****

**DATA 624 | Predictive Analytics**

**Project 2 - ABC Beverage Prediction Analysis**

[GitHub](https://github.com/gabbypaola/DATA624/tree/main/Project%202) [RPubs](https://rpubs.com/gpmartinez/901220)

**Company: ABC Beverage**

**Director of Production: Jeff Neiman**

**Production Data Science Team: Gehad Gad, Karim Hammoud, Gabriella Martinez**

**Purpose:**

Leadership at ABC Beverage is concerned about new regulations and is requiring the team to understand the manufacturing process, factors, and provide a predictive model of pH.

**Summary and Key Takeaways:**

Using the historical dataset multiple predictive models were created to understand the process and key factors, using statistical and machine learning models like linear, nonlinear, and tree-based then choose the best model to improve the beverage process and ensure compliance to new regulations. Major findings from our models and analysis include:

* The Cubist model found the following variables to be important components predictive of pH values: Mnf. Flow, Alch. Rel, Bowl.Setpoint, Pressure.Vacuum, and Temperature.
* It is recommended for production to prioritize these factors in the manufacturing process.
* It is also recommended to conduct periodic revisions and refinement of the model as new and other relevant data to the process becomes available.

**Details on Process and Models:**

The response variable, `PH` represents the pH level of each beverage. Our approach includes generating multiple models and selecting the best performing model for the data, followed by prediction. We were provided with test and train data. The train data was used to create and evaluate the models, and the test data was used to generate the predictions resulting from our modeling process. Our approach consisted of the following steps:

* Exploratory Data Analysis
* Data Preparation
* Model Building
* Model Evaluation and Selection
* Prediction

Pre-processing of the data was needed based on the distributions and missing values noted in the training data set. The training data for linear and non-linear needed to be normalized whereas the data did not need normalization for the tree based models.

**Models used:**

1. Linear Regression Models
   1. Multi-linear regression
   2. Partial least squares
   3. AIC optimized
2. Non-Linear Regression Models
   1. K-Nearest Neighbors
   2. Support Vector Machines
   3. MARS
   4. Neural networks
3. Tree-Based Models
   1. Random Forest
   2. Boosted Trees
   3. Cubist

RMSE was used as the metric to evaluate and choose the best model, because it penalizes large errors compared to MAE. The linear models' performance is almost the same as there are very small differences in the RMSE. Yet, the best model based on this iteration in the linear regression with RMSE 0.12761

In the non-linear models, the RMSE has been reduced by an average of 9% compared to the linear models. The best non-linear model to be used to make prediction is Neural Network model as the RMSE for this model is 0.112126

The tree models performed better than linear and non-linear models. The RMSE was reduced by an average of 12% compared to the non-linear models and 20% compared to linear models. The best model to be used for prediction for this data is one of the tree-based models which is Cubist with RMSE of 0.0952903.

**Limitations:**

The training data exhibited high correlations between predictor variables which hinder performance. The removal of highly correlated variables in the pre-processing stage results in a loss of potentially informative features in order to reduce the effects of multicollinearity. Moreover, the training set was relatively small and could potentially have benefited from added observations. Despite these concerns, the preprocessing measures taken into account mitigate the effects of multicollinearity in the final model.

**Recommendations:**

Some recommendations from the Production Data Science team for future iterations of the predictive model and the manufacturing process include:

1. Feature Engineering
2. Advanced model Tuning/optimizing
3. Increase the iteration number for each model
4. Controlled studies to uncover causal relationships with variables to pH
5. Create a process that ensures production is prioritizing the key factors affecting pH