Formal Results for Estimating Beverage PH Values

Joshua Hummell, Bharani Nittala, Eric Lehmphul

Problem statement

We are a team of Data Scientists working for a PCG company. Due to regulation, we need to better predict the PH in our beverages.

Summary

With the given data, we cannot confidently predict the PH for our line of variables, the issues begin with data gathering. We were able to understand 67.8% of the reasons for what makes up the PH value with error of about 1 PH value. Below is an overview for each individual Brand Code. We can see the highest is B and only has an understanding of 74%, when predicting we aim to achieve a minimum threshold of 85%.

Brand Code	Error	Understanding
A	0.10	53%
B	0.08	74%
C	0.15	26%
D	0.07	67%

The forecasting methods we tried included Linear Modelling and Non-Linear Methods. Linear included multi-linear regression, partial least squares, AIC optimized. Non Linear included k-nearest neighbors (KNN), support vector machines (SVM) and multivariate adaptive regression splines (MARS), rpart, and XGBoost. With XGBoost offerinng the highest RSqaured and lowest RMSE. Essentially a wide spread of tools that allowed for use to predict the data.

Conclusion and Next Steps

This brings us to the issues with the data and how we can resolve them in the face of government regulation. The first and foremost issue is that we lost about 5% of our data because the beverages were not properly labelled. Then, we noticed within the sensors themselves there were several missing values and values with wide ranges. This begs the question, how often are our sensors calibrated or tested?

The first steps we should take given the pressure for governmental regulation should be to build out better data QA procedures. We should teach the importance of labeling data so that any bottles we are tested have the full name and data.

Second, we should build out a new process to ensure our sensors are working properly and that we have the information for when they were last calibrated/inspected in a table where we can properly ensure the quality of the data we are receiving.