The Effect of COVID-19 on our Economy

Williams Chen, chenw23@rpi.edu

Heman Kolla, kollah@rpi.edu

Zachary Niles Peretz, nilesz@rpi.edu

KC Shashwot, kcs@rpi.edu

Rijul Verma, vermar@rpi.edu

CSCI/ITWS 4350/6350, Data Science

Professor Eleish

December 8th, 2024

1a: Abstract

The COVID-19 pandemic profoundly influenced the economy across the United States, with significant disparities across different industries. This study investigates the impact of COVID-19 case trends on the performance of major exchange traded funds, focusing on two hypotheses: (1) The healthcare sector was the most profitable industry during the observed period due to increased demand for medical services and products; and (2) web-based businesses experienced significant growth and thrived during the observed period, while non-web-based businesses faced challenges and struggled to maintain profitability.

To test these hypotheses, we constructed three models: Linear Regression, Logistic Regression, and a Support Vector Machine (SVM). The linear regression model served as a baseline to compare against more complex approaches, despite its limitations in capturing nonlinear time-series dynamics. Logistic regression, the primary model, focused on binary classification, predicting whether stock prices in different sectors increased or decreased in response to fluctuations in COVID-19 case counts. The SVM model was included as an exploratory tool, and was not relevant to any conclusions that we drew.

These models supported the first hypothesis, indicating that the healthcare sector was the most profitable industry. However, our models proved inconclusive on the second hypothesis.

1b: Introduction

The COVID-19 pandemic brought unprecedented disruptions to global healthcare systems and economies. As governments implemented lockdowns, social distancing measures, and travel restrictions to curb the spread of the virus, businesses across the world were forced to adapt to ever-changing customer behaviors and operational constraints. The pandemic highlighted vulnerabilities in traditional economic structures while simultaneously creating opportunities for certain sectors to thrive.

This study examines the economic impact of the COVID-19 pandemic on various industries in the United States, with a focus on two key hypotheses:

- The healthcare sector was the most profitable industry during the observed period due to increased demand for medical services and products.
- Web-based businesses experienced significant growth and thrived during the observed period, while non-web-based businesses faced challenges and struggled to maintain profitability.

The rationale for these hypotheses comes from observable trends from the pandemic. The healthcare sector faced an immediate surge in demand for vaccines, treatments, personal protective equipment, and hospital services, driving significant investment and revenue growth. Simultaneously, the necessity for social distancing and remote interactions accelerated a digital transformation of many aspects of day-to-day life, including shopping, entertainment, and work, leading to a boom in web-based industries. In contrast, non-digital businesses, particularly those reliant on in-person interactions such as brick-and-mortar retail, faced steep declines in revenue and, in many cases, permanent closures.

To investigate these hypotheses, we gathered COVID-19 case data alongside the performance metrics of several major exchange traded funds, with particular attention to sector-specific trends. We then developed and analyzed three predictive models: a linear regression, to serve as a baseline and comparative metric; a logistic regression, to classify stock price movements in response to COVID-19 case trends; and a SVM, a probe to offer a complementary perspective.

This paper proceeds with a detailed explanation of our methodology, followed by an in-depth discussion of our results. By identifying patterns in economic growth and decline during the pandemic, this research seeks to provide insight into how industries can adapt to future crises and inform decisions aimed at promoting economic stability in uncertain times.

1c: Literature Review

According to Mazur et al. (2020), there was significant market volatility experienced by the global markets in response to COVID-19. The study attributes much of the volatility to the uncertainty produced by COVID. Changes in consumer behavior and government interventions, such as lockdowns and stimulus packages, also contributed significantly to the volatility. The paper discusses the impact across different economic sectors, such as healthcare, technology, energy, etc. Travel and energy sectors faced significant declines due to COVID-19, whereas technology and healthcare sectors experienced substantial growth. This study is directly relevant to our research topic because it supports our first hypothesis regarding the profitability of the healthcare sector. It provides valuable insights into how COVID-19 impacted sector-specific stock performance, which is a key focus of our study as well.

