Антипов Евгений Александрович, ГУ-ВШЭ (Санкт-Петербург)
Ведущий аналитик Исследовательской компании «КОМКОН – Санкт-Петербург»

Совершенствование моделей предсказания ухода клиентов в случае неодинакового влияния факторов на поведение разных сегментов клиентской базы

Введение

Скоринговые системы, уже получившие достаточно широкое распространение в оценке рисков кредитными учреждениями и страховыми компаниями, также начинают получать распространение в маркетинге, где используются для оценки вероятности отклика (покупки) со стороны клиента или отказа от услуг компании клиента с определенными характеристиками.

В случае маркетинговых систем они позволяют направлять усилия только на целевые сегменты, избегая лишних расходов на коммуникации с менее важными сегментами потребителей (клиентов). Кризисное время должно начать активнее использовать такого рода подходы в бизнесе.

Как правило, при решении задачи классификации создается единая для всей рассматриваемой выборки модель. Между тем, такая модель может в одних сегментах наблюдений давать предсказание, существенно менее точное, чем для других. В результате точность модели ниже потенциально достижимой с использованием доступных объясняющих переменных. С точки зрения менеджмента и экономики это ведет к систематическим ошибкам в принятии решений. Например, в случае неодинакового в разных сегментах влияния факторов, обуславливающих уход клиентов, не достигается потенциально возможное качество предсказания (в результате, усилия по удержанию клиентов неоптимально распределены между клиентами).

В своем исследовании автор использует подход к диагностике логистической регрессии и повышению качества классификации, основанный на алгоритме построения деревьев классификации CHAID (Antipov, Pokryshevskaya, 2009). Подход позволяет решить ряд важных, но, по всей видимости, еще не решенных проблем:

1. Как автоматически выявить и наглядно описать сегменты, в которых исходная модель предсказывает существенно хуже, чем в среднем?

2. Как учесть информацию о неоднородности качества предсказания между сегментами, чтобы разбить наблюдения на несколько сегментов для достижения более точной классификации?

Тот факт, что диагностика получается наглядной, делает процедуру построения модели максимально прозрачной, что является необходимым условием для моделей для предсказания ухода клиентов. Кроме того, подход привлекателен и возможностью реализовать его с помощью наиболее часто используемых в российских фирмах статистических пакетов SPSS и Statistica без необходимости написания специальных программ.

Подход был применен к классическому набору данных о клиентах телекоммуникационной фирмы. Путем разбиения набора данных на 3 части (опираясь на дерево классификации) и построения отдельных логит-моделей для каждого сегмента, удалось повысить долю верно классифицированных наблюдений более чем на 5 процентных пунктов на обучающей и тестовой выборках.

Использованные данные

Для иллюстрации предложенного подхода был использован набор данных о клиентах телекоммуникационной фирмы. Раннее обнаружение клиентов, которые потенциально прекратят пользовать услугами фирмы, помогает компаниям таргетировать клиентов с помощью специальных мероприятий и, соответственно, сохранять прибыль. По биллинговой базе (базе продолжительности звонков и затрат на услуги) требуется предсказать, уйдет абонент или нет в ближайшее время.

Зависимая переменная – это то, ушел клиент или нет. Объясняющие переменные включают длительность звонков в разное время суток, количество SMS-сообщений, расходы на различные виды услуг.

Перед построение логистической регрессии, мы случайным образом делим выборку на обучающую (2000 наблюдений) и тестовую (1333 клиентов).

Эмпирическая оценка

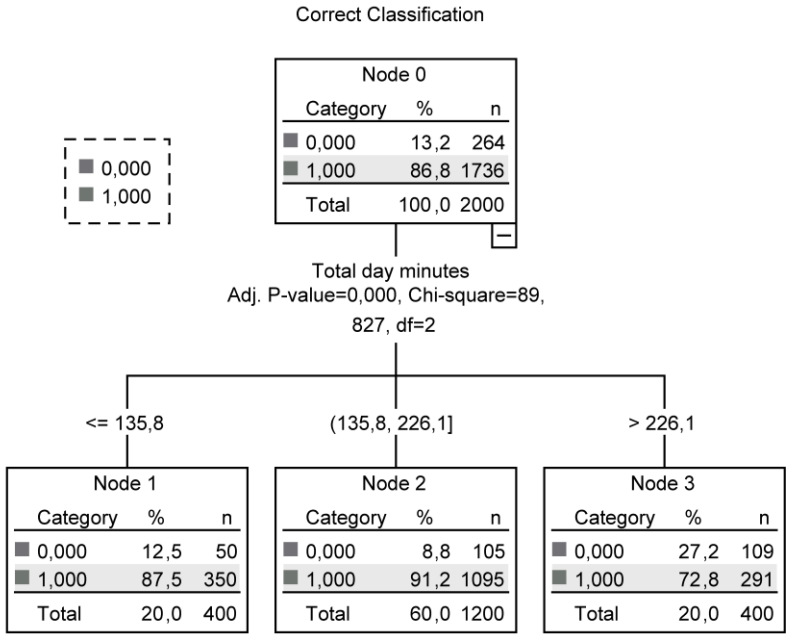
В Таблице 1 представлены коэффициенты Модели 1 (общей для всей выборки). Поскольку мы обсуждаем скорее методологический вопрос, мы не будем пытаться объяснить возможные причины получаемых взаимосвязей между поведением абонента и вероятности ухода для конкретной фирмы, предоставившей данные.

Таблица 1. Оценки параметров для Модели 1

Variable	B	Wald	Sig.	Exp(B)
Intercept	-8.13	163.35	0.00	
VMailMessage	-0.02	15.83	0.00	0.98
DayMins	0.01	101.44	0.00	1.01
EveMins	0.01	20.36	0.00	1.01
NightMins	0.00	7.04	0.01	1.00
IntlCalls	-0.07	5.70	0.02	0.93
IntlCharge	0.47	24.33	0.00	1.60
CustServCalls	0.42	66.65	0.00	1.52

Далее создается переменная C (индикатор верной классификации). После этого мы строим диагностическое дерево, беря C в качестве зависимой переменной, а все предикторы, использованные в построении модели – в качестве объясняющих.

Рис. 1. Дерево классификации CHAID decision tree: точность Модели 1 в разных сегментах



Диагностическое дерево классификации позволило автоматически выявить 3 сегмента, в которых доля верно классифицированных наблюдений сильно различается: от 72,8% в группе наиболее активно разговаривающих днем, до 91,2% в группе средней длительности разговоров в дневное время.

На основании проведенной диагностики построили отдельную модель (Модель 2) для каждого сегмента (использовали пошаговый метод отбора переменных). Отобранные переменные различны для каждого из трех сегментов.

Таблица 2. Оценки параметров Модели 2

Segment	Variable	B	Std. Error	Wald	Sig.	Exp(B)
Total day minutes<=135.8	**Intercept**	-3.45	0.51	45.19	0.00	
	IntlCalls	-0.24	0.09	6.38	0.01	0.79
	CustServCalls	1.17	0.16	56.09	0.00	3.22
135.8<Total day minutes<=226.1	**Intercept**	-5.09	0.50	103.09	0.00	
	CustServCalls	0.45	0.08	35.38	0.00	1.57
	IntlCharge	0.64	0.15	18.40	0.00	1.90
Total day minutes>226.1	**Intercept**	-34.09	3.86	78.06	0.00	
	VMailMessage	-0.14	0.02	44.83	0.00	0.87
	DayMins	0.07	0.01	57.99	0.00	1.08
	EveMins	0.04	0.01	63.21	0.00	1.04
	NightMins	0.02	0.00	25.31	0.00	1.02
	IntlMins	0.26	0.07	15.38	0.00	1.30

Модели 1 и 2 сравнили (Таблица 3) по двум аспектам: однородность качества предсказания между различными сегментами (Индекс Джини и энтропия) и точность

классификации (доля верно классифицированных наблюдений, доля верно определенных как «уход» от общего числа классифицированных как «уход», доля верно определенных как «уход» от общего числа тех, кто в действительности ушел).

