ИКОНОМИЧЕСКИ ИЗСЛЕДВАНИЯ
ECONOMIC STUDIES

ГОДИНА XXII, 2013, 3

Тодор Кръстевич[1]

ПОВИШАВАНЕ ЕФЕКТИВНОСТТА НА ДИРЕКТНИЯ МАРКЕТИНГ ЧРЕЗ ПРЕДИКТИВНО МОДЕЛИРАНЕ

Постоянният стремеж към привличането на нови и задържането на съществуващите клиенти не е нова маркетингова парадигма. Пазарната ѝ реализация обаче в условия на относително висока честота на купуване на продукти и услугите, както и при относително ниски бариери пред потребителите за промяна на доставчик и продуктова марка е сериозно предизвикателство. В редица стопански сектори регистрирането, съхраняването и използването на информационни масиви със записи, идентифициращи пазарните субекти и тяхното поведение е съществена и неразделна част от бизнес операциите. В изследването е направен опит за компаративен анализ на класификационни модели и техники за предиктивно извличане на знания от клиентски бази данни и възможностите им за планирането на директни маркетингови кампании и по-конкретно, при избиране на „оптимален“ списък с целеви клиенти за непосредствено персонифицирано въздействие.
JEL: M31; C52; C53

Въведение

Само по себе си наличието на данни за клиентите е необходимо, но не и достатъчно условие за опознаване и предсказване на тяхното бъдещо поведение. Извличането на знания от клиентските бази данни с помощта на предиктивни модели, изграждани с помощта на статистически методи и компютърни алгоритми, разкрива нови възможности за повишаване на ефективността на програмите за директен маркетинг и за подобряване на маркетинговата продуктивност. Аналитичните задачи по съставянето, валидирането и използването на предиктивни модели за оптимално таргетиране, предсказване и в крайна сметка управление на взаимоотношенията с клиентите, поставят нови предизвикателства пред маркетинговите мениджъри. Те са провокирани от нуждата за нов тип знания, от познаването на нови методи, техники и технологии за решаването на познатия стар проблем – задоволяването на потребностите на потребителите чрез предоставяне на допълнителни изгоди и повече ползи, и то по печеливш и рентабилен за фирмата начин.

[1] *Тодор Кръстевич е доц. д-р в СА „Д. А. Ценов” – Свищов, катедра „Маркетинг”, тел: 0631-66298.*

Обект на това изследване е изразеното изборно поведение и пазарните реакции на потребителите на бързооборотни потребителски стоки и услуги, а негов предмет – възможностите за повишаване на ефективността на програми за директен маркетинг на услуги или на бързооборотни потребителски стоки с помощта на методи за аналитично извличане на знания от клиентски бази данни. Обектът на изследване не е подбран случайно. Докато през 2009 г. вътрешният пазар на бързооборотни потребителски стоки възлиза на стойност около 6.5 млрд. лв. (Ватева, 2009), то година по-късно, в края на 2010 г. тази величина се оценява вече на 11 млрд. лв., като между 22 и 25% от него се падат на т.нар. модерна търговия (Кожухаров, 2010). Очакванията са през следващите години тя да се развива бурно и да достигне 60% дял от пазара. Очаква се също до пет години пазарът на бързооборотни потребителски стоки у нас да достигне насищане и да има фалити. Това е един от малкото незасегнати сериозно от финансовата и икономическа криза сектори в страната, бележещи ръст за последните години. Същевременно паралелни проучвания в друг бурно развиващ се пазарен сектор, този на банковите и финансовите услуги, нареждат България на последно място сред страните от Централна и Източна Европа по дял на населението, ползващи банкови продукти и услуги (GfK Bulgaria, 2009, с. 7).

В трети ключов за икономиката сектор, този на мобилните комуникационни услуги, осигуряващ около 6% от брутния вътрешен продукт на страната (срещу средно 2.9% за Европа) и сред най-конкурентните в Европа(Европейска комисия, 2010, с. 3-4), се регистрира пазарна дифузия, далеч надхвърляща стоте процента. На пръв поглед тези три сектора (освен регистрирания ръст) като че ли нямат сходни пазарни характеристики. Общо предизвикателството за маркетинга в тези (на пръв поглед) различни пазари е възможността за масово персонализиране (разбирано като възможност за пренастройване на офертата и подхода на оферирane към индивидуалните изисквания на клиента) и пряко въздействие и взаимодействие с отделния потребител чрез съставяне, поддържане и използване на клиентските бази данни.

Основната цел на изследването е компаративен анализ на някои класификационни модели и техники за предиктивно извличане на знания от клиентски бази данни, които да бъдат приложими при планирането на кампании за директен маркетинг и по-конкретно, при избор на „оптимален“ целеви пазар за пряко, индивидуализирано въздействие. Във връзка с постигането на целта са формулирани следните задачи, условно групирани в четири основни посоки: теоретични, методически, емпирични и технически. В *теоретичен* аспект, задачата ни е да предложим работна процедура, основана на обстоен литературен анализ на процесите в директния маркетинг, теорията на потребителското поведение при избор и покупка и на инструментите за ефективно привличане на нови потребители, както и задържане и разработване на съществуващи клиенти. В *методически* аспект, работната процедура да бъде осъществима чрез модели за аналитично извличане на знания от данни. На сравнителен анализ да бъдат подложени избрани методи и алгоритми за извличане на знания от многомерни масиви с маркетингови данни. Чрез съпоставянето им с традиционните класически RFM-техники за оценка на клиенти, да бъде обоснован подходът на избиране на най-добър метод за

конкретна ситуация и/или пазарен сектор. В *емпиричен* аспект, задачите са да бъдат подложени на валидиране различни алтернативни статистически методи и алгоритми за машинно учене с реални клиентски (транзакционни) бази данни от сектора на бързооборотните потребителски стоки. Най-добре представилият се от тях да бъде използван като ядро при съставяне на комплексен предиктивен модел за оптимално планиране на програма за директен маркетинг. В *технически* аспект, задачата ни е да проучим, съставим, тестваме и дадем предложения и насоки за интерпретация на софтуерни решения, подходящи за реализацията и внедряване на теоретичния модел.

При методическото и техническото разработване и тестване на различни модели за аналитично извличане на знания от данни се ограничаваме методологически върху две концептуални аналитични рамки: SEMMA (акроним от Sample, Explore, Modify, Model и Access), развита от SAS и на алтернативната аналитична рамка CRISP-DM.

1. От масов маркетинг през масово индивидуализиране към аналитично управление на взаимоотношенията с клиентите – теоретична перспектива

За привличането на нови и задържането на съществуващи клиенти са възможни принципно два основни подхода – масов маркетинг и директен маркетинг. При първия се цели индиректно целенасочено разпространяване на търговски апел посредством комуникационни канали за масово осведомяване. Чрез него се таргетират големи групи потребители (формирани на база сходни признаци), но без да се прави разлика между членовете на тези групи (сегменти), разпространяват се и еднотипни информации и/или послания. В съвременната динамична пазарна среда обаче поведението на клиентите е все по-нестабилно и непредвидимо. По-трудно се установяват устойчиви във времето пазарни сегменти. Изграждането на силни, дългосрочни и контролируеми връзки с масовия потребител е на практика невъзможно.

За разлика от масовия маркетинг, директният подход за маркетингово въздействие е насочен непосредствено върху отделния индивид или домакинство и цели еднократно въздействие и/или изграждането на трайни взаимоотношения с клиента. При него, върху всеки отделен клиент е възможно да се въздейства с различна информация, респ. с различни апел, послание и/или оферта. За реализирането на този подход на индивидуализирано и диференцирано комуникиране е необходимо систематично използване на дезагрегирани данни от контакти със съществуващи или с потенциални клиенти, позволяващ достигането на количествени маркетингови цели. В такава среда е необходимо да се използват техники за аналитично извличане на знания от маркетингови бази данни с цел проактивно оценяване на промените в поведението на потребителите и предиктивно моделиране на техния избор. Това улеснява процеса на изграждане и поддържане на дългосрочни, силни и контролируеми връзки между предложителите и техните клиенти. Икономическата логика и обосновката на този подход се крие в постулата, че е много по-скъпо да се привлекат нови клиенти, отколкото за да се задържат съществуващите.

Тази революционаризираща маркетингова парадигма, лансирана в началото на 90-те години на миналия век от Дон Пепърс и Марта Роджърс (Peppers, Rogers, & Dorf, 1999; Pappers & Rogers, 1998), подлага на преосмисляне класическата парадигма на масовия маркетинг. В основата на тази иновативна (за времето си) идея стои презумпцията, че масовото индивидуализиране е не само по-ефикасно, но и икономически по-ефективно (особено днес, в ерата на Интернет). От идеите на Пепърс и Роджърс се развиват съвременната маркетингова концепция за ефективно масово индивидуализиране чрез управление на връзките с клиентите и чрез управление на жизнения цикъл на клиента.

Използването на компютъризирани релационни клиентски бази данни в маркетинга (databasemarketing) цели повишаване на маркетинговата продуктивност и ефективност чрез по-ефикасно привличане, задържане и развитие на клиентите (Blattberg, Kim & Neslin, 2008, p. 4-5). Три са ключовите фрази, с които се описва тази съвременната парадигма на директния маркетинг: използване на бази данни за клиентите (реални и потенциални), маркетингова продуктивност (количествено оценяване маркетинговата ефикасност и ефективност с цел постоянно подобряване на възвращаемостта на маркетинговите усилия) и управление на клиенти (в смисъл привличане, задържане и развиване на техния потребителски потенциал).

Директният маркетинг, управлението на връзките с клиентите и маркетингът чрез използване на релационни клиентски бази данни в много голяма степен се припокриват като концепции, въпреки че всяка една от тях има своя различен нюанс (Blattberg, Kim & Neslin, 2008, p. 5). Управлението на връзките с клиенти е фокусирано към взаимоотношения. Такива обаче е възможно да се поддържат и развиват и без аналитичното използване на бази данни (например поддържане на връзки чрез познанства на търговските представители при малките фирми). Маркетингът чрез използване на релационни клиентски бази данни може да бъде разглеждан като концепция, типична за големите компании, стремящи се да развиват идентифицируеми връзки с многобройните, анонимни потребители). Оттук и самото наименование на софтуерните технологии, обслужващи тази концепция (CRM).

