Applied Biostatistics (MATH-493) - Final Report

Cosmin Rusu
EPFL, Switzerland
cosmin.rusu@epfl.ch

***Abstract*—This report aims to present a model to study the effect of pH, temperature, and other factors to the growth of the Alicyclobacillus acidoterrestris in apple juice. Such a model can have industrial applications such as monitoring infrastructure, food quality management, and cost reduction. The main reference of this report is the paper by Pena et Al [1], and we propose a simpler model that have similar performance.**

I. Introduction

Alicyclobacillus acidoterrestris is a thermoacidophilic, Gram positive, strictly aerobic bacterium. It was first isolated and identified from soil [2], but then was found from several pasteurized and contaminated fruit juices such as orange, apple or passion fruit juice. It was shown that the bacteria can grow in the 35-55 °C range, and in pH 2.2-5.8. However, the optimum growth temperature for A. acidoterrestris is 42-53 °C [3].

The growth prediction of this bacteria can have many practical applications such as food quality control, quality management, risk assessment, and cost reducing. Companies can quickly act on the development process and mitigate the risk of producing contaminated food, by spotting the growth probability early on.

Since evaluating the growth potential of the bacteria become difficult when studying more than one variable, a logistic regression model will be used to find the effects of temperature (25 to 50 °C), pH (3.5 to 5.5), soluble solids concentrations (11 to 19), and nisin concentration (0 to 70 IU/mL) on the probability of Alicyclobacillus acidoterrestris growth in apple juice.

The logistic regression model has been successfully applied in the past to numerous similar problems. For example, Presser et al [4] model the growth probability of E. coli as a function of pH, temperature, and water activity using logistic regression. By a similar approach, we can estimate the growth probability of Alicyclobacillus acidoterrestris given enough information about the storage conditions and product characteristics.

II. Exploratory analysis

A. Response and Predictable Variables

The dataset contains 5 columns, with 74 rows (tuples) in total. Temperature, pH, brix concentration and nisin concentration are our predictable variables and as the response variable we have a binary variable indicating weather or not, under those specific conditions, the growth of the bacteria was observed. There are 48 observations for the the negative class (growth column is 0) and 26 observations for the positive class (growth of the bacteria has been observed and the value is 1). A sample of the data can be seen in Table I. To make sure we do not introduce bias in our model, we will use the subsampling method to balance the two classes.

pH	NisinConc	Temperature	BrixConc	Growth
5.5	70	50	11	0
5.5	70	43	19	0
5.5	50	43	13	1
5.5	35	50	15	1
5.5	30	35	13	1
5.5	30	25	11	0

TABLE I
6 SAMPLES FROM THE DATASET

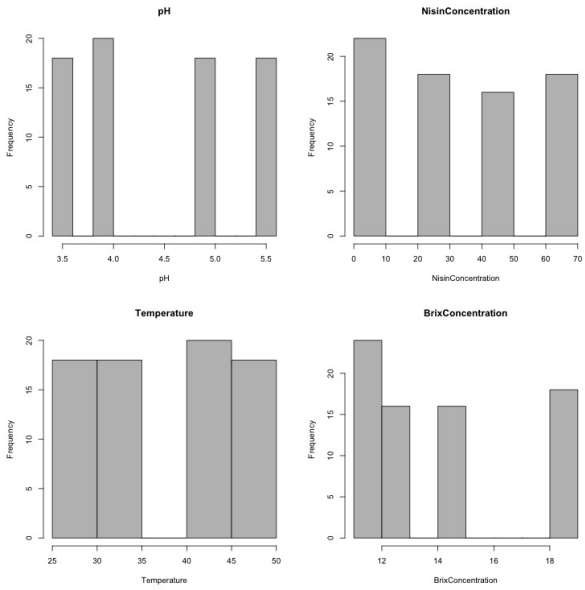

Fig. 1. A histogram visualization of the predictable variables distribution

In Figure 1 we can see a histogram of the predictable variables and in Figure 2 we can see a box plot visualization of the features. There are no missing or undefined values in our dataset, which makes the cleaning part of the analysis much simpler. Clearly, the features lie in different scales. And since logistic regression usually performs better when the features are on the same scale, we'll scale the features by subtracting the mean and dividing by their respective standard deviation. After scaling, a scatter plot for all the variable pairs is displayed in Figure 3. There is not obvious separation between the two, indicating that interaction and polynomial terms should be added to the model.

