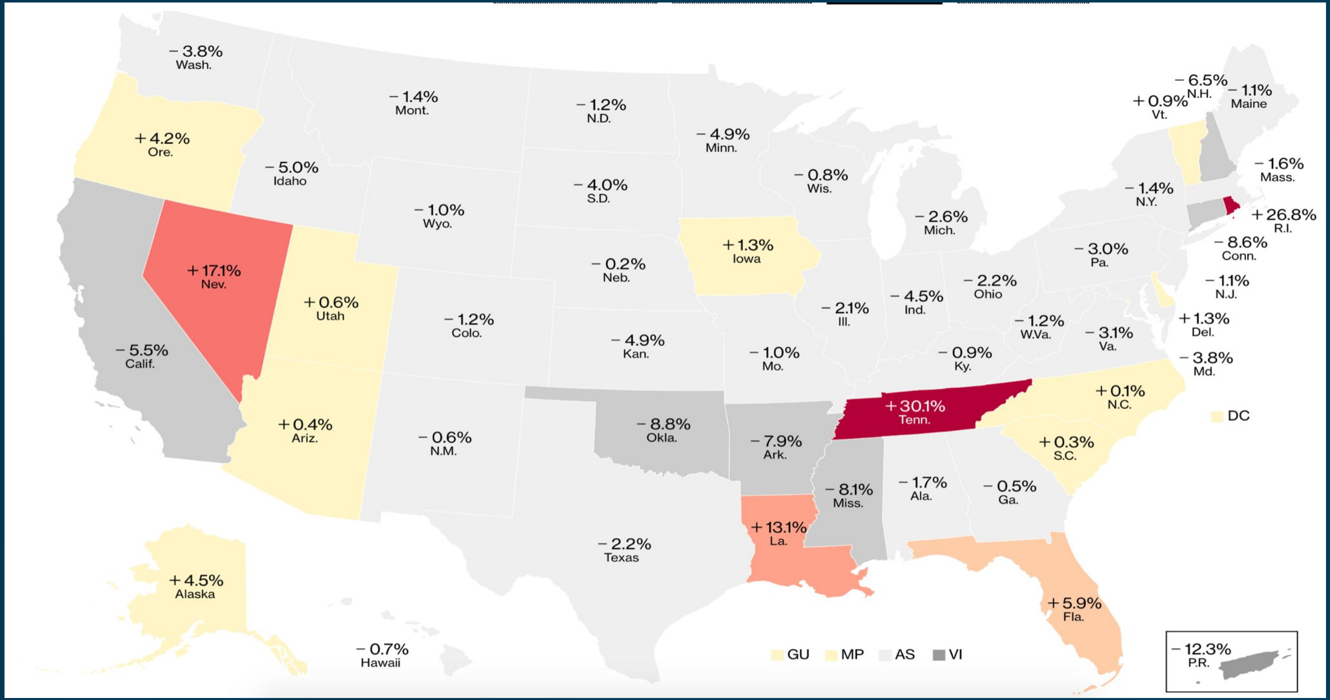
THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

U.S Covid-19 Data Mining Analysis



Present by: Yuwen Luo; Yihang Zhao; Runzhe Tang

Content

1. Introduction
2. Problem of Interest
3. Process of Modeling
4. Conclusion

3.1 Linear

3.1.1 Feature Selection

3.1.2 Model Fitting

3.1.3 Finding the most important variables

3.2 Classification

3.2.1 Data Preperation

3.2.2 Features Selection & Model Fitting

3.2.3 Model Accuray Comparison

1. Introduction

- Data Source: The Covid Tracking Project (https://covidtracking.com/data)
- The original data: national-history.csv
- originally data: 17 variables (15 variables which we are using)
- Split the data set into Training set and Test set
- Then, working on the training set

2. Problem of Interest

Which variables are most important in predicting the dependent variable (death Increase/ death Direction)?

Dependent Varibale

- death Increase (Linear Regression)
- death Direction (Classification)

