## horizontal line

Data 624 Predictive Analytics

Final Project



**─Prediction of PH model of Beverages**

Niteen Kumar

Hovig Ohannessian

Gurpreet Singh

Peter Goodridge

4/30/2019

**Contents**

**Executive Summary 2**

**Research Statement 2**

**Data Collection 2**

**Data Exploration and Visualization 2**

Variable Structure/Summary

Missing Values

Correlation

Normality

Zero/Near Zero Variance

Outliers

**Data Preparation 12**

Missing Values

Zero/Near Zero Variance

Correlation

Outliers

Splitting the Data

**Build Models 1**3

Model 1 - MARS

Model 2 - Random Forest

Model 3 - Cubist

Model 4 - XGB Trees

Model 5 - XGB Dart

[**M**](#_yyrhu7ml5bea)**odel Selection 16**

**Results**

**Appendix**

**References** 2

# 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 |

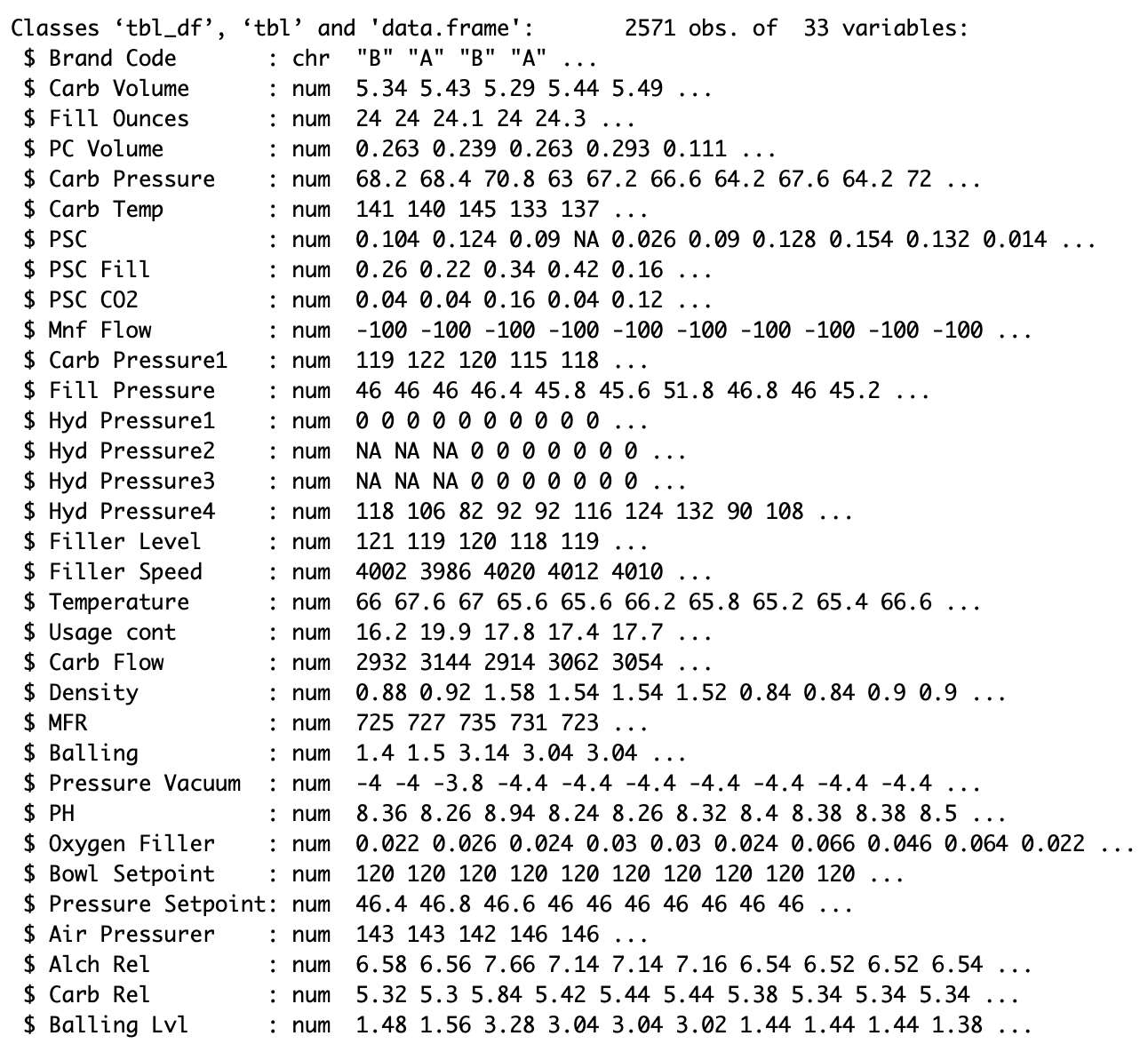
## 

## 

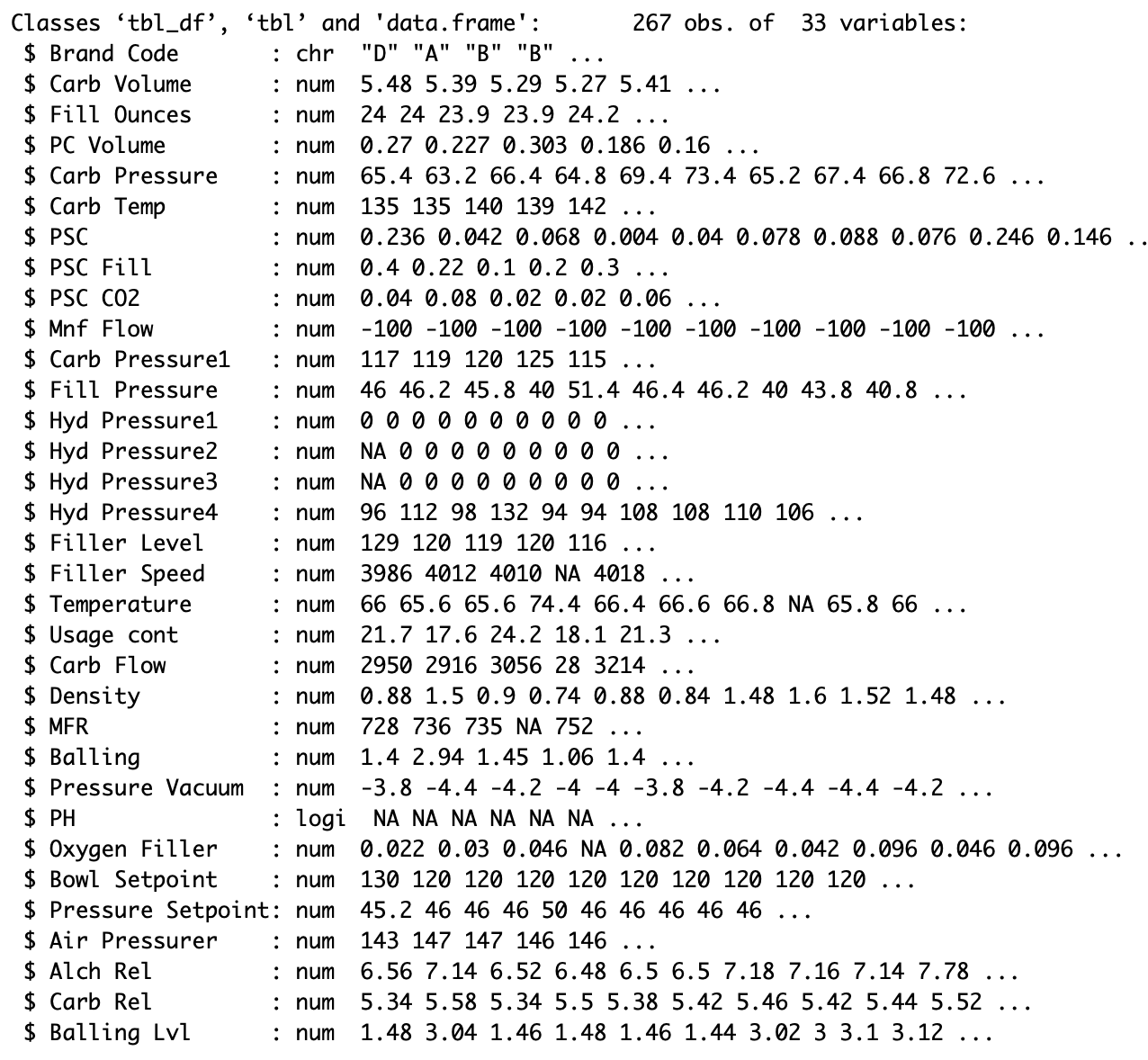
## Content of Datasets

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:



Summary content of test dataset:

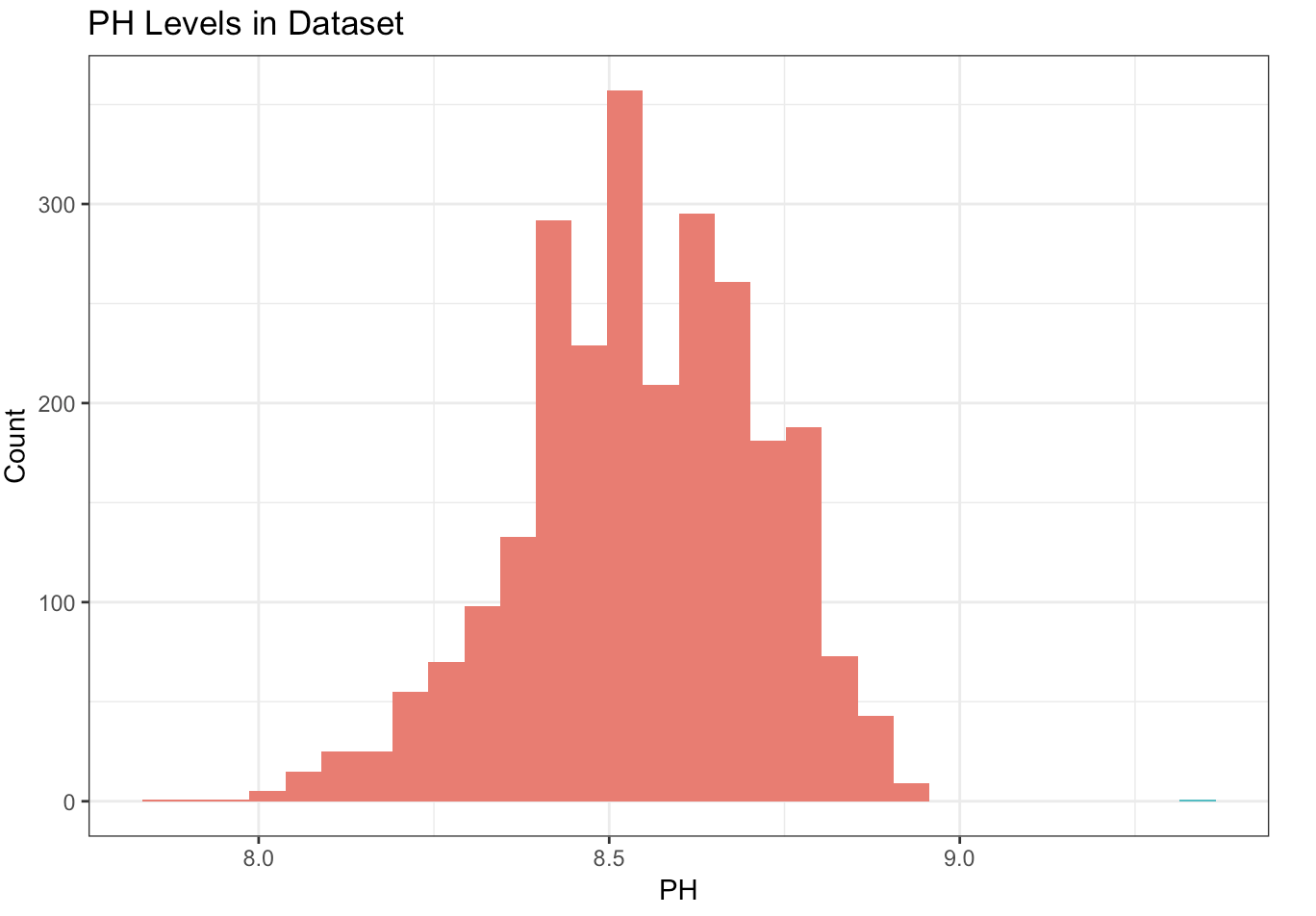


## pH

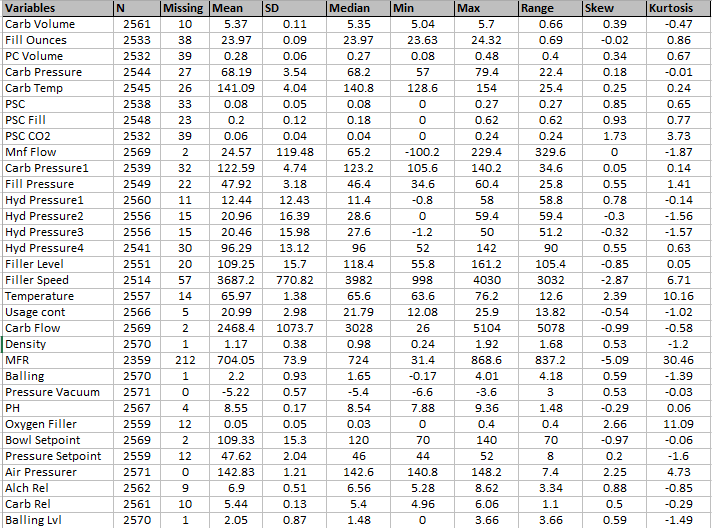
Our dataset include value variables for pH but first let’s determine how pH measurements, scale or definition:

* 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:

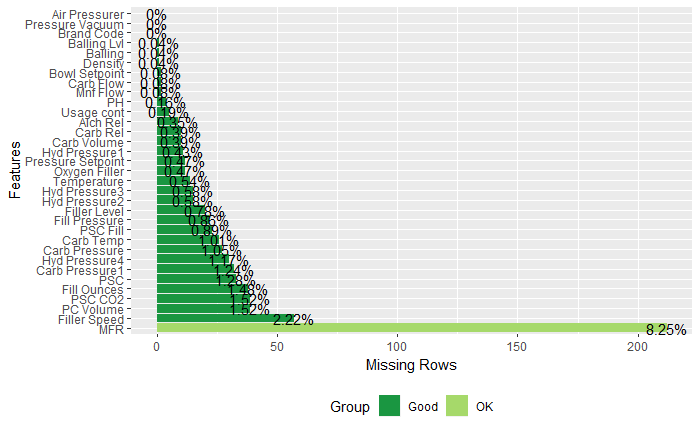


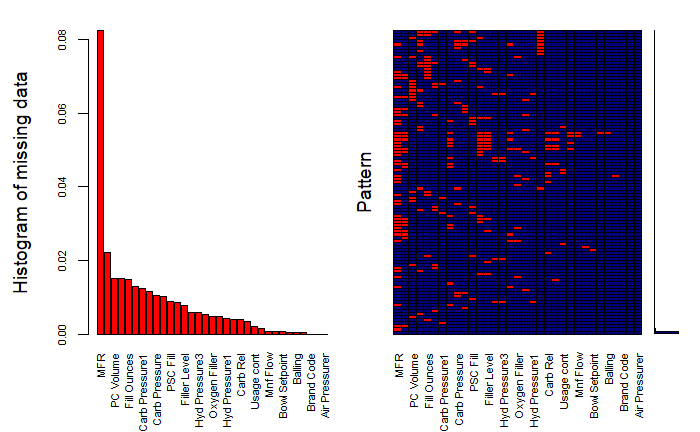
## Summary of Dataset



## 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% |

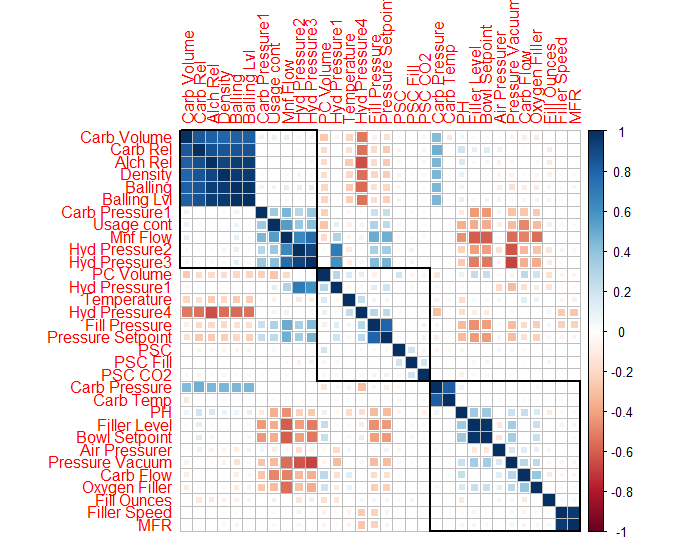




## 

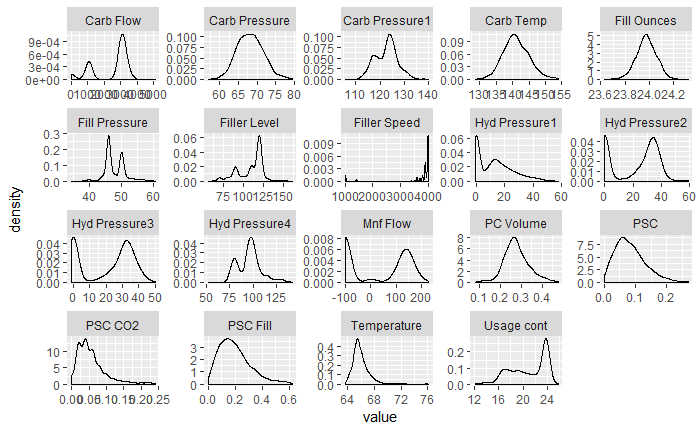
## 

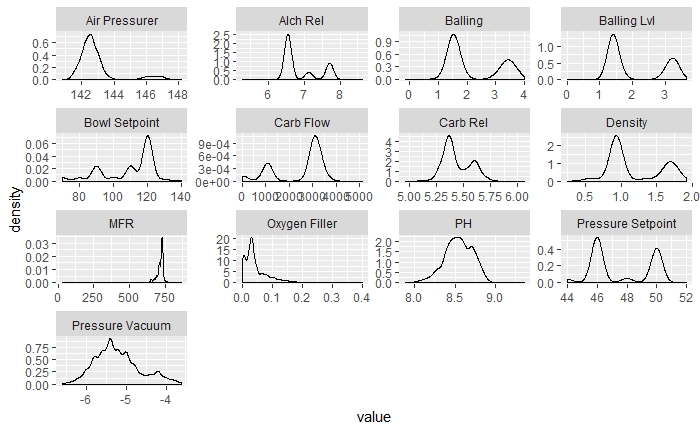
Correlation



