Robustness Validation Of Predictive Models In Database Marketing

Charles Chen, Ph.D

TD Marketing, Customer Analytics, Senior Manager

November 04, 2015

Agenda

- Introduction
- Predictive Model Robustness
- Resampling
- Holdout Validation
- Enhanced Holdout Validation By Jackknifing
- Enhanced Holdout Validation By Bootstrapping
- Comparison of Two Enhanced Methods
- Proposed Model Robustness Index
- Examples

Introduction

- The main use of a predictive model is to apply it on a new data to generate prediction
- The performance and robustness of the model are two major factors to determine the quality of the prediction
- Processes and indexes (such as K-S D, c-statistics, Lift, etc.) have been successfully established and employed to assess model fit
- The robustness of a model has not yet been addressed sufficiently in model validation
- In this study, the methods of robustness validation based on resampling technique will be proposed and discussed

Predictive Model Robustness

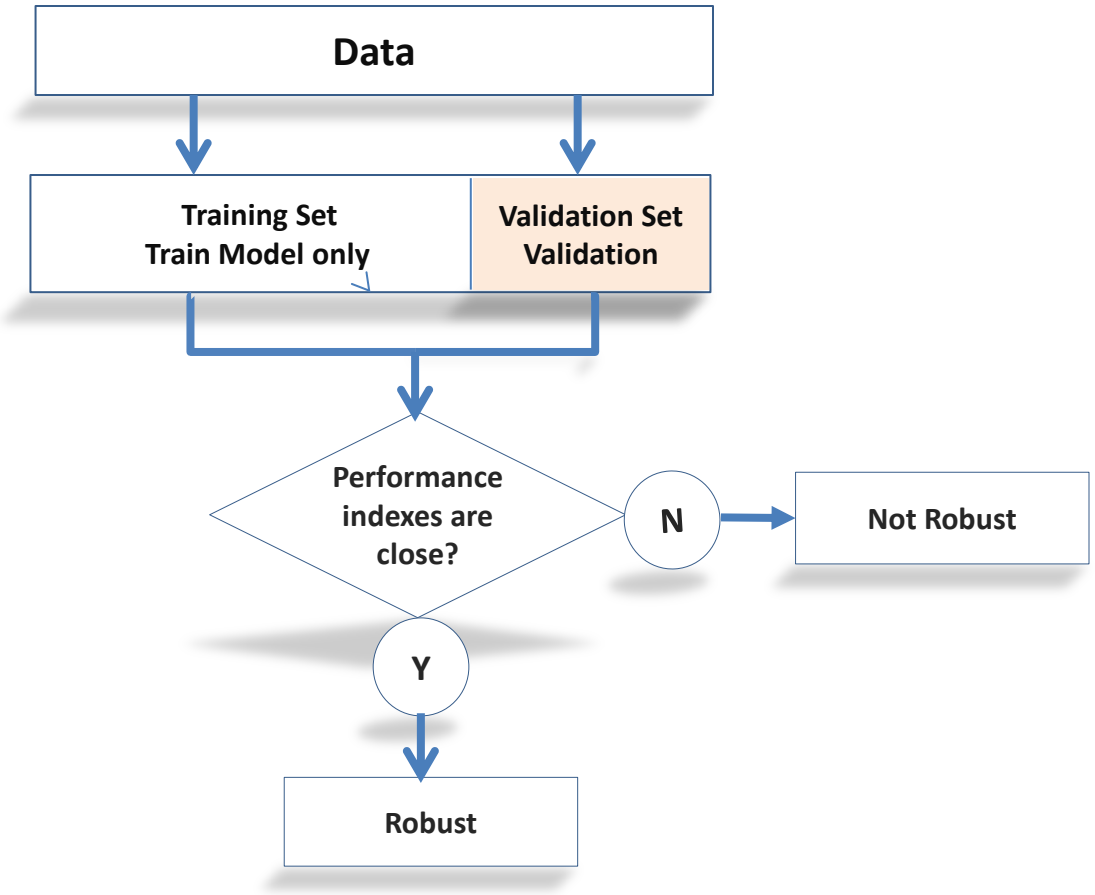
- Robustness of a predictive model refers to how well a model works on alternate data
- Robustness is subject to uncertainties in modeling due to:
  - The population of interest may be underrepresented in the modeling data
  - Model variables may be retained by chance when they are irrelevant, or relevant variables may be omitted
  - data problems

