Improving information quality in loan approval processes for fair lending and fair pricing

M. Cary Collins, Ph.D.

University of Tennessee
4th Floor, Stokely Management Center
College of Business Administration
Department of Finance
Knoxville, TN 37996-0540
Tel: 865.974.1715
Fax: 865.974.1716
mcollin6@utk.edu

Frank M. Guess, Ph.D.

University of Tennessee
3rd Floor, Stokely Management Center
College of Business Administration
Department of Statistics
Knoxville, TN 37996-0532
Tel: 865.974.1637
Fax: 865.974.2490
fguess@utk.edu

Abstract: Current banking data management on loan approval processes has great room for improvements of information quality and prevention of data problems in general, but especially with regards to fair lending and fair pricing practices. We first review briefly typical data collection protocols deployed at many financial institutions for loan approval and loan pricing. Portions of these data protocols are mandated by federal regulation. In discussing the data capture and analysis for fair lending, we illustrate some initial key steps currently needed in improving information quality for all parties involved.

Refereed paper to be presented at the Massachusetts Institute Technology's Information Quality Conference October 20-22, 2000 of 80 speakers world wide.

Acknowledgements

Dr. Guess gratefully acknowledges financial support from the University of Tennessee College of Business Administration Scholarly Research Grant Program and from the system-wide University of Tennessee Professional Development Award Program. Drs. Collins and Guess express cheerful appreciation to Dr. Paul Bowen, University of Queensland, Australia, Dr. Keith Harvey, Boise State University, and Mr. Ned Clapp, Oak Ridge National Laboratories, for a thorough reading and comments on an earlier draft of this paper.

1. **Introduction**

Bank data management and capture practices, particularly with respect to loan underwriting, have significant room for improvement. For example, most loan underwriting processes require information on the borrower's annual income, his/her past credit experiences, any previous financial defaults or bankruptcies, current & previous address information, and current & previous employment information. These pieces of the underwriting puzzle are, however, routinely expunged from the data warehouse once the credit application is acted upon – leaving only a physical record of these facts in the loan application file.

The weaknesses in these capture and archiving processes are evident in the operational risk area of fair lending compliance. This manuscript describes the most common data capture/management techniques in loan underwriting, as well as the fair lending tests of the underwriting process. We present some recommendations for improving the quality of information capture/management efforts and thus enhancing the fair lending tests as a result. These recommendations, plus the overview of the data processes, should be of benefit to senior risk managers, MIS/IT professionals, financial-services professionals, and regulators.

The remainder of the paper is organized as follows. Section 2 briefly describes the incentives for improving the data collection/management processes at financial-services firms. In Section 3, we briefly describe the typical data collection protocols deployed at most financial institutions for loan underwriting and pricing practices, as well as the regulatory incentives for improving the quality of the information. In Section 4, we relate the most common ways in which these financial institutions determine fair lending and fair pricing compliance, as well as the means by which the data collection phase and the information quality are improved as a result. We summarize and conclude the paper in Section 5.

2. Incentives for improving the data collection/management process

As further evidence of financial-services data shortcomings from a customer's perspective, Dr. Sviokla (1999) described the incredible lack of basic, crucial personal and financial information on him during standard transactions with a New England-based financial firm. Frustrations such as Dr. Sviokla experienced could be frequently prevented with improved information quality systems and a well-deployed strategy of individual service. In relating this banking experience Dr. Sviokla juxtaposed General Motors, L.L. Bean, Amazon.com, *et cetera* and their use (or lack) of information generated in real time or utilized in creating clear market-service advantages.[1]

To overcome information quality deficiencies such as those described by Dr. Sviokla and to create a market-service advantage, Bowen, Fuhrer, and Guess (1998) discuss usage of a continuous process improvement approach of data quality in an accounting and auditing setting. They emphasize the importance of a strong corporate culture and techniques of continuously improving information quality.[2] For a practical application, Dobbins and Guess (1999) discuss early employment of an information quality strategy in HealthMagic, Inc., with their information technology (IT) product HealthCompass™. This IT product was employed

---

1 See Sviokla (1999) and Sviokla (1998).

2 For those new to information quality and its process improvements, please see Redman (1992), Redman (1995), Redman (1997) and Huang, Lee, and Wang (1999).

by clients such as, drkoop.com, Walt Disney World's resident villages, and several hospitals to better monitor and improve the quality of information assimilated on clients, customers and patients.[3]

Regarding client or customer information quality improvements, Collins (2000) points out that the regulators and enforcement agents have escalated testing for data quality in loan underwriting, including tests for patterns in missing data elements by ethnicity, gender, age and income strata. In this compliance environment, the only means by which to reduce the likelihood of unfavorable examination findings for this type of test is to electronically capture the needed data in the underwriting process so as to monitor and improve its quality.

As support for the value of collecting these additional data elements, Collins, Harvey & Nigro (2000) show that data loss, particularly among credit underwriting factors, works to the detriment of low-to-moderate income borrowers relative to upper-income borrowers for home improvement credits. That is, neglecting some qualitative underwriting factors in favor of a "streamlined" credit scoring process raises the denial rate for low-to-moderate income borrowers more than for upper-income borrowers. This disparity in denial rates generates a potential fair lending violation due to poor data warehousing.

There are several benefits to the financial-services firm for retaining information on all credit applicants and for building a better archive – inclusive of the credit history information that is frequently discarded. The principal benefit of retaining these data elements for all credit applications lies in the risk manager's ability to efficiently ascertain compliance and to

[3] Guess (2000) presents more general IQ and IT situations and comments, while Guess and Bowen (2000) focus on the growth and reliability of IT systems.

proactively prevent non compliance with fair lending laws. The disbenefit of retaining this information, the cost of disk storage, has diminished in importance in recent years.[4] Other benefits include building better assessments of the profitability of current and future customers and creating better direct credit marketing approaches that filter on credit quality and past lending experiences.

3. Loan underwriting & pricing policies: Improving data processes

Since financial institutions underwrite several types of loans (e.g., first mortgages, home equity loans, automotive loans, personal unsecured loans, credit card loans), we will focus our illustrations and descriptions on the most homogenous of these loan types, first mortgages. Please keep in mind the mortgage loan underwriting and pricing processes are the "best" lending processes. The other lending channels have problems similar to mortgages but with further quality complications.

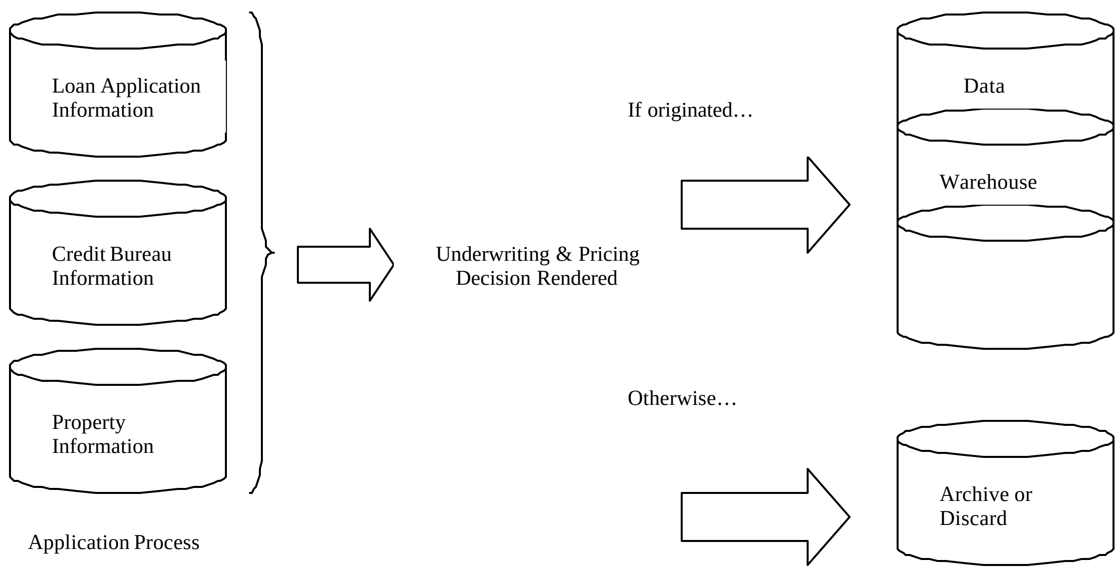
With respect to first mortgages, several data items must legally be captured, including race and gender information on both the borrower and the co-borrower. Additional required information regards the property location, loan amount and application income. A portion of these data protocols is mandated by the Community Reinvestment Act of 1977 and by the Home Mortgage Disclosure Act of 1979. These data elements are typically captured as part of the underwriting process for first mortgages.

