**PH Predictive Model Guide**

By: Melissa Bowman, Frederick Jones, Shoshana Farber, Gabriel Campos

Table of Contents

[Introduction: 2](#_Toc166165904)

[Exploratory Data Analysis: 2](#_Toc166165905)

[Missing Variables: 2](#_Toc166165906)

[Distributions: 3](#_Toc166165907)

[Correlation between Variables: 5](#_Toc166165908)

[Correlation of Variables with PH: 6](#_Toc166165909)

[Data Pre-Processing: 8](#_Toc166165910)

[Data Partition: 8](#_Toc166165911)

[Modeling: 9](#_Toc166165912)

[Comparing Models for Best Fit: 9](#_Toc166165913)

[Model Type Tested and Brief Explanation: 9](#_Toc166165914)

[Linear Model: 9](#_Toc166165915)

[PLS: 10](#_Toc166165916)

[KNN Model: 10](#_Toc166165917)

[MARS Model: 10](#_Toc166165918)

[SVM Model: 11](#_Toc166165919)

[Random Forest Model: 11](#_Toc166165920)

[Results: 12](#_Toc166165921)

[Linear Model Results: 12](#_Toc166165922)

[Diagnosis of Linear Stepwise Model: 12](#_Toc166165923)

[PLS Results: 14](#_Toc166165924)

[Diagnostic plots of the PLS: 15](#_Toc166165925)

[KNN Model Results: 15](#_Toc166165926)

[Diagnostic plots of the KNN: 15](#_Toc166165927)

[MARS Model Results: 16](#_Toc166165928)

[Diagnostic plots of the MARS: 17](#_Toc166165929)

[SVM Model Results: 17](#_Toc166165930)

[Diagnostic plots of the SVM: 18](#_Toc166165931)

[Random Forest Model Results: 18](#_Toc166165932)

[Results Comparison of Each Model: 20](#_Toc166165933)

[Conclusion: 20](#_Toc166165934)

# Introduction:

Considering new regulations, our leadership has highlighted the need to grasp the intricacies of our manufacturing processes, especially regarding the predictive factors affecting pH levels. pH holds significant sway over numerous manufacturing operations, impacting everything from product quality to regulatory compliance. Fluctuations in pH can trigger adverse consequences like defects in products, equipment corrosion, and environmental hazards. We must explore the drivers behind pH fluctuations within our manufacturing setting. In response, we're undertaking the development of a predictive model for pH, enabling us to stay ahead of the curve in monitoring and managing this critical aspect of our operations. Our approach involves a thorough analysis of historical data to pinpoint the key factors influencing pH and to establish a robust predictive model.

# Exploratory Data Analysis:

Upon initial examination of the dataset, we found it to consist of 2571 observations across 33 column variables. All variables are numeric with one column named Brand Code, comprising of categorical values. Our response variable is PH.

## Missing Variables:

In model-making, identifying missing variables is crucial because missing data can significantly impact the accuracy and reliability of the model's predictions. Failure to account for missing variables can lead to biased estimates, incorrect inferences, and reduced predictive power. By recognizing and addressing missing variables, modelers can ensure that their models are more comprehensive, robust, and better equipped to make accurate predictions based on the available data. Additionally, understanding the nature and extent of missing data allows for appropriate strategies to handle them, such as imputation techniques or model adjustments, ultimately improving the overall quality and effectiveness of the model.

The analysis revealed that the variable MFR values were absent in about 8% of cases, with Brand Code showing around 5% missing entries, signaling potential areas for data refinement and imputation. Meanwhile, variables like Filler Speed had approximately 2% missing values, while many other variables had minimal missing data, usually less than 1%. Intriguingly, Pressure.Vacuum and Air.Pressurer exhibited no missing values, indicating their suitability for further analysis.

A graph of a number of rows

Description automatically generated with medium confidence

Figure 1: Percentage of variables missing from the historical data.

