**COVID-19 Recovery Time Model Prediction Report**

1. **Introduction**

The “recovery.RData” dataset is designed to study the factors that predict recovery time from COVID-19 illness. Dataset contains 13 prediction variables (non-significant predictor “id” being removed), including Gender, Race, Smoking, Height, Weight, BMI, Hypertension, Diabetes, Systolic blood pressure, LDL cholesterol, Vaccination status at the time of infection, Severity of COVID-19 infection, Study. The outcome variable will be Time to recovery.

10000 participants’ recovery info data were collected from three existing cohort studies and 2000 participants were randomly selected from dataset for data analysis. Data partition will be introduced to split data into two sets: 70% training set and 30% testing set. Training data is used to train the predictive model, while the testing set is used to evaluate the model's performance. In order to maintain the reproducibility of the analysis, `set.seed()` function will be used to ensure that the random number generator produces the same sequence of random numbers each time.

1. **Exploratory analysis and data visualization**
2. **Model training**

The goal of the study is to develop a prediction model for time to recovery from COVID-19, using a variety of regression models. The selection of the model takes into account three different flexibility levels: *low* flexibility model [captures simple linear relationship] -- Least Squares (LS); *medium* flexibility models [capture non-linear relationships with limitations] -- Principal Components Regression (PCR), Partial Least Squares (PLS), Generalized Additive Models (GAM); and *high* flexibility models [capture complex non-linear relationships] -- Ridge Regression, LASSO, Multivariate Adaptive Regression Splines (MARS). However, high flexibility models may be more prone to overfitting and hard to interpret.

After splitting dataset into training and testing sets, each model will be trained using 10-fold cross-validation, which involves splitting the training data into 10 equal folds. The models are trained on the first 9 folds and validated on the last fold. This procedure will repeat 5 times (`repeats=5`) with each fold used for validation once and the average performance of the model across the 10 validation sets is calculated to obtain a relatively robust performance.

Tuning parameters are used in model training to optimize the performance of the model. Tuning parameters are selected using hyperparameter tuning to determine optimal lambda values. `tuneGrid` function generates a grid of specified values of lambda that are logarithmically spaced between optimal range. This study will first select a larger lambda distribution interval, and gradually reduce the value range of lambda according to the RMSE distribution of the fitted model.

The performance of each model is evaluated using RMSE, which measures the difference between the predicted values and observed values of the response variable. The model with lowest RMSE will be selected as the best model for predicting Covid recovery time.

* 1. The Least Squares model assumes that the relationship between the predictors and the response variable is linear, and the errors are normally distributed with constant variance. It also assumes that there is no multicollinearity among the predictors and errors are independent. The method argument is set to “lm” function from `caret` package to specify that a Linear Model should be used.
  2. The Ridge model adds a penalty term to the sum of squared residuals to reduce the variance of the estimates, where the LASSO model adds a penalty term to the sum of absolute values of the regression coefficients. In Ridge model, the alpha parameter, which controls the balance between the penalties, is set to 0, which corresponds to the Ridge penalty. Ridge model training uses `glmnet` method with cross validation, and `tuneGrid` argument specifies the grid of tuning parameters for cv. In this case, the grid consists 100 values of lambda range from 6 to -8 on log scale. LASSO model uses similar procedure, in LASSO model, the alpha parameter is set to 1, because LASSO uses L1 penalty which has the effect of forcing some of the coefficient estimates to zero when lambda is sufficiently large. The LASSO grid consists 100 values of lambda range from 1 to -6 on log scale. These two models both assume that the correlation between predictors and outcome is linear, but the errors are not necessarily normally distributed with constant variance. It is worth noting that although the lambda value uses for LASSO training range from 1 to -6, the lambda range that present the best local minimum RMSE is from -2.5 to -6. However, both lambda values to be negative means that the model is not penalizing any variables at all. This also means that the LASSO model will no longer have effect on variable selection and include all predictors.
  3. The GAM model assumes that the relationship between predictors and outcome is non-linear and can be modeled using smooth functions. It assumes normally distributed residuals with constant variance, and independent predictors with no multicollinearity. Method option in `tuneGrid` argument is for `mgcv` implementation to tune the method for selecting optimal number of basis function and estimate tuning parameters and coefficient of splines, and select option helps selecting variables. `bestTune` will indicate that adding an additional penalty term and using GCV approach is good.
  4. The MARS model also capture non-linear relationships using piecewise linear functions. It assumes linearity between predictors and outcome variable in each region, the effects of the predictors are additive, and predictors are independent of each other. In MARS model, degree is num possible products hinge functions in one term, nprune is upper bound of terms in model.
  5. The PCR model and PLS model assume that the predictors are highly correlated and few core latent variables explain most of the variability in data. These two models both assume the relationship between predictors and outcome variable to be linear, and the distribution of residuals to be normal with constant variance across all levels of predictors. Their differences is that PCR assumes no linear combination of the predictor is the same as another predictor, while PLS can handle situations of a predictor shares same linear combination as another predictor. For both trainings, it’s necessary to identify number of principal components `ncomp` first which can help balance the tradeoff between model complexity and performance. `selectionFunction=“best” gives these two models a new trainControl method which gives the best performance on the validation set.

1. **Results**

RMSE measures the average difference between predicted values of the model and actual observed values of target population, therefore the model with lowest RMSE indicates the model that gives the most accurate predictions. According to RMSE summary table for the ten selected models (Figure ?), MARS model has shown to provide *lowest* RMSE. With 10-fold cross validation, the nprune with lowest RMSE will be 13 and product degree will be 1. Assess testing performance, the model has RMSE of 22.49153 using test data, suggesting that on average the model’s prediction on test data are off by 22.49 units, indicating a good fit to the data. The model shows that bmi is the most important predictor, followed by vaccine1 and severity1. Based on this information, the *final model* for predicting time to recovery from COVID-19 will be using MARS model.

The final model will be:

recovery\_time = -41.526084-4.942170\*gender1+4.616207\*smoking1+7.657797\*smoking2+4.535753\*hypertension1-7.436460\*vaccine1+9.131057\*severity1+4.155071\*study2+10.051484\*h(bmi-23.9)+7.357588\*h(bmi-29.5)+ 9.948860\*h(31.7-bmi)+ 35.195751\*h(bmi-34.5)

The coefficients of the model suggest that, *holding all other predictors constant\**:

1. Being male is associated with shorter recovery time than being female.
2. Being former and current smoker is associated with longer recovery time than non-smoker.
3. Having hypertension is associated with longer recovery time than not having hypertension.
4. Being vaccinated is associated with shorter recovery time than not receiving vaccine.
5. Having severe disease is associated with longer recovery time than not severe disease.
6. **Conclusions**

MARS model will be considered as best fitted model. The model’s analysis provides valuable insights into predicting time to recovery from COVID-19. For instance, smoking, hypertension, and severity of the condition were identified as significant predictors of recovery time, which highlights the importance of managing these factors during the recovery process. Furthermore, the model’s inclusion of BMI, weight, and height as significant predictors suggests that managing participants’ weight and maintaining healthy BMI may help reduce recovery time, which emphasizes the importance of keeping a healthy lifestyle, especially during pandemic.