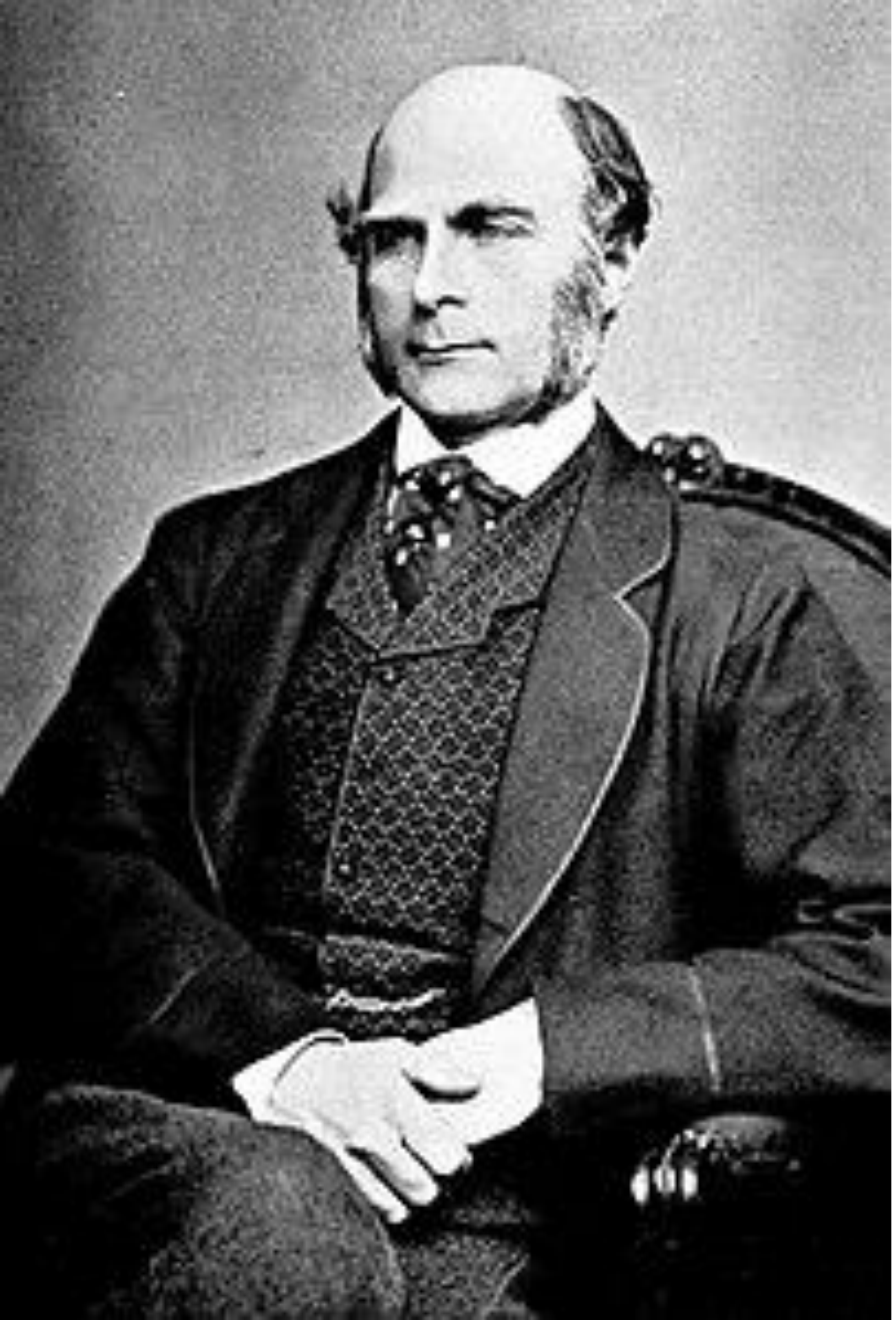
Loan Portfolio Model Management

A Cautionary Tale

David Romoff
djr2132@Columbia.edu

Outline

- Logistic Regression: Keeping the basics basic
- It's a probability … Probably
- Frank Knight's great distinction: Risk vs Uncertainty
- Risky business: Prediction Intervals.
- But wait! There's more! … Backtesting!!
- What's important?: Model management
- The fine print. Details, Details, Details…
- CODA: Play it again Sam! Summary and Conclusions.