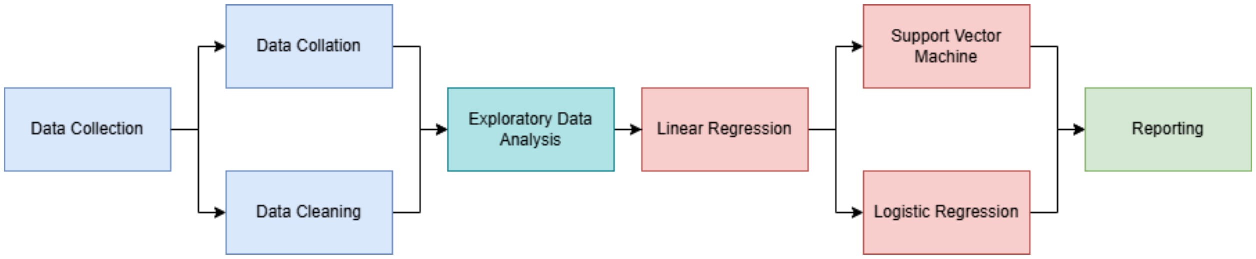
Smales (2020) uses Google Search Volume (GSV) as a proxy for investor attention and investigates how increased attention on GSV influenced stock returns. During COVID-19, heightened investor attention had a negative relationship with overall market returns. Healthcare and consumer staples sectors performed well due to the increased demand for these products. This paper supports and provides evidence for our first hypothesis as well. The use of GSV in the paper to analyze investor behavior aligns with the data-driven approach of our study.

Szász et al. (2021) explores the acceleration of e-commerce adoption during the pandemic. The paper highlights how web-based businesses adapted to the increased online demand, while traditional retailers faced declines in revenues and customers. It emphasizes the advantages of digital transformation, such as scalability, convenience, and resilience against physical restrictions, such as those posed by COVID-19. Although the paper primarily focuses on the retail sector, it forms a foundation for our second hypothesis. Our second hypothesis examines whether web-based businesses performed better than their brick-and-mortar counterparts, similar to the study's findings that web-based businesses experienced significant growth during the pandemic.

1d: Project Workflow Diagram

Our project workflow is designed to systematically process the data and apply machine learning models to analyze stock market trends. The workflow is structured as follows:

Figure 2: Project Workflow with Groupings for Each "Phase"



Workflow elements were grouped together so information isn't redundant

Data Collection, Collation, and Cleaning

The project began with collecting stock market data using the yfinance package, a Python library that pulls data from Yahoo Finance. We focused on Exchange-Traded Funds (ETFs) from various sectors, ensuring a diverse dataset. Once the data was collected, it was cleaned and collated by removing any null or missing values to ensure that the models would run smoothly without issues related to incomplete data.

Exploratory Data Analysis

During the EDA phase, we visualized stock data trends using line plots and boxplots to identify distributions, potential outliers, and general trends across the dataset. This step allowed us to better understand the data's structure and make informed decisions about which features to include in the models.

Model Development And Analysis

In this stage, three different machine learning models were implemented: linear regression, support vector machines (SVM), and logistic regression. Linear regression was used first as a benchmark to compare the performance of the other models. Each model has its own performance benchmarks which ensured that any results we gathered were valid and could be used to support our hypotheses.

Reporting

Finally, the findings from the analysis were compiled into a comprehensive report. This report (the one being read right now) summarized the insights gained from the data analysis and model performance, addressing the research questions and hypotheses. The report also included visualizations, performance metrics, and recommendations based on the results of the models.

2: Data Description and Methodology

The purpose of this work is to understand the effects of the COVID-19 pandemic on the stock market. To achieve this, we collected various stock performance metrics, including daily price changes, trading volumes, and market volatility. Instead of evaluating a handful of individual stocks from the tech and non-tech sectors, we opted to use Exchange-Traded Funds (ETFs). ETFs offer several advantages over individual stocks, as they provide a diversified portfolio of assets, which helps reduce risks associated with the volatility of individual stocks. This allowed us to capture the performance of entire sectors or industries, offering a broader view of market trends, as opposed to focusing on company-specific risks.

