Project #2 ABC Beverage PH Levels

DATA624 Spring 2018

Group 2

BRIAN KREIS

Bin Lin

Chris Martin

Asher Meyer

Contents

[ABC Beverage: PH Predictive Factors 2](#_Toc513921973)

[Executive Summary / Non-Technical Report 2](#_Toc513921974)

[Technical Report (with R Code) 2](#_Toc513921975)

[Data Exploration and Statistic Measures 3](#_Toc513921976)

[Missing and Zero Values 3](#_Toc513921977)

[Descriptive Statistics and Data Exploration 3](#_Toc513921978)

[Correlation 5](#_Toc513921979)

[Data Manipulation 6](#_Toc513921980)

[Dummy Variables 6](#_Toc513921981)

[Handling Missing Values 7](#_Toc513921982)

[Predictions 7](#_Toc513921983)

[Test & Training Sets 8](#_Toc513921984)

[Linear Regression Models 8](#_Toc513921985)

[Nonlinear Regression Models 8](#_Toc513921986)

[Regression Tree / Rule-Based Models 9](#_Toc513921987)

[Conclusion 9](#_Toc513921988)

[Appendix: Project Prompt 10](#_Toc513921989)

# ABC Beverage: PH Predictive Factors

# Executive Summary / Non-Technical Report

As data scientists for ABC Beverage, our task is to discover the predictive factors of a batch’s PH levels and create a predictive model.

Our project team has obtained a data set containing 2,571 observations with each observation representing a batch of beverages that were produced. Variables included in the data set provide information on the manufacturing process that resulted in the creation of each batch. The data contains 32 predictor variables, as well as a training and test set.

Our team used this to build multiple models on the training data and run analysis to determine which model performed best. After performing a thorough comparison of several types of predictive models, we determined that a \_\_\_\_\_\_\_\_\_\_ model has the best predictive ability, and the features most predictive of the PH level of a given batch using this model are:

* One
* Two
* Three
* Four
* Five

Using this model,

# Technical Report (with R Code)

The models and analysis used a variety of packages which will need to be installed and loaded:

library(psych)

library(ggplot2)

library(ggthemes)

library(dplyr)

library(DataExplorer)

library(knitr)

library(GGally)

library(missForest)

library(Hmisc)

library(earth)

library(caret)

For reproducibility of the results, the data was loaded to and accessed from a Github repository.

#Data to build models

data1 <- read.csv("https://raw.githubusercontent.com/624-Group2/Project-2/master/StudentData.csv", header=TRUE, sep=",")

#Data to make predictions on

toPred <- read.csv("https://raw.githubusercontent.com/624-Group2/Project-2/master/StudentEvaluation-%20TO%20PREDICT.csv")

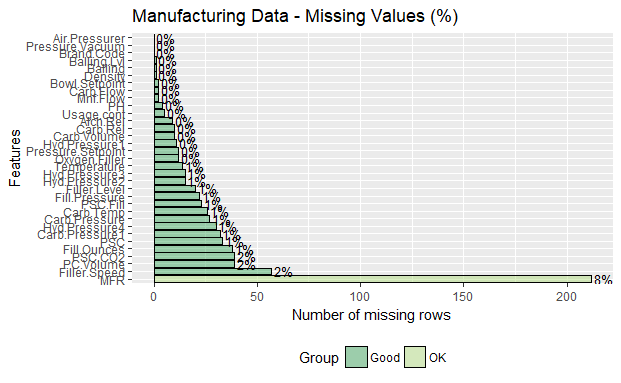
## Data Exploration and Statistic Measures

Data exploration and statistical measures were used to understand the data to determine how to process the dataset for modelling.

### Missing and Zero Values

Plotting the data as done below shows that the data has a number of features with missing values. Other methods of exploration used included counting the missing values by feature, understanding the distribution of missing values by Brand Code, and counting the missing values by Brand Code.

plot\_missing(data1, title="Manufacturing Data - Missing Values (%)")



kable(sapply(data1, FUN = function(x) sum(is.na(x))))

data1 %>%

group\_by(Brand.Code) %>%

summarise\_all(funs(sum(is.na(.))))

table(data1$Brand.Code)

Presumably, because we have identified NA values across various brands, both named and unnamed, we would expect that the NA values are not informative. In other words, if information for a particular brand is usually recorded, we would expect failure to record information to be the result of an error and not a typical process change for a particular brand. As our knowledge of the production process itself is limited, we will rely on the subject matter expertise to let us know if this presumption is incorrect.

