Analysis of the Beverage Production Factors that Impact Product pH at ABC Beverage Company

Zach Herold, Anthony Pagan, Betsy Rosalen

5/10/2020

Project Description

Data Analysis of the impact of ABC Beverage Manufacturing Process on pH

This report contains the findings of the data analysis undertaken by the data science team, lead by Zach Herold, Anthony Pagan, and Betsy Rosalen at ABC Beverage Company in order to better understand the impact of manufacturing processes on the pH level in our products and to comply with new federal regulations. The report has the following aims:

- to further senior management's general understanding of the ABC Beverage manufacturing process that impact pH;
- to internally prepare for inquiries and procedures pursuant to recent changes in the regulatory environment;
- specifically, to explicate the effect manufacturing processes have on beverage pH and a present a generalized model for predicting pH levels from input and process calibrations.

This report details the results and conclusions reached from the analysis and excludes technical details. For technical details about steps taken in our analysis, including the assumptions made, the methodology used, the models tested, and the model selection process please see the technical report, "Analysis of the Beverage Production Factors that Impact Product pH at ABC Beverage Company - Technical Report".

Data Description

Variable Summary Statistics and Distributions

We were given a dataset that consisted of 31 numerical predictor variables detailing a wide range of production processes, 1 categorical variable `Brand.Code`, and our numerical target variable, `PH`. Summary statistics for these variables are provided in Tables 1 and 2 on the next page and histograms of their distributions follow the tables.

Summary Statistics

Table 1: Summary of categorical variable, Brand.Code

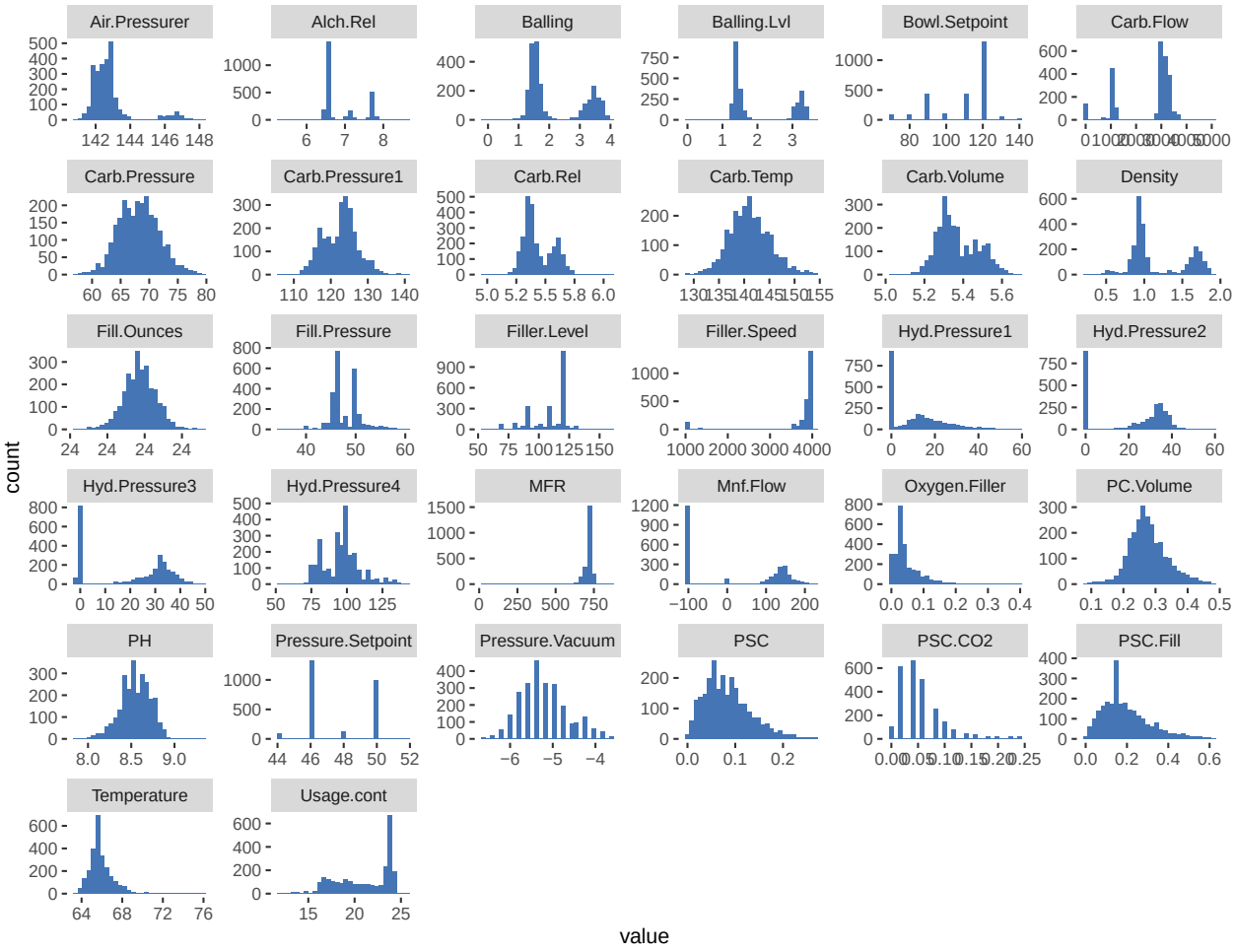
	Brand.Code
	: 120
	A: 293
	B:1239
	C: 304
	D: 615

