Handelsrådet

Dataanalys för ökad kundförståelse

Författare:
Ulf Johansson
Malin Sundström
Håkan Sundell
Rikard König
Jenny Balkow

Forskningsrapport 2016:6

Forskningsrapport 2016:6
Dataanalys för ökad kundförståelse,
ingår i Handelsrådets rapportserie.
Rapporten är finansierad av Handelsrådet,
men där forskarna själva är ansvariga
för rapportens innehåll. Rapporten är läst och
godkänd av Handelsrådets vetenskapliga råd.
Publiceringsår 2016.
Grafisk produktion: Fotoskrift AB
Tryck: Typografiska Ateljén AB
www.handelsradet.nu
ISBN: 978-91-86508-35-7

Förord

För handelsföretagen är kundförståelse avgörande, och metoderna för att skaffa denna kunskap blir alltmer analytiska. Djupare analyser av kundbeteende blir vanligare under de närmaste åren, och kvaliteten avgör företagens konkurrenskraft. Dataanalys (”data mining”) är därför för många företag redan en prioriterad aktivitet.

Denna skrift utgör slutrapporten av projektet *Framtidens Business Intelligence*. Projektet har finansierats av Handelsrådet och genomförts på Högskolan i Borås, med start 2013. Projektgruppen har bestått av forskare från datavetenskap och marknadsföring, och vi ser denna flervetenskapliga ansats som nödvändig för att ta sig an handelns komplexa frågeställningar. Lite tillspetsat anser vi att modern handelsforskning inte kan nöja sig med att beakta den digitala aspekten ”från sidan”, utan faktiskt måste involvera datavetenskaplig expertis.

Vi är övertygade om att projektet har genererat ny och värdefull kunskap, och det är vår förhoppning att kunna inspirera svenska handelsföretag till att uppskatta möjligheterna med dataanalys. En konkret målsättning med rapporten är därför att ge handfasta råd till företag som i dag vill införa dataanalys eller förbättra sina analysmetoder.

En av projektets centrala slutsatser är att det inte är förmågan att samla in, lagra och bearbeta stora mängder data som är avgörande, utan kvaliteten på den efterföljande analysen. En viktig konsekvens av ett dylikt skifte, från ”big data” till ”smart data”, blir att dataanalys inte behöver kräva enorma investeringar i hårdvara, datasystem och konsulter. Faktum är att avancerade analyser kan utföras med fritt tillgänglig mjukvara på standardmaskiner. Rent praktiskt kan dataanalys stödja många olika processer, så den verkliga nyckeln blir att kunna identifiera möjligheterna i den egna organisationen.

Forskningsmässigt har kunskapsbidrag levererats inom både datavetenskap och marknadsföring. En majoritet av projektets resultat är tekniska, ofta i form av algoritmutveckling. Lika viktiga resultat är dock hur man säkerställer datakvalitet, hur dataanalysen bör organiseras och integreras i företagen, samt aspekter kring hur kunder uppfattar att deras data och beteende analyseras.

Vi tackar Handelsrådet för att ha möjliggjort projektet.

Borås, augusti 2016

Ulf Johansson, professor i datavetenskap, Malin Sundström, docent i företagsekonomi, Håkan Sundell, docent i datavetenskap, Rikard König, teknologie doktor och Jenny Balkow, ekonomie doktor, samtliga från Högskolan i Borås

Sammanfattning

Projektet *Framtidens Business Intelligence* har studerat utveckling och användning av moderna tekniker, främst dataanalys, inom handeln. En utgångspunkt för hela projektet har varit att handeln är digital, varför forskning kring handeln måste beakta den digitala dimensionen. Specifikt har projektet fokuserat på de möjligheter som fenomenet ”big data” ger aktörer inom handelsområdet. Projektet har därmed i första hand utgått från företagens behov, men även konsumentperspektivet och konsekvenser för samhället har beaktats. Projektet har resulterat i sju övergripande slutsatser och rekommendationer, vilka vi redovisar nedan.

