Paper 132

# Population Stability and Model Performance Metrics Replication for Business Model at SunTrust Bank

Bogdan Gadidov, Kennesaw State University; Benjamin McBurnett, Georgia Institute of Technology

## Abstract

Board of Governors of the Federal Reserve System has published Supervisory Guidance on Model Risk Management (SR Letter 11-7) emphasizing that banks rely heavily on quantitative analysis and models in most aspects of financial decision making. Ongoing monitoring and maintenance (M&M) is essential for timely evaluation of model performance to determine whether changes in business strategies and market conditions require adjustment, redevelopment, or replacement of the model. A typical M&M plan includes tracking of Population Stability Index (PSI), Rank Ordering Test, and Kolmogorov-Smirnov Statistic (KS). As part of an internship program at SunTrust bank, I was able to track these key metrics for one business critical model. The model uses a logistic regression to predict the probability of default for a given customer.

To track the three metrics stated above, data from quarter 1 of 2014 is compared with a baseline distribution, generally the dataset which is used to create the model. PSI quantifies the shift in the distribution of the population between the baseline and current time periods. Rank Ordering Testing involves comparing the expected default rate, predicted by the model, to the actual default rate in the current quarter. The KS statistic assesses model performance by measuring the model's ability to discern defaults from non-defaults. The npar1way procedure was used in SAS® to calculate KS. Reports and charts presented in this poster will be sanitized due to the confidential nature of the data, but methodology and step-by-step procedures represent actual research results.

## Introduction

The three metrics will be used together to assess how well the model is performing in the validation period. Each observation in the datasets contains a score which is equal to the probability of default multiplied by 10,000. The score is created in such a way to be similar to a FICO score. For the calculation of PSI and rank ordering testing, the observations are grouped into bins based off of the score of each observation. 8 bins are created by using predefined cutoff points as the intervals. The number of observations in each bin does not necessarily have to be the same. The percentage of defaulting customers in each of these bins will be used in calculations in further sections.

## PSI

PSI quantifies shifts in population dynamics over time. As models are based on historical datasets, it is necessary to ensure that present-day population features are sufficiently similar to the historical population from which the model is based. A higher PSI corresponds to greater shifts in population. Generally, PSI values greater than 0.25 indicate a significant shift in the

population, while values less than 0.10 indicate a minimal shift in the population. Values between 0.10 and 0.25 indicate a minor shift. The formula for PSI is shown below, where $n_{di}$ is the number of observations in the $i^{th}$ bin of the development dataset, $n_{vi}$ is the total number of observations in the $i^{th}$ bin of the validation dataset, and $N_{di}$ and $N_{vi}$ are the total number of observations in the development and validation datasets respectively. This formula is used to calculate the PSI for each of the 8 bins, and the total PSI is the summation of the individual PSI's from each bin.

$$PSI = \sum ((\frac{n_{di}}{N_d}) - (\frac{n_{vi}}{N_v})) * \ln ((\frac{n_{di}}{N_d})/(\frac{n_{vi}}{N_v}))$$ Eq. 1.

Table 1 below shows the individual calculations of the PSI for each bin. The percentage of the observations which lie in each bin are shown for both the development and validation datasets in the Dev Percent and Val Percent columns. The PSI column shows the calculated PSI for each bin, using the formula from above. The value which is bolded in the Cumulative PSI column shows the overall value for the PSI across all 8 bins. Using the guidelines as defined earlier, this value is much less than 0.1, indicating a minimal shift in the population between development and validation periods.

| Score Range | Bin | Dev Frequency | Dev Percent | Val Frequency | Val Percent | PSI | Cumulative PSI |
|---|---|---|---|---|---|---|---|
| >1400 | 1 | 8846 | 1.34 | 8074 | 1.62 | 0.00023 | 0.00023 |
| 440-1400 | 2 | 18990 | 2.88 | 15241 | 3.05 | 0.00004 | 0.00027 |
| 200-440 | 3 | 35537 | 5.39 | 26272 | 5.26 | 0.00001 | 0.00028 |
| 90-200 | 4 | 74324 | 11.28 | 54985 | 11.01 | 0.00003 | 0.00031 |
| 40-90 | 5 | 92214 | 14.00 | 68979 | 13.81 | 0.00001 | 0.00032 |
| 18-40 | 6 | 105203 | 15.97 | 79916 | 16.00 | 0.00000 | 0.00032 |
| 8-18 | 7 | 223414 | 33.91 | 169095 | 33.85 | 0.00000 | 0.00032 |
| <=8 | 8 | 100347 | 15.23 | 76954 | 15.41 | 0.00001 | ***0.00033*** |

