Predicting pH Levels: A Data-Driven Approach to Process Monitoring

Report for ABC Beverage || Production Data Science Team || May 2025

DATA 624 || Ali Ahmed, Andreina Arias, Kaylie Evans, Naomi Buell, and Zaneta Paulusova

The Goal:

This report dives into our findings for the requested predictive components of pH based on the new regulations which now require us to understand our manufacturing process. The goal is to understand what factors influence pH and to develop a model that could accurately forecast it based on our available production data. We decided the MARS model offers the best balance of accuracy and consistency.

The Process:

The project began with data cleaning and preparation. We worked with historical manufacturing data that included both product and process variables. This step involved removing duplicate or uninformative columns, handling missing values, and converting all data into a format suitable for modeling. Ensuring data quality is essential to making sure the models would be both accurate and trustworthy.

Next, we explored the data to look for patterns and relationships. Several variables emerged as important predictors of pH, including flow rates (MnF Flow), temperature, alcohol release (Alch Rel), and balling. These findings suggest that pH is influenced by a complex but quantifiable combination of process inputs.

With clean, structured data in hand, we trained and tested the following six predictive models to find the highest performing model:

- Partial Least Squares (PLS)
- Elastic Net
- K-Nearest Neighbors (KNN)
- MARS (Multivariate Adaptive Regression Splines)
- Decision Tree
- Model Tree

Each model was evaluated using industry-standard accuracy metrics like RMSE (Root Mean Square Error) and R-squared. This helped us objectively compare performance across models.[1]

The Results:

After a thorough evaluation, we selected the MARS model as it offers the best balance of accuracy and consistency. 46% of the variation in pH can be explained by the model (R squared of 0.46). This model was then used to generate pH predictions on a new set of production data. The predictions on the provided data are saved in an Excel file (`pH_forecasts.xlsx` linked here) for easy review.

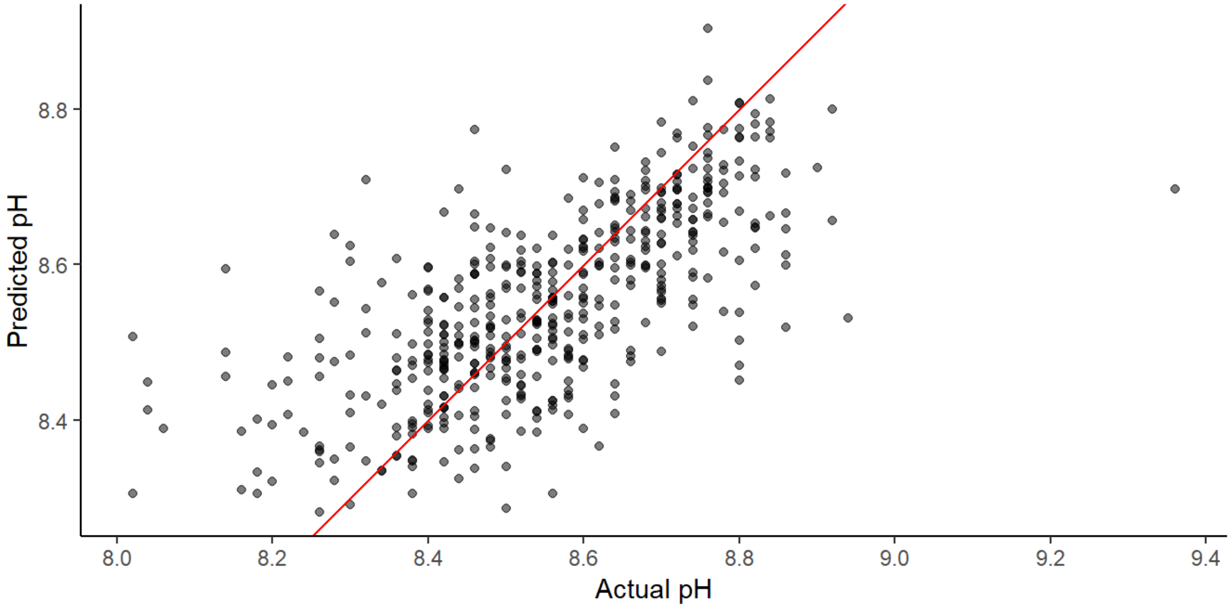

Figure 1. Model performance

As far as the factors that most correlation with the pH, Figure 1 below visualization may help.