Table 2: Summary statistics for numerical variables

	n	mean	sd	min	median	max	range	skew	kurtosis	se
PH	2567	8.55	0.17	7.88	8.54	9.36	1.48	-0.29	0.06	0.00
Carb.Volume	2561	5.37	0.11	5.04	5.35	5.70	0.66	0.39	-0.47	0.00
Fill.Ounces	2533	23.97	0.09	23.63	23.97	24.32	0.69	-0.02	0.86	0.00
PC.Volume	2532	0.28	0.06	0.08	0.27	0.48	0.40	0.34	0.67	0.00
Carb.Pressure	2544	68.19	3.54	57.00	68.20	79.40	22.40	0.18	-0.01	0.07
Carb.Temp	2545	141.09	4.04	128.60	140.80	154.00	25.40	0.25	0.24	0.08
PSC	2538	0.08	0.05	0.00	0.08	0.27	0.27	0.85	0.65	0.00
PSC.Fill	2548	0.20	0.12	0.00	0.18	0.62	0.62	0.93	0.77	0.00
PSC.CO2	2532	0.06	0.04	0.00	0.04	0.24	0.24	1.73	3.73	0.00
Mnf.Flow	2569	24.57	119.48	-100.20	65.20	229.40	329.60	0.00	-1.87	2.36
Carb.Pressure1	2539	122.59	4.74	105.60	123.20	140.20	34.60	0.05	0.14	0.09
Fill.Pressure	2549	47.92	3.18	34.60	46.40	60.40	25.80	0.55	1.41	0.06
Hyd.Pressure1	2560	12.44	12.43	-0.80	11.40	58.00	58.80	0.78	-0.14	0.25
Hyd.Pressure2	2556	20.96	16.39	0.00	28.60	59.40	59.40	-0.30	-1.56	0.32
Hyd.Pressure3	2556	20.46	15.98	-1.20	27.60	50.00	51.20	-0.32	-1.57	0.32
Hyd.Pressure4	2541	96.29	13.12	52.00	96.00	142.00	90.00	0.55	0.63	0.26
Filler.Level	2551	109.25	15.70	55.80	118.40	161.20	105.40	-0.85	0.05	0.31
Filler.Speed	2514	3687.20	770.82	998.00	3982.00	4030.00	3032.00	-2.87	6.71	15.37
Temperature	2557	65.97	1.38	63.60	65.60	76.20	12.60	2.39	10.16	0.03
Usage.cont	2566	20.99	2.98	12.08	21.79	25.90	13.82	-0.54	-1.02	0.06
Carb.Flow	2569	2468.35	1073.70	26.00	3028.00	5104.00	5078.00	-0.99	-0.58	21.18
Density	2570	1.17	0.38	0.24	0.98	1.92	1.68	0.53	-1.20	0.01
MFR	2359	704.05	73.90	31.40	724.00	868.60	837.20	-5.09	30.46	1.52
Balling	2570	2.20	0.93	-0.17	1.65	4.01	4.18	0.59	-1.39	0.02
Pressure.Vacuum	2571	-5.22	0.57	-6.60	-5.40	-3.60	3.00	0.53	-0.03	0.01
Oxygen.Filler	2559	0.05	0.05	0.00	0.03	0.40	0.40	2.66	11.09	0.00
Bowl.Setpoint	2569	109.33	15.30	70.00	120.00	140.00	70.00	-0.97	-0.06	0.30
Pressure.Setpoint	2559	47.62	2.04	44.00	46.00	52.00	8.00	0.20	-1.60	0.04
Air.Pressurer	2571	142.83	1.21	140.80	142.60	148.20	7.40	2.25	4.73	0.02
Alch.Rel	2562	6.90	0.51	5.28	6.56	8.62	3.34	0.88	-0.85	0.01
Carb.Rel	2561	5.44	0.13	4.96	5.40	6.06	1.10	0.50	-0.29	0.00
Balling.Lvl	2570	2.05	0.87	0.00	1.48	3.66	3.66	0.59	-1.49	0.02

