Machine Learning Performance over Long Time Frame

Yazhe Li, Tony Bellotti, Niall Adams

Imperial College London

y.li16@imperial.ac.uk

Credit Scoring and Credit Control Conference, Aug 2017

Acknowledgments

Yazhe Li is a Ph.D student from Department of Mathematics, Imperial College London.

This work is supervised by Dr Tony Bellotti (Imperial College London) and Professor Niall Adams (Imperial College London).

Machine Learning in Consumer Credit Risk

Common machine learning methods in the credit risk industry include:

- Logistic Regression
- Penalized Logistic Regression
- Decision Trees

Various studies have interested in machine learning algorithms in the credit risk industry:

- Random Forests
- Boosted Regression Trees

Methods Background

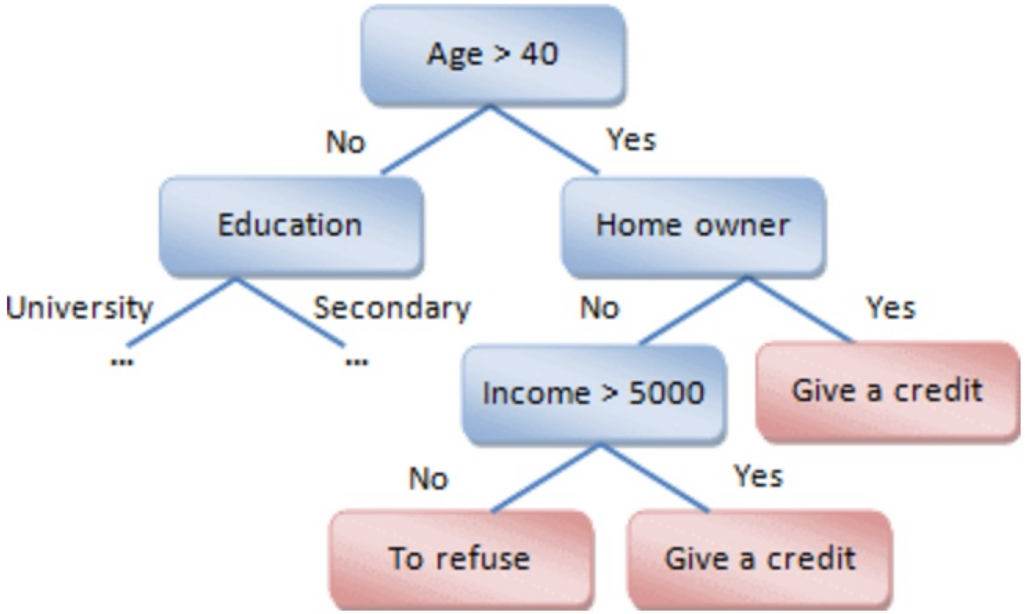
- Penalized Logistic Regression: penalized logistic regression adds penalty terms to the likelihood function of logistic regression.

$$\textit{Objective Function} = L(\beta; x) - \lambda\left[(1-\alpha)\frac{1}{2}|\beta|_2^2 + \alpha|\beta|_1\right], \quad \text{where } \lambda > 0 \text{ and } 0 \leq \alpha \leq 1. \tag{1}$$

It is designed for parameter shrinkage and variable selection.

Methods Background

- Decision Trees



Methods Background

Although decision trees have a good interpretability; decision trees also have an unstable nature.

Several ensemble methods based on the tree model, like boosted regression trees and random forests, are designed.

- Random forests: build approximately uncorrelated trees, and average them.
- Boosted regression trees: sequentially fit many trees to the training set and combine them with their learning rates.

Research Gaps in Current Literature

Several remaining research gaps which are relevant to credit risk issues:

1 Temporal issue:

The relationship between the **distribution changes** in the portfolio (i.e. **population drift**) and the credit risk model performance is an area need investigation [5].

2 Extreme class imbalance:

High imbalance (one class is rare, compared to the other) is a common problem in the credit risk industry. For example, mortgage default rate could be as low as 0.5% in some data sets. How **extreme imbalance** will influence **model behavior** in the financial industry.

Research Hypotheses

Two hypotheses prior to our experiment:

- Non-linear models (machine learning algorithms) are generally superior than linear models in credit risk modelling. Since non-linear models can capture the non-linear pattern in the credit data set.
- Parsimonious models are more robust than complex models over time. Because high model complexity can lead to overfitting.

