Ciljevi knjige.

Moć data mining-a je otkriće za veliki broj kompanija. Data mining jeste izvlačenje informacija iz bitnih podataka koji se generišu svakog momenta u svakom delu našeg života. U svakodnevnom radu sa podacima mogu se pronaći skriveni šabloni i značenja. Ova knjiga je rezultat detaljnih proučavanja podataka i prikazuje naučene lekcije kada se suočavamo sa podacima i kada ih koristimo da učinimo stvari boljim. Postoji mnogo trikova u trgovini koji osiguravaju efektivne rezultate. Statistička analiza koja je uključena u data mining ima karakteristike koje su drugačije od drugih tipova statistike.

Unutar knjige, pominjemo tipične probleme u prodaji i marketingu, kao što su pronalazak 10% korisnika koji će najverovatnije kupiti specijalan proizvod. Sadržaj se fokusira na prodaju i marketing zato što svi imaju osnovno znanje iz tih oblasti. Svi ljudi su obuhvaćeni kao korisnici i mušterije, tako da se svi možemo pronaći u problemima prodaje i marketinga.

Kontekst Data Mining-a

Moderni menadžment se vodi uz pomoć podataka. Kupci i korporativni podaci se prepoynaju kao strateška sredstva. Odluke koje se donose na osnovu objektivnih mera su bolje od odluka koje se donose na osnovu subjektivnih mišljenja koja mogu biti pogrešna i pristrasna. Podaci se prikupljaju iz mnogo ulaznih uređaja. Prikupljeni podaci se moraju analizirati, obraditi, i pretvoriti u informaciju koja obaveštava, upućuje, odgovara ili na neki drugi način pomaže da se podaci razumeju i da olakšaju donošenje odluka.

Ulazni uređaji podrazumevaju kase, dnevnike podataka, revizije skladišta i ERP sisteme. Sposobnost da se izvuku korisna, uglavnom skrivena znanja, iz podataka, postaje svakodnevno sve bitnija u kompetetivnom svetu. Kada se podaci koriste za predviđanje, buduće ponašanje posla postaje manje nesigurno i to može biti samo prednost ('Unapred upozoren, unapred naoružan'). Kao što je prikazano na slici, vredni resursi arhivskih podataka mogu voditi do prediktivnog modela i načina na koji će se prihvatiti novi aplikanti u poslovnoj šemi.

Slika 1.1

Kratki proces data mining-a

Sa tehnološkim napretkom, kompjuterska industrija je bila svedok neverovatnom rastu sektora za hardver i softver. Sofisticirane baze podataka su ohrabrile čuvanje masivnih setova podataka i ovo je otvorilo potrebu za data mining-om u opsegu poslovnih konteksta. Data mining sa svojim korenima u statistici i mašinskom učenju vodi računa o prikupljanju podataka, njihovom opisu, analizi i predikciji. Korisna je za donošenje odluka kada se sve činjenice i podaci ne mogu prikupiti ili su nepoznati. Ljudi se danas, više nego ikad, zanimaju za otkrivanje znanja (inteligenciju) i davanje smisla terabajtima podataka koje čuvaju u bazama, kako bi uvideli važne šablone iz tih podataka pomoću pouzdanih alata i metoda onda kada ljudi više nisu u mogućnosti upravljaju sa tim podacima i analizama u svojim glavama.

Slika 1.2.

Povećanje profita uz pomoć data mining-a

Domen znanja

Koncept domena znanja ćemo često pominjati u narednom tekstu. Domen znanja jesu sve dodatne informacije koje imamo u vezi sa određenom situacijom. Na primer, može postojati rupa u podacima i naš domen znanja će možda biti u mogućnosti da nam kaže kako su u tom periodu procesi prodaje i produkcije bili zaustavljeni. Tako možemo tretirati podatka ne kao nula jer nedostaju, već kao nula iz negog posebnog razloga. Domen znanja uključuje meta podatke. Na primer mi ćemo da nadgledamo prodaju proizvoda i naš glavni interes će biti u broju prodatog proizvoda i njegova cena. Međutim, meta podaci o nivou osoblja u prodajnom centru nam takođe mogu dati informaciju koja će poboljšati razumevanje.

Reči koje bi trebalo zapamtiti

Rezultati analize se mogu referencirati na različite načine. Sam model se može nazivati karta rezultata (scorecard) analize. Svaki kupac ima svon rezultat baziran na karti rezultata koja se implementira. Na primer kupac može imati rezultat za njegov afinitet da kupi šolju kafe i za to će postojati karta rezultata koja će pokazivati na strukturu modela koji predviđa afinitet. Termin karta rezultata dolazi iz ranijih dana kada su modeli bili jednostavniji, i tipično je kupac kreirao svoj rezultat kada je pokazivao određeno ponašanje. Primer ovog modela je RFM metoda segmentacije (RFM - Recency, Frequency and Monetary Value – Skoro, Frekventno i Monetarna vredost) u kojoj se rezultati dodeljuju za kupčev RFM i rezultati se kombinuju kako bi se razlikovao više bitan kupac od manje bitnog kupca.

