ABC BEVERAGE

DATA 624 - PROJECT 2

pH

Manufacturing Process and Predictive Factors

JOHN CRUZ
ANDREW BOWEN
JOSH FORSTER

ABC BEVERAGE

Introduction

New regulations are requiring ABC Beverage to understand its manufacturing process better. We will look into the predictive factors that may influence **pH** in this process.

ABC BEVERAGE

Objectives

DATA EXPLORATION 01

- Explore the raw data
- Impute missing information
- Transform skewed distributions

MODELING 02

- Multiple Linear Regression
- Random Forest
- XGBoost
- Neural Net

EVALUATION & RESULTS 03

Compare our models and determine which would be best for predictive capabilities

ABC BEVERAGE

Data Exploration

There are 32 variables in our dataset related to our manufacturing process. We can use these predictors in our models for **pH**

PH

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

ABC BEVERAGE

Data Wrangling

IMPUTATION

Replace missing values using **Predictive Mean Matching**

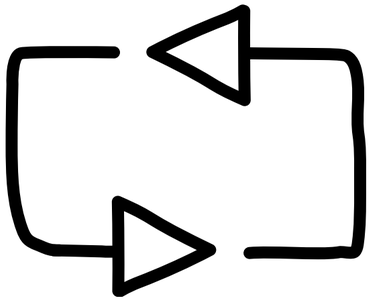
TRANSFORMATION

Transform our predictors to improve their distributions using **Box-Cox** method.

ABC BEVERAGE

Imputation

Having missing values will develop problems in our models. We used ***predictive mean matching*** which essentially draws real values found within the sampled data to replace them.



BEFORE

Brand.Code	Carb.Volume	Fill.Ounces	PC.Volume	Carb.Pressure	Carb.Temp	PSC	PSC.Fill	PSC.CO2	Mnf.Flow	Carb.Pressure1	Fill.Pressure	Hyd.Pressure1	Hyd.Pressure2	Hyd.Pressure3	Hyd.Pressure4
120	10	38	39	27	26	33	23	39	2	32	22	11	15	15	30

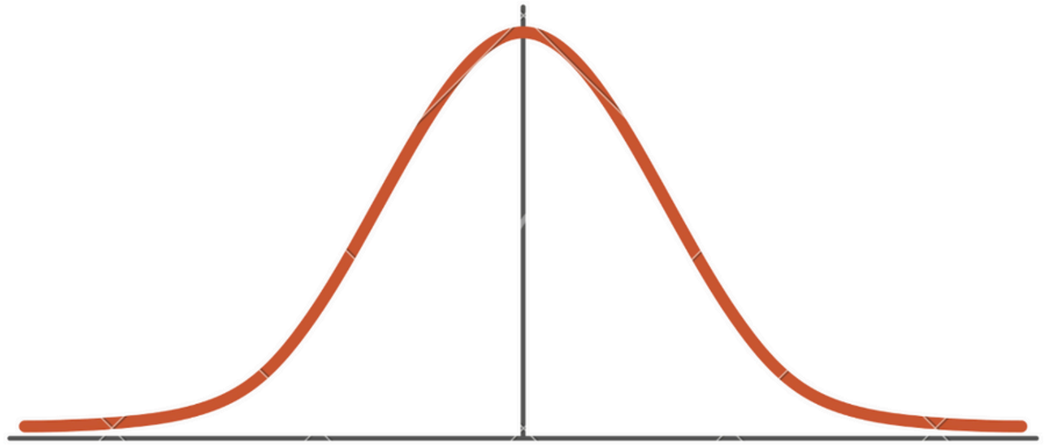
Filler.Level	Filler.Speed	Temperature	Usage.cont	Carb.Flow	Density	MFR	Balling	Pressure.Vacuum	Oxygen.Filler	Bowl.Setpoint	Pressure.Setpoint	Air.Pressurer	Alch.Rel	Carb.Rel	Balling.Lvl
20	57	14	5	2	1	212	1	0	12	2	12	0	9	10	1

