PRISONERS of the DATA

The Mortgage Industry and Predictive Modeling

By Jonathan Glowacki and Eric Wunder

THE GROWTH IN COMPUTING POWER AND DATA STORAGE capabilities over the past decade has led to the widespread adoption of predictive modeling—the use of historical data, statistics, and other mathematical techniques to estimate and explain past and future events. No longer confined to a few industries, predictive analytics has proven its usefulness in a variety of applications, including determining the effectiveness of new medications and medical treatments, assessing credit risk, and estimating how consumers may react to sales and other marketing efforts.

While the rapid expansion of predictive analytics has led to increased confidence in decision-making, it also has introduced some new risks. A key assumption in predictive analytics, for instance, is that all the applicable information necessary for modeling an event of interest is contained in the historical data. But if the data used in the analysis aren't representative, there's a chance that the results can lead to decisions that do more harm than good (the "prisoner of the data" axiom). The recent mortgage crisis is a textbook example.

Mortgage Data

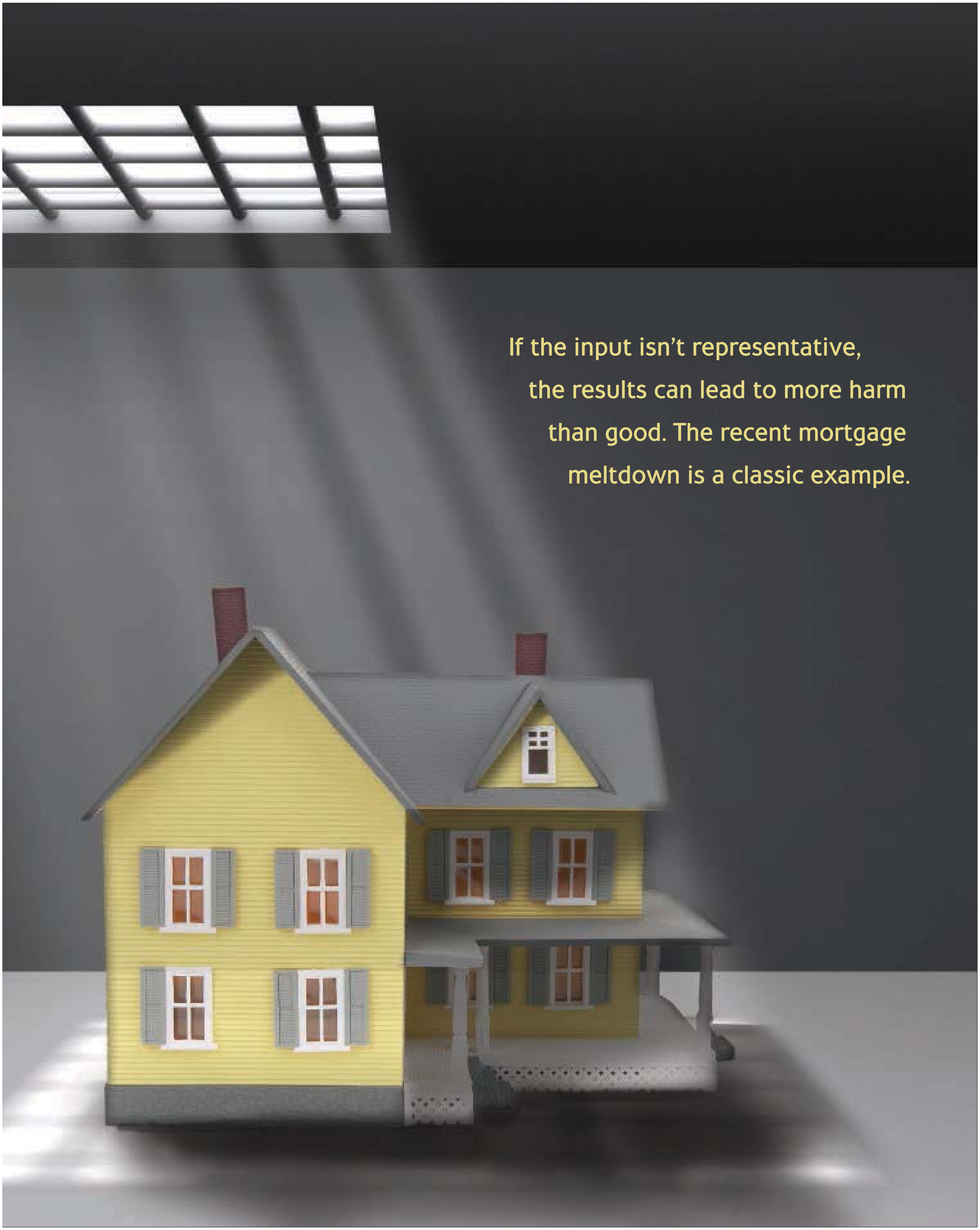
In the mortgage industry, predictive analytics currently is being used to develop underwriting guidelines, estimate losses on legacy business, allocate servicing resources, and calculate capital requirements. It's not uncommon for mortgage originators and investors to purchase third-party software rather than develop the models in-house.

