Data 624 - Predictive Analytics
Project 2

Table of Contents

Team Members

Mengqin Cai

Fan Xu

Sin Ying Wong

Overview

This is role playing in this project#2 assignment. Our team is a team of data scientists reporting to a new boss, the Head of Production Department at ABC Beverage. New regulations are requiring us to understand our manufacturing process, the predictive factors and be able to report to them our predictive model of PH.

By using the given historical data set, we will build and report the factors in both a technical and non-technical report. This report is the non-technical report, a business-friendly readable document, and our predictions will be provided separately in an Excel readable format. The technical report will clearly show the models we tested and how we selected our final approach.

Deliverables

A business-friendly readable document in Word.

A technical report in R.

A readable Excel with original predictors and our calculated prediction.

Load Package

We have used multiple packages in R: *tidyverse, rio, skimr, corrplot, VIM, Amelia, caret, recipes,* and *rsample.*

Load Data

We have two datasets. One is the training dataset `StudentData.xlsx`, and the other is the evaluation dataset `StudentEvaluation.xlsx`.

Exploratory Data Analysis

Both datasets include 31 numerical predictors and 1 categorical predictor in the dataset.

The responsible variable [PH] is continuous, therefore regression model is expected to be built.

From our study, only 1% of the data are missing, the predictor that contains most missing value is [MFR], this missing ratio is 212/2571 = 8.25%. Therefore, no predictor is suggested to be removed, imputation is to be included in the later data preprocess.

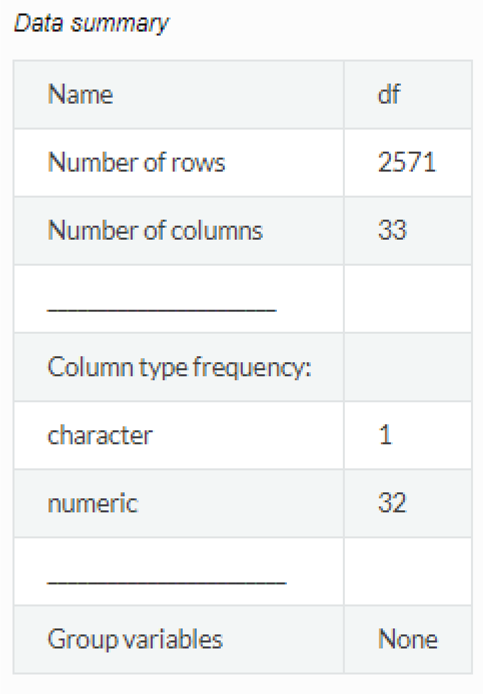
There are 4 rows in the training set which [PH] is missing, as imputing responsible variable is not meaningful in training set, therefore these 4 rows are suggested to be removed.

The majority of the continuous numerical predictors in both training set and evaluation set demonstrated skewed distribution, also some of the predictors contain negative values, therefore `*Yeo-Johnson*` transformation is used to remove the skewness.

A dummy variable will be created for categorical predictor [Brand.Code].

The pairwise correlation of predictors [Balling], [Hyd.Pressure3], [Density], [Balling.Lvl] and [Filler.Level], after missing value imputation, are greater than 0.9, therefore, they are suggested to be removed to avoid multicollinearity.

Training Data Summary



Data summary

Name	df
Number of rows	2571
Number of columns	33
________________	
Column type frequency:	
character	1
numeric	32
________________	
Group variables	None

Variable type: character

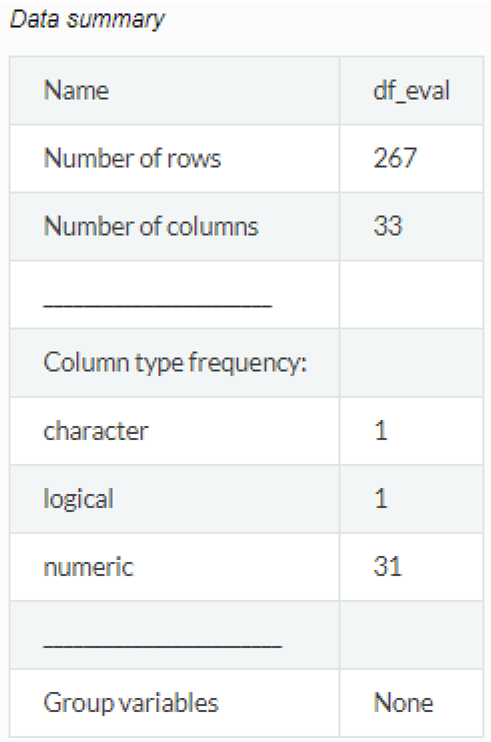
skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Brand.Code	120	0.95	1	1	0	4	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Carb.Volume	10	1.00	5.37	0.11	5.04	5.29	5.35	5.45	5.70	▁▆▇▅▁
Fill.Ounces	38	0.99	23.97	0.09	23.63	23.92	23.97	24.03	24.32	▁▂▇▂▁
PC.Volume	39	0.98	0.28	0.06	0.08	0.24	0.27	0.31	0.48	▁▃▇▂▁
Carb.Pressure	27	0.99	68.19	3.54	57.00	65.60	68.20	70.60	79.40	▁▅▇▃▁
Carb.Temp	26	0.99	141.09	4.04	128.60	138.40	140.80	143.80	154.00	▁▅▇▃▁
PSC	33	0.99	0.08	0.05	0.00	0.05	0.08	0.11	0.27	▆▇▃▁▁
PSC.Fill	23	0.99	0.20	0.12	0.00	0.10	0.18	0.26	0.62	▆▇▃▁▁
PSC.CO2	39	0.98	0.06	0.04	0.00	0.02	0.04	0.08	0.24	▇▅▂▁▁
Mnf.Flow	2	1.00	24.57	119.48	-100.20	-100.00	65.20	140.80	229.40	▇▁▁▇▂
Carb.Pressure1	32	0.99	122.59	4.74	105.60	119.00	123.20	125.40	140.20	▁▃▇▂▁
Fill.Pressure	22	0.99	47.92	3.18	34.60	46.00	46.40	50.00	60.40	▁▁▇▂▁
Hyd.Pressure1	11	1.00	12.44	12.43	-0.80	0.00	11.40	20.20	58.00	▇▅▂▁▁
Hyd.Pressure2	15	0.99	20.96	16.39	0.00	0.00	28.60	34.60	59.40	▇▂▇▅▁
Hyd.Pressure3	15	0.99	20.46	15.98	-1.20	0.00	27.60	33.40	50.00	▇▁▃▇▁

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Hyd.Pressure4	30	0.99	96.29	13.12	52.00	86.00	96.00	102.00	142.00	▁▃▇▂▁
Filler.Level	20	0.99	109.25	15.70	55.80	98.30	118.40	120.00	161.20	▁▂▅▇▁
Filler.Speed	57	0.98	3687.20	770.82	998.00	3888.00	3982.00	3998.00	4030.00	▁▁▁▁▇
Temperature	14	0.99	65.97	1.38	63.60	65.20	65.60	66.40	76.20	▇▂▁▁▁
Usage.cont	5	1.00	20.99	2.98	12.08	18.36	21.79	23.75	25.90	▁▃▅▃▇
Carb.Flow	2	1.00	2468.35	1073.70	26.00	1144.00	3028.00	3186.00	5104.00	▂▅▅▇▁
Density	1	1.00	1.17	0.38	0.24	0.90	0.98	1.62	1.92	▁▅▇▂▅
MFR	212	0.92	704.05	73.90	31.40	706.30	724.00	731.00	868.60	▁▁▁▂▇
Balling	1	1.00	2.20	0.93	-0.17	1.50	1.65	3.29	4.01	▁▇▇▁▇
Pressure.Vacuum	0	1.00	-5.22	0.57	-6.60	-5.60	-5.40	-5.00	-3.60	▂▇▅▂▁
PH	4	1.00	8.55	0.17	7.88	8.44	8.54	8.68	9.36	▁▅▇▂▁
Oxygen.Filler	12	1.00	0.05	0.05	0.00	0.02	0.03	0.06	0.40	▇▁▁▁▁
Bowl.Setpoint	2	1.00	109.33	15.30	70.00	100.00	120.00	120.00	140.00	▁▂▃▇▁
Pressure.Setpoint	12	1.00	47.62	2.04	44.00	46.00	46.00	50.00	52.00	▁▇▁▆▁
Air.Pressurer	0	1.00	142.83	1.21	140.80	142.20	142.60	143.00	148.20	▅▇▁▁▁
Alch.Rel	9	1.00	6.90	0.51	5.28	6.54	6.56	7.24	8.62	▁▇▁▂▁
Carb.Rel	10	1.00	5.44	0.13	4.96	5.34	5.40	5.54	6.06	▁▇▇▂▁
Balling.Lvl	1	1.00	2.05	0.87	0.00	1.38	1.48	3.14	3.66	▁▇▂▁▆

Evaluation Data Summary



Data summary

Name	df_eval
Number of rows	267
Number of columns	33
_______________________	
Column type frequency:	
character	1
logical	1
numeric	31
________________________	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Brand.Code	8	0.97	1	1	0	4	0

Variable type: logical

skim_variable	n_missing	complete_rate	mean	count
PH	267	0	NaN	:

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Carb.Volume	1	1.00	5.37	0.11	5.15	5.29	5.34	5.47	5.67	▂▇▃▅▁
Fill.Ounces	6	0.98	23.97	0.08	23.75	23.92	23.97	24.01	24.20	▁▅▇▃▁