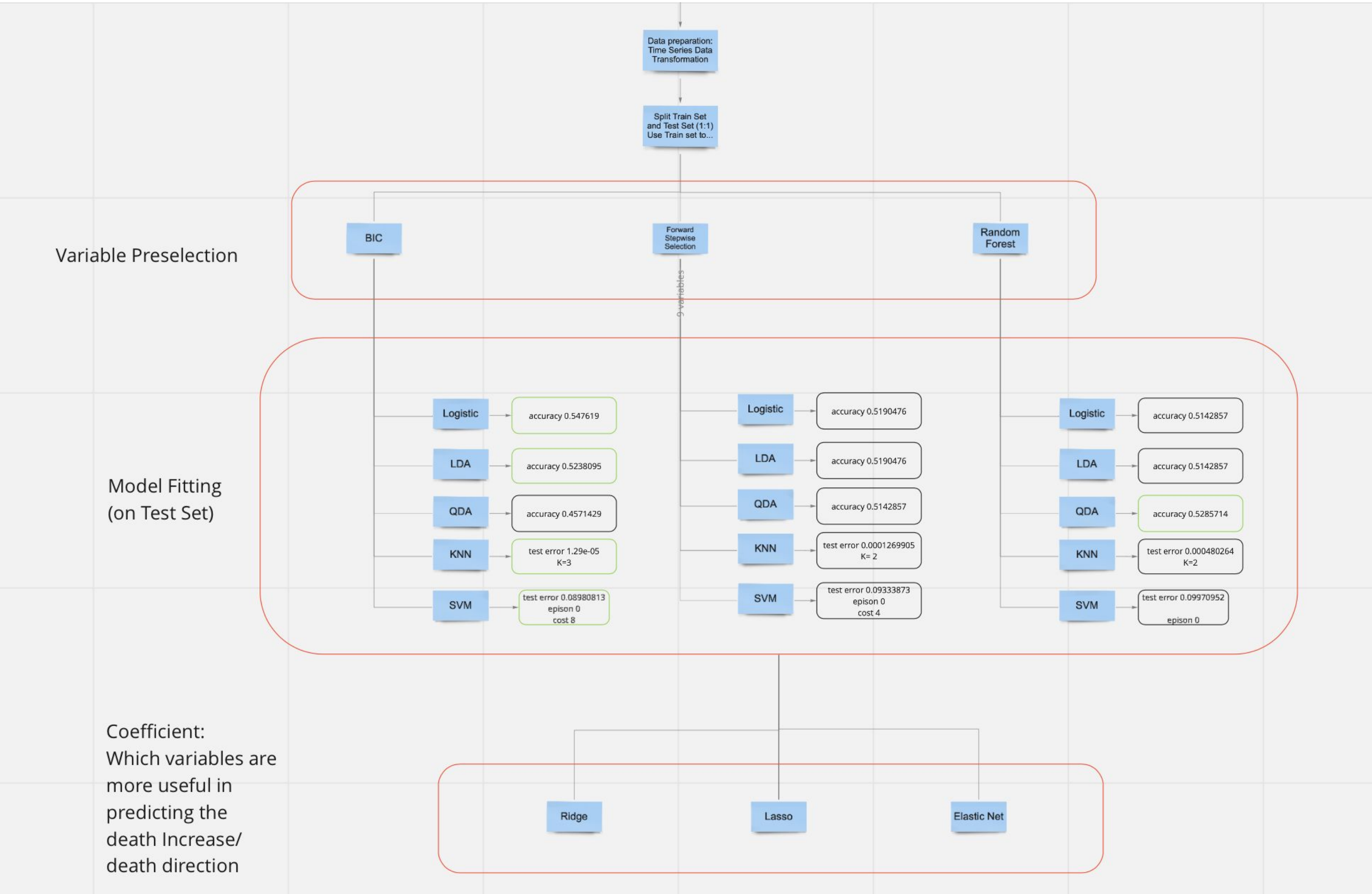
Dataset Description

- Two Predictors: deathincrease ; death direction
- date (YYYY/MM/DD)
- death
- states(name of state in the United States)
- hospitalized cases (Currently, Increase,Cumulative)
- In ICU (Cumulative, Currently)
- Negative
- Negative (Increased cases)
- Ventilator (Currently, Cumulative)
- Positive
- PositiveIncrease
- TotalTestResults
- TotalTestResultsIncreas

Potential Challenge & Solution

1. Many variable columns have empty values at the start of Covid-19, our plan is to set these values as 0.
2. Some variables have large values and some have small values, so we standardize the dataset at the beginning.
3. Variable ‘date’ is not helping in predicting the dependent variable and variable ‘death’ is highly correlated with the dependent variable. So, we delete both of them.