## Distributions:

Understanding the distribution of variables helps in selecting appropriate statistical techniques, determining the suitability of modeling assumptions, and identifying potential outliers or anomalies. By analyzing distributions, modelers can make informed decisions about data transformations, feature engineering, and model selection, ultimately leading to more accurate and robust predictive models. Moreover, examining distributions enables the identification of skewness, kurtosis, and other distributional properties, which are essential for interpreting model results and making reliable predictions.

Upon examining the distribution patterns within the dataset, several distinct trends emerged. Variables such as Carb Pressure, Carb Temp, Fill Ounces, PC Volume, and PH appeared to follow relatively normal distributions, suggesting a balanced distribution of data points across their respective ranges. Conversely, Hyd Pressure 1, PCS, PSC CO2, PSC Fill, Air Pressurer, Oxygen Filler, Pressure Vacuum, and Temperature exhibited a right skew, indicating a concentration of values towards the higher end of the scale. On the other hand, Hyd Pressure 2, Hyd Pressure 3, and Mnf Flow displayed left-skewed distributions, with a notable number of entries clustered at lower values, although instances of zero values were also observed. Similarly, Filler Speed and MFR showcased left skewness, suggesting a prevalence of lower values. Additionally, variables like Balling, Balling Lvl, Carb Rel, and Density demonstrated bimodal distribution patterns, implying the presence of two distinct peaks in their data distribution. These observations provide valuable insights into the distributional characteristics of our variables, informing subsequent analytical approaches.

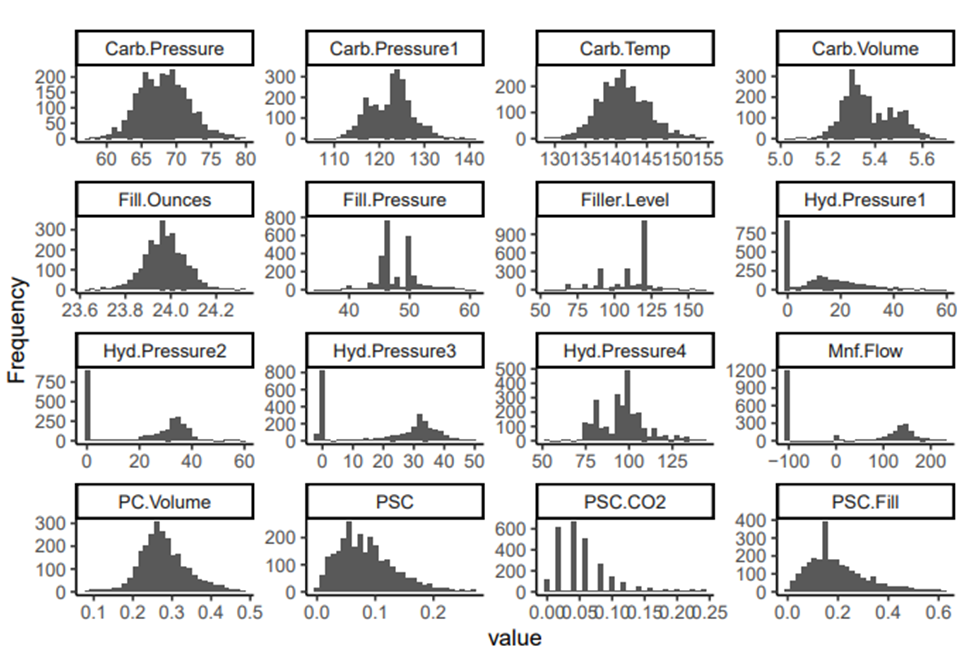


Figure 2: Distribution of variables from the historical data.

A screenshot of a graph

Description automatically generated

Figure 3: Continuing distribution of variables from the historical data.

## Correlation between Variables:

It’s crucial to examine the correlation between variables because it helps identify relationships and dependencies among the predictors. Understanding these correlations allows us to gauge the strength and direction of the relationships, which is essential for selecting appropriate features for the model. Highly correlated variables can lead to multicollinearity issues, where predictors convey redundant information, potentially distorting model estimates and reducing interpretability.