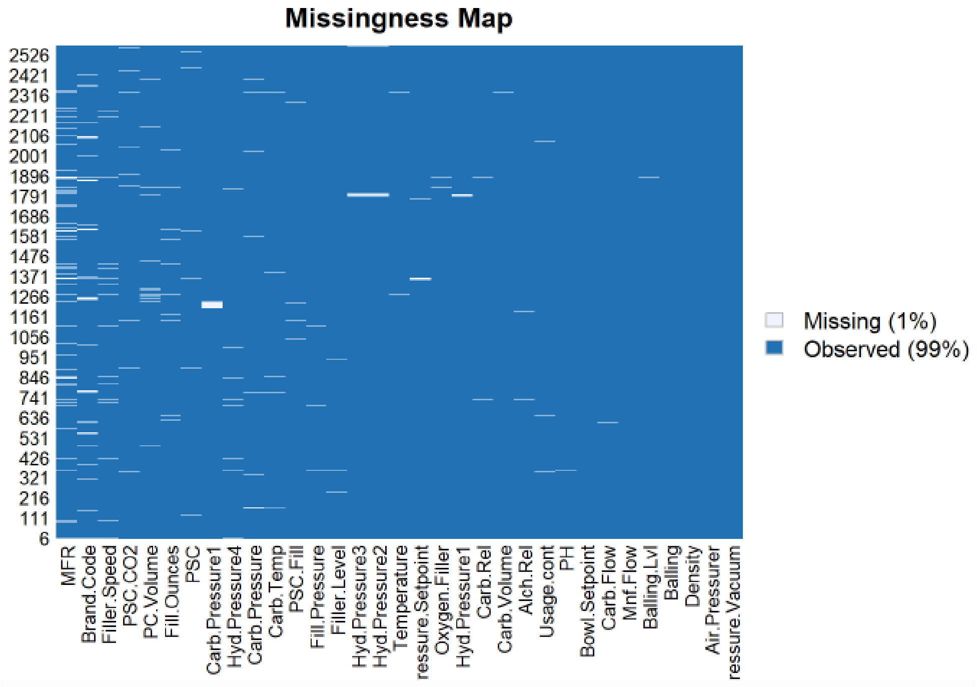
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
PC.Volume	4	0.99	0.28	0.06	0.10	0.23	0.28	0.32	0.46	▁▆▇▅▁
Carb.Pressure	0	1.00	68.25	3.86	60.20	65.30	68.00	70.60	77.60	▃▆▇▃▂
Carb.Temp	1	1.00	141.23	4.30	130.00	138.40	140.80	143.80	154.00	▁▆▇▃▁
PSC	5	0.98	0.09	0.05	0.00	0.04	0.08	0.11	0.25	▆▇▃▂▁
PSC.Fill	3	0.99	0.19	0.11	0.02	0.10	0.18	0.26	0.62	▇▇▃▁▁
PSC.CO2	5	0.98	0.05	0.04	0.00	0.02	0.04	0.06	0.24	▇▃▂▁▁
Mnf.Flow	0	1.00	21.03	117.76	-100.20	-100.00	0.20	141.30	220.40	▇▁▁▆▂
Carb.Pressure1	4	0.99	123.04	4.42	113.00	120.20	123.40	125.50	136.00	▃▃▇▂▁
Fill.Pressure	2	0.99	48.14	3.44	37.80	46.00	47.80	50.20	60.20	▁▇▇▂▁
Hyd.Pressure1	0	1.00	12.01	13.53	-50.00	0.00	10.40	20.40	50.00	▁▁▇▆▂
Hyd.Pressure2	1	1.00	20.11	17.21	-50.00	0.00	26.80	34.80	61.40	▁▁▆▇▁
Hyd.Pressure3	1	1.00	19.61	16.56	-50.00	0.00	27.70	33.00	49.20	▁▁▆▃▇
Hyd.Pressure4	4	0.99	97.84	13.92	68.00	90.00	98.00	104.00	140.00	▆▆▇▂▁
Filler.Level	2	0.99	110.29	15.50	69.20	100.60	118.60	120.20	153.20	▂▃▇▇▁
Filler.Speed	10	0.96	3581.39	911.19	1006.00	3812.00	3978.00	3996.00	4020.00	▁▁▁▁▇

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Temperature	2	0.99	66.23	1.69	63.80	65.40	65.80	66.60	75.40	▇▅▁▁▁
Usage.cont	2	0.99	20.90	3.00	12.90	18.12	21.44	23.74	24.60	▁▃▃▃▇
Carb.Flow	0	1.00	2408.64	1161.36	0.00	1083.00	3038.00	3215.00	3858.00	▂▃▁▇▇
Density	1	1.00	1.18	0.38	0.06	0.92	0.98	1.60	1.84	▁▁▇▁▅
MFR	31	0.88	697.80	96.40	15.60	707.00	724.60	731.45	784.80	▁▁▁▁▇
Balling	1	1.00	2.20	0.92	0.90	1.50	1.65	3.24	3.79	▅▇▁▂▅
Pressure.Vacuum	1	1.00	-5.17	0.58	-6.40	-5.60	-5.20	-4.80	-3.60	▁▇▆▃▁
Oxygen.Filler	3	0.99	0.05	0.05	0.00	0.02	0.03	0.05	0.40	▇▁▁▁▁
Bowl.Setpoint	1	1.00	109.62	15.02	70.00	100.00	120.00	120.00	130.00	▁▂▁▃▇
Pressure.Setpoint	2	0.99	47.73	2.06	44.00	46.00	46.00	50.00	52.00	▁▇▁▆▁
Air.Pressurer	1	1.00	142.83	1.23	141.20	142.20	142.60	142.80	147.20	▅▇▁▁▁
Alch.Rel	3	0.99	6.91	0.50	6.40	6.54	6.58	7.18	7.82	▇▁▂▁▂
Carb.Rel	2	0.99	5.44	0.13	5.18	5.34	5.40	5.56	5.74	▂▇▂▃▂
Balling.Lvl	0	1.00	2.05	0.88	0.00	1.38	1.48	3.08	3.42	▁▃▇▁▇

Missing Value View

As mentioned above, there are only 1% of the data are missing. Thus, no predictor was removed, but 4 rows with missing [PH] value were removed.

Below is a plot of missing value distribution in the training dataset `*df*`.



Numerical Predictor Correlation after Missing Data Imputation

We used kNN to impute missing values of the training dataset `df` and compute pair-wise correlations and locate the predictors with pair-wise correlation greater than 0.9.

The pairwise correlation of predictors [Balling], [Hyd.Pressure3], [Density], [Balling.Lvl] and [Filler.Level] are greater than 0.9. Therefore, they are suggested to be removed to avoid multicollinearity.

```
findCorrelation(df %>%
                  kNN() %>%
                  select(!ends_with('imp'), -c(Brand.Code, PH)) %>%
                  cor(),
                cutoff = 0.9,
                names = TRUE,
                verbose = TRUE)
```

```
## [1] "Balling"       "Hyd.Pressure3" "Density"       "Balling.Lvl"
## [5] "Filler.Level"
```

Data Preprocess

From the dataset summary sections above, we know that most of the continuous numerical predictors in both training set and evaluation set demonstrated skewed distribution. Also, some of the predictors contain negative values. Therefore, `*Yeo-Johnson*` transformation is used to remove the skewness.

A dummy variable will be created for categorical predictor [Brand.Code].

For the training dataset `df`:

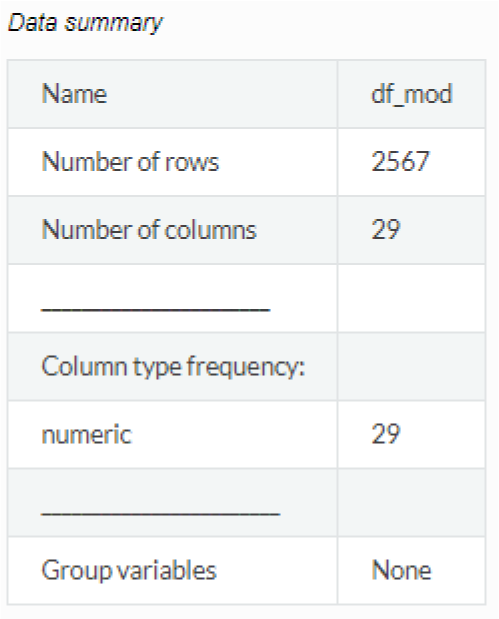
- Remove rows where PH is empty or NA
- Perform train-test-split at ratio 4:1

For both training and evaluation datasets:

- Impute missing values using bag trees
- create dummy variable for categorical variables
- center and scale numerical variables
- remove skewness of numerical variables
- remove predictors with near zero variance
- remove predictors with correlation greater than 0.9

Note that, although data preprocess can be performed during model training, however, as there are multiple models to be built in the later section, preprocessing data in advanced is more efficient than doing it during each model run.

Below is the data summary of the pre-processed training set (before train-test-split) `*df_mod*`.



