Modeling and Prediction pH level in Beverage Production Process

Team 4

03/12/2019

Table of Contents

Overview 1

Introduction 1

Data 2

Data Overview 2

Data Preparation 8

Approach 10

Strategy 10

Analysis 10

Results 11

Conclusion 13

# Overview

pH plays an important role in food processing and beverages in particular. This report will explain the factors used in the production of beverages and how these factors have an impact on the pH level of these beverages. More specifically, we are examining the underlying factors of the pH level of beverages that could potentially be a powerful predictor of the pH level and finally make a predictions of the pH values of beverages based on these factors.

# Introduction

A definition of pH is the measurement of the acidity or alkalinity of a solution commonly measured on a scale of 0 to 14. pH 7 is considered neutral, with lower pH values being acidic and higher values being alkaline or caustic. pH is the most common of all analytical measurements in industrial processing and since it is a direct measure of acid content [H+], it clearly plays an important role in food processing. Among the reasons for measuring pH in food processing include:

•To produce products with consistent well defined properties

•To efficiently produce products at optimal cost

•To avoid causing health problems to consumers

•To meet regulatory requirements

Due to the logarithmic nature of the measurement, even small changes in pH are significant. The difference between pH 6 and pH 5 represents a ten-fold increase in acid concentration; a change of just 0.3 represents a doubling of acid concentration. Variations of pH can impact flavor, consistency, and shelf-life.

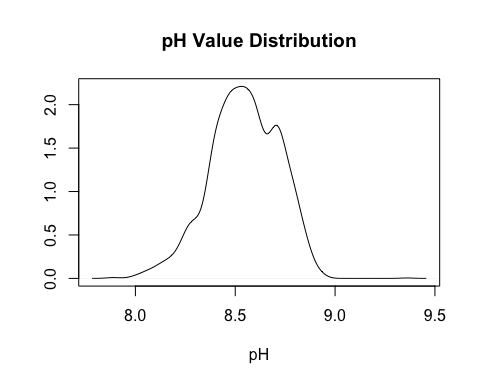
# Data

## Data Overview

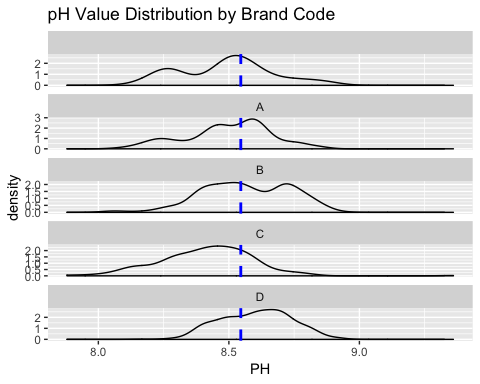
The data set was taken from a beverage manufacturing company. It consists of 2,571 rows/cases of data and 33 columns / variables. The data is represented as follows: the rows are the baverages (2,571) and the columns are the features (33 features).

The distribution of the target variable - pH level by beverage brand is shown on the histograms below. The overall pH level among the beverages is in the range of 8-9.

Target variables distribution (overall).



Target variables distribution (by Brand Code).



Our variable set includes both numeric and classification variables involved in the production process of beverages and which potentially may cause pH level variations. The full set of variables and their distribution presented below:

• Brand Code

• Carb Volume

• Fill Ounces

• PC Volume

• Carb Pressure

• Carb Temp

• PSC

• PSC Fill

• PSC CO2

• Mnf Flow

• Card Pressure1

• Full Pressure

• Hyd Pressure1

• Hyd Pressure2

• Hyd Pressure3

• Hyd Presure4

• Filler Level

• Filler Speed

• Temperature

• Usage Cont

• Carb Flow

• Density

• MFR

• Balling

• Pressure Vacuum

• Oxygen Filler

• Bowl Setpoint

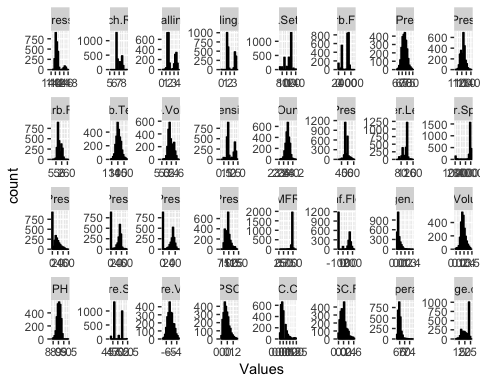
• Pressure Setpoint

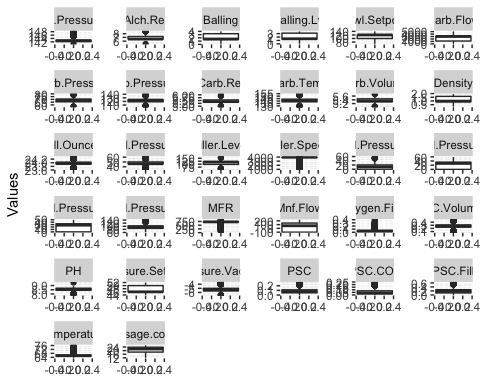
• Air Pressureer

• Alch Rel

• Carb Rel

• Balling Lvl





Many variables are highly skewed (MFR, Filler.Speed,Oxygen.Filler,Air.Pressurer). Some of the variables have close to normal distributions, for example: Fill.Ounces, Carb.Temp etc. or follow log-normal distribution, for example: PSC.Distribution and PSC.C02.Distribution.

The majority of variables are continious, but some of the predictors appear to be discrete: Pressure.Setpoint, Alch.Rel.

Hyd.Pressure1 to 3 have similar patterns with large number of 0 values. In case of collinearity between them and depending on their relationships to a target variable, they may be candidates for a removal.

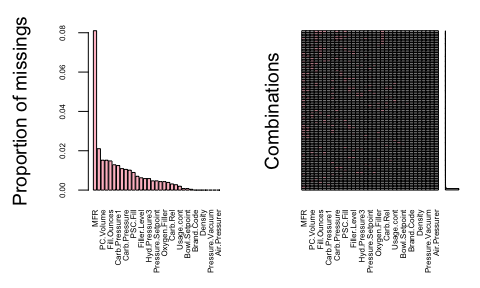
As we already detected many variables are highly skewed with large variance and we expect that many variables have outliers. Box Plots have confirmed that assumption.

As our main focus will be on tree based models, we think that data normalization is not necessary.

## Data Preparation

Main steps that were taken at the data preparation stage were handling missing values, eliminating near-zero variance predictors and checking highly correlated predictors.

Missing Values



Missing values of the target variable.

PH variable contains just 4 missing values, we have removed rows with these missing values as we think this is the most optimal solution.

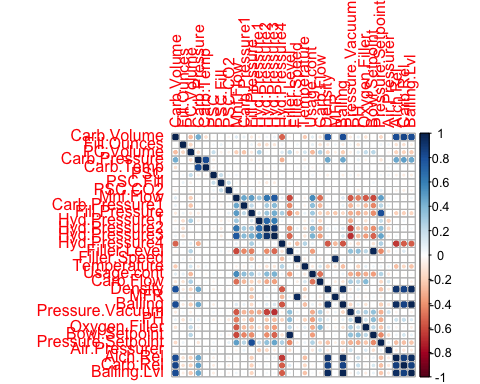
Missing values of the predictors variables.

The maximim number of missing values among all the variables of the data is 8% (MFR variable), which is not critical and imputation methds have been considered. One character variable - Brand.Code has 120 missing values. The missing values of the variables will be imputed using MICE (Multiple Imputation by Chained Equations). This is one of the most advanced methodology for performing missing data imputation. Creating multiple imputations, as opposed to single imputations to “complete” datasets, accounts for the statistical uncertainty in the imputations. In general, the limitation with single imputation is that because these techniques find maximally likely values, they do not generate entries which accurately reflect the distribution of the underlying data. Moreover, the chained equations approach is very flexible and can handle variables of varying types (e.g., continuous or binary).

Near-Zero Variance Variables

Near-Zero variance variables are variables that have one unique value or predictors that are have both of the following characteristics: they have very few unique values relative to the number of samples and the ratio of the frequency of the most common value to the frequency of the second most common value is large. One near-zero variance variable was detected - Hyd.Pressure1. This was considered as uninformative predictor and was removed from the list of predictors.

Highly correlated predictors



Correlation analysis showed that there are lots of predictors that are highly correlate between each other. Our consideration with regards to this problem is the following: correlated features will not always worsen a model, but they will not always improve it either. As speed is not an issue for this particular project, perhaps not removing these features right away can be a good choice, unless these these features are uninformative. Also main focus of the modeling part of these project will be on the tree-based models which cope with highly correlated features well.

# Approach

## Strategy

Initially, we simply run various models to obtain a better understanding of our data. Main focus was on tree based models taking into account the following properties of the data: non-linear relationships, collinearity, skewness and presence of the outliers. Also rule- and tree-based models implicitly conduct feature selection. Non-linear models such as MARS and Support Vector Machine were included for comparison reason. We trained our algorithms on the training data set and obtained our values using k-folds cross validation (setting k equals to 10). After assessing the performance of the algorithms we selected the best ones (based on the lowest average RMSE value). Best performed model was further tuned and then assessed on the test set.

