Team members: Andy Foo, Daniel Lok, Ngoc Nguyen
Team number: 3

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

The data we will be using is from Kaggle (Covid-19 cases daily in the USA) and investing.com (historical data for the S&P healthcare sectors). In detail, the Covid-19 dataset contains 'date', 'hospitalizedCurrently', 'recovered','deathIncrease', 'hospitalizedIncreased', 'positiveIncrease', 'negativeIncrease' and 'totalTestResultsIncrease' columns. As for the historical healthcare data, the columns which are present in the dataset are 'Date', 'Price', 'Open', 'High', 'Low', 'Close', 'Volume' and 'Change %'. These datasets will undergo the cleaning and combining process. Due to the numerous amounts of variables coming from two datasets, we've applied feature selection techniques to select columns that are important for the modelling process. For more information about the data, please check out the links attached above.

Purpose

The purpose of our project is to build a model that will help us predict the stock trend of the healthcare sector based on the data acquired from the COVID-19 pandemic. We will use the various columns in the COVID dataset as independent variables and the change in the stock trend of the healthcare sector as the dependent variable.

Methodology

For the purpose of finding the relation between COVID-19 and healthcare stocks,we will begin with cleaning and combining the data. We will need to combine them into a single dataset by arranging the rows of the datasets based on matching dates. Then, a new binary variable named 'price change' will be created as a dependent variable, where the value 0 indicates that the price has dropped while 1 indicates that the price has increased. To prevent any form of biases towards the data, the data will be splitted into training and testing data for model validation purposes. For the modelling process, different supervised learning models such as Logistic Regression, Random Forest will be created .We will need to compare each variable's correlation to our dependent variable and select predictor variables that have the strongest effect on our dependent variable. Having all the models created, we will then evaluate the performance of our model based on the R squared value and the accuracy using a confusion matrix of each model on the testing data. Parameters for each model will also be tweaked to obtain the best model possible.

Who cares? Importance?

The findings in this project would be of great relevance to the government, healthcare industries, epidemiologists, economists and individual investors. The stock market is one of the most important markets in the world, and as the covid pandemic spread around the world, it is important to examine its impact on the stocks. As covid is a healthcare issue, the healthcare stocks are very much related to the impact of the covid-pandemic. For instance, the government would be interested in using the findings to determine policies on healthcare such as whether the healthcare sector needs more funding. The healthcare industry could make better financial judgement based on the datas. Epidemiologists and economists would also be concerned at the relationship that a pandemic has on the healthcare sector.

Meanwhile individual investors would want to use the findings to their own advantage when trading shares.

What are the risks?

The risks are that the covid data may not accurately predict the stock price of the healthcare sector as there are many other variables that will affect the stock prices as well. Also correlation is not causation, therefore even if there is a correlation between covid cases and stock prices, it is possible that it could be just coincidence.

What is the payoff?

If the project is successful then many parties can use findings to make better financial decisions not only during this pandemic and other pandemic as well. People in general could also better prepare for the next pandemic financially based on the data.

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

Impact it may have?

It is in the last two years and people around the world assume that the stock market's phenomenon trend is caused by the Covid-19 Spread. So our project will help people somehow conclude and confirm the phenomenon.

Limits of current practice

Our trained model generalizes the trend of stocks by considering the sectors of healthcare the data obtained from the COVID pandemic. Given the large amounts of factors that may contribute to the trend of stocks such as weather, politics (incentives and policy changes to companies), unpredicted situations (Suez Canal incident affects exports/imports), the model that we have trained may not reflect the actual trend of healthcare stocks. If we were to actually consider all these variables, we may not be able to generate a prediction model due to the limited computing resources and insufficiencies in our understanding of advanced data science concepts.

How much will it cost?

The first dataset is free and open-sourced from Kaggle while the second dataset is also free from investing.com. Therefore the project is totally free financially.

How long will it take?

The project will take up to 3 weeks to complete. Combining and cleaning the data should take one week, meanwhile visualizing the data and building the model should take another week. Model evaluation and selection will take up to 1 week for completion.

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