Tri-Credit Reporting Company - Modeling Data Design White Paper

May 2006

Overview

Traditional multi-credit reporting company (CRC) model developments have involved one of two scenarios:

a. Extracting distinct samples from each CRC at different times and using those samples in separate development efforts, resulting in different algorithms that are then aligned on the back-end to have the same scale.

b. Extracting a single sample from one CRC and using that sample in a mono-CRC development effort, resulting in a single algorithm that is then "translated" to apply to the other CRC's data on the back-end.

The development of VantageScore featured a modeling data design superior to that of traditional tri-CRC model developments because it employed an equitable and consistent contribution of data from each CRC. Consequently, biases and variability inherent in the traditional tri-CRC data design were eliminated, resulting in the ability to create a single tri-CRC scoring algorithm that requires neither alignment nor translation among the three CRCs.

By facilitating a single tri-CRC scoring algorithm, the data design allows for a more seamless scoring strategy implementation for credit grantors and easier score interpretation for consumers. It also allows for true tri-CRC characteristic leveling (see related white paper). A detailed description of the VantageScore data design and its advantages over traditional model development data designs will be discussed in the remainder of this paper.

The Importance of Data Design for Tri-CRC Model Development

As with the development of any product, the use of flawed inputs results in a flawed end product. As the saying goes, "garbage in, garbage out." This applies to the traditional data design for the development of tri-CRC credit models. The common practice by credit grantors of using three CRC scores to make credit decisions highlights the need for a CRC-based score that is as consistent among the three CRCs as possible. Ideally, the score should:

a. Be based on a single algorithm common to the three CRCs so that there are no biases or variability due to differences in point assignment for a given credit characteristic.

b. Be based on data from all three CRCs so that no biases can be attributed to the contribution, sourcing, or timing of the data by any one CRC.

The traditional data design for the development of tri-CRC models does not meet these requirements. Consequently, the meaning of tri-CRC scores developed using this data design is not as clean as possible. Credit grantors do not have a tool that can be used to gauge risk with consistency and consumers do not have a score that they can interpret easily among the three CRCs.

Traditional Data Design for Tri-CRC Model Development

As described previously, the data design for tri-CRC models had been done in one of two ways:

a. Extracting distinct samples from each CRC at different times and using those samples in separate development efforts, resulting in different algorithms that are then aligned on the back-end to have the same scale.

b. Extracting a single sample from one CRC and using that sample in a mono-CRC development effort, resulting in a single algorithm that is translated and applied to the other CRC's data on the back-end.

The first of the traditional data design methods involves the developer independently extracting data from potentially different time frames. The data are then used to create independent models that will contain different characteristics and different point assignments between the three CRCs was used. The resulting models are then aligned to have the same score range and score-to-odds interpretation.

There are several problems with this data design method. First, the data extracted by each CRC may represent different points in time for each CRC, resulting in a bias where seasonality and different credit file compositions at different points in time are represented differently by each CRC. For example, one time period may have substantially more recently opened mortgages than another, resulting in a sub-optimal level of predictive power for that characteristic. Second, the characteristics and associated points that make up the three scores are not consistent. The differences in the data between the three CRCs will result in potentially different characteristics and different point values for the characteristics. This could result in a consumer potentially getting widely different adverse action reason codes between the three CRCs, even with scores that may be close to each other. Third, score alignment is an exercise that requires estimation, thus introducing additional variability to the aligned score.

The second of the traditional data design methods involves the development of the model using a single CRC's data, then force-fitting the remaining CRCs' data into the developed model. As with the first method, there are problems with this method as well. First, the model is biased toward the sampling routine used by the contributing CRC's data, as the other CRCs did not contribute to the development data. Second, the characteristics in the developed model are biased toward the contributing CRC's data. As such, equitable characteristic leveling is not attained because the non-contributing CRC's data are being forced to conform to the contributing CRC, when such conformation may not be possible given the data differences between CRCs.

A New Method: VantageScore Data Design for Tri-CRC Model Development

The development of VantageScore introduced a new and superior data design for tri-CRC model development, which eliminates the drawbacks of the traditional data design. The data design is as follows:

1. Each CRC created an X-record randomly selected source file of consumers with unique sequence number and identification information. Because the developed score is generic, a random selection of consumers across applications (account management, acquisitions, collections, etc.) and industries (e.g., sub-prime credit card lenders to prime mortgage lenders) was used.

2. The source lists are merged, resulting in 3X records. This is a key differentiator of the VantageScore Data Design from the traditional design.

