Technical report: beverage pH prediction

CLAUDIO, Mauricio

2023-04-03

Data Overview & Transformation

The training dataset consists of 2,571 observations, 32 predictor variables and one predictor variable, *pH*, related to the composition of the beverages manufactured by ABC Beverage. Of the 32 predictor variables, 31 predictors are continuous numerical and one predictor is a categorical beverage brand code denoting individual beverage products. The dataset is summarized in the table below.

No	Variable	Stats / Values	Freqs (% of Valid)	Valid	Missing
1	Brand.Code [character]	1. (Empty string) 2. A 3. B 4. C 5. D	120 (4.7%) 293 (11.4%) 1239 (48.2%) 304 (11.8%) 615 (23.9%)	2571 (100.0%)	0 (0.0%)
2	Carb.Volume [numeric]	Mean (sd) : 5.4 (0.1) min ≤ med ≤ max: 5 ≤ 5.3 ≤ 5.7 IQR (CV) : 0.2 (0)	101 distinct values	2561 (99.6%)	10 (0.4%)
3	Fill.Ounces [numeric]	Mean (sd) : 24 (0.1) min ≤ med ≤ max: 23.6 ≤ 24 ≤ 24.3 IQR (CV) : 0.1 (0)	92 distinct values	2533 (98.5%)	38 (1.5%)
4	PC.Volume [numeric]	Mean (sd) : 0.3 (0.1) min ≤ med ≤ max: 0.1 ≤ 0.3 ≤ 0.5 IQR (CV) : 0.1 (0.2)	454 distinct values	2532 (98.5%)	39 (1.5%)
5	Carb.Pressure [numeric]	Mean (sd) : 68.2 (3.5) min ≤ med ≤ max: 57 ≤ 68.2 ≤ 79.4 IQR (CV) : 5 (0.1)	106 distinct values	2544 (98.9%)	27 (1.1%)
6	Carb.Temp [numeric]	Mean (sd) : 141.1 (4) min ≤ med ≤ max: 128.6 ≤ 140.8 ≤ 154 IQR (CV) : 5.4 (0)	123 distinct values	2545 (99.0%)	26 (1.0%)
7	PSC [numeric]	Mean (sd) : 0.1 (0) min ≤ med ≤ max: 0 ≤ 0.1 ≤ 0.3 IQR (CV) : 0.1 (0.6)	129 distinct values	2538 (98.7%)	33 (1.3%)
8	PSC.Fill [numeric]	Mean (sd) : 0.2 (0.1) min ≤ med ≤ max: 0 ≤ 0.2 ≤ 0.6 IQR (CV) : 0.2 (0.6)	32 distinct values	2548 (99.1%)	23 (0.9%)

No	Variable	Stats / Values	Freqs (% of Valid)	Valid	Missing
9	PSC.CO2 [numeric]	Mean (sd) : 0.1 (0) min ≤ med ≤ max: 0 ≤ 0 ≤ 0.2 IQR (CV) : 0.1 (0.8)	13 distinct values	2532 (98.5%)	39 (1.5%)
10	Mnf.Flow [numeric]	Mean (sd) : 24.6 (119.5) min ≤ med ≤ max: -100.2 ≤ 65.2 ≤ 229.4 IQR (CV) : 240.8 (4.9)	487 distinct values	2569 (99.9%)	2 (0.1%)
11	Carb.Pressure1 [numeric]	Mean (sd) : 122.6 (4.7) min ≤ med ≤ max: 105.6 ≤ 123.2 ≤ 140.2 IQR (CV) : 6.4 (0)	140 distinct values	2539 (98.8%)	32 (1.2%)
12	Fill.Pressure [numeric]	Mean (sd) : 47.9 (3.2) min ≤ med ≤ max: 34.6 ≤ 46.4 ≤ 60.4 IQR (CV) : 4 (0.1)	108 distinct values	2549 (99.1%)	22 (0.9%)
13	Hyd.Pressure1 [numeric]	Mean (sd) : 12.4 (12.4) min ≤ med ≤ max: -0.8 ≤ 11.4 ≤ 58 IQR (CV) : 20.2 (1)	245 distinct values	2560 (99.6%)	11 (0.4%)
14	Hyd.Pressure2 [numeric]	Mean (sd) : 21 (16.4) min ≤ med ≤ max: 0 ≤ 28.6 ≤ 59.4 IQR (CV) : 34.6 (0.8)	207 distinct values	2556 (99.4%)	15 (0.6%)
15	Hyd.Pressure3 [numeric]	Mean (sd) : 20.5 (16) min ≤ med ≤ max: -1.2 ≤ 27.6 ≤ 50 IQR (CV) : 33.4 (0.8)	192 distinct values	2556 (99.4%)	15 (0.6%)
16	Hyd.Pressure4 [integer]	Mean (sd) : 96.3 (13.1) min ≤ med ≤ max: 52 ≤ 96 ≤ 142 IQR (CV) : 16 (0.1)	40 distinct values	2541 (98.8%)	30 (1.2%)
17	Filler.Level [numeric]	Mean (sd) : 109.3 (15.7) min ≤ med ≤ max: 55.8 ≤ 118.4 ≤ 161.2 IQR (CV) : 21.7 (0.1)	288 distinct values	2551 (99.2%)	20 (0.8%)
18	Filler.Speed [integer]	Mean (sd) : 3687.2 (770.8) min ≤ med ≤ max: 998 ≤ 3982 ≤ 4030 IQR (CV) : 110 (0.2)	244 distinct values	2514 (97.8%)	57 (2.2%)
19	Temperature [numeric]	Mean (sd) : 66 (1.4) min ≤ med ≤ max: 63.6 ≤ 65.6 ≤ 76.2 IQR (CV) : 1.2 (0)	56 distinct values	2557 (99.5%)	14 (0.5%)

No	Variable	Stats / Values	Freqs (% of Valid)	Valid	Missing
20	Usage.cont [numeric]	Mean (sd) : 21 (3) min ≤ med ≤ max: 12.1 ≤ 21.8 ≤ 25.9 IQR (CV) : 5.4 (0.1)	481 distinct values	2566 (99.8%)	5 (0.2%)
21	Carb.Flow [integer]	Mean (sd) : 2468.4 (1073.7) min ≤ med ≤ max: 26 ≤ 3028 ≤ 5104 IQR (CV) : 2042 (0.4)	533 distinct values	2569 (99.9%)	2 (0.1%)
22	Density [numeric]	Mean (sd) : 1.2 (0.4) min ≤ med ≤ max: 0.2 ≤ 1 ≤ 1.9 IQR (CV) : 0.7 (0.3)	78 distinct values	2570 (100.0%)	1 (0.0%)
23	MFR [numeric]	Mean (sd) : 704 (73.9) min ≤ med ≤ max: 31.4 ≤ 724 ≤ 868.6 IQR (CV) : 24.7 (0.1)	587 distinct values	2359 (91.8%)	212 (8.2%)
24	Balling [numeric]	Mean (sd) : 2.2 (0.9) min ≤ med ≤ max: -0.2 ≤ 1.6 ≤ 4 IQR (CV) : 1.8 (0.4)	217 distinct values	2570 (100.0%)	1 (0.0%)
25	Pressure.Vacuum [numeric]	Mean (sd) : -5.2 (0.6) min ≤ med ≤ max: -6.6 ≤ -5.4 ≤ -3.6 IQR (CV) : 0.6 (-0.1)	16 distinct values	2571 (100.0%)	0 (0.0%)
26	PH [numeric]	Mean (sd) : 8.5 (0.2) min ≤ med ≤ max: 7.9 ≤ 8.5 ≤ 9.4 IQR (CV) : 0.2 (0)	52 distinct values	2567 (99.8%)	4 (0.2%)
27	Oxygen.Filler [numeric]	Mean (sd) : 0 (0) min ≤ med ≤ max: 0 ≤ 0 ≤ 0.4 IQR (CV) : 0 (1)	338 distinct values	2559 (99.5%)	12 (0.5%)
28	Bowl.Setpoint [integer]	Mean (sd) : 109.3 (15.3) min ≤ med ≤ max: 70 ≤ 120 ≤ 140 IQR (CV) : 20 (0.1)	11 distinct values	2569 (99.9%)	2 (0.1%)
29	Pressure.Setpoint [numeric]	Mean (sd) : 47.6 (2) min ≤ med ≤ max: 44 ≤ 46 ≤ 52 IQR (CV) : 4 (0)	8 distinct values	2559 (99.5%)	12 (0.5%)
30	Air.Pressurer [numeric]	Mean (sd) : 142.8 (1.2) min ≤ med ≤ max: 140.8 ≤ 142.6 ≤ 148.2 IQR (CV) : 0.8 (0)	32 distinct values	2571 (100.0%)	0 (0.0%)

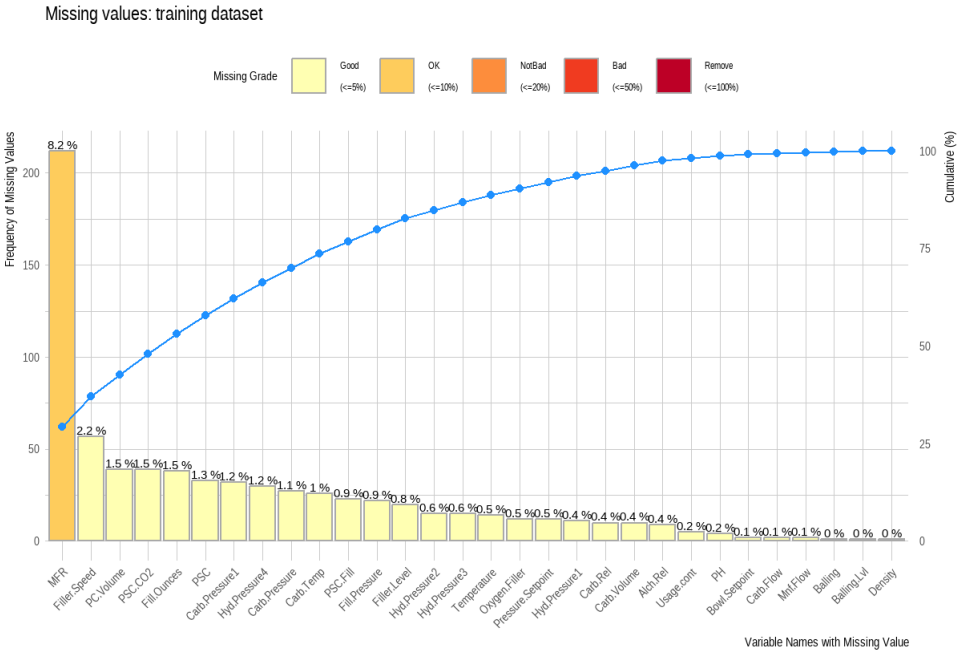
No	Variable	Stats / Values	Freqs (% of Valid)	Valid	Missing
31	Alch.Rel [numeric]	Mean (sd) : 6.9 (0.5) min ≤ med ≤ max: 5.3 ≤ 6.6 ≤ 8.6 IQR (CV) : 0.7 (0.1)	53 distinct values	2562 (99.6%)	9 (0.4%)
32	Carb.Rel [numeric]	Mean (sd) : 5.4 (0.1) min ≤ med ≤ max: 5 ≤ 5.4 ≤ 6.1 IQR (CV) : 0.2 (0)	42 distinct values	2561 (99.6%)	10 (0.4%)
33	Balling.Lvl [numeric]	Mean (sd) : 2.1 (0.9) min ≤ med ≤ max: 0 ≤ 1.5 ≤ 3.7 IQR (CV) : 1.8 (0.4)	82 distinct values	2570 (100.0%)	1 (0.0%)

Generated by summarytools (https://github.com/dcomtois/summarytools) 1.0.1 (R (https://www.r-project.org/) version 4.1.0)
2023-04-04

Missing Values

Missing values are present in 30 variables, including the target variable *pH*, ranging in prevalence from 0.001% to 8.2% of observations. In addition, there are 120 observations for which the grouping variable *Brand.Code* is empty, denoting no beverage product. Absent explicit guidance from management on how to treat these observations with a missing brand code, we do not consider these 120 observations further.

Therefore, to correct for missing values we perform the following transformations on the training dataset:
1. Drop 120 observations for which the beverage **Brand.Code** variable is empty.
2. Drop four (4) observations for which the target variable **PH** is empty.
3. Impute missing values with the Bagging Trees method.



variables	types	missing_count	missing_percent
MFR	numeric	212	8.2
Filler.Speed	integer	57	2.2
PC.Volume	numeric	39	1.5
PSC.CO2	numeric	39	1.5

variables	types	missing_count	missing_percent
Fill.Ounces	numeric	38	1.5
PSC	numeric	33	1.3
Carb.Pressure1	numeric	32	1.2
Hyd.Pressure4	integer	30	1.2
Carb.Pressure	numeric	27	1.1
Carb.Temp	numeric	26	1.0
PSC.Fill	numeric	23	0.9
Fill.Pressure	numeric	22	0.9
Filler.Level	numeric	20	0.8
Hyd.Pressure2	numeric	15	0.6
Hyd.Pressure3	numeric	15	0.6
Temperature	numeric	14	0.5
Oxygen.Filler	numeric	12	0.5
Pressure.Setpoint	numeric	12	0.5
Hyd.Pressure1	numeric	11	0.4
Carb.Volume	numeric	10	0.4
Carb.Rel	numeric	10	0.4
Alch.Rel	numeric	9	0.4
Usage.cont	numeric	5	0.2
PH	numeric	4	0.2
Mnf.Flow	numeric	2	0.1
Carb.Flow	integer	2	0.1
Bowl.Setpoint	integer	2	0.1
Density	numeric	1	0.0
Balling	numeric	1	0.0
Balling.Lvl	numeric	1	0.0

Variance

A check of variable variance reveals no predictor variable with zero or near zero variance.

	freqRatio	percentUnique	zeroVar	nzv
Brand.Code	2.0	0.2	FALSE	FALSE
Carb.Volume	1.1	4.4	FALSE	FALSE
Fill.Ounces	1.2	4.9	FALSE	FALSE
PC.Volume	1.1	19.4	FALSE	FALSE
Carb.Pressure	1.0	4.9	FALSE	FALSE
Carb.Temp	1.0	5.4	FALSE	FALSE
PSC	1.2	6.5	FALSE	FALSE
PSC.Fill	1.1	2.1	FALSE	FALSE
PSC.CO2	1.1	2.0	FALSE	FALSE
Mnf.Flow	1.1	19.6	FALSE	FALSE
Carb.Pressure1	1.0	6.3	FALSE	FALSE
Fill.Pressure	1.8	4.9	FALSE	FALSE
Hyd.Pressure1	30.7	10.1	FALSE	FALSE
Hyd.Pressure2	7.6	8.4	FALSE	FALSE
Hyd.Pressure3	12.1	7.8	FALSE	FALSE
Hyd.Pressure4	1.0	2.0	FALSE	FALSE
Filler.Level	1.1	11.7	FALSE	FALSE
Filler.Speed	1.1	10.9	FALSE	FALSE
Temperature	1.1	2.7	FALSE	FALSE
Usage.cont	1.1	19.6	FALSE	FALSE
Carb.Flow	1.4	21.5	FALSE	FALSE
Density	1.1	3.1	FALSE	FALSE
MFR	1.5	23.9	FALSE	FALSE
Balling	1.2	8.4	FALSE	FALSE
Pressure.Vacuum	1.4	0.7	FALSE	FALSE
PH	1.1	2.1	FALSE	FALSE
Oxygen.Filler	1.3	13.8	FALSE	FALSE
Bowl.Setpoint	3.0	0.5	FALSE	FALSE

	freqRatio	percentUnique	zeroVar	nzv
Pressure.Setpoint	1.3	0.5	FALSE	FALSE
Air.Pressurer	1.1	1.3	FALSE	FALSE
Alch.Rel	1.2	2.1	FALSE	FALSE
Carb.Rel	1.0	1.9	FALSE	FALSE
Balling.Lvl	1.3	3.3	FALSE	FALSE

Duplicates

A check for duplicate observations reveals that the dataset is free of duplicates.

```
### Check for duplicate observations ###
sum(duplicated(StudentData))
```

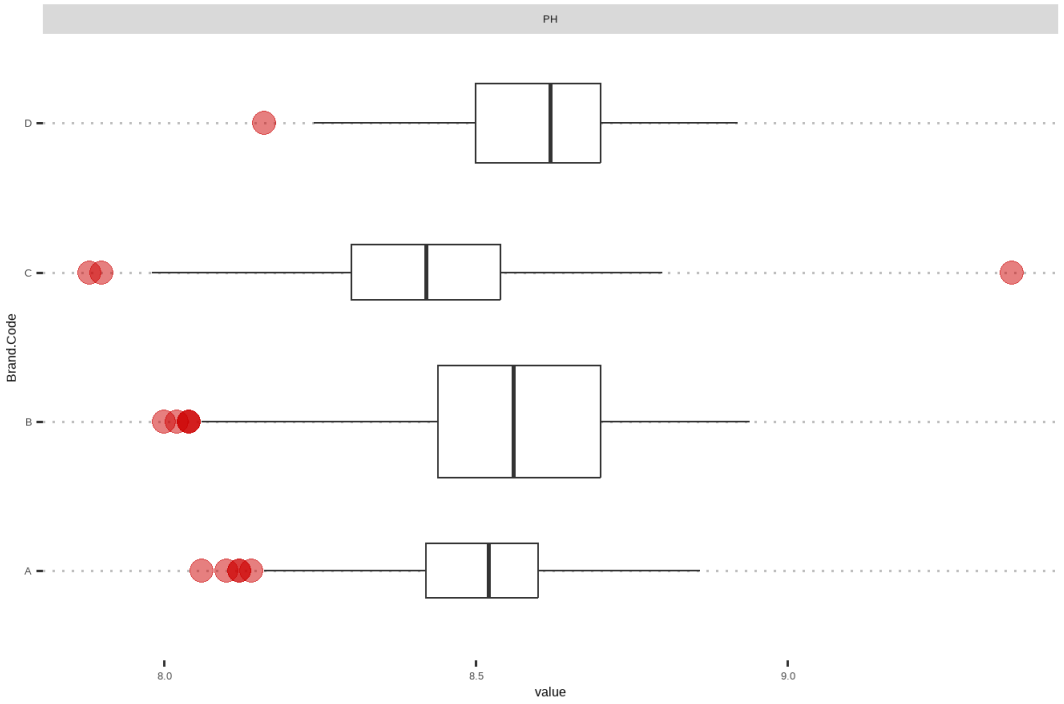
```
## [1] 0
```

Outliers

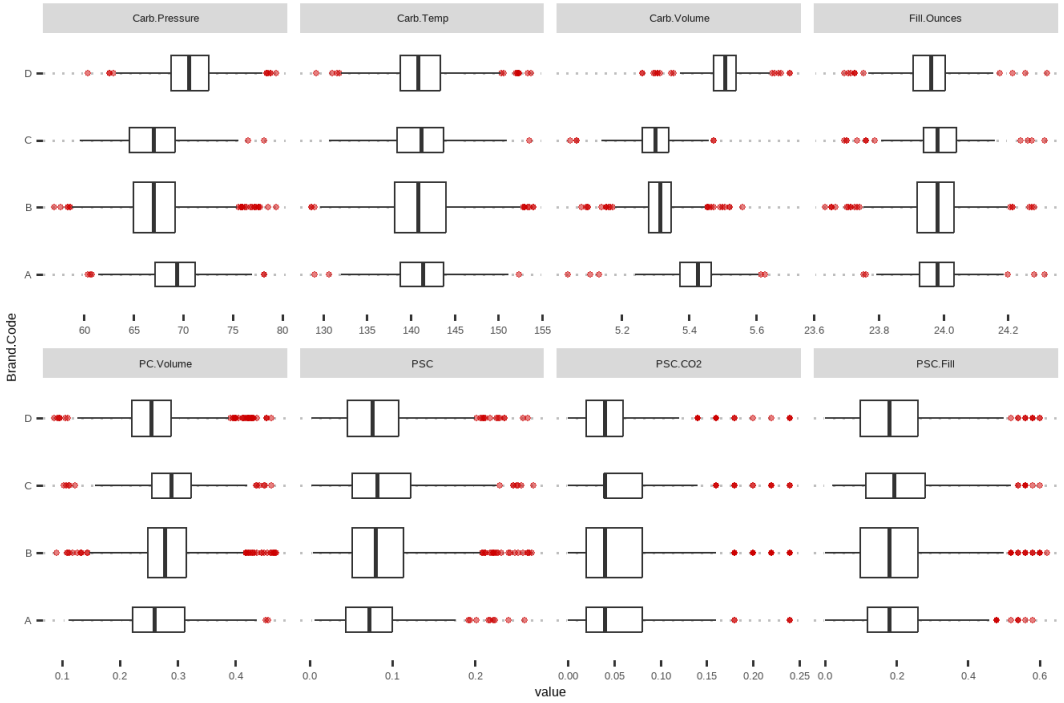
Boxplots of target and predictor variables grouped by beverage brand reveal outliers (shown in red) across predictors.

Nevertheless, because the task is at hand is predictive rather than inferential modeling, we leave outliers intact lest we risk model overfitting and diminished predictive performance later.

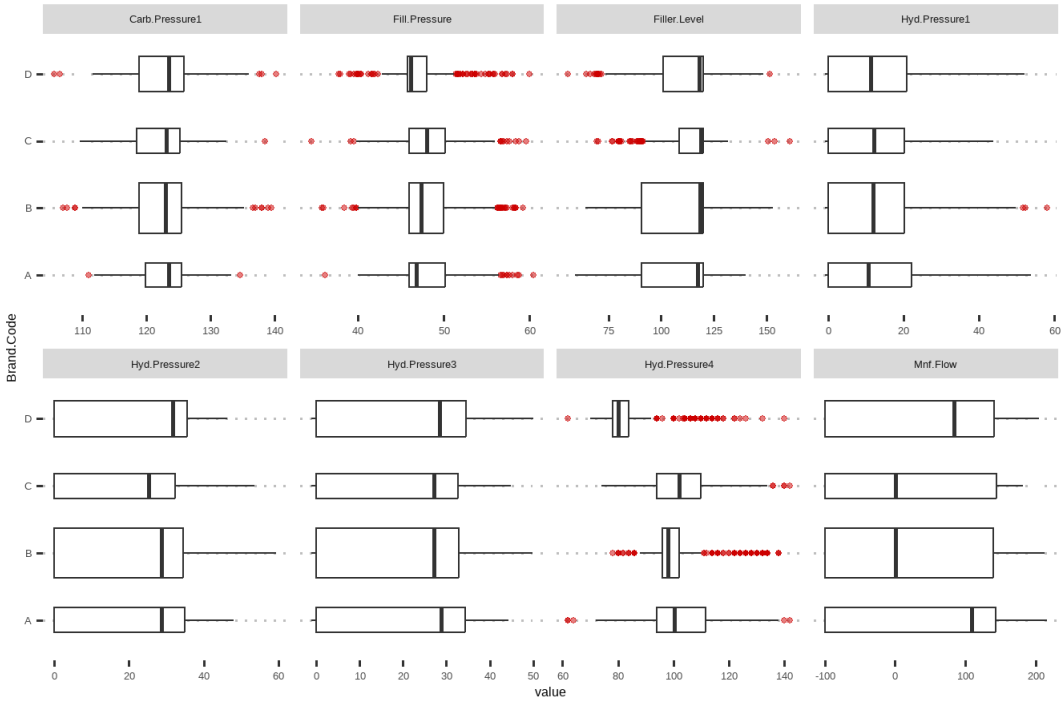
ABC Beverage: target variable by brand



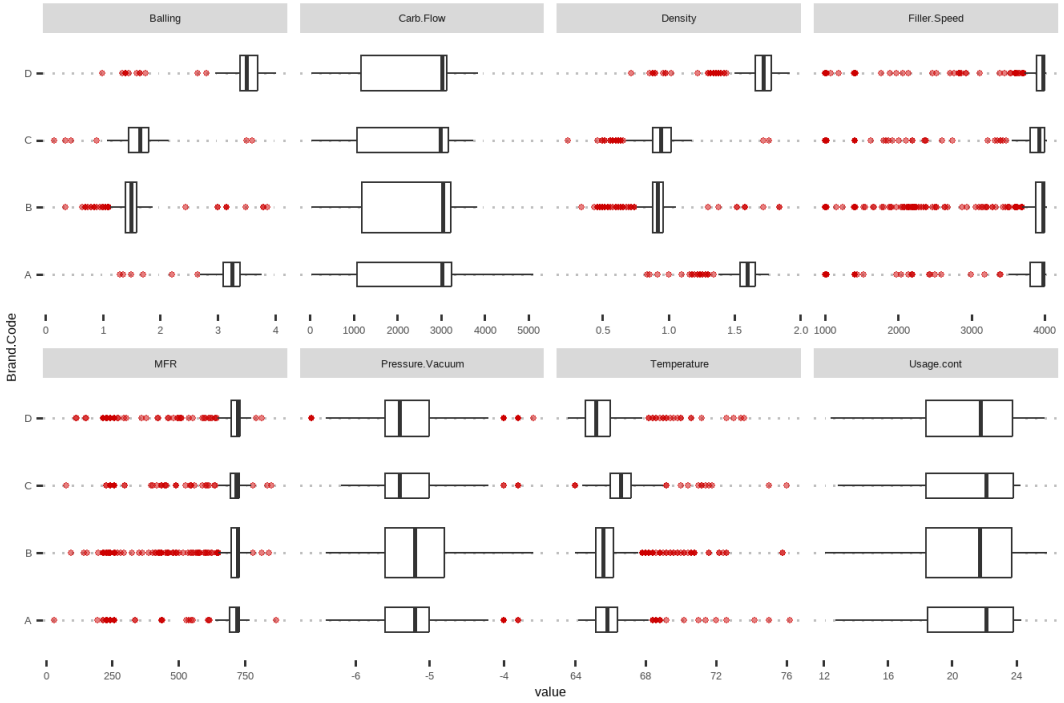
ABC Beverage: predictor variable by brand