Data Description

- Freddie Mac (a U.S. federal government sponsored enterprise) provides decade-long U.S. mortgage credit information and contains several extreme low default rate years.
- The characteristics of Freddie Mac data typically address the research gaps: **high imbalance** and **temporal issues**.
- Mortgage default status is defined as when a borrower is greater than 180 days due in making a repayment on their home loan.
- In our experiment, the target variable is whether those mortgages moved to the default status in the following **two years** after the first payment date.

Data Description

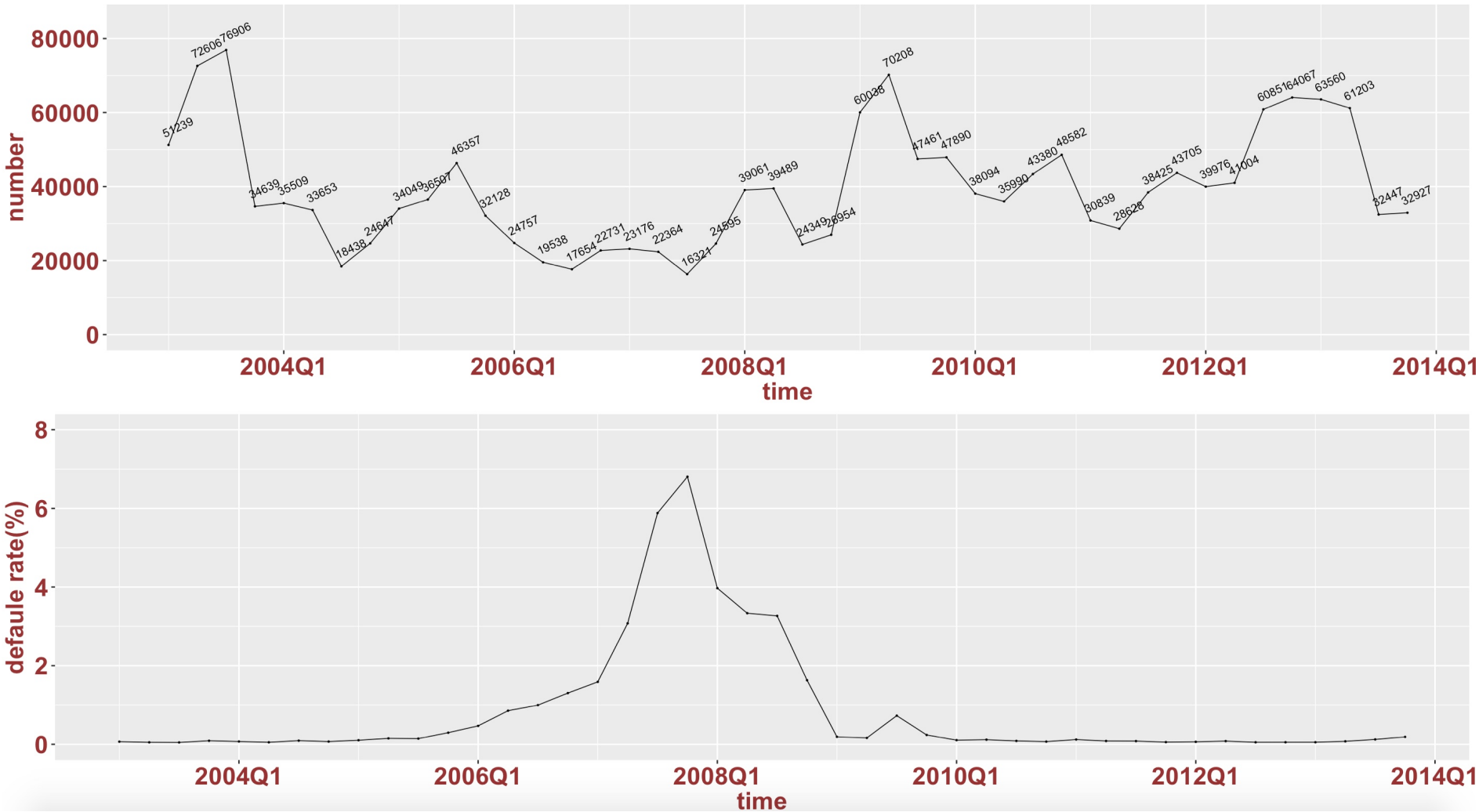


Figure: Sample size and default rate from 2003 to 2013.

Experiment Description

After data preparation process, we deploy five models: Balanced Random Forest (**BRF**) [1], Boosted Regression Trees (**BRT**), Undersample Boosted Regression Trees (**BRTU**) [3], Logistic Regression (**LR**) and Lasso Penalized Logistic Regression (**LLR**).