Data summary

Name	df_mod
Number of rows	2567
Number of columns	29
________________________	
Column type frequency:	
numeric	29
________________________	
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Carb.Volume	0	1	0.00	1.00	-3.11	-0.72	-0.22	0.81	3.10	▁▅▇▃▁
Fill.Ounces	0	1	0.00	1.00	-3.93	-0.63	-0.02	0.60	3.97	▁▂▇▂▁
PC.Volume	0	1	0.00	1.00	-3.28	-0.63	-0.10	0.58	3.31	▁▃▇▂▁
Carb.Pressure	0	1	0.00	1.00	-3.16	-0.74	0.00	0.67	3.15	▁▅▇▃▁
Carb.Temp	0	1	0.00	1.00	-3.08	-0.67	-0.08	0.66	3.17	▁▅▇▃▁
PSC	0	1	0.00	1.00	-1.69	-0.71	-0.14	0.56	3.78	▅▇▂▁▁
PSC.Fill	0	1	0.00	1.00	-1.67	-0.81	-0.13	0.55	3.62	▅▇▂▁▁
PSC.CO2	0	1	0.00	1.00	-1.32	-0.85	-0.39	0.55	4.29	▇▅▂▁▁
Mnf.Flow	0	1	0.00	1.00	-1.04	-1.04	0.38	0.97	1.71	▇▁▁▇▂
Carb.Pressure1	0	1	0.00	1.00	-3.60	-0.76	0.14	0.60	3.75	▁▃▇▂▁
Fill.Pressure	0	1	0.00	1.00	-4.19	-0.60	-0.48	0.66	3.93	▁▁▇▂▁
Hyd.Pressure1	0	1	0.00	1.00	-1.06	-1.00	-0.08	0.63	3.67	▇▅▂▁▁
Hyd.Pressure2	0	1	0.00	1.00	-1.27	-1.27	0.47	0.84	2.35	▇▂▇▅▁
Hyd.Pressure4	0	1	0.00	1.00	-2.62	-0.80	-0.04	0.42	3.46	▂▆▇▂▁
Temperature	0	1	0.00	1.00	-1.71	-0.56	-0.27	0.30	7.34	▇▃▁▁▁
Usage.cont	0	1	0.00	1.00	-3.00	-0.88	0.26	0.92	1.65	▁▂▃▂▇
Carb.Flow	0	1	0.00	1.00	-2.29	-1.22	0.52	0.67	2.46	▂▅▅▇▁
MFR	0	1	0.00	1.00	-5.13	0.17	0.38	0.45	1.55	▁▁▁▂▇

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Pressure.Vacuum	0	1	0.00	1.00	-2.43	-0.67	-0.32	0.38	2.83	▂▇▅▃▁
Oxygen.Filler	0	1	0.00	1.00	-0.98	-0.55	-0.29	0.30	7.83	▇▁▁▁▁
Bowl.Setpoint	0	1	0.00	1.00	-2.57	-0.61	0.70	0.70	2.00	▁▂▃▇▁
Pressure.Setpoint	0	1	0.00	1.00	-1.77	-0.79	-0.79	1.17	2.16	▁▇▁▅▁
Air.Pressurer	0	1	0.00	1.00	-1.68	-0.52	-0.19	0.14	4.42	▅▇▁▁▁
Alch.Rel	0	1	0.00	1.00	-3.20	-0.71	-0.67	0.66	3.41	▁▇▃▂▁
Carb.Rel	0	1	0.00	1.00	-3.70	-0.75	-0.28	0.80	4.85	▁▇▆▂▁
PH	0	1	8.55	0.17	7.88	8.44	8.54	8.68	9.36	▁▅▇▂▁
Brand.Code_B	0	1	0.00	1.00	-1.00	-1.00	1.00	1.00	1.00	▇▁▁▁▇
Brand.Code_C	0	1	0.00	1.00	-0.41	-0.41	-0.41	-0.41	2.43	▇▁▁▁▂
Brand.Code_D	0	1	0.00	1.00	-0.56	-0.56	-0.56	-0.56	1.78	▇▁▁▁▂

Model Building

Three categories of regression models are to be built in this section, including Linear Regression Models, Non-linear Regression Models and Tree-based Models. The model with best performance in the test dataset will be selected as the final model.

The models to be built are as below:

- Linear Regression Models: PLS, Ridge, LASSO and Elastic Net
- Non-linear Regression Models: KNN, SVM-Linear, SVM-Radial, MARS and Neural Network
- Tree-based Regression Models: Random Forest, Gradient Boosting Machine and Cubist

Linear Regression Models

PLS Regression

Using 10-fold cross-validation as train control, the PLS regression model gives:

- The 7th model is the optimal model.
- The corresponding resampled estimate of RMSE and R2 are 0.1362656 and 0.3739715 respectively.

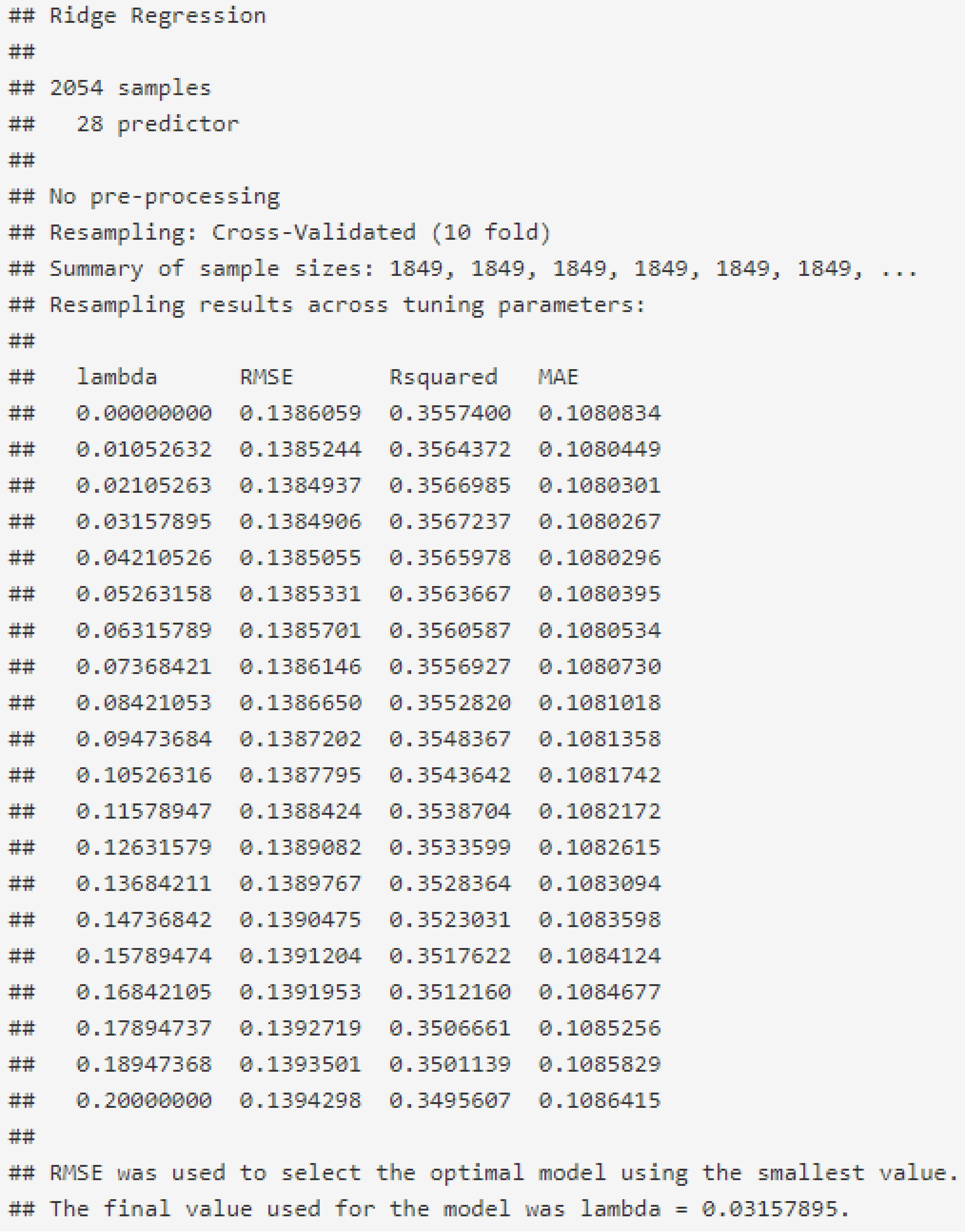
```
## Partial Least Squares
##
## 2054 samples
##   28 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1849, 1849, 1849, 1849, 1849, 1849, ...
## Resampling results across tuning parameters:
##
##   ncomp  RMSE       Rsquared   MAE
##    1     0.1497005  0.2470540  0.1176600
##    2     0.1430215  0.3139339  0.1116965
##    3     0.1413576  0.3297154  0.1108805
##    4     0.1396517  0.3458175  0.1093216
##    5     0.1390031  0.3516492  0.1085059
##    6     0.1384918  0.3566973  0.1080004
##    7     0.1384305  0.3573092  0.1081537
##    8     0.1384597  0.3570316  0.1080082
##    9     0.1385041  0.3566531  0.1080056
##   10     0.1385358  0.3563692  0.1080224
##   11     0.1385680  0.3560643  0.1080587
##   12     0.1385836  0.3559539  0.1080834
##   13     0.1385914  0.3558839  0.1080780
##   14     0.1385636  0.3561045  0.1080470
##   15     0.1385706  0.3560451  0.1080609
##   16     0.1385782  0.3559804  0.1080730
##   17     0.1385796  0.3559731  0.1080733
##   18     0.1385952  0.3558288  0.1080827
##   19     0.1386021  0.3557690  0.1080826
##   20     0.1386004  0.3557876  0.1080814
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was ncomp = 7.
```

```
##      RMSE  Rsquared       MAE
## 0.1362656 0.3739715 0.1064367
```

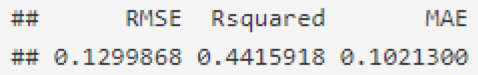
Ridge Regression

Using 10-fold cross-validation as train control, the ridge regression model gives:

