**Final Non-Technical Report DATA624**

**May 15, 2022**

**Authors: Jack Wright, Richard Zheng, Mario Pena**

Project Overview:

Leadership at ABC Beverage has informed the head of production that new regulations are requiring them to understand the manufacturing process, the predictive factors and be able to report to them our predictive model of PH.

We will use the historical data set we have been provided. Build and report the factors in BOTH a technical and non-technical report.

Background:

The data we were provided include our target variable (PH), and other components such as “hydrogen pressure”, “density”, “carb volume”, “oxygen filler”, and “temperature” among others that will help us understand, model and predict (PH). There are 2571 observations of 32 predictors and the response (PH). All data are continuous except the "Brand Code" which is unordered categorical.

Our analysis begins with the usual first steps, examining data for missing observations, data exploration, and preparation. The data is split into a train and test set. Models are trained on the train set using various parameters and tuning techniques to provide an unbiased estimate of the accuracy of each model.

After analyzing the missing values from the data we found that 94% of the time when `Filler Speed` is missing, `MFR` is also missing. It was also found that `Brand Code` can be predicted with high accuracy using the other predictors and we advised to impute the missing values from this predictor for our models. Additionally, `MFR` and `Brand Code` account for 90% of the missing data and some preliminary modeling showed that `Brand Code` was very responsive to imputation as well. However, we did note that there is missing data in the test set. In order to minimize leakage of bias from the training to the test set, we will impute the training set with the mean values of the test set. Since the dataset is fairly large and the remaining data points seem missing at random, we will drop the remaining observations with missing data.

Our exploratory analysis showed that a large number of the predictors were skewed (21 of the 32 numeric predictors), so we will center and scale the data for models that do not handle skewed data or unscaled predictors well, like nearest neighbor or support vector machine. As for the other methods such as random forest, we will leave the data as is.

Methodology:

Since the response variable (PH) is continuous, we will use a series of regression models as candidates. We will train a random forest and XGboost because they have generally proven to be excellent at modeling continuous variables. Our exploratory analysis pointed us towards some potential clustering in the most important predictors, so we will train a radial SVM model. Additionally, the exploratory analysis didn't point us towards any strong linear correlations, and the potential clustering of predictors led us to include a nearest neighbor model for performance comparison. We also explored a linear regression model for purposes of comparison between all our models. Lastly, we used 5 cross fold validation to validate models.

Furthermore, the principal components analysis we implemented on the data allowed us to improve the performance of our models by virtue of its feature extraction, eliminating predictors that do not contribute in any decision making. PCA, also gave us insights as to the most important variables that can potentially be manipulated to achieve a desired (PH) value, and it was found that about 90% of the models’ total variance can be explained with the first 16 principle components.

Project Findings:

As discussed above, the results from principle components analysis pointed us to some of the important variables in the data that can be predictive of (PH), and some of these include `Mnf Flow`, `Hyd Pressure2`, `Fill Pressure`, `Pressure Setpoint`, `Pressure Vacuum`, `Filler Level`, `Usage Cont`, and `Oxygen Filler`. The positive elements are `Mnf Flow` `Hyd Pressure2`, `Fill Pressure`, `Pressure Setpoint` and `Usage Cont`. The negative elements are `Pressure Vaccum`, `Filler Level` and `Oxygen Filler`. In this context positive means that as that variable increases, we move in a positive direction along this principle component, while negative means that when that variable increases we move in the negative direction along that principle component.

After training various models we found that the nearest neighbor model was the best performing on both the unseen test data and the training data. The performance metrics of this model are also within a range that over-fit would not be a concerned, while giving us confidence that it can predict values accurately.