The S&P 500 in Context: Predicting Stock Movements using Generalized Macroeconomic Variables

Yashna Gupta

Anika Bastin

Tanvi Pabbathi

May 4, 2025

Stat 4710

Table of Contents

Background

The Standard and Poor's 500, or S&P 500, is one of the best-known market indices around, having been around since 1957. It tracks the financial performance of the 500 leading U.S. publicly traded companies. It is widely regarded as an excellent measure of financial health and consumer trust in the overall economy, as well as a representative for nearly 80% of the total market capitalization of public companies in the United States. Companies on the S&P are represented with tickers and are weighted as per their market capitalization; indeed, the ten largest members of the S&P alone account for nearly 35% of the index. However, while the S&P has provided a great basis for the market, it has been extremely volatile in the last few months due to the confluence of several economic and social issues. However, there are few truly rigorous, statistically sound predictors for the direction of the S&P based on macroeconomic variables, which are potentially less endogenous to the S&P and could thus serve as better predictors.

Thus, we went into this analysis with several motivating questions.

- Can we identify how macroeconomic sectors move with and influence each other? Can we isolate sectors that correlate heavily, and what are the logical reasons for this correlation?

- Can we identify macroeconomic characteristics or sectors that heavily influence or correlate with the performance of the S&P 500? For example, can we isolate sectors that have spending that tends to influence the S&P? Are there any sectors that we can track that will allow us to make sound investment strategies?

- Can we isolate specific sectors (e.g. commodities, leisure spending) that are heavily correlated with the SPX? We would also like to be able to make comparisons between sectors to see which will result in a better model for outcome prediction.

- Can we create a model that will simply be able to predict whether the S&P increased in the next day? We typically view stocks as a random walk, with there being some p that they increase in a given instant, in a fairly Bernoulli manner. Thus, we are interested in seeing whether we can predict this seemingly random sequence with high probability.

These questions are incredibly valuable to investors and individuals in the finance industry. Especially due to recent economic uncertainties, individuals in the finance industry are in dire need of a good predictor of how stocks will move. Thus, our research has several specific implications for such investors. However, they are also very valuable for individual investors. While most people don't have access to specialized data like Bloomberg Second Measure, many news sources still report on more general macroeconomic indices, like the CNN Fear and Greed Index or general reporting on interest rates. Separate indices are very heavily correlated with the factors that we discuss, which means that there is significant importance in individual investors understanding how different macroeconomic factors are correlated with the performance of the generic SPX.

Thus, some specific questions we believe we can address are as follows.

- How should I best balance the indicators through which I watch the market to ensure that I have a good perspective on the potential behaviors of the market?

- How should I best balance my own portfolio? Is the SPX a very risky asset to be holding, and are there any sectors that are very complementary with it (move with it) or some that move against it? This question is especially helpful to building a countercyclical portfolio, to ensure that the portfolio minimizes risks in the market.

- Are stocks truly a completely random walk? Are there any indicators that are consistently reliable predictors despite the recent fluctuations? This question is especially helpful due to recent speculations about the true efficacy of portfolio managers.

Description of the Problem and Tested Solutions

The first method we used to approach this problem was PCA and k-means Clustering. We thought that this would be a good way to separate out the data to determine which variables move together, as well as what types of relationships may be present in the macroeconomic factors. We first performed PCA on a subset of indices consisting only of commodities and another subset consisting of leisure goods, which we break out in further detail later in the report. We were able to find quite clear splits in PCs for both subsets, with one PC typically consisting of more day-to-day, smaller-ticket or necessity purchases like medicine and clothing. The other PC often included larger-ticket purchases, such as Hotel purchases and catering. Thus, we can observe that similar categorical purchases do tend to move together. We then used k-means to cluster the graph against PC scores and months, which helped give us good intuition for seasonal patterns. The three clusters we elected to use could also be clearly differentiated across PCs, which indicates that the clustering seems to have picked on to the big-ticket/small-ticket spending patterns we were noticing. Additionally, it confirmed our intuition that spending seemed to pick up around holidays, confirming the existence of the well-known "Christmas Effect."

Having explored the relationship between variables in more depth, we then attempted to create a good predictive model for SPX. As we had a very large number of variables, we thought the data was very well suited to LASSO and Multiple Regression methods of modeling. Indeed, with LASSO alone, we were able to create a good prediction model, with a R^2 of around 0.95 and a relatively low AIC compared to some of our other models, including the model consisting of the backwards selection that we performed to ensure that all variables were significant. Our final model was quite a bit more parsimonious though, at 23 variables compared to the 113 variables that we started with. In this model, we found that leisure spending categories seemed to have particularly strong p values to predict model fit, as some of the most significant spend variables included specialty food stores, discount department stores, clothing, and more. However, macroeconomic indicators retained their relevance as the Discount Federal Funds Rate and General Consumer Spend tracking lines also remained significant. Thus, the LASSO model demonstrates good model parsimony and fit.

