pH Manufacturing Proposal:

By: Ron Balaban, Brandon Chung, Yanyi Li, Chi Hang(Philip) Cheung, Jiaxin Zheng

Abstract:

In this predictive model report, the best-fit model was trained by a historical dataset consisting of 2571 samples and 32 variables. Since the training dataset has varying degrees of missing data and multiple predictors, several data preprocessing methods, such as imputation and high-correlative predictor exclusion, were employed for data cleaning. While numerous predictive models were tested, such as linear, non-linear, and different tree models, the best predictive model identified in this report was the RandomForest (RF) model generated using R. The best-tuned hyperparameters were *mtry = 29* and *ntree = 1000*. The R-squared and RMSE were calculated to be 0.66 and 0.103, respectively.

Data Exploration:

In our data, there are 31 predictors and one target variable, PH. Through our exploration we found there to be missing data in the majority of variables; our team has addressed these missing values in the preprocessing portion of this report. In addition, there are outliers in the distribution plots seen in ***Figure 1***. However, none seem to be irregular or due to recording mistakes, so we will keep all values. Notably, in the target variable PH, there is a left-sided tail to the distribution, and so, how to maintain a sufficiently high pH level as a business question is of interest.

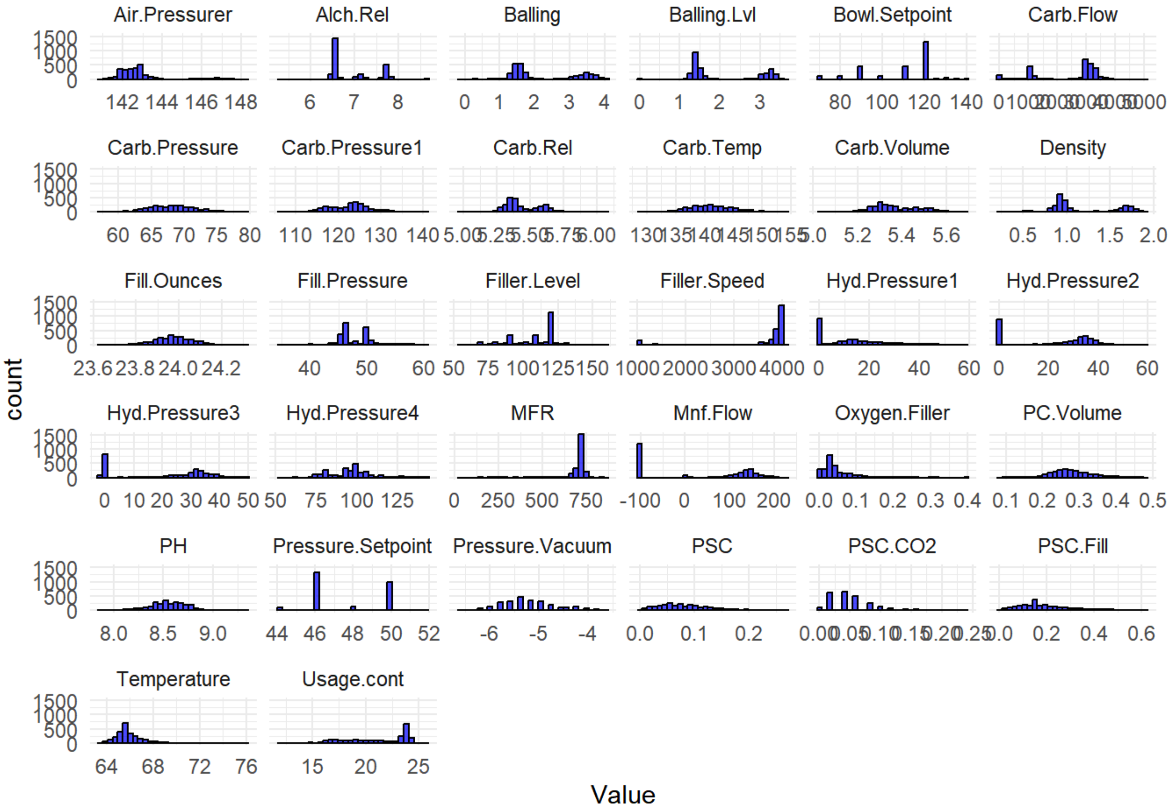

Figure 1: Distributions of each variable in the data.

