of the data the test set. In choosing whether or not to randomly select the data set or use the first 75% of the rows, we looked at a time series plot of foot traffic. Since the data had some seasonality and unexplained behavior with COVID-19, we decided to randomly select the rows to include in the training data.

4 Data Analysis

We constructed five different models to predict foot traffic and compared them to see what one performed best. After making a correlation matrix of all the covariates with foot traffic, Hits was the most correlated with foot traffic with a value of .602 (Fig 1).

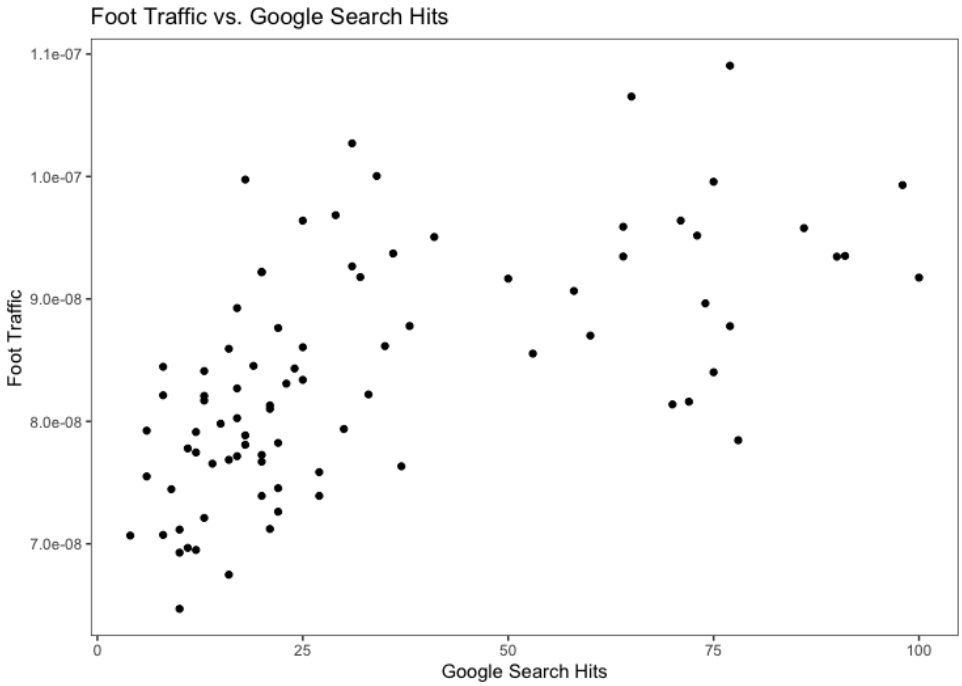

Figure 1: Dot plot of foot traffic in United States seafood restaurants versus google search hits for the term 'seafood takeout'

We then built a model containing Hits, Unemployment Rate, and Covid Cases. This model was chosen by adding all the covariates to the model and reducing it by using backwards AIC. AIC looks at a model's MLE and uses it as a measure of fit. One reason it is a favored method to rate models is because it adds a penalty term if the model contains a lot of parameters, which helps assess overfitting (Zajic, 2019).Another GLM was created using the previous week's Hits as a predictor as well as unemployment rate and the factor variable of pre COVID-19 or not. The last GLM created started with all pairwise interactions and was reduced using AIC. Finally, a ridge regression model was created and added to the mix.

Before looking at which model performed best, we decided to look at the effect size of each of the potential covariates before reducing the model (Fig 2).