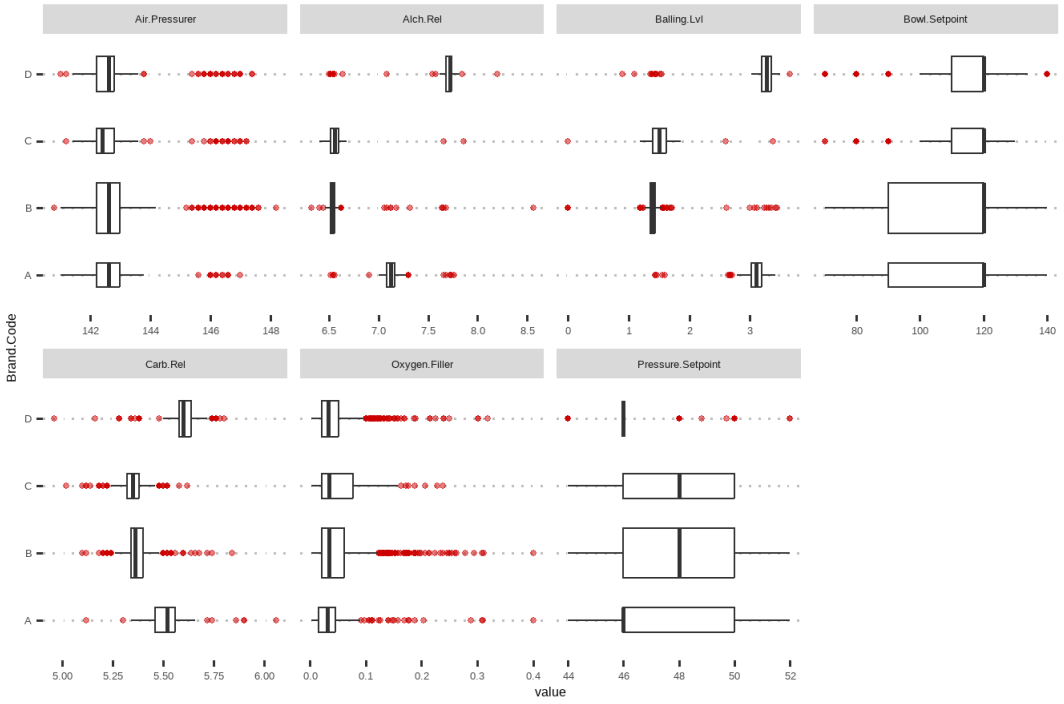
ABC Beverage: predictor variable by brand



ABC Beverage: predictor variable by brand



ABC Beverage: predictor variable by brand



Variable	n	% n	Outlier mean	Variable mean w/ outliers	Variable mean w/o outliers
Filler.Speed	363	14.8	2124.8	3677.8	3948.3
MFR	300	12.3	362.7	674.5	718.0
Air.Pressurer	220	9.0	146.4	142.8	142.5
Oxygen.Filler	178	7.3	0.2	0.0	0.0
Pressure.Vacuum	116	4.7	-4.0	-5.2	-5.3
Hyd.Pressure4	111	4.5	127.8	96.4	95.0
Temperature	104	4.3	70.3	65.9	65.7
PC.Volume	79	3.2	0.3	0.3	0.3
Fill.Pressure	72	2.9	51.7	47.9	47.8
PSC.CO2	69	2.8	0.2	0.1	0.1
PSC	54	2.2	0.2	0.1	0.1
PSC.Fill	52	2.1	0.6	0.2	0.2
Fill.Ounces	45	1.8	23.9	24.0	24.0
Carb.Temp	34	1.4	147.2	141.2	141.1
Carb.Pressure	21	0.9	74.4	68.3	68.3
PH	18	0.7	8.1	8.5	8.6
Carb.Pressure1	17	0.7	127.2	122.5	122.5
Hyd.Pressure1	6	0.2	53.3	12.5	12.4
Filler.Level	4	0.2	84.2	109.1	109.1
Carb.Rel	4	0.2	5.7	5.4	5.4
Carb.Volume	1	0.0	5.0	5.4	5.4
Mnf.Flow	0	0.0	NaN	24.5	24.5
Hyd.Pressure2	0	0.0	NaN	21.2	21.2
Hyd.Pressure3	0	0.0	NaN	20.5	20.5
Usage.cont	0	0.0	NaN	21.0	21.0
Carb.Flow	0	0.0	NaN	2468.5	2468.5
Density	0	0.0	NaN	1.2	1.2
Balling	0	0.0	NaN	2.2	2.2
Bowl.Setpoint	0	0.0	NaN	109.2	109.2
Pressure.Setpoint	0	0.0	NaN	47.6	47.6
Alch.Rel	0	0.0	NaN	6.9	6.9
Balling.Lvl	0	0.0	NaN	2.1	2.1

Collinearity

The dataset has eight predictor variable pairs with correlation ranging from moderate to very high, pointing some collinearity in the data.

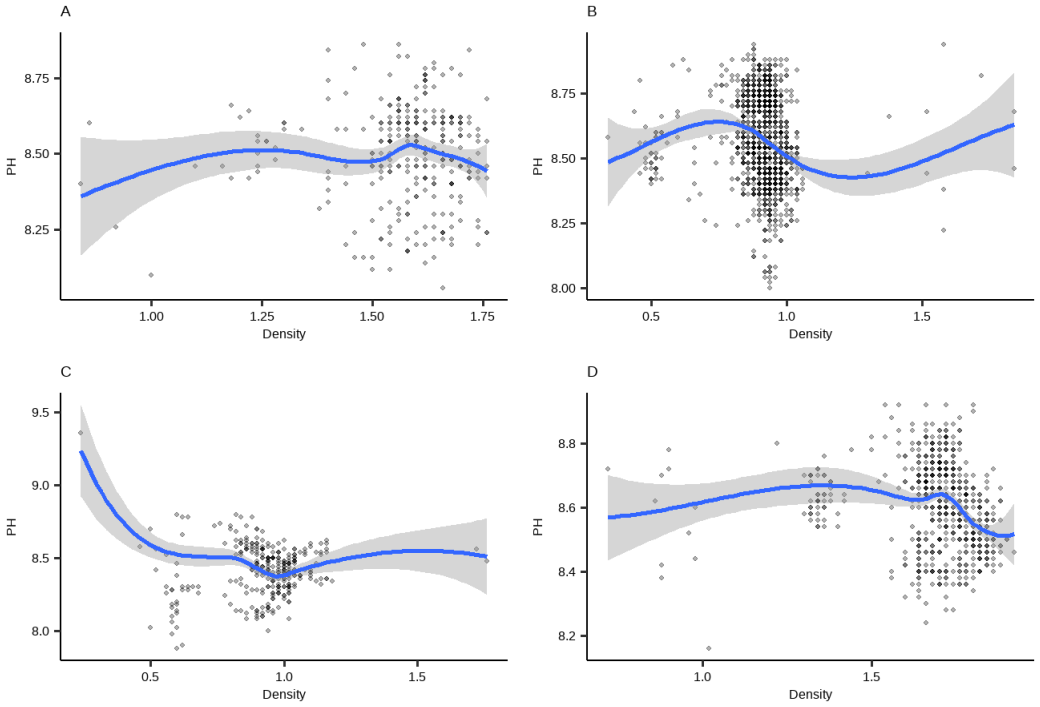
We do not effect any trasformations to decorrelate predictor variables due to the uncertain, premature nature of this task at this point and to the relatively high number of predictors involved.

Predictor 1	Predictor 2	Correlation	p-value
Carb.Pressure	Carb.Temp	0.91	0
Filler.Level	Bowl.Setpoint	0.87	0
Filler.Speed	MFR	0.84	0
Hyd.Pressure2	Hyd.Pressure3	0.78	0
Hyd.Pressure1	Hyd.Pressure2	0.68	0
Density	Balling	0.68	0
Mnf.Flow	Hyd.Pressure3	0.66	0
Alch.Rel	Balling.Lvl	-0.78	0

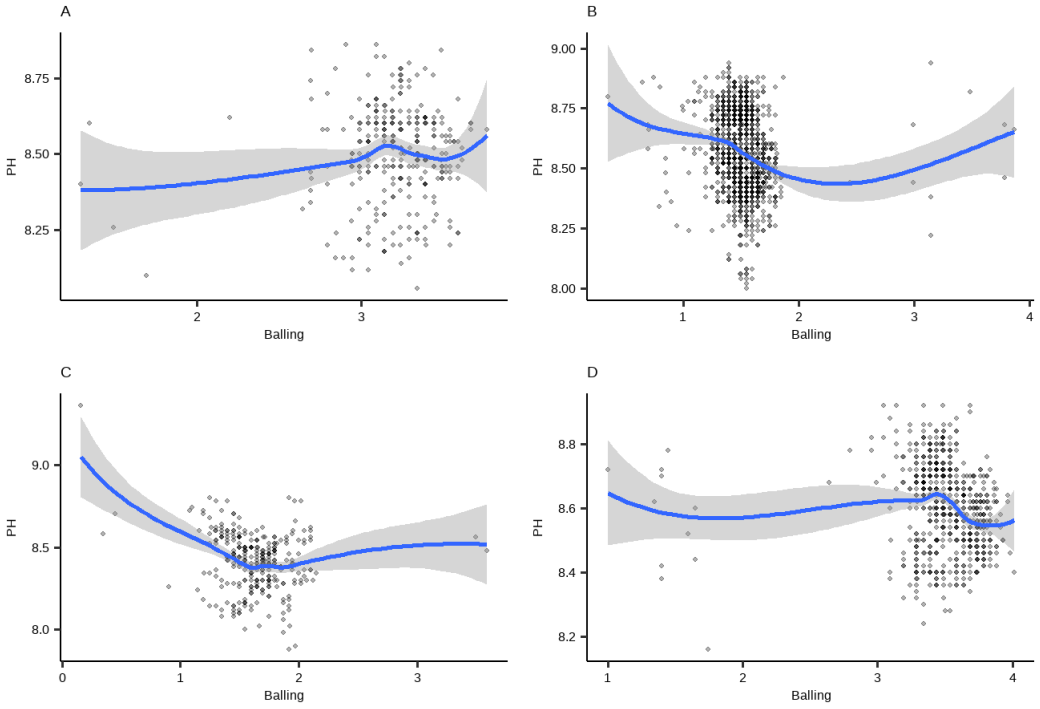
Non-linearity

The dataset shows nonlinear relationships in varying degrees between the predictor and target variable for most predictors. The graphs below for five selected predictor vs. target variable *PH* pairs illustrate this phenomenon.

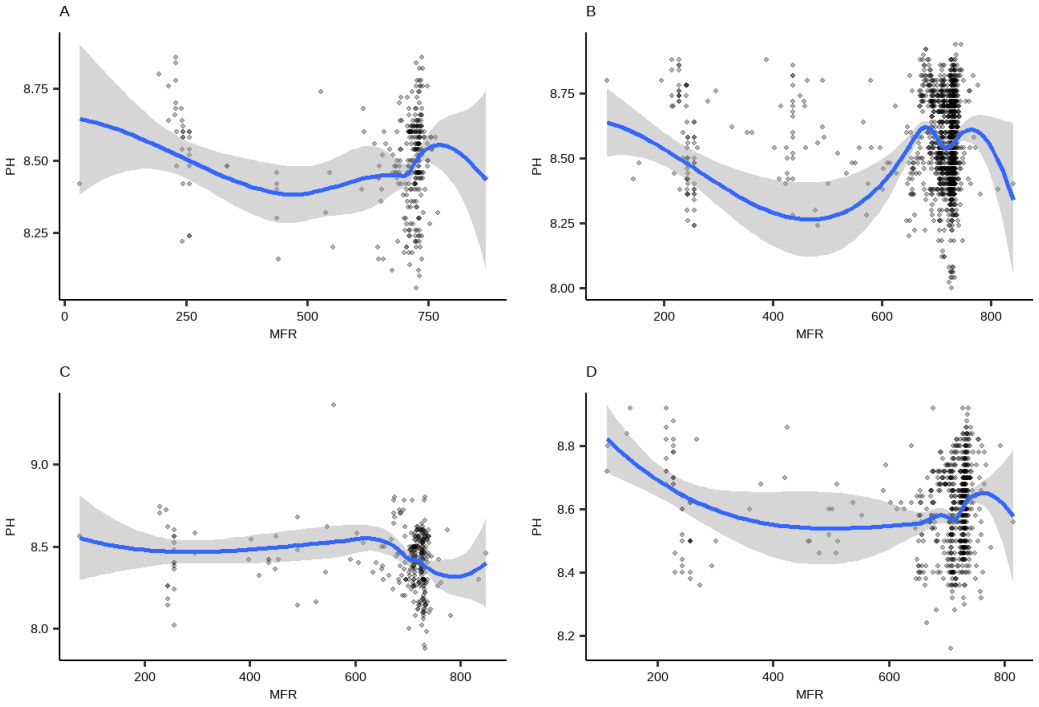
Training data: Density vs. PH



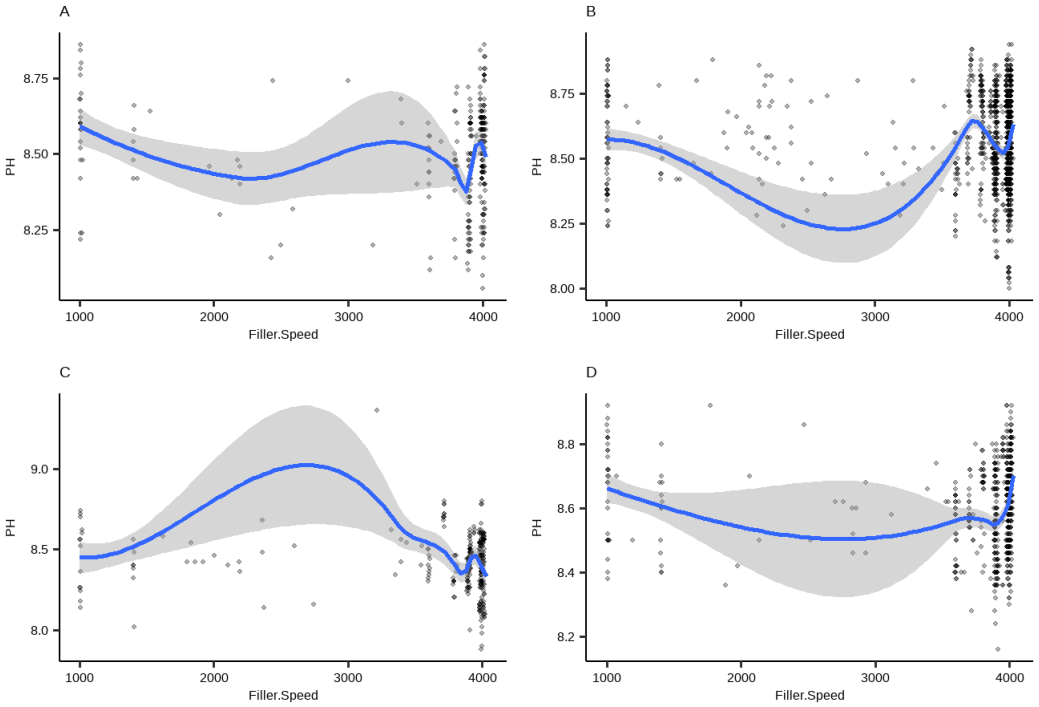
Training data: Balling vs. PH



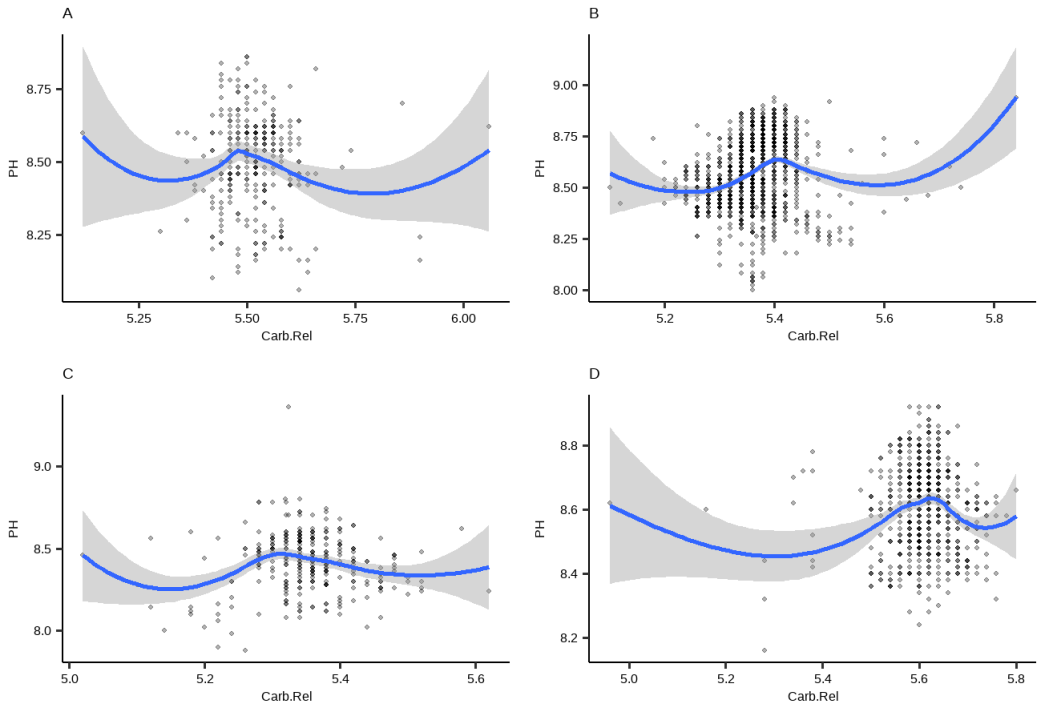
Training data: MFR vs. PH



Training data: Filler.Speed vs. PH

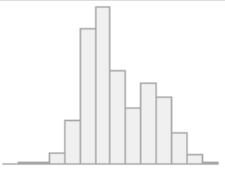
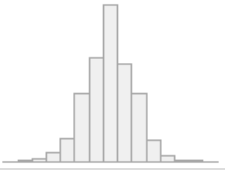
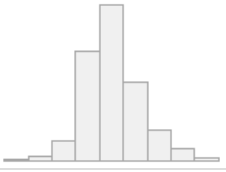
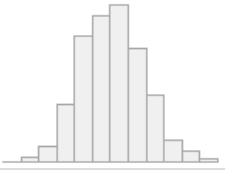


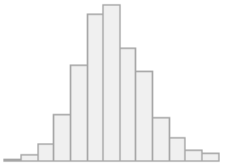
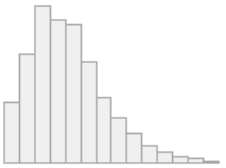
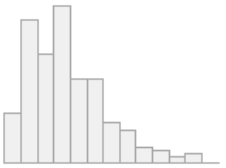
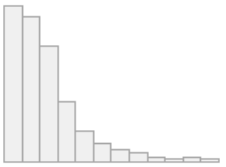
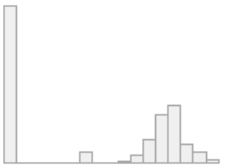
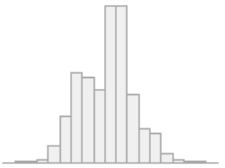
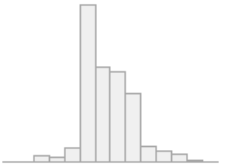
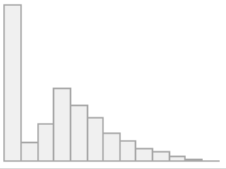
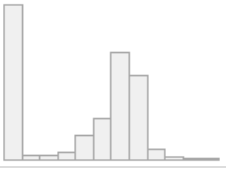
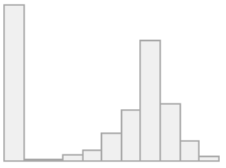
Training data: Carb.Rel vs. PH

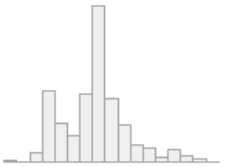
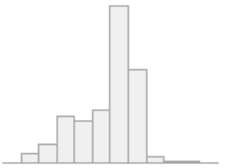
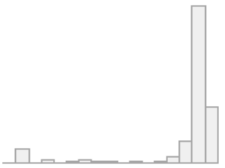
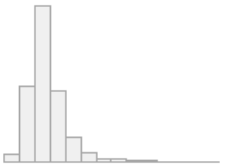
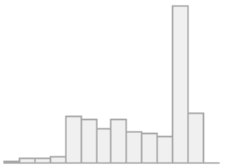
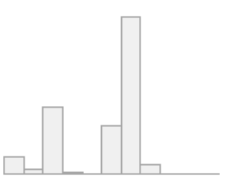
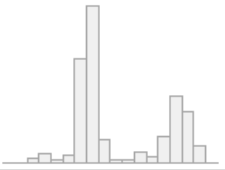
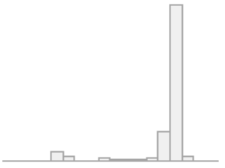
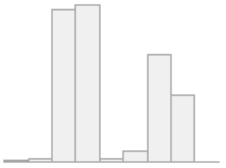
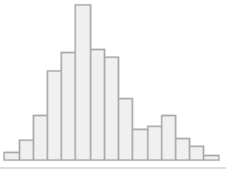


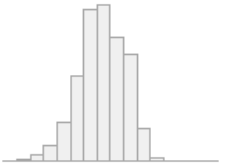
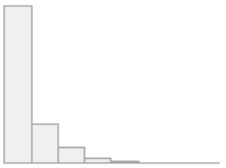
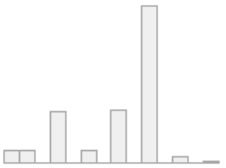
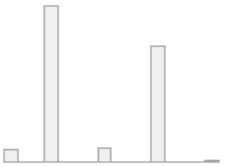
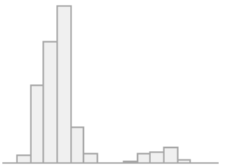
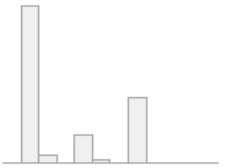
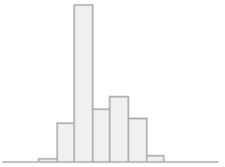
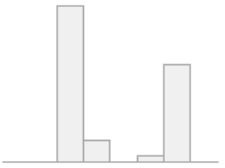
In summary, the training dataset is characterized by 1.) a substantial number of outliers and by 2.) collinearity across several predictor variables, and 3.) non-linearity between the predictors and the target variable.

The final training data after the foregoing transformations, now ready for next step of model fitting, is summarized below.