Prior to the 2007-2008 mortgage crisis, most third-party models centered on estimating when a mortgage would pay off through refinance rather than when it might default. The biggest perceived risk to investors was getting their money back before they had a chance to earn interest on the mortgage and, possibly, needing to reinvest that money at a lower yield. Default rates were at low levels, and recent significant annual appreciation in home prices generally meant that even if a loan did default, the loss severity was minimal.

As we now know, the default risk was significantly understated. One of the major contributors to this underestimation of risk was limitations in the data. Even if a modeler was able to accurately forecast the depth of the crisis in terms of the decline in home prices, models that lacked a catastrophic event in the development data still underestimated significantly the losses in a mortgage portfolio.

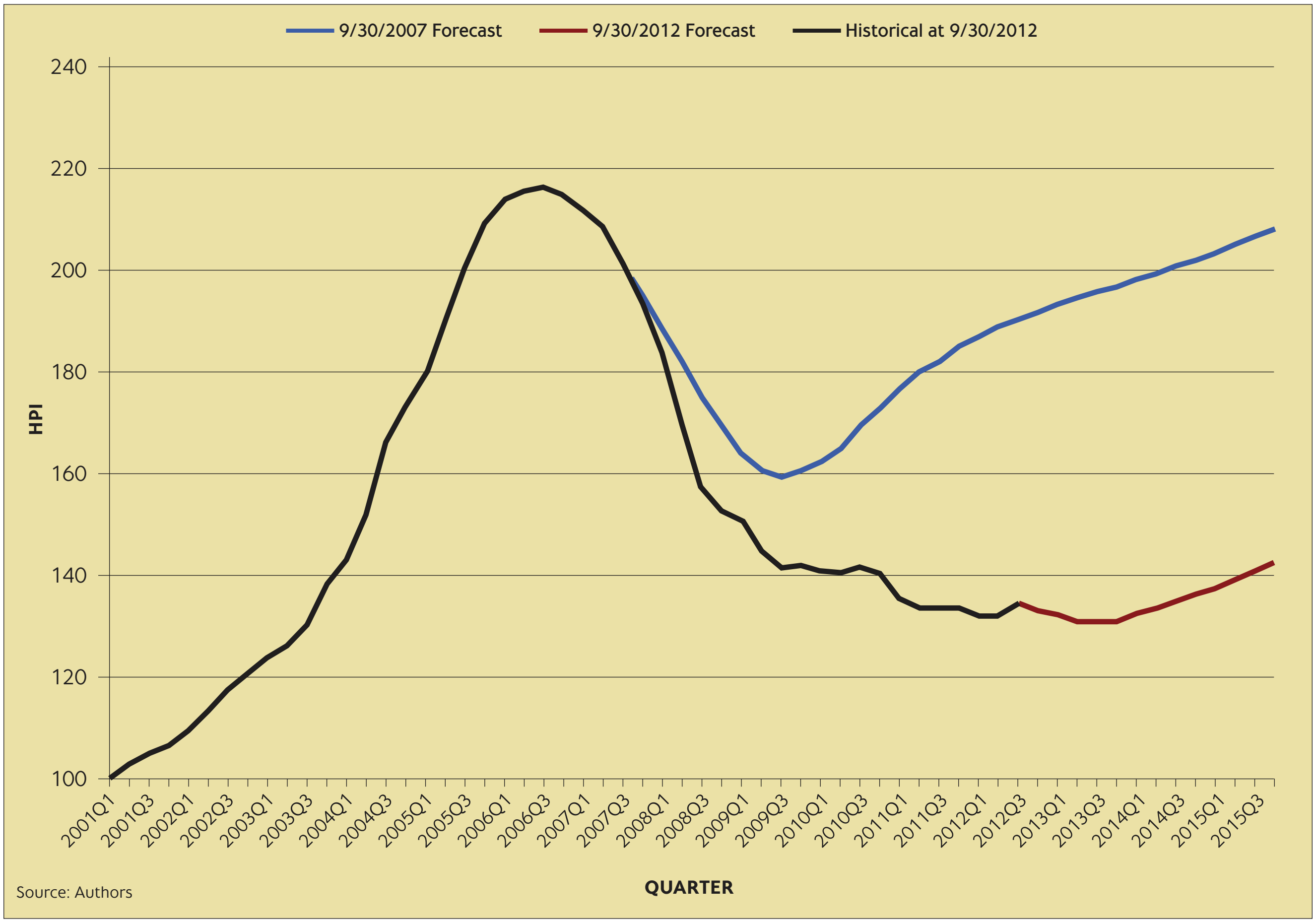
Third-party models for estimating the performance of mortgage collateral often were developed using loan origination and mortgage performance data that only began in 1990. Comprehensive mortgage data prior to 1990, including information about housing losses due to the collapse of oil prices in the 1980s, were sparse and not shared publicly.