Povezani koncepti

Veliki broj CRM analiza (Customer Relations Management) je komplementaran informacijama kompanijskih izveštaja i marketinškim kontrolnim tablama. Na primer, marketinške kontrolne table (Marketing Dashboard) tipično sadrže pregled kupovine mušterija u različitim grupama i način na koji se te table razlikuju od tabla iz prethodnih kvartala ili godina. Ovi brojevi mogu biti aktuelni, predviđeni ili kombinacija od oba. Rezultat grupnih rezultata kupaca mogu biti, na primer, kupci koji kupuju na leto ili kupci koji imaju stopu odgovora od 20%. Ovo grupisanje može biti vezano samo za određenu kampanju ili može biti prosek za duži period.

Ključni indikator performansi (KPI) je grupa mera i brojeva koji pomažu da se kontroliše posao, i mogu biti definisani do detalja nivoa kampanje i specijalnih marketinških aktivnosti. Tipični primeri za KPI su stopa klikova, stopa odgovora, stopa oscilacija i cena po narudžbini. Oni su pouzdan način za prezentovanje ukupnih performansi na sažet način, neizostavljajući pri tome važne detalje.

Analitika je generalno ime za analizu podataka i donošenje odluka. Opisna analitika se fokusira na opis karakteristika podataka, dok se prediktivna analitika referira kao modelovanje.

Globalna privlačnost

U poslovnom svetu, metode komunikacije sa mušterijama se konstantno menjaju. U ovoj knjizi, najviše pažnje usmerava ka poslovima koji imaju direktnu komunikaciju sa klijentima. Direktnom komunikacijom se smatra kada kompanija aktivno promoviše svoj proizvod. Promocije mogu biti putem mail-a, brošura, predstavnika prodaje, web stranica i društevnih mreža. Nezavisno od načina komunikacije, kompanije postaju svesnije da njihove ogromne rezerve podataka sadrže bogatstvo informacija. Velike kompanije kao što su supermarketi i maloprodajni giganti su eksploatisali ovaj izvor informacija više godina, a sada i manji biznisi postaju svesni tih mogućnosti. Osim marketinga i oglašavanja, produkcije i finansije imaju benefit od data mining-a. Ovi sektori koriste iste metode i mehanizme kao marketing i oglašavanje, međutim mi imamo tendenciju da koristimo marketinške podatke kako bi ilustrovali metode, zato što je razumljivije i ne zahteva specifične tehničke detalje o proizvodu ili poznavanje samog procesa produkcije. Prodaja je svima bliska zato što na rezultati data mining-a koje iznose velike kompanije utiču na svakoga. Institucije kao što su zdravstvene ustanove ili vlade država takođe pune svoja skladišta podacimai pronalaze uz pomoć data mining-a nove načine da unaprede svoje servise i povećavaju svoju efikasnost.

Pravilno izvlačenje podataka za upotrebu, zahteva naučni pristup i određenu količinu tehničkih sposobnosti. Međutim ljudi koji danas rade u bilo kojoj oblasti postaju sve veštiji kada je u pitanju manipulacija podacima. Tehnike opisane u ovoj knjizi dostupne su svim biznisima, velikim i malim.

Primeri setova podataka korišćenih u knjizi

Iako postoji mnogo setova podataka svi dele zajedničke karakteristike u vidu zahtevanog izlaza i objašnjenog ulaza. U ilustrativne svrhe pre-analitike i analitike u drugom delu knjige su primenjene na tipične setove podataka. Jedan set podataka je iz skladišta za slanje pošiljki. Ono je odabrano zato što je skladište poznat koncept čak iako data mining primenjujete u nekoj drugoj oblasti. Ovaj set podataka uključuje detalje kupovine, informacije o komunikaciji, demografske podatke, i podskup je stvarnog velikog skupa podataka koji se koristi za data mining vežbu. Postoji 50 000 kupaca koji su uzorak iz celokupnog skupa podataka i videćete u narednim koracima kako se skup podataka priprema za efektivni data mining.

Slika 1.3. Primer podatka 50 000 uzoraka kupaca

Inicijalno postoji oko 200 promenljivih, ali one će biti uvećane na način koji je opisan u sledećem tekstu.

Drugi set podataka je iz web analitika. Ovo je osnova za izazov Evropske Mreže za poslovne i industrijske statistike (ENBIS) 2012-te godine. Većina kalkulacija u ovoj knjizi je izneto putem JMP softvera i alatima iz SAS analitičkog softverskog paketa. JMP i SAS su odlični analitički softveri, međutim dostupni su i mnogi drugi softveri. Vodič za odabir softvera dat je u poglavlju 12.

Struktura recepta

Kuvar bi trebalo da ima instrukcije koje se lako prate, i naš cilj je da pokažemo kako se metode data mining-a mogu primeniti u praksi. Metode analitičkog data mining-a se blago razlikuju od statističkih analiza, i te razlike su naglašene u sledećem tekstu, pomoću vodiča za pripremu podatka i za metode.

Postoje standardne analize koje se zahtevaju iznova i iznova, i treći delu knjige se više govori o tome. Recepti su grupisani u četiri dela:

1. Predikcija
2. Analiza unutar klijenata
3. Učenje iz malih setova podataka
4. Ostali recepti

Svaki od ovih generičkih recepata je opisan u potpunosti i u okviru njih postoje modifikacije koje su dodate kao adaptacija postojećeg recepta. Čitava struktua recepta je opisana do detalja u tekstu ispod. Komponente koje su navedene ne moraju biti prisutne u svakom receptu, a adaptacije imaju komponente koje ih razlikuju od generičkih recepata.