Experiment Procedure:

1. We use data from an individual year as a training set to train five models (year 2000).
2. Five models are used to forecast the data for the four quarters in the **following third year** (year 2003).

Experiment Description

Experiment notes:

- The **"two-year gap"** in our procedure is designed for recording default status of mortgages in the training set.
- We use **AUC** as performance metric.
- In forecast process, we bootstrap each quarters data 100 times, in order to calculate the mean and the standard deviation of AUC.
- The efficacy of these models for mortgage default forecasting are observed over a **11-year** long time frame (includes the financial crisis period), which allow us to observe performance over an extended period.
- LR is regarded as a reference benchmark, because it is in common use now [6].

Empirical Results

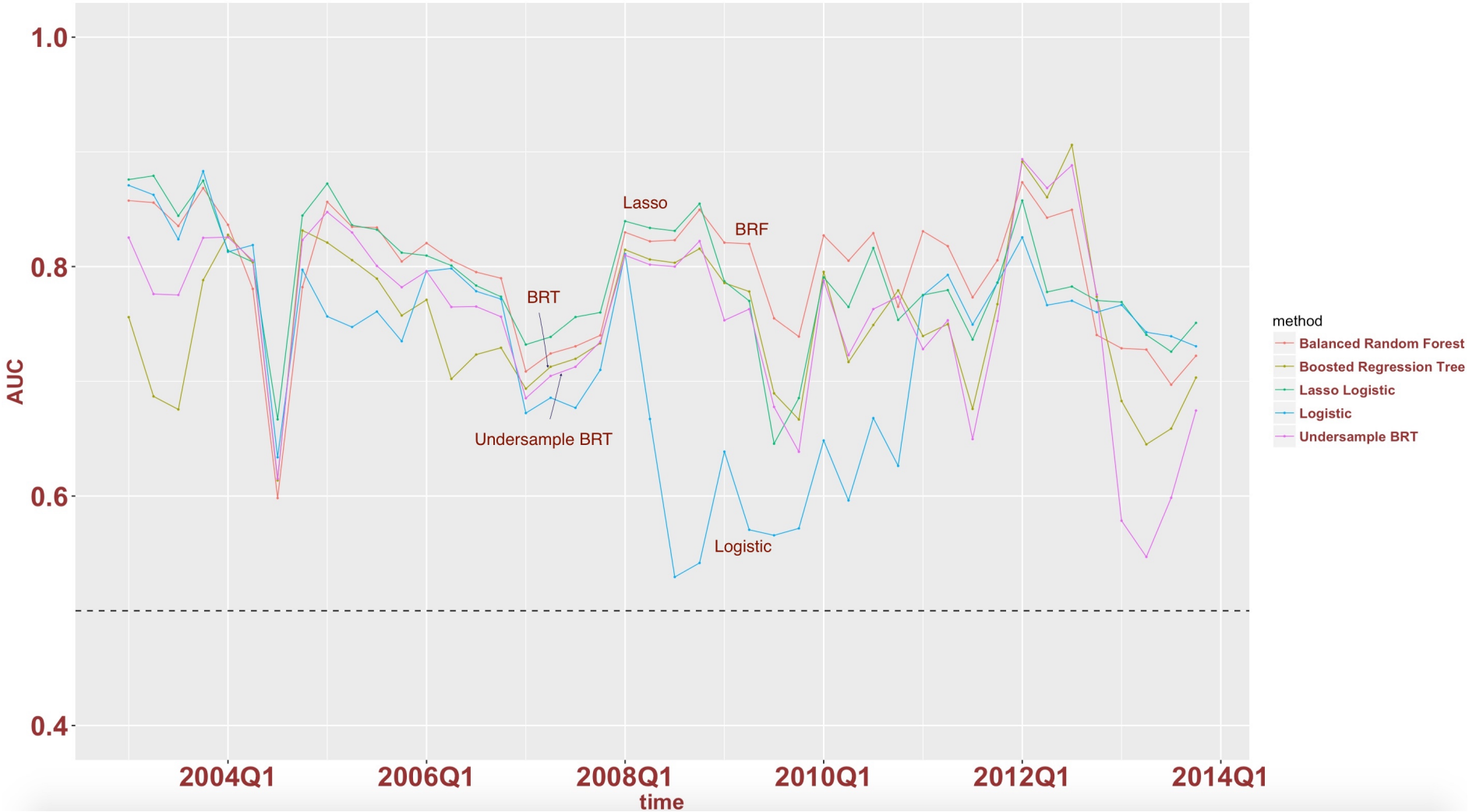


Figure: Forecast AUC from 2003 to 2013.

