# Documenting PH levels

(Non-technical report)

Jun Yan, Ritesh Lohiya, Justin Herman

May 9, 2019

## Problem

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.

## Process

At ABC beverage we take our porducts and government compliance issues extremely seriously. As such, my data science team, has conducted a thourough analysis of the PH levels in our mineral water. You can see the detailed approach my team took here . The purpose of this report is to document some of the findings in a digestible manner for both the executives and government compliance officials.

We collected over 2800 samples of our beverages and documented over 33 different measurements of our manufactoring processes. Below I present to you a chart as 10 examples of our processes are recorded and measured

| Brand Code | Carb Volume | Fill Ounces | PC Volume | Carb Pressure | Carb Temp | PSC | PSC Fill | PSC CO2 | Mnf Flow | Carb Pressure1 | Fill Pressure | Hyd Pressure1 | Hyd Pressure2 | Hyd Pressure3 | Hyd Pressure4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| B | 5.34 | 23.97 | 0.26 | 68.2 | 141.2 | 0.10 | 0.26 | 0.04 | -100 | 118.8 | 46.0 | 0 | NA | NA | 118 |
| A | 5.43 | 24.01 | 0.24 | 68.4 | 139.6 | 0.12 | 0.22 | 0.04 | -100 | 121.6 | 46.0 | 0 | NA | NA | 106 |
| B | 5.29 | 24.06 | 0.26 | 70.8 | 144.8 | 0.09 | 0.34 | 0.16 | -100 | 120.2 | 46.0 | 0 | NA | NA | 82 |
| A | 5.44 | 24.01 | 0.29 | 63.0 | 132.6 | NA | 0.42 | 0.04 | -100 | 115.2 | 46.4 | 0 | 0 | 0 | 92 |
| A | 5.49 | 24.31 | 0.11 | 67.2 | 136.8 | 0.03 | 0.16 | 0.12 | -100 | 118.4 | 45.8 | 0 | 0 | 0 | 92 |

| Filler Level | Filler Speed | Temperature | Usage cont | Carb Flow | Density | MFR | Balling | Pressure Vacuum | Oxygen Filler | Bowl Setpoint | Pressure Setpoint | Air Pressurer | Alch Rel | Carb Rel | Balling Lvl | PH |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 121.2 | 4002 | 66.0 | 16.18 | 2932 | 0.88 | 725.0 | 1.40 | -4.0 | 0.02 | 120 | 46.4 | 142.6 | 6.58 | 5.32 | 1.48 | 8.36 |
| 118.6 | 3986 | 67.6 | 19.90 | 3144 | 0.92 | 726.8 | 1.50 | -4.0 | 0.03 | 120 | 46.8 | 143.0 | 6.56 | 5.30 | 1.56 | 8.26 |
| 120.0 | 4020 | 67.0 | 17.76 | 2914 | 1.58 | 735.0 | 3.14 | -3.8 | 0.02 | 120 | 46.6 | 142.0 | 7.66 | 5.84 | 3.28 | 8.94 |
| 117.8 | 4012 | 65.6 | 17.42 | 3062 | 1.54 | 730.6 | 3.04 | -4.4 | 0.03 | 120 | 46.0 | 146.2 | 7.14 | 5.42 | 3.04 | 8.24 |
| 118.6 | 4010 | 65.6 | 17.68 | 3054 | 1.54 | 722.8 | 3.04 | -4.4 | 0.03 | 120 | 46.0 | 146.2 | 7.14 | 5.44 | 3.04 | 8.26 |

*Figure 1-First Five Observations*

Our main target (variable we want to explore) is PH. Looking at PH above, it seems like it is mostly in the 8-9 range. As you can see on figure 2 below, this is an accurate assumption. Our average PH is around 8.5, and the target is normally distributed. Normally distributed variables are great for predictive purposes, as prediction is a mathematically intensive process and normally distributed variables meet many of the assumptions needed to run complex models

