

Predictive Analysis of Recovery Time from Covid-19

Chenyao Ni (cn2604) Wenjing Yang(wy2369) Yeque Sun(ys3508)

Introduction

The Covid-19 pandemic has affected millions of people worldwide, with many experiencing prolonged recovery times after contracting the virus. This project aims to develop a prediction model for Covid-19 recovery time and identify associated risk factors using data from three cohort studies. Additionally, a secondary analysis will present a binary outcome model for recovery time (>30 days vs. ≤ 30 days), providing a concise understanding of prolonged recovery and identifying at-risk individuals. The report will feature updated findings using merged data and more analysis techniques to enhance the accuracy and relevance of the results.

Variables Summary

- ID: Participant ID
- Gender: 1 = Male, 0 = Female
- Race/ethnicity: 1 = White, 2 = Asian, 3 = Black, 4 = Hispanic
- Smoking: Smoking status; 0 = Never smoked, 1 = Former smoker, 2 = Current smoker
- Height: Height (in centimeters)
- Weight: Weight (in kilograms)
- BMI: Body Mass Index; BMI = weight (in kilograms) / height (in meters) squared
- Hypertension: 0 = No, 1 = Yes
- Diabetes: 0 = No, 1 = Yes
- Systolic blood pressure (SBP): Systolic blood pressure (in mm/Hg)
- LDL cholesterol: LDL (low-density lipoprotein) cholesterol (in mg/dL)
- Vaccine: Vaccination status at the time of infection - 0 = Not vaccinated, 1 = Vaccinated
- Severity: Severity of COVID-19 infection-0 = Not severe, 1= Severe
- Study: The study (A/B/C) that the participant belongs to
- Time to recovery: Time from COVID-19 infection to recovery in days
- Length indicator: A binary variable indicating whether the participant's recovery time (in days) was longer than 30 days - 1 = Recovery time longer than 30 days, 0 = Recovery time no longer than 30 days

Exploratory analysis and data visualization

The scatterplots show that **SBP**, **LDL**, and **Age** are the continuous variables linearly correlated (small correlation coefficient though) to **Recovery time** while no apparent discrepancy is found when finding the association between **Recovery time** and the categorical variables of different levels. The correlation plot suggests that **SBP** is positively correlated to **Hypertension** while the correlation among **BMI**, **Weight**, and **Height** are foreseen. In short, there is no obvious collinearity among most of the covariates (none of them exceed the threshold of strong correlation 0.8).

The LDA partition plots show that the response variable is **Length_ind** and the predictor variables are **age**, **BMI**, **SBP**, and **LDL**, which are continuous variables. From the plots, we would like to visualize the classification performance of a linear discriminant analysis (LDA) model based on every combination of two variables. However, separation of classes is not effective in the plots.

Model Training (Primary analysis)

(Note: 5 times repeated 10 folds cross-validation is used across the training process)

- Ordinary Least Squares Regression (OLS)

While ordinary least squares regression assumes linear relationship, normality of residuals, homoscedasticity, and independence of residual error terms, it can be seen from the diagnostic plots that the assumptions of normally distributed residuals and homoscedasticity are violated. Therefore, regularization methods and nonlinear methods should be introduced.

- Elastic Net Regression & Partial Least Squares

Elastic net regression has no other assumption as it is a parameter regularization method to handle the multicollinearity of OLS. Through the process of cross-validation with alpha from 0 to 1 and a positive range for penalty coefficient, the resulting elastic net model gives an alpha at 1, which is equivalent to the Lasso model with a penalty coefficient of 0.135. As a dimension reduction method, partial least square regression assumes the independence between components and homogeneity and function nature of the system while making no assumption to the observed data. The partial least square model gives 14 components after cross-validation for best predictability within a tuning range from 1 to 18 since there are maximum 18 predictors in OLS.

For the non-parametric models below, there is no assumption of the underlying data.

- Generalized Additive Model (GAM)

For a generalized additive model, the tuning parameter is a decision for adding a penalty to each of the terms. The selected model after cross-validation has no penalty to each term, 6 out of 18 predictors get smooth terms while the rest of predictors are linearly related to the response variable.

- Multivariate Adaptive Regression Spline Model (MARS)

In the selected multivariate adaptive regression spline model after cross-validation, there are 3 out 18 predictors with 5 terms and a degree of 2 while the tuning parameters for terms range from 2 to 18 and degree from 1 to 3. Within the 3 predictors in the model, BMI is of most significance as it is included in 3 out of 5 terms.

- K-Nearest Neighbors Regression (KNN)

The selected KNN model has the tuning parameter k = 13 after cross-validation, with a tuning range of k from 5 to 15.

- Regression Tree (CART)

The selected model after cross-validation has a complexity parameter (cp) of 0.00778, which is chosen from a candidate grid of cp values ranging from $\exp(-6)$ to $\exp(-2)$ with 50 equally spaced values on a logarithmic scale.

- Random Forest

The mtry is the number of variables randomly sampled at each split when building a decision tree in the forest, and its optimal value depends on the size and complexity of the dataset, as well as the number of trees in the forest. A smaller mtry value can result in lower correlation between the trees in the forest in order to prevent overfitting. Here, the tuning range for mtry is set to 4 to 6, which is appropriate for a dataset with 18 predictors. The min.node.size parameter was also tuned between 3 and 8 to control the minimum size of terminal nodes in each tree, based on the number of observations in the training set. After cross-validation, the optimal number of randomly selected predictors is 6 and the minimal node size is 7.

- Generalized Boosted Regression (GBM)

The final generalized boosted regression model was created using 300 trees and 3 splits per tree, with a range of tree numbers from 100 to 1200 in increments of 100. The split depth of each tree was set to 1, 2, and 3. The learning rate was specified as 0.01 and 0.03, and each terminal node had at least 5 observations based on the relative size of the training set. (Based on Rule of Thumb: learning rate = $\max(0.01, 0.1 * (\min(1, nl/10000)))$, default terminal nodes as 10 observations for a training set size of 10000).

Model Training (Secondary analysis)

(Note: 10 folds cross-validation is used across the training process, ROC-AUC is selected as metric because of the imbalanced data)

- Penalized Logistic Regression

The best tune of the selected penalized logistic regression model within the tuning range from $\exp(-10)$ to $\exp(-2)$ is 0.000069. The lambda is close to zero, which indicates penalization term is not needed. From the results, the mixing proportion (alpha) is 1 so the model might have a lasso penalty term.

- Multivariate Adaptive Regression Spline Classifier (MARS)

The MARS model here has 9 terms within the range from 2 to 20 with 1 degree of freedom within the range from 1 to 3.

- Quadratic Discriminant Analysis (QDA)

Unlike Linear Discriminant Analysis (LDA), which assumes that the covariance matrix is the same for all classes, QDA allows for different covariance matrices for each class.

- Naive Bayes (NB)

The Native Bayes classifier can be used for nonparametric estimators. The Laplace Correction is implemented by the parameter fL, which is equal to 1. The bandwidth of the kernel density estimates is adjusted from 0.2 to 3. We also assume features in each class are independent in the NB model.

- Classification Tree

The reported model with highest cross-validation AUC gives a complexity parameter at 0.007, within the tuning range from $\exp(-6)$ to $\exp(-3)$.

- Adaptive Boosting

The resulting Adaptive Boosting model has 400 trees with 1 split performed on a tree with a tuning range of tree number from 100 to 1200 and split depth from 1 to 3, as the learning rate is specified as 0.01 and 0.03 and each terminal node has at least 5 observations based on the relative size of training set (Rule of Thumb used: learning rate = $\max(0.01, 0.1 * (\min(1, n_l / 10000))$), default terminal nodes = 10 observations for size 10000).

