# Data 624 Project #2

Group 3: Shaya Engelman, Julia Ferris, Amanda Fox, Jean Jimenez

07-13-2024

## Introduction

pH is the potential for hydrogen. It is a measurement that tells us how acidic or basic a liquid is. As a beverage manufacturing company, pH is an important metric (and considered a KPI). We have a responsibility as a company to ensure that all our products are safe to consume. As a result, the pH of our product must fall within the acceptable range. pH is influenced by how many hydrogen ions (charged particles) are present in the liquid. The measurement is on a logarithmic scale that goes from 0-14. Low pH (many hydrogen ions) and high pH (many negatively charged particles present) can both be equally dangerous. Both high and low pH can lead to the deterioration and corrosion of materials. In our stomachs, stomach acid (which has a pH of 1.5-3.5) is responsible for breaking down the protein in our food. Similarly, highly concentrated acids and bases are used to dissolve metals and materials. Therefore, it is our responsibility as a company to constantly monitor this metric to ensure that it is in the acceptably safe range.

While pH is directly affected by how many ions are in a solution, there are many other factors that can influence the concentrations of these ions. Temperature, pressure, concentrations of different chemicals/ingredients, carbonation, and volume all affect the pH of a liquid. For example, at high temperatures hydrogen ions might be more likely released due to the breaking of the hydrogen bonds; lowering the pH of the liquid. If the liquid is carbonated, changing the temperature changes the concentration of carbonic acid in the beverage, changing the pH. At high pressures, gases can be dissolved in liquids, changing the pH.

Due to the complexity of pH and the numerous factors that can influence it, each beverage possesses a unique pH profile (e.g., soda has a low pH, whereas alkaline water has a high pH). By leveraging the data we have on our beverages, we can analyze which factors or combinations of factors most significantly impact pH. This analysis will enable us to create a predictive model for determining the pH of future products. Such a model is important for developing new beverages, allowing us to adjust different variables to achieve the desired pH and ensure the final product meets safety and quality standards. By understanding and controlling these influencing factors, we can consistently produce beverages that are not only safe but also meet the taste and quality expectations of our consumers.

## Data Exploration and Preparation

To figure out what factors influence pH the most, we will analyze our company’s historic beverage data. This data has many observations of different metrics, as well as the pH of the product. Specifically, we were provided with a dataset containing 2,571 observations of 33 variables (one outcome, pH, and 32 predictors). There is one categorical variable, Brand Code, with five values, and the remaining 31 variables are all numeric. We will use this data to fit a model to A white background with black and white text

Description automatically generatedpredict pH. The following **figure 1** shows the summary statistics for the 33 variables.

Figure 1- Summary Statistics of 33 variables in Beverage dataset

Most columns in **Figure 1** have some missing values, which can be a problem when we try to create accurate models. While missing data might not seem like a big issue, it's important to address them to avoid bias in our predictions. In total, there are 844 missing values, as shown in **Figure 2**. Fortunately, the 844 missing values is only ~1% of our total data.

A graph with numbers and symbols

Description automatically generated with medium confidence

Figure 2- Distribution of Missing values by Variable

If we don't fill in the missing data, our models might give inaccurate or skewed results. By estimating and filling in these gaps, we can make sure our predictions are more reliable and accurate. This process helps us understand the true patterns and relationships in our data, leading to better decisions and outcomes for our products.

How do we fill in missing data? The approach varies depending on the variable. Since each variable is unique, we may need different strategies to handle the missing values. To determine the best method, we first visualized the distributions of these variables. By examining these visualizations, we gained insights into the patterns and characteristics of the data. Based on this analysis, we selected appropriate imputation methods to accurately fill in the missing data.

**Figure 3** shows the distribution of these variables. Some variables, such as *Fill Ounces*, *Carb Pressure*, and *Carb Temp*, are unimodal and nearly normally distributed. Others, like the three *PSC* variables, exhibit strong skewness. Additionally, many variables, such as the *Hyd Pressure* variables, *Balling*, and *Carb Rev* variables, are multi-modal.

A screenshot of a graph

Description automatically generated

Figure 3- Distribution of Variables in dataset

In terms of product branding/ code, there are 4 different products that were included in this analysis. Some observations in our dataset lacks a *brand code*. The following **figure 4** shows the number of observations for each brand.

A graph of a bar graph

Description automatically generated

Figure 4- Distribution by Brand Code

Next, we wanted to see how related some of the predictor variables are to each other. What I mean by this is that some variables might be directly linked. For example, an increase in temperature might cause an increase in pressure. To figure out which predictors are correlated, we conduct a correlation analysis. This analysis graphs each predictor against the others to see if there is a correlation, helping us understand the relationships between variables.

**Table 1** shows the results of the correlation analysis. 25 predictor pairs were highly correlated with one another: they will also need to be handled properly, depending on our model.