1. Dataanalys är centralt för handelns konkurrenskraft
Dataanalys (”data mining”) har under ett antal år varit ett prioriterat område för företag inom en mängd olika branscher. I handeln har begrepp som ”big data” och ”data analytics” i många fall ersatt traditionella CRM-program och business intelligence som det huvudsakliga verktyget för ökad kundförståelse. Det är därför ingen överraskning att ett flertal av de tunga e-handelstrender för 2016 som lyfts fram av Faring (2015) är direkt kopplade till dataanalys. Farings utgångspunkt är att samtidigt som e-handeln kommer att likriktas, så sker en utveckling mot mer personaliserade sortiment och erbjudanden, vilka är styrda av beteende och köphistorik. Det blir därmed helt avgörande för företagens konkurrenskraft att de förmår höja sin målsättning från att samla in och bearbeta stora mängder data till att skapa en högkvalitativ dataanalys som möjliggör framgångsrik personalisering.

2. Fokus för handeln bör flyttas från ”big data” till ”smart data”
Inte ens de största kundregistren inom svensk handel är ”big data” i den meningen att de kräver speciell infrastruktur eller anpassade algoritmer. En viktig konsekvens är att dataanalys blir tillgängligt även för mindre aktörer, och att det inte kräver gigantiska investeringar i hårdvara, analysverktyg och konsulttjänster. Dataanalys kan också stödja många olika beslutsprocesser – där mindre aktörer kan välja exakt vilka.

3. Att skapa förståelse i organisationen för hur prediktiv modellering kan stödja en mängd centrala uppgifter är ett viktigt första steg för att kunna utnyttja dataanalys
Prediktiv modellering är en generisk uppgift där en algoritm utifrån tillgänglig historisk data skapar en modell som senare används för förutsägelser (prediktioner) eller förklaringar. För handeln kan prediktiv modellering utnyttjas för, bland annat, respons-modellering, churn-prediktion, försäljningsprognoser och kundvärde. Här är det viktigt att inse att det tekniskt är exakt samma metoder och algoritmer som används för alla dessa (och många andra liknande) uppgifter. Det svåra är därmed inte att kunna välja rätt algoritm eller system, utan snarare att ha tillräcklig kunskap om möjligheterna för att kunna identifiera lämpliga användningsområden för prediktiv modellering i den egna verksamheten.

4. De aktörer som i dag använder dataanalys för beslutsstöd bör överväga införandet av conformal prediction för att öka kvaliteten på beslutsunderlagen
Dataanalys och prediktiv modellering utgör beslutsunderlag för många centrala processer inom handeln, specifikt finns möjligheten att via simuleringar uppskatta det ekonomiska utfallet av olika alternativ. Tyvärr är prediktioner i praktiken alltid osäkra, och framför allt kan den osäkerheten inte kvantifieras, vilket gör att beslutsunderlagen egentligen är mycket mer osäkra än vad de kanske framstår. Ramverket conformal prediction löser precis det här problemet då det ger *matematiska garantier* för andelen prediktionsfel som kommer göras.

5. Att säkerställa tillgång till data av hög kvalitet är avgörande för all dataanalys
Data är hårdvaluta – och det är kvaliteten på den som är avgörande, inte förmågan att samla in och bearbeta stora mängder. Insamlad data kommer få ständigt ökande betydelse, men mycket av den data som företagen saknar är information som kunden av olika anledningar kan uppfatta som känslig eller helt enkelt inte vill dela med sig av. Med nya lagar och regler kring hantering av data bör företag därmed acceptera att kunden de facto äger sin egen data. I förlängningen kan detta innebära en marknad där kunder säljer sin personifierade data till olika företag. Företag bör i det läget inte stirra sig blinda på att data måste köpas in av kunderna, det vill säga att det ger upphov till kostnader som inte finns i dag, utan snarare se de möjligheter som så högkvalitativ data skulle innebära. Specifikt ger detta förstås tillgång till individualiserad data som är helt omöjlig att komma åt i dag, samtidigt som det antagligen stärker relationen mellan företaget och kunden.

