Abgangsprognose in der Gesetzlichen Krankenversicherung und Variablenselektion

Lars Cyriacks-Honebein
Techniker Krankenkasse
Bramfelder Strasse 140
22305 Hamburg
lars.honebein@tk-online.de

Martin Zieger
Techniker Krankenkasse
Bramfelder Strasse 140
22305 Hamburg
martin.zieger@tk-online.de

Zusammenfassung

Mit Hilfe der Abgangsprognose wird für Krankenkassenmitglieder eine Kündigungswahrscheinlichkeit erstellt, welche u.a. als Controllingtool eingesetzt wird.
In einem ersten Schritt werden die potentiell interessanten Variablen gesammelt, so wie neue Variablen generiert. Bevor der SAS Enterprise Miner zum Einsatz kommt, werden alle Merkmale univariat gescored und analysiert. Der Scorewert liefert einen Anhaltspunkt dafür, welche Variablen im späteren Modell relevant werden könnten. Zusätzlich können die inhaltlichen Aussagen der einzelnen Variablen plausibilisiert werden.
Im Enterprise Miner hat man bei der Realisierung das Problem, dass die Kündigerquote für manche Fragestellungen unter einem Prozent liegt. Auf Grund dessen wird eine 50/50 Stichprobe generiert, welche es ermöglicht, dass die spezifischen Kündigermerkmale dieser kleinen Gruppe besser differenziert werden können. Die Konsequenz hieraus sind verzerrte Wahrscheinlichkeiten, welche durch Einsatz eines Prior-Vektors im Entscheidungsbaum wieder richtig gestellt werden können.

Schlüsselworte: Kündigerprognose, Variablenselektion, Prior-Vektor Data Mining, Enterprise Miner, Gesetzliche Krankenkasse

1 Einleitung

Im Jahr 1996 hat sich die Marktsituation für gesetzliche Krankenkassen grundlegend geändert. Mit Einführung der freien Kassenwahl gab es eine Menge neuer direkter Konkurrenten und als Konsequenz dessen auch einen erhöhten Bedarf an Haltearbeit unter den eigenen Mitgliedern. Diese Situation zieht natürlich sofort die Frage mit sich „Wer wird wahrscheinlich kündigen?“ Um diese Frage fundiert zu beantworten, hat die Techniker Krankenkasse ein Kündigerprognosemodell entwickelt. In diesem konkreten Fall

ist unter anderem auf ein paar Besonderheiten zu achten, welche die spezielle Situation einer Gesetzlichen Krankenkasse widerspiegeln.

Ein Beispiel für eine Besonderheit, zeigt sich in dem mit einer Kündigung verbundenen Zwang, sich bei einer anderen Krankenkasse zu versichern, da es eine gesetzliche Krankenversicherungspflicht gibt. Im Unterschied zu beispielsweise einem Zeitschriftenabonnement reicht also nicht allein der Wunsch zu kündigen, sondern es muss eine geeignete Alternative geben. Als Folge dessen, und auf Basis einer guten Datenqualität – unter Berücksichtigung der gesetzlichen Datenschutzbestimmungen - kann man die Kündiger nach ihrer neuen Krankenversicherung kategorisieren und sich auf spezielle Gruppen konzentrieren. So werden finanziell attraktive Mitgliedergruppen vorselektiert und spezielle Modelle für diese entwickelt.

Ein Vorteil, der sich u.a. indirekt aus der gesetzlichen Versicherungspflicht ableitet, ist eine sehr gute Datenlage, welche sich zum einen aus einer breiten Datenmasse und zum anderen auch in einer guten Datenqualität dieser Masse widerspiegelt. Gleichzeitig stellt dies natürlich auch die Gefahr dar, dass man sich in zu vielen Daten und Variablen verirrt und den Überblick für Zusammenhänge verliert oder die wahre Aussage von Variablen falsch einschätzt. Aus diesem Grund ist eine ausführliche Datenaufbereitung und -analyse, so wie anschließend eine teilweise Vorselektion der Inputvariablen empfehlenswert.