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	Brand.Code [character]	1. A 2. B 3. C 4. D	293 (12.0%) 1235 (50.5%) 304 (12.4%) 615 (25.1%)		2447 (100.0%)	0 (0.0%)
2	Carb.Volume [numeric]	Mean (sd) : 5.4 (0.1) min ≤ med ≤ max: 5 ≤ 5.3 ≤ 5.7 IQR (CV) : 0.2 (0)	107 distinct values		2447 (100.0%)	0 (0.0%)
3	Fill.Ounces [numeric]	Mean (sd) : 24 (0.1) min ≤ med ≤ max: 23.6 ≤ 24 ≤ 24.3 IQR (CV) : 0.1 (0)	121 distinct values		2447 (100.0%)	0 (0.0%)
4	PC.Volume [numeric]	Mean (sd) : 0.3 (0.1) min ≤ med ≤ max: 0.1 ≤ 0.3 ≤ 0.5 IQR (CV) : 0.1 (0.2)	475 distinct values		2447 (100.0%)	0 (0.0%)
5	Carb.Pressure [numeric]	Mean (sd) : 68.3 (3.5) min ≤ med ≤ max: 57 ≤ 68.2 ≤ 79.4 IQR (CV) : 4.8 (0.1)	120 distinct values		2447 (100.0%)	0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
6	Carb.Temp [numeric]	Mean (sd) : 141.2 (4.1) min ≤ med ≤ max: 128.6 ≤ 140.8 ≤ 154 IQR (CV) : 5.4 (0)	131 distinct values		2447 (100.0%)	0 (0.0%)
7	PSC [numeric]	Mean (sd) : 0.1 (0) min ≤ med ≤ max: 0 ≤ 0.1 ≤ 0.3 IQR (CV) : 0.1 (0.6)	158 distinct values		2447 (100.0%)	0 (0.0%)
8	PSC.Fill [numeric]	Mean (sd) : 0.2 (0.1) min ≤ med ≤ max: 0 ≤ 0.2 ≤ 0.6 IQR (CV) : 0.2 (0.6)	51 distinct values		2447 (100.0%)	0 (0.0%)
9	PSC.CO2 [numeric]	Mean (sd) : 0.1 (0) min ≤ med ≤ max: 0 ≤ 0 ≤ 0.2 IQR (CV) : 0.1 (0.7)	48 distinct values		2447 (100.0%)	0 (0.0%)
10	Mnf.Flow [numeric]	Mean (sd) : 24.5 (119.7) min ≤ med ≤ max: -100.2 ≤ 70.2 ≤ 216.2 IQR (CV) : 241.2 (4.9)	479 distinct values		2447 (100.0%)	0 (0.0%)
11	Carb.Pressure1 [numeric]	Mean (sd) : 122.5 (4.7) min ≤ med ≤ max: 105.6 ≤ 123.2 ≤ 140.2 IQR (CV) : 6.6 (0)	154 distinct values		2447 (100.0%)	0 (0.0%)
12	Fill.Pressure [numeric]	Mean (sd) : 47.9 (3.1) min ≤ med ≤ max: 34.6 ≤ 46.4 ≤ 60.4 IQR (CV) : 4 (0.1)	121 distinct values		2447 (100.0%)	0 (0.0%)
13	Hyd.Pressure1 [numeric]	Mean (sd) : 12.5 (12.4) min ≤ med ≤ max: -0.8 ≤ 11.6 ≤ 58 IQR (CV) : 20.4 (1)	247 distinct values		2447 (100.0%)	0 (0.0%)
14	Hyd.Pressure2 [numeric]	Mean (sd) : 21.2 (16.4) min ≤ med ≤ max: 0 ≤ 28.8 ≤ 59.4 IQR (CV) : 34.8 (0.8)	206 distinct values		2447 (100.0%)	0 (0.0%)
15	Hyd.Pressure3 [numeric]	Mean (sd) : 20.5 (15.9) min ≤ med ≤ max: -1.2 ≤ 27.8 ≤ 50 IQR (CV) : 33.2 (0.8)	191 distinct values		2447 (100.0%)	0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
16	Hyd.Pressure4 [numeric]	Mean (sd) : 96.4 (12.9) min ≤ med ≤ max: 62 ≤ 96 ≤ 142 IQR (CV) : 14 (0.1)	49 distinct values		2447 (100.0%)	0 (0.0%)
17	Filler.Level [numeric]	Mean (sd) : 109.1 (15.7) min ≤ med ≤ max: 55.8 ≤ 118.4 ≤ 161.2 IQR (CV) : 22.7 (0.1)	286 distinct values		2447 (100.0%)	0 (0.0%)
18	Filler.Speed [numeric]	Mean (sd) : 3677.8 (771.9) min ≤ med ≤ max: 998 ≤ 3982 ≤ 4030 IQR (CV) : 126 (0.2)	266 distinct values		2447 (100.0%)	0 (0.0%)
19	Temperature [numeric]	Mean (sd) : 65.9 (1.3) min ≤ med ≤ max: 63.6 ≤ 65.6 ≤ 76.2 IQR (CV) : 1.2 (0)	65 distinct values		2447 (100.0%)	0 (0.0%)
20	Usage.cont [numeric]	Mean (sd) : 21 (3) min ≤ med ≤ max: 12.1 ≤ 21.8 ≤ 25.9 IQR (CV) : 5.4 (0.1)	480 distinct values		2447 (100.0%)	0 (0.0%)
21	Carb.Flow [numeric]	Mean (sd) : 2468.5 (1070.1) min ≤ med ≤ max: 26 ≤ 3030 ≤ 5104 IQR (CV) : 2024 (0.4)	525 distinct values		2447 (100.0%)	0 (0.0%)
22	Density [numeric]	Mean (sd) : 1.2 (0.4) min ≤ med ≤ max: 0.2 ≤ 1 ≤ 1.9 IQR (CV) : 0.7 (0.3)	77 distinct values		2447 (100.0%)	0 (0.0%)
23	MFR [numeric]	Mean (sd) : 674.5 (131.4) min ≤ med ≤ max: 31.4 ≤ 722.2 ≤ 868.6 IQR (CV) : 33.4 (0.2)	586 distinct values		2447 (100.0%)	0 (0.0%)
24	Balling [numeric]	Mean (sd) : 2.2 (0.9) min ≤ med ≤ max: 0.2 ≤ 1.6 ≤ 4 IQR (CV) : 1.8 (0.4)	205 distinct values		2447 (100.0%)	0 (0.0%)
25	Pressure.Vacuum [numeric]	Mean (sd) : -5.2 (0.6) min ≤ med ≤ max: -6.6 ≤ -5.4 ≤ -3.6 IQR (CV) : 0.6 (-0.1)	16 distinct values		2447 (100.0%)	0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
26	PH [numeric]	Mean (sd) : 8.5 (0.2) min ≤ med ≤ max: 7.9 ≤ 8.5 ≤ 9.4 IQR (CV) : 0.2 (0)	52 distinct values		2447 (100.0%)	0 (0.0%)
27	Oxygen.Filler [numeric]	Mean (sd) : 0 (0) min ≤ med ≤ max: 0 ≤ 0 ≤ 0.4 IQR (CV) : 0 (1)	338 distinct values		2447 (100.0%)	0 (0.0%)
28	Bowl.Setpoint [numeric]	Mean (sd) : 109.2 (15.3) min ≤ med ≤ max: 70 ≤ 120 ≤ 140 IQR (CV) : 20 (0.1)	12 distinct values		2447 (100.0%)	0 (0.0%)
29	Pressure.Setpoint [numeric]	Mean (sd) : 47.6 (2) min ≤ med ≤ max: 44 ≤ 46 ≤ 52 IQR (CV) : 4 (0)	13 distinct values		2447 (100.0%)	0 (0.0%)
30	Air.Pressurer [numeric]	Mean (sd) : 142.8 (1.2) min ≤ med ≤ max: 140.8 ≤ 142.6 ≤ 148.2 IQR (CV) : 0.8 (0)	32 distinct values		2447 (100.0%)	0 (0.0%)
31	Alch.Rel [numeric]	Mean (sd) : 6.9 (0.5) min ≤ med ≤ max: 6.3 ≤ 6.6 ≤ 8.6 IQR (CV) : 1.1 (0.1)	51 distinct values		2447 (100.0%)	0 (0.0%)
32	Carb.Rel [numeric]	Mean (sd) : 5.4 (0.1) min ≤ med ≤ max: 5 ≤ 5.4 ≤ 6.1 IQR (CV) : 0.2 (0)	46 distinct values		2447 (100.0%)	0 (0.0%)
33	Balling.Lvl [numeric]	Mean (sd) : 2.1 (0.9) min ≤ med ≤ max: 0 ≤ 1.5 ≤ 3.7 IQR (CV) : 1.8 (0.4)	80 distinct values		2447 (100.0%)	0 (0.0%)

Generated by summarytools (https://github.com/dcomtois/summarytools) 1.0.1 (R (https://www.r-project.org/) version 4.1.0)
2023-04-04

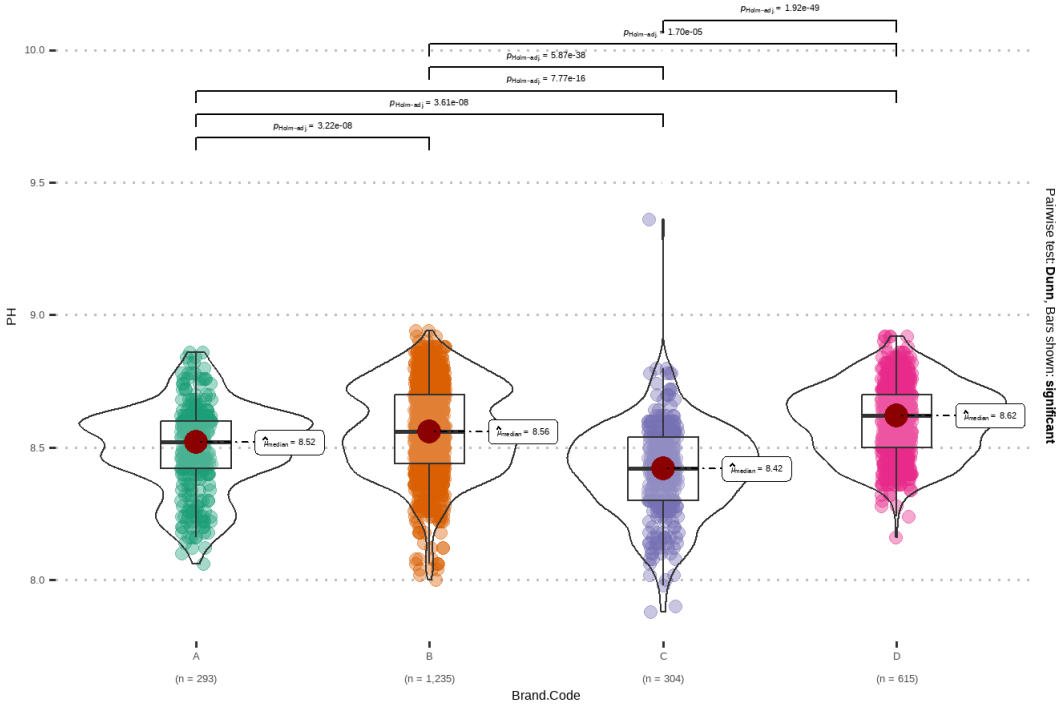
Model Fitting & Selection

Target variable **PH** displays statistically significant difference across the four A, B, C, and D beverage brands. This is demonstrated in the following graph showing pair-wise comparisons among the brands A, B, C and D for the **PH** target variable. First, a p-value of 2.01e-55 for the Kruskal-Wallis test (upper left in the plot) provides little evidence for the test null hypothesis that no one group median differs from the other, pointing to at least one group median differing from the others. Second the 0.11 value of the ϵ^2 statistic (upper center of the plot) points to the *moderate* impact of brand on *PH*. Finally, the p-values for the Dunn Test comparing each group against all other groups pair-wisely

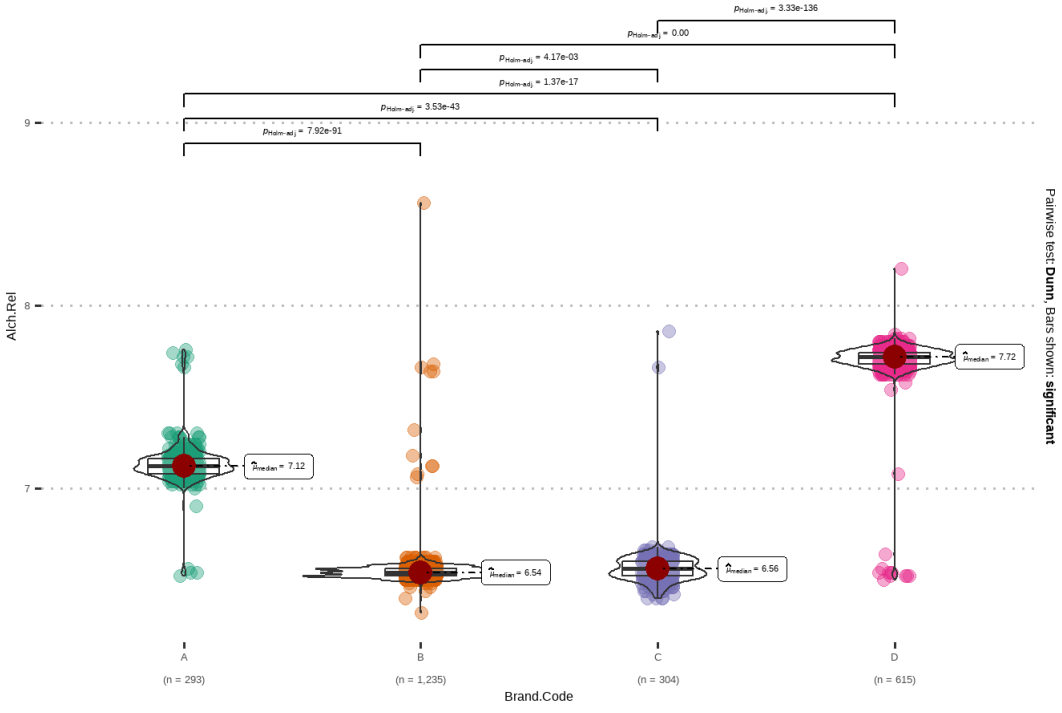
are below the 0.05 threshold, pointing to statitistcally significant group medians across all groups. Most predictor variables also show statistically significant difference across brand groups, often to a greater degree than the target variable. Three selected predictor variable plots illustrate this point. Note the very high ϵ^2 values.

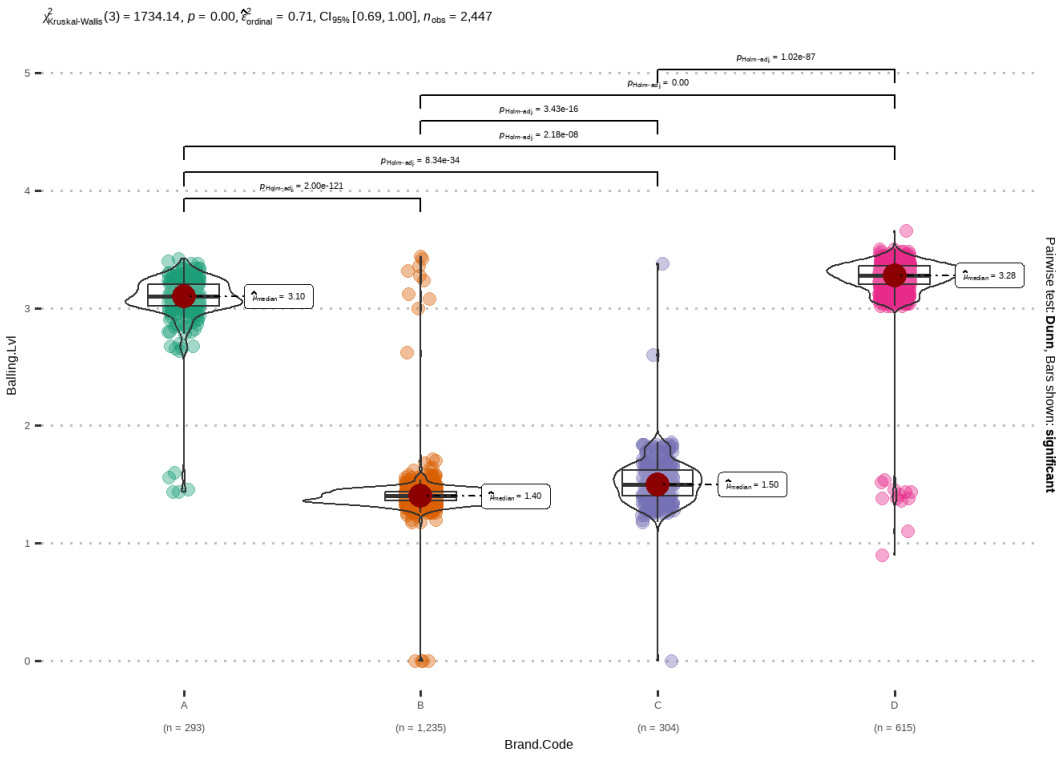
Consequently, we chose the model the target variable **PH** individually by brand A, B, C and D, rather than to model it colletively for all four brands together. Likewise, we model each brand individually so we can achieve a more granular, detailed insight into drivers of pH for each beverage and better understand the manufacturing process.

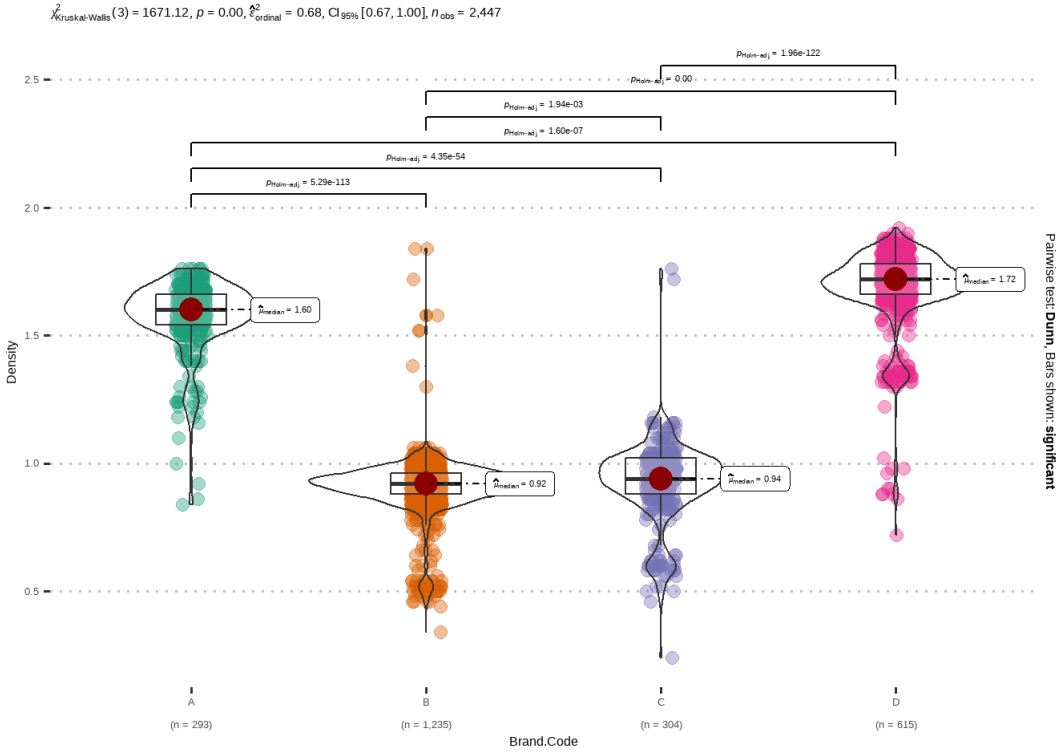
$\chi^2_{\text{Kruskal-Wallis}}(3) = 256.99$, $p = 2.01\text{e-}55$, $\hat{\epsilon}^2_{\text{ordinal}} = 0.11$, $\text{CI}_{95\%}\,[0.09, 1.00]$, $n_{\text{obs}} = 2{,}447$



$\chi^2_{\text{Kruskal-Wallis}}(3) = 1724.84$, $p = 0.00$, $\hat{\epsilon}^2_{\text{ordinal}} = 0.71$, $\text{CI}_{95\%}\,[0.69, 1.00]$, $n_{\text{obs}} = 2{,}447$







Due to the non-linearity, collinearity and outlier contamination in the data, we avoid fitting simple linear models like Ordinary Least Squares regression. Instead, we fit five more complex predictive models. The first, our baseline linear model, is the linear ElasticNet regression model. We expect the L1 and L2 regularization of ElasticNet to account for collinearity and outliers, but we expect it to lag in performance due to its linear natures.

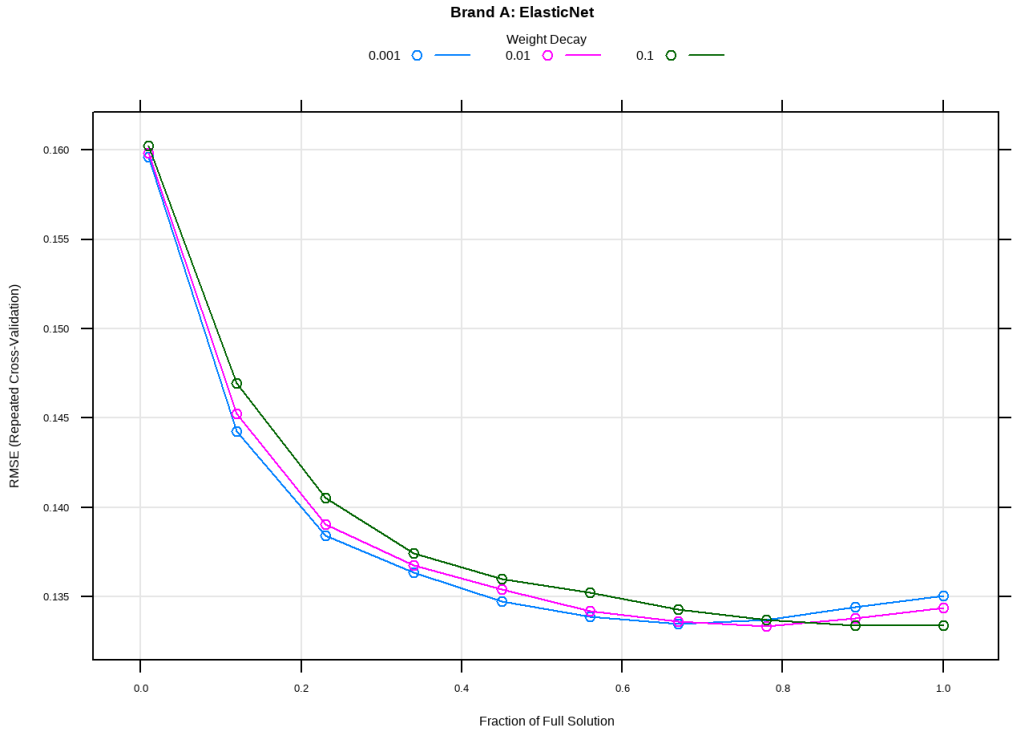
We fit also two non-linear regression models, Multivariate Adaptive Regression Splines (MARS) and Support Vector Machines (SVM). Finally, we fit two non-linear tree-based models, Gradient Boosted Machines (GBM) and Random Forest (RF). Together, we fit the five models to each one of the four beverage brands for a total of twenty different models.

Model resampled performance is determined by lowest estimate Root Mean Square Error (RMSE) metric via 10-fold cross-validation repeated three times for all models except for Gradient Boosted Machines. GBM performance is determined instead with bootstrap resampling with 25 repetitions. Each model is fitted with grid search hyper-parameter tuning in search of the model with lowest resampled RMSE value. For each

brand of the four brands, the model with the lowest resampled RMSE is selected for prediction from among its four competitors. All model training is conducted with `set.seed(111)`.