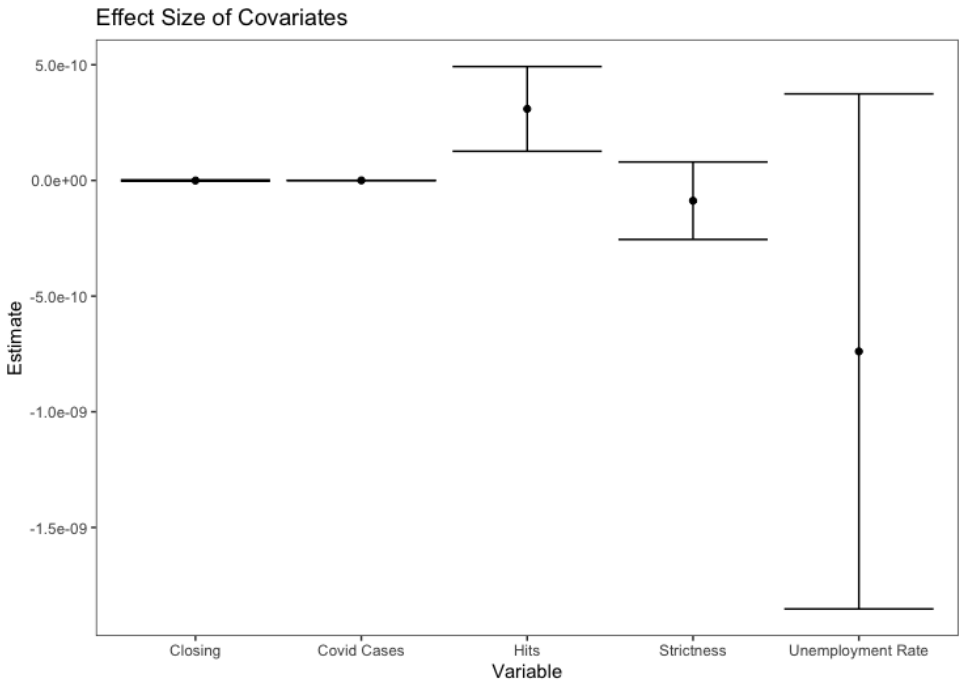

Figure 2: Chart of the coefficients in the model before model selection

When comparing the models, the mean squared error (MSE) and AIC were used (Table 1).. For both of these metrics, the lower the value, the better the performance. For all of the models except for the GLM with the previous week's Hits, the MSE was practically zero. The smallest MSE of the five models was the GLM with only Hits, and the GLM with Hits, Unemployment Rate, and Covid Cases. Since they were so close in MSE, we looked at AIC to compare them. The larger model had a lower AIC, so that is what we chose to be the best model (Fig 3). It is important to note that for different types of models, AIC cannot be used to compare them.

Table 1: MSE and AIC values for each model that was fit

	MSE	AIC
Hits as the only covariate	0.00	-2180.16
Best GLM chosen by AIC	0.00	-2180.49
Ridge Regression	0.00	13.50
Best GLM chosen by AIC using the previous week's Hits	1107.19	-2189.19
Best GLM using interaction terms	0.00	-2184.17