In selecting ETFs, we focused on funds representing major sectors of the economy that could easily be categorized as either tech-adjacent or non-tech. This approach allowed us to maintain a balanced dataset. While categorizing, we ensured an equal distribution of data points for each class. The ETFs used in the analysis and their corresponding classifications are as follows:

Figure 2: Breakdown of each Exchange Traded Fund used in our analysis

```
RWR - Real Estate: Non-Tech (Berkshire Hathaway etc.)
VCR - Consumer Cyclical: Non-Tech (Luxury Goods)
VDC - Consumer Defensive: Non-Tech (Survival Necessities)
VGT - Technology: Tech (FAANG Companies adjacent)
VHT - Healthcare: Non-Tech (United Health, J&J etc.)
VIS - Industrial: Tech (Boeing, Lockheed Martin etc.)
VOX - Communications: Tech (AT&T, Verizon etc.)
VTABX - World Bonds Hedged: Non-Tech (US Govt. Bonds)
```

For consistency, we collected stock performance data over a controlled four-month period during the pandemic. To complement the stock data, we gathered pandemic-related metrics, including infection rates, government lockdowns, and economic relief measures.

Data collection was facilitated by APIs: the Yahoo Finance API (via the "yfinance" Python package) for stock data, and the "Our World in Data" Python package for pandemic-related metrics. Data processing involved scraping and cleaning redundant data points, such as removing non-U.S. entries and null values.

All collected data was stored on Google Drive and GitHub, which allowed for easy version control and backup management. As the data underwent significant changes (e.g., filtering and processing), we maintained multiple versions to track how the dataset evolved. The data is publicly accessible through links provided in the citation section of this paper. On Google Drive, stock data is split by individual ETF into separate CSV files, while on GitHub, data is organized into pre- and post-processing categories.

For the stock data, we used the OHLC (Open, High, Low, Close) standard, a widely recognized format in financial markets for representing data over time. This format facilitated more straightforward analysis with existing methods. Metadata for the project is included in a metadata.md file in the GitHub repository, following the ISO 19115 standard. Although this ensured the dataset's usability, adopting a more widely used metadata standard, such as DataCite or Dublin Core, could have improved interoperability and discovery. All relevant data links are provided in the citation section of this paper.

3a: Hypotheses and Analysis Design

Our first hypothesis was that the healthcare sector was the most profitable industry during the observed period due to increased demand for medical services and products. To test this, we will begin by analyzing stock performance data for the healthcare sector (VHT ETF), focusing on periods of heightened pandemic activity, such as case surges and vaccine rollouts, and comparing these trends with COVID-19 metrics like case numbers and vaccination rates. In the full analysis, we will use statistical correlation and linear regression to examine the relationship between stock performance and pandemic metrics, particularly during peaks in COVID-19 cases and public health interventions. Finally, the post-analysis will summarize the key findings, determining whether the healthcare sector's profitability aligns with the pandemic's progression and providing insights for future investments and preparedness.

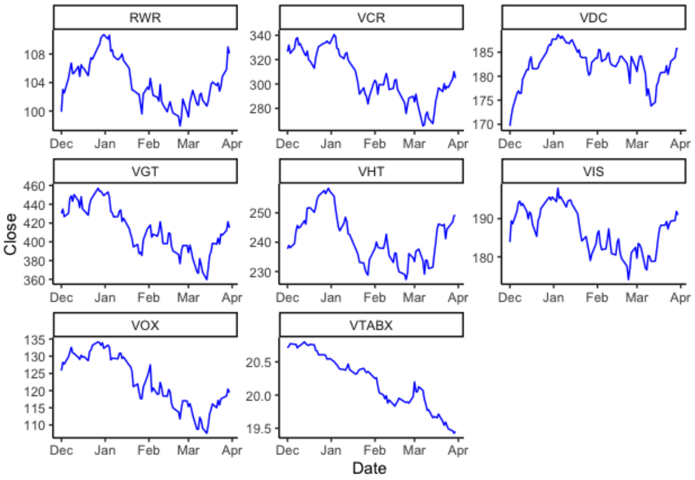
Our second hypothesis was that web-based businesses experienced significant growth, while non-web-based businesses struggled to maintain profitability during the observed period.

