Imputing Missing Values from a Model

Imputing missing values as a constant (the mean or median) or with a random value are is quick, easy, and often sufficient to solve the problem of missing data. However, better imputation can be achieved for non-MCAR missing values by other means.

The model approach to missing value imputation begins with changing the role of the input variable with missing values to now be a target variable. The inputs to the new model are other input variables that may predict this new target variable well. The training data should be large enough, and all of the inputs must be populated; listwise deletion is an appropriate way to remove records with any missing values. Keep in mind that even a moderately accurate model can still produce good imputations; the alternatives are either random or constant imputation methods.

Modelers have two problems with using models to impute missing values. First, if the software doesn't have methods to impute values from models automatically, this kind of imputation can take considerable time and effort; one must build as many models are as there are variables in the data.

Second, even if the imputation methods are done efficiently and effectively, they have to be done again when the models are deployed, any missing values in data to be scored must first be run through a model, adding complexity to any deployment process.

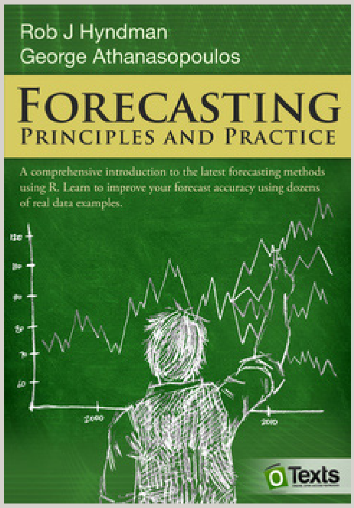
Nevertheless, the benefits may outweigh the problems. It seems that the modeling algorithms most often included in predictive analytics software to automatically impute missing values are decision trees and k- nearest neighbor, both of which can work well with imputing continuous and categorical values.

BOOK REVIEW: FORECASTING

While working on forecasting (understand "time series analysis") I found several interesting and state of the art articles from Rob J. Hyndman. He is the co-author, with George Athanasopoulos of *Forecasting: Principles and Practice*. This is an excellent concise and comprehensive text explaining concepts behind forecasting, common algorithms and how to implement them in R (for a business view of forecasting, I advise *Future Ready*).

The book presents key concepts of forecasting. From judgmental forecasting (which can be useful when you have no or few data) to simple/multiple regression, time series decomposition, exponential smoothing (ETS), ARIMA and a few more advanced topics such as Neural Networks. I would suggest to the author to add Support Vector Regression (SVR) and ensemble learning for the next edition of the book. Each concept of the book is covered through examples with real data. What is most appreciable about the book is how concise and readable it is. Each sentence is useful to understand the described concept, nothing "superflu".

The book contains a good overview and schema about each technique and how to set their meta-parameters. The R codes are well presented, concise and easy to implement and test. The book can easily be used to teach forecasting since each chapter contains exercises. In conclusion, *Forecasting: Principles and Practice* is THE book to learn time series analysis algorithms and how to implement them.



Review by Sandro Saitta, Data Scientist at Expedia