DS 201 FINAL PROJECT REPORT
Covid-19 Classification and Effect on the Health sector

I. Introduction

The goal of our project is to predict the trend of healthcare stocks using COVID-19 data. In detail, we try to find the relationship between the price index of stocks and daily covid data that contains several variables such as the (number of positive cases, negative cases, number of people admitted into the hospital, etc.) and try to build a machine learning model for prediction using logistic regression and random forest model. Having a successful model would be useful for investors, business analysts, and the government sector as they could utilize the model to determine their financing choices.

II. Problem

With the goal of the project in mind, the issue we encountered at the start of the project would be the lack of data. Upon searching the internet we have not found a dataset that consists of daily covid-19 data relating to the trend of stocks. Knowing that this is the issue, we have decided to construct our data by finding a dataset from the two sectors and performing the aggregation of the two datasets after performing cleaning, filtering, and some other form of processing (which will be discussed in the section below later).

III. Survey

No survey has been conducted for this project. We have acquired our data from online sources and combined them for the modeling process. More details about the data will be provided below.

IV. Proposed Method

A. Data Collection and Processing

For the purpose of our project, two separate datasets were acquired from two different sources. The first dataset (Covid-19 dataset) was acquired from Kaggle while the second dataset (S&P 500 healthcare data) was acquired from investing.com.

The first dataset is then processed. In general, nonessential columns were dropped from the covid dataset based on intuition. Knowing that the date is in reversed order (latest to oldest date), the order for the whole dataset is inverted. Some columns of the covid dataset contain null values because covid cases are not that widespread yet (no one has entered the hospital, or recovered from Covid at the beginning of the pandemic). So, rows with null values were dropped from the dataset to reflect normal Covid conditions. The processed dataset is then saved into a new .csv file.

For the second dataset, the whole dataset is also inverted to align with the order of the first dataset. Then, non-essential columns were dropped from the dataset. One problem with the dataset is that the prices index (high, low columns) has the string data type as the price indexes have the comma symbol between the numbers. To solve this problem, the coma from the price indexes is first removed and then converted into a float data type. Next, we make a new column called avg_price by taking the average of the sum between the high and low price columns. Another major problem we have with this dataset is that the dates shown in the dataset are not continuous. Having the goal of predicting the stock trend based on continuous dates/ covid conditions of the previous day, an algorithm was manually written only to include rows with dates that are only continuous. The filtered dataset is then saved into a .csv file.

Finally, the two datasets is then combined by aligning the dates for both datasets.

B. Data Analytics

A pair plot was used to visualize the relationship between each of the variables that are present in the dataset. Some variables in the dataset present a linear relationship (almost a straight line) while some of them present either a “W” or “E” shaped graph, showing that the relationship between some variables isn’t straightforward.

Further visualizations were made to understand the data better. For a clearer visualization of some of the variables, we have incorporated the date variable into graphs, making it a time series plot of variables. For instance, we have plotted the time series graph for the cumulative positive cases and found out that the cumulative positive cases are, in fact, increasing as time passes. But, knowing this from the graph isn’t very helpful (it's somewhat expected that the cumulative cases increase as time passes). So, through some additional research, we have decided to plot the logarithmic graph to better visualize the trend for the data. Plotting logarithmic graphs allow us to see the rate of change of the variables which gives us new information about the data. Several plots about other variables show unpredictable behavior/ non-linear relations, so there is no clear way of explaining the trend of the figures of some variables.

V. Experiments / Evaluation

In this part of our study, we will go through all the experiments and our explanations for them. First of all, we analyze the data and then train the two simplest machine learning models with Linear Regression and Random Forest algorithms.

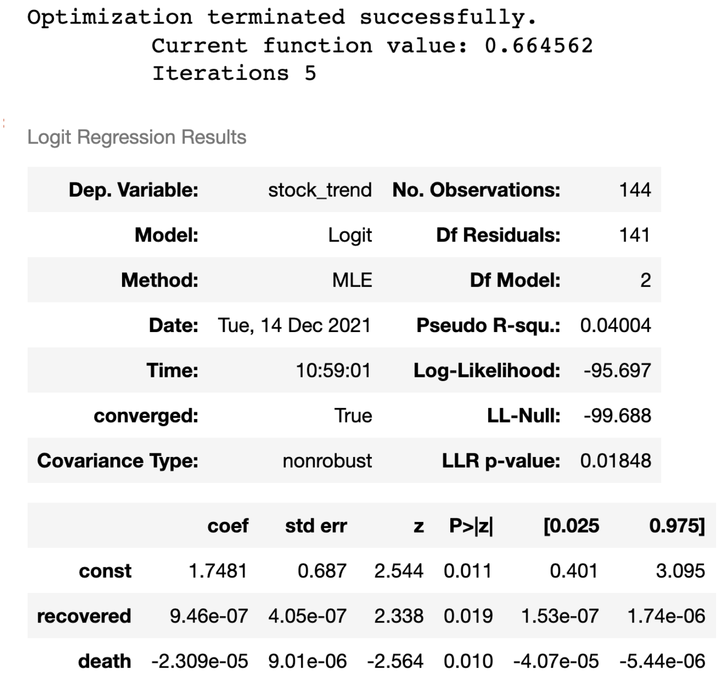
Clean data by formalizing and combining variables from multiple data sets.

