Evaluating the Stock Market on Externalities using Machine Learning

The Intersection between the Stock Market and Societal Conditions

Na’Sir Miller
Student
Rensselaer Polytechnic Institute
Troy, NY, USA
miller.j.nasir@gmail.com

ABSTRACT

In this project, the goal is to answer multiple questions centered around the stock market. The first question is, “Do external factors affect the stock market, if so, how much is performance affected”? The second is, “Are stock prices and performance indicators of the environment?”

To derive these insights, the datasets used were global stock indices (GSI), i.e. the S&P 500, core consumer price index (CCPI), environmental consumer price index (ECPI), environmental policy stringency index (EPSI), and global supply chain pressure index (GSCPI).

Before transforming the data, EDA was performed to potentially contextualize missing values and other future values that can be pivotal in answering and supporting the solution.

After wrangling the data, I utilized multiple regression-based models to extrapolate important missing values and answers to the above problem statement.

CCS CONCEPTS

•Stock Market + Machine Learning •Socioeconomics • Stock Market - Environment relationship

1 Data Description Initial Processing (1)

For GSIs, the dates are the data frame index, with float prices for open, high, low, and close prices. Furthermore, the price change is a percent change that is a representation of ((close - open) / open) * 100, in turn giving all indices, despite having different currencies, a universal measure. The percent change will be compared to a benchmark for each index, and if it is below or above the said benchmark, the model takes note of such and states whether or not the market was affected. The day-of-week feature will continue to give me the option to fine-tune the model if accuracy–among other things– is poor, while the month feature will provide a “grouping” feature to provide the potential for aggregating and viewing values in a broader sense.

For the CCPI, ECPI, EPSI, and GSCPI all of their features are as follows: Date (YYY-mm-dd), Country, and their respective measurement.
After cleaning, every dataset was joined by their common features.

2 Exploratory Data Analysis

The percent change for a month gives powerful insight into what’s happening in the market for a period, whilst being more stable than the volatility of a single day. The following is a prime example,

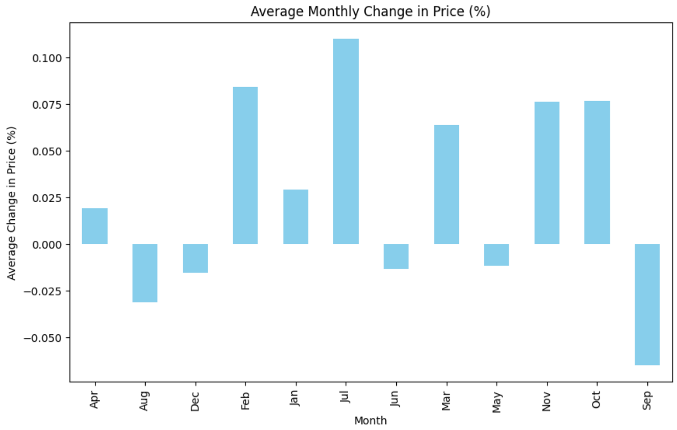
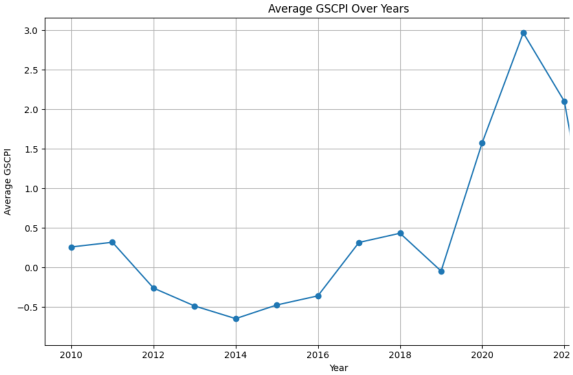


Figure 1: **Average Monthly Change in Price of S&P500**

Here we can see that February and July had the largest price changes, more than likely positive, meanwhile, September performed quite roughly with a negative difference. This tells us that February, July, and September had some events occur which is true. February was a continuation of the Ukrainian-Russian war, in turn raising crude oil prices, July had some pre-cursors to the Israeli-Palestinian turmoil, and September is historically considered a bad month due to the lack of activity on the market. Our descriptive statistics also provide some valuable insight. For instance, the mean price change is around .03, so this can be considered our benchmark percent change for a given month.

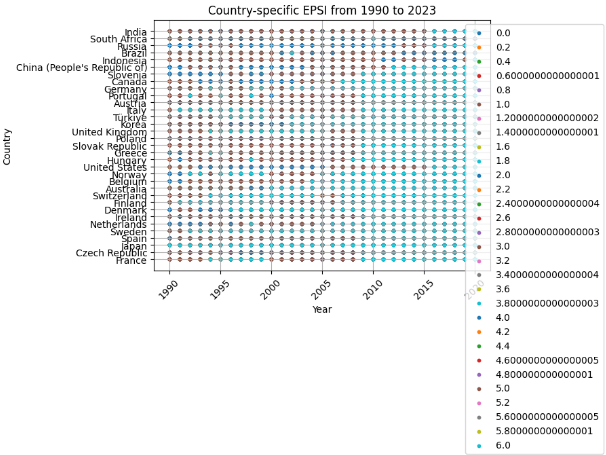
As we know, COVID-19 slowly broke out in late 2019 and ran rampant from 2020 to 2021. Our visualization supports the prior claims and can evaluate the supply chain amid the Israeli-Hamas war when the index is next updated. Overall, the long pressure period of mid-2019 and mid-2021 is a period focused on by the model, since it indicates a socioeconomic event.