**Table 1**: Correlation between Variables. *\*\* Indicate highly correlated variables.*

|  |  |  |
| --- | --- | --- |
| **Var1** | **Var2** | **Value** |
| *Balling Lvl* | *Balling* | 0.9782085\*\* |
| *Balling* | *Density* | 0.9551535\*\* |
| *Bowl Setpoint* | *Filler Level* | 0.9499345\*\* |
| *Balling Lvl* | *Density* | 0.9479189\*\* |
| *MFR* | *Filler Speed* | 0.9479115\*\* |
| *Balling Lvl* | *Alch Rel* | 0.9274995\*\* |
| *Hyd Pressure3* | *Hyd Pressure2* | 0.9248821\*\* |
| *Alch Rel* | *Balling* | 0.9244934\*\* |
| *Alch Rel* | *Density* | 0.9027378\*\* |
| *Balling Lvl* | *Carb Rel* | 0.8843844 |
| *Carb Rel* | *Carb Rel* | 0.8435578 |
| *Carb Rel* | *Balling* | 0.8415449 |
| *Carb Rel* | *Density* | 0.8224442 |
| *Carb Temp* | *Carb Volume* | 0.8171417 |
| *Carb Rel* | *Carb Volume* | 0.793304 |
| *Balling Lvl* | *Carb Volume* | 0.7827037 |
| *Alch Rel* | *Carb Volume* | 0.780748 |
| *Density* | *Carb Volume* | 0.7616051 |
| *Hyd Pressure3* | *Carb Volume* | 0.7590691 |
| *Hyd Pressure2* | *Mnf Flow* | 0.7237407 |
| *Pressure Setpoint* | *Fill Pressure* | 0.6721748 |
| *Hyd Pressure2* | *Mnf Flow* | 0.6561663 |
| *Hyd Pressure2* | *Hyd Pressure1* | 0.6341679 |
| *Hyd Pressure3* | *Hyd Pressure1* | 0.6154817 |
| *Pressure Vacuum* | *Hyd Pressure3* | -0.6028821 |

Next, we need to check the relationship between each of the 31 variables and the pH to show their relationship. Maybe there is one or two predictors that directly influence pH. This would have made things easier. However, when we checked, no single variable showed a clear relationship to pH.

Next, we checked for outliers. Outliers are values that are superhigh or low. Essentially, they are far away from the rest of the data. How far away? Greater than a standard deviation of +/-2. Having outliers is normal (depending on how many you have compared to the sample size). Many of our predictors had a significant number of outliers. Since we do not have context from our stakeholders, we will assume they are meaningful and leave them in the analysis.

Finally, we filled in the missing data. The imputation methods we chose were predictive mean matching (PMM) and polytomous logistic regression (Polyreg). We used PMM for filling in the numeric variables and Polyreg for filling in the two value categorical variable. PMM works by predicting the missing value and then finding the closest actual observed value to use as the fill-in. Polyreg models the probabilities of each category and fills in the missing values based on these probabilities. Now our dataset lacks missing values.

The last thing we do is split our data into a training and testing set. We will place 70% of our data in the training set and the other 30% goes in the test dataset. These values are placed in each dataset at random. Splitting the data this way helps us build and train our model on one portion (training set) and then evaluate its performance on another portion (testing set) to ensure it works well on new, unseen data. This step is important to verify that our model is accurate and appropriate.

## Model Selection and Evaluation

Now that we have split our data, we can use our training set to train a few models. What do I mean by training a model? As a metaphor, lets say we are shoemakers and we’d like to make the perfect sized shoe tailored to each individual. As a metaphor, think of it like teaching a student. The training set is like the set of textbooks and practice problems you give the student to learn from. Just as a student uses these resources to understand concepts and solve problems, our models use the training data to learn patterns, relationships, and how those influence pH.

Fitting a model is like evaluating how well the student has learned the material. During training, we might try different teaching methods (various models) to see which one helps the student understand the best. Similarly, we apply different algorithms or models to our training data to see which one captures the patterns most accurately. The models we used to train our data were OLS Regression, PLS Regression, Elastic Net Regression, Random Forest, cubist, and SVM. **Table 2** Shows the models used, as well as a snippet of how each works.

Once the student has studied, we test their knowledge with an exam (the test set) to see how well they can apply what they've learned to new problems. This helps us evaluate the effectiveness of our training and ensures the model performs well on new, unseen data. Similarly, we use each model to predict the pH of the training set. Since we already know the actual pH of the training set, we can compare our predicted and actual values to evaluate performance.

