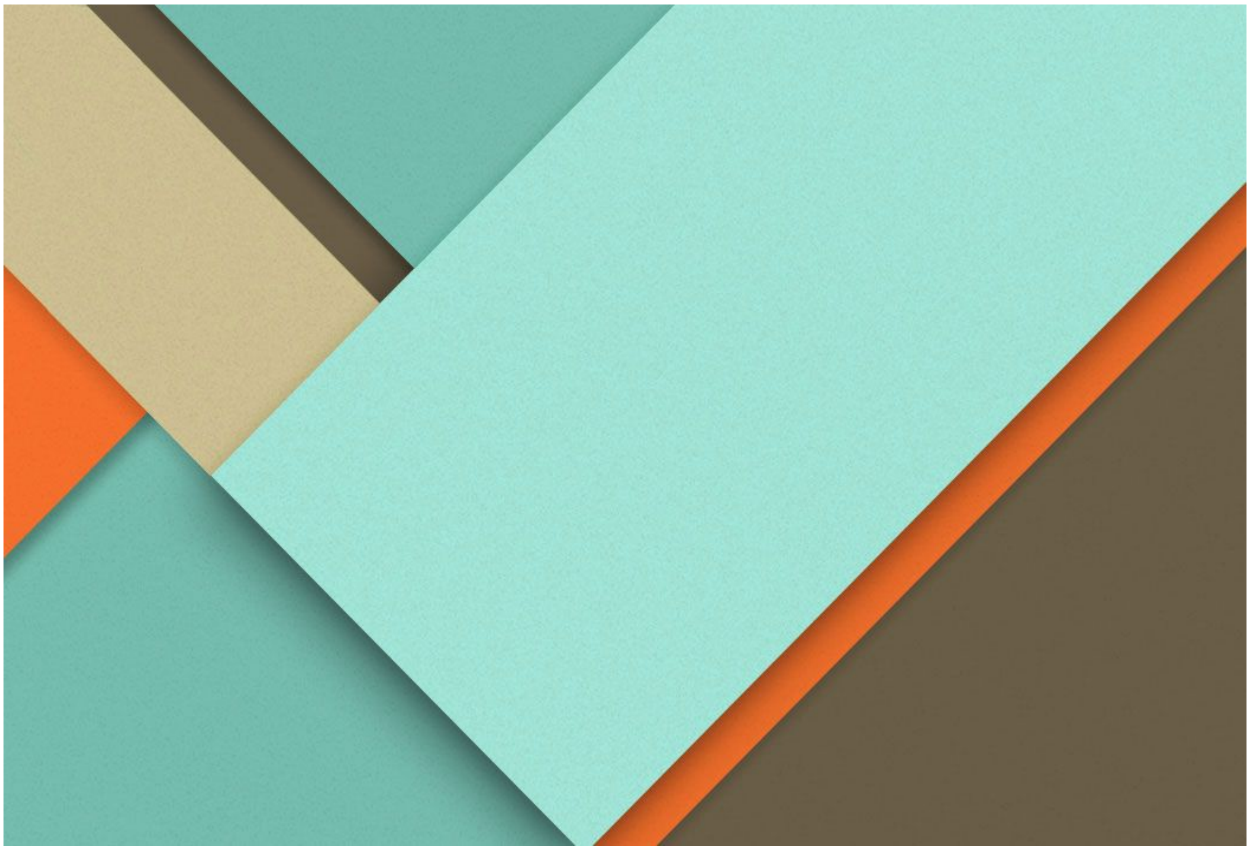
Data 624 Predictive Analytics

Final Project



—Prediction of PH model of Beverages

Niteen Kumar
Hovig Ohannessian
Gurpreet Singh
Peter Goodridge

4/30/2019

Contents

Project Description

Project #2 (Team) Assignment

This is role playing. I am your new boss. I am in charge of production at ABC Beverage and you are a team of data scientists reporting to me. My leadership has told me that new regulations are requiring us to understand our manufacturing process, the predictive factors and be able to report to them our predictive model of PH.

Please use the historical data set I am providing. Build and report the factors in BOTH a technical and non-technical report. I like to use Word and Excel. Please provide your non-technical report in a business friendly readable document and your predictions in an Excel readable format.   The technical report should show clearly the models you tested and how you selected your final approach.

Please submit both Rpubs links and .rmd files or other readable formats for technical and non-technical reports. Also submit the excel file showing the prediction of your models for pH.

Executive Summary

New Regulations by ABC beverage company leadership requires the company's production unit to better understand the manufacturing process, the predictive factors and their relationship to the PH of the beverages.

Research Statement

The research is an effort to find the predictive variables related to the ph of the beverages and build the predictive model for ph of beverages

Data Collection

The data set is a historic data containing predictors associated to the PH and is provided in an excel file. We will utilize this historic dataset to analyze and predict the PH of beverages. Two excel files are provided:

- The training data (StudentData.xlsx)
- The test data (StudentEvaluation.xlsx).

Data Exploration and Visualization

The data set consists of total variables of:

- Training dataset: **2,571** records and **33** predictors (pH included)
- Evaluation or test dataset: **267** records and **33** predictors (pH included)

In this section, we will explore the features found in the data set and analyze them for utilizing them in the model building section. PH variable will be our response variable and remaining 32 variables will be used for prediction.

Variable Structure

Majority of the variables found in the dataset are numeric or integer. The variables Brand Code's structure is character. It consists of four brand codes "A", "B", "C" and "D". In addition there are some records with missing brand codes. We will treat the values in the data preparation section.

Brand Code Distribution:

Brand Code	Number of Records
A	293
B	1,235
C	303
D	615
NULL	120

Content of Datasets

To have a better understanding on the namings and their relative values, the following summaries will display the predictors and some of its observations.

