Moody's | K·M·V

EXAMPLES OF OVERFITTING ENCOUNTERED WHEN BUILDING PRIVATE FIRM DEFAULT PREDICTION MODELS

OVERVIEW

AUTHOR

Douglas W. Dwyer

CONTACTS

André Salaam
+1.212.553.4517
andre.salaam@mkmv.com

The key to building default prediction models, if they are to be incorporated into credit risk management systems, is to build the most powerful model possible subject to the constraints that it is transparent, usable, and intuitive. In this process, we must constantly be on guard for whether or not we have overfit the data. In this paper, we present two examples of overfitting that we encountered while building private firm models. These issues, if not detected, would have reduced the usability of the model and overstated the true predictive power of the model in real credit decision-making.

TABLE OF CONTENTS

1 INTRODUCTION

Default probability models are playing an increasingly central role in the credit risk management systems of financial institutions. In order to be useful, such models need to be predictive of default for both a firm's current set of borrowers as well as for the new potential borrowers. Additionally, they need to be transparent, intuitive and understandable. In building such models, one needs to constantly be on guard against overfitting the data. A model has overfit the data if it contains an implication that is not reflective of an underlying behavioral relationship, but rather the result of the statistical method used to estimate the model, or an idiosyncratic feature of the data used to build the model, or both.

There are several types of overfitting. Perhaps the best understood is the so-called curse of dimensionality that results from estimating a model with a highly flexible functional form (cf., Haykin, 1999 or Hastie and Tibarshirani, 1990). Another type of overfitting results from testing many possible predictors of default and choosing the ones that work best on a specific sample.[1]

A third type of overfitting results from fitting a feature of one's data collection mechanism rather than an underlying economic relationship. This type of overfitting is quite tricky to detect with statistical hypothesis testing alone. If one fits a feature of one's data collection mechanism, then such a feature will be statistically significant with a sufficiently large sample. If the same data collection mechanisms are used to construct the validation (or hold out) sample, then out-of-sample testing may confirm the existence of the feature as well.

When building RiskCalc models, we uncovered two examples of overfitting that were due to specific features of the data collection mechanism. These examples are described in this paper. The first example involves fitting an asymmetry in the collection of specific fields for defaulting versus non-defaulting firms during certain time periods. The result of the issue would have been to include a seemingly quite powerful variable in the model that does not actually predict default, thereby reducing model power, transparency and usability. The second involves inadvertently incorporating a common sample selection bias into the default prediction. The result of this issue would have been to systematically understate the risk of certain loans in one's portfolio.

---

[1]These two types of overfitting are not unrelated. For example, one could estimate a second order approximation to a general function with 10 independent variables (estimate 10 linear terms, 10 quadratic terms and 45 interaction terms), which would be a highly flexible functional form. Alternatively, one could perform a stepwise procedure over these 65 independent variables and only keep the ones that are statistically significant. While the final regression estimated under the second approach may not contain a large number of parameters, the probability of a false positive (keeping an independent variable in the model when in fact it does not represent an actual behavioral relationship) has clearly increased. The test statistics no longer have the distributions that are assumed for standard hypothesis testing (cf., Greene, 2000, page 334).

2   OVERVIEW OF THE RISKCALC MODELING PROCESS

Currently, there is a common RiskCalc modeling framework that has been implemented for more than 20 countries collectively representing over 80% of the world's GDP. Each model implementation is based on actual private firm data from each respective country. Each model takes into account the local accounting practices of the country. We have updated seven models with new versions of RiskCalc fitted on larger and more current databases using an enhanced modeling methodology. The scope of the RiskCalc models is private firms excluding finance, real estate, and insurance companies, as well as not-for-profit and government agencies.

Before beginning the modeling process for a new default prediction model, we first build a database that combines firm-level financial statement information with the corresponding default history for the firm. Each data set has its own characteristics. Understanding the strengths and weaknesses of these data sets is a key part of the modeling process.