Second, we decided to answer the question of whether we could create a better model to predict stock market movement from the day before, as compared to simply assuming that stocks are a completely "random walk." For this section, we decided to use logistic regression models to approach this question. We also used boosting to see if we could create a better model. We thought this was particularly important as being able to predict relative movement of the SPX can occasionally be more helpful then being able to determine magnitude of change, especially as there is so much more uncertainty surrounding true magnitude. As is evident, the performance of our boosted model was better; indeed, our boosted model only misclassified stock movement 12 times, with a total accuracy of 94.3%, as compared to the 88.6% of Logistic Regression. Additionally, we found very different relationships between our predictor variables and actual stock movement between Logistic and Boosting techniques, as Logistic seemed to place more importance on predictors based in construction, with many manufacturing-based variables being in our Top 10. Alternatively, boosting had a much more general array of variables, with no seeming pattern to them, which could have been helpful since there might be less collinearity between these variables. It also utilized more macroeconomic indicators, like TSA and In Store, which could have been helpful to pushing up performance.

Description of the Data

The original basis for the dataset was collected from the closing price index for the S&P 500 over the last eight months. This data was obtained from the Bloomberg Terminal. One note to this analysis is that the S&P is not open to trading over the weekends and on holidays. As a result, prices tend to remain constant over those days, which may have affected the interplay between our dependent and independent variables in our final results. We then collated independent variables during these days.

- We pulled macroeconomic indicators, most notably debit card data, from the Bloomberg Second Measure Data. This data provides

insights from customer transaction data. Additionally, while Second Measure has an overall index category, it also contains debit card data broken into several different categories, such as Motor Vehicles, Food and Beverage, and more. This could allow more more specifications on what industries are key to driving stock performance. However, there are also a wealth of categories which makes this dataset ideal for analysis with dimensionality reduction tools like LASSO. We also had some intuition on important categories, like tech and food/commodities, which allowed us to perform some subset analysis on thie data.

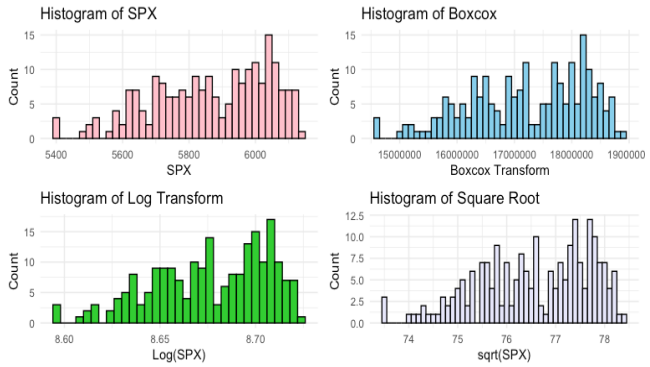
- We utilized data on other notable macroeconomic data indices, collated from various sources. One notable to this category included the number of people entering TSA checkpoints daily, which is an excellent indicator of the health of the overall leisure spending category. We also included data on job rates and important economic indices like the Federal Funds Rate, both of which were pulled from the FRED list of indices collated by the St. Louis Federal Reserve Bank. These more generic indices reflect the state of the economy, especially how "hot" it is / how much activity we can observe.

Section 1. EDA

We began by looking at some of the basics of the data, including the distribution of the S&P itself. Additionally, we broke out the data into subsets that only included information relevant to tech and commodities for use in later analysis. We started by cleaning the data, which included converting it, as it needed to be transposed before it could be used. We also removed duplicate columns, as Second Measure has the tendency to include one piece of data several times. This left us with around 113 variables, which I will explore more in depth throughout the EDA process.

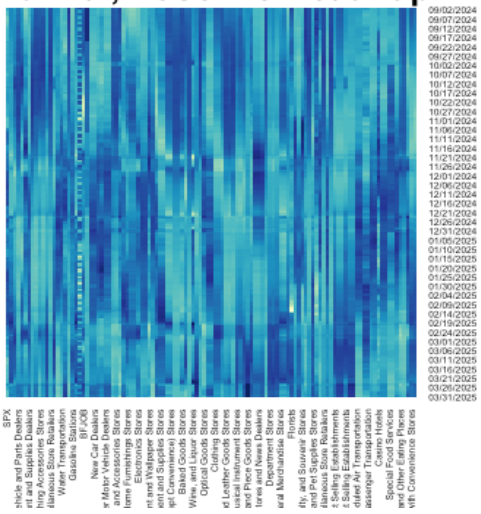
As we can see from the graph of SPX, there seems to be a somewhat rightwards skew on the distribution, with a much larger, more variable tail on this side. This could reflect the fact that the SPX has had a rather tumultuous last couple of months, which has led to there being longer periods of it deviating downwards as compared to normal, more stable economic periods where it may display more normal behavior.

Thus, we consider some other methods of making this data appear more normal. Some of the methods we attempted, which are common for addressing the type of skew that we observed, were a logarithmic transform, a square root transform, and a Boxcox Transform. Most of these seemed to somewhat retain the same type of skew that we observed from the untransformed dataset, so we elected to retain the untransformed variable to improve parsimony and simplicity of analysis conclusions. However, we have included below the ggplot histogram representations of all of the transforms to display the similarities



We also attempted to create a normalized heatmap of our data. However, as we have an extremely large number of variables, we decided to carry out a SRS on the variables to plot a representative heatmap of the data. Additionally, we included a single normalized, base heatmap for reference. This heatmap continaed the variables that were deemed most important by hierarchichal clustering. However, this heatmap is difficult to interpret due to the complexity. We can still see by the colors that many of these variables tend to move together, which makes logical sense. This could have effect on our final model, however, as this could indicate that several of our baseline independent variables may display some high amount of collinearity. We will have to check correlation between the variables, as well as LASSO choices of variables, in order to determine this. However, overall, SPX does seem to move somewhat differently from the other variables displayed here, which could mean that our predictive model will still hold value.