Resampling

- The most practical use of resampling methods is to derive confidence intervals and test hypotheses
- In model development, resampling is commonly used in two ways:
  - It avoids over-fitting by calculating model coefficients or estimates based on repeated sampling
  - It detects over-fitting by using repeated samples to validate the results of a model
- This study will only focus on the use of resampling as a validation technique

Holdout Validation

- Currently holdout validation is most commonly used in predictive model validation

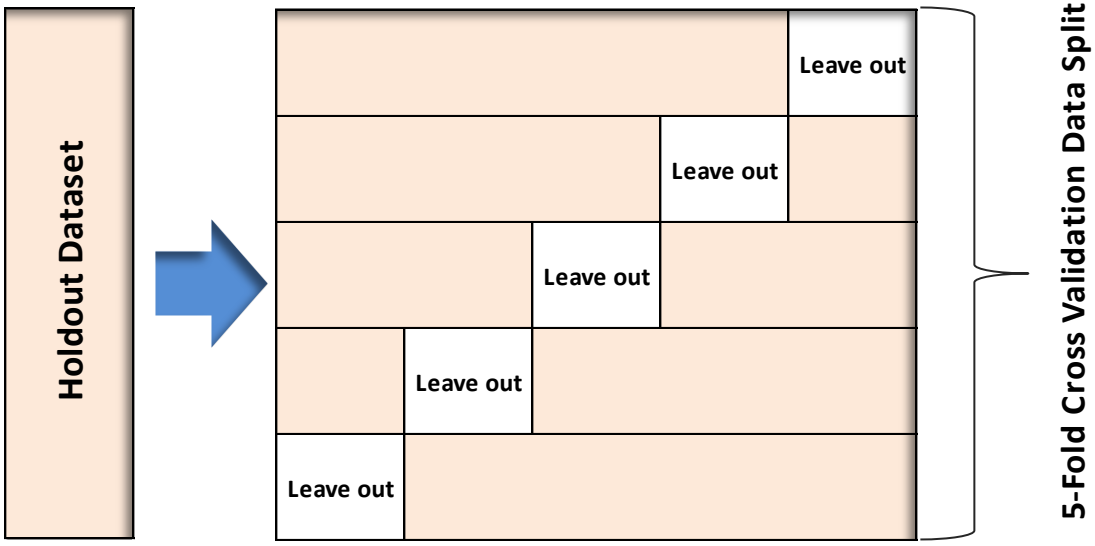


Holdout Validation Contd.

- **Pros**:
  - No parametric or theoretic assumptions
  - Easy to do and cheap
  - Conceptually simple
- **Cons**:
  - Have a high variance
  - Depending on how the division is made to create training and validation sets
  - Cannot provide the confidence interval of the model fit
  - Tends to overestimate the test error for the model fit on the entire data set
- **Enhanced Holdout Method**
  - Based on resampling techniques, propose enhanced holdout method to overcome the drawbacks stated as above

Enhanced Holdout Validation By Jackknifing

- Jackknifing is a resampling technique based on the "leave-one-out" principle. From this, a sample statistic can be used to estimate the bias and variance of the statistic
- Enhanced holdout validation by Jackknifing works on the same principle as leave-one-out. Instead of just one record, it leaves out one group of records



Enhanced Holdout Validation By Jackknifing Contd.

- **Steps** (Assume that original observed dataset has *N* records)
  - Generate n groups (each group has *N*/*n* records) , and leave one group out at a time
  - Create a set of groups, $X = \{X_1 \dots \dots X_n\}$
  - Compute $\hat{\theta}(X^{-i})$ a function of the data which estimates some parameter θ of the model as follows:

    For *i =1* to *n*

    - generate a jackknife sample $X^{-i} = \{X_1, \dots X_{i-1}, X_{i+1} \dots X_n\}$ by leaving out the *ith* group
    - calculate $\hat{\theta}_{-i}$ by applying the estimation process to the jackknife sample
  - Calculate the jackknifed estimates of mean, variance and confidence interval to judge stability

Enhanced Holdout Validation By Jackknifing Contd.

