Jordan Roberts

July 31, 2020

BAN 530

Predictive Model Building

**Country Model Building**

The “full\_grouped” dataset was the first of the three sets that I built models for. This dataset shows each country’s daily statistics in terms of cases, deaths, and recoveries. I also added in averages cases and deaths for each day on record, as well as columns designating if the country was over or under that global average for recorded cases or deaths.

In JMP Pro 15, I created a validation column splitting the data into three groups: Training, Validation, and Test. The models use the training and validation data to predict test data, which is not used in the model’s creation. The better it can predict the unseen test data, the better the predictive abilities and more reliable the model.

I created models of the following types: adaptive lasso, adaptive elastic net, neural network (boosted and unboosted), and random forest (boosted and unboosted). Creating predictions for both deaths and confirmed cases, I then compared all the models to each other to see which served as the best predictor for the unseen Test data:

*Confirmed Cases*

| **Validation** | **Creator** |  | **Entropy RSquare** | **Generalized RSquare** | **Misclassification Rate** | **N** | **AUC** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 2 | Neural |  | 0.6562 | 0.7404 | 0.0531 | 5760 | 0.9882 |
| 2 | Neural |  | 0.5355 | 0.6312 | 0.0710 | 5760 | 0.9936 |
| 2 | Fit Generalized Adaptive Elastic Net |  | 0.4973 | 0.5945 | 0.0795 | 5760 | 0.9712 |
| 2 | Fit Generalized Adaptive Lasso |  | 0.4953 | 0.5924 | 0.0795 | 5760 | 0.9703 |
| 2 | Bootstrap Forest |  | 0.5011 | 0.5982 | 0.0892 | 5760 | 0.9793 |
| 2 | Boosted Tree |  | 0.6018 | 0.6924 | 0.0712 | 5760 | 0.9772 |

*Confirmed Deaths*

| **Validation** | **Creator** |  | **Entropy RSquare** | **Generalized RSquare** | **Misclassification Rate** | **N** | **AUC** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 2 | Fit Generalized Adaptive Lasso |  | 0.5202 | 0.6042 | 0.0484 | 5760 | 0.9388 |
| 2 | Fit Generalized Adaptive Elastic Net |  | 0.5240 | 0.6079 | 0.0476 | 5760 | 0.9370 |
| 2 | Neural |  | 0.4058 | 0.4892 | 0.0655 | 5760 | 0.9748 |
| 2 | Neural |  | 0.7520 | 0.8116 | 0.0224 | 5760 | 0.9944 |
| 2 | Bootstrap Forest |  | 0.3274 | 0.4049 | 0.0524 | 5760 | 0.9772 |
| 2 | Boosted Tree |  | 0.3843 | 0.4664 | 0.0714 | 5760 | 0.9691 |

In many cases with model comparison, we look to the R-Squared function to determine predictive power. This tells us how close the model’s predictions were to the actual test data in terms of value differences. However, I’ve chosen to look at this as a classification problem, meaning that the value will either be “YES” or “NO”, depending on whether the country is above or below the global average for a certain date. To see the predictive capabilities of our models in these terms, we look at the ROC curve. This tells us the true and false positive rates, or how well it predicts that an actual YES will be a YES. The larger the area under the ROC curve (AUC), the better the model was at making predictions.

For confirmed cases, our boosted and unboosted neural networks outperformed all the other models. In addition, they also had the lowest misclassification rates, meaning they made wrong predictions the least. For deaths, our boosted neural network again outperformed the rest, but our unboosted fell behind. In either case, I can confidently say that the boosted neural network was the most accurate, so this will be my model for analysis.