Empirical Results

1. We notice the declining performance of LR in the financial crisis period; however other advanced methods still perform well.
2. We never observe one classifier continuously dominates LR performance; there is **“no clear winner"** in this experiment.

Empirical Results

We also use the average rank $\frac{\sum(rank\ \ in\ \ each\ \ quarter)}{(number\ \ of\ \ the\ \ quarters)}$ to evaluate these algorithms' performance (from 1 best to 5 worst), based on their AUC. The rank is:

- LLR (2)
- BRF (2.13)
- BRT (3.52)
- BRTU (3.56)
- LR(3.77)

Friedman's test [2] shows that in our experiment, there is a significant difference in different model's performance ranks, *Friedman* $\chi^2 = 51.727$ and $p - value < 10^{-9}$.

Empirical Results

It is important to check to what extent machine learning algorithms perform better than the benchmark algorithm LR.

Thus the highest rank technique **LLR** and second best performance **BRF** are compared with **LR** (worst performance) by using a permutation test [4], to check whether there is a significant difference in the **mean AUC** .

Table: Permutation test $p - value$ table.

Methods	p-value	AUC Difference
LLR vs BRF	0.3385	0.0049
LR vs BRF	10^{-4}	-0.0663
LR vs LLR	10^{-4}	-0.0614

$p - value$ table shows that both **LLR** and **BRF** appear to have better performance than **LR**. However, there is no apparent difference between **LLR** and **BRF**.

Discussion

Overall, the results indicate that over long time frame, machine learning algorithms efficacy varies.

Both **LLR** and **BRF** provide a comparatively reliable prediction, significantly outperform **LR**.

Discussion

1 LLR:

- capture important variables.
- it is easily interpreted. LLR extends the existing credit scoring standard model (i.e. LR).

2 BRF:

- ability to select important variables.
- capacity to handle highly imbalanced data [1]. Our initial experiment results show that BRF outperform RF in all 44 quarters.

Discussion

Table: Lasso coefficient table (2005).

Variable	Coefficient	Variable	Coefficient	Variable	Coefficient
score	-0.0073	number.borrowers	-0.2688	servicer	-0.5721
LTV	-0.0211	occupancy.statusS	0.7250	OIR	0.4608
Intercept	-4.7054	other variables	0		

