FICO Decisions

WHITE PAPER

Homecourt Disadvantage: Truncation Bias and the Art of Comparing Consumer Credit Scoring Models

This paper is organized as follows:

Generally available consumer credit scoring models can provide great value to lenders in evaluating the risk of loan applicants. These models are designed to distill the predictive power of a large number of consumer factors into a single, easily understood score. While the credit scoring models work well in estimating default likelihood over time, all models eventually may need to be evaluated to determine if an update to the existing model is needed, or whether the existing model should be replaced by a new credit score model. However, these evaluations can be challenging. For example, comparing the performance of the new credit scoring model to an existing model requires meticulous attention to detail as subtle statistical quirks can easily bias results.

A fundamental challenge in comparing the performance of a new credit scoring model to an existing model is that we only know the behavior of borrowers who have actually been granted credit based on the existing model. We cannot know how rejected applicants would have actually performed on a loan if they had been approved instead. While credit scoring models aim to estimate what that behavior would have been, by definition, there is no performance data on rejected applicants to verify that estimate.

When evaluating the relative performance of a new credit scoring model to an existing model, we must deal with the fact that we only know the subsequent repayment behavior of borrowers that the existing model had scored above the minimum cutoff threshold established for that market. As we will show in this paper, this creates a subtle but critical bias in statistical measures that falsely increases the reported accuracy of a new model compared to a model currently being used in a market. This is known as “truncation bias” and can lead to inaccurate conclusions when assessing new or competing models.

Credit Scoring Fundamentals

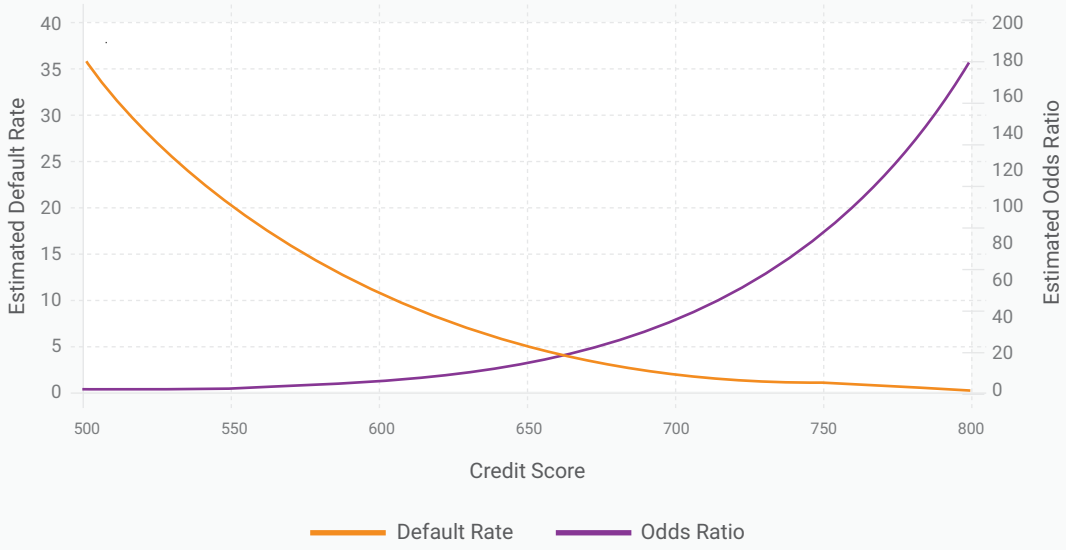
As an example, a lender may have decided that a portfolio with an average default rate of 2% would be profitable given current loan rates. In order to keep performance at or better than the target default rate, the lender must have 49 paid-in-full borrowers for every one borrower that defaults. The 49:1 relationship is known as the "odds ratio" in credit scoring and is simply calculated as:

$$Odds\ Ratio = \frac{1}{Default\ Rate} - 1$$

Odds ratios (and thus implied default rates) vary by credit score range. Exhibit 1 shows a typical pattern of odds ratios and default rates on a hypothetical credit score ranging from 500 to 800.

Exhibit 1

Example Relationships Between Odds Ratio, Default Rate and Credit Score



Lenders look at credit scoring models and pick a cutoff score above which they will approve applicants. In order to choose this cutoff score, lenders must understand the distribution of the likely applicant population along with the expected odds-to-score relationship.

The goal of a credit score is to statistically separate likely customer payment behavior according to observable characteristics available at the time of loan application. Credit scoring models group prospective borrowers into cohorts by score range. Each cohort contains borrowers with similar characteristics, such as payment history, credit utilization and credit capacity. Looking at historical default behavior, the model assigns scores according to the ratio of borrowers who have defaulted or gone seriously delinquent (known as a "default") to those who have paid their debts on time (known as a "paid"). A paid observation is typically defined as a borrower who has not gone 90 or more days past due on an obligation over a certain time period. A "default" observation is typically defined as a borrower with at least one 90+ day delinquency. Credit scoring models attempt to create cohorts that maximize the separation of paids and defaults across the score spectrum. More detail on this process is contained in the Appendix.

There are several important items to note about Exhibit 1 that will help as we delve into the intricacies of score development and truncation bias. First, the default rate and odds ratios have inverse relationships. Second, the odds ratio varies materially by credit score. In the example we have constructed, at a score of 800 we expect approximately 180 borrowers to consistently pay their loans on time for every one borrower that defaults. This compares quite favorably to consumers with a score of 600, where one out of every 11 borrowers is expected to have payment problems[1].

