Predictive Modeling: COVID-19

With LDA and QDA Machine Learning Techniques

Main Objectives

Main Objective: Process

1. **Exploratory Data Analysis** of Kaggle's Covid-19 Dataset
2. **Perform Variable Selection** to select the best variables for an LDA and QDA model
3. **Fit LDA and QDA** models to find which variables are associated with death by COVID-19
4. **Determine Accuracy** of the model

Main Objective: Research Question

Primary Research Question: Discover what variables in Kaggle's COVID-19 dataset are associated with death by COVID-19

Background: Logistic Reg., LDA, QDA

Connections between Log. Reg, LDA, QDA

LDA, QDA, and Logistic Regression, attempt to predict the probability of a categorical outcome variable based on a set of input variables. The primary difference between the three forms of regression lie in their assumptions:

- **Logistic Regression**: Does not have any distributional assumptions, but requires a categorical outcome variable.

Connections between Log. Reg, LDA, QDA

LDA, QDA, and Logistic Regression, attempt to predict the probability of a categorical outcome variable based on a set of input variables. The primary difference between the three forms of regression lie in their assumptions:

- **Logistic Regression**: Does not have any distributional assumptions, but requires a categorical outcome variable.

- **LDA**: Assumes that the predictor variables are normally distributed, that there is no heteroscedasticity in the outcome variable, and that the outcome variable is categorical.
    - This produces a linear decision boundary

Connections between Log. Reg, LDA, QDA

LDA, QDA, and Logistic Regression, attempt to predict the probability of a categorical outcome variable based on a set of input variables. The primary difference between the three forms of regression lie in their assumptions:

- **Logistic Regression**: Does not have any distributional assumptions, but requires a categorical outcome variable.

- **LDA**: Assumes that the predictor variables are normally distributed, that there is no heteroscedasticity in the outcome variable, and that the outcome variable is categorical.
  - This produces a linear decision boundary.

- **QDA**: A version of LDA allows each class to have its own covariance matrix.
  - This produces a quadratic decision boundary
  - **Covariance Matrix: A matrix that describes how much a set of features varies together

Main Objectives

Main Objective: Research Question

Primary Research Question: Discover what variables in Kaggle's COVID-19 dataset are associated with death by COVID-19

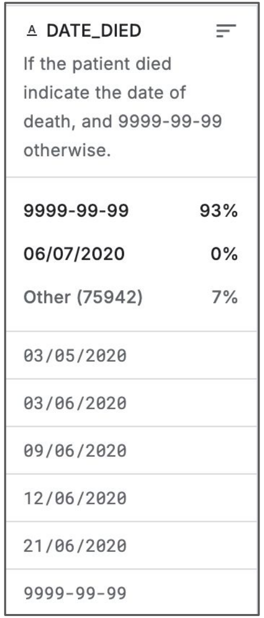
Main Steps

Primary Steps

1. Data Wrangling
2. Exploratory Data Analysis
3. Variable Selection
4. Predictive Modeling
5. Conclusions

Data Wrangling

Data Cleaning: Binary Response Variable

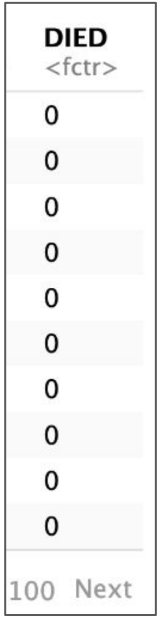


DATE_DIED

If the patient died indicate the date of death, and 9999-99-99 otherwise.

9999-99-99	93%
06/07/2020	0%
Other (75942)	7%

```
03/05/2020
03/06/2020
09/06/2020
12/06/2020
21/06/2020
9999-99-99
```

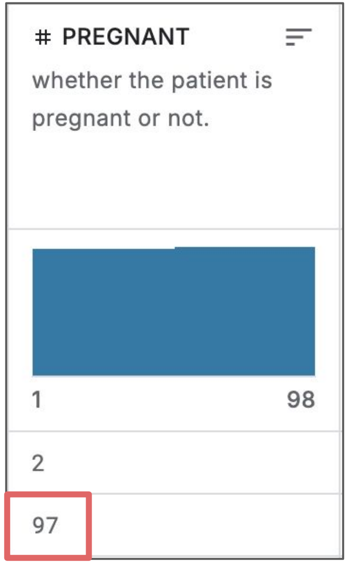


DIED <fctr>
0
0
0
0
0
0
0
0
0
0

- Both LDA and QDA take a binary response variable as an output variable
- We converted DATE_DIED to the binary response variable DIED

Data Cleaning: Missing Value Removal

Missing Value Example



PREGNANT

whether the patient is pregnant or not.