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





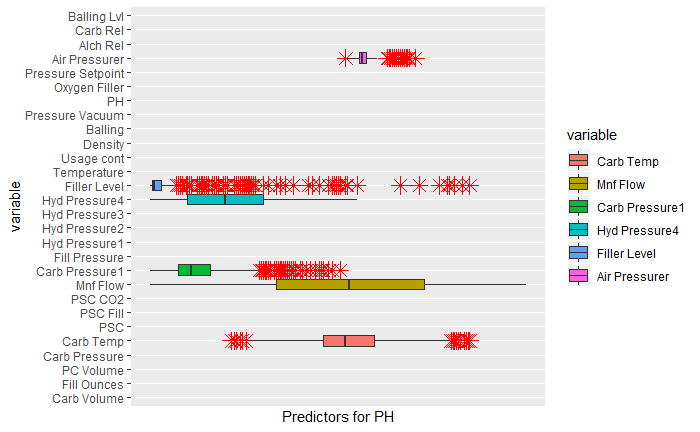
## Zero Variance/Near Zero Variance Predictors

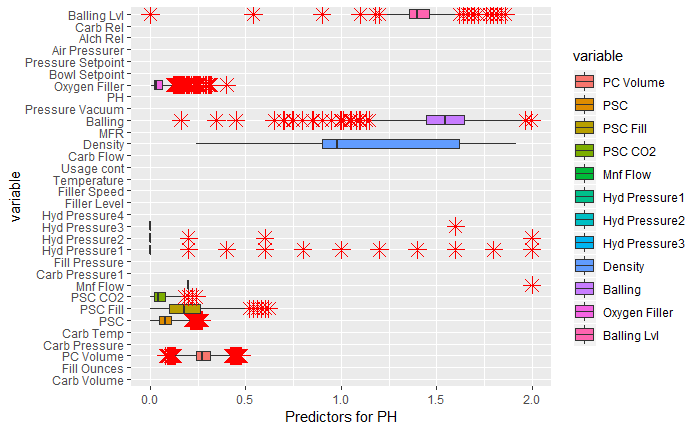
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Freq Ratio | Percent Unique | ZeroVar | Near Zero Var |
| Hyd Pressure 1 | 31.1 | 9.52 | FALSE | TRUE |