We've noticed some interesting relationships among our predictor variables. For instance, Carb Rel, Alch Rel, Density, Balling, and Balling Level all seem to be positively linked with each other. Similarly, Hyd Pressure2 appears to have a positive connection with Hyd Pressure 3. Furthermore, there are positive associations between Carb Temp and Carb Pressure, MFR and Fill Speed, Bowl Setpoint and Fill Level, as well as Pressure Setpoint and Fill Pressure. On the flip side, we've observed some variables that show strong negative correlations, like Pressure Vacuum with Hyd Pressure2 and Hyd Pressure3, Mnf Flow with Filler Level and Bowl Setpoint, and Hyd Pressure4 with Alch Rel. Additionally, some other variables display moderate correlations, as seen from the medium blue and medium red squares in our correlation plot.

A colorful triangle with numbers and letters

Description automatically generated with medium confidence

Figure 4: Correlation between Variables

## Correlation of Variables with PH:

Assessing the correlation of variables with the target variable is necessary as it helps identify which predictors are most strongly associated with the outcome we are trying to predict. By understanding these correlations, we can prioritize features that have a significant impact on the target variable, enhancing the predictive power of the model. Variables with high correlations with the target are likely to be more informative and influential in making accurate predictions. Additionally, examining these correlations enables us to identify potential predictors that may not be relevant or contribute little to the model's performance, helping streamline the feature selection process.

When looking at the variables individually, none show an extremely strong correlation with PH. Mnf Flow has the highest correlation at about -0.46. The variables most positively correlated with PH are Bowl Setpoint and Filler Level. On the other hand, the most negatively correlated variables with PH, aside from Mnf Flow, are Usage cont, Fill Pressure, and Pressure Setpoint.

A graph of a graph

Description automatically generated with medium confidence

Figure 5: Correlation of Variables with pH.

# Data Pre-Processing:

Pre-processing data ensures that the data is in a suitable format and quality for building accurate and reliable models. This involves steps such as handling missing values, scaling or normalizing features, encoding categorical variables, and detecting and addressing outliers. Pre-processing helps to improve the performance of the model by removing noise, reducing bias, and improving the interpretability of the results. Additionally, it helps to mitigate common issues such as overfitting and multicollinearity, leading to more robust and generalizable models.  
  
Steps in Preprocessing:  
 1. The column names of the dataframe were converted to the snake case format. Snake case is a naming convention where words are separated by underscores, and typically all lowercase.

2. Since pH is our target variable, we need to remove any rows that do not have a value for this column. We will also transform our brand\_code variable to categorized factors, replacing any NA value with Unknown.

3. We identify unhelpful columns in the dataset, such as any variables with zero variance or near-zero variance. The only variable with near-zero variance is hyd\_pressure\_1. Therefore, we will not include this variable in our modeling.

4. We preprocess the data using the preProcess function from the caret package. We utilize the preProcess function from the caret package to transform data using methods such as knnImpute for nearest neighbor imputation of missing data, nzv for removing near-zero values identified above, corr for filtering out highly correlated values to address multicollinearity, center for subtracting the mean of the predictor's data from the predictor values, scale for dividing by the standard deviation, and BoxCox for normalizing data.

5. We use the predict function to process the list variables created with preProcess() to recreate the dataframe.

6. Lastly, we will verify that there are no NA values remaining in the preprocessed data.

# Data Partition:

Splitting the data into a training and validation set allows us to assess the performance of our model on unseen data. The training set is used to train the model, enabling it to learn the underlying patterns and relationships within the data. The validation (test) set serves as an independent dataset to evaluate the model's performance and generalization ability. By validating the model on unseen data, we can detect overfitting, where the model performs well on the training data but poorly on new data. Splitting the data helps to ensure that the model's performance estimates are more reliable and representative of its true predictive ability, thereby improving its reliability and usefulness in real-world applications.

After splitting the data into an 80:20 training and validation set, we will proceed to construct multiple models using the training data. Subsequently, we will evaluate each model to ascertain which one best fit our dataset.