1

98

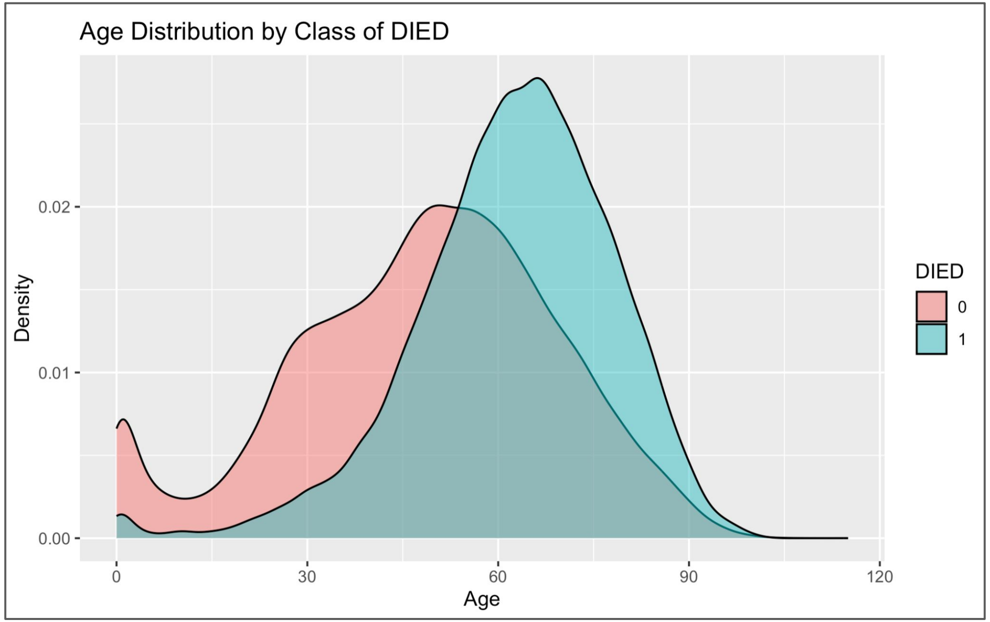
2

97

- Missing values (marked as 97, 98, 99) were removed from all rows
- This caused the SEX variable to only include the class of female, so it became useless after data removal and was dropped