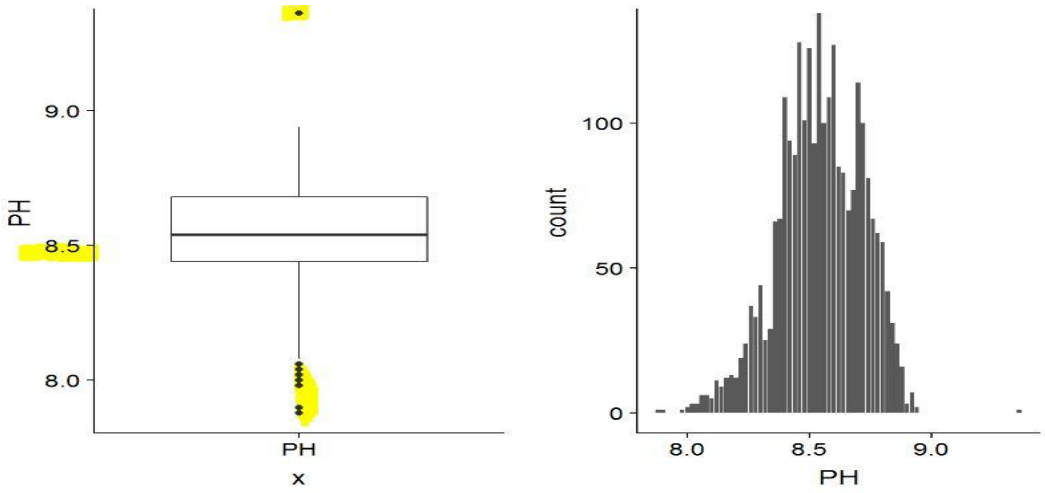


*Figure 2-Boxplot and Histogram of PH*

As you are aware, due to government regulations we need to give a report on our entire manufacturing process and how it effects our PH. Our goal is to use the data in the tables above, to predict future levels of PH.

## Exploratory Data Analysis (EDA)

Through exploratory data analysis, we can better understand our data. Healthy predictions are dependent upon healthy data. The first step we take in our EDA is understanding how our predictors (our measurements) correlate with our target.

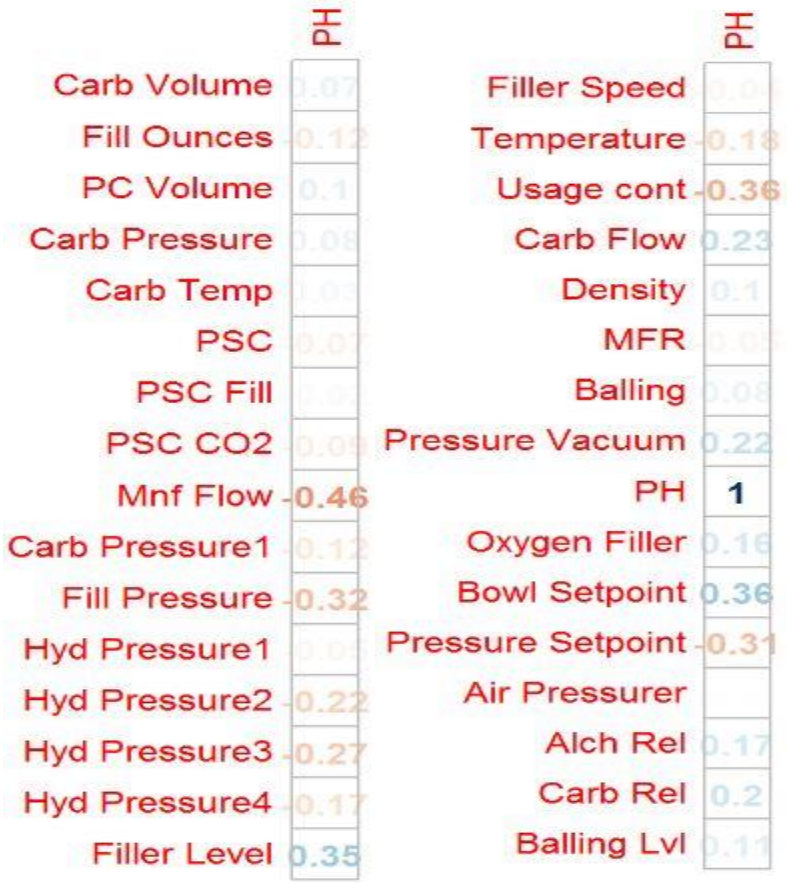


*Figure 3- Measuring PH Correlations*

Understanding correlation is more of an art than a science, but correlations that matter less, are faded out in the above graph. We can assume that MNF flow, Fill. Pressure, Filler Level, , Usage.Cont, Bowl. Setpoint and Pressure. Setpoint are all very impactful on our PH levels. We should keep this in mind when looking at the results of our models.