There are several steps to our modeling process. The first step is to transform financial statement ratios using a mini modeling procedure. The second step is to select a set of 7-15 financial statement ratios from a large set of possible financial statement ratios.[2] The third step is to estimate a probit (or a logit) model, which is perhaps the core step of the modeling process. At this step, we regress a default flag onto a set of independent variables that are typically transformations of financial statement ratios. The final step is to perform an empirical mapping of the estimated default likelihoods into an actual Expected Default Frequency (EDF™).

Our standard practice is to take an annual financial statement as the unit of observation and to define the default flag on the basis of a default event occurring during a certain window following the financial statement. For example, we may set the default window at 3-15 months following the financial statement date.[3] The implicit assumption is that the financial statement is released to the financial institution approximately three months after the end of the fiscal year. We eliminate financial statements that are dated either within three months of the default event, as the firm would have already been in default by the time the bank received the financial statement; we also eliminate financial statements that are dated after the default event. We test many variations of this approach.

Model validation plays a key role in this process. We first validate the model by looking for consistency in the rank ordering ability (the power) of the model across different cuts of the database: these tests determine whether the model effectively differentiates between high and low risk firms.[4] But because we can get apparently strong power that later fails to materialize out of sample, we analyze model power across time, industries, and size classifications. We also compare power across data sources when appropriate (often our analysis data set is the combination of multiple data sets provided by different financial institutions). We focus on power relative to a benchmark rather than the absolute power, as the measure of absolute power cannot be compared across datasets. Useful benchmarks are those models with well-established performance histories, such as Altman's Z-score and earlier generations of the model. Additionally, we test for multicollinearity through correlation matrices and variance inflation factors to understand if several input ratios are capturing the same information and muddying our ability to interpret drivers of risk. After we are confident that we have a solid model, we do out of sample testing using walk-forward and K-fold analyses.

---

[2]Any textbook on financial statement analysis will contain numerous financial ratios that can be used to assess the risk of a firm. Our working list contains over 100 such ratios.

[3] Depending on the data set, we will often work with a somewhat longer window for our 1-year model (e.g., 3-24 months). This longer window manages two data issues: (i) frequently, firms do not submit their last financial statement prior to default to their bank; and (ii) often loans have been non-performing for some time prior to when they are classified as a default.

[4]One of our tools for measuring the rank ordering ability of a model is the cumulative accuracy profile (CAP) and the corresponding accuracy ratio (AR). It has become the convention in the field to refer to the rank ordering ability of the model as the discriminatory power (or simply the power) of the model. Engelmann et. al., (2003) provide a description of CAP curves and their relationship with ROC curves. Regardless, CAP and ROC curves provide essentially the same characterization of a model's power.

Moody's | K·M·V

3   OVERFITTING DUE TO AN IMPROVEMENT IN DATA COLLECTION

When building a private firm model, we try to use as long a time series as possible to incorporate a full credit cycle into the analysis. Often, there is a decade or more between peaks in the credit cycle. In the case of RiskCalc 3.1 France, we had 10 years of data (Dwyer and Wang, 2005). One issue with a long time series, however, is that the data collection mechanism will almost always evolve over time. With the advent of modern credit practices and the push of Basel II, data collected in more recent years is consistently more complete. Such changes in the data collection mechanism must be well understood or the default model may fit these changes rather than an actual underlying behavioral relationship.

We encountered a fascinating example of this phenomenon in developing RiskCalc 3.1 France. Several accounting fields were consistently collected for non-defaulting firms, but were only collected for defaulting firms after 1997. We inadvertently used this information and, not surprisingly, it was a powerful predictor in the sample data set of whether or not a firm would default. Our validation process identified the issue.