Summary content of training dataset:

```
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':      2571 obs. of  33 variables:
 $ Brand Code       : chr  "B" "A" "B" "A" ...
 $ Carb Volume      : num  5.34 5.43 5.29 5.44 5.49 ...
 $ Fill Ounces      : num  24 24 24.1 24 24.3 ...
 $ PC Volume        : num  0.263 0.239 0.263 0.293 0.111 ...
 $ Carb Pressure    : num  68.2 68.4 70.8 63 67.2 66.6 64.2 67.6 64.2 72 ...
 $ Carb Temp        : num  141 140 145 133 137 ...
 $ PSC              : num  0.104 0.124 0.09 NA 0.026 0.09 0.128 0.154 0.132 0.014 ...
 $ PSC Fill         : num  0.26 0.22 0.34 0.42 0.16 ...
 $ PSC CO2          : num  0.04 0.04 0.16 0.04 0.12 ...
 $ Mnf Flow         : num  -100 -100 -100 -100 -100 -100 -100 -100 -100 -100 ...
 $ Carb Pressure1   : num  119 122 120 115 118 ...
 $ Fill Pressure    : num  46 46 46 46.4 45.8 45.6 51.8 46.8 46 45.2 ...
 $ Hyd Pressure1    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Hyd Pressure2    : num  NA NA NA 0 0 0 0 0 0 0 ...
 $ Hyd Pressure3    : num  NA NA NA 0 0 0 0 0 0 0 ...
 $ Hyd Pressure4    : num  118 106 82 92 92 116 124 132 90 108 ...
 $ Filler Level     : num  121 119 120 118 119 ...
 $ Filler Speed     : num  4002 3986 4020 4012 4010 ...
 $ Temperature      : num  66 67.6 67 65.6 65.6 66.2 65.8 65.2 65.4 66.6 ...
 $ Usage cont       : num  16.2 19.9 17.8 17.4 17.7 ...
 $ Carb Flow        : num  2932 3144 2914 3062 3054 ...
 $ Density          : num  0.88 0.92 1.58 1.54 1.54 1.52 0.84 0.84 0.9 0.9 ...
 $ MFR              : num  725 727 735 731 723 ...
 $ Balling          : num  1.4 1.5 3.14 3.04 3.04 ...
 $ Pressure Vacuum  : num  -4 -4 -3.8 -4.4 -4.4 -4.4 -4.4 -4.4 -4.4 -4.4 ...
 $ PH               : num  8.36 8.26 8.94 8.24 8.26 8.32 8.4 8.38 8.38 8.5 ...
 $ Oxygen Filler    : num  0.022 0.026 0.024 0.03 0.03 0.024 0.066 0.046 0.064 0.022 ...
 $ Bowl Setpoint    : num  120 120 120 120 120 120 120 120 120 120 ...
 $ Pressure Setpoint: num  46.4 46.8 46.6 46 46 46 46 46 46 46 ...
 $ Air Pressurer    : num  143 143 142 146 146 ...
 $ Alch Rel         : num  6.58 6.56 7.66 7.14 7.14 7.16 6.54 6.52 6.52 6.54 ...
 $ Carb Rel         : num  5.32 5.3 5.84 5.42 5.44 5.44 5.38 5.34 5.34 5.34 ...
 $ Balling Lvl      : num  1.48 1.56 3.28 3.04 3.04 3.02 1.44 1.44 1.44 1.38 ...
```

Summary content of test dataset:

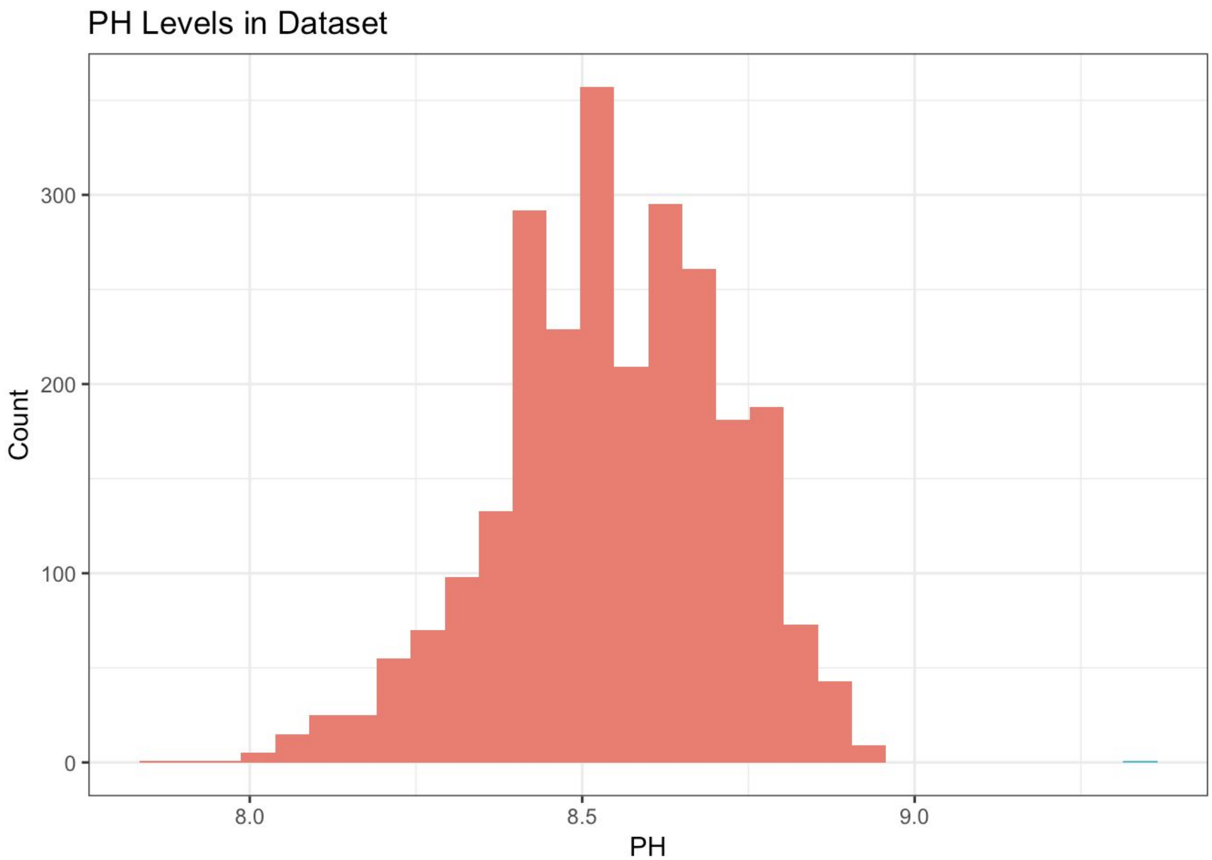
```
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':	267 obs. of  33 variables:
 $ Brand Code       : chr  "D" "A" "B" "B" ...
 $ Carb Volume      : num  5.48 5.39 5.29 5.27 5.41 ...
 $ Fill Ounces      : num  24 24 23.9 23.9 24.2 ...
 $ PC Volume        : num  0.27 0.227 0.303 0.186 0.16 ...
 $ Carb Pressure    : num  65.4 63.2 66.4 64.8 69.4 73.4 65.2 67.4 66.8 72.6 ...
 $ Carb Temp        : num  135 135 140 139 142 ...
 $ PSC              : num  0.236 0.042 0.068 0.004 0.04 0.078 0.088 0.076 0.246 0.146 ..
 $ PSC Fill         : num  0.4 0.22 0.1 0.2 0.3 ...
 $ PSC CO2          : num  0.04 0.08 0.02 0.02 0.06 ...
 $ Mnf Flow         : num  -100 -100 -100 -100 -100 -100 -100 -100 -100 -100 ...
 $ Carb Pressure1   : num  117 119 120 125 115 ...
 $ Fill Pressure    : num  46 46.2 45.8 40 51.4 46.4 46.2 40 43.8 40.8 ...
 $ Hyd Pressure1    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Hyd Pressure2    : num  NA 0 0 0 0 0 0 0 0 0 ...
 $ Hyd Pressure3    : num  NA 0 0 0 0 0 0 0 0 0 ...
 $ Hyd Pressure4    : num  96 112 98 132 94 94 108 108 110 106 ...
 $ Filler Level     : num  129 120 119 120 116 ...
 $ Filler Speed     : num  3986 4012 4010 NA 4018 ...
 $ Temperature      : num  66 65.6 65.6 74.4 66.4 66.6 66.8 NA 65.8 66 ...
 $ Usage cont       : num  21.7 17.6 24.2 18.1 21.3 ...
 $ Carb Flow        : num  2950 2916 3056 28 3214 ...
 $ Density          : num  0.88 1.5 0.9 0.74 0.88 0.84 1.48 1.6 1.52 1.48 ...
 $ MFR              : num  728 736 735 NA 752 ...
 $ Balling          : num  1.4 2.94 1.45 1.06 1.4 ...
 $ Pressure Vacuum  : num  -3.8 -4.4 -4.2 -4 -4 -3.8 -4.2 -4.4 -4.4 -4.2 ...
 $ PH               : logi  NA NA NA NA NA NA ...
 $ Oxygen Filler    : num  0.022 0.03 0.046 NA 0.082 0.064 0.042 0.096 0.046 0.096 ...
 $ Bowl Setpoint    : num  130 120 120 120 120 120 120 120 120 120 ...
 $ Pressure Setpoint: num  45.2 46 46 46 50 46 46 46 46 46 ...
 $ Air Pressurer    : num  143 147 147 146 146 ...
 $ Alch Rel         : num  6.56 7.14 6.52 6.48 6.5 6.5 7.18 7.16 7.14 7.78 ...
 $ Carb Rel         : num  5.34 5.58 5.34 5.5 5.38 5.42 5.46 5.42 5.44 5.52 ...
 $ Balling Lvl      : num  1.48 3.04 1.46 1.48 1.46 1.44 3.02 3 3.1 3.12 ...
```