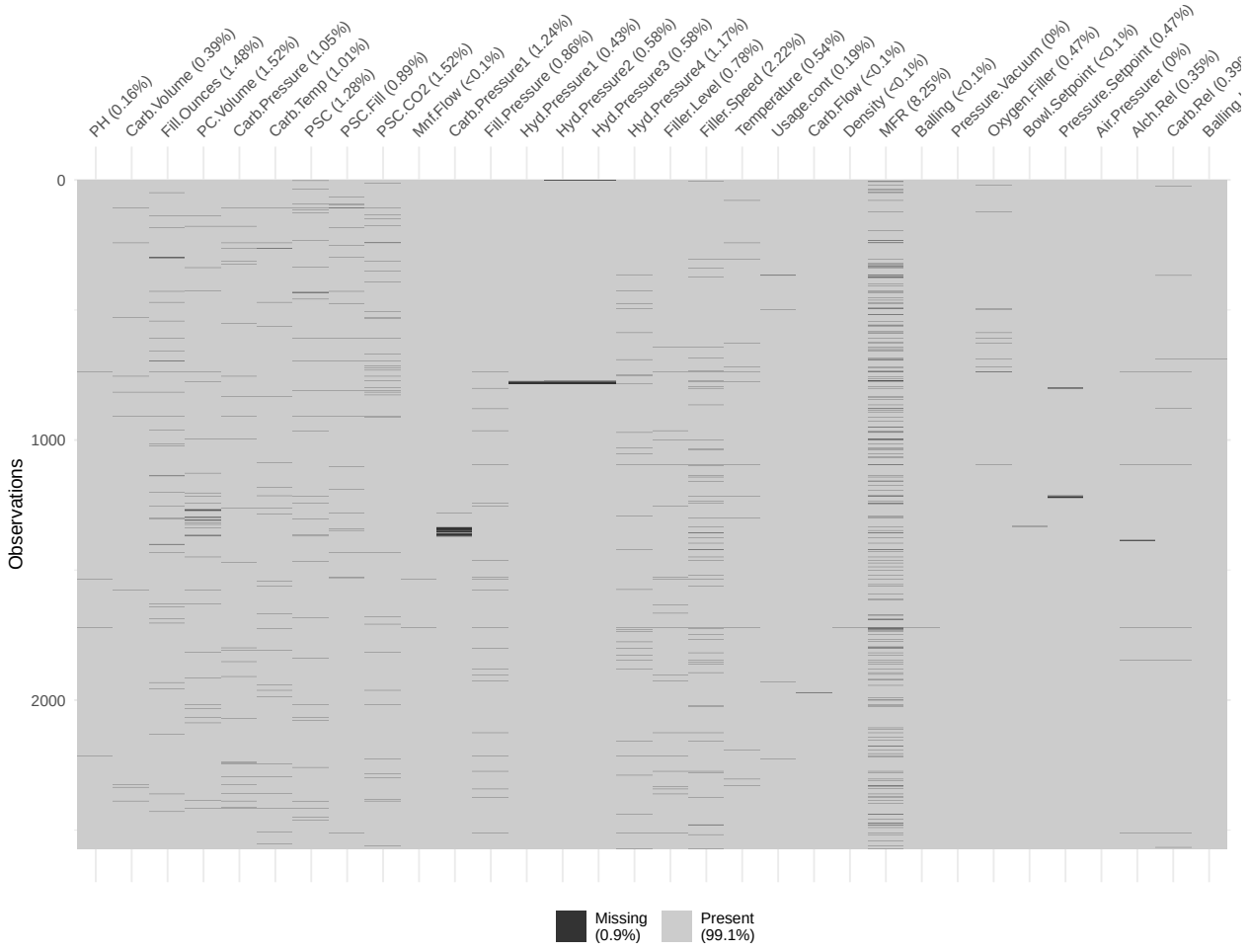
Distributions

Our predictors have a wide range of distributions with some normal, some skewed, some multi-modal, and some with high zero inflation. Our target, `PH`, has a mostly normal distribution.



Missing Values

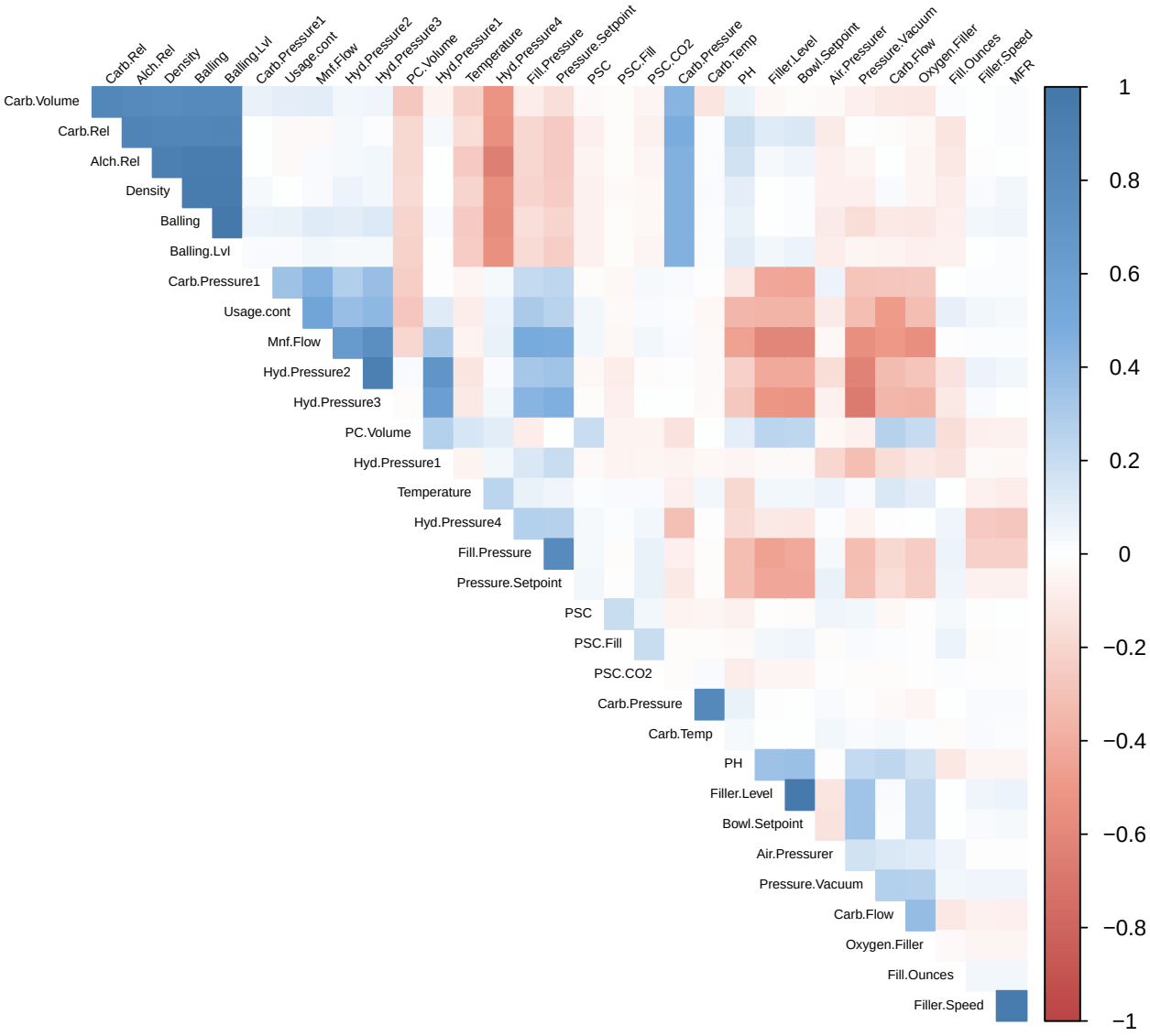
There was some missing data in our predictors most noticeably in `MFR`, which had 8.25% missing values as can be seen in the plot below. There didn't immediately seem to be any pattern in the missingness however, after our analysis we discovered some patterns which will be detailed in the section on "Understanding `Mnf.Flow`". Missing values were handled automatically in our final model algorithm, so they were not imputed or removed from the dataset.