*Confirmed Cases*

| **Validation** | **Creator** |  | **Entropy RSquare** | **Generalized RSquare** | **Misclassification Rate** | **N** | **AUC** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 2 | Fit Generalized Adaptive Lasso |  | 0.4384 | 0.5329 | 0.0800 | 8039 | 0.9895 |
| 2 | Fit Generalized Adaptive Elastic Net |  | 0.4377 | 0.5323 | 0.0789 | 8039 | 0.9879 |
| 2 | Neural |  | 0.4572 | 0.5521 | 0.0988 | 8039 | 0.9587 |
| 2 | Neural |  | 0.7248 | 0.7961 | 0.0262 | 8039 | 0.9914 |
| 2 | Bootstrap Forest |  | 0.3724 | 0.4637 | 0.0858 | 8039 | 0.9760 |
| 2 | Boosted Tree |  | 0.5509 | 0.6432 | 0.0546 | 8039 | 0.9851 |

*Confirmed Deaths*

| **Validation** | **Creator** |  | **Entropy RSquare** | **Generalized RSquare** | **Misclassification Rate** | **N** | **AUC** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 2 | Fit Generalized Adaptive Lasso |  | 0.5863 | 0.6608 | 0.0429 | 8039 | 0.9802 |
| 2 | Fit Generalized Adaptive Elastic Net |  | 0.5850 | 0.6596 | 0.0429 | 8039 | 0.9803 |
| 2 | Neural |  | 0.5079 | 0.5860 | 0.0535 | 8039 | 0.9919 |
| 2 | Neural |  | 0.7025 | 0.7652 | 0.0409 | 8039 | 0.9959 |
| 2 | Bootstrap Forest |  | 0.4577 | 0.5360 | 0.0551 | 8039 | 0.9799 |
| 2 | Boosted Tree |  | 0.5111 | 0.5891 | 0.0516 | 8039 | 0.9869 |

For “covid\_19\_clean\_complete”, our results are similar. Boosted neural networks seem to be the best models for our predictions, though they are all quite effective. We will do our analysis on country data using boosted neural networks. All random seeds used are 123 throughout this project.

**Country Data Analysis**

*(full\_group) Confirmed Cases*

| **Column** | **Main Effect** | **Total Effect** |  |
| --- | --- | --- | --- |
| New cases | 0.079 | 0.741 |  |
| Confirmed | 0.062 | 0.66 |  |
| Active | 0.046 | 0.576 |  |
| New recovered | 0.027 | 0.342 |  |
| Recovered | 0.027 | 0.328 |  |
| New deaths | 0.008 | 0.037 |  |
| WHO Region | 0.005 | 0.022 |  |
| Deaths | 0.001 | 0.004 |  |

*(full\_group) Confirmed Deaths*

| **Column** | **Main Effect** | **Total Effect** |  |
| --- | --- | --- | --- |
| New deaths | 0.088 | 0.83 |  |
| New recovered | 0.049 | 0.659 |  |
| Deaths | 0.042 | 0.642 |  |
| Confirmed | 0.022 | 0.399 |  |
| Active | 0.01 | 0.093 |  |
| New cases | 0.01 | 0.074 |  |
| Recovered | 0.007 | 0.049 |  |
| WHO Region | 0.003 | 0.018 |  |

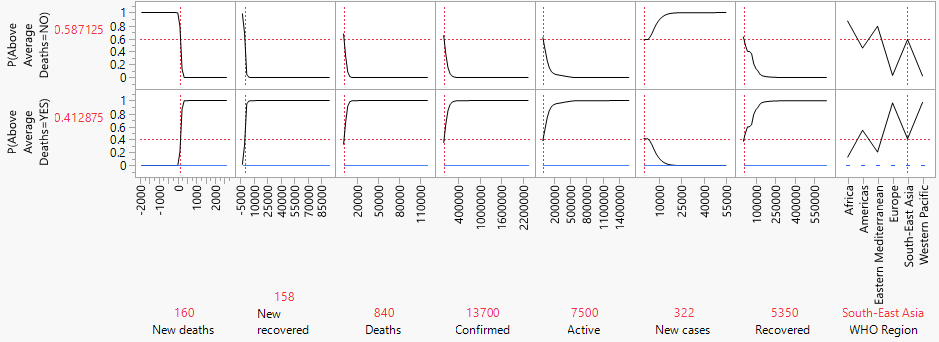
For country data, it appears the number of new cases is the best predictor of a country’s status as above or below the average for the world. In other words, a high increase from day to day tends to signify runaway COVID cases. Countries that have big spikes are likely to cross the threshold and struggle to contain the virus. Similarly, for deaths, above average deaths are prefaced by high spikes in deaths. Again, this suggests that once the ball gets rolling, countries are likely to be unable to slow or stop the spread. Better preventative measures may have helped stem the effects.

