DNSC 6279 Data Mining Final Report

Influential Factors to Mortality in Covid-19

Team Member: Yuwen Luo, Yihang Zhao, Runzhe Tang

1. Introduction

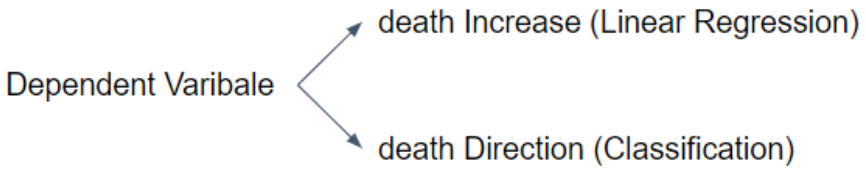
Due to the breakout of the Covid-19, the pandemic is affecting the normal life of all human beings all around the world. In this report, the main target is to give a more specific understanding of the influences of Covid-19 on different perspectives; including, the death cases, the number of patients in ICU, the hospitalized cases, the negative cases, the positive cases as well as the total results of Covid-19. Our data source was found from the covidtracking.com called national-histroy.csv. The original data includes 17 variables. After the filtering, we obtain the variables that are more meaningful. Then, we split the data set into training set and test set; basically, focusing on the training set. The whole process of data mining on Covid-19 can be divided into two parts; numerical with three minor sectors, feature selection; model fitting; finding the most important variables. Meanwhile, the categorical part will contain three minor sectors with data preparing; feature selection & model fitting and model accuracy comparison.

To briefly look at our data set description. There are two predictors, deathIncrease and variable death direction which was made by us in order to make our analysis easier to compare for further research. Other than this, we have date; state; death increase; hospitalized cases (Currently, Increase,Cumulative); In ICU (Cumulative, Currently); Negative; Negative (Increased cases); Ventilator (Currently, Cumulative); Positive; PositiveIncrease; TotalTestResults; TotalTestResultsIncreas.

2. Problem of Interest

We also have some potential challenges which we solved at the data preparation process. Firstly, there are many variable columns that have empty values. Also, some of the variable columns show NAs at the start of covid-19. Our plan is to set these values as 0. Furthermore,

some variables have large values, and some have small values, we standardize the data set at the beginning. So that, the different digits of the numbers would not influence the data too much. Thirdly, variable date is not helping in predicting the dependent variable and variable death is highly correlated with the dependent variables. So, we delete both of them. During our data mining progress, we are concentrating on one question: which variables are most important in predicting the dependent variables; which are, death direction and death direction. DeathDirection will be used in the classification part; meanwhile, Death Increase will be used in the linear model.



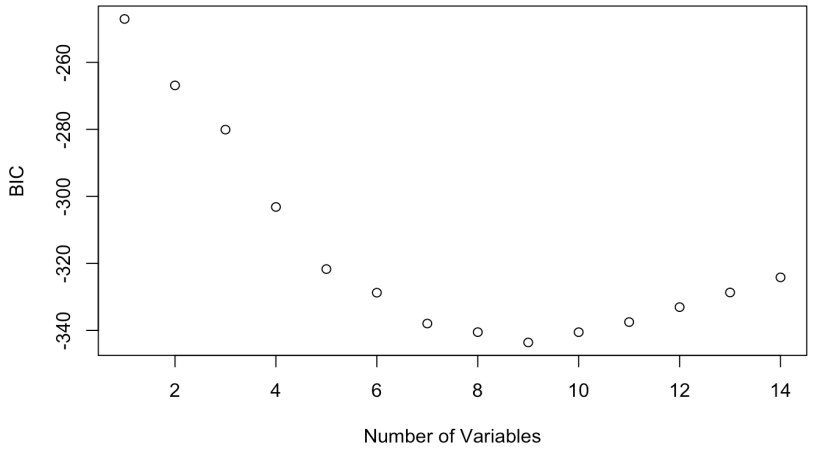
(Figure 1)

3. Process of Modeling

3.1 Linear - Death Increase

3.1.1 Feature Selection

In this modeling part, we started with using numeric dependent variable -- “DeathIncrease” . In the feature selection step, we apply three methods (BIC, Stepwise and Random Forest). In the first two methods, we accept all variables recommended. But in the Random Forest method, we accept six most important variables based on %IncMSE metric. This metric indicates how much our model accuracy decreases if we remove the variable.