Ключовото при традиционната концепция на директния маркетинг е адресността, т.е. възможността за пряка връзка с клиента. Адресността означава обаче идентифицируемост, която пък е ключов компонент при маркетинга чрез релационни клиентски бази данни. Докато при директния маркетинг е възможно да се работи просто с адресен списък на реални и/или потенциални клиенти (без формален анализ на данни), то маркетингът чрез релационни клиентски бази данни се опира на методи, техники и технологии за аналитично и предиктивно извличане на знания от тези данни. Такъв подход може да се реализира както чрез индивидуализирано изучаване, предсказване и въздействие върху поведението на всеки отделен клиент (типично за директния маркетинг), така и чрез оптимално профилиране, асоциативен анализ и ефективно таргетиращо обвързване на продукти с клиенти.

Повечето изследвания, свързани с аналитично извличане на знания от данни за целите на директния маркетинг са насочени към теоретичните и/или към

алгоритмичните аспекти на технологията. Един от ключовите въпроси, който остава не достатъчно изследван и структуриран като процес, е какви методи за извличане на данни е уместно да се използват за оценяване на клиентите и при какви обстоятелства, дейности, комбинация и последователност да се прилагат в процеса на планиране на програмите за директен маркетинг. Тук акцентът е поставен върху методическите и инструменталните аспекти на оптималното таргетиране, т.е. върху избора на целеви клиенти от релационни клиентски бази данни с помощта на предиктивни техники за аналитично извличане на знания от данни. Другите етапи от процеса на планиране на директните маркетингови кампании, предлагани често в литературата (Jonker, Franses & Piersma, 2002, p. 2-10; Flici, 2009), остават извън фокуса на изследователските ни усилия.

Един от най-важните проблеми при практикуването на директен маркетинг е подборът на клиенти от наличните релационни адресни бази данни, които да бъдат обект на пряко въздействие. При наличието на съхранени масиви от социодемографски, географски и поведенчески данни за клиентите, единственият възможен подход е да се използват количествени модели за описание и предсказване на техните реакции (и респ. предвиждане на тяхното бъдещо поведение). В контекста на това изследване дефинираме понятието „количествен модел" като модел, изграден или на статистически принципи (параметричен модел), или на принципите на машинното учене (модел, съставен от правила), който бива използван за аналитично и предиктивно извличане на знания от бази данни за целите на директния маркетинг. Когато параметрите (респ. правилата) на използваните модели се оценят (респ. създадат и/или обучат), подборът и съставянето на оптимален списък от клиенти (адреси) за директно маркетингово въздействие е възможно да бъде съставен по рентабилномаксимизиращ или по печалбомаксимизиращ начин. Обикновено параметрите (или правилата) не са известни и трябва да бъдат оценени на дезагрегирано равнище. Оценките се използват за степенуване на потенциалните адресати и на тази база се избира най-подходящата целева аудитория (таргет) за въздействие.

2. Планиране на кампания за директен маркетинг – системна перспектива

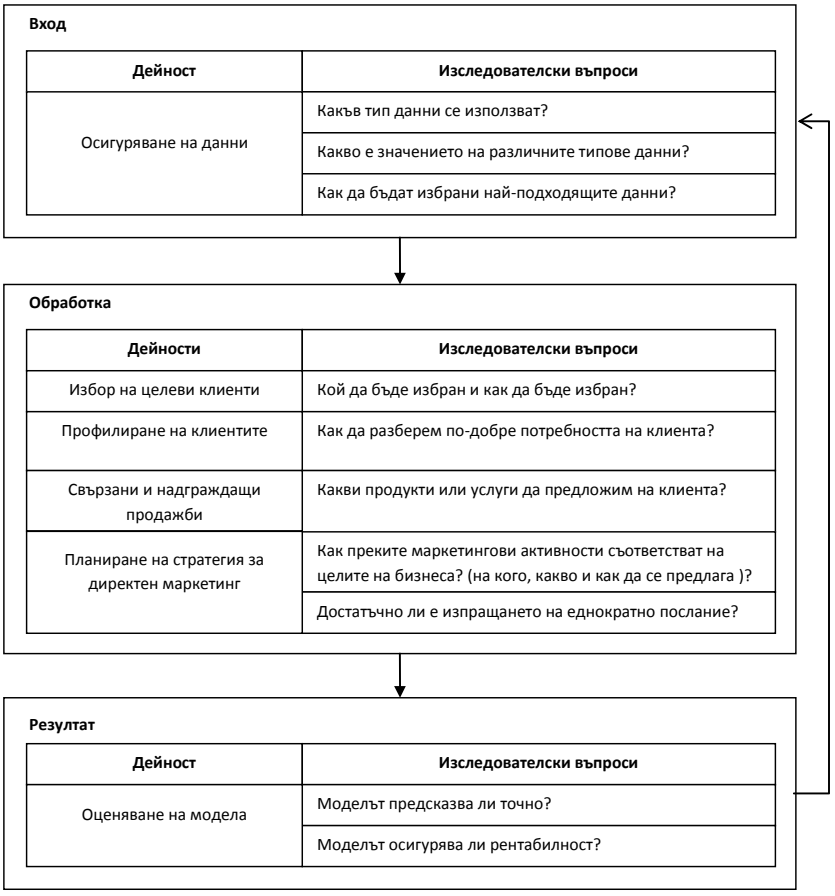
При планирането на кампанията е необходимо да се следва системен подход (Wasson, 2006, p. 22), преминаващ през организирането на изходните данни (вход), оценяване и обучаване на алтернативни модели с помощта на тези данни и съставяне на вероятностни оценки за поведението на клиентите със същите модели (обработка) и валидиране на последните чрез оценяване на тяхната предиктивна способност и икономическа рентабилност (резултат). Крайната цел на подобен стъпков процес е разработването на подходящ инструмент (или набор от инструменти) за подбор на оптимален списък от адресати, на които да бъдат изпращани еднократни или циклични маркетингови въздействия (послания, оферти, информации и др.).

По-широкото и свободното разбиране за стратегия на директен маркетинг се свързва и с поддържането на постоянно взаимодействие между фирмата и

клиентите. Типичната форма на това взаимодействие е изпращането на послание (търговско предложение) до отделния клиент (реален или потенциален), съдържащо покана за покупка на продукт или услуга. След получаването на посланието клиентът трябва да реши дали да приеме или отхвърли търговското предложение. Чрез наблюдаване на реакциите на клиента (респ. купува или некупува), маркетинговите специалисти коригират и настройват тяхната стратегия (на кого да се предлага, какво да се предлага и как да се предлага) и планират (вкл. обезпечават ресурсно) следващата кампания на изпращане на търговски предложения. С други думи, кампаниите за директен маркетинг са циклични и всеки един цикъл е свързан с набиране на данни, тяхната обработка и оценяване на резултатите. Този процес в неговата системна перспектива описва в обобщен вид всеки един модел за директен маркетинг (вж. Фигура 1).

Фигура 1

Системна перспектива при планиране на кампания за директен маркетинг



Източник: Bose & Chen. (2009). Quantitative models for direct marketing: A review from systems perspective. p. 2.

Необходими данниза директен маркетинг

Количествените модели за аналитично извличане на знания от клиентски бази данни изискват два типа данни – описателни и поведенчески. Под *описателни* се има предвид всякакъв вид географски, демографски, социографски характеристики, както характеристики, свързани със стила на живот. От гледна точка на достъпността, този тип данни могат да бъдат дефинирани като външни (van der Sheer, 1998, p. 11-12) и обикновено трудно се поддават на свободно набиране, съхраняване и използване (от правни и/или етични съображения).

Под *поведенчески* данни се разбират всякакви регистрирани взаимодействия между клиента и предложителя (например осъществени транзакции, обратна връзка, регистри от активности върху уеб сървъри и др.). Класическите поведенчески данни се свързват с променливите „време от последната покупка", „честота на купуване" и „монетарната стойност" на покупките на всеки наблюдаван потребител. Чрез съчетаването на тези променливи се дефинира популярния RFM-модел („R" = recency, „F" = frequency, "M" = modetary). С помощта на поведенчески данни е възможно да се оцени както вероятността, с която даден клиент би закупил (или не би закупил) даден продукт, така и потребителската нагласа, интерес и предпочитание към предлаганите продукти или услуги (Blattberg, Kim & Neslin, 2008, p. 325). Особено важна роля при изучаването на динамиката на последните три конструкции и влиянието им върху потребителския избор играят регистрираните данни от поведението на потребителите в Интернет (van den Poel & Buckinx, 2005). В общия случай, поведенческите данни имат вътрешен (за фирмата) характер и се използват като целеви (зависими) променливи при съставянето на предиктивни модели за директен маркетинг.

Предсказването на изборното поведение и респ. решението за покупка се влияе и от характеристиките на предлаганите продукти и услуги. Освен традиционните поведенчески променливи от RFM-модела е възможно да се следи и анализира информация за профила на купуваните едновременно и/или последователно продукти и профила на купуващите ги потребители (например чрез асоциативен анализ, анализ на връзките или анализ на пазарната кошница). Необходимите данни за това са отново записите на отделните транзакции по време и продуктови категории. Според някои автори (van der Sheer, 1998, p. 136; Hansotia & Wang, 1997) върху поведението и реакцията на купувачите оказва влияние дори спецификата, технологията и характеристиките на самото комуникационно въздействие.

Поведенческите данни са естествен фокус при изследванията, свързани с изграждането на предиктивни модели за директен маркетинг. Високата им значимост е породена не само поради все по-достъпните и евтини технологии за тяхното набиране и съхраняване, но и от пряката им съотносимост с целите на маркетинговите кампании. Затова е естествено да се започне с тяхната обработка. Основен проблем обаче е обстоятелството, че този тип данни стават налични едва след като е направена първата поръчка (респ. покупка), т.е. набирането на достатъчно голям обем подобни данни, освен че изисква време и ресурси, е свързано и с пропуснати ползи. Когато целта на

маркетинговия специалист е да привлича нови клиенти, единствената алтернатива е той да стъпи на външни, описателни данни (например социодемографски или географски). Единственият източник на поведенчески данни на практика са тези от следенето на поведението на потенциалните потребители и клиенти в интернет.

Въпреки че основната цел на директния маркетинг е масовото персонализиране, което изисква възможно най-ниско ниво на дезагрегиране на данните (т.е. ниво респондент), понякога се налага (и е единствено възможно) използването и на агрегирана географска информация, като например пощенски кодове или локация на домейни, чрез които да се идентифицират потребителите. Тази информация не е съвършена и не позволява персонализирано обръщение в строг смисъл, но е лесно достъпна и позволява профилиране на клиентите (например обръщение на съответния език на потребителя, в зависимост международната локацията).

Обобщена характеристика на данните (вж. Таблица 1), необходими при изграждането на количествени модели за директен маркетинг, е възможно да бъде изведена на четири различни равнища – клиент (описание), транзакция (поведение), предмет на оферирање (продукт, услуга) и характер на комуникационно средство (тип на съобщението).

