**Model introduction**

Logistic regression model is a statistical method that statisticians and data scientists use to classify people, products, entities, etc. It is mostly used to analyze data that produces a binary classification based on one or many independent variables. This means, it produces two clear classifications (Yes or No, 1 or 0, etc).

Logistic regression model is widely used in the healthcare field. For example, healthcare providers use logistic regression to accurately target at risk individuals who should receive a more tailored and detailed behavioral health plan to help improve their daily health habits. This in turn opens the opportunity for better health for patients and lower costs for hospitals. For insurance companies, logistic regression helps them detect where there is potential fraud of insurance plans by measuring certain characteristics of their insurance payers. This will help companies reduce huge financial risks.

Since logistic regression is often used in binary classification, in this case, it is quite appropriate to predict mortality after learning features such as age, gender and different types of diseases.

**Model Interpretation**

After running the logistic regression model 200 times, we mainly get predictor coefficients, confusion matrix and feature importance to evaluate our model’s predictive performance.

1. Top 10 Predictor Coefficients

There are 226 predictors and we ranked them according to their corresponding coefficients. Below is table of the Top 10 predictors, which all have positive signs of their coefficients. Therefore, for people with those ten diseases are *more likely* to die due to covid-19, holding other features constant.

We can see that 5 of them are related to cardiac, 2 of them are related to chest. The result is aligned with the news that high-risk people have complications such as Heart disease, Lung disease and Respiratory disease. [[1]](#footnote-1) To be specific, the top 5 most impactive predictor are CVASC\_Cardiac\_B (related to heart disease), DERMA\_Whole\_Body\_Nos\_B (related to body burn), ENDOC\_MET\_Diabetes (related to diabetes), CVASC\_Heart\_Rhythm\_A (related to heart disease) and CHEST\_Status\_A (related to chest disease).

One thing needs to mention is that 14 of them have the coefficients of zero and we discover that these ICD-10 codes represent diseases that are not correlated with covid-19, such as adjusting artificial arm, Corrosion of mouth and pharynx and Injury of optic chiasm.

Table 1 Top 10 Predictors and corresponding Coefficients of Logistics regression model

|  |  |
| --- | --- |
| **Predictor** | **Coefficient** |
| CVASC\_Cardiac\_B | 3.33686838 |
| DERMA\_Whole\_Body\_Nos\_B | 2.77104041 |
| ENDOC\_MET\_Diabetes | 2.15934205 |
| CVASC\_Heart\_Rhythm\_A | 2.03867024 |
| CHEST\_Status\_A | 2.00502955 |
| CHEST\_Airway\_Lungs\_B | 1.9542987 |
| CVASC\_Venous\_B | 1.92617972 |
| CVASC\_Heart\_Rhythm\_B | 1.88779993 |
| FGENT\_Screening | 1.79267164 |
| CVASC\_Other\_Nos\_B | 1.7159024 |

2. Confusion Matrix

As we can see from the confusion matrix, the overall accuracy is 0.94, which represents all the right predictions (real no death, predictive no death & real death, predictive death) divided by all the predictions. The high accuracy means our model made few wrong predictions.

To be specific, the recall value for label 0 is 0.94, which means among total actual no death (82436+4954), 94% can be detected as no death (82436). The precision value for label 0 is 1, which means among total predicted no death people (82436+352), nearly 100% could be detected as no death.

The recall value for label 1 is 0.93, which means among total actual death (352+4917), 93% of them can be detected as death. As for precision, it measures the percentage of predicted death among total actual death. In this case, our model may have a half-half probability to mistake no death to death.

Table 2 Confusion Matrix for Logistics Regression Model

|  |  |  |
| --- | --- | --- |
|  | **Negative (No Death)** | **Positive (Death)** |
| **False (No Death)** | 82436 | 4954 |
| **True (Death)** | 352 | 4917 |

3. Classification Report

The classification report, which includes indicators such as precision, recall, f1-score, support for each label, and accuracy, macro avg and weight avg. The report is used for model comparison.