Relationships Between Variables

Correlations

The correlation plot on the next page shows some strong correlations between predictors. Correlated predictors share the same predictive value and so can often be used interchangeably in a model. As a result, one from each pair is often removed from the model formula in order to increase model stability.



Our analysis of the highly correlated variables found 13 pairs of variables that had a correlation of 0.85 or more. These pairs are shown in Table 3 on the next page.

What we found is that there were exactly 5 variables that were most frequently associated with highly correlated pairs. Removing these variables, `Alch.Rel`, `Balling`, `Balling.Lvl`, `Carb.Rel`, and `Density` as well as `Filler.Speed`, `Hyd.Pressure2` and `Filler.Level` would eliminate all 13 highly correlated pairs in or data, however, after tuning models we discovered that we got better performance from a model type that is not as influenced by correlated predictors, called Random Forest, and by using the full set of predictors. So no variables were removed from our final model formula.

Table 3: Highly Correlated Variable Pairs

Var1	Var2	Correlation
Balling	Balling.Lvl	0.99
Filler.Level	Bowl.Setpoint	0.98
Density	Balling.Lvl	0.96
Density	Balling	0.95
Filler.Speed	MFR	0.95
Alch.Rel	Balling.Lvl	0.94
Balling	Alch.Rel	0.94
Hyd.Pressure2	Hyd.Pressure3	0.92
Density	Alch.Rel	0.92
Alch.Rel	Carb.Rel	0.88
Carb.Rel	Balling.Lvl	0.87
Balling	Carb.Rel	0.85
Density	Carb.Rel	0.85

Models

We trained and tuned a full range of model types including: Linear Regression, Ridge Regression, Lasso, Random Forest, Tree Bag, CTree, Classification and Regression Tree (CART), Multivariate Adaptive Regression Splines (MARS), K-Nearest Neighbors (KNN) and Support Vector Machine (SVM).

We chose a model that minimized error (measured as root mean squared error or RMSE) and maximized predictive value (measured as R^2 R-squared). The Random Forest model achieved this. The Random Forest regression model is a machine learning technique that randomly leaves out candidate features from each decision tree split, run on multiple iterations. In doing so, it "decorrelates" the trees, such that the averaging process can reduce the variance of the resulting models.

The accuracy measures, RMSE, R^2, and MAE, for all of the models tested are presented in the Table 4. It is ordered by the lowest (best) RMSE to highest and thus from best predictive performance to worst.

Table 4: MODELS

	RMSE	Rsquared	MAE
Random Forest	0.120	0.559	0.088
Tree Bag	0.134	0.455	0.102
SVM	0.136	0.444	0.097
MARS	0.137	0.413	0.104
KNN	0.142	0.369	0.106
Lasso	0.145	0.349	0.111
Ridge Regression	0.145	0.348	0.111
Linear Regression	0.145	0.348	0.111
CTree	0.151	0.292	0.116
CART	0.159	0.205	0.125

Random Forest Model

Top 10 Variables in the Random Forest Model by Importance Score

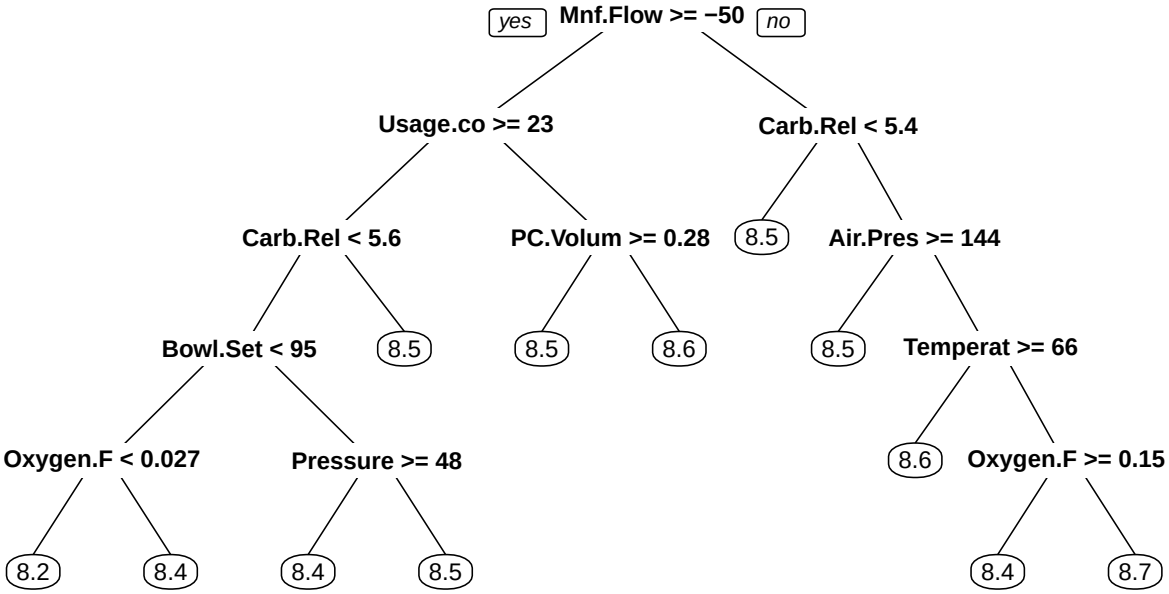
We ranked the top-ten most important variables in determining pH according to the Random Forest Model. They are shown in the Table 5.