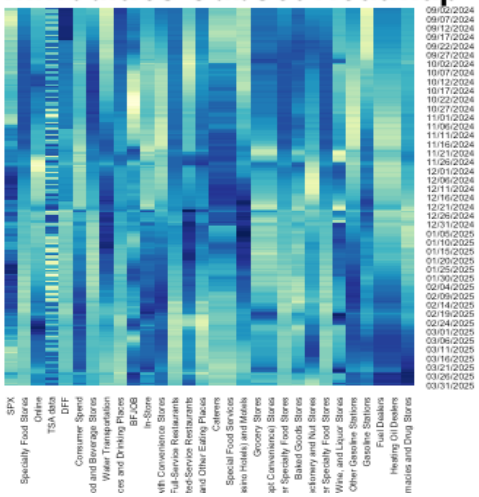
Normal, Baseline Heatmap



We were also interested in taking a subset only associated with commodities like Food, Oil, and other major necessities for life in America. We also retained some general macroeconomic indices, most notably TSA, the Federal Funds Rate, and statistics related to job seeking activity. We decided to test these as we believed that they would provide more stability than the entire set of data. Thus, we decided to perform some correlation tests, heatmaps, and EDA on this data specifically. It's also a smaller subset of variables, which we would be especially helpful for our logistic regression analysis as compared to the full dataset.

As we can see from the heatmap, this set of data still seems to retain significant variation in values as many columns of the heatmap do not move together. However, there are still some sections that could retain some potentially problematic collinearity, which we can remove during backwards selection processes. However, some of them do seem to exhibit some quite strong correlation with the SPX data.

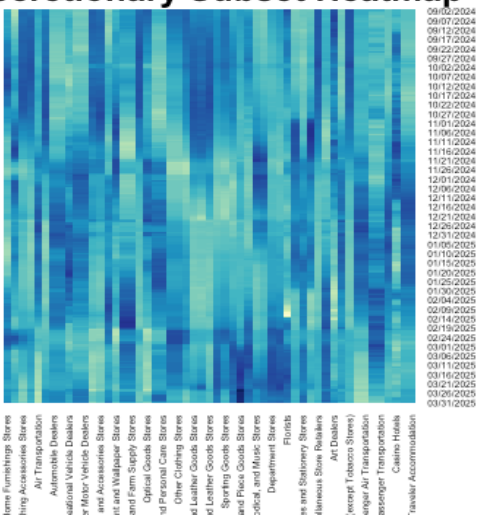
Commodities Subset Heatmap



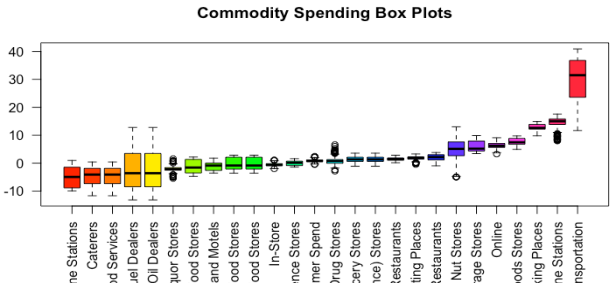
We then broke out another dataset associated solely with discretionary purchases, as we thought this could also be an interesting comparative to the commodities set that we broke out right before This subset included purchases like Furniture, Clothing, and Jewelry, which are not as necessary to daily life. As a result, this set of purchases fluctuates much more than some of our other predictors, since they are often the first types of spending to be cut during times of economic risk. We thought that up and downturns in this set of variables would be very sensitive to consumer sentiment.

We can see that this subset seems to be much more consistent in movement. This makes sense as stocks are also seen as a relatively risky, discretionary investment. Thus, downwards movement in categories like leisure spending should reasonably be correlated with downturns in investing sentiment.

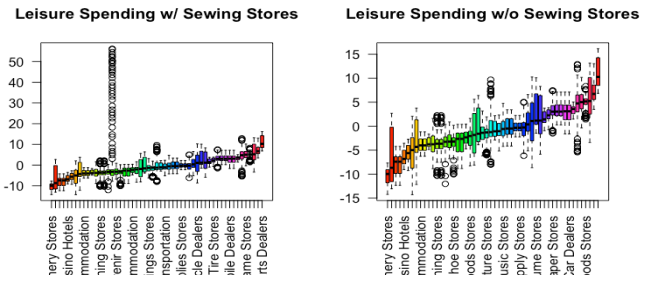
Discretionary Subset Heatmap



Having some idea about the interplay of variables, we then moved into looking more at our baselines for variables. We arranged the commodity and Leisure Spending Boxplots. From seeing these, it is evident that the Commodity Spending categories tend to move together relatively tightly. The major exception to this is the Transportation variable, which has a much higher median than the others. Even so, it has a relatively large spread, which indicates more variability in the variable overall. Additionally, we can see variables associated with transportation in general, such as gas stations, tend to have higher levels of spending. This makes logical sense as oil/transportation support does tend to be costlier than food expenses.