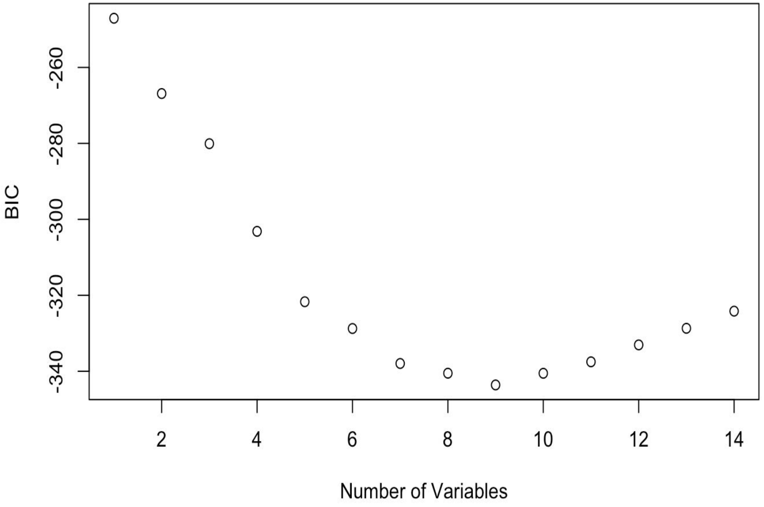
Road Map


Data preparation: Time Series Data Transformation
Split Train Set and Test Set (1:1) Use Train set to...
Variable Preselection
BIC
Forward Stepwise Selection
Random Forest
9 variables
Model Fitting (on Test Set)
Logistic
accuracy 0.547619
LDA
accuracy 0.5238095
QDA
accuracy 0.4571429
KNN
test error 1.29e-05 K=3
SVM
test error 0.08980813 epison 0 cost 8
Logistic
accuracy 0.5190476
LDA
accuracy 0.5190476
QDA
accuracy 0.5142857
KNN
test error 0.0001269905 K= 2
SVM
test error 0.09333873 epison 0 cost 4
Logistic
accuracy 0.5142857
LDA
accuracy 0.5142857
QDA
accuracy 0.5285714
KNN
test error 0.000480264 K=2
SVM
test error 0.09970952 epison 0
Coefficient: Which variables are more useful in predicting the death Increase/ death direction
Ridge
Lasso
Elastic Net

3.1.1 Feature selection (BIC)

Based on the pre-selection approach BIC,

the best model contains 9 independent variables, which include inIcuCumulative, hospitalizedCurrently, onVentilatorCurrently, positive, totalTestResults...



3.1.1 Feature selection (Stepwise)

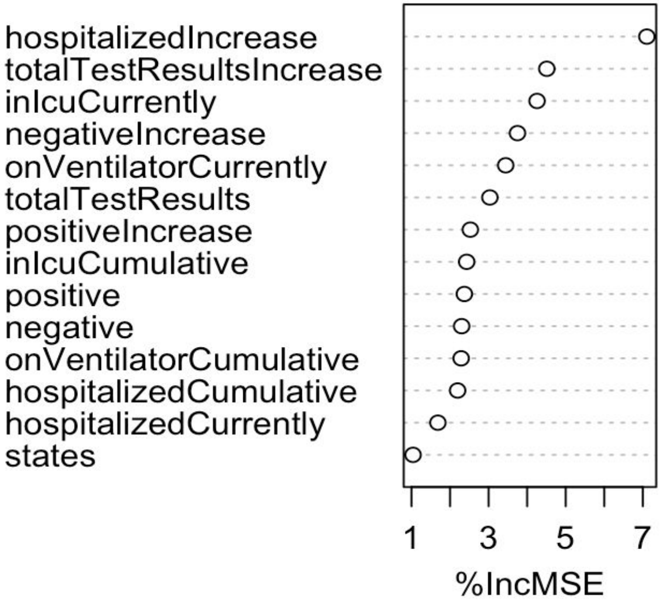
Based on the Stepwise approach, we find the model which contains 8 variables

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	***

3.1.1 Feature selection (Random Forest)

Because this is a linear model, we use the metrics %IncMSE to pick the variables we want. We use variable ‘totalTestResults’ and above 6 variables as our independent variables.



3.1.2 Model Fitting (KNN)

Based on the K-Nearest Neighbors, BIC Model has the lowest test error with K = 3

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

3.1.2 Model Fitting (SVM)

Based on Support Vector Machines technique, we found BIC model with epsilon = 0 and cost = 8 has the lowest error.

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

3.1.2 Model Fitting (SVM -- Kernel)

Then, we fitted BIC Model with 3 SVM kernels and found the linear kernel has the lowest test error. So, the decision boundary between classes is more likely to be linear.

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

3.1.2 Model Fitting

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

3.1.3 Finding most important variables

We use 3 methods: Ridge, Lasso and Elastic Net to find the coefficients of variables within the BIC model.