3.1   The Finding

In the process of developing RiskCalc 3.1 France, we tested a model that yielded a huge increase in overall power relative to the earlier version of the model (RiskCalc 1.0 France). The gain in Accuracy Ratio (AR) of over 10 points (an AR of 81.3% versus an AR of 70.2% for RiskCalc 1.0 on the same sample) caught the immediate attention of our validation team. The first clue came when we tested for this increase across different cuts of the data. The data on power changes over time showed that the power increase was entirely due to model performance prior to 1997 (Figure 1). The new model offered almost no improvement in the more recent time periods.
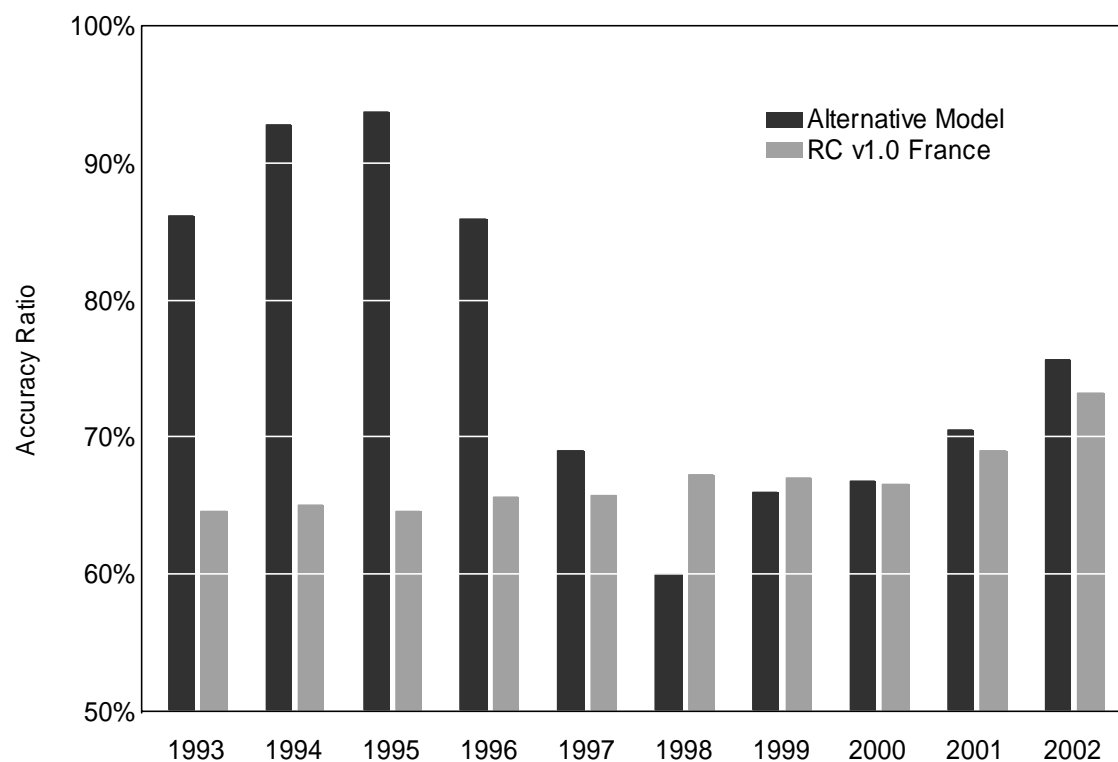

FIGURE 1   Accuracy Ratio by Year for an Alternative Model versus RiskCalc v1.0 France

3.2   The Explanation

We quickly identified that the variable "C08" was the source of the phenomenon. When this variable was dropped from the model, the power increase prior to 1997 disappeared. Further, the overall performance of the model was comparable with that of RiskCalc France v1.0. C08 is a measure of operational efficiency of the firm that captures how much additional expense (beyond materials) the firm needs to employ to generate a unit of sales. In accounting terms this variable is defined as:

$$C08 = \frac{\text{Operating Expenses - Sum(Materials, 0)}}{\text{Sum(Net Sales, } \Delta \text{Inventories, Capitalized Work, Other Operating Revenue)}}$$

where the sum function treats the missing values as 0. When the materials field is missing, operating expenses are typically close to net sales and C08 is close to one. The decision to treat missing values as 0 is one of the numerous small decisions that are made during the modeling process. Alternatively, we could have treated the ratio *C08* as missing any time any of the inputs were missing, but we chose not to at that time.