pH

Our dataset include value variables for pH but first let’s determine how pH measurements, scale or definition:

- pH = potential of Hydrogen
- 1 <= pH <= 14
- pH < 7 indicates acidity increase
- pH = 7 indicates neutrality
- pH > 7 indicates alkalinity

pH in the dataset is visualized as the following:



Summary of Dataset

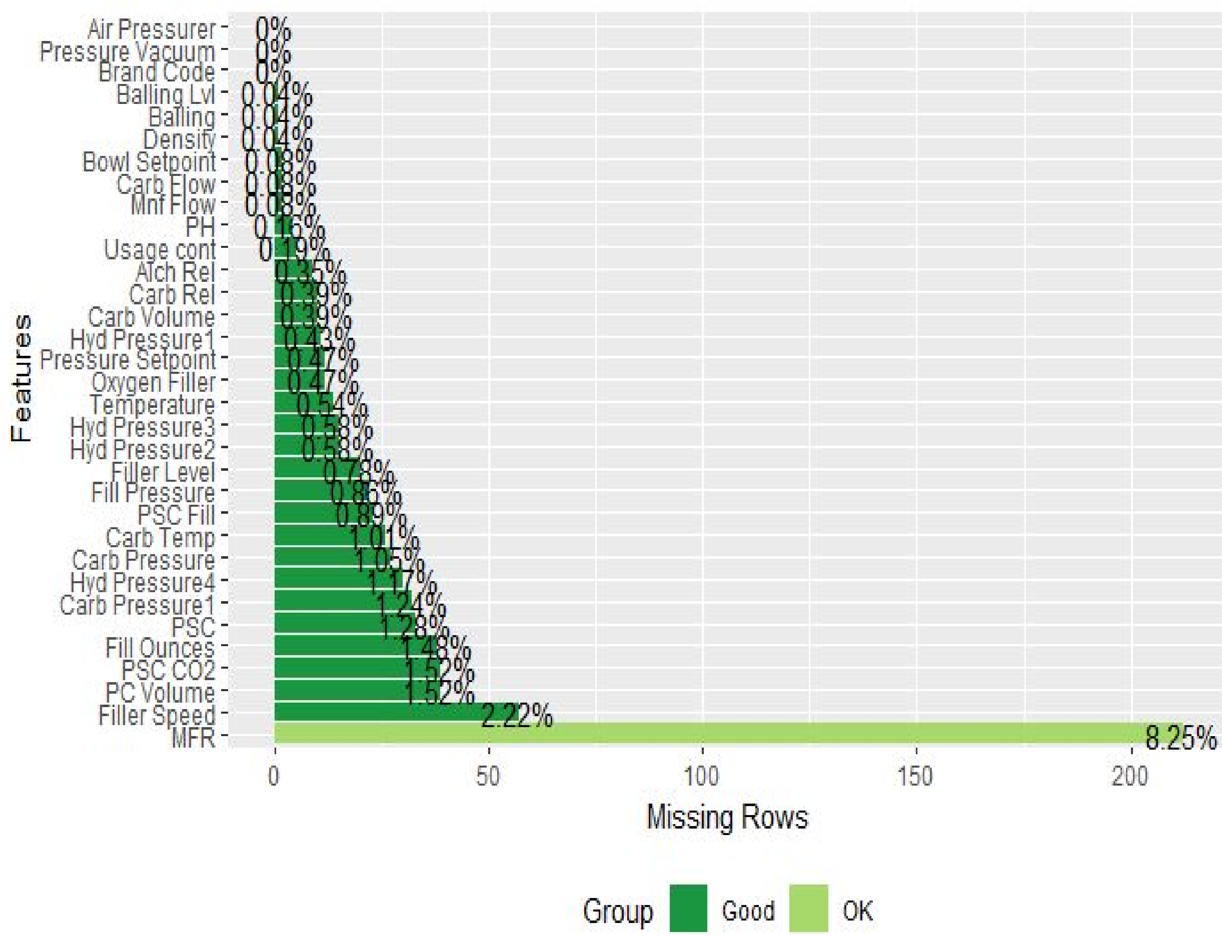
Variables	N	Missing	Mean	SD	Median	Min	Max	Range	Skew	Kurtosis
Carb Volume	2561	10	5.37	0.11	5.35	5.04	5.7	0.66	0.39	-0.47
Fill Ounces	2533	38	23.97	0.09	23.97	23.63	24.32	0.69	-0.02	0.86
PC Volume	2532	39	0.28	0.06	0.27	0.08	0.48	0.4	0.34	0.67
Carb Pressure	2544	27	68.19	3.54	68.2	57	79.4	22.4	0.18	-0.01
Carb Temp	2545	26	141.09	4.04	140.8	128.6	154	25.4	0.25	0.24
PSC	2538	33	0.08	0.05	0.08	0	0.27	0.27	0.85	0.65
PSC Fill	2548	23	0.2	0.12	0.18	0	0.62	0.62	0.93	0.77
PSC CO2	2532	39	0.06	0.04	0.04	0	0.24	0.24	1.73	3.73
Mnf Flow	2569	2	24.57	119.48	65.2	-100.2	229.4	329.6	0	-1.87
Carb Pressure1	2539	32	122.59	4.74	123.2	105.6	140.2	34.6	0.05	0.14
Fill Pressure	2549	22	47.92	3.18	46.4	34.6	60.4	25.8	0.55	1.41
Hyd Pressure1	2560	11	12.44	12.43	11.4	-0.8	58	58.8	0.78	-0.14
Hyd Pressure2	2556	15	20.96	16.39	28.6	0	59.4	59.4	-0.3	-1.56
Hyd Pressure3	2556	15	20.46	15.98	27.6	-1.2	50	51.2	-0.32	-1.57
Hyd Pressure4	2541	30	96.29	13.12	96	52	142	90	0.55	0.63
Filler Level	2551	20	109.25	15.7	118.4	55.8	161.2	105.4	-0.85	0.05
Filler Speed	2514	57	3687.2	770.82	3982	998	4030	3032	-2.87	6.71
Temperature	2557	14	65.97	1.38	65.6	63.6	76.2	12.6	2.39	10.16
Usage cont	2566	5	20.99	2.98	21.79	12.08	25.9	13.82	-0.54	-1.02
Carb Flow	2569	2	2468.4	1073.7	3028	26	5104	5078	-0.99	-0.58
Density	2570	1	1.17	0.38	0.98	0.24	1.92	1.68	0.53	-1.2
MFR	2359	212	704.05	73.9	724	31.4	868.6	837.2	-5.09	30.46
Balling	2570	1	2.2	0.93	1.65	-0.17	4.01	4.18	0.59	-1.39
Pressure Vacuum	2571	0	-5.22	0.57	-5.4	-6.6	-3.6	3	0.53	-0.03
PH	2567	4	8.55	0.17	8.54	7.88	9.36	1.48	-0.29	0.06
Oxygen Filler	2559	12	0.05	0.05	0.03	0	0.4	0.4	2.66	11.09
Bowl Setpoint	2569	2	109.33	15.3	120	70	140	70	-0.97	-0.06
Pressure Setpoint	2559	12	47.62	2.04	46	44	52	8	0.2	-1.6
Air Pressurer	2571	0	142.83	1.21	142.6	140.8	148.2	7.4	2.25	4.73
Alch Rel	2562	9	6.9	0.51	6.56	5.28	8.62	3.34	0.88	-0.85
Carb Rel	2561	10	5.44	0.13	5.4	4.96	6.06	1.1	0.5	-0.29
Balling Lvl	2570	1	2.05	0.87	1.48	0	3.66	3.66	0.59	-1.49