full\_grouped Above Average Cases



WHO Region is low on the list, but still holds some significance. The Americas and Europe have the highest chance of being above the worldwide average up until a certain point, where the chance reaches high levels and plateaus. Deaths are the least predictive for confirmed cases, but it is almost reversed for average deaths.

full\_grouped Above Average Deaths



Europe and the Western Pacific are the worst hit in terms of deaths.

full\_grouped Above Average Deaths Simulated



An African country with the above statistics has a 36.4% chance of being above the worldwide average. However, the same statistics in Europe give the country a 99% chance. The situation is clearly worse up north.

*(covid\_19\_clean\_complete) Confirmed Cases*

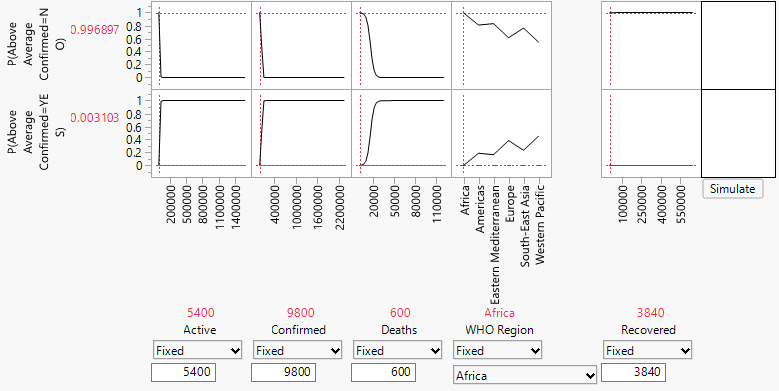
| **Column** | **Main Effect** | **Total Effect** |  |
| --- | --- | --- | --- |
| Active | 0.408 | 0.47 |  |
| Confirmed | 0.348 | 0.413 |  |
| WHO Region | 0.033 | 0.089 |  |
| Recovered | 0.01 | 0.031 |  |

*(covid\_19\_clean\_complete) Confirmed Deaths*

| **Column** | **Main Effect** | **Total Effect** |  |
| --- | --- | --- | --- |
| Confirmed | 0.734 | 0.791 |  |
| Active | 0.098 | 0.119 |  |
| Recovered | 0.08 | 0.108 |  |
| WHO Region | 0.013 | 0.041 |  |

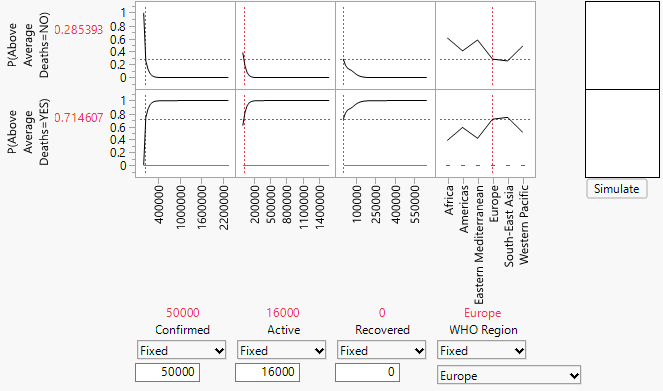
Active and Confirmed Cases were the highest predictors for our “covid\_19\_clean\_complete” dataset. Confirmed cases were huge for our above average death rating, but only about half as potent for our cases rating. WHO Region and Recovered patients round out the bottom, but WHO Region does have an effect for sure. COVID-19 cases went high and stayed high for some countries. It is clear that there were issues with containment, and we can see which areas struggle the most.

covid\_19\_clean\_complete Above Average Cases Simulated



Europe and Asia rule the roost for confirmed case rating, unlike America’s top spot in “full\_grouped”. This is possibly because this dataset has multiple statistics for Chinese and other provinces, giving the areas more weight overall than the Americas.

covid\_19\_clean\_complete Above Average Deaths Simulated



Europe and Southeast Asia are the hardest hit for deaths this time. This is possibly because this dataset has multiple statistics for Chinese provinces, giving the area more weight overall. The Western Pacific is still high up there.

