Insufficient Data for Forecasting Probability of Default - Challenges and Suggested Solutions

By Chandrakant Maheshwari and Bhaswati Sengupta

Table of Contents

Abstract

Precise calculation of reserve is very important for a loan portfolio to cover the expected loss within next one year. If we overestimate the reserve we lose the profitability of our business and if we under estimate it we may risk in ending up in a financial ruin. For this very reason a risk manager should have complete historical loss data of the loan portfolio to have an accurate measure of reserve. In spite of having complete and accurate information about losses in history; due to the way we define data collection we have incompleteness in the data. We highlight and explain this problem and suggest approaches the risk manager can take to overcome this problem.

Introduction

Financial Institutions allocate a reserve for their loan portfolio. This reserve is allocated to cover their expected losses incurred in the portfolio for the next one year. Expected losses occur due to change in market conditions, credit risk and operational risks. From the credit risk perspective expected loss is because of the possibility of default from the obligors, delay in interest payments and change in credibility of obligors within next one year. Hence expected loss of a loan portfolio is dependent on Probability of Default (PD), Loss Given Default (LGD) and Exposure at Default (EAD). Clearly PD is dependent on the change in possibility of default from the obligors due to any delay in interest payments and change in credibility of obligors within next one year.

Given that the expected loss is for next one year, the variables (PD, LGD and EAD) on which it depends on are projected for the same time. To find the projected value we have to implement a forecasting process. This will involve choosing a methodology/model and calibrate its parameters using available historical data. For correct implementation of the forecasting process, it is important that the historical data we have is accurate and complete or else the whole exercise will go for a toss.

As reserve is calculated for yearly horizon; PDs are forecasted for yearly horizon. For this reason we need historical PDs on yearly bases. We can use monthly data loss to project yearly PD but monthly loss data has high volatility. This may be due to seasonality in the data and many other reasons. Going deeper into the reasoning of volatility is not the scope of this article. Overall it is preferred to have historical PD data on yearly basis.

In the current article our focus will be looking into the completeness of historical PD data. We will see that how in spite of having complete and accurate data for the available history; just because of the definition of measurement, data in hand is incomplete. We will see how this incompleteness arises and how this issue can be addressed. Our focus will be to overcome this limitation but not in suggesting which is the best forecasting methodology.

Data Measuring Methodology and Inbuilt Limitations

The loss data is collected on monthly basis based on events like default, delay in interest payments and change in obligor ratings occurred on the given month in the loan portfolio in hand. This data is converted to the yearly data by summing up the loss numbers of next 12 months. Once we have the actual loss numbers for next one year we derive actual PD for next one year associated with each month.

Suppose we are in month of January 2014, we have next 12 month loss data for the month for December 2012. But for the month of January 2013 we have loss data only for next 11 months, for the month of

February 2013 we have the data for next 10 months and so on for the month of November we have just the data for 1 month. Hence yearly loss data from January 2013 to November 2013 is not complete.

When we are in January 2014, we have to forecast the PD values for next one year. Limitations arrive because we have complete yearly data for timestamp till December 2012 only. In other words we are currently in January 2014 beginning and have to forecast the PD for 2014 but because of the definition we only have complete data till December 2012.

Solutions to Handle the Problem of Incomplete Data

There are some solutions which can be employed to encounter the problem discussed above. Assume we are at January 2014; we have monthly loss data till December 2013.

Situation in hand:

1) We have complete yearly data till Dec 2012 (without loss of generality all that matters to us is when the data ends without bothering when it starts)
2) January 2013 time stamp is incomplete by one month loss data
3) February 2013 time stamp is incomplete by 2 month loss data
4) Similarly November 2013 time stamp is incomplete by 11 month loss data

This incomplete data can be approximated by the following steps:

1) For January 2013 time stamp: We use the historical monthly loss till December 2012, we take the median of the historical data to this time stamp.
2) For February 2013 time stamp: We make a series of continuous two monthly loss values from the historical data till January 2013. Take the median of this series and add to the February time stamp.
3) For March 2013 it will be 3 monthly loss series till February 2013 and its median will be added, for April 2013 it will be 4 monthly and so on.

Justification of the assumptions:

1. *We do not require 'n' monthly loss data to be non-overlapping. That is because if we see the two consecutive months of yearly loss data (of time stamps with complete data); it already has an overlap of 11 months.*
2. *Reason of choosing median: Every monthly loss is considered as an event. Every event occurred in history is assumed to have equal probability. The events are sorted as per the size of the loss amounts and median is chosen as it represents the average value from the perspective of probability.*
    a. *We may choose event in the form of loss value or loss percent.*

The other option we have is if we observe the seasonality in this historical data we can take average of all the January loss number, February loss numbers and so on and add to the incomplete time stamps appropriately.

Other approach can be, as discussed we go for yearly data because we have high volatility in monthly data. Instead of choosing yearly loss numbers, we can have a series of quarterly or semiannually loss numbers that ensure manageable volatility in the series to implement the forecasting methodology. This way we will

have lesser time stamps with incomplete data. The rest of the process remains the same as discussed above.

Once we have complete historical data series we can choose any forecasting methodology which satisfies the calibration criteria.

An Alternate Solution

An alternate solution will be to use vintage analysis in projecting PD. This is a common methodology which is used industry wide to handle the situation on incomplete data described. Here we club all the loss data which occurred after *nth* month of origination. For example loss after first month of loans originated on January 2013 will be clubbed with losses after first month of loans originated on February 2013 and so on for every month on books. Hence we have curve for PDs with respect to time. The strength of the methodology is: no approximation of data. We do not want to focus on this methodology in detail, but we want to point out one drawback. The methodology is not sensitive to capture sudden market changes and hence the market impact on forecasted PDs is seen with a higher lag.

Conclusion

In the article we have explained how incomplete data for forecasting PDs is a problem faced by every credit risk manager. This problem is independent of the size/type of the loan portfolio. The interesting point is that this occurs even when we have the most complete and accurate data about losses of a given portfolio. We suggested how this problem can be tackled and gave the justification of the solutions. We do not present this article as a critique on any methodology but for completeness we discussed an alternate solution briefly pointing out its inherent drawback.

Chandrakant Maheshwari

Manager – Genpact Smart Decision Services
Financial Services Analytics
Chandrakant.Maheshwari@genpact.com

Bhaswati Sengupta

Manager – Genpact Smart Decision Services
Financial Services Analytics
Bhaswati.Sengupta@genpact.com

About the Authors

Chandrakant Maheshwari is a manger in the Financial Services Analytics division at Genpact. He has nine years of experience in the area of financial risk management. He is an alumnus of IIT Delhi with a degree in Mathematics and Computing, a 5 yr Integrated M. Tech course.

Bhaswati Sengupta is a manger in the Financial Services Analytics division at Genpact. She has eight years of experience in the area of financial risk management. She is an alumnus of Lady Shriram College of Commerce with a degree in Statistics.