Digital Tools for Finance

Final Project: Causalities of the Covid Pandemic

Yung-Hsin Chen 20-744-322
Haoxin Cai 20-742-151

Department of Informatics
University of Zurich

Dec. 19, 2022

Outline

Introduction

The COVID-19 pandemic has severe impacts on almost every aspect around the world: it causes not only social and economic disruption, but also drastic death rates. Inevitably, people are curious about the causalities of COVID-19 and how to prevent the pandemic from getting worse. Therefore, the report aims to **determine the correlation between the severity of the pandemic and other information of a country**. [1]

[1] Please note that the time parameters are taken out of consideration for simplicity.

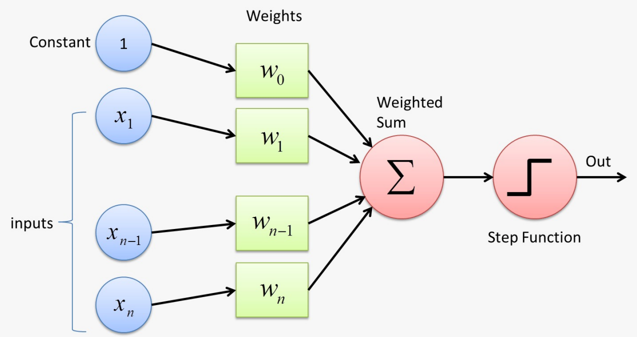
Outline

Domain knowledge

In this section, tools and domain knowledges used in the task and reasons of choosing them will be briefly explained including models used (logistic regression, linear perceptron, XGBoost), evaluation metrices (micro-f1, macro-f1) and visualisations (confusion matrix, normalised confusion matrix).

Comparison of Models: Linear Perceptron

Linear Perceptron is a very simple model for binary classification. After each data are multiplied by weights and biases, the result (scalars) will be passed to the activation function. In the case of unit-step function as the activation function, numbers larger than 0 will be assigned to a class, and vice versa. The whole process can be summarised by the flow chart below.[2]



[2] https://deepai.org/machine-learning-glossary-and-terms/perceptron

Comparison of Models: Linear Perceptron

Pros:

- Fast and easy to implement;
- Less susceptible to overfitting in dealing with small datasets.

Cons:

- High sensitivity to outliers;
- Limitations of linearity decision surface nature;
- Non-multicollinearity between independent variables needed;
- Binary outputs only and no calibrated probabilities possible.

Comparison of Models: Logistic regression

Logistic regression is a linear model commonly used for classification problems as well. It is very similar to the linear perceptron model except for the activation function. The activation function for logistic regression is Sigmoid function. For logistic regression, regularisation terms are also added to the loss function as a penalty for overfitting.

Pros:

- Regularisation terms helps prevent overfitting;
- Continuous probabilities for outputs allowed thanks to Sigmoid function.

Cons:

- Sensitive to outliers;
- Multicollinearity not allowed.

Comparison of Models: XGBoost

XGBoost is a gradient boosting model with extra structures that has been proved well performed on classification tasks. The boosting model creates weak learner (decision tree) one by one. The previous weak learner set a larger weight for the wrong answers as the input to the next weak learner. By going through the sequence, the classification will eventually be finalised correctly. Each weak learner will have their own predictions, and together they will have to vote to decide what the final answer is. The less accurate weak learners get less votes. **Pros**:

- Good at tackling non-linear problems;
- Non-multicollinearity on data not required;
- High efficiency due to the parallel tree boosting and less parameters to be trained.

Cons:

- High chances of overfitting due to its greediness, but can be solved by setting the correct maximum depth of the trees.

Since it is a classification task, f1-score is commonly used for model evaluation. However, f1-score is only calculated per class. The multi-class case will require the aggregation of f1-scores of each class to determine the overall performance of the models on all classes. In this case we use Micro-f1 and macro-f1.

F1-score is defined as equation 1.

$$\begin{aligned} \text{f1-score} &= 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \\ &= \frac{\text{TP}}{\text{TP} + \frac{1}{2}(\text{FP}+\text{FN})} \end{aligned} \tag{1}$$

Evaluation: Metrics

Macro-f1 calculates the average of f1-scores of all classes as equation 2. The equation shows that macro-f1 **gives same weights to each class no matter how large the class is**.

$$\text{macro-f1} = \frac{\text{sum(f1-score)}}{\text{number of classes}} \tag{2}$$

Evaluation: Metrics

On the other hand, micro-f1 is defined by equation 3. This equation is the same as the one for f1-score (1). However, the TP, FP and FN stands for the sum of the metrices for all classes. It gives **equal weights to each data points, resulting in the results dominated by the performance of large classes**.

$$\text{micro-f1} = \frac{\text{TP}}{\text{TP} + \frac{1}{2}(\text{FP}+\text{FN})} \tag{3}$$

Outline

Methods: Data Summary

Data used for this task is from covid-19-data from Our World of Data. A summary table of the data is shown in Table 1.