Table 1 presents a position analysis for two fields (Materials and Capitalized Work) in our French private company database. Each row reports the percent missing for financial statements in each year with separate columns for non-defaulting firms and firms that eventually would default. These two fields are four to nine times more likely to be missing for defaulting firms than for non-defaulting firms prior to 1997. This is a data collection issue that occurs because these statement item fields were not archived for defaulting firms prior to the statements being purged from the system. Since statements for non-defaulting firms remain in the system, these fields were not missing for active firms. These fields were not missing for recent defaults because the archiving process was expanded to capture these fields after 1997. Given the special nature of this issue, it can easily go undetected when looking at the data prior to beginning model estimation.

TABLE 1   Position Analysis for Two Fields in our French Database

	Materials		Capitalized Work	
	% Missing For Non-Defaults	% Missing For Defaults	% Missing For Non-Defaults	% Missing For Defaults
1993	16.4%	74.5%	16.4%	74.5%
1994	7.1%	60.8%	7.1%	60.8%
1995	5.1%	44.0%	5.1%	44.0%
1996	3.3%	25.9%	3.3%	25.9%
1997	1.0%	6.4%	1.0%	6.4%
1998	9.2%	1.0%	9.2%	1.0%
1999	0.4%	0.2%	0.4%	0.2%
2000	0.0%	0.0%	0.0%	0.0%
2001	0.0%	0.0%	0.0%	0.0%
2002	0.0%	0.0%	0.0%	0.0%

The distribution for the variable C08 for defaulting and non-defaulting firms pre and post 1998 is represented with box plots in Figure 2. Note that for defaulting firms the pre-1998 distribution is *above* the other distributions and the inter-quartile range is narrowly spread around one. The distributions are *lower* (typically less than one) and more disperse for both defaulting and non-defaulting firms after 1998 and for non-defaulting firms prior to 1998. This result is directly related to missing the materials and capitalized work fields for defaulting firms prior to 1998 (Table 1). When the materials field is missing, operating expenses is typically close to net sales and C08 is close to one. In other words, a large share of defaulted firms gets flagged with what is effectively a dummy variable for default. Therefore, the predictive power of this variable is largely the result of the data collection issue. Including this variable in the model yields a model

whose predictive power estimated on this data set will substantially overstate its predictive power in practice. Our validation methodology identified the problem. The solution to the problem was to revisit the variable selection process excluding all ratios that included either materials or capitalized work in their definitions. The final result was a model (RiskCalc 3.1 France) that yielded a more reasonable three-point increase in power over the previous generation of the model, (RiskCalc 1.0 France).
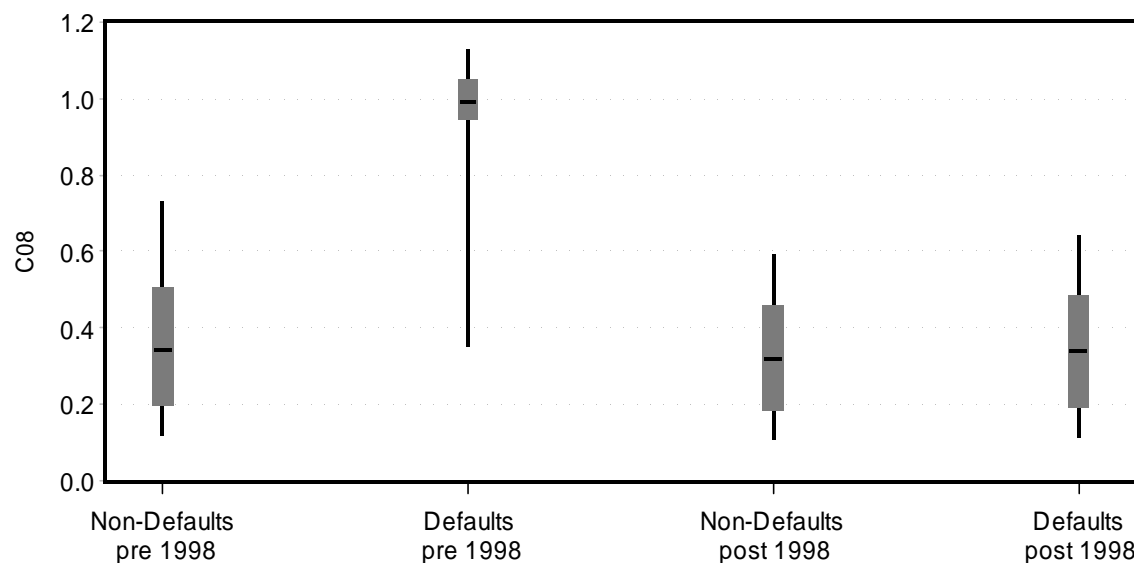