- The optimal model with lambda = 0.03157895
- The corresponding resampled estimate of RMSE and R2 are 0.1299868 and 0.4415918 respectively.



```
## Ridge Regression
##
## 2054 samples
##   28 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1849, 1849, 1849, 1849, 1849, 1849, ...
## Resampling results across tuning parameters:
##
##   lambda      RMSE       Rsquared   MAE
##   0.00000000  0.1386059  0.3557400  0.1080834
##   0.01052632  0.1385244  0.3564372  0.1080449
##   0.02105263  0.1384937  0.3566985  0.1080301
##   0.03157895  0.1384906  0.3567237  0.1080267
##   0.04210526  0.1385055  0.3565978  0.1080296
##   0.05263158  0.1385331  0.3563667  0.1080395
##   0.06315789  0.1385701  0.3560587  0.1080534
##   0.07368421  0.1386146  0.3556927  0.1080730
##   0.08421053  0.1386650  0.3552820  0.1081018
##   0.09473684  0.1387202  0.3548367  0.1081358
##   0.10526316  0.1387795  0.3543642  0.1081742
##   0.11578947  0.1388424  0.3538704  0.1082172
##   0.12631579  0.1389082  0.3533599  0.1082615
##   0.13684211  0.1389767  0.3528364  0.1083094
##   0.14736842  0.1390475  0.3523031  0.1083598
##   0.15789474  0.1391204  0.3517622  0.1084124
##   0.16842105  0.1391953  0.3512160  0.1084677
##   0.17894737  0.1392719  0.3506661  0.1085256
##   0.18947368  0.1393501  0.3501139  0.1085829
##   0.20000000  0.1394298  0.3495607  0.1086415
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was lambda = 0.03157895.
```



```
##      RMSE  Rsquared       MAE
## 0.1299868 0.4415918 0.1021300
```

Lasso

The Lasso regression model gives:

- The optimal model with fraction = 0.1
- The corresponding resampled estimate of RMSE and R2 are 0.1561285 and 0.2961838 respectively.

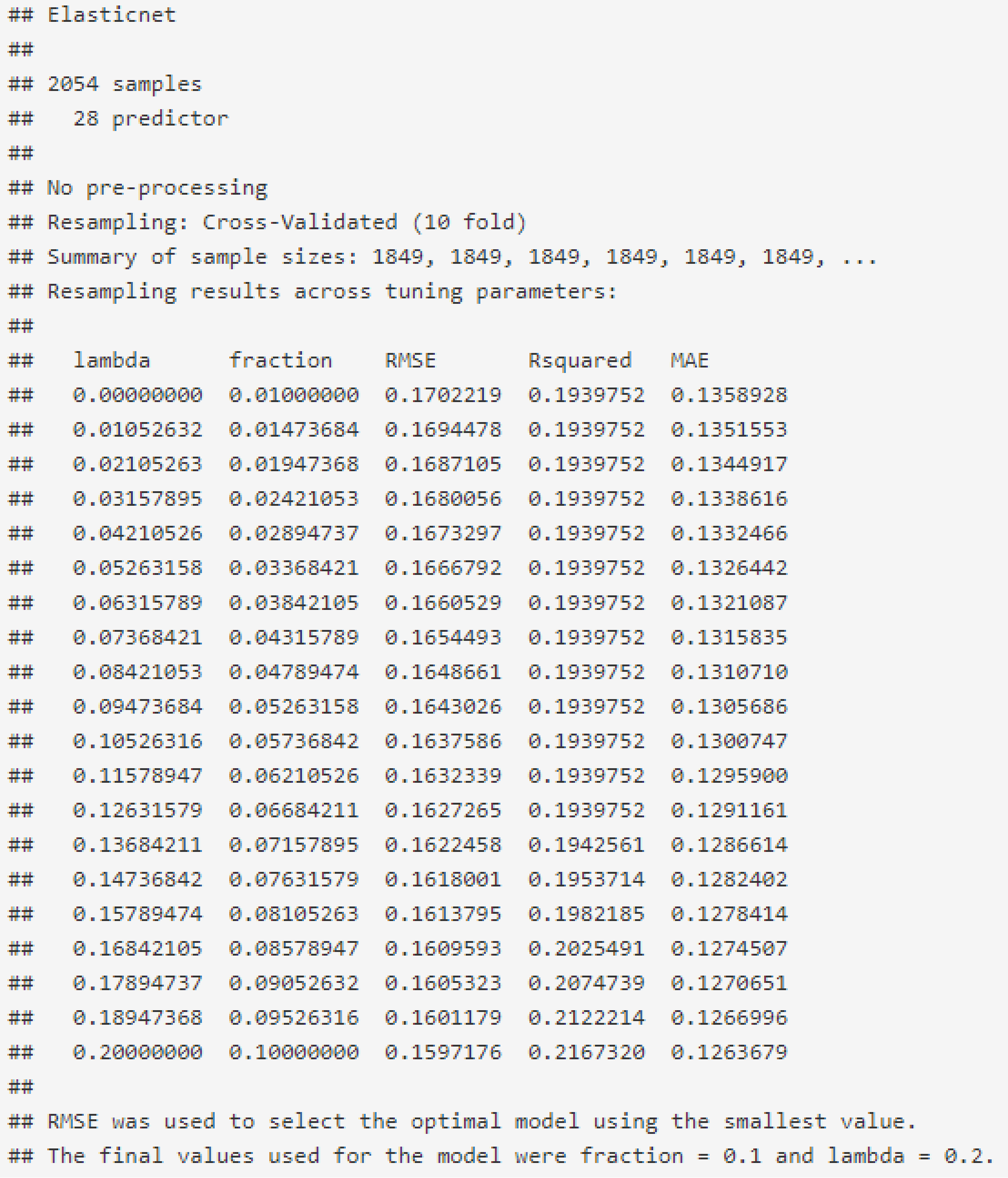
```
## The lasso
##
## 2054 samples
##   28 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1849, 1849, 1849, 1849, 1849, 1849, ...
## Resampling results across tuning parameters:
##
##   fraction    RMSE       Rsquared   MAE
##   0.01000000  0.1702219  0.1939752  0.1358928
##   0.01473684  0.1693460  0.1939752  0.1350627
##   0.01947368  0.1684926  0.1939752  0.1342976
##   0.02421053  0.1676620  0.1939752  0.1335504
##   0.02894737  0.1668545  0.1939752  0.1328056
##   0.03368421  0.1660705  0.1939752  0.1321218
##   0.03842105  0.1653103  0.1939752  0.1314585
##   0.04315789  0.1645743  0.1939752  0.1308063
##   0.04789474  0.1638627  0.1939752  0.1301621
##   0.05263158  0.1631759  0.1939752  0.1295284
##   0.05736842  0.1625142  0.1939752  0.1289066
##   0.06210526  0.1619037  0.1954704  0.1283338
##   0.06684211  0.1613301  0.1989758  0.1277900
##   0.07157895  0.1607555  0.2047889  0.1272578
##   0.07631579  0.1601689  0.2114635  0.1267473
##   0.08105263  0.1595945  0.2174196  0.1262680
##   0.08578947  0.1590325  0.2227269  0.1257959
##   0.09052632  0.1584829  0.2274517  0.1253301
##   0.09526316  0.1579511  0.2316237  0.1248748
##   0.10000000  0.1574345  0.2354022  0.1244261
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was fraction = 0.1.
```

```
##      RMSE  Rsquared       MAE
## 0.1561285 0.2961838 0.1274395
```

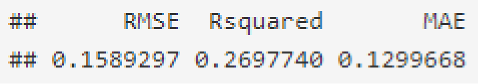
Elastic Net

The elastic net regression model gives:

- The optimal model with fraction = 0.1 and lambda = 0.2
- The corresponding resampled estimate of RMSE and R2 are 0.1589297 and 0.2697740 respectively.



```
## Elasticnet
##
## 2054 samples
##   28 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1849, 1849, 1849, 1849, 1849, 1849, ...
## Resampling results across tuning parameters:
##
##   lambda      fraction    RMSE       Rsquared   MAE
##   0.00000000  0.01000000  0.1702219  0.1939752  0.1358928
##   0.01052632  0.01473684  0.1694478  0.1939752  0.1351553
##   0.02105263  0.01947368  0.1687105  0.1939752  0.1344917
##   0.03157895  0.02421053  0.1680056  0.1939752  0.1338616
##   0.04210526  0.02894737  0.1673297  0.1939752  0.1332466
##   0.05263158  0.03368421  0.1666792  0.1939752  0.1326442
##   0.06315789  0.03842105  0.1660529  0.1939752  0.1321087
##   0.07368421  0.04315789  0.1654493  0.1939752  0.1315835
##   0.08421053  0.04789474  0.1648661  0.1939752  0.1310710
##   0.09473684  0.05263158  0.1643026  0.1939752  0.1305686
##   0.10526316  0.05736842  0.1637586  0.1939752  0.1300747
##   0.11578947  0.06210526  0.1632339  0.1939752  0.1295900
##   0.12631579  0.06684211  0.1627265  0.1939752  0.1291161
##   0.13684211  0.07157895  0.1622458  0.1942561  0.1286614
##   0.14736842  0.07631579  0.1618001  0.1953714  0.1282402
##   0.15789474  0.08105263  0.1613795  0.1982185  0.1278414
##   0.16842105  0.08578947  0.1609593  0.2025491  0.1274507
##   0.17894737  0.09052632  0.1605323  0.2074739  0.1270651
##   0.18947368  0.09526316  0.1601179  0.2122214  0.1266996
##   0.20000000  0.10000000  0.1597176  0.2167320  0.1263679
##
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were fraction = 0.1 and lambda = 0.2.
```



