Nontechnical Summary and Recommendations

DATA 624 - Predictive Analytics - (Project 2)

Authors: Beshkia Kvarnstrom, Nikoleta Emanouilidi, Evan McLaughlin, Victor Torres & Vladimir Nimchenko

Fall 2024

Problem Statement

Our assignment with this analysis is to predict pH levels in beverages, crucial to ensuring product standards, by applying concepts from our Predictive Analytics course to a beverage data set, assessing the data provided and determining whether we might employ a linear or non-linear regression mode and which features heave the most meaningful impact on pH levels.

Approach

We proceeded with an examination of the dataset, yielding important insights regarding the distribution of key variables and their relationships with pH. An exploration of the dataset revealed valuable insights, such as the distribution of key variables and their relationships with pH. Features like Mnf Flow, Carb Volume, and PC Volume showed potential relevance based on their variability and patterns. We employed boxplots and histograms highlighted outliers and right-skewed distributions in variables like PSC Fill and PSC CO2, prompting transformations to reduce skewness for better model performance.

Model Building and Comparison:

We tested several models, such as K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Multivariate Adaptive Regression Splines (MARS), Random Forest, Simple Linear Regression, Robust Linear Models (RLM), Partial Least Squares (PLS), Elastic Net (Enet), and Ridge Regression. Each model offers unique advantages: for example, RLM provides robustness against outliers, PLS is suitable for high-dimensional or collinear data, and Enet combines feature selection with regularization. Models were evaluated using three metrics: RMSE (Root Mean Squared Error), R^2 (R-squared), and MAE (Mean Absolute Error). These metrics measure the accuracy and goodness-of-fit of the predictions.

Variable Importance:

Our Random Forest model generated insights into feature importance. Key features contributing to pH levels include Mnf Flow (54.42), Carb Pressure1 (27.02), and Carb Volume (21.26), indicating their strong influence on pH prediction. Features like PSC Fill and PSC CO2 could be deprioritized in future analyses due to their low importance.

Models Chosen:

The Random Forest model was selected as the optimal predictive model because it outperformed other models, enjoying the lowest RMSE (0.62), highest R^2 (0.63), and lowest MAE (0.45), indicating that it explains a substantial portion of the variance in pH with superior accuracy. In contrast, models like KNN, SVM, and MARS exhibited higher error rates and lower explanatory power.

Recommendations:

- **Operational Focus:**

Production steps associated with Mnf Flow, Carb Pressure1, and Carb Volume ought to be prioritized for attention, as they were the most important features impacting pH levels. For example, improving Mnf Flow could lead to the meaningful improvements in pH maintenance. Features with low importance can be ignored for next steps.

- **Adjustments:**

Since pH decreases with increasing temperature, as we observed during our exploratory analysis, we must also assess temperature calibration for better control of pH levels.

- **Next Steps:**

Our recommendations to management would include ongoing monitoring of the performance of the Random Forest model with new infusions of data to improve robustness and maintain relevance. We would also want to investigate interactions among top variables to uncover nonlinear relationships that could also enhance our predictive capabilities.

- When we focus on key predictors and incorporating insights from our Random Forest model, we can gain greater, more precise control over pH levels which has the opportunity to enhance overall product quality, consistency, and standards.