An Analysis of ABC Beverage Manufacturing Process

Salma Elshahawy, John K. Hancock, Farhana Zahir

May 15, 2021

Executive Summary

ABC is a beverage manufacturer. The team was given historical data on the manufacturing processes of five brands. The objective is to determine which of the processes can help predict the PH level. Several linear, non-linear and tree models have been tested and it was found that the Cubist model works best in predicting the PH. The top five predictors are found to be Mnf.Flow, Density, Temperature, Pressure.Vacuum, and Filler Level. Two discrete categorical factors, Brand Codes C and D, are also in the most important predictors.

Problem Statement

Due to new regulations, the ABC Beverage co management requires the production team to have a better understanding of the manufacturing processes and their relationships to the PH of the beverages produced. This project attempts to find the optimal predictive variables and evaluate the accuracy of the models with statistical testing.

Data Analysis

In the data exploration process, the team has looked for missing values, outliers, data distribution and correlations between predictor variables.

In the training set, there are 2,571 observations consisting of 32 predictor variables and one dependent variable, PH. We also see that “Brand Code” is a factor variable that will need to be handled as well as several observations with several NAs. The test set used for prediction has 267 observations, the 32 predictors, and the dependent variable PH which is all NAs.

The detailed process followed is outlined below:

A. *Isolate predictors from the dependent variables* – Separate PH from the other variables

B. *Correct the Predictor Names* – Remove space between names, for e.g. Brand.Code instead of Brand Code

C. *Create a data frame of numeric values only*- Process numerical and categorical variables separately.

D. *Identify and Impute Missing Data*

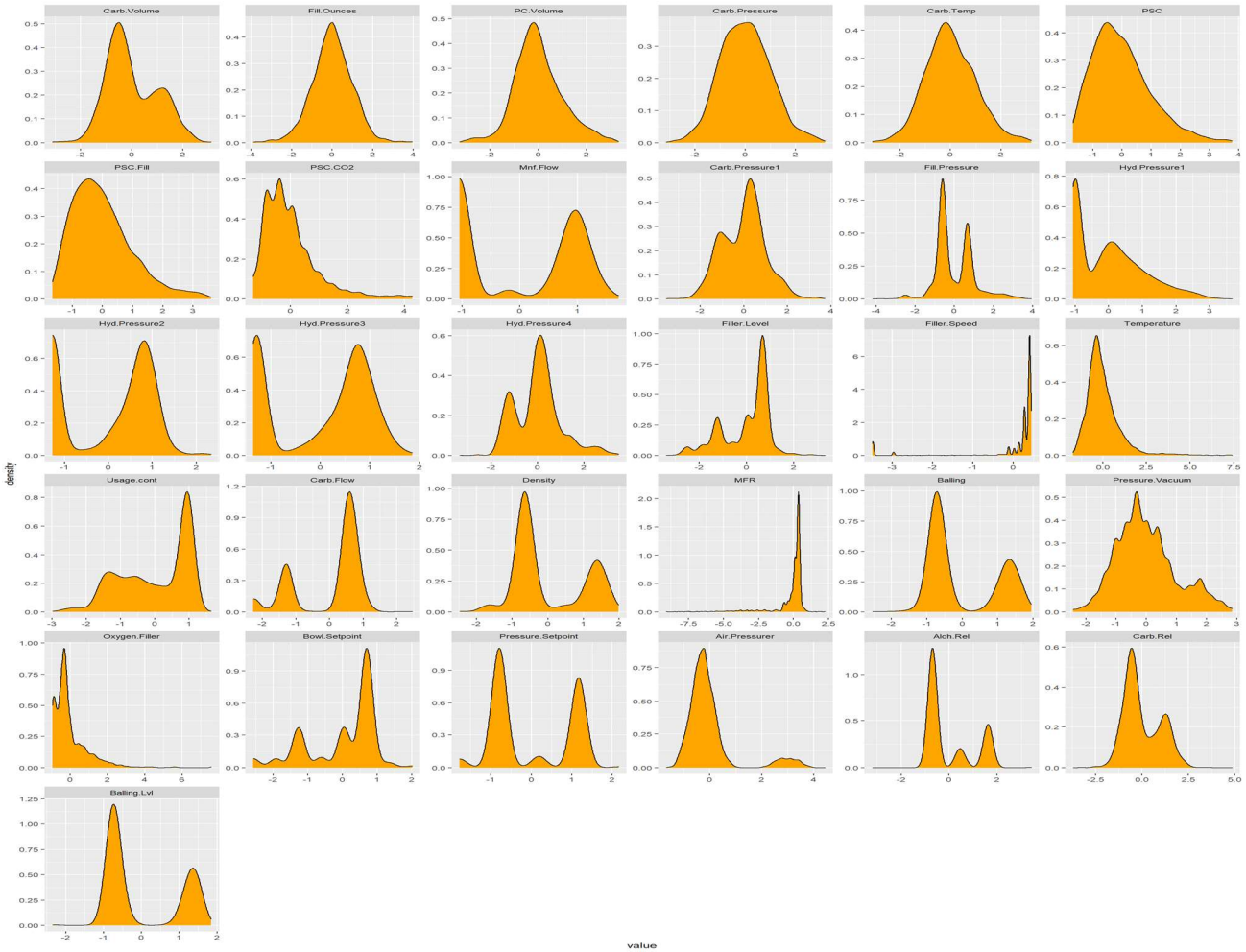
There are 212 missing values for MFR. KNN imputation was used to impute as in the case of further information about the processes and the distributions the variables are supposed to value,