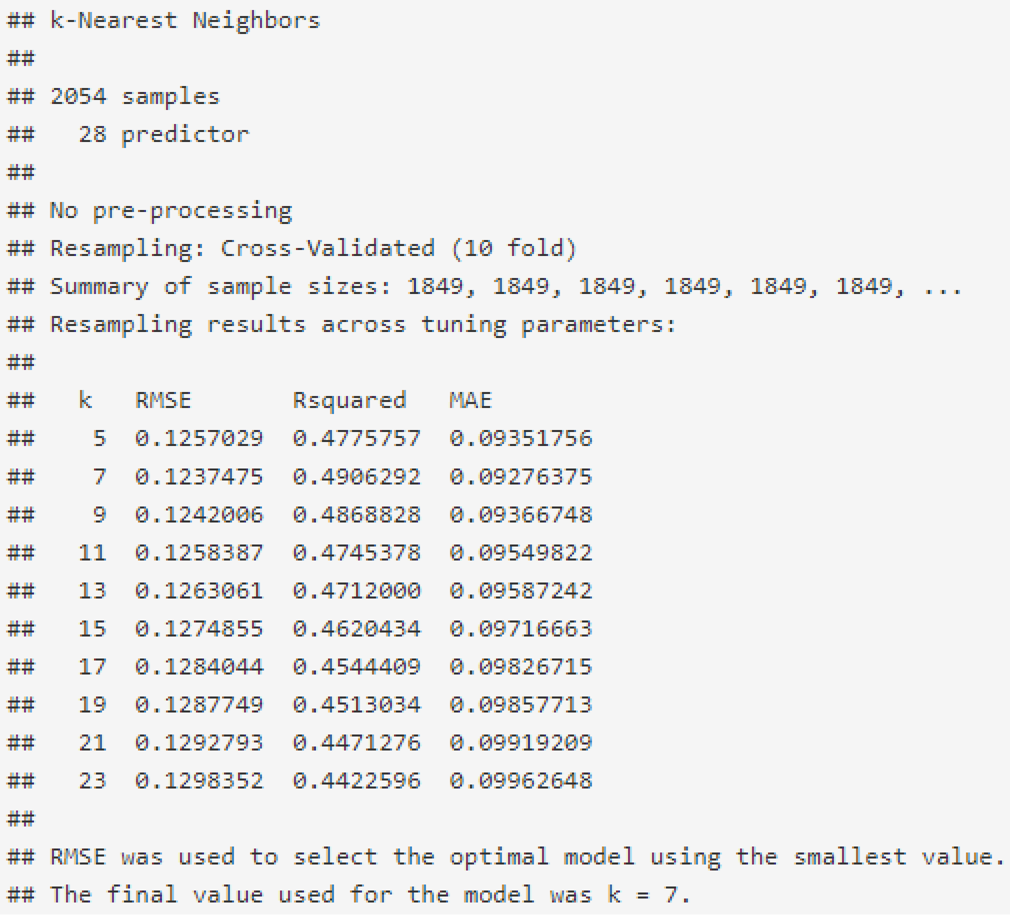
```
##      RMSE  Rsquared       MAE
## 0.1589297 0.2697740 0.1299668
```

Non-Linear Regression Models

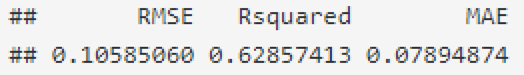
KNN

The kNN regression model gives:

- The optimal model with k=7
- The corresponding resampled estimate of RMSE and R2 are 0.10585060 and 0.62857413 respectively.



```
## k-Nearest Neighbors
##
## 2054 samples
##   28 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1849, 1849, 1849, 1849, 1849, 1849, ...
## Resampling results across tuning parameters:
##
##   k   RMSE       Rsquared   MAE
##    5  0.1257029  0.4775757  0.09351756
##    7  0.1237475  0.4906292  0.09276375
##    9  0.1242006  0.4868828  0.09366748
##   11  0.1258387  0.4745378  0.09549822
##   13  0.1263061  0.4712000  0.09587242
##   15  0.1274855  0.4620434  0.09716663
##   17  0.1284044  0.4544409  0.09826715
##   19  0.1287749  0.4513034  0.09857713
##   21  0.1292793  0.4471276  0.09919209
##   23  0.1298352  0.4422596  0.09962648
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was k = 7.
```

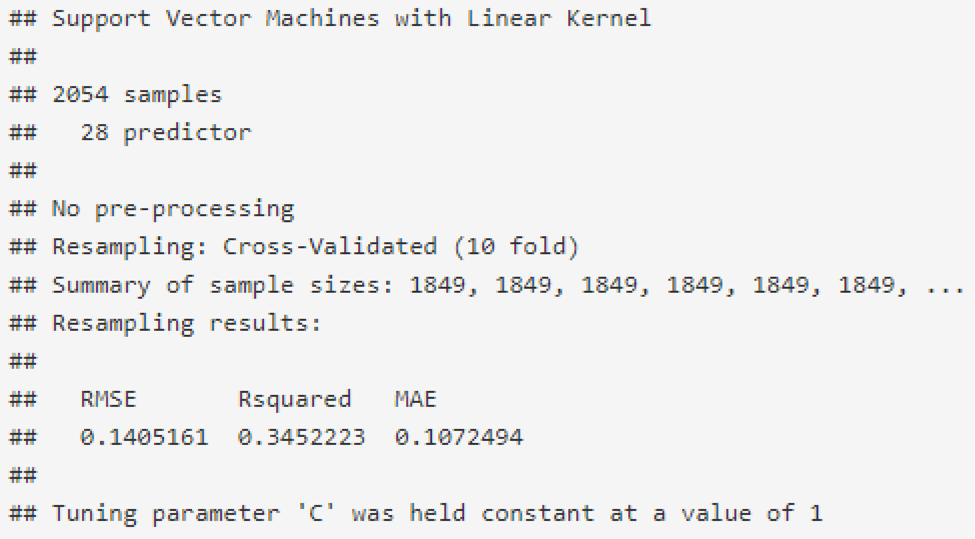


```
##       RMSE   Rsquared        MAE
## 0.10585060 0.62857413 0.07894874
```

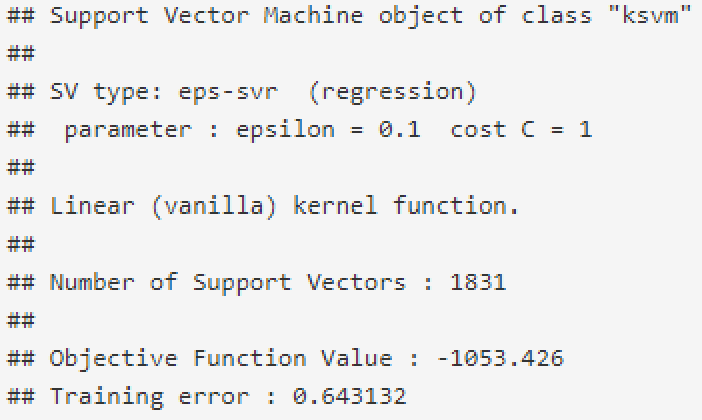
SVM-Linear

The SVM-Linear regression model gives:

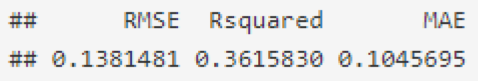
- The optimal model with epsilon = 0.1 and cost C = 1
- The corresponding resampled estimate of RMSE and R2 are 0.1381481 and 0.3615830 respectively.



```
## Support Vector Machines with Linear Kernel
##
## 2054 samples
##   28 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1849, 1849, 1849, 1849, 1849, 1849, ...
## Resampling results:
##
##   RMSE       Rsquared   MAE
##   0.1405161  0.3452223  0.1072494
##
## Tuning parameter 'C' was held constant at a value of 1
```



```
## Support Vector Machine object of class "ksvm"
##
## SV type: eps-svr  (regression)
##  parameter : epsilon = 0.1  cost C = 1
##
## Linear (vanilla) kernel function.
##
## Number of Support Vectors : 1831
##
## Objective Function Value : -1053.426
## Training error : 0.643132
```



```
##      RMSE  Rsquared       MAE
## 0.1381481 0.3615830 0.1045695
```

SVM-Radial

The SVM-Radial regression model gives:

- The optimal model with sigma = 0.0242724 and cost C = 4
- The corresponding resampled estimate of RMSE and R2 are 0.08011998 and 0.79263724 respectively.

```
## Support Vector Machines with Radial Basis Function Kernel
##
## 2054 samples
##   28 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1849, 1849, 1849, 1849, 1849, 1849, ...
## Resampling results across tuning parameters:
##
##   C        RMSE       Rsquared   MAE
##      0.25  0.1286431  0.4526820  0.09577483
##      0.50  0.1256923  0.4758057  0.09278004
##      1.00  0.1231104  0.4952829  0.09035109
##      2.00  0.1210941  0.5106732  0.08867772
##      4.00  0.1204826  0.5158644  0.08851988
##      8.00  0.1212283  0.5141755  0.08924725
##     16.00  0.1224728  0.5116971  0.09033769
##     32.00  0.1258334  0.4986777  0.09296903
##     64.00  0.1326503  0.4687005  0.09806454
##    128.00  0.1389973  0.4449388  0.10296902
##    256.00  0.1452495  0.4218464  0.10818173
##    512.00  0.1510565  0.4016687  0.11316640
##   1024.00  0.1519305  0.3984248  0.11383537
##   2048.00  0.1519305  0.3984248  0.11383537
##   4096.00  0.1519305  0.3984248  0.11383537
##
## Tuning parameter 'sigma' was held constant at a value of 0.0242724
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were sigma = 0.0242724 and C = 4.
```