## Internal Notes on Data Collection (EDA)

My team was not in charge of the data collection. In terms of missing values, the data collection team did an excellent job. There were only 4 observations where our PH wasn't recorded in the entire dataset. Below you can see a printout of our dataset, where the red values are our missing values. The chart of the left indicates that out MFR measurement(predictors) had about 8% of the data missing. Typically, this would be on the threshold for discarding a predictor. Overall, the other predictors are well within statistical limits. We decided not to drop any of this data, and we ran KNN imputation to fill in the missing data. This method can predict how to best fill in our missing data based upon KNN classification. Please see the Technical report for further details on how we tested imputation methods and their application in our models.

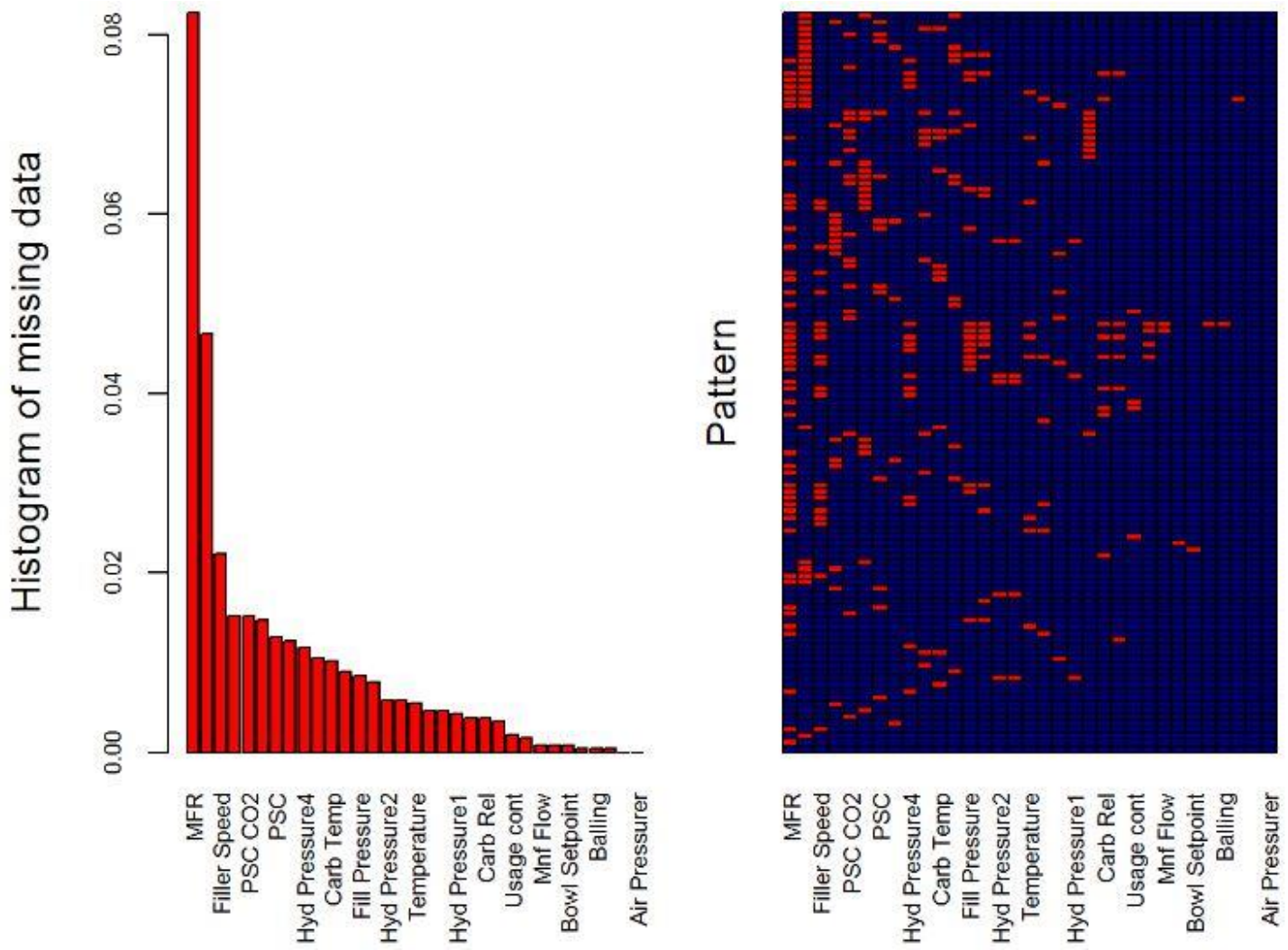


*Figure 4- Missing Data*

## Potential Issues with Data Collection

(This section of the report is for internal executives only)

We attempted to contact the data collection team, but we were unable to receive a satisfactory answer to some of the questions we have found within the data. On the following page in figure 5, Bowl. Setpoint seems to be measured in intervals of 10 from 70-120, and then the variable begins to take on several different even integer values. Pressure set point seems to be measured by integers, and then there are 3 observations between 46 and 48. PSC.CO2 seems to be suffering from a clear rounding issue. Observations of .1999999996 and of .02000000005 are very likely to both represent a measurement of .2. We want to document these issues, so that when future data collection occurs, we can prevent such issues. In terms of our current modeling approach, we kept the data as is, and believe it is unlikely that correcting the data would have resulted in materially different results