- **What should number of groups be?**
  - Basically, the smaller the number of groups, the higher biased the parameters estimates but the lower variances they will be
  - Understanding the Bias-Variance Tradeoff is important when making the decision
  - In practice, 100 (to 1000) - groups are generally effective group

Enhanced Holdout Validation By Bootstrapping

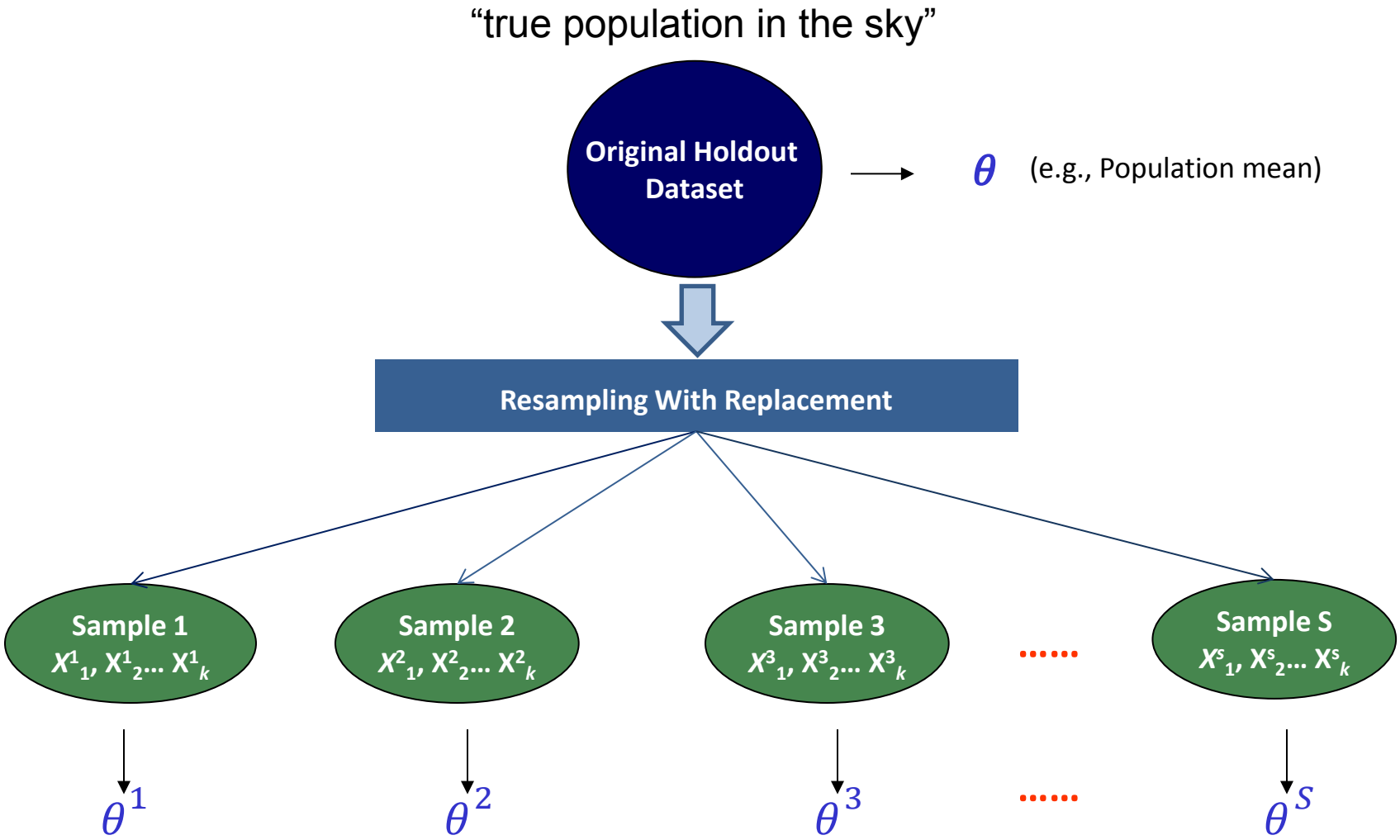
- Bootstrapping is a statistical method for estimating the sampling distribution of an estimator by sampling with replacement from the original sample
- Its purpose is to derive robust estimates of standard errors and confidence intervals of a population parameter
- **Steps** (Assume that original observed dataset has ***N*** records)
  - Create ***s*** bootstrap samples, each has ***k*** *(<=**N**)* records, denoted by $X = \{X^1 \dots X^s\}$
  - Compute $\hat{\theta}(X)$ a function of the data which estimates a parameter θ of the model as follows:

For *i =1* to ***s*** *(where s is the number of bootstrap samples being generated)*

- generate a bootstrap sample $X^i = \{x_1^i \dots \dots x_k^i\}$ by sampling with replacement from original observed dataset. Generally, let ***k*** = ***N***.
- Compute $\hat{\theta}^i = \hat{\theta}^i(X^i)$ in the same way that you calculated the original estimate θ

  - Calculate the bootstraps estimates of mean, standard error and C.I. of $\hat{\theta}$ (e.g., R^2, RMSE, etc.) to judge stability

Enhanced Holdout Validation By Bootstrapping Contd.



