ABC Beverages
Non-Technical Report

Problem Statement
New regulations mandate that we document the relationship between the ingredients as well as the properties of our manufacturing process and the final acidity (PH) level of the beverages produced by our company across all brands.

Data
We used a dataset comprising 2571 records spanning 32 features and the target PH values across 4 brand codes (A, B, C and D) to train our models.
We are submitting predicted PH values on an evaluation dataset comprising of 267 records.

Data Preparation
We cleansed and prepared the data to account for missing data, less predictive features, skewed and correlated features. Based on the exploratory data analysis of the training dataset, we decided to prepare the data - this included dropping some columns and rows, creating a separate "Unknown" group for missing brand codes, creating dummy variables for the single categorical variable i.e. Brand Code, imputing missing variable data and box-cox transforming the predictors to make them less skewed.

The details of the exploratory data analysis and data preparation are available in the detailed technical report in R markdown format.

Model Building
We considered models across 3 categories: Linear, Non-Linear and Tree-based models. We trained 2 models from each category for a total of 6 models.

Model Selection
We used a combination of Root-mean-squared-error (RMSE) and R-squared as the performance metrics to decide the final model. We decided to go with the Cubist model because it's metrics were clearly better than the other models. This model has the lowest overall error and also explains the variability in the PH levels better than the other models. This is not surprising given that these models handle non-linear relationships and multi-collinearity better. This comes across in the list of top predictors selected by this model, as described in the technical report.

For the final model selected, we see that it considers the following as the top 5 predictors in terms of importance: Mnf Flow, Alch Rel, Balling Lvl, Pressure Vacuum and Brand Code C. Finding Brand Code as a top predictor is interesting because at the end of the day, Brand is a not a physical/chemical construct that can be linked to PH levels. but we think it must be best encapsulating other chemical features collectively that are in turn helpful in explaining the PH levels.

We see that the range of the predicted values in the evaluation data is in line with the range of the predicted values in the training data, which gives us confidence that the selected model seems generalizable. Besides, the general shape of the distribution of the predicted values is approximately normal.

Conclusion and Next Steps

Based on performance metrics detailed in the technical report, we recommend proceeding with the tree-based Cubist model because it provides the best accuracy coupled with more robust handling of multi-collinearity. The analysis of key features is included in the detailed technical report.

As with any real-world data science process, the logical next step would be to calculate better accuracy metrics by comparing the predicted values to the actual PH values for the evaluation data. Our recommendation is to also put in place an on-going process to keep monitoring the model and fine-tuning in case the model metrics show any deterioration.