MFR stands out as a variable with a significant number of missing values; however, the percentage of missing values is still low enough where imputation is not unreasonable. We will likely find it necessary to drop the few observations that have an NA value for the response variable.

### Descriptive Statistics and Data Exploration

Descriptive statistics were performed for all predictor and response variables to explore the data. The ‘describe’ package helps in this by automatically calculating descriptive statistics for each feature and outputting metadata on each such as the number missing, number of distinct values in the feature, etc.

#Use Describe Package to calculate Descriptive Statistic

(CC\_des <- describe(data1, na.rm=TRUE, interp=FALSE, skew=TRUE, ranges=TRUE, trim=.1, type=3, check=TRUE, fast=FALSE, quant=c(.1,.25,.75,.90), IQR=TRUE))

We have significant skewness in several variables, and, depending on our choice of regression technique may require transformation.

vis1 <- data1 %>% select(-PH -Brand.Code)

par(mfrow=(c(1,1)))

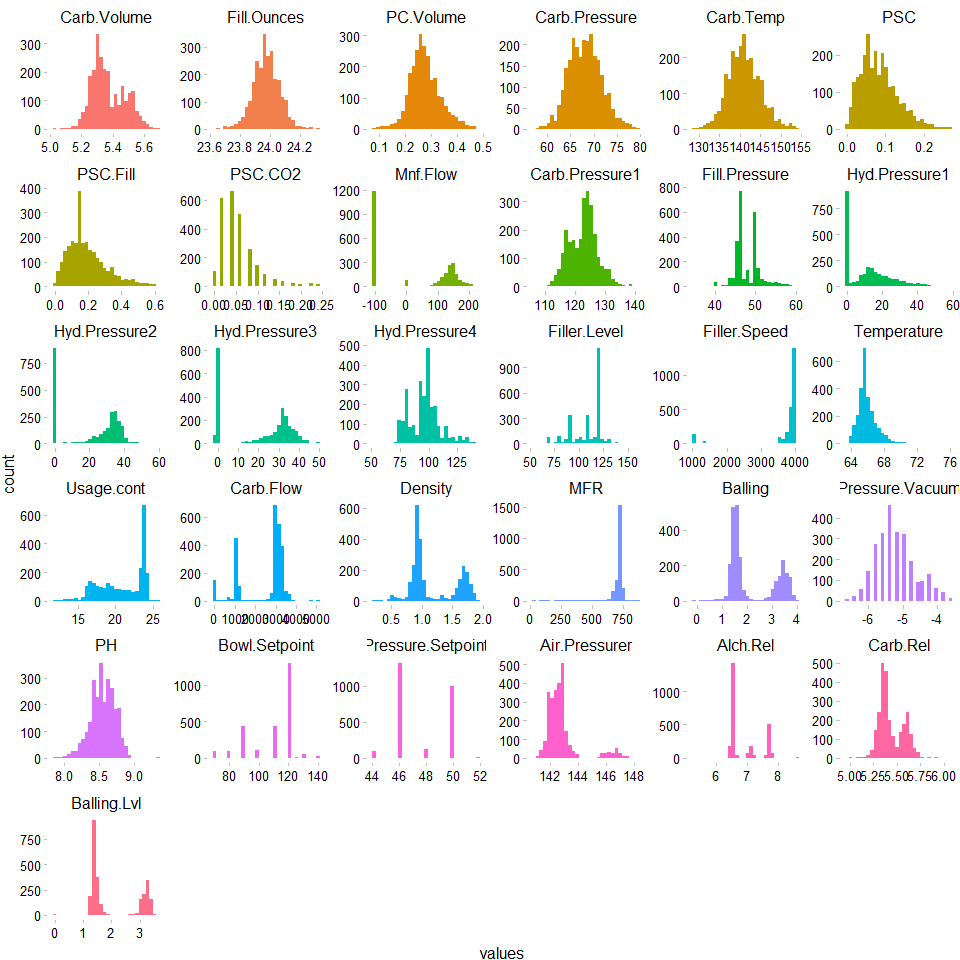
ggplot(stack(vis1), aes(values, fill=ind))+

facet\_wrap(~ind, scales = "free") +

geom\_histogram() +

theme\_pander()+

theme(legend.position="none")