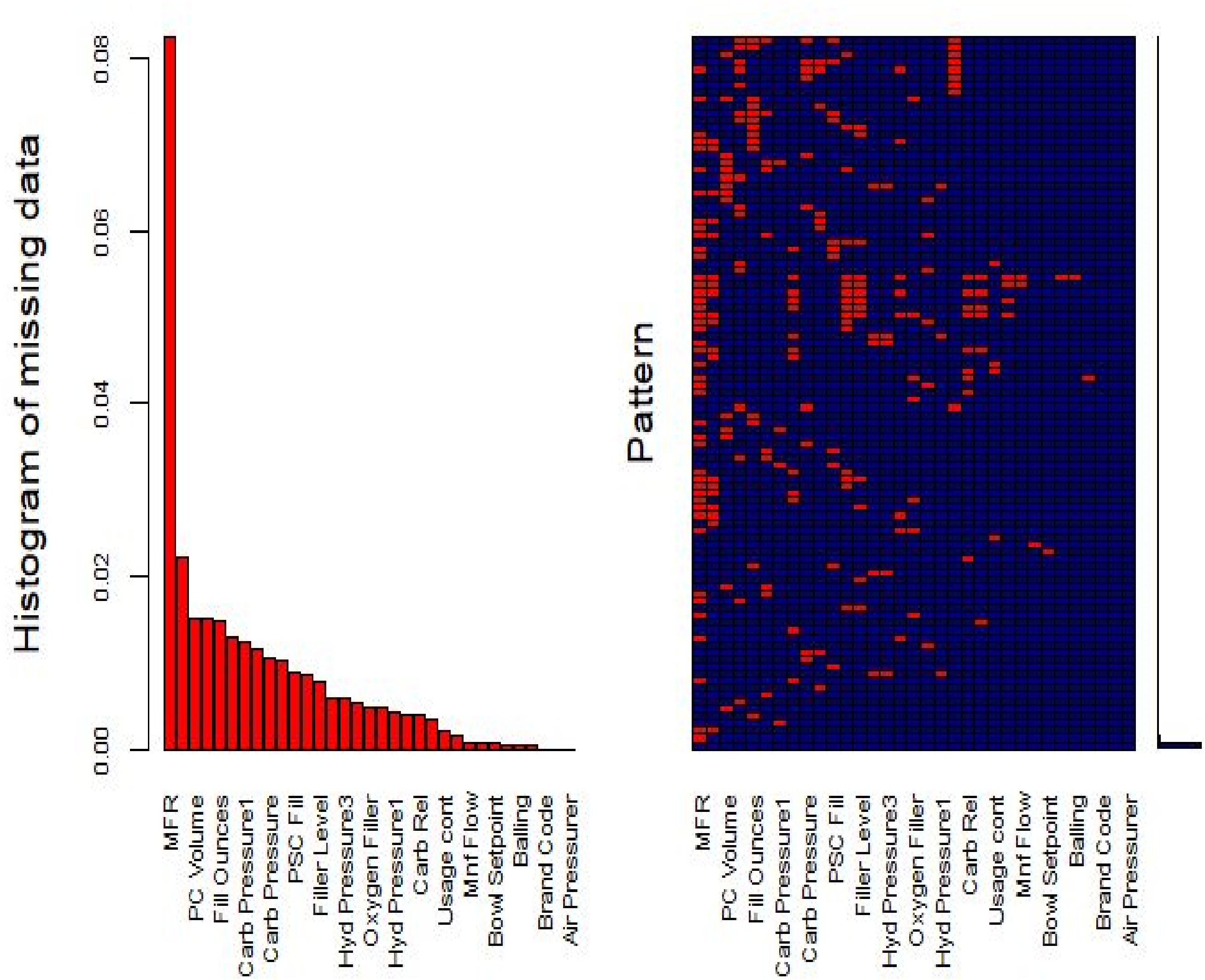
Missing Values

Variable	Missing	Missing %
MFR	212	8.2%
Filler Speed	57	2.2%
PC Volume	39	1.5%
PSC CO2	39	1.5%
Fill Ounces	38	1.5%

PSC	33	1.3%
Carb Pressure 1	32	1.2%
Hyd Pressure4	30	1.2%
Carb Pressure	27	1.1%
Carb Temp	26	1.0%
PSC Fill	23	0.9%
Fill Pressure	22	0.9%
Filler Level	20	0.8%
Hyd Pressure 2	15	0.6%
Hyd Pressure 3	15	0.6%
Temperature	14	0.5%
Oxygen Filler	12	0.5%
Pressure Setpoint	12	0.5%
Hyd Pressure 1	11	0.4%
Carb Volume	10	0.4%
Carb Rel	10	0.4%
Alch Rel	9	0.4%
Usage cont	5	0.2%
PH	4	0.2%
Mnf Flow	2	0.1%
Carb Flow	2	0.1%
Bowl Setpoint	2	0.1%
Density	1	0.0%
Balling	1	0.0%
Balling Lvl	1	0.0%
Brand Code	0	0.0%