Таблица 1

Специфика на данните, необходими при изграждане на модели за директен маркетинг

Обект	Тип данни	Значимост	Достъпност	Променливост
Клиент (описание)	Демографски, социографски, лайфстайл, географски	ниска	ниска (външни данни, трудни за набиране)	висока
Транзакция (поведение)	Регистър на транзакциите, записи от обратна връзка с клиенти, регистри от сърфиране в уеб	висока	висока (вътрешни данни, натрупващи се при увеличаване на транзакциите)	висока
Предмет (продукт, услуга)	Напр. размер, цвят, цена, дизайн, стил, сложност, съвместимост и др.	висока	висока (вътрешни данни)	ниска
Комуникационно средство (съобщение)	Дизайн, стил, технология	ниска	висока (вътрешни решения)	ниска

Източник: адаптирано от Bose & Chen. (2009). Quantitative Models for Direct Marketing: A Review from Systems Perspective. p. 5.

Съвременното равнище на информационните и комуникационните технологии в търговията позволява сравнително евтино и безпроблемно събиране на данни за клиентите. Дори външните данни (например социодемографски) се събират масово чрез различни програми за лоялност, промоционални механизми или в

следствие на промени в законодателството.[2] Огромните масиви от съхранявани данни съдържат стотици променливи, които биха могли да се използват като предиктори при изграждането на количествени модели за директен маркетинг. От информационен аспект това е познавателна възможност, но от методически аспект поражда редица въпроси и проблеми, свързани с осигуряване на статистически предпоставки, необходими за някои от методите. Осигуряването на висока степен на предиктивна точност на използваните модели налага използването на специфични методи и алгоритми за селектиране на най-значимите предиктори на поведението на клиентите (например чрез стъпков подход при регресионно аналитичните методи или чрез априорно филтриране (Bose & Chen, 2009, с. 5). Изборът на правилните данни зависи и от конкретните цели на кампанията. Така при изпращане на послания за (ре-)активизиране на съществуващи клиенти се използват обикновено агрегирани данни за транзакциите на всеки отделен клиент за фиксиран период. При профилирането на потребителите с цел изучаване на предпочитанията им и по-адекватно оферирane е възможно да се търсят асоциативни връзки между отделните обекти (продукти) на транзакциите на отделен купувач (кои продукти се купуват едновременно, кои се купуват в определена последователност и др.). Установяването на закономерности в тези асоциативни връзки е полезно при вземането на решения за това, кои продукти на кои клиенти да бъдат препоръчвани или оферирани едновременно (cross-selling) или кои продукти на кои клиенти да бъдат предлагани като надграждащо оферирane (up-selling).

Аналитично извличане на знания от данни (datamining)

Подобно извличане на знания е процес на автоматично или полуавтоматично разкриване на смислени асоциации, зависимости, повтарящи се образци или структури, тенденции и аномалии в големи масиви от данни, чрез използване на методи от областта на статистиката и математиката, както и техники, алгоритми и технологии от областта на машинното обучение (т.нар. изкуствена интелигентност), разпознаването на образи и визуализация на данни. Процесите на извличане на знания от данни генерират имплицитно формулирана, неочевидна, предварително неизвестна и потенциално полезна информация.

Чрез аналитичното извличане се цели установяването на шаблони или модели, отразяващи фрагментираните многоаспектни взаимоотношения между данните (респ. между различните променливи и/или между отделните наблюдения). Моделите и шаблоните отразяват сложни закономерности или многомерни структури, но подходящо представени в достъпни и интерпретируеми форми.

Аналитичното извличане на знания от данни се опира на съвременни информационни технологии за многомерен анализ, като например

[2] Вж. чл. 83, ал.(1) от Преходните и заключителни разпоредби на Закон за електронните съобщения.

многомерните интерактивни аналитични кубове (OLAP-технологии). За разлика от тях обаче целта е не да се търси просто обобщения в предварително дефинирани многомерни разрези, а да се установят причините и логиката на тези обобщения с цел предсказване на бъдещо поведение или повлияване върху него. Важна особеност на подобен процес е неочевидността и непредвидимостта в издирваните модели и шаблони на поведение и взаимодействия в различни извадки, както и търсенето на по-дълбоки пластове от скрити знания, които могат да бъдат разкрити само при едно детайлно проучване в дълбочина.

Аналитичното извличане на знания от данни е една от най-бурно развиващите се интердисциплинарни области в икономиката на знанието. Най-важният фактор, обясняващ тази тенденция е експлозивното нарастване на обемите съхранени данни. Така най-голямата търговска верига в света Wal-Mart обслужва над 200 млн. пъти седмично своите клиенти в над 8000 търговски обекта в 15 страни в света. Дневно се съхраняват повече от 20 млн. транзакции в обем над 10 терабайта (Shmueli, Patel & Bruce, 2007, p. 3). В телекомуникационния бранш обемите на регистрирани търговски и оперативни транзакции са много по-големи. За сравнение, преди 30-40 години дори и в най-големите световни фирми информацията, съхранявана в електронни бази данни рядко е надхвърляла обеми от няколко десетки мегабайта. Според някои автори (Lyman & Varian, 2003) общият обем на произведена и записана информация в света надхвърля 5 екзабайта (1 екзабайт = 1 млн. терабайта) и темпът този ръст е експоненциален.

Аналитичното извличане на знания от клиентски бази данни цели добиването на нова, полезна информация чрез разкриване на дълбоките и скрити взаимоотношения между на пръв поглед неизвестни и несвързани една с друга величини и аспекти на потребителско поведение. Чрез инструментите му е възможно да се обработват многомерни масиви и да се извличат многомерни зависимости, като същевременно автоматично се разкриват отдалечени случаи и ситуации, отличаващи се от общата, типична закономерност. В аналитичния процес автоматично се извеждат хипотези за разкриване на зависимости между различни компоненти и параметри. Работата на изследователя се свежда до проверка и доуточняване на получените хипотези.

С обекта на аналитичното извличане на знания често се свързват и някои други изследователски дейности, сходни по съдържание, но с различна терминология, като например експлораторен анализ на данни, бизнес интелигентност, дедуктивно машинно обучение, извличане на шаблони от бази данни. В повечето случаи тези дейности включват същото множество от техники и методи, които се използват при добиването на знания от данни за директен маркетинг. Поради своята универсалност, аналитичното извличане на знания от данни намира приложение в различни области, като военно дело, сигурност, комуникации, медицина, публичен сектор, образованието. Най-типичното му и естествено приложение е обаче при подпомагането на маркетингови решения. На практика, най-често срещаните маркетингови проблеми, които могат да бъдат решени с помощта на този аналитичен подход са от областта на директния маркетинг. Ключовите изследователски проблеми

тук обикновено гласят: (1) „Как от целия списък с данни клиенти, как да се установят точно тези, които с най-голяма вероятност биха реагирали на определено маркетингово въздействие?“ или (2) „Кои клиенти е най-вероятно да бъдат изгубени (да се откажат от употреба, да сменят доставчика, да прекратят абонаментен план или да злоупотребят с отпуснат кредит)?“ За решаването на подобни проблеми е възможно да се използват класификационни техники (като например логистрична регресия, класификационни и регресионни дървета – CART), да се използват изкуствени невронни мрежи, C.5 алгоритми и др.), с които да се идентифицират именно тези индивиди, чийто социодемографски, психографски и/или поведенчески профил най-точно съответства на профила на най-ценните съществуващи купувачи или респ. да се идентифицират „рисковите“ случаи (например клиенти, склонни да напуснат, върху които да се упражни персонализирано промоционално въздействие). Възможно е да се приложат и предиктивни модели, с чиято помощ да се прогнозират сумите, които ще похарчат потенциалните купувачи.

Могат да се разграничат и приведат като примери и много други проблемни области и полета в маркетинга, базиран на данни. Във формален план, типичните проблеми на интелектуалният анализ на данни (както често се нарича аналитичното извличане на знания от данни) са свързани с класифициране, моделиране и прогнозиране. По-конкретно, типичните задачи е възможно да се обобщят в следния ред:

- Задачи за класифициране – съотнасяне на информации от входящи вектори (за обекти, събития, наблюдения) към някой предварително известен клас обекти (потребителски сегменти, продуктови групи, типове обекти);
- Задачи за клъстеризиране – подразделяне множеството на входните вектори в групи (клъстери) на базата на изчислена или оценена степен на сходство между обектите;
- Задачи за асоцииране – търсене на повтарящи се образци в поведението. Например, търсене на устойчиви причинно-следствени връзки в пазарната кошница на купувача (marketbasketanalysis);
- Задачи за оценяване, предсказване и прогнозиране – например предсказване на бъдещо поведение, тенденции, величини и събития чрез прилагане на „обучени” модели върху нови данни за непознати клиенти, случаи, наблюдения, събития.
- Задачи за анализ на отклоненията – например, идентифицирането на нетипична мрежова активност позволява изолирането на вредни случаи (потребители, групи).
- Задачи за визуализиране и съкратено описанието – за визуализация на данни, за съставяне на лаконични модели, улесняващи измерването и интерпретацията, за компресиране на събраната и съхранена информация за поведението на потребителите.

3. Методи за избиране на целеви клиенти при директния маркетинг

Основният проблем при директния маркетинг е свързан с предсказване на потребителските реакции на дезагрегирано (индивидуално) равнище. Операционалното решаване на този проблем е свързано с индивидуалното оценяване на клиента, класифициране (клъстеризиране) на клиентите в зависимост от индивидуалните им оценки, тяхното профилиране с цел по-добро предсказване на реакциите им и агрегиране на резултатите от тези реакции на ниво маркетингови цели (например печалба, пазарен дял, ефективност) при използване на определен набор от индивидуални комуникационни въздействия.

Тук ограничаваме вниманието си само върху един, на пръв поглед тривиален въпрос, за оптималния подбор на целеви клиенти, върху които да упражним индивидуално маркетингово въздействие в рамките на ограничен бюджет. Но дори и при този тясно дефиниран проблем (свързан с оценяване, предиктивно моделиране и отсяване), е възможно боравене с голямо разнообразие от познати алтернативни методи и инструменти на математическата статистика. Като прибавим към това множество и познатите инструменти на машинното (само-)обучение, неминуемо възниква въпросът, кой или кои от тях са най-подходящи за решаването на проблема на таргетирането при директния маркетинг?