Logistic Regression: Keeping the basics basic

- Start with Good old linear regression.
- Remember the OLS assumptions?*
- Remember confidence intervals vs prediction intervals?
- We need to do the same thing with categorical predicions!

*

Linearity

Constant Error Variance

Normally Distributed Errors

Mean 0 Errors

Steps to Logistic Regression*

Step 1: Get the estimates onto the real line.

Step 2: Set regressors to the real line values.

Step 3: Postulate a model to solve for the coefficients

* “Logical-regression” get it??? Haaaa!

Step 1: Get the estimates onto the real line.

- y ∈ {no default, default}
- y ∈ {0, 1}
- p ∈ [0, 1]
- p/(1-p) ∈ [0, inf)
- ln(p/(1-p)) ∈ (-inf, inf)

Step 2: Set regressors to the real line values.

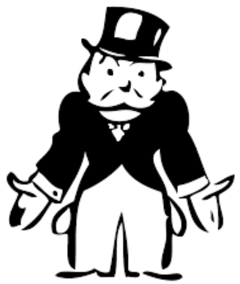
- $\ln(p/(1-p)) \in (-\text{inf}, \text{inf})$
- $\text{LogOdds}(p) \in (-\text{inf}, \text{inf})$
- $\text{LogOdds}(p) \equiv X\beta$
- $\text{LogOdds}^{-1}(\text{LogOdds}(p)) = \text{LogOdds}^{-1}(X\beta)$
- $p = \text{LogOdds}^{-1}(X\beta)$

Step 3: Postulate a Model to Solve

- $p = \text{Sigmoid}(X\beta)$
- $\text{argmax}(\beta) \prod (p^y * (1-p)^{1-y})$

It's a probability … Probably.

- What would it take to truly get a probability?
  - Span the domain of possible loan qualities.



  - Have even proportionality throughout the sample.
- How would we know?
  - Backtesting
- And what if we don’t know?
  - Then, we’re just rank ordering.

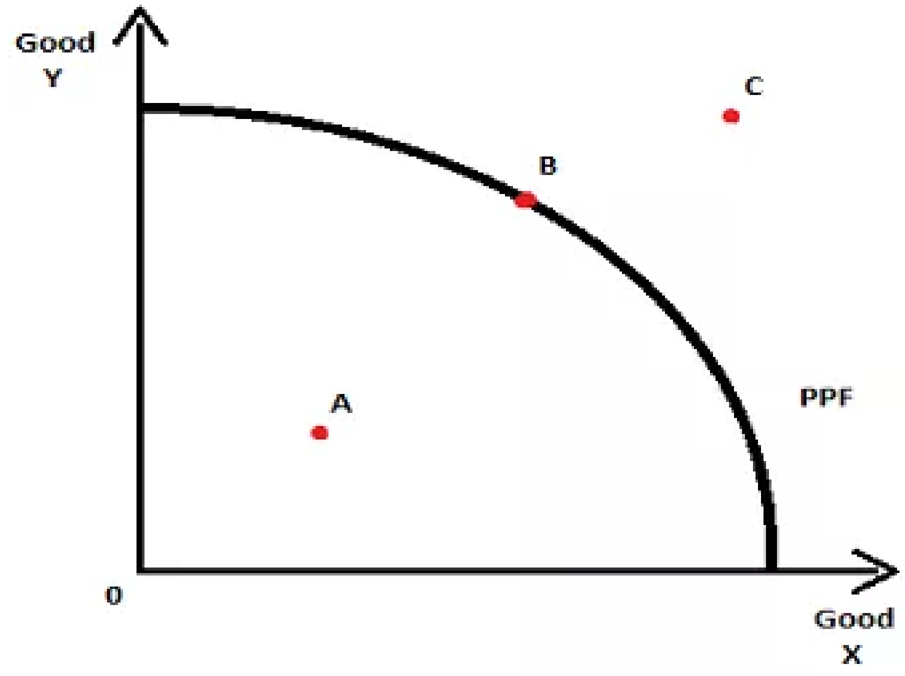
It's a probability ... Probably.

- The Receiver Operator Characteristic
- Birdie?