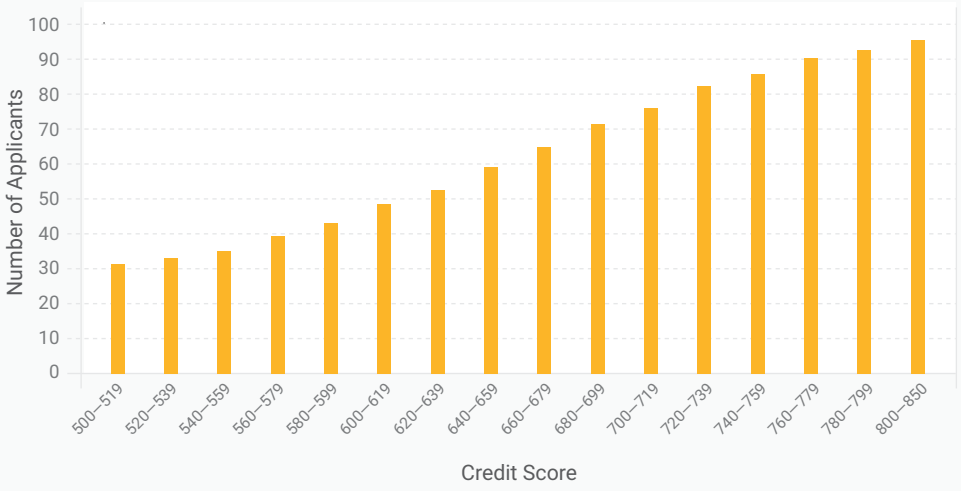
The whole purpose of a credit scoring model is to separate out groups of consumers according to their likelihood to repay. A good model has a very steep odds ratio slope as credit scores increase. This implies that the model has found characteristics that provide excellent separation of borrowers who subsequently paid on time vs. those who paid late or defaulted.

Returning to the lender who is trying to construct a portfolio with an average default rate of 2%, we see that the credit score can help make decisions but is not sufficient for portfolio construction. In addition to the likelihood of any individual consumer defaulting, the lender must also know what the likely distribution of scores will be across the entire applicant pool. This determines the cutoff score for loan approvals.

Exhibit 2 shows an example credit score distribution on 1,000 consumers likely to apply for loans.

Exhibit 2

Example Credit Score Distribution For Loan Applicant Pool



[1] The data behind Exhibit 1, presented in tabular form in Exhibit 3 below, is simply an illustration and not based on a commercially available credit score.

Exhibit 3 presents the data from the prior charts and also calculates the weighted average default rate at various lending cutoffs.

Exhibit 3

Weighted Average Default Rate at Various Lending Cutoffs

Exhibit 3

Credit Score Range	Number of Applicants	Odds Ratio	Interval Default Rate	Wgt Avg Cumulative Default Rate
800-850	95	177.6	0.6%	0.6%
780-799	93	130.6	0.8%	0.7%
760-779	90	96.1	1.0%	0.8%
740-759	86	70.7	1.4%	0.9%
720-739	82	52.0	1.9%	1.1%
700-719	76	38.2	2.5%	1.3%
680-699	71	28.1	3.4%	1.6%
660-679	65	20.7	4.6%	1.9%
640-659	59	15.2	6.2%	2.2%
620-639	53	11.2	8.2%	2.6%
600-619	48	8.2	10.8%	3.1%
580-599	43	6.1	14.2%	3.7%
560-579	39	4.5	18.3%	4.3%
540-559	35	3.3	23.4%	5.0%
520-539	33	2.4	29.3%	5.8%
500-519	32	1.8	36.1%	6.8%

Total # Applicants >= 660: 658
Wgt Avg Cumulative Default Rate: 1.9%

If the lender wanted to approve only those applicants who were projected to default at or below the target default rate of 2%, then all applicants with scores lower than 720 would have to be rejected. However, if the applicant pool has the overall credit profile given in Exhibit 2, the lender can accept scores down to 660 and still maintain an expected portfolio default rate below 2%. This is key to increasing overall profitability because the too restrictive cutoff of 720 would have qualified only 446 applicants, corresponding to approximately a 45% acceptance rate. The more appropriate cutoff of 660 would result in 212 additional loans being made – an increase of over 47%.

While the above explanation of credit application management may seem obvious, it is fundamental to issues of credit scoring model construction and evaluation. The two most important points are:

1. Models that provide better separation of paids and defaults allow lenders to accept more applicants.
2. Performance of credit scoring models near the lending cutoff level is essential to good portfolio construction and should be a major consideration in evaluating competing models.

Truncation bias is especially important with regard to the second point.

Defining Truncation Bias

Truncation bias, also known as selection bias, refers to false signals that model fit measures can deliver due to truncation of the sample of observable consumers. As an example, assume that the lender referenced above used a cutoff score of 660 supplied by a credit scoring model we will refer to as the Champion model. Further, assume that this lender's applicant pool consistently exhibited the distribution of credit scores as shown in Exhibit 2. During each year of lending, the bank would have evaluated 1,000 loan applications and accepted 658 of the applicants based on the 660 cutoff score. We would expect approximately 12 to default, consistent with our 1.9% expected default rate.

After several years of lending using the Champion model, the lender decides to evaluate a new credit scoring model we will refer to as the Challenger model. Having their historical loan application data in hand, the lender would like to obtain scores generated by the Challenger model, as though the model had been available at the time of application. The lender will then evaluate how well the Challenger model would have predicted defaults compared to the Champion.

On the face of things, this seems to be a perfectly logical approach. However, there are two significant problems – one easily solved and one much more difficult.

The easily solved problem is that the Challenger model must have been calculated only on data that was available prior to loan decisioning. Specifically, it should use only data that was available at the same point in time that the Champion model was calculated. To be clear, the Challenger could use different ***types*** of data than the Champion, but nothing that became available after the training period.

Why is this so important? First, if the Challenger used data from after the loans were granted, that would be an "in sample" analysis. In sample analysis simply ***describes*** the current behavior of the loans, it does not ***predict*** the future performance of those same loans. In fact, that performance to a certain extent has already occurred and has been incorporated into the score calculation. The lender wants a model that predicts performance of loans yet to be made, which requires testing models using "out of sample" data. This problem is easily fixed by restricting the data used to calculate the Challenger in order to make it comparable to the Champion.

The second problem is subtler and far more difficult to fix. If a lender looks at their portfolio of loans, they will only see loans that were already subject to a lending cutoff – in this case a Champion score of 660 at the time of application. The lender knows nothing about the behavior of loans scoring below 660 because ***those loans were never approved***. While the Champion model was originally applied to the entire pool of 1,000 loan applicants, the Challenger model is only being asked to evaluate the performance of the 658 approved loans. These 658 loans represent a ***truncated sample*** of the original applicant population.

The truncation problem presents several difficulties. First, there is no way to measure the performance of loans never made, so we cannot "un-truncate" the sample with perfect confidence.

The second issue with using truncated samples is both more subtle and potentially more misleading. When evaluating the power of a model to separate consumers into paids and defaults, statistical fit metrics will be inflated on the Challenger model relative to the Champion model ***regardless of how well the two models fit the entire population of loan applicants***. Remember that calculating scores and assessing default risk on only previously approved applicants is not our goal. We must score the ***entire applicant pool*** effectively in order to make future lending decisions.

The math that proves the bias in model fit statistics is complex. We have included references to academic articles on the topic in the References section. In fact, it can be mathematically proven that two identical credit scoring models will result in different relative Ginis when Champion/Challenger roles are reversed.[2]

[2] An excellent mathematical discussion of truncation bias can be found in David J Hand & Niall M Adams (2014) Selection bias in credit scorecard evaluation, Journal of the Operational Research Society, 65:3, 408-415

A Stylized Example of Truncation Bias

In order to demonstrate the concept and consequences of truncation bias, we created a stylized model of credit scores. The model results will look familiar to users of traditional scores, but all of the data is simulated in order to provide clear explanations of key concepts.

We start by creating a population of one million consumers with scores assigned from 500 to 850. The consumers are distributed across the credit spectrum according to the distribution in Exhibit 4, which is similar to published tables of FICO® Score distributions and was the data used in creating Exhibit 2. We will call the score assigned "Model 0".

Exhibit 4

Exhibit 4

Credit Score Range	Percent of Population	Default Rate	Paids	Defaults	Odds Ratio
500-519	3.2%	36.1%	20,457	11,543	1.8
520-539	3.3%	29.3%	23,321	9,679	2.4
540-559	3.5%	23.4%	26,814	8,186	3.3
560-579	3.9%	18.3%	31,849	7,151	4.5
580-599	4.3%	14.2%	36,905	6,095	6.1
600-619	4.8%	10.8%	42,800	5,200	8.2
620-639	5.3%	8.2%	48,652	4,348	11.2
640-659	5.9%	6.2%	55,361	3,639	15.2
660-679	6.5%	4.6%	62,002	2,998	20.7
680-699	7.1%	3.4%	68,562	2,438	28.1
700-719	7.6%	2.5%	74,063	1,937	38.2
720-739	8.2%	1.9%	80,452	1,548	52.0
740-759	8.6%	1.4%	84,800	1,200	70.7
760-779	9.0%	1.0%	89,073	927	96.1
780-799	9.3%	0.8%	92,293	707	130.6
800-850	9.5%	0.6%	94,468	532	177.6

The individual consumers within each 20-point score band are randomly assigned as a paid or default according to the odds ratio at each score band in Exhibit 4, which matches up to the odds ratios graphically displayed in Exhibit 1.

We now create two very similar score models (Model 1 and Model 2) to be used in the truncation bias testing. For ease of explanation, let's assume that Models 1 and 2 use the same score range of 500 to 850. Each of the new models takes each individual consumer and creates a score that is equal to the Model 0 score plus a normally distributed random perturbation. Importantly, the only difference between each of

the two new models and Model 0 is the small symmetrical random adjustment. This process creates Model 1 and Model 2 scores that have the same odds-to-score relationship as Model 0.

Although the Model 1 and 2 scores in this example are simulated, they can be thought of as having been generated by two scoring models that use similar approaches but have one or more unique explanatory variables that cause the score differences between the models.

In our stylized example, the same score level represents the same odds ratio in each model . This results from introducing only normally distributed random variation in the two models. What is different between the models is who is assigned to a particular score. For instance, consumer #1 may have a score of 720 in both models while consumer #2 has a Model 1 score of 750 and a Model 2 score of 740.

Exhibit 5 shows how consumers are mapped to the two different models. In this case, we have selected consumers in the Model 0 660-669 score band.

Exhibit 5

Dispersion Observations of Model 0 Score Band = 660-669

		Model 2 Score		
		650-659	660-669	670-679
Model 1 Score	650-659	433	1,500	1
	660-669	1,511	28,397	1,548
	670-679	-	1,578	433

(Note that Exhibit 5 uses 10-point score bands rather than 20. This is simply to highlight the different model mappings for observations with scores near the cutoff point.)

As we look at truncated samples, we will be more interested in how observations are scored near the cutoff score. Exhibit 6 shows the mapping of Model 2 scores from the Model 1 660-669 score band. If Model 1 is used to establish a cutoff at 660, over 19% of the observations[3] in that band alone will be scored at less than 660 by Model 2, which would not be constrained by the cutoff. Given that we constructed the example so that the odds-to-score ratios were the same for both models, we can conclude that Model 2 saw something in the data that indicated these particular consumers have a higher likelihood of default than predicted by Model 1 even if, on average, the odds-to-score mapping is the same for both models. The spread of observations from one model to another as shown below is key to the concept of truncation bias. From this point on, in this paper, we will refer to Model 1 as the

[3] As shown in Exhibit 6, there are 6,824 (193+6,631) observations from the Model 1 660-669 score band that are scored by Model 2 in either the 640-649 band or 650-659 band. These observations divided by the total in the Model 1 660-669 score band (6,824/35,586) gives 19.18% movement in Model 2 relative to Model 1.

"Champion" model and Model 2 as the "Challenger" model. This naming convention reflects the fact that in this simulated example, we will assume that the lender decision was driven by the consumer's Model 1 (Champion) score.

Exhibit 6

Dispersion of Observations
Model 1 Score Band = 660-669

		Model 2 Score					
		640-649	650-659	660-669	670-679	680-689	Total
Model 1 Score 660-669	Count	193	6,631	21,733	6,816	213	35,586
	Percent	0.5%	18.6%	61.1%	19.2%	0.6%	100.0%

The very existence of these consumers mapped below the Champion cutoff score will, by definition, give Challenger more separation than Champion on the truncated sample. That will result in better fit statistics, leading the model evaluator to favor Challenger. Of course Champion has some consumers who scored below 660 who have been scored above 660 by Challenger but, importantly, **those consumers have been eliminated in the truncated sample.**

We will use the Gini coefficient as our measure of model fit. A detailed explanation of how the Gini coefficient is calculated and interpreted is provided in Appendix B. In credit scoring, Gini measures the ability of a model to distinguish between paids and defaults. For our purposes here, it is enough to know that Gini ranges from 0 to 1 with 0 representing a model with no separation power and 1 representing a model that perfectly separates paids and defaults. Putting these extremes in credit scoring terms, a model with a Gini coefficient of 0 would give no information about the relative likelihood of one consumer defaulting vs. another, even after taking into account all of the predictive information the model had on the consumers. A Gini coefficient of 1 would imply a model that only had two scores – one for consumers who had a 100% likelihood of default and the other for consumers who had no likelihood of default. Credit scoring models generally exhibit Gini coefficients in the .40 to .80 range with .80 being considered a very good fit.

With the analysis population established and scored, we can now observe the effect of truncation bias through a series of simulations. In the analysis that follows, we used the base population of one million simulated consumers. We then drew 1,000 uniform random samples from that population, with each sample containing 100,000 consumers. The random draws for Models 1 and 2 are independent. Results are averaged across the 1,000 samples throughout the examples described below.

Importantly, Models 1 and 2 have been constructed to have nearly the same goodness of fit as measured by the Gini coefficient. Across the 1,000 samples drawn from the population, the average Gini coefficient for Model 1 is 0.569 compared to a Model 2 average Gini of 0.571.

We assigned the role of Champion to Model 1. Assuming the lender had accepted all applicants with a Champion score at or above 660 and rejected all other applicants, we can create a subset of the population with Champion score >= 660. No performance information on lower scoring consumers would be available because they would not have been granted loans.

We now draw 1,000 random samples of 100,000 observations from this subset. The average Champion Gini coefficient is 0.271.

Note that the average of the Gini values for the 1,000 truncated samples is considerably lower than for the entire population. By design, most of the defaulting consumers have lower scores. Cutting off the lower scoring segment of the population removes a lot of information from the model results. Generally speaking, when explanatory information is removed from a model data set, model performance suffers. To be clear, Champion does not perform any worse on the above 660 segment of the population than it did when all of the population was included. Rather, the statistical fit is simply reported as worse, meaning there is less separation of paids and defaults across the restricted score range than there was across the entire population range.

Now that we have the approved subset of consumers identified, we can evaluate the fit of Challenger on that group. This group of consumers is exactly the same as was used for calculating the Gini for Champion. However, we now have new scores for those consumers based on Challenger. As previously discussed, we expect the Challenger score range to extend below the Champion cutoff of 660 for some consumers.

The average Gini coefficient for Challenger using the cutoff segment is 0.307, which is higher than the average Gini of 0.271 that we calculated for Champion across the same 1,000 truncated samples. More importantly, Challenger had a better Gini than Champion in 80% of the 1,000 simulations. This would lead some model reviewers to mistakenly favor the Challenger model over the Champion model.

As shown in Exhibit 7, Challenger, due to its random variation from Champion, has a wider score range across which the same consumers are spread. This allows for more mathematical dispersion, which results in a higher Gini coefficient relative to the restricted-range Gini on Champion.

Exhibit 7

Exhibit 7				
	Score Range	Observations < 660	Average Gini Coefficient	% Wins
Champion (Model 1)	660 - 800	0.00%	0.271	20%
Challenger (Model 2)	640 - 800	1.11%	0.307	80%

Could the difference shown above be due to truly superior model performance by Challenger? After all, some models are, indeed, better than others. In this case, the answer is clearly no because we constructed the models to have the same odds-to-scores ratios over the entire population. We only introduced slight random variation in the assignment of individuals to one bucket vs. another.

In this analysis, Challenger and Champion are essentially the same model in terms of predictive power on the overall population. Yet, Challenger consistently beats Champion on the truncated population. That is the definition and impact of truncation bias: when present in a dataset, fit statistics on two models with identical accuracy can differ enough to create preference for the Challenger model, ***even when there is no material difference in the effectiveness of the underlying models.***

Before we move on to different techniques we can use to address this bias – and the implications of not addressing it – let us look at one more example to confirm that these results are not spurious.

Starting with the same population of simulated consumers, we will now simply reverse the roles of Model 1 and Model 2. Model 2 will now become the Champion model and the population will be truncated at a Model 2 cutoff of 660. Model 1 will now be the Challenger. Exhibit 8 shows the results of this new analysis.

Exhibit 8

Exhibit 8				
	Score Range	Observations < 660	Average Gini Coefficient	% Wins
Challenger (Model 1)	640 - 800	1.15%	0.314	87%
Champion (Model 2)	660 - 800	0.00%	0.263	13%

Model 1, now the Challenger, prevails in 87% of the samples. The average Gini coefficient for the Challenger is .314 on these samples, 19.4% higher than the average Gini coefficient of .263 for the Champion.

Remember that the fundamental data has not changed between the two examples. The models are the same, only the role they play has changed.

Comparing Models Appropriately

With all of this seeming statistical chicanery, you may be wondering if all models are the same or if model fit measures are useless. Fortunately, the answers are much easier than understanding the subtleties of truncation bias.

All models are clearly not the same. First, even a single modeling team generally produces improvements to existing versions of its own models. They will sometimes find previously undiscovered explanatory value in new variables or new functional forms of transforming old variables. Also, models can lose predictive power as they age because consumer behavioral patterns change. This can happen for many reasons, including a different macroeconomic environment, new sources of lending and changes in the actual pool of consumers. Taking all of these factors into account, a good modeling team can generally improve its models over time. However, as a model becomes increasingly refined and tuned to larger data sets, subsequent improvements will generally be small.

Model fit measures are also definitely useful and generally unbiased ***if they are used correctly***. Truncation bias is an example of incorrect reliance on standard measures of fit. Those same measures would be appropriate if they had been applied over a sample of the population that was not truncated. Recall that Models 1 and 2 have nearly identical fit measures when applied to the full range of scores. That's to be expected because the models were designed to be essentially the same.