| Bowl.Setpoint | Freq |
|---|---|
| 70 | 99 |
| 80 | 96 |
| 90 | 434 |
| 100 | 112 |
| 110 | 437 |
| 120 | 1307 |
| 122 | 1 |
| 126 | 10 |
| 130 | 51 |
| 134 | 2 |
| 140 | 20 |

| Pressure.Setpoint | Freq |
|---|---|
| 44 | 96 |
| 46 | 1322 |
| 46.4 | 1 |
| 46.6 | 1 |
| 46.8 | 1 |
| 48 | 125 |
| 50 | 1002 |
| 52 | 11 |

| PSC.CO2 | Freq |
|---|---|
| 0 | 108 |
| 0.0199999999999996 | 325 |
| 0.0200000000000005 | 288 |
| 0.0399999999999991 | 37 |
| 0.04 | 624 |
| 0.0599999999999996 | 283 |
| 0.0600000000000005 | 219 |
| 0.0799999999999992 | 23 |
| 0.08 | 234 |
| 0.0999999999999996 | 79 |
| 0.100000000000001 | 65 |
| 0.119999999999999 | 10 |
| 0.12 | 72 |
| 0.14 | 30 |
| 0.140000000000001 | 19 |
| 0.159999999999999 | 3 |
| 0.16 | 36 |
| 0.18 | 12 |
| 0.180000000000001 | 8 |
| 0.199999999999999 | 2 |
| 0.2 | 16 |
| 0.22 | 22 |
| 0.24 | 17 |

*Figure 5- Irregularities in Data*

## Modeling

We attempted several different modeling solutions. All our models are fivefold cross validated and tuned for best performance. Cross validation statistically tests our model against different samples of the data so that we can be confident the model will be predictive of unforeseen data. In terms of Linear models, we ran a PLS and an Elastic Net model. We use RMSE as our error measurement, but others were reported.

Linear models suffer from an inability to generalize nonlinear trends. These linear models allow us more flexibility to fit our models to the data. As seen above in figure 6, the PLS RMSE error is around.137 and

the elastic net is around .140. These models have been tuned to select their best hyperparameters. Hyperparameters allow us to determine optimal points in which our model has minimized bias (prediction accuracy on our dataset), while minimizing variance (ability to predict unforeseen data)

We tested two nonlinear models as well (KNN(RMSE=.127), SVM (RMSE= .135)).