SHUTTERSTOCK



If the input isn't representative, the results can lead to more harm than good. The recent mortgage meltdown is a classic example.

FIGURE 1 Moody's FHFA/OFHEA house price index comparison of California forecasts as of Sept. 30, 2007, and Sept. 30, 2012, reindexed to 2001 Q1 = 100



Source: Authors

Although the period from 1990 through the mid-2000s contains millions of records and was thought to be robust by market participants, it covers only a limited amount of data relating to severe stress periods in the mortgage market—and those are concentrated in specific regions rather than representative of national housing market stress. Furthermore, "subprime" and "Alt A" mortgages that collateralized private-label mortgage securities (those not guaranteed by the government) didn't become common until the mid-2000s, and the amount of exposure on these loans, although they were known to be riskier, was poorly understood because of the lack of historical data.

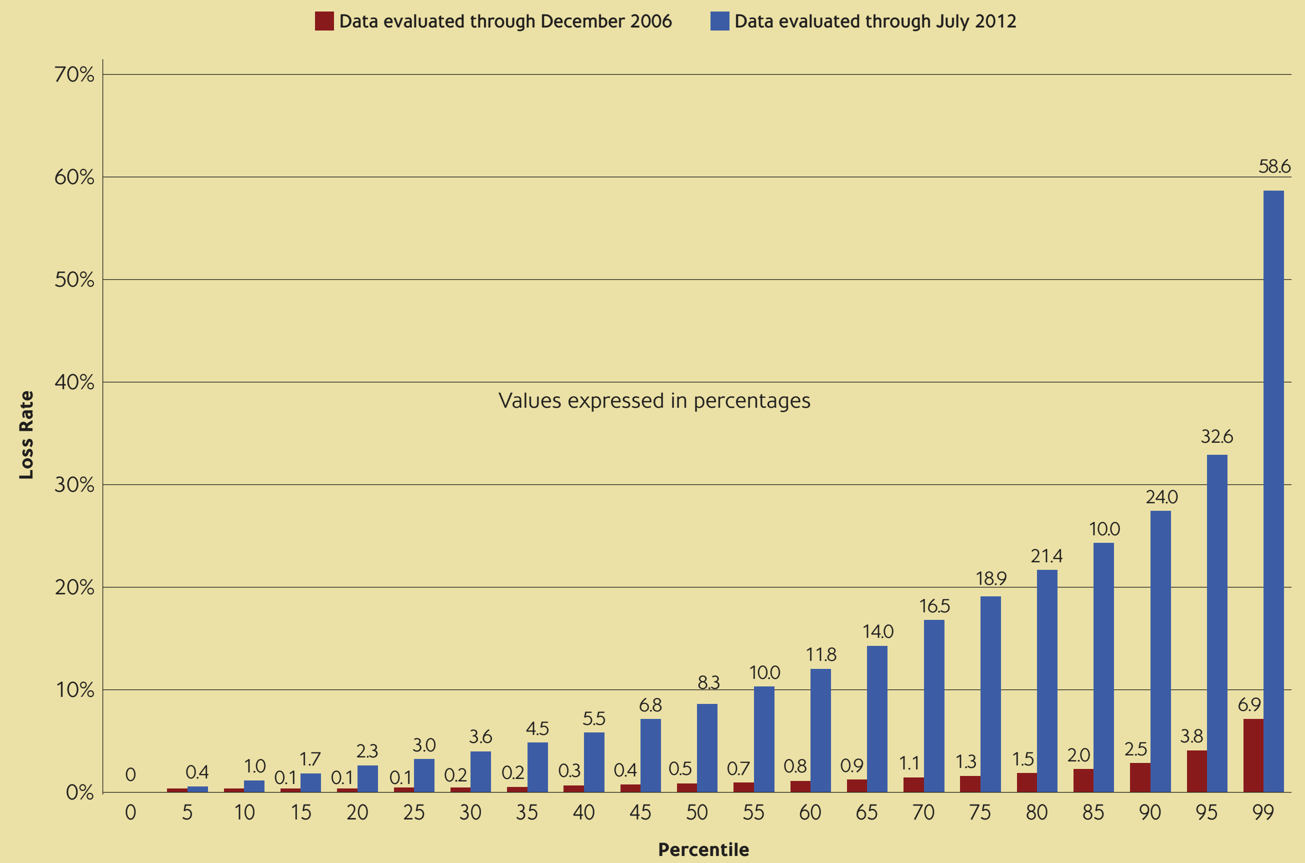
As a result, the mortgage industry (including third-party modelers and rating agencies) was forced to make assumptions about both the ultimate performance of loans during stressful periods and about loans with exotic characteristics. Mortgage models generally included "adjustment factors" or "default multiples" for economic stress or exotic product types. The adjustment factors ranged from 1.05 to perhaps several multiples depending on the scenario or product feature. Liar loans, loans without verification of income or assets, were generally given an adjustment factor between 1.25 and 1.75.

Home Price Indexes

Home price indexes (a geographic map of home prices over time) typically are tied to some base period with a value of 100. The index value for 1980, for example, may be set at 100 with the current value equaling 200. This would indicate that homes valued at $100,000 in 1980 would be worth, on average, $200,000 today. Home price indexes and forecasts of future home prices can be used, among other things, to estimate the current and forecast equity position of a borrower and as a gauge of the housing market sentiment. Indexes forecasting average home prices are an integral part of most mortgage default models today. But before the mortgage crisis, home price indexes weren't commonly used as an indicator of future default performance because the majority of data at the time included consistent home price appreciation. There were little publicly available data that could be used to model the relationship between home price declines and the performance of mortgage collateral.

By 2007, with the housing crisis imminent and home prices already beginning to plunge in value, mortgage models had two major issues to combat with respect to home price indexes:

FIGURE 2 **Actual loss rate by percentile for all 4,175 deals available at both evaluation dates, data as of December 2006 and July 2012**



Source: CoreLogic Securities data

- The error in economic forecasts regarding the depths of home price depreciation (particularly in the base case scenario);
- The scarcity of historical home price decline data and its effect on default rates.

