Data 624 PROJECT 2: Business Report

Joint Work

Alexander Ng, Philip Tanofsky

Due 12/13/2021

# Problem Statement

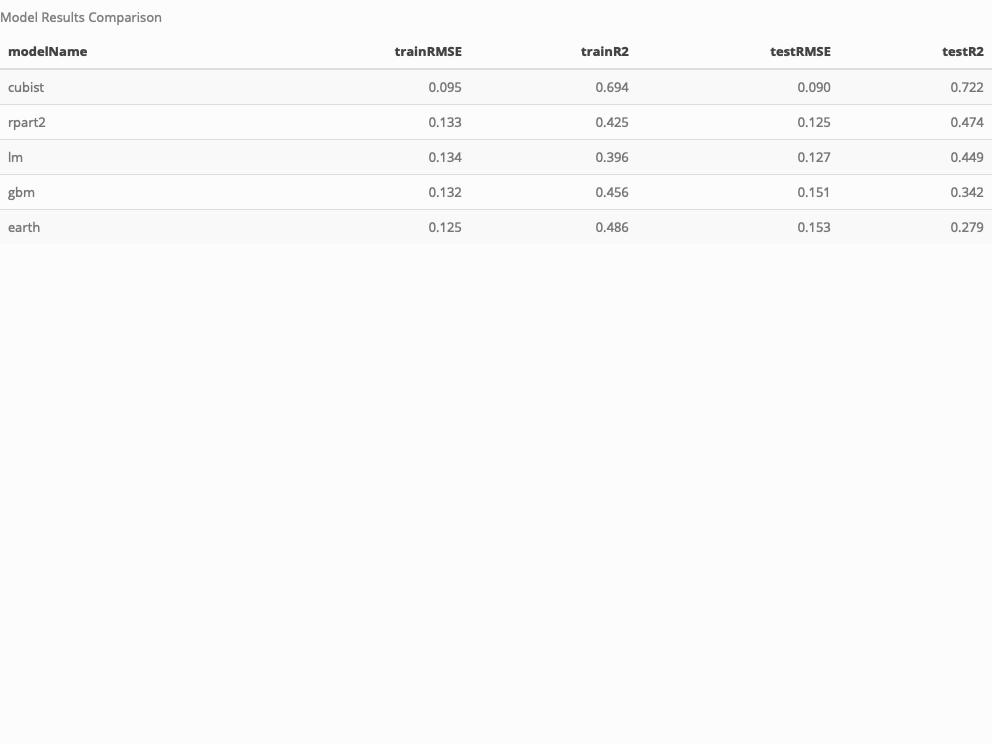
I am in charge of production at ABC Beverage and you are a team of data scientists reporting to me. My leadership has told me that new regulations are requiring us to understand our manufacturing process, the predictive factors and be able to report to them our predictive model of PH.

# Data Wrangling and Imputation

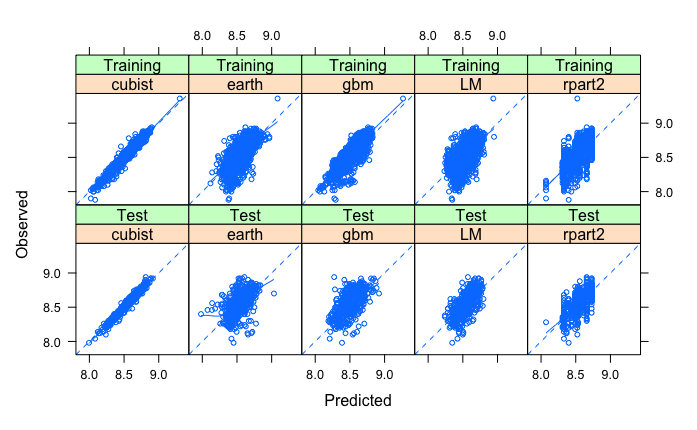
Defense of using all the variables, missing data imputation, replace outliers and zeroes

# Model Building and Results

We generated models based on five approaches, Gradient Boosted Trees, Cubist, CART, MARS, and Linear Regression. TheCubist model clearly outperforms the other four approaches based on RMSE and evaluation metrics. From the data provided the PH value, we trained on 80% and tested on 20% to ensure a confident result to the model. The table below indicates the results on the training and test data for the five models.