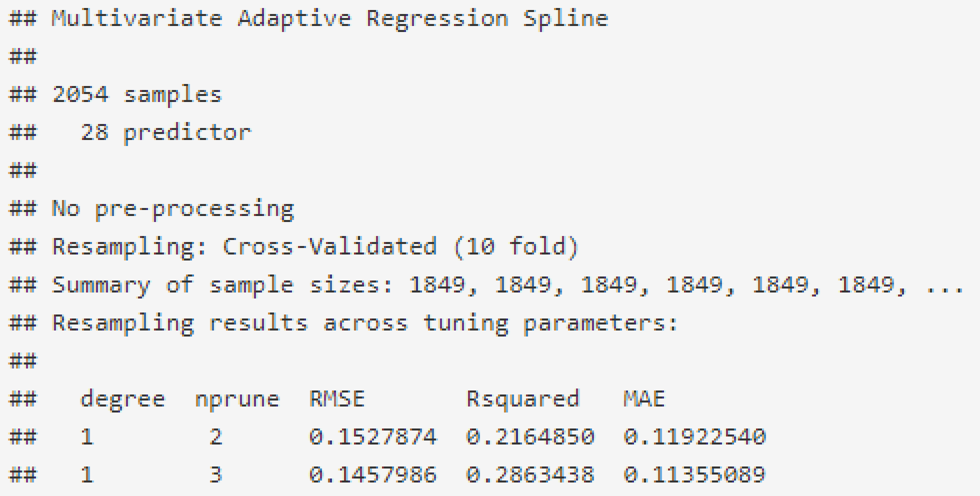
```
## Support Vector Machine object of class "ksvm"
##
## SV type: eps-svr  (regression)
##  parameter : epsilon = 0.1  cost C = 4
##
## Gaussian Radial Basis kernel function.
##  Hyperparameter : sigma =  0.0242723997688406
##
## Number of Support Vectors : 1748
##
## Objective Function Value : -2289.491
## Training error : 0.216318
```

```
##       RMSE   Rsquared        MAE
## 0.08011998 0.79263724 0.05028598
```

MARS

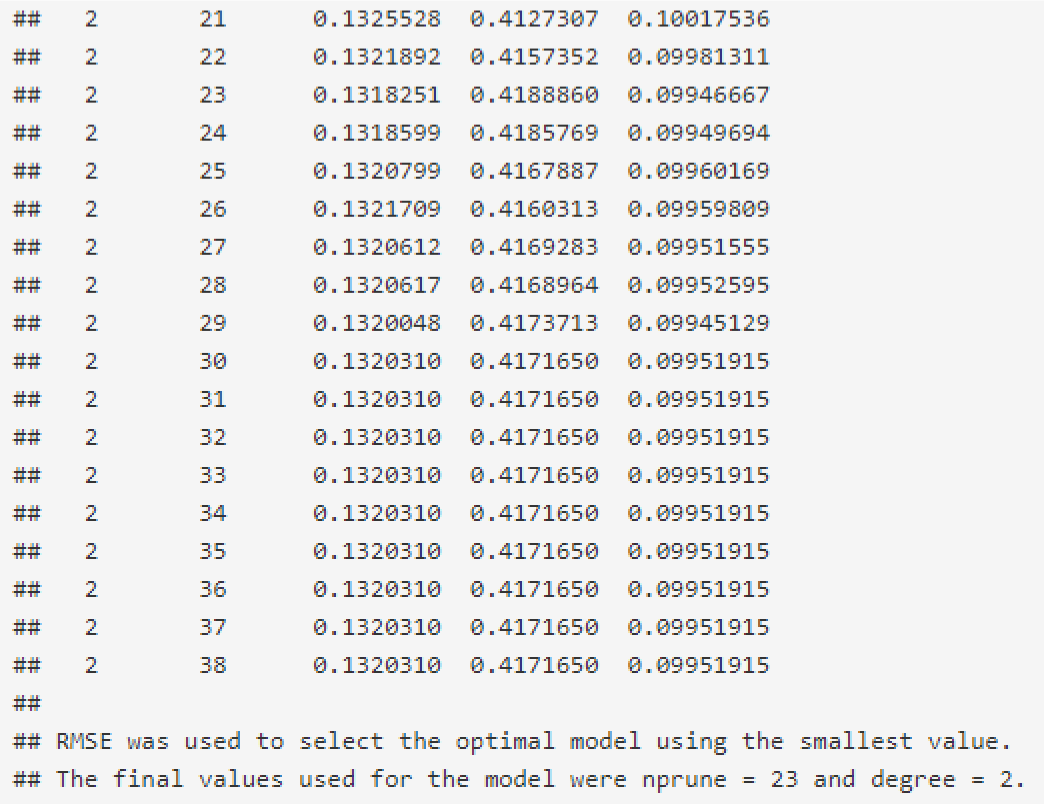
The MARS regression model gives:

- The optimal model with nprune = 23 and degree = 2
- The corresponding resampled estimate of RMSE and R2 are 0.12396741 and 0.49036903 respectively.

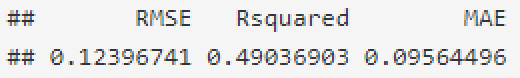


```
## Multivariate Adaptive Regression Spline
##
## 2054 samples
##   28 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1849, 1849, 1849, 1849, 1849, 1849, ...
## Resampling results across tuning parameters:
##
##   degree  nprune  RMSE       Rsquared   MAE
##   1        2      0.1527874  0.2164850  0.11922540
##   1        3      0.1457986  0.2863438  0.11355089
```

… … …



```
##   2       21      0.1325528  0.4127307  0.10017536
##   2       22      0.1321892  0.4157352  0.09981311
##   2       23      0.1318251  0.4188860  0.09946667
##   2       24      0.1318599  0.4185769  0.09949694
##   2       25      0.1320799  0.4167887  0.09960169
##   2       26      0.1321709  0.4160313  0.09959809
##   2       27      0.1320612  0.4169283  0.09951555
##   2       28      0.1320617  0.4168964  0.09952595
##   2       29      0.1320048  0.4173713  0.09945129
##   2       30      0.1320310  0.4171650  0.09951915
##   2       31      0.1320310  0.4171650  0.09951915
##   2       32      0.1320310  0.4171650  0.09951915
##   2       33      0.1320310  0.4171650  0.09951915
##   2       34      0.1320310  0.4171650  0.09951915
##   2       35      0.1320310  0.4171650  0.09951915
##   2       36      0.1320310  0.4171650  0.09951915
##   2       37      0.1320310  0.4171650  0.09951915
##   2       38      0.1320310  0.4171650  0.09951915
##
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were nprune = 23 and degree = 2.
```

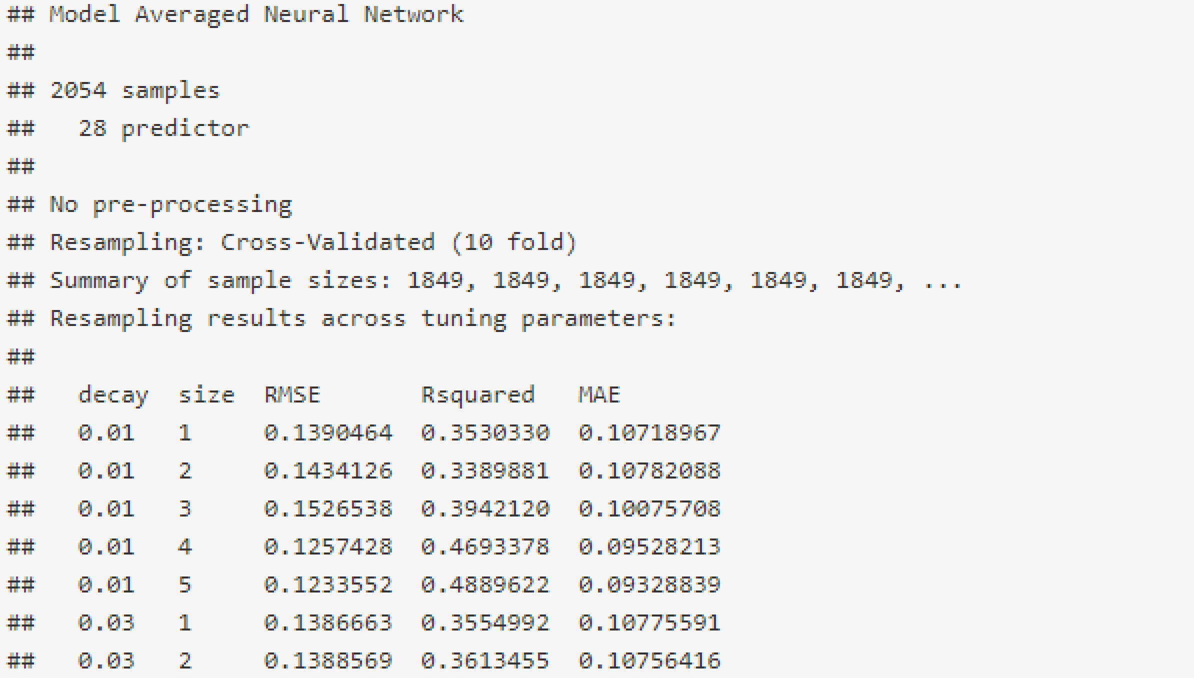


```
##       RMSE   Rsquared        MAE
## 0.12396741 0.49036903 0.09564496
```