The preliminary analysis will compare stock performance trends of web-based businesses (e.g., Amazon) with non-web businesses (e.g., traditional retailers) during key events such as lockdowns and COVID-19 case spikes. The full analysis will involve comparing growth rates and applying clustering techniques to evaluate the performance differences between web-based and non-web businesses, particularly during critical periods of the pandemic. In the post-analysis phase, we will assess how web-based businesses contributed to the overall consumer sector growth, examining the factors that supported their success, such as digital transformation, while contrasting them with the challenges faced by non-web businesses.

3b: Exploratory Data Analysis

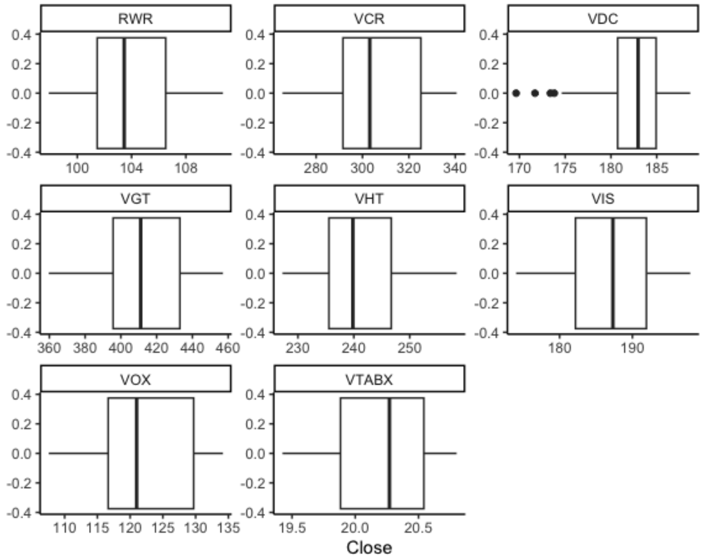
We can explore what our data looks like before we do an in-depth analysis on the data itself. Using R, we managed to combine all of the data sets into one giant dataset, and then used facet_wrap in order to see all of the different stocks. This led us with 8 different graphs for each stock that we decided to evaluate. The graphs is as follows:

Figure 3: Facet Wrap of the Closing Price v. Time (MM/DD/YYYY) of Each ETF Used



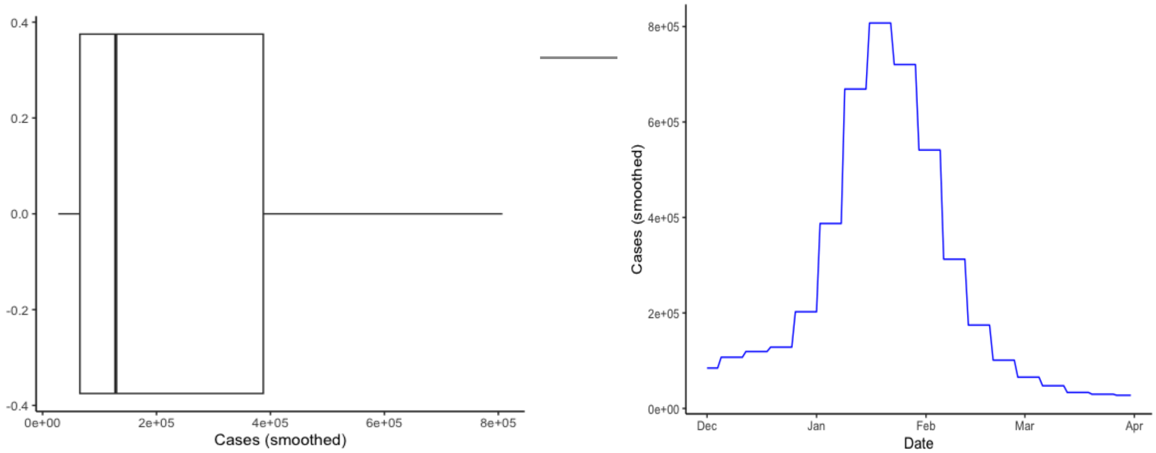
We can already see some information from looking at these graphs. The data is not linear, as stock data jumps up and down day from day. We can see a general trend with most of the data, as it goes down in February and sometimes March and then rises again in April, with VTABX being the only exception, which was our World Bonds Hedged Stock. We can then check for any outliers in our data, using a boxplot with all of the graphs.