*Industrija* – odnosi se na oblast ili sektor aplikacija, npr. Biznisi gde postoji naručivanje putem mail-a, izdavači, online prodavnice, robne kuće ili supermarketi (sa loyality karticama) ili ostali koji koriste direktnu komunikaciju kako bi unapredili posao.

*Oblasti interesovanja* – specifično, npr. Marketing, prodaja ili online promocija.

*Izazov* – ovo može biti npr. Da se nađe pravi broj kupadca da bi se optimizovao povraćaj investicije u marketinšku kampanju.

*Tipična aplikacija* – specifično, npr. Da se pripremi letnja promocija prodaje

*Neophodni podaci* – Svi podaci koji su vitalni za analizu. Podaci moraju da poseduju neke direktne veze ka reakcijama klijenata ili moraju dolaziti direktno od kupaca. (npr. Podaci koji dolaze direktno iz procesa prodaje ili marketinških aktivnosti)

*Populacija* – definisano u skladu su problemom i sa poslovnim informisanjem. Uzmite u obzir da kampanje mogu biti isključivo sezonske i u tom slučaju moramo uzeti populaciju bar jednog ciklusa.

*Ciljana promenljiva* – varijabla od interesa npr. Binarna promenljiva kao što je **kupuje** ili **ne kupuje**, ili može biti količina na nivou metrike kao što je broj ili vrednost prodaje.

*Obavezni ulazni podaci* – ključne promenljive od kojih analiza zavisi

*Neobavezni ulazni podaci* – druge promenljive koje mogu unaprediti modelovanje ali mogu biti teže za pronalaženje i konstruisanje.

*Data mining metode* – često postoji više različitih metoda koje će se korisiti i one se ovde navode.

*Kako da se odradi* – sekcije od pripreme podataka do implementacije daju detalje o tome šta treba da se odradi

*Priprema podataka* – specifične karakteristike za pripremu svakog recepta su date ovde

*Poslovni problemi* – Ovde se mogu navesti promene strategije koje uključuju npr. Kanale prodaje, lokacije, raznolikost ili proizvode. Ovo treba uzeti u obzir kada se analiziraju podaci.

*Transformacija* – npr. Ciljne i/ili ulazne promenljive bi možda trebalo nekako klasifikovati ili pretvoriti u indikativne promenljive. Druge promenljive možda zahtevaju transformaciju kako bi poboljšale asimetriju.

*Marketing baza podataka* – odnosi se na kreiranje seta podataka iz kojih se analiza može izvesti

*Analitika* – sekcije od particionisanja do validacije su postepeni račun analize

*Particionisanje podataka* – Ovo može da obuhvata razmatranje veličine uzorka, razlaganje u grupe ili druge probleme

*Pre-analitika* – Opisuje posao koji bi trebalo odraditi pre analize. Može uključiti projekciju nekih promenljivih, npr. Promenljive koje imaju vrednost 0 ili ako su svi jedna ista vrednost. Sekcija karakteristika se može izvršiti u ovoj fazi.

*Kreiranje modela* – Modeli se prave tako što dobijaju formule koji im najviše odgovaraju ili pravila članstva npr. Analiza klastera

*Evaluacija* – fokusira se na to kako dobro će se izvesti analitički procesi u smislu koliko će nakon što se odrade biti važni biznisu. Takođe uzima u obzir kvalitet modela u pogledu korisnosti za donošenje odluka. Validacija modela je bitan aspekt evaluacije i zbog toga se ove dve stvari posmatraju zajedno.

*Validacija* – osigurava da se rešenje vezuje za poslovni problem. Može uključiti i validaciju lica, što znači da se slične tačke gledišta porede sa rezulatitma modelovanja. Takođe uzima u obzir način na koji se model uklapa sa podacima. Ovo uobičajeno uključuje primenu modela na različite podsetove seta podataka i poređenje dobijenih rezultata.

*Implementacija* – ovde se obraćamo originalnom iskazu recepta, npr. Kako imenujemo i adresiramo precizan broj kupaca i diskutujemo o tome kako se model može primeniti u praksi.

*Saveti* – specifični za određeni recept, mogu uključiti sugestije kako bi se model osvežio.

*Kako prodati menadžmentu* – važan deo koji uključuje tabele i načine na koji rezultate možemo učiniti zanimljivijim i primamljivijim.

Definicija Data Mining-a

Data mining pokriva širok opseg aktivnosti. On teži da obezbedi odgovor na pitanja kao što su:

* Šta postoji u prikupljenim podacima
* Koje vrste šablona se mogu razlikovati u lavirintu podataka
* Kako se svi ovi podaci mogu koristiti za benefit u budućnosti

Populacija i uzorkovanje

U data mining-u setovi podataka mogu biti ogromni – može postojati na milione slučajeva. Različiti tipovi industrije međutim mnogo se razlikuju u broju slučajeva koje prouzrokuju poslovni procesi. Web aplikacije, naprimer, mogu prikupljati podatke od miliona cookie-ja, dok druge aplikacije, kao što su loyality klubovi ili CRM programi mogu imati više ograničenih slučajeva.