The density distribution of each variable broken down by

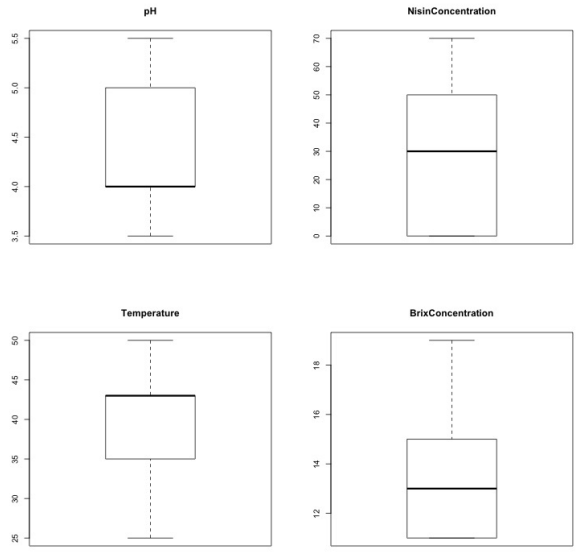

Fig. 2. A boxplot visualization of the predictable variables distribution

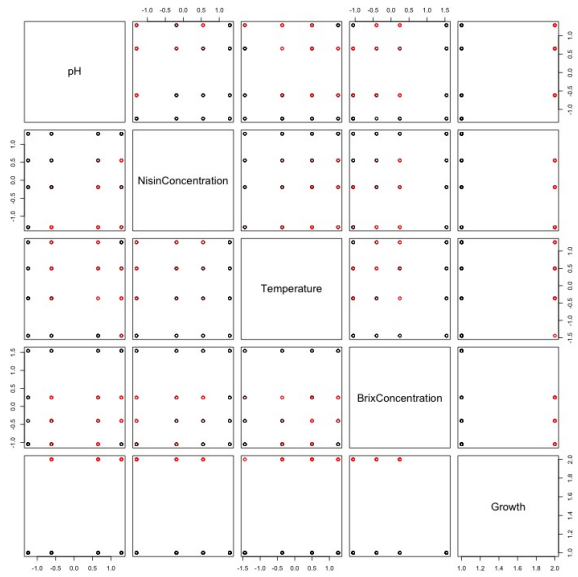

Fig. 3. Scatter plot of the normalized data, color indicating the response variable.

Growth value can be seen in Figure 4. Again, there seems to be no obvious separation for each individual variable. This is another indication that more complex terms must be added to a potential logistic regression model. As we will see in the model part of this report, our proposed model will have less interaction terms than the one found in [1].

B. Correlations

The features do not correlate with each other, as we can see from the correlation plot. This was expected since all of these variables are independent (hence they do not change together). The motivation behind checking this was to make sure the model will not suffer from bias.

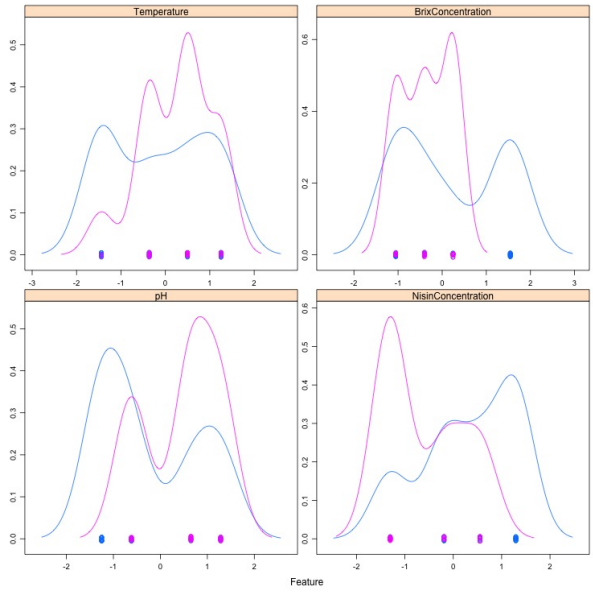

Fig. 4. Density plot of each variable with respect to each value of the Growth response variable.

