Modeling in the Models Application

Table of Contents

Overview 1
Using the Models Application 1
Evaluating Model Performance 2
Model Score Card 2
Variable's Percent Contribution Chart 3
Cumulative Lift and Gains Charts 5
Key Points for Lift and Cumulative Gains Charts 5
Lift Chart 5
Cumulative Gains Chart 6
Gains Tables 8

Overview

Using the Models Application

The Models application gives you the ability to quickly and easily work with predictive models in your analytic sandbox environment with little more than a few clicks of the mouse. In addition to the extensive Acxiom Audience Propensities catalog, available in the Audience Portrait application for evaluation and purchase, you can create look-alike models in this application.

Look-alike models are used to find larger audiences of consumers possessing characteristics similar to your best customers, or perhaps the characteristics of any segment in which you have an interest. Upload a file, create an insight file from it, step into the modeling center, and choose build a look-alike model. Your model is automatically created and you're provided with *Lift Charts and Cumulative Gains Table* reports to help you evaluate its performance.

The following is an overview of the techniques you need to properly evaluate the models you build, and an introduction to the reports Models application provides. Check it out, and then jump into the application to put the techniques into practice.

Evaluating Model Performance

Model Scorecard

The Model Scorecard helps you understand the relative importance of each variable specified in the model. The modeling build process uses an ensemble technique, so for each dependent variable, ten models are built from ten random samples. Ensemble modeling* refers to techniques where the predictors from a group of models are combined to produce a more accurate composite prediction.

* Ensemble Methods in Data Mining Improving Accuracy Through Combining Predictions, Giovanni Seni, John Elder, Synthesis Lecture on Data Mining and Knowledge Discovery, Morgan & Claypool Publishers, 2010.) Logistic regression is the fundamental statistical method used across the ensemble of models. The methodology is a binary logistic regression, performed using a forward stepwise variable selection procedure.

There are four elements in the Scorecard:

1. Variable ID/Name. This is the name of a specific variable that is used in the predictive model.
2. Count. This is the number of time each variable has been statistically included in one or more of the ten ensembles.
3. Rank. Rank is the stepwise order of importance in which the variable comes into a model, with 1 indicating most important: this is the average rank of the variable across one or more of the ten ensembles.
4. Percent Contribution. Percent Contribution measures the relative contribution each variable makes to one or more of the ten ensemble models. It is the result of first calculating the Absolute Contribution of each variable in each model as ABS (Standardized Estimate/Sum of ABS Standardized Estimates). Then, calculating the Total Absolute Contribution of each variable for all models as the sum of the Absolute Contributions for that variable divided by the number of models: these percent contribution sum to 100%.

The following is a Model Scorecard sample.

Models

Overview > Models > Scorecards

Model Scorecard - Demo Model

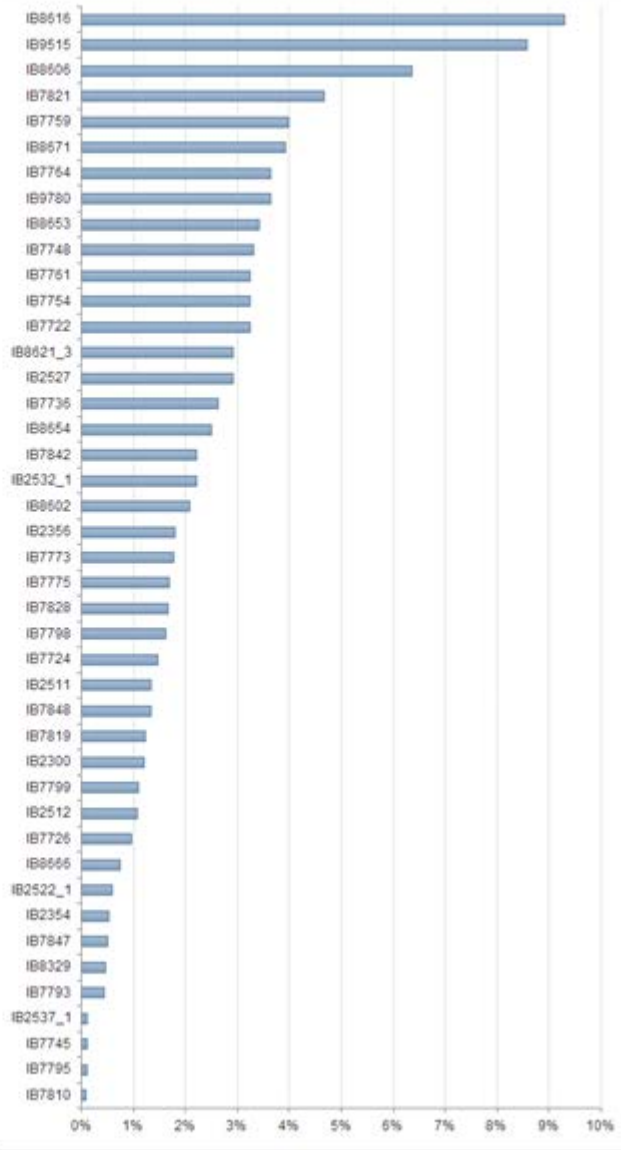
VARIABLE	DESCRIPTION	COUNT	RANK	% CONTRIBUTION
IB8616	Age in Two-Year Increments - 1st Individual	10	1.00	9.30%
IB9515	Gender - 1st Individual	10	2.00	6.57%
IB8606	Home Owner / Renter	10	3.00	6.36%
IB7821	Sweepstakes / Contests	10	4.20	4.68%
IB7759	Games - Board Games / Puzzles	10	6.10	3.99%
IB8671	Income - Estimated Household - Narrow Ranges	10	6.30	3.92%
IB7764	Movie Collector	10	7.70	3.65%
IB9780	eMail Append Available Indicator	10	8.00	3.64%
IB8653	Online Purchasing Indicator	10	10.10	3.43%