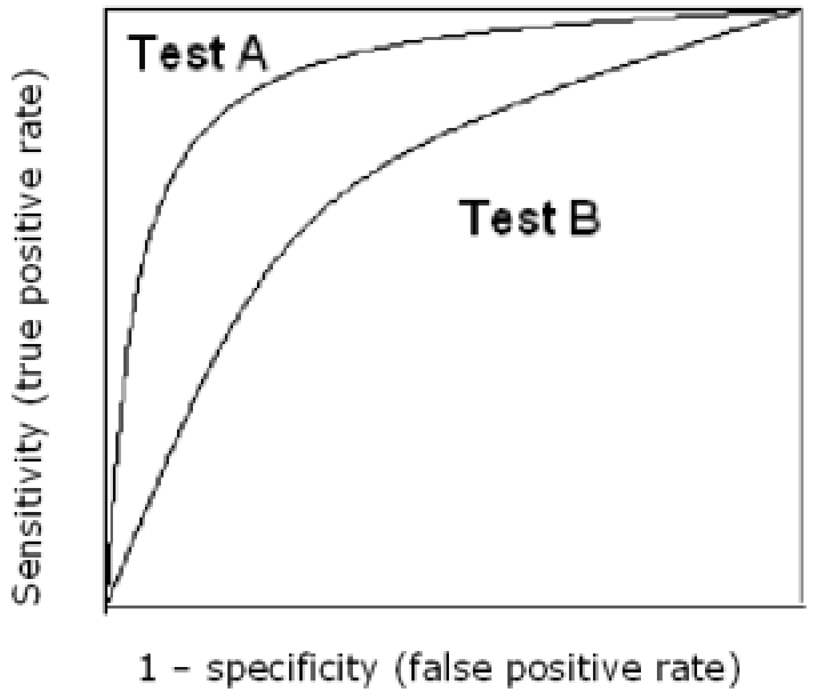
- or Bomber!!??

It's a probability … Probably.

Production Possibilities Frontier



ROC Curve



Frank Knight’s Great Distinction: Risk vs Uncertainty

- Risk: You know the distribution
- Uncertainty: You don’t know anything

Backtesting

- High Risk: wide confidence intervals and successful backtest.
- Low Risk: narrow confidence intervals and successful backtest.
- Uncertainty: unsuccessful backtest.

Pop Quiz!!!

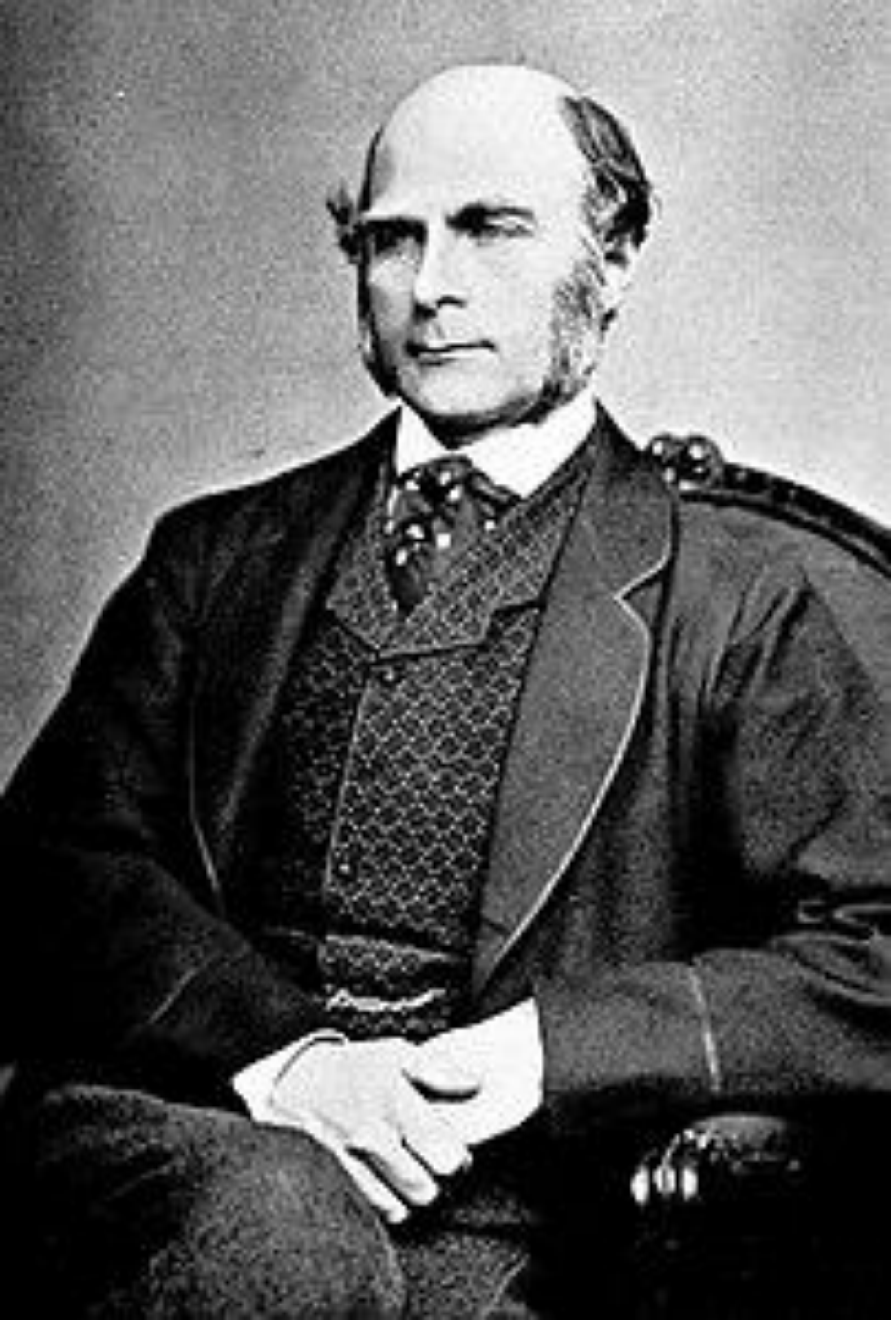
- Wide confidence intervals and successful backtest equals???
  - Bet on it! 