Data mining koristi naučne metode istraživanja i primene. Prezentuje se veliki broj podataka koje u jednom slučaju smatramo za čitavu populaciju. Drugim rečima koristimo sve informacije na raspolaganju. U drugom slučaju, naš set podataka može da se smatra kao veliki uzorak. Ako radimo sa manjom količinom podataka (do 10 000 slučeva), onda se možemo odlučiti za rad sa čitavim setom podataka. Ako radimo sa velikim setom podataka, možda ćemo se odlučiti za podset podataka radi lakše manipulacije. Ako se analiza vrši na uzorku, implikacija je da će rezultati biti reprezentativni za čitavu populaciju. Drugim rečima, rezultati analize nad uzorkom su relevantni za čitavu populaciju.

Ovo znači da uzorak mora biti dobar. Dobar uzorak je onaj koji je reprezentativan i nepristrasan. Uzorkovanje je samo po sebi predmet za izučavanje. Kako uglavnom radimo sa velikim populacijama i možemo priuštiti velike uzorke, možemo uzeti proizvoljan uzorak u kome će svi članovi populacije imati jednake šanse da budu izabrani. Vratićemo se kasnije na praktične probleme u vezi uzorkovanja u drugoj sekciji knjige. Takođe možemo podeliti set podataka u nekoliko uzoraka na kojima možemo testirati naše rezultate. Ako imamo mali set podataka, onda možemo ponovo uzorkovati taj set, uzimajući proizvoljne podsetove u okviru istog uzorka. Ovo se naziva bootstrapping. Tada možemo uzeti u obzir način na koji proveravamo da su rezultujući uzorci reprezentativni.

Ponekad, uzimamo u obzir samo deo populacije, za odredjenu analizu, na primer, mi možda bude zainteresovani samo za ponašanje kupaca oko Božića ili u toku letnjih meseci. U ovom slučaju, podset se referiše kao ram uzorka (sampling frame), i samo će se iz ovog podseta na dalje birati uzorci.

Priprema podataka

Priprema podataka za data mining je vitalni korak koji je ponekad zanemaren. Od najranijih godina, učeni smo da su 2 i 2 jednako 4. Brojevi se smatraju konkretnim, opipljivim, čvrstim, neizbežnim alatom koji se koristi da se izmeri sve. Međutim brojevi imaju inherentne varijacije, naprimer, dva proizvoda se mogu prodati na isti dan, ali njihova cena može biti drugačija. Interpretacije napravljene po nominalnoj vrednosti možda nisu tačne. Neki poslovi koriste podatke za pravljenje odluka bez provere da su podaci značajni, bez prethodne transformacije podataka u znanje i kasnije u inteligenciju. ‘Inteligencija’ dolazi od podataka kojima je potvrđena validnost kroz prethodno iskustvo i opisano je iz razmatranje njegovog konteksta.

Nadgledani i nenadgledani metodi

Data mining je proces koji koristi raznolike metode analize podataka da bi otkrio nepoznate, neočekivane i relevantne šablone i veze među podacima koji se mogu koristiti da bi se napravila tačna i precizna predviđanja. Generalno, postoje dve metode analize podataka: nadgledane (supervised) i nenadgledane (unsupervised). Vidi sliku 2.1 i 2.2. U oba slučaja, uzorak posmatranih podataka se zahteva. Ovi podaci se mogu nazvati kao trening uzorak (training sample). Trening uzorak se koristi u data mining-u kako bi se otkrili šabloni u podacima.

Nadgledane analize podataka se koriste za procenu nepoznatih zavisnosti iz poznatih ulazno izlaznih podataka. Ulazne promenljive mogu uključiti kvanititet različitih artikala koje je kupio određeni kupac, datum kada su obavili kupovinu, lokaciju i cenu koju su platili. Izlazne promenljive mogu uključiti indikaciju od toga da li je kupac odgovorio na kampanju prodaje ili ne. Izlazne promenljive su poznate i kao targeti u data mining-u. U nadgledanim okruženjima, uzorak ulaznih promenljivih se prosledjuje sistemu za učenje i naredni izlaz sistema za učenje se poredi sa izlaznim promenljivim uzorka. Drugim rečima, pokušavamo da predvidimo ko će odgovoriti na prodajnu kampanju. Razlika između izlaza sistema za učenje i izlaza uzorka se može smatrati za signal greške. Signali greške (Error signals) se koriste pri prilagođavanju sistema učenja. Ovaj proces se izvršava više puta sa podacima iz uzorka i sistem učenja se prilagođava dok se izlaz ne dođe do minimalnog praga greške. Isti proces se koristi za završno štimovanje tek kupljenog klavira. Fino podešavanje se može izvršiti uz pomoć eksperta ili korišćenjem nekog elektronskog instrumenta. Ekspert obezbeđuje note za trening uzorak, a tek kupljeni klavir je sistem za učenje. Melodija je usavršena kada se vibracije žica klavira podudaraju sa vibracijama u uhu eksperta.

Nenadgledana analiza podataka ne uključuje nikakava fina podešavanja. Algoritmi data mining-a vrše pretragu kroz podatke kako bi otkrili šablone, bez postojanja targeta ili ciljne promenljive. Samo ulazne vrednosti se prezentuju sistemu za učenje bez potrebe za validacijom bilo kog izlaza. Cilj nenadgledanih analiza podataka je da se otkriju prirodne strukture u ulaznim podacima. U biološkim sistemima, percepcija se uči putem nenadgledanih tehnika.