As a result, the extent of the decline in housing value was underestimated by most home price experts relative to what we now know with the benefit of hindsight. The chart in Figure 1 compares two forecasts of California home price indexes at different valuation dates. (We selected California because of the large concentration of loans that originated there during the housing boom. Many other states and large cities show a similar trajectory.) The Sept. 30, 2007, forecast was produced by Moody's Analytics using actual data through that date, adjusted to reflect changes made by the Federal Housing Finance Agency/Office of Federal Housing Enterprise Oversight (FHFA/OFHEA) to the Homeowners Protection Act of 1998, as seen in the Sept. 30, 2012, forecast. The Sept. 30, 2012, forecast was created using actual data through that date. The 2007 forecast predicted a steep decline in California house prices (approximately 26 percent). The actual decline has been quite a bit worse: approximately 38 percent. Although the 2007 forecast did anticipate significant declines in home prices, it underestimated the ultimate severity of the declines.

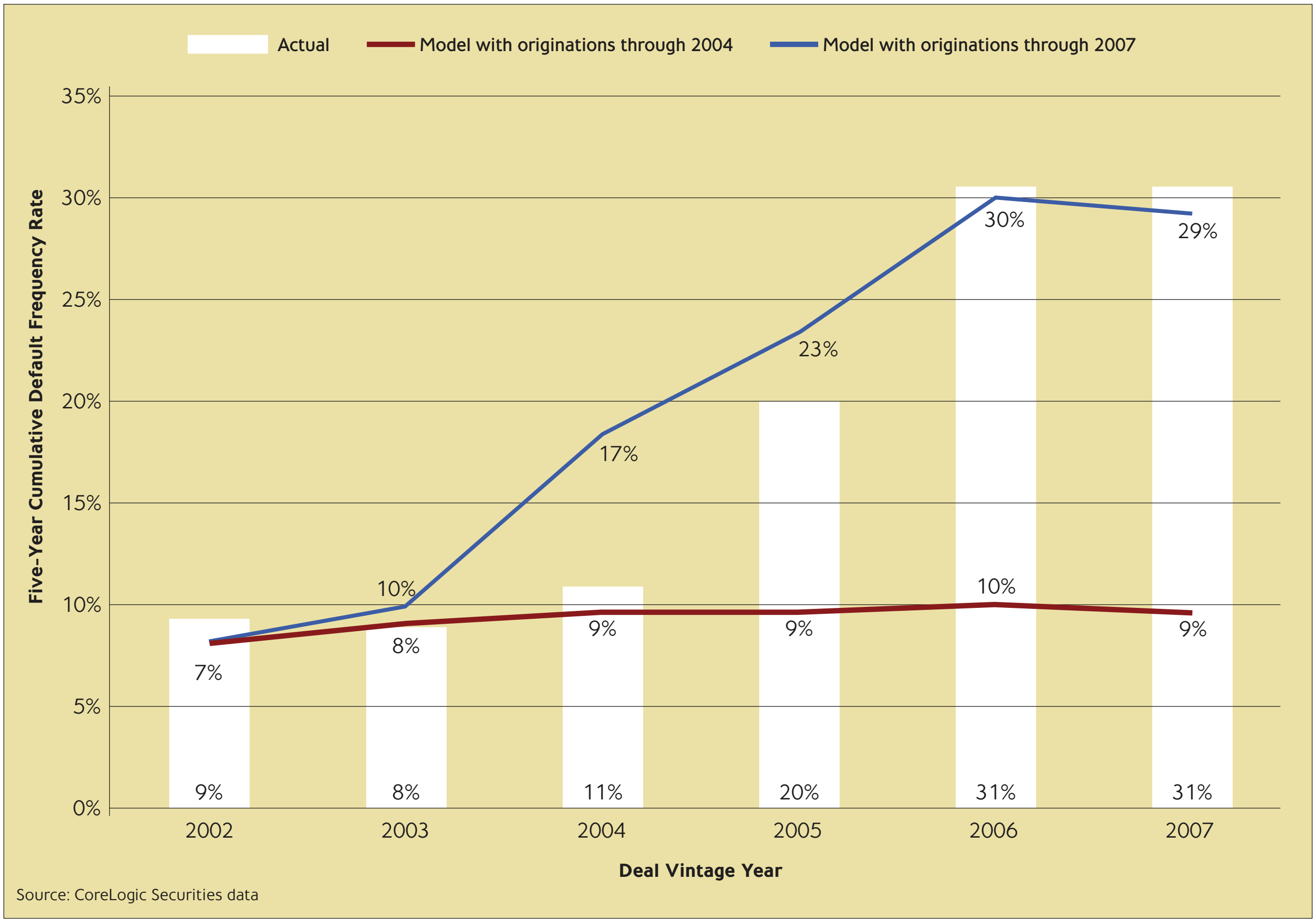
Comparing Data

To demonstrate the influence of the time period and data in developing a mortgage model, we first reviewed the empirical loss data of mortgage loans underlying private-label residential mortgage-backed securities (RMBS) at two different times: one excluding the recent stress period and one including it. Specifically, we examined the loss data on private-label RMBS by vintage year of the securities using data from CoreLogic evaluated from January 1990 through two distinct dates: December 2006 and July 2012.

The data contained more than 4,000 unique deals with some 17 million mortgages underlying the deals. We defined the loss rate as the realized loss on the collateral divided by the original balance of the mortgages. The chart in Figure 2 shows the loss rate by percentile for the same universe of deals evaluated as of each date (December 2006 and July 2012).

By the end of 2006, for example, less than 1 percent of all RMBS deals in the CoreLogic data had loss rates of greater than

FIGURE 3 **Five-year modeled cumulative default frequency rate comparison by deal vintage year**



Source: CoreLogic Securities data