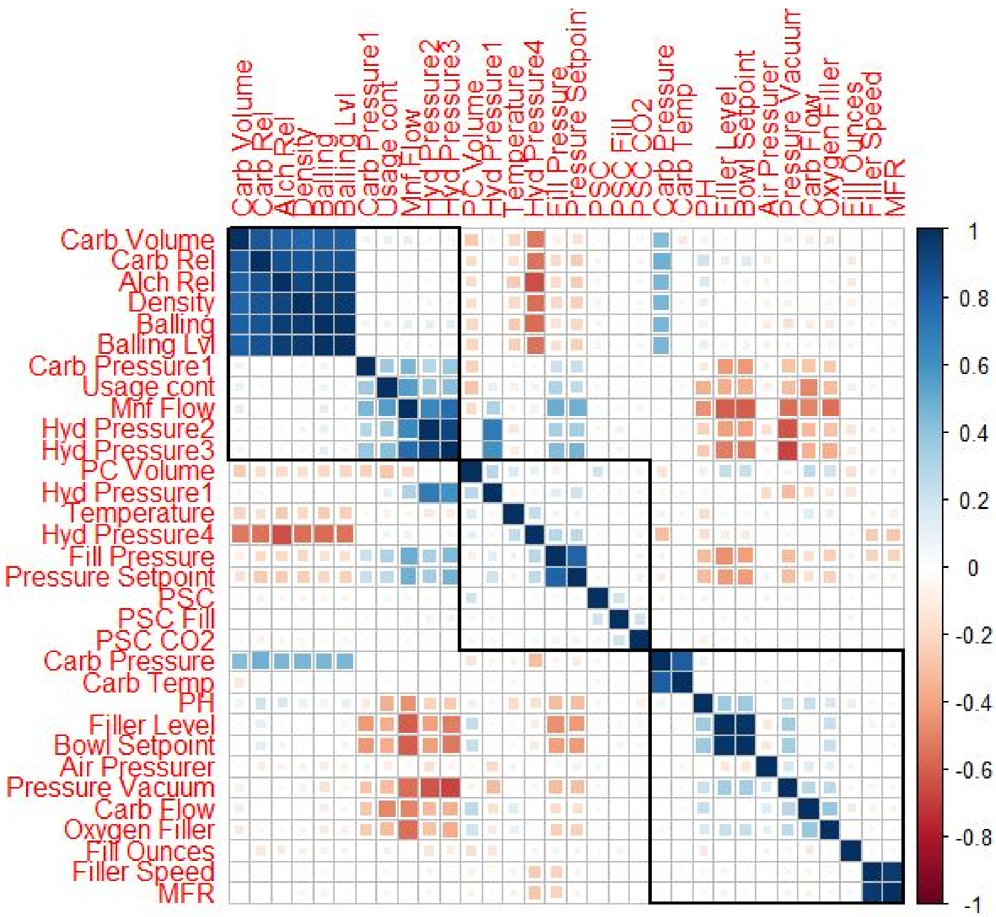
Pressure Vacuum	0	0.0%
Air Pressurer	0	0.0%





Histogram of missing data
0.00
0.02
0.04
0.06
0.08
MFR
PC Volume
Fill Ounces
Carb Pressure1
Carb Pressure
PSC Fill
Filler Level
Hyd Pressure3
Oxygen Filler
Hyd Pressure1
Carb Rel
Usage cont
Mnf Flow
Bowl Setpoint
Balling
Brand Code
Air Pressurer
Pattern
MFR
PC Volume
Fill Ounces
Carb Pressure1
Carb Pressure
PSC Fill
Filler Level
Hyd Pressure3
Oxygen Filler
Hyd Pressure1
Carb Rel
Usage cont
Mnf Flow
Bowl Setpoint
Balling
Brand Code
Air Pressurer

