Modeling Survival Rates of COVID-19 Using Neural Networks

By Christian Stone

April 25, 2021

Main objective:

The main objective of this report is to show the results of using neural networks to build models in order to predict the likelihood of a person dying from COVID-19. Many different stakeholders can benefit from such information. It can give hospitals an idea of how to use resources knowing what patients may be most at risk and how much medical help may be required from an outbreak. This information can also be helpful in managing expectations of having the virus. It can be helpful to know the likelihood of survival as it can be extremely important for people to know how likely they would be to survive catching the virus.

While this report is being made as an exercise in neural networks, it will treat the scenario as if it was to be used in a professional setting.

Data Description:

The dataset used in this analysis consists of various data of over 8 million people who contracted, or where thought to have contracted COVID-19. The features in the dataset consisted of important dates within one's timeline with COVID-19, demographic information such as sex, age group, and race/ethnicity. The data also included binary information as to if the person was hospitalized, put in an Intensive Care Unit, if they died, and if they had a medical condition.

Data Exploration:

With over 8 million samples, the first step in data exploration was finding ways to reduce this number to more easily create and train the models. In order to do this, any data sample that had missing or uncertain data was taken out.

Most of the data features were Boolean and so they were ready to be used for the models. In order to implement the race/ethnicity data, however; one hot encoding was used to create separate Boolean features for each individual race/ethnicity. Since there was a logical order to the age grouping data, label encoders were used.

An important aspect to the data is the imbalance in the results of deaths. Only about 5% of cases resulted in deaths which can result in complications when trying to create an accurate model. As a way to get around this issue, some of the models were trained with oversampled data that was created by using the borderline SMOTE method.

A heat map was generated in order to determine what attributes are correlated with having more severe results from a case of COVID-19:

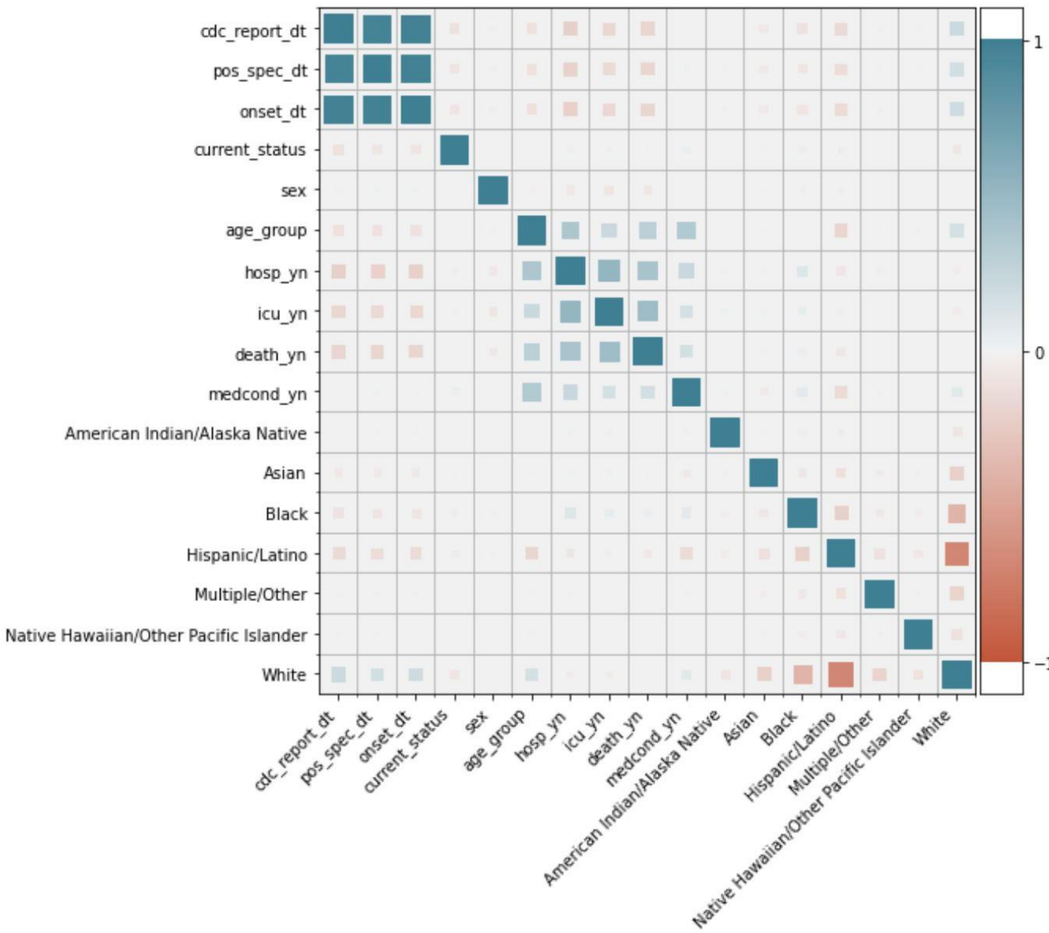

Figure 1: heat map showing the correlations of different data from COVID-19 cases

Model Descriptions:

4 different models were developed in order to compare how modifying the data and how modifying the model architecture could impact results. The first model used sigmoid activation and had a single layer of 12 nodes. The second model was identical to the first, except that it used the modified

oversampled dataset. The third model, instead used the tanh activation function and was built with 2 layers of 12 nodes instead of one layer. The final model used the modified dataset and more complex architecture from the third model.