2.1.4 Tehnike otkrivanja znanja

U zavisnosti od karakteristika problema biznisa i dostupnosti ’čistih’ i prigodnih podataka za analizu, analitičar mora napraviti odluku o tome koja će se tehnika otkrivanja znanja koristiti da bi se prikupio najbolji izlaz. Među dostupnim tehnikama su:

* Statističke metode
* Stabla odluke i pravila odluke
* Analize klastera
* Pravila asocijacije
* Veštačka neurološka mreža
* Genetski algoritmi
* Sistemi zamućenog (fuzzy) zaključivanja
* Metodi N-dimenzionalne vizualizacije
* Rasuđivanje na bazi slučajeva

Lista nije iscrpna i redosled ne sugeriše prioritet primena ovih tehnika. Fokus ove knjige će više biti na potrebama biznisa nego na naučne aspekte.

Proces data mining-a

Potreba za data mining-om nastaje iz saznanja da postoji mogućnost da se unapredi način na koji se vodi posao. Prvi logični korak jeste da se razumeju potrebe biznisa i da se identifikuje i da prioritet oblastima kojima je neophodna pažnja. Te oblasti uglavnom podrazumevaju: Povećan broj kupaca koji više ne kupuju, razočaravajuća prodaja ili na pozitivniju stranu, kako pretvoriti potencijalne kupce u kupce ili razviti područja sa mogućnostima. Sva ova pitanja se mogu rešiti posmatranjem relevantnih podataka.

Sva analitika data mining-a bi trebalo da prati definisani proces koji osigurava kvalitet rezultata. Postoje različite dostupne definicije data mining procesa koje su u suštini slične, npr. CRISP-DM i SEMMA.

Generalno, data mining procesi su prikazani na slici 2.3.

Bez obzira na oblasti primene ili definicija konkretnih problema, teoretski procesi se mogu mapirati sledećim koracima.

1. Poslovni problem: Razjašnjenje poslovnog pitanja koje stoji iza problema
2. Podaci: Obezbeđivanje i obrađivanje zahtevanih podataka
3. Modelovanje: Analiza podataka
4. Evaluacija i Validacija kroz faze analize
5. Primena rezultata data mining-a i učenje iz iskustva

Ovi koraci su primer unapređenja biznisa ili ciklusa revizije.

Slika (figure 2.3)

Poslovni zadatak: Razjašnjenje poslovnog pitanja iza ovog problema

Uz sve naučne napore, najvažnije je da se precizno zna šta je potrebno da se uradi. Ovde razmatramo definiciju i specifikaciju problema cilja akcije, planirane primene i period primene.

Vitalni deo problema definicije je da se ima pismeni ili usmeni brifing sa klijentom o planiranim marketniškim akcijama. Ovo bi trebalo da uključi predmete kao što su:

* Planirana ciljna grupa
* Budžet ili planirana produkcija
* Obim i vrsta promocije (broj stranica, sa dobrom prezentacijom, kuponima, popustima..)
* Uključene industrije i odeljenja
* Dobra uključena u promociju
* Scenario prezentacije npr. Baštenska zabava
* Slika koja se prenosi, npr. Agresivna cena, kompetencija brenda ili inovacije.
* Struktura cene

Pretpostavimo da je planirana marketinška aktivnost da se reaktiviraju frekventni kupci koji nisu ništa kupili u protekloj godini. Potrebno je da se pažljivo definiše ciljna grupa, prvenstveno da se utvrdi šta znači frekventno, ko je kupac, da li se uključuju oni koji su kupili pa vratili, da li se u ciljanu grupu uključuju i ljudi koji nisu platili, koji artikli su uključeni, da li postoji najniža/najviša cena, da li je lokacija gde su artikli kupljeni bitna itd. Kako klasifikujemo kupce koji su bili frekventni 10 godina, a onda su prestali da kupuju pre 3 godine, a kako one koji su kupili samo 3 puta, a prestali su da kupuju od skoro? Ova pitanja nisu nepremostiva, ali se moraju dogovoriti sa klijentom, i definicije se moraju upamtiti i sačuvati kao referenca. Glavni razlog je to što ove definicije utiču na ciljnu grupu, a mogu uticati i na model.

Potrebne su sledeće informacije:

* Zajedničke specifikacije ciljane grupe, npr. Aktiviranje prometa, reaktiviranje neaktivnih kupaca ili unakrsna prodaja
* Razjašnjenje mogućih primena, npr. da se proceni potencijalna ciljna grupa
* Posvećenost periodima akcije i primene
* Uzimanje u obzir bilo koji sezonski uticaj
* Uzimanje u obzir slične akcije iz prošlosti

Česte zamke se pojavljuju kada klijent ne popravi sve detalje na vreme za inicijalnu diskusiju ili kada se stvari promene između brifinga i akcije, a analitičar podataka se ne obavesti o tome. Ponekad kolege u marketingu preferiraju da ne budu previše precizni, kako ne bi previše ograničili svoju fleksibinost. Ali bez poznavanja svih detalja, mala je verovatnoća da će se napraviti dobar model. Na primer, klijenti mogu reći da je je akcija brendiranje, i da je ciljana grupa šira i manje orijentisana ka ceni, uključujući ljude koji nisu naročito osetljivi na cenu. Nakon toga, kampanja se menja u kampanju sa agresivnim cenama, a ciljna grupa je manje optimalna zato što nisu naročito osetljivi na cenu. Tako da akcija ne ispunjava očekivanja. Iskusni analitičari podataka znaju da je definicija problema ključni korak u dodavanju vrednosti i odlučivanju da li će akcija biti uspešna. Može zahtevati dosta vremena ali je veoma isplativo.