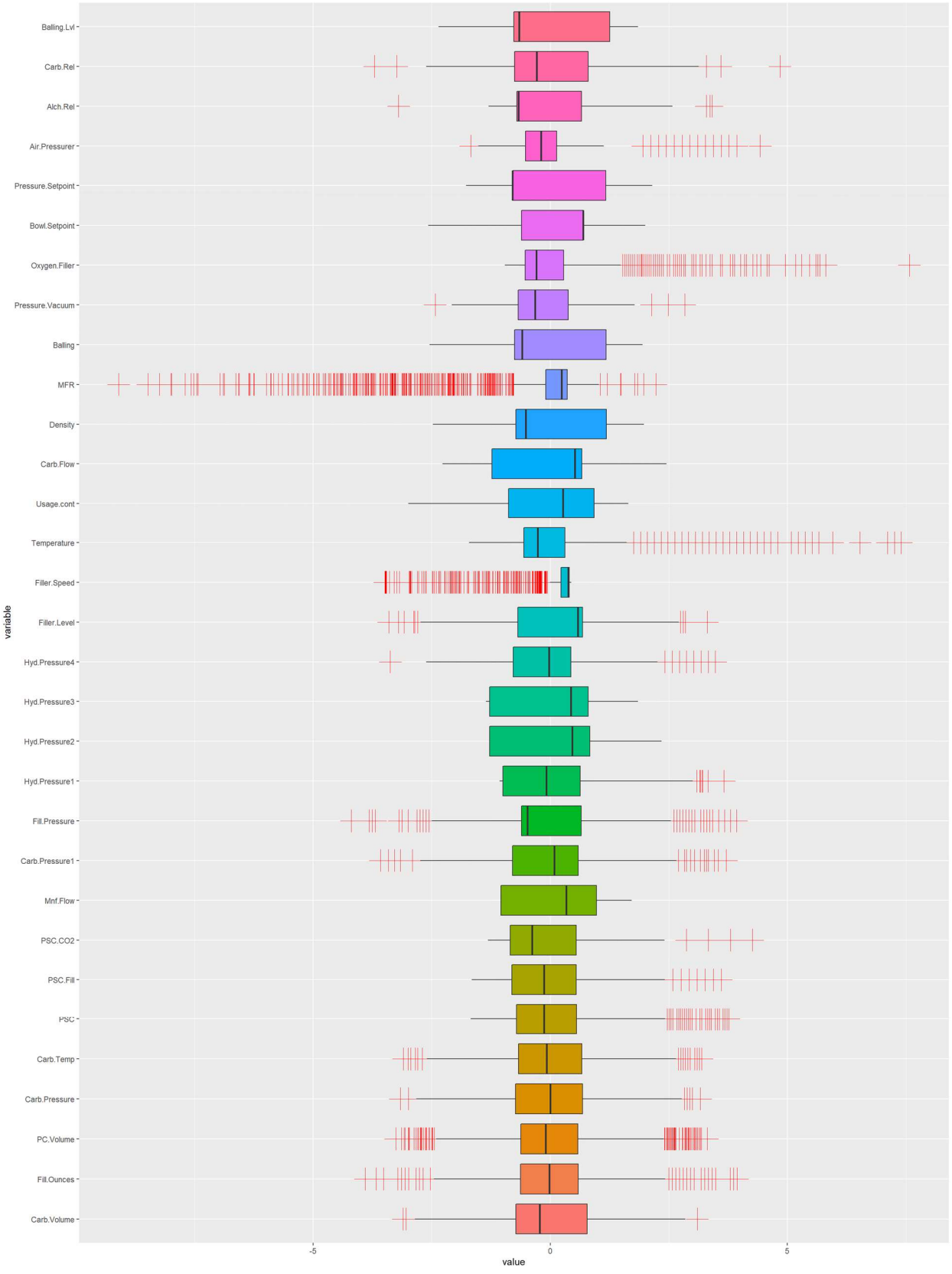
kNN of using nearest neighbors is a safe bet. The following table shows the variables with top missing observations.