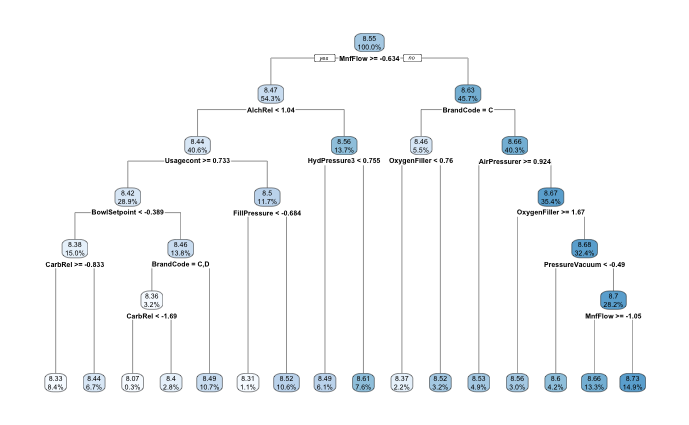


The following plot of the 5 models displays the results of the training and test data for the predictions against the observations and confirms Cubist as the clearly better model.



## Decision Tree for Context (Better Word)

CART tree: Help to comprehend which variables are valuable



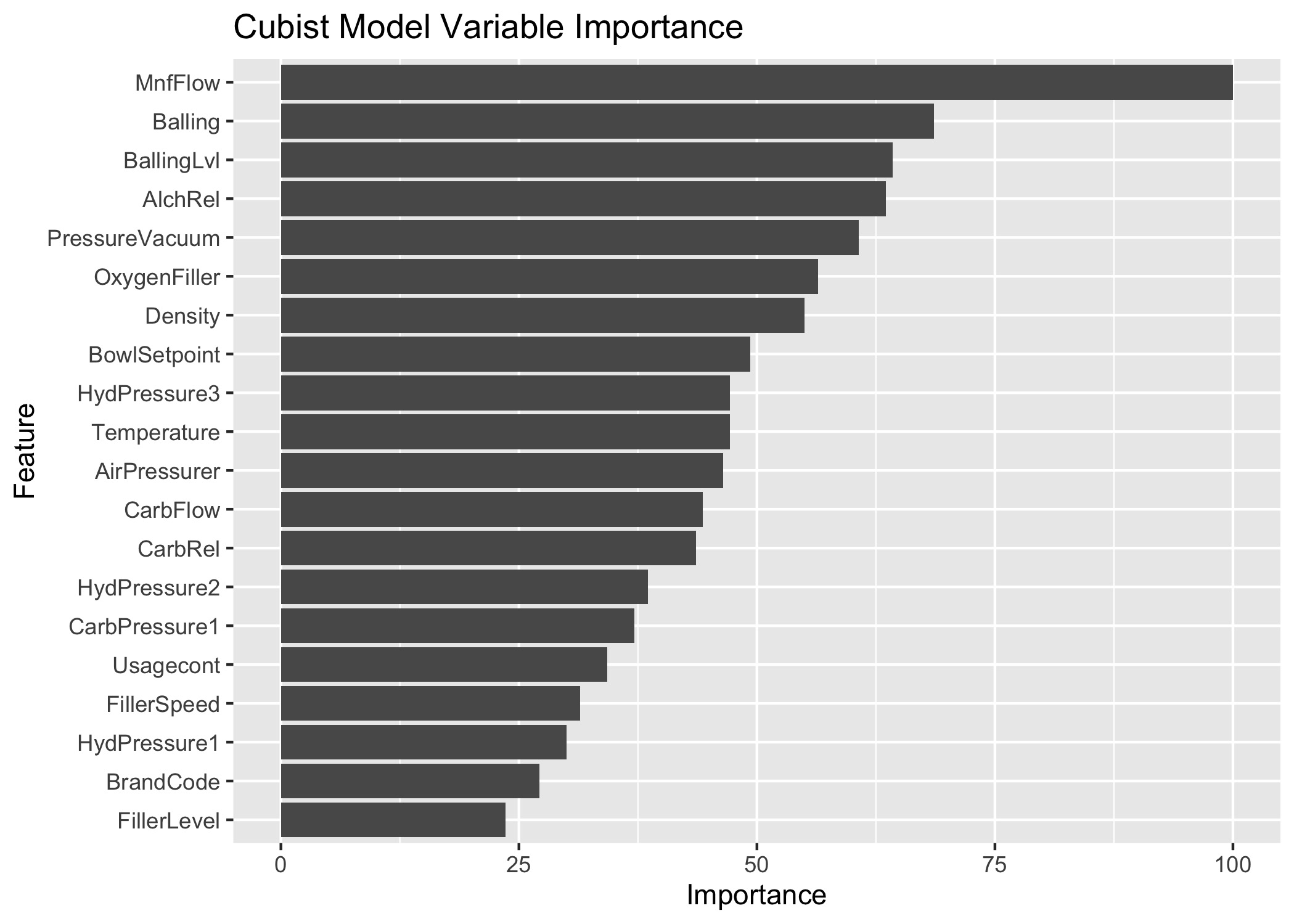
# Model Selection

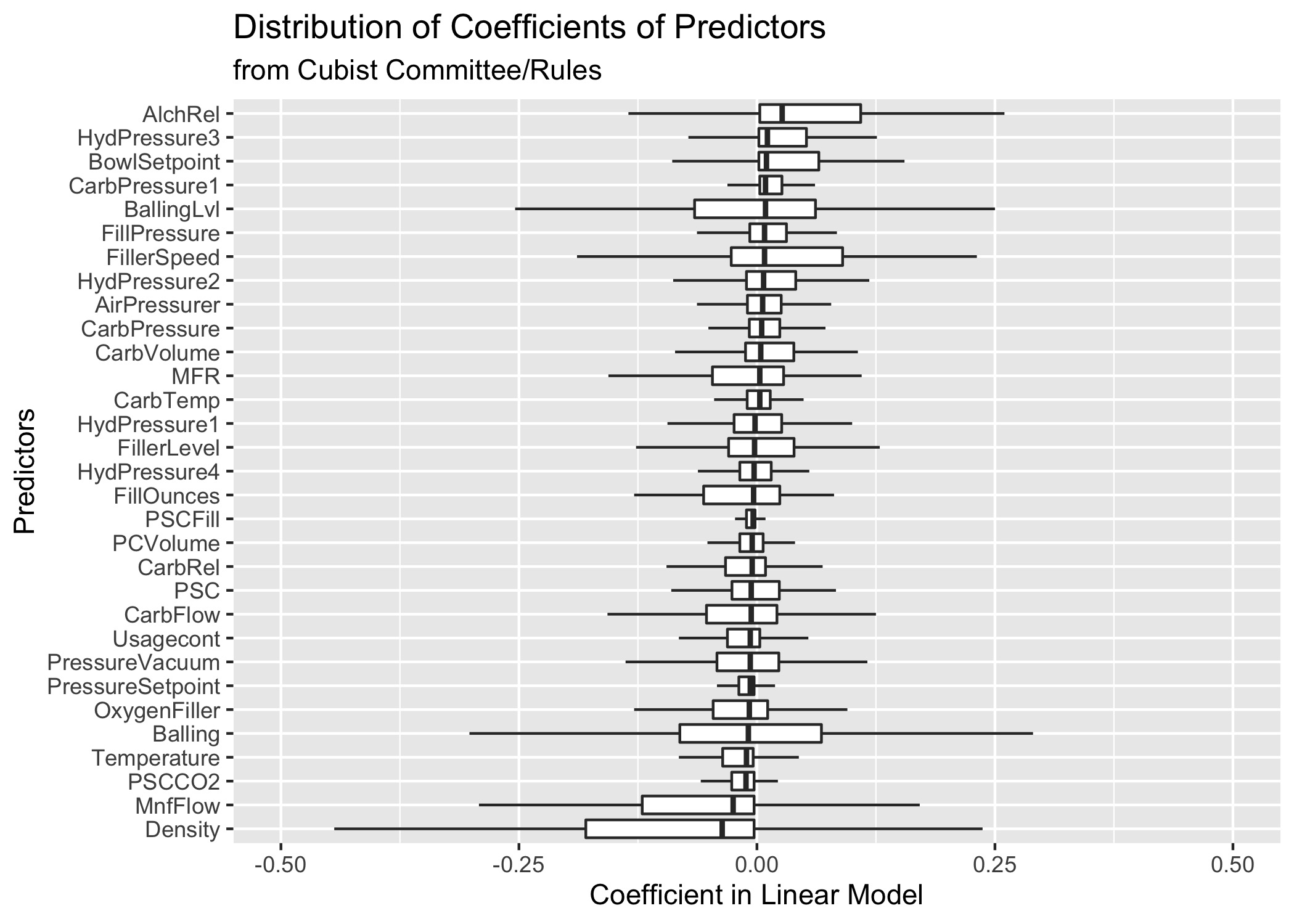
## Cubist

Results of the full train - similar results emerge and thus the model is not overfitting

### Variable Importance

The variable importance of the five models confirmed the overall value of one variable in particular, Mnf Flow. XXXPSTXXX Confirmed variables through linear regression and other model approaches





## Recommendations

Given the Cubist model the top 5 recommendations for understanding the manufacturing process impact on the PH value.

XXXDouble check these against the final, all-in model results XXX

* Mnf Flow: Inverse Correlation
* Balling Lvl: Positive Correlation
* Balling: Slight negative correlation
* Alch Rel: Positive Correlation
* Pressure Vacuum: Slight negative correlation

# Conclusion

The Cubist model produces the best RMSE and results over other advanced model techniques of Gradient Boosted Trees (GBM), CART, MARS, and baseline linear regression models with a focus on the Mnf Flow predictor for biggest impact on the PH value given the current manufacturing process.