Is Problem of Limited Data for Forecasting a Fact of Life for Credit Risk Managers?

Reasons and Solutions to This Problem

By

Chandrakant Maheshwari

Astha Sarda

Anand Dev Singh

Table of Contents

Abstract

In today's dynamic world, businesses are done at the speed of light. Every day new business ideas are generated. At the same speed many of those and existing ideas are exiting the market. Over all the business cycles are getting shorter and shorter. Due to this issue, financial institutions struggle to find enough relevant data to forecast risk related parameters for reserve calculation of their loan portfolio. It is not that they have flaw in their data collection methodology, but due to short business cycles the size of relevant data is not appropriate enough to come up with a good forecasting methodology. Also we cannot ignore the fact that the way data is collected and assumptions about the data keep on changing in to such dynamic environment. This further reduces the size of relevant data points. In this paper to address this issue we come up with forecasting methodologies which are efficient in forecasting when the data size is small.

1. Introduction

With growing market dynamics through technical innovations and change in the choices of people, a typical behavior and composition of loan portfolio of a financial institution keeps on changing. A typical example of dynamicity is social networking. Eight years back Orkut was the market leader in social networking, which was soon replaced by Facebook and now its position is being threatened by new mobile based networking applications. Similar trends can be seen in various other industries where the latest innovations are eclipsing the older technologies. Due to high level of networking and speed of information travel in the world, a good technology is quickly adopted and that again is replaced quickly by another better technology. This poses a serious challenge to the risk managers in the financial markets to forecast the risk numbers related to their portfolio. Their portfolio may contain counterparties who demonstrate varied behavior with respect to credit-worthiness within very short duration of time. These challenges result in limitations in the availability of relevant risk data to do any type of forecasting.

The objective of this paper is to discuss the issues faced by the risk practitioners when they are attempting to forecast risk parameters like (but not limited to) Probability of Default (PD) and Loss Given Default (LGD), when there is limited data availability. We present forecasting methodologies which are simple and intuitive. They do not require a mathematical rigor to be understood and prove to be efficient. They have the capability to follow any trend reversal observed in the subject risk parameter with minimum lag. The methodologies prove to be efficient even when we have historical data as small as just 4 readings.

In Sections 2 and 3, we discuss the limitations of conventional methodologies for forecasting purpose. While methodologies like ARIMA are highly data intensive, other simpler techniques like fitting of linear/nonlinear regression fail in capturing any trend reversal. Section 4 discusses the methodologies proposed by us and Section 5 and 6 demonstrates their performance measurement and strengths. To measure the robustness of our methodologies we test them with a simulator developed using the Ornstein–Uhlenbeck process. This process was chosen because we can make the process of any risk parameter path to auto-correlate enough to capture the reality and random enough so the robustness of the methodology regarding trend reversal can be tested. We check the robustness of the methodology by Minimum Absolute Percentage Error (MAPE) values under various values of Ornstein–Uhlenbeck process parameters.

2. Challenges of Conventional Forecasting Methodologies

For parametric time series approach we use different iterations of ARIMA and estimate model coefficients for best fit of data.

As per SAS Support- PROC ARIMA can handle time series of moderate size; there should be at least 30 observations. With fewer than 30 observations, the parameter estimates might be poor.

In ARIMA procedure partial auto correlation is used to produce confidence interval. SE=1/sqrt(N); if sample size is small SE will be high though there is no specific guideline for sample size.

Residual normality is an assumption for ARIMA approach-If the sample size is small, it may be difficult to detect assumption violations. Moreover, with small samples, non-normality is difficult to detect even when it is present.

Even if none of the test assumptions are violated, a normality test with small sample sizes may not have sufficient power to detect a significant departure from normality, even if it is present. Power decreases as the significance level is decreased (i.e., as the test is made more stringent), and increases as the sample size increases.

Low sample size (less than 20) doesn’t gives confidence on assumptions in parametric approach, so a non-parametric approach is recommended for time series forecasting like, distances-percentile, Rolling average etc.

As conventional high end methodologies like ARIMA etc need at-least 30 lengths of data series to come up with a proper forecasting model, hence if data for risk parameters like PD/LGD is taken in quarterly basis then we need at least 7.5 years of data. Any risk manager can understand that in 7.5 years the behavior of an entire portfolio of loan instruments of a financial institution changes which makes the relevant data about it an appropriate size but not relevant. Typically monthly calculations of PD/LGD values are not practical for big organizations and it is beyond the scope and capability of small organizations. For the former, because of the size of the portfolio they handle and for later, because of the limitation of resources (man power and system limitations).

