DATA 624: Project 2

DATA 624 - Predictive Analytics

Group 2

Group Members:

Vinicio Haro

Sang Yoon (Andy) Hwang

Julian McEachern

Jeremy O'Brien

Bethany Poulin

10 December 2019

Contents

File for final submission of Project 2.

1 Will Update Sunday Morning

[CONTENT EDITORS: ADD IN SUMMARY OF CONCLUSION AND ANY SUPPORTING INSIGHT FROM LAST SECTION]

2 Introduction

pH is a central component to the manufacturing of a commercial beverage as it is an indicator of both the process health and the ultimate flavor appeal of the final product. In fact pH plays a role in multiple facets of the a drinks appeal. The flavor, mouthfeel and the aesthetic experience of a given product is distinctly tied to the pH relative to other beverage qualities that brands use distinguish themselves from other liquid refreshments.

Because it is to central to the design of a product's drinking experience, pH is a key performance indicator in the beverage manufacturing process and is tested for and tracked diligently, as the final pH is dependent on and vulnerable to even slight changes in production methods.

Having monitored and recorded these production variables, as well as the final pH, we have the opportunity to improve production outcomes by more closely controlling pH in our beverages with predictive modeling with the potential to catch and correct variations in process which negative impact our taget pH.

Each group member worked individually to experiment with preprocessing while exploring a distinct set of model methods. Upon review, our team created a singular preprocessing protocol and data set, based on our most successful methods and evaluated our most performant models built over these data.

Links to Work Product [lame title - do we keep this section?]

- 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

3 Data Exploration

The beverage dataset includes 2,571 cases, 32 predictor variables, and a single response variable. One of these predictor variables (Brand Code) is categorical with four levels - A through D; for the purpose of our analysis we interpreted these to represent four distinct beverage brands.

While we found missing observations in both response and predictor variables, in our assessment the extent of NAs did not suggest a systemic issue in measurement or recording that imputing values could not remedy. For context: - The response variable (PH) is missing a total of four observations (< 1%). - Most (30) predictor variables are missing at least one observation, but only eleven are missing more than 1% of total cases and only three are missing more than 2% of total cases. These are: 1. MFR (continuous, 8.2%) 2. BrandCode (categorical, 4.7%) 3. and FillerSpeed (continuous, 2.2%)

[CONTENT EDITORS: DO WE STILL WANT TO CREATE MISSING DATA TABLE? IF SO, MISSINGDATA OBJECT NEEDS TO BE REBUILT IN MODEL_PREP.R]

Response Variable

[Density Plots]

Our target variable pH, is a continuous variable. pH is the inverse logarithmic scaled measure of hydrogen ions in solutions and reflects how acidic or basic a water-based solution is. Centered around a neutral value of 7, pH ranges from highly acidic 0 and to highly alkaline at 14.

In total, the pH variable is approximately normally distributed, centered around 8.546 (i.e. slightly base), with some negative skew / outliers. When evaluated by BrandCode: - A (293 observations) appears to be multimodal and have the most outliers, with a mean slightly lower than the aggregate (8.495) - B (1293 observations) appears to be bimodal with a number of outliers, as well as a mean nearest the aggregate (8.562) - C (304 observations) appears to be bimodal and is the most acid (8.419) - D (615 observations) is the most normal distribution and also has the highest alkalinity (8.603) - Missing Values

Predictor Variables

We examined the density of our variables to visualize the distribution of the predictors. Many of these variables contain outliers and present with a skewed distribution. The outliers fall outside the red-line boundaries, and highlight which predictors have heavier tails.

The density plots also contain an overlay of the only categorical indicator, `BrandCode`. This view shows us that some variables, including `AlchRel`, `CarbRel`, `CarbVolume`, `HydPressure4`, and `Tempature`, are strongly influenced by brand type.

[CONTENT EDITORS: DO WE STILL WANT TO CREATE THESE TABLES? IF SO, OUTLIER_WITH OBJECT NEEDS TO BE REBUILT IN MODEL_PREP.R]

[is this going to be evidenced by our graphs? As no predictor variable shows a particularly pronounced monotonic linear relationship with response, a non-linear approach to modeling seems warranted.]

[FIGURE SUCH AND SUCH] helps to further visualize the effect `BrandCode` has on our predictor and `pH` values. For example, `AlchRel` shows distinct `BrandCode` groupings. Other variables, such as `PSCO2`, `BowlSetpoint`, `MinFlow`, and `PressureSetup` show unique features likely related to system processes.