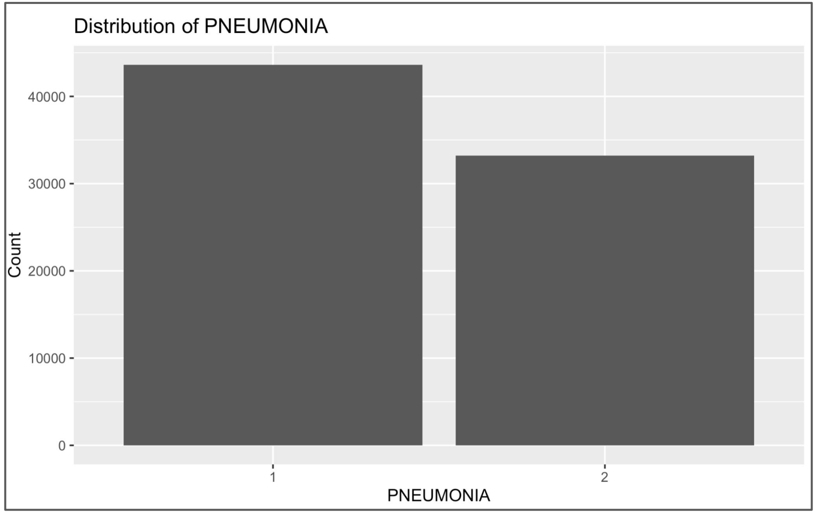
EDA: Distributions

Data Cleaning: Binary Response Variable



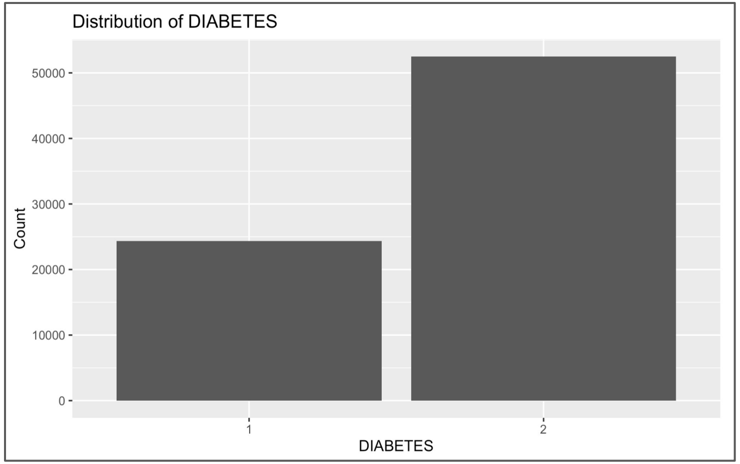
- Both distributions have a relatively bell shaped curve, suggesting normality
- The age of patients who died (mean = 62.44) is greater than the age of patients who did not die (mean = 48.5)

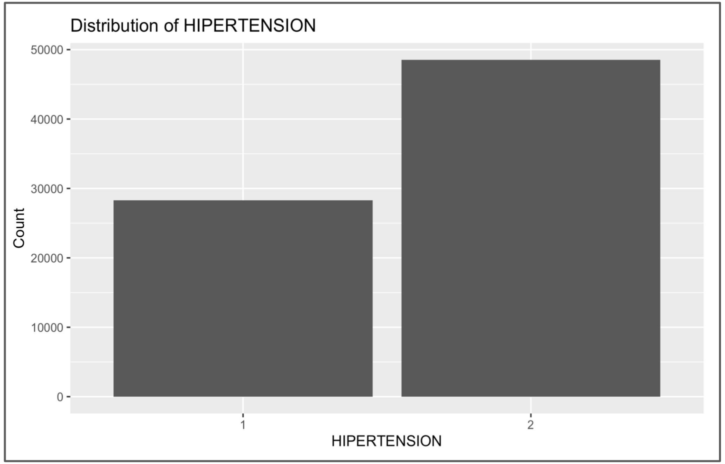
Exploratory Analysis: Bar Plots



- Next, we decided to understand distributions of variables that are often associated with death by COVID-19

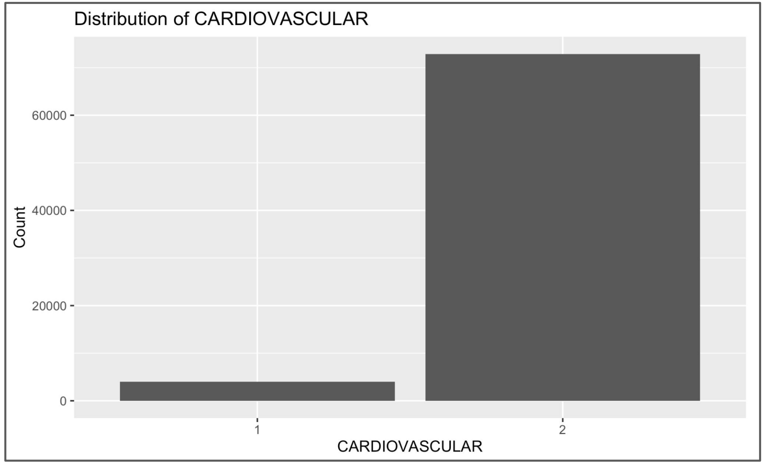
Exploratory Analysis: Bar Plots

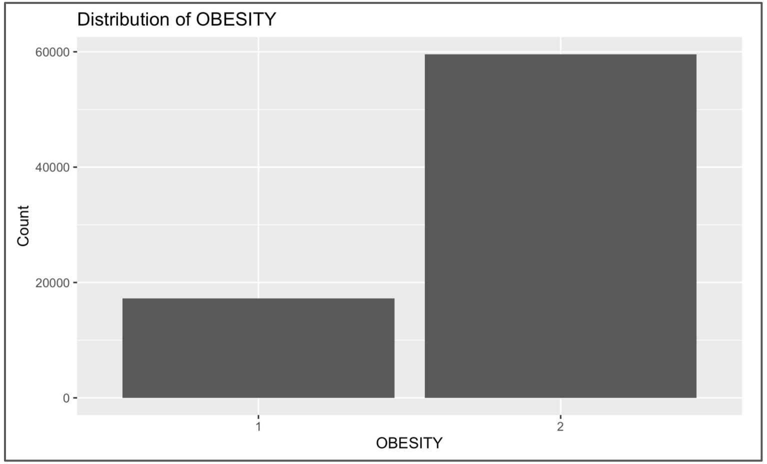




- Diabetes and hypertension have similar proportional representation within the population, suggesting that the two variables might be related

