DS 201 FINAL PROJECT PROPOSAL

Covid-19 Classification and Effect on Health sector

https://www.kaggle.com/sudalairajkumar/covid19-in-usa?select=us_covid19_daily.csv
https://www.investing.com/indices/s-p-500-health-care-historical-data

Data

We will be utilizing data from Kaggle (Covid-19 cases are available daily in the United States) and investing.com (historical data for the S&P healthcare sectors). The Covid-19 dataset includes the columns 'date,' 'hospitalizedCurrently," recovered,' 'deathIncrease,' 'hospitalizedIncreased,' 'positiveIncrease,' 'negativeIncrease,' and 'totalTestResultsIncrease.' The columns in the dataset for historical healthcare data are 'Date,' 'Price,' 'Open,' 'High,' 'Low,' 'Close,' 'Volume,' and 'Change percent'. These datasets will be cleaned and combined. Because of the large number of variables from the two datasets, we used feature selection approaches to pick covariates. Because of the large number of variables from the two datasets, we used feature selection techniques to pick columns that are significant for the modeling process. Please see the links provided above for further details on the data.

Purpose

The purpose of our study is to create a model that will assist us in predicting the stock trend of the healthcare industry using data from the COVID-19 pandemic. We will utilize the COVID dataset's numerous columns as independent variables and the change in the healthcare sector's stock trend as the dependent variable.

Methodology

To determine the relationship between COVID-19 and healthcare stocks, we will first clean and combine the data. We'll need to integrate them into a single dataset by organizing the datasets' rows based on matching dates. Then, as a dependent variable, a new binary variable named 'price change' will be generated, with the value 0 indicating that the price has decreased and 1 indicating that the price has increased. To avoid biases in the data, the data will be divided into training and testing data for model validation purposes. Various supervised learning models, such as Logistic Regression and Random Forest, will be developed for the modeling process.We will need to examine the correlations of each variable to our dependent variable and choose the predictor variables that have the greatest influence on our dependent variable. After we have generated all of the models, we will evaluate their performance based on the R squared value and the accuracy using a confusion matrix of each model on the testing data. To achieve the best model feasible, parameters for each model will be modified.

Who cares? Importance?

This project's findings would be extremely useful to the government, healthcare companies, epidemiologists, economists, and private investors. The stock market is one of the world's most significant marketplaces, and as the covid epidemic spreads over the world, it is critical to evaluate its influence on the stock market. Because covid is a healthcare issue, healthcare equities are inextricably linked to the impact of the covid-pandemic. For example, the government could be interested in utilizing the data to develop healthcare policy, such as whether the healthcare industry needs greater financing. Based on the data, the healthcare business might make better financial decisions. Epidemiologists and economics would be concerned about the impact of a pandemic on the healthcare industry. Individual investors, on the other hand, would seek to exploit the results to their advantage when trading stocks.

What are the risks?

The concerns include that the covid data may not precisely forecast the stock price of the healthcare sector because there are several other variables that impact stock prices. Furthermore, correlation is not causation, so even if there is a link between covid instances and stock prices, it is conceivable that it is only coincidental.

What is the payoff?

If the initiative is successful, numerous parties will be able to utilize the information to make better financial decisions not just during this pandemic but also throughout future pandemics. Based on the statistics, people, in general, should better prepare financially for the next epidemic.

How is it done today, and what are the limits of current practice?

Impact it may have?

It has been in the previous two years, and individuals all around the world believe that the Covid-19 Spread is to blame for the stock market's anomalous trajectory. As a result, our initiative will assist individuals in reaching a conclusion and validating the phenomena.

Limits of current practice

Our trained model generalizes the stock trend by taking into account the healthcare sectors and data from the COVID pandemic. Given the large number of factors that may contribute to the trend of stocks, such as weather, politics (incentives and policy changes to companies), and unexpected events (Suez Canal incident affects exports/imports), the model that we have trained may not accurately reflect the trend of healthcare stocks. If we took into account all of these variables, we might not be able to construct a prediction model due to restricted computational resources and gaps in our grasp of sophisticated data science ideas.

How much will it cost?

The first dataset is open-sourced and free from Kaggle, while the second is equally free from investing.com. As a result, the initiative is completely free of charge monetarily.

How long will it take?

The project is expected to take up to three weeks to complete. Combining and cleaning the data should take one week, followed by displaying the data and developing the model. The examination and selection of models might take up to a week.

How will progress be measured?

Progress can be measured based on the completion of several checkpoints.

1. Data selection

2. Data cleaning and aggregation

3. Visualization of data

4. Regression, Random Forest modeling

5. Model tweaking and performance comparison

6. Model selection

7. Report