Полезността и предимствата на един метод или техника над друг може да зависи от конкретния контекст – например типа и от размера на базата данни и типовете променливи, от типа на съществуващите скрити профили и структури, от начинът, по който изпълняват определени логически и/или аксиоматични допускания за статистическото разпределение на данните, от структурата на липсващите данни, както и от конкретните цели на анализа. Възможно е използването на различни методи при решаването на конкретен проблем да води до получаване на различни (понякога противоречиви) резултати. Затова при аналитичното извличане на знания от данни за директен маркетинг е типично да се експериментират различни методи при решаването на даден проблем и да се избира този от тях, който доставя най-полезни (за целите на анализа) резултати. Освен това е типично да бъдат съчетавани и едновременното използвани различни методи и техники в сложни съставни аналитични модели. Като обичайно използвани техники за аналитично извличане на знания от данни могат да се посочат дървото на решенията, асоциативните правила, методът на k-тите най-близки съседи, невронни мрежи, методите на размита логика, наивния бейсов класификатор, генетичните алгоритми, клъстерния анализ по „метода на k-средните, регресионни техники и др. Някои от тези методи са използвани при емпиричните анализи в проекта и на по-късен етап в изложението ще бъдат разгледани по-подробно тези от тях, които имат най-сериозно приложение при решаването на предиктивни и прогнозни проблеми на директния маркетинг. За момента можем да обобщим, че аналитичното извличане на знания от клиентски бази данни представлява съчетаното използване на аналитични техники и инструменти, подпомагащи аналитично ориентирания маркетингов

мениджър в извличането на смисъл от огромни масиви с данни с цел оптимизиране (най-често в смисъл печалбомаксимизиране) на своите решения.

Според предназначението си, инструментите, методите и техниките за аналитично и предиктивно извличане на знания от клиентски бази данни за целите на директния маркетинг е възможно да се класифицират и са удобство в четири основни категории (Wilson & Keating, 2009, p. 443-444):

- Предсказващи
- Клъстеризиращи
- Класифициращи
- Асоцииращи

Предсказващите инструменти за интелектуален анализ на данни се доближават най-силно до конвенционалните методи за прогнозиране на динамични редове, тъй като при тях се цели предсказване на стойността на някаква метрична променлива. Терминът класифициращи се използва в случаите, когато целта е да се предскаже определена категория на дадена неметрична променлива (номинална, по рядко ординална). Така цел на възможен анализ би могло да бъде предсказването на сумата (метрична променлива), която отделните потребители се очаква да изхарчат за закупуването на интересуваща ни марка (или продуктова категория) в рамките на бъдещ период. Възможна цел би могло да бъде обаче и предсказване на вероятността, дали даден потребител ще купи или не от интересуващата ни продуктова категория в рамките на периода, и ако да, колко средства би похарчил. В този случай обект на предсказване са както неметрична (да/не), така и метрична променлива (сума, които ще похарчи).

Клъстеризиращите аналитични инструменти целят установяването на предварително неизвестни за анализатора групи обекти (например потребители), които по естествен път (излизайки от наличните данни) би трябвало да се разглеждат заедно. Доколко идентифицираното чрез оптимални клъстеризационни алгоритми групиране е смислено, зависи от субективната интерпретация на анализатора. Някои от идентифицираните клъстерни решения е възможно да са интересни, но не и полезни (в смисъл информативни) за извеждане на адекватни маркетингови решения. Този клас инструменти се използват преди всичко за постериорно сегментиране и за т.нар. *ненасочено* извличане на знания от данни, т.е. не е налице целева променлива, която да бъде предсказвана (Collica, 2007, с. 77).

Едни от най-често използваните методи за аналитично извличане на знания от клиентски бази данни да целите на директния маркетинг са данните за *класифициращите* методи. Главната им цел е да разграничат различни класове (продукти, клиенти, въздействия). Класифициращите методи са основният инструмент за априорно сегментиране, както и за съставяне на потребителски профили. Те представляват инструменталната база за т.нар. насочено извличане на знания от данни (т.е. задължително трябва да присъства целева променлива, обект на предсказване).

Асоциращите инструменти (наричани още асоциативни правила) се използват за т.нар. асоциативен анализ (анализ на привличането). Например при анализирането на голяма база данни с данни от продажбени транзакции от голям брой различни продуктови марки, интерес би представлявал въпросът, дали при покупката на марка X се купува често и марка Y. Ако се установят подобни устойчиви връзки и същите се опишат чрез т.нар. логически асоциативни правила, е възможно да се планират свързани оферти или да се използват промоционални механизми към определен продукт с цел да се повиши и търсеното на друг, свързан с него продукт. Големи електронни магазини като Amazon.com или Netflix.com използват тези техники като ядро на системите им за индивидуализирани „препоръчителни" оферти.

Класифициращите, предсказващите и асоциращите техники и инструменти съставляват аналитичните методи за предиктивен анализ на данни (често биват наричани „предиктивни модели"). Предикативните модели са основните инструменти за избор на целеви клиенти в директния маркетинг. Освен тях, други типични инструменти за аналитично извличане на знания от клиентски бази данни са техниките за редуциране (консолидиране) на данните, за експлораторен анализ на данни и тези за визуализиране на данни.

Таблица 2

Обзор на основните техники за изграждане на предиктивни модели за директен маркетниг

Статистически техники	Техники за машинно обучение
За наблюдавано (администрирано) обучени	
RFM (обикновен и стохастичен)	Изкуствени невронни мрежи (ArtificialNeuralNetworks = ANN)
Линейна регресия	Дървета на решенията (DT) с автоматични класификатори (C5.0)
Логистична регресия	Наивен бейсов класификатор
Обобщени линейни модели (GLM)	Метод на k-тите най-близки съседи (K-NearestNeighbor = KNN) с целева променлива
Изборни модели (Logit/Probit/Tobit)	Генетични алгоритми (GA) и еволюционно програмиране (EP)
Дискриминантен анализ	Метод на опорните вектори (SupportVectorMachines = SVM)
Дърво на решенията (DecisionTrees) с CHAID/CART/QUEST	Автоматични класификатори, базирани на правила (C5.0)
За ненаблюдавано (неадминистрирано) обучение	
Двустъпковаклъстеризация	Асоциативни правила
	Метод на k-тите най-близки съседи (K-NearestNeighbor = KNN) без изрично дефиниране на целева променлива
Съчетани техники (EnsembleModeling)	
Например съчетаване на (ANN + Logit + RFM) или (ANN + DT + Logit) при изчисляването на оценките на клиентите – за подробности вж. Seni&Elder, 2010; Suh, Lim, Hwang & Kim, 2004 или Seni & Elder, 2010.	

Списъкът с възможни методи и техники за аналитично извличане на знания от данни е дълъг, но може да се обобщи в два големи класа (вж. Таблица 2).

Първият съществен клас представляват тези от тях, които могат да бъдат използвани за съставянето на предиктивни модели на индивидуалните потребителски реакции (т.нар. насочени администрирани методи за аналитично извличане на знания от данни), при които задължително присъства целева променлива. Вторият важен клас са техниките за ненаблюдавано (неадминитстрирано) обучение. При тях не е необходимо изричното посочване на зависима (целева) променлива. И в двата контекста, методите и техниките за съставяне на предиктивни модели могат да се разглеждат и в друг разрез – статистически техники и техники, базирани на машинното обучение (Bose & Chen, 2009, p. 6).

Ограниченията и целите на този аналитичен обзор не предполагат детайлното и изчерпателно представяне на всички класове техники, принципно приложими при аналитично извличане на знания от данни за целите на директния маркетинг. Според нас, много по-голямо значение за потребителите на това знание е формалната логика и особеностите на изграждането на един предиктивен модел, неговата оценка, както и насоките за интерпретация на аналитичните резултати.

4. Логика на съставянето, интерпретация и оценка на предиктивните модели за директен маркетинг

Независимо от сравнително голямото разнообразие на методи, техники и алгоритми за аналитично извличане на знания от клиентски бази данни, логиката на тяхното оценяване и използване е много сходна. Тази логика, както и някои насоки за интерпретация и оценка са апробирани върху реална клиентска база данни, с помощта на някои от най-разпространените статистически техники за съставяне на предиктивни модели – дърво на решенията (т.нар. класификационни и регресионни дървета), метода на k-тите най-близки съседи, наивния бейсов класификатор и логистичната регресия.

Особености при съставянето на предиктивни модели с дърво на решенията

Основен клас техники за аналитично извличане на знания от данни и за съставяне на предиктивни модели за директен маркетинг се обобщават под общото наименование „дърво на решенията". Под дърво на решенията се разбира група класификационни и регресионни алгоритми (наричани често класификационни и регресионни „дървета"). Както самото родово понятие подсказва, тук става дума за итеративно подразделяне на голяма съвкупност от обекти (например, наблюдения с данни за потребители) в последователни по-малки и подмножества (групи), в които те стават все по-сходни помежду си на базата на „правила". Класификационните дървета се използват за предсказване на категории (например купил/некупил) и по-рядко за прогнозиране на непрекъснати метрични променливи. Специалистите по директен маркетинг обикновено използват различни метафори за обясняване на механизма и резултатите от този клас техники. Регресионните дървета се използват по-често при предсказване на метрични данни. Когато се използват дървовидни техники за класифициране и/или предсказване на категории, обикновено те се

наричат „класификационни дървета" (или дървета на решенията). Последното понятие е наложено в жаргона на машинното обучение, както и в теорията на управлението, докато понятието „регресионно дърво" е по-скоро със статистически произход.

С понятието „дърво на решенията" обозначаваме цял популярен непараметричен клас класификационни алгоритми, чието прилагане за аналитично извличане на знания от данни не изисква от анализатора приемането на предварителни допускания, хипотези, знания и оценяване на параметри. Методите от тип „дърво на решение" спадат към методите за машинно обучение от клас контролирано (администрираното) обучение, при което даден модел принудително се „обучава" с помощта на примери от типа „стимул-реакция". При този тип методи, интересуващата ни зависима променлива е позната и алгоритъмът „учи" как да предскаже нейните стойности на базата на нови наблюдения, за които разполагаме с данни за независимите променливи. С помощта на такива данни (наричани „обучаващо множество") е възможно да се състави дървовидна йерархична структура от правила, с чиято помощ да бъдат предсказвани стойности (респ. принадлежност към априорно известни категории) за всеки нов непознат случай – наблюдение).

Класификационното дърво представлява непараметричен модел с йерархична дървовидна структура, който с помощта на верига от въпроси (или правила) класифицира/предсказва случаи на базата на известна информация за техните описателни признаци (атрибути). Признаците (независими променливи) могат да бъдат с различно равнище на скалиране – номинални (в т.ч. дихотомни), ординални или метрични, докато зависимата променлива е необходимо да бъде неметрична (номинална или ординална). При наличието на достатъчно по обем наблюдения за атрибутите (независимите променливи) и за класовете (категориите) на зависимата променлива, с помощта на този метод е възможно да се създаде поредица от правила (или серия от въпроси), с чиято помощ да се предсказва категоризацията на класовата принадлежност на наблюденията.

Традиционните статистически методи за решаване на предиктивни проблеми (например регресионен и дискриминантен анализ) се свеждат до разработване на модел, който да изглажда изходните данни. За тях е прието, че следват определено вероятностно разпределение, оценка на качеството на апроксимация (степента на приближение на модела към изходните данни) и извеждане на оценките на параметрите на модела, които в последствие се използват за предсказване. При моделите от типа „дърво на решенията" се използва друг подход. При тях, на база връзката между независимите и зависимите променливи, се извършва последователно разчленяване на обекти. Полученото дърво на решенията показва кои независими променливи в най-висока степен се корелират със зависимата променлива, както и подгрупите (крайните възли), в които биха могли да са концентрирани случаите с желаните характеристики.

Процесът на съставяне на класификационни дървета може да се нарече индуктивен, тъй като йерархичните дървовидни модели се изглаждат и „обучават" на базата на изходните данни. Преди да демонстрираме един от

популярните алгоритми за индукция на класификационно дърво е необходимо да се дефинират някои основни показатели (мерки) за хетерогенност, които се използват за оптимизиране на процеса.

При наличието на изходна база от данни, съдържаща значения за атрибутите (предиктори) и значения на категориите (класовете) на зависимата променлива, е възможно да се измери степента на хомогенност (респ. на хетерогенност) на базата на тези класове. Дадена изходна база данни се нарича хомогенна (или „чиста"), ако зависимата променлива съдържа случаи, попадащи само в един единствен клас (т.е. само в една категория на зависимата променлива). Ако наблюденията попадат в повече от една категории, изходната база данни се нарича хетерогенна. Познати са различни показатели за количествено измерване на степента на хетерогенност. Най-популярните от тях са индекса на ентропия, индекса на Джини и индекса на класификационната грешка. Тези три измерителя се изчисляват по следния начин:

$$Entropy = \sum_j -p_j \log_2 p_j$$

$$Gini\ index = 1 - \sum_j pj_j^2$$

$$Classification\ Error = 1 - \max_j[p_j]$$

И трите формули съдържат вероятността pj, с която дадено наблюдение би попаднало в клас (категория) j на зависимата променлива. Стойността на ентропията при една съвършено хомогенна таблица е 0 (тъй като вероятността за принадлежност е 1 и $\log_2(1)=0$), а максимално възможната й стойност се постига, когато всички категории на зависимата променлива имат еднаква вероятност (Teknomo, 2009).

Стойността на индекса на Джини при съвършено хомогенна таблица е 0, тъй като вероятността за принадлежност ще бъде 1 при една категория и $1 - (1)^2 = 0$. По подобие на коефициента на ентропия и този индекс получава максимална стойност тогава, когато всички категории на зависимата променлива имат еднаква вероятност. За разлика от коефициента на ентропия обаче индексът на Джини заема стойности в диапазона от 0 до 1, независимо от броя на категориите.

Третият алтернативен показател за измерване на хетерогенността на изходната база данни е класификационната грешка. Подобно на индексите на ентропия и на Джини, индексът на класификационната грешка приема стойност нула в само случаите на една категория (перфектна хомогенност), тъй като $1 - \max(1) = 0$. Диапазонът на възможните стойности е между 0 и 1. На практика максимално възможната стойност на индекса на Джини при даден брой категории е винаги равна на максимално възможната стойност на индекса на класификационната грешка.

Таблица 3

Разграничаване на основните алгоритми за конструиране на дърво на решенията

Характеристика	CHAID	Изчерпателен CHAID	CART	QUEST	C5.1
Тип разчленяване	Многочленно (повече от две субгрупи)	Многочленно (повече от две субгрупи)	Двучленно/ бинарно (две субгрупи)	Двучленно/ бинарно (две субгрупи)	Многочленно (повече от две субгрупи)
Непрекъсната (метрична) зависима променлива	Да*	Да*	Да	Не	Не
Непрекъснати (метрични) независими променливи	Да**	Да**	Да	Да	Да
Анализ на грешните класификации (разрастване на дървото)	Не	Не	Да	Да***	Да
Статистически тест при подбора на независимите променливи	Да	Да	Не	Да	Не
Статистически тест при разчленяването на обектите	Да	Да	Не	Не	Не
Скорост	Средна	Средна	Ниска	Средна/Ниска	Висока
Използване на априорни вероятности	Не	Не	Да	Да	Не

* В повечето достъпни софтуерни пакети (SPSS, SAS EM, DTREG) логиката на метода CHAID е доразвита, като се предлага възможност за включване на ординални и непрекъснати независими променливи.

** Стандартно, непрекъснатите независими променливи се конвертират в ординални променливи, съдържащи 10 приблизително еднакви по размер категории.

*** При наличието на симетрични грешни класификации, същите могат да бъдат включени посредством априорните вероятности.

Познати са различни алгоритми за оптимизиране на модел на класификационно дърво, като например CHAID, CART, C5.0, QUEST и т.н., всеки от които има специфични особености (вж. Таблица 3). Общото между тях е, че са рекурсивни (базират се на рекурсивния алгоритъм на Хънт, при който се използва само локалния оптимум на всеки възел, без да се прави изчерпателно обратно проследяване). Тази подход не е оптимален в математически смисъл, но е изключително бърз. Ключов контролен параметър при него е т.нар. информационен принос. Последният е сравнителен критерий

за равнището на хетерогенност преди и след съответната итерация на разчленяване в класификационното дърво. Стойността му се изчислява за всеки предиктор, като степента на хетерогенност на изходната база данни се намали с претеглената сума от степените на хетерогенност на дъщерните таблици. Като тегло се използва броят на наблюденията на всеки атрибут (променлива). Ако за измерване на степента на хетерогенност се прилага индекса на ентропия, то информационният принос се изразява по следния начин:

$$\text{Инф.принос}_i = \text{Ентропия}(D) - \sum_i \frac{n_k}{n} \cdot \text{Ентропия}(S_i)$$

Чрез сравняването на информационните приноси между отделните предиктори се идентифицира т.нар. оптимален предиктор, т.е. атрибута (независима променлива), чрез който се постига най-висок информационен принос при разчленяването. Процесът се повтаря рекурсивно до достигане на зададените от изследователя стоп-критерии (например невъзможност да се подобри стойността на информационния принос или брой случаи в дъщерната матрица).

Технологията и алгоритмите за съставяне на предиктивни модели за директен маркетинг са различни, но логиката на тяхното оценяване и използване е една и съща. Методът CHAID (Chi-squaredAutomaticInteractionDetection) предлага възможност за ефективно проучване и идентифициране на съществуващи зависимости между независимите променливи и категорийната зависима променлива(Kass, 1980). При прилагането му се генерира дървовидна диаграма, чрез която се изобразяват комбинациите от категории, имащи най-голямо значение за резултата. CHAID е ефективен инструмент за идентифицирането на сегменти, които максимизират желания резултат (Magidson, 1994).

Методът CART (Classification and Regression Tree) е предложен за първи път през 80-те години на миналия век (Breiman, Friedman, Olsen & Stone, 1984) и представлява мощен, самостоен статистически алгоритъм за структуриране на оптимални класификационни и регресионни дървета. Той позволява провеждането на два типа анализи: класификационни дървета за категорийните и регресионни дървета за метричните (непрекъснатите) зависими променливи. Този метод извършва последователни бинарни разчленявания на изходната извадка на базата на дефиниран критерий. Противно на CHAID, CART не се основава на статистически тестове при вземане на решение за включването на една независима променлива в модела. Вместо това се прилага друг подход. За всеки възел за разчленяването на данните се използва онзи предиктор, който осигурява най-голямо подобряване на критерия.

Конструирането на дърво при CART с номинална зависима променлива се ръководи от критерия на хетерогенност. С него се улавя степента, в която случаите от даден възел са концентрирани в отделна целева категория. Използването на равнището на хетерогенност при изграждането на дървото е съпроводено с два проблема. Хетерогенността почти винаги може да бъде

редуцирана чрез увеличаване на дървото и всяко дърво ще има нулева хетерогенност, ако е достатъчно голямо. За справянето с такива проблеми се използва измерителят „разходи/сложност", който налага „наказание", нарастващо с увеличаване размера на дървото. Тази функция за цялото дърво или за част от него се изчислява както следва:

Cost-Complexity = R(T) + a.|T|,

където:

R(T) е измерител на ресубституционния риск на дърво (или клон) T;

a – наказателен коефициент;

|T| – броят на терминалните възли в дърво (или клон) T.

Алтернатива на методите CHAID и CART представлява сравнително по-новия статистически алгоритъм QUEST (от англ. Quick, Unbiased and Efficient Statistical Tree), предложен от Ло и Ши (Loh & Shih, 1997), чиято логика е много сходна с тази на CART. Основното предимство на QUESTе бързината, ефективността и възможността за работа с липсващи стойности. Ограничение е възможността му да съставя само двоични класификационни дървета.

Сходна логика за конструиране на дърво на решенията използва и компютърния алгоритъм за машинно обучение, основан на изкуствен интелект и предложен през деветдесетте години от Джон Рос Куинлън (Quinlan, 1993). Този алгоритъм е претърпял многократни подобрения и актуалната версия се нарича C5.1. Често пъти алгоритъмът на Куинлан се нарича „статистически класификатор" и по подобие на методите CHAID и CART генерира множество от правила, максимизиращи информационния принос.

Ключов момент при използването на предиктивни модели за директен маркетинг е генерирането на правила, с чиято помощ да се оцени вероятностно и да се предскаже поведението на отделния клиент. В зависимост от спецификата на използваната техника (статистическа или машинно обучение), правилата се съставят по различен алгоритъм, стъпвайки на информация от всичките или от избрани предиктори. Така при прилагането на класификационни дървета правилата имат логически характер (например „…ако клиентът е над 25 и под 35 г., и ако е женен, и ако получава месечна заплата над 1300 лв. и ако е клиент на банката от поне 5 г. то….". вероятността му да използва продукта е 0.79). При дихотомна целева променлива, всички клиенти с оценка на вероятност над 0.5 се класифицират в категорията 1 (приема офертата), а останалите в категория 0 (не приемат). Тъй като вероятностната оценка се получава на индивидуално равнище и би могла да бъде използвана по по-прецизно сегментиране (например клиентите, които имат вероятност около критичната стойност от 0.5 да бъдат обект на специално внимание, докато от тези, с вероятност да закупят над 0.8 да бъде спестен маркетингов бюджет за въздействие. По същата логика е възможно да бъдат „пропуснати" всички, които имат вероятност за въздействие под определена критична вероятност или до изчерпване на разполагаемия бюджет.