We have a varying mix of distributions for the different data points. For example,

* Discrete: The variable Pressure.Setpoint appears to represent a categorical setting ranging from 44 to 52; however, we do not know at first glance if our training set contains all possible values that we might see in the test set or in general practice.
* Normal Continuous: Carb.Temp appears to be normally distributed.
* Multi Modal data: Density appears to have many values around 1 and another large concentration of values around 1.5.

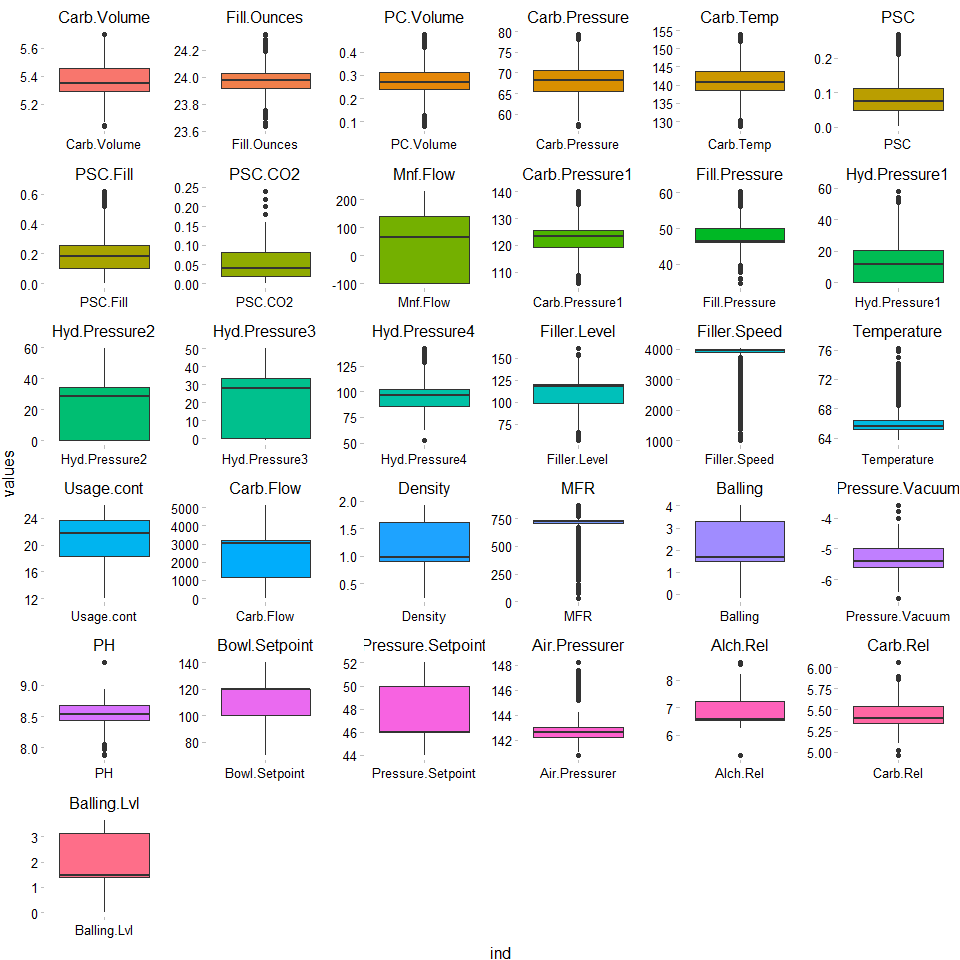
ggplot(stack(vis1), aes(x = ind, y = values, fill=ind))+

facet\_wrap(~ind, scales = "free") +

geom\_boxplot() +

theme\_pander() +

theme(legend.position="none")