Exploratory Analysis: Bar Plots



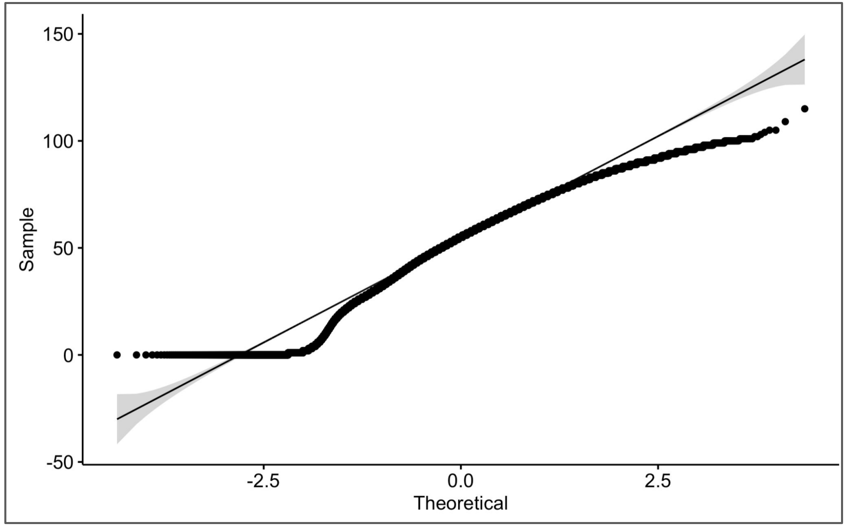


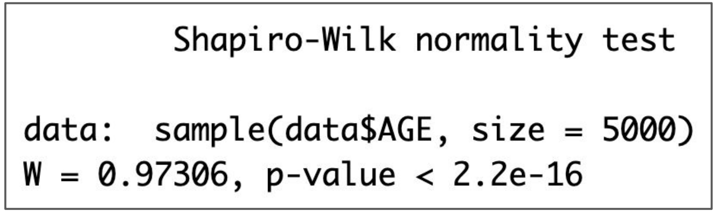
- As do cardiovascular and obesity

EDA: LDA + QDA Primary Assumptions

Assumption: Normality of Cont. Pred. Vars.

QQPLOT: AGE



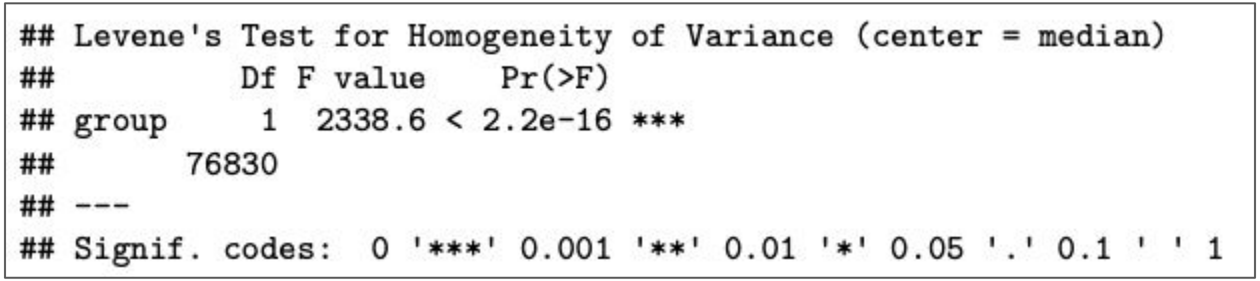


```
            Shapiro-Wilk normality test

data:  sample(data$AGE, size = 5000)
W = 0.97306, p-value < 2.2e-16
```