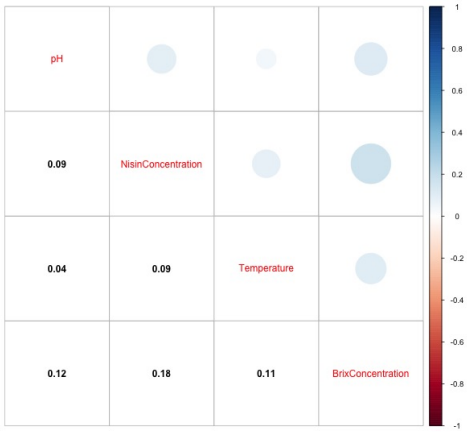

Fig. 5. There is no correlation between the features.

III. Models

A. Model fitting

Since the positive class has less observations than the negative class, we will always sample the negative class to match the number of positive class tuples. Moreover, because the dataset is quite small, we will choose a slightly modified leave one out strategy to asses the performance of our model. The strategy will be to take one tuple out, to subsample the negative class, to train the model using that subsample and the whole positive class, and to predict on the left-out tuple. Then, repeat this process for each tuple and take the average of these results.

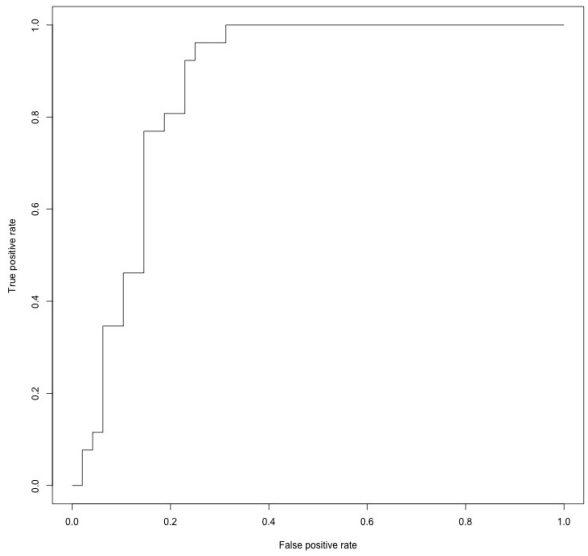

Fig. 6. The ROC of the simple model.

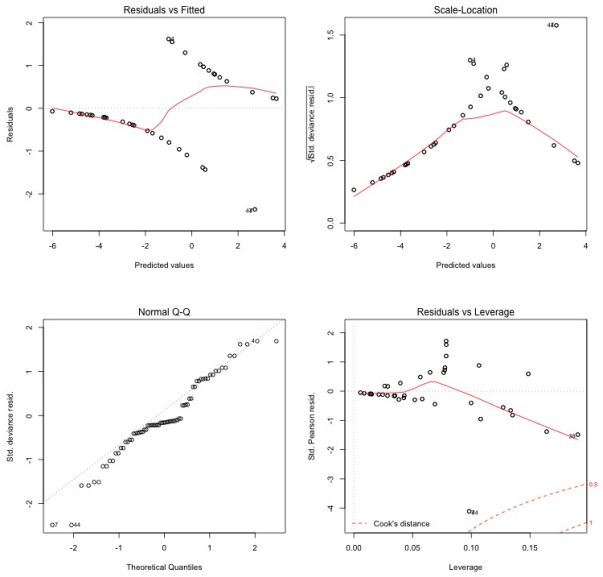

Fig. 7. R diagnostic plot for the simple model.

B. Simple model

The first model considered was a linear combination of all the response variables plus an intercept (bias) variable. The variables pH and Nisin Concentration are the most significant ones, with a p-value less than 0.001. While pH has a positive coefficient, Nisin Concentration has a negative one, suggesting that all other variables being equal, the higher the concentration the less likely it is for the bacteria to grow.

In terms of the performance of this model, we've achieved an accuracy score of 81.08% and an F1-score of 0.758. To overcome overfitting the data and the unbalanced dataset, we have combined the leave one out approach with sub sampling. Concretely, at every step, we take one data tuple out, balance the dataset (by subsampling the most frequent class - the non-growth class), and predict on the data that was removed. In the end, we compute the average and F1-score based on all these 74 predictions (since every tuple was left out exactly once).