# Modeling:

By building and testing models, we can identify patterns, trends, and relationships that might not be immediately obvious. This helps us to understand the underlying factors driving certain outcomes and enables us to make more accurate predictions about future events or trends. Additionally, testing our models allows us to assess their performance and reliability, ensuring that they provide accurate and actionable results that can be trusted for decision-making purposes.

Comparing Models for Best Fit:  
  
RMSE, Rsquared, and MAE value metrics help determine which model works best by providing quantitative measures of prediction accuracy and model fit. Models with lower RMSE and MAE values and higher R-squared values are generally preferred, as they indicate better predictive performance and a better fit of the model to the data.

RMSE measures the average deviation of predicted values from actual values. It calculates the square root of the average of squared differences between predicted and actual values. Lower RMSE indicates better model performance, as it signifies smaller prediction errors.

R-squared represents the proportion of variance in the dependent variable (target) that is explained by the independent variables (predictors) in the model. It ranges from 0 to 1, with higher values indicating a better fit of the model to the data. R-squared values closer to 1 imply that the model explains a larger proportion of the variability in the target variable.

MAE measures the average absolute difference between predicted and actual values. It provides a straightforward interpretation of prediction errors, regardless of their direction. Like RMSE, lower MAE indicates better model performance, as it signifies smaller prediction errors.

Model Type Tested and Brief Explanation:

Linear Model:  
Linear models assume relationships between the predictors and the target, which means that changes in the predictors result in proportional changes in the target. The core assumption of linear models is that this relationship is linear, meaning it can be represented by a straight line or plane in higher dimensions. Linear models are versatile and widely used in various fields for prediction, inference, and understanding of relationships in data. The purpose of linear models is to provide insights into how changes in predictor variables affect the target variable. By fitting a linear equation to the data, linear models allow us to make predictions about the target variable based on the values of the predictors. Additionally, linear models enable us to assess the significance of each predictor variable, identify patterns, and make inferences about the underlying relationships in the data.

### PLS:

The primary aim of PLS is to distill a concise set of latent variables, termed components, capable of encapsulating the utmost variation across both predictor and target variables. Its efficacy shines particularly in scenarios where conventional regression approaches encounter hurdles, such as instances with a surplus of predictors relative to observations or substantial collinearity among predictors. PLS operates through an iterative process, crafting linear combinations of predictors that exhibit high correlation with the target variable while maintaining orthogonality among themselves. These resultant linear amalgamations manifest as the latent variables or components. Key advantages of PLS encompass its prowess in handling multicollinearity, navigating high-dimensional datasets, and furnishing resilient predictions even amidst diminutive sample sizes. Moreover, it demonstrates a heightened resilience to outliers when juxtaposed with certain alternative modeling methodologies.

### KNN Model:

In KNN, the algorithm functions by identifying the K nearest data points (neighbors) to the new data point within the feature space, then determining its class or value based on the majority class (for classification) or the average value (for regression) of these K nearest neighbors. The key components of the KNN model encompass the K parameter, representing the number of nearest neighbors considered for predictions, chosen based on the dataset and problem at hand to influence the model's performance and generalization ability; the distance metric, measuring similarity or dissimilarity between data points, typically employing metrics like Euclidean distance, Manhattan distance, or cosine similarity; and the decision rule, guiding the aggregation of the class or value of the K nearest neighbors for predictions, often selecting the majority class for classification tasks or the average value for regression tasks.

MARS Model:  
The purpose of Multivariate Adaptive Regression Splines (MARS) is to capture nonlinear and non-monotonic relationships in the data by partitioning the feature space into smaller regions and fitting simple linear models within each region. MARS achieves this by iteratively adding basis functions, called "spline functions," to the model, which are piecewise linear segments allowing adaptation to data nonlinearities. Starting with a simple linear model, MARS adds spline functions stepwise, selecting optimal locations and shapes based on their contribution to model performance. Key components of the MARS model include basis functions, typically hinge functions representing predictor-target relationships, knots defining points where basis functions join, and pruning, which prevents overfitting by removing unnecessary functions. MARS offers advantages such as interpreting complex relationships, handling numerical and categorical predictors, robustness to outliers and collinearity, and providing feature selection by identifying important predictors.

