Print

Building a Predictive Model

The first part of the predictive modeling process is model building. You start by fitting a variety of models, and then you assess their performance and select the best model. You want to select a model that generalizes well that is, the model that's flexible enough to accurately predict new data. The key is to not overfit the training data set. The classic example of overfitting is selecting linear regression models based on R-square because the R-square metric improves or fits the model at least as well each time a predictor is added to the model.

How can you select the best model?

To ensure that the chosen model generalizes well, you use honest assessment, where you perform the assessment on a different data set than the one you use to build the model. Partitioning the data enables you to train the model and assess its performance on new cases. Using honest assessment, you partition or split the available data into a data set for training, and one for validation, and sometimes a third data set for testing. All partitions contain the predictors and the response.

The training data set is used to fit a variety of different models. The validation data set is a holdout sample that's used to compare model performance and select the best performing model. Using a holdout sample is a way of assessing how well the models generalize to new data. If created, the test data set is used to give a final honest estimate of generalization for the chosen model. In practice, many analysts see no need for a final assessment. Instead, the model assessment measured on the validation data is reported as an upper bound on the performance that is expected when the model is deployed.

Unfortunately, there's no globally optimal percentage to use when you partition the data. Common partitions include 70% training and 30% validation, or an 80/20 split, or 90/10 split, but the decision is the responsibility of the analyst.

Predictive modeling is typically used when we have very large data sets and partitioning the data enables us to still build models on an adequate amount of data. However, if you start with a small or medium-size data set, partitioning the data might not be efficient, because the reduced sample size can severely degrade the fit of the model. In fact, computer-intensive methods, such as cross validation were developed so that all the data can be used for both fitting and honest assessment.

Statistics 1: Introduction to ANOVA, Regression, and Logistic Regression

Close