Correlations

The plot below shows that BallingAlch, RelBalling, LvlDensityCarb, RelBrand, CodeDCarb., VolumeCarb, Pressure are all highly correlated with each other, but not particularly highly correlated with the outcome, pH variable. They are all 25% or less correlated with pH, as pH is with most other variables both positive and negative. No extreme heroics were necessary here, despite their being some variables which are highly correlated with each other, because they were sufficiently uncorrelated with the outcome variable and it is not clear how much these are influenced by Brand Code such that removing some may preferentially bias certain brands.

[NEW CHART WITH 2 Sets of Labels?]

4 Data Preparation

Preparing the data was the most discussed and influential part of our modeling process. It was clear from early on that in order to build a useful model with such a narrow range of expected pH values, how we groomed our data and the decisions we made would likely be as or more influential than the model we ultimately chose.

Train/Test Splits:

Prior to all pre-processing, we divided the production dataset using an 80/20 split to create a train and test sets.

All training models incorporated k-folds cross-validation set at 10 folds to protect against overfitting the data. We set up unique model tuning grids to find the optimal parameters for each regression type to ensure the highest accuracy within our predictions.

For both KNN and SVM models a grid of seeds was created from our original seed to ensure that out repeated cross validation would be repeateable. The same seeds were used in both the SVM and KNN.

]

Data Imputation:

Missing values are imputed using the `caret` package so that the same range of imputed values could be applied to the test and validation sets without confounding our training data and a bagging algorithm was used to impute all continuous variables.

Because we were convinced that the ’brand variable `BrandCode` may be one of the strongest predictors of pH, after much discussion, we decided not to impute the Brand Code variable, so that each of the observations with a known brand would be more accurately described by the other variables relative to pH.

Instead the missing labels were replaces with Unknown and the variables were converted to dummies of 0 and 1 to ensure that all modeling methods would be able to consider Brand Code.

Test data is imputed with the same model, with that target variable `PH` removed from the set.

Pre-Processing:

Most of the models concidered in our modeling process require scaling and centering, so we included this in our prepartions. Although, only one variable showed near-zero variance, Hyde.Pressure_1 we opted to remove it from all models during preprocessing and likewise applied Box-Cox conversions to the data to compensate for andy skews and non-normal modaliteis in the variables which might confound our models. Again, the preprocessing model was saved so that the test and validation sets could be consistenly transformed using caret’s predict method.

5 Modeling

We assessed the effectiveness of more than ten different non-linear regression models in our exploratory process. We settled on four models that exhibited the most favorable test metrics, tuned those models, and then chose the best performing model of that set to use in our final analyses (all performance results from the other five are included in [TABLE BLAH BLAH] in [APPENDIX BLAH BLAH]).

[BETHANY: INSERT SIDE-BY-SIDE TRAINING / TESTING METRICS FOR PREPRCESSED MODELS (SET 2) HERE].

- Model 1: Support Vector Machines Regression
- Model 2: Cubist Tree Regression
- Model 3: Multivariate Adaptive Regression Splines Regression
- Model 4: Random Forest Regression

6 Model Performance

- Set1 = Caret: bagImputed; no additional pre-processing
- Set2 = Caret: bagImputed; PreP `method=c('center', 'scale', 'nzv', 'BoxCox')`

Train Performance:

Table 6.1: Train1 Performance

MAPE	RMSE	RSquared	MAE	Method
1.0948	0.5916	0.6757	0.4319	rf
1.1629	0.5500	0.6979	0.3938	cubist
1.2664	0.6985	0.5165	0.5037	svmRadial
1.5170	0.6921	0.5261	0.5153	earth

Table 6.2: Train2 Performance

MAPE	RMSE	RSquared	MAE	Method
0.0081	0.0965	0.6904	0.0688	cubist
0.0088	0.1030	0.6735	0.0751	rf
0.0104	0.1224	0.5042	0.0883	svmRadial
0.0112	0.1265	0.4690	0.0954	earth

Test Accuracy:

Table 6.3: Test1 Performance

MAPE	RMSE	Rsquared	MAE	Method
0.4930	0.2234	0.9539	0.1578	cubist
0.5351	0.3265	0.9568	0.2466	rf
0.9987	0.6076	0.6385	0.4101	svmRadial
1.6688	0.7260	0.5226	0.5269	earth

Table 6.4: Test2 Performance

MAPE	RMSE	Rsquared	MAE	Method
0.0029	0.0367	0.9586	0.0250	cubist
0.0037	0.0436	0.9596	0.0311	rf
0.0084	0.1064	0.6307	0.0717	svmRadial
0.0107	0.1193	0.5257	0.0908	earth