Table: Data Summary Table

Information	**Description**
number of columns	67
number of rows	236386
number of countries	248
key	{location:date}
starting date	01.Jan.2020

Methods: Data Cleaning

After data cleaning, we select certain attributes for our classification task as shown in Table 2, and categorize total_cases_per_million into 4 levels of severity, namely: [0, 50000, 200000, 400000, 700000].

Table: Data Summary Table

Selected Attributes	
aged_65_older	cardiovasc_death_rate
aged_70_older	diabetes_prevalence
gdp_per_capita	hospital_beds_per_thousand
population_density	human_development_index
life_expectancy	median_age
people_fully_vaccinated_per_hundred	

Methods: Classifiers

Classifiers are used for classification tasks. Our data is splitted into training (80%) and testing (20%) only due to its small size. A summary of models used for the task and the best performing hyperparameters chosen by applying grid search cross validation are listed out in Table 3.

Table: Model Summary

Model	**Details**
Logistic Regression	
penalty	l2
solver	newton-cg
Linear Perceptron	
tolerance	0.001
random state	0
XGBoost	
learning rate	0.1
loss	deviance
max depth	3
n_estimators	100
random state	21

Methods: Feature Importance

The feature importance is generated automatically. The higher the weight, the more significant its effect on the accuracy.

Table: Feature Importance from XGBoost

	Feature	Importance
1	human_development_index	0.462551
2	life_expectancy	0.462551
3	population_density	0.138737
4	hospital_beds_per_thousand	0.082238
5	gdp_per_capita	0.065485
6	people_fully_vaccinated_per_hundred	0.059118
7	cardiovasc_death_rate	0.045620
8	aged_70_older_percentage	0.033860
9	median_age	0.029610
10	diabetes_prevalence	0.026468
11	aged_65_older_percentage	0.025800

Outline

Conclusion: Model Performance

By applying the test dataset on the model, the performance of the models can be evaluated. The accuracy summary of each model and the feature importance table from the best performing model is shown in 5.

Table: Model Accuracy Summary

Model	**Accuracy**	**Micro-f1**	**Macro-f1**
Logistic Regression	64.10%	0.6410	0.5863
Linear Perceptron	48.72%	0.4872	0.2976
XGBoost	76.92%	0.7692	0.7749

Conclusion: Visualization

Since XGBoost is the best performing model, its visualisation will be shown.
The confusion matrix shows that most classes are classified correctly with the diagonal squares brighter than the others. However, due to the fact that level 4 has fewer data size than the others, it is hard to tell if it actually performs worse than the other levels.

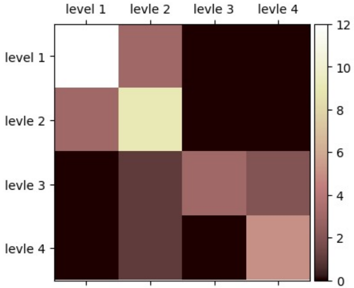


Figure: Confusion Matrix

Conclusion: Visualization

In the normalised confusion matrix, it is clear that the darker square of level 4 is caused by the smaller data size of the class. The model is actually performing well even on smaller data size classes. That is why the macro-f1 is larger than micro-f1.

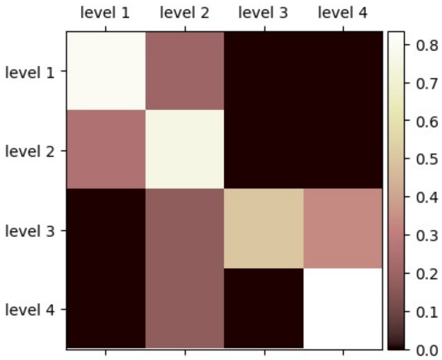


Figure: Normalized Confusion Matrix

Conclusion and Discussion

From the feature importance table, it can be concluded that the severity of covid is more related to the human development index. The human development index is a score on the quality of living in the country including average life expectanvy, GDP, education opportunity, etc. The number of people fully vaccinated, the population composition or the diabetes prevalence do not play the most important role in the total cases as expected.