PH Production Report

Jeff Shamp, Misha Kollontai, John Kellogg

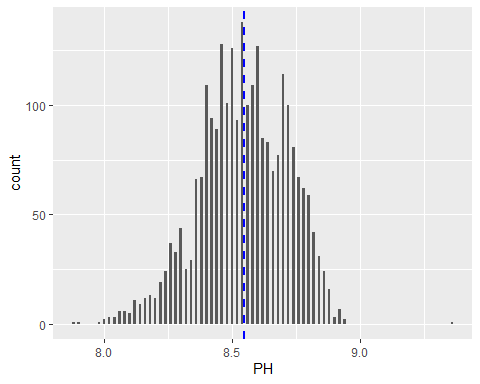
2021-05-17

## Intro

In order to better understand the manufacturing process within ABC Beverage, we endeavored to model the pH output within our plant based on all available predictive factors. The goal was to identify the predictors that are most indicative of the final pH value of our product as well as providing a model that sufficiently accurately predicts pH based on given values of predictors. This report could then be used to more efficiently steer manufacturing processes towards a desired pH level with (ideally) the minimum required changes to setting within the overall system.

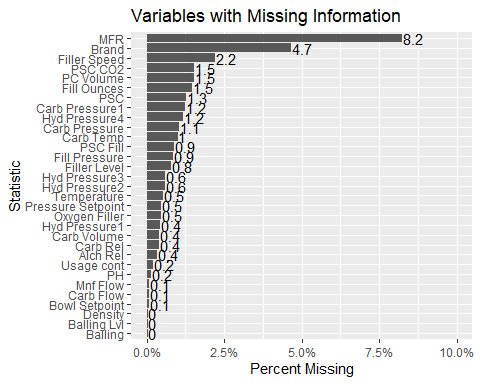
## Data Exploration

The available data for our analysis included the output pH value along with 32 predictor variables for a set of 2,571 observations. The predictor variables include one categorical variable for the Brand associated with each observation; each other predictor is a numerical setting within our manufacturing process including various pressure measures, temperature settings, fill levels, etc. In a significant portion of the available data, there is a measurement missing for at least one predictor of a given observation (1,483 of 2,571). The output variable PH is fairly normally distributed around a mean of 8.55 as can be seen in the graph below.

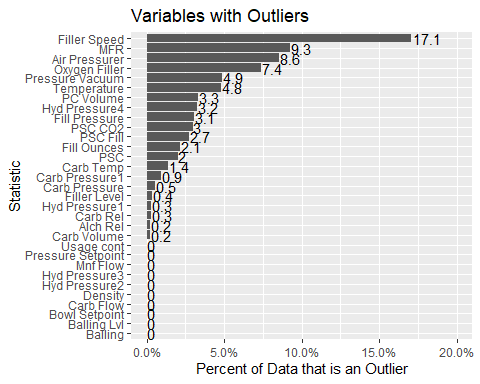


### Outliers/Erros

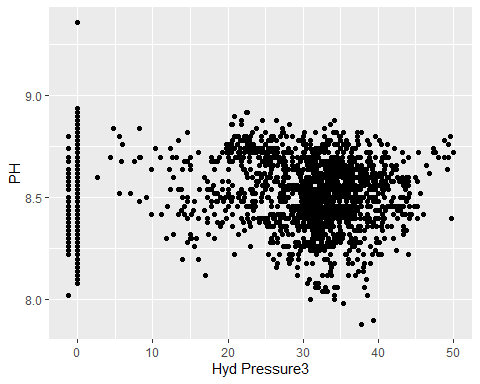
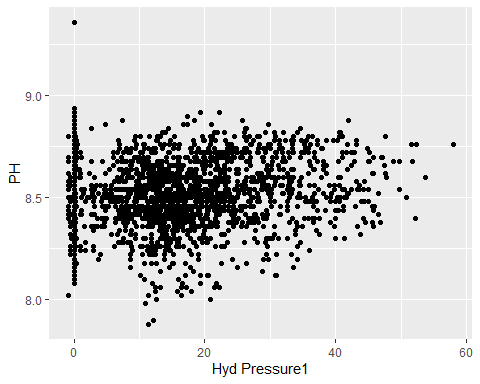
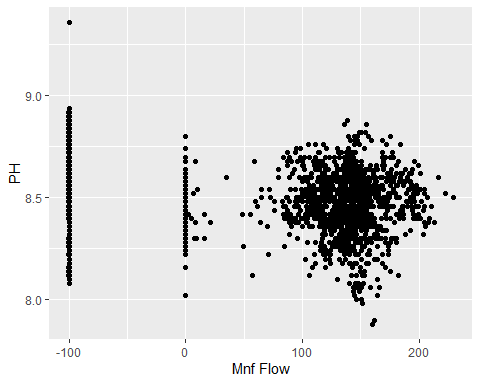
Before any models could be generated from this dataset, the large number of missing data needed to be addressed. Below we can see that 3 predictors were missing data in more than 2% of observations and 11 in at least 1% of observations.