Table 5: Variable Importance Scores

	%IncMSE	IncNodePurity
Mnf.Flow	0.01220	6.7313
Usage.cont	0.00579	4.5364
Bowl.Setpoint	0.00538	2.7553
Temperature	0.00259	2.6779
Carb.Rel	0.00380	2.4227
Filler.Level	0.00314	2.3245
Balling.Lvl	0.00311	2.1875
Oxygen.Filler	0.00262	2.1504
Alch.Rel	0.00382	2.0364
Carb.Pressure1	0.00126	1.8438

Sample Decision Tree

For comparison we also plotted a decision tree diagram which gave us similar results with the top three predictors also taking the top 3 nodes in the tree. The decision tree below is one snapshot of what a random forest model might look like. Here, we can observe that the most critical factor `Mnf.Flow` is at the top and largely negative values are associated with higher pH. To achieve lower pH we have `Usage.cont` $>= 23$, `Carb.Rel` < 5.6, `Bowl.Setpoint` < 95 and `Oxygen.Filler` < 0.027 resulting in a pH of approximately 8.2.



It's notable, however, that the tree only predicts pH levels in the middle of it's range. `PH` in our dataset ranges from 7.88 to 9.36 however our tree only predicts from 8.2 to 8.7.

Comparison of Random Forest to Standard OLS Linear Regression

When we compare the Adjusted-R^2 'goodness of fit' measure of a conventional linear regression model that uses the ten most important variables from our random forest model as predictors to one which uses all the variables as inputs, we note a very minor loss in goodness-of-fit. By removing two more statistically insignificant variables, `Pressure.Vacuum` and `Hyd.Pressure1`, we arrive at the model below:

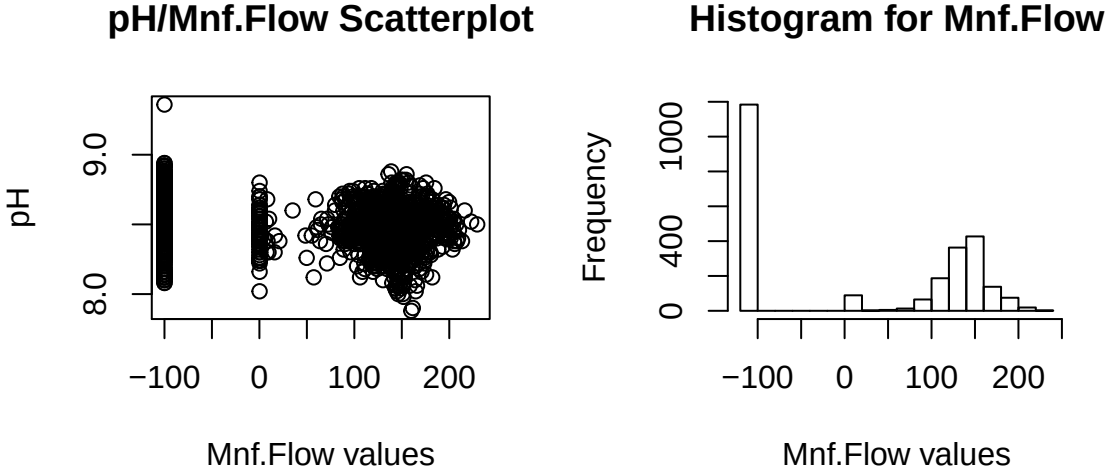
```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5903 -0.0800  0.0121  0.0939  0.3733
##
## Coefficients:
##                     Estimate Std. Error t value             Pr(>|t|)
## (Intercept)        9.4968043  0.3338057   28.45 < 0.0000000000000002 ***
## Mnf.Flow          -0.0007197  0.0000614  -11.72 < 0.0000000000000002 ***
## Usage.cont        -0.0064591  0.0014898   -4.34      0.0000155215769 ***
## Carb.Rel           0.1779166  0.0305671    5.82      0.0000000071794 ***
## Bowl.Setpoint      0.0011607  0.0003121    3.72              0.00021 ***
## Temperature       -0.0243467  0.0034953   -6.97      0.0000000000049 ***
## Oxygen.Filler     -0.4484166  0.1105615   -4.06      0.0000525618345 ***
```

```
## Pressure.Setpoint -0.0064261  0.0021279   -3.02             0.00257 **
## Hyd.Pressure3      0.0022136  0.0003730    5.93      0.0000000036612 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.139 on 1481 degrees of freedom
## Multiple R-squared:  0.322,  Adjusted R-squared:  0.318
## F-statistic: 87.8 on 8 and 1481 DF,  p-value: <0.0000000000000002
```

One interesting finding from this experiment was that we were able to determine that the impact of `Mnf.Flow`, `Usage.conf`, `Temperature`, `Oxygen.Filler`, and `Pressure.Setpoint` are negative due to the negative coefficients (in the "Estimate" column above) and the impact of `Carb.Rel`, `Bowl.Setpoint`, and `Hyd.Pressure3` are positive due to positive coefficients. So there is a balancing act between these most influential variables with some pulling in one direction on the pH and some in the other. Thus a change in one may necessitate a change in the others.

