**Data 624 Predictive Analytics**

**Project #2**

**Summer 2020**

**Project Format: Team, See Below**

**Project Due: July 18, 2020; Midnight ET**

**Project #2 - Requirements**

You are given a simple data set from a beverage manufacturing company.  It consists of 2,571 rows/cases of data and 33 columns / variables. Your goal is to use this data to predict PH (a column in the set).  Potential for hydrogen (pH) is a measure of acidity/alkalinity, it must conform in a critical range and therefore it is important to understand its influence and predict its values.   This is production data.  pH is a KPI, Key Performance Indicator.

You are also given a scoring set (267 cases).  All variables other than the dependent or target.  You will use this data to score your model with your best predictions.

**Submission**

You are to submit a professional, easy to read report. The consumers of this report are executives, data scientists and engineers. You need to ***communicate to all audiences***; therefore you cannot just present a technical report. You should provide commentary on your approach, why you are taking this approach and your findings along the way. The report should be very easy to navigate, follow and understand. You must explain what/how/why. And submit your scored results.   Your representative will submit the materials to me in an email with a minimal of two attachments – A Word readable doc (Report), An Excel readable doc (my XLS to you with the predictions), all as before.  An R Markdown file is appreciated, but not required. I will need your code either in the Word document **or** the markdown so I can reproduce results. Please include all libraries you are using – all code from A to Z and the code should be well documented as if you are passing off to a production engineer.

Note the modeling/scoring in this exercise is not really that difficult, you can differentiate your team in your report.

**Grading Rubric**

Report Clarity / Readability - 40

Report Content (Discussion) - 20

Report and Project Completeness – (weighted across other categories)

Report and Project Technical Competency - 20

Scored Results (MAPE) - 20

**Some additional thoughts**

How will you handle new predictions with missing values?  Never seen inputs?

What are the most important predictors?  Can you represent visual relationships?

If you are an engineer, how would you change your work given your analytic results?