Data 624 PROJECT 2: Business Report

Joint Work

Alexander Ng, Philip Tanofsky

Due 12/13/2021

## Problem Statement

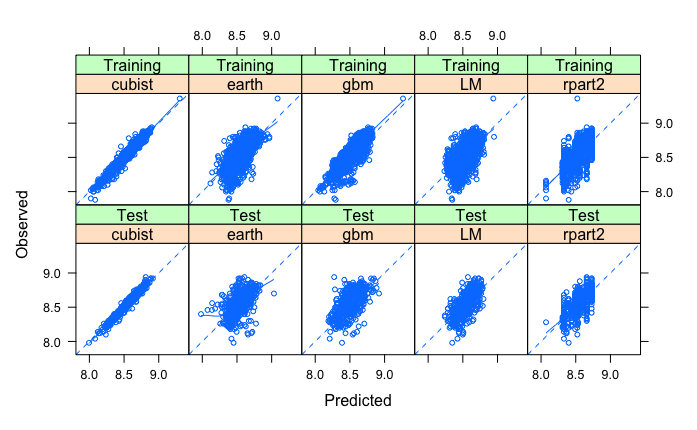
I am in charge of production at ABC Beverage and you are a team of data scientists reporting to me. My leadership has told me that new regulations are requiring us to understand our manufacturing process, the predictive factors and be able to report to them our predictive model of PH.

## Model Selection

Generated models based on four advanced techniques and one simple technique as a baseline. One model approach Cubist, clearly outperformed the other four approaches based on RMSE and evaluation metrics.