Един от най-важните въпроси, които обикновено си задава маркетинговият мениджър при избиране на списък от клиенти (или сегменти от клиенти) за таргетиране , е каква печалба би могла да се очаква след провеждането на съответната кампания за директен маркетинг. При съставянето на какъвто и да е предиктивен модел от интерес би могло да бъде възможността за специфициране на (очакваните) печалби за категориите на целевата променлива. Това е обичайна практика за директния маркетинг и маркетинга чрез бази данни, въпреки че би било ценно и за всяка друга приложна област.

Особености при съставянето на предиктивни модели с метода на k-тите най-близки съседи

Този метод (k-nearest neighbor = KNN) е администриран (контролиран) самообучаващ се класификационен алгоритъм, при който класифицирането се извършва на базата на мнозинството на k-тите най-близки съседни категории. Целта на алгоритъма е да класифицира нов обект на базата на информация от неговите атрибути. За класифицирането не се използва (изглажда) конкретен параметричен модел, а се извършва на базата на заучен критерий. При зададена отправна точка (обект) се търсят k-броя обекти, прилежащи най-плътно до отправния (като мярка за проксимация се използва дистанцията на Евклид). При неметрично скалирани независими променливи е възможно да бъдат използвани други мерки за сходство (например, дистанция на Минковски, блокова дистанция, дистанция на Чебишев дистанция на Ръсел и Рао и др.). За класифициране се използва преобладаващата класификационна категория сред k-тите обекти. Алгоритмът използва класифицирането на съседните случаи като предсказваща стойност за класифициране на новопоявяващи се случаи.

Алгоритмът на KNN е пределно прост. Той се базира на минималната дистанция между даден обект до обучаващата извадка с цел определянето на k-тите най-близки съседни случаи. След идентифицирането на оптималното k (като критерий се използва минималната грешка), на принципа на простото мнозинство, се определя принадлежността на интересуващия ни обект (т.е. обектът се причислява към мажоритарната подкатегория на зависимата променлива). За провеждането на KNN са необходими набор от предиктори X_i (независими променливи) и една зависима променлива Y. Променливите е възможно да бъдат както метрични, така и номинално или ординално скалирани.

Особености при съставянето на предиктивни модели с наивен бейсов класификатор

Набираща популярност техника за класифициране и предсказване при аналитичното извличане на знания от данни е наивният бейсов класификатор (от англ. Naïve Bayes). С тази техника е възможно да се предскаже вероятността, с която дадено наблюдение (случай) би попаднал в предварително известна категория (клас). Техниката се базира на популярната

бейсова теорема от теорията на вероятностите. Бейсовият класификатор често пъти дава сравними и дори по-добри резултати от класическите класификационни и регресионните дървета (Rish, 2001). Освен това той е изключително бърз, което му дава предимство при използването на големи бази данни. Моделът се нарича наивен, тъй като по наивен начин допуска, че предикторите (атрибутивните променливи) са независими помежду си. Това допускане, разбира се, никога не може да се изпълни изцяло, но въпреки нарушаването му, класификационните схеми, които се получават, са с висока степен на коректност.

Съставянето на модела започва аналогично на разгледаните предходни техники: данните е необходимо да бъдат разделени на обучаващо и на валидиращо подмножества. Споменахме, че в основата му стои теоремата на Бейс (по името на Томас Бейс). Тя се използва в теорията на вероятностите за изчисляване на вероятността за настъпване на дадено събитие (неречена постериорна или условна вероятност), след като вече е известна част от информацията за него (априорна вероятност). За яснота, нека си представим следния изследователски въпрос: каква би била вероятността даден клиент да приеме оферта за нов продукт, ако е известно, че той вече притежава предходен вариант на продукта? (обръщаме внимание на обстоятелството, че притежаването на предходен вариант предшества решението за адаптиране на новия вариант). В този пример, постериорната (нарича се още „условна") вероятност се обозначава като P(A|B). P(A|B) е вероятността за адаптиране на новия продукт при положение, че клиентът вече притежава предходния вариант. P(A) се нарича априорна вероятност на събитието A и изразява вероятността, че който и да е клиент би приел новия продукт, независимо от това дали е притежавал до момента предходен вариант на същия продукт. С помощта на тези символи, теоремата на Бейс може да се изрази по следния начин:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

където P(A) е априорната вероятност на A. Нарича се априорна в смисъл, че не отчита някаква информация за B. P(A|B) се нарича условна (постериорна) вероятност на A, при дадено B. Нарича се постериорна, тъй като се извежда (или зависи) от конкретната стойност на B. Това е вероятността, която обикновено се търси. P(B|A) се нарича условна вероятност на B при известно A. P(B) се нарича априорна вероятност на B.

Особености при съставянето на предиктивни модели с изкуствени невронни мрежи

Изкуствените невронни мрежи – ИНМ (от англ. Artificial Neural Networks – ANN) са сравнително малко познат клас аналитични техники, чиито възможности за приложение в директния маркетинг се коментират поляризирано. От една страна, в тях се търси граничещ с мистицизъм аналитичен потенциал, а от друга, краен скептицизъм относно валидността на

прилагането им, породен може би от високата степен на непрозрачност и контрол на обработката на данните. Въпреки това, ИНМ имплицитно присъстват в арсенала с аналитичен инструментариум за извличане на знания от клиентски бази данни, подпомагащи решенията за оптимално таргетиране.

ИНМ, наричани от някои и „конексионистични" мрежи, представляват клас съвременни, компютърно базирани инструменти за извличане на данни от големи, полуструктурирани информационни масиви и тяхното трансформиране в знание за подпомагане вземането на управленски решения. Тези мрежи са опит за опростено аналогизиране на мозъчната структура на бозайниците с математически средства и със средствата на компютърната и математическата логика. Те се състоят от стоящи в йерархични зависимости системи от взаимосвързани изкуствени елементи (изкуствени неврони) за симултантна обработка на данни, което наподобява начина на обработване на информацията от човешкия мозък. Силата на зависимостите се определя чрез процес на обучаване (метод за параметризиране). Въпреки редицата ограничения (в следствие на опростяването в сравнение с биологичните невронни мрежи), този клас аналитични техники все по-успешно и широко се прилагат през последните години при решаването на практически, технически или икономически проблеми.

ИНМ са модерно допълнение към класическите многомерни аналитични методи. Понякога чрез тях е възможно да бъдат решавани по-сложни предиктивни проблеми, с които класическите статистически методи не се справят успешно.

Невронните мрежи са в състояние самостоятелно да извличат закономерности и знания от налични масиви с маркетингови или финансови данни и с това разкриват нова форма за анализ на данни. Основното им предимство се проявява при анализа и решаването на лошо структурирани (или трудно поддаващи се на структуриране) изследователски проблеми.

Основните елементи за обработка на информацията в невронни мрежи са изкуствените неврони ("нервни" клетки), обвързани помежду си и организирани на слоеве. По такъв начин е възможно да се пресъздават чрез модел сложни, нелинейни връзки и зависимости между голям брой променливи, и то без наличието на предварителни знания и допускания за очакваната посока и интензивност на тези връзки. Съставената невронна мрежа е необходимо най-напред да бъде "обучена" с помощта на наблюдавани (налични) данни. При това се прави разграничение между фазата на обучение (при която истинските резултати са известни от данните и се прави опит да се репродуцират чрез администрирано, т.е. контролируемо обучение на модела) и фазата на предсказване, при която истинските резултати не са известни и се достигат чрез имитиране на същата схема за обработка на информацията (неадминистрирано, респ. неконтролируемо обучение). След фазата на обучение изкуствената невронна мрежа е конфигурирана и може да се използва за анализ на данни и/или прогнозиране (Backhaus, Erichson & Weiber, 2011, p. 173).

Съществуват различни типове изкуствени невронни мрежи. Като най-подходящи за целите на предиктивното моделиране на потребителския избор се оказват т.нар. многослойни невронни мрежи с обратно разпространение на грешката. Специфичното при използването на многослойните невронни мрежи е процесът на тяхното параметризиране. Последното представлява по същество процес на машинно учене, изразяващо се в решаването на нелинеен оптимизационен проблем, целящ достигането до глобален минимум на функция (Смокова, 2007). С помощта на „обучената" невронна мрежа е възможно да се предсказва постериорната групова принадлежност на обектите (клиентите) към предварително известни категории (например купувачи и некупувачи). За разлика от коментираните вече алтернативни методи, груповата принадлежност се извежда по детерминистичен (а не по стохастичен) начин, което предполага априорното задаване на абсолютна или относителна прагова стойност от страна на анализиращия (Lackes & Mack, 2000, p. 96).

Особености при съставянето на предиктивни модели с логистична регресия

Като специфична класификационна и предиктивна техника за аналитично извличане на знания от клиентски бази данни за целите на оптималното таргетиране е възможно да се използва и стандартна логистична регресия (Wilson & Keating, 2009; Komarek & Moore, 2005). Тази техника е естествено допълнение към линейната регресия с метода на най-малките квадрати. За разлика от линейната регресия обаче логистичната изисква номинално равнище на скалиране на зависимата променлива. Независимите променливи (предиктори) е възможно да бъдат както метрични, така и неметрични.

Целта на логистичната регресия е да се предскаже вероятността за настъпване (P) на интересуващата ни категория на целевата променлива в предиктивния модел, описан чрез съвкупност от предиктори X_1, X_2,…,X_k. При дихотомна целева променлива общият вид на модела е:

$$\mathrm{Log}\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots + \beta_k X_k$$

Тъй като логистичната регресия е по същество статистически метод, прилагането му върху твърде големи бази данни понякога е проблематично (или неефективно от гледна точка на изчислителен процесорен ресурс).

Независимо от използваните предиктивни модели за директен маркетинг, логиката за оценяване на техните предиктивните способности е една и съща. При обследване пригодността на предиктивните модели за директен маркетинг се извежда рискът от грешка в предиктивната (класификационната) способност на модела. Обикновено се изчислява общият процент на грешка (процентът на грешно класифицираните случаи) и неговата стандартна грешка. Моделите (най-често при аналитично извличане на знания от клиентски бази данни се разчита едновременно на няколко техники), които се представят най-добре по желан от анализатора критерий за предиктивна точност (или

ефективност), се използват като основа на механизма на планиране на съответната кампания за директен маркетинг.

