IEOR 4728 Big Data and Investment Research Final Project

Summary

We forecast Same Store Sales Growth (SSSG) of five fast food restaurant brands operating predominantly in the United States using foot traffic, Google Trends, US retail sales, and credit card spending. The best (and most parsimonious) model achieves an R^2 of 0.75 using only a single linear predictor. We also investigate whether a profitable trading strategy can be created using the model and observe excess returns only using publicly available data. We buy (short) if SSSG forecast is more than (less than) 0.5% as compared to the Bloomberg analyst consensus. Trading positions are initiated one month before quarter end and closed one day after quarter end.

Thesis

Foot traffic data obtained from SafeGraph can forecast changes in SSSG for publicly-traded fast food restaurants which obtain the majority of their revenue in the United States. Restaurants chosen for our analysis include Chipotle, Del Taco, Jack in the Box, Shake Shack, and Wingstop.

Signal

Investment Universe

We evaluated 10 limited service restaurants and ultimately decided on 5 companies to be the investable universe for this project. We choose restaurants that obtain over 90% of their global revenue from the United States. This was done because SafeGraph data is only available from US customers. Further, we discarded restaurant chains like Domino’s since a significant portion of their revenue is from deliveries.

Table 1: Investment Universe

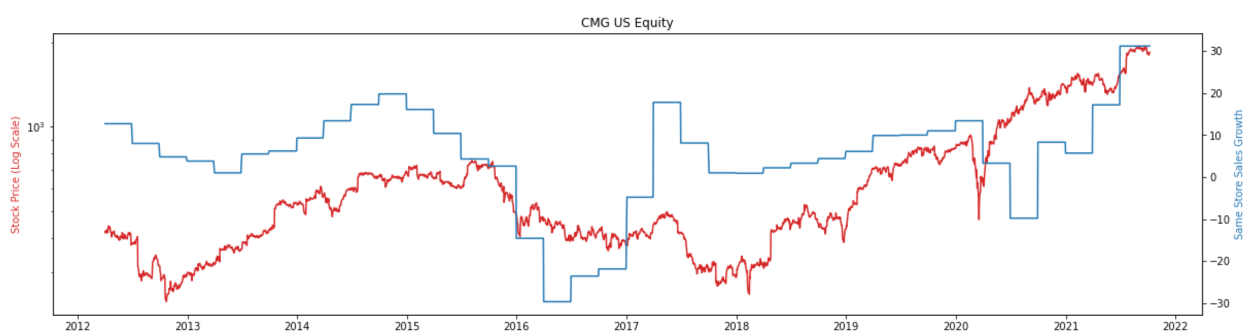
Firm	Ticker	US Revenue / Total Revenue	Annual Revenue ($M)	Market Cap ($M)	Include in Universe
Chipotle	CMG	99%	5,985	51,882	Yes
Del Taco	TACO	100%	492	332	Yes
Jack in the Box	JACK	100%	1,022	2,167	Yes
Shake Shack	SHAK	95%	523	3,282	Yes
Wingstop	WING	89%	249	4,778	Yes
Domino's	DPZ	93%			No
Wendy's	WEN	83%			No
Starbucks	SBUX	72%			No
Papa John's	PZZA	38%			No
McDonalds	MCD	38%			No

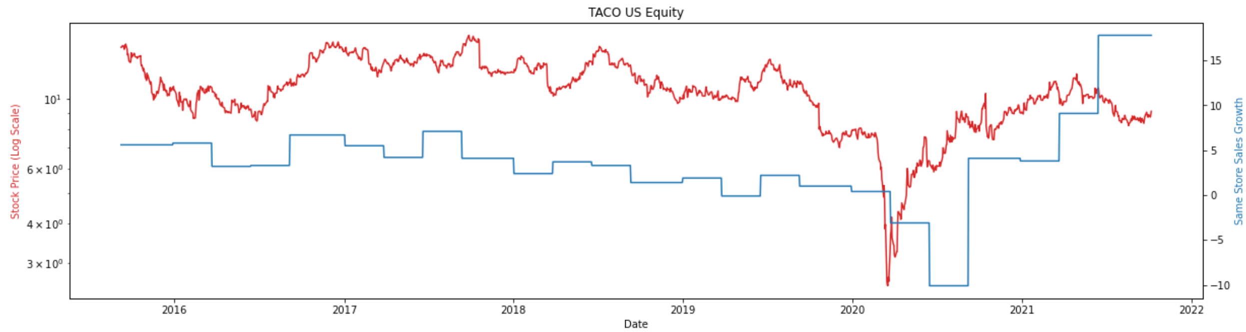
Source: Bloomberg Terminal accessed October 2021

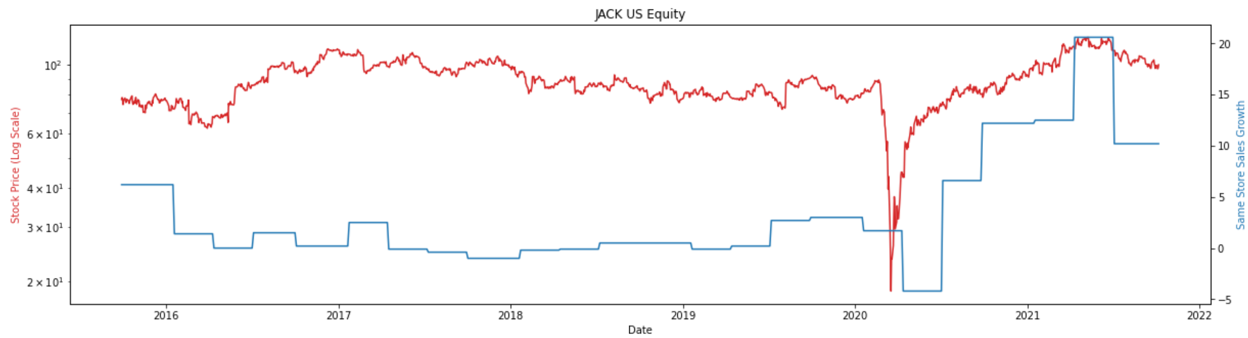
Same Store Sales Growth (SSSG)