Model Gains - Demo Model

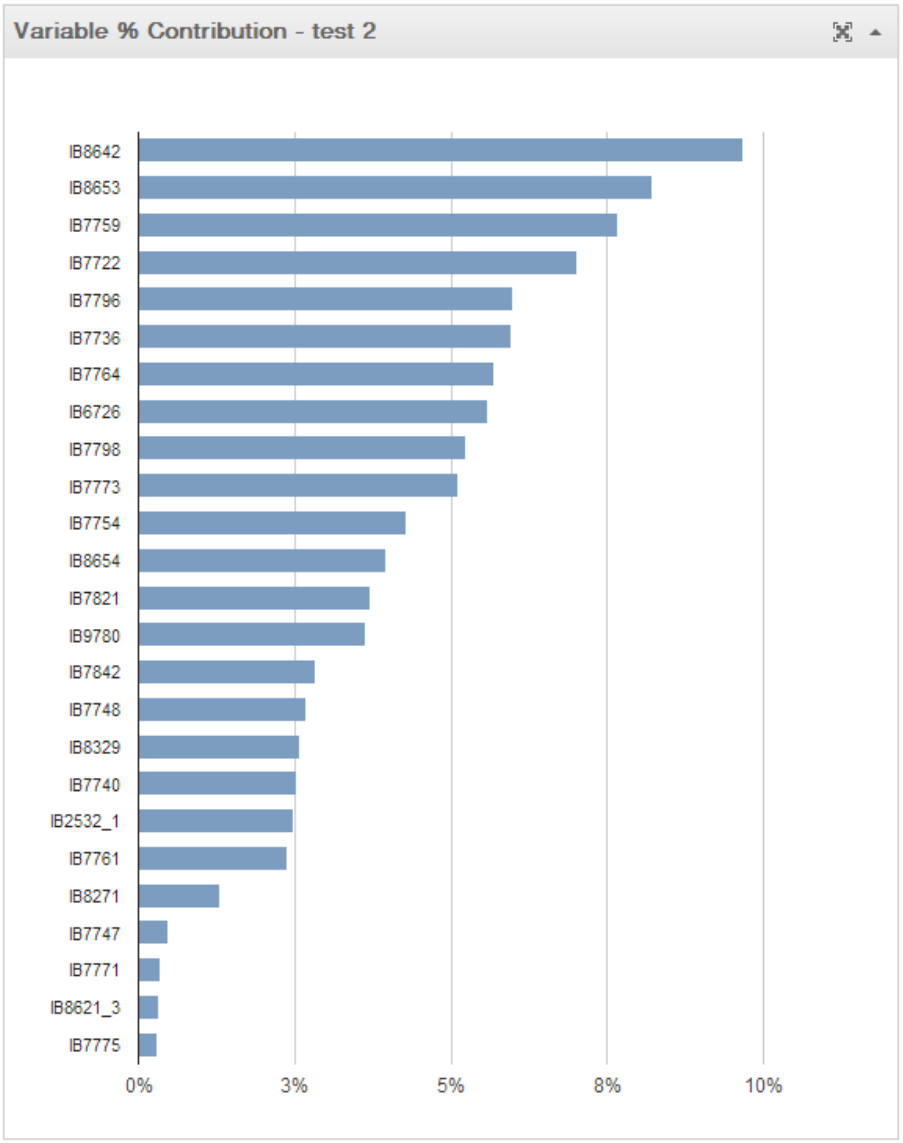
Model Validation Gains - Demo Model

Est. InfoBase Household Counts - Demo Model

Variable % Contribution - Demo Model



The following is a Variable's Percent Contribution chart sample.