**County Model Building**

I employed the same methodologies to the US county data as I did with the worldwide datasets. I created a validation column and created several predictive models looking at whether a county was above or below the national average for cases and/or deaths on a particular date. Once all my models were in play, I used JMP’s Model Comparison function to see which performed the best.

*Confirmed Cases*

| **Validation** | **Creator** |  | **Entropy RSquare** | **Generalized RSquare** | **Misclassification Rate** | **N** | **AUC** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 2 | Fit Generalized Adaptive Lasso |  | 0.1620 | 0.2219 | 0.1256 | 97205 | 0.9829 |
| 2 | Fit Generalized Adaptive Elastic Net |  | 0.2181 | 0.2920 | 0.1186 | 97205 | 0.9896 |
| 2 | Neural |  | 0.7237 | 0.7993 | 0.0407 | 97205 | 0.9915 |
| 2 | Bootstrap Forest |  | 0.2876 | 0.3747 | 0.1230 | 97205 | 0.9713 |
| 2 | Boosted Tree |  | 0.2739 | 0.3588 | 0.1377 | 97205 | 0.9641 |
| 2 | Neural |  | 0.6759 | 0.7598 | 0.0445 | 97205 | 0.9878 |

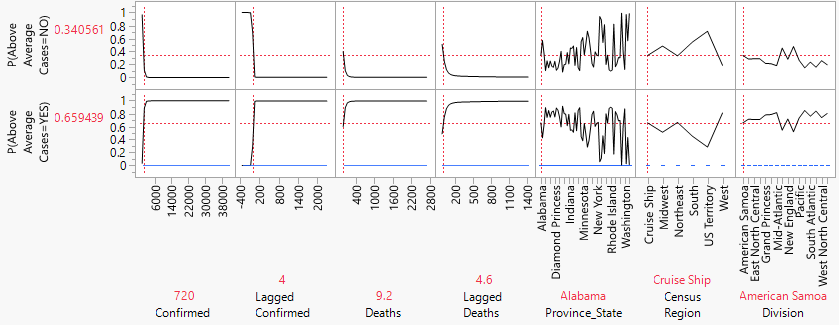
*Confirmed Deaths*

| **Validation** | **Creator** |  | **Entropy RSquare** | **Generalized RSquare** | **Misclassification Rate** | **N** | **AUC** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 2 | Neural |  | 0.6989 | 0.7733 | 0.0362 | 97205 | 0.9919 |
| 2 | Neural |  | 0.4711 | 0.5643 | 0.0831 | 97205 | 0.9621 |
| 2 | Bootstrap Forest |  | 0.2885 | 0.3689 | 0.0825 | 97205 | 0.9541 |
| 2 | Boosted Tree |  | 0.1812 | 0.2410 | 0.1179 | 97205 | 0.9667 |
| 2 | Fit Generalized Adaptive Lasso |  | 0.3193 | 0.4038 | 0.0924 | 97205 | 0.9733 |
| 2 | Fit Generalized Adaptive Elastic Net |  | 0.3164 | 0.4005 | 0.0926 | 97205 | 0.9729 |

In the data cleaning step, I pruned the county dataset quite a bit to remove outliers and unhelpful data. However, the dataset still remains very large, with close to 500,000 observations. The boosted neural network models took over five hours to complete with my computer, and the unboosted just over one hour. Neural networks take data points and run them through nodes that transform the data before producing a final output. With so many variables and so many observations, it makes sense that the process was so computer intensive. In the end, our neural networks came out ahead of the other models, much like with our global data.

For confirmed cases, our unboosted neural network actually outperformed our boosted, while the opposite was true for deaths. Looking again at our ROC curves, our highest AUC (highest predictive accuracy) was linked to one of our two neural networks in each case, as were low misclassification rates. These will be the models we use for analysis and predictions.