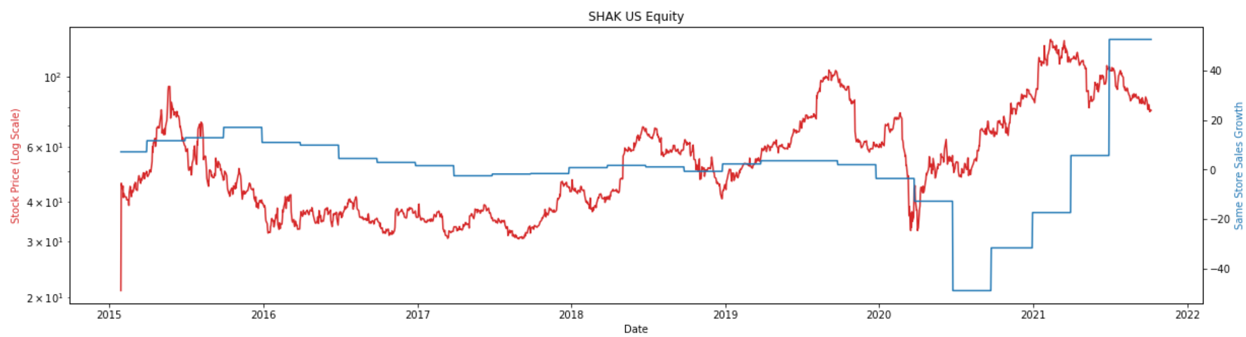
We began by plotting SSSG against the log stock price. The relationship is fairly similar between the two, lending credence to the idea that an increase in SSSG should drive an increase in stock price.

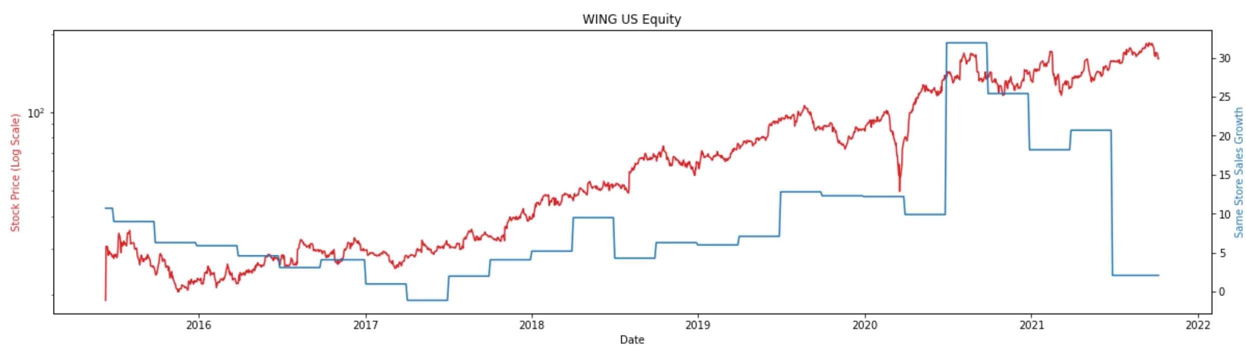
Charts 1-5: Log Stock Price vs SSSG











Data Sources & Transformations

We looked at four datasets to predict the same-store sales growth. The principal big data source is the foot traffic from SafeGraph. We tried to supplement it with keyword popularity from Google Trends, retail sales from the Census Bureau, and credit card spending from Affinity Solutions.

Foot Traffic

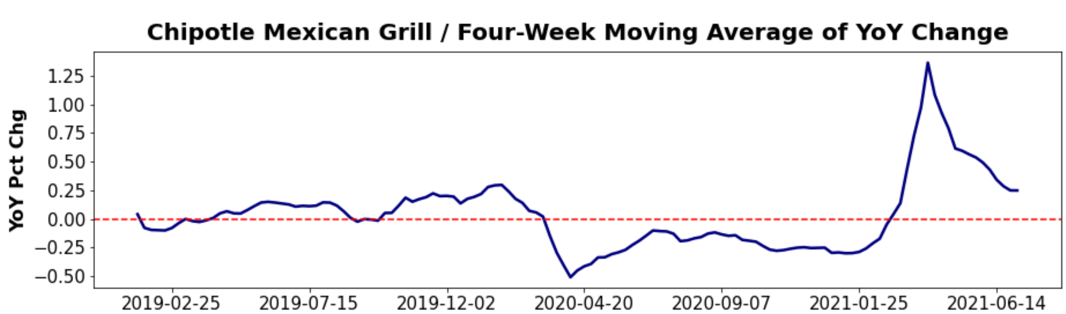
Safegraph's Weekly Patterns dataset provides foot traffic data for points of interest in the United States by tracking a panel of mobile devices. Safegraph publishes the numbers from Monday to Sunday on the following Wednesday, and the history extends to January 1, 2018. The fields include the place key, location, brand, dates, raw visit counts, and raw visitor counts **[1]**.

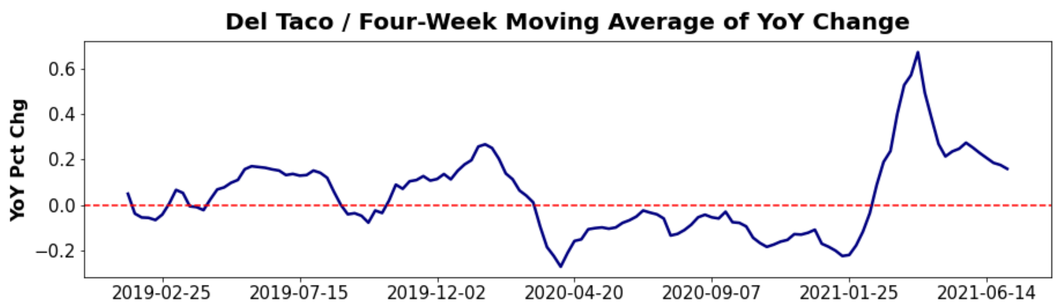
We extracted the entire history for points of interest linked to five restaurant chains: Chipotle Mexican Grill, Jack in the Box, Wingstop, Del Taco, and Shake Shack. Our primary goal was to estimate foot traffic by summing the raw visit counts across points of interest for each brand. However, the dataset has a few challenges. First, the panel changes. The number of devices tracked by SafeGraph trends upward with periods of increases and decreases. Second, the raw visit counts have outliers with abnormally low or high readings. Finally, restaurants open and close, changing the set of locations by brand. Therefore, we cleaned the data using the following steps:

- We normalized the data by applying weekly state multipliers to the raw visit counts. We calculated the multipliers by dividing the population by the number of devices tracked in the region during the period. The aim is to over-sample to compensate for increases or decreases and underrepresentation in the panel.
- We removed all places with at least one reading falling outside of 1.5 times the interquartile range. This eliminated 20% of restaurants on average for the five brands.
- We created a constant pool of locations for each chain by keeping only restaurants that were present from January 1, 2018 to August 31, 2021. This eliminated the impact of openings and closings on the total raw visits count. It was necessary if we wanted the changes in foot traffic to approximate the changes in same-store sales. After this step, forty-five to sixty-five percent of the places remain depending on the brand.

