And the winner is…? How to Pick a Better Model

Ben Williams, Director, Willis Towers Watson
Hoi Leung, Director of Predictive Analytics, AIG

CAS Anti-Trust Notice

- The Casualty Actuarial Society is committed to adhering strictly to the letter and spirit of the antitrust laws. Seminars conducted under the auspices of the CAS are designed solely to provide a forum for the expression of various points of view on topics described in the programs or agendas for such meetings.
- Under no circumstances shall CAS seminars be used as a means for competing companies or firms to reach any understanding – expressed or implied – that restricts competition or in any way impairs the ability of members to exercise independent business judgment regarding matters affecting competition.
- It is the responsibility of all seminar participants to be aware of antitrust regulations, to prevent any written or verbal discussions that appear to violate these laws, and to adhere in every respect to the CAS antitrustcompliance policy.

Agenda

- Motivation
- What do we mean by fit?
  - Goodness of fit vs performance
  - Qualitative vs quantitative
- Sampling
- Various measures in more detail
- Comparisons
- Bootstrap
- Conclusion
  - Considerations on Context and Usefulness
  - Business-related Metrics

Motivation

- Is the model good enough?
- Is the fit too good to be true (training, test, holdout data)?
- Is the prediction stable across data subsets or over time
- Is the model be highly influenced by outliers?

Goodness-of-fit vs Performance

- Goodness-of-fit are statistics that measure the distance between the actual and the predicted
- Performance statistics measure other qualities of the models. For example:
  - ordering of the observations (segmentation)
  - comparisons of models

Qualitative vs Quantitative measures

- Quantitative measures
  - Provide a number that assesses the model
  - Provide a conclusion
  - Don’t help to diagnose the model
  - Are often used to optimize model fitting
- Qualitative measures
  - Provides more information (usually a chart)
  - Not as easy to declare a winner

Sampling

- Sampling Setup is very important
- Breaking data into training and/or test sample and out-of-sample validation
- Sampling method for train/test split and out-of-sample split doesn’t need to be consistent
- It is important to assess final model fit on validation data that was not touched in model construction
- Sample should be based on the question to be answered

Training/Test Sample

- Used for model building including hyper-parameters selection
- Two approaches:
  - Initially split dataset into training and test, build model on training, and measure fit on test
  - Cross-validate –repeatedly use one subset to build and one to test
- Can randomly split dataset, or can split based on a control variable (like year)

Out-of-sample Validation

Sample used for final model validation and can be done based on the question to be answered

- Loss Cost Model with Rating Variables
  - Random Sample on policy or endorsement?
  - Make sure no overlapping policies?
  - Just take the available latest year(s)?
- Territory Model
  - withhold a random list of zip codes?
- Vehicle Symbol Model
  - withhold a random list of vehicle types?

Measures considered

	Performance	Goodness of fit
Quantitative	• Gini Index • Normalized Gini Index • Precision and recall	• Squared Error & Absolute Error • Likelihood • AIC & BIC • Deviance
Qualitative	• Gini index plots • Lift Charts or Quantile Plots • Double lift charts • Loss ratio charts • Precision and recall plots	• Actual vs Predicted Target Plot • Residuals

2 Candidate Models for 2 Problems

To illustrate some of these measures, we have defined 2 problems:

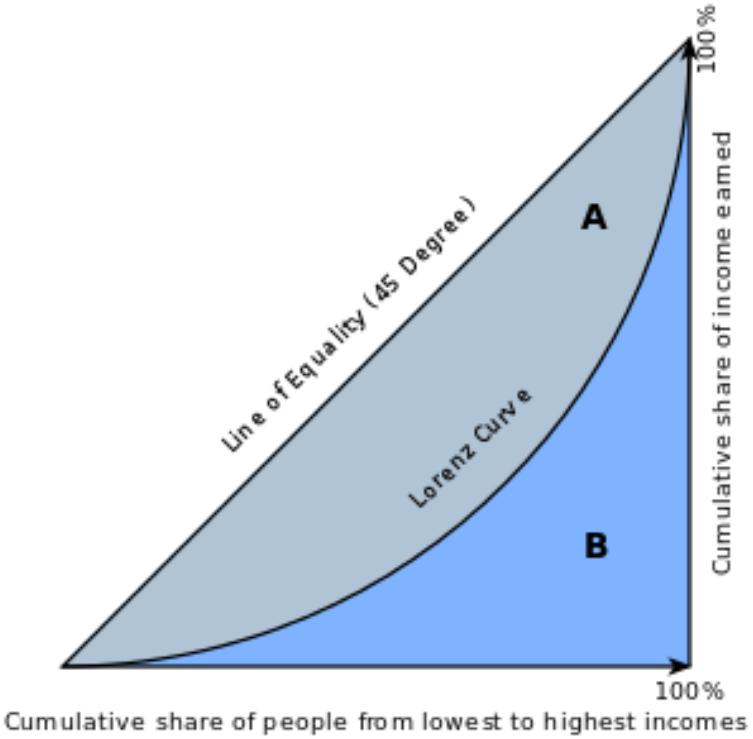
- Problem 1 is a binomial classification problem (classifying claims according to some measure of complexity)
- Problem 2 is a regression problem (severity)

For each problem we have fit 2 models, which we'll designate as Model A and Model B. At the end we'll reveal some information on the models

Gini Index

Measure of how well model classifies risks based on model prediction (e.g. predicted pure premium)

1. Sort holdout data by predicted pure premium and random number
2. Horizontal axis = cumulative percentage of weight (e.g. earned car years)
3. Vertical axis = cumulative percentage of actual response (e.g. reported loss)
4. A = Area between line of equality and Lorenz Curve
5. B = Area beneath Lorenz Curve
6. Gini index = A / (A + B)



How do our models perform?