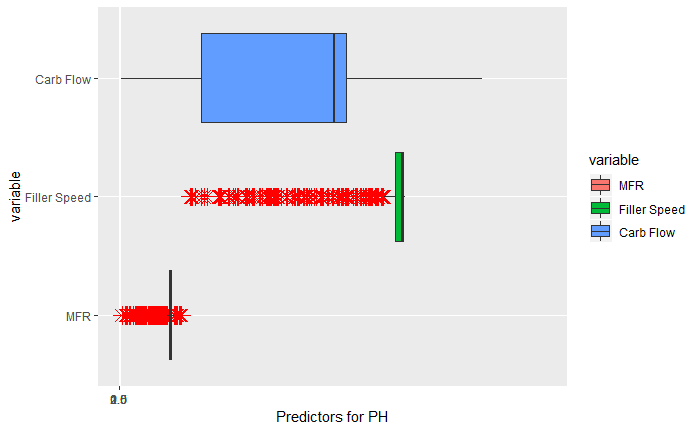
Using caret package and near zero variable function, only one variable is picked up as near zero predictor.

## 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.



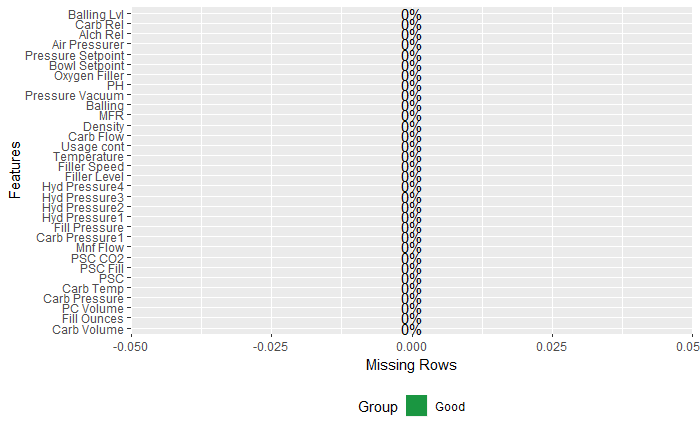




# Data Preparation

## 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.

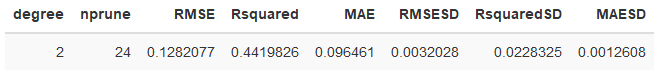
# Build Models

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.

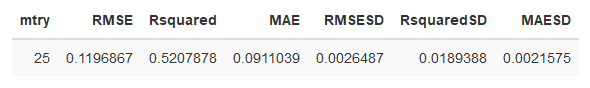
## MARS



Prediction Matrix



## Random Forest



Prediction Matrix



## Cubist



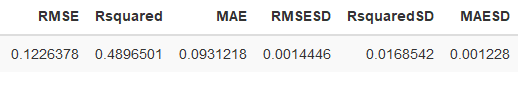
Prediction Matrix



## 

## XGB Trees





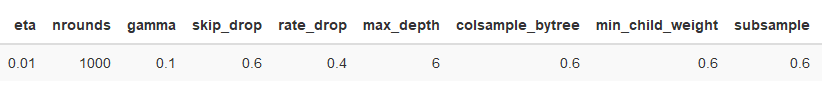
Prediction Matrix

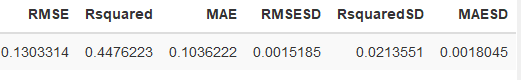


## 

## 

## XGB Dart

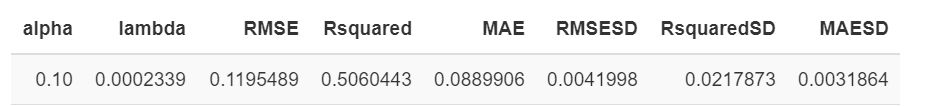




Prediction Matrix



## Model Stack

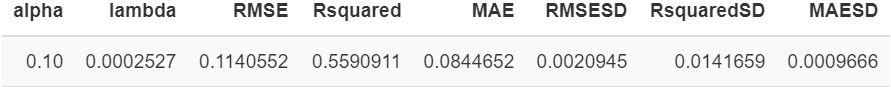




Our combined model improves performance by several percentage points over the best individual model, Random Forest.