1) Previewing the data. Then reprocessed data by adjusting the data type (converting date to DateTime format), removing null values and nonessential columns, then inverting the dataset due to the reversed date order. Then we monitor the relationship between covid and other variables in

the dataset. To avoid overfitting, we will remove all the unrelated variables. Next, create the new column called ‘avg_price’, in the meantime, we realized the dates aren't continuous. Some dates are missing in between. Hence, we will first create a stock_trend column and denote all of them with the value 3 then depending on the situation, we will denote the value with 1 or 0. The first dataset is ‘covid_df_preprocessed.csv’ then ‘hc_df_preprocessed.csv’ and ‘FinalData.csv’ is the elaborate data sets that we use for further experiments.

2) Building our Linear Regression and Random Forest model:

a) Linear Regression: First, we used OLS to filter out any variables that have p-values more than 0.05. After dropping all of them, the remaining variables are ‘recovered’ and ‘death’ which seems suitable for our LR model.



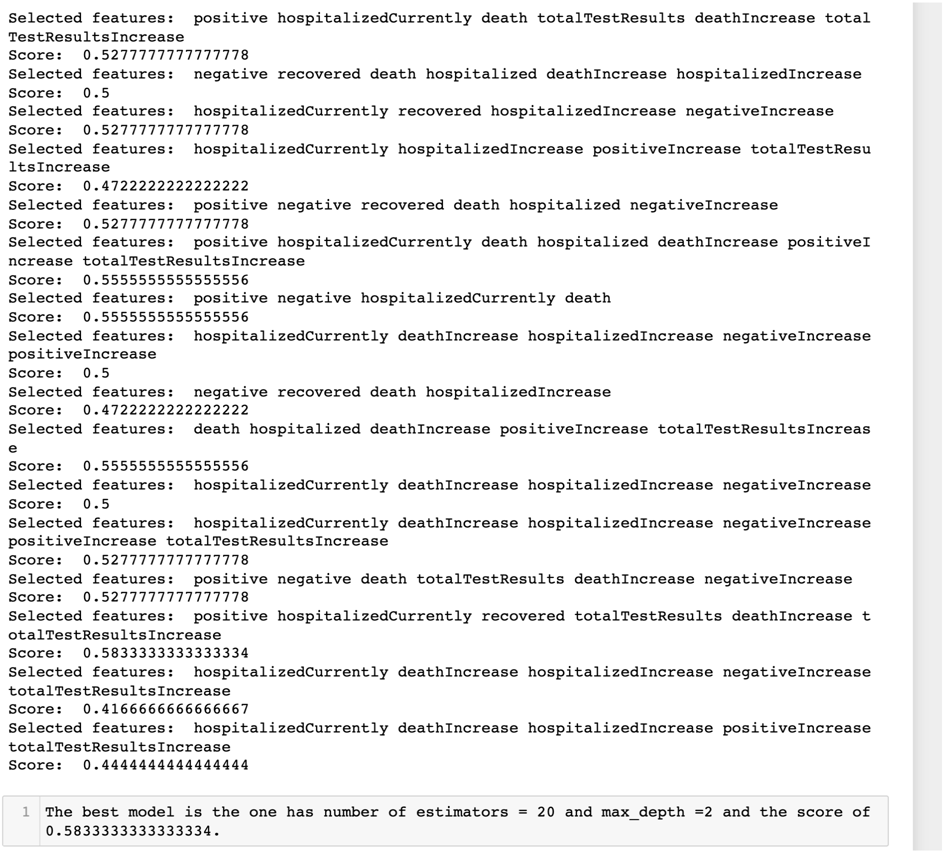
```
Optimization terminated successfully.
         Current function value: 0.664562
         Iterations 5
```

Logit Regression Results

Dep. Variable:	stock_trend	**No. Observations:**	144
Model:	Logit	**Df Residuals:**	141
Method:	MLE	**Df Model:**	2
Date:	Tue, 14 Dec 2021	**Pseudo R-squ.:**	0.04004
Time:	10:59:01	**Log-Likelihood:**	-95.697
converged:	True	**LL-Null:**	-99.688
Covariance Type:	nonrobust	**LLR p-value:**	0.01848

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	1.7481	0.687	2.544	0.011	0.401	3.095
recovered	9.46e-07	4.05e-07	2.338	0.019	1.53e-07	1.74e-06
death	-2.309e-05	9.01e-06	-2.564	0.010	-4.07e-05	-5.44e-06

b) Random Forest: We use a nested for loop for n_trees in [10, 20, 50, 100] and depth in [1, 2, 5, 10] and print out all the scores from the models.