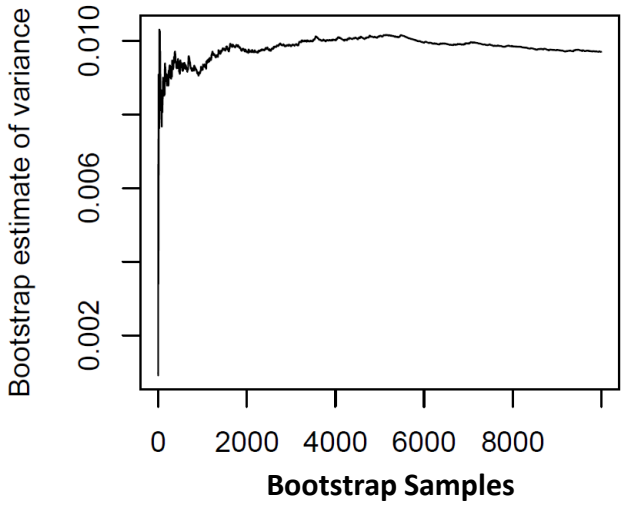
Enhanced Holdout Validation By Bootstrapping Contd.

Key Assumptions

- The bootstrap is nearly assumption free. The key assumption is that the distribution of your sample, ***s***, is a close approximation to the population distribution
- Tends to be true as size of bootstrap ***k*** becomes large. Max(***k***) = ***N***
- But if ***k*** is small, this is precisely when violations of assumptions for traditional statistics matters most
- Note that inference will always be suspect for small samples. No statistical approach can make that go away

Enhanced Holdout Validation By Bootstrapping Contd.

- **How many bootstrap samples to use?**
  - **Method A:** Determine the number of bootstrap samples (***s***)--- plot the value of the bootstrap estimate of variance against ***s*** to see whether it has settled down to some value



  - **Method B**[1,2]**:** most practitioners suggest that number of bootstraps should be between 1000 and 2000 to compute 95% confidence intervals

1: Efron B, Tibshirani RJ. An Introduction to the Bootstrap. Chapman and Hall: London, 1993

2: Davison AC, Hinkley DV. Bootstrap Methods and their Application. Cambridge Univ. Press: Cambridge, 1997

Enhanced Holdout Validation By Bootstrapping Contd.

Confidence Interval and Bias Correction

- When the bootstrap distributions are not skewed , the 95% confidence interval can be obtained by taking 2.5% and 97.5% quantiles of the bootstraps as the lower and upper bound respectively
- The bootstrap distribution may be skewed, which can cause a bias in the estimation of confidence interval
- When the bootstrap distributions are skewed , the *BCa* bootstrap[3] can adjust for both bias and skewness in the bootstrap distribution. (*BCa* bootstrap is beyond the scope of this presentation but further information can be found in the reference

3: Efron, B. (1987). "Better Bootstrap Confidence Intervals". Journal of the American Statistical Association, Vol. 82, No. 397, p171–p185.

Comparison of Two Enhanced Methods

- Each can be seen as approximation to the other, though there are significant theoretical differences in their mathematical insights
- Jackknifing is primarily concerned with calculating standard errors of statistics of interest and for reducing bias. However, if understanding the distribution of the parameter is a primary goal, Bootstrapping is generally superior
- The Jackknife method is suitable for smaller original data samples
- Bootstrap can estimate not only the standard error but also the distribution of a statistic
- The bootstrap method handles skewed distributions better
- The bootstrap estimate of model prediction bias is more precise than Jackknife estimates with linear models such as linear discriminant function or multiple regression[4]

4. Verbyla, D.; Litvaitis, J. (1989). "Resampling methods for evaluating classification accuracy of wildlife habitat models". Env. Management. 13 (6): 783–787

Proposed Model Robustness Index

About Robustness Index

- The bandwidth of 95% confidence interval of a performance index (e.g., statistic-c, KS, and Lift) can be used as a predictive model robustness index to measure the stability
- In addition to performance indexes, the robustness indexes should be reported in model documentation to provide the entire view of model's quality
- For model comparison, the lower the robustness measure value, the higher robust the model!