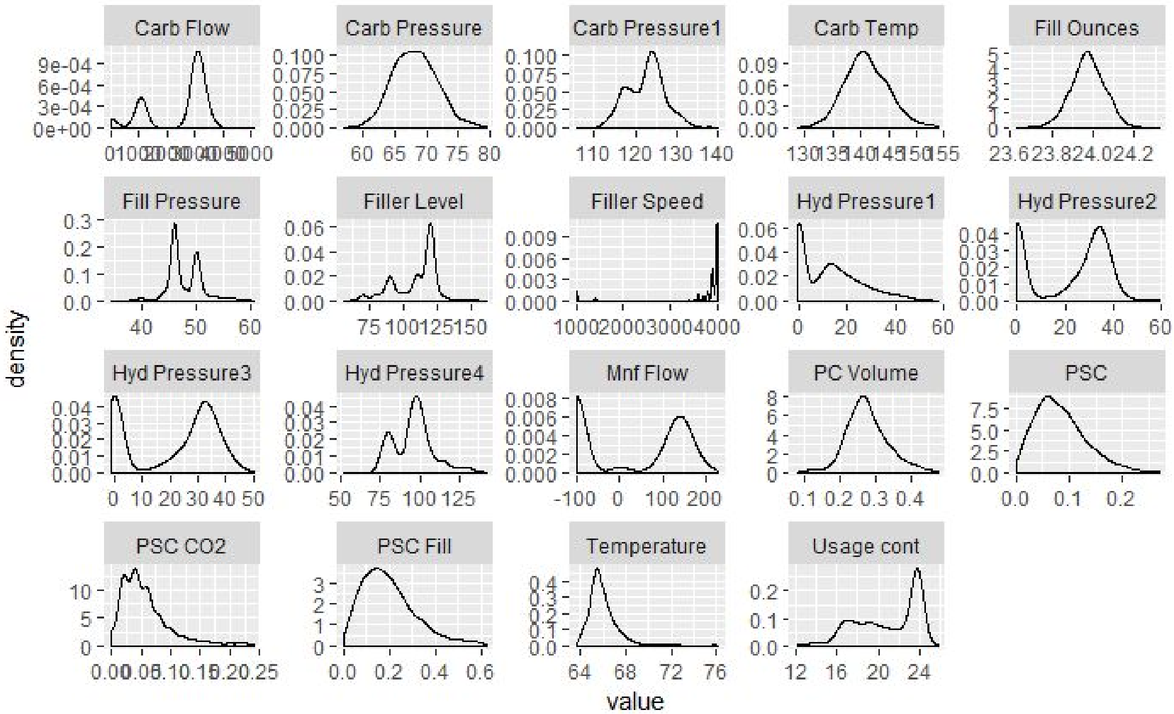
Correlation

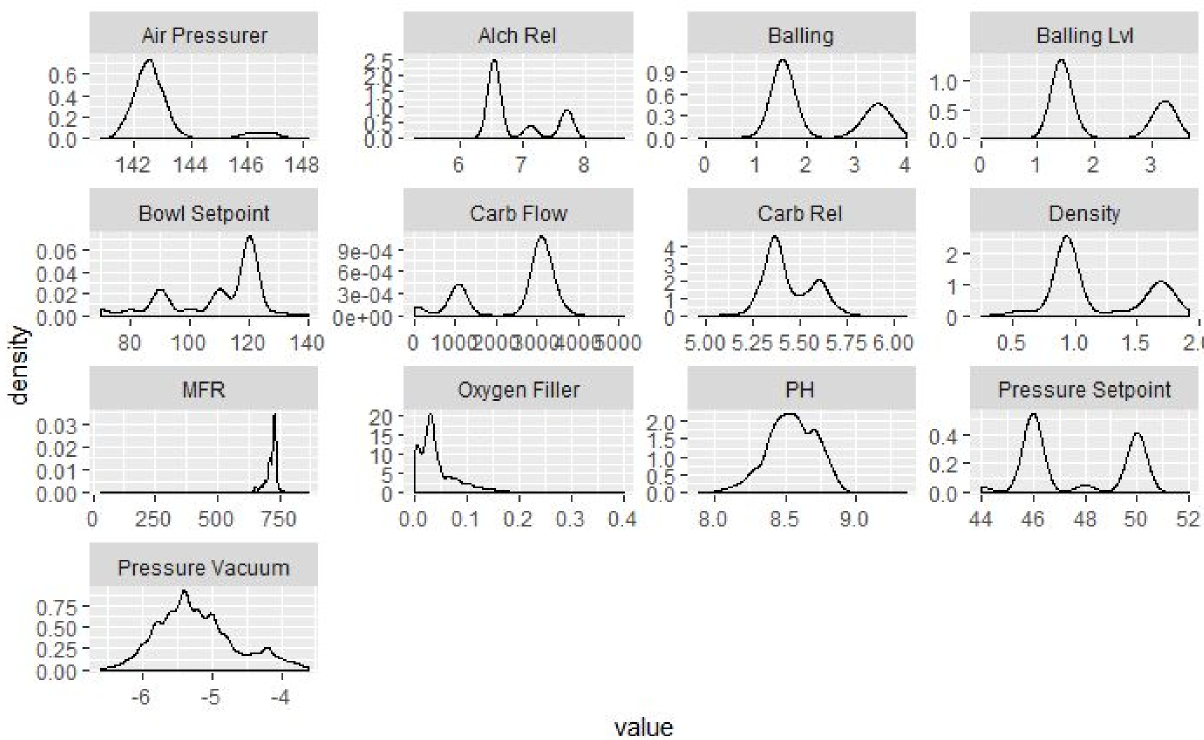


Breaking the variables into 3 clusters, we see the group in the top left having a potential problem with multicollinearity. With our candidate methods, this will not be a major problem for prediction, but it should be accounted for when examining variable importance.

Normality

Normality is one of the most widely used technique to understand the continuous predictors. In the below plot we can see normal distribution behavior of the given dataset for different features.


Carb Flow
Carb Pressure
Carb Pressure1
Carb Temp
Fill Ounces
Fill Pressure
Filler Level
Filler Speed
Hyd Pressure1
Hyd Pressure2
Hyd Pressure3
Hyd Pressure4
Mnf Flow
PC Volume
PSC
PSC CO2
PSC Fill
Temperature
Usage cont
density
value


Air Pressurer
Alch Rel
Balling
Balling Lvl
Bowl Setpoint
Carb Flow
Carb Rel
Density
MFR
Oxygen Filler
PH
Pressure Setpoint
Pressure Vacuum
density
value

Zero Variance/Near Zero Variance Predictors

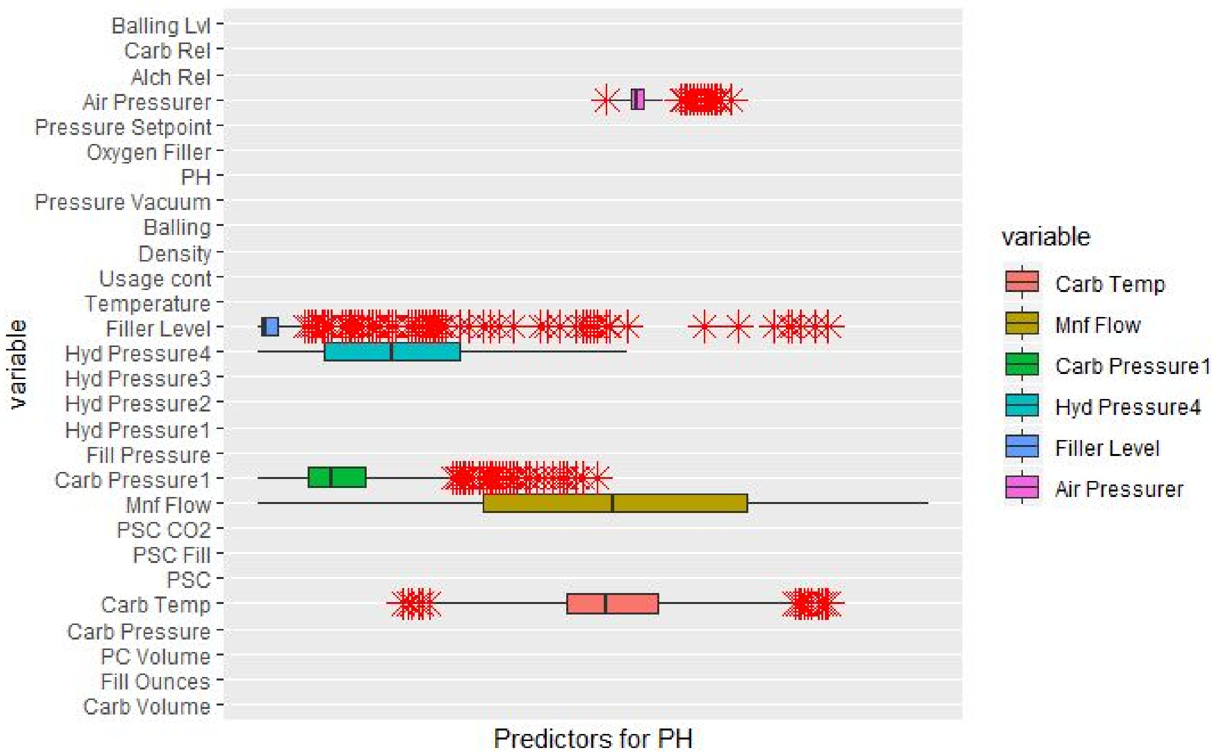
To filter for near-zero variance predictors, the caret package function nearZeroVar will return the column numbers of any predictors that fulfill the conditions outlined. A zero variance predictor will never be chosen for a split since it offers no possible predictive information.

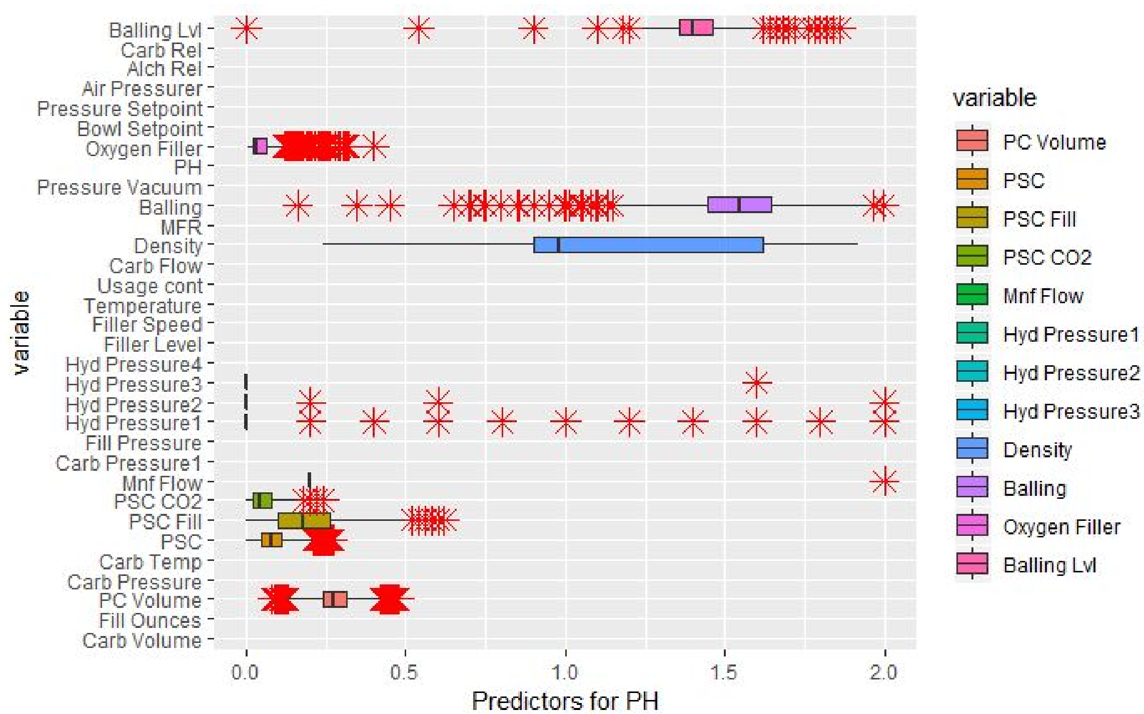
	Freq Ratio	Percent Unique	ZeroVar	Near Zero Var
Hyd Pressure 1	31.1	9.52	FALSE	TRUE