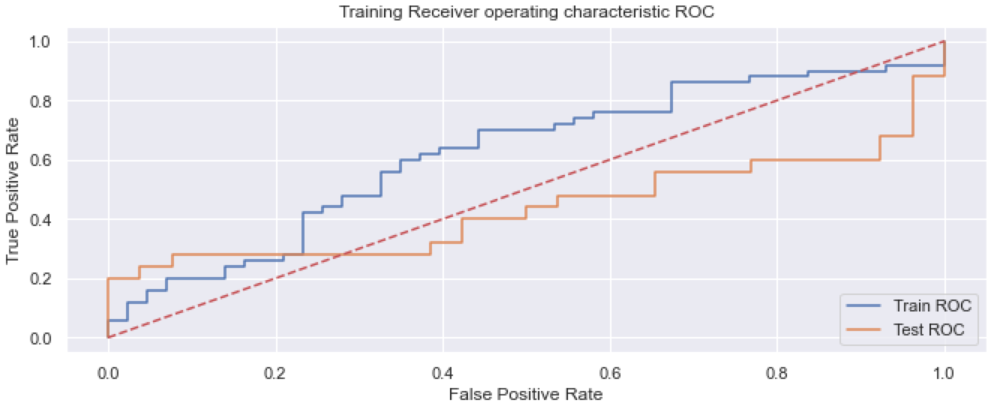


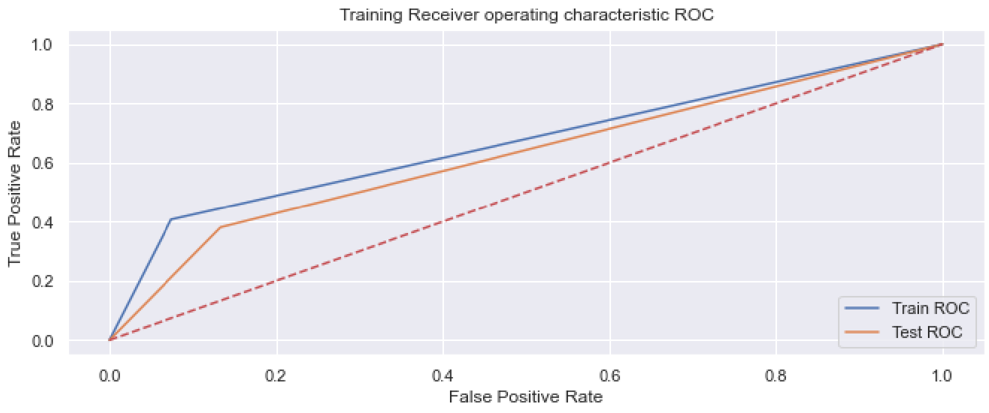
```
Selected features:  positive hospitalizedCurrently death totalTestResults deathIncrease total
TestResultsIncrease
Score:  0.5277777777777778
Selected features:  negative recovered death hospitalized deathIncrease hospitalizedIncrease
Score:  0.5
Selected features:  hospitalizedCurrently recovered hospitalizedIncrease negativeIncrease
Score:  0.5277777777777778
Selected features:  hospitalizedCurrently hospitalizedIncrease positiveIncrease totalTestResu
ltsIncrease
Score:  0.4722222222222222
Selected features:  positive negative recovered death hospitalized negativeIncrease
Score:  0.5277777777777778
Selected features:  positive hospitalizedCurrently death hospitalized deathIncrease positiveI
ncrease totalTestResultsIncrease
Score:  0.5555555555555556
Selected features:  positive negative hospitalizedCurrently death
Score:  0.5555555555555556
Selected features:  hospitalizedCurrently deathIncrease hospitalizedIncrease negativeIncrease
positiveIncrease
Score:  0.5
Selected features:  negative recovered death hospitalizedIncrease
Score:  0.4722222222222222
Selected features:  death hospitalized deathIncrease positiveIncrease totalTestResultsIncreas
e
Score:  0.5555555555555556
Selected features:  hospitalizedCurrently deathIncrease hospitalizedIncrease negativeIncrease
Score:  0.5
Selected features:  hospitalizedCurrently deathIncrease hospitalizedIncrease negativeIncrease
positiveIncrease totalTestResultsIncrease
Score:  0.5277777777777778
Selected features:  positive negative death totalTestResults deathIncrease negativeIncrease
Score:  0.5277777777777778
Selected features:  positive hospitalizedCurrently recovered totalTestResults deathIncrease t
otalTestResultsIncrease
Score:  0.5833333333333334
Selected features:  hospitalizedCurrently deathIncrease hospitalizedIncrease negativeIncrease
totalTestResultsIncrease
Score:  0.4166666666666667
Selected features:  hospitalizedCurrently deathIncrease hospitalizedIncrease positiveIncrease
totalTestResultsIncrease
Score:  0.4444444444444444
```