Table 1. PSI Calculations

## Rank Ordering Testing

The second metric also identifies shifts in population. Rank ordering is calculated by considering the percentage of “bads” (typically defaults, delinquencies, etc.) per given score ranges within the development and validation datasets. The expected event rate is calculated for each bin from the development dataset, using the calculated probabilities of default from the logistic regression. The actual event rate is calculated from the validation dataset by finding the percentage of defaults within each bin of observations. A monotonically decreasing pattern should be seen among the bins, as the higher score ranges have higher rates of default, which can be seen in Table 2. A 95% confidence interval is then calculated between the difference of the expected and actual event rates. Confidence intervals which do not contain 0 have a statistically significant difference between the expected and actual default rates.

| Score Range | Bin | Expected Event Rate | Actual Event Rate | Expected vs. Actual % Difference | Lower 95% CI | Upper 95% CI | Statistically Significant Difference |
|---|---|---|---|---|---|---|---|
| >1400 | 1 | 45.47% | 40.07% | 5.40% | 3.91% | 6.89% | Yes |
| 440-1400 | 2 | 10.54% | 8.07% | 2.47% | 1.86% | 3.09% | Yes |
| 200-440 | 3 | 4.43% | 3.34% | 1.09% | 0.79% | 1.40% | Yes |
| 90-200 | 4 | 1.87% | 1.48% | 0.39% | 0.25% | 0.53% | Yes |
| 40-90 | 5 | 0.96% | 0.71% | 0.25% | 0.16% | 0.34% | Yes |
| 18-40 | 6 | 0.41% | 0.51% | -0.10% | -0.16% | -0.03% | Yes |
| 8-18 | 7 | 0.15% | 0.21% | -0.06% | -0.09% | -0.03% | Yes |
| <=8 | 8 | 0.09% | 0.10% | -0.02% | -0.04% | 0.01% | No |

Table 2. Rank Ordering Testing

## KS Statistic

The KS statistic is used as a measure of the ability of the model to separate good and bad accounts. The KS statistic is calculated manually and also through the proc npar1way procedure in SAS. To calculate KS manually, each dataset is divided into 10 groups (deciles). Since the score values occur in discrete intervals, each decile contains approximately 10% of the dataset. Once the deciles are obtained, the cumulative percentage of defaults and non-defaults is calculated across the 10 deciles. The KS is the maximum difference between the cumulative percentage of defaults and non-defaults.

| | | | | Cumulative | | |
|---|---|---|---|---|---|---|
| Decile | Total | # Default | % Default | % Default | % Non-Default | KS |
| 1 | 59893 | 29965 | 50.03 | 62.06 | 4.90 | 57.16 |
| 2 | 55529 | 8713 | 15.69 | 80.10 | 12.57 | **67.53** |
| 3 | 63440 | 2912 | 4.59 | 86.13 | 22.48 | 63.65 |
| 4 | 59832 | 2404 | 4.02 | 91.11 | 31.89 | 59.22 |
| 5 | 51742 | 1915 | 3.70 | 95.08 | 40.05 | 55.03 |
| 6 | 60834 | 1745 | 2.87 | 98.69 | 49.73 | 48.96 |
| 7 | 43753 | 556 | 1.27 | 99.84 | 56.80 | 43.04 |
| 8 | 51317 | 0 | 0.00 | 99.84 | 65.20 | 34.64 |
| 9 | 139490 | 76 | 0.05 | 100.00 | 88.03 | 11.97 |
| 10 | 73045 | 0 | 0.00 | 100.00 | 100.00 | 0.00 |

Table 3. KS for Development Dataset

| | | | | Cumulative | | |
|---|---|---|---|---|---|---|
| Decile | Total | # Default | % Default | % Default | % Non-Default | KS |
| 1 | 45194 | 5230 | 11.57 | 69.91 | 8.12 | 61.79 |
| 2 | 42454 | 753 | 1.77 | 79.98 | 16.60 | **63.38** |
| 3 | 47540 | 462 | 0.97 | 86.16 | 26.17 | 59.99 |
| 4 | 45680 | 233 | 0.51 | 89.27 | 35.41 | 53.86 |
| 5 | 39389 | 199 | 0.51 | 91.93 | 43.37 | 48.56 |
| 6 | 46376 | 215 | 0.46 | 94.80 | 52.75 | 42.05 |
| 7 | 35763 | 84 | 0.23 | 95.92 | 60.00 | 35.92 |
| 8 | 42552 | 109 | 0.26 | 97.38 | 68.63 | 28.75 |
| 9 | 98651 | 147 | 0.15 | 99.34 | 88.65 | 10.69 |
| 10 | 55917 | 49 | 0.09 | 100.00 | 100.00 | 0.00 |

Table 4. KS for Validation Dataset

Table 3 and Table 4 show the calculations for the KS statistic in both the development and validation datasets. Each table shows the 10 deciles and the number of defaults within each decile. Using this, the cumulative percentage of the defaults and non-defaults can be calculated for each decile, and the difference is shown in the last column. The point where the difference is the largest is the value of the KS statistic. A graph in Figure 1 is used to visualize where this maximal KS occurs.