5. Избор на оптимален алгоритъм

Методите и техниките за аналитично извличане на знания от клиентски бази данни е възможно да бъдат използвани съчетано, под формата на цялостен, комплексен модел за оптимизиране на програмите за директен маркетинг. Тук ще демонстрираме един от възможните подходи за реализирането на този процес. Като изходни данни са използвани реални бази данни за годишните активности на електронен магазин, предлагащ няколко продуктови категории техника и електроника за бита. Базата данните отразяват всички 69216 регистрирани транзакции за календарната 2009 г. от страна на 12589 клиента. През наблюдаваната година фирмата е провеждала постоянно промоционални кампании и е регистрирала тези клиенти, които са откликнали (дали поръчка) поне веднъж. При регистрацията на сайта, както и при провеждането на периодични проучвания за удовлетвореност, от клиентите систематично са набирани и някои социодемографски данни. Подобни информации обикновено се организират в таблици от базата данни на фирмата. С случая данните за транзакциите представляват таблица с 46920 реда (някои клиенти са осъществили повече от една поръчка през периода). В нея са регистрирани също моментът (датата) на направената поръчка и заплатената цена по фактура. В отделна таблица са съхранени данните за клиентите (с техния индивидуален клиентски номер, семейно положение, предпочитан начин на плащане, средно време, прекарано на сайта, годишни доходи на домакинството). Тази таблица съдържа 12589 записа (съответстващи на 12589 регистрирани клиенти). През наблюдавания период фирмата е провела една мащабна промоционални кампании. В отделна, трета таблица е съхранена информацията за реакцията на клиентите на промоционални активности (дихотомна променлива, със значение 1 = ”реагирал” и 0 = „нереагирал”). Като предиктори са тествани различни комбинации от независими променливи, като в окончателните модели присъстват само тези, които са статистически значими, респ. които имат достатъчно висока дискриминантна способност.

С данните от описаната база, прилагайки най-напред конвенционална RFM-методология и на базата на установените RFM-оценки, съставихме пирамидален модел на клиентите (Curry & Curry, 2000, p. 9), подразделящ ги на топ клиенти (1%), големи клиенти (5%) средни клиенти (20%), дребни клиенти (80%) и неактивни клиенти. Идентифицираните категориите е възможно да се използват в последствие за различни типове анализи. Предимството на този подход е фокусирането върху категории и терминологии, които са близки и разбираеми за мениджърите. Така ако целта е привличането и задържането на най-доходоносните клиенти, терминология от типа „топ X% важни за таргетиране клиенти” лесно се комуникира сред вземащите решенията. Ако акцентът е в друга посока, например осигуряване на бъдещи постъпления, концептуалният модел помага при анализа на движението на клиентите между категориите (фокусирайки вниманието на изследователя върху проучване на причините, т.е. предикторите, водещи до

това преместване). И не на последно място, ако целта на фирмата е задържането на спечелени вече клиенти, моделът фокусира вниманието върху изследването и оценяването на риска и момента на напускането, респ. преминаването към друг доставчик, предложител или марка.

Хронологията на реализация на използвания комплексен модел е реализиран в следната последователност:

(1) Агрегиране на продажбените транзакции на ниво клиент;

(2) Изчисляване на оценките на трите основни параметъра на RFM модела за всеки клиент;

(3) Претегляне и агрегиране на общата RFM-оценка за всеки клиент (като тегла на отделните критерии са използвани R(100) с пет категории, F(10) с пет категории и M(10) с пет категории);

(4) Сортиране на клиентите според изчислената оценка и групирането им на топ клиенти, големи клиенти, средни клиенти, дребни клиенти и неактивни клиенти;

(5) Обвързване на базата данни с изчислените RFM-оценки с данните от проведената промоционална кампания и с разполагаемите социодемографски и поведенчески данни на клиентите;

(6) Подразделяне на базата данни на обучаваща и валидираща извадки;

(7) Балансиране на категорията „реагира на кампании" с тегло 10 в обучаващата извадка (това се налага поради относително малкия относителен дял на реагиращите в базата данни);

(8) Оценяване на няколко алтернативни предиктивни алгоритъма – четири варианта на дърво на решенията с методите CART, CHAID, QUEST и C5.1, наивен бейсов класификатор и логистична регресия;

(9) Съставяне на съчетан модел с трите най-добре представящи се предиктивни модели (по критерия предиктивна точност на целевата категория на зависимата променлива над 90%).

(10) Сравнителен анализ на способностите на съчетания модел с общия RFM анализ.

Общата аналитична и процесна рамка на модела[3] е представена на Фигура 2.

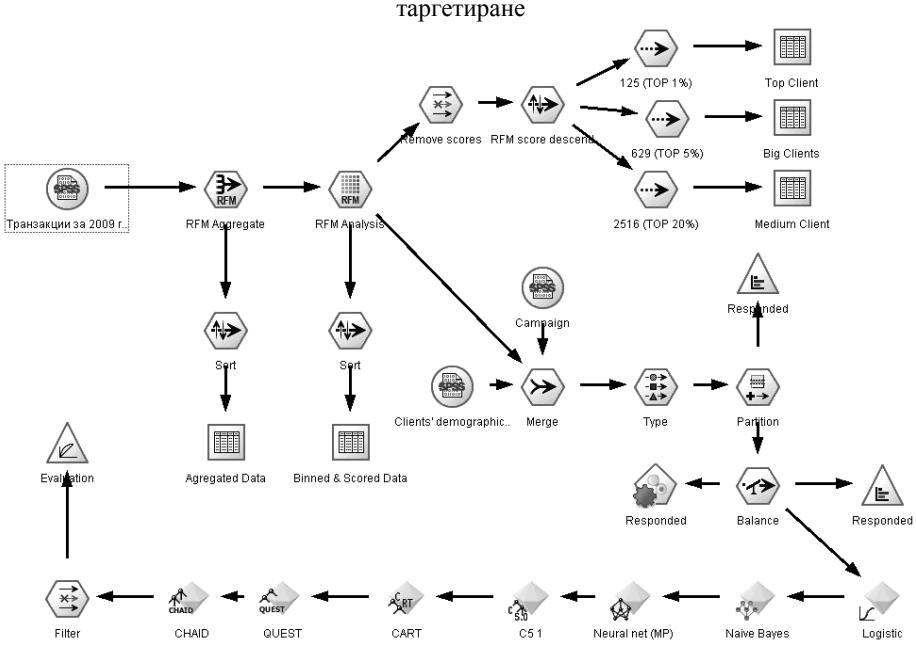
Моделът е интерактивен и позволява лесна адаптация към всякакъв тип клиентски бази данни, организирани по описания начин: транзакции, потребители, кампании. На всеки един от етапите в реализацията му възможно да се извличат клиентски списъци, оптимизирани за поддържането на различен тип маркетингови решения (и в частност – таргетиране). Възможно е например извеждането на поименни списък на клиентите с изчислени оценки по трите ключови критерия на RFM-модела – време от последната покупка, честота на купуване и похарчени средства е само (и то с различни сценарии) за

[3] Техническата реализация и оценяването на модела е извършено с IBMSPSSModeler.

претегляне на RFM-критериите. Аналогично е възможно да се генерират групи клиенти на база на агрегираната им RFM-оценка. Проблемът при този тип „априорно“ сегментиране на клиентите е нестабилната хипотеза между оценка и реакция, което е основната слабост на RFM-модела. По тази причина RFM-оценките на клиентите бяха използвани като база за сравнение при тестването на предиктивните възможности на целия комплекс от описани техники и алгоритми за предиктивно извличане на знания от данни.

Фигура 2

Аналитична процесна рамка на комплексен предиктивен модел за оптимално таргетиране



В това изследване бяха тествани седем алтернативни модела – четири за съставяне на дърво за решенията (CHAID, CART, QUESTи C5.1), две статистически техники (логистична регресия и наивен бейсов класификатор), класификационния алгоритъм на k-тите най-близки съседи и невронна мрежа с многослоен перцептрон и обратно разпространение на грешката. Само един от тези алгоритми (сходно на RFM-модела) се представи с предиктивна точност под 50% (критично равнище, тъй като целевата променлива е дихотомна) – това е методът на k-най-близки съседи. Останалите, издържали „теста“ модели са представени на Таблица 4.

От таблицата е видно, че според критерия „Обща предиктивна точност“, пет от тестваните алгоритми са на практика идентични (>92% правилно класифицирани случаи от валидиращата извадка). Логично е в случая да се търси „най-добре представящ се” модел по други критерии (например информационен индекс). За целта може да се състави лифтова диаграма

(вж.Фигура 3), от която ясно се вижда преимуществото на метода CHAID. Визуалният анализ показва например, че ако бъдат таргетирани с търговско предложение топ 20% от списъка с клиентите, подредени в низходящ ред, на базата на предсказаните от CHAID вероятностни оценки биха откликнали около 42% от тях (при 20% очаквана вероятност при случаен списък). Предиктивната способност на CHAID дори се увеличава при таргетирането на по-висок процент от клиентелата (напр. при таргетиране на топ 30% биха откликнали 55%), но намалява след 40-тия процентил.

Таблица 4

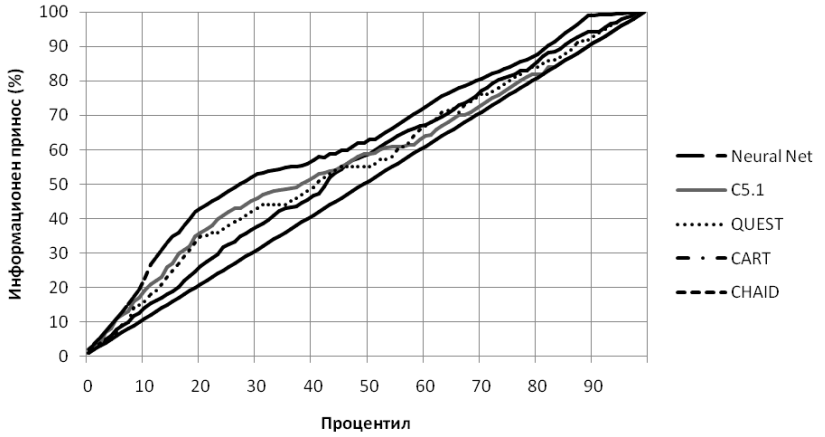
Сравнителни резултати от представянето на предиктивните алгоритми* върху тестовата извадка

Ранг	Аналитична техника	Предиктивна точност (%)
1	Neuralnet (MP)	92.609
2	C5.1	92.609
3	CART	92.609
4	QUEST	92.609
5	CHAID	92.609
6	Logistic regression	61.155
7	BayesianNetwork (NaïveBayes)	61.078

* Резултатите от останалите алгоритми, включени в експеримента (в. т.ч. методът на k-тия най-близък съсед и методът на поддържащите вектори, както и „класическия“ дискриминантен анализ не са рапортувани поради ниската им обща предиктивна точност.