2 Variablenselektion

In einem ersten Schritt bevor der SAS Enterprise Miner (EM) eingesetzt wird, kümmert man sich zunächst um die später als Input benutzten Daten. Hierbei sind drei Punkte besonders hervorzuheben, welche Kernaufgaben darstellen:

- Daten sammeln
- Datenqualität prüfen
- Aussagekraft analysieren.

Auf den Punkt „Datenqualität prüfen“ soll hier nicht weiter eingegangen werden, in der Praxis sollte er aber natürlich auf keinen Fall ignoriert werden. Statt dessen wird im folgenden auf die Punkte „Daten sammeln“ und insbesondere „Aussagekraft analysieren“ näher eingegangen.

2.1 Daten sammeln

Vorausgesetzt man weiß aus welcher Quelle die Daten kommen sollen (z.B. Datawarehouse), untergliedert sich dieser Teil nochmals in zwei Teilaufgaben.
Zum einem müssen die Variablen festgelegt werden, welche man analysieren möchte bzw. welche später als erklärende in dem Prognosemodell auftauchen könnten. Zum anderen muss den genaue Umfang der Stichprobe festgelegt werden, also mit wie viel

Datensätzen später den EM gefüttert werden soll. Hier soll nun insbesondere der erste Teil erläutert werden, also die Untersuchung, welches potentiell interessante und aussagefähige Merkmale sind.

Die in Betracht kommenden Variablen für den EM können grob in drei Kategorien aufteilt werden:

1. Basisvariablen
2. Gruppierte Variablen
3. Abgeleitete Variablen

Unter Basisvariablen kann man alles subsumieren, was eins zu eins aus dem Datawarehouse übernommen werden kann. Hierzu gehören insbesondere Soziodemographische Ausprägungen wie Alter und Geschlecht, aber auch allgemeine krankenversicherungsspezifische Merkmale wie freiwillig oder pflichtig versichert, Anzahl Mitversicherter Familienangehörige, usw.

Die anderen beiden Kategorien haben gemeinsam, dass sie aus Basisvariablen konstruiert werden müssen, da sie in der gewünschten Form nicht vorgehalten werden. Das heißt, man muss sich zunächst überlegen, welche Informationen oder indirekten Merkmale noch von Interesse sein könnten. Im darauffolgenden Schritt muss noch überprüft werden, ob einen Datenbasis vorhanden ist, aus der diese Informationen gefiltert, gruppiert oder auf einen andere Art und Weise generiert werden können.

Zum Gruppieren bieten sich Merkmale mit extrem vielen Ausprägungen an (z.B. Postleitzahlen, Krankenhausausgaben o.ä.). Bei diesen Variablen kann es beispielsweise Sinn machen, sie in hohe, mittlere und geringe Ausprägung einzuteilen. Dies bietet zusätzlich den Vorteil, dass man bei einer späteren Analyse das Ergebnis im Entscheidungsbaum besser einordnen kann. Ein Beispiel hierzu wäre die Angabe Krankenhausausgaben: Eine schlichte Zahl wie 3.147 Euro sagt nur fachlichen Experten mehr – eine Klassifizierung in mittelhohe Ausgaben wäre für alle Anwender sprechender.

Ferner kann auch eine abgeleitete Variable betrachtet werden. Hiermit sind Merkmale gemeint, welche sich implizit aus anderen herleiten lassen oder auch eine Veränderung über einen definierten Zeitraum repräsentieren können. So kann man beispielsweise (näherungsweise) abbilden, ob ein Versicherter in den letzten 12 Monaten umgezogen ist, in dem seine Postleitzahlen an zwei Stichtagen verglichen werden und diese Information in eine neue Variable packt. Weiter könnte zusätzlich danach differenziert werden, wie viele Stellen der Postleitzahl sich verändert haben und in große und kleine Umzüge unterscheiden.
Eine weiteres Beispiel für eine abgeleitet Variable ist eine Aufteilung der Versicherten an Hand von Einwohnerzahlen in die Klassen Großstadt, Kleinstadt und Dorf.
Zusammenfassend gesagt, sind bei der Definition neuer Variablen der Phantasie letztlich keine Grenzen gesetzt – vorausgesetzt, dass die zur Herleitung benötigten Daten

vorgehalten werden. Ob diese Merkmale später tatsächlich ins Kündigerprognosemodell einfließen ist natürlich nicht gesagt, aber das ist auch bei den Basisvariablen nicht garantiert und diese neuen Variablen bieten zumindest das Potential neue Erkenntnisse zu gewinnen.

2.2 Aussagekraft analysieren

Ein wichtiger Schritt nach der Definition der in Betracht gezogenen Variablen und vor dem Einsatz des EM ist eine ausführliche Analyse der später als Inputvariablen vorgesehenen Merkmale. Hierbei steht zum einen die reine Aussagekraft der einzelnen ausgewählten Variablen im Interesse und zum anderen sollte die inhaltliche Plausibilität dieser überprüft werden.

Mit Hilfe einer univariaten Analyse der einzelnen Merkmale, kann man die Antwort auf beide Fragestellungen gemeinsam vorbereiten. Hierbei wird zunächst für jede Ausprägung jeder Variable ein Scorewert berechnet. Der Scorewert für die Variable selber ergibt sich aus der Addition der Scores der einzelnen Ausprägungen. Bei der Suche nach einer geeigneten Formel zur Berechnung des Scorewertes ist dann natürlich zu berücksichtigen, dass Merkmale mit vielen Ausprägungen (z.B. Alter) nicht gegenüber Variablen mit wenig Ausprägungen (z.B. Geschlecht) überbewertet werden.

Als Basiswert für den Scorecode bietet sich sie Standardabweichung an. Je stärker also einzelne Ausprägungen einer Variablen vom Durchschnitt abweichen, desto mehr Aussagekraft wird ihr zugeordnet. Gleichzeitig wird noch die Klassengröße als Faktor genommen.

Tabelle 1: Scoringbericht der Variable Mitversicherte

Mitversicherte	Mitglieder	Relative Kündigerquote	Score
0	2.500.000	145,0%	3
1	500.000	95,8%	1
2	300.000	108,3%	2
3	180.000	99,2%	0
4	50.000	84,7%	1
5+	10.000	70,8%	0
Gesamt	3.542.000	100,0%	7

In Tabelle 1 findet man ein Beispiel zum Scoring einer Variablen (hier: Anzahl Mitversicherte). Da der Score eine relative Größe ist, kann der gesamte nur im Verhältnis zu dem anderer Variablen interpretiert werden. Direkt aus der Tabelle können allerdings

die einzelnen Ausprägungen dieses Merkmals analysiert werden. So kann man erkennen, wie die relative Kündigungsgefahr von der Anzahl der Mitversicherten abhängt. Mitglieder mit mehr Mitversicherten scheinen relativ betrachtet, seltener zu kündigen, als solche ohne. An dieser Stelle ergibt sich noch mal die Möglichkeit einer fachlichen Überprüfung der Ergebnisse aus der univariaten Analyse. Ist es inhaltlich nachvollziehbar oder gibt es evtl. noch Probleme mit den Daten?

Anhand der gesammelten Scorewerte aller Variablen, kann man zwar nicht abschließend selektieren, welche Merkmale in den Entscheidungsbaum einfließen, aber man kann zumindest eine Vorselektion durchführen und Variablen mit fachlich nicht nachvollziehbaren Ergebnissen ausschließen, Datenfehler entdecken oder auf Basis der Erkenntnisse alternative Variablen definieren. So haben wir zum Beispiel in einem Bundesland eine auffällig hohe Kündigerquote gesehen. Diese konnte fachlich mit einer lokal tätigen anderen Krankenkasse mit einem sehr niedrigen Beitragssatz erklärt werden. Mittlerweile hat diese Kasse jedoch ihren Beitragssatz deutlich angehoben und somit ist dieser regionale Effekt nicht mehr zeitgemäß und würde im Modell zu fragwürdigen Schlüssen führen. Auf Grund dieser Erkenntnisse wurde die Variable so angepasst (Bundesländer zusammengefasst), dass dieser Effekt durch die neue Variable nicht mehr repräsentiert wird.