3. Each CRC appends anonymized raw credit data to the 3X records for the agreed upon common performance and observation dates. Typically, each CRC uses name, address, SSN, and CRC-specific unique identifiers to obtain the credit data. A total of 9X records are available:
    a. CRC A characteristics and performance for CRC A source list
    b. CRC A characteristics and performance for CRC B source list
    c. CRC A characteristics and performance for CRC C source list
    d. CRC B characteristics and performance for CRC A source list
    e. CRC B characteristics and performance for CRC B source list
    f. CRC B characteristics and performance for CRC C source list
    g. CRC C characteristics and performance for CRC A source list
    h. CRC C characteristics and performance for CRC B source list
    i. CRC C characteristics and performance for CRC C source list

4. Leveled performance and characteristics are aggregated. Leveling, or normalization, is the process that yields consistent and equitable performance and characteristic definitions across multiple sources of information. Simply put, leveling ensures that when the same data are present for multiple sources (here, two or more CRCs) they are interpreted in the same manner, keeping in mind that differences in the data itself may still be present. The leveling process is possible because each consumer in the 3X sample has credit data from each CRC. (See the white paper on true tri-CRC leveling for a more detailed description of the process.)

5. Modeling populations are chosen:
   a. "Stacked" development sample chosen from the 3X records. By a "stacked" sample, we refer to the characteristics and performance aggregated in Step 4 being used from each CRC's source list and then being set together (or stacked). In this scenario, each source list represents one-third of the sample and characteristics and performance come from the same CRC. This type of sample affords the ability to create a single, composite scoring algorithm because there is an equal representation of credit data by each CRC. The stacked development sample consists of

      1. CRC A sourced characteristics and performance
      2. CRC B sourced characteristics and performance
      3. CRC C sourced characteristics and performance

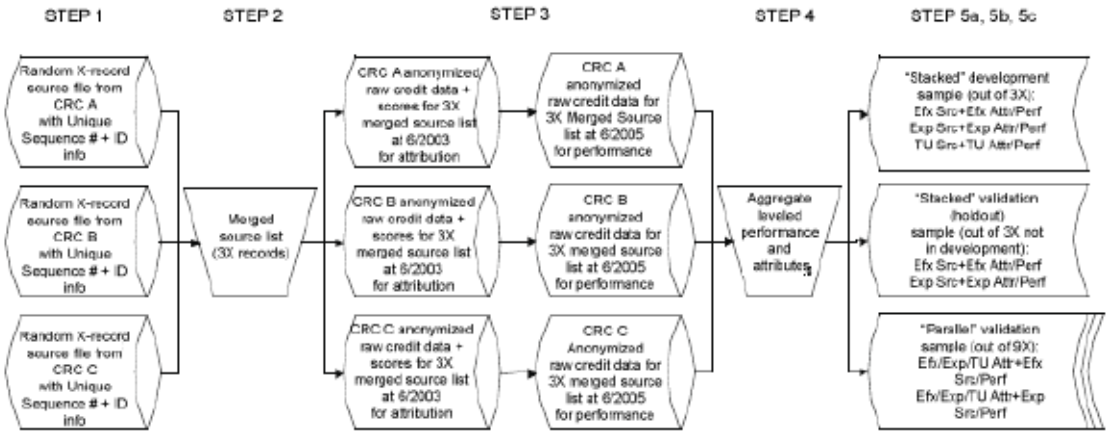
   b. Stacked validation (holdout) sample from the 3X records (not in the development sample). The stacked holdout sample is used to ensure model performance results consistent with those obtained using the stacked development sample. The "stacked" holdout sample consists of

      1. CRC A sourced characteristics and performance
      2. CRC B sourced characteristics and performance
      3. CRC C sourced characteristics and performance

   c. "Parallel" validation sample of 9X records. By a "parallel" sample, we refer to the performance aggregated in Step 4 for each CRC's source list to be crossed with the characteristics in Step 4 for the corresponding consumers from each of the three CRCs. So the same performance for all three source lists is paralleled in the sample three times, once with each CRC's version of the aggregated characteristics. The parallel sample consists of

1. CRC A sourced performance and CRC A, B, C characteristics
2. CRC B sourced performance and CRC A, B, C characteristics
3. CRC C sourced performance and CRC A, B, C characteristics

Graphically, the VantageScore data design can be represented as follows:



The VantageScore data design is preferable to traditional data design for the following reasons:

A. **Consistent Seasonality —** The extracted data are taken from the same points in time by all three CRCs, eliminating seasonality biases and ensuring consistent credit file composition across CRCs.

B. **Ability to Level Characteristics —** With equal sourcing and representation by each CRC, the attributes can be leveled, eliminating the bias that would be present if only one CRC's characteristics were used.