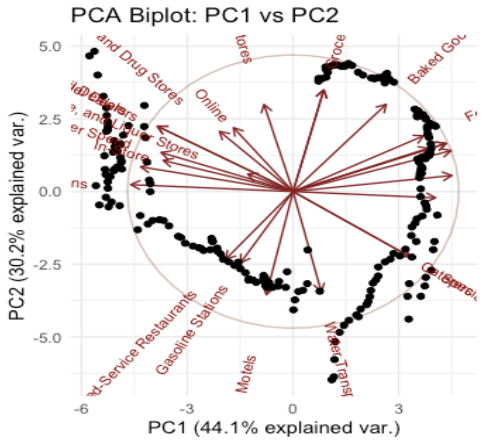
The Leisure spending boxplot also seemed somewhat similar, in terms of a relatively tight median distribution even though it had significantly more variables in its category. However, one variable notably had an immense number of outliers, which was Sewing and Clothing Goods Stores. This industry is likely relatively volatile, as crafting goods can only be pursued by people with significant amount of leisure time to learn, or those who are already very familiar with the art. However, most of the outliers were in the upwards range. Thus, there could have just been certain days with major upticks, such as holidays or weekends where people had a lot more time and could visit such stores, which resulted in the skewed distribution. We have included another boxplot without that variable to make the image more digestible.



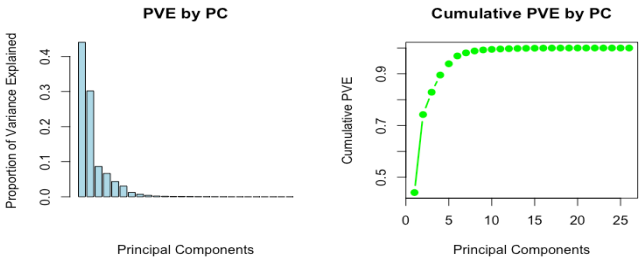
Section 2. PCA and Cluster Analysis

We then moved towards determining the groups of variables and spending that seemed to move together, as part of our first question. We utilized PCA and Clustering to determine meaningful results.

First, we performed PCA on the commodities subset of the data. This subset includes variables like food, gasoline, and pharmacies, as described before. We used these indicators as they are relatively stable with the year, as people rely on them so they do not cut spending during economic stress.



First, we looked at the PC scores and were able to confirm that they are orthogonal and independent. Then, looking at the PVE curve, we see that using two PC scores is enough as it encompasses 69.8% of the variability of the data, which we determined was a sufficient cutoff for our needs. When interpreting the different PC scores, we look at the variables that are segmented by PC1 and PC2 scores, as well as the relative directions of the biplot.



Top 10 Positive PC2 Variables: Everyday/Cyclic Spending

Variable	PosPC2
Grocery Stores	0.312
Supermarkets and Other Grocery (except Convenience) Stores	0.312
Baked Goods Stores	0.268
Confectionery and Nut Stores	0.268
Fuel Dealers	0.201
Heating Oil Dealers	0.201
Online	0.196
Pharmacies and Drug Stores	0.185
Food and Beverage Stores	0.171
Other Specialty Food Stores	0.148

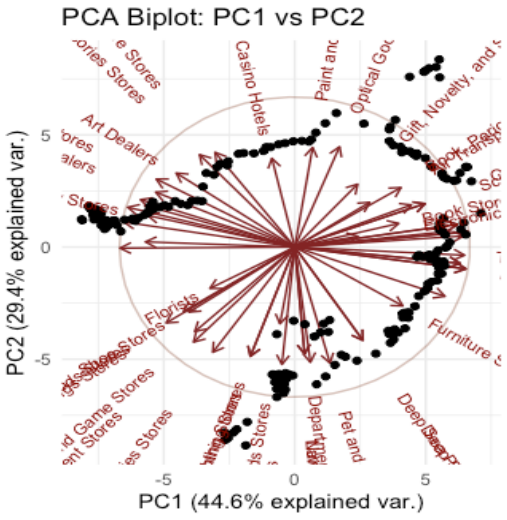
Top 10 Negative PC2 Variables: Leisure/Larger Ticket Purchases

Variable	NegPC2
Hotels (except Casino Hotels) and Motels	-0.317
Water Transportation	-0.307
Gasoline Stations	-0.219
Limited-Service Restaurants	-0.207
Caterers	-0.200
Special Food Services	-0.200
Restaurants and Other Eating Places	-0.019

We believed that PC2 lended itself to especially interpretable results, which are clearly outlined in the charts and in our analysis.

The PC2 scores tend to be negative with more leisure and luxury purchases (such as art dealers and casino hotels), whereas positive PC2 scores tend to be more everyday purchases (such as department stores, clothing stores, and pet supplies). To show this, the first ten most positive and negative variables are plotted.

Thus, through this split, we were able to make our first logical distinction on spending. Rationally, it makes sense for spending categories to correlate with categories that are similar to themselves in purchase characteristics and types. Additionally, as our data is timing based on a day-to-day basis, it makes sense that the fluctuations in types of spending correlate more heavily with similar types of purchases across category as compared to being restricted to the exact same sector as itself.