- LDA and QDA work best when continuous variables are normally distributed
- There was only one continuous variable in the dataset after pruning, age
  - It was not normal

Assumption: Homoscedasticity

Levene’s Test for Homoscedasticity

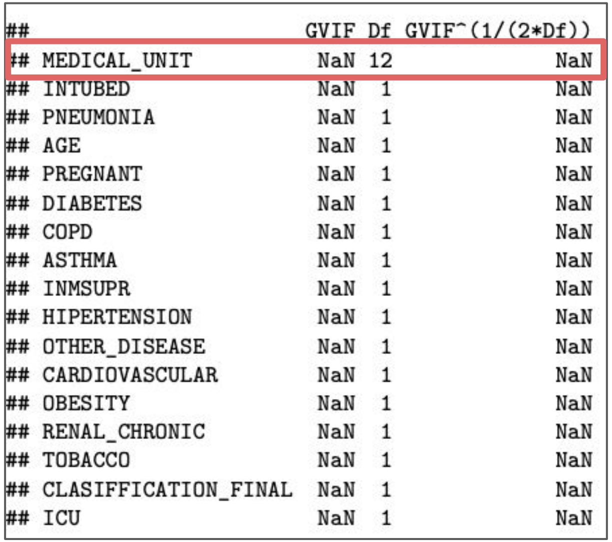


```
## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value    Pr(>F)
## group     1  2338.6 < 2.2e-16 ***
##       76830
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The most important assumption of LDA and QDA is Homoscedasticity (variance is constant among classes in the outcome variable)
- The assumption was not violated

Variable Selection

Prerequisite: Multicollinearity Check

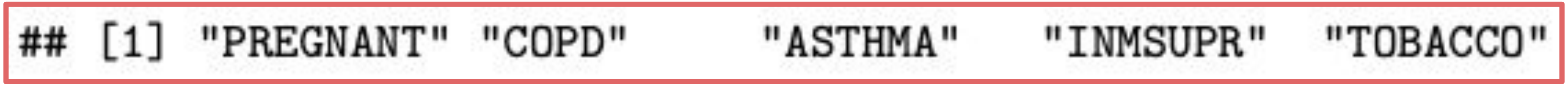


```
##                         GVIF Df GVIF^(1/(2*Df))
## MEDICAL_UNIT             NaN 12             NaN
## INTUBED                  NaN  1             NaN
## PNEUMONIA                NaN  1             NaN
## AGE                      NaN  1             NaN
## PREGNANT                 NaN  1             NaN
## DIABETES                 NaN  1             NaN
## COPD                     NaN  1             NaN
## ASTHMA                   NaN  1             NaN
## INMSUPR                  NaN  1             NaN
## HIPERTENSION             NaN  1             NaN
## OTHER_DISEASE            NaN  1             NaN
## CARDIOVASCULAR           NaN  1             NaN
## OBESITY                  NaN  1             NaN
## RENAL_CHRONIC            NaN  1             NaN
## TOBACCO                  NaN  1             NaN
## CLASIFFICATION_FINAL     NaN  1             NaN
## ICU                      NaN  1             NaN
```

- Initially, we tried to perform variable selection before checking for multicollinearity
- This produced bugs and bad results, so **we opted to remove multicollinear variables first with the vif function**

Prerequisite: Near Zero Variance Check

Variables output by nearZeroVar



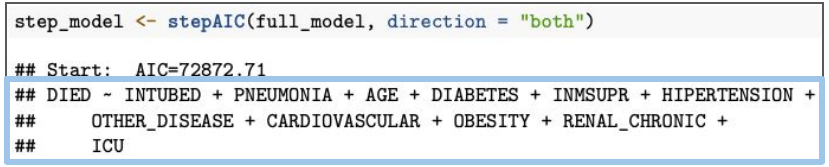
```
## [1] "PREGNANT" "COPD"      "ASTHMA"    "INMSUPR"   "TOBACCO"
```

- Similarly, we checking for near zero variance was necessary prior to variable selection and model fitting
- These variables, alongside the multicollinear variable (MEDICAL_UNIT), and other troublesome variables like SEX

Variable Selection

Stepwise Variable Selection

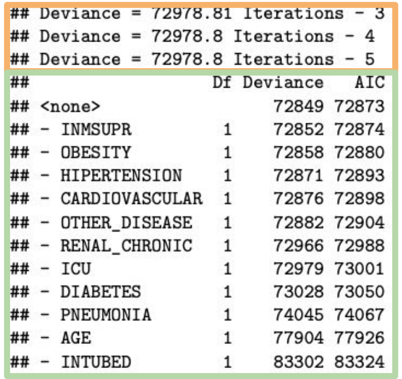
Stepwise Selection