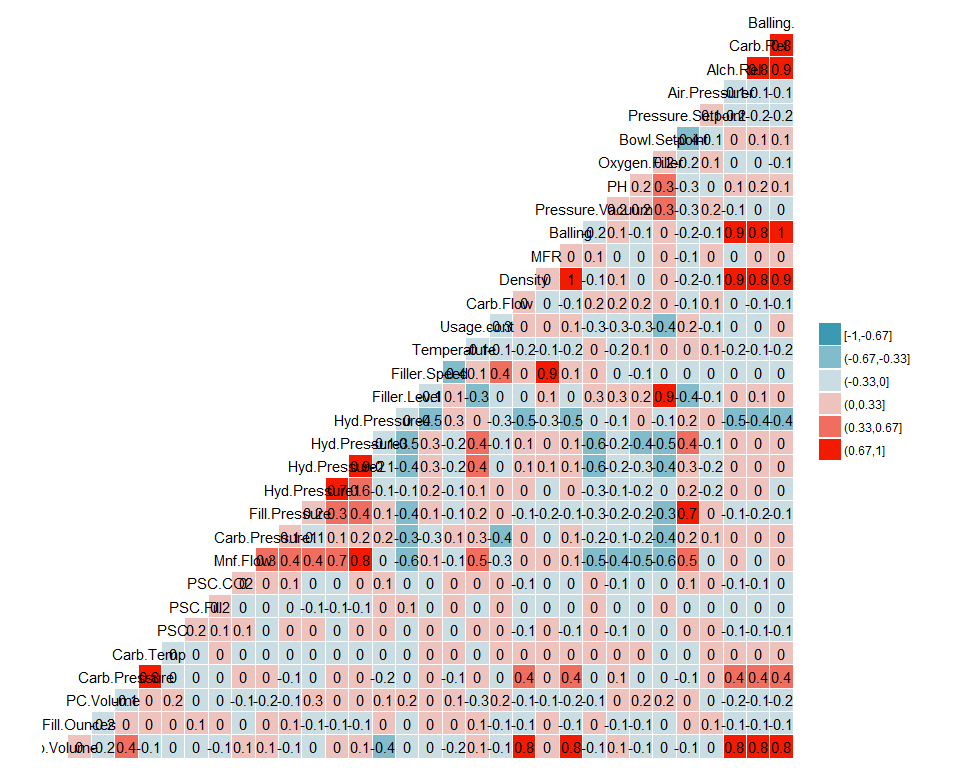


The boxplots above demonstrate the skewness of a number of variables as we have discussed.

### Correlation

The tables below represent correlation between response and predictor variables. There is high collinearity among certain variables. For instance, in Density and Alch.Rel. This is also true of Alch.Res and Brand D.

ggcorr(data1, method = "pairwise", label=TRUE, nbreaks=6)



Below is a look at the unique values for each variable. As mentioned, it appears that some of these variables represent various non-continuous settings.

apply(data1, 2, function(x)length(unique(x)))

## Data Manipulation

Data manipulation techniques were used to adjust the data for use in predictions.

### Dummy Variables

Dummy variables are used to replace categorical values, specifically where Brand Code is used.

data1$A <- ifelse(data1$Brand.Code == "A", 1, 0)

data1$B <- ifelse(data1$Brand.Code == "B", 1, 0)

data1$C <- ifelse(data1$Brand.Code == "C", 1, 0)

data1$D <- ifelse(data1$Brand.Code == "D", 1, 0)

data1 <- data1 %>% select(-Brand.Code)

toPred$A <- ifelse(toPred$Brand.Code == "A", 1, 0)

toPred$B <- ifelse(toPred$Brand.Code == "B", 1, 0)

toPred$C <- ifelse(toPred$Brand.Code == "C", 1, 0)

toPred$D <- ifelse(toPred$Brand.Code == "D", 1, 0)

toPred <- toPred %>% select(-Brand.Code)

### Handling Missing Values

As previously discussed, missing values are important to either skip or impute. After analysis and getting a deeper understanding of our data, the decision has been made to impute missing values to preserve the data and records.