FIGURE 2   Distribution of C08 for Defaulting and Non-defaulting Firms

Presents box plots representing the distributions of C08 for defaulting and non-defaulting firms pre- and post-1998 (post-1998 is inclusive of 1998). The bottom and top of the solid black lines of the box plot represent the 10th and 90th percentiles for the distribution. The box represents the inter-quartile range, and the horizontal bar in the middle represents the median of the distribution. The box plot reveals that the variable C08 was very predictive of default prior to 1998 due to a data collection issue.

3.3   The Implication

Incorporating a model that used C08 into a credit risk management system would have two negative consequences. First, the variable was not predictive of default after 1998, so incorporating it into an estimated probability of default is effectively adding noise to the model's output, which would make the model less predictive than it could be. Second, the variable C08 is difficult to understand and to explain. Lenders and credit analysts will do pro-forma analyses and discover that minor changes in operational efficiency overwhelm leverage, size, and cash flow impact on creditworthiness. Having such a variable in the model would unnecessarily reduce the model's usability and transparency and increase the support cost for the internal ratings process.

4 CORRELATION IS NOT CAUSATION

Imagine a risk management or internal rating system where new borrowers that only provide one financial statement are automatically upgraded by about three rating notches above equivalent firms that gave a three year history. In such a system, a borrower that would otherwise be a Ba2 becomes a Baa2, solely because it is a new borrower with only one financial statement. Such a system seems very counterintuitive and opens the door to some serious adverse selection behavior. Nevertheless, such a system would yield a significant increase in power as measured by standard practices.

4.1 The Finding

We could create a substantial increase in the power of RiskCalc simply by flagging whether or not the statement is the first, second, or third financial statement for each firm in our database (Figure 3). Specifically, we estimate two models (Model 1 and Model 2) using the most recently available Credit Research Database (CRD) database for the U.S. following the methodology used in developing RiskCalc 3.1 (Dwyer and Kocagil, 2004). Model 2 uses the same independent variables and the same transforms as RiskCalc 3.1, but on the new data set.[5] Model 1 is identical to Model 2 except that Model 1 adds the following 3 indicator (dummy) variables:

$$\text{FirstObs} = \begin{cases} 1 & \text{if no prior financial statements for firm in database} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{SecondObs} = \begin{cases} 1 & \text{if one prior financial statement for firm in database} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{ThirdObs} = \begin{cases} 1 & \text{if two prior financial statements for firm in database} \\ 0 & \text{otherwise} \end{cases}$$

Each of these three indicator variables has a negative coefficient that is highly significant. The negative coefficient implies that if RiskCalc gets a flag that it is being run with input from a first statement, for example, the probability of default is lowered significantly. Including an indicator variable for the fourth financial statement yields an insignificant coefficient. The accuracy ratio of Model 1 is 3.5 points larger than Model 2 (Figure 3).