### SVM Model:

Support Vector Machine (SVM) serves to find the optimal hyperplane that effectively separates data points of different classes in the feature space or approximates the decision boundary for regression tasks. In classification, SVM endeavors to identify the hyperplane maximizing the margin, representing the distance between the hyperplane and the nearest data points of each class, termed support vectors. By maximizing this margin, SVM enhances generalization performance, facilitating accurate classification of new, unseen data. Key components of the SVM model include the hyperplane, acting as the decision boundary for binary classification, the support vectors crucial in determining the hyperplane's position and orientation, and the kernel trick, employed to map data into a higher-dimensional space for linear separability when necessary. SVM offers advantages such as handling high-dimensional data, flexibility in selecting kernel functions to address nonlinear relationships, and resilience against overfitting, particularly in small sample sizes or complex, nonlinear decision boundaries.

Random Forest Model:

Random Forest aims to enhance predictive accuracy and mitigate overfitting by amalgamating predictions from numerous individual decision trees. Within a Random Forest model, an extensive array of decision trees undergoes training on varied subsets of the training data and features. Each tree operates autonomously, generating its own predictions. The final prediction of the Random Forest materializes through aggregating the forecasts of all constituent trees. Key elements of the Random Forest model encompass decision trees, where each tree is trained on bootstrapped samples of the training data and considers only random subsets of features at each node for splitting, thus fostering diversity and reducing correlation between trees. Aggregation of predictions involves techniques like majority voting for classification tasks and averaging for regression tasks. Random Forest presents advantages including adeptness with high-dimensional data, capacity to capture nonlinear relationships, and resilience against noise and feature correlation, alongside reduced susceptibility to overfitting compared to individual decision trees and typically requiring less hyperparameter tuning.

# Results:

## Linear Model Results:

The first model test represents a basic linear regression model where the target variable pH is predicted based on all available predictor variables in the training dataset. This model aims to find the best-fitting linear relationship between the target variable and all predictors simultaneously, without considering the potential redundancy or irrelevance of individual predictors. The R-squared for this model is 0.394 and there are several insignificant variables in the model.

The second model tested involves a stepwise regression approach to model selection. The step function iteratively evaluates the addition or removal of predictor variables to the initial linear model, based on a specified criterion. In this case, a direction set to both means that the function considers both forward and backward steps, adding or removing variables one at a time. The trace parameter set to zero suppresses output, indicating that no information about the stepwise process is displayed.

The output of the Linear Stepwise Model from the test set:

|  |  |  |
| --- | --- | --- |
| RMSE | Rsquared | MAE |
| 0.1305888 | 0.3934348 | 0.1043181 |

Diagnosis of Linear Stepwise Model:

A graph of a number of dots

Description automatically generated with medium confidence

Figure 6: Residuals vs. Fitted Values Plot: This plot examines whether the residuals (the differences between observed and predicted values) exhibit a pattern or structure. A random scatter of points around the horizontal line at zero indicates that the assumptions of constant variance and linearity are met.

A graph of a graph

Description automatically generated

Figure 7:Normal Q-Q (Quantile-Quantile) Plot: This plot assesses whether the residuals follow a normal distribution. If the residuals closely follow the diagonal line, it suggests that the assumption of normally distributed errors is reasonable.

## PLS Results:

The best results were achieved when using 9 components. These components are essentially combinations of the original variables that the model uses to understand the relationships in the data.

The output of the PLS Model from the test set:

|  |  |  |
| --- | --- | --- |
| RMSE | Rsquared | MAE |
| 0.1297584 | 0.4008291 | 0.1034359 |

### Diagnostic plots of the PLS:

A graph with blue dots and white text

Description automatically generated

Figure 8: mnf\_flow is the most important variable. This essentially highlighting its significance in of that variable in the PLS model.