Podaci: Obezbeđivanje i obrada zahtevanih podataka

Da biste obezbedili neophodne podatke za analizu, morate uzeti u obzir period analize, osnovnu jedinicu interesa i metode za procenu, promenljive i particije podataka za generisanje podataka za učenje/testiranje i za nasumične uzorke.

Popravka perioda analize

Po rasporedu, očekivano je da postoji vremenski razmak između korišćenja modela koji je analitičar podataka proizveo i izvođenja aktivnosti. Na primer, možemo koristiti naš data mining model kako bi odredili mailing listu kupaca koji bi najverovatnije kupili, ali ti kupci dobijaju mail-ove tek kroz nekoliko dana (ili sati ili meseci).

Period analize se sastoji od osnovnog perioda (za unos promenljivih) i target perioda (za targete ili izlazne promenljive). Osnovni period je uvek ispred target perioda, i reflektuje vremenski razmak između izvršenja modela do korišćenja rezultata izvršenog modela.

Iz prošlih aktivnosti, odlučujemo koliki će vremenski razmak biti, nakon toga uključujemo privremenu neusklađenost u podacima koji se modeluju, tako da npr. ulazne promenljive kao što su godine, lokacija, navike nisu kasnije od jednog perioda, recimo perioda broj 10, i ciljne promenljive kao što su kupovna akcija ili aktivnosti oscilacije prihoda jesu iz kasnijeg perioda, recimo perioda 14 ali i ne pre. Vremenski period se razlikujeu zavisnosti od tipa posla, i može se reprezentovati u danima, mesecima, kvartalima ili nekim drugim vremenskim jedinicama poslovnog ciklusa.

Ova privremena neusklađenost promenljivih je najveća razlika u odnosu na druga statistička modelovanja, ali ne prikazuje najveće metodološke izazove. Umesto toga najveći izazovit su pitanja pažljivog razmatranja i precizne pripreme podataka.

Da bi se omogućilo za sezonu, često se period koji je godinu dana pre aplikacionog perioda bira kao target period. Odgovarajući osnovni period je tipično nekoliko nedelja pre target perioda, a kada će on biti se odlučuje pažljivim razmatranjem vremena koje je potrebno da se preuzme aktivnost koja uključuje objavljivanje i štampanje. Na primer, u brifingu nam klijent saopštava da želi Božićnu mail listu, što znači da bi kupci trebalo da prime promocionalne literature do kraja Novembra. Zato što znamo koja je sezona, odlučujemo da nam target period bude od prvog do 31. decembra prošle godine. Pošto je poznato da obrada i dostavljanje traju 4 nedelje, kraj osnovnog perioda je 31. oktobar prošle godine. Tako da u pripremi modela, koristimo ulazne promenljive do 31. oktobra, a ciljne promenljive su od prvog do 31. decembra. U aplikacionom periodu model se koristi sa ulaznim promenljivim trenutne godine do 31 Oktobra kako bi utvrdili kome bi trebalo da se pošalje promociona literatura ove godine. Takođe procenjujemo kupovna ponašanja kupaca od 1. do 31. decembra ove godine. Uzimamo u obzir da smo ostavili čitav Novembar za obrađivanje i prosleđivanje promocionih mail-ova.

Osim privremenih pomeranja podataka, dostupnost podataka takođe mora da se uzme u obzir. Tipine zamke uključuju:

* Kada u primeni jedan od neophodnih setova podataka nije dostupan
* Promene odeljenja i industrija između perioda (uključuje osnovni i target) analiza i perioda primene.

Napomena: Target vrednosti uobičajeno se odnose na drugo vreme u odnosu na ulazne promenljive. Problem vremenske skale je prikazan na slici 2.4.

Slika 2.4. Vremenska skala data mining procesa

Osnovna jedinica interesa

Potrebno je odlučiti šta čini jedan slučaj, tj. osnovnu jedinicu interesa, bilo da je to postojeća osoba, kompanija ili lokacija, faktura itd. Za marketing, jedinica je uobičajeno osoba, zato što osoba koja donosi odluku o kupovini. U drugim industrijama, slučaj može biti dan (npr. slučaj može sadržati dnevnu vrednost podataka), i osnovni i target periodi su verovatno istovremeni. U produkcionom procesu, jedinica mogu biti svi materijali koji čine jedan proizvod, a cilj kvalitet proizvoda. Interpretacija privremene neusklađenosti je ta da je potrebno da postoji period uključenja pre nego što ulazni materijali utiču na izlaz, i ovo treba uzeti u obzir kod modelovanja.

Ciljne promenljive

Ciljna promenljiva je fiksna, prema brifingu. Ponekad se razumna ciljna promenljiva ne može dobiti direktno iz dostupnih podataka već je bolje da se izvede na neki drugi način. Možemo koristiti iznos kupovine ili promet na različitim nivoima, ne u zavisnosti od broja prodatih stavki, već koliko generičkih stavki je prodato, npr. ne koliko roze šolja, već koliko šolja, ili čak koliko keramičkih posuda se prodalo. Ovo se uglavnom odlučuje na brifingu i cilju akcije, ali takođe i od dostupnih podataka i količine postojećih podataka koji se uklapaju u postojeći model . Na primer, ukoliko se prodao mali broj roze šolja, model možda neče biti pouzdan.