Neural Network

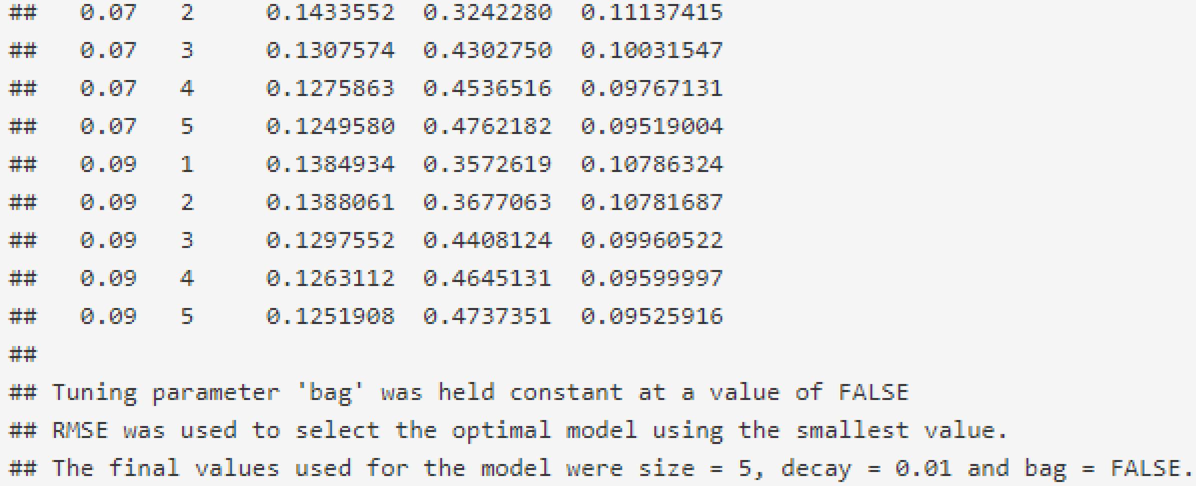
The neural network regression model gives:

- The optimal model with size = 5 and decay = 0.01
- The corresponding resampled estimate of RMSE and R2 are 0.11423783 and R2 0.56938536 respectively.

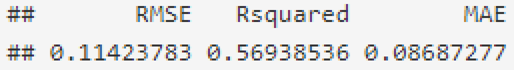


```
## Model Averaged Neural Network
##
## 2054 samples
##   28 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1849, 1849, 1849, 1849, 1849, 1849, ...
## Resampling results across tuning parameters:
##
##   decay  size  RMSE       Rsquared   MAE
##   0.01   1     0.1390464  0.3530330  0.10718967
##   0.01   2     0.1434126  0.3389881  0.10782088
##   0.01   3     0.1526538  0.3942120  0.10075708
##   0.01   4     0.1257428  0.4693378  0.09528213
##   0.01   5     0.1233552  0.4889622  0.09328839
##   0.03   1     0.1386663  0.3554992  0.10775591
##   0.03   2     0.1388569  0.3613455  0.10756416
```

… … …



```
##   0.07   2     0.1433552  0.3242280  0.11137415
##   0.07   3     0.1307574  0.4302750  0.10031547
##   0.07   4     0.1275863  0.4536516  0.09767131
##   0.07   5     0.1249580  0.4762182  0.09519004
##   0.09   1     0.1384934  0.3572619  0.10786324
##   0.09   2     0.1388061  0.3677063  0.10781687
##   0.09   3     0.1297552  0.4408124  0.09960522
##   0.09   4     0.1263112  0.4645131  0.09599997
##   0.09   5     0.1251908  0.4737351  0.09525916
##
## Tuning parameter 'bag' was held constant at a value of FALSE
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were size = 5, decay = 0.01 and bag = FALSE.
```



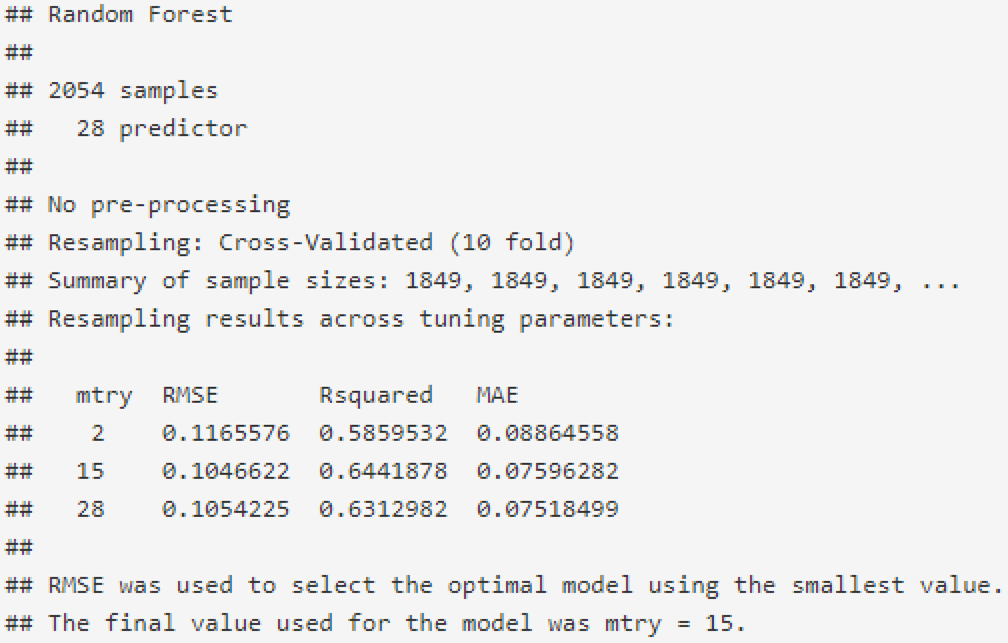
```
##       RMSE   Rsquared        MAE
## 0.11423783 0.56938536 0.08687277
```

Tree-Based Regression Models

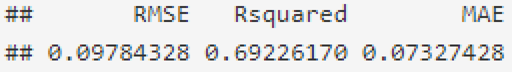
Random Forest

The random forest regression model gives:

- The optimal model with mtry = 15
- The corresponding resampled estimate of RMSE and R2 are 0.09784328 and 0.69226170 respectively.



```
## Random Forest
##
## 2054 samples
##   28 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1849, 1849, 1849, 1849, 1849, 1849, ...
## Resampling results across tuning parameters:
##
##   mtry  RMSE       Rsquared   MAE
##    2    0.1165576  0.5859532  0.08864558
##   15    0.1046622  0.6441878  0.07596282
##   28    0.1054225  0.6312982  0.07518499
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 15.
```

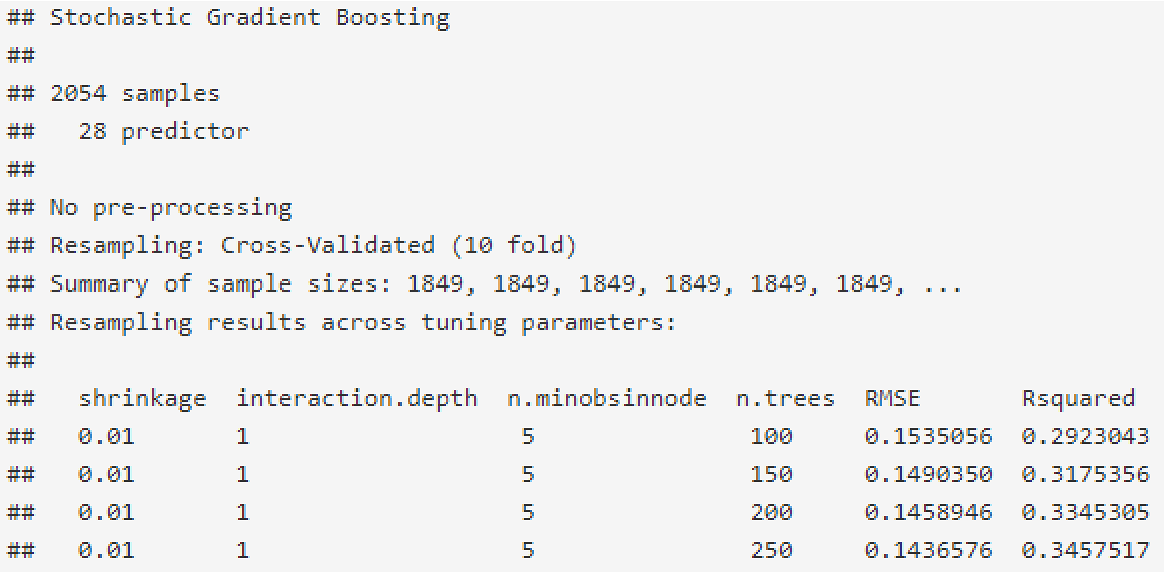


```
##       RMSE   Rsquared        MAE
## 0.09784328 0.69226170 0.07327428
```