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

3.1.3 Finding most important variables

Lasso and Elastic Net give similar results which indicate variables

‘inIcuCumulative’,‘hospitalizedCumulative’,
‘onVentilatorCumulative’ and ‘positive’ are the most important varibles in predicting ‘deathincrease’.

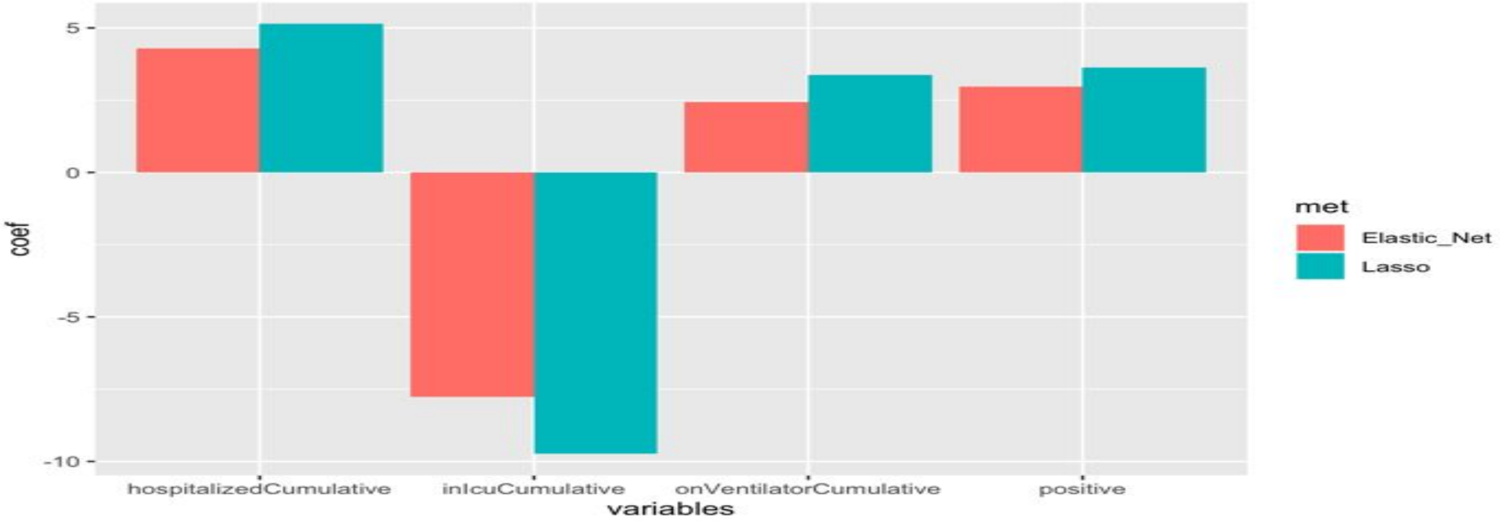
Lasso

(Intercept)	inIcuCumulative	hospitalizedIncrease	hospitalizedCurrently
0.002399336	-9.740818357	0.090069788	-0.748414966
hospitalizedCumulative	onVentilatorCumulative	onVentilatorCurrently	positive
5.135366588	3.381048660	0.609301088	3.630977597
positiveIncrease			
0.968815059			

Elastic Net

(Intercept)	inIcuCumulative	hospitalizedIncrease	hospitalizedCurrently
0.003479313	-7.763770724	0.099781111	-0.631721424
hospitalizedCumulative	onVentilatorCumulative	onVentilatorCurrently	positive
4.291198511	2.428861047	0.574974601	2.978116143
positiveIncrease			
0.850557145			

3.1.3 Finding most important variables



THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

3.2.1 Data Preparation

Origin Dataset

3/7/21	515151	842	45475	8134	726
3/6/21	514309	1680	45453	8409	503
3/5/21	512629	2221	45373	8634	2781
3/4/21	510408	1743	45293	8970	1530
3/3/21	508665	2449	45214	9359	2172

Time Series Dataset

			Y_t	Y_{t-1}	Y_{t-2}	Y_{t-3}	Y_{t-4}		
date	death	deathIncrease	deathDirection	deathDirection1	deathDirection2	deathDirection3	deathDirection4	inIcuCumulative	inIcuCurrently
3/7/21	515151	842	Down	Down	Up	Down	Up	45475	8134
3/6/21	514309	1680	Down	Up	Down	Up	Up	45453	8409
3/5/21	512629	2221	Up	Down	Up	Up	Up	45373	8634
3/4/21	510408	1743	Down	Up	Up	Up	Down	45293	8970
3/3/21	508665	2449	Up	Up	Up	Down	Down	45214	9359