3. Other Basic Forecasting Approaches

Linear trend fitting models might seem appropriate when dealing with small data sets; however they are often inappropriate for business and economic data. Most naturally occurring business time series do not behave as though there are straight lines. In reality their trends changes over time reflected in change of slope or intercept or both. The linear trend model tries to find the slope and intercept that give the best average fit to all the past data, and unfortunately its deviation from the data is often greatest near the end of the time series, where the forecasting starts.

Since in a linear trend model, the next few values of the series will be slightly above the last observed value, they eventually lie outside the 95% confidence limits for the predictions.

If the residual plots and diagnostic tests are studied, it can be noticed that the errors are severely auto correlated: there are long runs of negative errors alternating with long runs of positive errors.

More generally, exponential and quadratic trend based models have similar disadvantages and are not suitable especially when limited data is available for the dependent variable.

4. Our Forecasting Approaches

In our paper, we suggest some approaches which follow simple thumb rule and work well with the small data set. They require data sets of length as small as of size 4. The following are the forecasting approaches discussed in this paper

1. **Method 1:** ***Modified Step AR approach***. This methodology is a slight modification of already existing Step AR methodology. The steps are:
   a. Fit a time trend model to the data series using ordinary least squares regression. The predicted values are the trend component for the forecast.
   b. Calculate the residuals by taking the difference between the predicted and actual values.
   c. Calculate auto correlation at lag 1 for the residuals. This gives the AR component for the forecast.
   d. Add the AR component and the trend component to get the forecasted values.
2. **Method 2:** ***Comparing Moving Average and Percentile approach.*** Steps are as follows:
   a. Compare the average of last two and last four actual values.
   b. If average of last 4 is greater than average of last 2, then there is a falling trend, hence the forecast is the 10%ile of the last 4 actual values.
   c. If average of last 4 is lesser than average of last 2, then there is rising trend, hence the forecast is the 90%ile of the last 4 actual values.
3. **Method 3:** ***Comparing Moving Median and Percentile approach.*** Steps are as follows:
   a. Compare the median of last two and last four actual values.
   b. If median of last 4 is greater than median of last 2, then there is a falling trend, hence the forecast is the 10%ile of the last 4 actual values.
   c. If median of last 4 is lesser than median of last 2, then there is rising trend, hence the forecast is the 90%ile of the last 4 actual values.
4. **Method 4:** ***Comparing Moving Average and Min/Max approach.*** Steps are as follows:
   a. Compare the average of last two and last four actual values.
   b. If average of last 4 is greater than average of last 2, then there is a falling trend, hence the forecast is the minimum of the last 4 actual values.
   c. If average of last 4 is lesser than average of last 2, then there is rising trend, hence the forecast is the maximum of the last 4 actual values.
5. **Method 5:** ***Comparing Moving Median and Min/Max approach.*** Steps are as follows:

a. Compare the median of last two and last four actual values.
b. If median of last 4 is greater than median of last 2, then there is a falling trend, hence the forecast is the minimum of the last 4 actual values.
c. If median of last 4 is lesser than median of last 2, then there is rising trend, hence the forecast is the maximum of the last 4 actual values.

5. Performance Measurement

We measured the performance of the methodologies by building the test cases using Ornstein–Uhlenbeck process. This process was chosen because after ample amount of iterative research we found that the behavior of this process resembles PD/LGD behavior of a typical portfolio. This process also gave enough randomness so that we could measure the performance of our forecasting approaches for any trend reversal. Also the process allows the autocorrelation in the sample paths which is typically observed in the risk parameters.

In each simulation step we first simulated first 14 steps and started forecasting the next 14 steps using the discussed approaches. Performance comparison criterion used was MAPE.

The step size in simulation has nothing to do between the periods between the two readings.

For Step AR 14 data points were used and for rest of the methodologies just 4 data points were enough. The following tables were developed after 2000 simulations. The first column refers to methodology, second refers to the average MAPE after 2000 simulations, third column refers to the standard deviation of the 2000 MAPEs and the forth column refers to that among 2000 simulations, how many times the MAPE of the particular methodology was minimum in comparison to the rest of the methodologies.

