DATA 624 - Predictive Analytics

Fall 2020 - Project #2

ABC BEVERAGE NON-TECHNICAL REPORT

Team Members:

Vijaya Cherukuri, Samantha Deokinanan, Habib Khan, Priya Shaji

Table of Contents

Problem Statement

Due to new regulations by ABC Beverage, the company leadership requires that the production team have a better understanding of the manufacturing process, the predictive factors and their relationship to the pH of the beverages. Therefore, this project is an effort to find the optimal predictive variables related to the pH of the beverages and evaluate the accuracy of the same with rigorous statistical testing.

Analysis

Analysis of the project consists of data exploration which includes exploring missing values, outliers, data distribution and correlation of variables. It is then followed by data preparation which comprises of preparing dataset based on the points of data exploration, that is, replacing missing data points (Fig 1) with an appropriate value, normalizing data distribution (Fig 2), removing outliers etc.

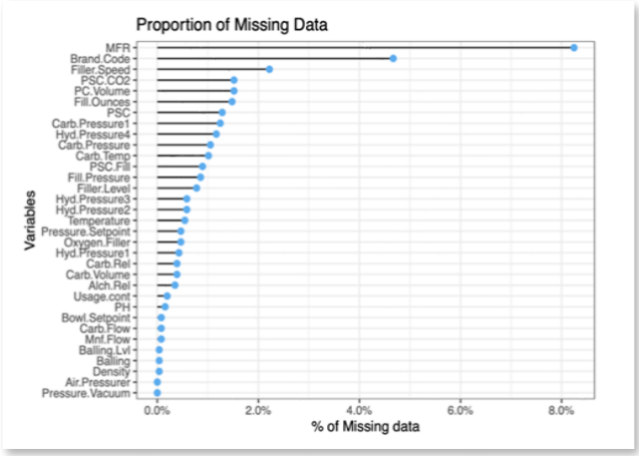

Fig 1

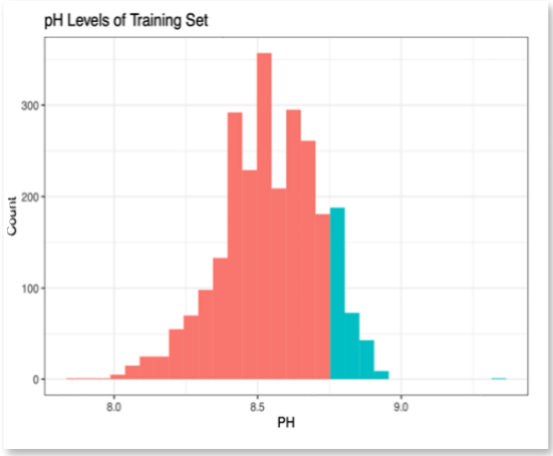

Fig 2

The final pre-processed dataset is then split into test and train datasets to build models and evaluate them.

Building Models

Following models are build using training dataset: Baseline Model, Multivariate Adaptive Regression Splines (MARS), Cubist model, Partial Least Squares, Gradient Boosting, Random Forest (RF), Ensemble Regression model out of which RF has the lowest RMSE or error rate value of 0.104. Fig 3 shows model importance of above mentioned regression models.

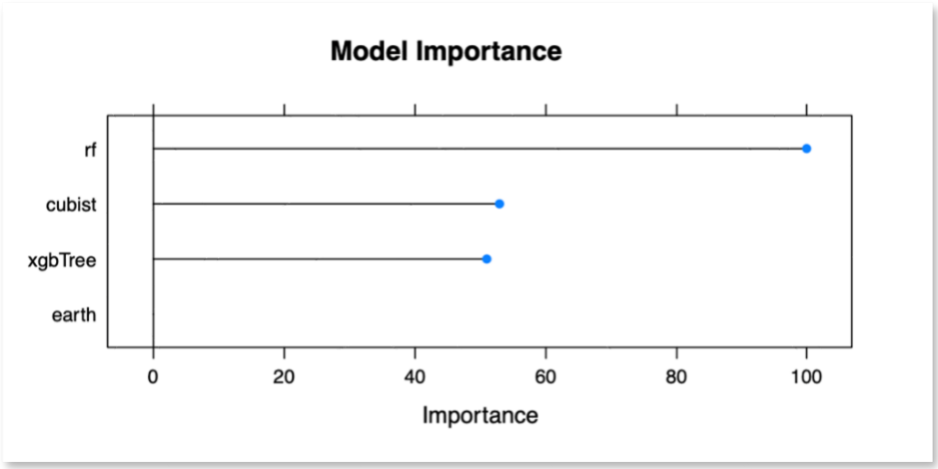

Fig 3

Important Variables

Model selection is followed by important feature extraction. According to the selected model, which is random forest, top three informative variables are Mnf.Flow, Brand.Code and Usage.cont (Fig 4).

According to the model, minimum night flow (Mnf.Flow) affects pH value of beverages to a high extent followed by brand codes and usage content.