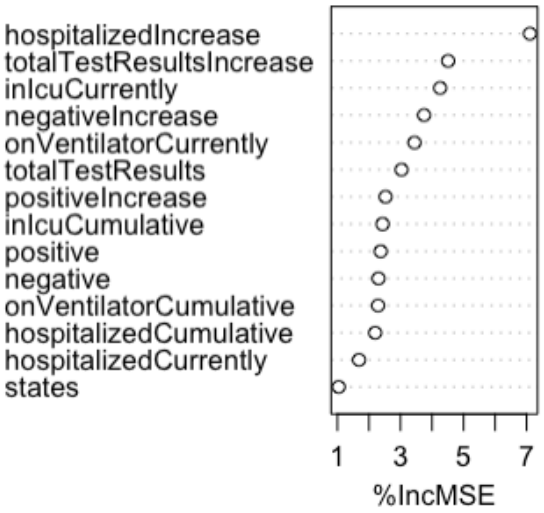
(Figure 2 - BIC)

```
[1] 9
          (Intercept)        inIcuCumulative   hospitalizedIncrease  hospitalizedCurrently
            0.001535654          -10.974443856             0.086549621           -0.855903276
hospitalizedCumulative onVentilatorCumulative  onVentilatorCurrently              positive
            5.493554713            4.249111496             0.644277193            4.386473675
      positiveIncrease       totalTestResults
```

(Figure 3 - BIC)

