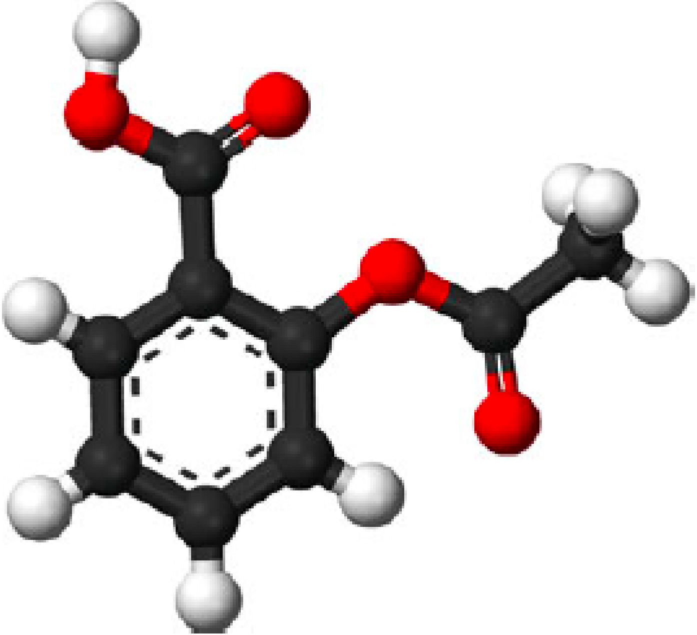
Grant Case Study

Kuhn and Johnson, Ch. 9

Brooke Anderson

April 6, 2016

Drug development



Chemical descriptors

Chemical descriptors: “There are myriad types of descriptors that can be derived from a chemical equation.”
Examples:

- number of carbon atoms
- molecular weight
- electrical charge
- surface area

QSAR

A key task in drug development is determining which compounds have certain biological activity.

Quantitative structure-activity relationship (QSAR) modeling: Try to determine something like biological activity (e.g., if it can inhibit production of a specific protein) from chemical descriptors of the chemical (including for componds that don't yet exist).

This is usually determined through experimentation.

Other compound characteristics

In addition to a compound's activity, you also need to figure out other properties, to determine if it could be used for a drug:

- Solubility
- Toxicity

Purpose of model

Model aim: Predict whether the solubility of a chemical compound based on its *chemical descriptors*.

- Training data: 1,267 chemical compounds

For each compound, solubility was determined experimentally.

Predictive variables

Predictive variables include:

- 208 binary “fingerprints” indicating the presence or absence of a particular chemical structure
- 16 count variables (e.g., number of bonds, number of bromine atoms)
- 4 continuous descriptors (e.g., molecular weight, surface area)

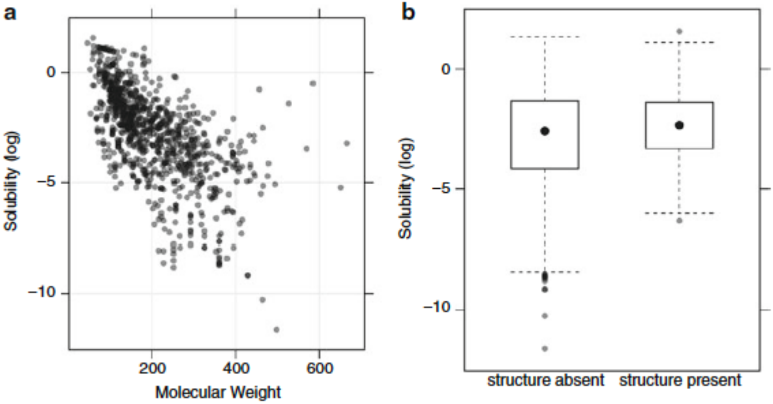
Evaluation

Solubility was converted to the $\log_{10}$ scale.

Models were assessed using root mean squared error (RMSE).

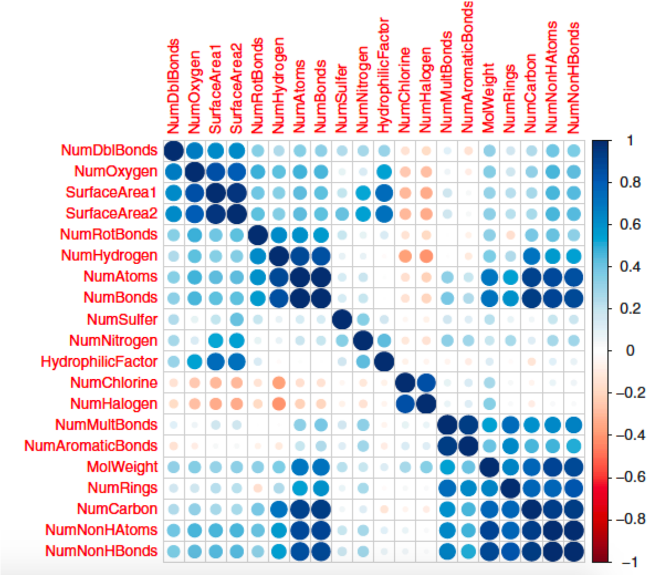
Exploratory data analysis

Univariate plots:



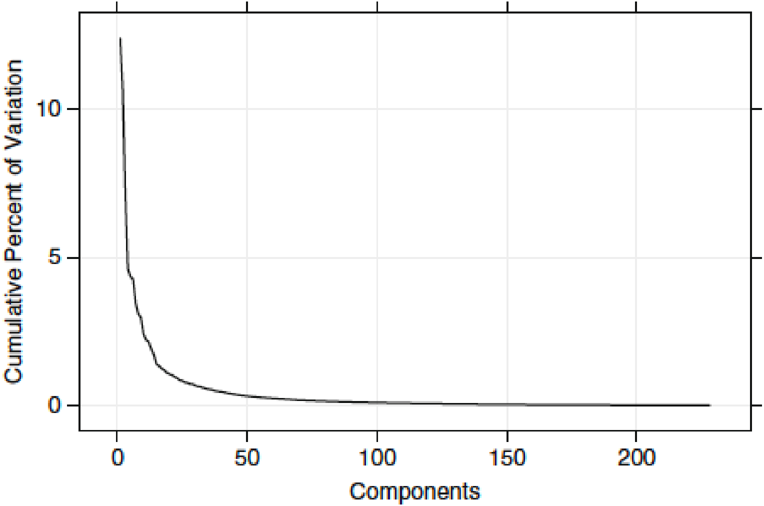
Exploratory data analysis

Pair-wise correlation of count / continuous predictive variables:



Exploratory data analysis

Results of a PCA:



Exploratory data analysis

Shape of distributions for predictors:

> *"The count-based descriptors show a significant right skewness."*

They assessed this by measuring an average skewness statistic.

Splitting data

They randomly split the data into training and testing sets:

- Training dataset: 951 chemical compounds
- Test dataset: 316 chemical compounds

> *"The training set will be used to tune and estimate models, as well as to determine initial estimates of performance using repeated 10-fold cross-validation. The test set will be used for a final characterization of the models of interest."*

Pre-processing

For the binary variables, “there is very little that pre-processing will accomplish”.

Pre-processing

For continuous variables, two of the issues of concern are:

- Skewness
- Between-predictor correlations

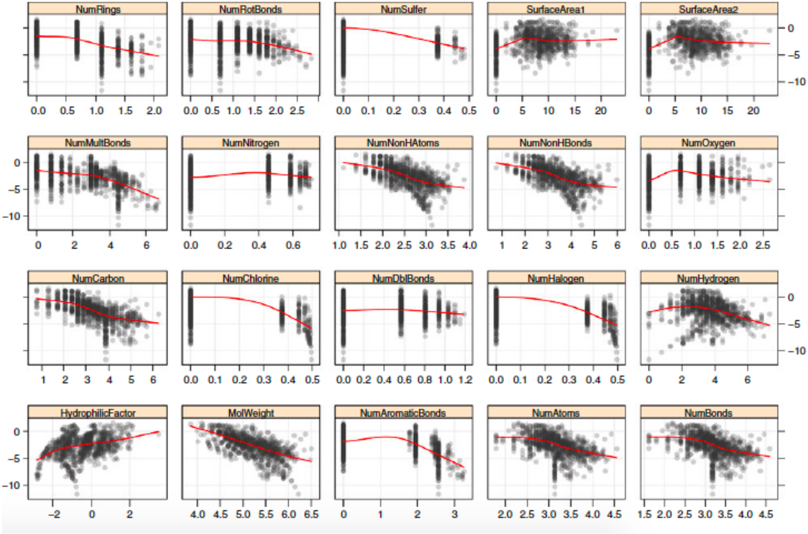
Pre-processing

To account for skewness, they applied a Box-Cox transformation to all of the continuous predictors.

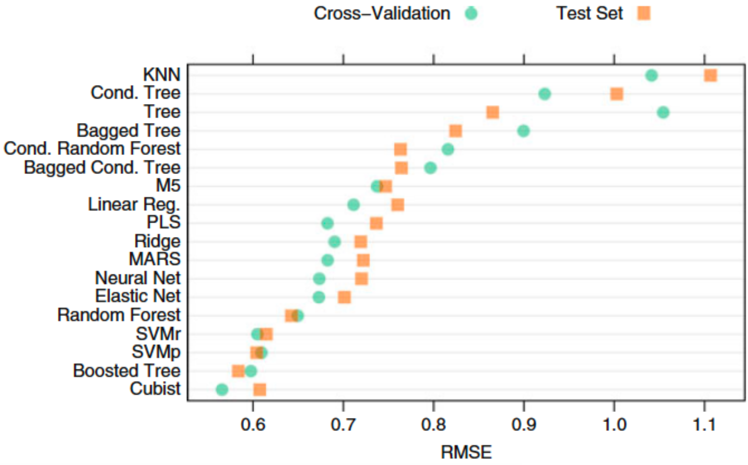
In `caret`, you can do this using the `preProcess` function, with `"BoxCox"` as one of the options in the `method` argument.

Feature engineering

Added quadratic terms for continuous variables.



Overall results


Cross–Validation ●
Test Set ■
KNN
Cond. Tree
Tree
Bagged Tree
Cond. Random Forest
Bagged Cond. Tree
M5
Linear Reg.
PLS
Ridge
MARS
Neural Net
Elastic Net
Random Forest
SVMr
SVMp
Boosted Tree
Cubist
RMSE