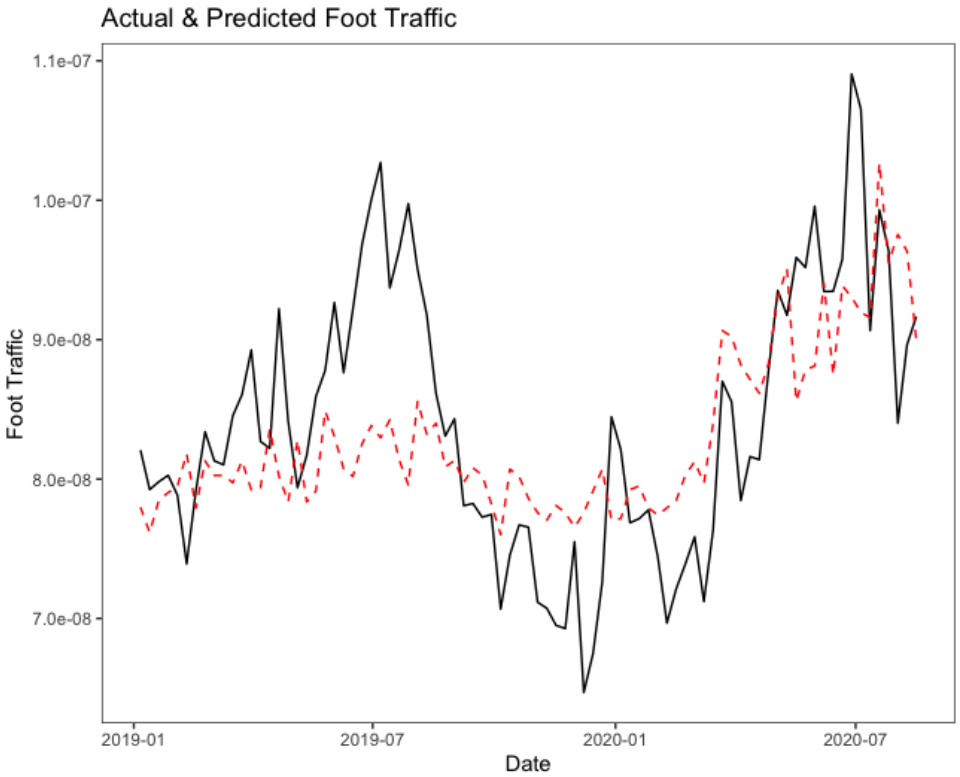

Figure 3: Actual foot traffic (black line) plotted with how the best model predicts foot traffic (dashed red)

After choosing the best model, we predicted weekly foot traffic from August 23, 2020 to January 30, 2022 (Fig 4). As noted previously, foot traffic data is hard to get and not available until

after it is recorded. We still do not have access to the current data, but have the predicted values stored for when that becomes available.

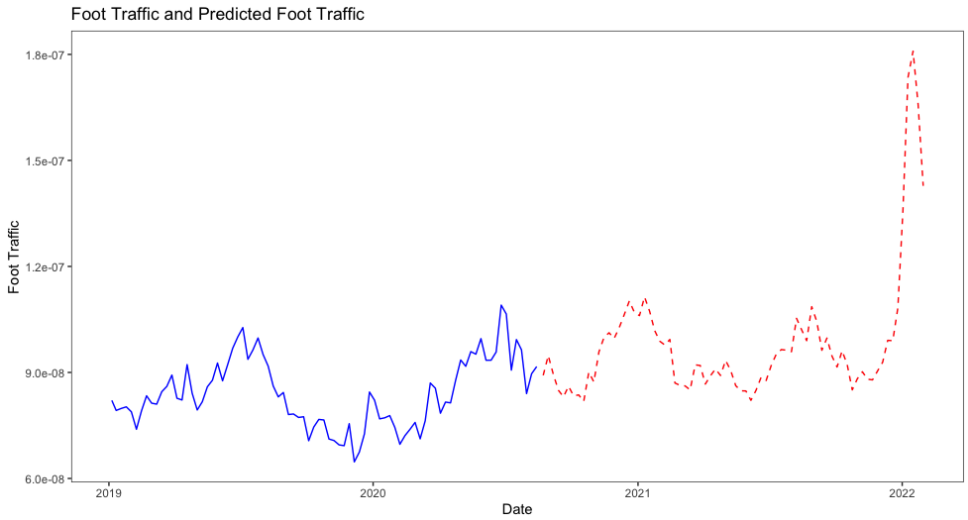

Figure 4: Time series plot of foot traffic (solid blue line) alongside the predicted foot traffic based on our best model (dashed red line)

We observed a large spike in foot traffic leading up to 2022 (Fig 4). After looking at plots of the three predictors, it appears that the downward trend in unemployment rate and upward trend in covid cases could contribute to this steep spike, but nothing can be certain until we receive the actual foot traffic data (Fig 5).

One model that is worth looking at even though it had the worst MSE is the model using the previous week's Hits to predict foot traffic. We decided to build this model to see if previous week's data would be a good predictor since companies might need to plan in advance and will not have time to wait until the present data is available. It was disappointing that the MSE was the worst, but surprisingly the AIC was the best. Looking at the actual versus predicted plot, it looks very similar to that of the best chosen model (Fig 6).