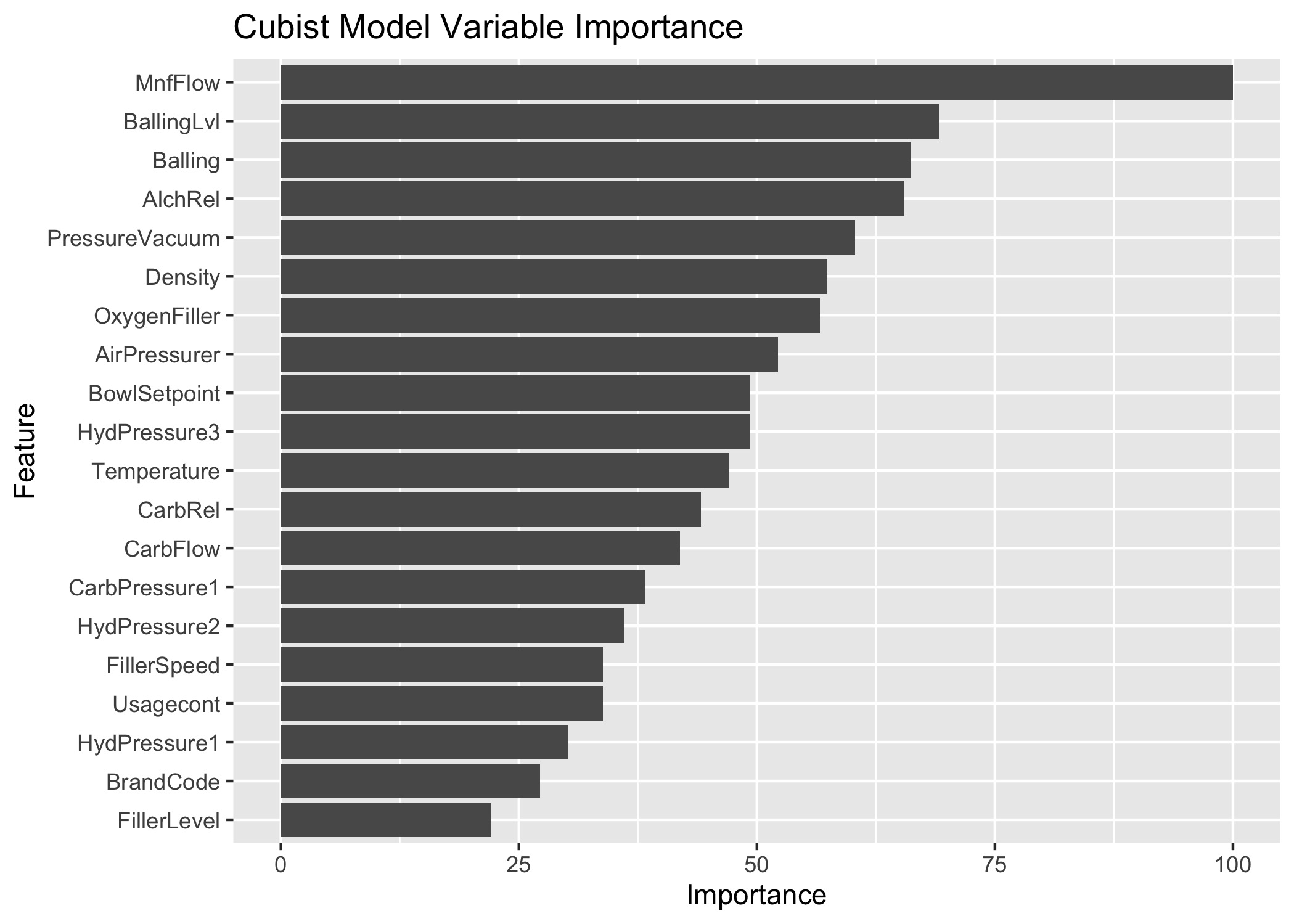


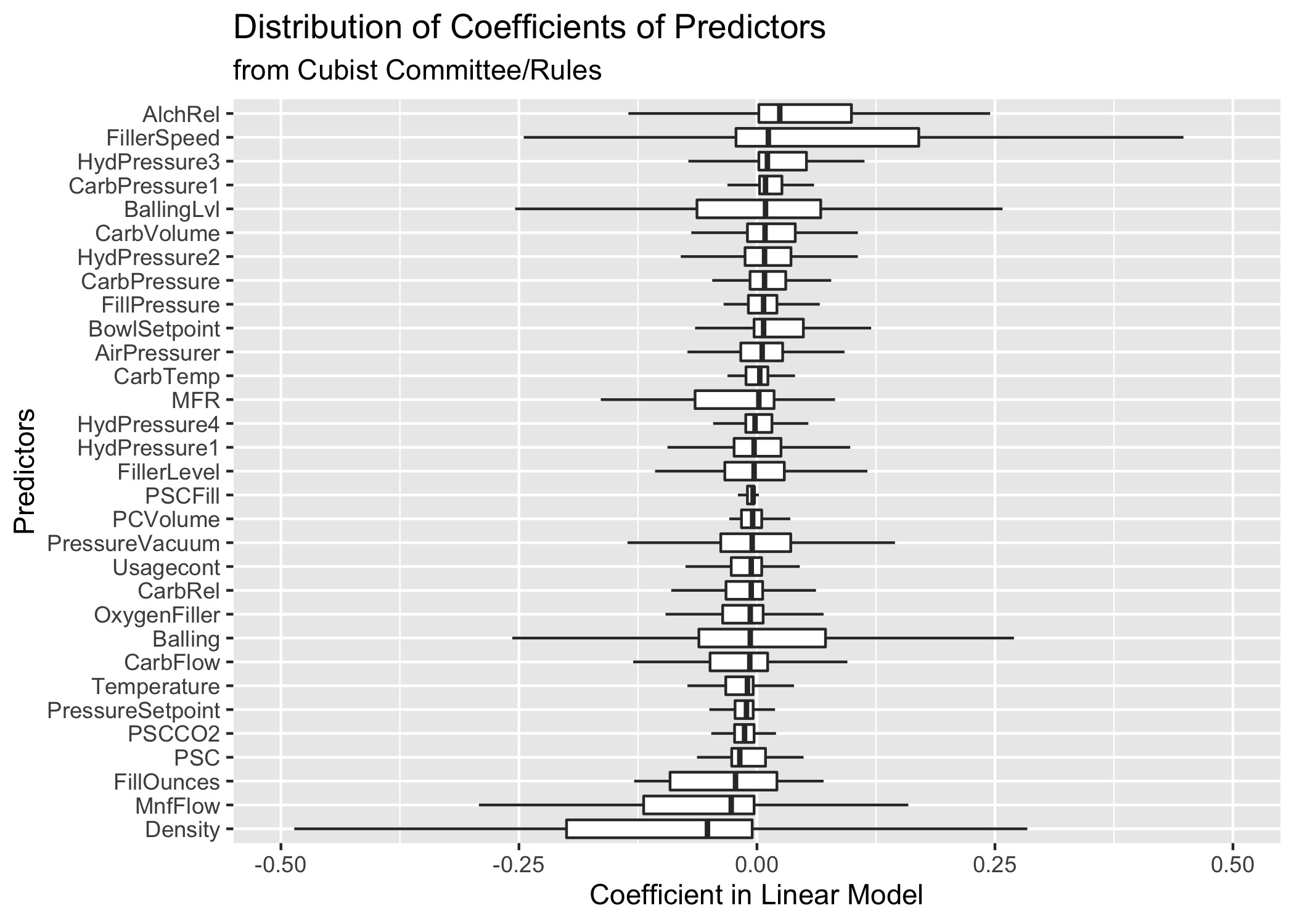
Of the 5 models, here’s a high level view of the results of the training and test data of the predictions against the observations.



## 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





Given the Cubist model the top 5 recommendations are

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

# BLUF

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.

Data pre-processing to address missing values and outliers.

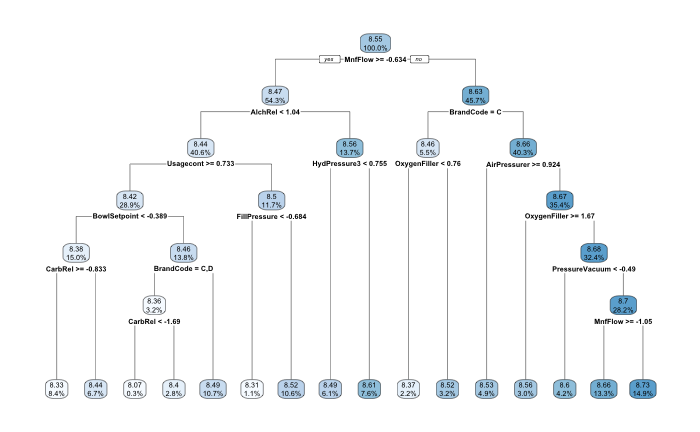
Of the XX (32?) provided predictor variables, variable importance

## Preprocessing info

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

## Decision Tree to provide more context

CART tree: Help to comprehend which variables are valuable



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

Variable Importance and impact on PH - copy from the final section of the larger doc