**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. ***Primary Analysis*** (Regression analysis for Recovery Time)
2. **Exploratory analysis and data visualization**
3. Correlation plot (Figure 1) illustrates the visual representation of the correlation between predictors. The Correlation plot of the training dataset shows that a negative correlation exists between BMI and height, with positive correlations exist between SBP and age, BMI and weight.
4. The function “featurePlot()” generate visualization of the relationship between prediction and the outcome variable in the training dataset. According to the distribution presented by lattice plots (Figure 2), the following correlation can be summarized between continuous predictors and Covid time to recovery:

**(a)**There is little or no correlation between SBP, LDL, age with recovery time. **(b)**Participants with Body Mass Index range from 20 to 30 do not indicate apparent correlation between their BMI and recovery time. However, participants with a BMI over 30 had a significant increase in recovery time as their BMI increased. **(c)**There is a slightly inverse correlation between height and recovery time. **(d)**Participants with weight range from 60 to 80 do not indicate apparent correlation between their weight and recovery time. However, participants with a BMI over 80 had a slightly increase in recovery time as their weight increased.

1. The boxplot (Figure 3) provides insights into the connection between recovery\_time and categorical predictors. Boxplots suggest that a female who doesn’t smoke, has no hypertension or diabetes, is vaccinated, and is not severely infected with Covid, has a lower recovery\_time. While patients in study group 2 have indicates slightly shorter recovery time compared to other study groups.
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 combining with Tree-based methods: *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); *high* flexibility models [capture complex non-linear relationships] -- Ridge Regression, LASSO, Multivariate Adaptive Regression Splines (MARS), Regression Trees, Random Forest, Boosting. 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 a linear relationship between recovery\_time and the predictors, 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. LASSO sets alpha to 0 and the grid consists 100 values of lambda range from 6 to -8 on log scale. Ridge sets alpha to 1 and the grid consists 100 values of lambda range from 1 to -6 on log scale.
  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. Model will use the “GCV.cp” method for model selection, and the intercept term should be included while the second term should not be included in the model. The tuning parameter for the GAM model is tuneGrid, `GCV.Cp` is specified as the method, which stands for generalized cross-validation with a complexity penalty. `Select` is a logical vector that indicates which of the input features should be included in the GAM model.
  4. The MARS model captures 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. The grid of “degree” ranges from 1 to 3 and the grid of “nprune” ranges from 2 to 25.
  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. In both cases, the tuning grid of “number of principal components (ncomp)” contains value from 1 to 18, and the pre-processing steps are centering and scaling.
  6. Regression tree assumes a linear and addictive relationship between predictors and the recovery\_time, and the errors are normally distributed with constant variance. The tunning grid of consists 50 values of “cp” range from -6 to -2 on log scale.
  7. Random forest assumes a non-linear relationship between recovery\_time and the predictors, and the errors are normally distributed with constant variance, it is also robust to outliers. The tuning grid of “the number of variables randomly sampled at each split(mtry)” contains values from 1 to 16, and “the minimum number of observations required in a terminal node of the tree(min.node.size)” were tested using values from 1 to 6.
  8. Boosting assumes that the observations in the data set are independent of each other. The tuning grid includes four parameters: The number of trees varies from 100 to 3000, the depth of the trees ranges from 1 to 3, the shrinkage ranges from 0.0005 to 0.002, and the minimum number of observations in a terminal node is set to 1.

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 4), 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.

1. ***Secondary Analysis* (**Consider time to recovery as binary outcome (>30 days vs. <= 30 days) and develop prediction model for binary outcome**)**
2. **Exploratory analysis and data visualization**
3. Correlation plot (Figure 1) displays pairwise correlations between continuous variables which should produce the same results as primary analysis since predictors have not been modified.
4. The featured plot (Figure 5) captures the relationship between recovery\_time and the continuous variables. It indicates that there is a linear relationship between SBP, LDL, and age, while a non-linear trend appears in BMI, height, and weight.
5. **Model training**

A variety of machine learning models are introduced to develop a predictive model for a binary outcome of time to recovery (>30 days vs. <= 30 days). The models used in this study include Generalized Linear Models (GLM), Penalized Logistic Regression, Generalized Additive Models (GAM), Multivariate Adaptive Regression Splines (MARS), Quadratic Discriminant Analysis (QDA) & Linear Discriminant Analysis (LDA), Naïve Bayes (NB), Support Vector Machines (SVM), Classification Trees, Boosting, and Random Forests.

In this case, the `control` object for binary outcome model training is different as the previous analysis. `ctrl2` is using k-fold cross-validation, where the data is split into k equally-sized folds, and the process is repeated k times. In this case, `method=cv` specifies the type of cross-validation, and `classProbs=TRUE` indicates that class probabilities will be calculated based on classification. Additionally, `twoClassSummary` specifies a summary function to be used for classification problems, which calculates performance metrics such as sensitivity, specificity, and area under the ROC curve.

