Project Problem Statement

This is role playing. I am your new boss. I am in charge of production at ABC Beverage and you are a team of data scientists reporting to me. My leadership has told me that new regulations are requiring us to understand our manufacturing process, the predictive factors and be able to report to them our predictive model of PH. Please use the historical data set I am providing. Build and report the factors in BOTH a technical and non-technical report. I like to use Word and Excel. Please provide your non-technical report in a business friendly readable document and your predictions in an Excel readable format. The technical report should show clearly the models you tested and how you selected your final approach. Please submit both Rpubs links and .rmd files or other readable formats for technical and non-technical reports. Also submit the excel file showing the prediction of your models for pH.

Executive Summary

New Regulations by ABC beverage company leadership requires the company’s production unit to better understand manufacturing processes, the predictive factors and its relationship to the PH values of the beverages.

Research Statement

The research is to develop a model to find the predictive variables related to the PH of beverages

Data Collection

The data set is a historic data containing predictors associated to the PH and is provided in an excel format. We will utilize the training dataset to analyze the PH of beverages and test dataset to predict PH. Two excel files are provided:

* The training data (StudentData.xlsx)
* The test data (StudentEvaluation.xlsx).

Model Selection

We will attempt multiple machine learning approaches and use RMSE and R2 to determine the best approach. The reason to use these measures is because they are widely used for finding the accuracy in regression models which uses Ordinary Least Squares. Both uses SST (difference in distance of the actual value and the baseline model which is generated from the mean) and SSE (difference in distance between the actual value and predicted model which we consider to be the best fit model). We will also review the performance of our best model in consideration to the accuracy gained.

Linear Regression Models

Linear regression is a [linear](https://en.wikipedia.org/wiki/Linearity) approach to modeling the relationship between a scalar response (or [dependent variable](https://en.wikipedia.org/wiki/Dependent_variable)) and one or more [explanatory variables](https://en.wikipedia.org/wiki/Explanatory_variable) (or [independent variables](https://en.wikipedia.org/wiki/Independent_variable)). In this instance, due to multiple dependent variables, we use a process called multiple linear regression. Three different approaches are selected, generalized linear model (GLM), glmnet and partial least squares (PLS).

The performance of each linear regression model is discussed within the technical document. The model chosen is the generalized linear model (GLM), although not the best performing, was very close in performance. The best model showed only a 0.001 improvement in the RMSE values. Due to computing burdens, we elect the simpler GLM model.

Non-linear Regression Models

Tree Models

Conclusion