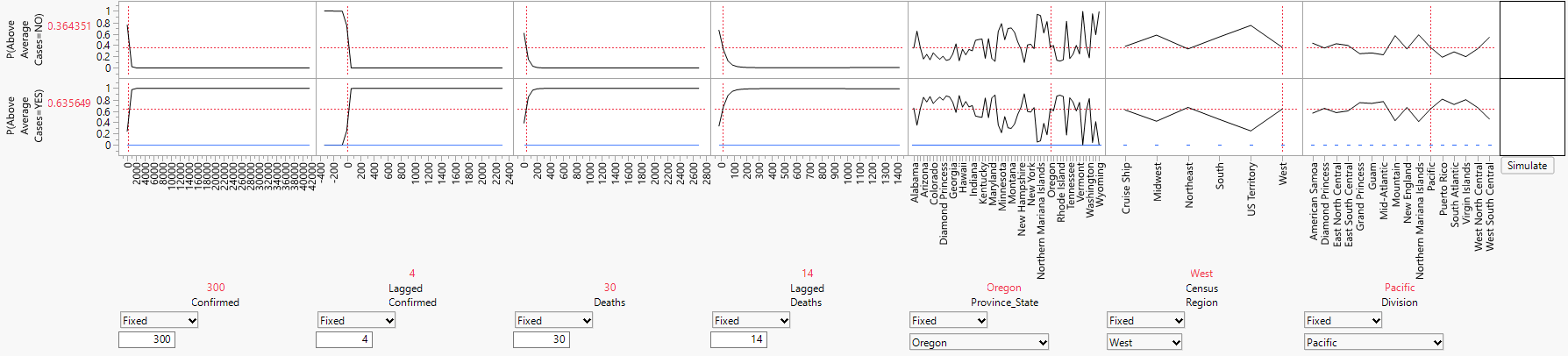
**County Data Analysis**



| **Column** | **Main Effect** | **Total Effect** |  |
| --- | --- | --- | --- |
| Confirmed | 0.064 | 0.986 |  |
| Lagged Confirmed | 0.035 | 0.802 |  |
| Deaths | 0.01 | 0.597 |  |
| Lagged Deaths | 0.03 | 0.322 |  |
| Province\_State | 0.014 | 0.194 |  |
| Census Region | 0.005 | 0.051 |  |
| Division | 0.006 | 0.047 |  |

With our “above average cases” metric, we see confirmed and lagged confirmed cases come in highest as predictors. As total cases reach a certain point, the chance of the county being above the average skyrockets. As many counties never saw a single corona case, the national average ended up hovering around a set point as growth only exploded in a few areas. Following the confirmed cases are deaths and lagged deaths. Deaths tend to rise when cases do, so this makes sense as well.

Province/State comes in after these stats, explaining about 19.4% of variation in the data. We can see from the profiler graphs that chances vary wildly from state to state. A county in Alabama with all other stats being equal has a lower chance than one from New York of being above the national average. After state comes census region (general area of the country), and finally division (more specific). The reason that location of the county is less important than other stats is that each area of the country has many counties that have zero cases/deaths. This is more that the USA has a *lot* of counties than the fact that the coronavirus is not effective. As seen in our density map from previous visualizations, many counties are far apart, and transmission is difficult in these cases.



When simulated we can see whether a county is likely to be above average for confirmed cases. For example, this county has 300 confirmed cases, 30 deaths, and is located in Oregon. The likelihood given is 63.6% that this county is struggling to contain the virus’s spread compared to the rest of the country.

For our “above average deaths” metric, our lagged statistics came in handy as predictors. Confirmed cases from two days prior were highly correlated with the chance of being above the national average. This lagged data can be an indication of runaway infections, leading to later numbers being higher as cases rise exponentially. More infections unfortunately means more deaths, as the death rate is fairly consistent. Lagged deaths follows lagged confirmed cases, which also makes sense as deaths rose at an exponential rate alongside cases as the virus spread.

Following these lagged variables are the Province/State, Census Region, and Division, just like with the confirmed cases. It makes sense that deaths would follow the trends of confirmed cases. Inability to contain the spread of the virus likely illustrates a general unpreparedness that would lead to more deaths from the virus.