1. GLM assumes a linear relationship between recovery\_time and predictors, observations are independent of each other, the variance of the response variable is constant, and the residuals are normally distributed.
2. Penalized logistic regression assumes that the relationship between log odds of the response and the predictors is linear and there is no multicollinearity among the predictors. The tuning grid consist of 21 values of alpha from 0 to 1 and 50 values of lambda from -8 to -1.
3. The MARS model assumes linearity between predictors and outcome variable in each region, the effects of the predictors are additive, and predictors are independent of each other. The tuning grid of degree consists of values from 1 to 4 and nprune from 2 to 20.
4. LDA and QDA both assume that a linear relationship between recovery\_time and the predictors and the distribution of each predictor variable is normal. It also assumes that the variance of the predictor variables is equal across all groups or classes and the predictor variables are independent of each other. The default tuning parameters are typically the prior probabilities of each class.
5. NB assumes that the predictor variables are conditionally independent given the outcome variable besides the independence of each predictor. It also assumes that the predictor variables are normally distributed, and the variance of the predictor variables is equal across all classes. The tuning grid consists values of adjust from 0.2 to 3 with an increment of 0.2.
6. SVM assumes that the data is linearly separable. The pre-processing steps are centering and scaling and the tuning grid contains 50 values of “cost” from -3 to 2.
7. Classification tree assumes that the predictors are independent and identically distributed and there is no multicollinearity among predictors. We used both “rpart” and “ctree” methods to train the model. The “rpart” method contains a tuning grid of 50 “cp” from exp(-6) to exp(-3).
8. Classification Random forests assume that the predictors are independent of each other and have equal importance without outliers or influential observations in the data. The tuning parameter grid for random forest includes 9 values for `mtry` ranging from 0 to 8, “gini” as the split rule and 5 values for `min.node.size` ranging from 2 to 10 with a step size of 2.
9. Classification Boosting assumes that the predictors are independent of each other and have equal importance without outliers in the data. Tuning parameter grid includes 6 values for n.trees (100, 250, 500, 1000, 2000, 3000), 6 values for `interaction.depth` (1 to 6), 3 values for shrinkage (0.002, 0.005, 0.008), and 1 value for `n.minobsinnode` (1).
10. **Results**

Based on the resulting table and boxplot (Figure 6) of the “resamples” function, we concluded that GAM model displayed highest accuracy. The training error rate is 26.55% and the corresponding CI is (24.83%, 28.33%). The testing error rate is 29.4% and the corresponding CI is (24.7%, 30.12%). The R-squared value of 0.146 indicates that the model explains about 14.6% of the variation in the outcome.

The final model will be:

log(p(recovery\_time = “long”) = 2.28480 - 0.32518 \* gender1 - 0.11553 \* race3 - 0.05404 \* race4 + 0.42913 \* smoking1 + 0.54608 \* smoking2 + 0.32009 \* hypertension1 + 0.04432 \* diabetes1 - 0.72124 \* vaccine1 + 0.79673 \* severity1 - 1.56884 \* study2 - 0.37745 \* study3 + s(age) + s(SBP) + s(LDL) + s(bmi) + s(height) + s(weight)

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, compared to 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. Being Black or Hispanic is associated with shorter recovery time compared with being White or Asian.

The smooth functions suggest that age, SBP, LDL, bmi, height and weight all have nonlinear relationship with the recovery\_time. The smooth term “s(bmi)” is significant (p-value < 0.05). GAM Model Feature Plot (Figure 7) indicates feature selections has higher accuracy compared to non-feature selections. Plot of Relationship Plot of Nonlinear Terms (Figure 8) shows the relationship between the non-linear terms and the binary outcome variable recovery\_time. We can conclude that the smooth term of BMI has a significant non-linear relationship with the outcome.

1. **Conclusions**

In conclusion, our secondary analysis suggests that the GAM model is most suitable for predicting the time to recovery from COVID-19. The insights gained from this analysis are significant, with body mass index (BMI) and smoking status being identified as two of the most crucial predictors. The findings emphasize the importance of maintaining a healthy lifestyle during pandemics, as it may potentially reduce recovery time. Therefore, it is highly recommended that individuals prioritize healthy habits such as maintaining a healthy BMI and refraining from smoking.

1. ***Appendix***

**Figure 1. Primary and Secondary Analysis Correlation Plot**

**图表, 气泡图

描述已自动生成**

**Figure 2. Primary Analysis Lattice Plots for Continuous Predictors**

**图片包含 图形用户界面

描述已自动生成**

**Figure 3. Primary Analysis Boxplot for Categorical Predictors**

**图示, 箱线图

描述已自动生成**

**Figure 4. Primary Analysis RMSE Summary Table**

**图表, 箱线图

描述已自动生成**

**Figure 5. Secondary Analysis Feature Plot**

**图表, 图示, 箱线图

描述已自动生成**

**Figure 6. Secondary Analysis Accuracy Summary Table**

**图表

描述已自动生成**

**Figure 7. GAM Model Feature Plot**

**图表, 折线图

描述已自动生成**

**Figure 8. GAM Relationship Plot of Nonlinear Terms**

**图形用户界面, 图表, 图示

描述已自动生成**