Classification		
	Model A	Model B
Gini	0.8939	0.9079

Regression		
	Model A	Model B
Gini	0.2776	0.3736

Gini Index

- A difficulty with the Gini is to understand how much better a model with a Gini of 0.8939 is than a model with a Gini of .9079.
- A way to develop some intuition about this is to remove “important” and “unimportant” variables from the model and see how the Gini changes.

Normalized Gini Index

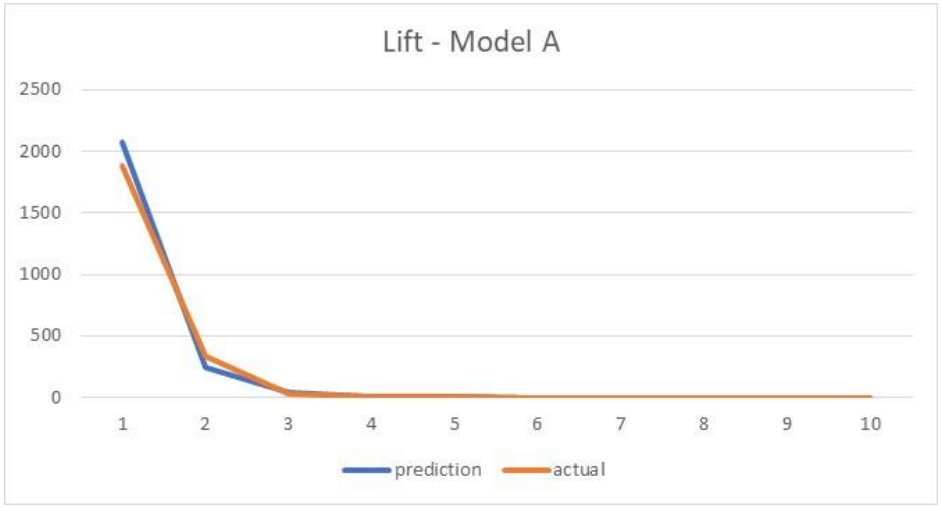
- This is the Gini index divided by the best possible Gini index
- Instead of sorting holdout data by model prediction, best possible Gini sorts holdout data by the actual responses
- This helps to put different types of model (loss cost, retention, etc.) into similar scale

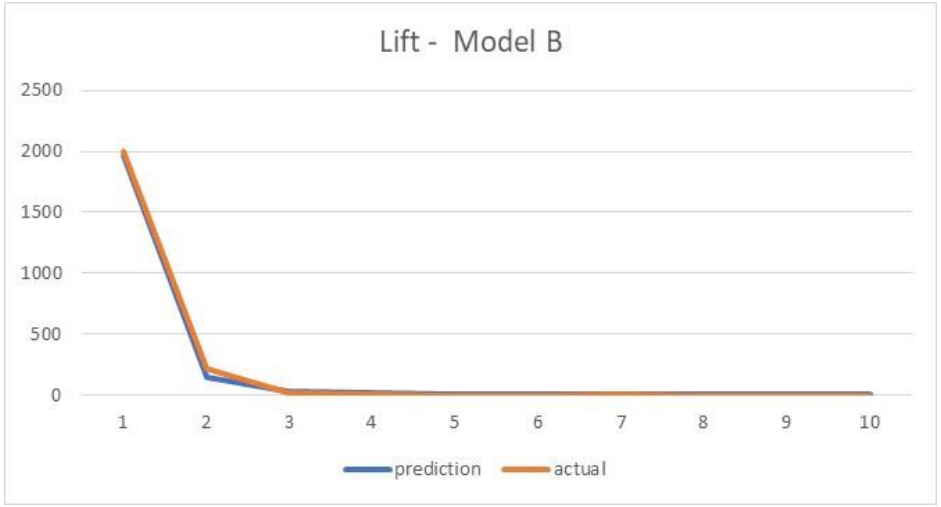
Lift Charts or Quantile Plots

1. Sort holdout data based on predicted values.
2. Subdivide sorted data into quantiles with equal weight.
   - Use exposure weights for frequency and pure premium.
   - Use claim count weight for severity.
3. Calculate average actual value and average predicted value for each quantile and (optionally) index to their overall average.

How do our models perform?