Gradient Boosting Machine

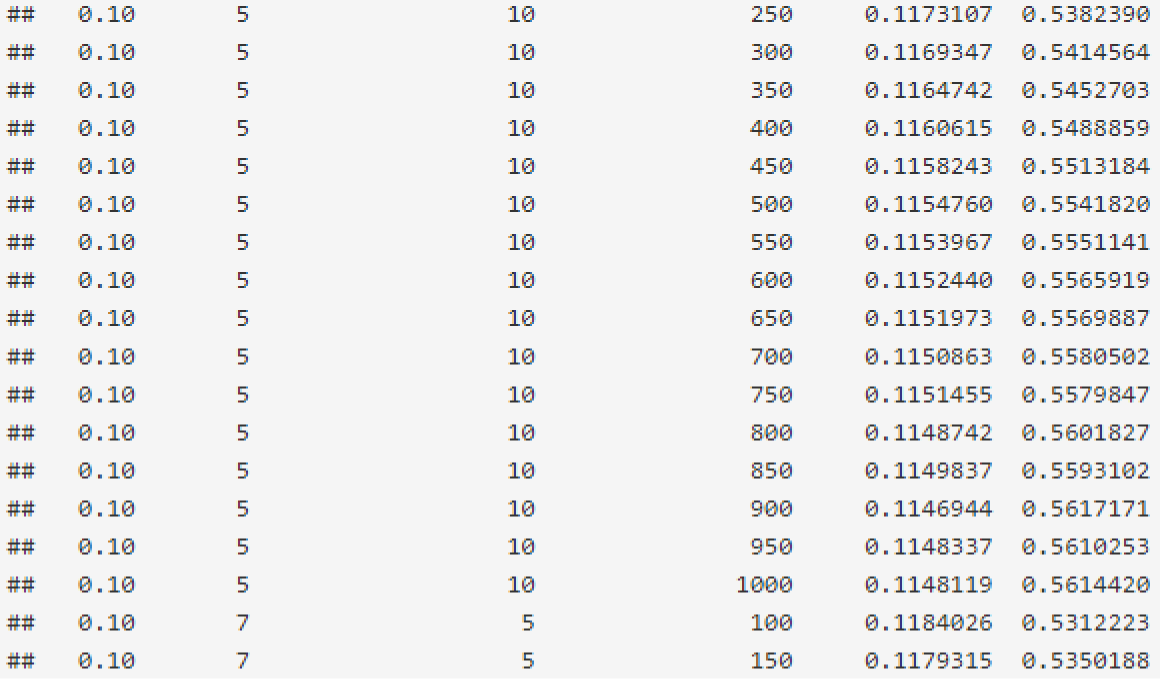
The gradient boosting machine regression model gives:

- The optimal model with shrinkage = 0.1, interaction.depth = 5, n.minobsinnode = 10, and n.trees = 900
- The corresponding resampled estimate of RMSE and R2 are 0.1104675 and 0.5972602 respectively.



```
## Stochastic Gradient Boosting
##
## 2054 samples
##   28 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1849, 1849, 1849, 1849, 1849, 1849, ...
## Resampling results across tuning parameters:
##
##   shrinkage  interaction.depth  n.minobsinnode  n.trees  RMSE       Rsquared
##   0.01       1                   5               100     0.1535056  0.2923043
##   0.01       1                   5               150     0.1490350  0.3175356
##   0.01       1                   5               200     0.1458946  0.3345305
##   0.01       1                   5               250     0.1436576  0.3457517
```

… … …



```
##   0.10       5                  10               250     0.1173107  0.5382390
##   0.10       5                  10               300     0.1169347  0.5414564
##   0.10       5                  10               350     0.1164742  0.5452703
##   0.10       5                  10               400     0.1160615  0.5488859
##   0.10       5                  10               450     0.1158243  0.5513184
##   0.10       5                  10               500     0.1154760  0.5541820
##   0.10       5                  10               550     0.1153967  0.5551141
##   0.10       5                  10               600     0.1152440  0.5565919
##   0.10       5                  10               650     0.1151973  0.5569887
##   0.10       5                  10               700     0.1150863  0.5580502
##   0.10       5                  10               750     0.1151455  0.5579847
##   0.10       5                  10               800     0.1148742  0.5601827
##   0.10       5                  10               850     0.1149837  0.5593102
##   0.10       5                  10               900     0.1146944  0.5617171
##   0.10       5                  10               950     0.1148337  0.5610253
##   0.10       5                  10              1000     0.1148119  0.5614420
##   0.10       7                   5               100     0.1184026  0.5312223
##   0.10       7                   5               150     0.1179315  0.5350188
```

```
##
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were n.trees = 900, interaction.depth =
##  5, shrinkage = 0.1 and n.minobsinnode = 10.
```

```
##      RMSE  Rsquared       MAE
## 0.1104675 0.5972602 0.0845282
```

Cubist

The cubist regression model gives:

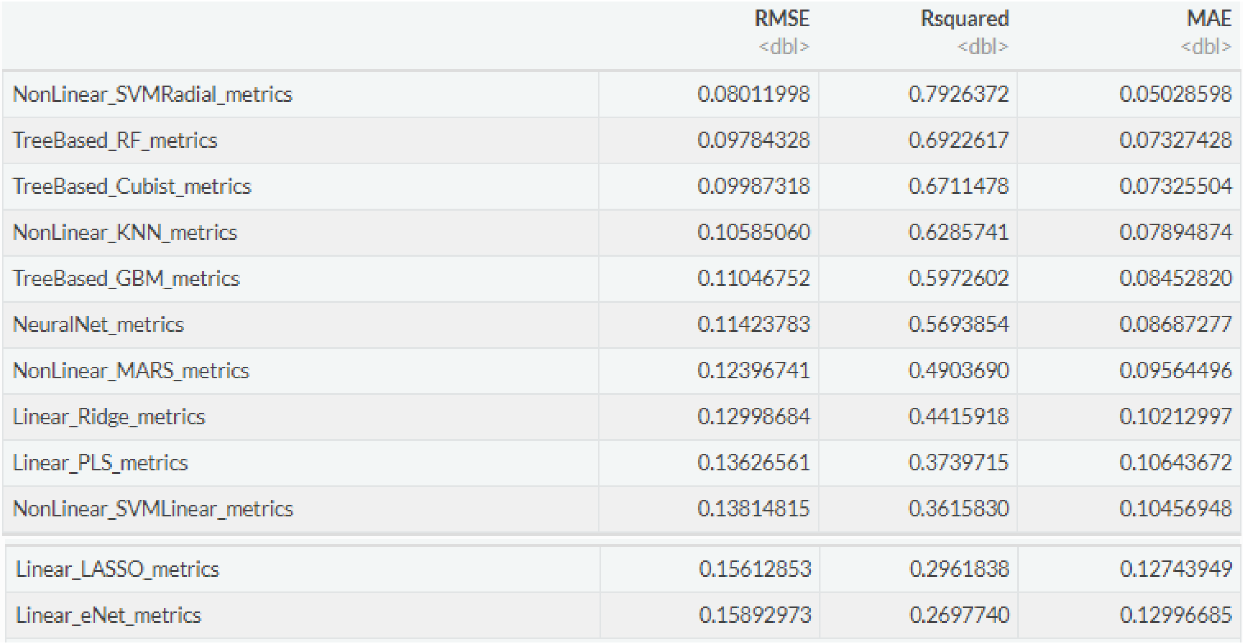
- The optimal model with committees = 20 and neighbors = 5
- The corresponding resampled estimate of RMSE and R2 are 0.09987318 and 0.67114775 respectively.

```
## Cubist
##
## 2054 samples
##   28 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1849, 1849, 1849, 1849, 1849, 1849, ...
## Resampling results across tuning parameters:
##
##   committees  neighbors  RMSE       Rsquared   MAE
##    1          0          0.1286602  0.4755080  0.08985362
##    1          5          0.1239531  0.5287986  0.08520446
##    1          9          0.1236001  0.5245257  0.08516113
##   10          0          0.1117588  0.5832072  0.08091727
##   10          5          0.1054796  0.6275221  0.07473292
##   10          9          0.1054210  0.6270252  0.07520324
##   20          0          0.1107426  0.5919454  0.08024516
##   20          5          0.1042786  0.6350231  0.07382106
##   20          9          0.1043447  0.6342982  0.07434306
##
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were committees = 20 and neighbors = 5.
```

```
##       RMSE   Rsquared        MAE
## 0.09987318 0.67114775 0.07325504
```

Model Selection

By comparing the RMSE and R-squared values from all above models, the SVM-Radial model has both lowest RMSE and highest R2. Therefore, it is selected to be the best model.



	RMSE	Rsquared	MAE
	<dbl>	<dbl>	<dbl>
NonLinear_SVMRadial_metrics	0.08011998	0.7926372	0.05028598
TreeBased_RF_metrics	0.09784328	0.6922617	0.07327428
TreeBased_Cubist_metrics	0.09987318	0.6711478	0.07325504
NonLinear_KNN_metrics	0.10585060	0.6285741	0.07894874
TreeBased_GBM_metrics	0.11046752	0.5972602	0.08452820
NeuralNet_metrics	0.11423783	0.5693854	0.08687277
NonLinear_MARS_metrics	0.12396741	0.4903690	0.09564496
Linear_Ridge_metrics	0.12998684	0.4415918	0.10212997
Linear_PLS_metrics	0.13626561	0.3739715	0.10643672
NonLinear_SVMLinear_metrics	0.13814815	0.3615830	0.10456948
Linear_LASSO_metrics	0.15612853	0.2961838	0.12743949
Linear_eNet_metrics	0.15892973	0.2697740	0.12996685

Prediction on Evaluation Data

Having the SVM-Radial model as the best model among all models, use this model to predict the evaluation dataset `StudentEvaluation.xlsx` after removing its empty [PH] column.

Once received the prediction result, we combined it to the evaluation dataset as the [PH] column.

Export Prediction as CSV

Export the above completed evaluation dataset as a csv file.

This is the readable Excel deliverable with our predictions.

Reference

Export the above completed evaluation dataset as a csv file.

This is the readable Excel deliverable with our predictions.

Thank You.