Data 624 – Predictive Analytics

Project 2

Non-Technical Report

Amanda Arce, Amit Kapoor, Jatin Jain

Table of Contents

Overview

ABC company is a beverage manufacturer and we have been given a task to analyze its data of manufacturing processes and predict the PH level. We have analyzed the data, build several models to improve the accuracy of prediction. At the end, we have determined the model with best prediction accuracy and found the important factors that should be considered to determine PH in future manufacturing.

Problem Statement

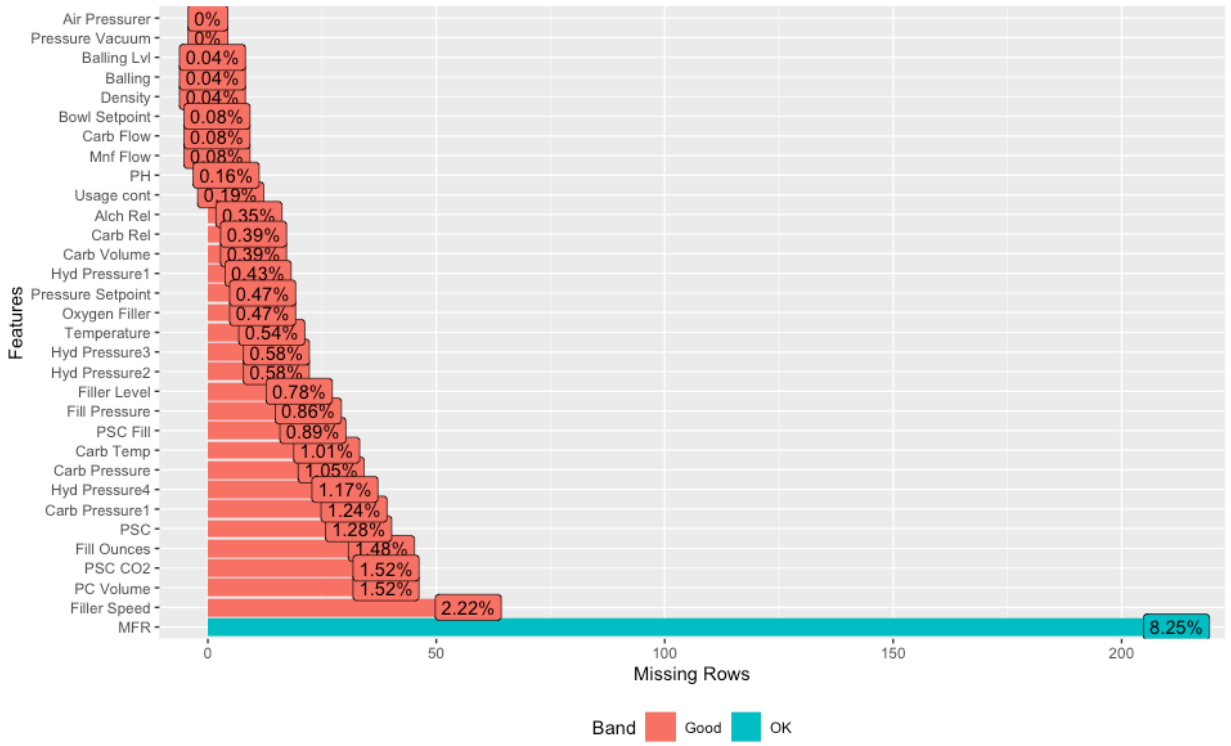
ABC Beverage has new regulations in place and the leadership team requires the data scientists team to understand the manufacturing process, the predictive factors and be able to report to them predictive model of PH. The selection of model depends upon various factors like model accuracy, data relevance, cross validation etc.