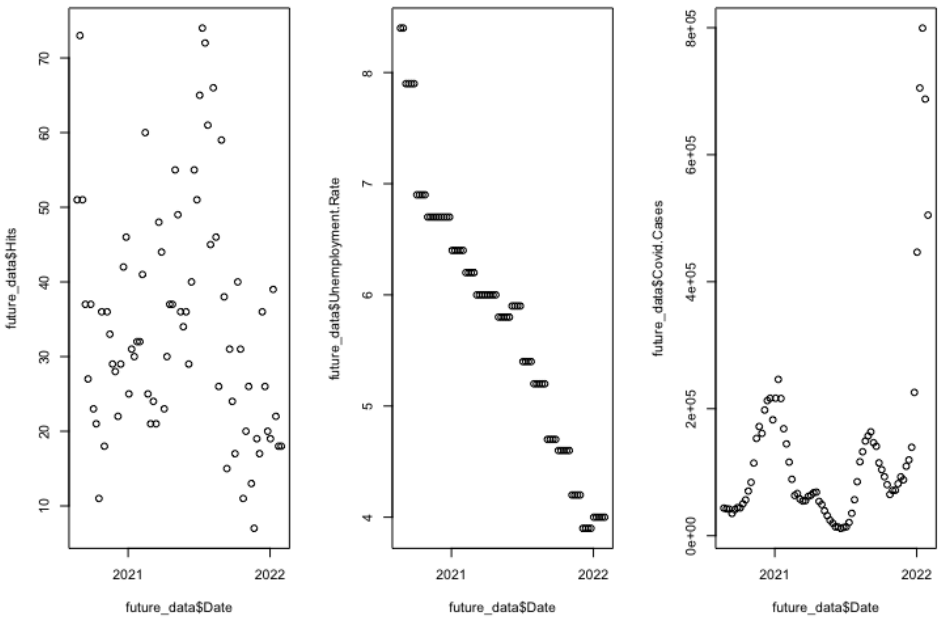

Figure 5: Time series plots of Hits, Unemployment Rate, and weekly average new Covid Cases from August 23, 2020 to January 30, 2022

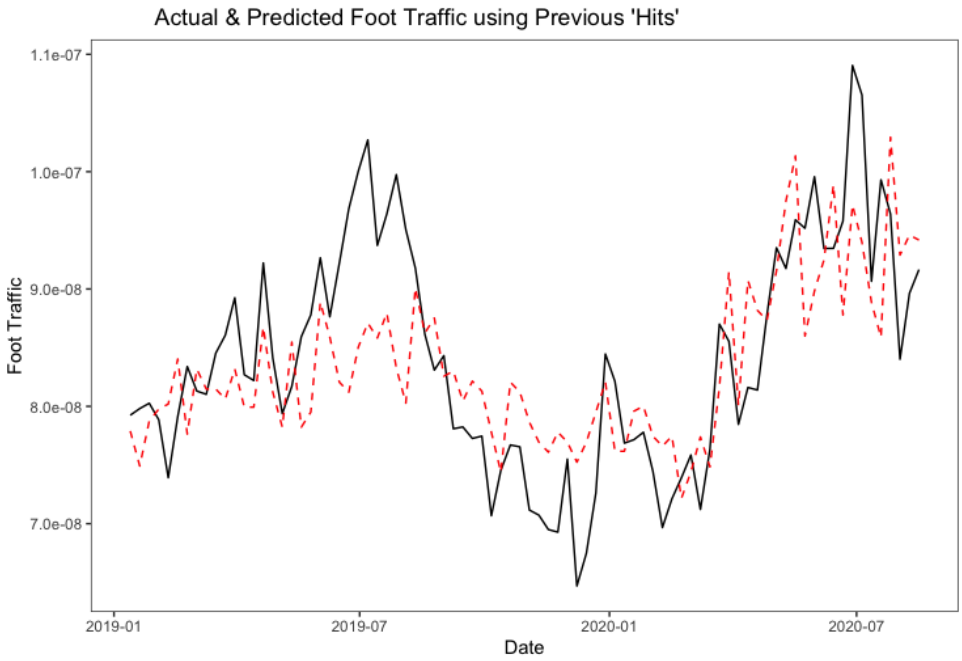

Figure 6: Actual foot traffic (black line) plotted with how the lagged model predicts foot traffic (dashed red)

5 Conclusion

In our study, we built five different models to find the best method to predict foot traffic in seafood restaurants in the United States. Because COVID-19's impact on the seafood industry and changes in consumer behavior, foot traffic is a valuable thing for business owners to be able to predict. Given that foot traffic data is extremely expensive, we proved that there are free and easily accessible predictors that could produce accurate results. The best predictors we found were Hits, Unemployment Rate, and Covid Cases.