6. De kampanjvariabler som bör finnas med vid prediktiv modellering ska vara sådana variabler som är samtida och som anger mottagarens inställning till olika media
Det finns inget behov av att lägga till avancerade variabler om hur mottagaren till exempel tittar på tryckt reklam. Det som däremot bör studeras är konsumenters inställning och attityd till budskap, särskilt om de är utformade utifrån ett one-to-one perspektiv eller ett one-to-many perspektiv. Det finns mycket som talar för att det är relevant att lägga in kampanjvariabler i prognosmodellerna som identifierar *typen av påverkan*, det vill säga om mediet är analogt eller digitalt.

7. Beakta att det finns en gräns mellan *cute* och *creepy* när det gäller personifierad reklam
Trots alla i huvudsak positiva reaktioner kring personifierad reklam finns även en motreaktion. När företagen når den gräns där analyser och prediktioner berättar mer om kunden än den själv vet, eller vill att företagen ska veta, uppfattas den som *creepy* snarare än *cute*. Gränsen för vad som uppfattas som *cute* eller *creepy* verkar vara beroende av personliga faktorer hos kunden, men även av såväl den uppfattade som den önskade relationen till företaget. Med rätt motivation anpassad till kundsegmentet (till exempel genom upplevelsebaserade incitament för det yngre segmentet) bör företagen ha möjlighet att förflytta denna gräns. Detta föreslås även vara möjligt genom att låta kunden se den egna nyttan av de analyser som görs av befintlig data.

Begreppslista

Big data: Då mängden data blir så stor att standardsystem inte klarar av att samla in, bearbeta och behandla den inom en rimlig tid, benämns dessa datamängder för *big data*. Att analysera och utnyttja dylika datamängder för att hitta intressanta mönster och utvinna värdefull information kallas *big data analytics*. Exakt vad som utgör *big data* varierar därmed utifrån uppgifterna, infrastrukturen samt den ackumulerade erfarenheten hos organisationen som äger datamängderna. Ofta kopplas big data samman med ”3V” det vill säga *Volume, Velocity,* och *Variety*. Volume innebär att datamängderna är stora, typiskt terabytes eller mer. Velocity handlar om att datamängderna är kontinuerligt växande i situationer där det ständigt tillkommer ny data. Variety, slutligen, beskriver det faktum att formatet på datan, liksom systemen som den hämtas ur, kan variera oerhört. All *big data analytics* kräver exceptionell teknik för att på ett effektivt sätt kunna behandla och analysera datan, inom givna tidsramar.

Conformal prediction: Problemet med alla prediktioner är att de riskerar att vara felaktiga. Specifikt ger de flesta metoder ingen möjlighet att kvantifiera säkerheten i en viss prediktion, alltså hur mycket de går att lita på. Conformal prediction är ett matematiskt ramverk som tillåter prediktioner med matematiska garantier för att de är korrekta. Användaren väljer en acceptabel signifikansnivå, exempelvis fem procent, och andelen prediktionsfel som kommer göras motsvarar då exakt fem procent. Priset man betalar för användandet av comformal prediction är att själva prediktionerna blir multivärda, det vill säga vid regression fås ett intervall och vid klassificering en mängd klasser. Conformal prediction kan användas ”ovanpå” vilka prediktiva modeller som helst, och är därmed extremt generellt.

Dataanalys: Termen dataanalys (vilket ofta används synonymt med uttrycket ”data mining”) är ett paraplybegrepp för en mängd aktiviteter som syftar till att på ett strukturerat sätt finna värdefulla mönster i datamängder. Normalt sett menas automatiska analyser, vilka typiskt genomförs med hjälp av olika algoritmer, men även tekniker som visualisering kan användas vid dataanalys.

Känslighetsanalys: Vid dataanalys skapas ofta modeller för hur en viss egenskap (*den beroende variabeln* eller *målvariabeln*) påverkas av andra variabler, vilka då benämns *oberoende* eller *inputvariabler*. Vid en känslighetsanalys varieras värdena för inputvariablerna på ett kontrollerat sätt och man noterar deras påverkan på målvariabeln. Resultatet av känslighetsanalysen är därmed en förståelse för vilka inputvariabler som har störst påverkan på målvariabelns värde.

Market basket analysis eller associationsregelanalys innebär att från kvittodata hitta regler som beskriver vanliga köpbeteenden med avseende på vilka varor som ofta köps tillsammans.

Maskininlärning: Ett delområde av artificiell intelligens som utgör ett samlingsnamn för en stor mängd tekniker av vilka vissa kan användas för dataanalys. De algoritmer från maskinininlärning som används för dataanalys lär sig från och gör förutsägelser gällande data. Typiskt byggs en modell från en mängd exempel för att möjliggöra datadrivna prognoser eller beslut. Maskininlärning kontrasteras ofta med mer rigida och hypotesdrivna statistiska metoder.

Prediktiv modellering: Prediktiv modellering är en generisk uppgift där en algoritm utifrån tillgänglig historisk data skapar en modell (funktion) som senare används för förutsägelser (prediktioner) eller förklaringar. Modellen beskriver sambandet mellan *målvariabeln* och en mängd *inputvariabler*. Om målvariabeln är kontinuerlig benämns uppgiften *regression*. Fallet där målvariabel är diskret och begränsad till en mängd alternativ kallas på motsvarande sätt *klassificering*. Modellen skapas genom att algoritmen utnyttjar en uppsättning instanser (observationer) av det modellerade sambandet där det korrekta värdet för målvariabeln är känt. Syftet är förstås att modellen senare ska kunna göra prediktioner på nya instanser där målvariabelns värde inte är känt, varför ett grundläggande antagande för all prediktiv modellering är att instanserna som modellen optimerades på verkligen innehåller det underliggande samband som man letar efter.

Smart data är ett populärt men inte helt etablerat begrepp, vilket ofta används som kontrast till big data. Enkelt uttryckt kan man kanske säga att smart data är den del av big data som är relevant för en viss organisation i ett visst tillfälle. När (big) data väljs ut, hanteras och analyseras så att den tillför direkt nytta och värde blir den smart data. En fara med termen big data är ett ensidigt fokus på begreppet ”big”, det vill säga det faktum att datamängderna är stora eller snabbt växande. Diskussionen om big data har därför i för hög utsträckning handlat om förmågan att samla in, lagra och bearbeta stora datamängder på rätt sätt. Naturligtvis kräver analys av verkligt stora datamängder utvecklad infrastruktur och kraftfulla analystekniker, men den typen av datamängder är betydligt mindre vanligt förekommande än man kan tro. Detta innebär sammantaget att det väldigt ofta är andra faktorer än förmågan att processera stora mängder data som avgör en organisations möjlighet att framgångsrikt utföra och utnyttja dataanalys. Specifikt utgår analys av smart data från betydligt mindre och mer hanterbara datamängder, och ofta är även själva analyserna mer fundamentala.

Innehållsförteckning

1 Inledning

Handelns digitalisering innebär så mycket mer än det som den oinsatte brukar sammanfatta till ökad e-handel. Handelns digitalisering utgör en drivkraft för innovation och bidrar till utveckling och bättre lönsamhet samtidigt som den leder till ökat kundvärde genom nya digitala tjänster och bättre produktinformation. Digitalisering innebär också att större och större datamängder samlas in och används på olika sätt. Det digitala handelslandskapet rymmer således många möjligheter som bör studeras och förstås. Vi har valt att studera handelns digitalisering via de stora datamängderna och den efterföljande affärsintelligensen som skapas, där vi fördjupar kunskaperna kring hur stora datamängder kan struktureras, analyseras och användas för att förbättra affärerna.