[1] A technical report detailing the full modelling process can be found here: https://github.com/naomibuell/624-project-2/blob/main/Technical_Report.qmd.

Figure 2. Correlation Matrix

Lots of process aspects our company tracks have a sway in what pH will turn out to be. The top 4 most important features in predicting pH are *Mnf Flow, Temperature, Alch Rel,* and *Balling,* where the former two have an inverse relationship with pH, and the latter two have a positive relationship with pH. That is, when *Mnf Flow* or *Temperature* increase, pH tends to decrease and when *Alch Rel* or *Balling* increases, pH tends to increase (when all other factors are constant).

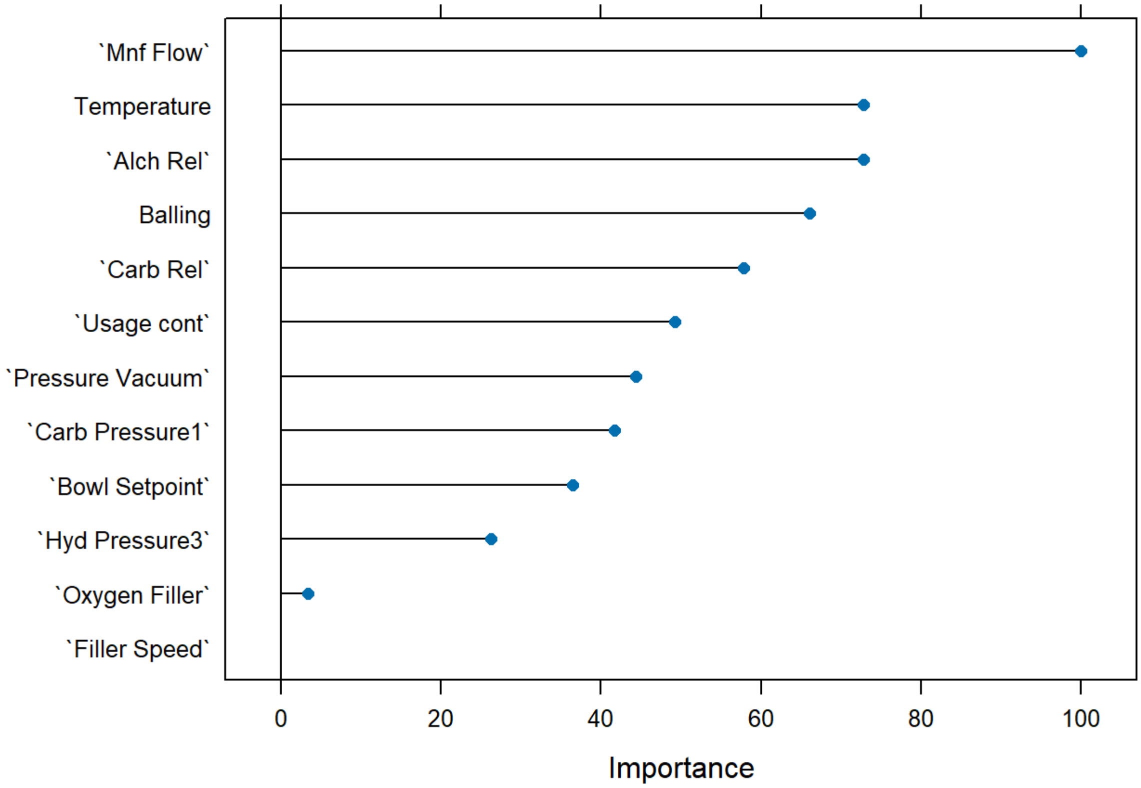

Figure 3. Important Model Variables

We recommend that ABC Beverage closely monitor and fine-tune key stages of the production process to maintain optimal pH levels. By leveraging this predictive model, the company can gain actionable insights into the drivers of pH variation which empowers smarter, data-driven decisions that ensure consistent product quality moving forward.