We conclude our modeling process by testing out tree-based models. Random Forest and XgBoost were chosen. XgBoost does not actually need data imputation, so we tested models with and without imputation. Preliminary results indicated that imputation models performed better (please see technical writeup). Our Tree models produced the best results of all models. Below you can see out Random Forest, which reaches our lowest error rate of .117

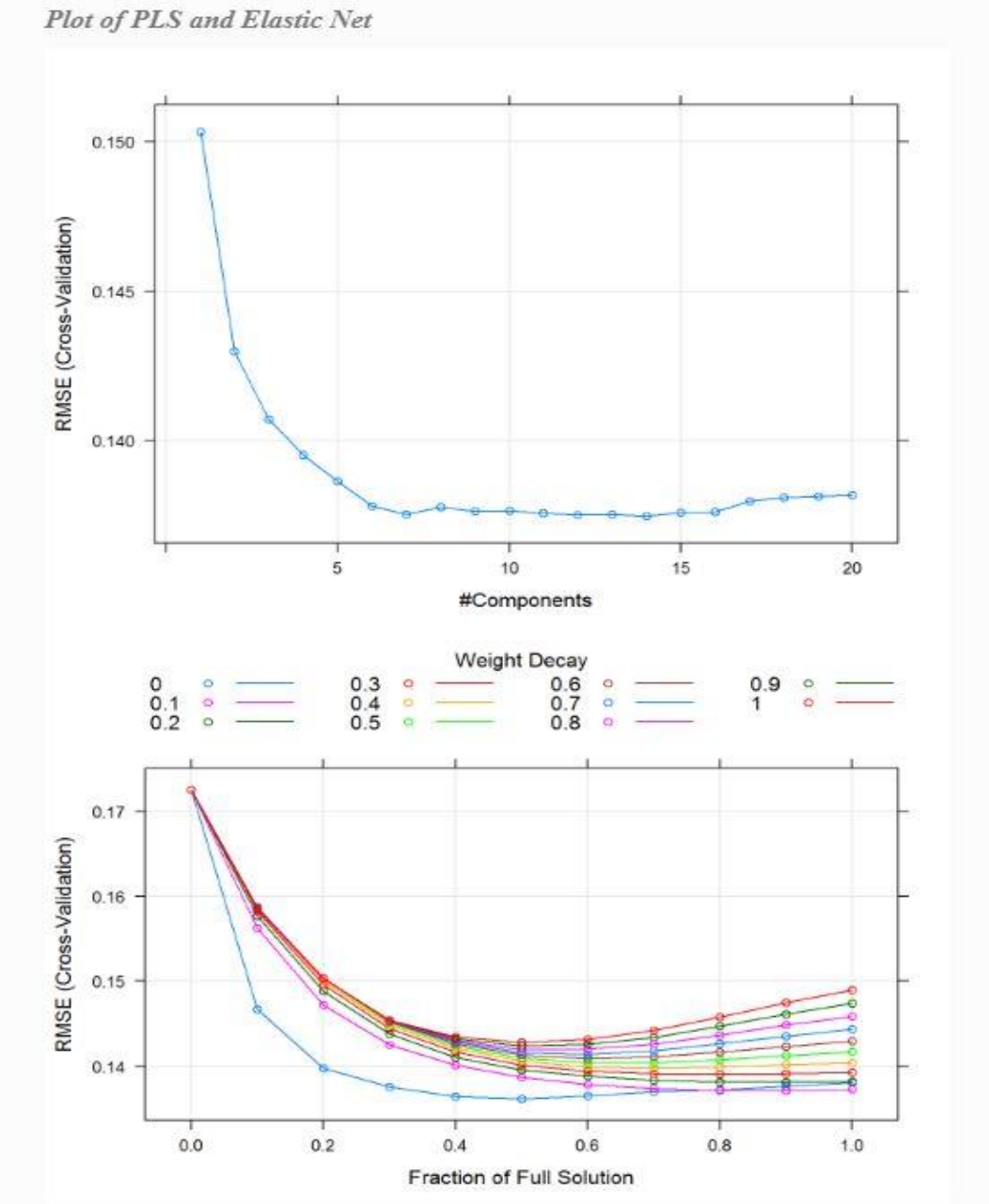


*Figure 6- PLS (top) & Elastic Net(bottom) CV tuning*

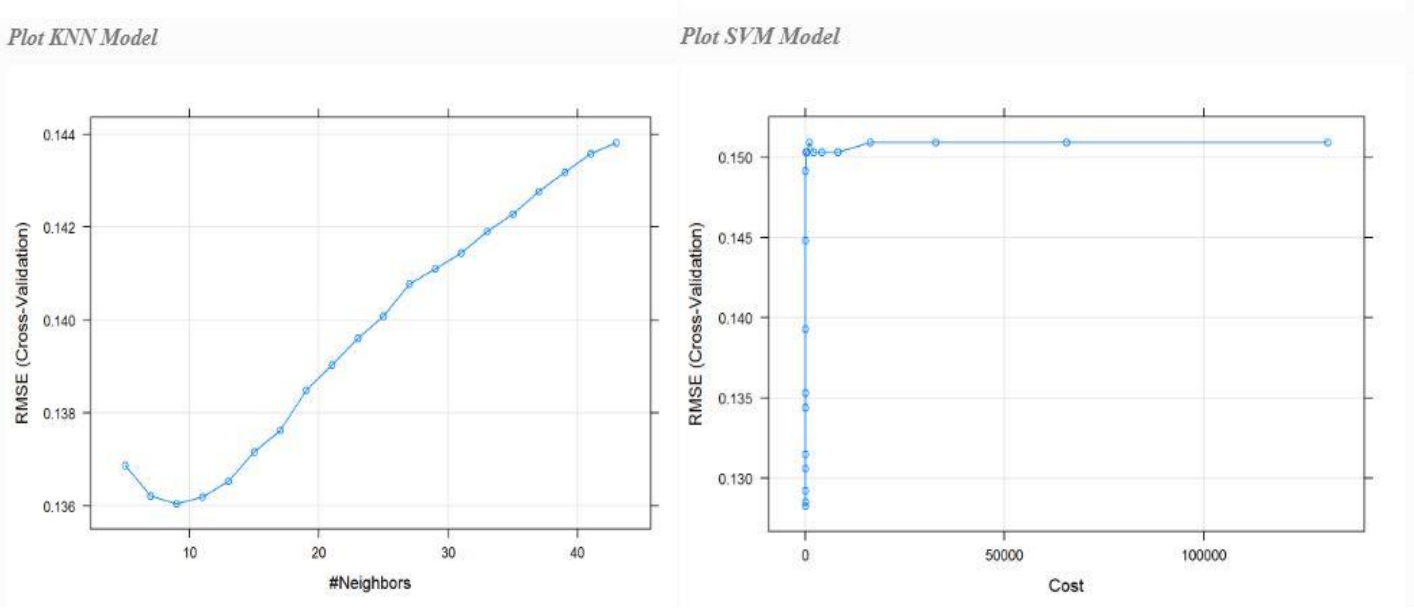


*Figure 7-Nonlinear Model Tuning-KNN(left) SVM(right)*

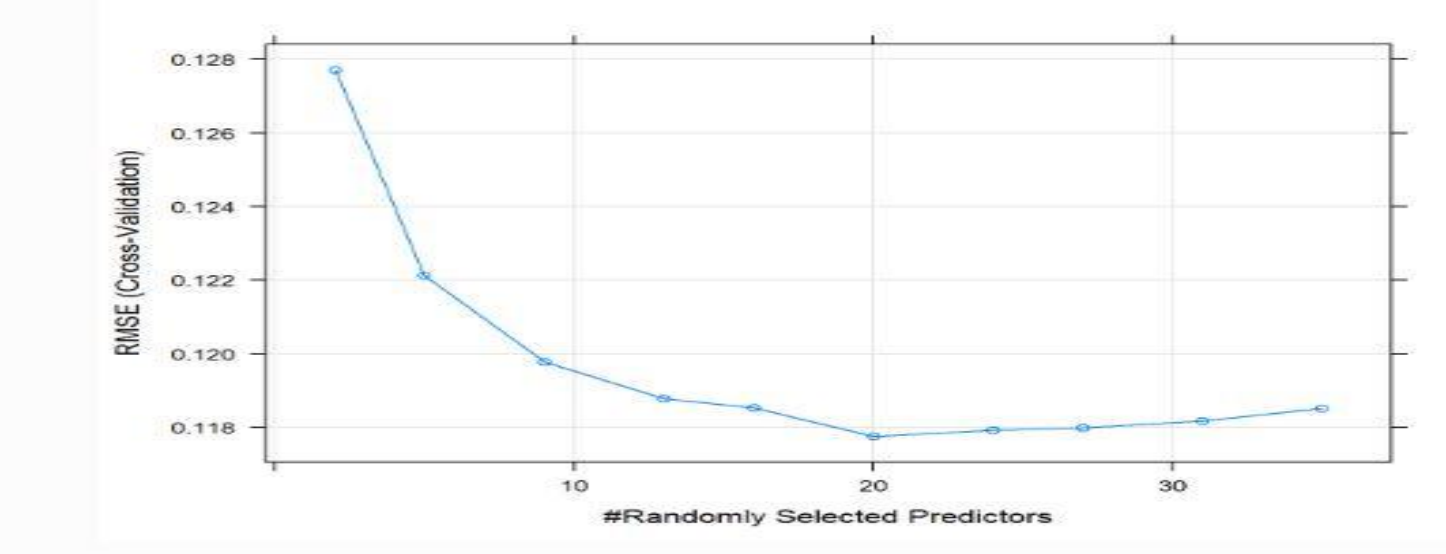


*Figure 8-Random Forest*

The XgBoost model (RMSE= .116) doesn't produce a graph, but as you can see from the table summarizing all the results from our models (figure 9), XgBoost produces the strongest cross validated performance.

```
Models: PLS, ENet, KNN, SVM, RF, XGB
Number of resamples: 4

MAE
          Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