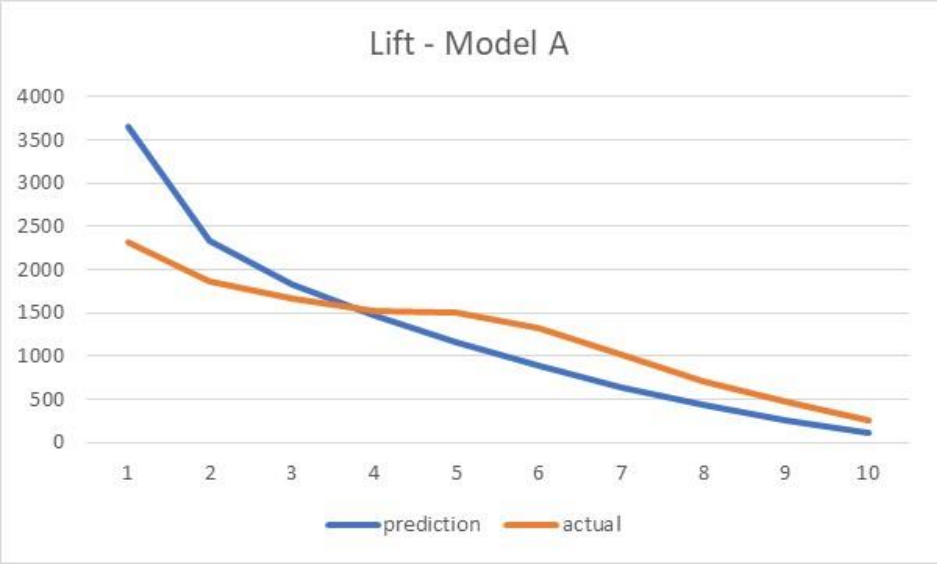
Classification Models

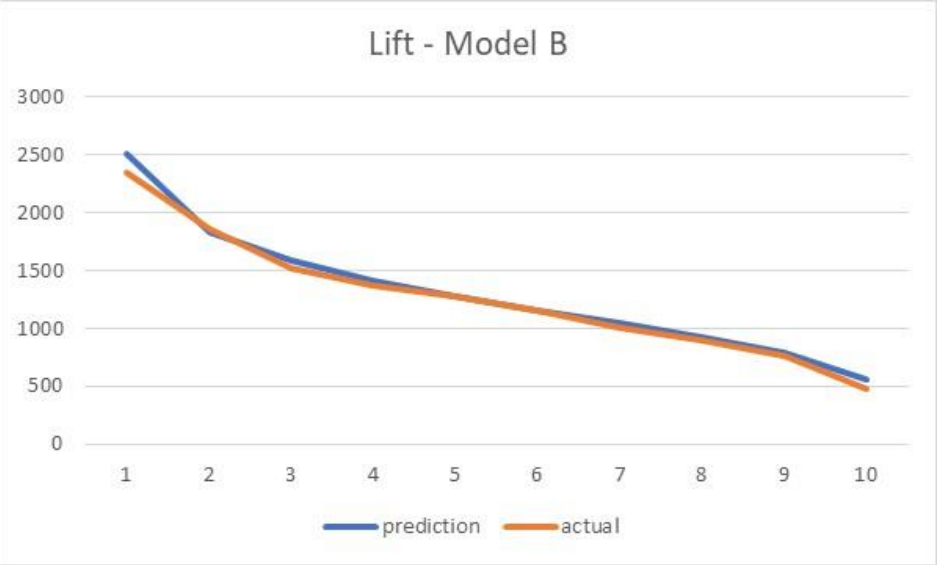




How do our models perform?