[4] For example, there were roughly 25 million mortgage applications filed as part of the Home Mortgage Disclosure Act for 1998. These data collectively occupy only 1.4 gigabytes (GB) of ASCII disk storage. With 20 additional variables appended to each record and pertaining to those critical credit history and financial characteristics, the entire file remains small enough to reside on a 2GB drive.

The mortgage underwriting process. Most consumers will experience the mortgage underwriting process for the first time as they purchase their first residence. The process involves several steps, including: (1) completing a loan application, with details regarding current income, bank balances, requested loan amount and the value of the desired residence; (2) choosing a mortgage loan type (e.g., conventional vs. government insured) and loan term (e.g., adjustable rate vs. 15-year fixed vs. 30-year fixed); (3) verification of the status of employment, residence, current debt payments and liquid assets; (4) assessing the overall creditworthiness of the borrower based on the items listed in (1) through (3) above; and (5) either approving or denying the mortgage credit application based on the overall assessment of worthiness.

The mortgage pricing process. For those approved applications, the next set of choices involves the pricing possibilities for the mortgage credit. Assuming in this example a 30-year fixed rate mortgage credit, the steps include, but are not limited to: (1) determining whether the borrower wishes to "buy down" the interest rate by paying an additional fee called discount points; and (2) deciding whether the closing costs for the loan (e.g., appraisal fees, discount points, origination fees) will be rolled into the total amount borrowed or paid by the borrower at closing.

The mortgage processes from a data perspective. The previous paragraphs provide a glimpse of the steps required in loan underwriting and pricing, as well as the information collected throughout the process. Those descriptions, however, do not explain the information loss associated with loans that are not approved and subsequently originated (i.e., the loan is actually made) nor do they highlight the fair lending compliance difficulties associated with missing or discarded data elements. For example, if the applicant agrees to borrow the funds under the offered price, then the loan is originated. A great deal of performance information is now available on the new loan and its applicant. Unfortunately, most of this information is either discarded or taken "off-line", never making it into the firm's data warehouse. For all other outcomes, there is a higher level of information loss, as shown in the process below:



As such, the data warehouse builds a weak base of information for originated applications, but no information base at all for other credit application outcomes.

In essence, the typical underwriting system and origination system are disjointed and continue to be driven by old habits related to disk storage costs, even in an age when storage is relatively inexpensive. The risk manager must access information on originations from the data warehouse, but must retrieve information on all credit applications other than originated applications from archived storage systems. Archived storage systems, however, rarely retain the most critical information components from a fair lending perspective, the credit history information. He/she must then supplement this electronic information for both originated and non originated credit applications from the physical credit files. This process is extremely expensive and time-consuming.

Since lenders must comply with fair lending and fair pricing laws and regulations, the data shortcomings create a critical risk management weakness. That is, gauging regulatory compliance requires a full set of information on credit applicants, including those who did not receive credit. In fact, the denied credit applications are the primary focus of fair lending rules and yet, these are the credit applicants for whom institutions have the least information under legacy systems.

4. Fair lending & fair pricing practices: Focus on data quality enhancement

In keeping with regulatory guidelines, risk managers create a database of critical information required to comply with the Home Mortgage Disclosure Act (HMDA) and the Community Reinvestment Act. These data are filed as of March 1, each year, with the federal regulators. The data items provide the regulator with a sense of the loan application denial disparities among protected classes of persons, including non minority versus minority borrowers and male versus female borrowers.

Methods of testing for fair lending. Most financial institutions create fair lending datasets using the HMDA data as the base from which to begin, providing a reasonable review of one of the core assets for most financial institutions. These data are enhanced with additional information from the electronic system for originated credits and either electronic or manual-entry information on the denials and other lending outcomes. Once the data are collected and entered, the compliance officer tests the data for integrity and then recreates the lending decisions using both regression analysis and comparative file analysis.[5] In recreating the lending decisions, the compliance officer is testing for two types of fair lending violations: disparate impact and disparate treatment.

Disparate impact in lending is revealed when two similar populations are treated differently as a result of characteristics of a particular loan program. For example, a lender might not capture information on non traditional credit sources – rent payments or utility payments – thereby denying more frequently the group with higher levels of non traditional credit in lieu of traditional credit card, automotive, and mortgage indebtedness. Disparate treatment in lending occurs when similarly-situated borrowers have different lending outcomes. In particular, regulators are wary of situations in which comparably qualified minorities are declined more frequently than non minorities.

Data integrity checks. The compliance officer or fair lending data manager uses simple statistics – average, minimum, maximum, and mode – to get a sense of the range of data in each quantitative element. For example, most loan data entry systems require a numeric entry in each specified field in order for the officer or underwriter to advance the process. If the most common value – the mode -- for a field is “9999”, the compliance officer could reasonably infer that this data element is routinely missing from the credit application field and the entry clerk used “9999” as a means of advancing to the next field. The compliance officer would also test for basic Pearson correlations among these lending variables. For example, the higher the credit

[5] See Black, Collins & Cyree (1997) for the regression model employed by The Board of Governors of the Federal Reserve in testing for surface level evidence of discriminatory behavior.

score, the higher the credit quality. As such, we should see lower overall interest rates for higher credit quality. We will focus our discussion on disparate treatment tests in fair lending.

Determining potential disparate treatment with regression. Risk managers use logistic regression analysis as the modeling, or refining, mechanism in analyzing credit application decisions.[6] The use of regression analysis allows for the contemporaneous control of underwriting characteristics of all applicants within a particular loan program and provides a probability of approval score for every application. This approach enables the risk manager to focus on only those marginal or potentially misclassified applications. That is, the situations in which the logistic regression outcome differed from the actual decision. A review of these situations aids in preventing problems when the internal banking data quality and the model are periodically refined and deployed.

For example, the risk manager would focus his/her time on minority applicants which "scored" more like the approved applicants and who were, in fact, denied credit or are currently slated for denial. The risk manager would use a simple test of proportions to determine whether proportionately more minority applicants fit this profile than non minority applicants. Please keep in mind that the underwriting process can lack consistency without being discriminatory. That is, there might be a high proportion of both minority and non minority applicants in this category, indicating a poor underwriting model with no discriminatory intent.

In pricing analyses, the use of regression analysis allows for the contemporaneous control of pricing attributes for a particular loan program, as well as recognizing the left-

---

[6] Collins & Guess (2000) provide a brief review of both the regression and the comparative file approaches that are most common in fair lending.

truncated nature of interest rates. The regression approach is similar to the one used in underwriting in that it also provides a "difference from expected price" score for every application, leading to an evaluation of significant mis-pricing situations. In particular, the risk manager would focus on potential disparities in the proportion of "over charged" minority versus non minority originations and separately, on the proportion of "under charged" non minorities versus minorities.