Second, we performed PCA on the Leisure subset of this data. This subset includes variables like art, gift and novelty spending, and transportation. We sectioned off these variables as they are more cyclic, as people do not need these types of spending in the same way that they need commodities. Thus, these types of spending reflect another, more cyclic variable which could be helpful as the SPX is also affected by macroeconomic movements.

Top 10 Positive PC2 Variables: Leisure/Occasion Spending

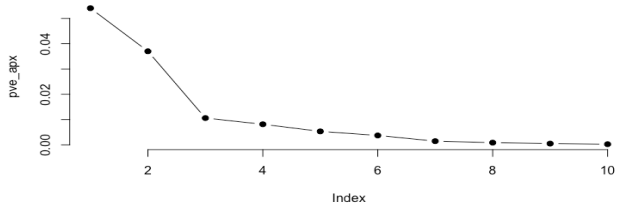
Variable	PosPC2
Optical Goods Stores	0.203
Paint and Wallpaper Stores	0.201
Automotive Parts, Accessories, and Tire Stores	0.192
Automotive Parts and Accessories Stores	0.187
Casino Hotels	0.179
Art Dealers	0.150
Luggage and Leather Goods Stores	0.138
Gift, Novelty, and Souvenir Stores	0.127
Air Transportation	0.120
Motor Vehicle and Parts Dealers	0.111

Top 10 Negative PC2 Variables: Everyday/Cyclic Spending

Variable	NegPC2
Department Stores Except Discount Department Stores	-0.237
Pet and Pet Supplies Stores	-0.235
Department Stores	-0.233
Clothing Stores	-0.221
Other Clothing Stores	-0.221
Lawn and Garden Equipment and Supplies Stores	-0.216
Nursery, Garden Center, and Farm Supply Stores	-0.216
Musical Instrument and Supplies Stores	-0.213
Sporting Goods, Hobby, and Musical Instrument Stores	-0.192

Variable	NegPC2
Deep Sea, Coastal, and Great Lakes Water Transportation	-0.189

We then wanted to see if the groups that we created were reliably determined. Thus, kmeans was used to group clusters, graphed against the PC1 and PC2 scores.



We used the elbow method on the PVE plot to determine that 3 clusters were optimal and captured a majority of the variance. Thus, we determined that it was the lowest number of clusters that would explain the data well.

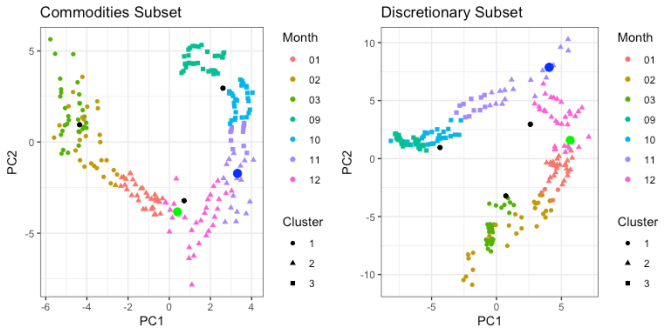
We then plot the clusters separately according to the data subsets that we broke out prior, commodities and leisure. However, we also included graph coloring by month in order to determine if we could capture any interesting seasonal effects.

As shown the PC2 scores tend to increase around November and December for the Discretionary subset and decrease for the Commodity subset. This is especially true in late November and December where the holiday season tends to ramp up and Black Friday also occurs. Travel for holidays like Thanksgiving and Christmas, holiday vacations, and gift spending would all contribute to these higher PC2 scores.

However, the fact that the scores seem relatively centered could potentially be because spending on items that we classified as daily necessary spending, like food and clothing, also increases substantially due to the holidays.

To further emphasize the potential for the Holiday Effect, Christmas is plotted in lime green, and Thanksgiving is plotted in dark blue. As shown, there are spikes in spending during these dates, likely because of increased travel during those holidays, as well as increased spending over several different categories. Indeed, the days leading up to and after Christmas and Thanksgiving are known to be the busiest travel times of the year.

As the holidays pass, the PC2 scores dip, showing that there is a decrease in occasion-dependent spending, likely as people are settling back into their lives in the New Year and there are not any major holidays coming up.



Thus, we were able to create a logically reasonable story to split our data based on the clusters and PC's, as we found that spending often split down the lines of daily, smaller items and bigger ticket purchases for both the commodities set and leisure spending subset.

On top of this, we were able to clearly demonstrate massive upticks in spending over the holiday season. These splits are extremely helpful to our understanding of the SPX may move, as it informs our understanding of how the economy moves. Traditionally, the stock market has been a good indicator of consumer sentiment; this is because people tend to pour money into it when they feel richer, as part of their other spending. Thus, when the economy is in a downwards spending cycle, spending on stocks tends to decline significantly.

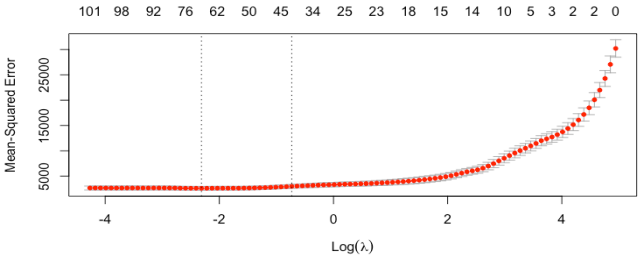
We observe that this is especially true in earlier months, as individuals seem to draw back on big ticket spending during this time, which is in line with the fact that we often experience (mildly) recessionary periods in the first few months of a new year.