[BETHANY: EACH OF US SHOULD WRITE BULLETS WITH REASONS TO CHOOSE THIS MODEL BY SAT EVENING 12/7 - BETHANY WILL FLESH OUT]

Model Selection Considerations

[BETHANY: NEED TO PICK OUR FIRST CHOICE MODEL, THINK IS SHOULD BE VARIMP-ABLE SO WE CAN USE THAT IN INTERPRETATION / CONCLUSIONS]

Model 1: Support Vector Machines (SVM) Regression

[BETHANY TO WORDSMITH RATIONALE FOR USING SVM MODEL] Support vector machine (SVM) regression with a radial bias functin kernel is a promising choice for predicting beverage pH because it excels when working with data which may not be linearly separable, which comes into play with this data specifically because pH is non-linear.

Although less efficient than the k-nearest neighbor and Multiple Adaptive Regression Splines to train, the SVM provided robust final model using a radial kernel with a cost of 10, passed as the tune length settling on $\sigma = 0.020$ and $cost = 8$ returning a $RMSE = 0.1127$

```
# remove echo/eval later [BETHANY: ADD IN VARIMP /
# PERFORMANCE CHART FOR SVM AS ALIGNED WITH GROUP]
```

Model 2: Cubist Tree Regression

[JEREMY: CONDENSING BELOW WITH RATIONALE FOR USING MODEL IN BULLET FORM BY EVENING SAT 12/7 - BETHANY WILL WORDSMITH]

· Cubist regression models provide a balance between predictive accuracy interpretability · For a continuous response variable, Cubist models functions like as piecewise linear model · The model creating rules (which can overlap) to subset the data and then regression models to each subset to arrive at a prediction. · They can also integrate instance-based, nearest neighbors ensembling and boosting using committees. · Based on cross-validation [ASSUMING THIS IS STILL CORRECT] and grid search across hyper-parameters, we found best RMSE with an instance-based model tuned to 5 neighbors and 50 committees. Based on cross-validation and a grid search across hyper-parameters, we found the best RMSE performance with an instance-based model that factoring in many neighbors built on non-pre-processed training data.

Model 3: Multivariate Adaptive Regression Splines (MARS) Regression

Multivariate regression splines (MARS) are more flexible about relationships between preditors and the outcome variable than linear regression models yet maintain their ease of interpretation.

MARS models also perform well without major pre processing steps with reasonable bias-variance trade-off and are computationally efficient as well as optimized work on very large data sets efficiently.

Model 4: Random Forest Regression

[ANDY: PLEASE ADD CONCISE BULLETS WITH RATIONALE FOR USING RF MODEL BY EVENING SAT 12/7 - BETHANY WILL WORDSMITH]

The optimal parameters for model was mtry = 31 and ntree = 2500. MAPE is **r s$MAPE** where as top 3 important predictors are `MnfFlow`, `BrandCode` and `PressureVacuum` for %incMSE and `MnfFlow`, `BrandCode` and `OxygenFiller` for IncNodePurity. Unlike `PLS`, `Random Forest` can produce 2 different variable importance plots.

The first graph shows how much MSE would increase if a variable is assigned with values by random permutation. The second plot is based on `node purity` which is measured by the difference between RSS before and after the split on that variable (`Gini Index`). In short, each graph shows how much MSE or Impurity increases when each variable is randomly permuted.

7 Interpretation

[BETHANY MAKING MAGIC HAPPEN WITH APPROPRIATE VARIMP GRAPH ONCE FINAL MODEL SELECTED]

8 Conclusion

[BETHANY CREATING NEXT STEPS FOR PRODUCTION PROCESS BASED ON FINAL MODEL]

Appendix

Code

Data Dictionary

Exploratory Plots and List Models

9 Citations

Shelton, Robert B. "PH Values Of Common Drinks." Robert B. Shelton, DDS MAGD Dentist Longview Texas, 2019, www.sheltondentistry.com/patient-information/ph-values-common-drinks/.

Cubist Model Background: https://www.rulequest.com/cubist-win.html Cubist Model Overview: https://static1.squarespace.com/static/51156277e4b0b8b2ffe11c00/t/56e3056a3c44d8779a61988a/1457718645593/cubist_BRUG.pdf Cubist Model Mechanics: http://ftp.uni-bayreuth.de/math/statlib/R/CRAN/doc/vignettes/caret/caretTrain.pdf]

(https://en.wikipedia.org/wiki/PH).