|  |
| --- |
| *Data 624 - Predictive Analytics Project 2* |

Table of Contents

[Team Members 3](#_Toc72702170)

[Overview 3](#_Toc72702171)

[Deliverables 3](#_Toc72702172)

[Load Package 3](#_Toc72702173)

[Load Data 3](#_Toc72702174)

[Exploratory Data Analysis 4](#_Toc72702175)

[Training Data Summary 4](#_Toc72702176)

[Evaluation Data Summary 7](#_Toc72702177)

[Missing Value View 10](#_Toc72702178)

[Numerical Predictor Correlation after Missing Data Imputation 10](#_Toc72702179)

[Data Preprocess 11](#_Toc72702180)

[Model Building 13](#_Toc72702181)

[Linear Regression Models 14](#_Toc72702182)

[PLS Regression 14](#_Toc72702183)

[Ridge Regression 15](#_Toc72702184)

[Lasso 16](#_Toc72702185)

[Elastic Net 17](#_Toc72702186)

[Non-Linear Regression Models 18](#_Toc72702187)

[KNN 18](#_Toc72702188)

[SVM-Linear 19](#_Toc72702189)

[SVM-Radial 19](#_Toc72702190)

[MARS 21](#_Toc72702191)

[Neural Network 22](#_Toc72702192)

[Tree-Based Regression Models 23](#_Toc72702193)

[Random Forest 23](#_Toc72702194)

[Gradient Boosting Machine 24](#_Toc72702195)

[Cubist 25](#_Toc72702196)

[Model Selection 26](#_Toc72702197)

[Prediction on Evaluation Data 26](#_Toc72702198)

[Export Prediction as CSV 26](#_Toc72702199)

[Reference 26](#_Toc72702200)

# 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.

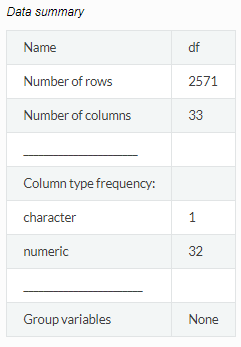
There are 4 rows in the training set which [PH] is missing, as imputing responsible variable is not meaningful in training set, therefore these 4 rows are suggested to be removed.

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



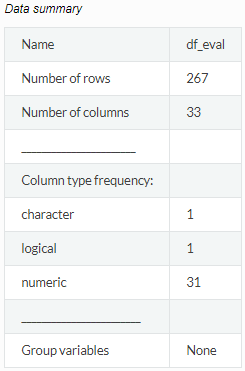
Variable type: character

| 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 | ▇▁▃▇▁ |
| 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



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 | : |

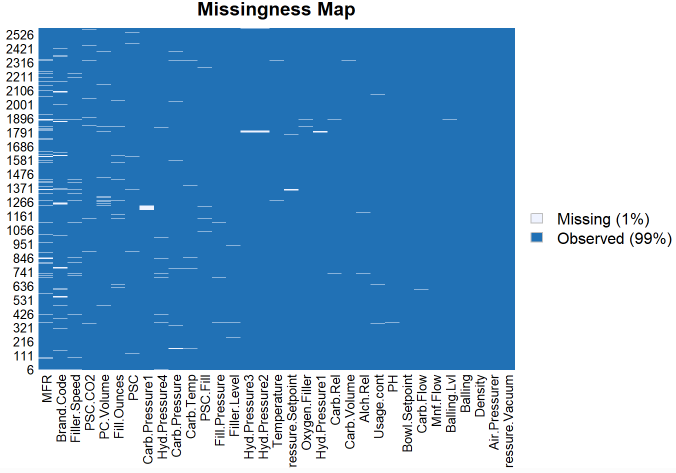
Variable type: numeric

| 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 | ▁▅▇▃▁ |
| 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 | ▁▁▁▁▇ |
| 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.

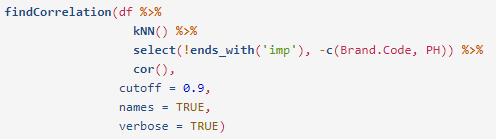
Below is a plot of missing value distribution in the training dataset *`df`*.



# 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.





# 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`:

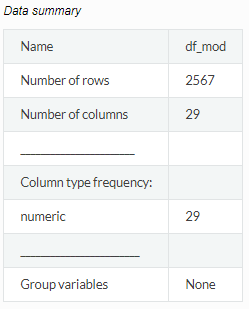
* Remove rows where PH is empty or NA
* Perform train-test-split at ratio 4:1

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*`.