When evaluating the relative performance of models, keep four things in mind:

1. Make sure the time periods for the training data sets are the same. New explanatory variables can be used, but they must have been available at a time prior to the performance period being used in the assessment of the models.
2. Do not use truncated samples. This means that lenders will have to go outside of their own data to evaluate the performance of models because the in-house data is, by definition, truncated.
3. If at all possible, do not compare a new model to the model that was used for lending decisions on the training set. Use a newer version of that model to mitigate some of the truncation bias. Ideally, use a model that was not involved in creating the sample.
4. Use caution in comparing nominal scores from two different models. Even if the score ranges are the same, the odds-to-score mapping could be quite different between the models. This is particularly important when setting lending cutoff levels.

Mitigating Truncation Bias

One question that often comes up with regard to truncation bias is: How is it possible to avoid truncation bias if the performance data is based on someone's selection of a cutoff score?

There are several valid approaches to credit scoring model comparison that reduce or eliminate the impact of truncation bias.

As one option, reject inference is a standard approach in developing application scoring models that can also be applied in validating and comparing models on a population affected by truncation bias. The aim of reject inference is to estimate or infer what the performance would have been on consumers who were rejected at the time of application. Reject inference, if done in a sound manner, is one of the best ways to combat truncation bias. Using reject inference allows for comparison of models on a full applicant population which is much more representative of the population on which the models are expected to be used. We discuss reject inference methods further in the Appendix.

A second option is to find a population that was unaffected by a particular cutoff score to perform the testing. As an example, while conforming mortgage lending is generally subject to hard cutoff scores, many lenders have non-conforming programs that cover a broader range of credit score levels. Using data from these non-conforming populations to compare two models is a good way to reduce truncation bias - although there may be other biases present due to differences in underwriting approaches between conforming and non-conforming lending. We use conforming mortgage underwriting solely for illustrative purposes; the principle applies to any situation where lenders or guarantors apply different score cutoffs to substantially similar populations of loan applicants.

A third option, as explained in point 3 above, is to use new models for Champion/Challenger analysis. Do not mix an older model that was used to determine lending cutoffs in comparison to a new model that was not. When doing this, take care to understand how correlated the results of the new models are with the older cutoff generating model. As an example, FICO regularly updates its models to reflect new data and new modeling techniques. That is one of the benefits of the FICO models having evolved steadily over 30 years. High correlation between new and existing models is by no means a bad thing. It simply makes model comparison more difficult.

A fourth option involves running a test program wherein standard underwriting criteria are adjusted to allow some loans to be approved below the traditional score cutoff level. The performance on these loans would then be tracked and used in comparing score models. While this option creates objective data for evaluating models with less truncation effect, it does so at the cost of taking on riskier loans than a lender would normally allow. This option also requires a substantial waiting period of perhaps 24 months for performance to develop.

Conclusion

Credit scoring models evolve over time as new data becomes available, consumer behavior changes and improved modeling techniques are deployed. New models, and updated versions of older models, have the potential to better predict consumer payment behavior. However, comparing credit scoring models is a complex statistical exercise that must take into account many nuances. In this paper, we outlined the challenge of truncation bias. We showed that even if two models are theoretically identical, the model that was used to approve booked loans in the validation sample will be at a distinct disadvantage in terms of standard model fit statistics. Without careful design of the model comparison approach, lenders face the real risk of accepting a new model that may actually have inferior predictive power compared to an existing model.

Statistical methods exist to mitigate truncation bias and ensure appropriate comparison of credit scoring models. We advise lenders and policymakers to consult with statisticians and data scientists who are familiar with these techniques prior to undertaking model validations. This can ensure that the conclusions drawn from a credit score assessment are robust. Making the correct decision about which score to use going forward will be borne out by the future results of lending based on that score.

Appendix: Measuring Model Goodness of Fit

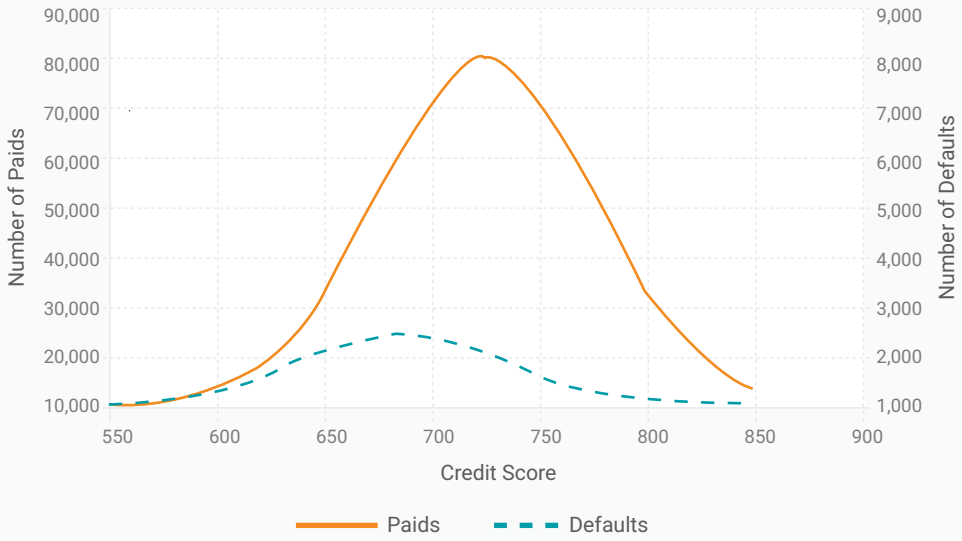
As described in the body of the paper, comparing credit scoring models requires one or more measures of **goodness of fit**. This appendix describes the basics behind two common measures of goodness of fit: the KS statistic and the Gini coefficient. The examples used here are purely illustrative.

Credit scoring models are designed to maximize the separation of paids and defaults. The greater the separation, the better fit a model has and the more useful it is in making credit granting decisions.

Exhibit A1 shows the distribution of paids and defaults for Model A, a purely hypothetical model built on the characteristics of one million fictional consumers. The solid line represents the distribution by score of those consumers who paid debts on time and are thus designated as paids. For instance, of those consumers scoring 800 or higher, there are approximately 25,000 paids. The dashed line represents the count of consumers who failed in their credit obligations. In this example, we use an average overall default rate of 2%. The solid line thus represents 980,000 consumers while the dashed line represents 20,000 consumers[4].