```
step_model <- stepAIC(full_model, direction = "both")

## Start:  AIC=72872.71
## DIED ~ INTUBED + PNEUMONIA + AGE + DIABETES + INMSUPR + HIPERTENSION +
##     OTHER_DISEASE + CARDIOVASCULAR + OBESITY + RENAL_CHRONIC +
##     ICU
```

...Many Steps...

Final Model

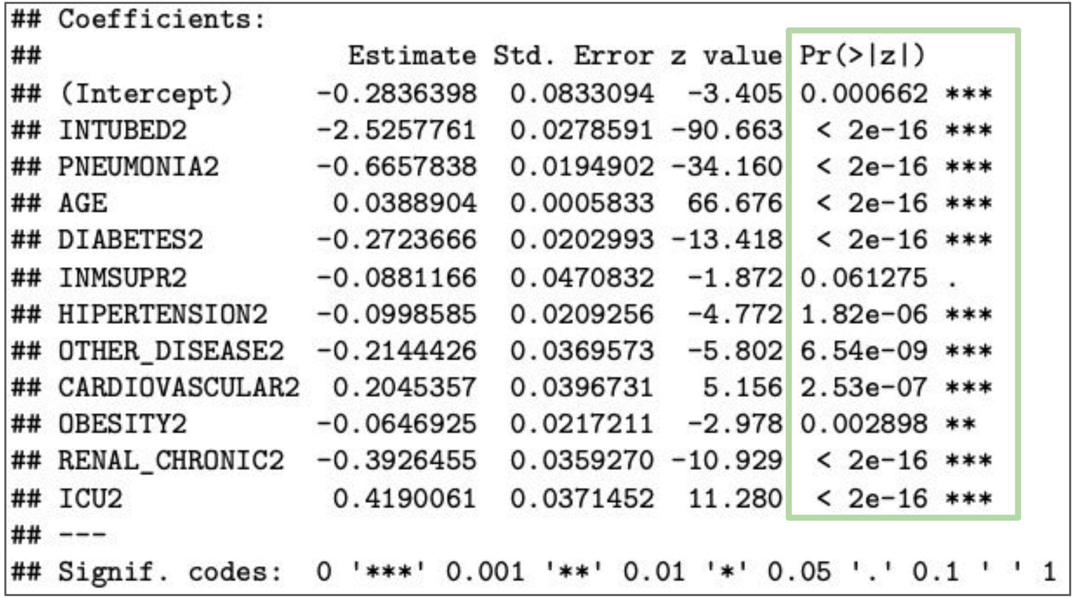


```
## Deviance = 72978.81 Iterations - 3
## Deviance = 72978.8 Iterations - 4
## Deviance = 72978.8 Iterations - 5
##
##                   Df Deviance   AIC
## <none>                 72849 72873
## - INMSUPR          1    72852 72874
## - OBESITY          1    72858 72880
## - HIPERTENSION     1    72871 72893
## - CARDIOVASCULAR   1    72876 72898
## - OTHER_DISEASE    1    72882 72904
## - RENAL_CHRONIC    1    72966 72988
## - ICU              1    72979 73001
## - DIABETES         1    73028 73050
## - PNEUMONIA        1    74045 74067
## - AGE              1    77904 77926
## - INTUBED          1    83302 83324
```

- Due to the large number of variables pruned in the earlier steps, there was no drop in deviance after variable removal

- This resulted in the new model being equalling the initial model

Stepwise Variable Selection



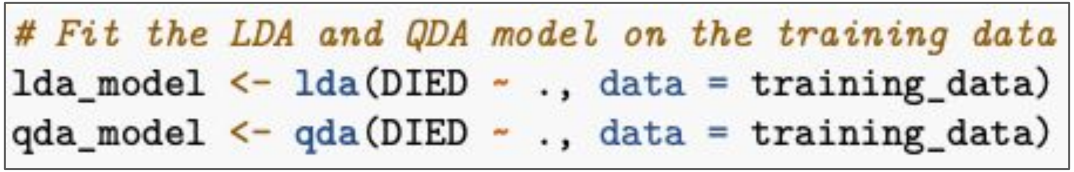
```
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)    
## (Intercept)      -0.2836398  0.0833094  -3.405 0.000662 ***
## INTUBED2         -2.5257761  0.0278591 -90.663  < 2e-16 ***
## PNEUMONIA2       -0.6657838  0.0194902 -34.160  < 2e-16 ***
## AGE               0.0388904  0.0005833  66.676  < 2e-16 ***
## DIABETES2        -0.2723666  0.0202993 -13.418  < 2e-16 ***
## INMSUPR2         -0.0881166  0.0470832  -1.872 0.061275 .  
## HIPERTENSION2    -0.0998585  0.0209256  -4.772 1.82e-06 ***
## OTHER_DISEASE2   -0.2144426  0.0369573  -5.802 6.54e-09 ***
## CARDIOVASCULAR2   0.2045357  0.0396731   5.156 2.53e-07 ***
## OBESITY2         -0.0646925  0.0217211  -2.978 0.002898 ** 
## RENAL_CHRONIC2   -0.3926455  0.0359270 -10.929  < 2e-16 ***
## ICU2              0.4190061  0.0371452  11.280  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Originally, this concerned us, but after viewing the statistical significance of each predictor, we decided to keep all predictors for accuracy