Regression Models





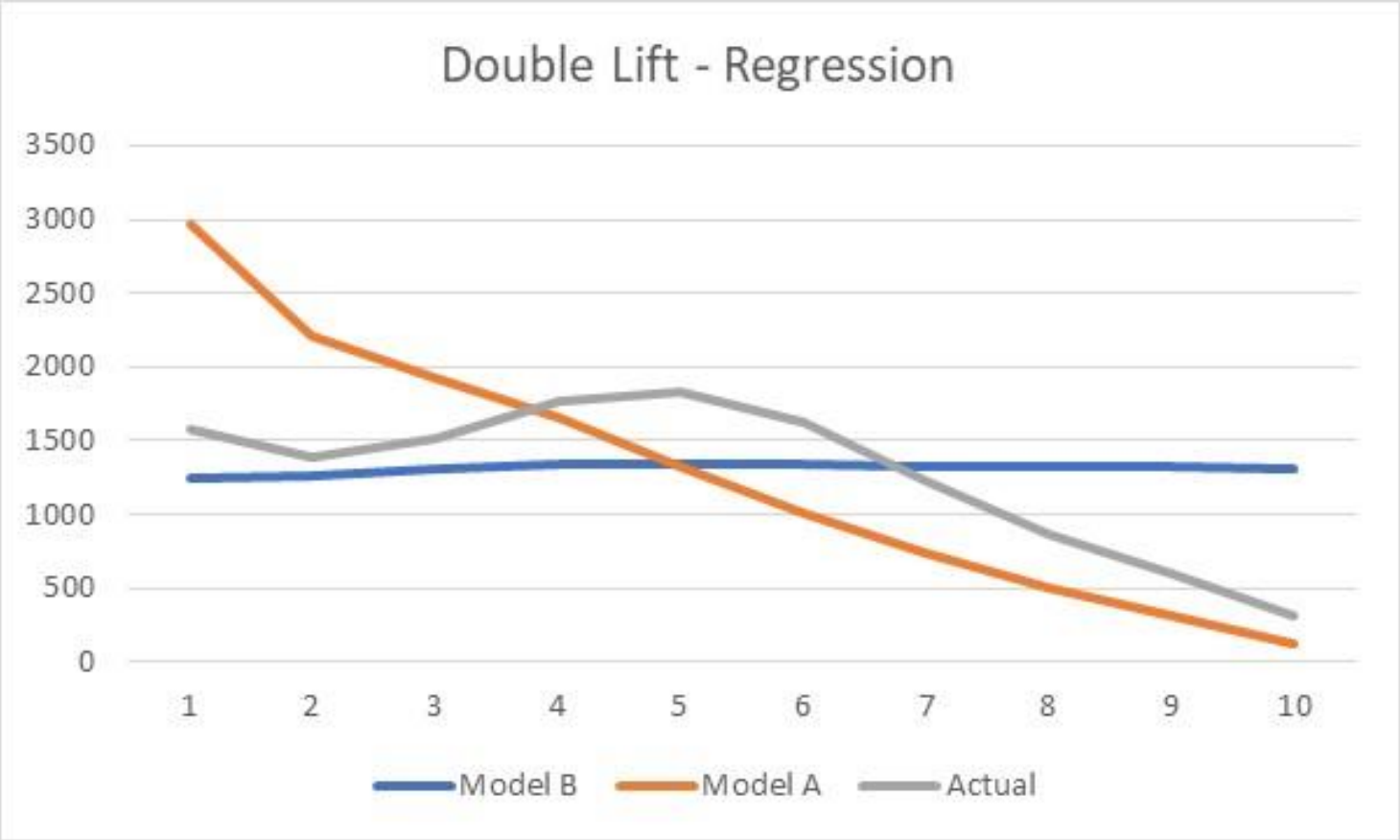
Double Lift Charts

Similar to Lift Charts, but allow us to compare model predictions

1. Sort holdout data by ratio of model predictions.
2. Subdivide sorted data into quantiles with equal exposure.
3. For each quantile calculate average actual value and average predicted value for each model.
4. (Optionally) index the quantile averages to the overall averages.

How do our models perform?

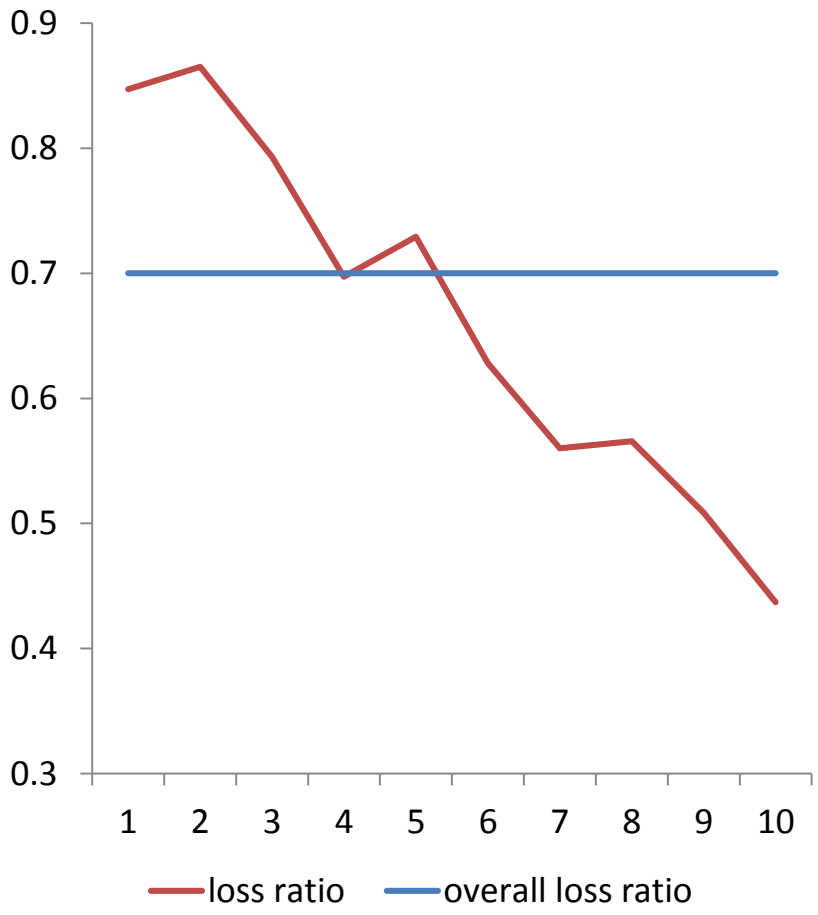
How do our models perform?



Loss Ratio Chart

Similar to double lift chart but more applicable to loss cost models

1. Sort holdout data by ratio of predicted pure premium to current premium
2. Subdivide sorted data into quantiles with equal exposure
3. For each quantile calculate loss ratio
4. (Optionally) index loss ratios to the overall average



Squared Error & Absolute Error

- For each record, calculate the squared or absolute difference between actual and predicted target variable
- Easy and intuitive, but generally inappropriate for insurance data, and can lead to selection of wrong model
- Squared error is appropriate for normally distributed data, but insurance data is generally not Normal

How do our models perform?