3.2.2 Feature Selection & Model Fitting

Classification - BIC

```
[1] 8
         (Intercept)       deathDirection2Up       deathDirection3Up       deathDirection4Up
inIcuCurrently
        9.095583e-01           -2.617984e-01           -2.069842e-01           -3.357001e-01
-2.081421e-05
  hospitalizedIncrease hospitalizedCumulative                negative                  states
        4.784880e-05           -3.999961e-06            4.627983e-08            2.161727e-02
```

Logistic

```
         ytest
glm.pred Down Up
    Down   75 47
    Up     48 40
[1] 0.547619
```

LDA

```
          ytest
lda.class Down Up
     Down   77 46
     Up     46 41
[1] 0.5619048
```

QDA

```
          ytest
qda.class Down Up
     Down   44 29
     Up     79 58
[1] 0.4857143
```

3.2.2 Feature Selection & Model Fitting

Classification - Stepwise

```
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -1.140e+01  6.996e+00  -1.630 0.103119
deathDirection1Up    -7.888e-01  4.508e-01  -1.750 0.080132 .
deathDirection2Up    -2.212e+00  5.173e-01  -4.276 1.91e-05 ***
deathDirection3Up    -1.483e+00  4.316e-01  -3.437 0.000589 ***
deathDirection4Up    -2.473e+00  4.960e-01  -4.986 6.16e-07 ***
inIcuCumulative      -3.738e-04  1.115e-04  -3.354 0.000797 ***
hospitalizedIncrease  3.602e-04  1.623e-04   2.219 0.026455 *
hospitalizedCurrently -6.445e-05  2.819e-05  -2.287 0.022211 *
onVentilatorCurrently -3.740e-04  2.508e-04  -1.491 0.135990
positiveIncrease      3.671e-05  9.468e-06   3.877 0.000106 ***
states                3.093e-01  1.355e-01   2.284 0.022390 *
totalTestResults      4.790e-08  1.319e-08   3.632 0.000281 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Logistic

```
        ytest
glm.pred Down Up
    Down   75 47
    Up     48 40
[1] 0.547619
```

LDA

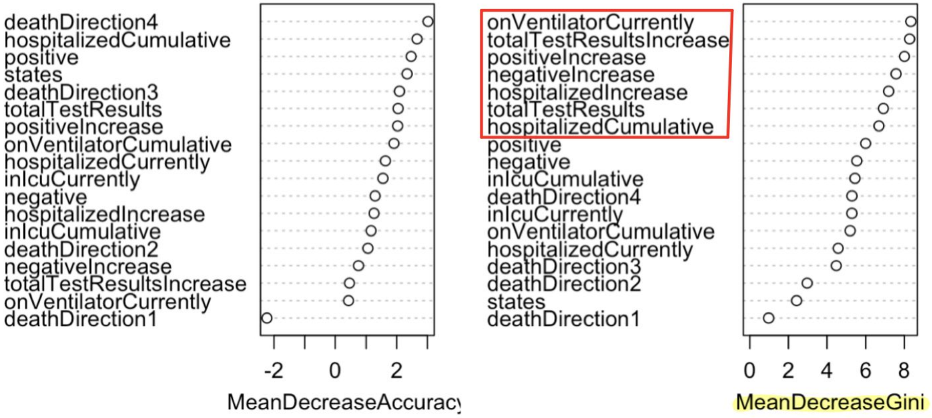
```
         ytest
lda.class Down Up
     Down   77 45
     Up     46 42
[1] 0.5666667
```

QDA

```
         ytest
qda.class Down Up
     Down   44 26
     Up     79 61
[1] 0.5
```

3.2.2 Feature Selection & Model Fitting

Classification - Random Forest



Logistic

```
          ytest
glm.pred Down Up
    Down   97 63
    Up     26 24
[1] 0.5761905
```

LDA

```
          ytest
lda.class Down Up
     Down   98 65
     Up     25 22
[1] 0.5714286
```

QDA

```
          ytest
qda.class Down  Up
     Down  102  67
     Up     21  20
[1] 0.5809524
```

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

3.2.3 Model Accuray Comparison

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

4. Conclusion:Respond to Problem of Interest

Which variables are most important in predicting the dependent variable (death Increase/ death Direction)?

Variable Seleted by Classification
deathDirection ~

onVentilatorCurrently
totalTestResultsIncrease
positiveIncrease
negativeIncrease

Variable Seleted by Linear Regression
deathIncrease ~

	Coefficients
inIcuCumulative	-9.7408
hospitalizedCumulative	5.1354
onVentilatorCumulative	3.381
positive	3.6309

Q&A

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC