Rank-Based Error Tracking for Agency MBS Prepayment Models

JIAWEI "DAVID" ZHANG

JIAWEI "DAVID" ZHANG
is a managing director and Head of Securitized Product Research at MSCI, New York, NY.
jzhang@alumni.princeton.edu or david.zhang@msci.com

KEY FINDINGS

- The proposed ranked-based error tracking provides a comprehensive and consistent measure for model performance across macroeconomic environment and across pool/loan attributes, superior to the prevailing one-dimensional error-tracking methodology.
- Two key measures are the steepness and accuracy of the ranking curve. The steepness measures quantify the model's ability differentiating across macro and pool variables. The accuracy measures quantify the model's ability forecasting the prepayment intensity.
- The rank-based methodology can be applied to any selected sample pool universe and performance periods. It can be systematically used for model review and model comparison.

ABSTRACT: *Mode testing, also called error tracking, is a key requirement for collateral behavior forecasting models in securitized product, including models of prepayment, default, response, utilization, etc. The high dimensionality and statistical noise associated with agency mortgage-backed securities prepayment behavior make error tracking a complex task. The traditional method focuses on single dimension, for example, along vintage coupon, and does not provide a clear measure of model accuracy and effectiveness.*

The new method, a rank-based error-tracking methodology, provides an efficient and comprehensive approach to measure model performance. It provides a clear definition of accuracy-which allows clear measure to compare among models. This is superior to the existing one-dimensional method, and other supplement methods (e.g., lift curve method for loans).

We discuss these issues as well as applicable statistical theory, and potential applications.

TOPICS: ***MBS and residential mortgage loans, factor-based models****

In securitized products investment and risk analysis, prepayment and default behaviors are key inputs. These forecasting models are often categorized as discrete choice models in econometrics. Model back testing, or error tracking in industry lingo, is often performed along single-risk factors, even though the numbers of risk drivers are large and risk-driver behavior is highly nonlinear and interactive. Exhibits 1 and 2 shows the most typical error-tracking method and reports in the industry.

Exhibit 1 shows a sample one-dimensional error tracking for a major vintage cohort, along the performance months, using a version of the MSCI agency prepayment model as example.

This type of error tracking for major vintage coupon cohort is the most typical method used by the mortgage-backed securities (MBS) investment market participants.

*All articles are now categorized by topics and subtopics. **View at PM-Research.com.**

Exhibit 1

A Typical One-Dimensional Error Tracking: 2012 FNCL 3.5s, Error Tracking Along Performance Month

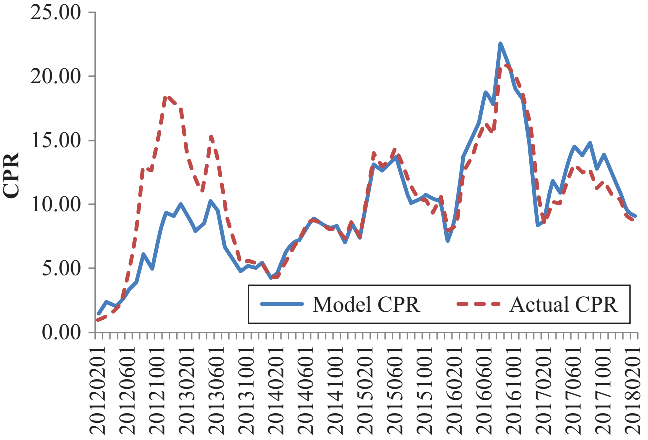


Exhibit 2 shows a sample page of a monthly error-tracking report also typical among market participants. Here, the error tracking is along the dimension of cohort type: major vintage-coupon cohorts as well as important specified pool-type cohorts.

However, this type of one-dimensional error tracking has major drawbacks.

Lack of Error Tracking over the High Dimensionality of Risk Drivers

Prepayment behavior has large numbers of risk drivers. They include macroeconomic risk drivers and loan/pool attributes. Macroeconomic risk drivers include interest rates and mortgage rates environment, house-price appreciation environment, mortgage credit policy and availability, technology advancement and business practices of servicers and originators, and consumer behavior regimes. There are also dozens of loan/pool attributes that are important indicators of prepayment behavior (e.g., loan size, WAC, age/weighted average loan age, geo, loan purpose, performance history). Investors generally own individual MBS securities that have varying attributes different from major cohorts. And prepayment behavior across varying macroeconomic risk factors are key input to MBS valuation. For these reasons, prepayment models need to be tested across this large number of risk factors. The one-dimensional error-tracking approach cannot be scaled for the high-dimensional requirement. For example, tasks of performing one-dimensional error tracking for FNCL 3.5s NY pools with loan size between $100–175k, FCO 700–720, refinance loans, and so on, would need to be repeated millions of times to cover the high-dimensional risk factor space. In addition, two issues, statistical noise and interaction between risk factors, would make this approach ineffective even when one is willing to perform and exam these large amounts of one-dimensional error tracking.

The prepayment is a doubly stochastic process. Prepayment models typically model the prepayment intensity process; prepayment itself is often a Poisson process (with the exception of curtailment, which is generally a very small part of prepayment).

The error tracking of a prepayment model, fundamentally, is to compare modeled distribution of the prepayment intensity against the true prepayment intensity, which is unknown in the prepayment data. The prepayment data are the realization of the doubly stochastic process, combination of the prepayment intensity process and the Poisson process.

We can estimate the statistical noise caused by the Poisson process in the following Equation (1):

$$\sigma = \frac{\sqrt{smm(1-smm)}}{\sqrt{N}} \tag{1}$$

where N is the number of loans in a pool, *smm* is single monthly mortality rate for the pool.

Exhibit 3 tables the statistical noise estimation for various combinations of conditional prepayment rate (CPR) and pool size, assuming $200k average loan size.

For example, for a typical pool size of USD 50 million, the statistical noise for a 10 CPR (0.87 *smm*) is 0.59 *smm*, a very high percentage of the total CPR print. Even for a USD 1 billion pool/security, the typical size of a collateralized mortgage obligation (CMO) deal, the statistical noise is substantial, on the same order as the potential error or bias of the prepayment model forecasts.

Exhibit 4 shows prepayment performance of a sample of 10,000 random agency 30-year pools with a 4% coupon and 90–100 basis point incentive between 2010 and 2018. In the scatter graph, actual prepayment speeds in the CPR are plotted against the sample model results, as a proxy for true prepayment intensity. The histograms (bar charts) show the distribution of the model and actual prepayment speeds. Even with the tight band of refinance incentives, model speeds ranged from 0 to 52 CPR because of variations in numerous risk drivers—for example, loan attributes and macroeconomic and regional economic variables. The range for the actual prepayment

Exhibit 2

Sample Page of Monthly Error Tracking Report Along Vintage, Loan Size, and Several Other Key Specified Pool Types

Issuer	Coupon	Year	Term	Spec	CBal ($B)	Factor	WAC	WALA	OLnsz	CLnsz	OLtv	Fico	Refi (%)	Tpo (%)	NOO (%)	CPR1	CPR3	CPR6	Err1	Err3	Err6
FNM	3.5	2017	360	all	169.1	0.965	4.06	7	279	274	78	755	31	49	2	4.5	4.8	5.3	0.6	0.5	0.5
FNM	3.5	2017	360	hlb	8.9	0.972	4.03	6	132	130	75	756	29	39	1	4	4	4	0.5	0.3	0.1
FNM	3.5	2017	360	hlb2	7.4	0.973	4.04	6	163	161	78	756	27	43	1	4.1	4	3.8	0.6	0.4	–0.2
FNM	3.5	2017	360	inv	0.1	0.921	4.17	13	322	315	63	766	70	51	100	4.2	7.6	7.5	–3.6	–1.3	–1.9
FNM	3.5	2017	360	jumbo	10.6	0.941	4.17	7	535	528	76	753	39	53	2	7.2	9.1	10.6	1.4	2.5	2.4
FNM	3.5	2017	360	llb	1.4	0.968	4.02	7	68	67	68	751	29	30	2	5	4.5	4.5	1	0.6	0.3
FNM	3.5	2017	360	mlb	2.8	0.97	4.03	7	99	97	70	755	32	35	2	4.2	4.4	4.4	0.3	0.5	0.3
FNM	3.5	2017	360	oLtv80	15.3	0.969	4.12	7	230	227	82	756	19	50	1	3.9	4.1	4.3	0.2	0.1	0.2
FNM	3.5	2017	360	oLtv90	0.6	0.981	4.2	7	231	229	92	726	8	80	0	3.9	3.3	2.5	1	0	–1
FNM	3.5	2017	360	oLtv95	11.5	0.977	4.08	6	299	294	95	754	2	43	0	3.9	3.6	3.7	1.2	0.8	0.6
FNM	3.5	2016	360	all	101	0.832	4.08	20	247	238	78	738	46	40	8	9	9.5	10.3	2.4	2.8	2.9
FNM	3.5	2016	360	hlb	11.2	0.885	3.97	19	131	126	77	744	39	36	6	7	7.2	7.5	0.5	0.6	0.6
FNM	3.5	2016	360	hlb2	6.4	0.876	4.04	19	162	156	79	740	42	37	7	7.8	7.8	8.3	1.3	1.3	1.4
FNM	3.5	2016	360	inv	1.2	0.896	4.08	16	316	307	65	766	62	56	100	3.3	5.4	6.8	–2	–0.1	0.9
FNM	3.5	2016	360	jumbo	3.8	0.759	4.18	19	535	519	75	742	56	58	9	12.7	14	16.3	5.8	5.7	6.5
FNM	3.5	2016	360	llb	2.3	0.881	3.96	19	68	65	71	744	37	30	7	6.6	7.4	7.7	–0.5	0.1	0
FNM	3.5	2016	360	mlb	4.3	0.883	3.96	19	99	95	73	745	39	33	7	7.7	7.8	7.9	0.8	0.8	0.5
FNM	3.5	2016	360	oLtv105	0.4	0.908	4.21	20	190	183	114	699	100	10	16	7.3	6.9	7.3	0.9	0.6	0.8
FNM	3.5	2016	360	oLtv125	0.2	0.93	4.25	20	171	165	146	700	100	10	19	7.7	6.2	5.9	1.7	0.3	–0.1
FNM	3.5	2016	360	oLtv80	19.9	0.823	4.15	21	258	248	82	734	39	48	7	9.9	10.4	11	3.1	3.3	3.2
FNM	3.5	2016	360	oLtv90	1.4	0.902	4.21	19	202	196	92	727	13	85	1	4.5	6.1	6.4	–1.1	0.5	0.4
FNM	3.5	2016	360	oLtv95	4.6	0.891	4.1	19	278	269	95	740	4	46	0	8.9	8.9	8.4	3.7	3.5	2.5
FNM	3.5	2015	360	all	146.6	0.689	4.1	31	258	243	78	751	42	39	6	10	10.7	11.9	1	0.8	0.9
FNM	3.5	2015	360	hlb	10.1	0.775	4.05	31	131	123	77	754	36	37	5	8.9	9.3	10.1	0.4	0.5	0.5
FNM	3.5	2015	360	hlb2	7.5	0.765	4.06	32	162	152	78	753	38	35	5	8.3	9.2	10.5	–0.1	0.5	0.9
FNM	3.5	2015	360	inv	0.1	0.707	4.22	33	304	286	65	774	60	41	100	0.4	8	9.8	–7.9	–1.1	–0.4
FNM	3.5	2015	360	jumbo	3	0.547	4.15	31	527	499	73	751	51	48	4	9.2	10.8	14.2	–1	–1.1	0.3
FNM	3.5	2015	360	llb	1.9	0.785	4.05	31	68	64	70	752	35	33	7	7.3	7.9	9.1	–2.3	–2	–1.6
FNM	3.5	2015	360	mlb	3.6	0.786	4.04	31	98	92	72	755	38	35	6	8.8	9	9.4	–0.4	–0.4	–0.8

speeds, 0–100 CPRs, was even wider, because of numerical noise. Prepayment modeling, distilling a very high-dimensional mathematical function from a large set of data amid large statistical noise, remains challenging to the modeling community after more than 20 years of effort.

Lack of Measure of Overall Model Performance or Accuracy and, as an Extension, Lack of Clear Measure of Comparing Among Different Models and Model Iterations

Given the high dimensionality of risk factors across millions of MBS pools and securities, and the high statistical noise, it is difficult to provide a summary set of statistics to measure model performance among the numerous one-dimensional error-tracking analyses. As the extension of this difficulty of summarily measuring model performance, it is also difficult to compare and rank model performance of two or more different models, or model iterations.

As a result of this shortcoming, there is generally no agreed-upon criteria to accept or reject a prepayment model among market participants and regulators. Although this problem is not common in financial modeling space, the MBS prepayment modeling community has lived with it for a long time. It is often part of the reason that market participants refer prepayment model as "partly arts, partly science." We apply new

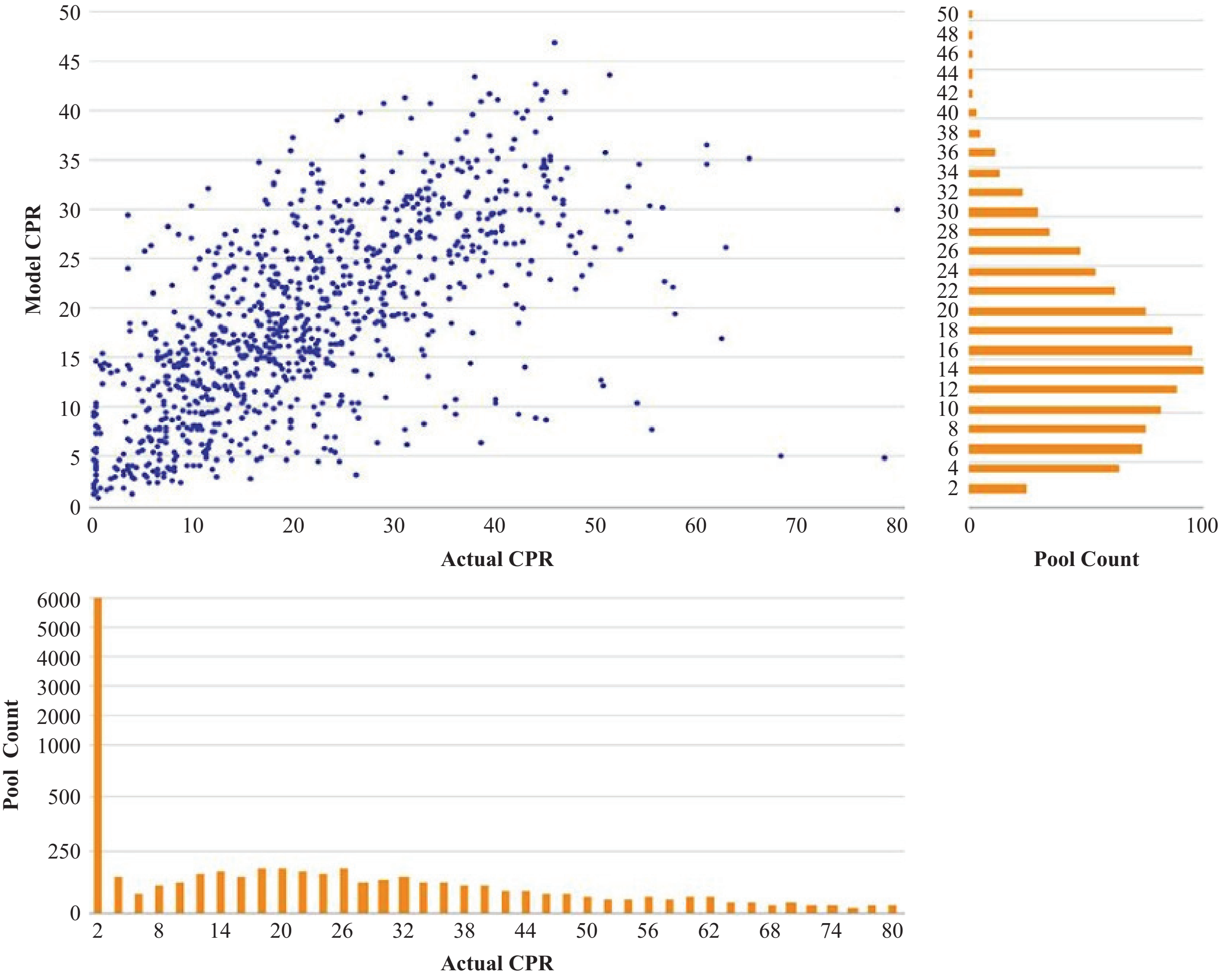
EXHIBIT 3

Measure of Statistical Noise for Various Combination of Conditional Prepayment Rate (CPR) and Pool Size, Assuming a 200k Average Loan Size

Pool Size	*smm*	0.43%	0.87%	1.84%	2.93%	5.61%	17.46%
Billion USD	**CPR**	**5**	**10**	**20**	**30**	**50**	**90**
0.05	**Statistical Noise Estimation**	0.41%	0.59%	0.85%	1.07%	1.46%	2.40%
0.1		0.29%	0.42%	0.60%	0.75%	1.03%	1.70%
0.5		0.13%	0.19%	0.27%	0.34%	0.46%	0.76%
1		0.09%	0.13%	0.19%	0.24%	0.33%	0.54%
2		0.07%	0.09%	0.13%	0.17%	0.23%	0.38%
3		0.05%	0.08%	0.11%	0.14%	0.19%	0.31%
4		0.05%	0.07%	0.10%	0.12%	0.16%	0.27%
5		0.04%	0.06%	0.09%	0.11%	0.15%	0.24%
10		0.03%	0.04%	0.06%	0.08%	0.10%	0.17%

EXHIBIT 4

Prepayment Model versus Actual Performance: Sample of 10,000 Random Agency 30-Year Pools with a 4% Coupon and 90- to 100-Basis-Point Incentive between 2010 and 2018



statistical techniques and propose the following rank-based error-tracking methodology that would move the "art" of error tracking to more firmly "science" footing.

THE RANK-BASED ERROR TRACKING METHODOLOGY

For a cohort of NP counts of prepayment forecasts/observation pairs, {P_model_i, P_actual_i, upb_i}, i = 1, 2, …, NP, the cohort can be any combination of pool/group of pools and observation month,

Rank the pairs based on model projections P_model_i into

{P_model_j, P_actual_j, upb_j}, j = 1, 2, …, NP, where

P_model_1 < P_model_2 < ⋯ < P_model_j < P_model_j + 1 < ⋯ < P_model—NP

Then create N equal size UPB buckets along this ranking dimension:

Bucket k: consist of pairs from j_k, j_k + 1, …, j_(k + 1), where

Sum(upb_j, j = j_k, j_k + 1, …, j_(k + 1)) = Total_UPB/N

The *N* is chosen so that the statistical noise, estimated by formula 1, and Exhibit 3, is below the model error range that the error tracking is targeting.

The error tracking can be performed in two ways.

Compare the Model Forecasts and Actual CPRs, Averaged Over Each Bucket

This measure shows two key model performance metrics:

1. The model's ability to differentiate over diverse risk factors is the slope of the ranking curve, including the overall slope, as well as slopes at different rankings.
2. The model's accuracy across the stratification of prepayment intensity shows whether model has error/bias when forecasting prepayment intensity at different ranking internal. The overall model accuracy measure can be summarized as weighted average absolute error

$$\text{Summary error} = \text{sum (abs [P\_model\_k} - \text{P\_actual\_k]})/N \quad (2)$$

Apply Kolmogorov-Smirnov Statistics to the Cumulative Distribution

The ranking curves described in previous section are essentially conditional probability functions using the model forecast as the distribution variable. If transformed into the cumulative density function, a Kolmogorov-Smirnov test can be performed using the maximum distance between the two conditional probability functions of model and actual, using the model forecast as the distribution variable. The KS distance can be used as an overall metrics for model accuracy performance measure.

The cohort that is applied of this analysis can take many forms, as shown in these examples:

- FNCL 30-year universe over an extensive period (Exhibit 5). This can be used as an overall metrics for model performance, model acceptance criteria, and for model comparisons.
- Certain vintage-coupon cohorts (Exhibit 6). This can be used as a measure for whether the model accurately differentiates a pool's performance between fastest to-be-announced (TBA) delivery segment and a specified pool segment.
- Cohort over various macroeconomic regimes, for example, major rates rallies and selloffs. This can be used as a measure for whether a model accurately prices the embedded call options, or whether model accurately reflects performance over stress periods.