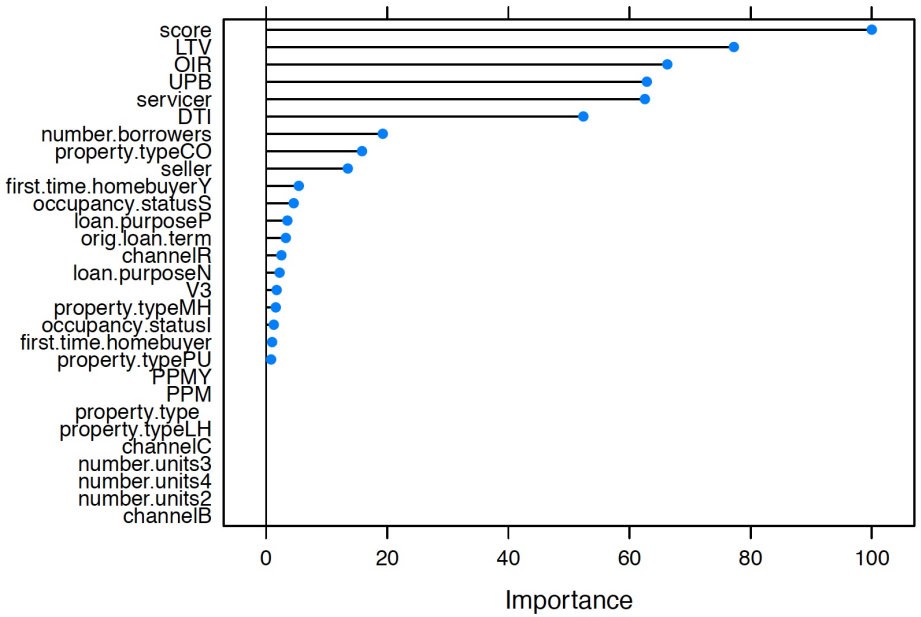


Figure: Variable importance of BRF in 2005

Discussion 3-Year Gap

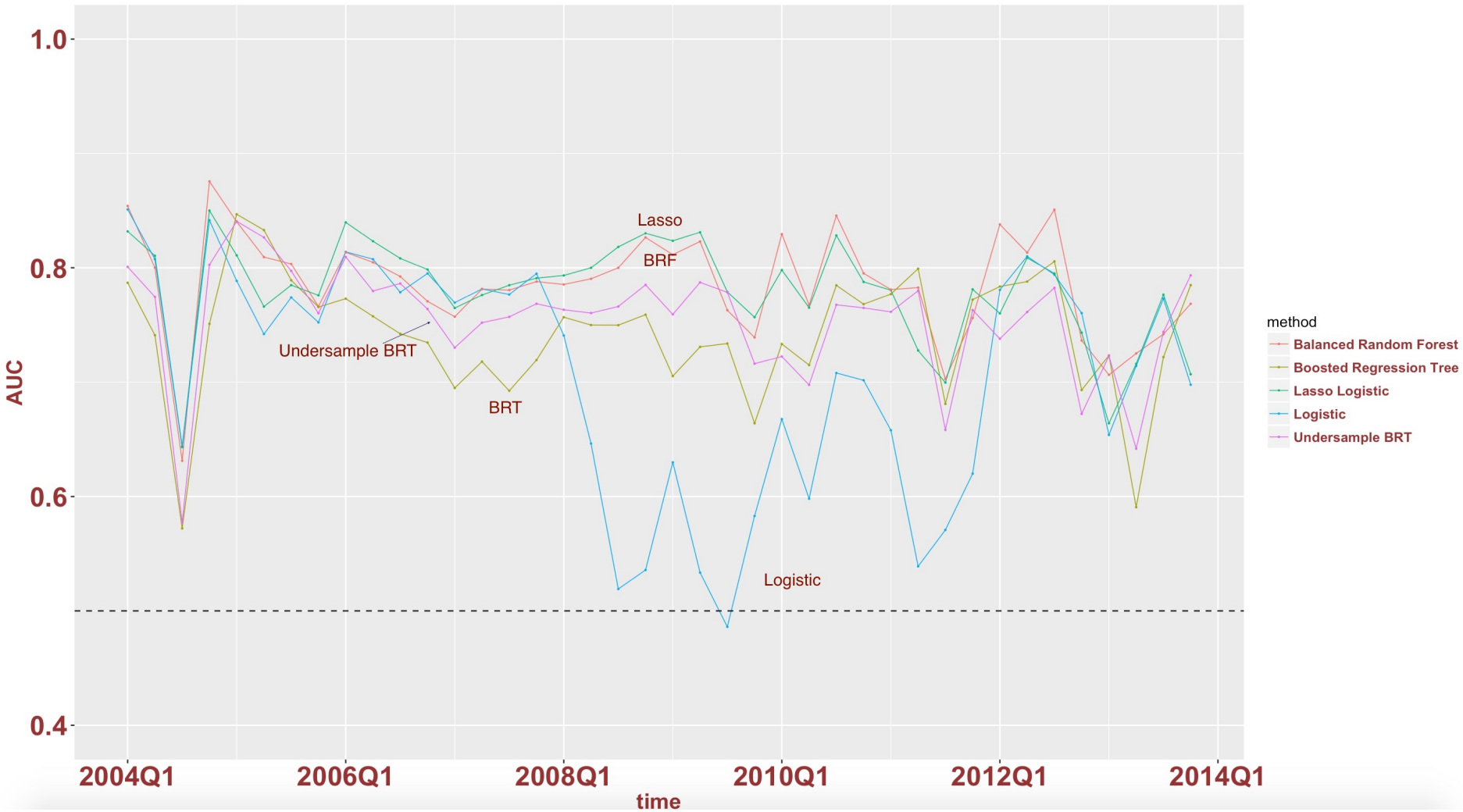


Figure: Forecast AUC from 2004 to 2013 (3-year gap).