Preprocessing methods:

Prior to imputation, one near-zero variance variable, “Hyd.Pressure1”, was identified and removed from the training set since it had an insignificant impact on the target variable. Predictors with high correlation with one another were also examined, and none were identified. Missing data was calculated and plotted as shown in ***Figure 2.*** The highest missing data was identified in column MFR. Four missing values were also found in the target column. Our team decided to remove these four rows of samples, where the missing pH values were, for two reasons. First, the missing values account for less than 5% of the pH column. Second, since pH is the target variable, any attempt to impute the missing data could potentially affect the accuracy of the model.

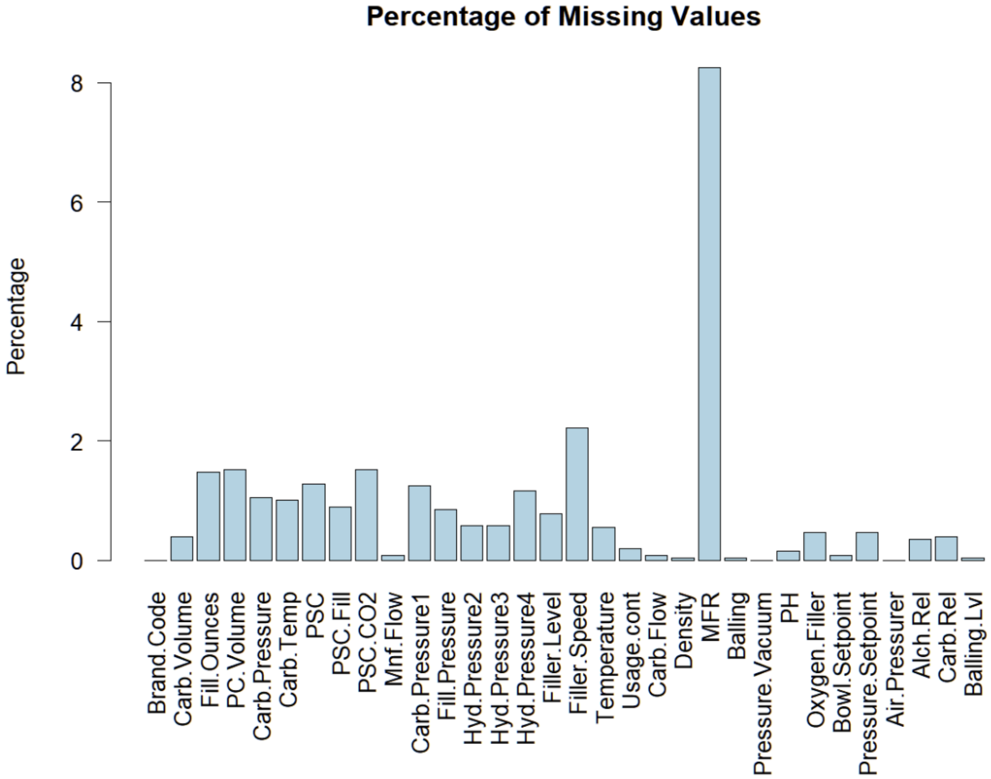

Figure 2: Percentage of missing data visualized.

To select the best imputation method between KNN, RandomForest, and MICE, the imputed results were plotted in a density diagram (***Figure 3-5***). As shown by the density plots and the t-test for mean values between the imputed data and the original dataset, only the KNN imputed values showed significant deviation. This is also seen by the red and the black lines that do not overlap in the density plot, and the t-test p-value is much less than 0.05 from the original dataset. In other words, KNN imputation introduces excessive bias and, therefore, was not an optimal choice for imputing this dataset. In contrast, the density plots for MICE and RandomForest showed almost complete overlapping of red and black lines, and the t-test p-values were greater than 0.05, indicating the acceptance of the Null-Hypothesis of difference. Since both of these methods are acceptable for imputing the missing data, our team selected MICE as our method of imputation.