- Narrow confidence intervals and successful backtest equals???
  - Bet on it according to your risk tolerance! 

- Unsuccessful backtest equals???
  - Don’t bet on it! Consider mitigating or transferring. 

- Narrow confidence intervals and unsuccessful backtest equals???
  - Alien / Zombie / Nuclear Apocalypse



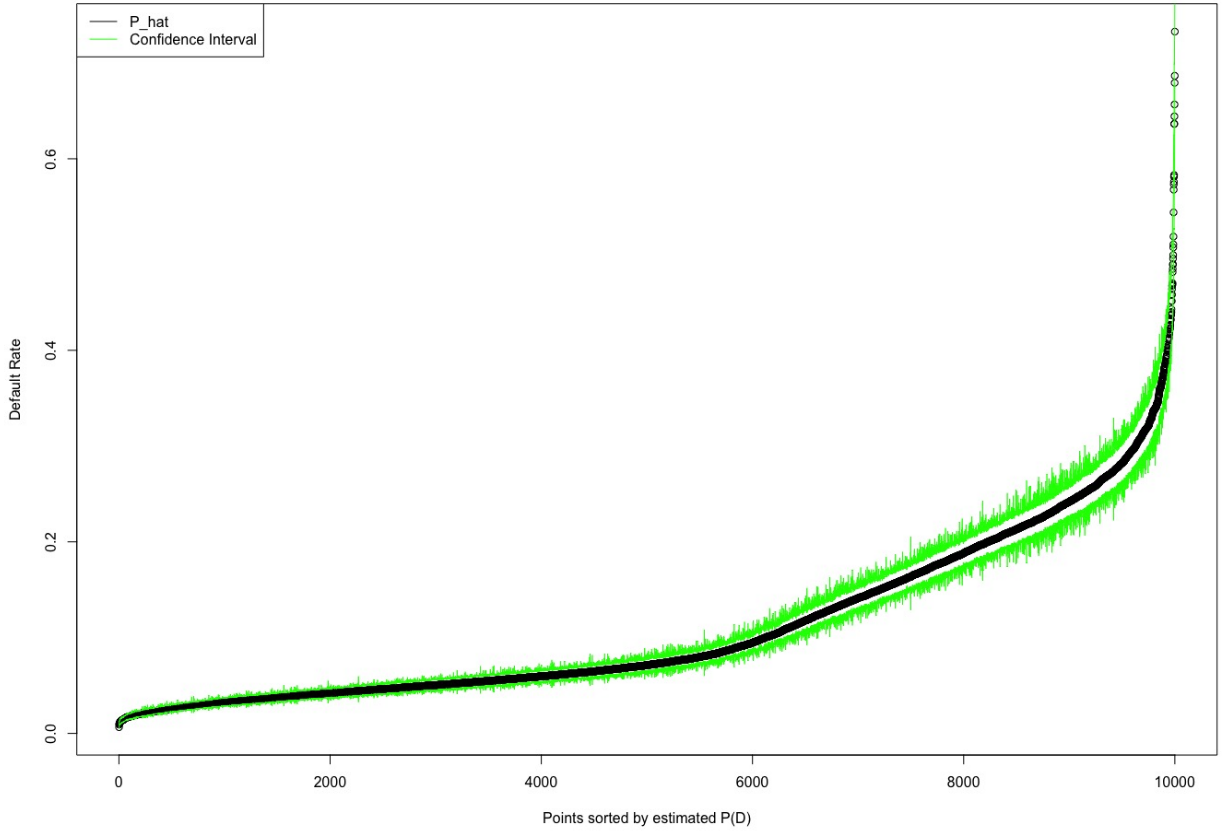



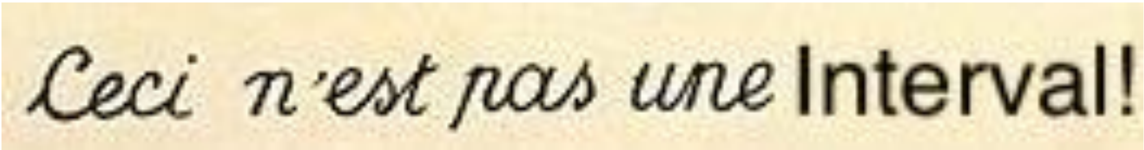
Risky business: Prediction Intervals

- Linear regression models population variance
- $y = X\beta + \varepsilon$

- Can we get a prediction interval for Logistic regression?
  - Short answer: No!!
  - Long answer: Yes!! (Denial always works. (Always!))

Note: Confidence Intervals do not Help





Making Our Prediction Interval

- Step 1: Get the variance of the expected estimates
- Step 2: Get the variance of the point estimates

Step 1: Get the variance of the estimates

- Simulate beta’s with multivariate normal
  - μ = beta’s
  - ∑ = cov(beta’s)

- Bootstrapping
  - This could take a while.

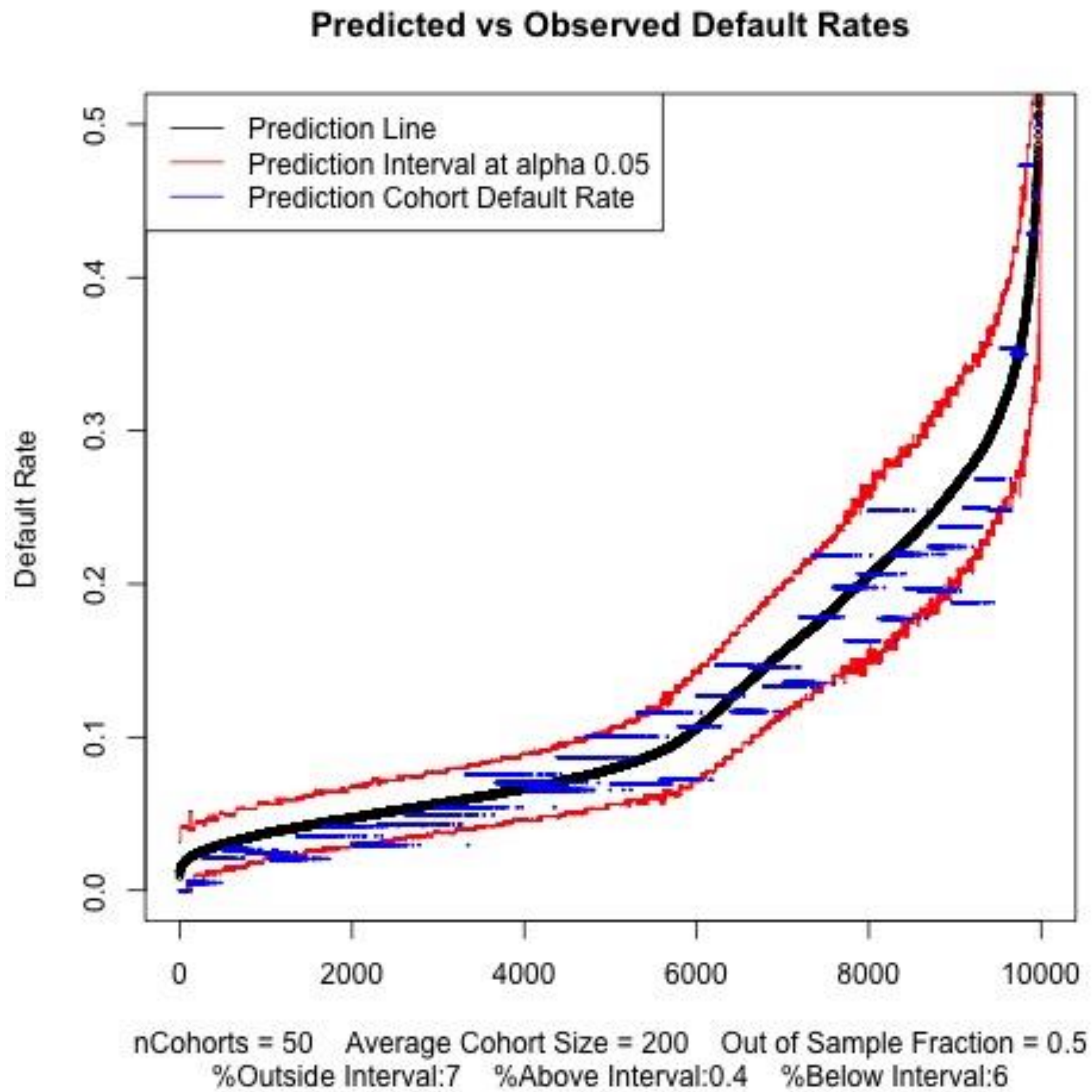
Step 2: Get the variance of point estimates