PLS  0.10401017 0.10573777 0.10715156 0.10662121 0.10803500 0.10817153
ENet 0.10325386 0.10502571 0.10631413 0.10584920 0.10713763 0.10751469
KNN  0.10357093 0.10386052 0.10401850 0.10396019 0.10411817 0.10423283
SVM  0.09480799 0.09557060 0.09584557 0.09591785 0.09619282 0.09717227
RF   0.08744219 0.08787527 0.08848266 0.08871182 0.08931921 0.09043978
XGB  0.08697277 0.08713229 0.08730968 0.08735522 0.08753261 0.08782876
     NA's
PLS     0
ENet    0
KNN     0
SVM     0
RF      0
XGB     0

RMSE
          Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
PLS  0.1356567 0.1371824 0.1377113 0.1374039 0.1379328 0.1385362    0
ENet 0.1339212 0.1356141 0.1362803 0.1357745 0.1364407 0.1366161    0
KNN  0.1344314 0.1345692 0.1349825 0.1357495 0.1361629 0.1386017    0
SVM  0.1259120 0.1265990 0.1273402 0.1278324 0.1285736 0.1307371    0
RF   0.1165398 0.1172607 0.1177933 0.1177503 0.1182830 0.1188748    0
XGB  0.1156593 0.1157390 0.1161317 0.1164846 0.1168773 0.1180158    0

Rsquared
          Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
PLS  0.3625243 0.3656883 0.3683503 0.3689665 0.3716286 0.3766411    0
ENet 0.3700989 0.3761005 0.3796023 0.3791235 0.3826253 0.3871907    0
KNN  0.3790417 0.3836021 0.3902105 0.3906420 0.3972504 0.4031051    0
SVM  0.4293764 0.4446522 0.4546945 0.4522908 0.4623331 0.4703980    0
RF   0.5379883 0.5399803 0.5430877 0.5458798 0.5489872 0.5593555    0
XGB  0.5351413 0.5400408 0.5417958 0.5459440 0.5476990 0.5650431    0
```

*Figure 9-Overall Model performance*

## Variable Importance

The results of our model are rather encouraging. Our RMSE is extremely low relative to our target variable. We are unsure as of now what regulatory burden will be placed on our company, however, we can confidentially predict our PH levels given the data. An RMSE of .116 means that on average the error squared of any prediction is within .116 of PH. Assuming we need to eventually adjust the PH levels of our mineral water, we have confidentially identified the predictors which most influence our PH. Below you can see the variable importance from all the models we ran.