Before imputing the “missing” data, we decided to look over some of the variables and determine if any of the information appeared erroneous. This would entail a datapoint where the value wasn’t missing, but simply did not make sense (i.e. a negative number for a variable where that is impossible). To start we looked over how many outliers were present in the predictors.



Though a significant number of predictors appears to contain a lot of outliers, not many of it appears to be “erroneous”. A slight dig into each predictor yielded three that contained datapoints that appeared to have either been recorded in error or incorrectly entered. The three variables in question are Mnf Flow, Hyd Pressure1 and Hyd Pressure3. For each of these 3 variables negative values don’t makes sense since neither flow rate nor hydraulic pressure are unlikely to be negative by design. Each negative entry for Mnf Flow is at -100 as opposed to a spread of values and the negative hydraulic pressure values are all slightly below 0.



Our solution to the seemingly erroneous data was to convert negative Hydraulic Pressure and ‘Mnf Flow’ readings with NA’s so they wiould be imputed along with the information that was missing from the get-go.

### Imputation

In order to prepare the data for modeling we first split the data into training/testing sets along an 80/20 percentage split, and utilized the tidymodels package within R to create what is known as a recipe. This recipe outlines the following steps that are then followed in order to pre-process a given dataset:

1. Specifies as the predicted variable within the dataset
2. Centers all numeric predictors, normalizing them around 0
3. Performs a BoxCox transformation on numeric predictors
4. Runs a k-Nearest Neighbor imputation on all missing data (NAs)
5. converts all nominal predictors (Brand) into dummy variables
6. Removes observations for which the predicted variable (PH) is missing

## Model Creation

After identifying erroneous data, replacing it with NAs and specifying the pre-processing approach, several models were generated. The models created and evaluated against the training data are as follows:

* Linear Regression
* Random Forest (randomForest engine)
* Random Forest (ranger engine)
* k-Nearest Neighbors
* Cubist

The models were fit to the training dataset and then each was evaluated against the test set for accuracy metrics. Below is a table summarizing 3 performance metrics for each of the 5 evaluated models: Root Mean Square Error (RMSE), R-Squared (RSQ) and Mean Absolute Error (MAE).

## RMSE RSQ MAE  
## lm 0.1304 0.3835 0.1036  
## randomForest 0.0386 0.9622 0.0286  
## ranger 0.0412 0.9599 0.031  
## kknn 0.0768 0.7982 0.0563  
## Cubist 0.0165 0.991 0.0115

Tuning the best-performing Cubist model in terms of the Committee and Neighbors parameters did not seem to improve the performance against our test set, suggesting some overfitting to the training data. We settled on using the original cubist model to move forward with highlighting the most important variables and final predictions.

### Variable Importance

## # A tibble: 34 x 2  
## Variable Importance  
## <chr> <dbl>  
## 1 Bowl.Setpoint 61   
## 2 Pressure.Vacuum 59   
## 3 Balling.Lvl 58.5  
## 4 Temperature 50.5  
## 5 Alch.Rel 47   
## 6 Density 43   
## 7 Usage.cont 42.5  
## 8 Balling 41.5  
## 9 Oxygen.Filler 40.5  
## 10 Carb.Rel 36   
## # ... with 24 more rows

Based on our model, it appears that the strongest predictors for predicting the output pH of our product are:

1. Bowl Setpoint
2. Pressure.Vacuum
3. Balling.lvl
4. Temperature
5. Alch.Rel

## Prediction

With our final Cubist model we will predict values given our test set. See attached .xlsx file.