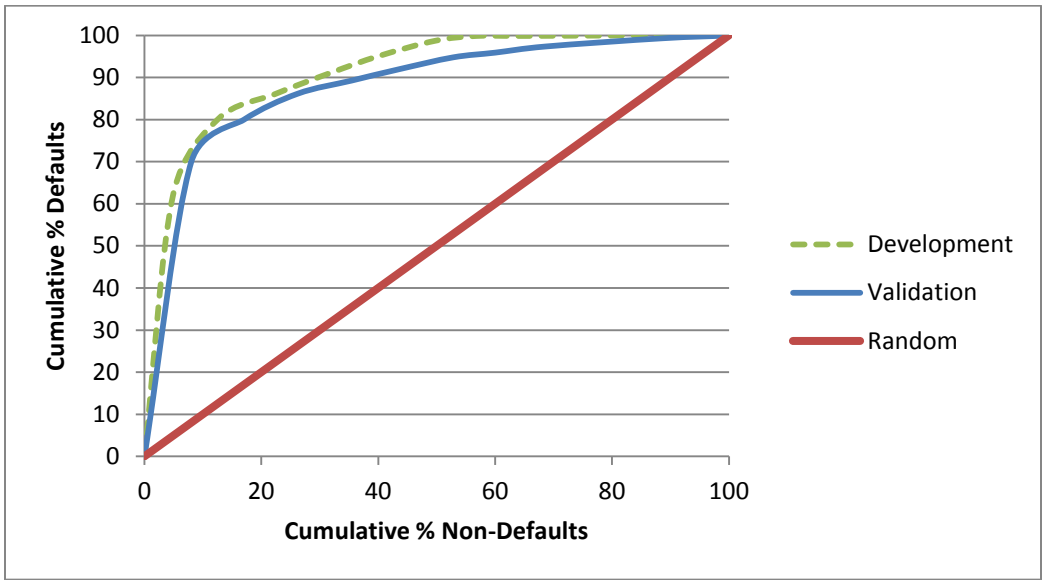


Figure 1. Graphical Comparison between Development and Validation KS

It is important to see from the figure above that the shape of the KS curves for both the development and validation datasets have approximately the same shape, indicating that the maximal separation occurs around the same point in the second decile. The red line represents "random" guessing at assigning defaults and non-defaults. The distance between the red line and the point on green or blue curve represents the value of the KS. It can be seen that the validation KS is slightly less than the development KS. This is expected as model performance will deteriorate over time. The difference, however, is relatively small. Some cutoff points for acceptable KS ranges are:

- Validation KS > 50 indicates excellent model performance

- Validation KS < 50 and less than 20% decrease from development KS indicates acceptable performance
- Validation KS < 50 and more than 20% decrease from development KS indicates a deterioration of model performance

In addition to calculating KS manually, the npar1way procedure can be performed in SAS to achieve similar results. The output from this procedure appear in Output 1 and Output 2 below. The value of the KS statistic calculated in tables 3 and 4 can be compared to the D value in the outputs below.

| Kolmogorov-Smirnov Two-Sample Test (Asymptotic) | | | |
|---|---|---|---|
| KS | 0.176977 | D | 0.679102 |
| KSa | 143.654206 | Pr > KSa | <.0001 |

Output 1. SAS Output from Proc Npar1way for Development Dataset

| Kolmogorov-Smirnov Two-Sample Test (Asymptotic) | | | |
|---|---|---|---|
| KS | 0.077963 | D | 0.641892 |
| KSa | 55.101723 | Pr > KSa | <.0001 |

Output 2. SAS Output from Proc Npar1way for Validation Dataset

## Conclusion

The three metrics discussed, PSI, rank ordering testing, and KS statistic, can be used in conjunction to assess model performance. PSI and rank ordering testing focus more on how the population may have shifted between development and validation periods, while the KS statistic is used to assess the predictive capability and performance of the model. Of the three metrics, PSI and KS should be more closely monitored when making decisions regarding the model and its performance. For the model evaluated in this paper, the PSI was well within acceptable ranges (<0.1). The KS statistic decreased by approximately 4 between development and validation periods, but using the guidelines described in the previous section, since the KS is greater than 50, the model is still performing well within acceptable range. The rank ordering testing should be used in conjunction with the other two statistics. The rank ordering testing suggests that the model is over-predicting the actual default rate, but since the other two statistics agree that the model performance has not deteriorated, it can be determined that the model is still performing within acceptable standards.

## Acknowledgments

Special thanks to Alex Shenkar at SunTrust Bank who provided materials and assisted with this project throughout the internship program.

## Contact Information

Your comments and questions are valued and encouraged. Contact the author at:
Bogdan Gadidov
Kennesaw State University
bgadidov@kennesaw.edu