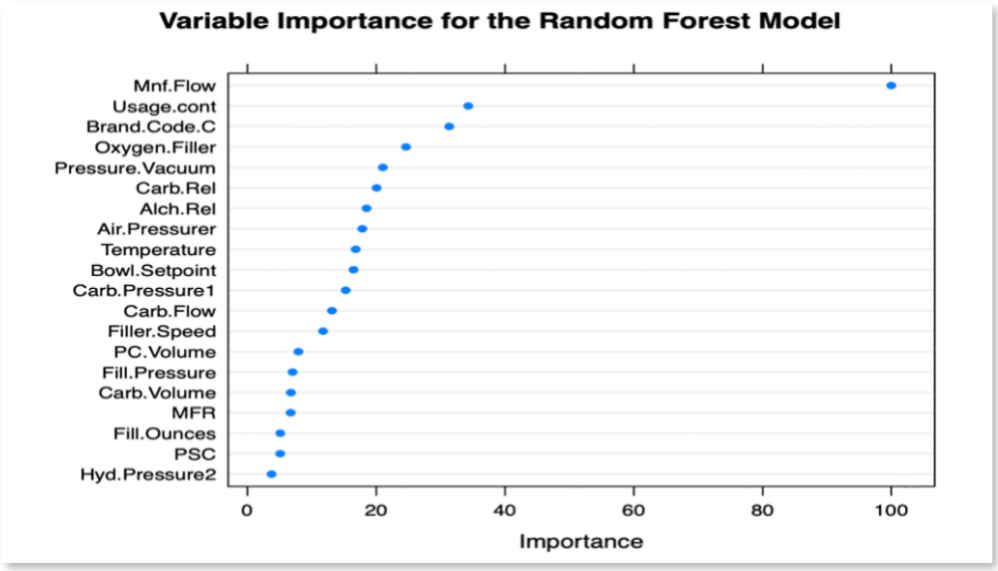

Fig 4

Model Evaluation

In this section, we selected the best model to make predictions.
For exceptional prediction quality, a meta-algorithm such as the `ensemble` method proved to be beneficial since it combines several learning techniques into one predictive model in order to decrease variance, bias, and improve predictions.

The model with both the smallest errors and also accounted for the largest proportion of the data variability is the ensemble model with an RMSE value of 0.100.

This meta-model outperformed other linear, and tree-based models in every resampling performance metric, and it proved to be superior to Random Forest, the best single model (Fig 5).

Performance metric for all models

	RMSE	Rsquared	MAE
MARS	0.126	0.476	0.096
CUBIST	0.103	0.649	0.075
PLS	0.139	0.367	0.108
GBM	0.109	0.612	0.080
RF	0.102	0.665	0.072
Ensemble	0.100	0.674	0.071

Fig 5

Prediction Result

Based on model evaluation, ensemble model is selected as the optimal model. Therefore, ensemble model is used to make the pH prediction of beverages.

Fig 6 shows plot of actual vs predicted values of ensemble model with RMSE value of 0.10 and Fig 7 shows data distribution of predicted pH values.

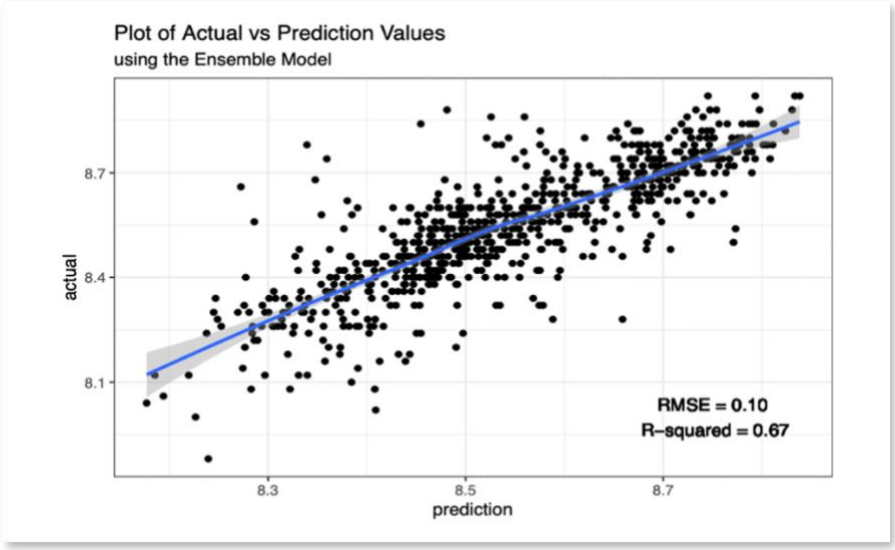

Fig 6

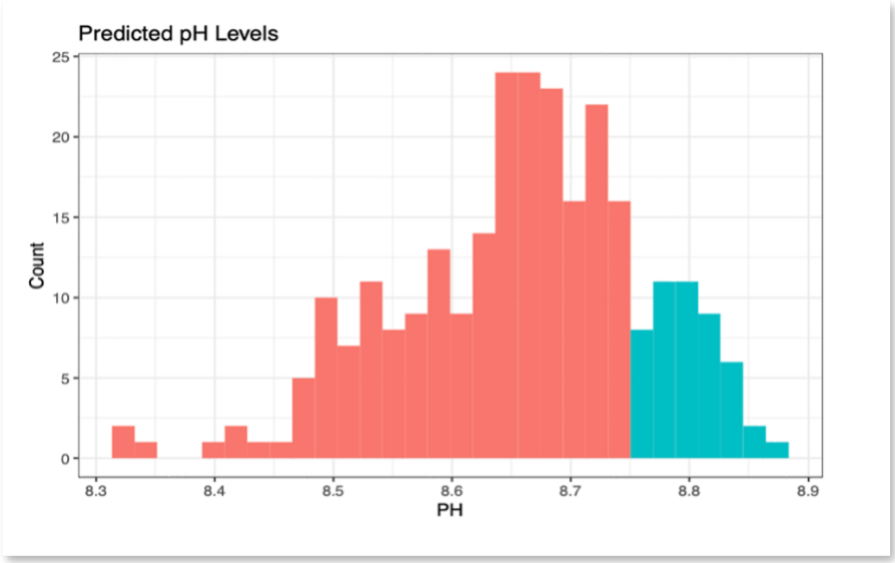

Fig 7