Figure 4: Facet Wrap of Boxplots for Each ETF to Analyze Outliers



We can see from the boxplot that we rarely had outliers, with the outliers being in VDC only. We can assume that the outliers will not affect our data analysis because the outliers are insignificant if we are looking at the data as a whole. Now, we can look at the COVID data.

Figure 5: OWID COVID-19 Boxplot and Graph graphing New Cases (smoothed) v. Date (MM/DD/YYYY)



The data for COVID data is cases (smoothed) because our dataset did not record data everyday, so we smoothed it across the days where no data is collected. This makes our data either more normal or linear, which was more normal in this case. We can see an interesting trend, where the stocks generally fall when there are a lot of COVID cases. However, we would need to further apply models in order to see if COVID had an actual significant difference on the stock prices. In particular, we are considering different sectors and how COVID affected them, so future analysis would most likely include the stock market as a whole.

It was not really important to check the boxplot of the COVID data, as once again even if there were outliers, they would be insignificant to consider in our questions. Just from the boxplots and line graphs, we are able to reconstruct our summary statistics, which can also be computed, but would not be useful to explore in our hypothesis. Overall, the most useful data to explore would be the two line graphs, and we can do further analysis and comparisons based on how the stock data reacts to the COVID data, if there are any reactions.

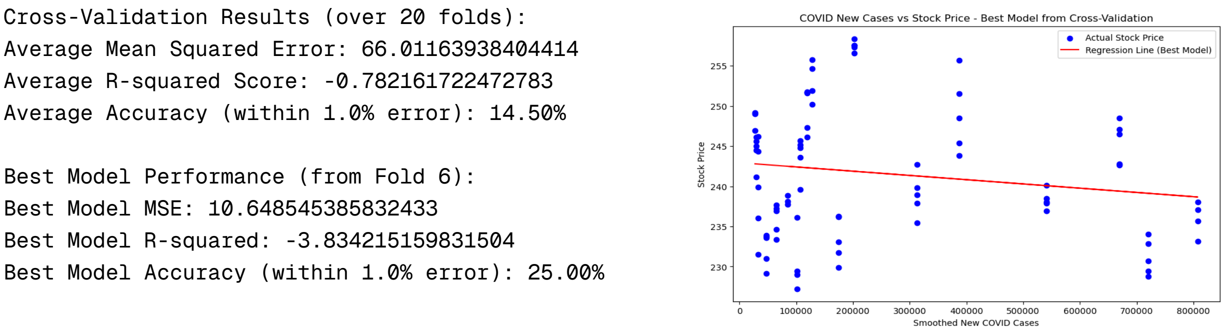
4: Model Development and Application of Models

We constructed three models: Linear Regression, Logistic Regression and a Support Vector Machine (SVM). The first two were used to support or reject our hypotheses and the SVM was done mainly out of curiosity to see how it would handle nonlinear relations.

Linear Regression

The baseline linear regression model as a comparison metric for any other models, as most other types of regression build on the linear regression equation. This model was not expected to work well at all, as we were essentially attempting a time-series analysis (which tends to be non-linear) on a dataset that wasn't inherently suited for linear regression. To improve the model's performance, we performed a 20-Fold Cross Validation test. K-Fold Cross Validation involves splitting the dataset into *K* subsets, training the model on *K−1* of these subsets, and testing it on the remaining one. This process is repeated for each subset, ensuring that each data point is used for both training and testing at least once. This approach allowed us to assess model performance more robustly and select the iteration with the best overall fit. Out of the 20 folds, we selected the model with the lowest Mean Squared Error (MSE) to determine which iteration provided the best predictive accuracy, aiming to minimize errors and optimize the model's generalization ability.

