Predicting BiodegradabilityChallenge

Contents

Report Background 1

Introduction 1

Data Description: 2

Baseline Results: 2

Feature Selection Results 2

Challenge Prediction 2

Comparison of Methods 3

Additional Analysis 3

Conclusion 3

Notebook Grading Rubrics [REMOVE FROM FINAL NOTEBOOK] 3

Report Background

This report was prepared for **ChemsRUs** by *YOUR NAME HERE*

This report embeds all of the R code necessary to produce the results described in this report. If non-R programs are used, then summarize the results here.

NOTE: THIS IS A TEMPLATE FOR THE REPORT WITH INSTRUCTIONS FROM ASSIGNMENTS INCLUDED FOR YOUR CONVENIENCE. ***DELETE THE INSTRUCTIONS GIVEN IN ITALICS IN YOUR FINAL REPORT NOTEBOOK.*** *THE NOTEBOOK SHOULD BE AS YOU WOULD GIVE CHEMS-R-US. CODE FRAGMENTS ARE PROVIDED. YOU CAN ADD OR DELETE R CODE BLOCKS AS NECESSARY. THERE IS SOME SAMPLE CODE AT THE END OF THIS NOTEBOOK. IT SHOULD BE REMOVED BEFORE SUBMISSION.*

Introduction

Provide an overview of your report

Chems-R-Us has created an entry to the challenge at https://competitions.codalab.org/competitions/22892 based on logistic regression. Based on the information in the leaderboard under bennek, their entry is not performing feature selection well. The purpose of this report is to investigate alternative approaches that achieve high AUC scores on the testing set while correctly identifying the releaves features.

The approaches tried were...

Data Description:

Provide a basic description of the data which includes: 1. number of attributes 2. number of points in each class

Describe the data preparation The data preparation should include:

- Read in the external train and external test datasets.
- Divide the external training set into an internal train and internal validation set using an 90% and 10% split.
- Chems-R-Us didnot scale any data in their entry. But you can add centering and scaling if desired.

Handy HINTS if you would like to scale: The following code scales data in matrix `tr` and then applies the same scaling to matrix `tst`.

'sc_tr <- scale(tr,center = TRUE, scale = TRUE) # scale tr means <- attr(sc_tr, 'scaled:center') # get the mean of the columns stdevs <- attr(sc_tr, 'scaled:scale') # get the std of the columns sc_tst <- scale(tst, center=means, scale=stdevs)#scale tst using the means and std of tr'

Baseline Results:

*Investigation of alternative models using all the features. ChemsRUs has asked you to evaluate how LDA and logistic regression performs on this problem using all the features. Divide the training data into 90% train and 10% validation splits. Set seed(300) before you split so you get the same train and validation splits. Train LDA and logistic regression on the training data and evaluate how well it does on the validation data. Compute the balanced accuracy and the AUC for the train and test results. Compare the results between the two models.

Feature Selection Results

Create an approach for selecting the relevant features. Describe your approach. Describe the features that you selected. Create a PCA biplot comparing the two classes with these two classes. Create a classifier using LDA and logistic regression using these features. Evaluate how they perform on the training and sets in terms of balanced accuracy and AUC. Compare your results with your prior results. Discuss your findings.

Challenge Prediction

My challenge ID is .... with an AUC score of ..... for prediction and .... for feature selection.

Pick your best shot classification and feature selection methods and then enter the contest. Provide your scores in the text. Discuss your results and the strengths and weaknesses of the different approaches. There should be a separate entry for each participant. Make sure that between all of your teams entries three different classification methods are used and three different feature selection methods are used. Using PCA with another method counts as a different classification method.

The contest can be found here:

https://competitions.codalab.org/competitions/22892

*Enter the contest. Prepare you entry by making the classification.csv and selection.csv and zipping them into a single file. See FinalProjChemsRUs.Rmd for an example and discussion of the format of the file.

Provide a csv file with your predictions of the biodegradibility of each data point in `chems_test.csv`. Chems-R-Us will use this to independently verify the quality of your results. These predictions should be given as a csv files with on column containing the prediction (`1` or `-1`) for each points in chemstest.csv. Provide a csv file with your prediction of which features are real. The feature predictions should be given as a csv files with

on column containing the prediction (`1` or `0`) indicating if each of the 168 features should be included. Create the files:

- `classification.csv`: Test target values (437 lines x 1 column)
- `selection.csv`: Solution indicating which variables are real and which are fake (168 lines x 1 column)

Your submission must be a zip archive containing the following files: - `classification.csv`, your predicted labels for test dataset. It should include plus or minus one values, one for each test sample, representing the class label predictions. - `selection.csv`, representing the features you selected as real or fake (ie `0` or `1`).

Comparison of Methods

*Create a **table** comparing the results between your group members on the contest data as reported on the leader board. Create an ROC curve plot of the test classification results that includes all of the methods in one plot. Include a table made by creating a data frame that summarizes the results of each team member and then nicely printing it with the `kable()` function; **Do not include a screen shot of "raw" R output!! in your presentations** Include the number of team member, the number of features used, training set AUC, internal testing set AUC, classification AUC from challenge, and feature selection AUC challenges. Discuss which method performed best for feature selection, and which method worked best for prediction. What method would you recommend overall? Why?*

Additional Analysis

Provide additional analysis and/or visualizations that may be insightful to Chems-R-Us. Use some method in R not covered by prior labs. Use your imagination, extra credit for creativity here! Discuss the insights your analysis provides. Be sure to title any figures! Comment your code so all can understand what you are doing. Feel free to use any R code from class or from the web.

Conclusion

Provide a conclusion which summarizes your results briefly and adds any observations/suggestions that you have for Chems-R-Us about the data, model, or future work.

Notebook Grading Rubrics [REMOVE FROM FINAL NOTEBOOK]

The project will be graded using the following rubric.

- (5 pts) Did the data description described both test and train? Did it explain data preparation.
- (5 pts) Were the baseline results properly done, presented, and discussed?
- (10 pts) Was the procedure for feature selection well thought-out, presented, described, and executed?
- (10 pts) Was the procedure for constructing the final challenge predictive model well thought-out, described, and executed?
- (5 pts) Did the individual enter the contest sucessfully and report their result. Were the results discussed?
- (5 pts) Were the additional visualizations and analysis effective? Was a method not covered in class included? Were the results discussed?This grade will be an individual grade.
- (5 pts) Was a conclusion provided that included a summary of primary findings and discussion of future work?
- (5 pts) What is the grammatical quality and clarity of the written report? Did you communicate your results effectively in written form. Did you discuss results
- (3 pts) Up to 3 points extra credit available for extra effort and very creative solutions. Get extra credit for being the top in classification, features selection, or overall ranking.