Predictors	NAs
MFR	212
Filler.Speed	57
PC.Volume	39
PSC.CO2	39
Fill.Ounces	38
PSC	33

E. *Identify and Address Skewness and Outliers*

There are only four predictors that are normally distributed. The box plots show a high number of outliers in the data. To correct for this, we centered and scaled these distributions. The following shows the distributions before the preprocessing.



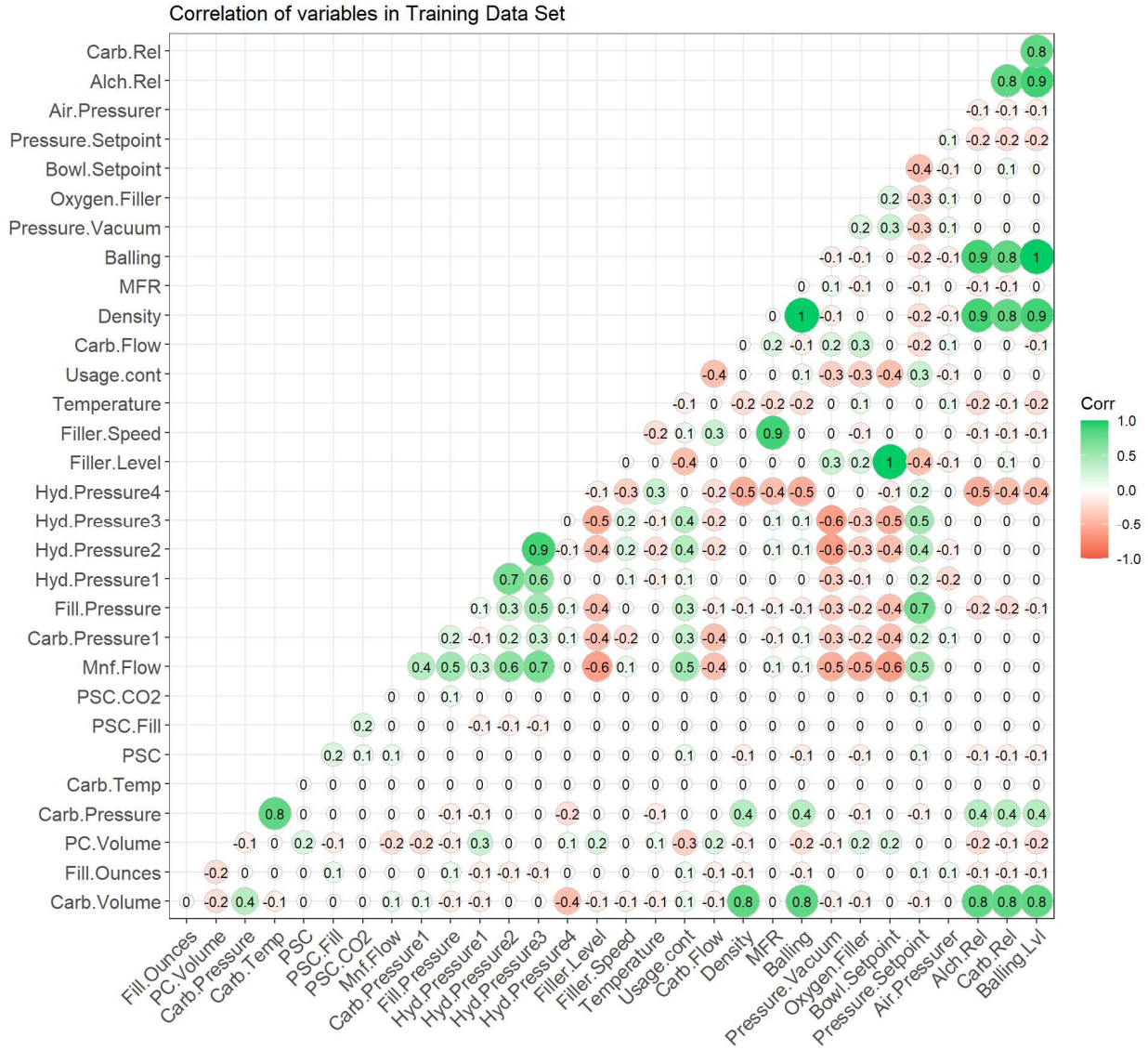

Balling.Lvl
Carb.Rel
Alch.Rel
Air.Pressurer
Pressure.Setpoint
Bowl.Setpoint
Oxygen.Filler
Pressure.Vacuum
Balling
MFR
Density
Carb.Flow
Usage.cont
Temperature
Filler.Speed
Filler.Level
Hyd.Pressure4
Hyd.Pressure3
Hyd.Pressure2
Hyd.Pressure1
Fill.Pressure
Carb.Pressure1
Mnf.Flow
PSC.CO2
PSC.Fill
PSC
Carb.Temp
Carb.Pressure
PC.Volume
Fill.Ounces
Carb.Volume
variable
-5
0
5
value

F. *Check for and remove correlated predictors*

We identified five variables that are highly correlated with other variables at above .9. These variables are ‘Balling’, Hyd.Pressure3’, Balling’Lvl’, ‘Alch.Rel’ and ‘Bowl.Setpoint’.

Highly correlated variables lead to Multicollinearity which reduces the precision of the estimate coefficients and weakens the statistical power of regression models. We therefore removed these variables from the dataset.

The following shows the correlation for all the predictors.



G. *Identify Near Zero Variance Predictors*

NearZeroVar diagnoses predictors that have one unique value (i.e. are zero variance predictors) or predictors that are have both of the following characteristics: they have very few unique values relative to the number of samples and the ratio of the frequency of the most common value to the frequency of the second. There were no zero variance predictors in the dataset.

H. *Impute missing values and Create dummy variables for Brand.Code*

Earlier, we saw that there are 120 missing values for Brand.Code, a factor variable. The imputation strategy here is to impute with the most frequent value, "B". After imputation, Brand.Code was converted to dummy variables. The converted Brand.Code predictor is joined to the num_predictors_02.

I. *Impute missing data for Dependent Variable PH*

The median values were used to impute the missing values for the dependent variable PH.

Models Used

The models tested can be divided into Linear, Non-Linear and Tree based.