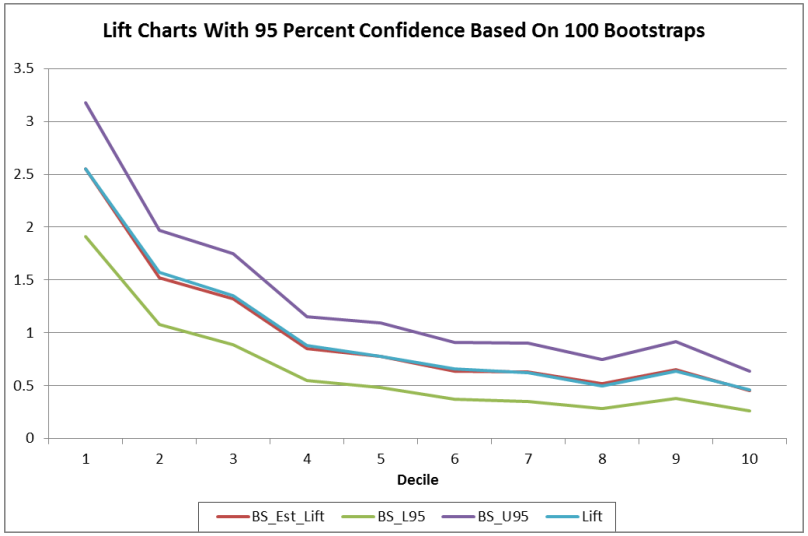
Examples

Response Model of Product A

- There are 492 responders in total giving a response rate about 0.71%
- 60% of responders (295) are used for model training, and the rest (197) for validation
- The Lift confidence interval is relatively wide, likely due to the limited responders in this case; implying that the model is moderately robust

	Training	Validation		
		L95_BS_Est	BS_Est	U95_BS_Est
c-statistic	0.691	0.658	0.688	0.721
KS	0.293	0.254	0.291	0.327

Decile	BS_Est_Lift	BS_L95	BS_U95	Lift
0	2.55	1.91	3.18	2.55
1	1.52	1.08	1.97	1.57
2	1.32	0.89	1.75	1.35
4	0.85	0.55	1.15	0.88
3	0.78	0.48	1.09	0.78
5	0.64	0.37	0.91	0.66
6	0.63	0.35	0.9	0.62
7	0.52	0.28	0.75	0.5
8	0.65	0.38	0.92	0.64
9	0.45	0.26	0.64	0.46



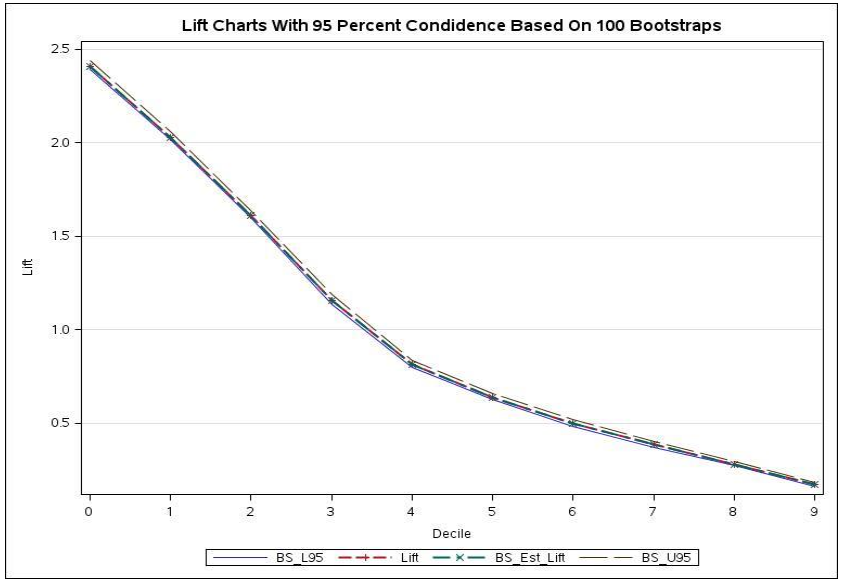
Examples Contd.

