# WHAT'S IN THE MODEL?

**By: Pete Barwis, Ph.D.**
*Senior Data Scientist, Capture Higher Ed*

## "What's in the model?"

This is a question we receive a lot. The question is asked in the spirit of wanting to know whether a predictive model can be trusted. It makes sense when approaching a predictive model as being a causal explanation of some behavior. Let's try to understand why we're asking this question in the first place, and whether it is the right question to be asking. Dr. Paul Allison nicely sums up the difference between causal and predictive models in the context of the dominant modeling approach, multiple regression, on which he literally wrote the book (Multiple Regression, 1999):

> *"There are two main uses of multiple regression: prediction and causal analysis. In a prediction study, the goal is to develop a formula for making predictions about the dependent variable, based on the observed values of the independent variables ... In a causal analysis, the independent variables are regarded as causes of the dependent variable. The aim of the study is to determine whether a particular independent variable really affects the dependent variable, and to estimate the magnitude of that effect, if any."*

### Causal vs. Predictive Models

In a causal model, predictive variables are hand-selected based on their theoretical contribution to actually causing the outcome to occur. For instance, visiting campus needs to be included in a causal model that predicts enrollment, because visiting campus has a considerable, psychological effect on students that increases their likelihood of enrolling.

In causal models, variables that may not turn out to be statistically significant predictors are included anyway, because they serve as controls, or because they are theoretically important measures that contribute to a larger causal explanation of a certain outcome. This is the dominant modeling paradigm in enrollment management, and it's where the question "what's in the model?" comes from. In a causal regression model, you get a coefficient size, the direction, and a p-value for each named input. Seeing the size, direction, and strength of those coefficients gives people who assess those models a sense of how trustworthy they are. Unfortunately, just because a model appears to make good theoretical sense does not mean that it is any good at actually making predictions.

As Dr. Paul Allison suggests, when specifying a causal model, the actual predictiveness of the model is not of paramount importance. Counter-intuitively, in a causal model, what is most important is determining the proper effect size for each independent variable:

> *"Even with a low $R^2$, you can do a good job of testing hypotheses about the effects of the variables of interest. That's because, for parameter estimation and hypothesis testing, a low $R^2$ can be counterbalanced by a large sample size … For predictive modeling, on the other hand, maximization of $R^2$ is crucial. Technically, the more important criterion is the standard error of prediction, which depends both on the $R^2$ and the variance of y in the population. In any case, large sample sizes cannot compensate for models that are lacking in predictive power."*

Rather than asking, "What's the $R^2$ (goodness of fit) of the model?" many connoisseurs of enrollment models tend to ask, "What's in the model?" The focus is on the inputs and their coefficients' size, significance, and direction, rather than on the outputs and their performance. This would be great if we were interested in hypothesis testing, but since we want to actually use and depend on our predictions in real life, we are arguably asking the wrong question.

## Inputs vs. Outputs

We wish it were as simple as only needing a single model to both explain and predict an outcome. But models are assembled and assessed differently depending on which use you prioritize. At Capture, we prioritize the predictive utility of our models, at the expense of not being able to point to specific model coefficients and exclaim with confidence that each one accurately identifies a true causal effect. Instead, we focus on how accurately a machine-learning model can produce predictions on data that were not used to train the model.

The fundamental difference between how Capture builds models and how most other enrollment management companies build models, is that *we determine the trustworthiness and utility of a model based on its outputs, rather than on its inputs*. Concentrating on inputs means that the consumer of the model needs to understand the full complexity of the model architecture, not simply which features are included. This is straightforward when the predictive algorithm is one logistic regression model with hand-selected features. It is much more difficult when the model is an ensemble of many different "base classifier" algorithms, each of which use different feature selection methods and dimensionality reduction techniques, which are then all stacked together to make a one set of predictions.

## But Really, What's in the Model?

The answer to the question "what's in my model?" is consistent at Capture. What is included is every piece of historical information that is provided to Capture is measured reliably over time and contributes to a more predictive model, without overfitting. That includes additional contextual and behavioral data (if available) that is collected by Capture. Information that:

- is unstable over time;
- does not increase predictive accuracy;
- results in overfitting;
- is redundant;
- is removed or "regularized," in some fashion, from the model.