Additionally, we were able to observe some potential sets of categories which could demonstrate high collinearity. This is helpful to understanding the output of the predictive models that we will be building in the next few sections, as we will be able to understand if one set of heavily colinear variables is significantly influencing the outcome of the model (as will happen with the Logistic

Regression). Being able to note this allows us to caveat the analysis that we build in the end.

Section 3. LASSO for Multiple Regression

We then proceeded with creating our first predictive model using LASSO and multiple regression, with the goal of predicting the outcomes of the SPX over the defined period. Since we have an extremely large number of features in our dataset, starting with over 100 different measures of consumer spending and general macroeconomic factors, we chose to start the analysis by using LASSO as a shrinkage method.



We elected the use the lambda with minimal standard error in this model, as we wanted to get the best possible performance in MSE for our initial model. Additionally, we noted that we would be removing variables through backwards selection to get a more parsimonious model for our second model in this section.

Above is the first curve of MSE v. Log Lambda for our Lasso Model. As is evident, a large number of variables was still retained. Thus, we only included a description of the top 10 most important variables, their values, as well as their significance in the original chart. We have excluded the intercept in order to focus on the values of the variables.

As is evident, the top ten variables were quite significant and had quite large magnitudes, in terms of their estimates. This makes sense since we were using the raw, untransformed values of the SPX indices over time. Additionally, the variables that are captured in this top 10 of the model seem to cover quite an expansive array of sectors in the economy.

Top LASSO-Selected Predictors for SPX

	Estimate	Std. Error	t value	Pr(>|t|)
Special Food Services	-29.7	6.76	-4.39	0.000
Sporting Goods Hobby Musical Instrument and Book Stores	-94.6	24.46	-3.87	0.000
Clothing and Clothing Accessories Stores	104.1	28.56	3.65	0.000
Lawn and Garden Equipment and Supplies Stores	-100.3	34.68	-2.89	0.004
Confectionery and Nut Stores	-55.6	19.45	-2.86	0.005
Other General Merchandise Stores	322.1	114.50	2.81	0.006
Luggage and Leather Goods Stores	56.1	20.09	2.79	0.006
Building Material and Garden Equipment	65.4	23.60	2.77	0.006
Restaurants and Other Eating Places	-688.3	259.74	-2.65	0.009
DFF	-176.3	66.70	-2.64	0.009

Indeed, based on the linear regression that we then performed, we see that our R^2 value is quite high (meaning our model was good at predicting the data) at 0.95 using just the variables that were selected with LASSO. However, the model is still extremely complex, compounded by the fact that we used lambda.min rather than the more parsimonious lambda.1se. Thus, we decided to create a second, simpler model that we hoped would yield clearer relationships.

Second, we use backwards selection to only keep the most significant variables for our model, after starting with the larger model that we built through using LASSO. This helps with parsimony.

However, this also helps build a more robust model. As we were able to observe that most of the features that were selected by LASSO, though lending the lowest MSE, were still statistically insignificant and thus did not reliably have an observable effect on the SPX.

Thus, we then elected to reduce the number of features we are modelling on following LASSO. This way, we are only keeping the most significant features for our model. The list to the right of this text shows the most significant values that we retain after backwards selection.

After performing backwards selection, some features that remained significant stayed somewhat similar to the previous list that we laid out through purely LASSO. However, there were some significant changes in magnitude of variable estimates, which is especially evident when directly comparing the tables that we include below for backwards selection and the table that was purely for all variables selected by LASSO.

While magnitudes changed significantly, degree of significance also adjusted for many of the predictive variables. For example, general merchandise stores became one of the most significant variables in our model. Some macroeconomic variables also gained more importance, such as consumer spend.

Thus, this set of variables seems to be somewhat better rounded out, since we have two macroeconomic predictors in the final mix, as well as a good mix of other, individual spending sectors in the top ten variable category.

Top Predicted Following Backwards Selection

	Estimate	Std. Error	t value	Pr(>|t|)
All Other General Merchandise Stores	206.2	21.90	9.42	0.000
Other Specialty Food Stores	-835.1	151.44	-5.51	0.000
Consumer Spend	288.7	54.67	5.28	0.000
Discount Department Stores	-221.0	44.76	-4.94	0.000
All.Other Specialty Food Stores	684.6	146.38	4.68	0.000
Sporting Goods Hobby Musical.Instrument and Book Stores	-65.1	15.77	-4.12	0.000
Hobby Toy and Game Stores	-53.5	13.70	-3.90	0.000
Furniture.Stores	-54.5	14.92	-3.65	0.000
DFF	-264.3	74.77	-3.53	0.001
Clothing and Clothing.AccessoriesStores	33.1	9.66	3.43	0.001

From the summary of the final model following backwards selection, we see that variables such as "Other Specialty Food Stores" have a large negative coefficient, meaning they are inversely proportional to SPX. This is quite an interesting relationship to observe, as specialty food spending doesn't seem to have much of a relationship to the financial market; however, it proves to be reliably significant in both this model and the total LASSO model. Thus, we

could view it in the lens of discretionary spending, where positive kickups in this type of spending could simply indicate less available capital to put in the stock market. However, "All Other Specialty Food Stores" is decidedly positive, which could indicate some collinearity between those two variables, where one is simply trying to tamp down the effects of the other. Indeed, we determined these two variables move together in our PCA and kmeans clustering section. Thus, the effects of the food store variables could be somewhat overstated.