Classification		
	Model A	Model B
Gini	0.8939	0.9079
RMSE (Root Mean Square Error)	0.2380	0.1793
MAE (Mean Absolute Error)	0.0566	0.0321

Regression		
	Model A	Model B
Gini	0.2776	0.3736
RMSE (Root Mean Square Error)	0.1203	0.1202
MAE (Mean Absolute Error)	0.0272	0.0277

Likelihood

- The probability, as predicted by our model, that what actually did occur would occur
- A GLM calculates the parameters that maximize likelihood
- Higher likelihood implies better model fit (in very simple terms)
- A problem with likelihood is that adding a variable always improves likelihood, so it isn't very useful. However…

AIC & BIC

- Akaike Information Criterion (AIC):

$$-2 \cdot (\log likelihood) + 2 \cdot (\#\ of\ Parameters\ in\ Model)$$

- Bayesian Information Criterion (BIC):

$$-2 \cdot (\log likelihood) + (\#\ of\ Parameters\ in\ Model) \cdot \log(\#\ of\ Records\ in\ Dataset)$$

- These are penalized measures of fit
- These can provide a rule for deciding which variables to include – unless a variables improves AIC or BIC, don't include it
- BIC is often too restrictive
- Requires that the number of parameters in the model can be counted

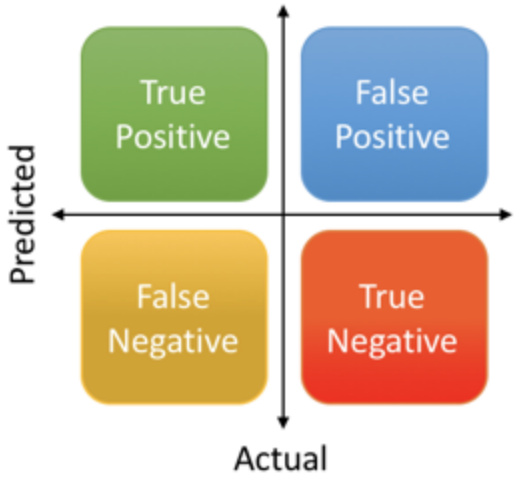
Precision and Recall

- For (binomial) classification models only…

$$\textbf{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

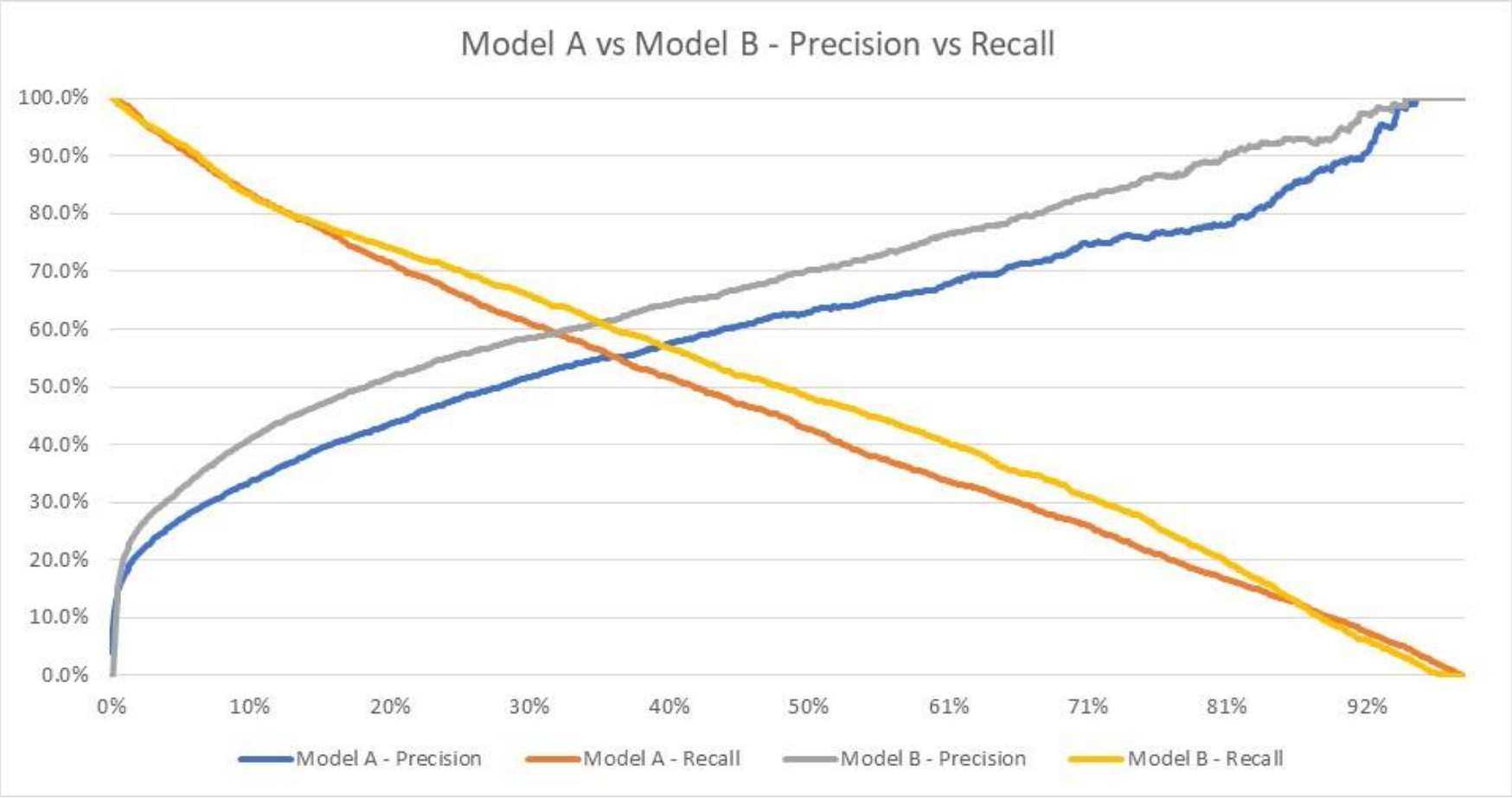
$$\textbf{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

$$\textbf{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{Total}}$$



- Each are defined for a given threshold of application (i.e. a probability above which a predicted value is assigned a 1)

How do our models perform?



