Article from
Predictive Analytics & Futurism
December 2017
Issue 16

Dangers of Overfitting in Predictive Analytics

By Rosmery Cruz

We're given a fixed dataset, and we are asked to build a predictive model. The problem is, how can we be sure that the model we're building with the data we have access to today, will allow us to make useful predictions in the future? All applied statistics practitioners face this problem, and without careful attention, they will build models that either don't predict new data well or find insights that aren't replicable. These scenarios occur when models overfit. This article is organized as follows. First, we will define the concept of overfitting, next, we will discuss when overfitting is likely to occur and provide some strategies to minimize overfitting.

Overfitting yields overly optimistic model results: 'findings' that appear in an overfitted model don't really exist.

OVERFITTING: A DEFINITION

Overfitting is defined in a variety of ways across many disciplines, however, Babyak (2004) provides an intuitive definition: "The problem of capitalizing on the idiosyncratic characteristics of the sample at hand, also known as overfitting, in regression-type models. Overfitting yields overly optimistic model results: 'findings' that appear in an overfitted model don't really exist in the population and hence will not replicate." Put another way, overfitted models will start picking up more of the noise in your sample data instead of the underlying process or pattern that exists in the world. As a result, these models will fail to provide accurate predictions or useful insights.

WHEN DOES OVERFITTING OCCUR?

Generally, there are two key areas where analyst oversight leads to overfitting: researcher degrees of freedom and asking too much from the data. The former concept relates to the number of unrestricted choices available to an analyst that leads to obtaining results that don't hold in future samples, while the latter concept is related to model complexity given the number of observations available in your data. The upcoming sections will focus on these two concepts and how they relate to overfitting.

When Does Overfitting Occur? Researcher Degrees of Freedom

Researcher degrees of freedom is receiving more and more attention as the replication crisis across many disciplines continues to unfold. The frequentist application of statistics assumes that there is a "true" model that exists in the world, and repetitions of the same experiment should generate similar findings (Gelman and Loken, 2013). Thus, the ability to replicate previous results is a critical component of the scientific process. However, third-party and original researchers alike fail to replicate many published findings. In the paper "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant," Simmons et al. (2011) provide computer simulations and experiments to show how easy it is to uncover relationships in the data that don't actually exist. In the experiments, they set out to prove that certain songs can change a listener's age. Through a series of data manipulations, and valid statistical techniques, they were able to show that listening to the song "Hot Potato," made people feel older than they were, while the song "When I'm Sixty-Four" by the Beatles made people feel younger than their actual age!

So how can this be? Gelman and Loken (2013) provide an answer: "Statistical significance can be obtained from pure noise, just by repeatedly performing comparisons, excluding data in different ways, examining different interactions and controlling for different predictors, and so forth." Given all the choices we have to make in our analysis, how can we be sure that the results we produce are sound, and more likely to be reproduced in the future?

There are a variety of strategies you can employ in your statistical analysis to reduce vulnerability to overfitting from researcher degrees of freedom. These strategies include: predetermine your analysis plan before exploring your data, rely on subject matter expertise to inform comparisons and grouping of data and limit the exclusion of observations from your dataset.

The first strategy is to create a framework for analyzing your data before the data exploration phase (Babyak 2004). You should have a clear question or problem that you'd like answered, and your analysis plan should reflect the steps you will take to answer that question. Here, you should outline if you will only be focusing on particular subsets of the data, potential predictors of interest, and methods you will use to select your model. Of course, it's difficult to a priori anticipate all issues/difficulties that may arise in data unexpectedly, but the more decisions you make beforehand and stick to, the less

likely you are to start making arbitrary choices contingent on you observing the data. The more decisions made that are contingent on the sample data, the more vulnerable you are to overfitting.

A second strategy to reduce researcher degrees of freedom is to rely on subject matter expertise or previous research to help inform comparisons or grouping of data (Babyak 2004). It's very simple, and tempting, to view your data to make decisions about how to bin age groups, group time points, etc. However, for increased robustness of your results, the more you can rely on previous research, or evidence from your own industry regarding appropriate data manipulation, comparisons and groupings, the less likely you are to produce results that don't replicate.

Limiting the exclusion of observations from your data is the third and final strategy presented here to reduce researcher degrees of freedom. To be sure, identifying and removing data entry errors is important and is not at question here. Instead, removing records due to cut points such as two or three standard deviations from the mean is arbitrary, and contingent on the distribution of the data itself (Simmons et al. 2011). It's important to spend some time determining if you truly understand the data generating process if you find a series of points that are falling further out from what you would normally expect. Only if you are absolutely sure that these data points are erroneous should they be excluded.