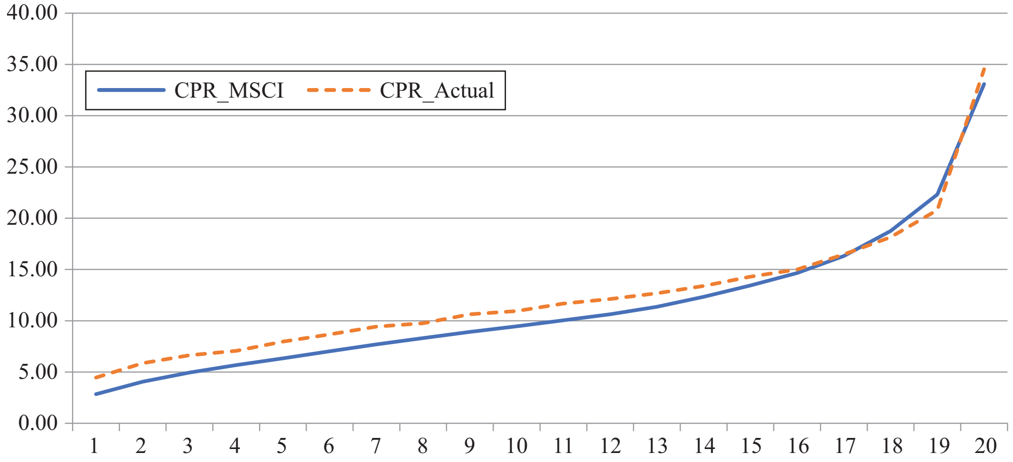
Exhibit 5 is an example of applying this rank-based methodology to the Fannie 30-year TBA universe in the 2019 January-to-September performance period. The 20 ranking buckets each have about USD 100 billion UPB, hence eliminate the statistical noise. The sample model generally slightly underpredicts prepayment over lower prepayment segment and slightly overpredicts prepayment over higher prepayment segment. The overall model error can be represented as

- Weighted average absolute *smm* error is 0.124 *smm*.
- The KS statistics for the cumulative distribution is 0.15.

Exhibit 6 is an example of applying this rank-based methodology to the Fannie 30-year TBA universe 2017

Exhibit 5

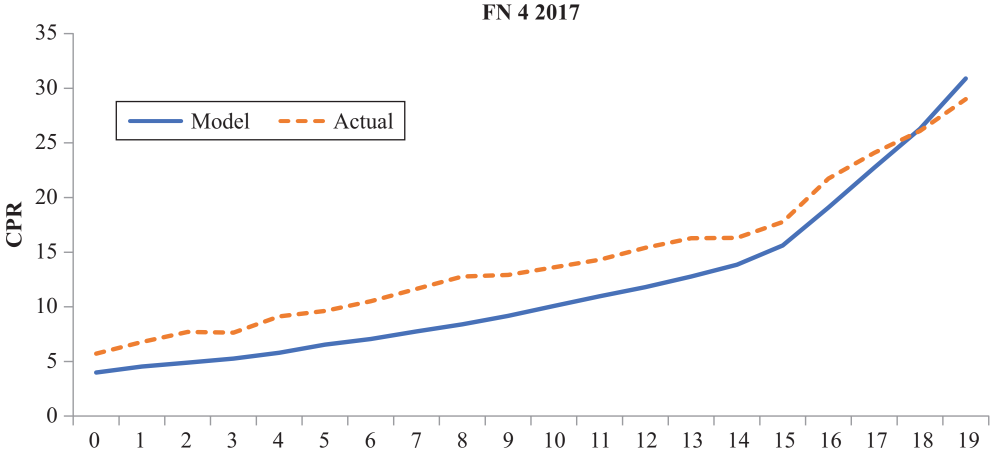
Sample Rank-Based Error Tracking for Fannie 30-Year To-Be-Announced Eligible Universe for 2019 January-to-September Performance Period



Note: Weighted average absolute smm *error is 0.124, and KS statistics is 0.15.*

Exhibit 6

Sample Rank-Based Error Tracking for Fannie 30-Year To-Be-Announced Eligible Universe 2017 Vintage 4s for 2019 January-to-September Performance Period



Note: Weighted average absolute smm *error is 0.26, and KS statistics is 0.29.*

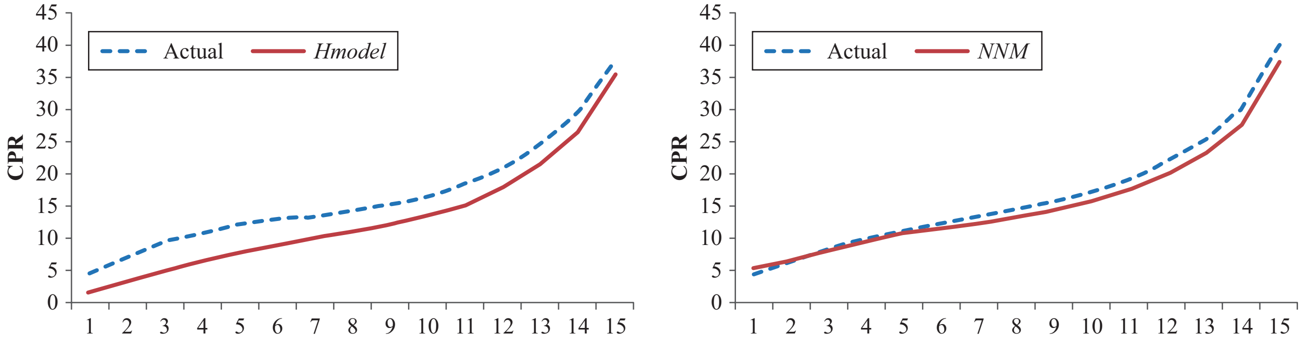
vintage 4s for the 2019 January-to-September performance period. The 20 ranking buckets each have about USD 5 billion UPB; hence, they have relatively small statistical noise. The sample model generally slightly underpredicts prepayment, suggesting an issue of modeling housing turnover segment of the prepayment. The overall model error can be represented as

- Weighted average absolute *smm* error is 0.26 *smm*.
- The KS statistics for the cumulative distribution is 0.29.

Exhibit 7 is an example of applying this rank-based methodology to compare two models: One is the sample

EXHIBIT 7

Sample Rank-Based Error Tracking for Fannie 30-Year To-Be-Announced Eligible Universe Coupon 4s



Note: Weighted average absolute smm *error is 0.13, and KS statistics is 0.12.*

model used as sample heretofore (the *human model* or *Hmodel*); the other is a corresponding model built with artificial intelligence (AI) technique (the *neural network model* or *NNM*; Zhang et al. 2019). Measured over the entire Fannie and Freddie 30-year TBA universe for prepayment behavior between 2003 and 2018, the 15 ranking buckets each have about USD 16 billion UPB. Hence, the buckets eliminate the statistical noise. The overall model comparison can be represented as follows:

- The AI model's ranking curve is generally steeper than the human model. Hence, the AI model generally outperform the human model by a higher degree of differentiate prepayment performance across the pools' attributes and across macroeconomic periods.
- The AI model is generally more accurate than the human model. (Note that this methodology focuses purely on fitting accuracy, and does not opine on whether the models are parsimonious, structurally sound, etc.)
    - Weighted average absolute *smm* error is 0.13 versus 0.33 *smm*, AI model versus human model
    - The KS statistics for the cumulative distribution is 0.12 versus 0.36, AI model versus human model.

These examples show the advantage of the rank-based error-tracking methodology, which produce a set of comprehensive and efficient modeling testing measures.

1. Measure the overall model performance in two key aspects: ability to differentiate performance across all risk factors (both macroeconomic drivers and collateral attributes) and combinations, and accuracy of model forecasts. This contrasts well with the simple one-dimensional error-tracking method's difficulty in representing high dimensional risk factors.
2. Summarize model performance in a small set of measures. The ability to differentiate across risk factors can be measured in slopes of the ranking curve. The accuracy of the model forecasts can be measured in distance between the model and the actual ranking curve, for example, through weighted average absolute *smm* error or KS statistics.
3. The small number of model test measures discussed previously can be used to benchmark model performances, set up clean model acceptance/rejection criteria, and compare models and model iterations.
4. This methodology allows efficient separation of modeling errors in prepayment intensity, the true modeling error, versus statistical noise caused by the Poisson process.
5. This methodology can be consistently applied across models, risk factors, and sample sizes.