AFTER

Brand.Code	Carb.Volume	Fill.Ounces	PC.Volume	Carb.Pressure	Carb.Temp	PSC	PSC.Fill	PSC.CO2	Mnf.Flow	Carb.Pressure1	Fill.Pressure	Hyd.Pressure1	Hyd.Pressure2	Hyd.Pressure3	Hyd.Pressure4
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Filler.Level	Filler.Speed	Temperature	Usage.cont	Carb.Flow	Density	MFR	Balling	Pressure.Vacuum	Oxygen.Filler	Bowl.Setpoint	Pressure.Setpoint	Air.Pressurer	Alch.Rel	Carb.Rel	Balling.Lvl
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

ABC BEVERAGE



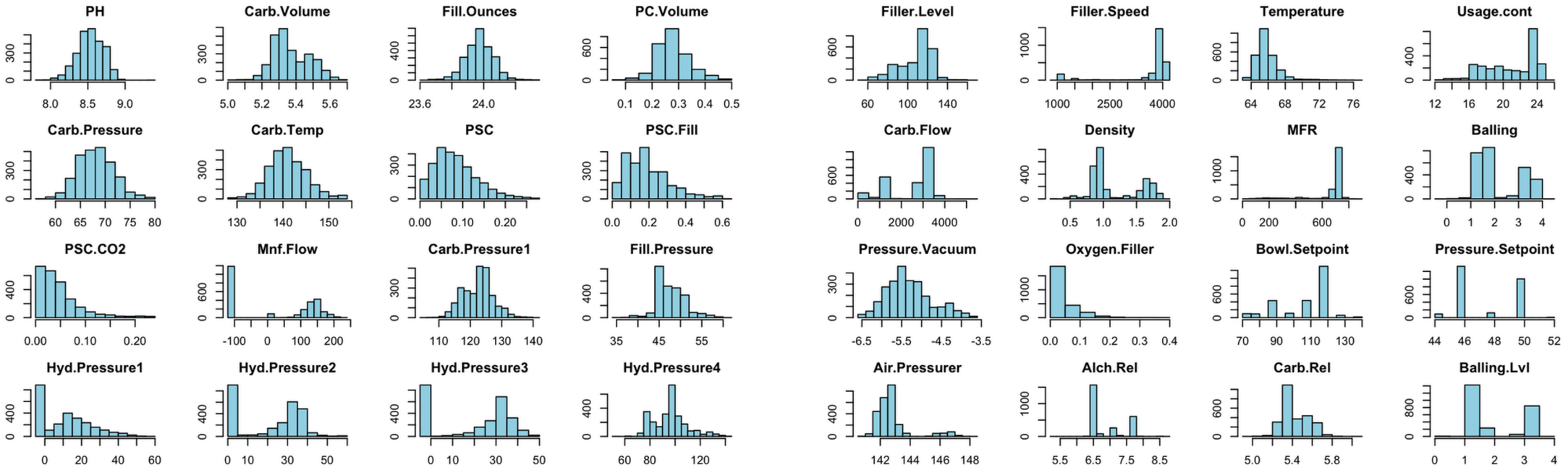
Transformation

We need to transform our variables to improve their distributions using the **Box-Cox** method. What this does is make our values closer to a normal distribution so that they do not violate assumptions we need later on to trust our model outputs.