All data was normalize using the standard scalar technique in order to ensure that any extreme values would not cause an excessive impact to the training of weights. All models were trained over 100 epochs of variations of the training data. Since accuracy would not be an ideal measurement of a model's success on its own (a model that would predict nobody to ever die would be about 95% accurate) all models were trained to try to maximize precision and recall as well.

Recommended Model:

The following are the results from each of the different models:

Figures 1-4: test scores from each of the 4 models

Model 1	Test Scores
Accuracy	.9601
Precision	.6127
Recall	.3881

Model 2	Test Scores
Accuracy	.9316
Precision	.9025
Recall	.9675

Model 3	Test Scores
Accuracy	.9608
Precision	.6205
Recall	.3881

Model 4	Test Scores
Accuracy	.9324
Precision	.9056
Recall	.9659

In terms of numerical results, models 1 and 3 and models 2 and 4 had nearly identical scores and can be treated somewhat interchangeably. While models 1 and 3 have the higher accuracy scores, the precision and recall scores show that these models may actually perform poorly and may benefit from the high probability that a COVID-19 case will not result in death.

The two models that used the oversampled data performed better in several categories and therefore would be recommended in most cases over the other two. However, this report only deals with techniques involving neural networks. A previous report found that using simple decision trees provided similar model scores.

Key Findings:

One interesting result from these models is the use of oversampled data had a large positive impact on the precision and recall scores of the models. Models 2 and 4 used data modified by the borderline SMOTE technique and shared similar high scores in these categories compared to the other models.

Surprisingly, the architecture of the models appears to have had an insignificant effect on the training of the models. Models 1 and 3 and models 2 and 4 were set up completely different from one another, and yet the pairs ended up with nearly identical scores between them.

While each model was trained using 100 epochs, validation loss data shows that this was far above the amount needed to reach these ranges of scores.

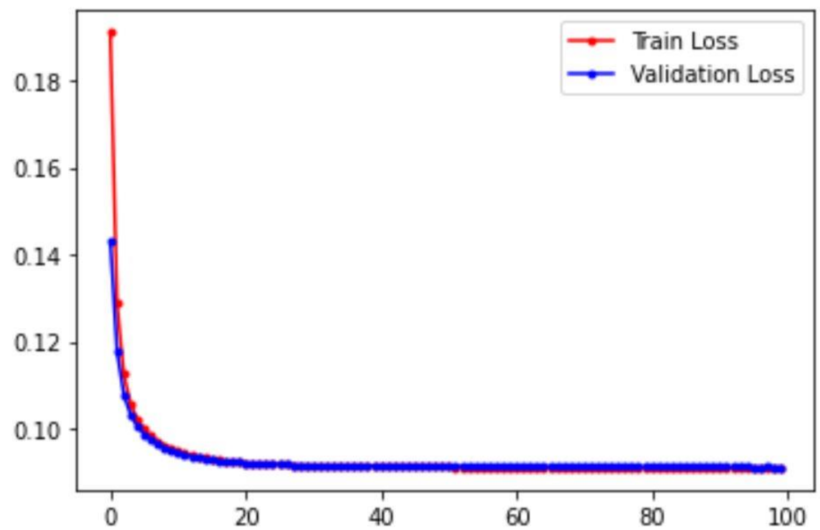

Figure 2: shows that the training and validation losses leveled out around 20 epochs for the training of model 1

At a certain point, additional epochs were not training the models to be any better and the different scores moved around a certain range.

Next Steps:

There are many different ways these models can be modified in order to try to optimize them further. The scores proved that modifying the data caused great changes in the results of the models and so different techniques such as eliminating outliers or even finding more reintroducing some of the data that was taken out initially could be useful in improving the models.

While this analysis focused on manipulating the data and changing the architecture of the models, further work can be done in order to train the models differently. While binary cross-entropy was used to calculate loss for each of the models, there are many different other methods that could be used as a means to train the models in a different way.

In the end, however; neural networks may not be the optimal tools for model generation of this dataset. Previously, models were created that scored better than the ones in this analysis while using far simpler techniques. As neural networks become more complex in order to optimize models, they also become harder to be explained to others and become harder to replicate. If simpler models can be created that perform well, it is possible that there is little benefit to using techniques as complicated as neural networks.