Following the cleaning steps, we summed the raw visit counts by brand for each week. Here are some graphs showing the 4-week moving average of the year-over-year change in weekly foot traffic for two of the five brands:

Chart: Evolution of Growth in Clean Visit Counts for Chipotle and Del Taco





We aggregated the weekly data by quarter to model the perfect signal because the companies release same-store sales growth quarterly. We calculated the year-over-year changes in the quarterly comparable foot traffic numbers to account for seasonality.

<u>Google Trends</u>

Google Trends' weekly indices measure the popularity of keywords in Google searches by geography using a sample of the daily queries. The weekly data preceding the last 36 hours is available going back to 2004. Google normalizes the data by adjusting for volume and rescaling values to a range of 0 to 100 **[2]**.

We hoped Google Trends would provide a measure of deliveries which contribute to sales without generating foot traffic. We extracted the weekly data in the United States for the five companies. Then, we averaged the values for each quarter, and we calculated the year-over-year change.

<u>Retail Sales</u>

The Census Bureau's Advance Monthly Retail Trade Report (MARTS) tallies monthly sales for retailers in the United States through surveys. It releases the estimates ten days after the month, and they are available by industry **[3]**. We used the seasonally adjusted (X-11 filter) numbers for the total sales and the food services and drinking places sales.

We expected the growth in total and food sales to provide general trends that would serve as baselines for the five chains' same-store sales growth. We aggregated the adjusted monthly estimates by quarter and computed the year-over-year change.

<u>Credit Card Spending</u>

Affinity Solutions discloses weekly credit card spending by North American Industry Classification System (NAICS) code and by Designated Market Area (DMA) **[4]**. The available dataset only went back to the end of August 2018. The brief history severely limited the usefulness of the data, and we discarded it. Credit card spending had the potential to provide a more direct indicator of same-store sales growth. If we had the data by brand for a longer history, we could supplement or replace the foot traffic data.

Fitting the Perfect Signal

We analyzed if we could model the quarterly same-store sales growth numbers using the aggregate foot traffic, keyword popularity, and the retail and food sales growths. We did not factor partial information at this stage.

We tried the year-over-year growth and one-quarter difference in year-over-year growth for the four data sources listed above. We only have nine or ten quarters for five companies because the foot traffic data starts in January 2018. This results in forty-five or fifty data points. This is a small dataset, and it warrants the use of a simple model. We opted for an adjusted autoregressive model fitted using ordinary least squares.

The retail sales were seasonally adjusted by the Census Bureau using the X-11 filter. For the rest, the year-over-year growth calculations addressed most of the seasonality issues.

<u>Model 1: Baseline Autoregressive Model</u>

We begin by modeling the autoregressive component to form a baseline before incorporating the independent variables. The model we want to fit is given by SSSGt=SSSGt-1+ where SSSG is the same-store sales growth. The table below lists the coefficient and explanatory power of the baseline model:

Table 1: Summary Table for Model 1

```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:                   SSSG   R-squared (uncentered):                   0.462
Model:                            OLS   Adj. R-squared (uncentered):              0.449
Method:                 Least Squares   F-statistic:                              37.72
Date:                Sun, 24 Oct 2021   Prob (F-statistic):                    2.09e-07
Time:                        03:37:11   Log-Likelihood:                         -177.48
No. Observations:                  45   AIC:                                      357.0
Df Residuals:                      44   BIC:                                      358.8
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
SSSG_prior     0.8166      0.133      6.142      0.000       0.549       1.085
==============================================================================
Omnibus:                       13.560   Durbin-Watson:                   1.544
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               44.495
Skew:                           0.357   Prob(JB):                     2.18e-10
Kurtosis:                       7.819   Cond. No.                         1.00
==============================================================================
```

The coefficient of the lagged value is highly statistically significant with a t-statistic of 6.142. However, the model is insufficient with an adjusted R-squared value of 0.45.

Model 2: Autoregressive Model with Foot Traffic Growth Adjustment

We now introduce our principal big data independent variable: year-over-year foot traffic growth. While the absolute value of the foot traffic growth does not match same-store sales growth, we believe that the change in foot traffic growth is a directional indicator for the change in same-store sales growth. In other words, we think the same-store sales growth should accelerate (decelerate) if the foot traffic growth speeds up (slows down) on average. Therefore, the model we want to fit is given by $SSSG_t = \beta_1 \times SSSG_{t-1} + \beta_2 \times (FTG_t - FTG_{t-1}) + \epsilon$ where SSSG is the same-store sales growth and FTG is the foot traffic growth. The table below lists the coefficients and explanatory power of the updated model:

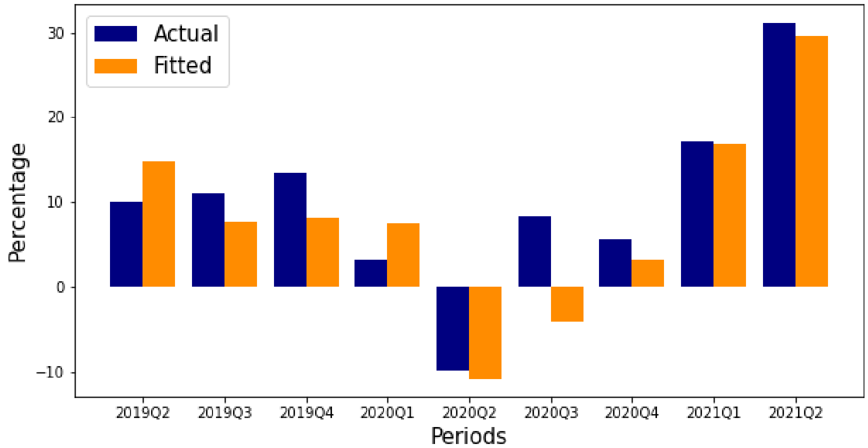
Table 2: Summary Table for Model 1

```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:                   SSSG   R-squared (uncentered):                   0.758
Model:                            OLS   Adj. R-squared (uncentered):              0.747
Method:                 Least Squares   F-statistic:                              67.44
Date:                Sun, 24 Oct 2021   Prob (F-statistic):                    5.51e-14
Time:                        03:37:12   Log-Likelihood:                         -159.46
No. Observations:                  45   AIC:                                      322.9
Df Residuals:                      43   BIC:                                      326.5
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
SSSG_prior     0.7273      0.091      7.996      0.000       0.544       0.911
FTG_diff       0.3900      0.054      7.265      0.000       0.282       0.498
==============================================================================
Omnibus:                       14.177   Durbin-Watson:                   1.424
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               36.718
Skew:                           0.585   Prob(JB):                     1.06e-08
Kurtosis:                       7.268   Cond. No.                         1.73
==============================================================================
```

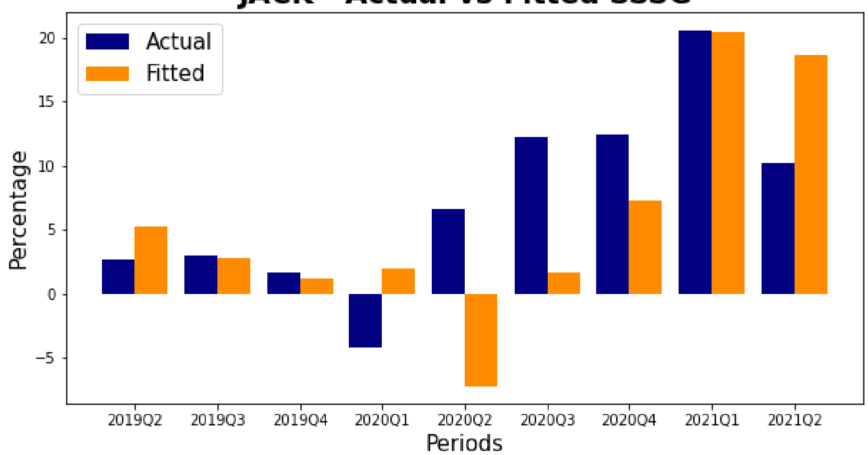
The results confirm our intuition. The coefficient of 0.39 for the differenced foot traffic growth is positive. Increasing (decreasing) foot traffic growth leads to increasing (decreasing) same-store sales growth. The contribution of the autoregressive component remains high, with a coefficient of 0.73. Both coefficients are highly statistically significant, with t-statistics of 7.996 and 7.265. Finally, the explanatory power rose to a respectable level with an adjusted R-squared of 0.747. The interpretation is clear. We can formulate a fair prediction using the previous quarter's same-store sales growth as a baseline and the change in foot traffic growth as an indicator for the evolution of fundamentals in the current quarter. We show the actual versus fitted values obtained with the second model for the five operators:

Graphs 2-7: Actual vs Fitted Values from Model 2 by Company

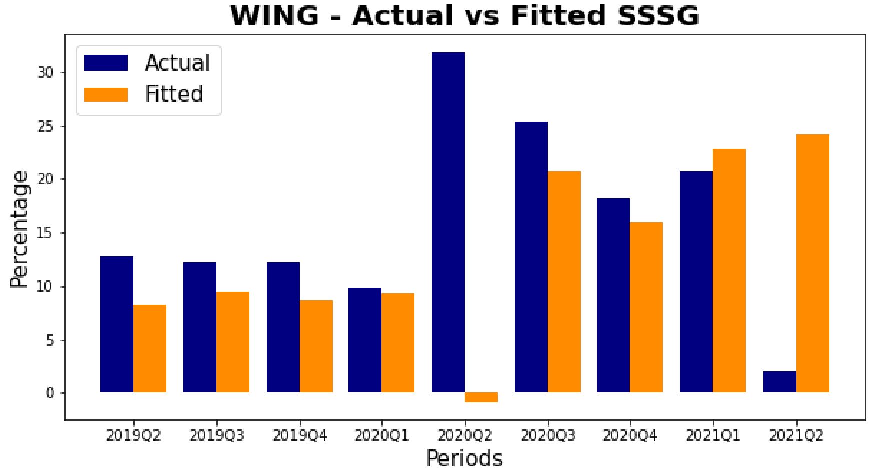
CMG - Actual vs Fitted SSSG

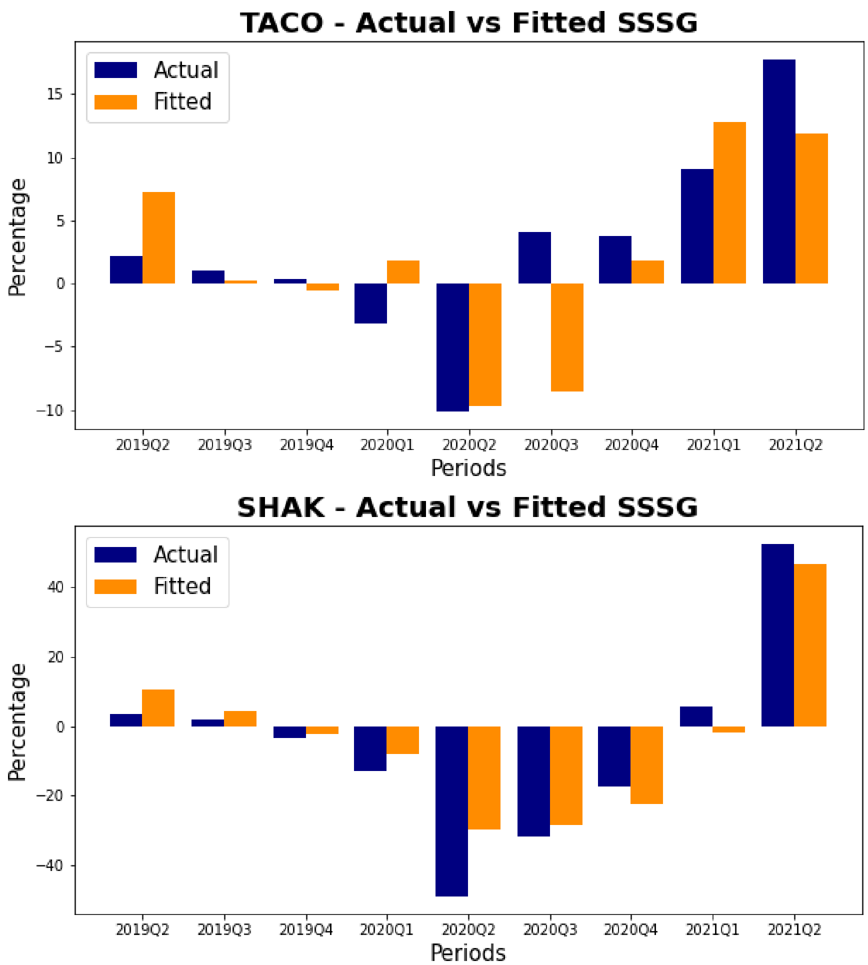


JACK - Actual vs Fitted SSSG

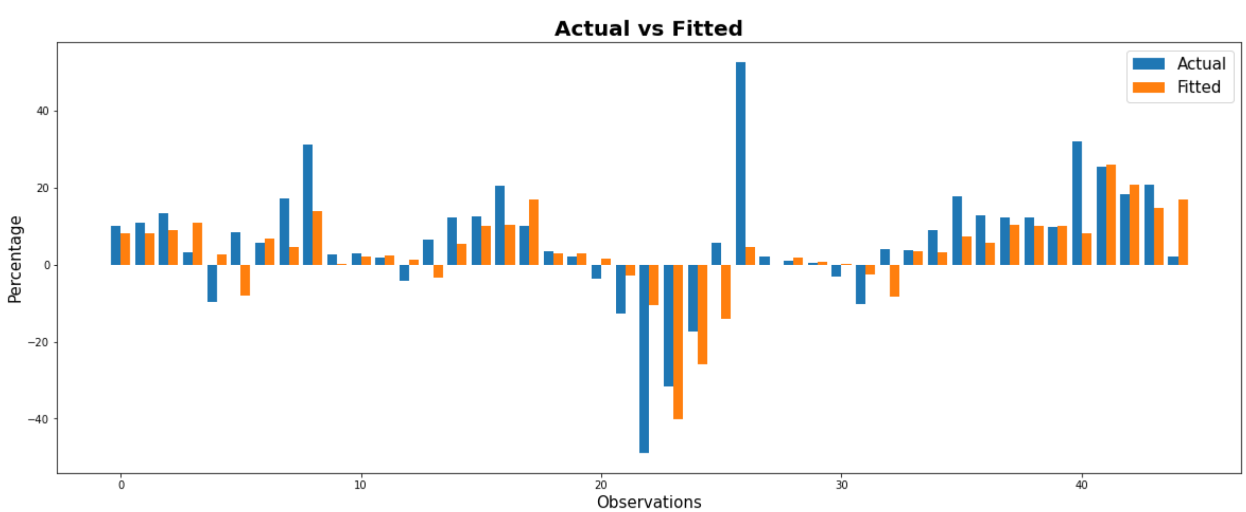


WING - Actual vs Fitted SSSG





Graph 8: Average Actual vs Fitted Values from Model 2



<u>Model 3: Autoregressive Model with Foot Traffic Growth and Other Adjustments</u>

Finally, we tried to include the supplemental independent variables into the previous model: keyword popularity indices, total retail sales, and food retail sales. We tested the growth and difference in growth for each. The coefficients were not statistically different from zero for the Google Trends indices and for the total retail sales. The coefficient for the food retail sales growth was statistically significant, but its value is hard to interpret. In this latter case, the model is given by $SSSG_t = \beta_1 \times SSSG_{t-1} + \beta_2 \times (FTG_t - FTG_{t-1}) + \beta_3 \times FRSG_t + \epsilon$ where SSSG is the same-store sales growth, FTG is the foot traffic growth, FRSG is the food retail sales growth. The table below lists the coefficient and explanatory power of model 3:

Table 3: Summary Table for Model 3

```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:                   SSSG   R-squared (uncentered):                   0.798
Model:                            OLS   Adj. R-squared (uncentered):              0.784
Method:                 Least Squares   F-statistic:                              55.40
Date:                Sun, 24 Oct 2021   Prob (F-statistic):                    1.19e-14
Time:                        03:37:13   Log-Likelihood:                         -155.39
No. Observations:                  45   AIC:                                      316.8
Df Residuals:                      42   BIC:                                      322.2
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
SSSG_prior     0.8102      0.089      9.119      0.000       0.631       0.990
FTG_diff       0.5104      0.065      7.872      0.000       0.380       0.641
FRSG          -0.1644      0.057     -2.885      0.006      -0.279      -0.049
==============================================================================
Omnibus:                       13.023   Durbin-Watson:                   1.866
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               19.345
Skew:                           0.827   Prob(JB):                     6.30e-05
Kurtosis:                       5.753   Cond. No.                         2.74
==============================================================================
```

The negative coefficient would suggest that same-store sales growth has a negative relationship with food retail sales growth in the same period. This is unlikely to be true. Meanwhile, the increase in adjusted R-squared was marginal. *We prefer model 2 because it is simpler and more interpretable with only marginally lower explanatory power.*

Potential Model Improvements

1. The process to extract weekly foot traffic data for brands with large numbers of locations is time consuming. This is the main reason we focused on five chains. However, we could improve the validity of our results by incorporating more companies to increase the size of the training set and to create a testing set.
2. We would obtain a better proxy for same-store sales growth if we could access a longer history of credit card spending data by brand.

3. We should adapt the model to run on partial information. We currently assume that we have access to the full quarter. Instead, we could use only the first two months of every quarter.

Strategy

Strategy with the Perfect Signal

Assuming that we have the perfect signal, that is we have the perfect knowledge of the full quarter's SSSG data at the beginning of the quarter for each of the fast food chains, we can implement a trading strategy based on the perfect signal and the market consensus and select the best knowledge period and holding period according to the overall performance.