Figure 3: Sample Output of Linear Regression Model Predicting VHT (Healthcare) ETF



```
Cross-Validation Results (over 20 folds):
Average Mean Squared Error: 66.01163938404414
Average R-squared Score: -0.782161722472783
Average Accuracy (within 1.0% error): 14.50%

Best Model Performance (from Fold 6):
Best Model MSE: 10.648545385832433
Best Model R-squared: -3.834215159831504
Best Model Accuracy (within 1.0% error): 25.00%
```

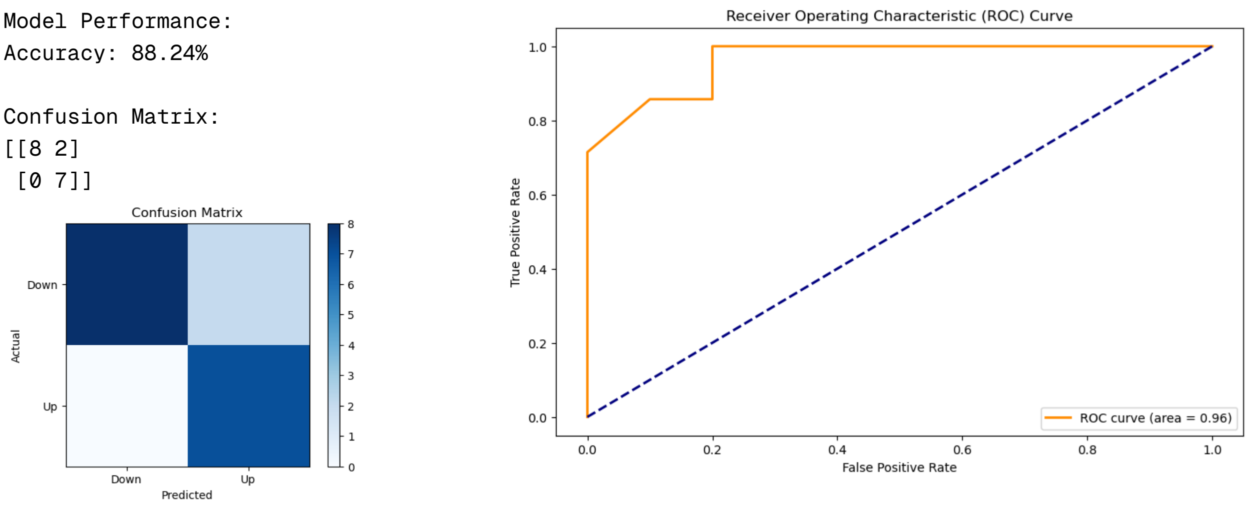
Logistic Regression

Having a baseline model, we decided to implement a logistic regression model next. The intuition behind this was that our end goal is to ultimately try and predict if the rise of COVID-19 cases had any significant impact on stock growth for a variety of sectors. Stock growth is mainly measured by the closing price of the stock, so if our model is able to predict an increase on a certain day based on the rise in COVID-19 cases, then we could confidently say, if it is of course not overfitting in any way, that the model is able to display a correlation between the rise in COVID-19 cases and an ETF's price increase or decrease.

Since logistic regressions are well-suited for binary classification tasks, they are ideal for situations where the goal is to predict one of two possible outcomes, such as whether a stock price will go up or down based on COVID-19 case trends. This is because logistic regression models the probability of an event occurring, outputting values between 0 and 1, which can easily be thresholded to assign a binary label (e.g., 1 for an increase and 0 for a decrease). This makes logistic regression both interpretable and effective for distinguishing between two distinct classes.