| Rank | pls | rf | xgbTree | enet | knn | svmRadial |
|---|---|---|---|---|---|---|
| 1 | MnfFlow | MnfFlow | MnfFlow | MnfFlow | MnfFlow | MnfFlow |
| 2 | BrandCodeC | BrandCodeC | Usagecont | MFR | MFR | MFR |
| 3 | BowlSetpoint | PressureVacuum | BrandCodeC | BowlSetpoint | BowlSetpoint | BowlSetpoint |
| 4 | Usagecont | OxygenFiller | OxygenFiller | FillerLevel | FillerLevel | FillerLevel |
| 5 | FillerLevel | BallingLvl | AlchRel | PressureSetpoint | PressureSetpoint | PressureSetpoint |
| 6 | HydPressure3 | AirPressurer | Temperature | Usagecont | Usagecont | Usagecont |
| 7 | PressureSetpoint | AlchRel | PressureVacuum | CarbFlow | CarbFlow | CarbFlow |
| 8 | FillPressure | CarbRel | FillerSpeed | BrandCodeC | BrandCodeC | BrandCodeC |
| 9 | BrandCodeB | Usagecont | CarbPressure1 | HydPressure3 | HydPressure3 | HydPressure3 |
| 10 | Temperature | Temperature | CarbRel | FillPressure | FillPressure | FillPressure |
| 11 | HydPressure2 | CarbFlow | AirPressurer | PressureVacuum | PressureVacuum | PressureVacuum |
| 12 | PressureVacuum | FillerSpeed | BowlSetpoint | HydPressure2 | HydPressure2 | HydPressure2 |
| 13 | CarbPressure1 | HydPressure3 | CarbFlow | CarbRel | CarbRel | CarbRel |
| 14 | OxygenFiller | CarbPressure1 | Balling | HydPressure4 | HydPressure4 | HydPressure4 |
| 15 | CarbFlow | Density | BallingLvl | Temperature | Temperature | Temperature |
| 16 | BrandCodeD | BowlSetpoint | FillerLevel | BrandCodeD | BrandCodeD | BrandCodeD |
| 17 | AlchRel | HydPressure2 | Density | OxygenFiller | OxygenFiller | OxygenFiller |
| 18 | CarbRel | Balling | PCVolume | FillerSpeed | FillerSpeed | FillerSpeed |
| 19 | BrandCodeA | PCVolume | MFR | AlchRel | AlchRel | AlchRel |
| 20 | BallingLvl | FillPressure | HydPressure2 | CarbPressure1 | CarbPressure1 | CarbPressure1 |

*Figure 10- Variable Importance*

Variable importance tells us the predictors which most influenced our PH. You can see that universally MNF flow was selected as our most important predictor. Bowl. Setpoint, Transcode, Filler Level, and Pressure Setpoint, all seem to be selected often in terms of importance in our models. Earlier on in the report, from simply looking at the correlations of PH to our predictors we speculated that (MNF flow, Fill. Pressure, Filler Level, Usage.Cont, Bowl. Setpoint and Pressure. Setpoint) would be important. This is useful in many ways.

Highly correlated variables being used in our models often is confirmation that our model is working correctly, as we assumed these variables should be important to our model. In the future we can develop simpler linear models using the predictors we know are likely to have predictive value. We can also better understand how to manipulate our future manufacturing process assuming new regulations require so. MNF flow is our minimum night flow. We know that with a correlation of -.46, as we increase MNF, our PH tends to decrease. Manipulation of this one measurement may in fact be enough to produce results to meet future regulatory burden

## Conclusion

XGBoost has produced an extremely accurate model with a cross-validated RMSE of .116 . We suggest that this report be used to initialize further experiments. We would like to test the effects of our importance predictors in a live setting. We also suggest that data collection errors that we noted be addressed in our future data collection. We are confident that by calibrating predictors in our manufacturing process to certain levels, Beverage ABC will be able to manipulate our PH levels to be prepared for whatever future regulatory burden is placed on us in the future.