Cumulative Lift and Gains Charts

Key Points for Lift and Cumulative Gains Charts

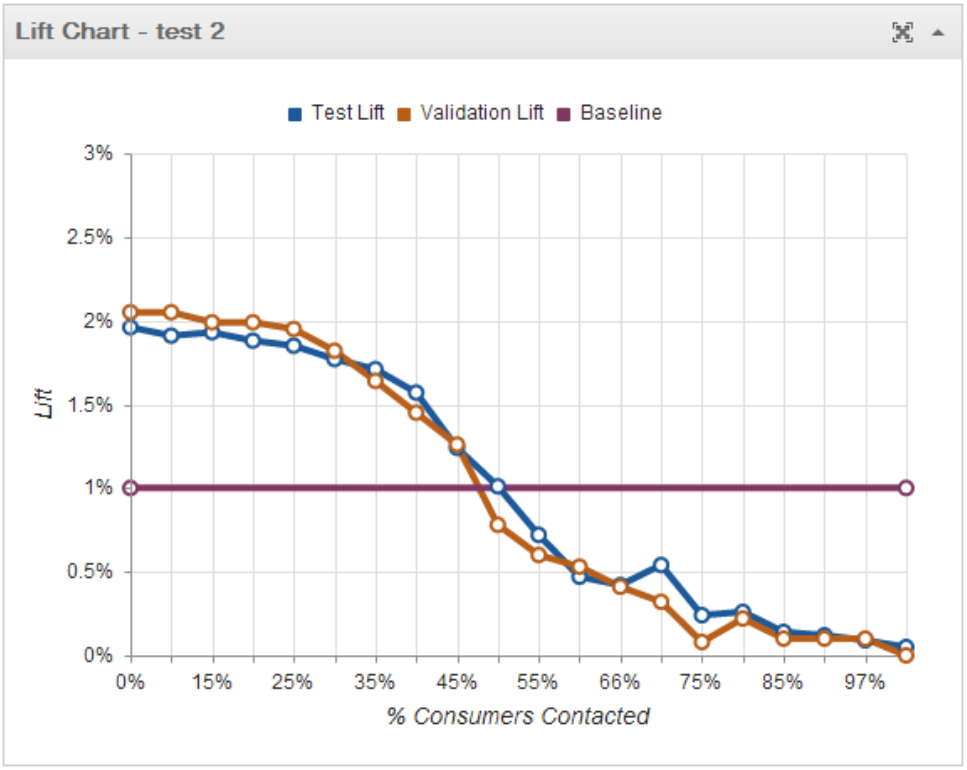
- Lift is a measure of a predictive model's effectiveness and is calculated as the ratio between the results obtained with and without a predictive model.
- Both the cumulative gains and lift charts are visual aids for measuring a model's performance.
- Both charts contain a lift curve and a baseline.
- The greater the area between the lift curve and the baseline, the better the model.

Lift Chart

A **Lift Chart** shows the actual lift produced from using a model. The points on the lift curve are determined by calculating the ratio between the result predicted by the model and the result using no model. The lift chart shows *how much more likely* you are to receive positive responses than if you just contact a random sample of customers.

As with any model development practice, Acxiom establishes a model build sample(s) and a holdout sample for immediate validation of the models using standard best practices: two almost identical lift curves indicate the model is well validated.

The following is an example of a Lift Chart.



Cumulative Gains Chart

The **Cumulative Gains Chart** helps you visually determine how effective you can be by selecting a relatively small number of consumers while getting a relatively large portion of the responders.

