COVID-19 Prediction Models

Covid-19 Modelling with Hyperparameter Optimization (Logistic Regression)

Rawinder Singh

Bachelor’s Software Engineering (Artificial Intelligence)

Media Design School

**Method employed and model building comments:**

COVID-19 is a pandemic that has hit the globe with significant social, economic, and environmental implications. Since the virus has the properties of fast transmission, a region needs to do its best to understand and predict the spread of the outbreak. With the projected model the aim to see if there is a correlation between the independent variables such as the demographic data and the dependent variables such as death. The model will be trained using Logistic Regression. Which is a statistical model that uses a logistic function to model a binary dependent variable. (Tolles & Meurer, 2016). The objective of this study is to use several COVID-19 patients' data to predict the severity of the cases and possible outcomes (e.g., recovery or death). The method used to produce this model is a follows: Load the data set with all the variables. Analyze the current data set and its structure and formatting. Pre-process the data, getting rid of all the unnecessary information, and clean/format data to suit the model. Then train the model using Logistic Regression. Perform hyperparameter optimization to generate better results from the model. Generate the evaluation metrics including accuracy, precision, recall, and F1 Score. Finally, the assess the results of the model have been generated.

**Results:**

The most important part of the Prediction model is essentially the results. The results are what drive the purpose behind such projects. Predictive models are designed to help predict future outcomes. (Choudhury, 2019) This helps an organization to make better decisions. The results of any prediction model are important, however in the case of Healthcare. There needs to be extra careful to ensure the results of the model. In healthcare, there are situations that deal with life and death. The models help to allocate resources, and thus there should be as little error as possible. (*Quality Data Critical to Healthcare Decision-Making*, 2004) A way that we can greater the results produced by the model is through Hyperparameter Optimization (HO). A process in which you adjust the parameters of the model. Towards the end of the associated code, there are results of the model that have been Hyperparameter Optimized. Evaluation metrics are a good way of evaluating how well a model can produce results. These metrics include Precision, Recall, and F1 Score. Precision is the calculation of how accurate results are produced. (Iwendi et al., 2020) The model trained produced a Precision value of 0.6 and 0.72 (HO). Another metric is Recall. Which identifies the correctly classified patients. (Iwendi et al., 2020) Our model produced Recall scores of 0.72 and 0.72 (HO). Finally, the last metric is the F1 Score. This provides the correct evaluation of the model performance. (Iwendi et al., 2020) In which the model scored 0.696 and 0.72 (HO). The accuracy of the models was 0.84 and 0.86 for HO.

**Issues of limitations of the prediction models, including a comparison of the model:**

With there being several contributing factors, not every model can be perfect. There are always limitations and issues within the prediction models. Predictive analytics is limited due to a couple of reasons. Firstly, hypothesis-driven responses, mean that the models are built to favor the hypothesis. Secondly, data-driven responses, which are driven by what the type of information is in the data. (Kearns, n.d.) With the Covid-19 predictive model, it may be possible that there may have been a slight lean towards hypothesis-driven responses. With the objective being to see the correlation between the variables, there may have been favoring decisions made unconsciously. Now diving into Linear Regression (LR). A major limitation of LR is the assumption of linearity between the two variables, independent and dependent. This again starts to favor the model towards a certain side, which is good in this case but it may not be in other cases. Lastly, if LR is compared against other classification algorithms then, it is possible to see that there are better performing algorithms such as Boosted Random Forest. (BRF) (Iwendi et al., 2020) Thus it can be concluded that BRF is a more optimal choice for this predictive model.

# References

Choudhury, A. (2019, January 28). *What Is Predictive Model Performance Evaluation And Why Is It Important*. Analytics India Magazine. https://analyticsindiamag.com/what-is-predictive-model-performance-evaluation-and-why-is-it-important/

F. (2020, March 7). *P2 : Logistic Regression - hyperparameter tuning*. Kaggle. https://www.kaggle.com/funxexcel/p2-logistic-regression-hyperparameter-tuning

Grace-Martin, K. (2018, January 5). *Model Building Strategies: Step Up and Top Down*. The Analysis Factor. https://www.theanalysisfactor.com/model-building-strategies/

Iwendi, C., Bashir, A. K., Peshkar, A., Sujatha, R., Chatterjee, J. M., Pasupuleti, S., Mishra, R., Pillai, S., & Jo, O. (2020). COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm. *Frontiers in Public Health*, *8*. https://doi.org/10.3389/fpubh.2020.00357

Kearns, S. (n.d.). *The Limitations of Predictive Analytics Tools and Why Execs Should Care*. RiverLogic. https://www.riverlogic.com/blog/the-limitations-of-predictive-analytics-tools-and-why-execs-should-care

Peshkar, A. (2020). *Atharva-Peshkar/Covid-19-Patient-Health-Analytics*. GitHub. https://github.com/Atharva-Peshkar/Covid-19-Patient-Health-Analytics/blob/master/COVID\_19\_Model\_Comparison.ipynb

*Quality Data Critical to Healthcare Decision-Making*. (2004, October 15). AHIMA. http://library.ahima.org/doc?oid=106428#.YJpbSLUzbZt

Tolles, J., & Meurer, W. J. (2016). Logistic Regression. *JAMA*, *316*(5), 533. https://doi.org/10.1001/jama.2016.7653