Exhibit A1

Model A: Distribution of Paids & Defaults



Model A shows reasonable separation of paids and defaults. Most of the paids are at the mid to higher end of the scoring spectrum while most of the defaults are at the lower end. Models are not good enough to achieve complete separation, which would mean that we could predict with certainty exactly who was going to pay and who was not. Therefore, it is important to be able to distinguish various levels of separation when comparing models.

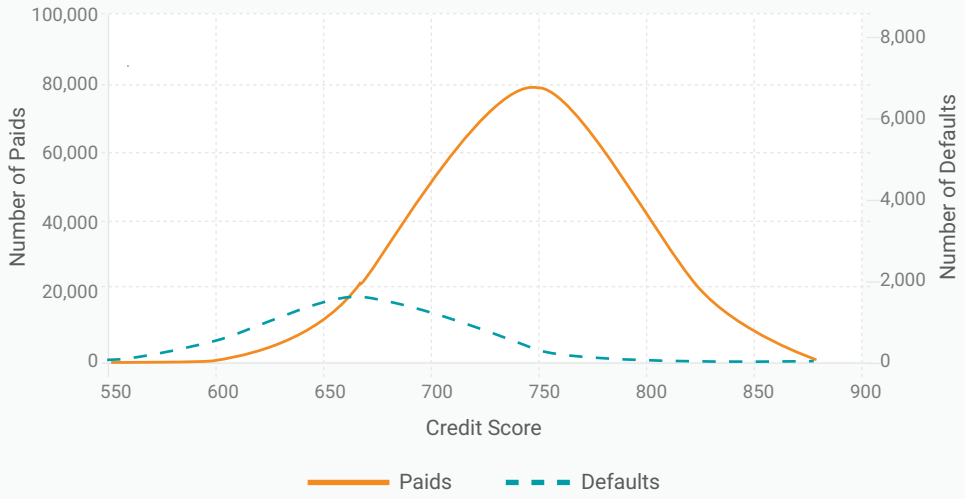
Of course, we want the separation to be as great as possible. Exhibit A2 shows the distribution by score generated by Model B. It is clear that there is greater differentiation in this model compared to Model A. A higher proportion of the defaults

[4] Note that while the curve looks continuous for visual clarity, consumers are actually grouped into discrete 10 point score bands for calculation. Different scales for the paids and defaults are used to make the illustration clearer.

are concentrated further down the scoring spectrum. Note that the consumers' behavior in the two models is the same. Only the mapping of each consumer to a score is different.

Exhibit A2

Model B: Distribution of Paids & Defaults



The default rate at any given score level is given by dividing the number of defaults at that level by the total number of consumers at the same score level. For instance, in Model B, there are 1,248 defaults in the 700-709 credit score range while there are 52,136 paids at that level. This translates to a default rate of 1,248/(1,248 + 52,136) or 2.3% for Model B in the 700-709 range. Model A, on the other hand has a default rate of 2.2% in the 700-709 score range. Exhibit A3 displays the default rate by score level for the two models.

Exhibit A3

Default Rate by Score Level

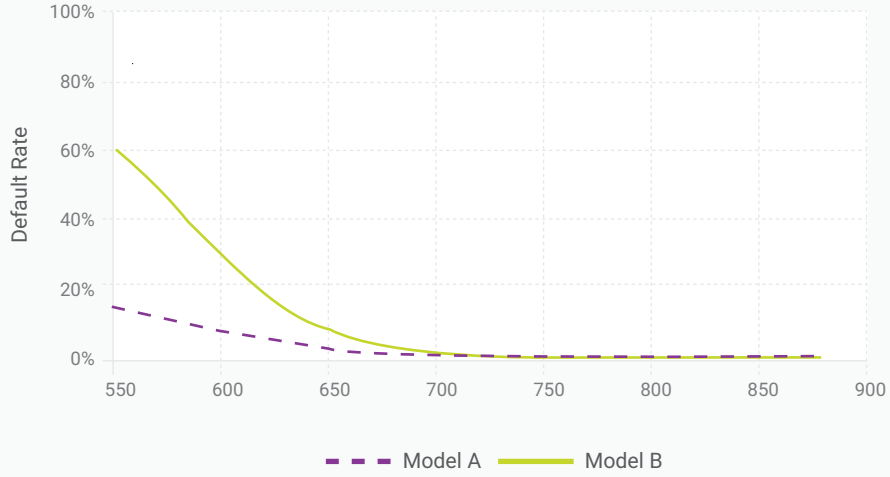


Exhibit A3 clearly shows the superior separation provided by Model B relative to Model A as the lower scores have higher defaults for Model B while the higher scores have lower defaults relative to Model A. Remember that the population default rate is 2%, only the mapping of the one million consumers to scores varies by model. Note that the distribution of scores, given in Exhibits A1-A2 is still necessary to make conclusions about the superiority of Model B.

We have constructed this example to make the superior separation of Model B readily apparent in the charts. Of course, real models do not generally display such obvious differences, so we need statistical measures to summarize the fits. We now explain how two of most common fit summaries can be interpreted.

Exhibit A4 displays the cumulative percentage of paids and defaults at different score levels, starting with the lowest score. For example, slightly less than 30% of the total defaults in the population are scored at 650 or less, while slightly less than 10% of the paids are at or below 650. In separating the paids and defaults, we would like to have most of the defaults occurring at the lower end, leading to a higher cumulative default curve, and most of the paids at the high end, leading to a lower cumulative paids curve. The KS statistic measures the maximum vertical separation between the curves on a scale of 0 to 1 with higher KS values representing better separation and thus better model fit. Model A has a KS of 0.31.