Understanding `Mnf.Flow`

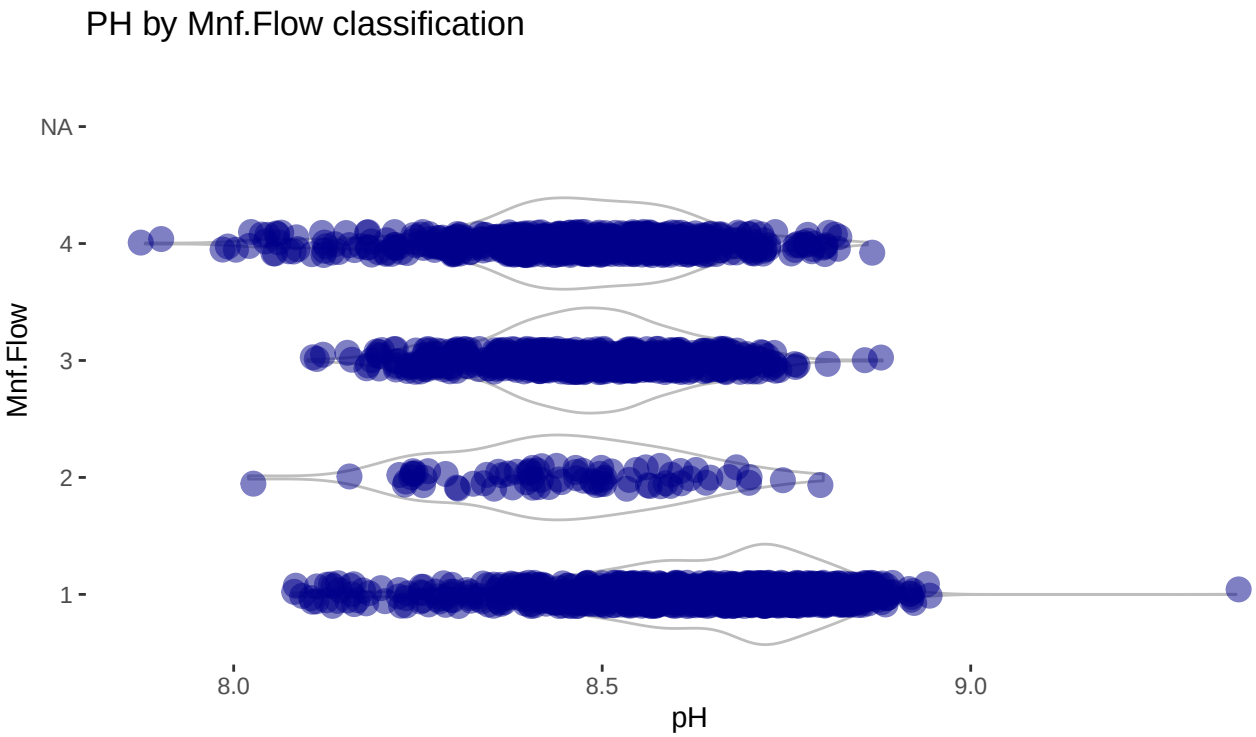
Since we found that several models, considered `Mnf.Flow` to be the most critical input, we visualized the distribution of the `Mnf.Flow` data in a histogram and in a scatterplot against `PH`:



We made the following observations about `Mnf.Flow`:

1. Of the 2567 observations in our training set, 1183 (46%) have a value of -100 or less which can be clearly seen in the histogram.
2. We note that the mean value of all non-negative `Mnf.Flow` data is 140.
3. A disproportional number of our missing values come from observations in which the `Mnf.Flow` is between 0 and 1. Although only 3% of all observations have this range of `Mnf.Flow`, 18% of the observations with missing values come from this subset.

The negative influence of `Mnf.Flow` is subtle but apparent in the violinplot below, separated by buckets of Mnf.FLow in the following ranges {1: [-1000, -1), 2: [-1, 1), 3: [1, 140), 4: [140, 1000]}



Modeling by Brand

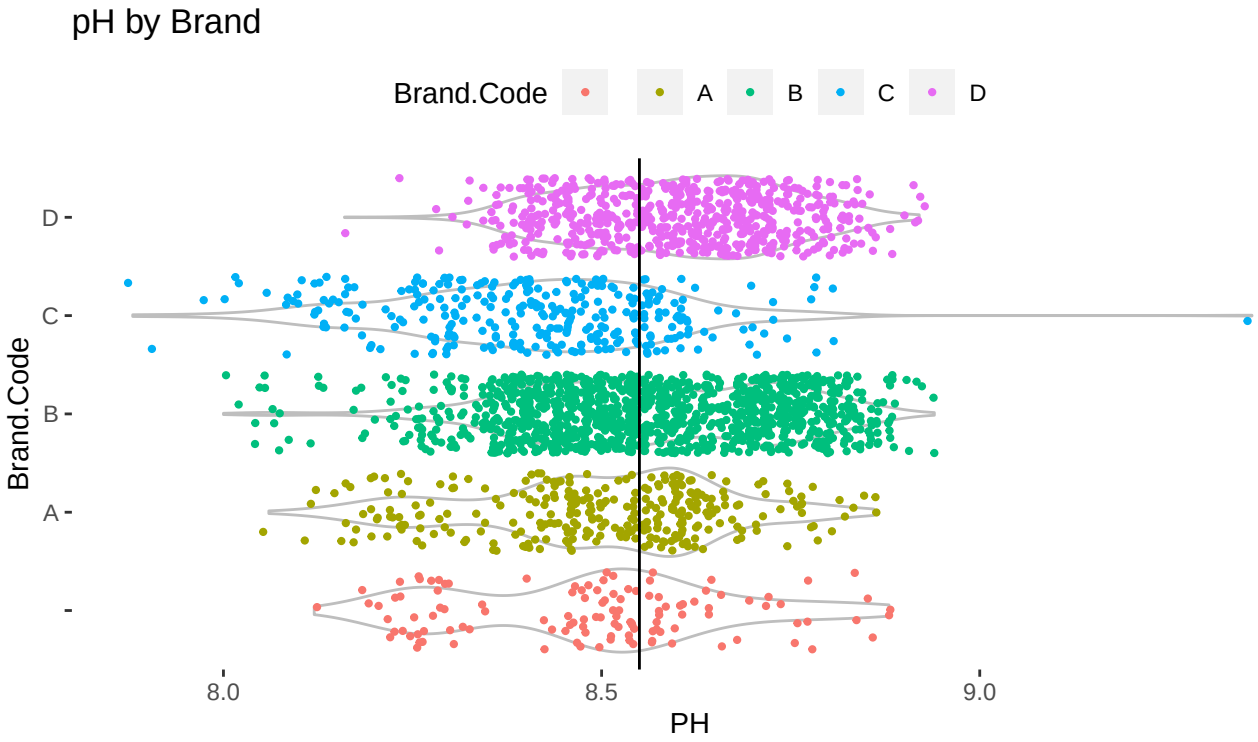
In another experiment we divided the dataset into subsets according to `Brand.Code` in order to assess what production processes are most relevant for each brand type. We imputed missing values by replacing them with the trimmed mean and then applied a random forest model to each of the four subsets. Our aim was to determine if the variables found to be most important for the whole dataset carry through to the subsets.