Observations with parameters values for mean reversion speed, long term mean, volatility and step size were 100%, 80%, 100% and 8.33% respectively.

Method	MAPE Mean	MAPE STDEV	%Success
Method 1	19.31%	14.75%	47.05%
Method 2	20.26%	21.82%	31.10%
Method 3	19.94%	20.10%	11.45%
Method 4	21.36%	23.43%	7.60%
Method 5	21.02%	21.46%	2.80%

Observations with parameters values for mean reversion speed, long term mean, volatility and step size were 100%, 80%, 150% and 8.33% respectively.

Method	MAPE mean	MAPE STDEV	%Success
Method 1	30.17%	69.84%	54.75%
Method 2	35.37%	153.08%	24.70%

Method 3	34.99%	152.93%	9.25%
Method 4	37.94%	203.75%	8.05%
Method 5	37.56%	203.60%	3.25%

Observations with parameters values for mean reversion speed, long term mean, volatility and step size were 100%, 80%, 200% and 8.33% respectively.

Method	MAPE mean	MAPE STDEV	%Success
Method 1	34.95%	63.10%	53.05%
Method 2	38.17%	99.91%	26.60%
Method 3	37.82%	99.23%	8.90%
Method 4	39.32%	71.05%	8.75%
Method 5	38.98%	70.14%	2.70%

6. Strength of Proposed Approaches/Methodologies

When the data is sparse, any mathematical/statistical rigor is of little use as for that the basic assumptions required are of equilibrium market hypothesis. Concluding that markets are in equilibrium, with just 4 data points will be a conceptually in-correct. There will be trend and very high chances that there will be trend reversal as well, in just 4 data points, so it is important that any methodology implemented should be able to capture that.

Our methodologies are not limited by the assumptions of stable market hypothesis. They simply extract the trends from last 4 data points. Mathematically they present no complexity in understanding to any practitioner who is more focused towards the business side rather than the quantitative flavor of any approach. Having the percentile approach enables the methodologies to follow any trend reversal with a minimum lag.

7. Conclusion

With growing market uncertainty, data limitations are the fact of the day for the financial institutions. In our paper we have presented some simple and intuitive approaches which show promising results in forecasting any given risk variable. To check their robustness we test them with a stochastic differential equation, which provided randomness enough to check trend reversal strength of the methodologies as well as autocorrelation which typically is demonstrated by risk parameters. We hope that using these approaches will give forecasting ideas to any risk practitioner who is facing the problem of limited data. He/she can also make changes/manipulations in these approaches to solve his/her business problem.

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

Astha Sarda
Manager – Genpact Smart Decision Services
Financial Services Analytics
Astha.Sarda@genpact.com

Anand Dev Singh
Manager – Genpact Smart Decision Services
Financial Services Analytics
AnandDev.Singh@genpact.com

About the Authors

Chandrakant Maheshwari, Manager, Financial Services Analytics, Genpact has 9 years' experience in the area of Market Risk focused towards Interest Rate Risk and Foreign Exchange Risk. He is an alumnus of IIT Delhi with a degree in Mathematics and Computing, a 5 yr Integrated M. Tech course.

Astha Sarda, Manager, Financial Services Analytics, Genpact has 7 years' experience in the area of Credit Risk, Investment banking and Market/Economic Research. She is an alumnus of JNU Delhi with a M.A. degree in Economics.

Anand Dev Singh, Manager, Financial Services Analytics, Genpact has 7 years' experience in the area of Fraud predictive analytics, Retail banking and Market research. He is an alumnus of ISI Kolkata with a M.Tech degree in Quality Reliability and Operations Research.

References

1. SAS/(ETS(R) 9.3 User Guide: The ARIMA Procedure

http://support.sas.com/documentation/cdl/en/etsug/63939/HTML/default/viewer.htm#etsug_arima_sect016.htm

2. Is Facebook Losing Its Cool? Some Teens Think So
http://business.time.com/2013/03/08/is-facebook-losing-its-cool-some-teens-think-so/

3. Ornstein-Uhlenbeck Theory

http://personal.lut.fi/wiki/lib/exe/fetch.php/en/technomathematics/time_series_research/start/ornstein-uhlenbeck_theory.pdf