Table 2 demonstrates the implications of these indicator variables on a typical firm. We take a financial statement that would have an EDF level of 1.7% from Model 2 (standard RiskCalc approach), which corresponds to an EDF-implied rating of a Ba3.[6] If RiskCalc had a flag that told it there were no prior financial statements in the database for the firm, as it does in Model 1, then the EDF level would be reduced by 70% and the EDF-implied rating would be upgraded by 3 notches (Ba3 to a Baa3). If Model 1 knew that there was just one preceding financial statement, the EDF level is still reduced by more than 58%, which is a two-notch upgrade. Given the information that there are two prior financial statements, the EDF level is reduced by 40%, which is also a two-notch upgrade. Using Model 1 would have a fairly dramatic impact on the credit risk of new exposures in a portfolio, all because most initial loans are made with a requirement of one to two trailing financial statements.

---

[5] There are two notable differences in the methodology implemented here versus how we developed RiskCalc 3.1 U.S. (Dwyer and Kocagil, 2004). First, here we included the first financial statement for each firm whereas in estimating RiskCalc 3.1 we eliminated it. Second, here we defined default as a default event within 3-15 months following the financial statement date. In estimating the one-year model in RiskCalc 3.1, in contrast, we defined default as a default event within 3-24 months following the financial statement date. In developing RiskCalc 3.1, we excluded the first financial statement and used a long window for our default definition to manage these sample selection issues.

[6] RiskCalc 3.1 U.S. is designed to produce an average EDF value of 1.7% in FSO mode for the one-year model. RiskCalc models map EDF credit measures to EDF-implied ratings using a static mapping. An EDF-implied rating allows one to interpret an EDF credit measure within a conceptual framework (an agency rating) that many RiskCalc users find more familiar than an EDF credit measure.

TABLE 2   Impact of Controlling for Position of Financial Statement in the Database
Holding All Other Variables Constant

Position of Financial Statement	Estimated EDF Credit Measure	EDF–Implied Rating
No Prior Financial Statement	0.49%	Baa3
One Prior Financial Statement	0.75%	Ba1
Two Prior Financial Statements	1.02%	Ba1
Other	1.70%	Ba3

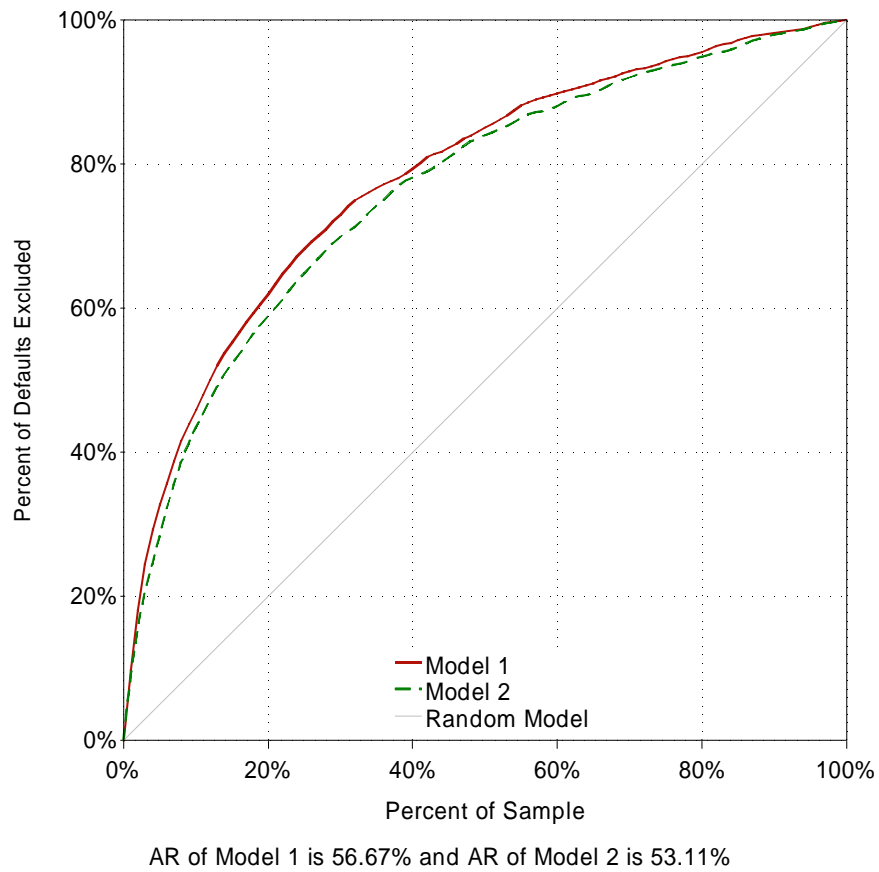AR of Model 1 is 56.67% and AR of Model 2 is 53.11%