Because the catalyst in this strategy is the quarterly released Same Store Sales Growth, our strategy conducts trades on a once per quarter basis. The market consensus

Strategy Description

In this strategy, we assume to have the perfect knowledge of the Same Store Sales Growth(SSSG) data a certain period of time before its release for each fast food chain. We compute the change in the current quarter's SSSG against the previous quarter's, which signifies the second derivative of Same Store Sales, and decide to long the particular stock if the change is above a predetermined threshold or to short the stock if the change is below the negative of the threshold. If the SSSG change is in between the threshold and the negative of the threshold, we do not do any trades.

We assume equal absolute weight in each stock that the perfect signal tells us to trade. For example, if the perfect signal gives 1 for three stocks, 0 for one stock, and -1 for one stock, we are supposed to long the three stocks with weight of 0.25 each, and short the last stock with a -0.25 weight.

Instrument

The investment instruments in this strategy are the 5 stocks "Chipotle"(CMG), "Jack in the Box"(JACK), "Shake Shack"(SHAK), "Del Taco"(TACO), "Wingstop"(WING). We do both long and short trades in the 5 stocks.

Catalysts

The default catalyst for this strategy is the quarterly Same Store Sales Growth data released by each fast food chain. Because the data gives the most straightforward information on the revenues generated by the existing stores, it is one of the most important price movers for the restaurant stocks and we use it as the catalyst for the strategy.

However, since other investors are also likely looking into macroeconomic data for signals, data such as the monthly Retail Sales and the weekly Credit Card Spending also indicate partially the performance of the restaurant industry and may act as catalysts that move the prices days or weeks before the release of the SSSG.

Threshold

One approach is to use a predetermined level for the change in SSSG as the threshold. For example, we can adopt 2% as the threshold and if the previous quarter's SSSG is 4% and the current quarter's SSSG is 6.5% the change is 2.5%, which is above 2%. Accordingly we should long the stock. The same logic follows for shorting if the change of SSSG is below -2%.

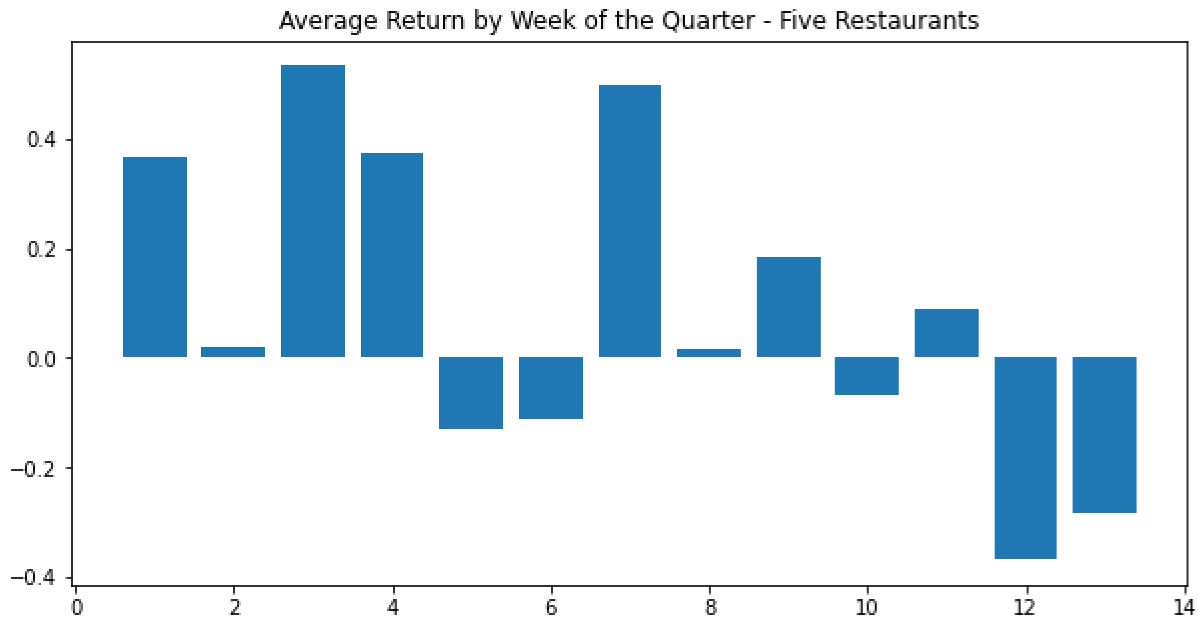
The other approach is to adopt a threshold for the standardized change in SSSG = (ΔSSSG - Mean of Historical ΔSSSG) / Standard Deviation of Historical ΔSSSG. For example, we can use 0.75 as the threshold and any trades would be desirable if standardized change in SSSG is greater than 0.75 or less than -0.75.

Timing

We test on different open dates of the trades separated by 1 week each, starting from 12 weeks before the release date. We also test on the holding period that leads to a close date a certain days before or after the release of SSSG. The combination of open and close dates are chosen based on the overall performance of the strategy compared to the S&P 1500 Restaurant Sub-industry Index.

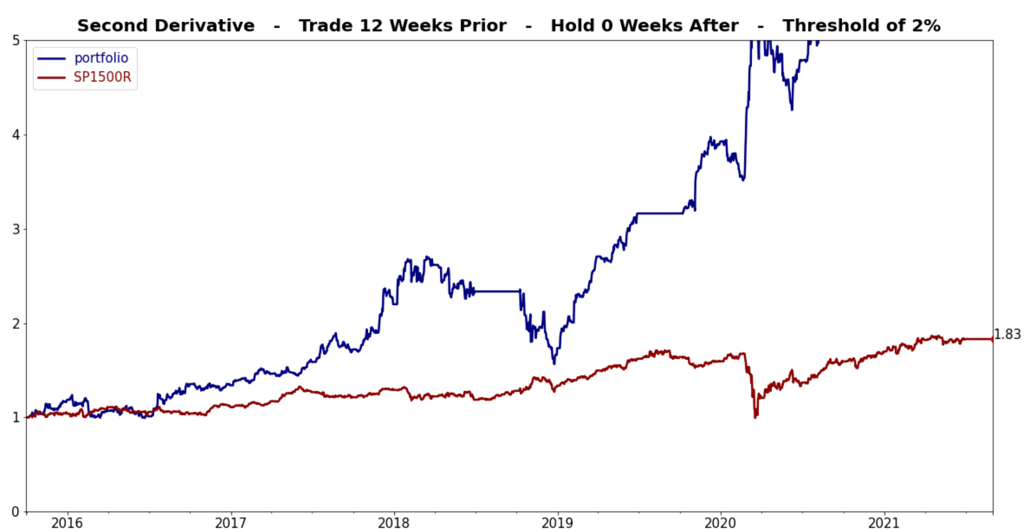
We also evaluated stock returns as the quarter progresses for each company. Here, we plot the simple weekly return[1] average between all five restaurants. The x-range is between 1 and 13 representing the 13 weeks in a calendar quarter. Large drops occur in the last two weeks of the quarter. Likewise a large drop occurs the first week of a calendar quarter. This behavior is consistent with informed trading occurring close to the quarter end.

