**City University of New York- School of Professional Studies**

**DATA 624 – Predictive Analytics**

**Project -2.0**

**Title: Prediction of factors affecting PH levels of Beverages**

**Presenters: Data Analytics Team – Group 4**

**(Ahmed Sajjad, Chad Smith, Harpreet Shoker, Jagruti Solao, Todd Weigel)**

**(In alphabetical order)**

**Abstract:**

As new regulations are being imposed by Food and Drug Administration (FDA) and Food Safety inspection Services (FSIS),the organization is required to have detailed study conducted towards understanding and improving the quality of manufacturing process for beverages. Factors affecting PH values are examined and predictions are made for an optimized PH value our company could offer for customers to help them achieve their health goals.

**Motivation:**

Sweetened and flavored beverage consumption has increased dramatically over the past 35 years in the United States with carbonated soft drinks being consumed the most frequently, and most often by children, teens, and young adults [1]. . The pH of commercial nonalcoholic, nondairy beverages ranges from 2.1 (lime juice concentrate) to 7.4 (spring water). Acids are added to beverages and compose a flavor profile giving the beverage a distinctive taste. Studies suggest that pH is the primary determinant of beverage erosive potential.[2] Due to the fact that the consumption of beverages is high in USA, the factors affecting PH values in manufacturing process are needed to be evaluated.

**Approach:**

To optimize and gain maximum accuracy, team worked with ‘------------’

Machine learning model under supervised learning umbrella.

Following are the independent variables we worked with to train our model.

1)Brand code – Only categorical variable

2) Carb Volume

3) Fill Ounces

4) PC Volume

5) Crab Pressure

6) Carb temp

7) Temp

8) PSC

9) PSC CO2

10)Mnf flow

11)Carb Pressure

12) Fill Pressure

13) Hyd Pressure 1

14) Hyd pressure 2

15) Hyd pressure 3

16) Hyd pressure 4

17) Filler Level

18) Filler Speed

19) Temperature

20)Usage cont

21) Carb Flow

22) Density

23) MFR

24) Balling

25) Pressure Vaccum

26) Oxygen Filler

27) Bowl Setpoint

28) pressure Setpoint

29) Air Pressure

30) Alch Rel

31) Crab Rel

32) Balling Lvl

The prediction variable to predict – PH

**Step I: Data collection and data cleaning:**

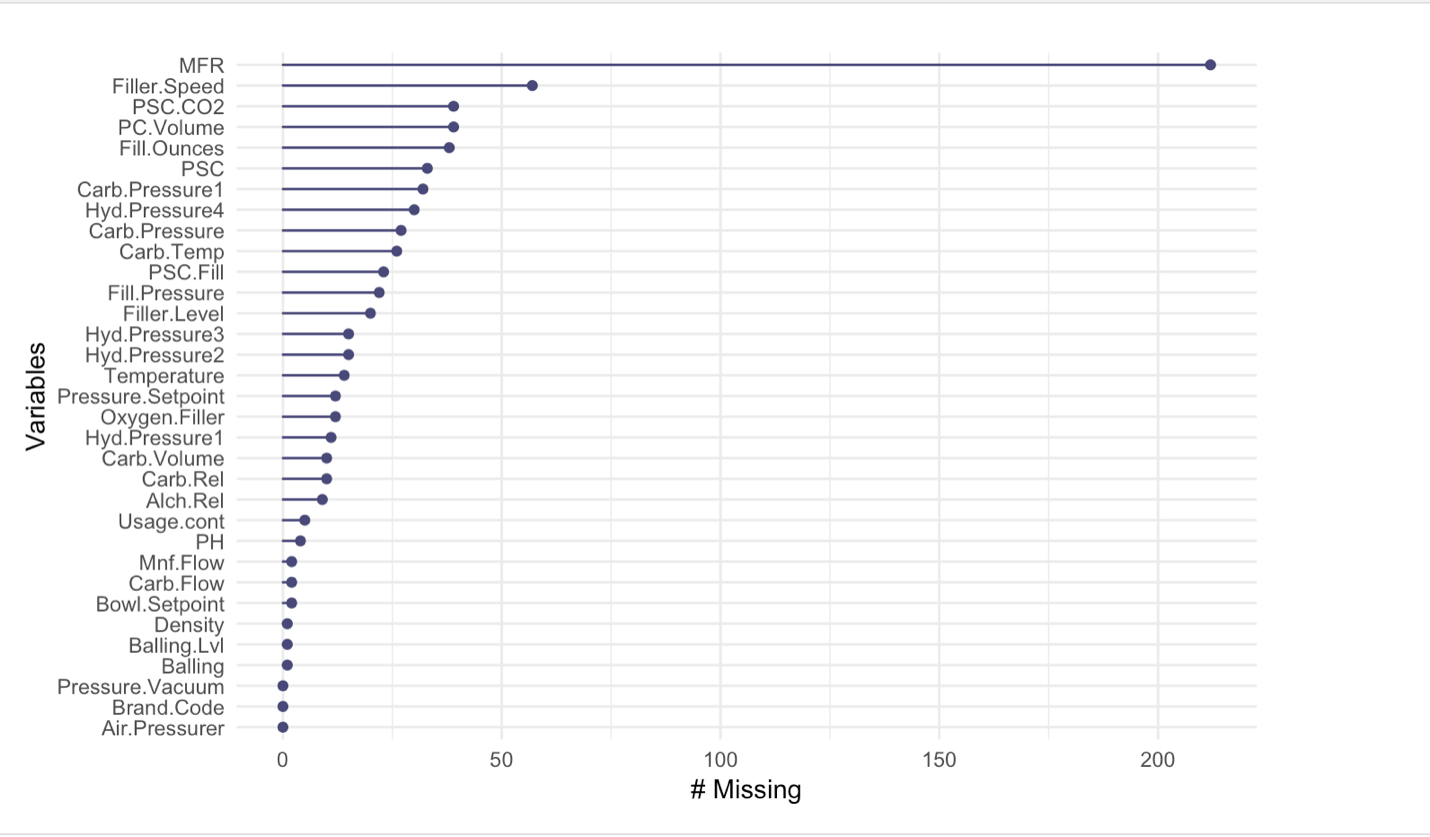
**Tool and Platform used: R studio Cloud**