It is worth mentioning that we've picked the threshold at 0.5. The AUC for the model was 0.8701%, indicating a good predictive ability. A plot of the ROC curve can be seen in the Figure 6.

A diagnostic plot for this simple model can be seen in 8. The normal Q-Q plot looks almost normal, with only a few outliers. From the Cook's distance plot, we can see that only two outliers heavily influence the regression (leverage points).

C. Better model

Motivated by the fact that interaction terms might be relevant, we performed a stepwise model selection by AIC starting with the basic model that contains only the interaction term. Our model is a linear combination of all the features plus the interaction terms (pH, BrixConcentration), (NisinConcentration, BrixConcentration), (Temperature, BrixConcentration), and (NisinConcentration, Temperature). The resulting model

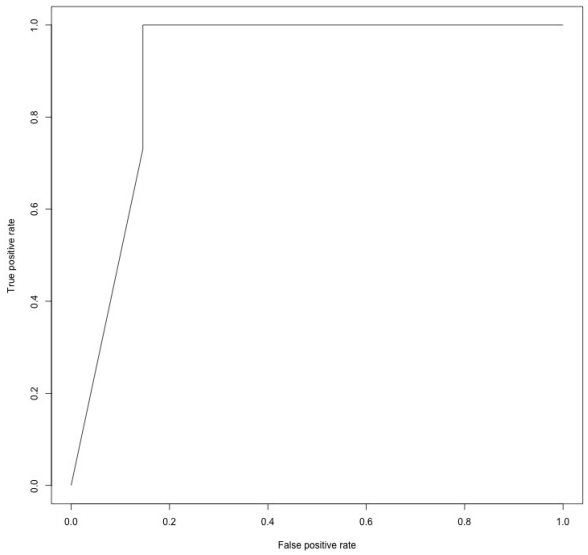

Fig. 8. R diagnostic plot for the better, more complex model.

has an accuracy of 93.24% and an F1-score of 0.912%. Again, the performance was computed in the same fashion as the simple model, using the leave one out with subsampling strategy. The AUC value was 0.9567, indicating, again, a strong predictive ability.

IV. Conclusions and Future Improvements

We've analyzed a simple logistic regression model to study how the growth of the Alicyclobacillus acidoterrestris bacteria is stimulated by the environment variables pH, temperature, brix and nisin concentration. We've proposed a model that is easier to debug and understand, and the performance we've achieved suggest a better real-life performance on unseen data.

The resulting model contains less interaction terms then the model found in [1], which makes it faster, and easier to understand when deployed at scale. An exact comparison in terms of the performance would be impossible given the fact that we don't know if they subsampled the data or not. Moreover, we suspect that their model is prone to overfitting by not addressing this unbalance setting.

A potential future improvement would be to see the effects of introducing polynomial features to the logistic model. We believe it might be able to produce better results, the only challenge would be to overcome overfitting.

REFERENCES

[1] W. Pena, P. De Massaguer, A. Zuniga, and S. Saraiva, "Modeling the growth limit of alicyclobacillus acidoterrestris cra7152 in apple juice: effect of ph, brix, temperature and nisin concentration," *Journal of Food Processing and Preservation*, vol. 35, no. 4, pp. 509–517, 2011.

[2] G. Deinhard, P. Blanz, K. Poralla, and E. Altan, "Bacillus acidoterrestris sp. nov., a new thermotolerant acidophile isolated from different soils," *Systematic and Applied Microbiology*, vol. 10, no. 1, pp. 47–53, 1987.

[3] J. D. Wisotzkey, P. Jurtshuk JR, G. E. Fox, G. Deinhard, and K. Poralla, "Comparative sequence analyses on the 16s rrna (rdna) of bacillus acidocaldarius, bacillus acidoterrestris, and bacillus cycloheptanicus and proposal for creation of a new genus, alicyclobacillus gen. nov." *International Journal of Systematic and Evolutionary Microbiology*, vol. 42, no. 2, pp. 263–269, 1992.

[4] K. Presser, "Ross. t.; ratkowsky. da 1998. modelling the growth limits (growth/no growth interface) of escherichia coli as a function of temperature. ph. lactic acid concentration and water activity," *Applied and Environmental microbiology*, vol. 5, pp. 1773–1779.