Discussion 4-Year Gap

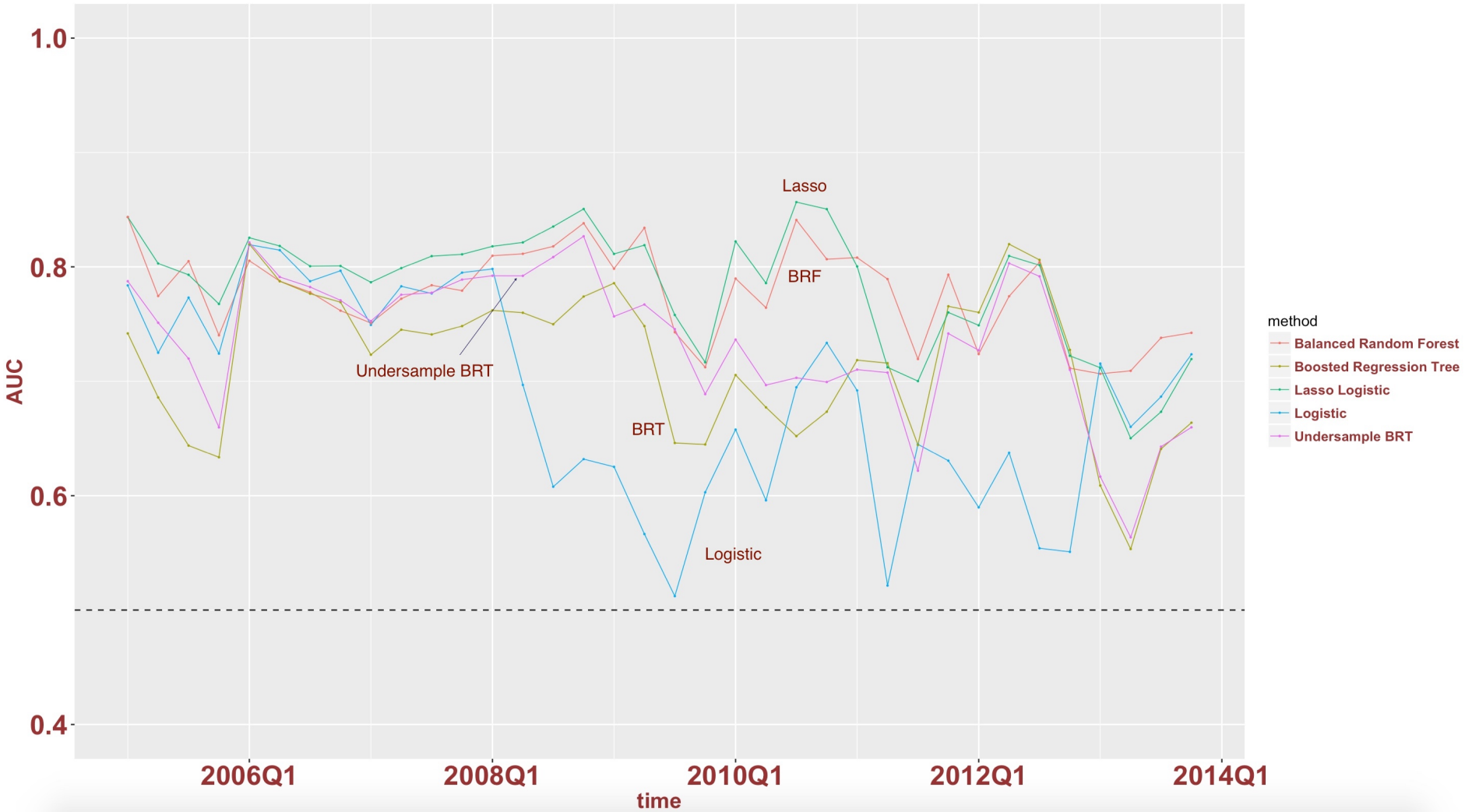


Figure: Forecast AUC from 2005 to 2013 (4-year gap).

Discussion

The two prior hypotheses are contrary to our results:

- If we use LLR as our linear model, both nonlinear model BRF and linear model LLR provide a reliable forecast.
- Parsimonious model (LR) is not more robust than a complex model (BRF) over time. If we increase the time gap to 3 years or 4 years, we find logistic regression still has a declining performance in the financial crisis.

Conclusion

- Machine learning algorithms' efficacy varies, which shows that continuing to use one kind of model is not appropriate.
- Overall, both LLR and BRF provide a comparatively reliable forecast.
- With gap time increasing, models' efficacy decreases.
- The declining performance of LR during the financial crisis is significant.

Future Work

- Issues of using logistic regression in highly imbalanced data set and remedies to fix its decline performance in the financial crisis. (will be discussed in another talk)
- In the financial application, the costs of false positive error and false negative error are different; which is critical in measuring models' effectiveness for operational purpose. Incorporating cost information into model building process is meaningful in the credit risk industry.

References I

[1] L. Breiman, C. Chen, and A. Liaw.
Using random forest to learn imbalanced data.
J. of Machine Learning Research, (666), 2004.