Exhibit A4

Model A: KS Statistic 0.31

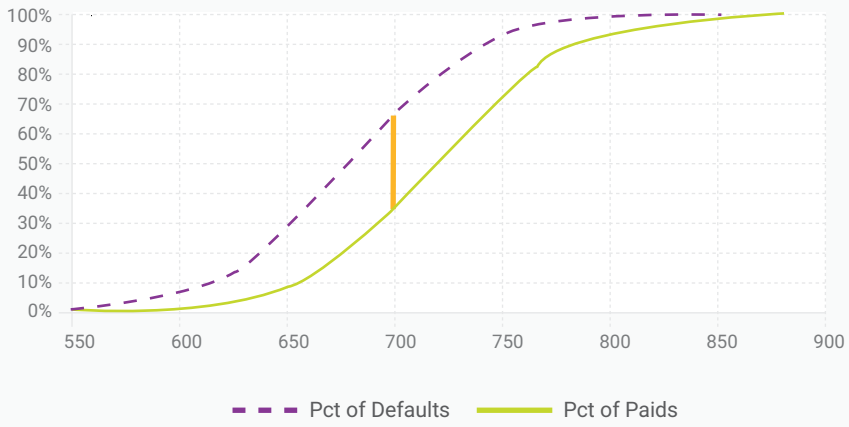
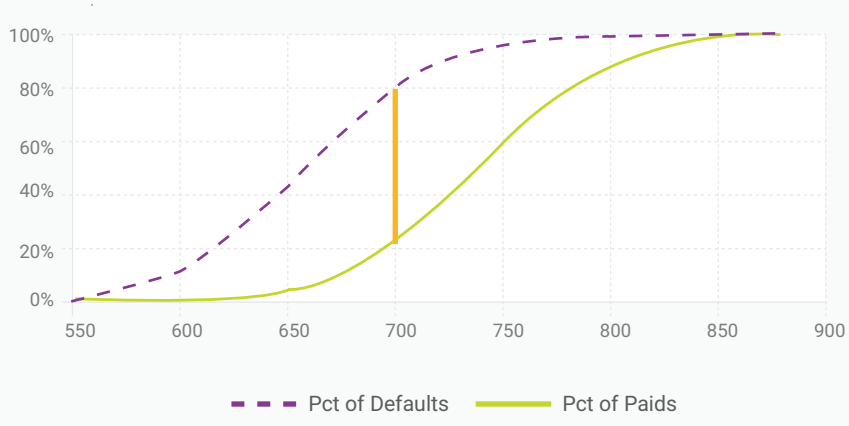


Exhibit A5 shows the results for Model B. As we saw earlier, Model B has better separation, and this is confirmed by the larger KS of 0.58.

Exhibit A5

Model B: KS Statistic 0.58

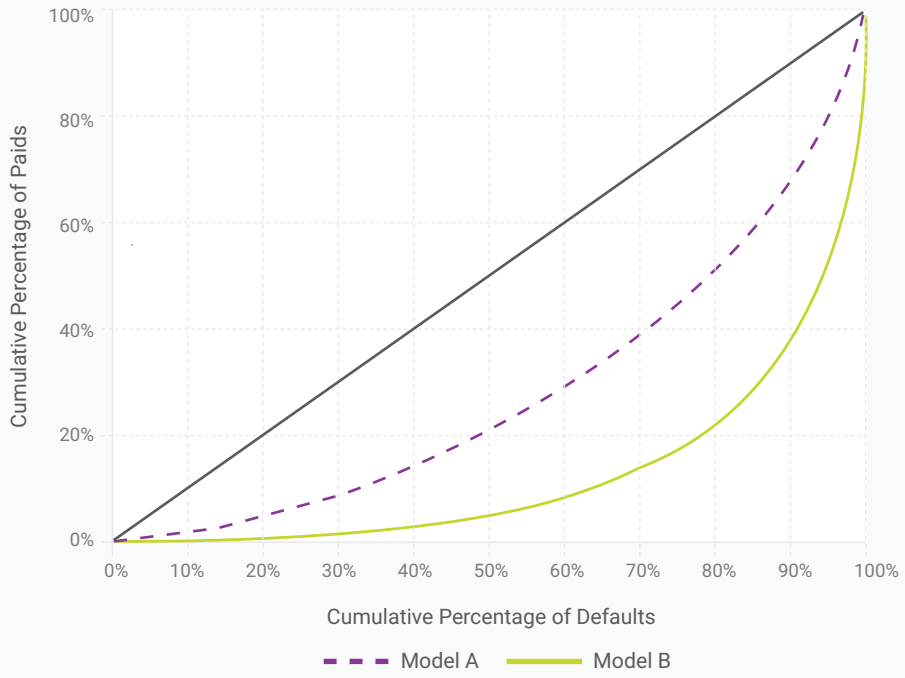


Another measure of model goodness of fit commonly used in credit scoring is the **Gini coefficient**. Similar to the KS statistic, the Gini coefficient measures separation on a scale of 0 to 1 with higher values indicating better separation. In order to visualize the separation used in calculating the Gini coefficient, we plot lines known

as Lorenz curves with cumulative percentage of defaults on the x axis and cumulative percentage of paids on the y axis. A model with high separation will have a fairly flat slope to start before it steepens dramatically. The actual Gini coefficient is calculated as the area between the Lorenz curve and the diagonal. The greater the area, the higher the Gini coefficient indicating better separation achieved by the scoring model. Exhibit A6 displays the Lorenz curves for Model A with a Gini of 0.682 and Model B with a Gini of 0.836, thus confirming our prior knowledge that Model B results in better separation of paids and defaults and is thus a better fit model than Model A.

Exhibit A6

Gini Coefficients
Model A: 0.682 Model B: 0.836



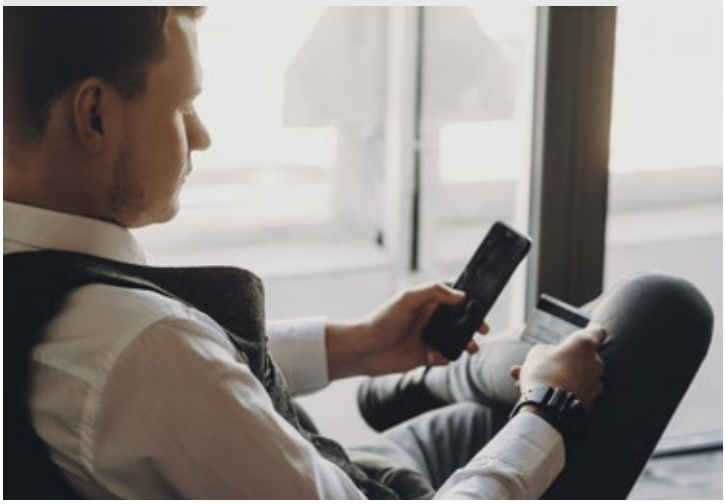
While the calculation and interpretation of goodness of fit measures is fairly technical, the point to remember is that you want a model that clearly distinguishes between consumers likely to pay their obligations from those who are likely to pay late or default. Looking for higher KS statistics or Gini coefficients is a first step in evaluating models, but as noted throughout the main article, all other biases must be accounted for before comparing goodness of fit measures.