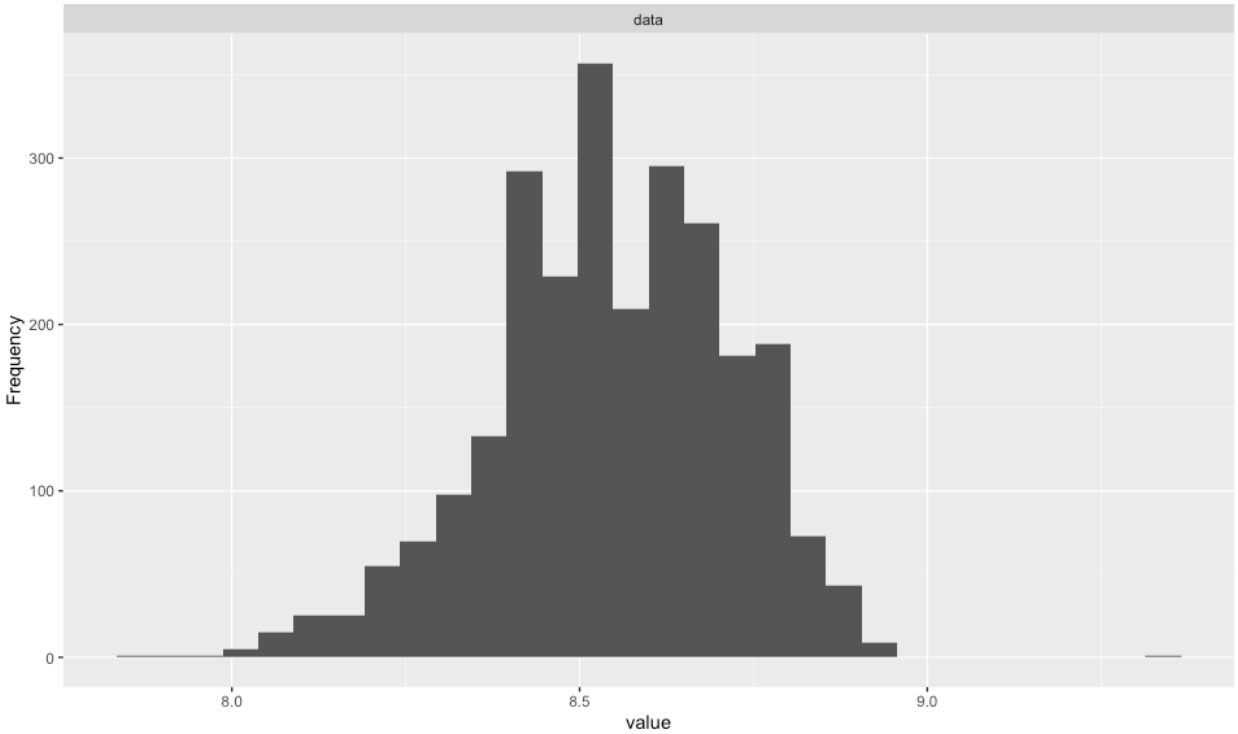
Data Analysis

The dataset contains 33 measures. All are numerical with the exception of Brand Code, which is categorical.

Test Time	Carb Pressure1	Density
Brand Code	Fill Pressure	MFR
Carb Volume	Hyd Pressure1	Balling
Fill Ounces	Hyd Pressure2	Pressure Vacuum
PC Volume	Hyd Pressure3	Oxygen Filler
Carb Pressure	Hyd Pressure4	Bowl Setpoint
Carb Temp	Filler Level	Pressure Setpoint
PSC	Filler Speed	Air Pressure
PSC Fill	Temperature	Alch Rel
PSC CO2	Usage cont	Carb Rel
Mnf Flow	Carb Flow	Balling Lvl

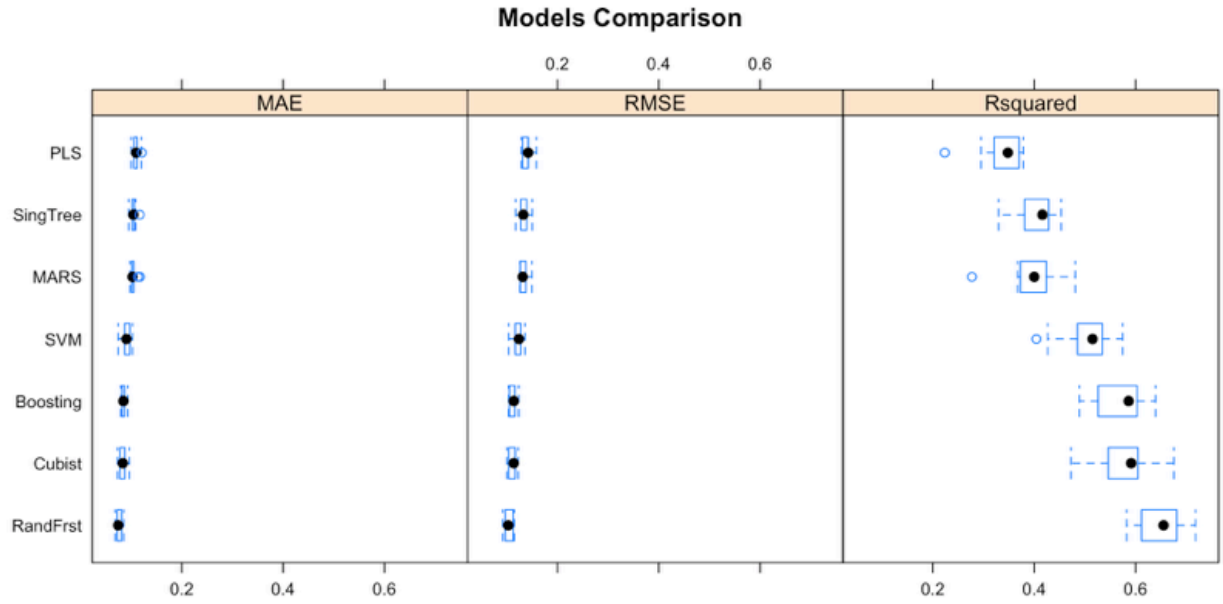
Exploratory data analysis includes finding the missing values, handling outliers, correlation among variables and variables distribution. Below 2 figures shows the missing values % in data and the target variable distribution. Once we handled all these cases, we split the training data into training and validation sets and evaluate their accuracy.


