Rich Gruss

Practical Data Science With R, Chapter 9

1. One way to avoid over-fitting is to use to a three-way data split: train, calibration, and test. The test set is reserved for evaluating the final model. We build the model on the training set, and then use the calibration set to see if we're over-fitting. If we are, we can re-train. Cross validation will help to estimate the effects of over-fitting.
2. A common method of using numerical data for prediction is to bin the data on ranges and use the range as a categorical variable. This is exactly what we're doing with our bicycle rental data project. We want to build a machine learning algorithm to predict the number of rentals per hour, and we're using a training set that gives labels to 100 bins of range 10. To do this in R, you would use the quantile() and cut() commands.
3. Feature/variable selection is the key to building accurate models. Each variable can contribute either substantial information or could introduce noise. One way of selecting variables is to use the AIC, the Akaike Information criterion. The AIC gives a score that evaluates the tradeoff between goodness of fit and the complexity of the model, so more variables might add to the accuracy of the model, but may not be worth the added complexity.