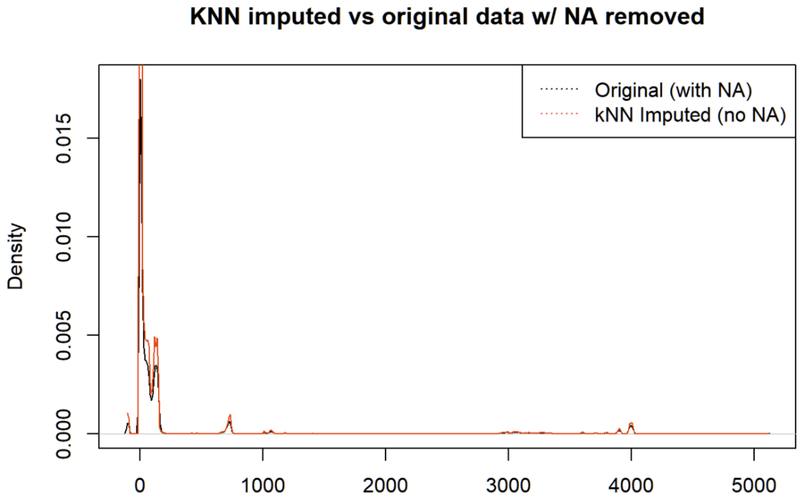

Figure 3: Density plot for the original data points(black) vs KNN Imputed data points(red)

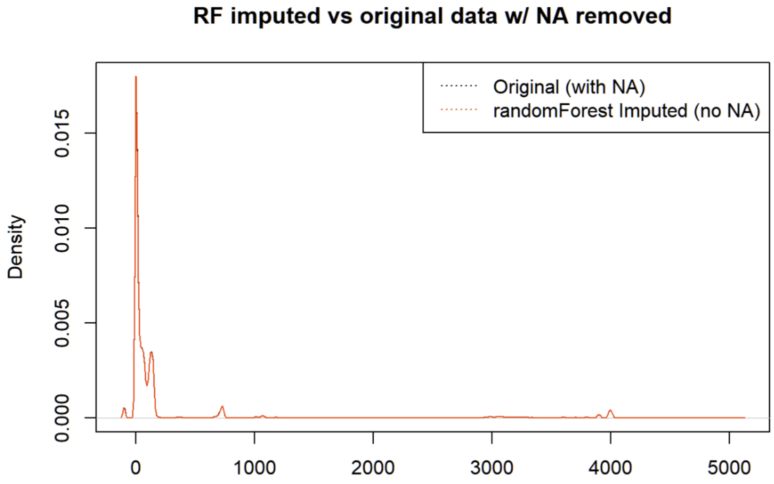

Figure 4: Density plot for the original data points(black) vs RandomForest Imputed data points(red)

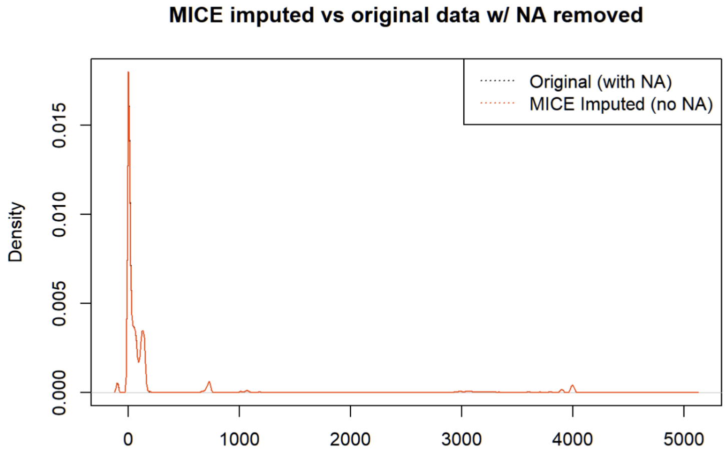

Figure 5: Density plot for the original data points(black) vs. MICE-imputed data points(red)

Predictive model results:

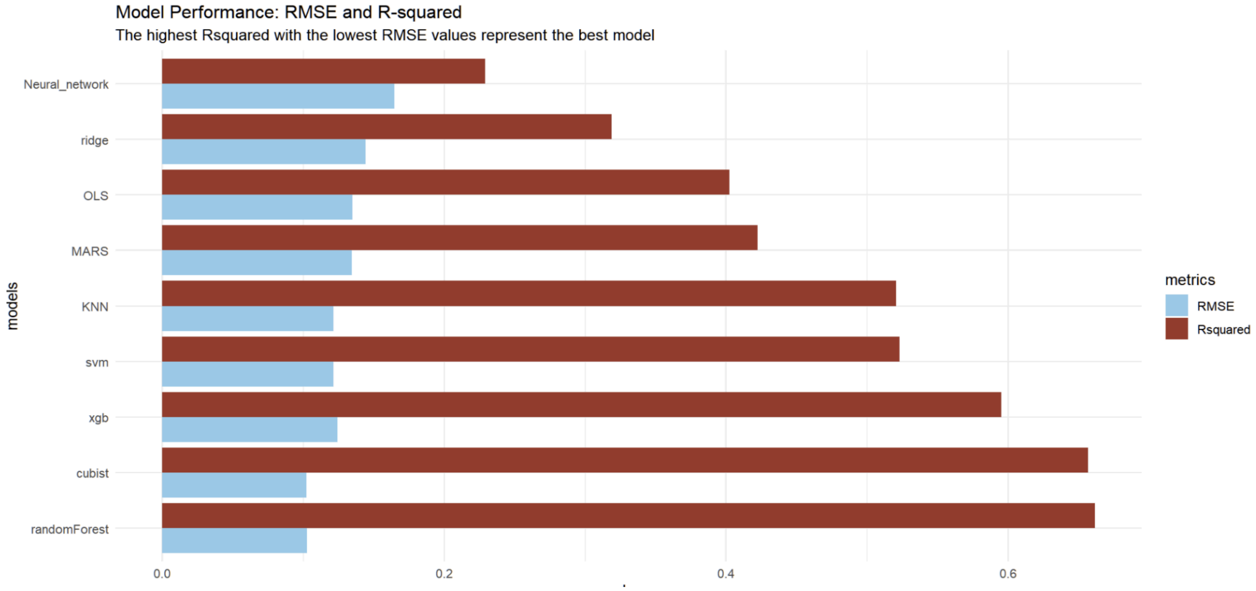

Figure 6: RMSE and R-squared metrics of various models

Several models were chosen to see the RMSE and R-squared, with the strongest one being a Random Forest model. The best RF model was tuned by testing the *mtry* hyperparameters from every other number in the range of 1 to 31. The upper range of 31 was determined by subtracting the total number of predictors by 1 to introduce randomness in predictor selection. The best R-squared and RMSE were determined at 1000 decision trees, 29 randomly assigned predictors, and a 10-fold cross-validation in both the training and the validation test metrics. According to the RF model, the most important predictor was the “Mnf. flow”, which accounted for almost 100% of the impact on the pH value (***Figure 7***). The second highest predictor was the categorical predictor “Brand.Code” and in particular, the brand of letter “C” played an important role in affecting the manufacturing process of pH.

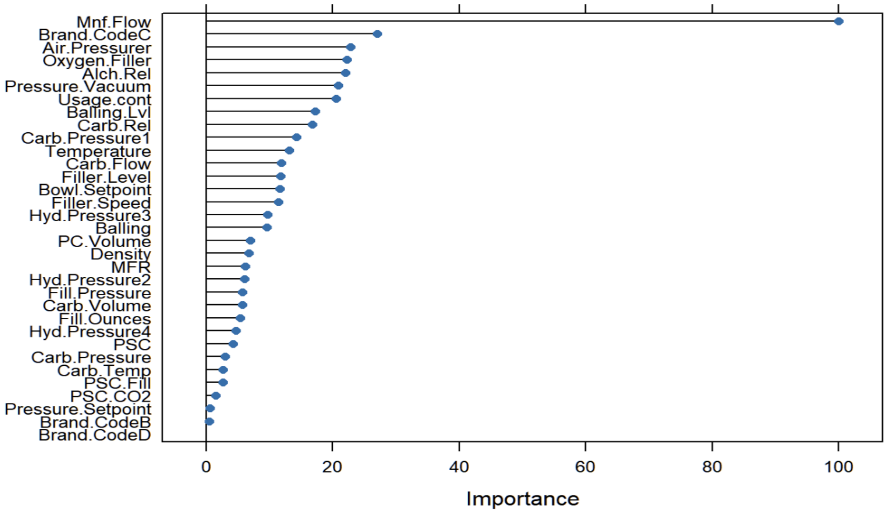

Figure 7: The variable importance plot generated from RandomForest.