- Support Vector Classifier (Linear kernel)

The support vector classifier with linear boundary has best cross-validation performance with a cost parameter at 0.49 within the tuning range from $\exp(-5)$ to $\exp(2)$.

- Support Vector Classifier (Radial kernel)

The support vector classifier with radial boundary has best cross-validation performance with a cost parameter at 0.19 within the tuning range from $\exp(-5)$ to $\exp(2)$, and gamma parameter (called sigma in kernlab package) at 0.0011 within the tuning range from $\exp(-8)$ to $\exp(-3)$, which controls the distance of the influence by a single training point.

Results

- Primary Analysis

The generalized boosted regression model is selected as the final model for its best training performance with the lowest cross-validation root mean square error (22.50) among all models. This boosted regression model has a test mean squared error of 492.57. The final model indicates that **BMI** is the most significant predictor, explaining over 61% of the reduction in the loss function with this set of predictors. **Study group**, **LDL**, and **weight** also play an essential role as predictors due to their mathematical relationship to **BMI**. A partial dependence plot shows that the recovery time is relatively short for individuals with a BMI between 23 and 30, whereas the obese patients with **BMI** greater than 30 tend to experience a remarkably long recovery period from Covid-19. **Vaccine**, another factor that draws the public's attention, does help to slightly shorten the recovery time as shown by the partial dependence plot while the recovery time increases mildly as **Age** increases for the elderly over 50. On the other hand, the variables such as systolic blood pressure (**SBP**), **Race**, and **Diabetes** have little influence on the recovery time from Covid-19.

- Secondary Analysis

The adaptive boosting model is selected as the final model for its best training performance with the highest mean cross-validation area under the ROC curve (0.71) among all models. This Adaboost model has a test accuracy of 0.72 and a kappa of 0.18. **Study group**, **BMI**, and **Vaccine** are believed to be the most important predictors, which account for more than 65% of the reduction to the loss function given this set of predictors. Partial dependence plot and

Individual Conditional Expectation plot show that the individuals with **BMI** between 23-30 have lower probability of getting long recovery time. Lower **SBP** level is associated with lower long recovery probability while the **LDL** level does not influence the long recovery probability significantly. **Vaccine** is also related to lower probability of having a long recovery time.

Conclusion

- Primary Analysis

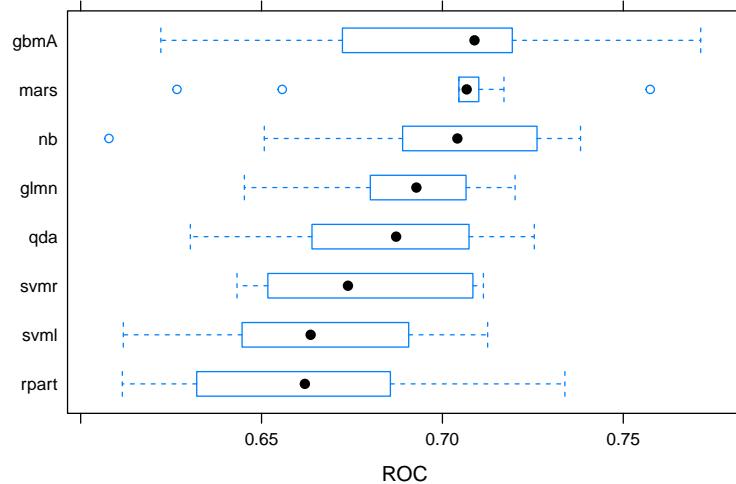
Based on the generalized boosted regression model with the best performance in predictions, **BMI** is the single most important factor that influences long recovery time (as well as study group, LDL, and weight due to their mathematical relationship to BMI). Obese patients with a **BMI** over 32 will experience a remarkably longer recovery period compared to patients with a lower BMI. Other common risk factors, such as symptom severity and age, do have a positive correlation with the length of recovery time, while vaccination does shorten the recovery period. However, there is no evidence to suggest that these factors have an effect on the scenario of long recovery time as they only associate with a mild difference in recovery time.

- Secondary Analysis

Although the adaptive boosting model lists **Study group B** as the most important predictor, it can be largely attributed to the imbalanced distribution of data from the three cohort studies (See Appendix). For a new observation, this predictor would be no longer valid. **BMI** is still the most significant predictive variable as the figures show that the individuals with relatively normal **BMI** (23-30) have remarkably lower risk of experiencing long recovery time from covid, compared to either underweight or obese individuals (reduction of 17% and 27% respectively). In contrast to the results from primary analysis, **SBP** becomes a more influential variable in predicting whether an individual would experience long recovery time from covid though it might be given little weight in predicting the exact recovery time.

Figures and Tables for Secondary Analysis

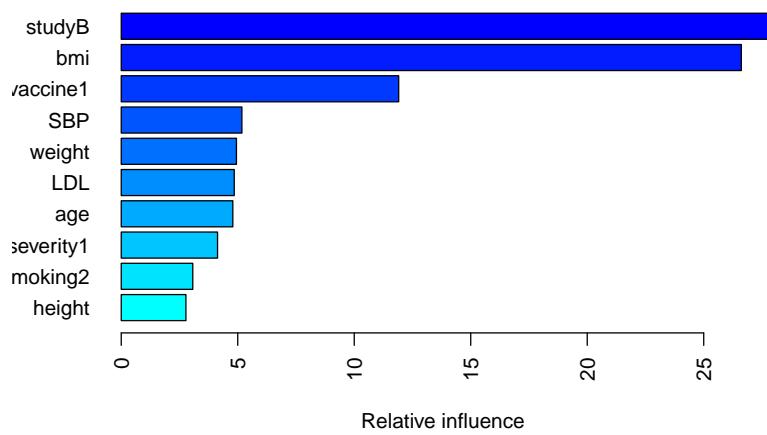
Models Comparsion based on Area Under Curves