Таблица 3. Сравнение моделей

	Модель 1	**Модель 2**	**Комментарий**
Индекс Джини (на обучающей выборке)	0,219	0,142	Снизилась неоднородность качества предсказания: различия качества классификации между сегментами стали существенно слабее
Энтропия (на обучающей выборке)	0,535	0,386	
Доля верно классифицированных наблюдений (на тестовой выборке)	85,3%	86,8%	Модель лучше улавливает
Доля верно определенных как «уход» от общего числа классифицированных как «уход» (на тестовой выборке)	55,6%	83,8%	В 83,8% случаев выявленные моделью 2 абоненты действительно уйдут в будущем. Модель 2 снизила долю средств, потраченных на тех, кто не собирался уходить.
Доля верно определенных как «уход» от общего числа тех, кто в действительности ушел (на тестовой выборке)	10%	46,5%	Модель 2 позволяет обнаружить 46,5% абонентов, которые уйдут в будущем: фирма сможет предотвратить уход почти половины абонентов с помощью грамотной программы лояльности

Заключение

Для решения многих задач, в связи с наличием гетерогенности в данных, полезно делать предсказания для отдельных сегментов, а не по всей выборке целиком. В работе предлагается подход для выявления неоднородности качества предсказания между сегментами наблюдений, основанный на применении алгоритма CHAID.

Подход был применен к классическому набору данных об оттоке клиентов телекоммуникационной фирмы (из UCI Repository of Machine Learning Databases). Путем разбиения набора данных на 3 части (на основании предварительно построенного диагностического дерева классификации) и построения отдельной логистической регрессии для каждого сегмента мы повысили точность общей модели (построенной по всей выборке в целом) более чем на 5 процентных пунктов на обучающей и тестовой выборке. С экономической точки зрения программа лояльности, охватывающая людей, выявленных такой моделью, скорее всего будет более эффективной.

Различные сегменты абонентов имеют в нашем примере совершенно разную важность объясняющих уход клиента переменных. Поэтому использованное разбиение позволяет не только повысить точность предсказания, но и получить более хорошее представление о движущих силах поведения клиентов. В дальнейшем планируется протестировать возможности применения других алгоритмов построения деревьев классификации для разбиения выборки на сегменты.

Литература

1. Antipov, Evgeny and Pokryshevskaya, Elena, Applying CHAID for Logistic Regression Diagnostics and Classification Accuracy Improvement (October 16, 2009). Available at SSRN: http://ssrn.com/abstract=1412208
2. Deodhar, M., Ghosh, J. (2007) *A framework for simultaneous co-clustering and learning from complex data*. Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining; 12-15 August 2007, San Jose, California, USA.
3. Hill, T. and Lewicki, P. (2007) STATISTICS Methods and Applications. StatSoft, Tulsa, OK.
4. Neslin, S., Gupta, S., Kamakura, W., Lu, J. and Mason, C. (2006) Detection defection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research* 43(2): 204–211.
5. Levin, N. and Zahavi, J. (1998) Continuous predictive modeling, a comparative analysis. *Journal of Interactive Marketing* 12: 5–22.
6. Kass, G.V. (1980) An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Journal of Applied Statistics* 29(2): 119-127.
7. Ripley, B.D. (1996) *Pattern recognition and neural networks*. Cambridge: Cambridge University Press.
8. Morgan, J.N. and Messenger, R.C. (1973) THAID: A sequential analysis program for the analysis of nominal scale dependent variables. Institute of Social Research, University of Michigan, Ann Arbor. Technical report.
9. Blake, C. L. and Merz, C. J., Churn Data Set, UCI Repository of Machine Learning Databases, http://www.ics.uci.edu/~mlearn/MLRepository.html. University of California, Department of Information and Computer Science, Irvine, CA, 1998.