A correlation plot from the top 10 predictors calculated by the RF model was also created (***Figure 8***). It was based on the original training data without performing a train-split method. The missing data was also imputed by the MICE method. The 'Mnf. Flow' exhibited a strong negative correlation to the pH outcome, while 'Usage.cont' and 'Temperature' had a minor negative correlation. The strongest positive correlations to pH were 'Carb.Rel' and 'Pressure.Vacuum'.

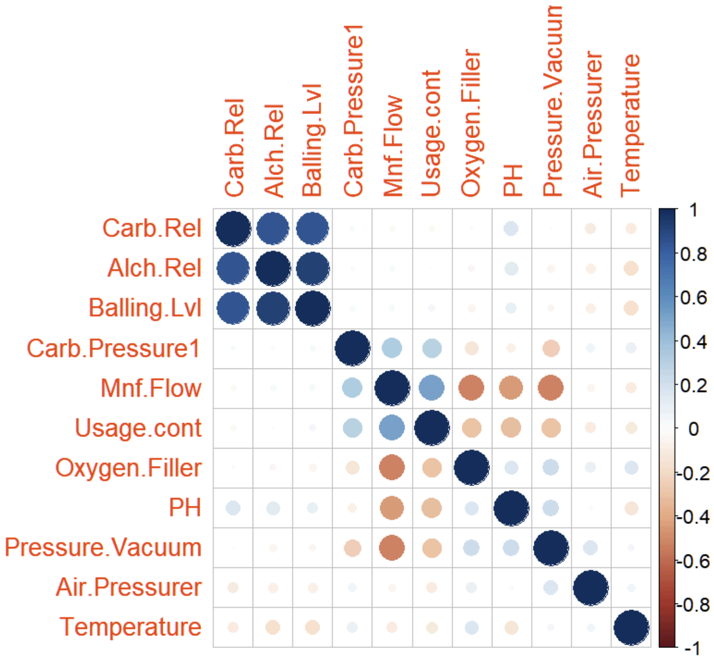

Figure 8: Correlation plot of the top ten predictors for pH calculated by the RF model.

Discussion/Conclusion:

Although the R-squared calculated from the model is 0.66, the predictive performance of the model reported here is still relatively strong. This is observed in the mean predicted pH value, 8.55, and the mean pH value of the training dataset, 8.54. Both are very similar values with a similar mean predictor value for 'Mnf.Flow' (~21, ~24, new test set and training set, respectively). As seen in ***Figure 7-8***, the input of the 'Mnf.Flow' is the most impactful factor on the pH value during the manufacturing process. There is a strong inverse relationship between 'Mnf.Flow' and pH. Therefore, a similar 'Mnf.Flow' input value would warrant a similar pH output value as we observed in the predicted pH value versus the training set pH value.

On the other hand, 'Pressure.Vacuum', 'Carb.Rel', and 'Alch.Rel' exhibited a positive correlation to pH, indicating the tendency for these values to rise or fall in the same direction. However, these positively correlated predictors are not nearly as strong as the effect of the 'Mnf.Flow' as shown in the importance plot in ***Figure 7***. If the manufacturing goal is to keep a high pH value for the final products, the 'Mnf.Flow' input is best kept at a low value during the production process, or vice versa for low pH. Therefore, a strict monitoring of 'Mnf.Flow' would seem to be crucial to the effect of pH. To increase pH, we would recommend increasing 'Pressure.Vacuum', 'Carb.Rel', and 'Alch.Rel', while minimizing 'Mnf.Flow', and vice versa.