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 | ▁▁▁▂▇ |
| 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:

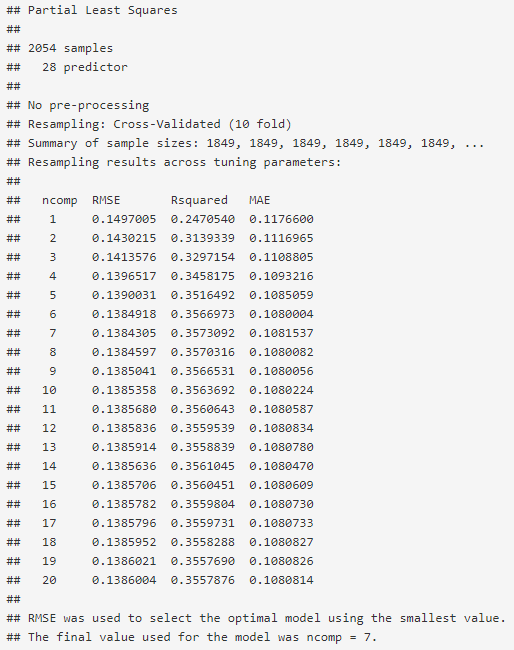
* Linear Regression Models: PLS, Ridge, LASSO and Elastic Net
* Non-linear Regression Models: KNN, SVM-Linear, SVM-Radial, MARS and Neural Network
* Tree-based Regression Models: Random Forest, Gradient Boosting Machine and Cubist

## 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.

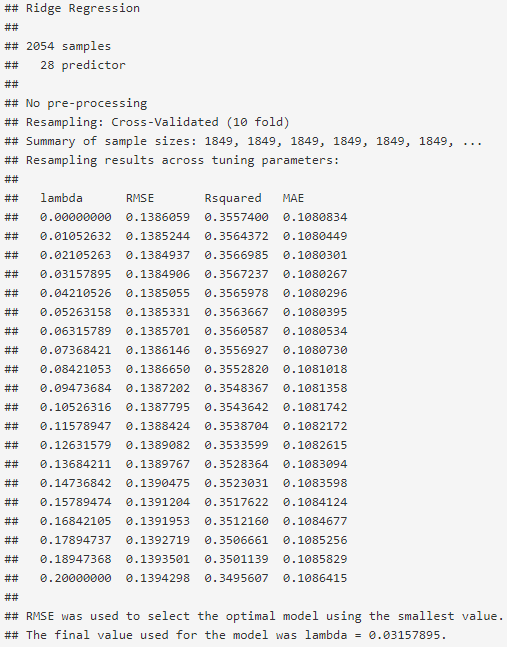




### Ridge Regression

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.

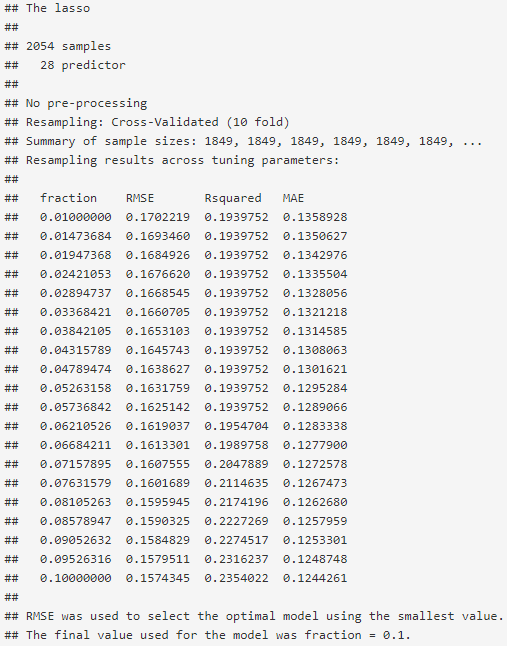




### Lasso

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.

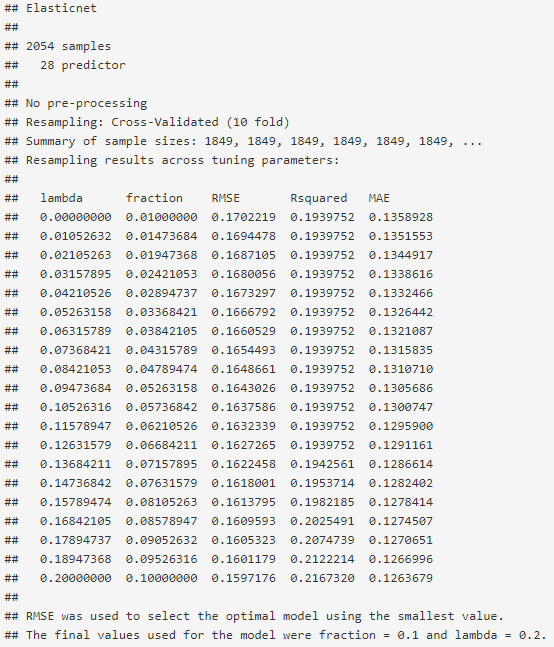




### Elastic Net

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.