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              4.495e-04  2.697e-02   0.017 0.986722
inIcuCumulative         -1.028e+01  1.336e+00  -7.691 6.69e-13 ***
hospitalizedIncrease     9.196e-02  3.035e-02   3.030 0.002768 **
hospitalizedCurrently   -1.004e+00  1.954e-01  -5.139 6.60e-07 ***
hospitalizedCumulative   4.955e+00  7.864e-01   6.300 1.88e-09 ***
negative                 2.238e+00  1.482e+00   1.509 0.132773
negativeIncrease         1.153e-01  7.760e-02   1.486 0.138968
onVentilatorCumulative   2.697e+00  1.410e+00   1.913 0.057197 .
onVentilatorCurrently    8.087e-01  1.312e-01   6.165 3.89e-09 ***
positive                 4.645e+00  8.257e-01   5.625 6.27e-08 ***
positiveIncrease         9.996e-01  1.277e-01   7.830 2.90e-13 ***
totalTestResults        -4.073e+00  1.189e+00  -3.426 0.000745 ***
```

(Figure 4 - Stepwise)



(Figure 5 - Random Forest)

3.1.2 Modele Fitting

In this step, we use two techniques (K-Nearest Neighbors and Support Vector Machines) to find the test error of three subset variables selected by previous methods. Based on the results of KNN and SVM, the BIC model has the lowest test error compared to other subset variables selected by methods Stepwise and Random Forest. Then, we apply different SVM Kernels to the BIC model; we found that the Linear Kernel has the lowest test error compared to Radial kernel and Polynomial Kernel. The value of epsilon defines a margin of tolerance where no penalty is given to errors. Since the test error is lowest when epsilon = 0, every error is penalized.

KNN			
	BIC Model	Stepwise Model	Random Forest Model
K value	K = 3	K = 2	K = 2
Test Error	1.29E-05	0.000126991	0.000480264

(Figure 6)

SVM			
	BIC Model (Linear)	Stepwise Model	Random Forest Model
Epsilon	epsilon = 0	epsilon = 0	epsilon = 0
Cost	cost = 8	cost = 4	cost = 4
Test Error	8.98E-02	0.09333873	0.09970952

(Figure 7)

BIC Model (SVM)							
	Linear Kernel			Radial Kernel			Polynomial Kernel
Epsilon	0		Gamma	0.5		Degree	2
Cost	8		Cost	10		Cost	5
Test Error	0.08980813		Test Error	0.08987863		Test Error	0.1071624

(Figure 8)

Since the BIC selected subset is better than other subsets, we compare the test error of the BIC model to the Full model. Using the KNN method, We found the test error has reduced from 3.06E-03 to 1.29E-05.

	Test error
	KNN
Full model	3.06E-03
BIC model	1.29E-05

(Figure 9)

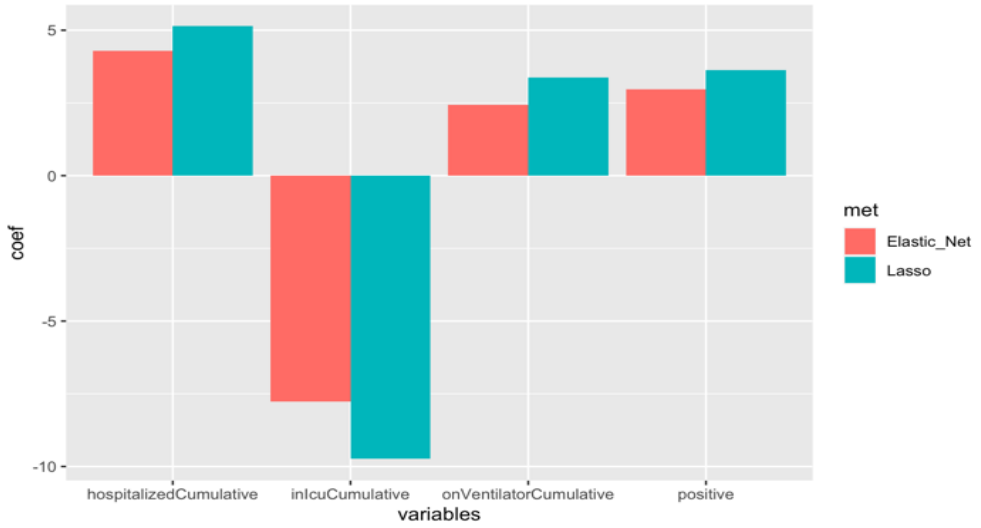
3.1.3 Finding the most important variables

