PROJECT 2: PREDICTING PH

DATA 624 - Predictive Analytics

Group 2

Group Members:

Juliann McEachern

10 December 2019

Contents

Introduction

This project is designed to evaluate production data from a beverage manufacturing company. Our assignment is to predict PH, a Key Performance Indicator (KPI), with a high degree of accuracy through predictive modeling. After thorough examination, we approached this task by splitting the provided data into training and test sets. We evaluated several models on this split and found that **what-ever-worked-best** method yielded the best results.

Each group member worked individually to create their own solution. We built our final submission by collaboratively evaluating and combining each others' approaches. Our introduction should further outline individual responsibilities. For example, **so-and-so** was responsible for **xyz task**.

For replication and grading purposes, we made our code avaliable in the appendix section. This code, along with the provided data, score-set results, and individual contributions, can also be accessed through our group github repository:

- Pretend I'm a working link to R Source Code
- Pretend I'm a working link to Provided Data
- Pretend I'm a working link to Excel Results
- Pretend I'm a working link to Individual Work

1 Data Exploration

The beverage manufacturing production dataset contained 33 columns/variables and 2,571 rows/cases. In our initial review, we found that the response variable, PH, had four missing observations.

We also identified that 94% of the predictor variables had missing data points. Despite this high occurance, the NA values in the majority of these predictors accounted for less than 1% of the total observations. Only eleven variables were missing more than 1% of data.

Table 1.1: Variables with Highest Frequency of NA Values

	MFR	BrandCode	FillerSpeed	PCVolume	PSCCO2	FillOunces	PSC	CarbPressure1	HydPressure4	CarbPressure	CarbTemp
n	212.0	120.0	57.0	39.0	39.0	38.0	33.0	32.0	30.0	27.0	26
%	8.2	4.7	2.2	1.5	1.5	1.5	1.3	1.2	1.2	1.1	1

Response Variable

Understanding the influence pH has on our predictors is key to building an accurate predictive model. pH is a measure of acidity/alkalinity that must conform in a critical range. The value of pH ranges from 0 to 14, where 0 is acidic, 7 is neutral, and 14 is basic.

Fig. 1.1: Distribution of Response Variable: pH

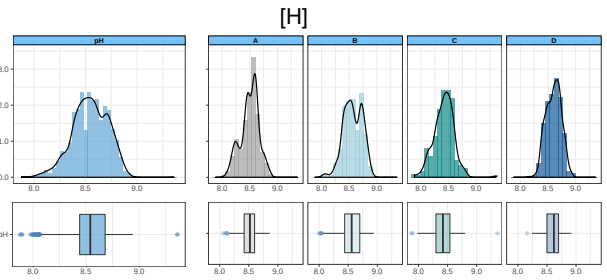


Figure 1.1 shows that our response distribution follows a somewhat normal pattern and is centered around 8.5. The histogram for pH is bimodal in the aggregate, but varies by brand. The boxplot view allows us to better visualize the effect outliers have on the skewness within our target variable.

Brand A has a negatively skewed, multimodal distribution, which could be suggestive of several distinct underlying response patterns or a higher degree of variation in pH response for

Fig. 1.2: Box-Plot Distribution of Numeric Predictor Variables

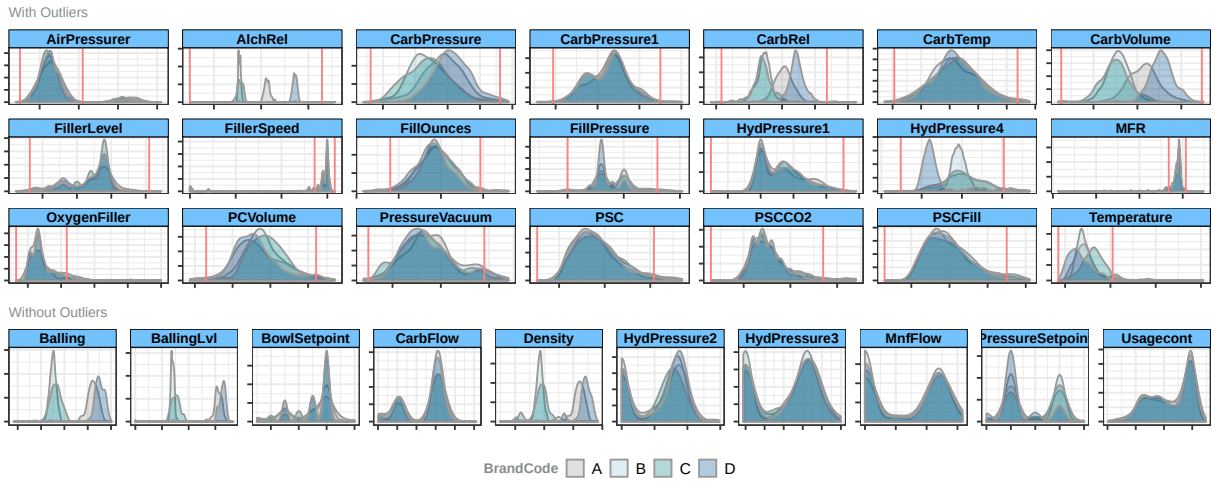
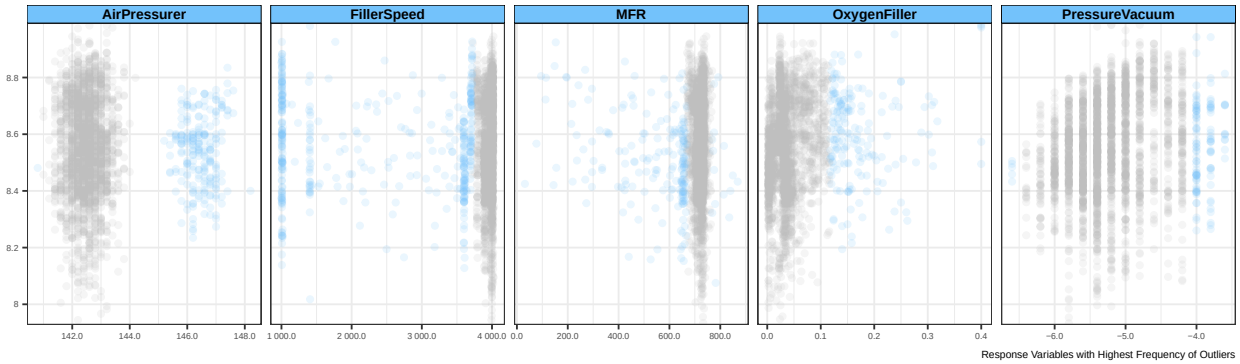


Fig. 1.3: Response~Predictor Scatterplots



this brand. The density plot and histogram for Brand B show two bimodal peaks with a slight positive skew. These peaks indicate that this brand has two distinct response values that occur more frequently. The distribution for Brand C and D are both more normal, with a slight negative skew. Brand D has the highest median pH value and Brand C has the lowest. Brand C also appears to have the largest spread of pH values.

Predictor Variables

Many of our predictors also contain outliers and have a skewed distribution. The boxplots below help us visualize this spread of our numeric predictor variables.

We examined the predictor variables with outliers in a scatterplot against our target, pH to better understand predictor and response relationship. The outliers, highlighted in blue, further show which predictors have a heavy-tail distribution. We can also identify many variables with strong outlier patterns, suggesting a high degree of variability within certain measurements.

For example, `AirPressurer` shows a very distinct, bifurcated pattern. This variable has a clear split between normal and extreme values. `MFR` also shows an interesting pattern. The outliers have a weak, negative linear relationship with pH, but the non-outliers have no linear relationship and follow a straight, vertical line.

2 Data Preparation

Decision models trees are robust against the affect of correlated variables, outliers, and missing values. We applied different tranformation to properly evaluate our three model types.

Data was divided using an 80/20 split to create a train and test set. All models incorporated k-folds cross-validation set at 10 folds to protect against overfitting the data.

Data Imputation

We choose to drop the complete cases of all `pH` observations with null data in the target as they accounted for such a small proportion (< 0.002%) of the observations. Doing such increased our non-linear modeling accuracy measures. For our predictor variables, we applied a Multiple Imputation by Chained Equations (MICE) algorthim to predict the missing data using sequential regression. This method filled in all incomplete cases, including `BrandCode`, our one unordered categorical variable.

We can also use this same approach to handle outliers (linear model) by setting their value to NA and predicticting a value within the expected range.

Pre-Processing

Correlation

We examined the relationship between our numeric predictors and found that 9 of the variables appear heavily related, with correlation values exceeding ± 0.75. The full correlation matrix can be viewed in the appendix section. *Revisit section to add more text*

Table 2.1: Highly Correlated Predictors

V1	V2	COR	V1	V2	COR
AlchRel	BallingLvl	0.93	CarbVolume	BallingLvl	0.78
AlchRel	CarbRel	0.84	CarbVolume	Density	0.76
Balling	BallingLvl	0.98	Density	Balling	0.96
Balling	AlchRel	0.92	Density	BallingLvl	0.95
Balling	CarbRel	0.82	Density	AlchRel	0.90
CarbPressure	CarbTemp	0.81	Density	CarbRel	0.82
CarbRel	BallingLvl	0.84	FillerLevel	BowlSetpoint	0.95
CarbVolume	CarbRel	0.79	FillerSpeed	MFR	0.93
CarbVolume	Balling	0.78	HydPressure2	HydPressure3	0.92
CarbVolume	AlchRel	0.78	MnfFlow	HydPressure3	0.76

Test the effect of pre-processing methods to maximize the success of our tree and non-tree models. Not currently adding data transformations but may revist: ie. scale data for PLS.

For linear models, we removed the predictor `HydPressure1` as it contained near-zero variance. `HydPressure3`, `Balling`, `BallingLvl`, `FillerSpeed`, `FillerLevel`, and `Density` were also removed due to large absolute correlations with other variables.

3 Predictive Modeling

We modeled the data using linear, non-linear, and tree-based regression. In this section, we discuss the attempts from our best model performance for each regression type.

Linear Regression

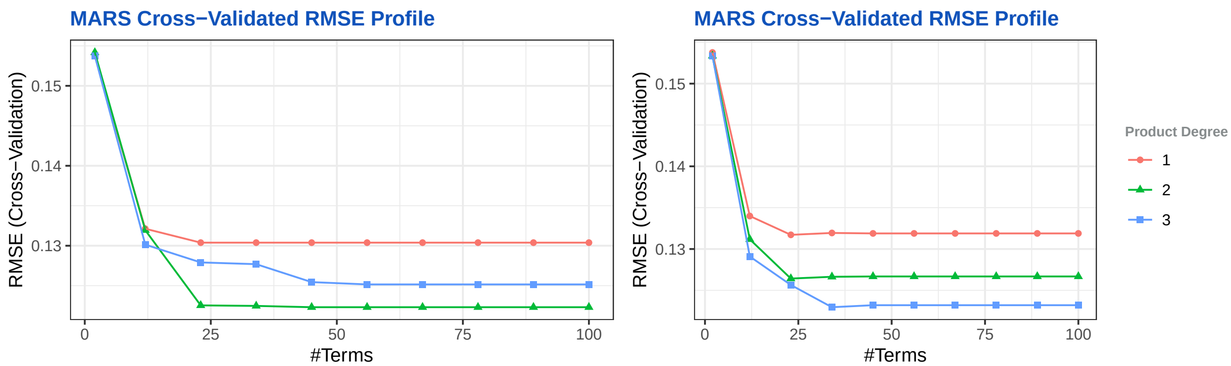
multicollineraty

Explain text here. Text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text.

Non-Linear Regression

Explain text here. Text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text.

Tree-Based Regression



MARS CV RMSE:

Train Accuracy:

	MARS1_Train	MARS2_Train	MARS1_Test	MARS2_Test
RMSE	0.1223	0.1230	0.1179	0.1269
Rsquared	0.5023	0.5034	0.5214	0.4577
MAE	0.0938	0.0934	0.0882	0.0897
MAPE	0.0105	0.0102	0.0103	0.0105

Variable Importance:

Train

Train text.

Table 3.1: Comparison of MARS Models Var Importance

Test

Test text.

[H]

Variable	MARS2	MARS1	diff
MnfFlow	100.00	100.00	0.00
BrandCodeC	78.60	77.38	-1.21
PCVolume	78.60	16.54	-62.06
HydPressure2	78.60	61.28	-17.31
AirPressurer	69.77	66.43	-3.34
Temperature	65.27	61.28	-3.99
MFR	65.27	0.00	-65.27
BrandCodeD	59.59	0.00	-59.59
CarbPressure1	54.94	39.12	-15.82
OxygenFiller	54.94	22.29	-32.66
Usagecont	51.02	33.70	-17.32
BowlSetpoint	47.93	42.65	-5.28
PressureVacuum	44.96	49.93	4.98
HydPressure3	39.72	39.12	-0.60
CarbFlow	36.19	0.00	-36.19
BrandCodeB	0.00	0.00	0.00
CarbVolume	0.00	0.00	0.00
FillOunces	0.00	0.00	0.00
CarbPressure	0.00	0.00	0.00
CarbTemp	0.00	0.00	0.00
PSC	0.00	0.00	0.00
PSCFill	0.00	0.00	0.00
PSCCO2	0.00	0.00	0.00
FillPressure	0.00	0.00	0.00
HydPressure4	0.00	0.00	0.00
PressureSetpoint	0.00	0.00	0.00
CarbRel	0.00	0.00	0.00
BallingLvl	0.00	0.00	0.00

4 Discussion

Eval text. The end.

5 Conclusion

fasdfs

Appendix

Summary Statistics

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
BrandCode*	1	2451	2.5	1.0	2.0	2.5	0.0	1.0	4.0	3.0	0.4	-1.1	0.0
CarbVolume	2	2561	5.4	0.1	5.3	5.4	0.1	5.0	5.7	0.7	0.4	-0.5	0.0
FillOunces	3	2533	24.0	0.1	24.0	24.0	0.1	23.6	24.3	0.7	0.0	0.9	0.0
PCVolume	4	2532	0.3	0.1	0.3	0.3	0.1	0.1	0.5	0.4	0.3	0.7	0.0
CarbPressure	5	2544	68.2	3.5	68.2	68.1	3.6	57.0	79.4	22.4	0.2	0.0	0.1
CarbTemp	6	2545	141.1	4.0	140.8	141.0	3.9	128.6	154.0	25.4	0.2	0.2	0.1
PSC	7	2538	0.1	0.0	0.1	0.1	0.0	0.0	0.3	0.3	0.8	0.6	0.0
PSCFill	8	2548	0.2	0.1	0.2	0.2	0.1	0.0	0.6	0.6	0.9	0.8	0.0
PSCCO2	9	2532	0.1	0.0	0.0	0.0	0.0	0.0	0.2	0.2	1.7	3.7	0.0
MnfFlow	10	2569	24.6	119.5	65.2	21.1	169.0	-100.2	229.4	329.6	0.0	-1.9	2.4
CarbPressure1	11	2539	122.6	4.7	123.2	122.5	4.4	105.6	140.2	34.6	0.1	0.1	0.1
FillPressure	12	2549	47.9	3.2	46.4	47.7	2.4	34.6	60.4	25.8	0.5	1.4	0.1
HydPressure1	13	2560	12.4	12.4	11.4	10.8	16.9	-0.8	58.0	58.8	0.8	-0.1	0.2
HydPressure2	14	2556	21.0	16.4	28.6	21.1	13.3	0.0	59.4	59.4	-0.3	-1.6	0.3
HydPressure3	15	2556	20.5	16.0	27.6	20.5	13.9	-1.2	50.0	51.2	-0.3	-1.6	0.3
HydPressure4	16	2541	96.3	13.1	96.0	95.5	11.9	52.0	142.0	90.0	0.5	0.6	0.3
FillerLevel	17	2551	109.3	15.7	118.4	111.0	9.2	55.8	161.2	105.4	-0.8	0.0	0.3
FillerSpeed	18	2514	3687.2	770.8	3982.0	3920.0	47.4	998.0	4030.0	3032.0	-2.9	6.7	15.4
Temperature	19	2557	66.0	1.4	65.6	65.8	0.9	63.6	76.2	12.6	2.4	10.2	0.0
Usagecont	20	2566	21.0	3.0	21.8	21.3	3.2	12.1	25.9	13.8	-0.5	-1.0	0.1
CarbFlow	21	2569	2468.4	1073.7	3028.0	2601.1	326.2	26.0	5104.0	5078.0	-1.0	-0.6	21.2
Density	22	2570	1.2	0.4	1.0	1.2	0.1	0.2	1.9	1.7	0.5	-1.2	0.0
MFR	23	2359	704.0	73.9	724.0	718.2	15.4	31.4	868.6	837.2	-5.1	30.5	1.5
Balling	24	2570	2.2	0.9	1.6	2.1	0.4	-0.2	4.0	4.2	0.6	-1.4	0.0
PressureVacuum	25	2571	-5.2	0.6	-5.4	-5.3	0.6	-6.6	-3.6	3.0	0.5	0.0	0.0
PH	26	2567	8.5	0.2	8.5	8.6	0.2	7.9	9.4	1.5	-0.3	0.1	0.0
OxygenFiller	27	2559	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.4	2.7	11.1	0.0
BowlSetpoint	28	2569	109.3	15.3	120.0	111.3	0.0	70.0	140.0	70.0	-1.0	-0.1	0.3
PressureSetpoint	29	2559	47.6	2.0	46.0	47.6	0.0	44.0	52.0	8.0	0.2	-1.6	0.0
AirPressurer	30	2571	142.8	1.2	142.6	142.6	0.6	140.8	148.2	7.4	2.3	4.7	0.0
AlchRel	31	2562	6.9	0.5	6.6	6.8	0.1	5.3	8.6	3.3	0.9	-0.9	0.0
CarbRel	32	2561	5.4	0.1	5.4	5.4	0.1	5.0	6.1	1.1	0.5	-0.3	0.0
BallingLvl	33	2570	2.1	0.9	1.5	2.0	0.2	0.0	3.7	3.7	0.6	-1.5	0.0

Correlation Matrix