Pieces of information that don't meet the aforementioned criteria but that you might think are causally related to an enrollment outcome, might not end up being included as inputs. Here are some reasons why they might not be included as individual features in a model:

1. Capture uses different tools to reduce the size of the inputs going into a model. These tools include PCA (Principal Components Analysis), which doesn't remove features or exclude them at all, but rather maps all of the raw data to a lower dimensional space.
2. Some features are just not nearly as predictive as assumed. This is by far the most typical reason a feature that seems theoretically useful doesn't make it through feature selection.
3. With some features, we prioritize reliability over sheer predictive power. Even with regularization techniques, the effect of select highly predictive features can be so large that practically imperceptible differences in the distributions between years results in extreme differences in predicted values.
4. Some features are not collected reliably over time, or their measurement changes over time in a way that compromises model performance.
5. Some features have high amounts of missing data or have very small amounts of variance, either of which can negatively impact the quality and reliability of predictions.
6. Some features are highly similar to other features, so much so that multi-collinearity makes it difficult or impossible for a model to attribute responsibility to one feature or the other. This is more commonly a problem in causal models, but in extreme cases, this can be a problem even in models that are optimized for making predictions.
7. Features that have a theoretical linear association with an outcome might be passed over by a feature selection algorithm that identifies and capitalizes on more predictive non-linearities provided by other features.

**This is the product and service Capture offers:** we process the data and assemble the best possible machine-learning algorithm for our partners. In short, you can be confident that if a column of data is useful for making better predictions, it is included, in some form, in the model.

## Outputs: The Currency That Matters

Many other industries are more comfortable relying on predictions without having a full understanding of model inputs in order to believe and rely on the predictions. Here is a short list of some of these predictive tools:

- **Weather predictions:** The average consumer of weather predictions doesn't need to understand the enormity of the data collection, processing, individual features, and ensemble model specification that goes into generating weather forecasts in order to believe that if the forecast says it's going to rain, it's a good idea to wear a rain coat or bring an umbrella. Evidence shows that weather forecasts are indeed useful and accurate.
- **Voice recognition:** Each time an iPhone user asks Siri to play a certain song, an Amazon user asks Alexa for today's weather forecast, or a Google user says "Ok Google, what is the difference between a causal and predictive model?" predictive algorithms go to work to accomplish the requested task. Predictive models turn speech into data that computers can understand. Models convert those data into characters, and those characters into text, and that text into computer commands. An iPhone user knows if these models do their job if the song they requested begins playing. They don't need to think through the hidden layers, the number of neurons, the activation functions, the backpropagation method etc. of the deep neural networks models under the hood. In fact, voice recognition is now actually more accurate and faster than humans transcribing on keyboards.
- **Game play:** Computer predictive models are now capable of beating the most capable humans at Jeopardy, at Atari Games, at chess and even at Go.

The list of predictive model technologies that don't have an easy answer to the question "what's in the model" goes on, and I'll push it a bit further to reinforce the notion that the best predictive models that people actually use (and will come to use) in real life are not used more often, nor used with greater confidence, because they are provided along with a set of named inputs and weights. In every case, what is more interesting and useful for assessing the trustworthiness of the model are the outputs, utility, and predictiveness of each technology:

- Search engines
- Online advertising
- Facial recognition
- Recommendations for Netflix, Spotify, Amazon, etc.
- Spam filtering
- Credit card fraud detection
- Stock market trading
- Self-driving vehicle algorithms
- Chat bots
- Credit scoring
- Driving directions / route-finding
- Ridesharing
- Mobile check deposits
- Object in image identification
- Facial filters

Like the predictive methods that drive the aforementioned technologies, Capture's Envision models are built to optimize their predictive utility. We think that rather than focusing on the inputs to the model, the more illuminating and useful question is, "What amount of confidence should I put in the model outputs?"

Capture provides a full set of predictive model performance information with each model, including training and holdout set predictive accuracy, RMSE, and AUC statistics, that when understood in context, can help you determine how much you should rely on the model, and how much you should rely on your institutional expertise.