Фигура 3

Диаграма на относителния информационен принос на сравняваните модели

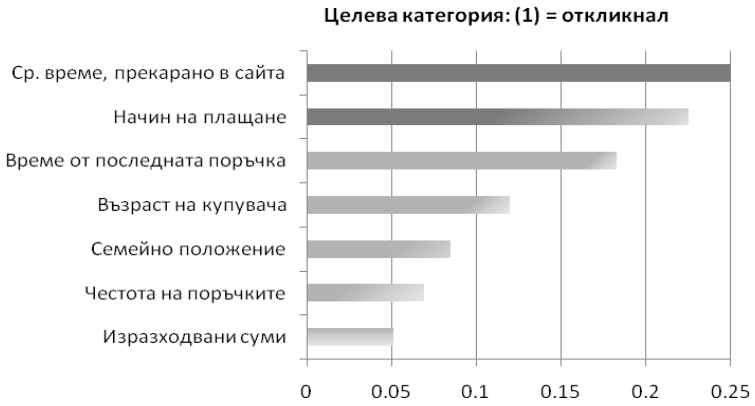


Интересен момент при анализа е оценяването на относителната важност на предикторите (независимите променливи). Всяко една предиктивна техника позволява подобни изводи. В случая с помощта на изкуствената невронна мрежа (многослоен перцептрон с обратно разпространение на грешката) са идентифицирани тези относителни тегла (вж. фиг. 3). Така най-важният

предиктор за предсказване реакцията на клиентите при бъдещи промоционални активности е средното време, прекарано на уеб-сайта на електронния магазин, следвано от начина на плащане, времето от последната поръчка и т.н. Най-нисък принос върху представянето на моделите има похарчената сума пари, от което може да се направи извод, че величината на цените на предлаганите продукти не оказва влияние върху ефекта на практикувания промоционален механизъм.

Фигура 4

Относителна важност на предикторите



Въпреки че в конкретния аналитичен контекст първите пет техники показват еднаква предиктивна способност (респ. еднаква осреднена класификационна грешка), експериментирането със съчетани модели (от ан. ensemble modeling) не е за пренебрегване. Редица автори сочат (Barai & Reich, 1999; van Wezel & Potharst, 2005), че чрез симултантното използване на няколко техники е възможно допълнително да се увеличи предиктивната валидност на общия модел. В конкретното изследване, на базата на петте най-добре представящи се алгоритъма, беше съставен и апробиран подобен съчетан модел, в който допълнително са включени и RFM оценките. В съответствие с очакванията, комбинираният модел се представя почти незабележимо по-добре при оценъчната, но не и при валидиращата (тестова) извадка. Основната изненада и ключова констатация обаче е свързана с предиктивната способност на RFM-модела. В проведения анализ последният се оказва с изключително ниска предиктивна способност, което е поредното доказателство, че този инструмент, използван от мнозина специалисти по директен маркетинг за генериране на „оптимални" списъци за избор на целеви клиенти е далеч от изискването за „оптимална" ефективност. На практика всички експериментирани предиктивни техники за аналитично извличане на знания от експерименталната база данни в изследването го превъзхождат.

*

В разработката е направен опит да се докаже емпирично, че използването на предиктивни модели и алгоритми за извличане на знания от клиентски бази данни за целите на директния маркетинг е по-ефективен подход, в сравнение с класическите RFM-модели. В приложен план е представен и демонстриран подход за априорно предсказване на ефективността на програма за директен маркетинг, който би могъл (1) да инспирира полезни идеи за индустрии, разполагащи с актуализиращи се клиентски бази данни (банков и финансов сектор, модерните вериги магазини, телекомуникации, електронен бизнес, каталожна търговия, ютилити бизнеса) и конкретно на практикуващите директен и интерактивен маркетинг; (2) да разкрие нови възможности за автоматизиране на маркетинговата дейност при планиране на кампании за директен маркетинг; (3) да подпомогне процеса на максимизиране на приходи и минимизиране на разходи на конкретна маркетингова кампания във фирмите; (4) да подпомогне процеса на разработването на програми за предотвратяване на загубата на клиенти; (5) да подобри (оптимизира) планирането на стратегии за привличане и задържане на клиенти, както и други дейности, свързани с оптимизирането на маркетинга и продажбите.

Използвана литература

Backhaus, K., Erichson, B. & Weiber, R. (2011). Fortgeschrittene Multivariate Analysemethoden. (1 Ausg.). Berlin: Springer.

Barai, S. V. & Reich, Y. (1999). Ensemble modelling or selecting the best model: Many could be better than one. – Artificial Inteligence for Engineering, Design, Analysis and Manufacturing, 13, p. 377-386.

Blattberg, R., Kim, B.-D. & Neslin, S. (2008). Database Marketing: Analysing and Managing Customers. NY: Springer.

Bose, I. & Chen, X. (2009). Quantitative models for direct marketing: A review from systems perspective. – European Journal of Operational Research, 195(1), p. 1-16.

Breiman, L., Friedman, R., Olsen, R. & Stone, C. (1984). Classification and Regression Trees. Pacific Grove, CA: Wadsworth.

Collica, R. S. (2007). CRM Segmentation and Clustering Using SAS Enterprise Miner. Cary, NC: SAS Institute Inc.

Curry, J. & Curry, A. (2000). A Customer Marketing Method: How to Implement and Profit from Customer Relationship Management. NY: The Free Press.

Flici, A. (2009, April 20). Business School, Brunel University, London, UK. Retrieved June 22, 2010, from www.brunel.ac.uk: http://www.brunel.ac.uk/329/BBS%20documents/PHD%20Doctoral%20Symposium%2009/Adelflici0630951.pdf.

GfK Bulgaria. (6 3 2009 г.). Спестявания и инвестиции в Централна и Източна Европа 2009. Изтеглено на 12 8 2010 г. от askgfk.bg: http://askgfk.bg/fileadmin/studies/bg/gfk_mood_barometer_charts_2009_bg.pdf.

Hansotia, B. J. & Wang, P. (1997). Analytical challenges in customer acquisition. – Journal of Direct Marketing, 11(2), p. 7-19.

Jonker, J.-J., Franses, P. & Piersma, N. (2002, Feb 21). Evaluating Direct Marketing Campaigns: recent findings and future research topics. Retrieved Oct 21, 2010, from ERIM: http://hdl.handle.net/1765/169.

Kass, G. V. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. Applied Statistics, 29(2), p. 119-127.

Komarek, P. & Moore, A. (2005). Making Logistic Regression A Core Data Mining Tool: A Practical Investigation of Accuracy, Speed, and Simplicity.Robotics Institute. Carnegie Mellon University.

Lackes, R. & Mack, D. (2000). Neuronale Netze in der Unternehmensplanung. München: Vahlen.

Loh, W.-Y. & Shih, Y.-S. (1997). Split selection methods for classification trees. – Statistica Sinaca, 7, p. 815-840.

Lyman, P. & Varian, H. R. (2003). How much information. Retrieved 5 25, 2011, from http://www2.sims.berkeley.edu: http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/.

Magidson, J. (1994). The CHAID approach to segmentation modeling: chi-squared automatic interaction detection. – In: Bagozzi, R. P. Advanced Methods of Marketing Research (p. 118-159). Oxford: Blackwell.

Pappers, D. & Rogers, M. (1998). The One-to-one Future: Building Business Relations One Customer at a Time. NY: Bantam Doubleday Dell.

Peppers, D., Rogers, M. & Dorf, B. (1999, Jan-Feb). Is your Company Ready for One-to-One Marketing?. – Harvard Business review, p. 151-160.

Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. San Francisco: Morgan Kaufmann Publishers.

Rish, I. (2001, 11 2). www.research.ibm.com. Retrieved 5 25, 2011, from An empirical study of the naive Bayes classifier: http://www.research.ibm.com/people/r/rish/papers/RC22230.pdf.

Seni, G. & Elder, J. F. (2010). Ensemble Methods in Data Mining: Improving Accuracy Through Combining. San Rafael, CA: Morgan & Claypool.

Shmueli, G., Patel, N. R. & Bruce, P. C. (2007). Data Mining for Business Inteligence. New Jersey: John Wiley & Sins, Inc.

Suh, S., Lim, S., Hwang, H. & Kim, S. (2004). A prediction model for the purchase probability of anonymous customers to support real time Web marketing: A case study. Expert Systems with Applications, 27(2), p. 245-255.

Teknomo, K. (2009). Tutorial on Decision Tree. Изтеглено на 8 12 2010 г. от http://people.revoledu.com/kardi/tutorial/DecisionTree/.

van den Poel, D., & Buckinx, W. (2005). Predicting online-purchasing behaviour. European Journal of Operational Research, 166(2), p. 557-575.

van der Sheer, H. R. (1998). Quantitative Approaches for Profit Maximization in Direct Marketing (PhD thesis ed.). Groeningen, Netherlands: University of Groeningen.

van Wezel, M. C. & Potharst, R. (2005). Improved Customer Choice Predictions using Ensemble Methods. Erasmus University, Faculty of Economics, Econometric Institute, Rotterdam.

Wasson, C. S. (2006). System Analysis, Design,and Development: Concepts, Prnciples, and Practices. NJ: John Wiley & Sons.

Wilson, J. H. & Keating, B. (2009). Business Forecasting. (6th ed.). NY: McGraw-Hill/Irwin.

Ватева, Д. (2009). Пазар за 6.5 млрд. лв. Изтеглено на 15 3 2010 г. от www.dnevnik.bg: http://www.dnevnik.bg/pazari/2009/10/13/798495_pazar_za_65_mlrd_lv/

Европейска комисия. (2010). Доклад за напредъка на единния европейски пазар на електронни съобщения през 2009 година. Брюксел: Европейска комисия.

Кожухаров, Г. (2010). Томаш Красни, GfK Австрия: До 5 години пазарът на бързооборотни стоки в България ще достигне насищане и ще има фалити. Изтеглено на 4 12 2010 г. от http://www.dnevnik.bg: http://www.dnevnik.bg/pazari/2010/11/24/999320_tomash_ krasni_gfk_avstriia_do_5_godini_pazarut_na/

Смокова, М. (2007). Предиктивно моделиране на пазарния дял с изкуствени невронни мрежи (Том 87). Свищов: Библ. Спопански свят.