Slično, sumiranje za ispitivanje ciljne promenljive može biti na nivou odeljenja, industrije ili polja potrošnje. Ciljna promenljiva mora biti merljiva, precizna, robustna kao i relevantna.

U prediktivnom modelingu, preferiraju se manje varijacije u ciljnoj promenljivoj. Ovo je velika devijacija u odnosu na uobičajeno statističko modelovanje gde se preferira više varijacija. Razlog za ovo je to što se mnogo uticajnih faktora ne mogu kontrolisati, čak iako pokušate najbolje što umete da izjednačite podatke, uvek su prisutni efekti vremena i nema kontrole podataka. Konkurencija i poslovno okruženje takođe utiču na ishod. To znači da je previše preciznosti beznačajno i pogrešno. U prediktivnom modelovanju, binarni i kategorički ciljevi su poprilično prihvatljivi, dok se u drugim statističkim modelovanjima uglavnom preferiraju kontinuirane promenljive. Statistički razlog za izbegavanje binarnih i kategoričkih ciljeva jeste to što zahtevaju više podataka. Sa data mining-om međutim dostupno je mnogo podataka i to ne predstavlja problem.

Ulazne (objašnjavajuće) promenljive

Sve ulazne promeljive su generisane samo za osnovni period. Suptilna ali važna činjenica je da ih treba koristiti u data mining procesu kao da su na kraju osnovnog perioda. Ovo može izazvati problem sa promenljivama koje nisu statičke već su podložne promenama, kao npr. bračni status ili adresa. Osim ukoliko nije moguće da se utvrdi da li su se promenile, ove promenljive bi trebalo koristiti sa obazrivošću, čak iako se one mogu posmatrati kao da su statičke ili sporo promenljive.

Stabilniji, unapređeni modeli se dobijaju klasifikovanjem kontinuiranih promenljivih. Kada su promenljive kao što su iznos prometa ili kupovine klasifikovane, one naglašavaju bolje promene u poslovnom procesu. Na primer, veoma se razlikuje to da li neko potroši 0 evra u poređenju sa nekim ko potroši 1 evro. Matematički, ovi kvantiteti su slični, ali u našoj biznis aplikaciji, svaka kupovina, pa i ona od 1 evra implicira nečiji interes za posao, dok nula znači da tu ne postoji interes. Bez klasifikacije, razlika između male kupovine i nekupovine bi se podcenila. Sa druge strane vage, važnije je da se zna da kupac spada u 10% ljudi koji troše najviše količine, a ne da on troši 2 500 evra umesto 5000. U ovom slučaju, važnost kupca koji troši 5000 evra je jedino ta da on spada u kategoriju visokih potrošača. Ova klasifikacija se može izvesti na mnogo načina koji će biti opisani kasnije.

Modelovanje: Analiza podataka

Očigledno je da postoji mnogo različitih data mining metoda koje su dostupne, i još mnogo njih se razvija u momentu. Jezgro data mining procesa jeste da se kreira dobar model. To znači da on predviđa dobro. Međutim iz razloga što se data mining razvija u dinamičkom i promenljivom okruženju, brzi model za prigodan poslovni problem koji se generiše brzo i razvija precizno i osetljivo može imati veću poslovnu vrednost čak iako predviđa nepreciznije od modela kome treba više vremena.

Data mining alati su relativno jednostavni za upotrebu. Važno je da se obrati pažnja na čitav data mining proces. Ovo uključuje prethodno navedene korake: definicija problema, pažljiva selekcija podataka, izbor promenljivih, kao i provere relevantnosti i preciznosti modela

Postoji mnogo data mining software-a koji su dostupni, koji nude najčešće metode. U zavisnosti od politike kompanije, algoritmi se mogu pisati ili naslediti putem freeware-a. Lične preferencije mogu težiti ka vizuelno orijentisanom pristupu procesu, i on zahteva minimalne programerske veštine.

Dobar data mining software bi trebalo da uključi zvučne alate za pripremu podataka i transformaciju. On bi trebalo da bude usmeren ka dobavljanju razvojnih modela ili skripti, koji bi se lako koristili u različitim okruženjima.

Evaluacija i validacija u toku faze analize

Procena kvaliteta izračunatog modela se može odraditi na 3 načina: koristeći testni uzorak koji ima istu podelu (između targeta = 0 i targeta = 1) kao uzorak za trening, koristeći testni uzorak koji ima različite stratifikacije, i korišćenje testnog uzorka koji ima istu podelu duž čitavog seta podataka. Možemo generisati veliki broj kandidata modela koristeći regresiju, drvo odluke itd. Modeli se mogu razlikovati u pogledu promenljivih koje su uključene u model. Nakon toga upoređujemo modele i ovo je odrađeno primenom svakog modela na testni uzorak i upoređujući dobijene rezultate. Neki data mining software-i omogućavaju automatska poređenja između modela, ili obezbeđuju alat koji upoređuje modele u toku obrade istih uzoraka.