Even with these results, there is still more that could be researched. We have yet to receive the foot traffic data for August 23, 2020 and onward, but once we do we will see how well our model performed. As COVID-19 is becoming more normal in our society and more people are vaccinated, it would be beneficial to research if there are other covariates that could predict foot traffic.

Our results using the previous week's Hits also raise a flag for further investigation. It would be interesting to see how many weeks of past Hits data produces accurate results and how far in advance we can accurately predict foot traffic. It would also be interesting to test combinations of previous weeks data for the other covariates as well.

6 References

CHOI, H. and VARIAN, H., 2012. Predicting the Present with Google Trends. Economic Record, 88(s1), pp.2-9.

Faraway, J., 2006. Extending Linear Model With R. London: Chapman & Hall/CRC.

Galanakis, C., Rizou, M., Aldawoud, T., Ucak, I. and Rowan, N., 2021. Innovations and technology disruptions in the food sector within the COVID-19 pandemic and post-lockdown era. Trends in Food Science & Technology, 110, pp.193-200.

Gawlik„ E., Kabaria, H. and Kaur, S., 2011. Predicting tourism trends with Google Insights.

James, G., Witten, D., Hastie, T. and Tibshirani, R., 2021. An Introduction to Statistical Learning with Applications in R. 2nd ed.

Liang, Y., Yin, J., Pan, B., Lin, M. and Chi, G., 2021. Assessing the validity of SafeGraph data for visitor monitoring in Yellowstone National Park. University of Massachusetts Amherst.

Love, D., Allison, E., Asche, F., Belton, B., Cottrell, R., Froehlich, H., Gephart, J., Hicks, C., Little, D., Nussbaumer, E., Pinto da Silva, P., Poulain, F., Rubio, A., Stoll, J., Tlusty, M., Thorne-Lyman, A., Troell, M. and Zhang, W., 2021. Emerging COVID-19 impacts, responses, and lessons for building resilience in the seafood system. Global Food Security, 28, p.100494.

Marx, V., 2013. The big challenges of big data. Nature, 498(7453), pp.255-260.

MIT Sloan Management Review, 2012. How 'Big Data' is Different. [online] Available at: https://www.hbs.edu/ris/Publication%20Files/SMR-How-Big-Data-Is-Different_782ad61f-8e5f-4b1e-b79f-83f33c903455.pdf.

Moon, S., 2020. Effects of COVID-19 on the Entertainment Industry. IDOSR JOURNAL OF

EXPERIMENTAL SCIENCES, 5(1).

Preis, T., Moat, H. and Stanley, H., 2013. Quantifying Trading Behavior in Financial Markets Using Google Trends. Scientific Reports, 3(1).

Ruiz-Salmón, I., Fernández-Ríos, A., Campos, C., Laso, J., Margallo, M. and Aldaco, R., 2021. The fishing and seafood sector in the time of COVID-19: Considerations for local and global opportunities and responses. Current Opinion in Environmental Science & Health, 23, p.100286.

Safegraph.com. 2022. Unique IDs for Store Locations | SafeGraph Places. [online] Available at: https://www.safegraph.com/product-info/store-id [Accessed 2 April 2022].

White, E., Froehlich, H., Gephart, J., Cottrell, R., Branch, T., Agrawal Bejarano, R. and Baum, J., 2020. Early effects of COVID-19 on US fisheries and seafood consumption. Fish and Fisheries, 22(1), pp.232-239.

Zajic, A., 2019. Introduction to AIC-Akaike Information Criterion. [online] Medium. Available at: https://towardsdatascience.com/introduction-to-aic-akaike-information-criterion-9c9ba1c96ced [Accessed 2 April 2022].