[1] As opposed to a log return.



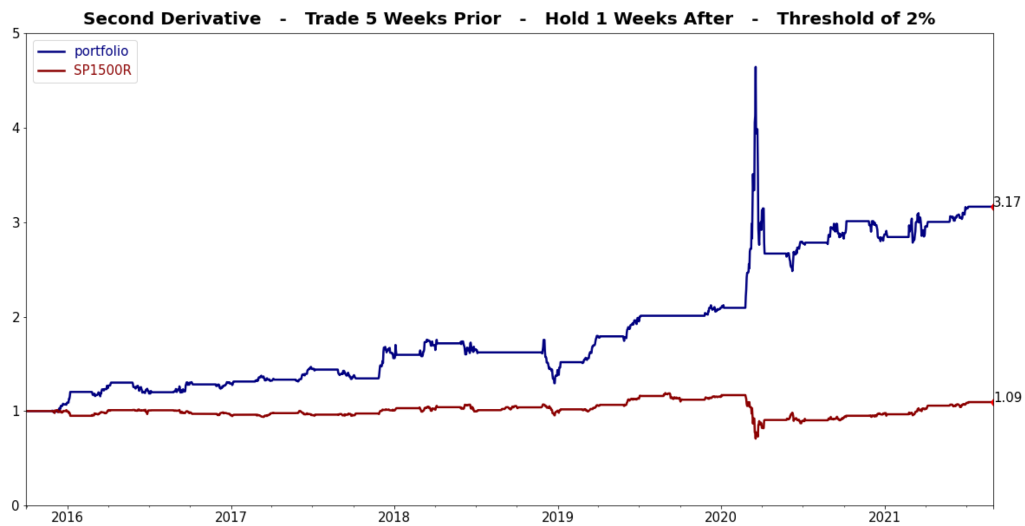
<u>Return Analysis</u>

The cumulative return of the strategy is tested on different combinations of knowledge period(the amount of time we have the perfect knowledge before the release of SSSG) and holding period(the amount of time we hold on to the trade after the release of SSSG).



Here we assume to have the knowledge of the current quarter's SSSG data right at the beginning of the quarter and hold onto the position until the end of the quarter. We use an absolute level of 2% change in Same Store Sales as the threshold. It turns out that the cumulative return ends up at 633% compared to the 183% of the SP1500R index, which clearly shows the advantage of having the prior knowledge 12 weeks in advance would render the strategy very effective. Under this scenario, the Sharpe ratio is 21.75, the Sortino ratio is 21.62, and the Max drawdown is 8.55%.

In general, the more we move the open date of the trade towards the end of the quarter, the less profit we are able to achieve out of the information advantage we have on the perfect signal. In the most extreme case where we do not have any prior knowledge of the SSSG and trade on the day the data is released and hold the position for one week, the cumulative return simply tracks the SP1500R index over time.



On the other hand, if we move the close date of the trade forward, we also observe that the earlier we close the trade the less profit we are likely to make on the perfect signal. It seems that the return has a positive correlation with the length of the holding period. In the graph below, we observe that closing the trade 8 weeks before the release date yields a significantly lower return than closing the trade 4 weeks before the release date.

Strategy with Big Data Signal

Strategy Description

So far, we have used the aggregate foot traffic, keyword popularity, and the retail and food sales growths to fit Same Store Growth, which make it possible to replace the assumed true knowledge of Perfect Signal with Big Data Signal. We define Big Data Signal with information from the alternative data that we use such as Foot Traffic, Google Trends, Retail Sales and Credit Card Spending which could provide information to fit Perfect Signal.

Just same as the Strategy with the Perfect Signal, we compute the change in the current quarter's SSSG against the previous quarter's, which signifies the second derivative of Same Store Sales, and decide to long the particular stock if the change is above a predetermined threshold or to short the stock if the change is below the negative of the threshold. And we still assume equal absolute weight in each stock that the perfect signal tells us to trade. The difference is that with the Big Data Signal, we use the fitted SSSG given by $SSSG_t = \beta_1 \times SSSG_{t-1} + \beta_2 \times (FTG_t - FTG_{t-1}) + \epsilon$in the Model 2: Autoregressive Model with Foot Traffic Growth Adjustment explained above.

Instrument

Long and short trade in "Chipotle"(CMG), "Jack in the Box"(JACK), "Shake Shack"(SHAK), "Del Taco"(TACO), "Wingstop"(WING)

Catalysts

The default catalyst for this strategy is the quarterly Same Store Sales Growth data released by each fast food chain, on the financial report, while the investors may use alternative data from monthly Retail Sales and the weekly Credit Card Spending to anticipate the performance of the restaurant industry.