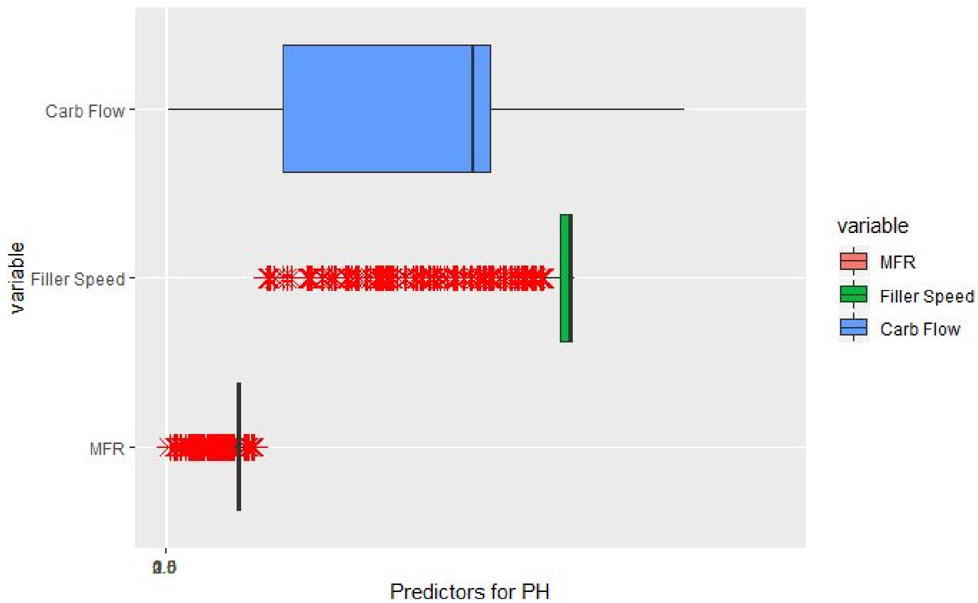
Using caret package and near zero variable function, only one variable is picked up as near zero predictor.

Outliers/ Box Plots

Box plots for the variables reveal, that besides having the outliers in the variables, most of the variables are skewed. For example: Variables density, carb flow, filler speed and oxygen filler are skewed providing us an opportunity to further check their distribution.


Balling Lvl
Carb Rel
Alch Rel
Air Pressurer
Pressure Setpoint
Oxygen Filler
PH
Pressure Vacuum
Balling
Density
Usage cont
Temperature
Filler Level
Hyd Pressure4
Hyd Pressure3
Hyd Pressure2
Hyd Pressure1
Fill Pressure
Carb Pressure1
Mnf Flow
PSC CO2
PSC Fill
PSC
Carb Temp
Carb Pressure
PC Volume
Fill Ounces
Carb Volume
variable
Predictors for PH
variable
Carb Temp
Mnf Flow
Carb Pressure1
Hyd Pressure4
Filler Level
Air Pressurer

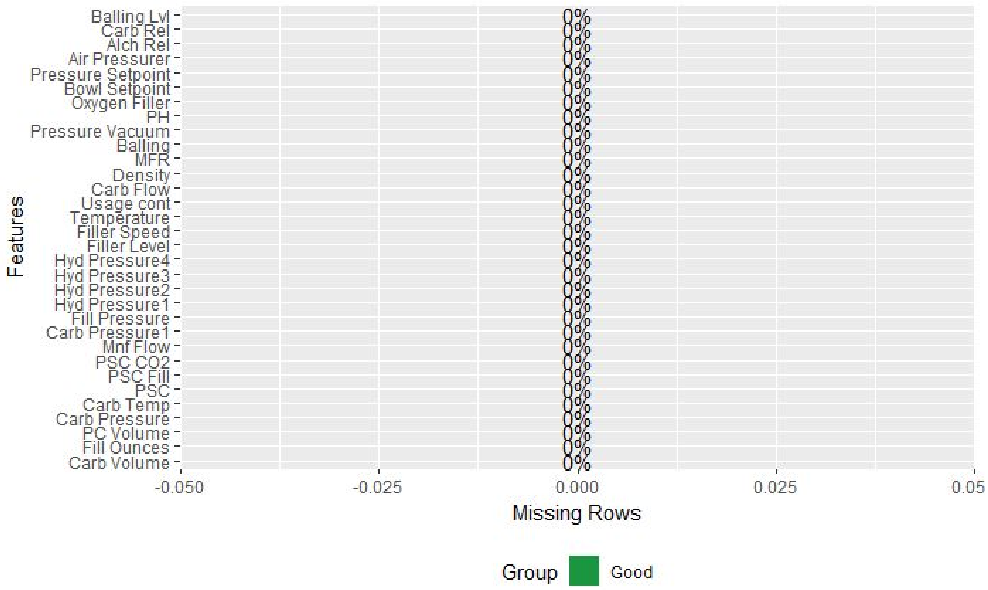

Balling Lvl
Carb Rel
Alch Rel
Air Pressurer
Pressure Setpoint
Bowl Setpoint
Oxygen Filler
PH
Pressure Vacuum
Balling
MFR
Density
Carb Flow
Usage cont
Temperature
Filler Speed
Filler Level
Hyd Pressure4
Hyd Pressure3
Hyd Pressure2
Hyd Pressure1
Fill Pressure
Carb Pressure1
Mnf Flow
PSC CO2
PSC Fill
PSC
Carb Temp
Carb Pressure
PC Volume
Fill Ounces
Carb Volume
variable
0.0
0.5
1.0
1.5
2.0
Predictors for PH
variable
PC Volume
PSC
PSC Fill
PSC CO2
Mnf Flow
Hyd Pressure1
Hyd Pressure2
Hyd Pressure3
Density
Balling
Oxygen Filler
Balling Lvl



Data Preparation

Missing Values Treatment

We will treat the missing values using random forest in the mice package..



Outliers Treatment:

Looking at the box plots in data exploration section, we noticed there were outliers in the data set. Although the outliers impact the result of analysis, it is not always the best approach to drop the outliers or imput them. We are assuming that these are the most legitimate and interesting observations which can provide better inference about the predictions.

Splitting Data:

We will split the data into training and testing data using 75:25 ratio by utilizing createDataPartition function in the caret package.

Build Models

For state-of-the-art prediction quality, we will use a model stack. This will consist of tuning models separately and then combining the candidate models in a manner that will make the whole greater than the sum of the parts.

