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

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.

Unsupervised data analysis does not involve any fine tuning.