ABC BEVERAGE

Before Transformation

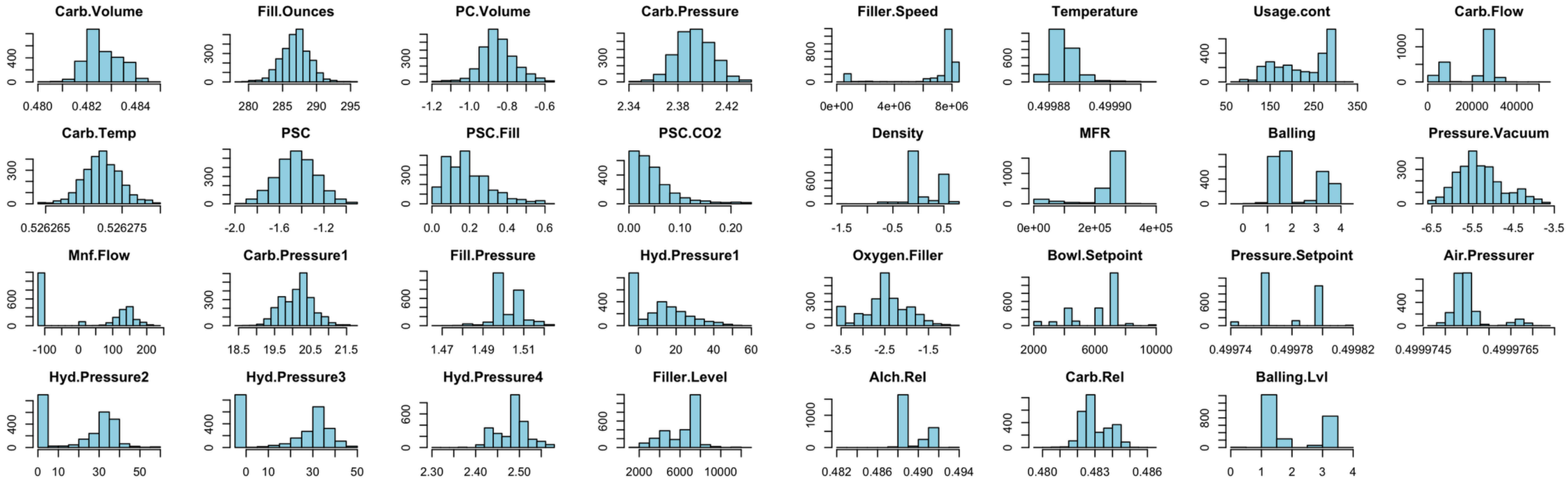
Here we can see variables such as ***PSC*** being right-skewed. We also see ***Hyd.Pressure*** variables are bimodal distributed.



ABC BEVERAGE

After Transformation

Now ***PSC*** is normally distributed, while we still are retaining bimodal distributions on ***Hyd.Pressure*** variables.



ABC BEVERAGE

Modeling

We developed four different types of models for evaluation. The goal is to find which model provides us the best option for predicting **pH**.

MULTIPLE LINEAR REGRESSION

Our benchmark model as it is the simplest way to evaluate **pH** using a “best-fit” line approach.

RANDOM FOREST

Produce a model that combines multiple decision trees to reach a single result, which is determining **pH.**

XGBOOST

Develop a model that iterates and reduces errors by the previous models it created.

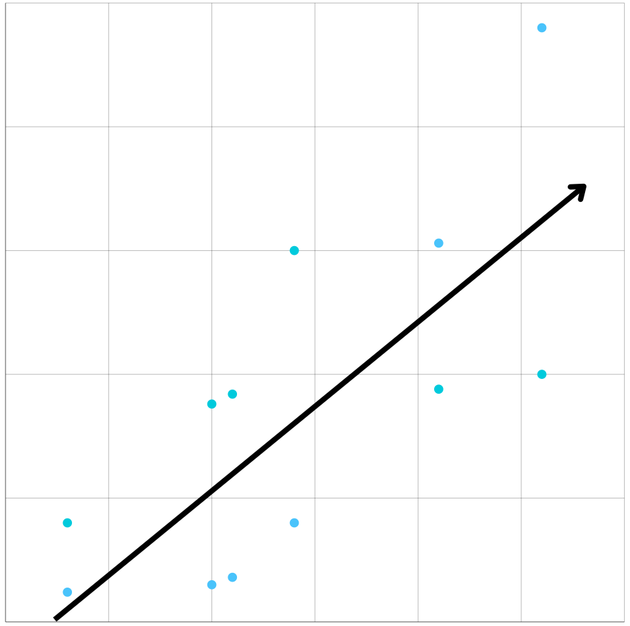
NEURAL NET

Similar to how the human brain works, we will develop a model that can look into more complex relationships than our previous models can.

ABC BEVERAGE

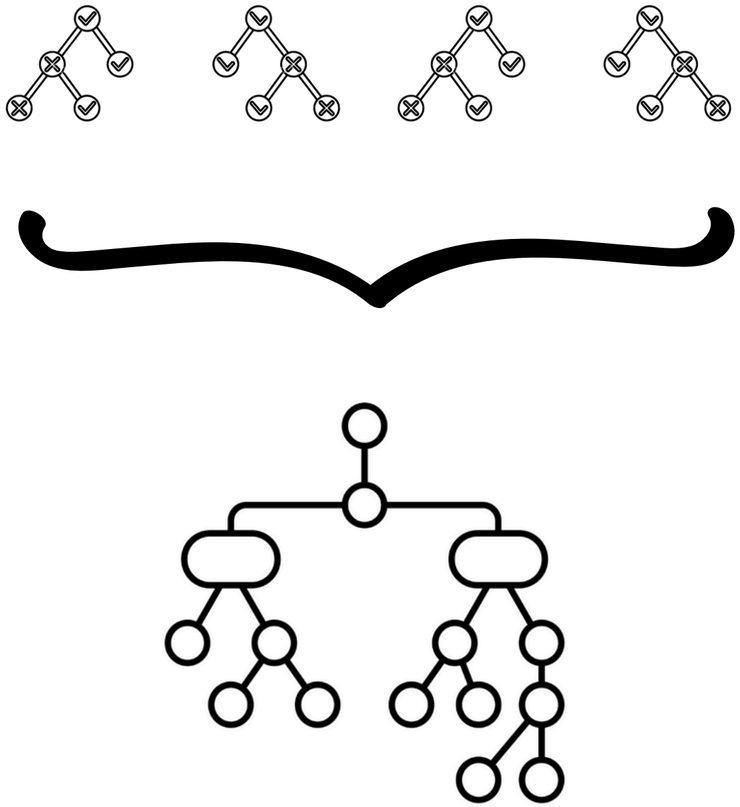
Multiple Linear Regression

The goal of this model is to estimate the relationship between **pH** and our predictors. This is approached by reducing the errors and finding a “best-fit” straight line between this relationship.



ABC BEVERAGE

Random Forest

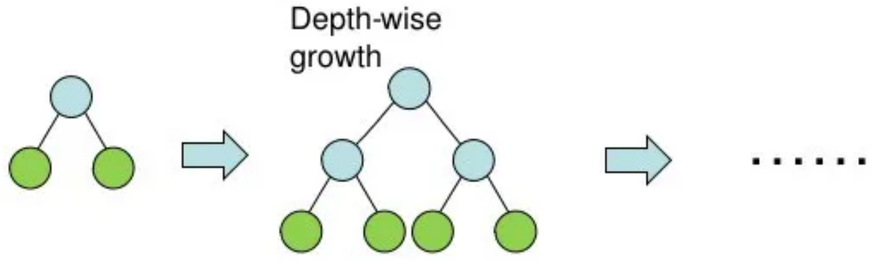


In this model it combines several decision trees, where our predictions for **pH** are more accurate than a simple tree. Essentially, it averages out these different trees into one larger tree for a better understanding of our predictors to **pH**

ABC BEVERAGE

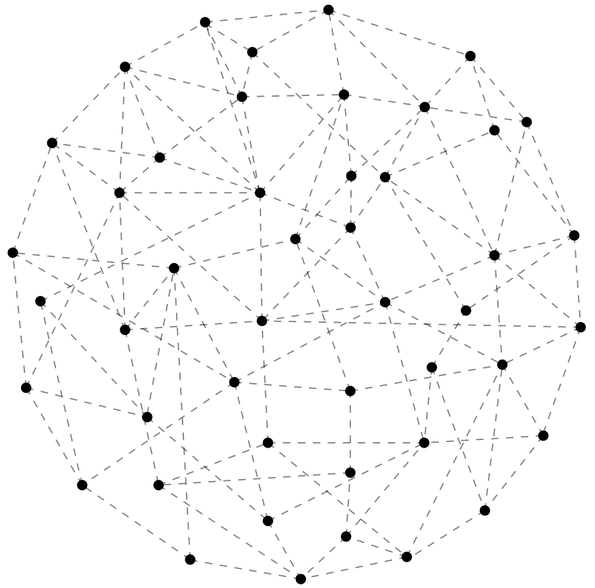
XGBoost

This model uses a gradient descent boosting method that works by iteratively adding models to an ensemble, where each subsequent model is trained to correct the errors made by the previous models created.



Neural Net

With this type of model we can develop more complex relationships that looks into how **pH** connects to our predictors. Unlike our baseline model, it is not a straight forward comparison from predictors to **pH**, but can develop more layers and patterns linear regression cannot. This model closely resembles how the human brain works and connects information that is gathered.



Evaluation

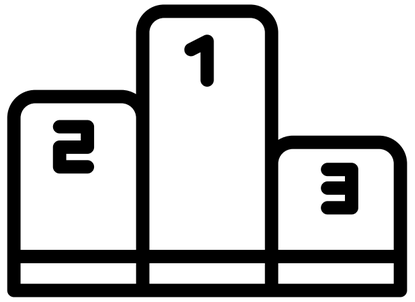
- The model with the best prediction performance is our **Random Forest** model.
- With an R^2 =0.6698, it tells us that **66.98%** of the variation of **pH** can be explained by our predictors.
- It also has the lowest RMSE = 0.1020 which measures the average difference between our model's predicted values and the actual values we have.

Model Performance

Model	RMSE	Rsquared	MAE
Random Forest	**0.1020**	**0.6698**	**0.0734**
XGBoost	0.1229	0.4976	0.0947
Neural Net	7.5522	NA	7.5503
Linear Reg	345758.9319	0.0168	345518.4869

ABC BEVERAGE

Variables Importance



Neural Net Variable Importance

	Overall	Rank
Filler.Speed	**52.8166**	**1**
MFR	5.7060	2
Carb.Temp	5.6210	3
Air.Pressurer	5.3660	4
Carb.Pressure1	4.9543	5
Filler.Level	4.1279	6
Bowl.Setpoint	4.1193	7
Hyd.Pressure4	3.4544	8
Carb.Pressure	2.8613	9
Temperature	2.5664	10

Random Forest Variable Importance

	Overall	Rank
Mnf.Flow	**6.3812**	**1**
Brand.Code	4.2171	2
Usage.cont	3.8414	3
Filler.Level	2.6751	4
Oxygen.Filler	2.2854	5
Temperature	2.1639	6
Carb.Rel	1.9371	7
Pressure.Vacuum	1.9206	8
Bowl.Setpoint	1.9115	9
Balling.Lvl	1.8785	10

XGBoost Variable Importance

	Overall	Rank
Mnf.Flow	**0.2004**	**1**
Usage.cont	0.1071	2
Bowl.Setpoint	0.0792	3
Brand.CodeC	0.0739	4
Carb.Pressure1	0.0635	5
Temperature	0.0577	6
Alch.Rel	0.0493	7
Oxygen.Filler	0.0432	8
Pressure.Setpoint	0.0393	9
Balling	0.0371	10

Linear Regression Variable Importance

	Overall	Rank
Mnf.Flow	**10.2168**	**1**
Carb.Pressure1	8.3684	2
Temperature	7.9653	3
Bowl.Setpoint	6.1482	4
Pressure.Setpoint	4.7046	5
Usage.cont	4.5990	6
Balling	3.6684	7
Carb.Flow	3.4208	8
Hyd.Pressure2	3.3508	9
Fill.Pressure	3.2368	10

In 3 out of our 4 models, within the top 10 predictors, **Mnf.Flow** is a key important variable in our models. Interestingly enough, our Neural Net model shows **filler.speed** as the most important variable, however, it does not fall into any of our other models top 10 at all.

Results

- Given the model results, using a Random Forest model to predict **pH** would be our best option.

- We can have a heavy focus within our manufacturing process on **Mnf.Flow** as it is the most important predictor across a few models.