Deviance

- Saturated model –the model with the highest possible likelihood
  - One indicator variable for each record, so model fits data perfectly
- Deviance = 2*(loglikelihood of saturated model – loglikelihood of fitted model)
- GLMs minimize deviance
- Like squared error, but reflects shape of assumed distribution
- We generally fit skewed distributions to insurance data (Tweedie, gamma, etc), and thus deviance is more appropriate than squared error

Deviance – in Math

Binary $$2\sum_i y_i \log\left(\frac{y_i}{\mu_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \mu_i}\right)$$

Poisson $$2\sum_i w_i \left(y_i \ln\frac{y_i}{\mu_i} - y_i + \mu_i\right)$$

Gamma $$2\sum_i w_i \left(-\ln\frac{y_i}{\mu_i} + \frac{y_i - \mu_i}{\mu_i}\right)$$

Tweedie $$2\sum_i w_i \left(y_i \frac{y_i^{1-p} - \mu_i^{1-p}}{1-p} - \frac{y_i^{2-p} - \mu_i^{2-p}}{2-p}\right)$$

Normal $$\sum_i w_i (y_i - \mu_i)^2$$

Benefit of Deviance over Square Error

- Since squared error is the deviance of a regression model with a Normal distribution, using squared error for non-Normal data can lead to the incorrect model being chosen
- We can run two loss cost models on a dataset – one with a Tweedie distribution and one with a Normal distribution
- While the data is far from Normal, but using squared error as a metric, the Normal GLM wins
  - Even absolute error shows the Normal winning

How do our models perform?

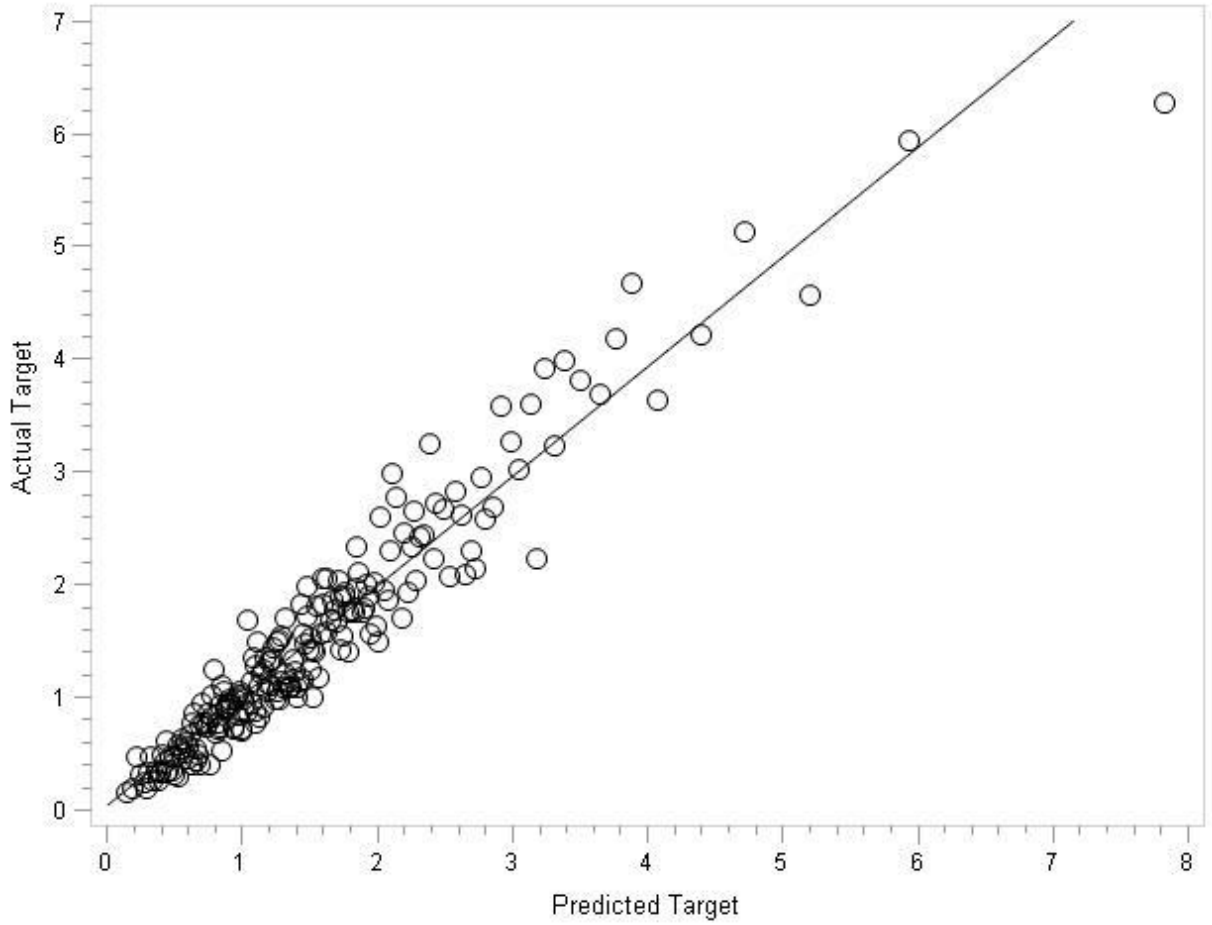
Classification		
	Model A	Model B
Gini	0.8939	0.9079
RMSE (Root Mean Square Error)	0.2380	0.1793
MAE (Mean Absolute Error)	0.0566	0.0321
R^2	0.2619	0.3055
Deviance	0.1638	0.1455