The code performs a logistic regression to analyze the relationship between COVID-19 data and stock market movements, with the goal of predicting whether a stock price will increase or decrease based on trends in new COVID cases. The model is trained on a dataset that merges COVID-19 case data with stock price data, where the target variable is binary, with one indicating a stock price increase and zero indicating a decrease. The dataset is split into training and testing sets to evaluate the model's performance, which is assessed using metrics such as accuracy, confusion matrix, Receiver Operating Characteristic (ROC) curve, and Area Under Curve (AUC).

Figure 4: Sample Output of Linear Regression Model Predicting VHT (Healthcare) ETF



```
Model Performance:
Accuracy: 88.24%

Confusion Matrix:
[[8 2]
 [0 7]]
```

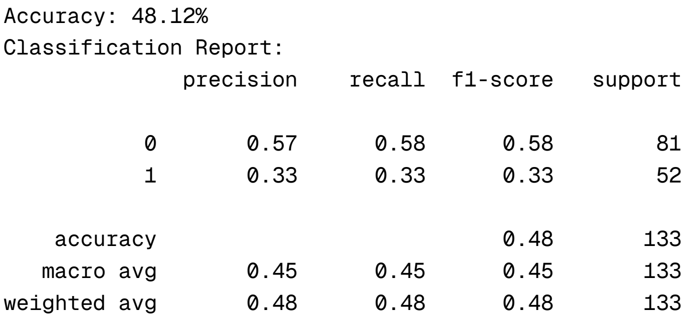
Support Vector Machine

The goal of the Support Vector Machine (SVM) in this task was to categorize Exchange-Traded Funds (ETFs) into two sectors: Tech (label 1) and Non-Tech (label 0) based on their stock data, specifically the closing price and the number of new COVID-19 cases. If the SVM model was able to conclusively and accurately classify every Tech sector ETF, it would suggest that the Tech sector stocks experienced a steady increase over time compared to the Non-Tech sector stocks during the observed period. This would provide valuable insights into how the performance of these sectors was influenced by COVID-19 metrics, helping to understand trends in the stock market during the pandemic.

However, the SVM struggled with this classification task. The SVM is designed to find the optimal hyperplane that separates the two classes (Tech and Non-Tech) in the feature space. While this method works well in many scenarios, particularly when there is a clear boundary between classes, it can face challenges when the data is noisy or the relationship between the features and the target variable is more complex. In this case, the SVM may have struggled to find a definitive and accurate separation between Tech and Non-Tech stocks, possibly due to overlapping patterns in the stock prices and COVID-19 cases or insufficient distinguishing features.

Despite this, using an SVM for this task was a good choice because SVMs are well-suited for classification problems, especially in cases where the data is high-dimensional and complex. In this case, the model was trained on a relatively simple two-feature dataset (closing price and new COVID-19 cases), which allowed for a clear interpretation of how these two factors might relate to sector classification. Moreover, the linear kernel used by the SVM helps to determine if a linear decision boundary can effectively separate the two classes. Here is a sample output:

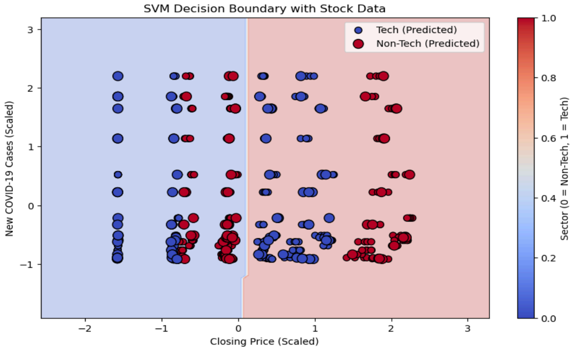
Figure 4: Sample Output of Linear Regression Model Predicting VHT (Healthcare) ETF



```
Accuracy: 48.12%
Classification Report:
              precision    recall  f1-score   support

           0       0.57      0.58      0.58        81
           1       0.33      0.33      0.33        52

    accuracy                           0.48       133
   macro avg       0.45      0.45      0.45       133
weighted avg       0.48      0.48      0.48       133
```



While the SVM didn't fully succeed in categorizing the ETFs accurately, with poor accuracy and F1 scores, it still provided a useful framework for attempting to identify patterns and relationships between stock performance and pandemic-related factors. Despite the challenges in classification, the SVM model offered valuable insights into the complexities of predicting sector performance during the pandemic. The lower performance may have been due to the limitations of the model in handling the nuances of stock data or the difficulty in separating the tech and non-tech sectors based solely on the available features.