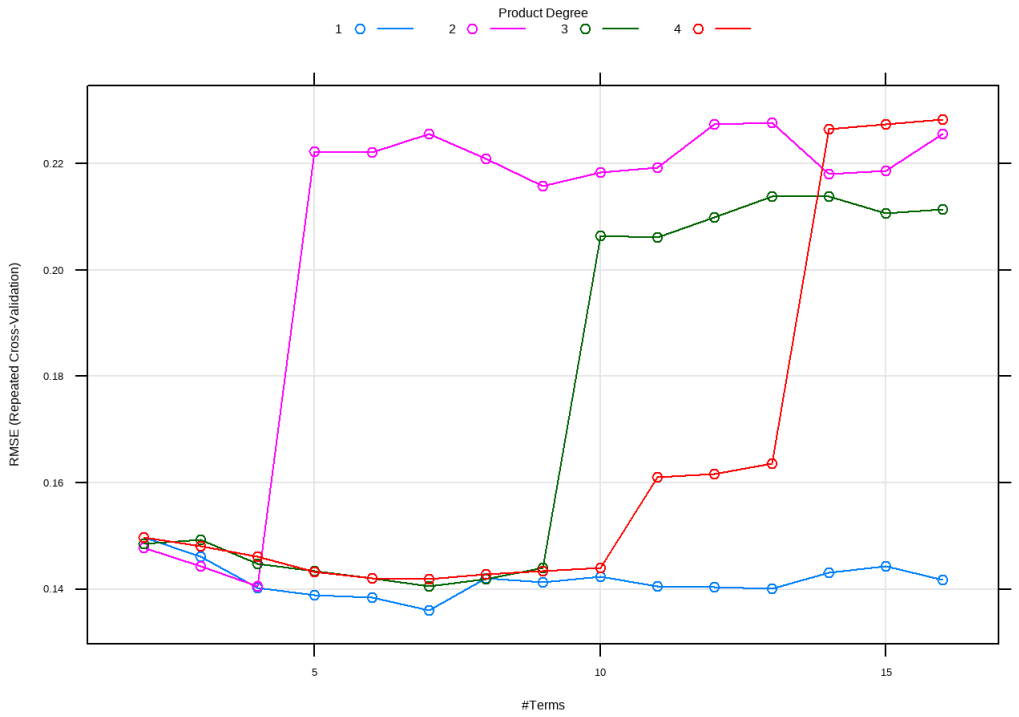
Brand A: Model Fitting

ElasticNet



```
## Elasticnet
##
## 293 samples
##  31 predictor
##
## Pre-processing: centered (31), scaled (31)
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 263, 264, 265, 264, 264, 263, ...
## Resampling results across tuning parameters:
##
##   lambda  fraction  RMSE       Rsquared   MAE
##   0.001   0.01      0.1596176  0.1935028  0.1290054
##   0.001   0.12      0.1442281  0.2649499  0.1177923
##   0.001   0.23      0.1383974  0.3157206  0.1125780
##   0.001   0.34      0.1363303  0.3334398  0.1103391
##   0.001   0.45      0.1347494  0.3481141  0.1090957
##   0.001   0.56      0.1338731  0.3557820  0.1087714
##   0.001   0.67      0.1334910  0.3609864  0.1086946
##   0.001   0.78      0.1336888  0.3627528  0.1089076
##   0.001   0.89      0.1344047  0.3586296  0.1096454
##   0.001   1.00      0.1350621  0.3548472  0.1102096
##   0.010   0.01      0.1598506  0.1920054  0.1291479
##   0.010   0.12      0.1452432  0.2574341  0.1186112
##   0.010   0.23      0.1390343  0.3112382  0.1131543
##   0.010   0.34      0.1367345  0.3291936  0.1107311
##   0.010   0.45      0.1354219  0.3416934  0.1096091
##   0.010   0.56      0.1341999  0.3522386  0.1088372
##   0.010   0.67      0.1336112  0.3582534  0.1087322
##   0.010   0.78      0.1333220  0.3629162  0.1086710
##   0.010   0.89      0.1337649  0.3626014  0.1090385
##   0.010   1.00      0.1343682  0.3594591  0.1095921
##   0.100   0.01      0.1602469  0.1907130  0.1293838
##   0.100   0.12      0.1469249  0.2440517  0.1198450
##   0.100   0.23      0.1405403  0.2949691  0.1145835
##   0.100   0.34      0.1374064  0.3212898  0.1117508
##   0.100   0.45      0.1359968  0.3339862  0.1102823
##   0.100   0.56      0.1352281  0.3423971  0.1095499
##   0.100   0.67      0.1342807  0.3519657  0.1090101
##   0.100   0.78      0.1337071  0.3583894  0.1087967
##   0.100   0.89      0.1333772  0.3630935  0.1087530
##   0.100   1.00      0.1333864  0.3658786  0.1088826
##
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were fraction = 0.78 and lambda = 0.01.
```

Multivariate Regression Adaptive Splines

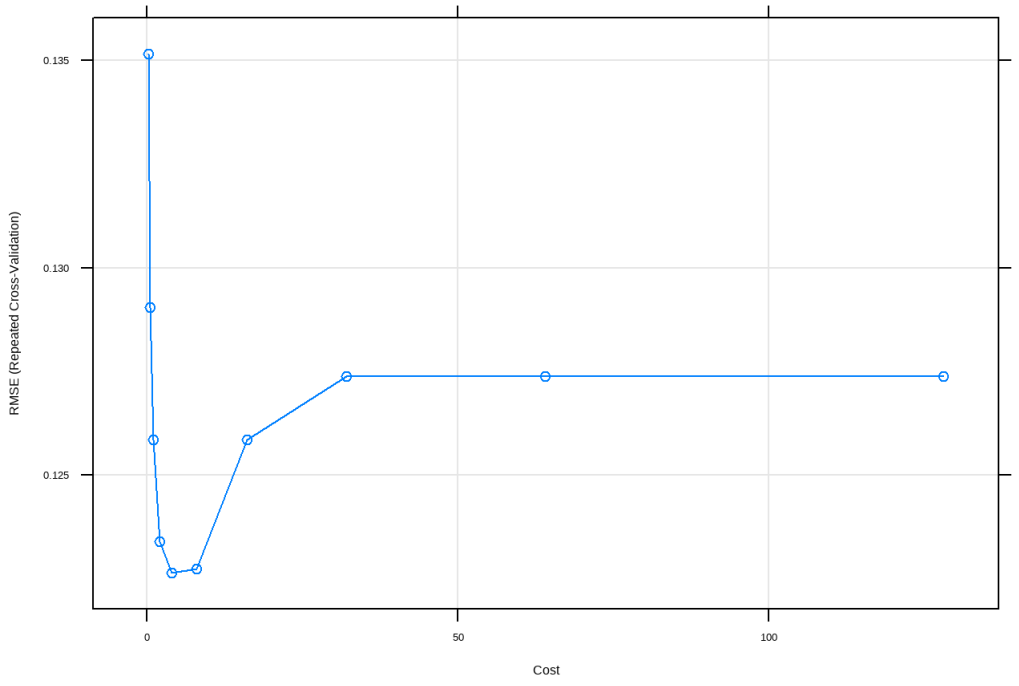



```
## Call: earth(x=data.frame[293,31], y=c(8.26,8.24,8.2...), keepxy=TRUE, degree=1,
##             nprune=7)
##
##                          coefficients
## (Intercept)                 8.3594019
## h(Usage.cont-23.08)        -0.3622916
## h(Usage.cont-23.74)         0.5792048
## h(Carb.Flow-1078)          -0.0000453
## h(48-Pressure.Setpoint)     0.0316661
## h(Balling.Lvl-2.8)          3.4661617
## h(Balling.Lvl-2.86)        -3.2738424
##
## Selected 7 of 50 terms, and 4 of 31 predictors (nprune=7)
## Termination condition: Reached nk 63
## Importance: Usage.cont, Balling.Lvl, Carb.Flow, Mnf.Flow-unused, ...
## Number of terms at each degree of interaction: 1 6 (additive model)
## GCV 0.01607962    RSS 4.302534    GRSq 0.3974131    RSq 0.4459231
```

```
## Multivariate Adaptive Regression Spline
##
## 293 samples
##  31 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 263, 264, 265, 264, 264, 263, ...
## Resampling results across tuning parameters:
##
##   degree  nprune  RMSE       Rsquared   MAE
##   1        2      0.1497895  0.1765104  0.1220520
##   1        3      0.1461241  0.2179489  0.1179056
##   1        4      0.1401672  0.2790081  0.1124763
##   1        5      0.1388269  0.3038616  0.1106929
##   1        6      0.1383952  0.3063864  0.1085944
##   1        7      0.1360784  0.3318833  0.1061964
##   1        8      0.1420504  0.3403795  0.1067228
##   1        9      0.1412433  0.3481863  0.1065976
##   1       10      0.1422962  0.3527384  0.1068634
##   1       11      0.1404718  0.3602394  0.1058330
##   1       12      0.1404604  0.3554585  0.1058684
##   1       13      0.1400636  0.3647653  0.1051840
##   1       14      0.1431069  0.3579358  0.1069793
##   1       15      0.1443706  0.3577208  0.1075569
##   1       16      0.1417314  0.3687306  0.1058541
##   2        2      0.1478067  0.2004894  0.1203446
##   2        3      0.1443139  0.2367403  0.1152442
##   2        4      0.1404713  0.2865896  0.1109984
##   2        5      0.2222557  0.2788055  0.1324524
##   2        6      0.2221659  0.3016988  0.1310854
##   2        7      0.2255985  0.3217352  0.1303869
##   2        8      0.2209616  0.3287764  0.1287254
##   2        9      0.2157660  0.3383243  0.1266696
##   2       10      0.2184277  0.3425960  0.1272096
##   2       11      0.2193643  0.3451729  0.1277245
##   2       12      0.2274581  0.3519932  0.1288637
##   2       13      0.2277323  0.3653638  0.1277046
##   2       14      0.2180609  0.3699529  0.1244608
##   2       15      0.2186685  0.3725972  0.1248910
##   2       16      0.2256669  0.3691158  0.1269161
##   3        2      0.1485152  0.1918464  0.1196063
##   3        3      0.1492309  0.2015505  0.1190837
##   3        4      0.1448037  0.2438274  0.1154733
##   3        5      0.1433823  0.2814787  0.1137647
##   3        6      0.1419791  0.3051124  0.1120085
##   3        7      0.1405001  0.3165243  0.1101833
##   3        8      0.1419150  0.3182626  0.1109443
##   3        9      0.1440282  0.3111473  0.1108471
##   3       10      0.2064421  0.3194179  0.1255848
##   3       11      0.2062345  0.3325774  0.1240654
##   3       12      0.2099551  0.3252820  0.1244176
##   3       13      0.2138250  0.3198227  0.1267056
##   3       14      0.2138233  0.3302131  0.1260539
##   3       15      0.2107035  0.3397736  0.1249467
##   3       16      0.2115291  0.3358571  0.1251702
##   4        2      0.1497441  0.1761908  0.1203147
##   4        3      0.1480237  0.2081231  0.1176112
##   4        4      0.1460975  0.2300717  0.1160110
##   4        5      0.1431923  0.2674400  0.1134285
##   4        6      0.1419831  0.2890076  0.1118377
##   4        7      0.1419197  0.3189492  0.1092421
##   4        8      0.1428714  0.3231410  0.1085312
##   4        9      0.1434773  0.3254148  0.1076015
##   4       10      0.1440525  0.3248637  0.1085567
```

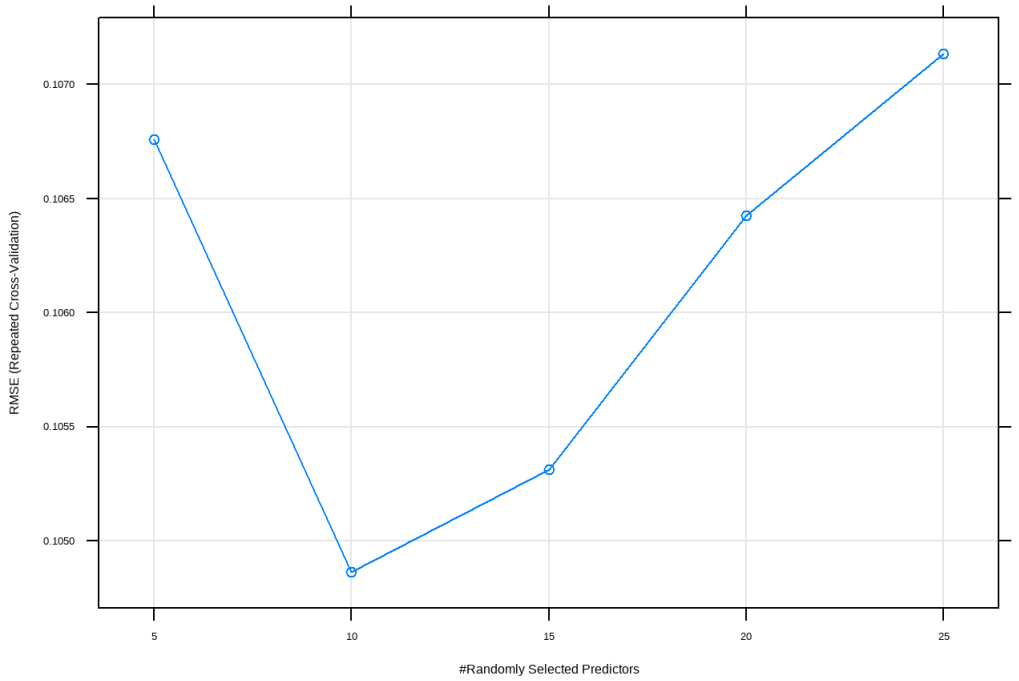
```
##   4       11      0.1610644  0.3278866  0.1120245
##   4       12      0.1616648  0.3363670  0.1115437
##   4       13      0.1636269  0.3337479  0.1106425
##   4       14      0.2264804  0.3288955  0.1258900
##   4       15      0.2274440  0.3313336  0.1260803
##   4       16      0.2283039  0.3399964  0.1264094
##
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were nprune = 7 and degree = 1.
```

Support Vector Machines



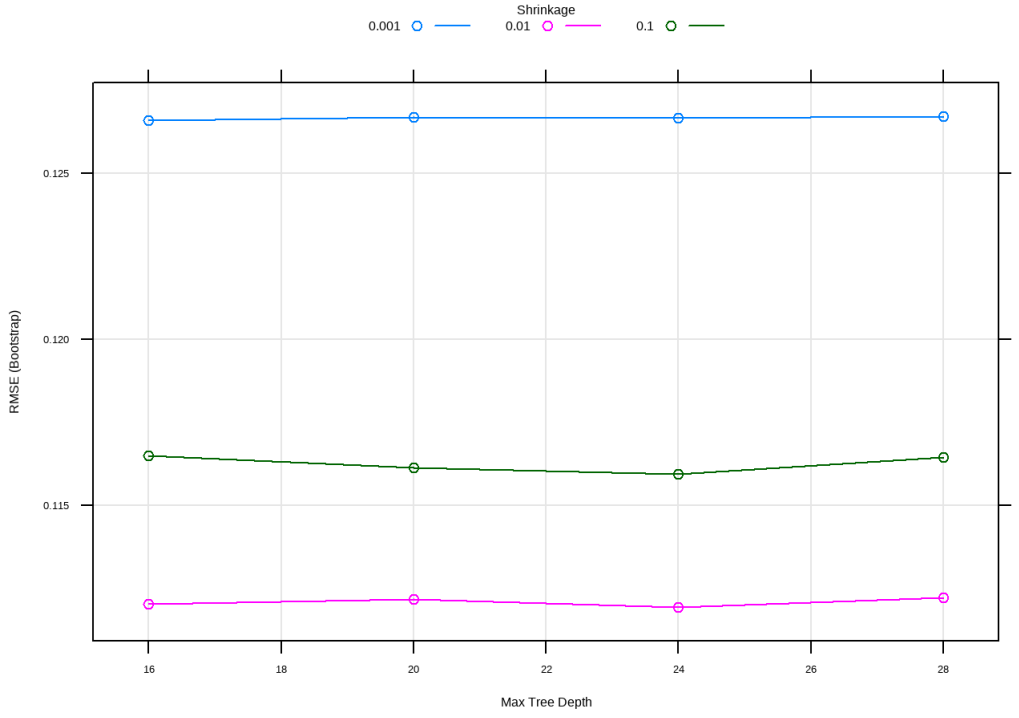
```
## Support Vector Machines with Radial Basis Function Kernel
##
## 293 samples
##  31 predictor
##
## Pre-processing: centered (31), scaled (31)
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 263, 264, 265, 264, 264, 263, ...
## Resampling results across tuning parameters:
##
##   C       RMSE       Rsquared   MAE
##     0.25  0.1351562  0.3697929  0.10701096
##     0.50  0.1290491  0.4101149  0.10092541
##     1.00  0.1258554  0.4238301  0.09663402
##     2.00  0.1234073  0.4371509  0.09456393
##     4.00  0.1226503  0.4427542  0.09365975
##     8.00  0.1227574  0.4460786  0.09358514
##    16.00  0.1258548  0.4263526  0.09616864
##    32.00  0.1273879  0.4165427  0.09777012
##    64.00  0.1273879  0.4165427  0.09777012
##   128.00  0.1273879  0.4165427  0.09777012
##
## Tuning parameter 'sigma' was held constant at a value of 0.02440697
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were sigma = 0.02440697 and C = 4.
```

Random Forest



```
## Random Forest
##
## 293 samples
##  31 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 263, 264, 265, 264, 264, 263, ...
## Resampling results across tuning parameters:
##
##   mtry  RMSE       Rsquared   MAE
##    5    0.1067583  0.6104512  0.08307805
##   10    0.1048628  0.6111798  0.08087744
##   15    0.1053113  0.5995171  0.08069480
##   20    0.1064259  0.5867083  0.08129418
##   25    0.1071341  0.5785370  0.08148475
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 10.
```

Gradient Boosting Machines



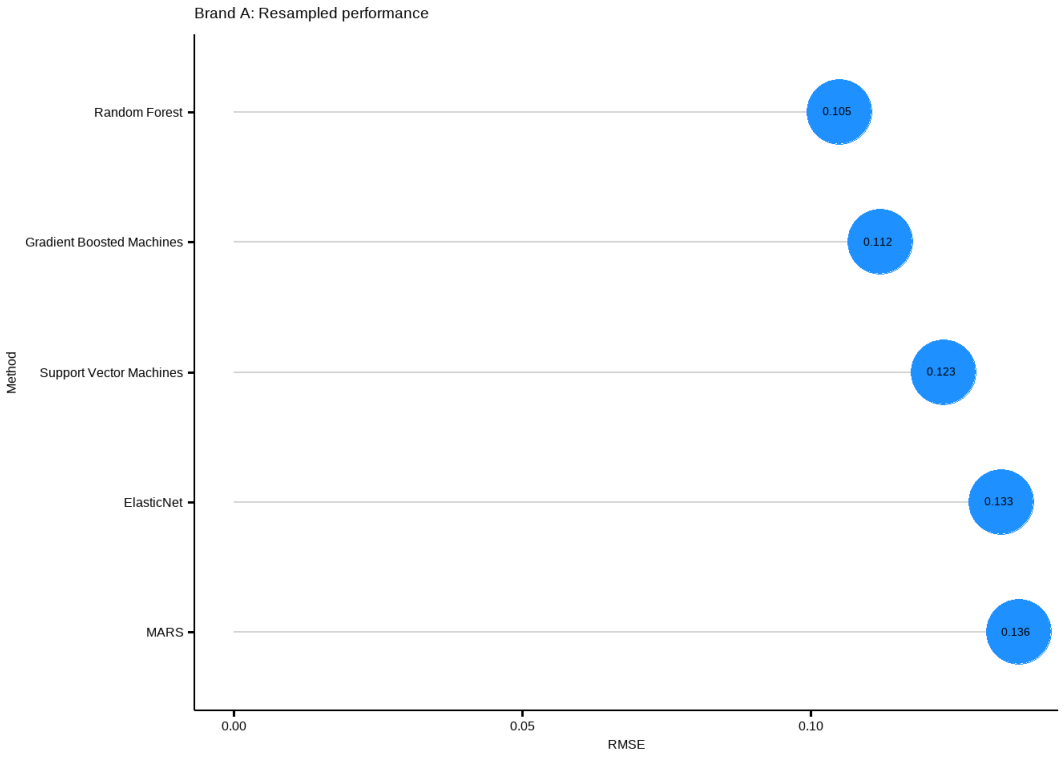
```
## Stochastic Gradient Boosting
##
## 293 samples
##  31 predictor
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 293, 293, 293, 293, 293, 293, ...
## Resampling results across tuning parameters:
##
##   shrinkage  interaction.depth  RMSE       Rsquared   MAE
##   0.001      16                 0.1265889  0.4911136  0.10044966
##   0.001      20                 0.1266912  0.4897948  0.10053831
##   0.001      24                 0.1266717  0.4897186  0.10058050
##   0.001      28                 0.1267118  0.4894791  0.10059477
##   0.010      16                 0.1120268  0.5314421  0.08675730
##   0.010      20                 0.1121656  0.5299483  0.08685698
##   0.010      24                 0.1119409  0.5316753  0.08685785
##   0.010      28                 0.1122142  0.5291169  0.08695866
##   0.100      16                 0.1164895  0.4960969  0.09015960
##   0.100      20                 0.1161320  0.4993099  0.09029883
##   0.100      24                 0.1159495  0.5013164  0.09026446
##   0.100      28                 0.1164530  0.4963438  0.09033184
##
## Tuning parameter 'n.trees' was held constant at a value of 1000
##
## Tuning parameter 'n.minobsinnode' was held constant at a value of 10
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were n.trees = 1000, interaction.depth
##  = 24, shrinkage = 0.01 and n.minobsinnode = 10.
```

Brand A: Model Selection

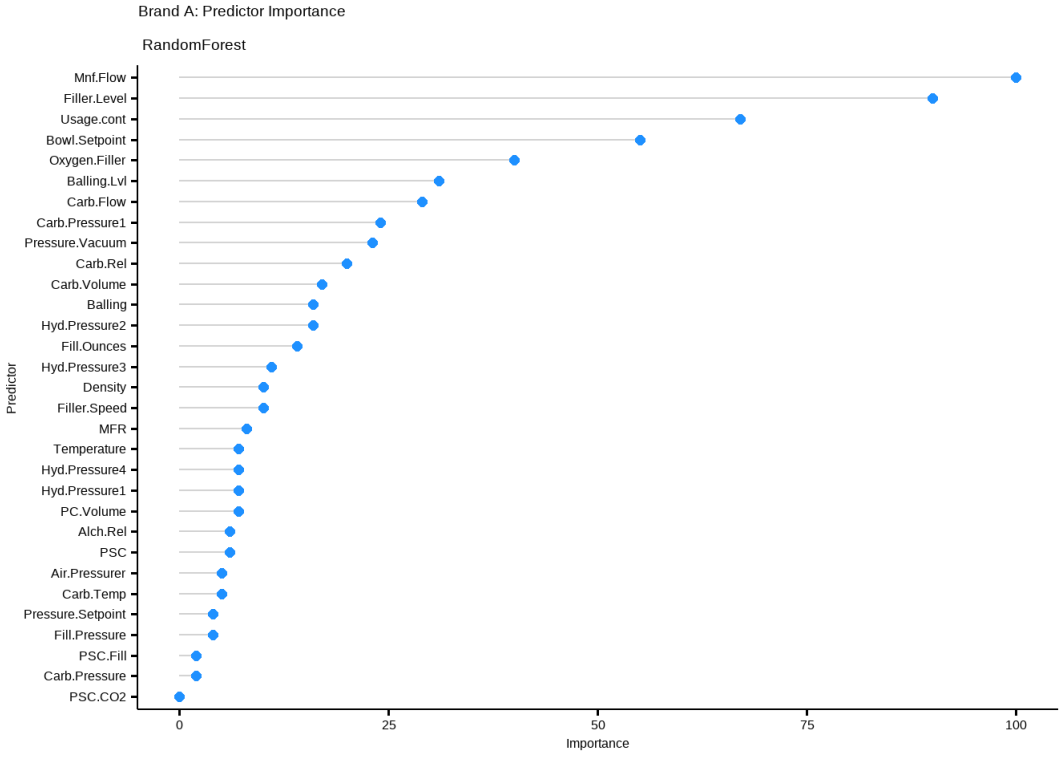
The **Random Forest** model achieves the lowest resampled RMSE on brand **A** among the five models with score of **0.105**.

Brand A: Resampled performance

Method	RMSE
Random Forest	0.105
Gradient Boosted Machines	0.112
Support Vector Machines	0.123
ElasticNet	0.133
MARS	0.136



Brand A: Model Variable Importance




```{=html}
```

Brand A: Predictor Importance

Predictor	Importance
Mnf.Flow	100
Filler.Level	90
Usage.cont	67
Bowl.Setpoint	55

Brand A: Predictor Importance

Predictor	Importance
Oxygen.Filler	40
Balling.Lvl	31
Carb.Flow	29
Carb.Pressure1	24
Pressure.Vacuum	23
Carb.Rel	20
Carb.Volume	17
Hyd.Pressure2	16
Balling	16
Fill.Ounces	14
Hyd.Pressure3	11
Filler.Speed	10
Density	10
MFR	8
PC.Volume	7
Hyd.Pressure1	7
Hyd.Pressure4	7
Temperature	7
PSC	6
Alch.Rel	6
Carb.Temp	5
Air.Pressurer	5
Fill.Pressure	4
Pressure.Setpoint	4
Carb.Pressure	2
PSC.Fill	2
PSC.CO2	0

```
## Brand B: Model Fitting

### ElasticNet

<img src="TechnicalReport_ABC_Beverage_pHprediction_files/figure-html/unnamed-chunk-25-1.png" width="672" />
```

Elasticnet

1235 samples

31 predictor

Pre-processing: centered (31), scaled (31)

Resampling: Cross-Validated (10 fold, repeated 3 times)

Summary of sample sizes: 1111, 1113, 1111, 1111, 1112, 1112, …

Resampling results across tuning parameters:

lambda fraction RMSE Rsquared MAE

0.001 0.01 0.1652423 0.3409641 0.13683314

0.001 0.12 0.1390092 0.3615839 0.11159151

0.001 0.23 0.1309643 0.4112674 0.10362063

0.001 0.34 0.1267166 0.4465020 0.09884304

0.001 0.45 0.1239159 0.4687172 0.09533422

0.001 0.56 0.1222295 0.4808906 0.09310141

0.001 0.67 0.1216164 0.4854818 0.09227329

0.001 0.78 0.1214609 0.4872509 0.09206097

0.001 0.89 0.1215219 0.4873384 0.09210781

0.001 1.00 0.1217524 0.4860900 0.09229237

0.010 0.01 0.1655206 0.3409641 0.13711036

0.010 0.12 0.1404062 0.3605443 0.11303416

0.010 0.23 0.1320246 0.4016948 0.10464566

0.010 0.34 0.1276288 0.4383887 0.09989257

0.010 0.45 0.1247811 0.4616765 0.09644070

0.010 0.56 0.1227853 0.4767621 0.09383240

0.010 0.67 0.1218127 0.4838273 0.09257317

0.010 0.78 0.1215211 0.4864472 0.09211097

0.010 0.89 0.1214453 0.4876158 0.09205813

0.010 1.00 0.1216071 0.4868443 0.09218934

0.100 0.01 0.1661947 0.3409641 0.13779035

0.100 0.12 0.1445916 0.3583536 0.11709061

0.100 0.23 0.1349209 0.3771477 0.10760925

0.100 0.34 0.1304576 0.4108670 0.10291880

0.100 0.45 0.1276366 0.4338144 0.09963974

0.100 0.56 0.1256710 0.4504125 0.09719450

0.100 0.67 0.1240111 0.4642712 0.09507435

0.100 0.78 0.1229946 0.4728028 0.09386922

0.100 0.89 0.1224307 0.4780204 0.09322736

0.100 1.00 0.1222296 0.4804798 0.09295743

RMSE was used to select the optimal model using the smallest value.