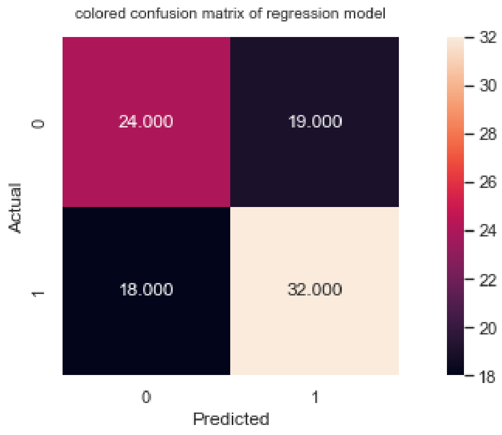
```
The best model is the one has number of estimators = 20 and max_depth =2 and the score of
0.5833333333333334.
```

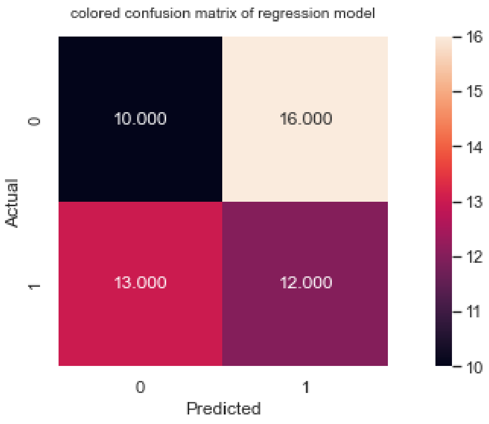
3) Model evaluation and comparison: Model3 with 65% train and 35% test is the best model with the highest score of 0.57 Meanwhile, the best Random Forest model is the one that has a number of estimators = 20 and max_depth =2 and the score of 0.583.

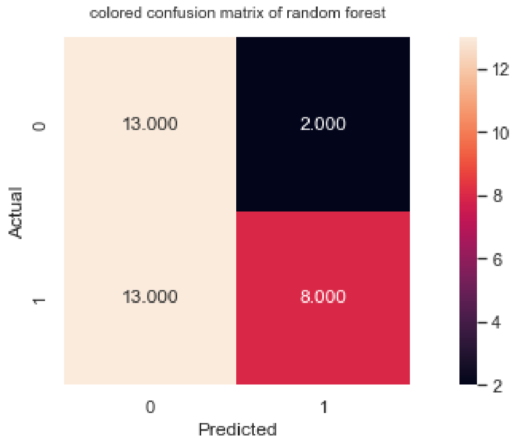
4) ROC curve and Colored Confusion Matrix Comparison











VI. Conclusion

In conclusion, we were able to define factors that do not assist anticipate stock trends; 'death' and 'recovered,' on the other hand, are more essential than other variables.

Based on the accuracy of the Logistic Regression and Random Forest (0.57 and 0.58, respectively), we can infer that all of the Covid-19 features, notably 'death' and 'recovered,' have no significant impact on the healthcare sector's stock trend. To explain, the stock market is a very complicated market that is impacted by many components, including human factors, and Covid-19 is only one of them, albeit a considerable one.

In order to dive deeper further into determining whether Covid-19 has a significant impact on the stock trend of healthcare sectors or not, we need a deeper investigation that requires more data and more time. Ideally, through our research, there will be plenty of prospects in the future that we can grow and continue to study and improve.

VII. References

Sevi, Semra, et al. “Logarithmic versus Linear Visualizations of COVID-19 Cases Do Not Affect Citizens’ Support for Confinement.” Canadian Journal of Political Science. Revue Canadienne De Science Politique, pp. 1–6. PubMed Central, https://doi.org/10.1017/S000842392000030X. Accessed 11 Dec. 2021.