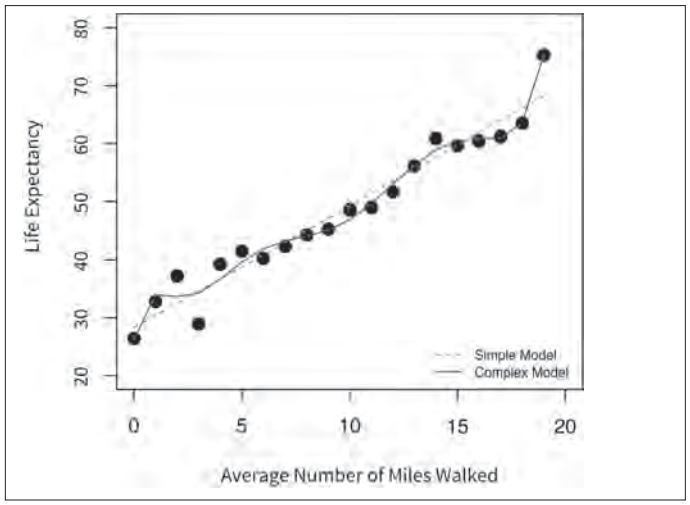
When Does Overfitting Occur? Asking Too Much From the Data

Generally, if a simpler model produces improved predictions over your more complex model, you've overfitted the data. Babyak (2004) provides an intuitive explanation of this phenomena: "Given a certain number of observations in a dataset, there is an upper limit to the complexity of the model that can be derived with any acceptable degree of uncertainty." An example with simulated data is provided below to illustrate overfitting due to model complexity.

Twenty data points are drawn from the same distribution as defined by the author. In this simulated example, the x-axis represents the average number of miles walked a week, and the y-axis represents life expectancy. The goal of this exercise is to estimate two models that relate life expectancy as a function of the weekly number of miles walked, and assess which model has improved predictive accuracy.

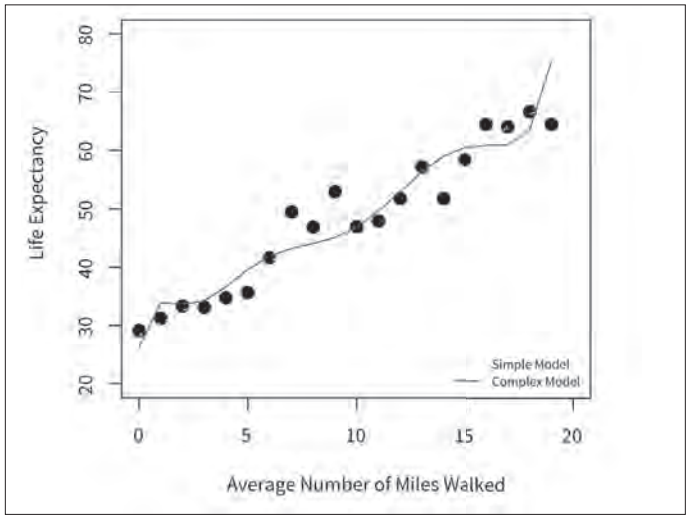
The two models are plotted in Figure 1. The simple model (Figure 1: dashed line) is estimated on the 20 data points and the formula is as follows: $Y = \beta_0 + \beta_1 X + \varepsilon$. A more complex model ($Y = \beta 0 + Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_8 X^8 + \varepsilon$) is estimated on the same points and is represented with a solid line in Figure 1.

Figure 1
Simulated Data



Visual inspection of both candidate models suggests that the complex model does a better job of fitting the sample dataset, and indeed it does. The simple model has a mean squared error of 8.45 compared to the complex model which has an MSE of 3.27. However, as Mosteller and Tukey (1977) state: "Testing the procedure on the data that gave it birth is almost certain to overestimate performance." Indeed, Figure 2 shows both models estimated once more on a new set of 20 data points generated from the same distribution to measure the out-of-sample performance of these models.

Figure 2
New Simulated Data



Now, visual inspection of both models on the new dataset paints a different picture. The complex model no longer appears to predict the data points as well as it did on the previous set of data. Comparing the mean squared errors of both models confirms this point. The simple model's MSE is 8.86 compared to the complex model's MSE of 17.76. This example illustrates two important points. First, it affirms earlier statements that model complexity is restricted by the sample size, and second, it is essential that candidate models are chosen based on out-of-sample performance, and not using the same dataset that was used to build the models. While outside the scope of this paper, there are a variety of statistical techniques that allow you to estimate the out-of-sample performance of your models without the need to gather more data. Some of those techniques include cross-validation, AIC/BIC[1], and bootstrapping.

CONCLUSION

Advancements in computing power allow analysts to quickly manipulate data and build models on small and large datasets alike to answer important business questions. However, with the increased number of choices available to analysts, comes greater exposure to build models that overfit the data. Consider making research design decisions a priori, and examine the number of observations you have available to avoid building overly complex models that your dataset cannot support. ■

Rosmery Cruz is a data scientist at RGA Reinsurance Company, in Chesterfield, Mo. She can be reached at *Rosmery.Cruz@rgare.com.*

ENDNOTE

1 AIC (Akaike information criterion) and BIC (Bayesian information criterion) estimates take the log likelihood and apply a penalty to it for the number of parameters being estimated. The specific penalties are explained for AIC by Akaike in his papers starting in 1974. BIC was selected by Gideon Schwarz in his 1978 paper and is motivated by a Bayesian argument.

REFERENCES

Babyak, M.A. 2004. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine, http://journals.lww.com/psychosomaticmedicine/Abstract/2004/05000/What_You_See_May_Not_Be_What_You_Get__A_Brief,.21.aspx.*

Gelman, Andrew, and Eric Loken. 2013. The garden of forking paths: why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. In technical report, Department of Statistics, Columbia University. Retrieved Aug. 17, 2017 from *http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf*

Simmons, Joseph P., Leif D.Nelson, and Uri Simonsohn. 2011. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. Psychological Science. *http://journals.sagepub.com/doi/pdf/10.1177/0956797611417632.*