Regression		
	Model A	Model B
Gini	0.2776	0.3736
RMSE (Root Mean Square Error)	0.1203	0.1202
MAE (Mean Absolute Error)	0.0272	0.0277
R^2	0.0028	0.0024
Deviance	0.2440	0.2431

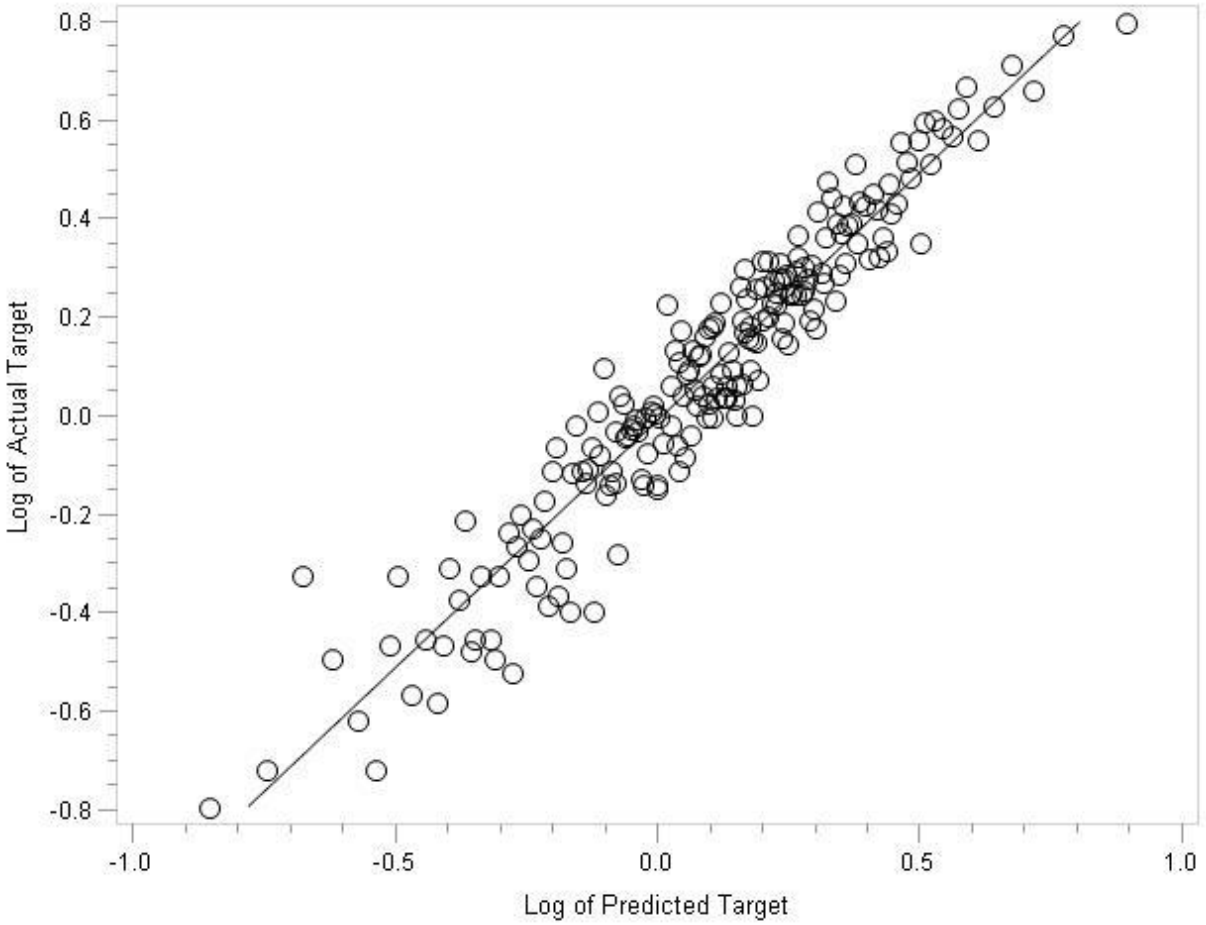
Actual vs Predicted Target

- Scatter plot of actual target variable (on y-axis) versus predicted target variable (on x-axis)
- If model fits well, then plot should produce a straight line, indicating close agreement between actual and predicted
  - Focus on areas where model seems to miss
- It may be necessary to bucket (such as into percentiles)
- Depending on scale, it may be necessary to plot on a log-log scale