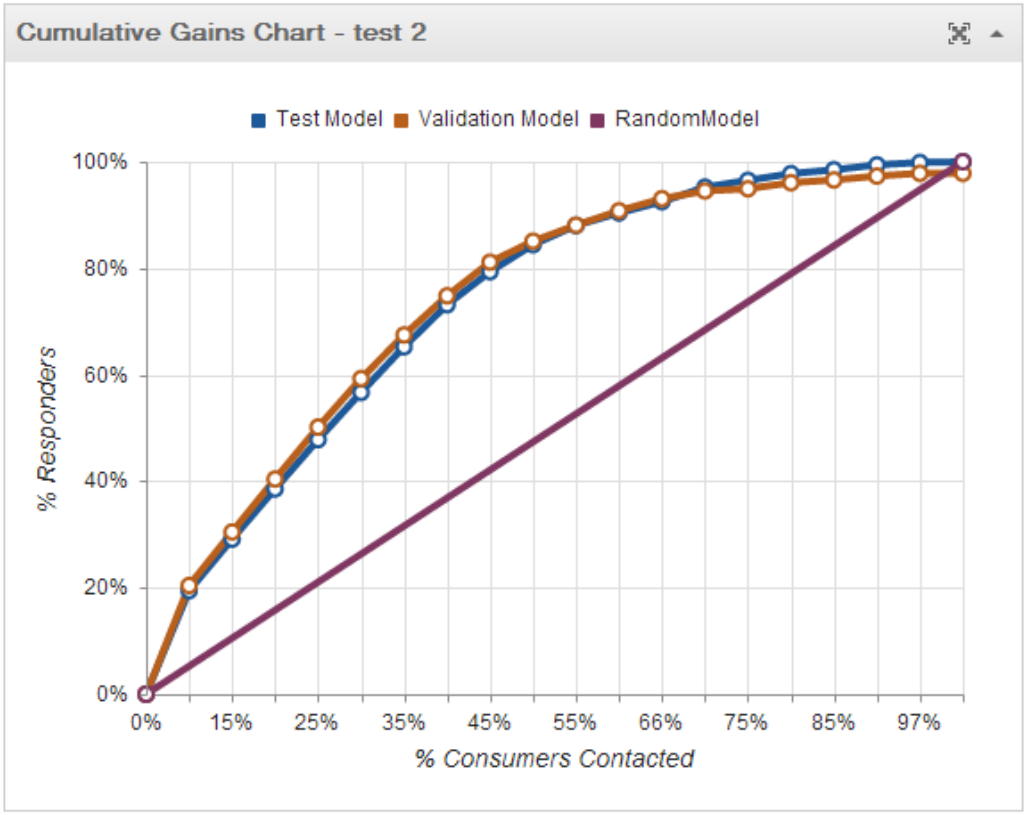
- The Y-axis shows the percentage of positive responses. This is a percentage of the total possible positive responses.
- The X-axis shows the percentage of consumers contacted, which is a fraction of all consumers.
- The Baseline or Random line in the Gains Chart represents the average or overall response rate - contacting **X**% of the customers will produce **X**% of the total positive responses.
- The Test Model lift curve represents the likely percent of responders for each percent of consumers contacted – contacting **X**% of customers will produce **Y**% of the total positive responses.
- The Validation Model lift curve represents the likely percent of responders for each percent of consumers contacted – contacting **X**% of customers will produce **Y**% of the total positive responses.

The greater the area between the two lift curves and the Random Model (baseline), the more the model is able to ***concentrate likely responses*** in the top deciles. The chart

shows how much more likely it is to receive responses from consumers selected from a predictive model than if you contact a random sample of consumers.

As with any model development practice, Acxiom establishes a model build sample(s) and a holdout sample for immediate validation of the models using standard best practices: two almost identical cumulative gains curves indicate the model is well validated.

The following is a sample of the Cumulative Gains Chart.



Gains Tables

The Gains Table and Validation Gains Table are another way of representing how well your model is likely to perform. Gains tables are typically organized in deciles (10 breaks), demi-deciles (20 breaks) or centiles (100 breaks). These tables contain important information about the target group in which you have interest and a comparison reference group.

These particular Gains Tables are composed of 14 columns and 20 rows (demi-deciles), with the following descriptions:

Rank

The gains table is organized into equal number of **breaks** (e.g., demi-eciles or 20 breaks).

Lift

Lift Index is the (Target % within a break, divided by the Total Target Rate)*100: the total target rate is calculated as the Total Cum Target/Total Cum Total (the counts found in the last break of each of the two columns respectively). The Lift Index/100 by break is also plotted along the Y-axis of the Lift Chart.

Total

Total number of Target records + Reference records by breaks used in the modeling sample.

Cume Total

Total number of records in the first break + the total number of records in the next break.

Cume Total %

Total % of records in the first break + the total % in the next break which total to 100% across all breaks. The **Cum Total%** creates the X-axis of the Lift Chart and the X-axis for the Cumulative Gains Chart.

Target

Total number of customers by break in your file.

Target %

Also known as Target Rate, is the total number of your customers in a break / by the total number of records in a break.

Cume Target

Total number of customers in the first beak + the total number of customers in the next break.