3 Prior-Vektor setzen

Im Anschluss an die Variablenselektion kommt dann der eigentliche Data Mining-Prozeß, was hier die Entwicklung eines Entscheidungsbaumes ist. Im folgenden Teil soll im Rahmen dieses Prozesses insbesondere auf das Problem einer kleinen Zielgruppe eingegangen werden. Am Beispiel des Kündigermodells wird ein Lösungsvorschlag unter Einsatz eines Prior-Vektors erläutert.

Die vorausgehende Datenanalyse hat eine Datei mit einer Stichprobe aus dem Bestand (nicht Kündiger), so wie allen Kündiger, inklusive der nach Abschluss der Variablenselektion in Betracht gezogenen, erklärenden Merkmalen ergeben. Das „Problem“ im folgenden sind eine sehr geringe Anzahl an Kündiger. Da man eine tatsächliche Kündigerquote im kleinen einstelligen Prozentbereich hat, wird das erste Ergebnis des EM (wenn man alle Einstellungen voreingestellt lässt) keinen Baum produzieren bzw. lediglich einen einblättrigen Baum, welcher sinngemäß sagt „zu fast 99% Wahrscheinlichkeit bist du kein Kündiger“. Dies ist natürlich wahr und isoliert betrachtet ein Superergebnis. Jedoch hilft es bei dem hier behandelten Problem leider nicht weiter.

Einen ersten Ausweg aus diesem Dilemma liefert eine 50/50 Stichprobe. Diese kann man ohne großen Aufwand in einen „Sampling-Node“ generieren. Hierdurch ergibt sich der Vorteil, das die differenzierenden Merkmale besser erkannt werden. Wenn der EM noch einmal läuft, erhält man nun tatsächlich einen Entscheidungsbaum, welcher differenzierte Kündigungswahrscheinlichkeiten liefert.

Nun hat taucht jedoch ein neues Problem auf. Die im Entscheidungsbaum angegebenen Kündigungswahrscheinlichkeiten sind deutlich zu hoch um realistisch zu sein. Hatte man ursprünglich noch eine Basiskündigungswahrscheinlichkeit im kleinen einstelligen Prozentbereich, hat man nun in fast jedem Blatt des Baumes zweistellige Kündigungswahrscheinlichkeiten stehen. Da es sich offensichtlich nicht um „die echten" Wahrscheinlichkeiten handelt - auch wenn sie relativ zueinander richtig sein mögen - hat man keine präsentierbaren Ergebnisse und eben auch keinen realen Ergebnisse.

Einen möglichen Ausweg aus diesem Dilemma bietet sich durch den Einsatz des Prior-Vektors an. Grob gesprochen, verschafft er die Möglichkeit, die ursprüngliche Kündigerquote anzugeben, im Ergebnis zu sehen und gleichzeitig die Vorteile der 50/50 Stichprobe weiterhin zu behalten. Seine Funktionsweise besteht im wesentlichen daraus, dass er die Stichprobe in einem vorgegebenem Verhältnis skaliert; im hiesigen Fall wäre dies also die reale Kündigerquote. Es mag zwar auf den ersten Blick fragwürdig erscheinen, zunächst die Daten auf 50/50 anzupassen um sie dann wieder zurück zu skalieren, allerdings werden dadurch für jeden einzelnen Schritt die besten Voraussetzungen geschaffen, um am Ende reale und präsentierbare Ergebnisse zu erhalten.