Figure 2: Average GSCPI over 2010 to 2022



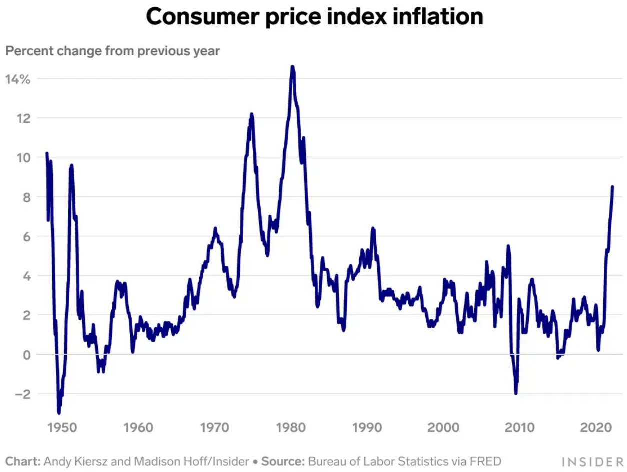
Unsurprisingly, the EPSI, also confirms what we already know about the increase in emphasis on environmental policy. Again, this allowed for the model to be easily evaluated, as well as reducing its potential complexity. I believe it will be interesting to see the relationship between stock prices as the EPSI increases since it will truly define whether or not we are making strides to improve the environment.

Figure 3: Country-specific EPSI from 1990 to 2023



As previously stated, for each country in the last three or so years, the EPSI is approaching strict environmental policy (6.0 being the most strict possible). In this period, we may see massive spikes in performance in E.V., sustainability securities like Lucid, Tesla, and Rivian, and growing companies like Tangible.

The CPIs and ECPIs found potentially tell us why indices are struggling and prove the claims that inflation is hindering the said market.


Consumer price index inflation

Percent change from previous year

Chart: Andy Kiersz and Madison Hoff/Insider • **Source:** Bureau of Labor Statistics via FRED INSIDER

Figure 4: Consumer Price Index Inflation from 1950 to 2020 (INSIDER)

From the transformed data, it can be seen that the U.S. evolved from a 91.5 CPI to a 126.1 CPI. Additionally, the only other countries with such significant leaps similar to the U.S. are based in South America and Africa– recognized as incredibly unstable economies (i.e. Venezuela). Moreover, the ECPI also displays astounding performance in socially renowned sectors and companies like E.Vs, NEP, NIO, etc. Again, over the past decade, this index has reached its peak of over 70 higher than the prior maxima, emphasizing the focus on purchasing environmentally sound goods at all levels of society.

3 Modeling

There were many important NaN values, i.e. EPSI values in 2023 or CCPI values in 2010 that linear regression was used to predict. The problem that occurred stemmed from the 2021 to 2023 data points that had NaN EPSI values, as well as select 2010 CCPI and ECPI. As a result, the CCPI and ECPI values with their respective means temporarily. After this, the EPSI was predicted. Then, they were used to help predict the actual NaNs in CPI; eventually, the CCPI was used to assist in predicting EPSI. Lastly, the ECPI and CPI re-predicted the ECPI. Again, this was a method to make sure that the missing CCPI and ECPI values were properly being predicted to remain as representative as possible of its time and country. Remember, EPSI is a measure of how strict a country is on environmental policy, so in the last three years, those values are much different than those in 2010.

4 Solutions

Instead of building a model, I analyzed the features of my primary dataset (market and external indices) and answered the problem aforementioned statements. In regards to problem one, the stock market–to no surprise–is affected by external factors with the daily change in price being at least six times worse than the average. For instance, from January 1st, 2020 to March 31st, 2023 (end of the dataset), the average daily return was about a 30 percent loss (-27.68%), and from February 24, 2022, to March 31st, 2023, the return was a 20 percent loss. These two time periods are COVID and the Russia-Ukraine War, respectively. However, times before and after these events boasted at worse a loss of 3 percent daily otherwise a gain of about 2 percent was imminent. Furthermore, these changes in prices, coupled with the GSCPI in these same periods, displayed that the stock market is an indicator of socioeconomic conditions. On average the GSCPI in “regular” market conditions is .20, with lows of .06 and highs of .33. On the contrary, the GSCPI rose on average 8 times its non-event amounts; the GSCPI in COVID was almost 27 times larger than its usual as well.

Drawbacks

Despite the insightful answers, this project has some noticeable drawbacks. First of all, non-market data –EPSI, ECPI, CCPI– had no data after 2022 due to the inability of the data to be properly joined with GSIs. Therefore, the Israel-Palenstine war was not documented in this project, despite its potential to solidify or reject findings. In addition to the loss of data in the merge of the GSIs and external indices, was the concerning performance of the linear regression model predicting the NaN EPSI, CCPI, and ECPI values. The model possessed a mean squared error of .1 and a correlation of .32. Furthermore, my approach was unable to implement advanced models to output whether or not any time given in the market was undergoing a socioeconomic event due to the lack of conformity between the features. Sadly, PCA and other feature engineering techniques were not factors due to the necessity of every feature to contextualize the market.

ACKNOWLEDGMENTS

Special thanks to Professor Lydia Manikonda, without her expertise, my growth throughout this project would not have been possible.

REFERENCES

[1] Na’Sir J. Miller, 2023. Project Github: https://github.com/NaSirMiller/introToMLapps

[2] Ben Winck and Maddison Hoff, 2022. Inflation Surged to its fastest rate since 1981 in March as food and energy prices soared. Figure 4.