7 percent (that is, the 99th percentile loss rate was about 7 percent). By contrast, the 99th percentile for the same pool of deals evaluated as of December 2012 was approximately 60 percent. A model fitted to the data through 2006 might have a difficult time estimating a loss rate as high as 60 percent because the data didn't include nearly enough actual loss data to result in such an estimate. The model would have to extrapolate beyond historical experience to get to a loss rate estimate that high. And generally, the more extrapolation that's needed in predictive modeling, the less accurate the model.

It's important to note many of the deals in the database were not developed as of December 2006. Because mortgage defaults and losses don't occur all at once but are incurred over many years, we would expect lower loss rates as of 2006 compared with 2012. But we wouldn't expect the 99th percentile loss rate to increase from 7 percent to 60 percent because of development alone.

Comparing Models

The same concepts also are applicable to predictive models. We developed two models to estimate the default rate for a cohort of loans based on the credit quality of the borrower, the underwriting characteristics of the loans, and the economic scenario presented to the underlying loans. In both models, the default rate is equal to the number of loans that defaulted (i.e., resulted in a loss) divided by the total number of loans. The 2004 model was developed using loans that originated between 1990 and 2004 with performance evaluated five years after loan origination. The 2007 model was developed using loans that originated between 1990 and 2007 with performance evaluated five years after loan origination. The additional origination years used in the 2007 model allow the credit crisis (our catastrophic event) to more directly affect our data rather than requiring the model to extrapolate to these points.

The economic scenario in each model is measured using the 2012 actual home price index data and forecast. This means that we assumed for each model that we estimated with perfect accuracy what the future home price paths would have been as of the model date. The reason for doing this was to remove economic forecast error from the comparison. Each model estimates the five-year cumulative default rate for a cohort of mortgages in a logistic regression framework. The explanatory variables in the model are exactly the same. The only difference is the time period used to estimate the coefficients of the model. On a loan level, the model estimates the probability of default within five years.

It's clear that relying too much on data alone poses significant risks. Professionals who utilize data analytics should be aware of the potential risks and limitations of such analysis.