Abschließend noch ein paar Anmerkungen zur Umsetzung. Es gibt verschieden Stellen im SAS Enterprise Miner (es wurde der EM für Windows NT, Rel 4.1 verwendet), an welchen man den Prior-Vektor anlegen kann. Die Erfahrung hat gezeigt, dass es sinnvoll ist ihn direkt im Entscheidungsbaumknoten für das Target zu definieren. Ferner sollte im Baum unter dem Reiter „advanced" das Kontrollkästchen für „use prior probability in spilt search" deaktiviert sein. Tut man dies nicht und belässt auch alle anderen Einstellungen unverändert, erhält man wieder einen einblättrigen Baum. Wenn man den Baum mit dem Prior aufbauen möchte hat man natürlich wieder das Ausgangsproblem – eine zu kleine Zielgruppe und genau dies wollte man ja vermeiden. Es sei an dieser Stelle kurz angemerkt, dass der Prior-Vektor auf jeden Fall beim Pruning benutzt wird, unabhängig davon, ob er beim Baumaufbau bereits verwendet wird oder nicht.

4 Controlling

Mit den Ergebnissen aus dem Data Mining konnten unsere Kunden in Kündigerklassen aufgeteilt werden. In der Praxis hat sich eine Aufteilung in 4 Klassen bewährt. Mit einer Kündigungswahrscheinlichkeit von z.B. 0,08 können Endanwender in der Regel nichts anfangen. Hilfreich ist hier anstatt 0,08 dem Endanwender die Kündigungswahrscheinlichkeit "Tief" zu nennen. Wie viele Kündigungsklassen (z.B. Tief, Mittel, Hoch, sehr Hoch) gebildet werden sollten, hängt davon ab, wie die Kündigerklassen eingesetzt werden.
In der Praxis kann eine weitere Aufteilung einer Kündigerklasse in Unterklassen hilfreich sein. So könnten z.B. für ein Call Center die sehr hoch abwanderungsgefährdeten Kunden nochmals unterteilt werden. Damit ist gewährleistet, dass diejenigen, die das allerhöchste Kündigerrisiko haben auch tatsächlich als erstes angerufen werden.

Insgesamt lassen sich mit Data Mining gute Vorhersagen treffen. Das kann einfach überprüft werden, indem man den Scorewert der tatsächlichen Kündiger untersucht.
Stellt man im Laufe der Zeit fest, dass die Menge der Kündiger mit einem sehr hohem Score-Wert anteilig kleiner wird, als die Menge der Kündiger mit einem niedrigerem Score-Wert, dann ergeben sich zwei Fragen.

1.) Ist das Modell noch richtig, d.h. werden die tatsächlichen Kündiger richtig vorhergesagt?
2.) Sind die entsprechenden Kundenbindungsmaßnahmen geeignet Kündiger zu halten?

Ob und wie lange ein Kündigermodell verwendet werden kann hängt von vielen Faktoren ab. Mögliche Einflüsse sind z.B. die Branche in der das Unternehmen tätig ist, gesetzliche Veränderungen, Konkurrenzsituation, ... Letztendlich sollte man ständig die tatsächlichen Kündiger nach ihrer vorhergesagten Kündigungswahrscheinlichkeit untersuchen, um zu sehen, wie gut die Trefferquote ist. In regelmäßigen Abständen empfiehlt es sich das Modell als ganzes zu überprüfen.

Klassisch wird die Beantwortung der 2. Frage mit der Implementierung einer Kontrollgruppe beantwortet. Dies scheidet bei uns aus folgenden Gründen aus:

1.) Eine Kontrollgruppe sollte so zusammengesetzt sein, wie die Maßnahmengruppe. Unsere Erfahrungen haben aber gezeigt, dass jeder Mensch bei der Wahl einer ("seiner") Krankenkasse die Entscheidung aus anderen Gründen fällt. Eigentlich bräuchte man ein "Zwillingspaar" (Ein "Zwilling" in der Maßnahmengruppe der andere "Zwilling" in der Kontrollgruppe). Solche Paare stehen uns nicht zur Verfügung.

2.) Kundenbindungsmaßnahmen haben nur über einen längeren Zeitraum Erfolg. In der Praxis bedeutet dies, dass man die Kontrollgruppe mind. über 2 - 3 Jahre von Kundenbindungsmaßnahmen fernhalten sollte. In diesem Zeitraum verändern sich jedoch bei vielen Teilnehmern (Test- und Kontrollgruppe) die Vorraussetzungen warum sie in der einen oder in der anderen Gruppe sind. Dies sind z.B. Umstufungen von Pflichtversicherung in eine freiwillige Mitgliedschaft, Anzahl an Kindern, Einkommensveränderungen, Morbidität, ...
Die Nachbildung aller Möglichkeiten und somit die Zerlegung der Test- und Kontrollgruppe ist zwar über den Beobachtungszeitraum möglich, führt aber zu immer kleineren Untergruppen.