C. **Ability to Create a Single, Composite Scoring Algorithm —** With an equal representation of credit data by each CRC, a single scoring algorithm can be created that reflects the combined level of predictive power of the leveled characteristics for the three CRCs (as opposed to the traditional method's reliance on the predictive power of only one CRC's characteristics) resulting in a true tri-CRC model. Additionally, the elimination of the need to scale three separate scores to each other removes another source of variability present in the traditional method.

D. **Ability to Test Stability of Score Performance Across CRCs —** With the parallel validation samples described above in 5c, the consistency of the score's predictive power across the three CRCs when using one CRC's characteristics with another CRC's performance can be tested and validated. This provides yet another way to ensure the score's "CRC-independent" level of predictive power.

A Demonstration of Consistent Tri-CRC Predictive Power: VantageScore Parallel Validation Results

The consistent effectiveness of VantageScore across all three CRCs can be illustrated in the graphs shown in Figures 1, 2, 3, and 4, which used the parallel validation sample. The lines or bars in each graph reflect the scores for one CRC and performance from each of the three CRCs, where each CRC contributes one-third of the performance to the population.

Figure 1. New Account 90+ Bad Rate by Score for each CRC

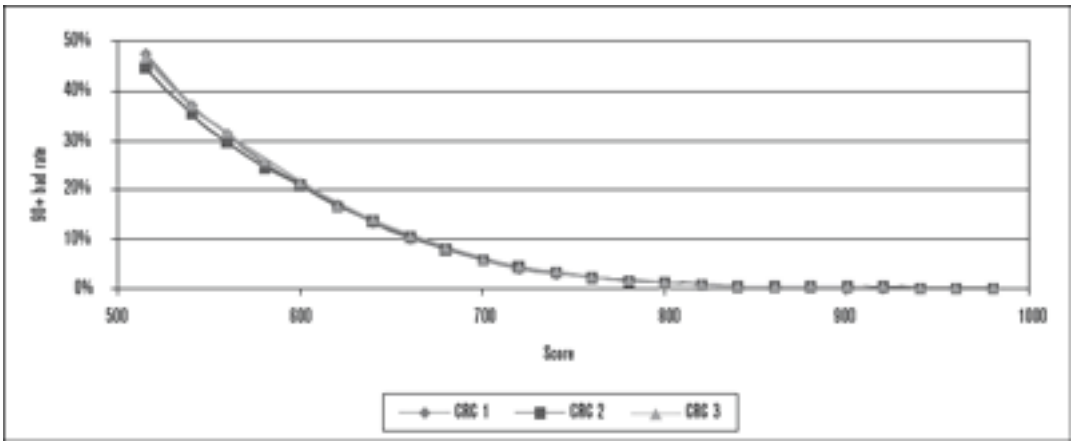


Figure 1 shows the consistent bad rate of new accounts across all three CRCs for any given score. The separation of the three CRCs at the lower scores can be attributed to the low population size of the new account population.

Figure 2. New Account KS by Industry for each CRC

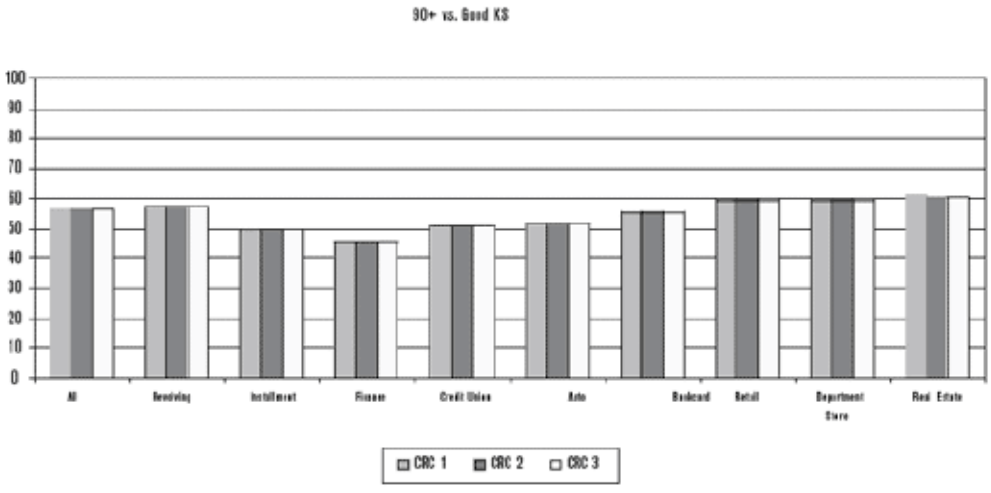


Figure 2 shows the KS for new accounts for each CRC across various industries. Within any given industry, each CRC performs virtually identically to the others.