Table 3 Classification Report for Logistics Regression Model

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **precision** | | **recall** | | **f1-score** | | **support** |
| **0** | 1.00 | 0.94 | | 0.97 | | 87390 | |
| **1** | 0.50 | 0.93 | | 0.65 | | 5269 | |
| **accuracy** |  |  | | 0.94 | | 92659 | |
| **macro avg** | 0.75 | 0.94 | | 0.81 | | 92659 | |
| **weighted avg** | 0.97 | 0.94 | | 0.95 | | 92659 | |

4. Feature Importance

Below is the feature importance graph, which is according to the rank of coefficients. We can see that CVASC\_Cardiac\_B and DERMA\_Whole\_Body\_Nos\_B are the most important predictors among top 10 predictors, which are more than 80%. The result is sensible since people with heart diseases are listed as high-risk people while patients with body burn are very sensitive to viruses and they have weak resistance towards a disease that can be spread by air.

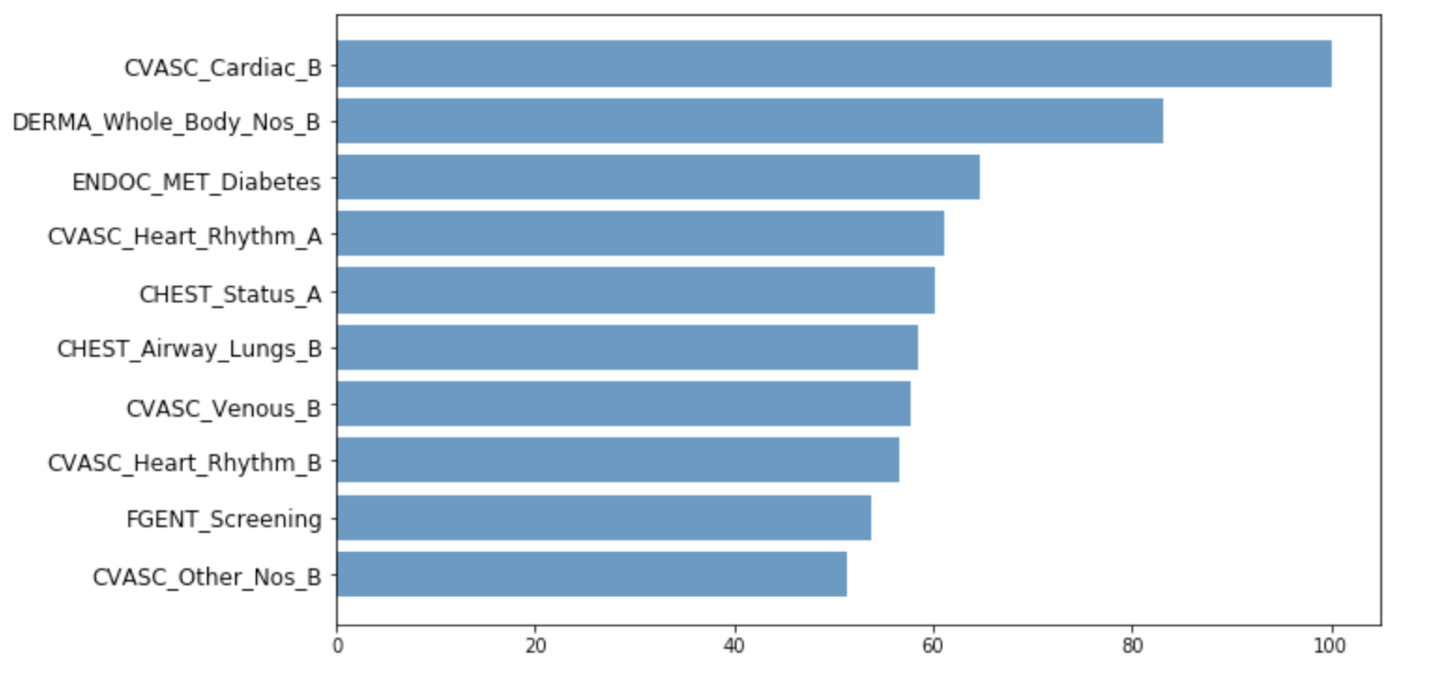


Figure 1 Feature importance for Logistics Regression Model

1. https://www.capitalcardiology.com/patient-education/covid-19/?lang=zh [↑](#footnote-ref-1)