From the perspective of understanding the manufacturing process, the model stack will also provide benefits. The stack is like a panel of experts, each looking at the data through slightly different lenses to form their diagnoses. By looking at the predictors each model uses, we can gather assemble a complete picture of the factors that affect our manufacturing process.

Below are the results of the individual model tuning. The First table shows the result of the tuning process, and the next show the prediction quality on completely unseen data.

MARS

degree	nprune	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
2	24	0.1282077	0.4419826	0.096461	0.0032028	0.0228325	0.0012608

Prediction Matrix

RMSE	Rsquared	MAE
0.12321748	0.50918351	0.09193757

Random Forest

mtry	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
25	0.1196867	0.5207878	0.0911039	0.0026487	0.0189388	0.0021575

Prediction Matrix

RMSE	Rsquared	MAE
0.09909556	0.69434421	0.07105279

Cubist

committees	neighbors	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
5	7	0.125044	0.4754669	0.0944422	0.0034074	0.0249831	0.0030823

Prediction Matrix

RMSE	Rsquared	MAE
0.10442588	0.64620868	0.07529925

XGB Trees

eta	nrounds	max_depth	gamma	colsample_bytree	min_child_weight	subsample
0.01	1000	6	0	0.8	0.8	0.8

RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
0.1226378	0.4896501	0.0931218	0.0014446	0.0168542	0.001228

Prediction Matrix

RMSE	Rsquared	MAE
0.10369057	0.66025197	0.07691495

XGB Dart

eta	nrounds	gamma	skip_drop	rate_drop	max_depth	colsample_bytree	min_child_weight	subsample
0.01	1000	0.1	0.6	0.4	6	0.6	0.6	0.6

RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
0.1303314	0.4476223	0.1036222	0.0015185	0.0213551	0.0018045

Prediction Matrix

```
      RMSE    Rsquared         MAE
0.11895212 0.57482820 0.09249779
```

Model Stack

alpha	lambda	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
0.10	0.0002339	0.1195489	0.5060443	0.0889906	0.0041998	0.0217873	0.0031864

```
##       RMSE   Rsquared        MAE
## 0.09036733 0.74478118 0.06319195
```

Our combined model improves performance by several percentage points over the best individual model, Random Forest.

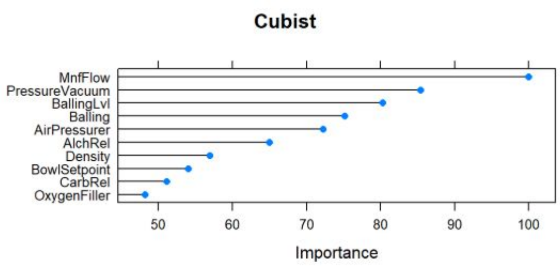
Final Model Evaluation

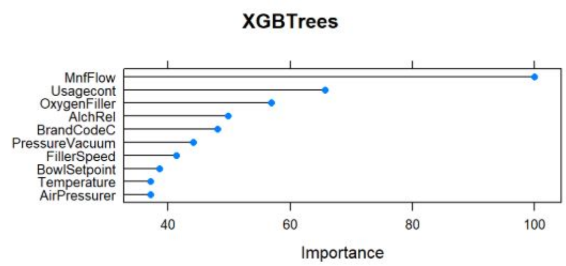
With the tuning parameters known, we built our final model using the full training dataset. The results of this model were even better than the last as a result of the additional data that was fed to the model.

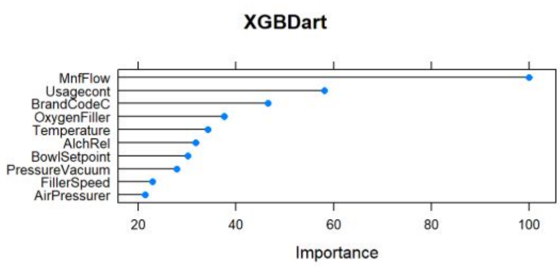
alpha	lambda	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
0.10	0.0002527	0.1140552	0.5590911	0.0844652	0.0020945	0.0141659	0.0009666

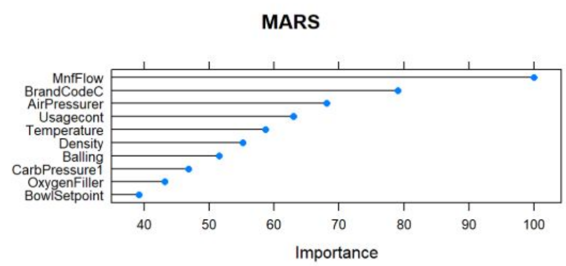
Manufacturing Process Factors

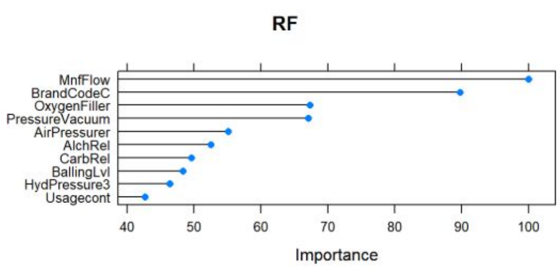
Each model present in the stack is able to attach a score to each feature that summarizes how important it was in the prediction of PH. The higher the score, the more the model used that feature in it's prediction.











To summarize factors relevant to our manufacturing process, below are the predictors present in each model's top 10, and the count of different models in which they appeared.

var	ModelCount
AirPressurer	5
MnfFlow	5
OxygenFiller	5
AlchRel	4
BowlSetpoint	4
BrandCodeC	4
PressureVacuum	4
Usagecont	4
Temperature	3
Balling	2
BallingLvl	2
CarbRel	2
Density	2
FillerSpeed	2
CarbPressure1	1
HydPressure3	1

Policy Recommendation

Using our highly tuned model, manufacturing can be adjusted to achieve the desired pH for each actively produced beverage line. The process for future products can be planned before any actual beverages are produced, saving costs on trial and error. The feature importances can be used as a priority list for equipment maintenance,, with the equipment related to more important predictors being tuned first.

Conclusion

After working on extracting the data from the given files, we processed data cleansing and handle the missing values along with the NAs. We trained and tested the data 75% to 25% respectively.

Our models were able to produce for us predicted values for **pH** which are also saved in **predictions.csv** file separately.

We notice that all the values predicted are greater than 7 and more specifically greater than 8. This scale translates into saying that the beverage made is **alkaline**.

At the beginning of this study, we were not informed about the nature of the ABC Beverage company, meaning of what type of beverage manufacturer it was. But from our studies we can conclude that this company produces alkaline beverages like water, dairy, tea, fruit drinks, etc.

Appendix:

Github : https://github.com/hovig/Team5-Data624-Project2

RPUBS: https://rpubs.com/hovig613/493738

Prediction Results:
https://github.com/hovig/Team5-Data624-Project2/blob/master/predictions.csv

References

https://towardsdatascience.com/the-use-of-knn-for-missing-values-cf33d935c637

Forecasting: Principles and Practice - 2nd edition: https://otexts.com/fpp2/

Applied Predictive Modeling:
https://github.com/hovig/Team5-Data624-Project2/blob/master/applied-predictive-modeling-max-kuhn-kjell-johnson_1518.pdf