[2] M. Friedman.
A comparison of alternative tests of significance for the problem of m rankings.
The Annals of Mathematical Statistics, 11(1):86–92, 1940.

[3] H. He and E. A. Garcia.
Learning from imbalanced data.
IEEE Transactions on knowledge and data engineering, 21(9):1263–1284, 2009.

[4] T. Hesterberg, D. S. Moore, S. Monaghan, A. Clipson, and R. Epstein.
Bootstrap methods and permutation tests.
Introduction to the Practice of Statistics, 5:1–70, 2005.

[5] G. Krempl and V. Hofer.
Classification in presence of drift and latency.
In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pages 596–603. IEEE, 2011.

[6] S. Lessmann, B. Baesens, H.-V. Seow, and L. C. Thomas.
Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research.
European Journal of Operational Research, 247(1):124–136, 2015.

Thanks for your attention! Any questions?

Appendix: Empirical Results SD

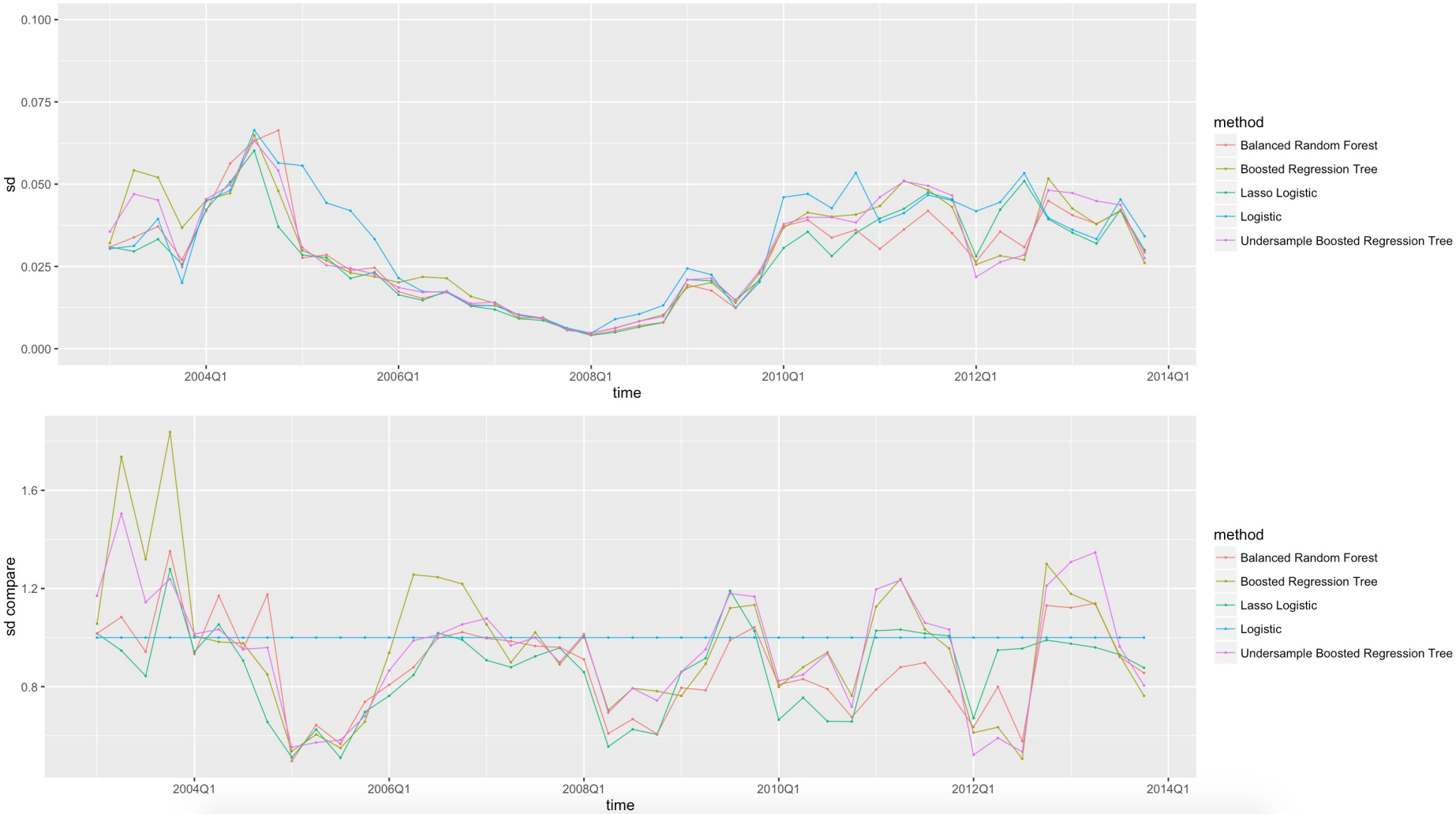


Figure: SD of forecast AUC from 2003 to 2013.