Propensity Model of Product B

- There are 134,313 responders in total giving a response rate about 32.86%
- 61.9% of responders (82,954) are used for model training, and the rest (51,359) of responders for validation
- The Lift confidence interval is quite narrow, implying that the model is robust

	Training	Validation		
		L95_BS_Est	BS_Est	U95_BS_Est
c-statistic	0.783	0.778	0.781	0.787
KS	0.432	0.429	0.431	0.434

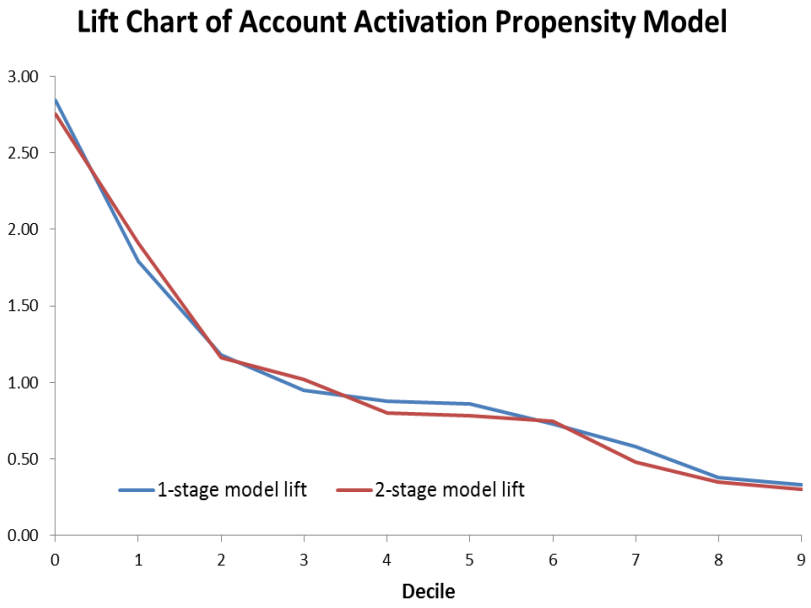
Decile	BS_Est_Lift	BS_L95	BS_U95	Lift
0	2.410	2.394	2.439	2.410
1	2.031	2.017	2.060	2.030
2	1.614	1.600	1.638	1.613
3	1.158	1.138	1.190	1.158
4	0.815	0.800	0.838	0.816
5	0.641	0.627	0.661	0.639
6	0.497	0.485	0.521	0.498
7	0.384	0.372	0.404	0.384
8	0.281	0.272	0.298	0.281
9	0.169	0.161	0.183	0.171

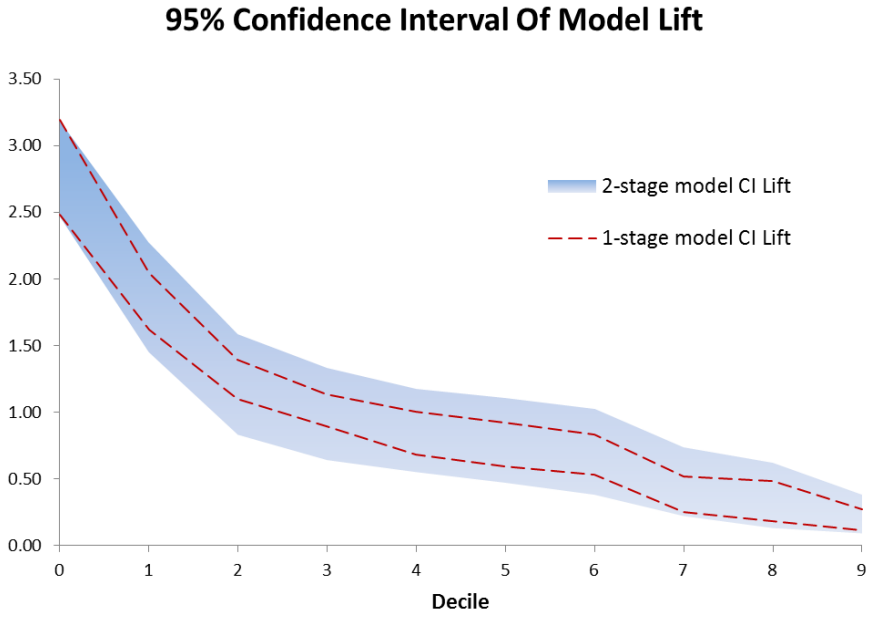


Examples Contd.

Propensity Model of Product C for Account Activation

- There are two options to calculate the probability of activation, build one model that predicts activation directly or build two models, one for response and one for activation given response
- It is difficult to determine which method works better since two methods produce very similar results. However, 1-stage model has higher stability due to the narrow confidence interval of lift





Questions and Comments