The final values used for the model were fraction = 0.89 and lambda = 0.01.

```
### Multivariate Regression Adaptive Splines
<img src="TechnicalReport_ABC_Beverage_pHprediction_files/figure-html/unnamed-chunk-26-1.png" width="672" />
```

Call: earth(x=data.frame[1235,31], y=c(8.36,8.94,8.3…), keepxy=TRUE,

degree=2, nprune=16)

coefficients

(Intercept) 8.6015845

h(Bowl.Setpoint-90) 0.0053199

h(146.8-Air.Pressurer) -0.0683323

h(0.2-Mnf.Flow) * h(66.4-Temperature) 0.0006741

h(Mnf.Flow-0.2) * h(Balling-1.498) 0.0011596

h(Mnf.Flow-0.2) * h(1.498-Balling) 0.0039373

h(0.2-Mnf.Flow) * h(Air.Pressurer-143.8) -0.0009162

h(0.2-Mnf.Flow) * h(143.8-Air.Pressurer) 0.0009451

h(0.2-Mnf.Flow) * h(5.4-Carb.Rel) -0.0080825

h(Mnf.Flow-0.2) * h(Carb.Rel-5.32) -0.0072044

h(120.2-Carb.Pressure1) * h(146.8-Air.Pressurer) -0.0033682

h(1734-Carb.Flow) * h(146.8-Air.Pressurer) 0.0000207

h(Carb.Flow-1734) * h(146.8-Air.Pressurer) 0.0000202

h(Oxygen.Filler-0.029) * h(Bowl.Setpoint-90) 0.0357323

h(Oxygen.Filler-0.036) * h(146.8-Air.Pressurer) -0.4298194

h(Bowl.Setpoint-90) * h(46-Pressure.Setpoint) 0.0039144

Selected 16 of 29 terms, and 10 of 31 predictors (nprune=16)

Termination condition: RSq changed by less than 0.001 at 29 terms

Importance: Mnf.Flow, Air.Pressurer, Bowl.Setpoint, Temperature,

…

Number of terms at each degree of interaction: 1 2 13

GCV 0.0117318 RSS 13.59951 GRSq 0.5894491 RSq 0.6140224

Multivariate Adaptive Regression Spline

1235 samples

31 predictor

No pre-processing

Resampling: Cross-Validated (10 fold, repeated 3 times)

Summary of sample sizes: 1111, 1113, 1111, 1111, 1112, 1112, …

Resampling results across tuning parameters:

degree nprune RMSE Rsquared MAE

1 2 0.1376590 0.3383746 0.10833970

1 3 0.1336080 0.3762613 0.10565443

1 4 0.1327781 0.3852614 0.10444933

1 5 0.1285716 0.4234443 0.10018887

1 6 0.1266557 0.4413159 0.09702042

1 7 0.1248935 0.4574990 0.09467981

1 8 0.1214034 0.4867291 0.09213142

1 9 0.1208215 0.4921892 0.09110060

1 10 0.1203172 0.4968168 0.09034941

1 11 0.1187077 0.5100245 0.08916183

1 12 0.1168915 0.5252552 0.08755587

1 13 0.1168058 0.5262437 0.08745611

1 14 0.1162239 0.5308531 0.08710429

1 15 0.1157749 0.5345730 0.08657140

1 16 0.1158079 0.5344282 0.08658049

2 2 0.1370339 0.3448825 0.10718563

2 3 0.1303615 0.4077230 0.09959505

2 4 0.1280788 0.4286184 0.09730094

2 5 0.1250684 0.4556830 0.09503998

2 6 0.1223057 0.4792355 0.09267291

2 7 0.1208348 0.4930166 0.09083409

2 8 0.1189300 0.5085630 0.08889793

2 9 0.1175218 0.5204787 0.08785422

2 10 0.1154574 0.5377138 0.08595411

2 11 0.1140075 0.5487134 0.08465550

2 12 0.1128512 0.5573491 0.08369692

2 13 0.1122029 0.5628858 0.08331160

2 14 0.1113621 0.5696090 0.08277665

2 15 0.1106083 0.5753214 0.08219012

2 16 0.1105250 0.5759511 0.08205863

3 2 0.1367048 0.3477184 0.10731205

3 3 0.1298599 0.4116674 0.09934184

3 4 0.1260093 0.4472454 0.09562663

3 5 0.1245898 0.4603259 0.09416319

3 6 0.1208320 0.4914175 0.09094513

3 7 0.1185882 0.5104602 0.08907551

3 8 0.1181712 0.5146173 0.08871058

3 9 0.1162828 0.5304720 0.08704796

3 10 0.1154730 0.5380710 0.08585513

3 11 0.1151552 0.5410036 0.08572203

3 12 0.1139981 0.5504968 0.08480900

3 13 0.1126208 0.5611833 0.08373807

3 14 0.1123451 0.5652425 0.08303995

3 15 0.1118023 0.5694382 0.08246029

3 16 0.1107164 0.5766264 0.08132145

4 2 0.1369323 0.3453940 0.10728729

4 3 0.1294951 0.4152926 0.09853600

4 4 0.1271269 0.4367082 0.09593112

4 5 0.1246192 0.4581024 0.09418562

4 6 0.1200031 0.4977890 0.09066310

4 7 0.1186430 0.5100276 0.08914485

4 8 0.1175547 0.5195261 0.08802284

4 9 0.1160215 0.5326032 0.08683248

4 10 0.1147167 0.5429736 0.08562247

4 11 0.1134417 0.5526235 0.08435578

4 12 0.1213555 0.5373360 0.08477349

4 13 0.1215669 0.5384856 0.08454100

4 14 0.1205200 0.5470036 0.08357502

4 15 0.1206672 0.5453831 0.08375969

4 16 0.1203285 0.5480727 0.08314397

RMSE was used to select the optimal model using the smallest

value.

The final values used for the model were nprune = 16 and degree = 2.

```
### Support Vector Machines
<img src="TechnicalReport_ABC_Beverage_pHprediction_files/figure-html/unnamed-chunk-27-1.png" width="672" />
```

Support Vector Machines with Radial Basis Function Kernel

1235 samples

31 predictor

Pre-processing: centered (31), scaled (31)

Resampling: Cross-Validated (10 fold, repeated 3 times)

Summary of sample sizes: 1111, 1113, 1111, 1111, 1112, 1112, …

Resampling results across tuning parameters:

C RMSE Rsquared MAE

0.25 0.1111147 0.5784868 0.08080179

0.50 0.1068685 0.6070192 0.07726026

1.00 0.1039448 0.6262299 0.07491341

2.00 0.1018102 0.6400147 0.07368793

4.00 0.1005462 0.6478477 0.07395346

8.00 0.1000319 0.6514387 0.07435215

16.00 0.1016850 0.6430623 0.07611450

32.00 0.1057505 0.6211409 0.07949674

64.00 0.1115816 0.5906531 0.08358961

128.00 0.1162263 0.5679631 0.08709812

Tuning parameter ‘sigma’ was held constant at a value of 0.02449773

RMSE was used to select the optimal model using the smallest value.

The final values used for the model were sigma = 0.02449773 and

C = 8.

```
### Random Forest
<img src="TechnicalReport_ABC_Beverage_pHprediction_files/figure-html/unnamed-chunk-28-1.png" width="672" />
```

Random Forest

1235 samples

31 predictor

No pre-processing

Resampling: Cross-Validated (10 fold, repeated 3 times)

Summary of sample sizes: 1111, 1113, 1111, 1111, 1112, 1112, …

Resampling results across tuning parameters:

mtry RMSE Rsquared MAE

5 0.08978463 0.7353356 0.06575872

10 0.08622702 0.7503952 0.06255348

15 0.08501646 0.7546690 0.06146220

20 0.08462013 0.7547007 0.06083610

25 0.08469223 0.7523943 0.06052719

RMSE was used to select the optimal model using the smallest value.

The final value used for the model was mtry = 20.

```
### Gradient Boosting Machines
<img src="TechnicalReport_ABC_Beverage_pHprediction_files/figure-html/unnamed-chunk-29-1.png" width="672" />
```

Stochastic Gradient Boosting

1235 samples

31 predictor

No pre-processing

Resampling: Bootstrapped (25 reps)

Summary of sample sizes: 1235, 1235, 1235, 1235, 1235, 1235, …

Resampling results across tuning parameters:

shrinkage interaction.depth RMSE Rsquared MAE

0.001 16 0.11658960 0.6599780 0.09151848

0.001 20 0.11520064 0.6706141 0.09022432

0.001 24 0.11434946 0.6759719 0.08930533

0.001 28 0.11368182 0.6811031 0.08863051

0.010 16 0.09347134 0.7038298 0.06859140

0.010 20 0.09280375 0.7081851 0.06778656

0.010 24 0.09245613 0.7104377 0.06722711

0.010 28 0.09241247 0.7104852 0.06704985

0.100 16 0.09620942 0.6845260 0.07094653

0.100 20 0.09561164 0.6879198 0.07024458

0.100 24 0.09546584 0.6893213 0.06990972

0.100 28 0.09551639 0.6889588 0.06985230

Tuning parameter ‘n.trees’ was held constant at a value of 1000

Tuning parameter ‘n.minobsinnode’ was held constant at a value of

RMSE was used to select the optimal model using the smallest value.

The final values used for the model were n.trees = 1000,

interaction.depth

= 28, shrinkage = 0.01 and n.minobsinnode = 10.

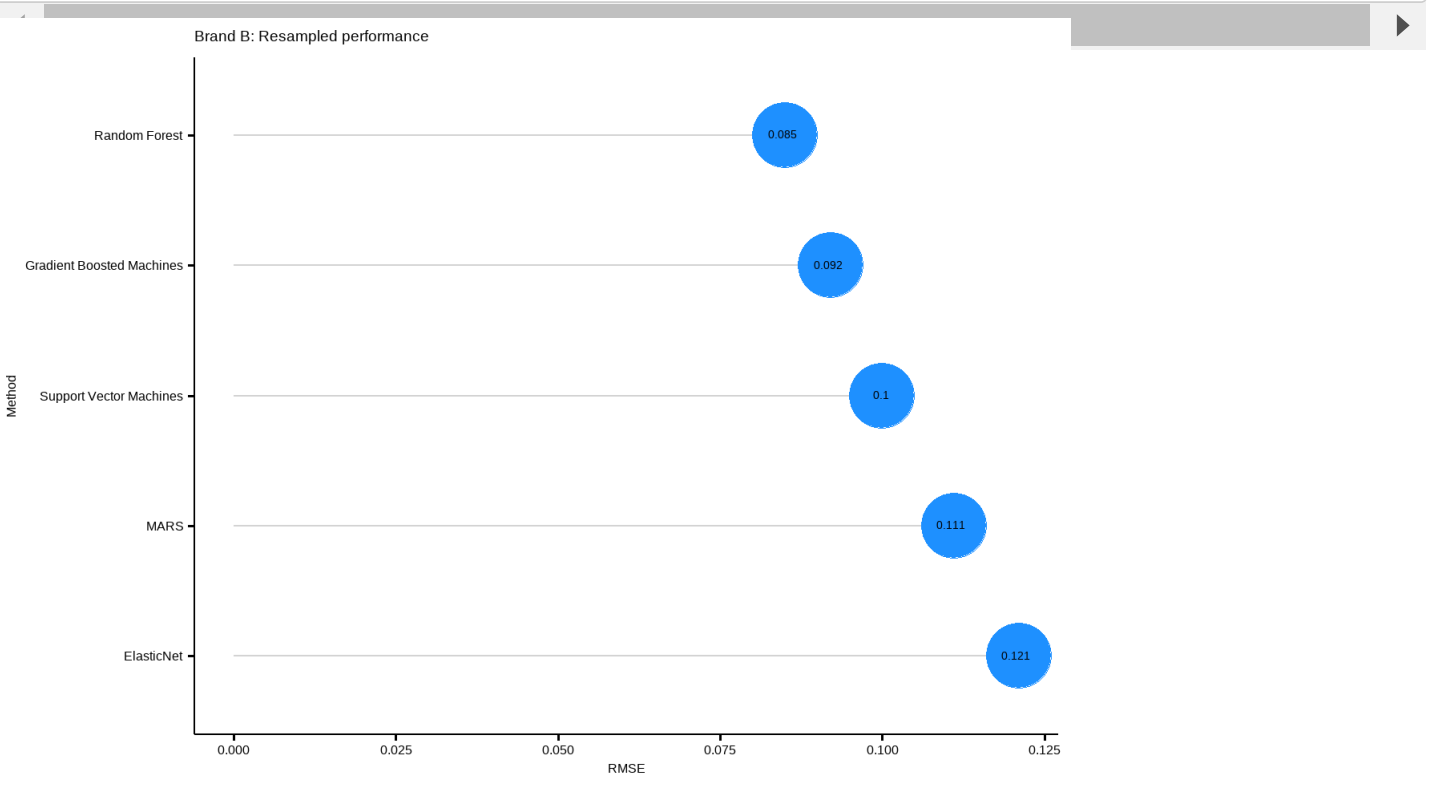
```
## Brand B: Model Selection

The **Random Forest** model achieves the lowest resampled RMSE on brand **B** among the five models with score of **0.085*
*.

```{=html}
<template id="19db8364-b70a-4b46-8b95-25e819ce2472"><style>
.tabwid table{
  border-spacing:0px !important;
  border-collapse:collapse;
  line-height:1;
  margin-left:auto;
  margin-right:auto;
  border-width: 0;
  display: table;
  border-color: transparent;
  caption-side: top;
}
.tabwid-caption-bottom table{
  caption-side: bottom;
}
.tabwid_left table{
  margin-left:0;
}
.tabwid_right table{
  margin-right:0;
}
.tabwid td {
    padding: 0;
}
.tabwid a {
  text-decoration: none;
}
.tabwid thead {
    background-color: transparent;
}
.tabwid tfoot {
    background-color: transparent;
}
.tabwid table tr {
background-color: transparent;
}
.katex-display {
    margin: 0 0 !important;
}
</style><div class="tabwid"><style>.cl-2f5de1cc{}.cl-2f56bafa{font-family:'Arial';font-size:14pt;font-weight:bold;font-sty
le:normal;text-decoration:none;color:rgba(0, 0, 0, 1.00);background-color:transparent;}.cl-2f56bafb{font-family:'Arial';fo
nt-size:11pt;font-weight:bold;font-style:normal;text-decoration:none;color:rgba(0, 0, 0, 1.00);background-color:transparen
t;}.cl-2f56bafc{font-family:'Arial';font-size:11pt;font-weight:normal;font-style:normal;text-decoration:none;color:rgba(0,
0, 0, 1.00);background-color:transparent;}.cl-2f591d4a{margin:0;text-align:left;border-bottom: 0 solid rgba(0, 0, 0, 1.0
0);border-top: 0 solid rgba(0, 0, 0, 1.00);border-left: 0 solid rgba(0, 0, 0, 1.00);border-right: 0 solid rgba(0, 0, 0, 1.
00);padding-bottom:5pt;padding-top:5pt;padding-left:5pt;padding-right:5pt;line-height: 1;background-color:transparent;}.cl
-2f591d4b{margin:0;text-align:right;border-bottom: 0 solid rgba(0, 0, 0, 1.00);border-top: 0 solid rgba(0, 0, 0, 1.00);bor
der-left: 0 solid rgba(0, 0, 0, 1.00);border-right: 0 solid rgba(0, 0, 0, 1.00);padding-bottom:5pt;padding-top:5pt;padding
-left:5pt;padding-right:5pt;line-height: 1;background-color:transparent;}.cl-2f591d4c{width:2.5in;background-color:transpa
rent;vertical-align: middle;border-bottom: 2pt solid rgba(102, 102, 102, 1.00);border-top: 2pt solid rgba(102, 102, 102,
1.00);border-left: 0 solid rgba(0, 0, 0, 1.00);border-right: 0 solid rgba(0, 0, 0, 1.00);margin-bottom:0;margin-top:0;marg
in-left:0;margin-right:0;}.cl-2f591d4d{width:2.5in;background-color:transparent;vertical-align: middle;border-bottom: 2pt
```
```

solid rgba(102, 102, 102, 1.00);border-top: 2pt solid rgba(102, 102, 102, 1.00);border-left: 0 solid rgba(0, 0, 0, 1.00);border-right: 0 solid rgba(0, 0, 0, 1.00);margin-bottom:0;margin-top:0;margin-left:0;margin-right:0;}.cl-2f591d4e{width:2.5in;background-color:transparent;vertical-align: middle;border-bottom: 0 solid rgba(0, 0, 0, 1.00);border-top: 0 solid rgba(0, 0, 0, 1.00);border-left: 0 solid rgba(0, 0, 0, 1.00);border-right: 0 solid rgba(0, 0, 0, 1.00);margin-bottom:0;margin-top:0;margin-left:0;margin-right:0;}.cl-2f591d4f{width:2.5in;background-color:transparent;vertical-align: middle;border-bottom: 0 solid rgba(0, 0, 0, 1.00);border-top: 0 solid rgba(0, 0, 0, 1.00);border-left: 0 solid rgba(0, 0, 0, 1.00);border-right: 0 solid rgba(0, 0, 0, 1.00);margin-bottom:0;margin-top:0;margin-left:0;margin-right:0;}.cl-2f591d50{width:2.5in;background-color:transparent;vertical-align: middle;border-bottom: 2pt solid rgba(102, 102, 102, 1.00);border-top: 0 solid rgba(0, 0, 0, 1.00);border-left: 0 solid rgba(0, 0, 0, 1.00);border-right: 0 solid rgba(0, 0, 0, 1.00);margin-bottom:0;margin-top:0;margin-left:0;margin-right:0;}.cl-2f591d51{width:2.5in;background-color:transparent;vertical-align: middle;border-bottom: 2pt solid rgba(102, 102, 102, 1.00);border-top: 0 solid rgba(0, 0, 0, 1.00);border-left: 0 solid rgba(0, 0, 0, 1.00);border-right: 0 solid rgba(0, 0, 0, 1.00);margin-bottom:0;margin-top:0;margin-left:0;margin-right:0;}</style><table class='cl-2f5de1cc'><thead><tr style="overflow-wrap:break-word;"><td colspan="2"class="cl-2f591d4c"> Brand B: Resampled performance </td></tr><tr style="overflow-wrap:break-word;"><td class="cl-2f591d4c"> Method </td><td class="cl-2f591d4d"> RMSE </td></tr></thead><tbody><tr style="overflow-wrap:break-word;"><td class="cl-2f591d4e"> Random Forest </td><td class="cl-2f591d4f"> 0.085 </td></tr><tr style="overflow-wrap:break-word;"><td class="cl-2f591d4e"> Gradient Boosted Machines </td><td class="cl-2f591d4f"> 0.092 </td></tr><tr style="overflow-wrap:break-word;"><td class="cl-2f591d4e"> Support Vector Machines </td><td class="cl-2f591d4f"> 0.100 </td></tr><tr style="overflow-wrap:break-word;"><td class="cl-2f591d4e"> MARS </td><td class="cl-2f591d4f"> 0.111 </td></tr><tr style="overflow-wrap:break-word;"><td class="cl-2f591d50"> ElasticNet </td><td class="cl-2f591d51"> 0.121 </td></tr></tbody></table> </template>

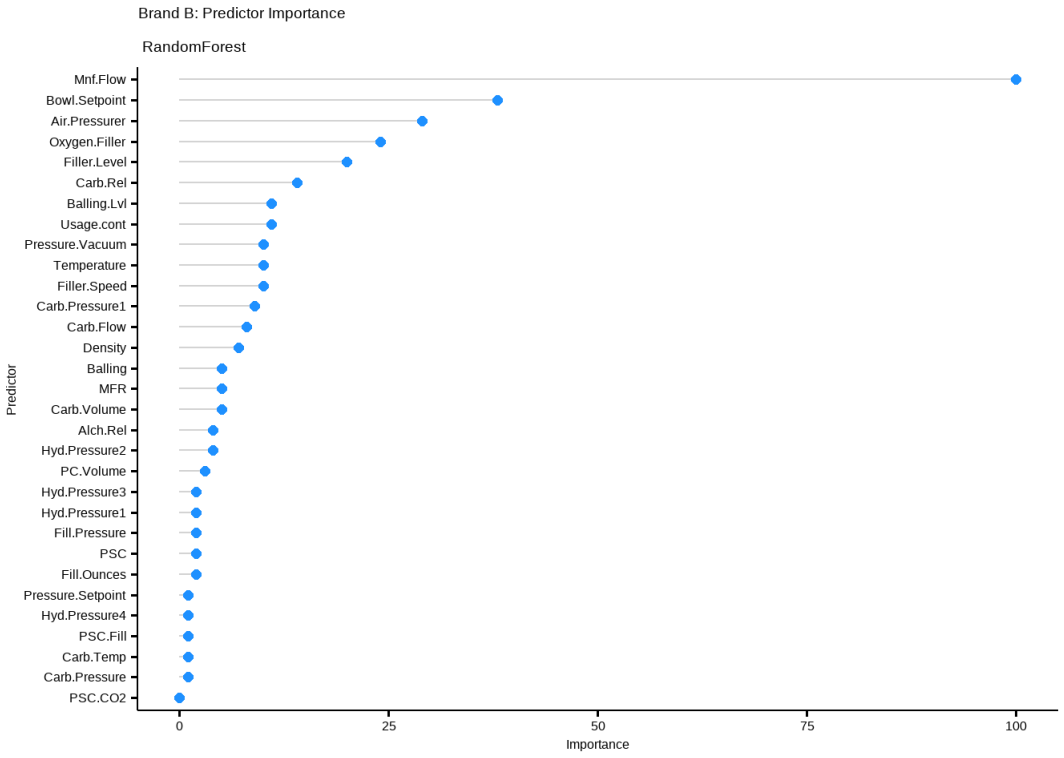
<script>
var dest = document.getElementById("efb8130c-75cc-47ff-bba6-aae157570e5b");
var template = document.getElementById("19db8364-b70a-4b46-8b95-25e819ce2472");
var caption = template.content.querySelector("caption");
var fantome = dest.attachShadow({mode: 'open'});
var templateContent = template.content;
fantome.appendChild(templateContent);
</script>



Brand B: Model Variable Importance

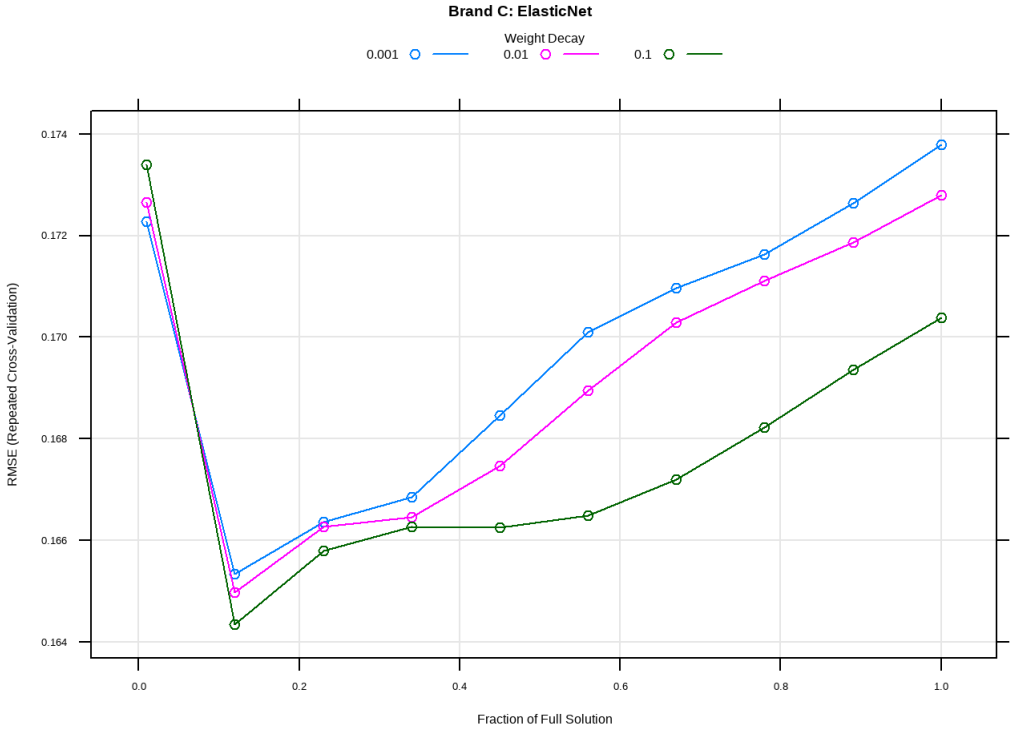
Brand B: Predictor Importance RandomForest