Interestingly the random forest model performed most poorly on the brand with the highest frequency in our dataset as can be seen in the table below.

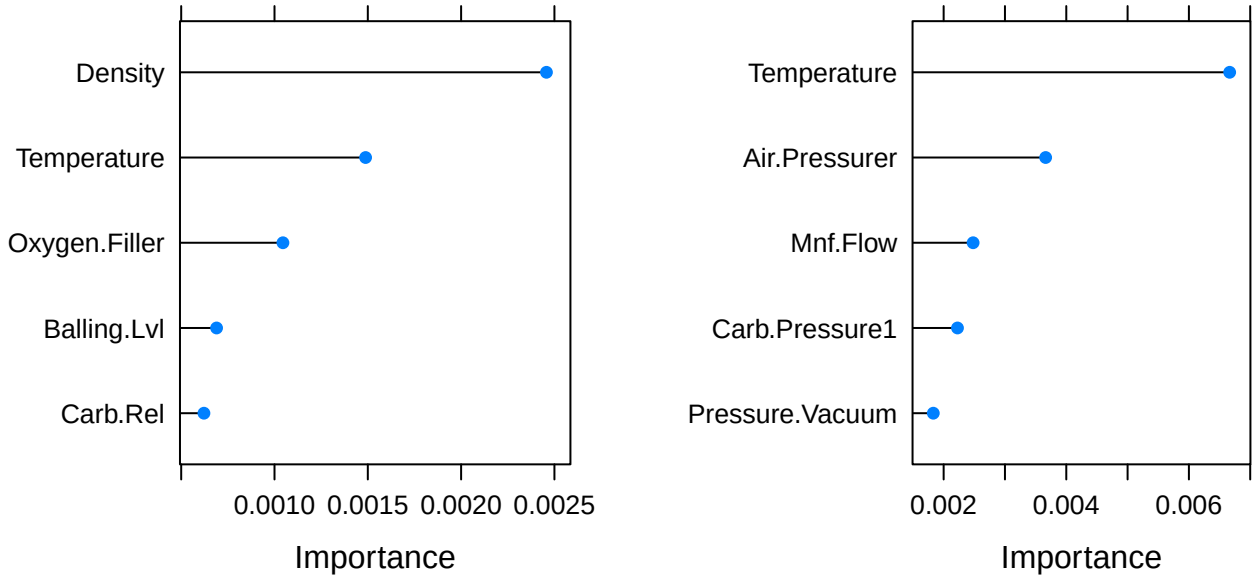
Table 6: BRANDS

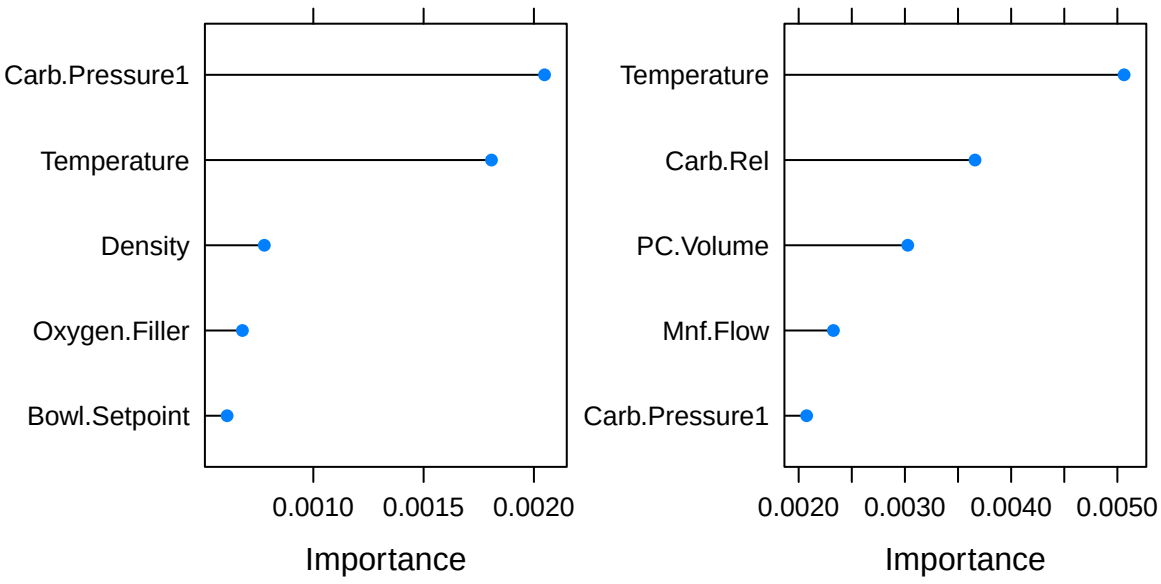
	RMSE	Rsquared	MAE	freq
C	0.16594	0.16568	0.13348	304
A	0.17294	0.12308	0.14015	293
D	0.17796	0.05285	0.13913	615
B	0.21086	0.02331	0.16769	1239