Figure 3 provides the results of two regression models. The bar chart shows the actual average default rate of the mortgages by vintage year, the dotted line shows the default rate predicted by the 2004 model, and the solid line shows the default rate predicted by the 2007 model. The model that was fitted with originations through 2004 predicts elevated default rates in 2006 and 2007, but ultimately less than half of actual default rates to date for the securities represented in our data set. In other words, during the years that the riskiest mortgages were being originated with greater frequency and home prices were rapidly increasing, mortgage models would have estimated elevated default rates, but those default rates would not have sounded the alarms they should have. These results could be expected because of the lack of catastrophic data underlying the 2004 model (similar to what was seen in Figure 2).

Mitigating Data Risk

This mortgage example highlights the potential risk in modeling from data when an abnormal experience may be outside the data that are used for the analysis. We aren't trying to dissuade anyone from using predictive analytics (in fact, we are huge proponents of it). But it's important to be aware of the risks, such as potential limitations in the data, when using models and other big data applications.

For industry models built in the years before the mortgage meltdown, the data surrounding relaxed underwriting, blatantly false mortgage data, and a bubble in property prices simply weren't available. Modelers had to make certain assumptions to estimate the additional risk in the mortgages, and those assumptions often understated the underlying risk of the mortgages.

It's critical in data analytics that the data reflect the risk the modeler is trying to evaluate. If the input doesn't reflect the underlying risk, those limitations need to be recognized and flagged. This can be accomplished by gaining a broad understanding of the risk that's being modeled. In the area of mortgages, for instance, most data didn't include regional stress periods in home prices prior to 1990 that resulted in significant losses to mortgage investors. Data risk could have been mitigated by understanding the losses incurred during these periods. Aggregate loss data could have been collected from historical financial statements of lenders and mortgage insurers and compared with the losses predicted by the models.

A second means for mitigating data risk is to have an independent validation performed on the model. Industry best practice is for all models to be validated by someone who is independent of the team that developed the model. This can be performed internally or externally. If an entity relies heavily on the model (for example, a pricing model for insurance), those who are doing the model validation should have expertise in that particular area of exposure. Before the mortgage crisis, for instance, it would have been prudent to have mortgage models evaluated by experts with knowledge from the oil patch crisis of the 1980s.

A more recent example comes from the rise of micro lending on the Internet, in which small consumer-to-consumer loans are brokered via an online exchange. Any models used to evaluate these opportunities should be validated by experts with unsecured consumer loan experience, such as credit card lending.

A third way to mitigate data risk is to limit a model to the use for which it was developed. Returning to the mortgage market as an example, models developed using fully documented, conforming loan data shouldn't have been used to analyze negative amortization loans requiring no down payment or no documentation. Separate models should have been developed to evaluate those mortgages. A fully documented, conforming loan model shouldn't have been assumed to mimic a non-conforming loan model. In retrospect, this seems obvious, but at the time it would have been a difficult argument to make. Organizations that rely on predictive models need to have rules and model governance policies that can provide guidance on ways to manage such risks.

It's clear that relying too much on data alone poses significant risks. Professionals who utilize data analytics should be aware of the potential risks and limitations of such analysis. Methods to minimize data risk include:

- Understanding the risk being modeled, such as potentially researching the risk over a longer period than is available in the data;
- Ensuring that the model being used is validated independently by an expert in the risk that is being modeled;
- Creating model governance policies that specifically address data risk. □

JONATHAN GLOWACKI, a fellow of the Society of Actuaries, a member of the Academy, and a chartered enterprise risk analyst, is a consulting actuary with the Milwaukee office of Milliman. He can be reached at jonathan.glowacki@milliman.com.

ERIC WUNDER, an associate of the Casualty Actuarial Society and a member of the Academy, is an associate actuary in the Milwaukee office of Milliman. He can be reached at eric.wunder@millman.com.

This article is solely the opinion of its authors. It does not express the official policy of the American Academy of Actuaries; nor does it necessarily reflect the opinions of the Academy's individual officers, members, or staff.