Final model performance on test dataset

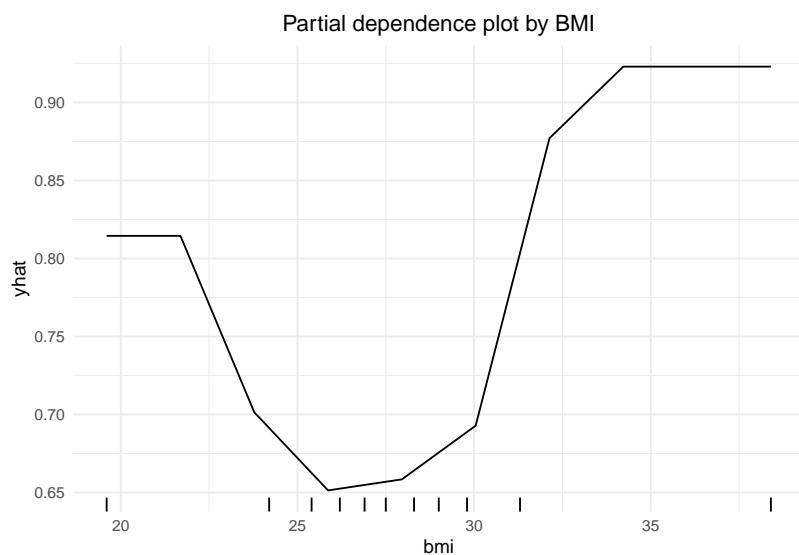
x	
Accuracy	0.7218359
Kappa	0.1815036

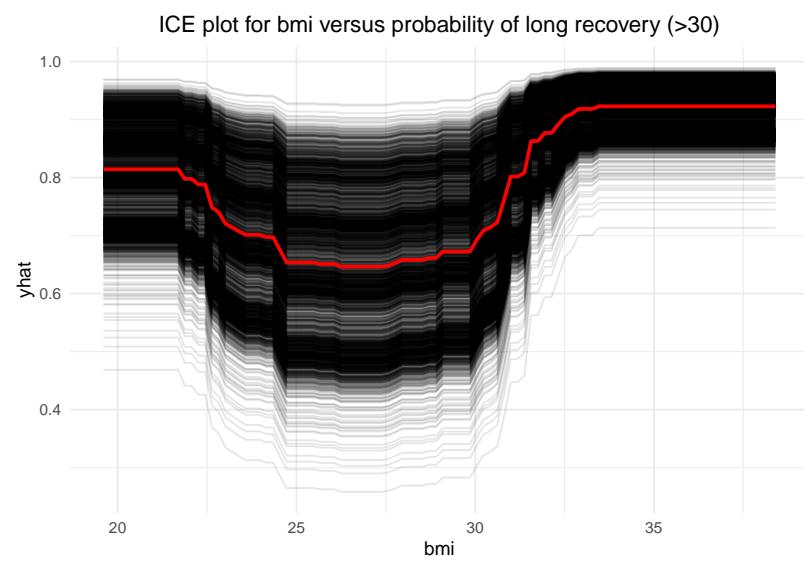
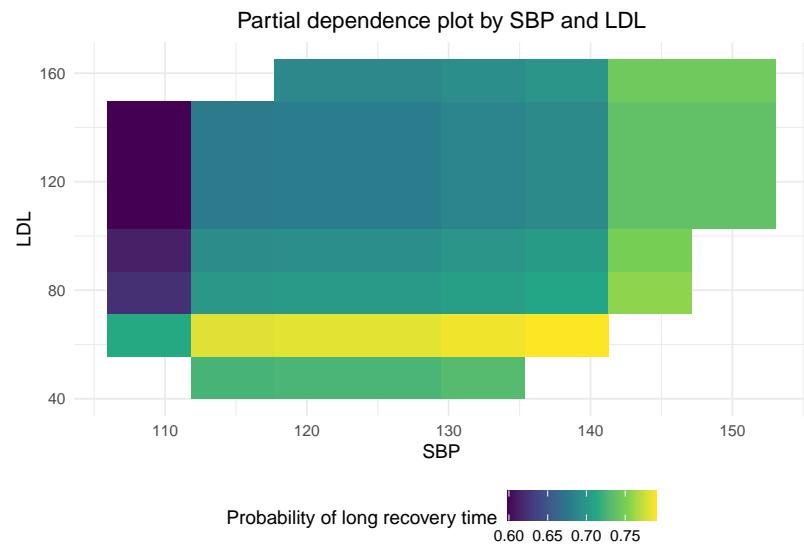
Variable importance plots of the final model



	rel.inf
studyB	27.8958031
bmi	26.6131604
vaccine1	11.9080853
SBP	5.1783551
weight	4.9451667
LDL	4.8495707
age	4.7895698
severity1	4.1331856
smoking2	3.0702806
height	2.7755020
gender1	1.1978980
smoking1	1.0717996
race4	1.0470291
race2	0.2303987
race3	0.1567322
studyC	0.1374633
hypertension1	0.0000000
diabetes1	0.0000000

Final Model interpretation

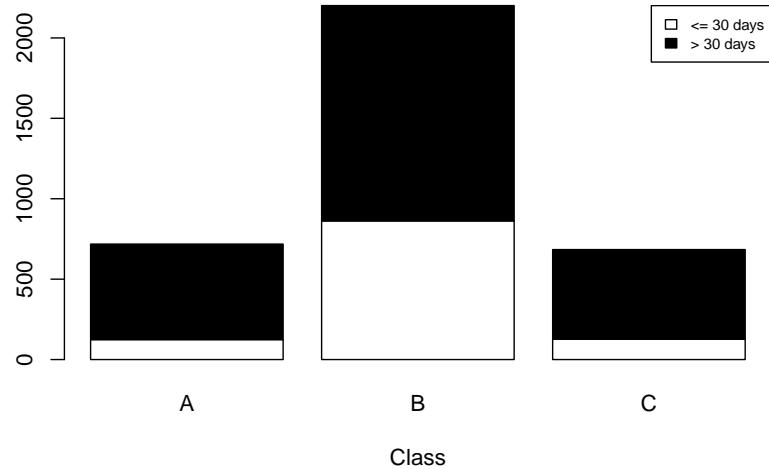




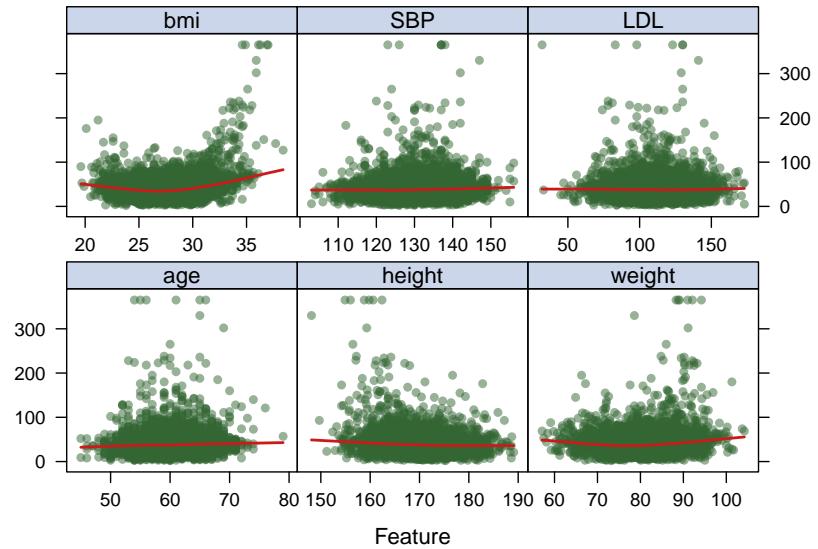
Appendix – Primary Analysis of Covid Recovery Time

Exploratory analysis and data visualization

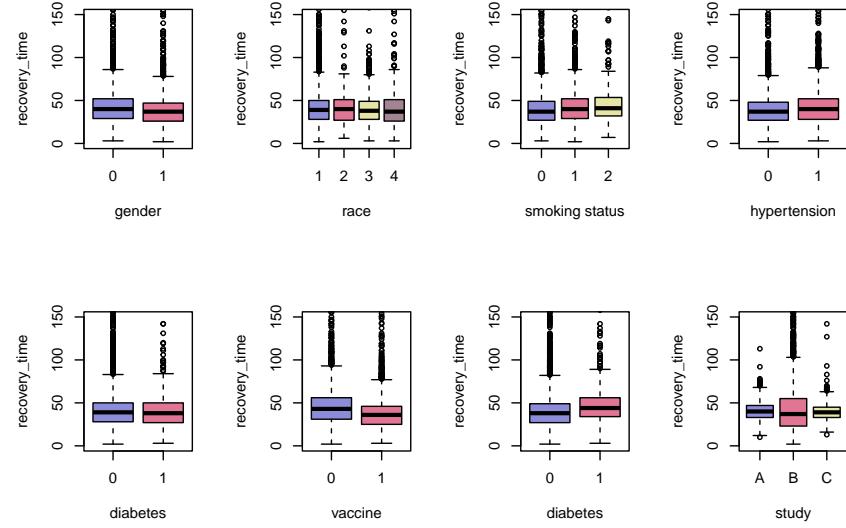
Number of cases seperated by 30 days in recovery by Study Group:



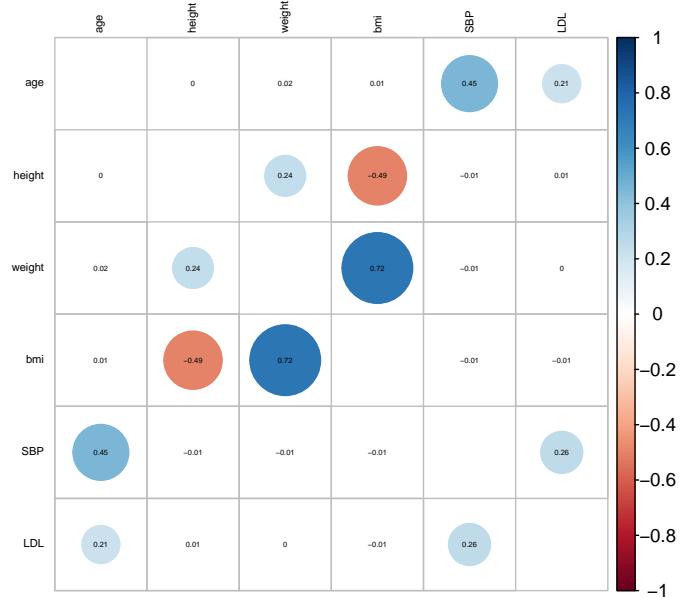
Visualize potential relationship between reponse variable and numeric predictors



Visualize potential relationship between response variable and categorical predictors

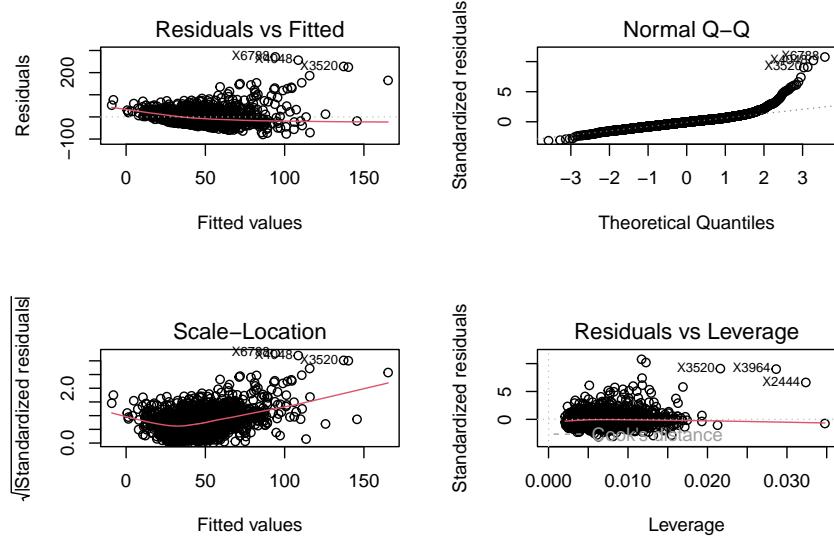


Correlation plot to check collinearity between covariates (based on training data)

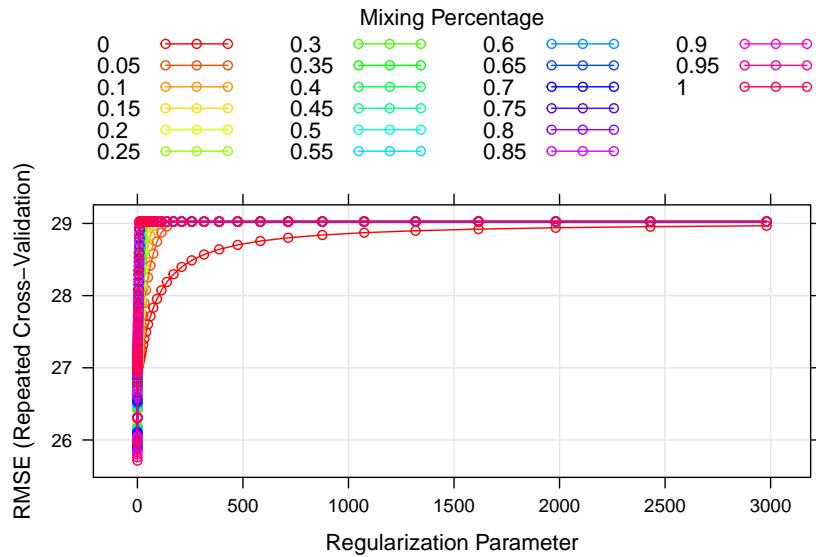


Model Training

Ordinary Least square

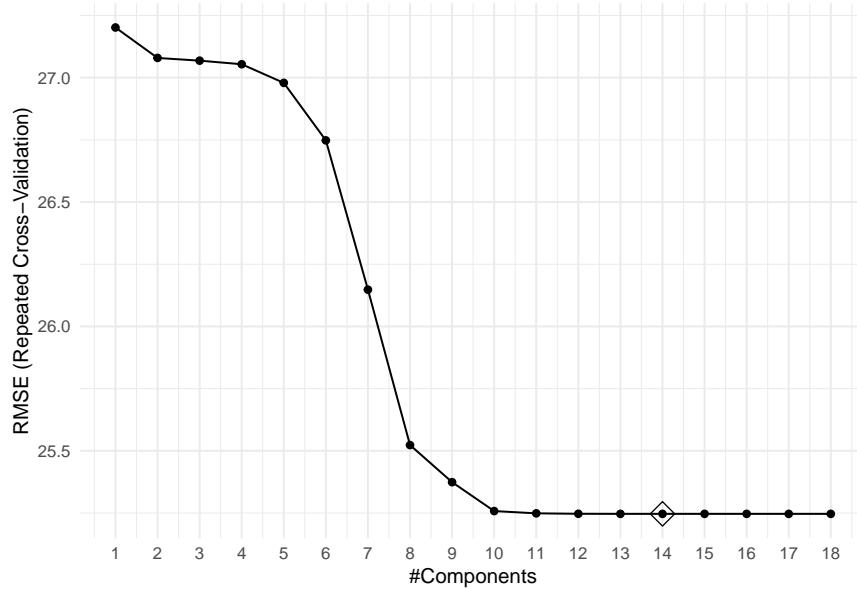


Elastic net regression

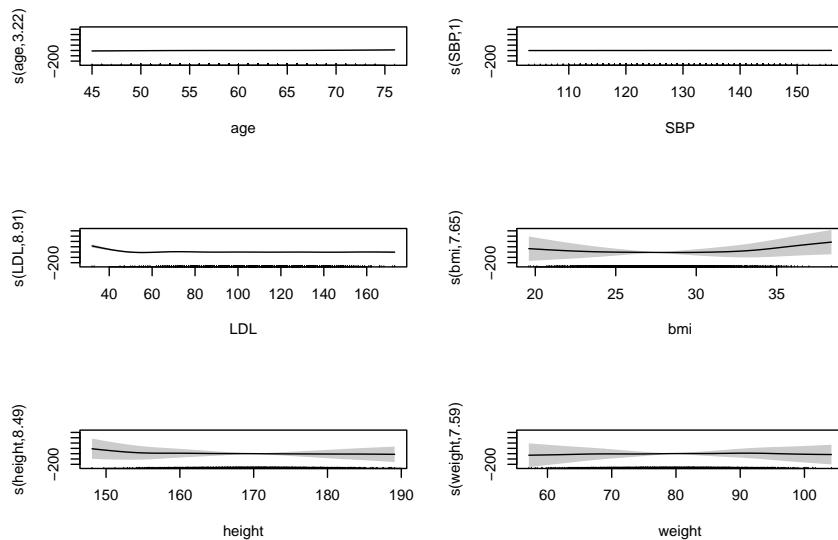


```
##      alpha     lambda
## 1001     1 0.1353353
```