The mean pH for our dataset is 8.55, however, from the violin plot below, we observe that the distribution of pH values for Brand D tends to be above mean, while that of Brand C is markedly below mean. We further investigated what factors determine the acidic signature of Brand C, with the conclusion that lower balling method levels (which promote solution alkalinity) may at least partially contribute.



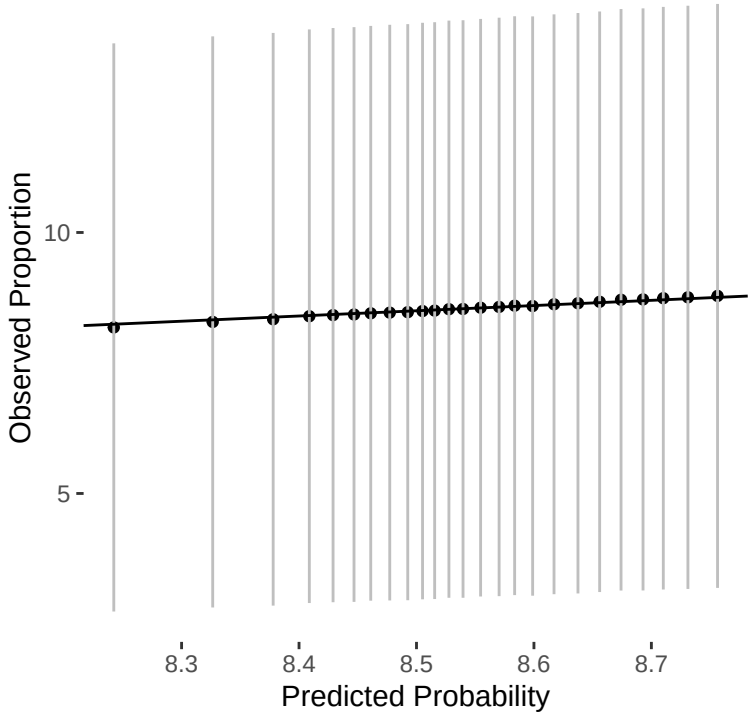
We discovered that `Mnf.Flow` is no longer the most important variable at the brand level; rather, `Temperature` is, ranking in the top five for each of the four brands. By contrast `Mnf.Flow` only shows up in the top 5 list for two brands and in the 3rd and 5th spots. These results suggest that `Mnf.Flow` may not be as robust a predictor as our other models indicated.





Predictions

The goodness of fit plot below shows that our predictors fall close to the fit line. See the accompanying csv file, "predicted_eval_values_PH.csv" for predictions of pH made by applying our Random Forest model to new data.



Conclusions

- The main processes putting downward (acetic) pressure on pH are `Mnf.Flow`, `Usage.conf`, `Temperature`, `Oxygen.Filler`, and `Pressure.Setpoint` when increased; Positive adjustment may be attained through increase in `Carb.Rel`, `Bowl.Setpoint`, and `Hyd.Pressure3`;
- There is strong correlation between several of the manufacturing processes, in particular: `MFR`, `Hyd.Pressure2`, `Carb.Rel`, `Air.Pressurer`, `Carb.Flow`, `Hyd.Pressure4`, and `Filler.Level`;
- Some of the observations have missing data in our predictors, most noticeably in `MFR`, which had 8.25% missing values, as well as `Mnf.Flow` when in the range of 0 to 1.
- The metric most highly-correlated with `PH`, `Mnf.Flow`, has an irregular tri-modal distribution, with approx. 46% of values -100 or less, indicative of a distinct qualitative process of itself. Barring the negative and near-zero values, the positive values, which are approximately normal in distribution, have little correlation to `PH`. `Mnf.Flow`'s statistically significant predictive value can be wholly distilled from its transformation into a three-class categorical variable.
- When the entire dataset is subsetted according to `Brand.Code`, a different series of critical variables emerges for each class from those of the general model. `Mnf.Flow` loses its force as a predictor, while `Temperature` and `Air.Pressurer` become key, ranking in the top five most important variables for each of the four brands under a random forest model.
- pH varies with brand profile, especially in the case of Brand D (tending to be above-mean) and Brand C (markedly below mean). We further investigate what factors determine the acidic signature of Brand C, with the conclusion that lower `balling` method levels (which promote solution alkalinity) may at least partially contribute.

Recommendations for Further Analysis

Since we had success in strengthening `Mnf.Flow`'s predictive value by transforming it into a categorical variable, we may want to investigate using the same transformation on some of the other variables with multi-modal distributions. Some of the variables with multi-modal distributions include: `Alch.Rel`, `Balling`, `Balling.Lvl`, `Carb.Flow`, `Carb.Rel`, `Density` and all three `Hyd.Pressure` variables. In addition, rather than transforming these variables, we may want to investigate using piecewise linear or MARS models with finer tuning in order to preserve the distributions in each bin.