- Segment your population of data into cohorts.
- Each cohort c has a member count n_c
- Use the variance of beta to simulate a probability of default
  - $p_i = x_i^t * N(\mu, \sum)$
- Simulate the number of defaults for that point
  - $d_i = B(n_c , p_i)$
- Divide d_i by n_c to model the observed default rate

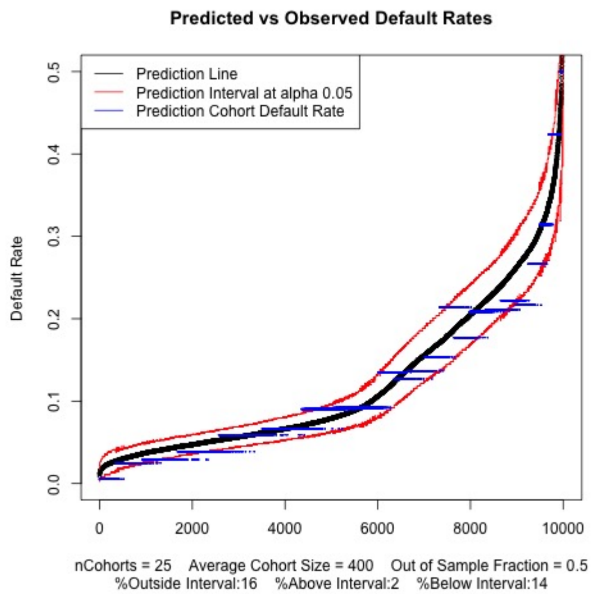
The Knobs We Turn

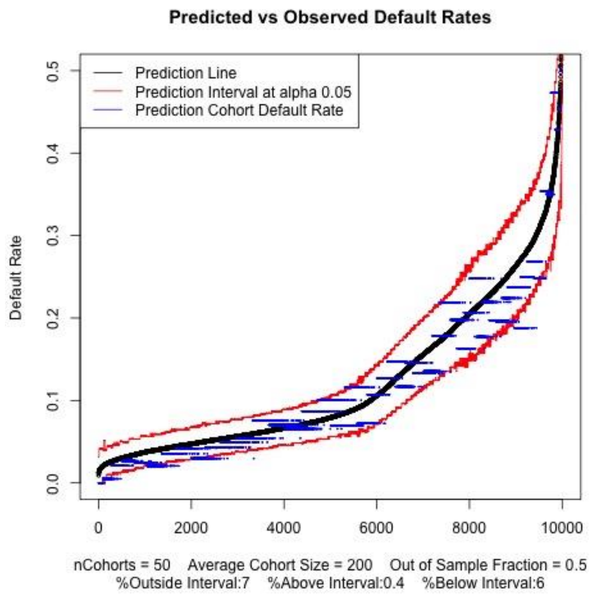
- Alpha
- Number of Cohorts
- Cohort Feature Space
- Out of Sample Size