In both the underwriting and pricing decisions, the compliance officer must utilize the available data to match the underwriter's decision. Without critical data elements like the applicant's credit characteristics readily available, the compliance officer's reconstruction of the underwriting and pricing outcomes will not "fit" the actual decisions with a high degree of confidence. Adding those additional elements to the data warehouse, rather than losing them once the loan is acted upon, will ensure that the compliance officer gains a better representation of the actual outcomes. With a poor-fitting model, the underwriter must manually evaluate a very high proportion of credits to determine whether fair lending violations occurred. This manual review is costly, entailing a thorough review of the physical credit file. The number of physical file reviews can be dramatically reduced with a better fitting model.

Detecting potential disparate treatment with cohort analysis. Compliance officers use cohort, or comparative, file analysis as the microscope in credit application decisions to evaluate each minority denial relative to comparable applicants. In each instance, the compliance officer is creating two test groups: (1) of applicants with characteristics similar to the target denied applicant and (2) of originated credits with credit and financial characteristics which are equal to or worse than those of the target denied applicant.

In this first test, the compliance officer is eager to determine whether the proportion of denials for the majority comparators is similar to the proportion of denials for the minority comparators. That is, if there are 40 majority denials among 100 majority comparators, then the proportion of denials is 40:100 or 2:5. If there are 20 minority denials among 30 minority comparators, then the proportion of denials is 20:30 or 2:3. The statistical question for the compliance manager is whether these two proportions are statistically different, indicating that the minority target received a treatment dissimilar from the treatment received by members of the majority.

In the second test, containing only those approved credits of lower quality than the denied target application, the compliance officer is eager to determine whether there are approvals for the majority comparators with poorer credit quality than our target minority denial. For example, the compliance officer might find that for a similar loan purpose, a target minority has a debt-to-income ratio of 42 percent and has no delinquencies in his/her credit history. If the examiner found an approved majority comparator with a debt-to-income rate of 50 percent and three credit delinquencies in his/her credit history, then an instance of inconsistent treatment is exposed. The root cause of the differential treatment might arise from poor loan officer training. In this instance, one loan officer better understands the guidelines than another. The outcome might stem from unclear underwriting guidelines allowing the underwriter some discretion in what constitutes an approval. The situation could obtain from a poorly structured test in which the tester lacked enough information on compensating factors to make an accurate determination of similarity. That is, the poor credit quality of the approved majority comparator might have been offset by other credit or financial strengths. Finally, the result may also derive from a discriminatory situation. Increasing the amount and quality of credit information available to the compliance officer would enhance the sensitivity of these two tests, while reducing the effects of different types of errors.

In pricing comparative file analysis, the risk manager's approach is similar to the comparative file approach detailed above. The risk manager's objective is to find similarly-situated, originated credits and test whether the cost of credit (e.g., interest rate, APR, appraisal fees, broker fees) charged to minority applicants is higher than the rates charged to majority applicants. The most commonly applied methods here involve simple difference of means tests (i.e., Do similarly situated minority borrowers pay more than majority borrowers on average?) and tests of individual applicants against the similar population (i.e., Did this minority borrower pay more for credit than the comparable majority borrowers?). Detecting and correcting pricing deficiencies is a continuous improvement process: As the pricing process is refined to eliminate disparate treatment, the volatility or variance in the distribution of interest rates charged will also decline. As the number of credits sampled is increased to an evaluation of the population, the significance of even these smaller variances will increase.

Evaluating the physical files. For the credit applications that warrant further review, either for underwriting deficiencies or for pricing practices, the compliance officer evaluates the physical file for any differences between the electronic information they have and the information the underwriter used to make his/her determination. It is possible that the underwriter adjusted figures on the Loan Application Worksheet, but did not make those changes on the electronic record, resulting in a misclassified applicant and additional work for the compliance officer due to poor information quality.

Recapitulation. The risk manager evaluates both the underwriting and pricing practices for two types of disparities. First, with respect to disparate impact, the risk manager works to determine whether the underwriting (or pricing) criteria systematically denied more (charge more to) minority applicants than majority applicants. This finding of disparate impact would begin an inquiry of the underwriting or pricing process to determine which factor or combination of factors is generating the penchant for denying or overcharging minority applicants.