Predictor	Importance
Mnf.Flow	100
Bowl.Setpoint	38
Air.Pressurer	29
Oxygen.Filler	24
Filler.Level	20
Carb.Rel	14
Usage.cont	11
Balling.Lvl	11
Filler.Speed	10
Temperature	10
Pressure.Vacuum	10
Carb.Pressure1	9
Carb.Flow	8
Density	7
Carb.Volume	5
MFR	5
Balling	5
Hyd.Pressure2	4
Alch.Rel	4
PC.Volume	3
Fill.Ounces	2
PSC	2
Fill.Pressure	2
Hyd.Pressure1	2
Hyd.Pressure3	2
Carb.Pressure	1
Carb.Temp	1
PSC.Fill	1
Hyd.Pressure4	1
Pressure.Setpoint	1
PSC.CO2	0



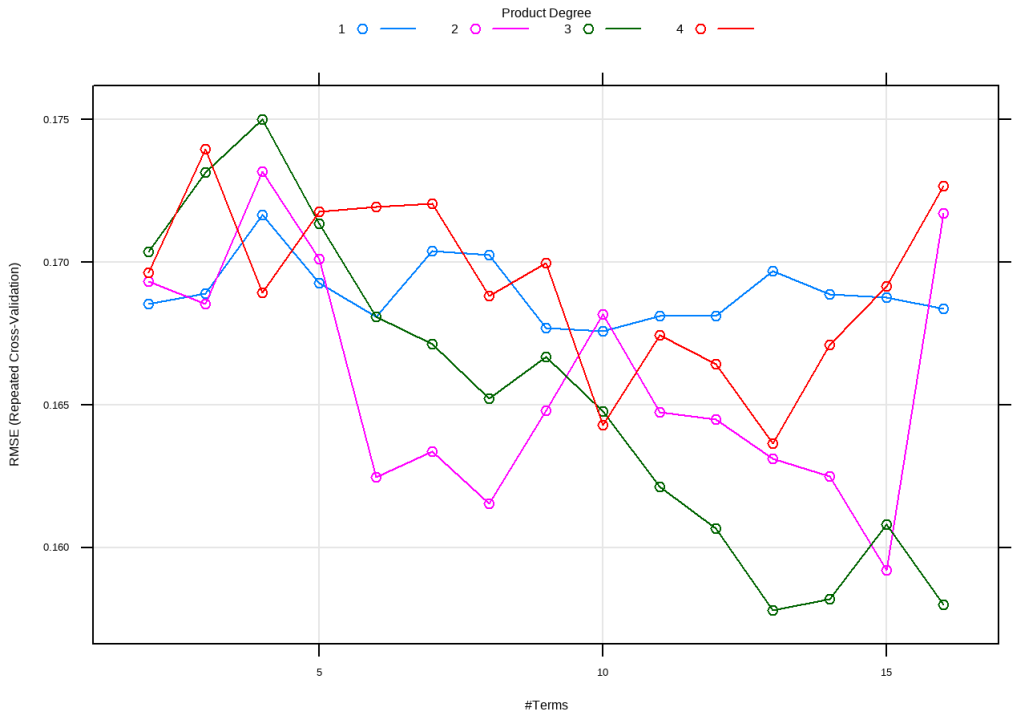
Brand C: Model Fitting

ElasticNet



```
## Elasticnet
##
## 304 samples
##  31 predictor
##
## Pre-processing: centered (31), scaled (31)
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 273, 274, 274, 274, 274, 272, ...
## Resampling results across tuning parameters:
##
##   lambda  fraction  RMSE       Rsquared   MAE
##   0.001   0.01      0.1722820  0.1593309  0.1342385
##   0.001   0.12      0.1653409  0.1283079  0.1273742
##   0.001   0.23      0.1663700  0.1247517  0.1278237
##   0.001   0.34      0.1668602  0.1302383  0.1277484
##   0.001   0.45      0.1684544  0.1303583  0.1286477
##   0.001   0.56      0.1701025  0.1289478  0.1296269
##   0.001   0.67      0.1709730  0.1305921  0.1301528
##   0.001   0.78      0.1716317  0.1308863  0.1306072
##   0.001   0.89      0.1726393  0.1292039  0.1313884
##   0.001   1.00      0.1737936  0.1268430  0.1322163
##   0.010   0.01      0.1726468  0.1593309  0.1345657
##   0.010   0.12      0.1649818  0.1332953  0.1273280
##   0.010   0.23      0.1662771  0.1231393  0.1278507
##   0.010   0.34      0.1664648  0.1288330  0.1276734
##   0.010   0.45      0.1674662  0.1306343  0.1281119
##   0.010   0.56      0.1689524  0.1299580  0.1289862
##   0.010   0.67      0.1702924  0.1293107  0.1298130
##   0.010   0.78      0.1711118  0.1303109  0.1302842
##   0.010   0.89      0.1718594  0.1299886  0.1308215
##   0.010   1.00      0.1728037  0.1284435  0.1315358
##   0.100   0.01      0.1733911  0.1593309  0.1352258
##   0.100   0.12      0.1643420  0.1467663  0.1271970
##   0.100   0.23      0.1657991  0.1231052  0.1276187
##   0.100   0.34      0.1662740  0.1237830  0.1278303
##   0.100   0.45      0.1662604  0.1291528  0.1277279
##   0.100   0.56      0.1664974  0.1331489  0.1277337
##   0.100   0.67      0.1671960  0.1337296  0.1281064
##   0.100   0.78      0.1682288  0.1327481  0.1287422
##   0.100   0.89      0.1693544  0.1316941  0.1294594
##   0.100   1.00      0.1703829  0.1306617  0.1301378
##
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were fraction = 0.12 and lambda = 0.1.
```

Multivariate Regression Adaptive Splines

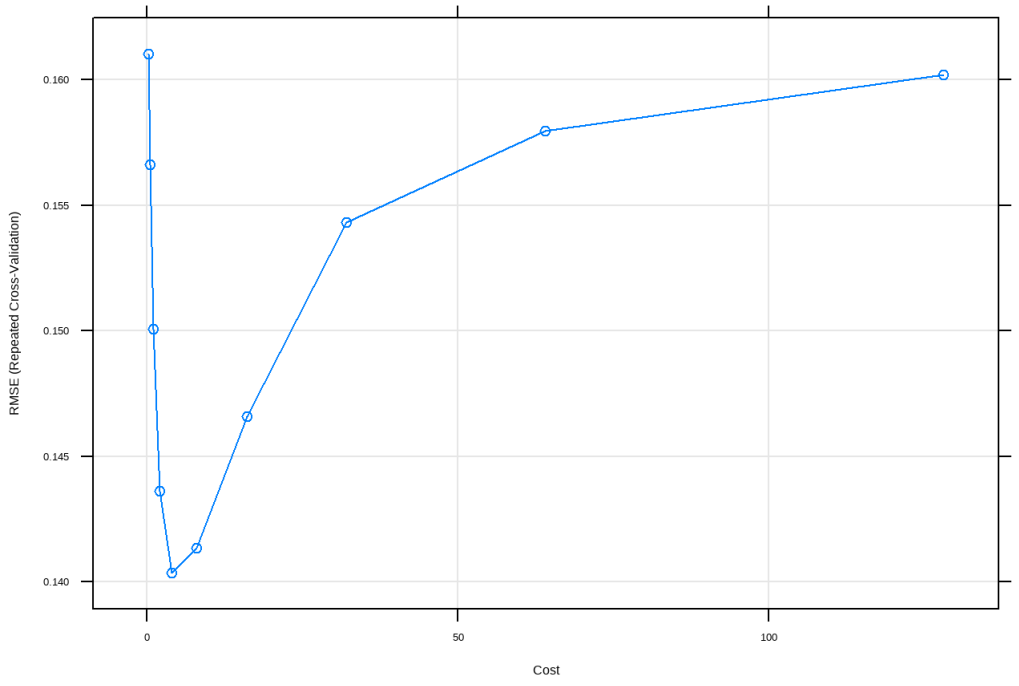



```
## Call: earth(x=data.frame[304,31], y=c(8.58,8.3,8.42...), keepxy=TRUE, degree=3,
##             nprune=13)
##
##                                                                   coefficients
## (Intercept)                                                          8.4486284
## h(150.4-Mnf.Flow)                                                    0.0004872
## h(Carb.Rel-5.3)                                                     -0.5697936
## PSC.Fill * h(65.4-Temperature)                                      -1.2393402
## h(Mnf.Flow-150.4) * h(12-Hyd.Pressure1)                              3.8756520
## h(37.6-Hyd.Pressure3) * h(1.448-Balling)                             0.0114570
## h(Temperature-65.4) * h(Pressure.Vacuum- -5.6)                       -0.0987472
## h(Pressure.Vacuum- -5.6) * h(1.58-Balling.Lvl)                       0.6347300
## h(0.064-Oxygen.Filler) * h(1.58-Balling.Lvl)                       -12.1126087
## Carb.Volume * h(Mnf.Flow-150.4) * h(12-Hyd.Pressure1)               -0.7325321
## h(PC.Volume-0.246667) * h(150.4-Mnf.Flow) * h(31.2-Hyd.Pressure2)    -0.0003400
## h(122-Carb.Pressure1) * h(Temperature-65.4) * h(Pressure.Vacuum- -5.6)  0.0097131
## h(Carb.Pressure1-122) * h(Temperature-65.4) * h(Pressure.Vacuum- -5.6)  0.0096896
##
## Selected 13 of 60 terms, and 14 of 31 predictors (nprune=13)
## Termination condition: Reached nk 63
## Importance: Pressure.Vacuum, Balling.Lvl, Temperature, PSC.Fill, ...
## Number of terms at each degree of interaction: 1 2 6 4
## GCV 0.01570104    RSS 3.849287    GRSq 0.501196    RSq 0.5950793
```

```
## Multivariate Adaptive Regression Spline
##
## 304 samples
##  31 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 273, 274, 274, 274, 274, 272, ...
## Resampling results across tuning parameters:
##
##   degree  nprune  RMSE       Rsquared    MAE
##   1        2      0.1685256  0.09927497  0.1300827
##   1        3      0.1689133  0.10501557  0.1317391
##   1        4      0.1716498  0.12650667  0.1318485
##   1        5      0.1692812  0.14657503  0.1298966
##   1        6      0.1680995  0.15075507  0.1305044
##   1        7      0.1704098  0.15503098  0.1320919
##   1        8      0.1702538  0.16731100  0.1302664
##   1        9      0.1676817  0.19780853  0.1281429
##   1       10      0.1675964  0.20788369  0.1287754
##   1       11      0.1681133  0.21284723  0.1274386
##   1       12      0.1681199  0.22739473  0.1266981
##   1       13      0.1696894  0.22010134  0.1261041
##   1       14      0.1688847  0.23160383  0.1257149
##   1       15      0.1687515  0.23932004  0.1248947
##   1       16      0.1683720  0.24290254  0.1242177
##   2        2      0.1693244  0.09237852  0.1310773
##   2        3      0.1685256  0.13267893  0.1298259
##   2        4      0.1731696  0.13026418  0.1306126
##   2        5      0.1701072  0.16143736  0.1269893
##   2        6      0.1624686  0.22176626  0.1242839
##   2        7      0.1633530  0.24359787  0.1238262
##   2        8      0.1615470  0.25536698  0.1217300
##   2        9      0.1647911  0.25062985  0.1236450
##   2       10      0.1681701  0.26468367  0.1235193
##   2       11      0.1647464  0.29185313  0.1217056
##   2       12      0.1644856  0.29472454  0.1220472
##   2       13      0.1631127  0.31342422  0.1210661
##   2       14      0.1624972  0.31621034  0.1213444
##   2       15      0.1591930  0.34804645  0.1198542
##   2       16      0.1717104  0.32748688  0.1226411
##   3        2      0.1703662  0.07775652  0.1319217
##   3        3      0.1731491  0.09424031  0.1330723
##   3        4      0.1750002  0.14022411  0.1324846
##   3        5      0.1713583  0.17502491  0.1296243
##   3        6      0.1680932  0.20021872  0.1268185
##   3        7      0.1671317  0.21190263  0.1266351
##   3        8      0.1652317  0.21957770  0.1248171
##   3        9      0.1666785  0.22028886  0.1257937
##   3       10      0.1647561  0.24084981  0.1234402
##   3       11      0.1621142  0.25907986  0.1211188
##   3       12      0.1606675  0.27790913  0.1197932
##   3       13      0.1578045  0.29187065  0.1175268
##   3       14      0.1581894  0.29582650  0.1176273
##   3       15      0.1607894  0.29377899  0.1188196
##   3       16      0.1579988  0.31719790  0.1161344
##   4        2      0.1696436  0.10057402  0.1309797
##   4        3      0.1739713  0.10151817  0.1338365
##   4        4      0.1689381  0.15072780  0.1299199
##   4        5      0.1717728  0.14427940  0.1312505
##   4        6      0.1719565  0.16088922  0.1298910
##   4        7      0.1720690  0.16279972  0.1289200
##   4        8      0.1688281  0.19521159  0.1261999
##   4        9      0.1699864  0.20161356  0.1251317
##   4       10      0.1642892  0.23820374  0.1216996
```

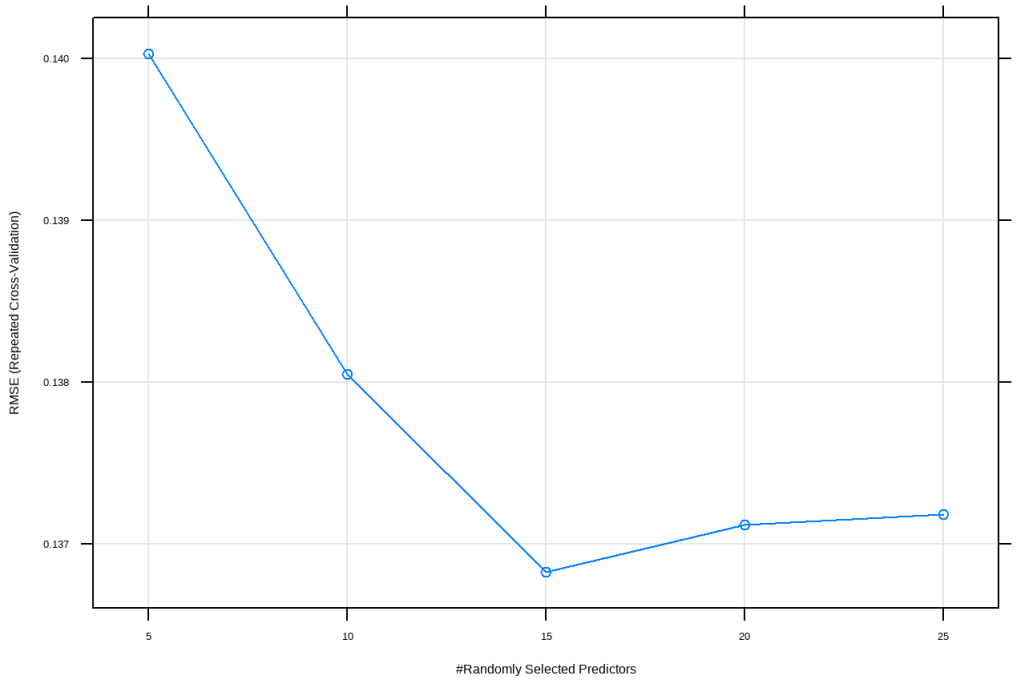
```
##   4       11      0.1674324  0.22520054  0.1221726
##   4       12      0.1664408  0.24413035  0.1205045
##   4       13      0.1636378  0.26657727  0.1190668
##   4       14      0.1670930  0.25788084  0.1200022
##   4       15      0.1691585  0.25925445  0.1199216
##   4       16      0.1726638  0.25913643  0.1195906
##
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were nprune = 13 and degree = 3.
```

Support Vector Machines



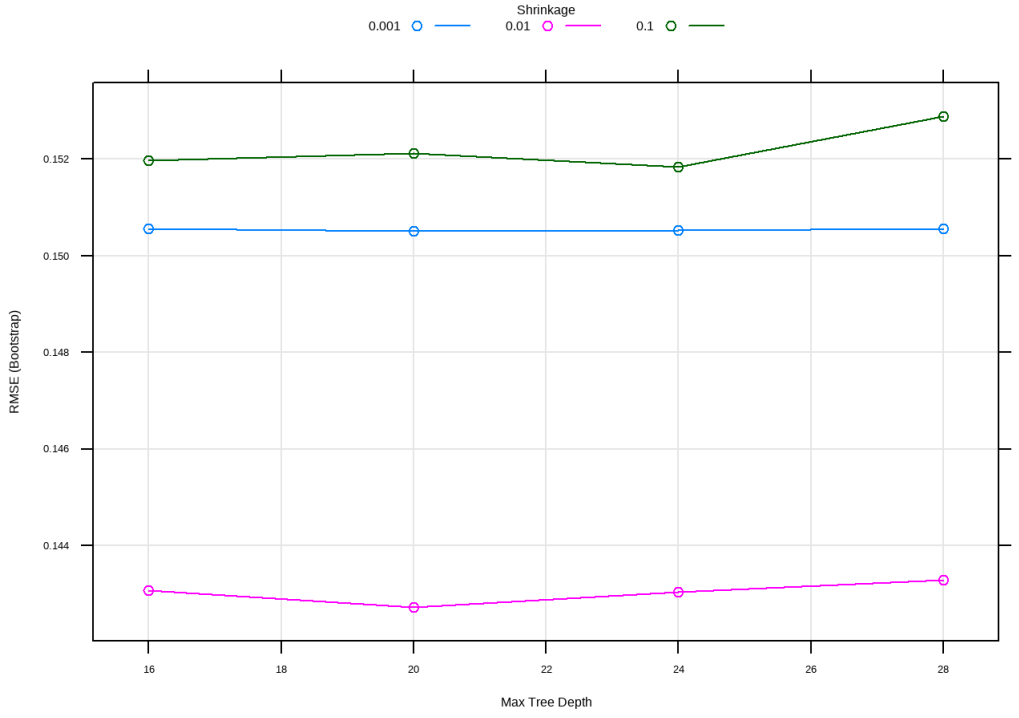
```
## Support Vector Machines with Radial Basis Function Kernel
##
## 304 samples
##  31 predictor
##
## Pre-processing: centered (31), scaled (31)
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 273, 274, 274, 274, 274, 272, ...
## Resampling results across tuning parameters:
##
##   C       RMSE       Rsquared   MAE
##     0.25  0.1610312  0.1997524  0.1167249
##     0.50  0.1566301  0.2381001  0.1129563
##     1.00  0.1500734  0.2949293  0.1090715
##     2.00  0.1436185  0.3476952  0.1040239
##     4.00  0.1403574  0.3762256  0.1019496
##     8.00  0.1413501  0.3750022  0.1036112
##    16.00  0.1465884  0.3495667  0.1085126
##    32.00  0.1543274  0.3213800  0.1144622
##    64.00  0.1579529  0.3203842  0.1170477
##   128.00  0.1601950  0.3138925  0.1190984
##
## Tuning parameter 'sigma' was held constant at a value of 0.02269464
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were sigma = 0.02269464 and C = 4.
```

Random Forest



```
## Random Forest
##
## 304 samples
##  31 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 273, 274, 274, 274, 274, 272, ...
## Resampling results across tuning parameters:
##
##   mtry  RMSE       Rsquared   MAE
##    5    0.1400274  0.4121574  0.10180045
##   10    0.1380463  0.4139678  0.09922102
##   15    0.1368251  0.4209485  0.09788084
##   20    0.1371173  0.4156262  0.09707782
##   25    0.1371805  0.4136596  0.09661835
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 15.
```

Gradient Boosting Machines



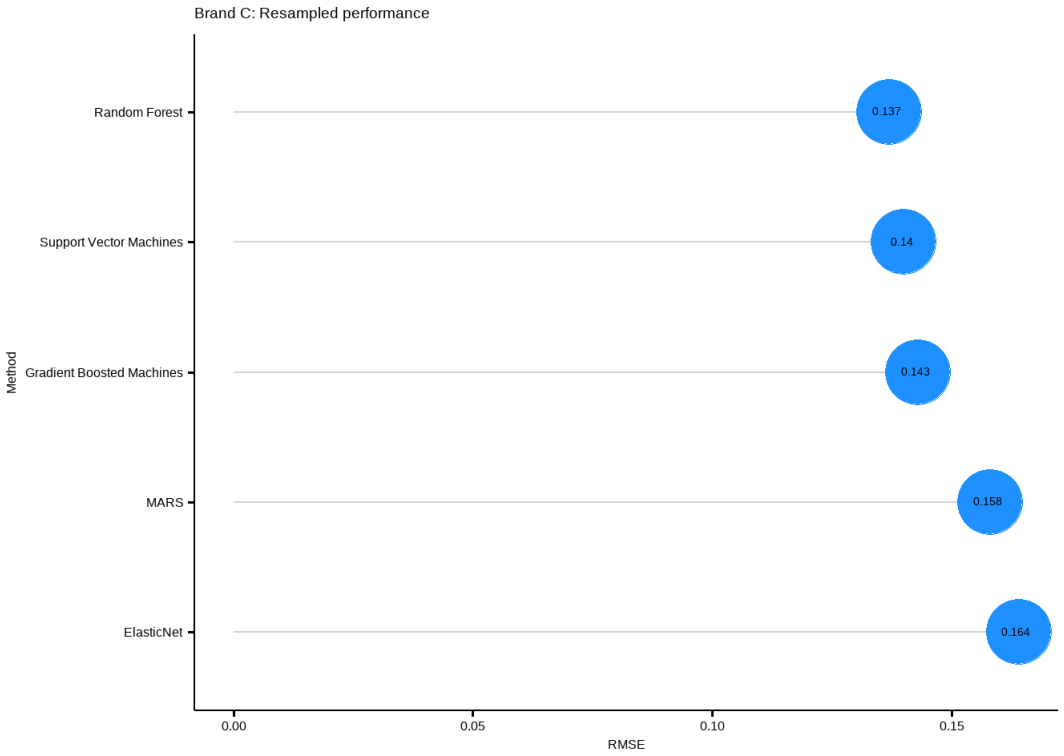
```
## Stochastic Gradient Boosting
##
## 304 samples
##  31 predictor
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 304, 304, 304, 304, 304, 304, ...
## Resampling results across tuning parameters:
##
##   shrinkage  interaction.depth  RMSE       Rsquared   MAE
##   0.001      16                 0.1505535  0.3105752  0.1158195
##   0.001      20                 0.1505114  0.3117369  0.1158472
##   0.001      24                 0.1505260  0.3110164  0.1157932
##   0.001      28                 0.1505606  0.3104246  0.1158420
##   0.010      16                 0.1430680  0.3289688  0.1060503
##   0.010      20                 0.1427292  0.3313323  0.1059183
##   0.010      24                 0.1430403  0.3291289  0.1058133
##   0.010      28                 0.1432944  0.3273150  0.1061345
##   0.100      16                 0.1519711  0.2843482  0.1130532
##   0.100      20                 0.1521117  0.2825452  0.1125439
##   0.100      24                 0.1518385  0.2856228  0.1124837
##   0.100      28                 0.1528770  0.2779153  0.1129275
##
## Tuning parameter 'n.trees' was held constant at a value of 1000
##
## Tuning parameter 'n.minobsinnode' was held constant at a value of 10
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were n.trees = 1000, interaction.depth
##  = 20, shrinkage = 0.01 and n.minobsinnode = 10.
```

Brand C: Model Selection

The **Random Forest** model achieves the lowest resampled RMSE on brand **C** among the five models with score of **0.139**.