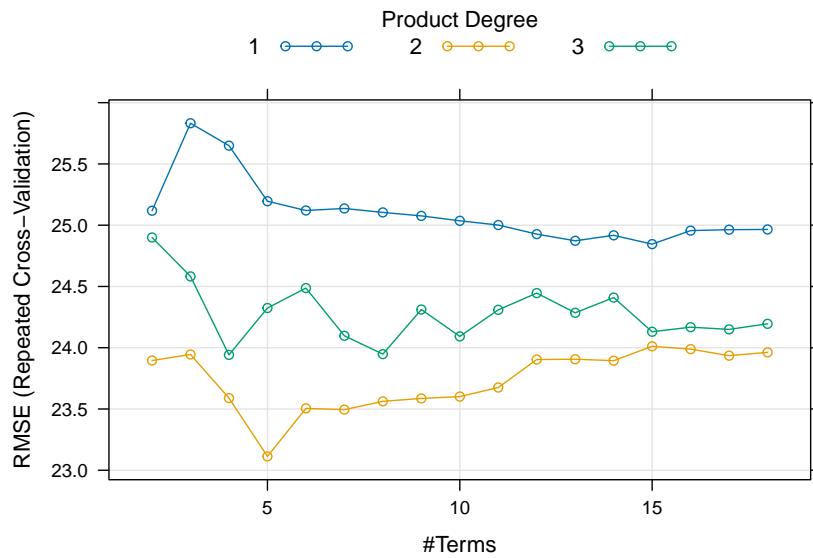
Partial least squares (PLS)



Generalized Additive Models (GAM)



Multivariate adaptive regression spline model (MARS)

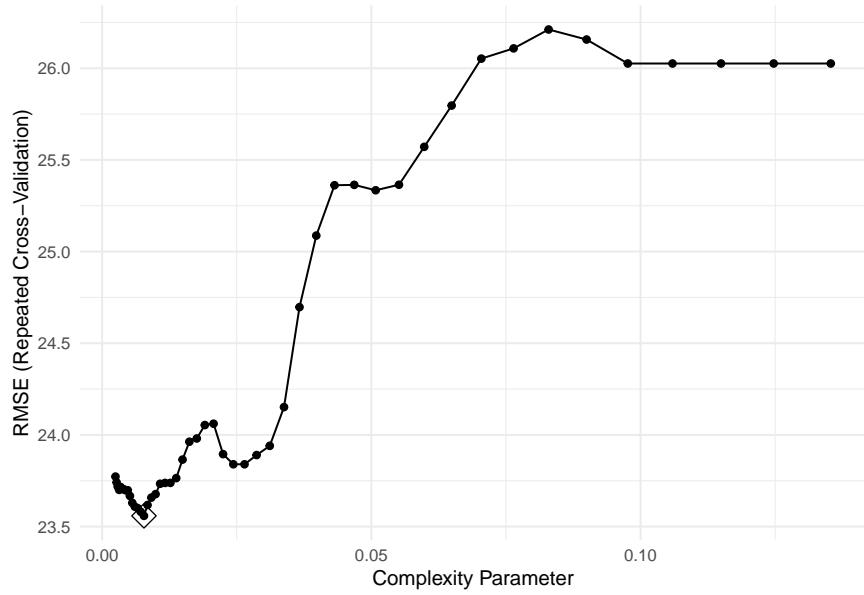


```
## Call: earth(x=matrix[2884,18], y=c(17,33,92,20,4...), keepxy=TRUE, degree=2,
##             nprune=5)
##
##                               coefficients
## (Intercept)              -2.941972
## vaccine1                 -9.506054
## h(bmi-24.5)                7.647737
## h(30.6-bmi)                7.320881
## h(bmi-30.6) * studyB    20.811559
##
## Selected 5 of 24 terms, and 3 of 18 predictors (nprune=5)
## Termination condition: Reached nk 37
## Importance: bmi, studyB, vaccine1, age-unused, gender1-unused, ...
## Number of terms at each degree of interaction: 1 3 1
## GCV 516.0731      RSS 1477023      GRSq 0.4041199      RSq 0.4082465
```

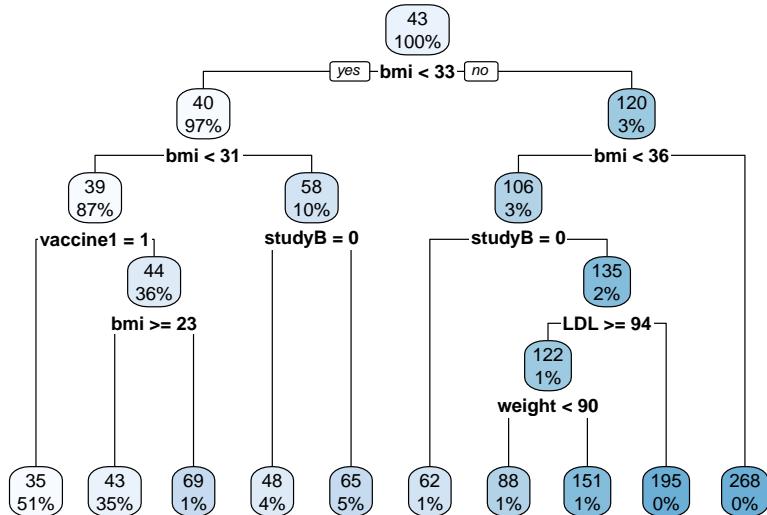
K-Nearest Neighbors (KNN)

```
##     k
## 9 13
```

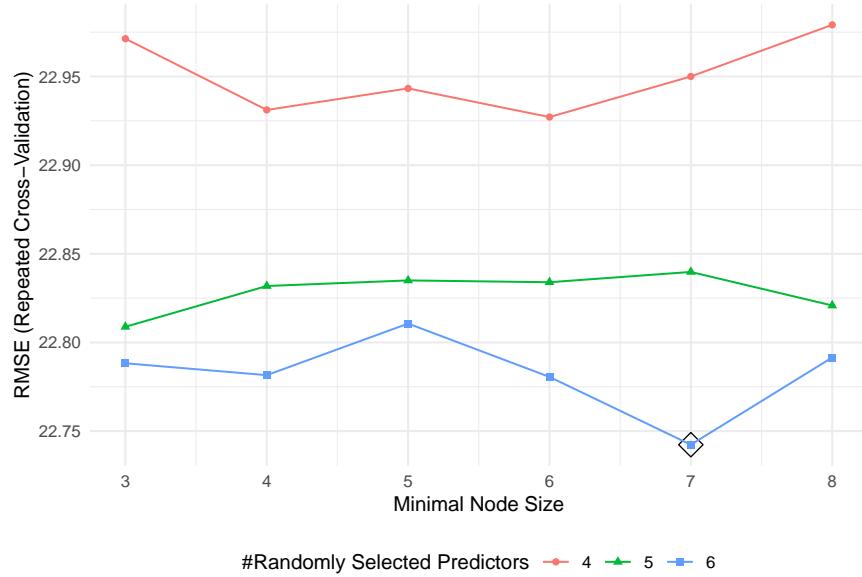
Regression Tree (CART)



