When Big Data Isn’t Enough: Solving the long-range forecasting problem in supervised learning

Joseph L. Breeden, PhD
CEO
breeden@prescientmodels.com

Eugenia Leonova
Senior Scientist
eugenia.leonova@prescientmodels.com

Prescient Models © 2019
Big Data is enabling advances in AI and Machine Learning
But we only have Big Data for a short slice of history
Critical information is available only in long, thin data
We need different types of models for these different types of data
If we start with a model of the long, thin data, then the AI / DM model can learn around that.
Ranking Risk
• Joint research with Casey Foltz, Oregon Community Credit Union

• Sought to predict credit card attrition using data from January 1st, 2014 through December 31st, 2017. Data on fees, account errors, and behavioral factors were tested.

• Tested 23 different scoring methods in R.
Comparing ROC Curves for Different Methods

Best Models

Results similar to Logistic Regression
Comparing ROC Curves for Different Methods

Discriminant Analysis

Classification Trees
By studying the best performing methods, we learned that their biggest gains were coming from optimal binning of the inputs – essentially linearizing the inputs.

The rest of the gain is coming from interactions between variables and validation arbitrage – less visibility for validators means more flexibility for the algorithm to operate without p-value tests.

Optimizing discrimination does not account for systematic shifts due to the economy. We need a different kind of solution for that.
Different Perspectives on Predictive Analytics

- In-sample fit and bias
- Rank ordering and out-of-sample testing
- Out-of-time backtesting
- Sensitivity testing

Optimizing rank ordering is different from capturing long-term trends. ML algorithms currently optimize discrimination over short time periods, not long range forecasting.
Blue line: Credit cycle as measured with an Age-Period-Cohort (Vintage) model applied to US mortgage performance.


Gray bars: Economic cycle as measured by change in Real GDP. Credit cycle leads economic cycle by an avg of 17 months.
The first model can be any kind of prior knowledge that we want to hold as a fixed driver of the result and learn around.

This is not a Bayesian prior, because the AI / machine learning (ML) model will not have data sufficient to revise it. The Given Knowledge is a fixed input.
Example from Consumer Lending
Experience has shown that Vintage models like Age-Period-Cohort models are very effective at making long range predictions at the total portfolio level.
Vintage-aggregate data is created for default rate, attrition rate, outstanding balance, and default balance ratio.
Each key rate is decomposed into a lifecycle versus age of the loan, credit quality by vintage, and environmental impacts by calendar date.

\[
\log \left( \frac{PD}{1 - PD} \right) = F(a) + G(v) + H(t)
\]

The environment is correlated to macroeconomic factors so that economic scenarios may be used to predict each vintage.
Each key rate is decomposed into a lifecycle versus age of the loan, credit quality by vintage, and environmental impacts by calendar date.

\[
\log \left( \frac{PD}{1 - PD} \right) = F(a) + G(v) + H(t)
\]

We want to replace the vintage function with an advance AI / DM model.
NN + APC Model Example

Build an APC model on vintage data.

Use APC lifecycle and environment as fixed inputs to a NN model built with all available behavioral factors.

Create forecasts with extrapolations of the APC lifecycle and environment functions.
We built a neural network with APC lifecycle and environment as fixed inputs.

In-sample the results improved because of the adjustment for age and environment.

(Not a fair test, because it was new information.)

Out-of-sample the results were significantly better in the long run because of the ability to steer the economic environment.

(Again, not a fair test, but it is a new capability.)
We should be able to make such combinations with any technique that allows for a fixed input.

- Adaboost
- Gradient Boosting
The use of fixed inputs from a model with long-term predictability is a natural fit for iterative tree-based methods that start from a base learner. We can make a vintage model the base learner and build the trees on top.

AdaBoost [T.Hastie, R.Tibshirani, J.Friedman, 2001] starts with a base learner, computes residuals, and then iteratively computes data weightings to compute successive refinements.

1. Initialize the observation weights $w_i = 1/N, \ i = 1, 2, \ldots, N$.

2. For $m = 1$ to $M$:
   (a) Fit a classifier $G_m(x)$ to the training data using weights $w_i$.
   (b) Compute
       $$\text{err}_m = \frac{\sum_{i=1}^{N} w_i I(y_i \neq G_m(x_i))}{\sum_{i=1}^{N} w_i}.$$  
   (c) Compute $\alpha_m = \log((1 - \text{err}_m)/\text{err}_m)$.
   (d) Set $w_i \leftarrow w_i \cdot \exp[\alpha_m \cdot I(y_i \neq G_m(x_i))], \ i = 1, 2, \ldots, N$.

3. Output $G(x) = \text{sign} \left[ \sum_{m=1}^{M} \alpha_m G_m(x) \right]$. 
Gradient Boosting [Friedman, 1999] starts with an initial model, computes pseudo-residuals, and iteratively creates refinements.

Boosting represents the final model as a sum of models, \( F(x) = \sum_{m=0}^{M} \beta_m h(x; a_m) \).

In our case, start with an initial model, \( F_0 \) as the lifecycle and environment from APC with \( \beta_0 = 1 \).

Begin the iterations by choosing a model type for the base learners, \( h(x; a_m) \).

At each modeling iteration, pseudo-residuals are computed given a loss function \( \Psi(y_i, F(x_i)) \)

\[
y_m = -\left[ \frac{\partial \Psi(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)} \right]
\]

So the expansion coefficients and model parameters are fit

\[
(\beta_m, a_m) = \arg\min \sum_{i=1}^{N} \Psi(y_i, F_{m-1}(x_i)) + \beta h(x_i, a))
\]

\[
F_m(x) = F_{m-1}(x) + \beta_m h(x_i, a_m)
\]
The given knowledge can come from any kind of model, not just APC

- The longest data sets often support only time series models

The AI / DM model can be of any type, so long as

- We can find a way to adjust the model to accept a fixed input
- The estimation function is statistically consistent with the given knowledge
The Big Picture

Combining long, thin data with short, big data is a common problem.

It is really a problem of combining different model types, and this approach will always work.
In many application areas, knowledge exists that cannot easily be converted to model coefficients or training data.

The Given Knowledge could be expert knowledge based upon decades of experience, like parents teaching children.

This may be the key ingredient in moving from Artificial Intelligence to Artificial Wisdom.
For training or consulting on these methods, contact:

Joseph L. Breeden, CEO
Prescient Models LLC
1600 Lena St., Suite E3
Santa Fe, NM 87505
+1-505-670-7670
breeden@prescientmodels.com