3.) Um statistisch eine Aussage treffen zu können, benötigt man eine gewisse Anzahl von Kündiger bzw. Nicht-Kündiger. Bei sehr kleinen Kündigerquoten müsste die Kontrollgruppe in ungünstigen Fällen die Größe der Testgruppe haben. Hier stellt sich die Frage, ob ein Unternehmen sich das leisten kann, einer Gruppe von Kunden keine Kundenbindungsmaßnahmen zukommen zu lassen.

4.) Vielfach stehen unsere Kunden untereinander in Verbindung (Ehepartner, Arbeitskollege, Freunde, Bekannte, . . .). Hier kann es zu Missverständnissen kommen, da Kunden sich unterschiedlich behandelt fühlen.

5.) Die Kontrollgruppe sollte vor sonstigen Kundenbindungsmaßnahmen abgeschirmt werden. In der Realität ist das kaum umsetzbar, da es innerhalb eines großen Unternehmens sehr schwierig ist in "reine" und in "sonstige" Kundenbindungsmaßnahmen zu unterscheiden. Oft richten sich bestimmte Angebote an alle Kunden. Für den einen kann ein solches Angebot eine kundenbindende Maßnahme sein, für den anderen ist es ein ganz "normales" Angebot von vielen.

Um dennoch Aussagen über die Wirksamkeit von Kundenbindungsmaßnahmen treffen zu können kann man als Kontrollgruppe diejenigen nehmen, die die gleiche Kündigungswahrscheinlichkeit haben, wie die Maßnahmengruppe, jedoch keine "reine" Kundenbindungsmaßnahmen erhalten. Diese werden in regelmäßigen Abständen hinsichtlich ihres Kündigungsverhaltens untersucht. Jede dieser beiden Gruppen besitzt einen eigenen Kündigerverlauf über die Zeit. Die Zahl der Kündiger kann als Anteil von einem Referenzzeitpunkt genommen werden. Es können dann Aussagen in der Form wie z.B. "im Vergleich zum Januar 2003 haben wir im November 2005 nur 48% der Kündiger in der Gruppe, die "reine" Kundenbindungsmaßnahmen erhalten haben und eine Kündigungswahrscheinlichkeit von "sehr hoch" besitzen.

Sowohl bei der Maßnahmen- als auch bei der Kontrollgruppe werden Veränderungen zwischen zwei Zeitpunkten analysiert. Die zuvor beschriebenen Probleme hinsichtlich einer exakten Kontrollgruppe bestehen zwar weiterhin, können aber außer Acht gelassen werden, da diese "Fehler" in beiden Gruppen und zu jedem Zeitpunkt enthalten sind. Man vergleicht praktisch pro Gruppe zwei Zeitpunkte miteinander. Kundenbindungsmaßnahmen und oder andere Ereignisse können so in der Interpretation der Ergebnisse berücksichtigt werden.

Um nun den Effekt der "reinen" Kundenbindungsmaßnahmen zu erhalten, kann man sich die Differenz zwischen Test- und Kontrollgruppe ansehen.
Wie zuvor beschrieben, ist hier die Kontrollgruppe nicht eine repräsentative Stichprobe aus der Testgruppe (was theoretisch ideal wäre), sondern besteht aus Kunden, die die gleiche Kündigungswahrscheinlichkeit haben.
Daher handelt es sich um Kunden, die sich hinsichtlich einiger Soziodemographischer Faktoren völlig von der Maßnahmengruppe unterscheiden. Wichtig bei der Überprüfung der Wirksamkeit von Kundenbindungsmaßnahmen ist, dass man zuerst jede Gruppe für sich alleine analysiert. Erst danach ist eine Differenzbetrachtung sinnvoll.