In order to find the most important variables in predicting the dependent variable 'Death Increase' , we apply 3 techniques : Ridge, Lasso and Elastic Net to the BIC model. Ridge regularization will reduce the impact of features that are not important in predicting the dependent variable; Lasso regularization will eliminate many features and the resulting model is sparse. Elastic Net combined both feature elimination from Lasso and feature coefficient reduction from Ridge to improve our model predictions. According to the results, Lasso and Elastic Net have better performance than Ridge.

	Ridge	Lasso	Elastic Net
Best_lamda	0.0804617	0.000106366	0.000160923
Test Error	0.2744617	0.1969197	0.2020113

(Figure 10)

Moreover, based on the value of coefficients, we found variables ‘inIcuCumulative’,‘hospitalizedCumulative’, ‘onVentilatorCumulative’ and ‘positive’ are the most important variables in predicting ‘Death Increase’. These 4 variables belong to the category of cumulative variables.



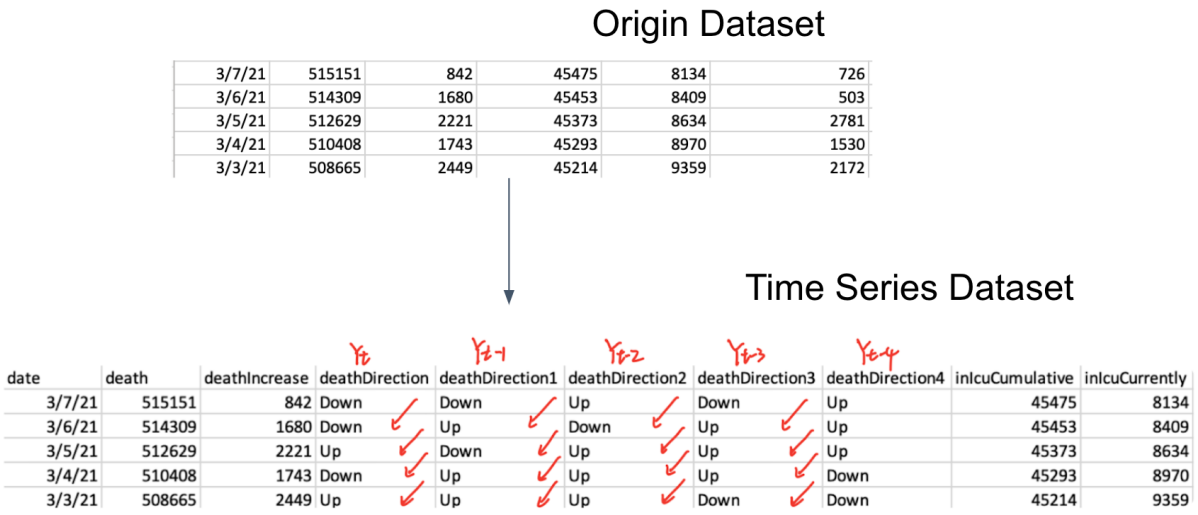
(Figure 11)

3.2 Categorical - Death Direction

3.2.1 Data Preparation

In terms of analysing categorical dependent variables, 2 steps of data preparation should be performed before the actual data mining process. The first step is to convert numerical dependent variables into categorical dependent variables. We calculate daily death changing rate by differencing "death increase" and get a list of positive and negative rates. Then the

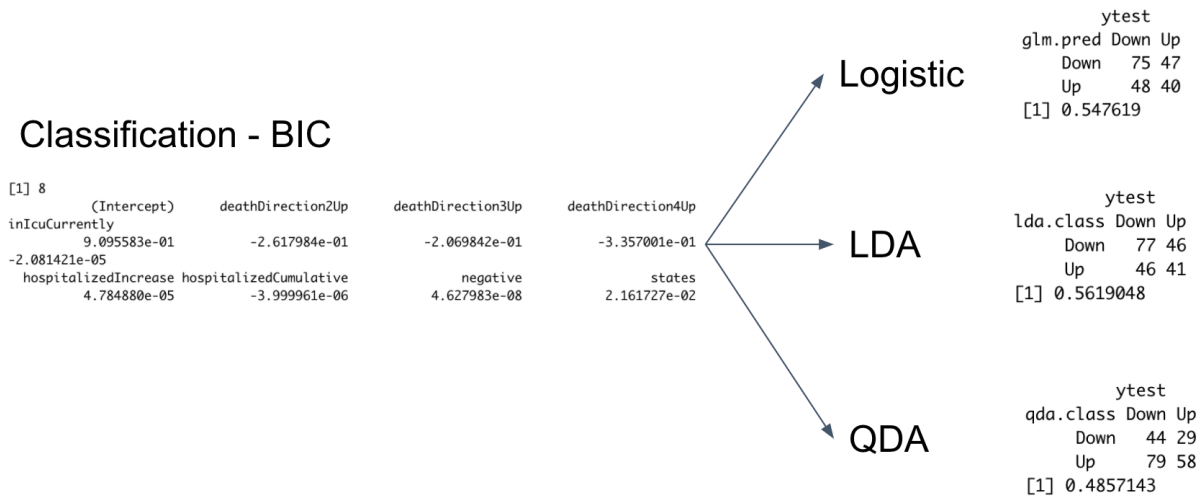
positive rate goes to "up trend", the negative rate goes to "down trend" and gets the list of "death Direction" to represent the daily death changing trend. The second step is to transform the dataset into a "Time Series" dataset in order to perform more accurate analysis based on previous data. Here we create four more dummy variables based on "death Direction", namely Yt. We define the new variables as Yt-1, Yt-2, Yt-3 and Yt-4, each of it is equal to the previous date value and the rest of the "undefined value" was considered as zero. After the above two steps, we finish our preparation of categorical dependent variable dataset.



(figure12)

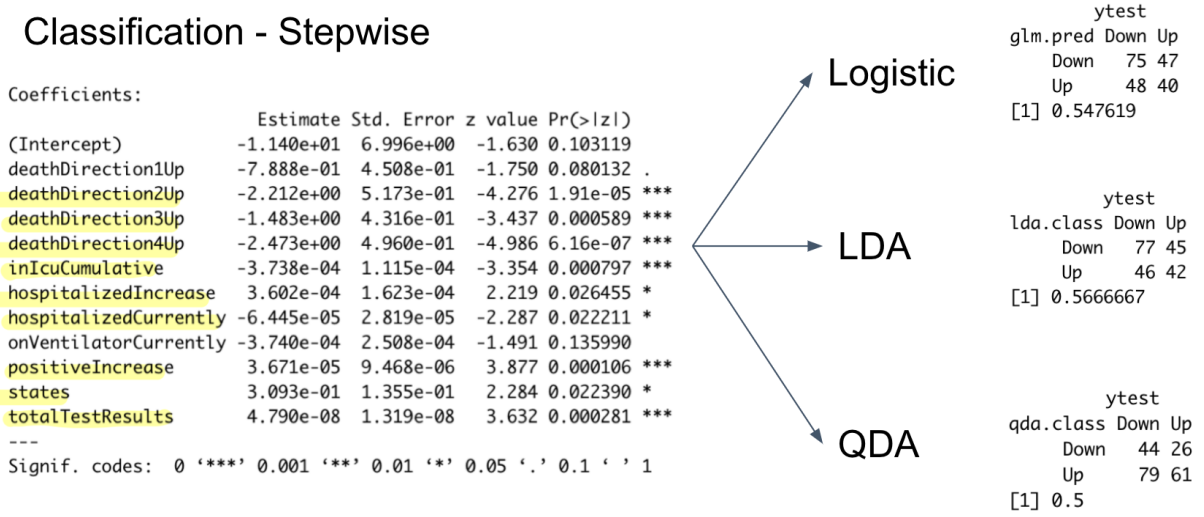
3.2.2 Feature Selection & Model Fitting

The first feature selection method we used is BIC. Eight variables are selected for the model fitting. And the accuracy of each model is Logistic 0.5476, LDA 0.5619 and QDA 0.5867. Figure 13 is the process of BIC and model selection.



(figure 13)

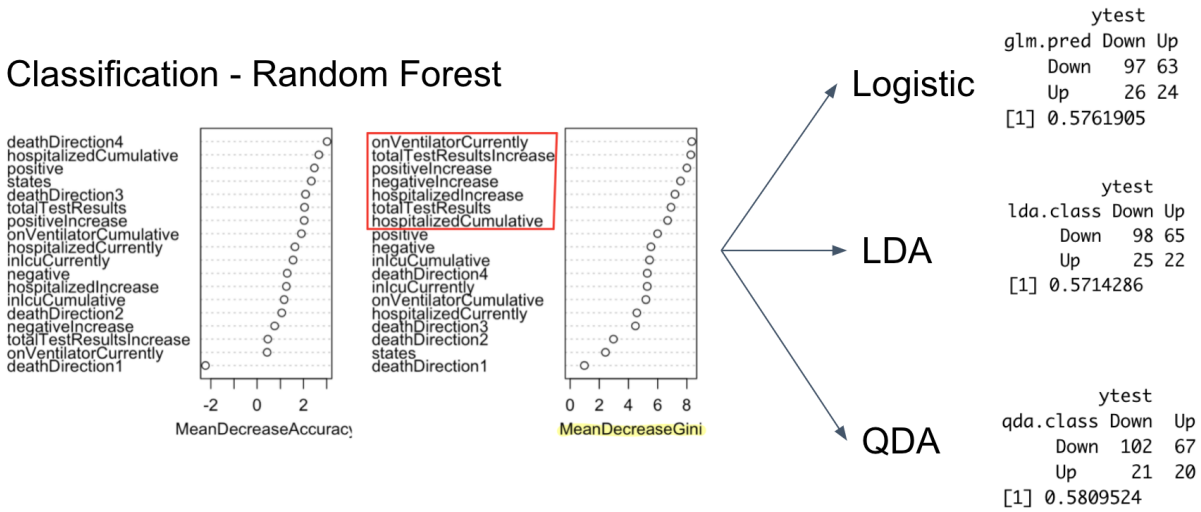