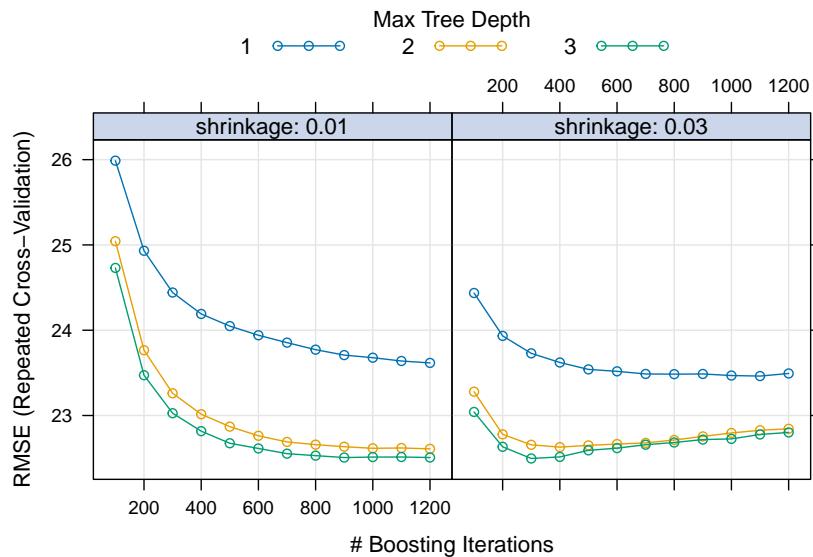
```
##          cp
## 15 0.00777266
```



Random Forest



Generalized Boosted Regression



Models comparsion based on cross validation error

```
## 
## Call:
## summary.resamples(object = resamp)
## 
## Models: lm, enet, pls, gam, mars, knn, gbm, cart, rf
## Number of resamples: 50
```

```

##  

## MAE  

##          Min. 1st Qu. Median   Mean 3rd Qu.   Max. NA's  

## lm    15.16547 16.10302 16.78177 16.83446 17.42209 19.04602 0  

## enet  14.56452 15.92803 16.52216 16.46752 17.01105 18.64227 0  

## pls   15.16548 16.10302 16.78177 16.83446 17.42208 19.04604 0  

## gam   14.35229 15.33438 15.83680 15.95120 16.56241 17.79521 0  

## mars  13.93931 14.88663 15.26803 15.37548 15.95721 17.55273 0  

## knn   13.77991 14.91692 15.66110 15.73718 16.40355 18.31290 0  

## gbm   13.16340 14.34295 14.85952 14.96472 15.45562 16.75934 0  

## cart  14.24679 14.87635 15.29857 15.47210 16.11928 17.75836 0  

## rf    13.66061 14.41903 14.83788 15.08850 15.72662 17.32203 0  

##  

## RMSE  

##          Min. 1st Qu. Median   Mean 3rd Qu.   Max. NA's  

## lm    19.74488 23.43892 25.17216 25.24668 27.21931 32.18024 0  

## enet  19.46546 23.14204 25.30888 25.71390 27.81579 33.96273 0  

## pls   19.74490 23.43892 25.17216 25.24668 27.21931 32.18025 0  

## gam   19.14744 22.27447 23.50243 24.21183 26.16733 31.05932 0  

## mars  18.46040 21.55759 22.41118 23.11308 24.68399 27.78102 0  

## knn   18.37738 23.09791 24.94296 26.01258 29.24863 36.95880 0  

## gbm   18.45737 21.10448 21.86910 22.49591 24.18224 27.22898 0  

## cart  18.92266 21.53195 22.80303 23.55843 25.57517 30.80530 0  

## rf    17.98350 21.08297 22.32057 22.74234 24.74996 28.19421 0  

##  

## Rsquared  

##          Min. 1st Qu. Median   Mean 3rd Qu.   Max. NA's  

## lm    0.08620897 0.1900059 0.2572871 0.2531044 0.3099187 0.4654627 0  

## enet  0.08078819 0.1811701 0.2277041 0.2290413 0.2731629 0.3775725 0  

## pls   0.08620900 0.1900061 0.2572879 0.2531045 0.3099185 0.4654625 0  

## gam   0.12269099 0.2416508 0.3294108 0.3350657 0.4183874 0.6223217 0  

## mars  0.07601256 0.2807337 0.3859944 0.3750641 0.4765129 0.6877643 0  

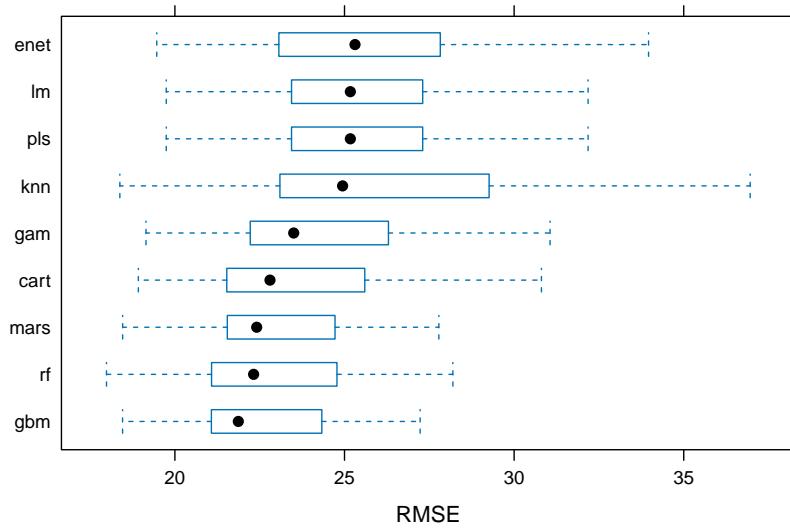
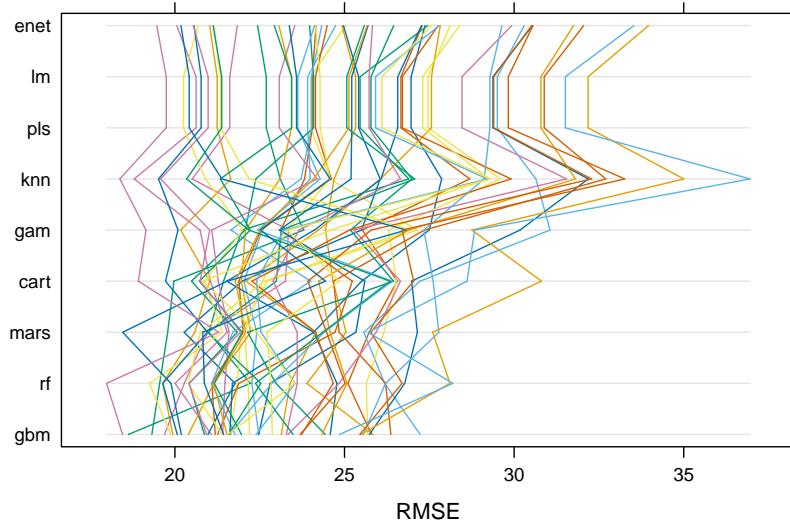
## knn   0.06268675 0.1685928 0.2095025 0.2280360 0.2800124 0.4727332 0  

## gbm   0.07050574 0.3111809 0.4021608 0.4026223 0.4940622 0.7041863 0  

## cart  0.05216633 0.2359045 0.3635301 0.3517752 0.4471816 0.6223333 0  

## rf    0.09295909 0.3246594 0.3947281 0.3929464 0.4732256 0.6063936 0