## KNN Model Results:

The best results were achieved when using 13 components. These components are essentially combinations of the original variables that the model uses to understand the relationships in the data.

The output of the KNN Model from the test set:

|  |  |  |
| --- | --- | --- |
| RMSE | Rsquared | MAE |
| 0.12001314 | 0.49682031 | 0.09114698 |

### Diagnostic plots of the KNN:

A graph with blue dots and black text

Description automatically generated

Figure 9: , oxygen\_filler is the most important variable and mnf\_filler is the second most important variable.

## MARS Model Results:

The MARS model has an optimal pruning level is 30. This refers to the number of times the model prunes or trims unnecessary elements during its construction, with a higher value indicating more aggressive pruning. The degree is equal to 2. This pertains to the complexity of the spline functions used in the model. A higher degree indicates more flexibility in capturing complex patterns in the data.

The output of the MARS Model from the test set:

|  |  |  |
| --- | --- | --- |
| RMSE | Rsquared | MAE |
| 0.1184498 | 0.5038378 | 0.0910705 |

### Diagnostic plots of the MARS:

A graph with blue dots and white text

Description automatically generated

Figure 10: This model has fewer important variables than the PLS and the KNN models. Like with the PLS model, mnf\_flow is the most important variable.

## SVM Model Results:

The optimal model has a sigma of about 0.024 and C = 4. Sigma determines the width of the Gaussian kernel, which controls the flexibility of the decision boundary. A smaller sigma results in a smoother decision boundary, while a larger sigma leads to a more complex and flexible boundary. C is the regularization parameter that balances the trade-off between maximizing the margin and minimizing classification errors. A smaller C value emphasizes a wider margin, potentially sacrificing some accuracy on training data, while a larger C value focuses on correctly classifying each data point, potentially leading to a narrower margin.

The output of the SVM Model from the test set:

|  |  |  |
| --- | --- | --- |
| RMSE | Rsquared | MAE |
| 0.1098260 | 0.5760762 | 0.0823950 |

### Diagnostic plots of the SVM:

A graph with blue dots and lines

Description automatically generated

Figure 11: oxygen\_filler is the most important variable and mnf\_filler is the second most important variable.

Random Forest Model Results:  
The output of the SVM Model from the test set:

|  |  |  |
| --- | --- | --- |
| RMSE | Rsquared | MAE |
| 0.09531386 | 0.68860455 | 0.07130942 |

A screenshot of a computer

Description automatically generated

Figure 12: For the Random Forest Model, brand\_code is the most important variable for this model and mnf\_flow is the second most important.

## Results Comparison of Each Model:

A table with numbers and letters

Description automatically generated

|  |  |  |  |
| --- | --- | --- | --- |
| Models | RMSE | Rsquared | MAE |
| Linear | 0.1305888 | 0.3934348 | 0.1043181 |
| PLS | 0.1297584 | 0.4008291 | 0.1034359 |
| KNN | 0.1200131 | 0.4968203 | 0.0911470 |
| MARS | 0.1184498 | 0.5038378 | 0.0910705 |
| SVM | 0.1098260 | 0.5760762 | 0.0823950 |
| Random Forest | 0.0953139 | 0.6886045 | 0.0713094 |

We can clearly see that the random forest model has the highest prediction accuracy when it comes to the evaluation set, with an 𝑅2 of about 69%.

Conclusion:  
Based on the evaluation metrics, it is evident that each model performs differently in predicting the target variable. While Linear Regression and Partial Least Squares models provide modest predictive power, K-nearest Neighbors and Multivariate Adaptive Regression Splines models offer improved accuracy, with KNN slightly outperforming MARS. However, Support Vector Machine demonstrates superior performance compared to all other models, followed by Random Forest, which exhibits the lowest prediction errors and explains the highest proportion of the variance in the target variable.

Given the new regulations prioritizing understanding the manufacturing process's key variables influencing the pH of beverages rather than emphasizing model transparency, the decision was made to select the Random Forest model for its superior predictive performance. Consequently, Random Forest emerges as the preferred choice for predicting pH levels.