Cume Ref

Total number of reference file records in the first break + the total number of reference consumers in the next break.

Cume Target%

The total target% for the first break+ the total target% for the next break which total to 100% across all breaks. The Cum Target% is also plotted up the Y-axis for the Cumulative Gains Chart.

Cume Ref %

The total % for the first break+ the reference % for the next break which total to 100% when cumulated across all breaks.

Lower Bound

The lowest model score for a particular break.

Upper Bound

The highest model score for a particular break.

KS

Stands for "Kolmogorov-Smirnov," a statistic that is a standard measure used in evaluating a model's ability to discriminate between the customer and reference file. In practice the range is generally from about 20 to 70: below 20 indicates questionable discrimination, above 70 is probably too good to be true.

An example of calculating the **target rate** for the first break (e.g.,demi-decile) is the target count divided by combining the target count and reference count: 98.6% is 100* (2364/2397).

The **lift index** for the first break is the target rate divided by the total target rate (penetration is 50.07%=100*(24007/47938): the 196 lift index for demi-decile one is 100* (98.6/50.07).

The Gains Tables are expected to show breaks (e.g., deciles) ranked in descending order of the lift index. The spread between the top decile and bottom decile indicates the degree to which your model is discriminating between the target and reference groups. A lift index around 200 at the top and below 50 at the bottom generally indicates a good discriminating model.

The Gains Tables (like the Validation Gains chart below) also show for each break, how much more the model is able to concentrate likely responses in the top deciles: in this case, contacting 10% of consumers is likely to produce 20% of your total responses. Contacting 20% of consumers is likely to produce close to 39% of your total response.

The following are samples of the Gains Table and the Validation Gains Table.

Gains Table

RANK	LIFT	TOTAL	CUME TOTAL	CUME TOTAL %	TARGET	TARGET %	CUME TARGET	CUME REF	CUME TARGET %	CUME REF %	LOWER BOUND	UPPER BO
1	205	99	99	5.00%	98	98.99%	98	1	10.16%	0.10%	0.98511120	1.00000
6	205	100	199	10.00%	99	99.00%	197	2	20.41%	0.19%	0.97361269	0.98511
11	199	101	300	15.00%	97	96.04%	294	6	30.47%	0.58%	0.95803431	0.97361
16	199	100	400	20.00%	96	96.00%	390	10	40.41%	0.96%	0.93668783	0.95803
21	195	100	500	25.00%	94	94.00%	484	16	50.16%	1.54%	0.90693082	0.93668
26	182	100	600	30.00%	88	88.00%	572	28	59.27%	2.69%	0.86191421	0.90693
31	164	100	700	35.00%	79	79.00%	651	49	67.46%	4.72%	0.80040274	0.86191
36	145	101	801	40.00%	71	70.30%	722	79	74.82%	7.60%	0.71204291	0.80040
41	126	100	901	45.00%	61	61.00%	783	118	81.14%	11.36%	0.59668168	0.71204
46	78	100	1,001	50.00%	38	38.00%	821	180	85.08%	17.32%	0.46682501	0.59668
51	60	100	1,101	55.00%	29	29.00%	850	251	88.08%	24.16%	0.35227023	0.46682

Validation Gains Table

RANK	LIFT	TOTAL	CUME TOTAL	CUME TOTAL %	TARGET	TARGET %	CUME TARGET	CUME REF	CUME TARGET %	CUME REF %	LOWER BOU
1	205	99	99	5.00%	98	98.99%	98	1	10.16%	0.10%	0.9
6	205	100	199	10.00%	99	99.00%	197	2	20.41%	0.19%	0.9
11	199	101	300	15.00%	97	96.04%	294	6	30.47%	0.58%	0.9
16	199	100	400	20.00%	96	96.00%	390	10	40.41%	0.96%	0.9
21	195	100	500	25.00%	94	94.00%	484	16	50.16%	1.54%	0.8
26	182	100	600	30.00%	88	88.00%	572	28	59.27%	2.69%	0.8
31	164	100	700	35.00%	79	79.00%	651	49	67.46%	4.72%	0.7
36	145	101	801	40.00%	71	70.30%	722	79	74.82%	7.60%	0.6
41	126	100	901	45.00%	61	61.00%	783	118	81.14%	11.36%	0.5
46	78	100	1,001	50.00%	38	38.00%	821	180	85.08%	17.32%	0.3
51	60	100	1,101	55.00%	29	29.00%	850	251	88.08%	24.16%	0.2

In summary, this collection of charts and tables provide you the ability to properly evaluate the performance of the models you build using the Models application.