Metode poređenja uključuju konstruisanje grafikona i pripremu konfuzionih matrica. Slika 2.5. prikazuje tipičan grafikon

Slika 2.5. Grafikon podizanja za poređenje modela

Najbolji model zavisi od poslovnog pitanja. Posmatrajte sliku 2.6. koja prikazuje dva modela sa sličnim rezultatima. Ukoliko želimo dobru diskriminaciju najboljih korisnika, biramo model sa tamnijom linijom iz razloga što prvih 20% korisnika imaju veću stopu odgovora. Ukoliko smo zainteresovani za u dobru diskriminaciju polovine ljudi, oba modela su slična. Ukoliko se odlučimo za najlošijih 10%, onda su takođe modeli slični.

Slika 2.6 Grafikon podizanja za poređenje modela

Ponekad se grafikon podizanja mora prikazati na finijoj skali, kako bi se primetile abnormalne regije. Na primer, može postojati nestabilna regija oko 40% i ukoliko je to regija od interesa onda modeli nisu pogodni za upotrebu. Slika 2.7 prikazuje model 1 sa slike 2.5. sa finijim detaljima. Taj model ima tri nestabilne regije u sredini. Međutim ukoliko nama treba najboljih 20% ili najgorih 40% slučajeva, model ostaje dovoljno stabilan za upotrebu.

Slika 2.7 Grafikon podizanja sa finijom skalom

Za dodatno poređenje modela data je konfuziona matrica. Dobar model poseduje sličnu osetljivost i u trening i u test fazi. Na slici 2.8, vrednosti su slične, što je dobro. Male razlike mogu se mogu tolerisati, a model sa velikim razlikama nije poželjan.

Slika 2.8. Konfuziona matrica za poređenje modela

Alternativni alat korišćenjem Excel-a je prikazan na slici 2.9

Slika 2.9. Primer kontrole modela u Excel-u

Ponekad, sposobnost modela da rangira kupce na relevantan način je važnije od statističkog kvaliteta modela. Koristan model je onaj koji daje kredibilnu rang listu korisnika u vidu pridruženih promenljivih.

Još jedan način na koji se mogu validirati modeli, koji je ponuđen od većine data mining software-a jeste unakrsna validacija. Na ovaj način se može proceniti kako će se rezultati analize generisati na nezavisnom setu podataka i uključuje veliki broj različitih metoda. Predlog je da se unakrsna validacija koristi kada je set podataka mali i kada sadrži pouzdane podatke vrhunskog kvaliteta.

Najvažniji deo validacije jeste da se utvrdi da model ili bilo koje drugo data mining rešenje imaju smisla iz poslovnog ugla, tj. da su rezultati kredibilni i korisni za razvoj posla.

Primena data mining rezultata i učenje iz iskustva

Poenta data mining-a jeste da se dobijeni rezultati iskoriste za novu akciju, npr:

* Da se pronađu najbolji kupci, naprimer, za plan distribucije brošura i kataloga
* Da se izračunaju relevantni faktori uticaja koji opisuju ciljnu grupu

Iako je prihvatljivo da se oslonimo na trening period analize kada se traže najbolji kupci, primena rezultata modela na planiran budući period se mora izvesti pažnjivo.

Sve promenljive se moraju transferovati kroz period između analize i vremena primene, i ocene kupca se moraju odlučiti na bazi trenutnog vremena. Na primer, uzmite u obzir promenljive kao što su *godine* ili *životni prihod*. Ukoliko je period između analize i aplikacije godinu dana, u vreme aplikacije moramo transformisati promenljivu *godine*, tako što ćemo kreirati novu promenljivu koja će reprezentovati odgovarajuće godine osobe u toku trening perioda analize, u ovom slučaju tako što ćemo oduzeti jednu godinu od promenljive *godine* ili tako što ćemo preračunavati godine od datuma rođenja ukoliko je datum dostupan. Ocena za tu osobu se računa korišćenjem nove godine, tj. nakon oduzimanja jedne godine. Za primer *životni prihod*, ne možemo koristiti trenutnu vrednost u vreme aplikacije modela već moramo ponovo računati rezimiranjem prihoda do trening perioda analize.

Naglašava se da informacija koja je bila dostupna u trening periodu možda neće biti dostupna u vreme primene modela. Ukoliko je ovaj problem poznat unapred, onda bi promenljive koje su vezane za tu informaciju trebalo izostaviti. Ukoliko je promenljiva vezana za specifičnu osobinu, koja je podložna promeni, onda bi trebalo nju zameniti opštijom promenljivom. Na primer, razmotrite kupovinu žute olovke. Ukoliko znamo da se u periodu aplikacije žute olovke neće prodavati, ili neće biti moderne, onda bi trebalo da se promenljiva *žuta olovka* zameni opštijom promenljivom, npr. *bilo koja olovka*, ili možda promenljivom *bilo koja moderna olovka* bez specifikacije boje. Nova promenljiva se mora kreirati tako da ima vrednost u setu podataka i u toku trening perioda i u toku perioda primene.

Uspeh prediktivnog modela se mora predvideti. Ovo se uglavnom radi poređenjem procena i ishoda. Mora se međutim gledati čitav proces kako bi se utvrdilo da je čitav proces vredeo. Data mining je skuplji proces za kompaniju, i menadžeri će očekivati značajne prednosti koje proizilaze iz troškova. Poboljšanje osnovnih ključnih indikatora performasi moraju biti jasno saopšteni. Lekcije naučene iz procesa bi trebalo da se vraćaju kako bi se osigurao kontinuirani ciklus poboljšanja.

Praktičnosti data mining-a