FIGURE 3   Power Curve Comparison Between Two Versions of RiskCalc 3.1

Both Model 1 and Model 2 are estimated on the most recently available data using the same variables included in RiskCalc U.S. 3.1. Model 1 includes controls for first, second, and third financial statement for the firm in the database, while Model 2 does not. Model 1 reduces the default risk for the first three financial statements of each firm in the database.

4.2　The Explanation

There is a simple explanation for the finding: sample selection bias. Thanks to common lending practices, firms cannot default on their first financial statement by construction and most cannot default on their second or third.

As a best practice, when banks originate a new loan, they will ask for the two most recent financial statements in the loan application and enter both into their systems.[7] In many cases it will be impossible for a firm to default within 15 months of the first financial statement, because there is not yet a loan outstanding to default on. Even if the firm did have existing loans, the possibility of a firm defaulting within 15 months of the second financial statement will often be limited because it would be defaulting within a few months of having applied for and received a line of credit. For example, suppose one applied for a loan at the end of September of 2003 with financial statements dated December of 2001 and December of 2002. In order to default within 15 months of the second financial statement, one would have to default before April of 2004, which is within 6 months of applying for the loan.

Immediately following the origination of a loan, defaults will typically be low. The reason for this is clear: if a borrower did not have the financial wherewithal to make the first 6-12 months of payments, then the financial institution probably would not have made the loan in the first place (cf., Keenan, 1999, and Duffie and Singleton, 2003 pages 79-84, discusses the issue). There are cases where this is not true (such as Just for Feet which defaulted before making the first coupon payment on its bond) but they are memorable for their rarity. So by looking at historical lender data to build models, we have sample bias in the statements available at the moment of origination. If we could observe those firms that were declined a loan on the basis of the same 2 statements, we would then find a biased sample of poorly performing firms. That is the job of the model *going forward*: to take a broad group of applicant firms, both those that should be accepted and those that should be declined, and accurately differentiate between them.

4.3　The Implication

The correlation between very low default rates and first, second, and third statements is absolutely real. It comes from trying to predict the future by fitting a model on a biased historical dataset, and letting predictive power override business intuition. Credit analysts, relying on common sense over statistical fitting, generally prefer a firm that can provide a longer financial history over one that cannot.

Implementing an apparently highly predictive model which prefers a borrower with one financial statement to one with three statements would create a number of problems:

- First, it is clear that one would systematically understate the risk of these firms due to the sample selection bias.

- Second, the credit risk assessment produced by the model would be counterintuitive to the users of the model, which would create confusion. Credit analysts would be reluctant to incorporate such a model into their decision making process.

- Third, because of the non-intuitive nature of this result, convincing a regulator of the validity of such a model would pose a challenge.

- Fourth, it would be easy for an opportunistic lender to "game the model" by neglecting to incorporate all the available financial statements into the analysis.

This issue could unintentionally impact a model in a number of ways. For example, one can set sales growth to 0 whenever it is missing and include a dummy variable for missing sales growth, which is one technique for handling

---

[7]The extent to which this practice is implemented varies both across financial institutions as well as within financial institutions.

Moody's | K·M·V

missing data.[8] Such a dummy variable would effectively proxy for the absence of a prior financial statement thereby generating the same issue. In developing our models, we manage this sample selection issue carefully. First, we do not use the number of financial statements available for a firm as a predictor of default. Additionally, we typically eliminate the first observation from the analysis. Finally, our models are built to be robust with respect to input data and generate PDs in the absence of prior year financial statement information without rewarding an analyst for failing to provide such information.