Appendix: Reject Inference

Reject inference is an approach that can be used to estimate or infer what the repayment performance would have been on consumers who were rejected at the time of application. Reject inference can be a complex process that has its challenges and may introduce its own biases. Thus, it is best undertaken by data scientists who have expertise in this area.

A few alternative techniques for reject inference are as follows:

1. Consider an applicant population for a particular product type, say mortgage. The lender has a number of rejected applicants whose performance behavior with the lender is thus unknown. For the purpose of ascertaining what their performance



behavior might have been had the lender approved them, it may be possible for the lender to get recent credit bureau data on these consumers. From there, the lender may observe whether many of these consumers showed that they obtained a similar loan in the application time period from some other lender who used different approval criteria. The lender would use performance behavior from such loans on the credit report as a surrogate for the performance the consumer would have demonstrated had the lender booked these rejected consumers. Thus the rejects could be classified as paids or defaults on the basis of these surrogate loans for the purpose of undertaking the calculation of model metrics and other reports constructed for a comparison of predictive models. Clearly, not all of the rejected applicants will have some surrogate loan on their credit report to use for this purpose, which may pose some limitations.

   a. Related, performance behavior on less similar products observed on the credit report over the same performance period of the approved loans could be used as surrogate performance. Of course, the lender wishes to understand how well the models fit to performance on their own product type, so using other product types to calculate surrogate performance is not ideal, and statistics derived from such an approach are less likely to be reflective of what the true statistics would have been.

2. A reject inference method that is commonly used in developing application scorecards can also be used to combat the truncation bias in a validation exercise. The approach is to develop a model or use an existing model that fits the approved population on their characteristics at the time of application. This model has an observed odds-to-score relationship. By then applying the model to the rejected applicants based on their data at the application point in time, each rejected applicant has its own score calculated from the model, and corresponding odds of default. From there, each applicant can be assigned to be wholly a paid or a default based on their projected odds, or each applicant can be "parceled" as being partly a paid and partly a default based on the projected odds. The paids and defaults as imputed on the rejects are then combined with the paids and defaults on the known booked population for the model evaluation analysis. An important caveat here is that a score that is used in the reject inference, as well as its component factors, may appear overly strong in a model validation on the resultant population. A knowledgeable analyst needs to review the results of a reject inference by this method for soundness and appropriateness before proceeding to the model evaluation exercise.

The same methods can be used on "uncashed" applicants, i.e., those who were approved at the time of application but who may have turned down the loan in order to take on a similar loan with a different lender.

For further information, please see the white paper *Building Powerful, Predictive Scorecards* cited in reference 15.

References

1. G Verstraeten & D Van den Poel (2005) The impact of sample bias on consumer credit scoring performance and profitability, Journal of the Operational Research Society, 56:8, 981-992, DOI:10.1057/palgrave.jors.2601920.
2. David J Hand & Niall M Adams (2014) Selection bias in credit scorecard evaluation, Journal of the Operational Research Society, 65:3, 408-415, DOI:10.1057/jors.2013.55.
3. D J Hand (2005) Good practice in retail credit scorecard assessment, Journal of the Operational Research Society, 56:9, 1109-1117, DOI:10.1057/palgrave.jors.2601932.
4. J Banasik, J N Crook & L C Thomas (1999) Not if but when will borrowers default, Journal of the Operational Research Society, 50:12, 1185-1190, DOI:10.1057/palgrave.jors.2600851.
5. B Baesens, T Van Gestel, S Viaene, M Stepanova, J Suykens & J Vanthienen(2003) Benchmarking state-of-the-art classification algorithms for credit scoring, Journal of the Operational Research Society, 54:6, 627-635, DOI:10.1057/palgrave.jors.2601545.
6. J Banasik, J Crook & L Thomas (2003) Sample selection bias in credit scoring models, Journal of the Operational Research Society, 54:8, 822-832, DOI:10.1057/palgrave.jors.2601578.
7. E Mays (Editor), Handbook of Credit Scoring (2001), Glenlake Publishing.
8. D.J. HAND W.E. HENLEY (1993), Can reject inference ever work? MA Journal of Management Mathematics, Volume 5, Issue 1, 1 January 1993.
9. www.fico.com/blogs/analytics-optimization/how-analytics-developers-can-game-model-results/
10. www.fico.com/blogs/risk-compliance/us-average-fico-score-hits-700-a-milestone-for-consumers/
11. M ŘEZÁČ, F ŘEZÁČ (2001) How to Measure the Quality of Credit Scoring Models, Finance a úvěr-Czech Journal of Economics and Finance, 61, 2011, no. 5.
12. www.eco.uc3m.es/~ricmora/MICCUA/materials/S23T43_English_handout.pdf.
13. N Kiefer & C Larson (2004), Specification and Informational Issues in Credit Scoring, OCC Economics Working Paper 2004-4.
14. S Glasson(2007) , Censored Regression Techniques for Credit Scoring http://researchbank.rmit.edu.au/eserv/rmit:6356/Glasson.pdf.
15. Building Powerful, Predictive Scorecards, An overview of Scorecard module for FICO® Model Builder (2014), www.fico.com/en/resource-download-file/3477, includes a section on Performance Inference beginning on pg. 29.

Authors

Ethan Dornhelm
Vice President, Scores and Predictive Analytics at FICO

Paul Panichelli
Principal Scientist, Scores and Predictive Analytics at FICO

Tom Parrent
Principal at Quantilytic

4682WP_EN 03/19 PDF