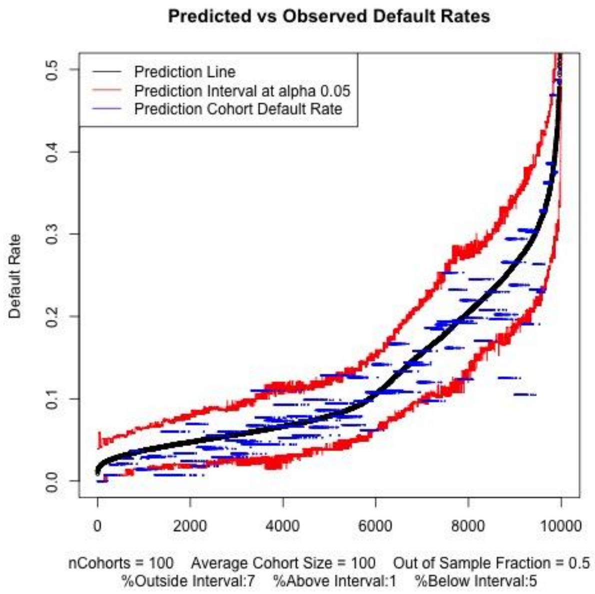

A Feel-Good Version

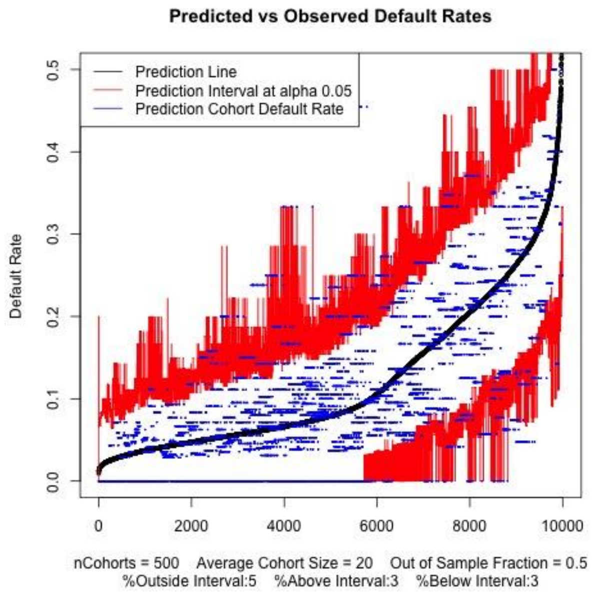


Cohort Count: 25, 50, 100, 500

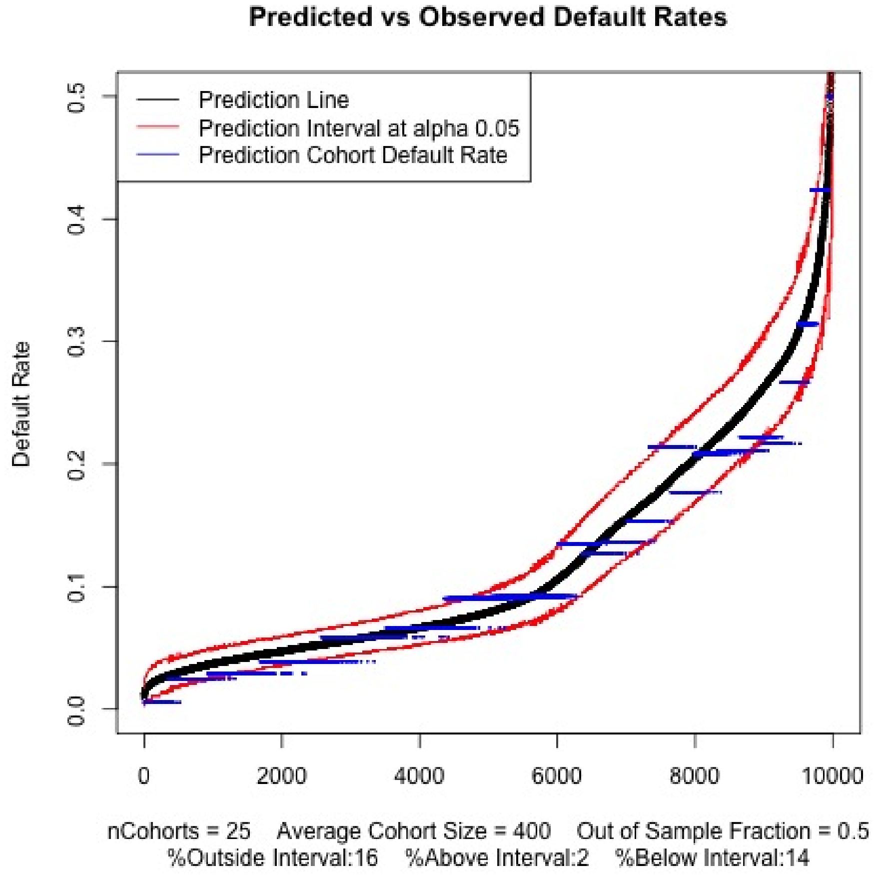


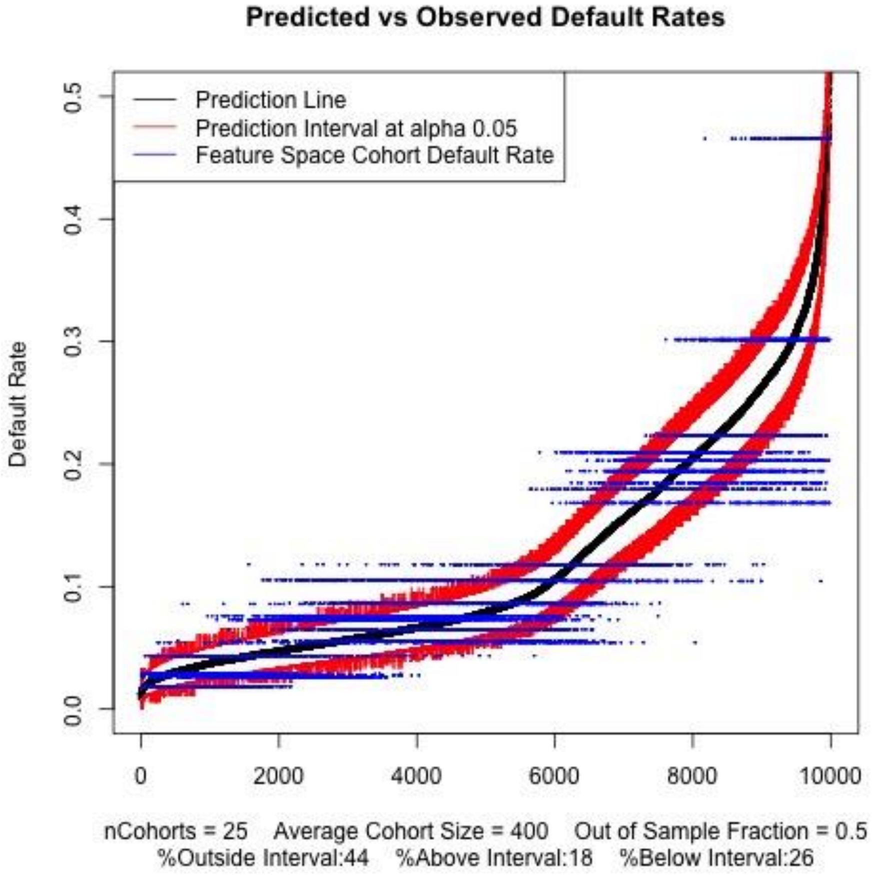






Identifying Cohorts from Feature Space





But wait! There's more! … Backtesting!!

- Keep the interval methodology constant and backtest.
- Track Risk and Uncertainty measures for cohorts.
  - Error bound width
  - Proportion of exceptions

The fine print. Details, Details, Details…

- Data
  - Data set should be made up of closed loans.
  - Maybe: Open loans can be added with an indicator variable.

- Cohorts
  - Cluster analysis can pick the cohorts.
    - You can use a scree plot to determine clusters.
    - In practice, you'll need enough clusters for performance to stabilize.

- Regularization
  - Regularized coefficients will be even harder to interpret.

What's important: Model management

- Know where you predict poorly.
- If data is scarce and the environment is changing, refit periodically.
  - At each refit:
    - Hold the data model and observe difference in predictions as function of data change.
    - Hold the data constant and observe difference in predictions as function of model change.

- Keep a baseline data set to offer absolute point of reference.
- Keep a simple transparent model running alongside a black box model.

CODA: Play it again Sam!
Summary and Conclusions

- Estimated probabilities might be nothing more than rank ordering.
- Failed probability estimation leads to failed expected loss calculation.
- You can use the Risk vs Uncertainty distinction for clarity of mind.
- Know where your model predicts poorly.
- Be mindful of closed vs open loans in your data set.

Appendix: Logistic Regression in context of GLM’s aka 'ruining the simplicity'

- Link function (logodds)
- Inverse link function (sigmoid)