## Final Model Evaluation

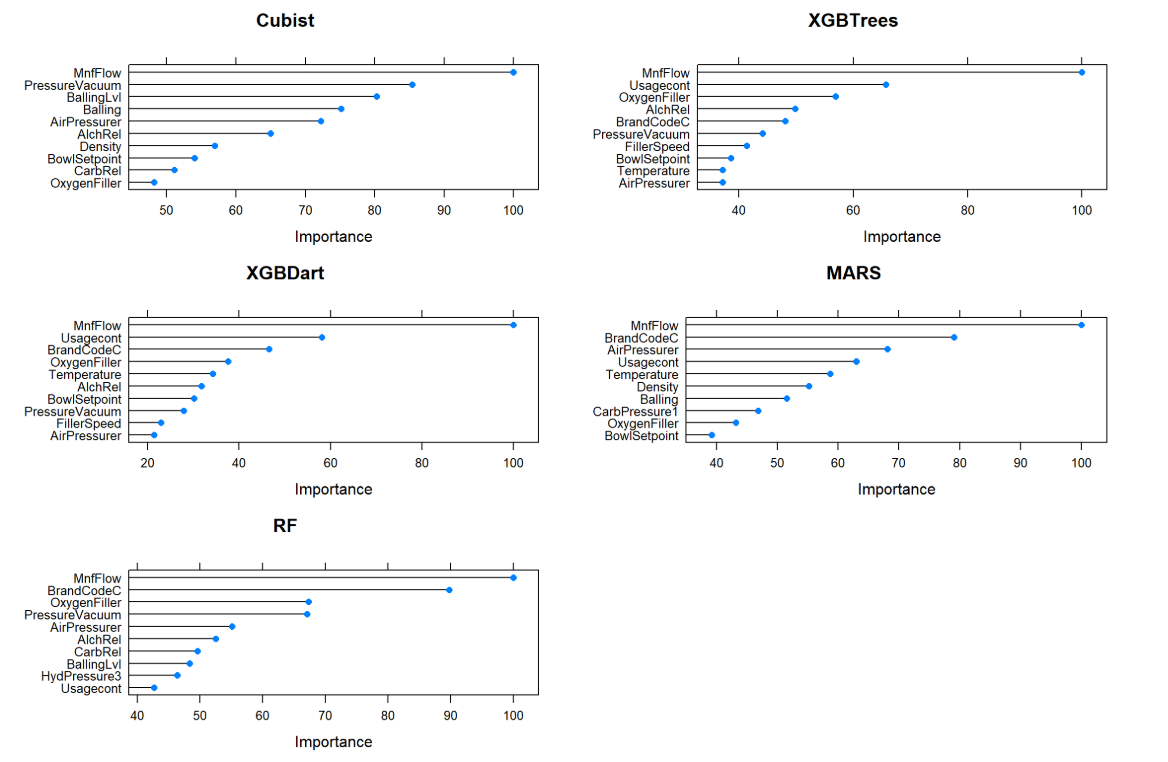
With the tuning parameters known, we built our final model using the full training dataset. The results of this model were even better than the last as a result of the additional data that was fed to the model.



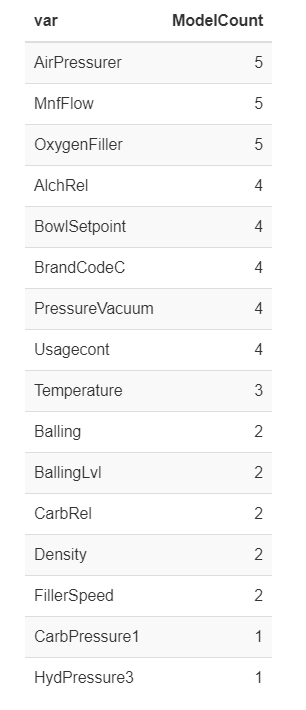
# 

# 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.



# 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>