Variables such as "Consumer Spend" have very large positive coefficients, meaning they are proportional to an increased SPX. This makes logical sense as we stated prior that the model was positively related to macroeconomic effects. Thus, we expect to be somewhat significant increases in the SPX during times of increased consumer spending

From our comparison, we actually see that the LASSO fitted model has a lower AIC, at 2191, than its counterpart after backwards selection, which has an AIC of 2378. This rather significant difference indicates that the LASSO model seems to have a better tradeoff between model fit and complexity. Thus, the first model is somewhat preference in terms of actual performance, which aligns with the fact that it had the lambda with the lowest MSE. However, the second model is much more robust and interpretable.

Thus, in the end, we were able to create with a well performing model with a high R^2. Additionally, we were able to reasonably interpret the results of the variables, which helped give us more insight into the logic of the SPX.

Section 4. Logistic Regression

We then sought to tackle the next question that we had laid out in our introduction. Namely, we wanted to see if we could create a reliable way to determine whether the SPX would move up or down in a given day. This is especially important as, often, it's hard to predict the actual magnitude of movements and we simply care more about whether the market is going to move up or down (as is more reliable and is enough to determine whether we will make profit or not). This is also a good way to approach whether we can guess the direction of the random walk of the stock market after a day of seemingly random movement.

Top Predicted Following Backwards Selection

	Estimate	Std. Error	z value	Pr(>|z|)
Manufactured Mobile Home Dealers	-1.72	0.628	-2.73	0.006
Hardware Stores	217.79	81.147	2.68	0.007
Pharmacies and Drug Stores	38.59	14.474	2.67	0.008
Building Materials and Supplies Dealers	-1159.13	436.447	-2.66	0.008
Home Centers	812.17	307.755	2.64	0.008
Other Building Material Dealers	81.06	31.043	2.61	0.009
Shoe Stores	31.35	12.092	2.59	0.010
All Other Miscellaneous Store Retailers	-953.81	372.375	-2.56	0.010
All Other Miscellaneous Store Retailers except Tobacco Stores.	450.21	184.813	2.44	0.015
Tobacco Stores	143.07	59.608	2.40	0.016

Our logistic regression has an AIC of 325. Additionally, we evaluate the performance of the model, such as ROC curve and accuracy, more directly with statistics in Section 4, where we are able to directly compare performance with the boosted model.

Overall, though, the model performs quite well, with an accuracy of around 88.6%. Additionally, we can seem to see some patterns in the regression, as it seems that there is much weight given to manufacturing related spending, as they make up 5 of the top 10 variables by significance in the regression. This focus on one industry paints an interesting picture, as we do not tend to think of manufacturing as extremely decisive towards the direction of the stock market—tech industries usually take that place. However, manufacturing has long been a pillar of the US economy and often serves as a spending multiplier in poor economic times. Thus, this nuance could be something the model picked up on.

This could also be an issue though, as we discussed in the PCA and kmeans section, these industries are quite heavily correlated with each other and thus display high collinearity. Thus, we may have a reason to worry if a majority of our predictors are concentrated in this one sector. Namely, we worry that there is not enough diversity in the sector of markets reflected in the logistic predictor. Indeed, as we will see by the confusion matrix that we discuss more in Section 5, it tends to misclassify at a much higher rate than the boosted model does.

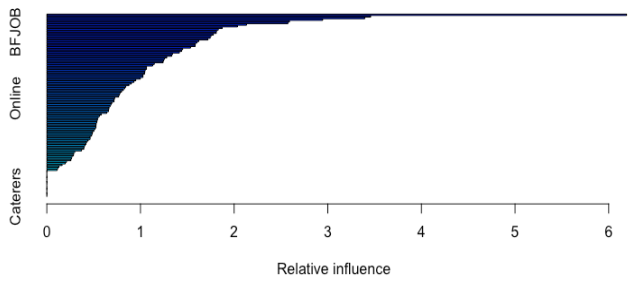
Indeed, as we discuss later, the boosted model, which covered a more diverse array of variables, did end up performing better in the end. Thus, this focus on one industry could have potentially been detractive to the final performance of the Logistic Regression. We delve more into this comparative in Section 4, however.

Section 5. Boosting

Second, as discussed, we wanted to see if we could create a better model to predict upwards and downwards stock movement, since the performance of the logistic model was good, but we believed that it could use some improvements.

Thus, we utilized the boosted model in order to hopefully create a more robust model. In this model, we used the Gradient Boosting Machine, or the gbm package in R. Additionally, we used cross validation with 5 folds in order to validate our results. This gave us a much more stable model, which also included a greater diversity of variables which is evident through the list of top predicted variables from boosting.

Additionally, we have plotted below the relative influence chart. This chart has helped rank the variables by their contribution to the model's final performance, which we will discuss more later in the section.



