White Paper

Charlie Batch and the Cost of Obfuscation

by Andrew Lutes

Data Scientist

March 2015

ELDER RESEARCH

DATA SCIENCE & PREDICTIVE ANALYTICS

The hot Florida sun shone down on fans, coaches, and players alike as Charlie Batch took the snap and kneeled down to seal a commanding 38-13 win for the Pittsburg Steelers. He may not know it, but with his performance in that 2010 game, the 35 year old NFL quarterback tied an all-time record for ESPN's total quarterback rating (QBR). Despite 2 interceptions in only 18 plays, but with 12 completed passes for 187 yards and 3 touchdowns, Batch somehow tied two other performances for the highest single game rating ever achieved by an NFL quarterback (a perfect 99.9). The goal of QBR is to distill the myriad facets of a quarterback's performance down to one number. QBR measures a quarterback's contribution to points through passing, running, penalties, and sacks—incorporating details such as distance a ball travels through the air on a pass, and even weighting for "clutch" game situations. However, the gulf, in this game, between Batch's perfect QBR and his pedestrian stats caused fans and analysts to openly question ESPN's metric. Many said that ESPN overcomplicated the analysis of a quarterback by using measurements to which most fans don't have access. Others said that ESPN hides behind proprietary algorithms to give their system the aura of mystery, and that it amounts to a bunch of "hog-wash".

Clearly, opacity can hinder acceptance, and therefore deployment, of an analytical model. Opaque models are liable to be misconstrued or misapplied, if they are even adopted at all by a potentially skeptical audience. Often, these difficulties have been a result of keeping analytic methods behind closed doors. However, with the great Cambrian Explosion of elaborate data science algorithms in the last few decades, complexity has become the new opacity. In a forthcoming article, I'll tackle the broader costs of complexity along with some heuristics for deciding how good is good enough. Today, I'll focus on a few questions every data scientist should keep in mind at the outset of each project - to design a model based on how it will be adopted and consumed.

Interpretability and Application

All models have limits. As a data scientist, it's important to know how wrong your model will be and why. This goes beyond accuracy and prediction intervals; it involves analyzing any systemic biases in the model or weakness in the underlying assumptions. Under what conditions does this model do well or poorly? What exactly are we measuring and what are we unable to measure?

The degree to which we care about these questions is based on how this model is going to be deployed. Assuming a model has been adopted by decision makers, then pushing its prediction to an automated system which will take an action then its inner workings don't need to be widely understood; the machine doesn't care about interpretation. However, if we are going to pass a set of predictions to a human decision maker, then we need to be clear about what exactly our model does and doesn't mean. That way, the human can mix in the metric with qualitative information about the situation and make a more informed decision. More notably, the human can ensure that we aren't asking our model a question that it isn't equipped to answer.

Back to ESPN's total QBR. This metric was made for consumption by sports analysts, coaches, fans, etc., but its lack of interpretability doomed it to criticism. We can't exactly see what is driving Batch's top rating. ESPN claims that Batch added 5.1 "expected points" (that model is the crux of the metric) through passing, but we can't see how his passing created those points. Was it the yards weighted by where he was on the field? Was it his third down conversions? How much was he penalized for the interceptions? If we knew what the model drivers were, we could begin to understand why a generally average quarterback performed so well that day -- and a coach could decide under what situations he should start again. A fan also can't see the model's shortcomings: QBR apparently doesn't control for how good a defense each quarterback was facing or how much help the quarterback had from good pass blocking and route running. Not to mention that in a game like Charlie Batch's with only 18 active plays (observations), our measurement isn't going to be very precise. Ratios can easily be extreme with small numbers of cases.

If fans knew the innards of the metric, the most interested could validate it. People often focus on specific cases where they can intuitively judge the output of the model. When fans saw Charlie Batch that day, they saw a mediocre quarterback put up an average looking game, which was contradicted by the perfect QBR. Often, such a surprise is exactly what we want data science to do—redefine and improve knowledge about the world with ideas more supported in evidence. However, QBR would need to be shown to be predictive of future results to overthrow our intuition. Without such evidence, the conflict is sign of a weakness of the metric.

Interpretability depends on how straightforward your model is as well as your ability to communicate the driving forces behind it. Before cloaking your model in a shroud of complexity, ask yourself two important questions: Who will need to apply the result and who will need to accept it? If you spend time at the front end of a project examining how the results could be misconstrued, misapplied, or misunderstood you can often meet interpretability challenges before they arise.

Implementing a model may be routine for Data Scientists, but relying on it in production can be a radical and scary step for our clients. We have to be aware of their state of mind

and gently convince them to adopt this new (to them) approach that seems like a big change. Our ability to influence actions depends on **truth** (solid, objective scientific experiments) but also on our **sensitivity** to the hidden fears and doubts of the other stakeholders. Interpretable models are a great help in obtaining model adoption, without which, all of our great science is wasted.

About the Author

Andrew Lutes is a data scientist for Elder Research Inc. He enjoys using analytic techniques to find order and meaning within noisy and complex systems. Andrew's technical focuses include causal analysis, experimental design, and predictive analytics. Andrew has applied analytics to support agencies within the Department of Homeland Security & Department of Labor, to aid corporate investigations, to inform strategic decision making for growing start-ups, and (in his spare time) to assess decision making in sports. Andrew earned a B.S. in Systems Engineering and a B.A. in Economics from the University of Virginia. Currently, he is pursuing an M.S. in Mathematics and Statistics from Georgetown University.

www.elderresearch.com

ELDER RESEARCH
DATA SCIENCE & PREDICTIVE ANALYTICS

National Capital Region
2101 Wilson Boulevard
Suite 900
Arlington, VA 22201

855.973.7673

Headquarters
300 W. Main Street
Suite 301
Charlottesville, VA 22903

434.973.7673

Maryland Office
839 Elkridge Landing
Suite 215
Linthicum, MD 21090

855.973.7673