Rich Gruss

Practical Data Science with R, Ch. 7

1. linear regression, useful if you would like to predict a numerical response variable given a vector of input explanatory variables, is such a central part of data analysis that R makes it laughably easy: R provides a two-letter function lm(), which a returns an object that can be passed to the predict() function along with a matrix of new values for testing. Important summaries for assessing the strength of the regression: a) R-squared (percentage of the variance explained by the linear model), and b) residuals quantiles (you want them to be centered around 0, and without too large a range).
2. Important points in reading the output for a linear regression: a) the coefficients give a measure of how influential a particular variable is (magnitude) and whether it affects the response variable positively or negatively (sign). b) Pr(>|t|) is the p-value, the probability that the variance occurs due to chance, and anything smaller than .05 indicates that the variable is significant. .005 is not necessarily "more significant" than .05; they're both significant.
3. Logistic regression is intended to state a probability that the input vector belongs to a certain class. The result of the linear combination of explanatory variables is passed through the logistic function (1/(1+E^x)), to yield a value between 0 and 1. The inverse of the logistic function is logit function(P/(1-p)), if you want to recover the initial result. Logistic regression is performed in R using the "generalized linear model" function glm and passing the parameter family=binomial(link="logit")
4. You can use machine learning techniques to test the validity of the regression. Create a train-test split (maybe 90-10), build the model on the train set, and then call predict(model, test) on the test set.