```

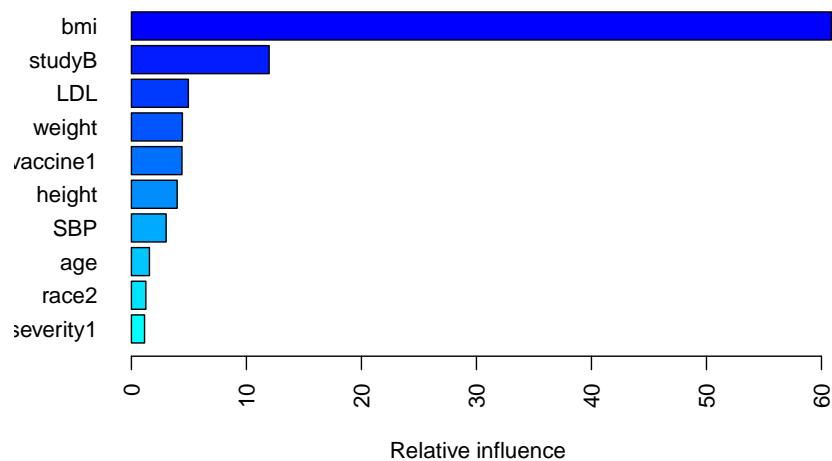


Results

Test Mean Squared Error

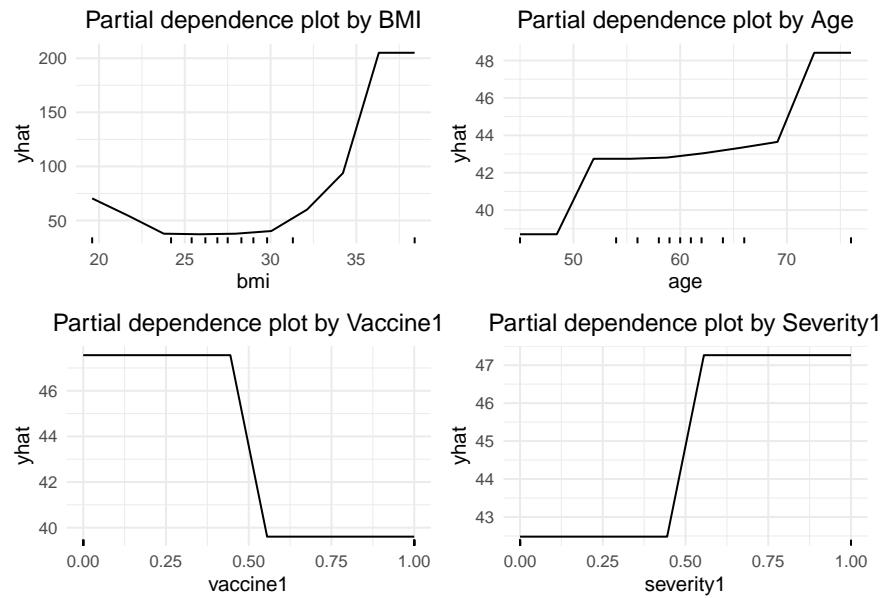
```
## [1] 492.5701
```

Variable importance plots

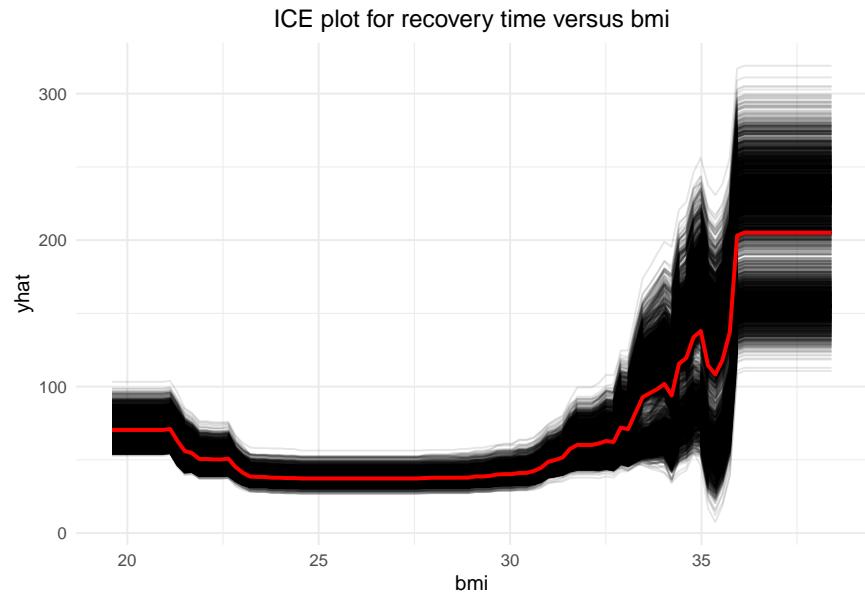


	rel.inf
bmi	60.8581546
studyB	11.9739399
LDL	4.9517050
weight	4.4347751
/vaccine1	4.3964715
height	3.9777372
SBP	3.0197676
age	1.5690459
race2	1.2559401
severity1	1.1426109
gender1	1.0789580
smoking2	0.5346932
smoking1	0.3760449
hypertension1	0.1605474
diabetes1	0.1503396
race4	0.1009740
race3	0.0182951
studyC	0.0000000

Partial dependence plots



Individual Conditional Expectation (ICE) plot



Secondary Analysis – Figures and outputs

Exploratory analysis for classification

Table 1: Data summary

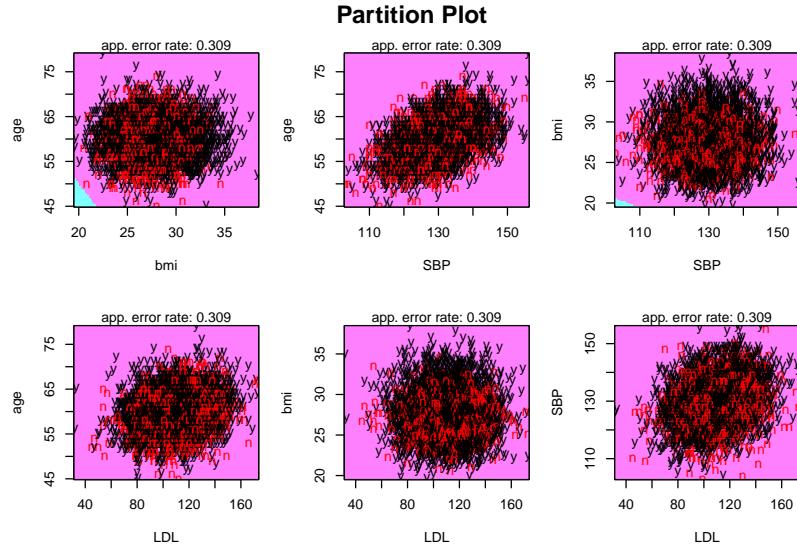
Name	Piped data
Number of rows	3603
Number of columns	17
Column type frequency:	
factor	9
numeric	8
Group variables	None

Variable type: factor

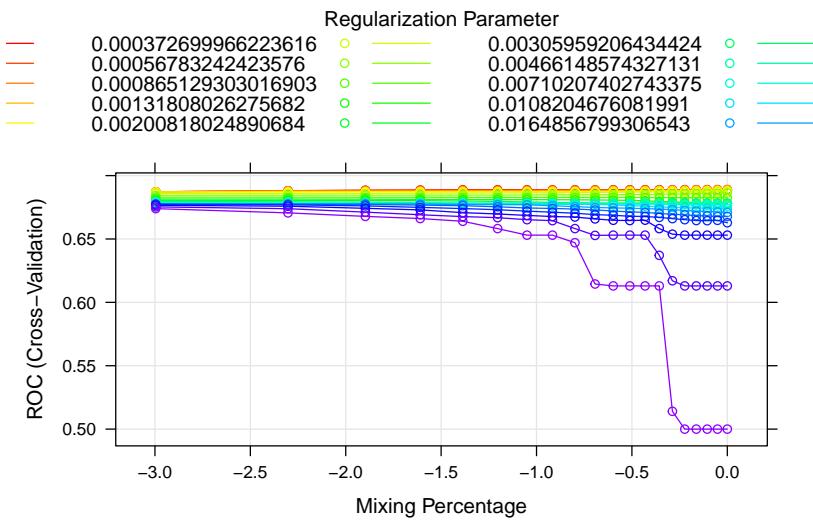
skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
gender	0	1	FALSE	2	0: 1867, 1: 1736
race	0	1	FALSE	4	1: 2350, 3: 720, 4: 358, 2: 175
smoking	0	1	FALSE	3	0: 2176, 1: 1072, 2: 355
hypertension	0	1	FALSE	2	0: 1886, 1: 1717
diabetes	0	1	FALSE	2	0: 3065, 1: 538
vaccine	0	1	FALSE	2	1: 2119, 0: 1484
severity	0	1	FALSE	2	0: 3252, 1: 351
study	0	1	FALSE	3	B: 2201, A: 718, C: 684
length_ind	0	1	FALSE	2	yes: 2491, no: 1112