Fitting LDA and QDA Models

Fitting LDA and QDA Models

Fitting LDA and QDA Model



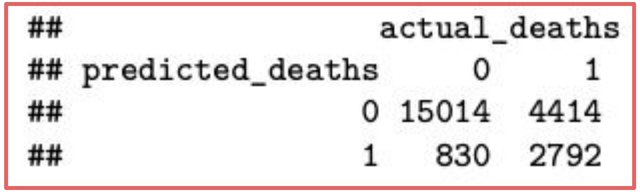
```r
# Fit the LDA and QDA model on the training data
lda_model <- lda(DIED ~ ., data = training_data)
qda_model <- qda(DIED ~ ., data = training_data)
```

- After a 70% Train 30% Test Split, we fit the LDA and QDA Models on the training data using the outcome variable of DIED

Choosing the Best Model

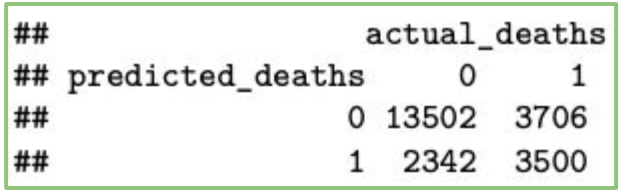
Accuracy, Recall, and Precision

LDA Confusion Matrix



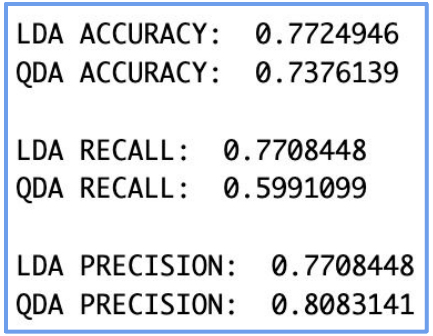
```
##                  actual_deaths
## predicted_deaths     0     1
##                0 15014  4414
##                1   830  2792
```

QDA Confusion Matrix



```
##                  actual_deaths
## predicted_deaths     0     1
##                0 13502  3706
##                1  2342  3500
```

Accuracy Metrics



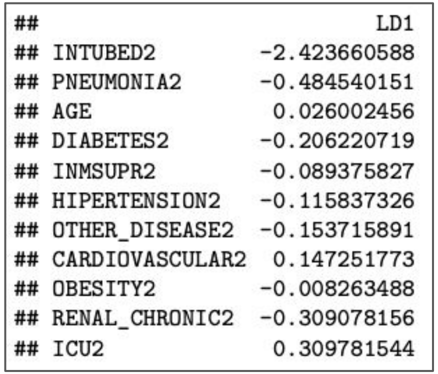
```
LDA ACCURACY:  0.7724946
QDA ACCURACY:  0.7376139

LDA RECALL:  0.7708448
QDA RECALL:  0.5991099

LDA PRECISION:  0.7708448
QDA PRECISION:  0.8083141
```

