Analysis *across the* BORDER

Can predictive models be applied across the U.S. and Canada? Here are some answers.

by Richard Boire

With the growth in predictive modelling, discussions have arisen concerning whether or not predictive models can be applied across different countries. The most common example of this is the application of U.S. models to consumers of Canadian subsidiaries.

As any statistician will tell you, the development of a Canadian model for the Canadian marketplace should always be superior to the application of a U.S. model within the Canadian marketplace. However, the decision to build Canadian-specific models should be an economic one by comparing the incremental revenues versus the incremental costs.

The incremental revenues are determined from the additional lift in performance while the incremental costs are due to increased resources such as manpower, software, and hardware. A case example later will better illustrate this concept.

Before we look at this example, it is important for the analyst to have a broad understanding of both the data environments and demographics in both countries. For instance, within the data environment, one major assumption is that the modelling fields or variables are available in both Canada and the U.S.

In fact, this is the usual case as Canadian subsidiaries simply represent extra records on the U.S. centralized database. If the databases are different, then the analyst needs to determine how many common variables can be constructed in both markets.

Common Variables

With a common set of variables in both markets, the analyst can compare the performance of U.S. models versus Canadian models. Once these data environment issues are addressed, some basic understanding of the demographic differences in both countries is required.

In many U.S. models, specific regional variables such as northwest, northeast, southwest, and southeast are created. In some cases, more specific geographic variables are fashioned at the state level.

Within Canada, it would be impractical to create the same kind of regional variables. One needs only to look at the population distribution of the two countries to understand this rationale.

The population in the U.S. is clustered into the northeast, southeast, and southwest with some milder concentrations in the northwest and midwest. In Canada, virtually 90 percent of the population lives within 200 miles of the U.S. border.

Linguistic Differences

Besides these population density differences, the linguistic and racial characteristics of the U.S. are also very different. In the U.S., several minority languages prevail with some languages having more prominence than others, in particular Spanish.

Within Canada, although French is a minority language, its huge significance has caused it to be recognized alongside English as one of Canada's two official languages. Another distinguishing feature is that 90 percent of the French-speaking population within Canada is clustered in Quebec, which is certainly unlike any minority group within the U.S.

The above differences represent the visible ones, yet there are the less visible differences such as those relat4ed to culture, attitudes and psychographics. Much could be written about what differentiates Canadians and Americans in these areas.

It is not my intent to delve into these differences since what we really want to understand is how culture, attitude, and psychographics differences alongside language and population density differences impact customer database behaviour differently in both countries. Once these database behaviour differences are identified, we can then determine whether it makes economic sense to build region-specific models.

The quantification of whether or not to build region-specific models (i.e. Canadian models) is best demonstrated through the example of a predictive model. Listed below are the respective models, their variables, their weights or coefficients, and their importance within the overall equation.

The U.S. Model Applied to Canadian Consumers – R^2 = .035

Variable	Weight of Variable	Importance of Variable in Equation
% change in spending between last 6 months and previous 6 months	+.035	1
Tenure	+.005	2

The Canadian Model Applied to Canadian Consumers – R^2 = .045

Variable	Weight of Variable	Importance of Variable in Equation
Tenure	-.0065	1
Live in Quebec	-.040	2
% change in spending between last 6 months and previous 6 months	+.0245	3

From the two models below, distinct differences exist. The regional variable is a significant characteristic within the Canadian model.

The percent change in spending, although exhibiting the same sign in both models, is strongest for the U.S. model and weakest in the Canadian model. The tenure variable also exhibits the opposite impact on response in both models.

Performance Levels

But the most important difference is the level of performance which can be seen by comparing he R^2 values (the statistical measure of the model's predictive capability) of both the U.S. and Canadian models when applied to Canadian consumers.

The Canadian model would seem to be better performing due to its higher R^2 (.045-Canada vs. .035-U.S.).

The accompanying gains chart further demonstrates the different performance levels of both models. In the accompanying chart, we report the lift within each interval which represents the increase in response rate at a given cutoff as a result of using the model. A lift of 100 percent represents the average response rate with no modelling.

As you can see from the gains chart results, the Canadian model outperforms the U.S. model from a lift perspective. The additional lift in performance by the Canadian model can be quantified in terms of the additional saved mailing costs.

Mailing Costs

Saved mailing costs represent the additional unmodelled names which would have to be mailed to achieve the same number of responders that are provided for by modeled names. Listed on the gains chart are the saved mailing costs for each model as well as the incremental saved mailing costs by using the Canadian model instead of the U.S. model for Canadian consumers. Note the results in the gains charts assume an available mailing universe of 500,000 names, an average response rate of one percent with no modelling and a cost per mailing of $1.00.

From the chart, we can see that the cost of not developing a Canadian specific model ranges anywhere from $10,000 to $32,500 (incremental saved mailing costs) depending on the cutoff. These opportunity costs are further magnified if the universe increases or if this type of campaign is repromoted.

Application of Canadian and U.S. Models to Canadian Consumers (Gains Chart)

% of list (ranked by model score)	Number of names mailed	Cum. Resp. Rate (U.S.)	Cum. Resp. Rate within each interval (U.S.)	Cum. Resp. Rate (Canada)	Cum. Resp. Rate within each interval (Canada)	Saved Mailing Costs (Canada)	Saved Mailing Costs (U.S.)	Incremental Saved Mailing Costs (Canada)
0-10%	50,000	1.80%	180	2.00%	200	$50,000	$40,000	$10,000
10-20%	100,000	1.65%	165	1.90%	190	$90,000	$65,000	$25,000
20-30%	150,000	1.55%	155	1.70%	170	$105,000	$82,500	$22,500
30-40%	200,000	1.45%	145	1.60%	160	$120,000	$90,000	$30,000
40-50%	250,000	1.35%	135	1.50%	150	$125,000	$87,500	$37,500
50-60%	300,000	1.28%	128	1.30%	130	$105,000	$84,000	$21,000
-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-
90-100%	500,000	1.00%	100	1.00%	100	$0	$0	$0

This kind of analysis reveals the inadequacies of blindly applying U.S. models to the Canadian market. Yet, another example could indicate the development of Canadian-specific models might be considered a waste of resources due to its minimal benefit.

Certainly, U.S. multi-national type companies should undergo this type of analysis before deciding whether to build specific models. If analysis cannot be conducted due to time and resources, a good rule of thumb is that Canadian subsidiaries with more than 250,000 customers need their own specific model.