Example of Actual vs Predicted

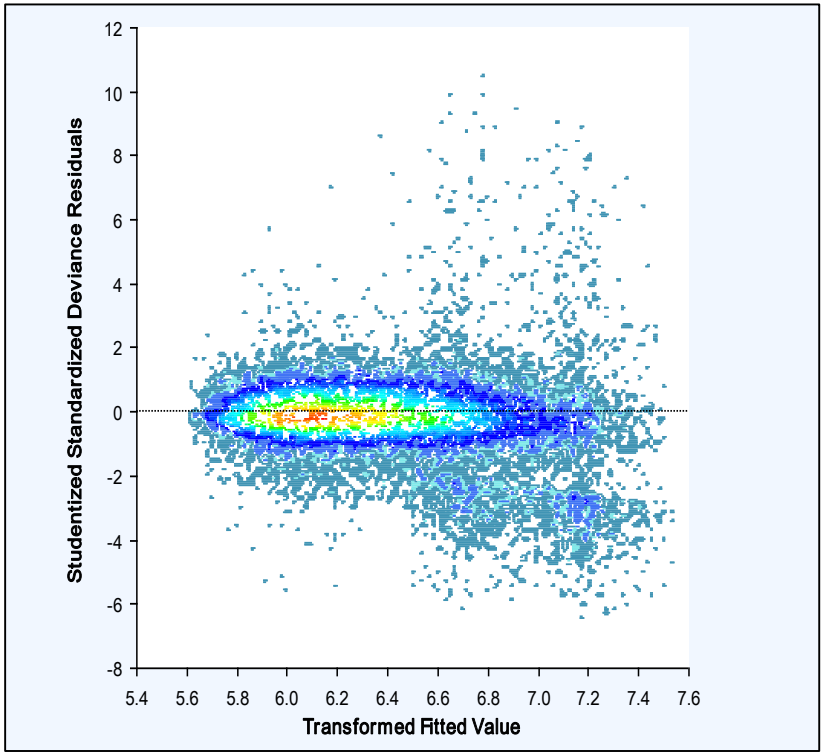


Example of Log of Actual vs Log of Predicted



Residuals

- Residuals are functions of actual and expected target variables
- There are various kinds of residuals (we won't explore them here)
- A pattern like the one here is desirable



Bootstrapping

- This is a re-sampling technique that allows us to get more out of our data
- Start with a dataset and sample from it with replacement
  - Some records will get pulled multiple times, and some will not get pulled at all
- Generally, we create a dataset with the same number of records as our original dataset
- We can create many bootstrap datasets, and each dataset can be thought of as an alternate reality
  - Since each bootstrap is an alternate reality, we can use bootstrapping to construct confidence intervals

Bootstrap CIs for Parameter Estimates

- Some parametrized models (e.g. elastic net) do not produce confidence intervals for parameters
- GLMs produce confidence intervals for parameter estimates, but it is valuable to get a second opinion
- To do this, create many bootstrap datasets, re-run the model on each dataset, and construct a confidence interval based on the distribution of parameter estimates
- If the bootstrap confidence interval is significantly wider than that produced by GLM, it is a sign that results are excessively-influenced by a few records

Confidence Intervals for Lift Measures

- We can use bootstrapping to create confidence intervals around lift measures, like Gini indices
- In measuring lift, we seek to answer the question: Does Model A outperform Model B?
- If the answer is yes, then the second question is: How significant is the win?
- Say Model A has a Gini index of 15.90 and Model B has a Gini index of 15.40
  - Model A has a Gini index that is 0.50 higher, but is that difference significant?
- We can also bootstrap quantile plots and double lift charts

Which Model Won?

	Model A	Model B
Classification	Won 1 metric	Won 4 metrics
Regression	Won 3 metrics	won 2 metrics

What are they?

For each problem,

- Model A is a GLM
- Model B is a GBM

Other considerations

- All these metrics are statistically based, when these models will have a business application.
- When possible, using metrics that reflect the application can be useful.
- For example, if a model is going to be used for rating purposes, in which real improvement in loss ratio is desired, then the decision as to the value associated with each model could be determined by simulating the impact on loss ratio.

Other considerations

This can be done by building a scenario testing environment which:

- Contemplates the characteristics of the corresponding business, at a granular level
  - E.g. quotes and existing business
- Includes material business features
  - For pricing purposes, these will be loss cost models, and either demand models or demand assumptions
  - Potentially capping algorithms for renewing business

This will give a view of the real difference in business value that may be realized using different models.

Considerations on Context and Usefulness

Up until now we have focused on statistical concepts, and ignored questions related to context, which can be really important, and might even change your decision on "which model is the winner".

Considerations on Context and Usefulness

For example, Model A might be better than Model B because:

- Model A includes predictors that can't be used when the model will be applied
  - This can be as simple as including some variable that isn't known at point of sale
- Model A cannot be programed to be used at the point at which the model will be applied
  - This can be as simple as including some interaction that can't be (easily) programmed into the existing rating engine
- Model A uses a methodology that results in a model that is difficult to explain
  - If a model needs to be filed, a certain amount of transparency is important

Any of these situations might mean you end up choosing Model B when it comes to deployment.