- Using the LDA confusion matrix and QDA confusion matrix, we produced accuracy metrics
- **We determined the LDA Model as the best model** due to it’s high accuracy (77%) and recall (77%)

Understanding the Best Model

Understanding the Best LDA Model

LDA Coefficients



```
##                          LD1
## INTUBED2        -2.423660588
## PNEUMONIA2      -0.484540151
## AGE              0.026002456
## DIABETES2       -0.206220719
## INMSUPR2        -0.089375827
## HIPERTENSION2   -0.115837326
## OTHER_DISEASE2  -0.153715891
## CARDIOVASCULAR2  0.147251773
## OBESITY2        -0.008263488
## RENAL_CHRONIC2  -0.309078156
## ICU2             0.309781544
```

$$\text{Outcome} = \text{Intubed} * \text{LD1}_{\text{Intubed}} + \text{Pneumonia} * + \text{LD1}_{\text{Pneumonia}} + \ldots + \text{ICU} * \text{ICU}_{\text{LD1}}$$

- The output of an LDA model is not as interpretable as other ML models

- If the **outcome** of the LDA algorithm is greater than the **cut off,** which is generally set to 0.5, the class of the outcome variable is set to 1 for the given observation
  - In our case, the DIED=1 (the patient died)

- Therefore, the negative coefficients reduce the probability of COVID-19, while the positive coefficients increase it

Understanding The Cut Off Variable

$$\text{Outcome} = \text{Intubed} * \text{LD1}_{\text{Intubed}} + \text{Pneumonia} * + \text{LD1}_{\text{Pneumonia}} + \ldots + \text{ICU} * \text{ICU}_{\text{LD1}} \; ?> \text{Cut Off}$$

The **Cut Off** parameter in LDA and QDA models determines the point (probability) at which an observation is considered one class of the outcome variable or another.

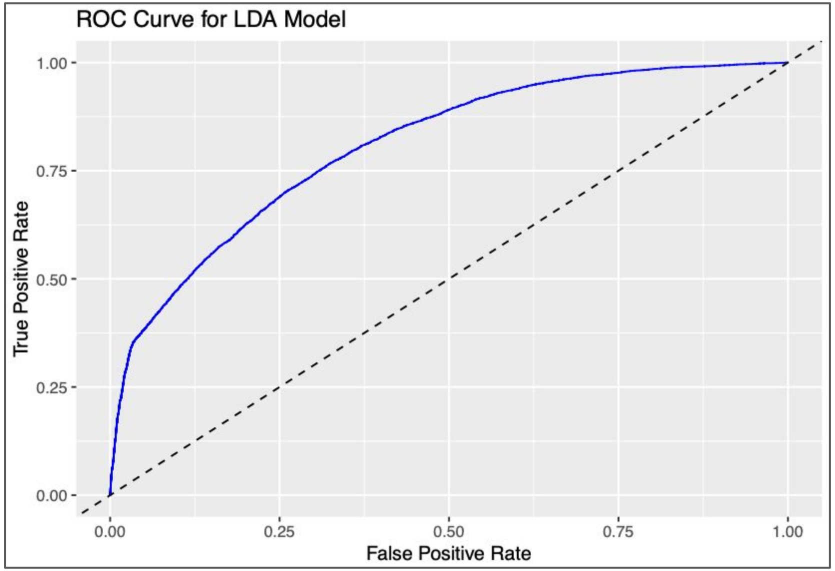
If the cutoff is increased

- Fewer deaths predicted → increased number of false negatives → reduced recall
- Fewer deaths predicted → reduced number of false positives → increased precision

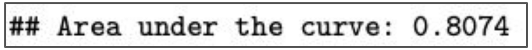
ROC Curve and Area Under The Curve

Final Test: ROC and AUC

ROC Curve



AUC Value



```
## Area under the curve: 0.8074
```

- The blue ROC Curve represents the True Positive Rate against the False Positive Rate
- It is compared to the dashed line, which represents typical random chance (an AUC of 0.5)

Results of ROC / AUC and Conclusion

Conclusion

AUC Value	80%
Accuracy	77%

- The final LDA model had an AUC of 80%, meaning that the model performs 30% better than random chance
    - An AUC value greater between 80% and 90% is considered ‘excellent’

- The final model had an accuracy of 77%, meaning that the model correctly predicts death by COVID-19 77% of the time