* Data provided is set in two parts

1. Student data
2. Evaluation data

As the data is loaded in R, missing values in data are replaced by computed values with average of values from previous and next field.

The variables with missing values are identified with graph below:



‘MFR’ is variable with highest number of missing values.

‘Brand code’ is the categorical variable which is converted to numerical one by replacing it with integers. As we name our integers in order (1,2,3,4) we need to keep in mind that categorical variables are never in increasing order of values, but numerical replacements are (4>3>2>1), this could add error while building the model. To avoid this scenario, four new dummy independent variables are created and added to overall dataset. Dummy variable are variables containing values such as 1 or 0 representing the presence or absence of the categorical value.

By including dummy variable while building model however, one should be careful of the ‘Dummy Variable Trap’. The Dummy Variable trap is a scenario in which the independent variables are multicollinear - a scenario in which two or more variables are highly correlated.

Student data provided is partitioned into train and test data in ration 80:20. Due to the fact that we will build our model on 80 % of student data ,20% can be used to test result. Achieving desired accuracy, we can use this model to get predictions for Evaluation data.

**Step II: Preprocessing Data**

|  |  |
| --- | --- |
| Step II is to center, scale and perform boxcox transform the data. Note  Each of these tests is performed individually, and highest performance result  was taken into account.   * Principle Component Analysis gave negative impact of RMSE even after   removal of any low performing predictors. If we want to improve  performance, we could do PCA at 95% and remove about a third of the  variables.   * Cross Validation is performed to improve results, by reducing chance of   poor randomness in any initial partitioning.   * Cubist Grid: The best performance from a number of different models were   compared. Cubist grid was the best of those tested. Neural Networks  were a close second but took forever (hours) to get close.  SVM was also close in performance.  **Procedure :** First set up the grid of test data. "Committees" are the sets of  models tested and chained together. Prior models affect future models, as  each data point's weight in the current model is adjusted based on its  performance in the prior model.  Neighbors is a KNN algorithm that can further adjust the model based on  each datapoints neighbors. The print of the cubist grid shows all the variations  that will be tested. Note, while developing, no test with small committees was  chosen as the best model, so we are only searching with committee size of 50  or above.  **Step III: Exploratory Analysis**  In cases with vast number of Independent variables, it is often difficult to get  spread of variables or the distribution of variable by taking a look at data.  Following distribution plot helps to check the distribution of all our parameters.  (Variables and their biases) |  |
| All the PSC variables are heavily biased towards left whereas parameters such  as ‘Usage Count’,’Carb.flow’, ’MFR’ are skewness to right. Strong skewness is  Sign for presence of outliers. In order check outliers’ boxplots are plotted  as shown in below.        \*\*\* Write about outlier observations   * Correlations between variables:     Correlation plot shows following pairs are highly corelated to each other.   1. Carb.Volume with Density, Balling, Alch.Rel, Carb.Rel, and Balling.Lvl 2. Carb.Pressure with Carb.Temp 3. Filler.Level with Bowl.Setpoint 4. Filler.Speed with MFR   Let us now look at the correlation between the target (pH) variable and the  predictors. |  |
|  | We then call the partition function, to create the dftrain, dftest and the PHTrain and PHTest datasets. |
| \*\*\* Add text here  **Step IV : Statistical inference and Model Building**  **\*\*\*** Add text here  **Step V: Predictions**  **\*\*\*** Add text here  **Step VI: Conclusion** |  |
|  | From results of the preprocessing, we see all values were scaled and centered, and 22 had a boxcox transformatapplied. |

**Citations:**

1. Heller KE, Burt BA, Eklund SA. Sugared soda consumption and dental caries in the United States. J Dent Res. 2001;80(10):1949-1953.
2. <https://www.ada.org/en/~/media/ADA/Public%20Programs/Files/JADA_The%20pH%20of%20beverages%20in%20the%20United%20States>