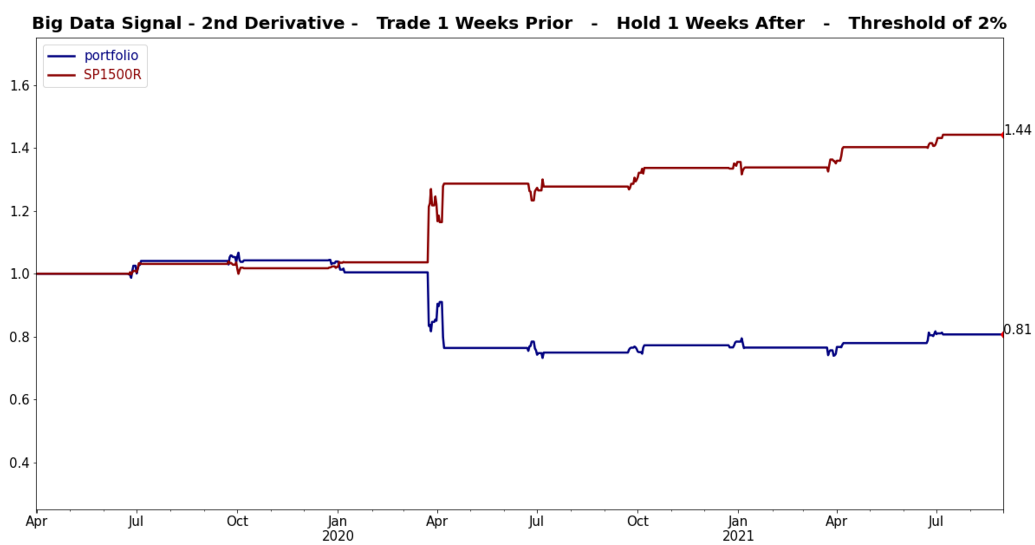
Timing

We test on different open dates of the trades separated by 1 week each, starting from 12 weeks before the release date. We also test on the holding period that leads to a close date a certain days before or after the release of SSSG.

Return Analysis

We choose different combinations of knowledge period and holding period before and after the catalyst. We find that the performance of the strategy is very sensitive to the period.

Here we assume we could use the Foot Traffic data of the whole quarter 5 weeks before the official financial report publication to fit the Same Store Sales and calculate Big Data Signal, and hold the portfolio until 1 week after the catalyst. The performance of the strategy is obviously better than the SP1500R benchmark.



When the knowledge period is reduced to 1 week, the performance of the Big Data Signal strategy is worse than the SP1500R benchmark, while the Perfect Signal Strategy is better than the benchmark. Since we try to fit the SSSG with 0.747 adjusted R-Square to generate Big Data Signal, this result is in accordance with expectation.

Contrary to the general observation in the Perfect Signal Strategy, we find that earlier knowledge doesn't definitely mean more profit. Here we assume that we got the predicted SSSG 10 weeks before the catalyst, the performance is not always better than SP1500R benchmark, especially during and after the pandemic.

With the knowledge from Perfect Signal Strategy, we don't hold the portfolio a long time after catalysts. We tested the two combinations of 11 weeks knowledge with 1 weeks holding and 12 weeks knowledge with 0 weeks holding to keep total length the same and found that though we got earlier knowledge, the performance is hardly better.

Big Data Strategy and Perfect Signal Strategy

After running the strategy using the Perfect Signal and the Big Data Signal, we found that there are differences in the performance and the big data outcome wasn't as expected:

<u>Perfect Signal Strategy</u>

Using the perfect signal we were able to outperform the benchmark using different combinations of knowledge period and holding period.We found out that the earlier we knew the prior knowledge the more effective the strategy was and the higher the profit.We also found out the earlier the trade is closed,the less effective the strategy was.
So in general, the strategy will be more effective when having an early prior knowledge and a long holding period which ends one day before the release date.

<u>Big Data Signal Strategy</u>

On the other hand using the big data signal the outcome wasn't as good as expected. Though we use regression to fit the Same Store Sales Growth with Foot Traffic Growth at an adjusted R-Square of 0.747, there is still a huge gap between the Big Data Signal Strategy backtest performance and Perfect Signal Strategy result. Extending the prior knowledge period didn't mean more profit for the portfolio. There could be several reasons for that:

1. One of the main reasons is that the history of the data we are using is too short. History Foot Traffic data from Safegraph is limited, thus enlarging the Confidence Interval for regression coefficients.
2. The Covid-19 pandemic is included in the period we are using, which could be a reason for not having a clear relationship.
3. The sample of the companies we are using is relatively small and could be exposed to idiosyncratic fluctuation.
4. If we can incorporate other signals such as credit card spending ,which can have a more precise indicator about the sales growth, the strategy will be more effective.
5. We assume that there should be a constant linear relationship between the volume of Foot Traffic and the Same Store Sales, while the correlation between them may vary with time. The reasons include temporary sales and seasonal customer preference.
6. The Foot Traffic data used in Big Data Signal cover the whole quarter, while we trade several days before the quarter-end financial report publication. That means we use some data after the day we decide to trade, casting doubt on the effectiveness of the Big Data Signal Strategy.

Final Discussion

It was valuable to differentiate between two alternative research strategies. The first strategy goes directly from data to forecasting security returns. This is labeled as the traditional “quant approach”. Another approach (taught in this class), begins with data and attempts to forecast company fundamentals such as revenue or profit margin. Then, this approach studies how the fundamentals impact security returns.

Another important takeaway is to evaluate the performance of the “perfect signal”. The perfect signal is some company fundamental that is known with 100% accuracy and ahead of time. Evaluating whether this perfect signal has predictive value is crucial before trying to predict the “perfect signal”. What use is predicting the perfect signal, if we cannot then use it to predict security returns? By adding an intermediate step (a fundamental indicator), we need to be careful to confirm causality.

For project improvement, the model could be refined as discussed earlier. We could expand the number of restaurants analized and even include retail stores. A longer credit card history would allow for this data to be included in our model. The model could also be tested on partial information; instead of using the full three months of big data, we could make a forecast after two months.

References

[1] “Weekly Patterns” *Safegraph*, https://docs.safegraph.com/docs/weekly-patterns

[2] “FAQ about Google Trends Data” *Google*, https://support.google.com/trends/answer/4365533?hl=en

[3] US Census Bureau: Rob Swartz (SSSD Division Security Coordinator), Paul Bucchioni (SSSD). “US Census Bureau Retail Trade Advance Monthly Retail Trade Survey Methodology Page.” *Advance Monthly Retail Trade Survey Methodology*, 16 Jan. 2009, https://www.census.gov/retail/marts/how_surveys_are_collected.html.

[4] “How We Do It.” *Affinity Solutions*, https://www.affinity.solutions/how-we-do-it#how-we-do-it-inside