Linear:

Basic linear, Partial Least Squares, Ridge Regression

Non-Linear :

kNN, Neural Network, MARS, Support Vector Machines

Tree based :

Single, Bagged, Random Forest, Gradient Boost, Cubist

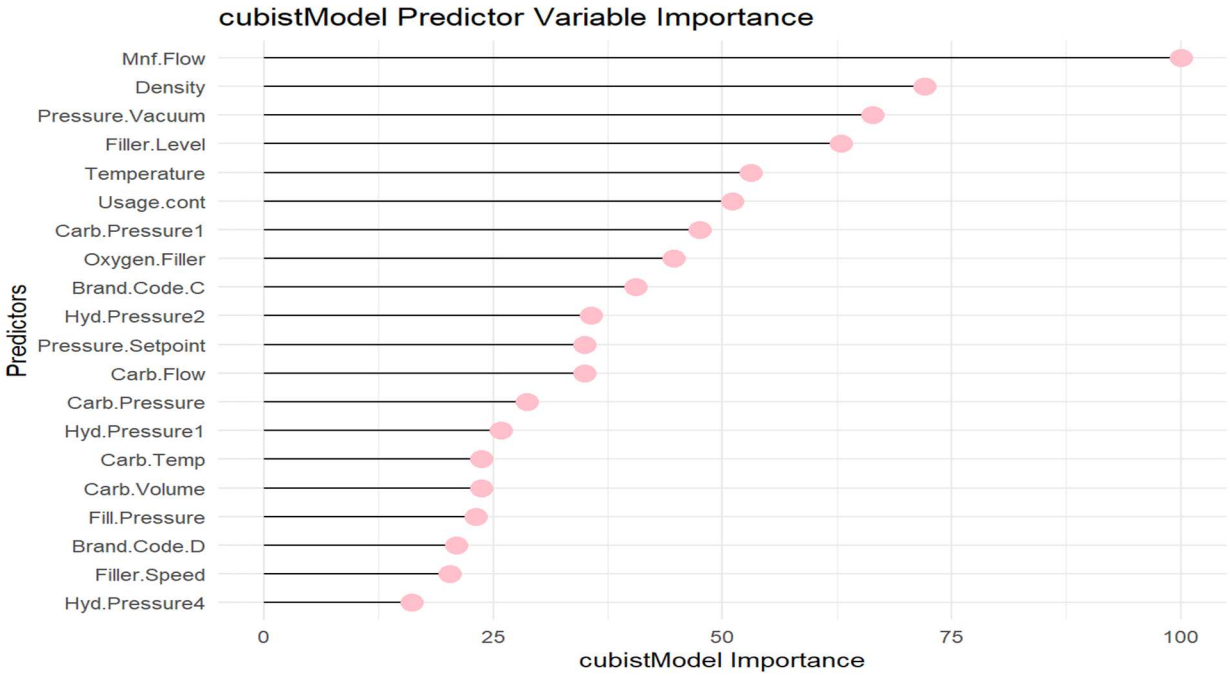
For each model, the best tuned parameters with lowest errors were chosen as the final.

Findings and Conclusion

It was found that the Cubist model had the lowest RMSE (0.10976) value as well as the lowest MAE value (0.081). It also had the highest Rsquared value (0.601).

Model	RMSE	Rsquared	MAE
baggedTree Model	0.110161222176994	0.587582121274168	0.0815608529120577
Cubist Model	0.114240479875035	0.584813533518392	0.082990152334238
Random Forest Model	0.117440026812849	0.589849374176466	0.086826432034632
Gradient Boost Model	0.121617924199404	0.537649903321855	0.091974418000837
KNN	0.127265605246562	0.443813404567053	0.0982312937654356
cTree Model	0.130743440773608	0.461731568560653	0.0976388347059531
Linear Model	0.134119241211064	0.384037147180449	0.105085743503772
Partial Least Square	0.134146647737492	0.383603238597427	0.105370300144996
Ridge Regression	0.134535682073648	0.376695740032475	0.105548235473887
Multivariate Adaptive Regression Spline	0.135101845051539	0.417088631109576	0.103248537986403
Support Vector Machines - Linear	0.141656699921108	0.359436260223437	0.105443174125596
Neural Network	7.55157663076987	NA	7.54950649350649

The data science team found that the Cubist model is the best for predicting the PH value. The most important predictors from this model are shown in the visualization below. The top five predictors are Mnf.Flow, Density, Temperature, Pressure.Vacuum, and Filler Level. Two discrete categorical factors, Brand Codes C and D, are also in the most important predictors.



We have exported the predicted PH values in the attached excel file.