Variable type: numeric

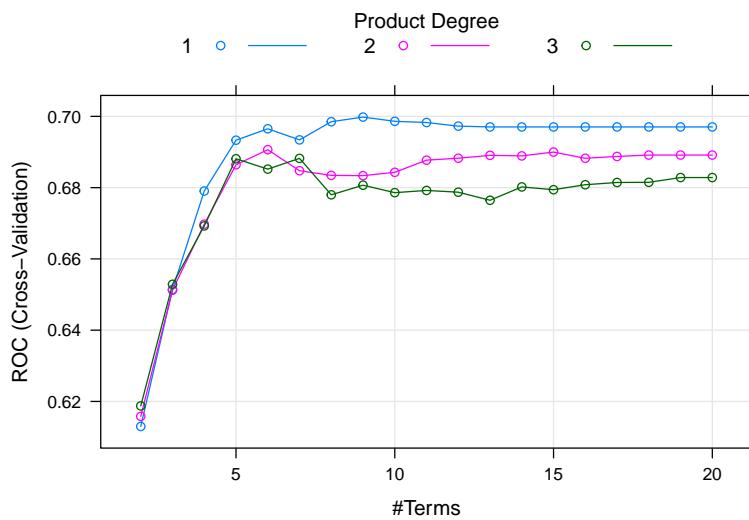
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
id	0	1	4979.49	2845.35	2.0	2528.0	4987.0	7438.0	9999.0	
age	0	1	60.10	4.46	45.0	57.0	60.0	63.0	79.0	
height	0	1	170.02	5.93	148.1	166.1	170.0	173.9	189.1	
weight	0	1	79.91	7.08	57.1	75.0	79.9	84.7	104.2	
bmi	0	1	27.71	2.77	19.6	25.8	27.6	29.4	38.4	
SBP	0	1	130.01	7.88	103.0	125.0	130.0	135.0	156.0	
LDL	0	1	110.51	19.83	32.0	97.0	111.0	125.0	173.0	
recovery_time	0	1	42.98	29.46	2.0	28.0	39.0	50.0	365.0	



Penalized logistic regression

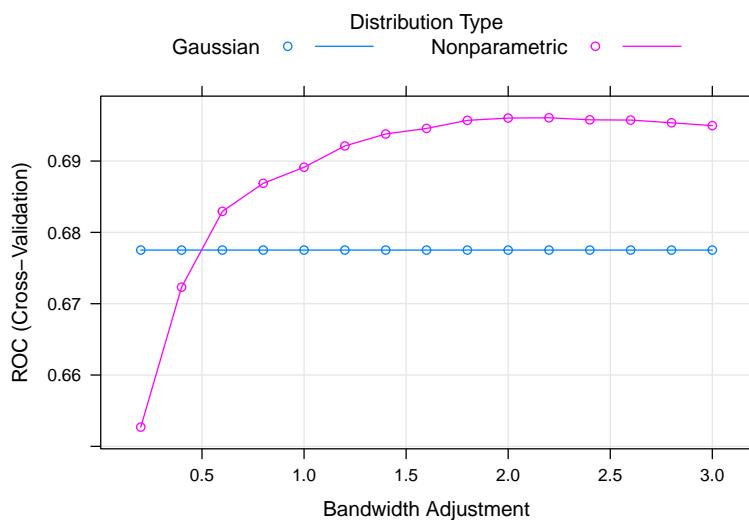


Multivariate adaptive regression splines classifier (MARS)

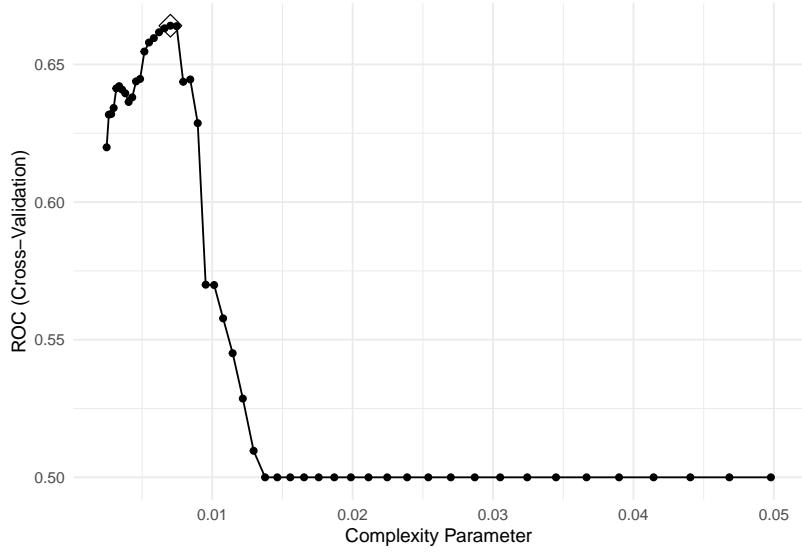


Quadratic discriminant analysis (QDA)

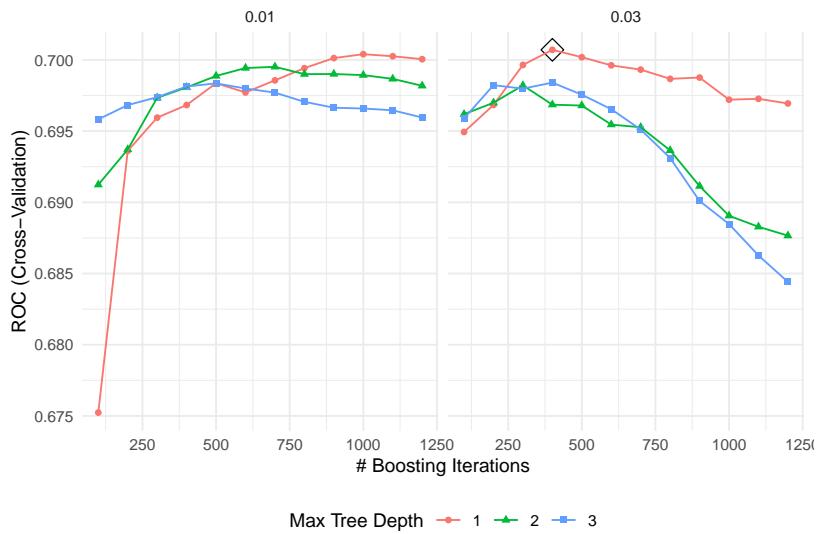
Naive Bayes Classifier



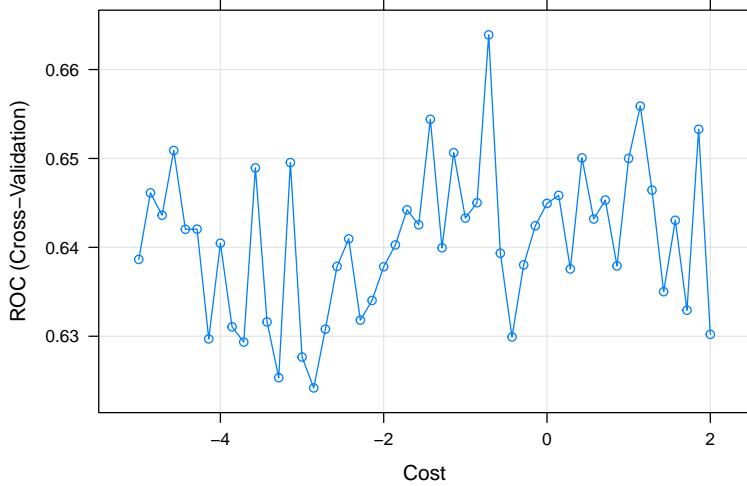
Classification Tree



Adaptive boosting classifier



Support vector classifier with linear kernel



Support vector classifier with radial kernel