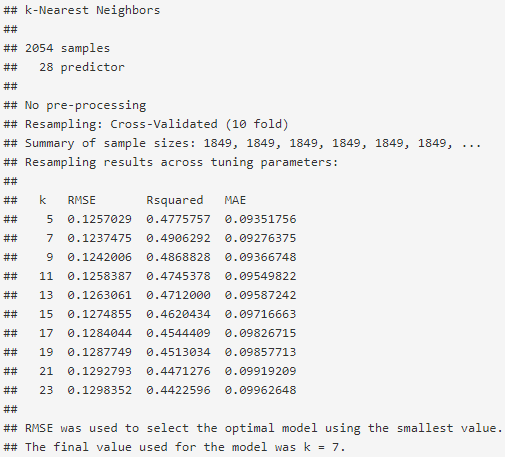


## Non-Linear Regression Models

### 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.

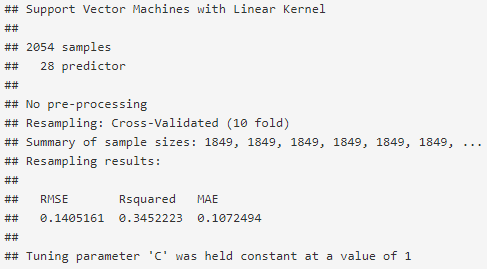


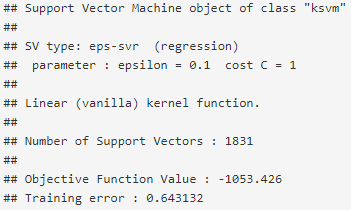


### SVM-Linear

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.



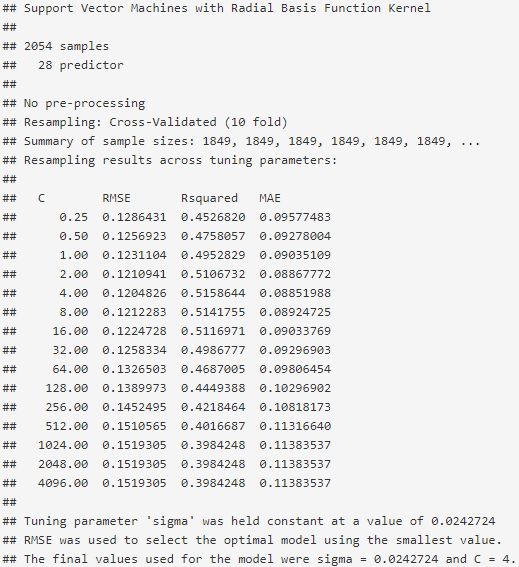


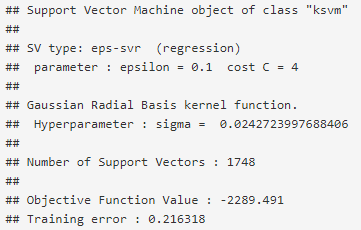


### SVM-Radial

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.



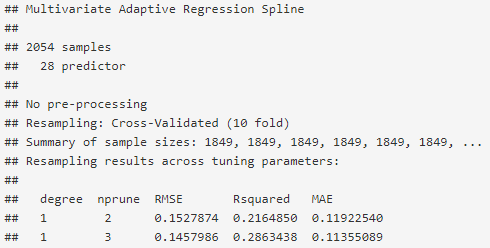




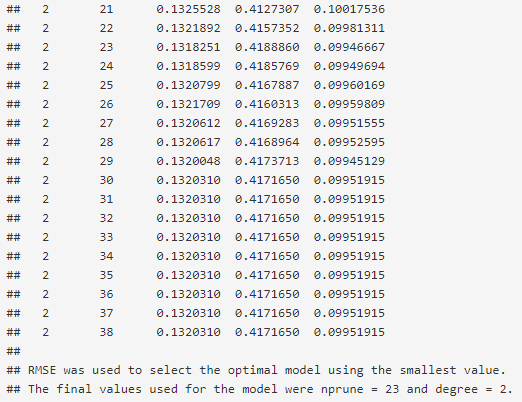
### MARS

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.