Brand C: Resampled performance

Method	RMSE
Random Forest	0.137
Support Vector Machines	0.140
Gradient Boosted Machines	0.143
MARS	0.158
ElasticNet	0.164



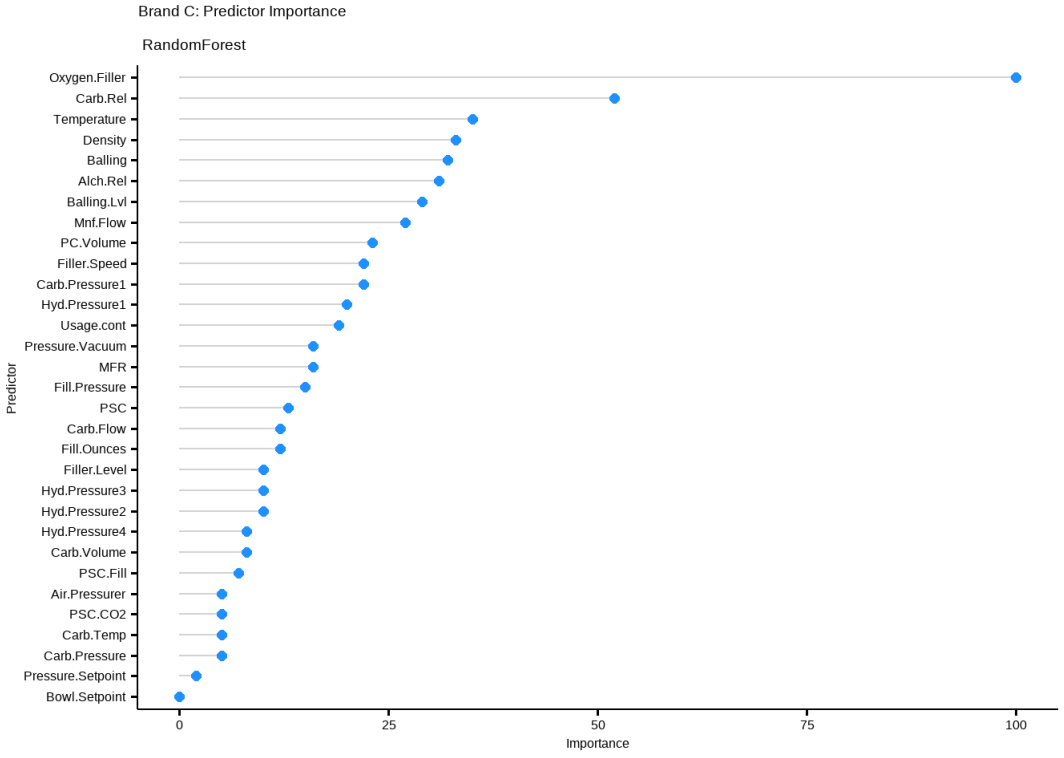
Brand C: Model Variable Importance

Brand C: Predictor Importance RandomForest

Predictor	Importance
Oxygen.Filler	100
Carb.Rel	52
Temperature	35
Density	33
Balling	32
Alch.Rel	31
Balling.Lvl	29
Mnf.Flow	27
PC.Volume	23
Carb.Pressure1	22
Filler.Speed	22
Hyd.Pressure1	20
Usage.cont	19
MFR	16
Pressure.Vacuum	16
Fill.Pressure	15
PSC	13
Fill.Ounces	12
Carb.Flow	12
Hyd.Pressure2	10
Hyd.Pressure3	10

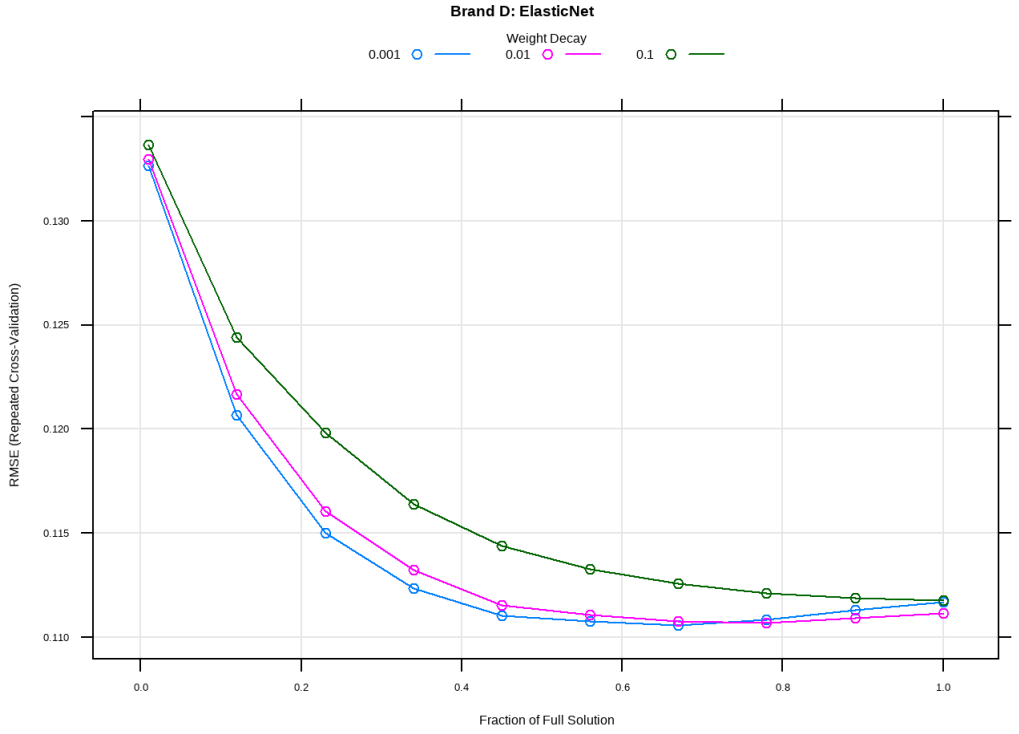
Brand C: Predictor Importance RandomForest

Predictor	Importance
Filler.Level	10
Carb.Volume	8
Hyd.Pressure4	8
PSC.Fill	7
Carb.Pressure	5
Carb.Temp	5
PSC.CO2	5
Air.Pressurer	5
Pressure.Setpoint	2
Bowl.Setpoint	0



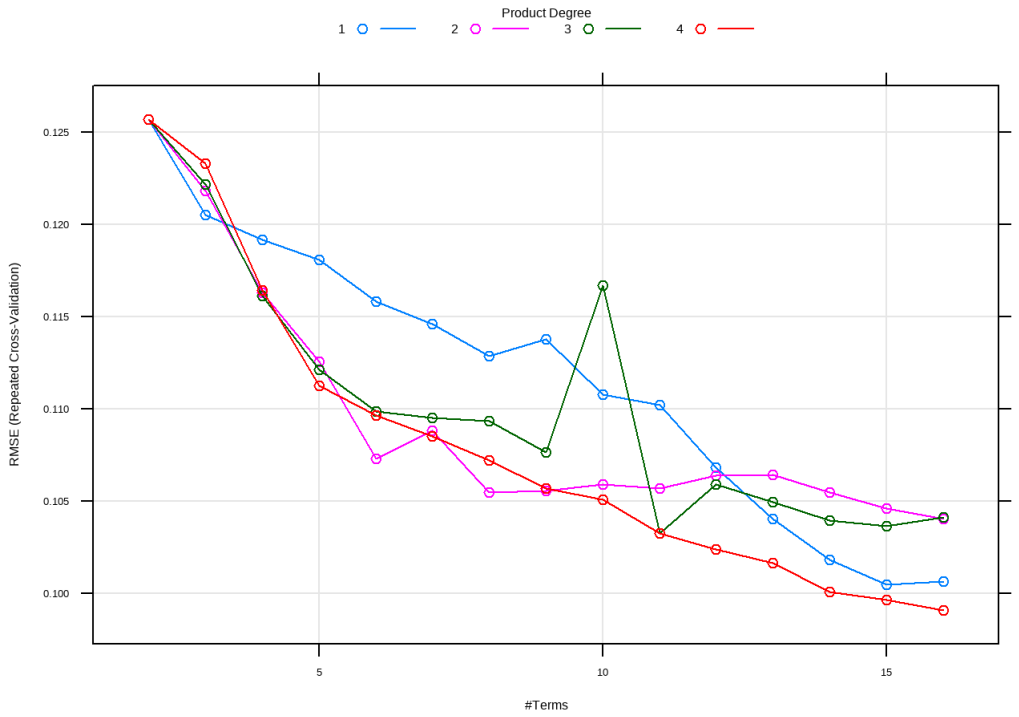
Brand D: Model Fitting

ElasticNet



```
## Elasticnet
##
## 615 samples
##  31 predictor
##
## Pre-processing: centered (31), scaled (31)
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 553, 553, 553, 554, 554, 552, ...
## Resampling results across tuning parameters:
##
##   lambda  fraction  RMSE       Rsquared   MAE
##   0.001   0.01      0.1326718  0.1493177  0.11000568
##   0.001   0.12      0.1206501  0.2364222  0.09859062
##   0.001   0.23      0.1149854  0.2984483  0.09247676
##   0.001   0.34      0.1123319  0.3208445  0.08925355
##   0.001   0.45      0.1110254  0.3329568  0.08712594
##   0.001   0.56      0.1107630  0.3354885  0.08643638
##   0.001   0.67      0.1105534  0.3389712  0.08609762
##   0.001   0.78      0.1108344  0.3380296  0.08617578
##   0.001   0.89      0.1112906  0.3354089  0.08641719
##   0.001   1.00      0.1117016  0.3332009  0.08670807
##   0.010   0.01      0.1329530  0.1493177  0.11022453
##   0.010   0.12      0.1216637  0.2219483  0.09966211
##   0.010   0.23      0.1160442  0.2889437  0.09367297
##   0.010   0.34      0.1132429  0.3115818  0.09033946
##   0.010   0.45      0.1115225  0.3283875  0.08809215
##   0.010   0.56      0.1110514  0.3320583  0.08694541
##   0.010   0.67      0.1107540  0.3357085  0.08643330
##   0.010   0.78      0.1106864  0.3374823  0.08626705
##   0.010   0.89      0.1108966  0.3367002  0.08634135
##   0.010   1.00      0.1111329  0.3355277  0.08647521
##   0.100   0.01      0.1336659  0.1493177  0.11076746
##   0.100   0.12      0.1244058  0.1801047  0.10250259
##   0.100   0.23      0.1198146  0.2439564  0.09764091
##   0.100   0.34      0.1163828  0.2825878  0.09395030
##   0.100   0.45      0.1143809  0.2995706  0.09156268
##   0.100   0.56      0.1132557  0.3082277  0.09005417
##   0.100   0.67      0.1125596  0.3151765  0.08898827
##   0.100   0.78      0.1121207  0.3197835  0.08834649
##   0.100   0.89      0.1118600  0.3226618  0.08794377
##   0.100   1.00      0.1117784  0.3236748  0.08779732
##
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were fraction = 0.67 and lambda = 0.001.
```

Multivariate Regression Adaptive Splines

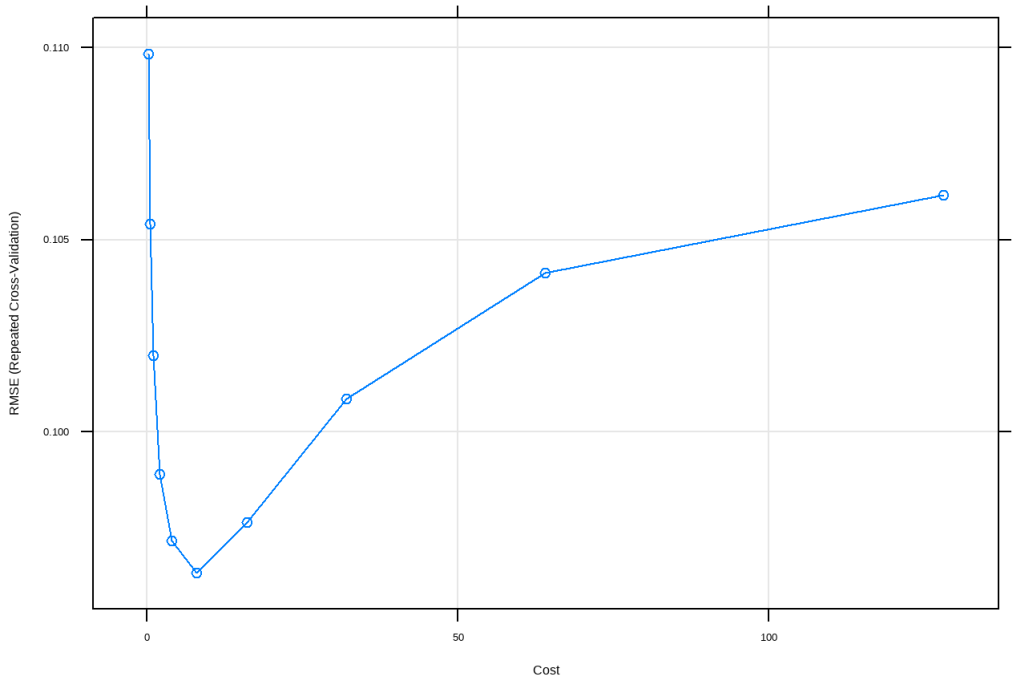



```
## Call: earth(x=data.frame[615,31], y=c(8.4,8.46,8.4,...), keepxy=TRUE, degree=4,
##             nprune=16)
##
##                                                          coefficients
## (Intercept)                                                 8.6431903
## h(4.8-Mnf.Flow)                                             0.0025236
## h(1.3-Density)                                             -0.3865072
## h(Density-1.3)                                             -0.2999558
## h(Carb.Rel-5.62)                                           -0.8316391
## h(Mnf.Flow-4.8) * h(Hyd.Pressure3-25.6)                     0.0001203
## h(Mnf.Flow-4.8) * h(Hyd.Pressure3-34.4)                    -0.0002137
## h(4.8-Mnf.Flow) * h(Pressure.Vacuum- -4.8)                 -0.0053392
## h(4.8-Mnf.Flow) * h(-4.8-Pressure.Vacuum)                  -0.0015907
## h(Temperature-64.6) * h(Density-1.3)                       -0.0359805
## h(Mnf.Flow-4.8) * h(17.8-Hyd.Pressure1) * h(Hyd.Pressure3-34.4)  0.0000091
## h(4.8-Mnf.Flow) * h(Filler.Speed-3998) * h(Pressure.Vacuum- -4.8)  0.0002087
## h(Mnf.Flow-4.8) * h(Usage.cont-20.4) * h(Carb.Rel-5.68)     0.0036719
## h(Mnf.Flow-4.8) * h(Usage.cont-20.4) * h(5.68-Carb.Rel)    -0.0016907
## h(4.8-Mnf.Flow) * h(Carb.Flow-3066) * h(Pressure.Vacuum- -4.8)  0.0000156
## h(4.8-Mnf.Flow) * h(3066-Carb.Flow) * h(Pressure.Vacuum- -4.8)  0.0000014
##
## Selected 16 of 61 terms, and 10 of 31 predictors (nprune=16)
## Termination condition: Reached nk 63
## Importance: Mnf.Flow, Pressure.Vacuum, Hyd.Pressure3, Usage.cont, Carb.Rel, ...
## Number of terms at each degree of interaction: 1 4 5 6
## GCV 0.007994328    RSS 4.320216    GRSq 0.5644023    RSq 0.6159857
```

```
## Multivariate Adaptive Regression Spline
##
## 615 samples
##  31 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 553, 553, 553, 554, 554, 552, ...
## Resampling results across tuning parameters:
##
##   degree  nprune  RMSE        Rsquared   MAE
##   1        2      0.12568386  0.1459621  0.10294369
##   1        3      0.12052375  0.2194563  0.09682545
##   1        4      0.11918639  0.2366935  0.09603555
##   1        5      0.11807813  0.2677916  0.09433780
##   1        6      0.11585539  0.2835943  0.09316252
##   1        7      0.11460858  0.3097197  0.09109369
##   1        8      0.11288526  0.3233605  0.08927234
##   1        9      0.11377935  0.3373861  0.08870644
##   1       10      0.11079433  0.3711457  0.08552829
##   1       11      0.11021686  0.3840006  0.08415124
##   1       12      0.10684912  0.4079204  0.08238088
##   1       13      0.10405722  0.4272698  0.08105452
##   1       14      0.10185889  0.4480770  0.07980938
##   1       15      0.10051819  0.4610666  0.07873152
##   1       16      0.10067823  0.4616079  0.07894967
##   2        2      0.12568386  0.1459621  0.10294369
##   2        3      0.12183645  0.1997265  0.09963625
##   2        4      0.11632280  0.2774404  0.09289400
##   2        5      0.11258597  0.3270387  0.08791474
##   2        6      0.10730735  0.3778700  0.08455789
##   2        7      0.10883534  0.3795673  0.08413130
##   2        8      0.10550446  0.4065302  0.08229607
##   2        9      0.10558387  0.4138697  0.08224906
##   2       10      0.10595016  0.4153583  0.08202908
##   2       11      0.10570924  0.4239886  0.08121160
##   2       12      0.10639228  0.4250948  0.08097206
##   2       13      0.10645593  0.4295891  0.08063537
##   2       14      0.10550298  0.4418366  0.07999898
##   2       15      0.10460999  0.4521742  0.07899337
##   2       16      0.10407054  0.4562597  0.07890501
##   3        2      0.12568386  0.1459621  0.10294369
##   3        3      0.12219960  0.1904007  0.09957487
##   3        4      0.11615998  0.2721740  0.09395333
##   3        5      0.11213234  0.3315950  0.08843170
##   3        6      0.10988510  0.3577819  0.08624080
##   3        7      0.10953568  0.3641387  0.08541644
##   3        8      0.10938047  0.3751489  0.08490157
##   3        9      0.10765212  0.3943247  0.08342802
##   3       10      0.11668659  0.3991147  0.08383107
##   3       11      0.10327966  0.4411157  0.08036230
##   3       12      0.10593832  0.4480605  0.07991445
##   3       13      0.10495251  0.4581704  0.07920561
##   3       14      0.10396661  0.4732596  0.07804429
##   3       15      0.10366129  0.4764838  0.07760869
##   3       16      0.10414194  0.4738000  0.07790820
##   4        2      0.12568386  0.1459621  0.10294369
##   4        3      0.12329152  0.1786597  0.10002445
##   4        4      0.11642559  0.2699720  0.09400693
##   4        5      0.11128578  0.3347299  0.08829794
##   4        6      0.10965555  0.3549411  0.08642219
##   4        7      0.10852786  0.3730125  0.08500347
##   4        8      0.10724070  0.3934362  0.08357648
##   4        9      0.10572893  0.4113085  0.08167721
##   4       10      0.10510021  0.4176637  0.08111383
```

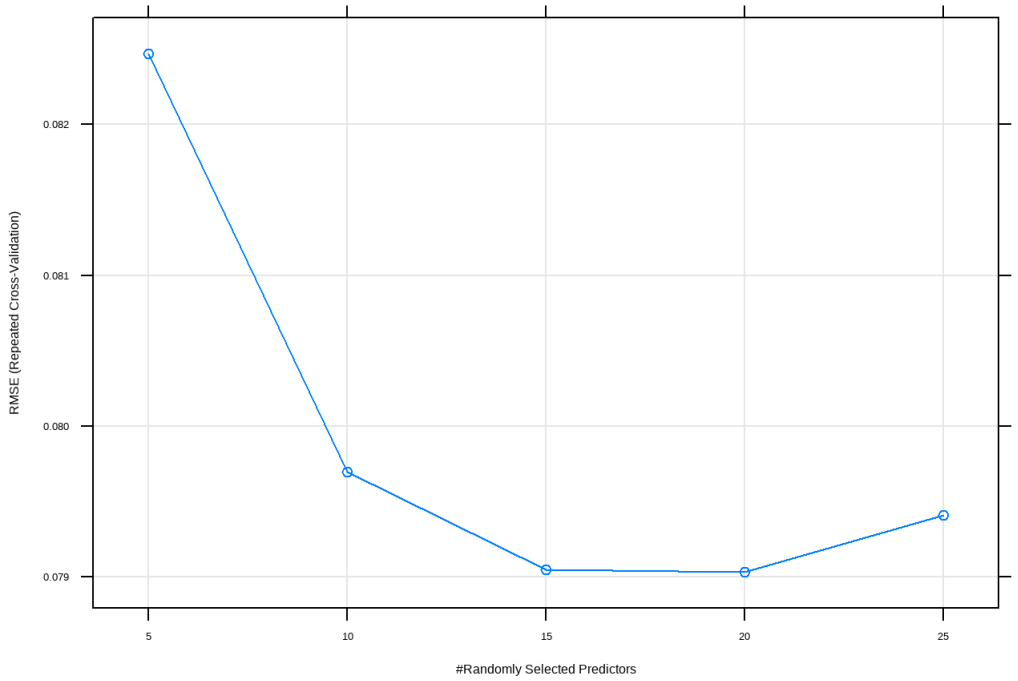
```
##   4       11      0.10327523  0.4352704  0.07985747
##   4       12      0.10239769  0.4452089  0.07935228
##   4       13      0.10167293  0.4542799  0.07867163
##   4       14      0.10010623  0.4718498  0.07775081
##   4       15      0.09968751  0.4824231  0.07712302
##   4       16      0.09912653  0.4904741  0.07679830
##
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were nprune = 16 and degree = 4.
```

Support Vector Machines



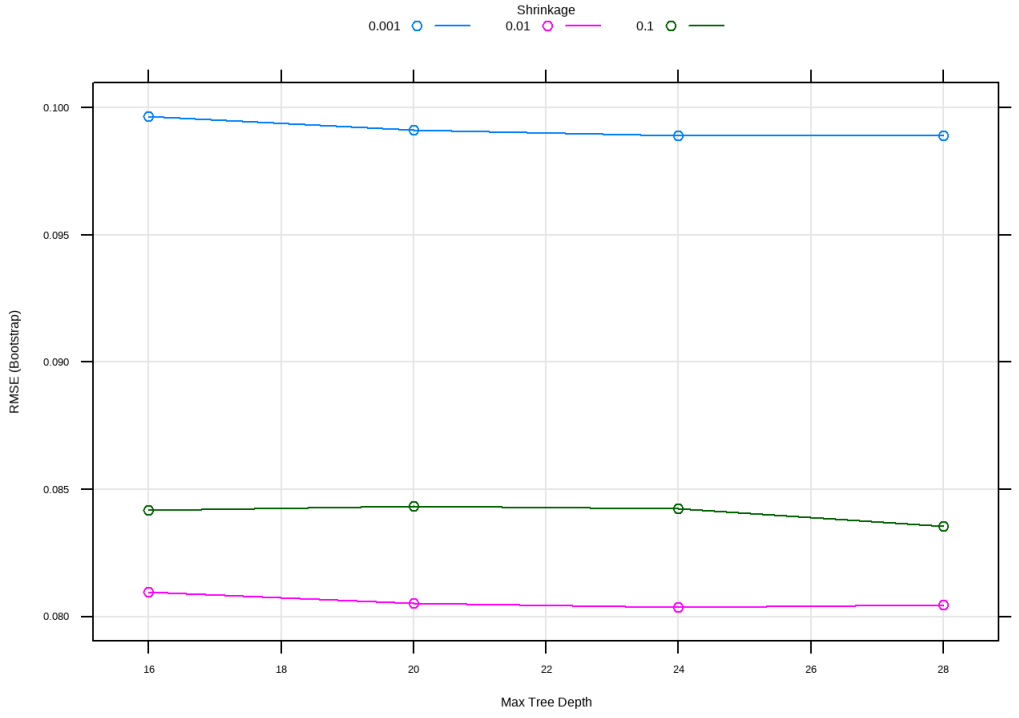
```
## Support Vector Machines with Radial Basis Function Kernel
##
## 615 samples
##  31 predictor
##
## Pre-processing: centered (31), scaled (31)
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 553, 553, 553, 554, 554, 552, ...
## Resampling results across tuning parameters:
##
##   C       RMSE        Rsquared   MAE
##     0.25  0.10983553  0.3669651  0.08417503
##     0.50  0.10540691  0.4069904  0.07966251
##     1.00  0.10198899  0.4392020  0.07662828
##     2.00  0.09889386  0.4705352  0.07366675
##     4.00  0.09716361  0.4887415  0.07169873
##     8.00  0.09631693  0.5002906  0.07102722
##    16.00  0.09763825  0.4940896  0.07236410
##    32.00  0.10084563  0.4752858  0.07511024
##    64.00  0.10413598  0.4572552  0.07794216
##   128.00  0.10615475  0.4487700  0.07982054
##
## Tuning parameter 'sigma' was held constant at a value of 0.02384996
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were sigma = 0.02384996 and C = 8.
```

Random Forest



```
## Random Forest
##
## 615 samples
##  31 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 553, 553, 553, 554, 554, 552, ...
## Resampling results across tuning parameters:
##
##   mtry  RMSE        Rsquared   MAE
##    5    0.08246790  0.6823233  0.06306530
##   10    0.07969753  0.6886952  0.06090016
##   15    0.07904839  0.6860032  0.06018081
##   20    0.07903369  0.6797659  0.05996094
##   25    0.07940961  0.6731217  0.05987187
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 20.
```