Air Pressurer
Pressure Vacuum
Balling Lvl
Balling
Density
Bowl Setpoint
Carb Flow
Mnf Flow
PH
Usage cont
Alch Rel
Carb Rel
Carb Volume
Hyd Pressure1
Pressure Setpoint
Oxygen Filler
Temperature
Hyd Pressure3
Hyd Pressure2
Filler Level
Fill Pressure
PSC Fill
Carb Temp
Carb Pressure
Hyd Pressure4
Carb Pressure1
PSC
Fill Ounces
PSC CO2
PC Volume
Filler Speed
MFR
0%
0%
0.04%
0.04%
0.04%
0.08%
0.08%
0.08%
0.16%
0.19%
0.35%
0.39%
0.39%
0.43%
0.47%
0.47%
0.54%
0.58%
0.58%
0.78%
0.86%
0.89%
1.01%
1.05%
1.17%
1.24%
1.28%
1.48%
1.52%
1.52%
2.22%
8.25%
Features
0
50
100
150
200
Missing Rows
Band
Good
OK


data
300
200
100
0
Frequency
8.0
8.5
9.0
value

Build Models

The models we have built using the training dataset were Simple Linear Regression, Partial Least Squares, Multivariate Adaptive Regression Splines (MARS), Support Vector Machine (SVM), Single Tree, Boosted Tree, Random Forest and Cubist. Below figure depicts that model's comparison on 3 measures: RSquared, RMSE and MAE.