Top Predicted Following Boosting

var	rel.inf
TSA data	6.19
Tire Dealers	3.46
Painting and Wallpaper.Stores	3.40
Recreational Vehicle Dealers	2.95
Beer Wine and Liquor Stores	2.59
Deep Sea Coastal and Great Lakes Water Transportation	2.57
Used Merchandise Stores	2.13
In Store	2.04
Book Stores	1.87
Accommodation	1.83

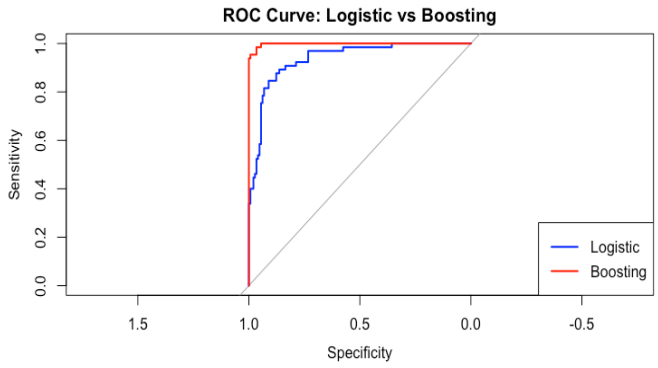
The variable importance graph highlights the most influential predictors in our Boosting model by measuring each variable's relative contribution to reducing classification error. Variables with higher importance scores had a greater impact on the model's ability to correctly predict whether the S&P 500 would increase on a given day.

Additionally, we can analyze the variables and their relative influence. The boosting model has a much larger spread of influential variables, which cover every industry from food and beverage to automotives. Additionally, we cover lots of influential macroeconomic indicators, most notably the In Store indicator and the TSA data indicators.

Thus, while this model has less of a story focused in one sector than some of the other models we have created, this is actually beneficial as we would like for our model to not be heavily weighted towards any one sector of the economy.

We then proceeded to compare the performance of the Boosted model and the Logistic Regression to determine which yielded a better final model in the end.

First, as is evident by the ROC curve, the boosted model has much better trade-offs between sensitivity and specificity pairs. Additionally, it clearly has a larger AUC which indicates better model performance. Interestingly though, both of the models seem to have somewhat high, constant performance on specificity, especially the boosted model. We are able to analyze this more in-depth when looking at the confusion matrices.



Model	Prediction	Reference	Freq
Logistic Regression	0	0	138
Logistic Regression	1	0	8
Logistic Regression	0	1	16
Logistic Regression	1	1	49
Boosting	0	0	146
Boosting	1	0	0
Boosting	0	1	12
Boosting	1	1	53

To evaluate the classification performance of our models, we compared logistic regression and gradient boosting using confusion matrices and standard metrics. The boosting model outperformed logistic regression across the board, achieving a higher overall accuracy (94.3% vs. 88.6%) and better-balanced accuracy (0.908 vs. 0.850). Notably, boosting achieved perfect specificity, correctly identifying all instances where the S&P 500 did not increase, and demonstrated stronger sensitivity (0.815 vs. 0.754), indicating more accurate detection of positive market movements. The model also exhibited a perfect positive predictive value, meaning every predicted increase in the S&P 500 was correct. The results highlight that boosting overall was a better model at predicting whether SPX would increase.

Indeed, as we discussed prior, boosting included a wider array of significant predictive variables, in terms of sector. Thus, this confirms our hypothesis that looking at the economy as a whole is necessary to understand how the SPX will move. Still, the Logistic Regression helped us understand one particularly valuable set of predictors, namely variables relating to the manufacturing sector, as honing in on that sector performed only moderately worse than allowing for a wider, more complex variety in our predictors.

Conclusion:

k-means clustering and PCA revealed that macroeconomic indicators for both commodity and discretionary spending were increased around the holidays. In the period studied, September 2024 to March 2025, there was a peak around November and December. These peaks are displayed on the graphs in that section 2.

In LASSO and backwards selection, general merchandise stores were seen to be a significant macroeconomic indicator for the SPX. Specialty food stores are negatively correlated with SPX, showing that a high value for general merchandise stores and a lower value for specialty food stores was indicative of a higher SPX value. Other variables were also identified as important. These trends are interesting as they are not indicators that would be usually recognized as important trends for the financial market, however our analysis proved they might be.

Another methods used to analyze this dataset was logistic regression. This model had an accuracy of 88.6% predicting the SPX value. The top variables

identified were manufactured mobile home dealers, hardware stores, and pharmacies. This model was then boosted to see if the model could be more accurate. With boosting, the model was 94.3%, and showed TSA data and tire dealers were the best predictors for the SPX data.

In the future, we would like to continue exploring the performance of this model over the next couple of months to see if it continues to retain its significance in the future. Additionally, we would like to consider using alternative models, such as different Boosting Techniques, to see if we could build out a better fit for our logistic regression.

Additionally, we would like to try breaking out our model to different tickers in the S&P to see if we could potentially create an optimal investing algorithm.

Use of Large Language Models:

For our project, we used large language models (LLMs) to aid in understanding the material. We focused on applying the concepts we learned in class to our project, particularly by using functions and processes that we've examined throughout the semester. Our use of LLMs was purely to understand the material at a deeper level and connect it to our models and results. Occasionally, LLM's were used to help debug code as well.