Figure 3. Existing Account 90+ Bad Rate by Score for each CRC

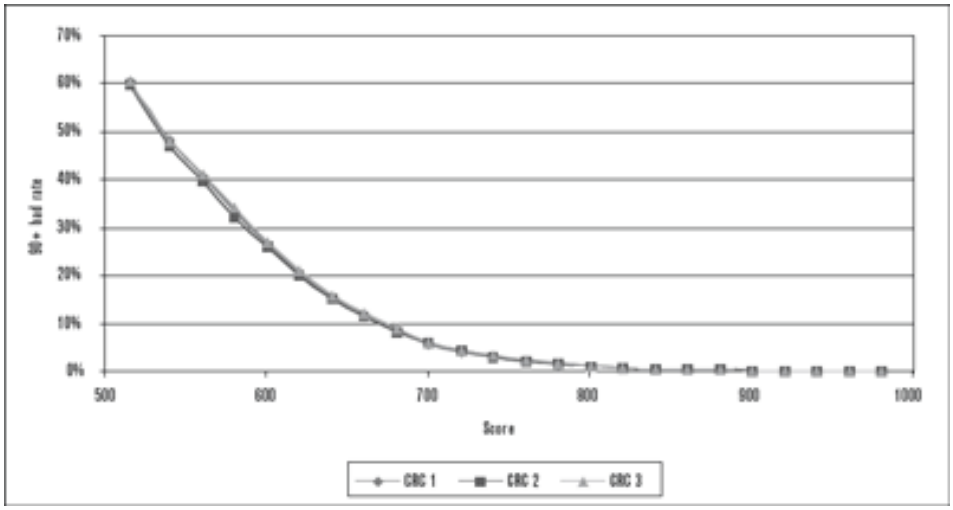


Figure 3 shows the consistent bad rate of existing accounts across all three CRCs for any given score. Because of the large sample size, the three CRCs all converge to a common bad rate at a given score.

Figure 4. Existing Account KS by Industry for each CRC

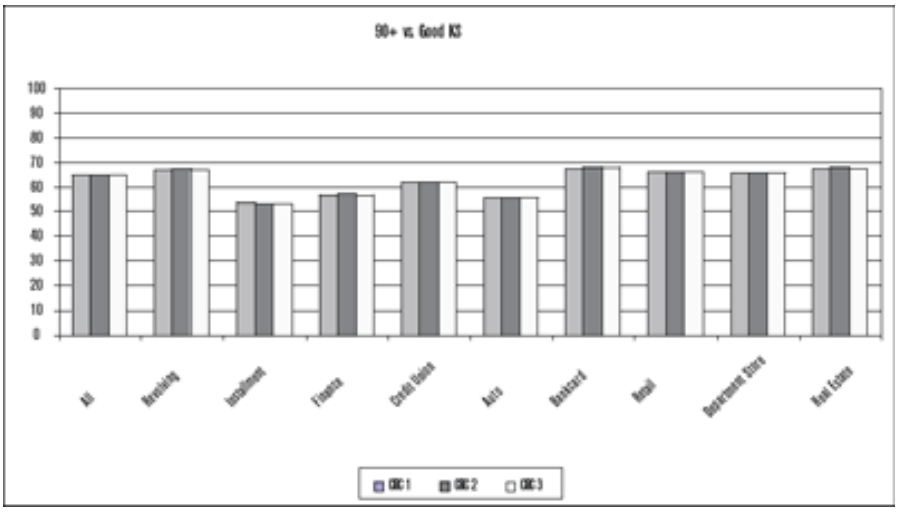


Figure 4 shows KS for existing accounts across all three CRCs for various industries. As with the new accounts, the CRCs are virtually identical in KS within any given industry.

The VantageScore data design allows for this type of CRC performance comparison because the consumers in the sample are present in all three CRC files. This type of comparison is not possible under the traditional tri-CRC design method because of the absence of the "three CRC files for the same consumer at the same time" feature present in the VantageScore design.

Superior Data Design Befitting a "True" Tri-CRC Score

The data design for traditional tri-CRC model developments have resulted in scores that have elements of a tri-CRC design. However, they fall short of the label "tri-CRC" because they lack an equitable contribution or use of CRC data, or they introduce unnecessary variability to the development process. The development of VantageScore introduces a much more desirable tri-CRC data design, resulting in the first true tri-CRC score consisting of one scoring algorithm for all three CRCs derived using an equitable contribution of data from all three CRCs. This true tri-CRC data design sets the foundation for any other tri-CRC development or analysis efforts, maximizing the benefits of tri-CRC data for credit grantors and consumers.