5   CONCLUSION

In this document, we provided two examples of overfitting that we encountered in our experience of building, calibrating, and validating private firm default prediction models. Fortunately, our model development and validation process ensures that such issues are detected prior to release of a model. There would have been significant costs to users had such models been widely rolled out into risk management systems.

As a postscript, we encountered the efficiency variable C08 in France in one other context: industry modeling. The RiskCalc 3.1 model was built using our standard approach to capture industry effects: a single indicator variable for each industry. After completing RiskCalc 3.1, we tested whether or not building individual industry-specific models would improve model performance, and if it did, whether or not the performance increase would warrant the difficulties for users to feed and implement such models and train users.

Unlike many regions, in France, we have sufficient data coverage to develop a separate model for each industry. We were able to estimate a specific model for each industry. By separate model, we mean a different set of independent variables and parameter estimates for each industry. On our first pass, we found a very substantial increase in power and the variable C08 showed up in every industry model. On our second pass, after removing C08 from the set of potential explanatory variables, we found only a minimal increase in model power.

There are other examples of overfitting that we have encountered. One was an apparent relationship between a macro factor and default that we encountered in the early stages of building RiskCalc U.S. 3.1. There were a substantial number of defaults from one financial institution that appeared to occur on a specific date in our database. These defaults did not actually occur on this date but at some time prior to this date. In the process of archiving the data, the actual default date was lost. The institution recorded all those defaults with a common default date at the time the bank began to use this data, the "as-of date." This data collection problem obviously led to a spike in the default rates in our sample. It was very easy to find a macro variable that would predict these defaults because they peaked at the date when the defaults were recorded as occurring, though in reality, they were distributed over a long period before that. This resulted in a very substantial increase in model power. After determining that the increase in model power was coming largely from these defaults, we decided to drop these defaults from the analysis database.

A certain amount of skepticism is appropriate when a new modeling methodology yields a large increase in power. The increase in power will often be the result of fitting the data collection mechanism rather than an actual underlying behavioral relationship. In addition, these issues are clear afterward, but were not turned up in the ordinary pre-modeling data-cleansing process that we had in place at the time. Work like this has guided us as we extend our cleansing and validation processes. Using the model that results from these problems within a risk management process can potentially be quite costly. At best, overfitting will introduce unneeded complexity into the model or discredit it with users. At worst, it can lead to systematic error in the risk assessment of a portfolio.

---

[8] This is one of many possible approaches for handling missing data in regression analysis (cf. Little and Rubin, 1987).

6 REFERENCES

1. Basel Committee on Banking Supervision. "International Convergence of Capital Measurement and Capital Standards (A Revised Framework)." Bank for International Settlements, 2004.

2. Duffie, Darrell and Kenneth J. Singleton. *Credit Risk: Pricing, Measurement and Management*, Princeton Series in Finance. Princeton: Princeton University Press, 2003.

3. Dwyer, Douglas and Ahmet Kocagil. "Moody's KMV RiskCalc 3.1 United States: Modeling Methodology." New York: Moody's KMV, 2004.

4. Dwyer, Douglas and Yi-Jun Wang. "Moody's KMV RiskCalc 3.1 France: Modeling Methodology." New York: Moody's KMV, 2005.

5. Engelmann, Bernd, Evelyn Hayden, and Dirk Taschen. "Testing Rating Accuracy." *Risk* (2003).

6. Greene, William H., *Econometric Analysis*. Fourth Edition, Upper Saddle River, NJ: Prentice Hall, 2000.

7. Hastie, Trevor J.; Rob J.Tibshirani. *Generalized Additive Models*. Vol. 43, Monographs on Statistics and Applied Probability. London: Chapman & Hall/CRC, 1990.

8. Haykin, Simon. *Neural Networks: A Comprehensive Foundation,* Prentice Hall, 1999.

9. Roderick, J.A. Little and Donald B. Rubin. Statistical Analysis with Missing Data, Wiley Series in Probability and Mathematical Statistics: John Wiley & Sons, 1987.

Moody's | K·M·V