#The following will impute the data. For efficiency the data sets can be loaded from github below.

# ##For Windows to run in parallel

# library(parallel)

# library(doParallel)

#

# cluster <- makeCluster(detectCores() - 1) # convention to leave 1 core for OS

# registerDoParallel(cluster)

#

# #impute missing training data

# set.seed(123)

# dfImputed <- missForest(data1, parallelize = 'forests')

# write.csv(dfImputed$ximp, "StudentDataImputedMF")

#

# #impute missing test data

# set.seed(123)

# predImputed <- missForest(toPred, parallelize = 'forests')

# write.csv(predImputed$ximp, "PredictImputedMF")

#

#

# #turn off parallel processing

# stopCluster(cluster)

# #resume use of the sequential backend

# registerDoSEQ()

#read imputed train set

imputed <- read.csv("https://raw.githubusercontent.com/624-Group2/Project-2/master/StudentDataImputedMF")

imputed <- imputed %>% select(-X)

sum(is.na(imputed))

#read imputed test set

predictImp <- read.csv("https://raw.githubusercontent.com/624-Group2/Project-2/master/PredictImputedMF")

predictImp <- predictImp %>% select(-X)

sum(is.na(predictImp))

## Predictions

With the data prepped, we used a variety of models to find the most predictive features in finding the PH levels of the batch.

### Test & Training Sets

Prior to testing our models on the actual prediction data set, it is prudent to evaluate our models against data where the response is known, so our predictions can be compared.

smp <- floor(0.70 \* nrow(imputed))

set.seed(123)

train\_index <- sample(seq\_len(nrow(imputed)), size = smp, replace = FALSE)

train\_set <- imputed[train\_index, ]

validation\_set <- imputed[-train\_index, ]

### Linear Regression Models

#### Ordinary Linear Regression

Using the train function, we can preprocess the data using methods that would reduce the issues we’ve identified in the previous stages, namely outliers and skew.

lmTune <- train(PH ~ ., data=imputed, method='lm', preProcess=c('BoxCox', 'center', 'scale'))

After training the dataset, we get an RMSE of 0.1341 and an R2 of 0.4003. The most predictive variables for this model are Hyd.Pressure4, Hyd.Pressure1, CarbPressure, Carb.Temp, and Carb.Rel.

#### Partial Least Squares

The optimal model using PLS shows an RMSE of 0.1473 and R2 of 0.213:

plsTune <- train(PH ~ ., data=imputed, method='pls', preProcess=c('BoxCox', 'center', 'scale'), tuneLength=5, trControl=ctrl)

plsTune

#### Ridge-Regression

The optimal ridge-regression model shows an RMSE of 0.1505 and R2 of 0.221:

ridgeGrid <- data.frame(.lambda=seq(0, 0.1, length=15))

ridgeTune <- train(PH ~ ., data=imputed, method='ridge', preProcess=c('BoxCox', 'center', 'scale'), tuneGrid=ridgeGrid, trControl=ctrl)

summary(ridgeTune)

### Nonlinear Regression Models

#### MARs

#### Support Vector Machines

#### Neural Network

### Regression Tree / Rule-Based Models

#### Cubist

#### Regression Tree

#### Bagged Tree

#### Random Forest

## Conclusion

# Appendix: Project Prompt

This is role playing. I am your new boss. I am in charge of production at ABC Beverage and you are a data scientist 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 report in a Word readable format and your predictions in an Excel readable format.

I also rely on a colleague for advice. She is very data savvy and can provide info on good code form to me, and just make me feel better about a technical solution. Please provide all your code and technical dialogue so she can review it. She should be able to quickly cut and paste into R studio. NOTE, include R library calls in your code.

Questions? We can discuss. As always, I am busy (boss role, not professor) so I really want you to take the ball and run with it the best you can. But, I will answer as I can. Let’s talk more in our weekly meeting next Tuesday.