The next feature selection method we used is Stepwise. 9 variables which have p-value less than 0.05 which is significant in the model, namly, variables have "*" on the right hand side are selected for the model fitting. And the accuracy of each model is Logistic 0.5476, LDA 0.5667 and QDA 0.5. Figure 14 is the process of Stepwise and model selection.



(figure 14)

The last method we used is Random Forest. In figure 15, these two pictures provide rankings of variables based on different criteria. The left picture shows the ranking of different variables based on how much the model accuracy reduces if the variables were excluded from the model. The right picture shows the ranking of variables based on the reduction of the Gini

index that comes from splits over the variable. The Gini index is used to measure node purity. We chose MeanDecreaseGini as our reference plot and picked the first seven variables. After feature selection, we performed model fitting and got accuracy of Logistic model 0.5762, LDA model 0.5714, and QDA model 0.5809.



(figure 15)

3.2.3 Model Accuarcy Comparison

Based on the conclusion of Feature Selection and Model Fitting, in terms of classification dependent variable, we compare the accuracy of each model and the conclusion are listed in figure 16. We can see that Random Forest + Logistic, Random Forest + LDA, Random Forest + QDA have the highest accuracy among others. Also under comparison with Full Dataset, we can say that the model accuracy increases indeed. The final variables we selected for are listed in figure 16. However the accuracy is only near to 60% which means those models are not good enough to do the prediction.

In terms of this situation, what we considered was to do the same process that we performed on the dependent variable of creating Yt-1 so on and so forth. Independent variables X1(t-1), X1(t-2), X2(t-1), X2(t-2) should also be included in the model.

Model Accuracy Comparison

Accuracy	BIC	Stepwise	Random Forest	Full Dataset
Logistic	0.5476	0.5416	0.5762	0.4952
LDA	0.5619	0.5667	0.5714	0.4905
QDA	0.4857	0.5	0.5809	0.4667

Variables Selected

onVentilatorCurrently
totalTestResultsIncrease
positiveIncrease
negativeIncrease

(figure 16)

4. Summary

Back to the problem of interest: Which variables are most important in predicting the dependent variable (death Increase or death Direction)? Our conclusion is, in linear regression, “inIcuCumulative”, “hospitalizedCumulative”, “OnVentilatorCumulative” and “positive” would be the most influential variables on death increase. In classification problems, “onVentilatorCurrently”, ‘totalTestResultsIncrease”, “positiveIncrease” and “negativeIncrease” are the selected variables.