Appendix: Empirical Results SD

Stability is another important issue to judge the performance of a classifier. We find:

- No algorithm has a continuous lower standard deviation
- All classifiers' standard deviation are relatively low in 2007/2008.

Appendix: Empirical Results SD

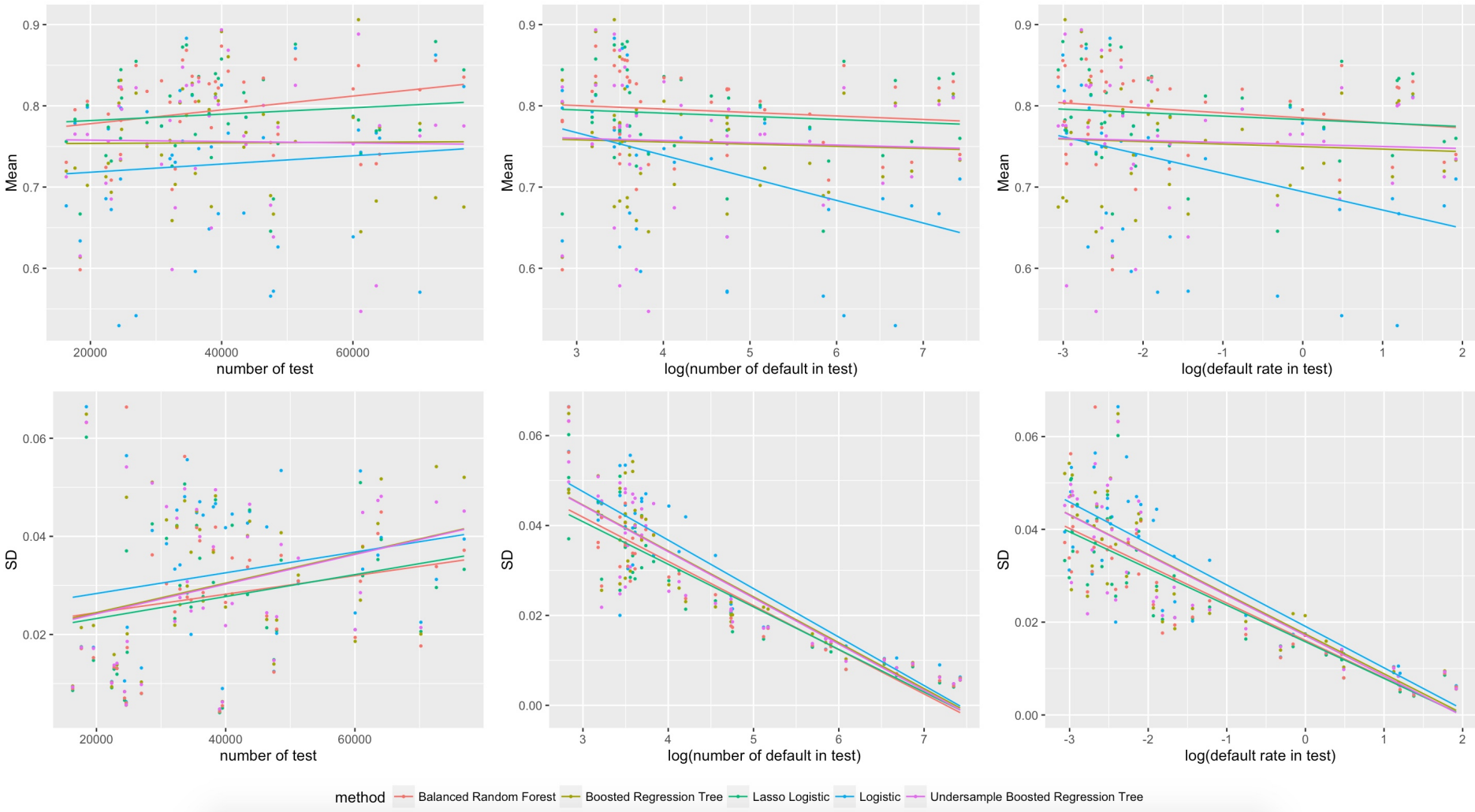


Figure: Mean and SD of AUC vs number of sample points.