**Table 2**: Initial Model Performance Summary

|  |  |  |
| --- | --- | --- |
| Model | Explanation | Performance |
| Random Forest | Uses many decision trees and combines their results to make more accurate and stable predictions. | **MAE**: 0.0675942  **RMSE**: 0.0889245  **R-squared**: 0.5855466 |
| Cubist | Creates rules and uses them to make predictions. | **MAE**: 0.0666379  **RMSE**: 0.0909819  **R-squared**: 0.5785172 |
| Elastic Net Regression | Improves prediction accuracy by combining two methods that help in managing data complexity. | **MAE**: 0.0909566  **RMSE**: 0.1132927  **R-squared**: 0.3069637 |
| OLS Regression | Finds the best-fit line by minimizing the differences between actual and predicted values. | **MAE**: 0.0949522  **RMSE**: 0.1239641  **R-squared**: 0.3104240 |
| PLS Regression | Simplifies the data into fewer components and then makes predictions based on these components. | **MAE**: 0.1111390  **RMSE**: 0.1388444  **R-squared**: 0.1417295 |
| SVM | Finds the best way to separate data into different groups. | **MAE**: 0.0801286  **RMSE**: 0.1086437  **R-squared**: 0.4212952 |

The model that performed the best was Random Forest and Cubist. They have the highest R-squared and lowest RMSE values, indicating that they have the best predictive performance. The Cubist model's better interpretability, due to its clear rules, combined with its slightly lower RMSE value, makes it the best model for predicting pH. This is because the Cubist model not only provides accurate predictions but also allows us to understand the underlying relationships and rules that it uses to make those predictions. This interpretability is important for making informed decisions and adjustments in the process.

No model is perfect the first time it is run. Models require constant fine-tuning and adjustments. This fine-tuning and adjustments allow us to get a more accurate model. To improve our Cubist model, we decided to use a grid search to find the optimal parameters for the number of committees and rules. This involved creating a grid of possible values for the number of committees, ranging from 10 to 100, and different settings for the number of neighbors, including values like 0, 3, 5, 7, and 9. We then used cross-validation, a method where the training data is split into multiple subsets, and the model is trained and validated on these subsets to ensure robustness. By training the Cubist model with each combination of parameters and evaluating its performance, we were able to identify the best settings that minimized the RMSE value. This process of grid search and cross-validation helped us refine the Cubist model to achieve even better predictive performance, making it accurate and reliable for predicting pH.

Finally, we evaluate our fine-tuned cubist model. This is like what we did before, where we calculate these error metrics by comparing actual pH with the calculated pH. This will help us understand in the future the performance of our predictive model. We compared this to the other models. Another metric MAPE can help us understand the performance of the model. **Table 3** shows us the performance of each model, including our fine-tuned cubist model.

**Table 3**: Fine-Tuned Model Performance Metrics

|  |  |  |
| --- | --- | --- |
| **Model** | **MAPE** | **RMSE** |
| **Cubist\_Tuned** | 0.0081899 | 0.0969781 |
| **Random\_Forest** | 0.0083083 | 0.0974769 |
| **Cubist\_First** | 0.0081982 | 0.097541 |
| **SVM** | 0.0103123 | 0.1190667 |
| **ElasticNet** | 0.0120861 | 0.1298058 |
| **OLS** | 0.0121501 | 0.1304081 |
| **PLS** | 0.0140257 | 0.1519543 |

Our fine-tuned Cubist model performed better than the other models. This gives us confidence that we can use the Cubist model to predict our missing pH values accurately. The way the Cubist model works, we must check each predictor and see how important it is in influencing pH. This is because understanding which variables most significantly impact pH helps us make more informed adjustments to our processes.

The most important variables in the Cubist model are manufacturing flow (*mnfFlow*), balling level (*ballingLvl*), balling, alcohol release (*alchRel*), and pressure vacuum. These variables have the most significant impact on predicting the pH value. By focusing on these key features, we can make adjustments to change the pH of the product, ensuring it meets the desired standards. Partial dependent plots were used to further understand the specific relationship between these predictors and pH.

## Making Predictions and Conclusion

Now that we have our model all fined tuned, and know that it performs well, we can use it on other data to predict/estimate the pH of various products. We import a dataset with predictor data and pH missing. We used our model to guess the pH and filled in the values accordingly.

In this project, we aimed to develop a reliable model for predicting the pH of our beverage products, which is essential for ensuring safety and quality. We analyzed historical data, addressed missing values, and investigated the relationships between various predictors and pH. After evaluating multiple models, we identified the Cubist model as a top performer due to its high predictive accuracy and interpretability.

We further refined the Cubist model using a grid search and cross-validation to optimize its parameters, which significantly improved its performance. Our fine-tuned Cubist model outperformed the others when evaluated with MAPE and RMSE, confirming its suitability for predicting pH. Key predictors like manufacturing flow, balling level, balling, alcohol release, and pressure vacuum were identified as the most significant factors influencing pH.

With our model validated and fine-tuned, we applied it to new data sets to estimate missing pH values, helping us maintain product quality and proactively control potential issues