… … …

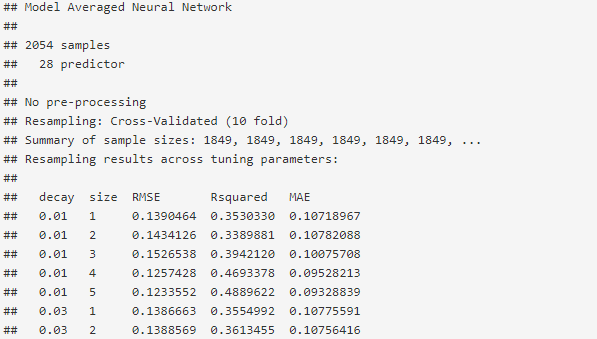




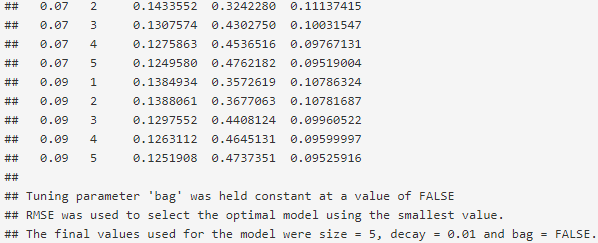
### Neural Network

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.



… … …



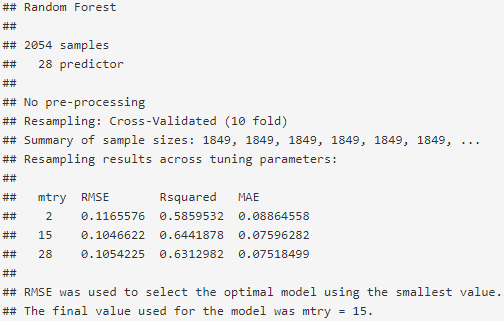


## Tree-Based Regression Models

### 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.

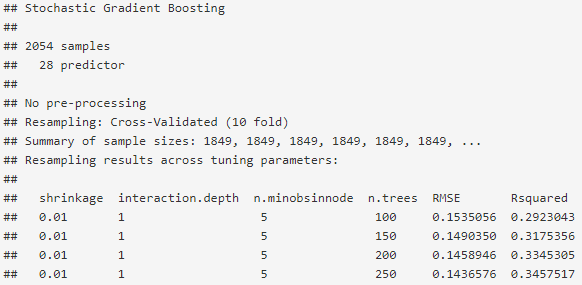




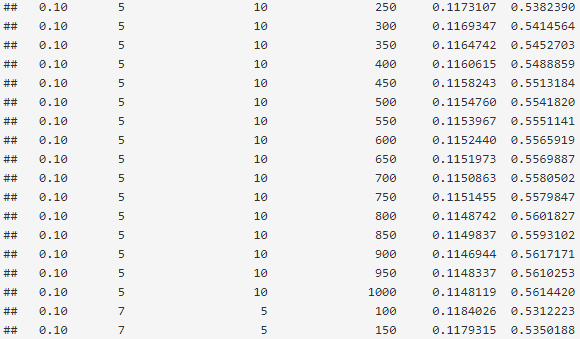
### Gradient Boosting Machine

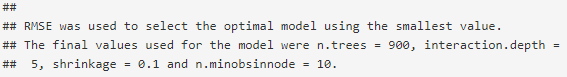
The gradient boosting machine regression model gives:

* 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.



… … …



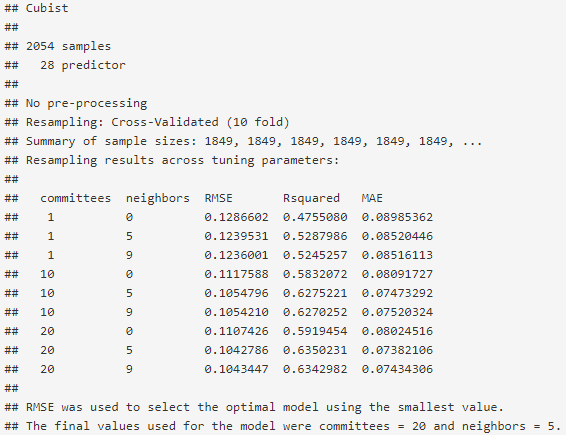




### Cubist

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.





# 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.





# 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.**