APPLICABLE STATISTICS BACKGROUND AND RELATIONSHIP TO OTHER MODEL TESTING METHODS

Here we briefly discuss this methodology's background in statistics and the relationship with the lift curve, which is popular with loan-level modeling.

Kolmogorov-Smirnov Test

In statistics, the Kolmogorov-Smirnov (KS) test is a nonparametric test of the equality of continuous one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution or to compare two samples.[1] The KS test quantifies a distance between the cumulative distribution functions of the two distributions being compared. The null distribution of this statistic is calculated under the null hypothesis that the sample is drawn from the reference distribution (in the one-sample case) or that the samples are drawn from the same distribution (in the two-sample case).

Despite its wide use, the KS method is seldom used in agency prepayment modeling. As discussed previously, the model error tracking aims to compare the modeled prepayment intensity versus the true intensity. However, owing to the Poisson process, the true intensity is a model construct and is not readily available from the prepayment performance data. Blindly applying the KS test to the prepayment data will yield a measure that will be dominated by the statistical noise, and hence will be of little use.

The innovation of the ranked-based method is to apply the model forecast as the distribution variable, and apply the KS test to the ranking curve.

Lift Curve

In data mining and association rule learning, lift is a measure of the performance of a targeting model (association rule) at predicting or classifying cases as having an enhanced response (with respect to the population as a whole), measured against a random choice targeting model. A targeting model is doing a good job if the response within the target is much better than the average for the population as a whole. Lift is simply the ratio of these values: target response divided by average response.[2]

The lift curve can also be considered a variation on the receiver operating characteristic curve and is also known in econometrics as the Lorenz or power curve.

Similar to the rank-based error-tracking method, lift curve measures or compares the model's ability to differentiate the intensity. On the other hand, the lift curve does not measure model accuracy. Rank-based error-tracking methodology is superior to lift curve, in that it measures the differentiation and model accuracy.

APPLICATION AND INDUSTRY DEVELOPMENT

Pre-2000s, the agency MBS issuers—Fannie Mae, Freddie Mac, and Ginne Mae—did not disclose many pool-level attributes, partly to support the liquidity of their corresponding TBA market by reducing the market's ability to differentiate the underlying pools and concentrating the trading in the TBA market. As a result, error tracking along vintage-coupon dimension became the standard prepayment error-tracking methodology and generally supports the main investment use case of trading agency TBA market.

However, with more and more pool-level and loan-level data disclosure, and growth of specified pool and CMO markets, this one-dimensional vintage-coupon approach is becoming inadequate for investment and risk analysis. Furthermore, the recent launch of the UMBS market provides the market unprecedented TBA market liquidity and may potentially drive further expansion in specified pool and CMO markets. The potential increases in electronic trading in MBS TBA and pools would require much more efficient prepayment modeling and model-testing methodologies. These factors are likely to afford much more usage for the rank-based error-tracking methodology.

REFERENCE

Zhang, D., X. Zhao, J. Zhang, F. Teng, L. Siyu, and H. Li. 2019. "Agency MBS Prepayment Model Using Neural Networks." *The Journal of Structured Finance* 24 (4): 17–33.

[1] See standard statistical textbooks or Wikipedia.

[2] See standard statistical textbook or Wikipedia.

ADDITIONAL READING

Agency MBS Prepayment Model Using Neural Networks

Jiawei "David" Zhang, Xiaojian "Jan" Zhao,
Joy Zhang, Fei Teng, Siyu Lin and Hongyuan
"Henry" Li

The Journal of Structured Finance

https://jsf.pm-research.com/content/24/4/17

ABSTRACT: *The authors apply deep neural networks, a type of machine learning method, to model agency mortgage-backed security (MBS) 30-year, fixed-rate pool prepayment behaviors. The neural networks model (NNM) is able to produce highly accurate model fits to the historical prepayment patterns as well as accurate sensitivities to economic and pool-level risk factors. These results are comparable with model results and intuitions obtained from a traditional agency pool-level prepayment model that is in production and was built via many iterations of trial and error over many months and years. This example shows NNM can process large datasets efficiently, capture very complex prepayment patterns, and model large group of risk factors that are highly nonlinear and interactive. The authors also examine various potential shortcomings of this approach, including nontransparency/"black-box" issues, model overfitting, and regime shift issues.*