**Purpose:**

New regulations in the pH level of beverages produced at the Cunyville plant required executive leadership to ask the data team to conduct a study analyzing the numerous factors within our manufacturing process. The team goal was to create a predictive model pertaining to which aspect of production would produce pH levels above the newly established regulations.

**Findings:**

Our predictions conclude **Bowl Setpoint**, **Vacuum Pressure, and Balling Level** are the most important predictors when a batch produces the adverse pH levels during the manufacturing process. Table of all important predictors at the end of this report.

**Data:**

The data used to build the model consisted of measurements from 32 aspects of the manufacturing process in 2,571 previous batches and their resultant pH level.

Data Risks:

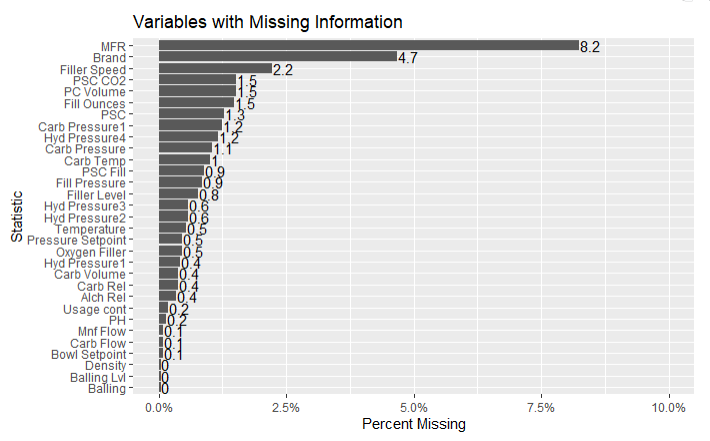
In the 32 predictors, we identified a large amount of missing data as well as a few variables with seemingly ‘erroneous’ data.

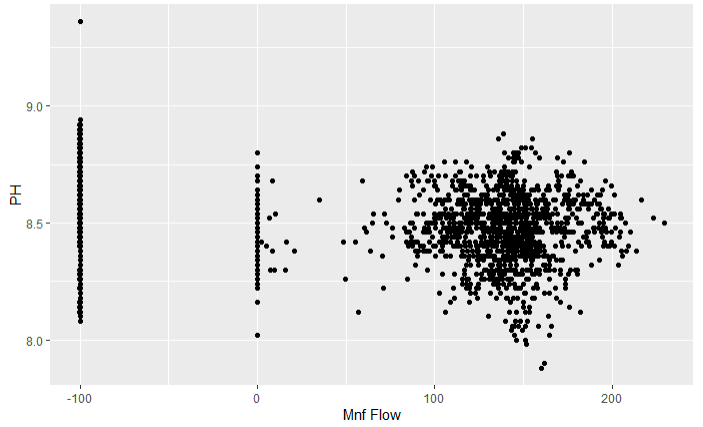
Example of erroneous/missing data

* Variable ‘Mnf Flow’ (Minimum Night Flow)
  + A large subset of the data showed a flow value of -100 even though no other measurements were negative.
* Hydraulic Pressures 1, 2, & 3
  + Large sections of the data has no value listed
  + Negative values existed where negative pressure is not possible.

Risk Resolution:

* To prevent such erroneous data from skewing our results, we removed these values and treated them as “missing” alongside the other set of missing data in other variables.
* Missing data was imputed using a mathematical technique called K-nearest neighbor. The technique takes an average of the values to either side of the missing value to create the missing one and ensure a complete dataset for our modeling.

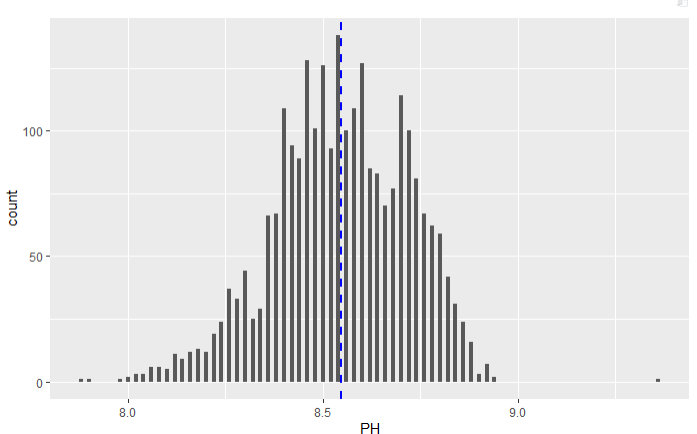




* Top graph shows the missing values by percent missing.
* Bottom graph shows the “erroneous” in Mnf Flow; make note the strong amount of data to the left of the ZERO line.

**Methodology:**

The pH level yield in the set of previous batches appeared fairly centralized (normally distributed) around an average just above 8.5 pH.

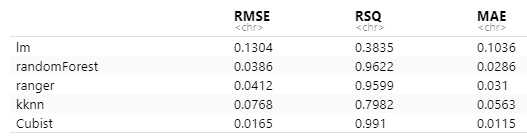


After preprocessing and cleaning as defined in the Data section, we prepared training/test sets for evaluation using the following predictive models types. We ran so many model types to ensure we could choose the best method to produce the most accurate results.

Models Ran:

* Linear Regression
* Random Forest
* Ranger (a type of Random Forest)
* K-Nearest neighbor
* Cubist

The cubist model appeared to perform the best along all three metrics evaluated:



* RMSE (Root Mean Square Error)
  + TARGET - lowest value
* RSQ (R squared)
  + TARGET - highest value
* MAE (Mean Absolute Error)
  + TARGET - lowest value

Using the cubist model, we evaluated the most important predictors as identified by our final model and ran the model against the latest set of predictors available. The list of most important predictors is shown below along with a predicted set of pH values.

|  |  |
| --- | --- |
| **Manufacturing Variable** | **Importance Score** |
| Bowl.Setpoint | 61 |
| Vacuum Pressure | 59 |
| Balling Lvl | 58.5 |
| Temperature | 50.5 |
| Alch.Rel | 47 |
| Density | 43 |
| Usage.cont | 42.5 |
| Balling | 41.5 |
| Oxygen.Filler | 40.5 |
| Carb.Rel | 36 |

\*\*NOTE\*\* Importance Score is defined by which variable had the highest effect on the model outcome. The Score is out of 100.