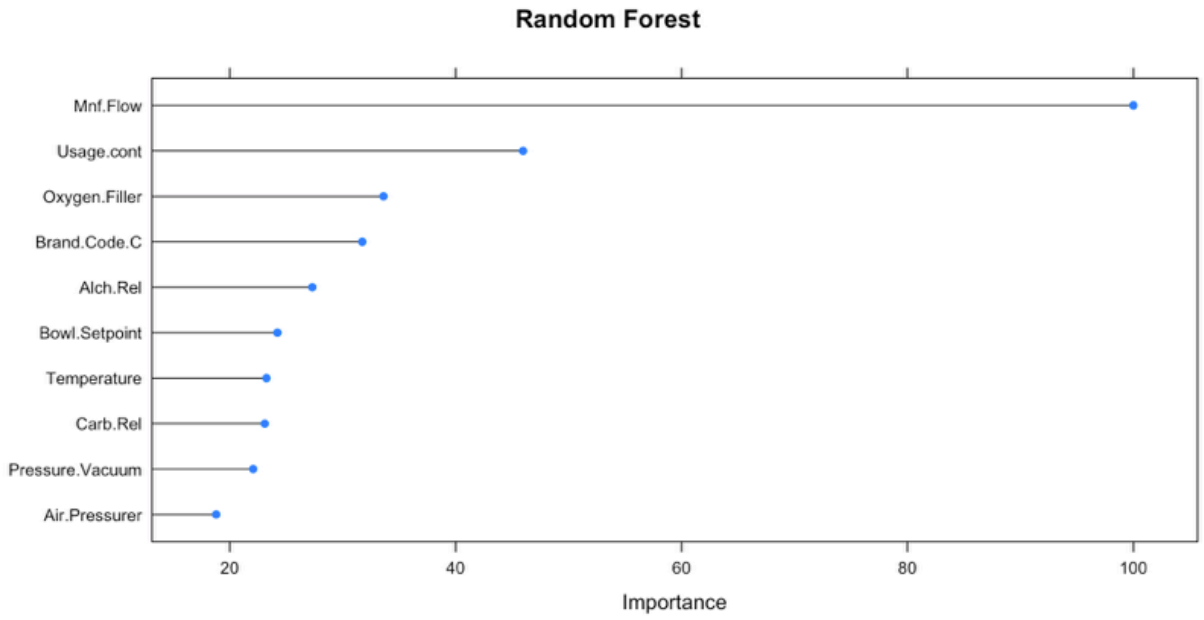
Model Selection

We can see here Random Forest performed the best among all the models tried considering the 3 metrics Rsquared, RMSE and MAE, we identified earlier.

	RMSE	Rsquared	MAE
PLS	0.1371313	0.3722429	0.10861129
MARS	0.1302255	0.4347408	0.10245641
SVM	0.1175582	0.5395039	0.08694041
SingTree	0.1314174	0.4238484	0.10420760
RandFrst	0.1004640	0.6737302	0.07310555
Boosting	0.1113162	0.5897395	0.08321032
Cubist	0.1029136	0.6505194	0.07567148

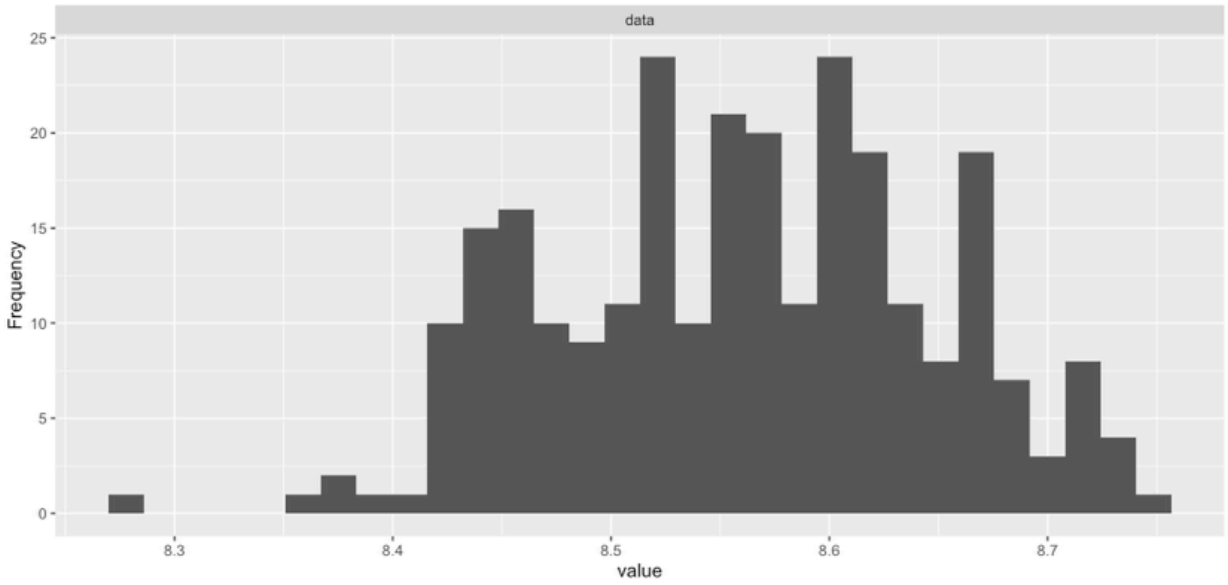
Important factors

Below are the informative variables found by Random Forest models. it is evident Mnf.Flow is the most informative variable for PH response variable.



Predictions

Based on the analysis so far Random Forest model has been selected as the optimal model. Here is the plot of the predicted values of PH on evaluation dataset.



Conclusion

After performing data exploration, analysis, final model selection and prediction we notice that all the values predicted are greater than 8. This value translates that the beverage made is alkaline. At the start of this study, we were not known about the nature of the ABC Beverage company i.e. what type of beverage manufacturer it was. But from this study we can conclude that this company mainly produces alkaline beverages like water, tea, fruit drinks and all.