PreMo
Predictive Modeling, *LLC.*

CASE STUDY
THE VALUE OF ACCURACY IN THE REGRESSION LIFT MODEL

Most users of predictive models are happy to have the benefit of a good model with which to target their marketing initiatives and do not ask the equally important question, is this the best model we can be using? This study will demonstrate that a very small improvement in the accuracy of the model can result in very large financial gains.

The accuracy of a model is controlled by three major variables: 1). First and foremost the ability of your data to be predictive. There is an unknown and fixed limit to which any data can be predictive regardless of the tools used or experience of the modeler. 2.) The experience and skill of the modeler. 3.) The tools selected. Some tools are designed to give very quick if somewhat approximate results. Other tools are inherently more accurate if somewhat slower.

Models can frequently be improved through better selection or preparation of the data including the addition of appended data. However, even when the data is exactly the same, the selection of the modeling tool can be critical.

Many modelers tend to utilize only one tool in creating their models, frequently the one they are most comfortable with or were initially trained on, logistic regression, neural nets, decision trees, Bayesian classifiers, support vector machines, or genetic programs. Not all tools create equally accurate answers when applied to the same data sets.

How important is accuracy? This case study illustrates that a change in fitness of only 0.01 point can mean a financial improvement of nearly 8%. Bigger increases in model quality will translate into higher percentages of financial improvement. The benefit each user actually receives will depend on how much the model can be improved and the financial details of the offer, but this example should make one thing clear, small increases in model quality can translate to large increases in financial performance.

Example:

This example is based on actual data from a major technology and services company pursuing cross sell or up sell opportunities with their existing customers. It would be equally true of initiatives aimed at new customer acquisition or customer retention (churn/defection prevention) campaigns, or to any of the other major uses of scoring (regression) models such as fraud detection, credit scoring, or billing review.

The data is from a large direct mail test where the overall response rate was found to be 1%, very typical for this type of campaign. In our example we assume a full mailing to all available targets would be 250,000 pieces at a cost of $3.00 per mailing, and with a gross profit of $300 per successful sale.

This means that a mailing to all 250,000 targets would require an investment of $750,000 and would return $750,000. Most business managers would regard this as a bad investment and would elect not to conduct the full mailing, counting the cost of the test mailing as the sunk cost of an unsuccessful promotion.

To illustrate the importance that small improvements in accuracy can make, we developed two models, one with a fitness measure of .195064 and the other with a fitness measure of .182995, only .012069 between them. The fitness measure is the remaining unexplained difference between the actual data and the model. Lower scores are better. A fitness measure of 0.00 means the model completely explains and predicts the actual data so both these models show good and useful predictive ability, explaining more than 80% of the difference between the actual and the model.

In the detailed table below, the business manager evaluates the less accurate of the two models and finds that his mailing can yield a good profit, $163,043 if he only mails to the top 50% of the list. The model has scored all prospects from 0 to 1 based on their likelihood to buy, and after evaluating the net profit (projected profit from sales less the cost of mailing) for each decile of the list (a decile equals 10% of the list, a very common division for this analysis) sees that the bottom half of the list is a money-losing proposition.

Worse Model .195064

Decile	predicted % buyers by decile	cum lift		expected buyers in a mailing of	gross profit at	cost of mailing	Net profit from decile	Net profit from mailing above breakeven
			pieces>	250,000		250,000		
			rate >	1.0%	$300	3.00		
1st	16.03%	16.03%		401	$120,245	75,000	$45,245	$45,245
2nd	14.95%	30.98%		374	$112,092	75,000	$37,092	$37,092
3rd	13.32%	44.29%		333	$99,864	75,000	$24,864	$24,864
4th	14.95%	59.24%		374	$112,092	75,000	$37,092	$37,092
5th	12.50%	71.74%		313	$93,750	75,000	$18,750	$18,750
6th	9.51%	81.25%		238	$71,332	75,000	-$3,668	
7th	6.79%	88.04%		170	$50,951	75,000	-$24,049	
8th	4.62%	92.66%		115	$34,647	75,000	-$40,353	
9th	4.35%	97.01%		109	$32,609	75,000	-$42,391	
10th	2.99%	100.00%		75	$22,418	75,000	-$52,582	
	100.00%			2,500	$750,000	$750,000	$0	$163,043

However, if the manager had the benefit of the better model (table 2), and better by only 0.012 points in fitness, he now forecasts a profit of $175,679, an improvement of 7.75%.

Better Model .182995

Decile	predicted % buyers by decile	cum lift		expected buyers in a mailing of	gross profit at	cost of mailing	Net profit from decile	Net profit from mailing above breakeven
			pieces>	250,000		250,000		
			rate >	1.0%	$300	3.00		
1st	16.58%	16.58%		414	$124,321	75,000	$49,321	$49,321
2nd	15.49%	32.07%		387	$116,168	75,000	$41,168	$41,168
3rd	13.86%	45.92%		346	$103,940	75,000	$28,940	$28,940
4th	12.77%	58.70%		319	$95,788	75,000	$20,788	$20,788
5th	13.32%	72.01%		333	$99,864	75,000	$24,864	$24,864
6th	11.41%	83.42%		285	$85,598	75,000	$10,598	$10,598
7th	6.52%	89.95%		163	$48,913	75,000	-$26,087	
8th	4.89%	94.84%		122	$36,685	75,000	-$38,315	
9th	3.80%	98.64%		95	$28,533	75,000	-$46,467	
10th	1.36%	100.00%		34	$10,190	75,000	-$64,810	
	100.00%			2,500	$750,000	$750,000	$0	$175,679

Small improvements in model accuracy can make big improvements in financial outcome. Be sure to ask the question: Is this really the most accurate model that can be created from my data?

PROCEDURES USED IN PREPARING THIS EXAMPLE:

This example is based on a data set used in a data mining competition and is extracted from a much larger data set of actual repeat buyers from a major Canadian technology company. In the test mailing buyers and non-buyers are known and coded 0 or 1. The overall response rate of the test mailing was about 1%. All 1079 responders were used, together with 1079 randomly-chosen non-responders, for a total of 2158 cases.

There are 200 explanatory variables in the file including prior product purchases, recency frequency, and size of purchase, and demographic data gathered by the company, demographics appended from census data, and demographics appended from "tax filer" data.

The 2158 cases were first divided into three randomized sets of approximately 720 each for the validation and training sets used to develop the model and one additional set held aside as 'unseen' data. The true test of a model is its ability to score approximately the same level of accuracy on data never seen during the development of the model and the 'unseen' data is used for this validation step. In this analysis the two models were evaluated based on the fitness of the best model as evaluated on the unseen data.

A number of modeling runs were conducted from subsets of the variables to determine which had predictive capability. For the final run, 55 of the 200 variables were selected. After the final model was developed 40 variables were determined to have predictive value, of which 27 had significant predictive value.

The first model was allowed to run until a fitness of .195064 on the unseen data had been achieved. For comparison, the model was allowed to continue to run until a fitness of .182995 had been achieved, a difference of 0.012069, an improvement of 6.2% over the first model.

The lift models were then constructed and compared under the realistic but hypothetical values for the overall promotion as described above to determine the dollar value and percentage of the improvement from the difference in the models.

PreMo 1514 Falling Star, Westlake Village, CA 91362
818.257.2035