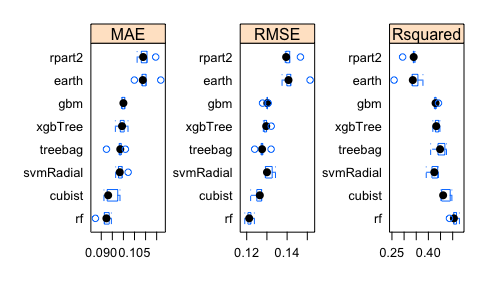
The list of models that were trained:

Non-Linear Models: MARS, Support Vector Machine (SVM with radial kernel)

Tree Based Modles: Classification And Regression Tree (CART), Bagged CART, Random Forest, Gradient boosting, Cubist, Extreme Gradient Boosting.

## Analysis

Analysis of the trained models presented below:



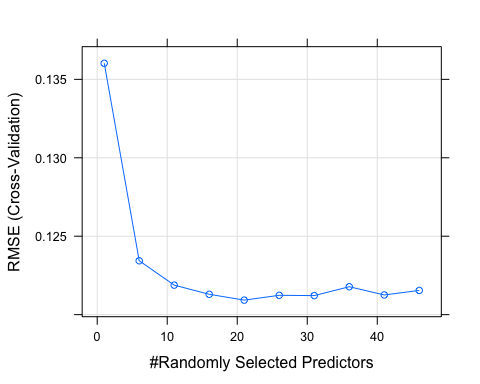
## Results

Random Forest is best performed model as it has provided optimal resampling: lowest RMSE with relatively low variation among the samples. This is quite expected result, because Random Forest is often a good choice when we need high predictive accuracy for a problem with highly correlated features.

RF models train each tree independently, using a random sample of the data. This randomness helps to make the model more robust than a single decision tree, and less likely to overfit on the training data.

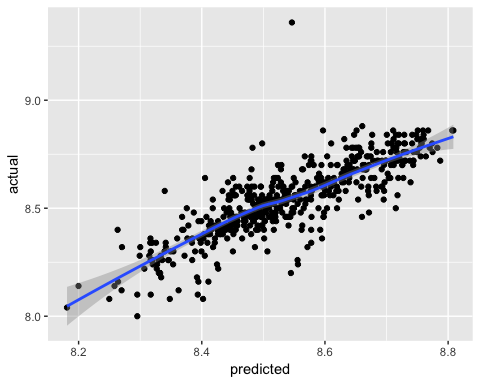
The predictions of our selected random forest model was further augmented through the use of tuning - grid search.

Best tuning parametrs are presented on the graphs below:

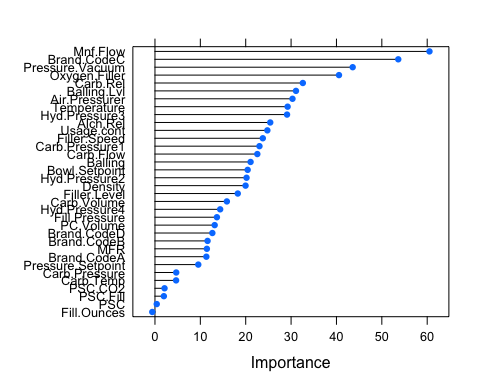


Performance of the selected tuned optimal model on the test set showed even better result - lower RMSE (0.09983016) compare to their performance on the training data, that means that most likely our model generalizes quite well and does not significantly overfit the data.

Predicted vs Actual results of the tuned Random Forest model on the test set are shown on the scatter plot below. Overall we see good results - points are pretty close to the diagonal line and are in the acceptable range of pH values.



Based on the best performed models we can conclude that the features presented on the graph below are the most important in causing variations in the level of pH in beverages.



# Conclusion

pH plays an important role in food processing and beverages in particular. Predicting pH and its maintanace on the same level is an important part of the productions process. Several predictive models were built and tuned in order to produce the most accurate pH predictions and understand most important vactors that influence pH level significantly. It has been determined that the Random Forest model provides the optimal resampling and test set performance. Random Forest model obtained the lowest value on RMSE of 0.0998 (in the units of pH) among all predictive models.

Based on the analysis performed we can conclude that the main driving factors (top 10) influencing PH are the following:

Mnf.Flow, Pressure.Vacuum, Oxygen.Filler, Carb.Rel., Balling.Lvll, Air.Pressurer, Temperature, Hyd.Pressure3, Alch.Rel, Usage.Count.

In order to maintain stable level of pH in beverages it is advisable to maintain stable work of the production processes listed above.

Predictions of the pH values are attached in the file to this report: “Team4\_Predictions.csv”