Second, with respect to disparate treatment, the risk manager uses the regression results as his/her backbone for investigation, but will additionally perform a cohort, or comparative, file analysis as affirmatory evidence of discrimination. At this individual credit level, the officer will review the physical file against the electronic record used in testing and note any differences. Additionally, the compliance officer will typically work to determine whether the deficiencies in the data form a pattern connected to a particular underwriting center, branch, loan officer or loan broker. These regulatory incentives create a great motivation for financial institutions to improve their information quality, reducing the likelihood of costly regulatory and legal inquiries for discrimination.

5. Summary & Conclusions

We stress the importance of a data quality strategy regarding the underwriting and pricing of mortgage applications. Improving information (knowledge) quality is an important strategy for financial institutions to embrace for better customer service, prevention of legal violations, building of brand loyalty, et cetera. We have presented the typical data capture process at lending institutions, demonstrating the deficiencies of such an approach from a data quality and efficiency perspective. We have also demonstrated the approach to fair lending risk management most often adopted by lending institutions and with an understanding that the most critical piece of the fair lending compliance puzzle is the capture and inclusion of information on credit history and related financial characteristics.

References

Black, H., Collins, M. C., and Cyree, K. B., (1997), "Do Black-owned Banks Discriminate Against Black Borrowers?", Journal of Financial Services Research, Volume 9, No. 1 & 2.

Bowen, P., Fuhrer, D., and Guess, F. (1998). "Continuously Improving Data Quality in Persistent Databases," *Data Quality* @ www.dataquality.com/998bowen.htm. (See www.dataquality.com/ in general.)

Collins, M., (2000), "A Note on Risk-based Pricing," www.bankinfo.com/compliance/compliance.html. (See www.bankinfo.com in general)

Collins, M. and Guess, F., (2000), "Statistical methods for improving information quality in loan approval processes," University of Tennessee Technical Report (forthcoming) Contact Collins for a pre-print.

Collins, M., Harvey, K., and Nigro, P., (2000), "The Influence of Bureau Scores, Customized Scores and Judgmental Review on the Bank Underwriting Decision Making Process," Working Paper, The Office of the Comptroller of the Currency and The University of Tennessee.

Dobbins, J. G. and Guess, F., (1999), "Developing a data quality strategy in a provider of Web based health information systems," Information Quality Conference in 1999 at MIT's Sloan School of Management Proceedings, pp. 176-184.

Guess, F., (2000), "Improving Information Quality and Information Technology Systems in the 21st Century," invited talk and paper for the International Conference Statistics in the 21st Century, June 29 to July 1, 2000.

Guess, F., and Bowen, P., (2000), "Growth and Reliability of Information Technology Systems," International Journal of Reliability and Application, invited paper (forthcoming).

Harvey, K., Collins, M. C., Nigro, P., and Robinson, B., (2000), “Determinants of Denial Disparities in Home Mortgage Lending: Bank Performance, Macro-economic Factors and Regulatory Influence,” Journal of Real Estate Finance & Economics (forthcoming).

Huang, K-T., Lee, Y. L. and Wang, R. Y., (1999), Quality Information and Knowledge. (New York: Prentice Hall).

Redman, T. C., (1992), Data Quality, Management and Technology, (New York: Bantam Books).

Redman, T. C., (1995), “Improve Data Quality for Competitive Advantage,” Sloan Management Review.

Redman, T. C., (1997), Data Quality for the Information Age. (Artech House Computer Science Library).

Sviokla, J. J., (1999), Keynote speech at the Information Quality Conference at MIT’s Sloan School of Management: “Customer Data and eCommerce.” [See web.mit.edu/tdqm/www/speech/sld001.htm for his talk’s PowerPoint slides or, in general, his web site: www.sviokla.com/.]

Sviokla, J. J., (1998), "Virtual Value and the Birth of Virtual Markets," Chapter 10, Sense & Respond: Capturing Value in the Network Era, coedited by Stephen P. Bradley & Richard L. Nolan, (Boston: Harvard Business School Press).