Gradient Boosting Machines



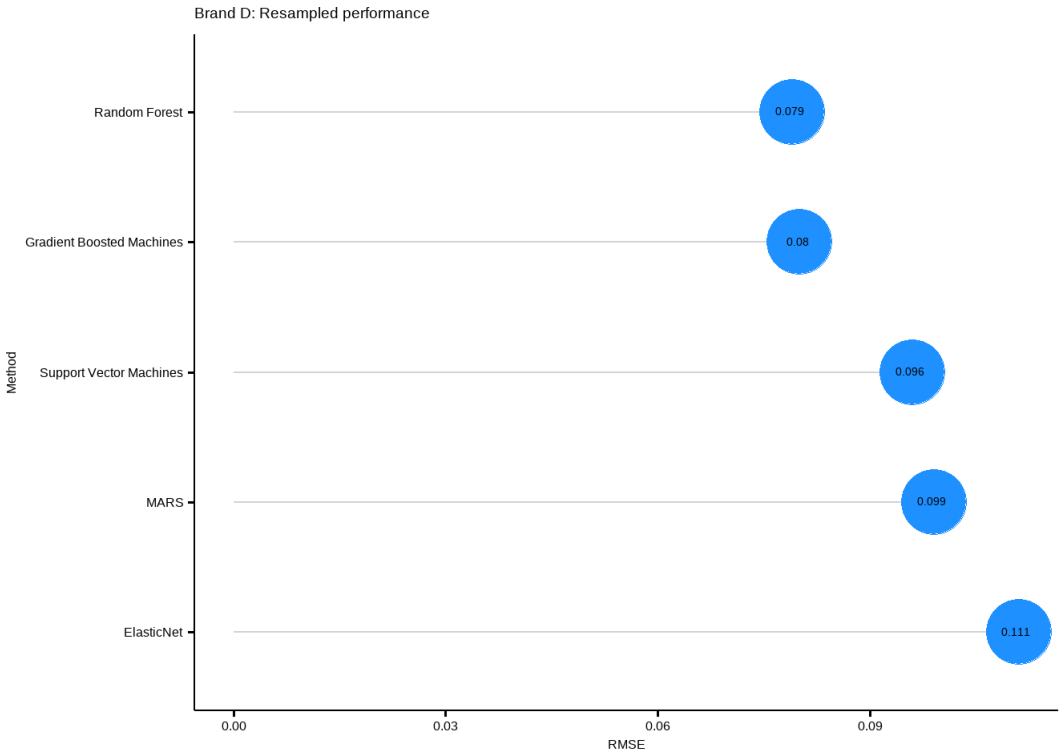
```
## Stochastic Gradient Boosting
##
## 615 samples
##  31 predictor
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 615, 615, 615, 615, 615, 615, ...
## Resampling results across tuning parameters:
##
##   shrinkage  interaction.depth  RMSE        Rsquared   MAE
##   0.001      16                 0.09965641  0.5916244  0.08004908
##   0.001      20                 0.09913682  0.5961648  0.07955028
##   0.001      24                 0.09892325  0.5975963  0.07936543
##   0.001      28                 0.09890768  0.5986099  0.07936132
##   0.010      16                 0.08096124  0.6493832  0.06147210
##   0.010      20                 0.08051882  0.6533230  0.06118585
##   0.010      24                 0.08037763  0.6548980  0.06103837
##   0.010      28                 0.08046348  0.6537886  0.06110303
##   0.100      16                 0.08418563  0.6139790  0.06410711
##   0.100      20                 0.08432291  0.6130808  0.06407920
##   0.100      24                 0.08424446  0.6145785  0.06412831
##   0.100      28                 0.08356151  0.6199137  0.06327144
##
## Tuning parameter 'n.trees' was held constant at a value of 1000
##
## Tuning parameter 'n.minobsinnode' was held constant at a value of 10
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were n.trees = 1000, interaction.depth
##  = 24, shrinkage = 0.01 and n.minobsinnode = 10.
```

Brand D: Model Selection

The **Random Forest** model achieves the lowest resampled RMSE on brand **D** among the five models with score of **0.079**.

Brand D: Resampled performance

Method	RMSE
Random Forest	0.079
Gradient Boosted Machines	0.080
Support Vector Machines	0.096
MARS	0.099
ElasticNet	0.111

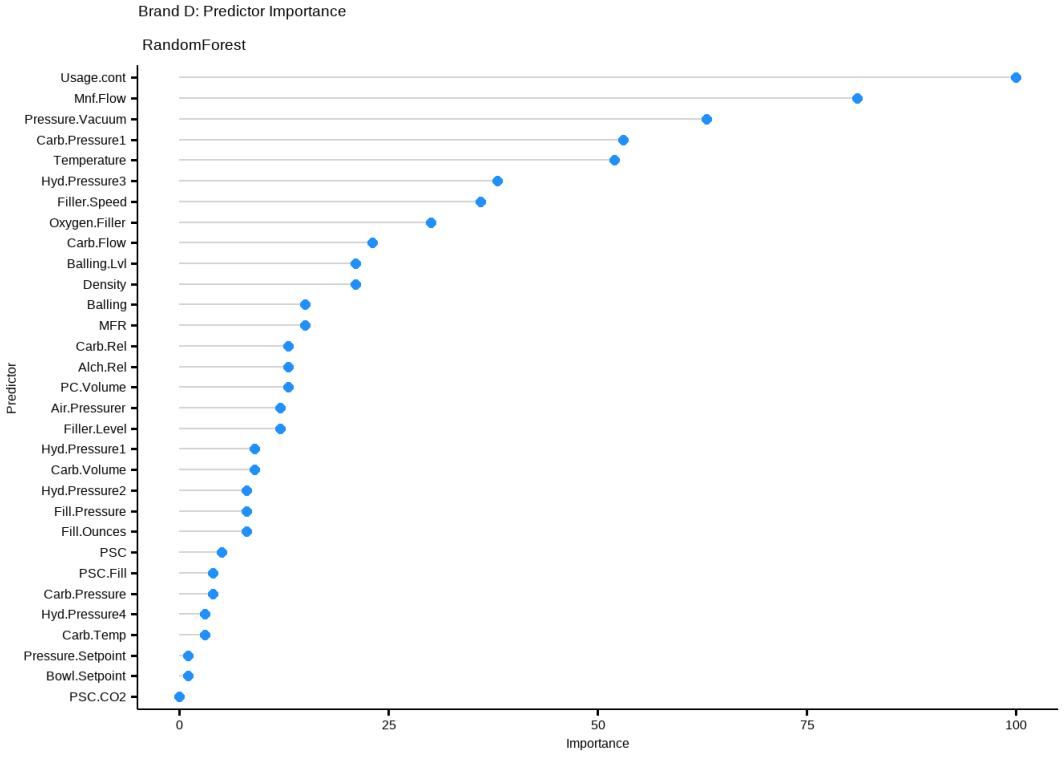


Brand D: Model Variable Importance

Brand D: Predictor Importance RandomForest

Predictor	Importance
Usage.cont	100
Mnf.Flow	81
Pressure.Vacuum	63
Carb.Pressure1	53
Temperature	52
Hyd.Pressure3	38
Filler.Speed	36
Oxygen.Filler	30
Carb.Flow	23
Density	21
Balling.Lvl	21
MFR	15
Balling	15
PC.Volume	13
Alch.Rel	13
Carb.Rel	13
Filler.Level	12
Air.Pressurer	12
Carb.Volume	9
Hyd.Pressure1	9
Fill.Ounces	8

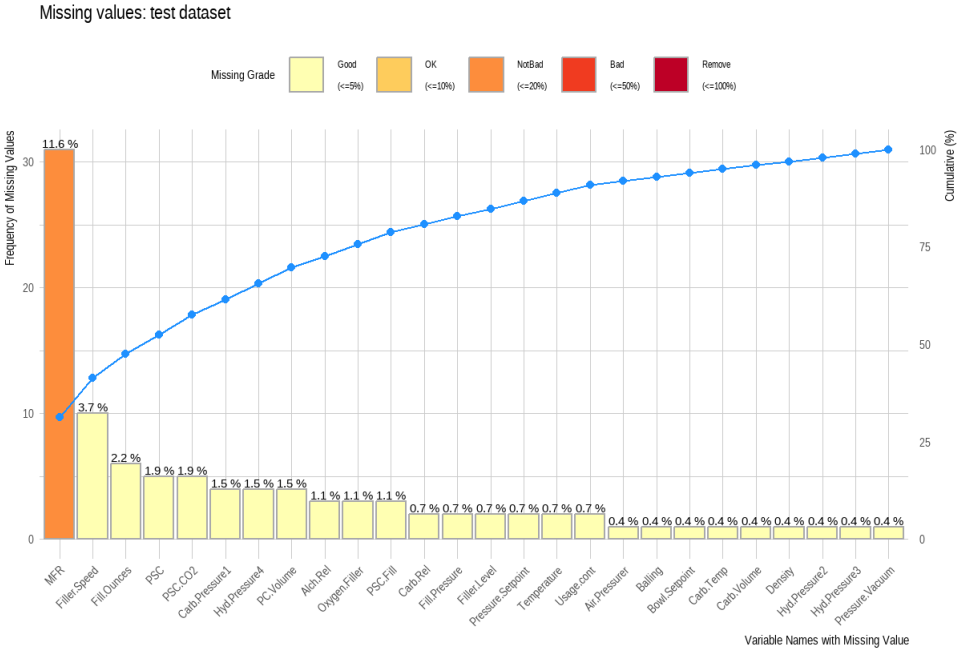
Brand D: Predictor Importance RandomForest	
Predictor	**Importance**
Fill.Pressure	8
Hyd.Pressure2	8
PSC	5
Carb.Pressure	4
PSC.Fill	4
Carb.Temp	3
Hyd.Pressure4	3
Bowl.Setpoint	1
Pressure.Setpoint	1
PSC.CO2	0



Model Prediction

The test data contains a fair number of missing values across several variables. As with the training data, we impute the missing values with the Bagged Trees method.

The final test data ready for prediction is summarized below.

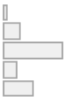
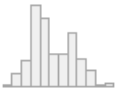
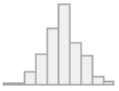
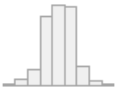
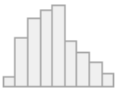
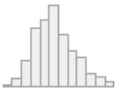
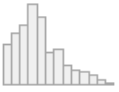
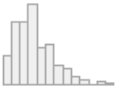
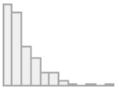
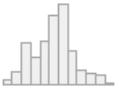


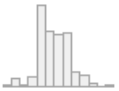
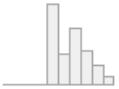
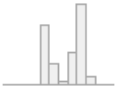
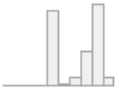
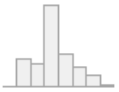
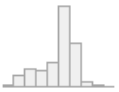
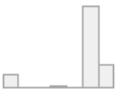
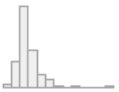
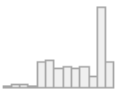
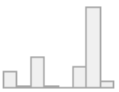
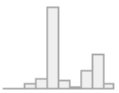
variables	missing_count	missing_percent
MFR	31	11.6
Filler.Speed	10	3.7
Fill.Ounces	6	2.2
PSC	5	1.9
PSC.CO2	5	1.9
PC.Volume	4	1.5
Carb.Pressure1	4	1.5
Hyd.Pressure4	4	1.5
PSC.Fill	3	1.1
Oxygen.Filler	3	1.1
Alch.Rel	3	1.1
Fill.Pressure	2	0.7
Filler.Level	2	0.7
Temperature	2	0.7
Usage.cont	2	0.7
Pressure.Setpoint	2	0.7
Carb.Rel	2	0.7
Carb.Volume	1	0.4
Carb.Temp	1	0.4
Hyd.Pressure2	1	0.4
Hyd.Pressure3	1	0.4
Density	1	0.4
Balling	1	0.4
Pressure.Vacuum	1	0.4
Bowl.Setpoint	1	0.4
Air.Pressurer	1	0.4

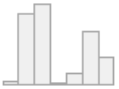
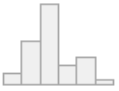
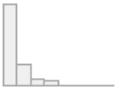
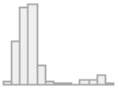
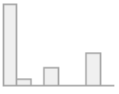
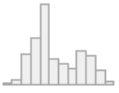
Data Frame Summary

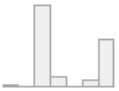
testData

Dimensions: 267 x 32
Duplicates: 0

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	Brand.Code [character]	1. (Empty string) 2. A 3. B 4. C 5. D	8 (3.0%) 35 (13.1%) 129 (48.3%) 31 (11.6%) 64 (24.0%)		267 (100.0%)	0 (0.0%)
2	Carb.Volume [numeric]	Mean (sd) : 5.4 (0.1) min ≤ med ≤ max: 5.1 ≤ 5.3 ≤ 5.7 IQR (CV) : 0.2 (0)	72 distinct values		267 (100.0%)	0 (0.0%)
3	Fill.Ounces [numeric]	Mean (sd) : 24 (0.1) min ≤ med ≤ max: 23.7 ≤ 24 ≤ 24.2 IQR (CV) : 0.1 (0)	58 distinct values		267 (100.0%)	0 (0.0%)
4	PC.Volume [numeric]	Mean (sd) : 0.3 (0.1) min ≤ med ≤ max: 0.1 ≤ 0.3 ≤ 0.5 IQR (CV) : 0.1 (0.2)	187 distinct values		267 (100.0%)	0 (0.0%)
5	Carb.Pressure [numeric]	Mean (sd) : 68.3 (3.9) min ≤ med ≤ max: 60.2 ≤ 68 ≤ 77.6 IQR (CV) : 5.3 (0.1)	76 distinct values		267 (100.0%)	0 (0.0%)
6	Carb.Temp [numeric]	Mean (sd) : 141.3 (4.3) min ≤ med ≤ max: 130 ≤ 140.8 ≤ 154 IQR (CV) : 5.5 (0)	89 distinct values		267 (100.0%)	0 (0.0%)
7	PSC [numeric]	Mean (sd) : 0.1 (0.1) min ≤ med ≤ max: 0 ≤ 0.1 ≤ 0.2 IQR (CV) : 0.1 (0.6)	101 distinct values		267 (100.0%)	0 (0.0%)
8	PSC.Fill [numeric]	Mean (sd) : 0.2 (0.1) min ≤ med ≤ max: 0 ≤ 0.2 ≤ 0.6 IQR (CV) : 0.1 (0.6)	31 distinct values		267 (100.0%)	0 (0.0%)
9	PSC.CO2 [numeric]	Mean (sd) : 0.1 (0) min ≤ med ≤ max: 0 ≤ 0 ≤ 0.2 IQR (CV) : 0 (0.7)	16 distinct values		267 (100.0%)	0 (0.0%)
10	Mnf.Flow [numeric]	Mean (sd) : 21 (117.8) min ≤ med ≤ max: -100.2 ≤ 0.2 ≤ 220.4 IQR (CV) : 241.3 (5.6)	114 distinct values		267 (100.0%)	0 (0.0%)
11	Carb.Pressure1 [numeric]	Mean (sd) : 123 (4.4) min ≤ med ≤ max: 113 ≤ 123.4 ≤ 136 IQR (CV) : 5.4 (0)	91 distinct values		267 (100.0%)	0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
12	Fill.Pressure [numeric]	Mean (sd) : 48.1 (3.4) min ≤ med ≤ max: 37.8 ≤ 47.8 ≤ 60.2 IQR (CV) : 4.2 (0.1)	63 distinct values		267 (100.0%)	0 (0.0%)
13	Hyd.Pressure1 [numeric]	Mean (sd) : 12 (13.5) min ≤ med ≤ max: -50 ≤ 10.4 ≤ 50 IQR (CV) : 20.4 (1.1)	115 distinct values		267 (100.0%)	0 (0.0%)
14	Hyd.Pressure2 [numeric]	Mean (sd) : 20 (17.2) min ≤ med ≤ max: -50 ≤ 26.8 ≤ 61.4 IQR (CV) : 34.8 (0.9)	96 distinct values		267 (100.0%)	0 (0.0%)
15	Hyd.Pressure3 [numeric]	Mean (sd) : 19.5 (16.6) min ≤ med ≤ max: -50 ≤ 27.6 ≤ 49.2 IQR (CV) : 33 (0.8)	90 distinct values		267 (100.0%)	0 (0.0%)
16	Hyd.Pressure4 [numeric]	Mean (sd) : 98 (13.9) min ≤ med ≤ max: 68 ≤ 98 ≤ 140 IQR (CV) : 14 (0.1)	38 distinct values		267 (100.0%)	0 (0.0%)
17	Filler.Level [numeric]	Mean (sd) : 110.3 (15.4) min ≤ med ≤ max: 69.2 ≤ 118.4 ≤ 153.2 IQR (CV) : 19.6 (0.1)	109 distinct values		267 (100.0%)	0 (0.0%)
18	Filler.Speed [numeric]	Mean (sd) : 3588.7 (901) min ≤ med ≤ max: 1006 ≤ 3952.8 ≤ 4020 IQR (CV) : 144 (0.3)	86 distinct values		267 (100.0%)	0 (0.0%)
19	Temperature [numeric]	Mean (sd) : 66.2 (1.7) min ≤ med ≤ max: 63.8 ≤ 65.8 ≤ 75.4 IQR (CV) : 1.2 (0)	38 distinct values		267 (100.0%)	0 (0.0%)
20	Usage.cont [numeric]	Mean (sd) : 20.9 (3) min ≤ med ≤ max: 12.9 ≤ 21.4 ≤ 24.6 IQR (CV) : 5.6 (0.1)	176 distinct values		267 (100.0%)	0 (0.0%)
21	Carb.Flow [integer]	Mean (sd) : 2408.6 (1161.4) min ≤ med ≤ max: 0 ≤ 3038 ≤ 3858 IQR (CV) : 2132 (0.5)	178 distinct values		267 (100.0%)	0 (0.0%)
22	Density [numeric]	Mean (sd) : 1.2 (0.4) min ≤ med ≤ max: 0.1 ≤ 1 ≤ 1.8 IQR (CV) : 0.7 (0.3)	54 distinct values		267 (100.0%)	0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
23	MFR [numeric]	Mean (sd) : 677.8 (115) min ≤ med ≤ max: 15.6 ≤ 723.1 ≤ 784.8 IQR (CV) : 35.2 (0.2)	167 distinct values		267 (100.0%)	0 (0.0%)
24	Balling [numeric]	Mean (sd) : 2.2 (0.9) min ≤ med ≤ max: 0.9 ≤ 1.6 ≤ 3.8 IQR (CV) : 1.7 (0.4)	83 distinct values		267 (100.0%)	0 (0.0%)
25	Pressure.Vacuum [numeric]	Mean (sd) : -5.2 (0.6) min ≤ med ≤ max: -6.4 ≤ -5.2 ≤ -3.6 IQR (CV) : 0.8 (-0.1)	16 distinct values		267 (100.0%)	0 (0.0%)
26	Oxygen.Filler [numeric]	Mean (sd) : 0 (0) min ≤ med ≤ max: 0 ≤ 0 ≤ 0.4 IQR (CV) : 0 (1.1)	152 distinct values		267 (100.0%)	0 (0.0%)
27	Bowl.Setpoint [numeric]	Mean (sd) : 109.6 (15) min ≤ med ≤ max: 70 ≤ 120 ≤ 130 IQR (CV) : 20 (0.1)	70.00 : 9 (3.4%) 80.00 : 15 (5.6%) 90.00 : 35 (13.1%) 100.00 : 12 (4.5%) 108.11 ! : 1 (0.4%) 110.00 : 49 (18.4%) 120.00 : 139 (52.1%) 130.00 : 7 (2.6%) ! rounded		267 (100.0%)	0 (0.0%)
28	Pressure.Setpoint [numeric]	Mean (sd) : 47.7 (2.1) min ≤ med ≤ max: 44 ≤ 46 ≤ 52 IQR (CV) : 4 (0)	44.00 : 9 (3.4%) 45.20 : 1 (0.4%) 45.43 ! : 1 (0.4%) 46.00 : 128 (47.9%) 47.78 ! : 1 (0.4%) 48.00 : 18 (6.7%) 50.00 : 106 (39.7%) 52.00 : 3 (1.1%) ! rounded		267 (100.0%)	0 (0.0%)
29	Air.Pressurer [numeric]	Mean (sd) : 142.8 (1.2) min ≤ med ≤ max: 141.2 ≤ 142.6 ≤ 147.2 IQR (CV) : 0.6 (0)	25 distinct values		267 (100.0%)	0 (0.0%)
30	Alch.Rel [numeric]	Mean (sd) : 6.9 (0.5) min ≤ med ≤ max: 6.4 ≤ 6.6 ≤ 7.8 IQR (CV) : 0.6 (0.1)	33 distinct values		267 (100.0%)	0 (0.0%)
31	Carb.Rel [numeric]	Mean (sd) : 5.4 (0.1) min ≤ med ≤ max: 5.2 ≤ 5.4 ≤ 5.7 IQR (CV) : 0.2 (0)	30 distinct values		267 (100.0%)	0 (0.0%)

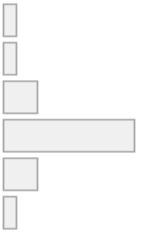
No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
32	Balling.Lvl [numeric]	Mean (sd) : 2.1 (0.9) min ≤ med ≤ max: 0 ≤ 1.5 ≤ 3.4 IQR (CV) : 1.7 (0.4)	53 distinct values		267 (100.0%)	0 (0.0%)

Generated by summarytools (https://github.com/dcomtois/summarytools) 1.0.1 (R (https://www.r-project.org/) version 4.1.0)
2023-04-04

Brand A

Test set prediction summary statistics for beverage brand A are provided below.

The predictions are available for viewing and for download in .csv format here (https://raw.githubusercontent.com/MauricioClaudio/DATA624-/main/predictions_BrandA.csv).

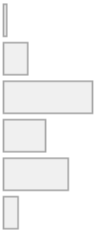
No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid
1	Prediction [numeric]	Mean (sd) : 8.5 (0.1) min ≤ med ≤ max: 8.2 ≤ 8.5 ≤ 8.7 IQR (CV) : 0.1 (0)	8.20 : 2 (5.7%) 8.30 : 2 (5.7%) 8.40 : 5 (14.3%) 8.50 : 19 (54.3%) 8.60 : 5 (14.3%) 8.70 : 2 (5.7%)		35 (100.0%)

Generated by summarytools (https://github.com/dcomtois/summarytools) 1.0.1 (R (https://www.r-project.org/) version 4.1.0)
2023-04-04

Brand B

Test set prediction summary statistics for beverage brand B are provided below.

The predictions are available for viewing and for download in .csv format here (https://raw.githubusercontent.com/MauricioClaudio/DATA624-/main/predictions_BrandB.csv).

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid
1	Prediction [numeric]	Mean (sd) : 8.6 (0.1) min ≤ med ≤ max: 8.3 ≤ 8.6 ≤ 8.8 IQR (CV) : 0.2 (0)	8.30 : 2 (1.6%) 8.40 : 13 (10.1%) 8.50 : 48 (37.2%) 8.60 : 23 (17.8%) 8.70 : 35 (27.1%) 8.80 : 8 (6.2%)		129 (100.0%)

Generated by summarytools (https://github.com/dcomtois/summarytools) 1.0.1 (R (https://www.r-project.org/) version 4.1.0)
2023-04-04

Brand C

Test set prediction summary statistics for beverage brand C are provided below.

The predictions are available for viewing and for download in .csv format here (https://raw.githubusercontent.com/MauricioClaudio/DATA624-/main/predictions_BrandC.csv).

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid
1	Prediction [numeric]	Mean (sd) : 8.4 (0.1) min ≤ med ≤ max: 8.1 ≤ 8.4 ≤ 8.7 IQR (CV) : 0.2 (0)	8.10 : 1 (3.2%) 8.20 : 1 (3.2%) 8.30 : 9 (29.0%) 8.40 : 9 (29.0%) 8.50 : 8 (25.8%) 8.60 : 1 (3.2%) 8.70 : 2 (6.5%)		31 (100.0%)

Generated by summarytools (https://github.com/dcomtois/summarytools) 1.0.1 (R (https://www.r-project.org/) version 4.1.0)
2023-04-04

Brand D

Test set prediction summary statistics for beverage brand D are provided below.

The predictions are available for viewing and for download in .csv format here (https://raw.githubusercontent.com/MauricioClaudio/DATA624-/main/predictions_BrandD.csv).

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid
1	Prediction [numeric]	Mean (sd) : 8.6 (0.1) min ≤ med ≤ max: 8.4 ≤ 8.6 ≤ 8.8 IQR (CV) : 0.2 (0)	8.40 : 3 (4.7%) 8.50 : 19 (29.7%) 8.60 : 22 (34.4%) 8.70 : 15 (23.4%) 8.80 : 5 (7.8%)		64 (100.0%)

Generated by summarytools (https://github.com/dcomtois/summarytools) 1.0.1 (R (https://www.r-project.org/) version 4.1.0)
2023-04-04