5: Conclusion and Discussion

The COVID-19 pandemic had a significant impact on the stock market, which differed greatly across industries. Our team sought to analyze these differences using predictive models to validate two hypotheses: the profitability of the healthcare sector during the pandemic and the comparative success of web-based companies over their brick-and-mortar counterparts.

The results of our project corroborated the first hypothesis, as the healthcare sector emerged as the most profitable sector in this timeframe during the pandemic. Increased demand for medicinal products such as vaccines, treatments, and medical supplies triggered consistent revenue growth, which was exemplified by the healthcare focused exchange traded funds (ETFs). The logistic regression models demonstrated how their stock prices increased with the COVID-19 case count rising—an evident positive correlation.

The results of our project, though, were inconclusive in the case of the second hypothesis. While identified trends suggested a transformation towards web-based, digital businesses, our models failed to confirm this suggestion with statistically significant evidence.

Throughout the project, our analysis evolved significantly. Initially, we employed linear regression as a baseline. However, its unsuitability for time-series data became apparent. With higher interpretability and robust performance metrics in mind, we implemented a logistic regression model for binary classification, which was successful. An SVM was also explored to be a potential model in order to consider nonlinear relationships. In the end though, this did not influence our conclusions.

Data handling was also an essential component. We collected and integrated both stock and COVID-19 data by leveraging two Python packages in particular—yfinance and Our World in Data. All generated data was stored before and after processing (unfiltered and filtered versions) in Google Drive and GitHub for easy collaboration, version control, and additionally easily discoverable. Adopting OHLC standards for stock data and strictly following ISO 1115 metadata conventions enhanced our datasets' usability.

Looking forward to future plans for exploration, our focus would be refining our models. This could extend to employing deep learning models, such as LSTMs A secondary focus would be to expand the dataset to include additional factors, such as vaccination rates, or to expand to include other countries for a global sector analysis. In terms of our data though, we would first look to improve data discoverability and interoperability by switching to a different metadata standard: Dublin Core.

In summary, this project explores how economic health can be largely impacted by public health crises. This makes adaptability vital. Industries like healthcare and web-based businesses are prime examples, showcasing strong resilience in their market values despite unprecedented market disruptions. Our hope is that these findings provide a foundation to understanding sector-specific dynamics during global crises.

Citations

Mazur, M., Dang, M., & Vega, M. (2021). COVID-19 and the March 2020 stock market crash: Evidence from S&P1500. *Finance Research Letters, 38,* 101690. https://doi.org/10.1016/j.frl.2020.101690

Smales, L. A. (2020). Investor attention and global market returns during the COVID-19 crisis. *International Review of Financial Analysis, 72*, 101617.

Szász, L., Bálint, C., Csíki, O., Nagy, B. Z., Rácz, B.-G., & Csala, D. (2022). The impact of COVID-19 on the evolution of online retail: The pandemic as a window of opportunity. *Journal of Retailing and Consumer Services, 66*, 103089. https://doi.org/10.1016/j.jretconser.2022.103089

Google Drive with Stock and COVID-19 Data

Google. (n.d.). *Stock data*. Google Drive. Retrieved December 8, 2024, from https://drive.google.com/drive/folders/1CdDCAzUwHsTB6UhxlpUTR7KoWbe5ffES?usp=drive_link

GitHub Repository with Stock and COVID-19 Data

Amrevr. (2021, October 5). *Metadata_Covid_and_Stocks*. GitHub. https://github.com/amrevr/Metadata_Covid_and_Stocks/tree/main

Chen, W., Verma, R. (2024, December 8). *ITWS-4350-Project* . GitHub. https://github.com/williamschen23/ITWS-4350-Project