Detaljhandelsföretag har sedan ett tiotal år bakåt börjat samla in stora datamängder baserat på kundernas affärstransaktioner i de så kallade lojalitetsprogrammen. Data som samlas in bidrar till förståelse av köpögonblicket, kundens identitet, tillfälle och tidpunkt för köp samt vilka produkter som varje kund köper. Med hjälp av data mining kan sedan företagen använda denna kunskap för att agera på olika sätt, till exempel genom att skräddarsy erbjudanden, effektivisera kampanjer och prognosticera volymer. En del menar till och med att det har blivit möjligt att skräddarsy *kunden* med hjälp av dessa verktyg (Coll, 2013). Projektets övergripande syfte har varit att utveckla, anpassa och pröva metoder för dataanalys inom handeln, men vi har även studerat aspekter kring integritet och datakvalitet. Den nya dataskyddsförordningen som EU:s medlemsländer måste inordna sig under, troligen år 2018, kommer att ställa nya krav på hur svenska företag hanterar, lagrar och redovisar personrelaterad data. Men en förändrad lagstiftning kan också innebära möjligheter att affärsutveckla sin verksamhet och förbättra affärsintelligensen. Genom att ta utgångspunkt i lagstiftningens grundantaganden, nämligen att *individen äger datan om sig själv*, finns möjligheter att omdefiniera grunderna för hur morgondagens lojalitetsprogram kan struktureras och fungera.

Projektets övergripande syfte har varit att utveckla, anpassa och pröva metoder för dataanalys inom handeln, men vi har även studerat aspekter kring integritet och datakvalitet.

Data mining utnyttjar och kombinerar tekniker och verktyg från flera olika discipliner. Många av de mest kraftfulla algoritmerna kommer från det område som benämns maskininlärning. I kontexten data mining, så består inlärningen oftast av att skapa en generell modell från en stor mängd exempel. Ett övergripande mål för projektet har varit att utveckla effektiva tekniker för prediktiv klassificering och regression. En nyckel till att skapa robusta och träffsäkra prediktiva modeller är att använda ensembler, det vill säga sammansatta modeller som utför prediktioner genom att kombinera en mängd enklare modeller. Att utveckla nya förbättrade ensembletekniker, specifikt beaktande

de höga krav som ställs vid big data analytics, har därför varit ett övergripande mål för projektet. Ensembler är överlag träffsäkrare än enskilda tekniker, men är samtidigt inte tolkningsbara – vilket kan vara en förutsättning i många situationer där dataanalysen används som beslutsstöd. Tolkningsbara modeller ger också en djupare förståelse för det modellerade sambandet då det möjliggör mänsklig analys av de funna modellerna. Att illustrera och föreslå lösningar på detta problem, ofta kallat the accuracy vs. comprehensibility trade-off, har varit ett annat mål för projektet.

Uppgiften att i nära-realtid kunna analysera eller modellera stora datamängder kräver uppenbarligen synnerligen effektiva tekniker för såväl modelleringen som själva prediktionerna. Det blir dock viktigt att i det sammanhanget undersöka vilka storlekar på datamängderna som egentligen kräver speciella big data – verktyg, och i vilken utsträckning den typen av datamängder och problem verkligen återfinns i handeln.

1.1 Disposition

Då arbetet inom projektet bedrivits som en mängd separata men relaterade studier har vi organiserat rapporten utifrån fem teman för att lättare kunna presentera empirin. De valda temana är följande:

- **Big data:** Vi har i projektet studerat modellering av big data, vilket enligt ansökan var ett av de huvudsakliga målen, ingående. Inom ramen för det här temat presenterar vi tre olika arbeten; en teknisk studie där vi föreslår en algoritm för analys av strömmande data, ett egenutvecklat nytt programverktyg för avancerad analys av kundkorgar samt en undersökning av den speciella formen av big data som innebär att det är själva mängden prognoser som utgör problemet.

- **Smart data:** Ett av de viktigaste övergripande resultatet för hela projektet är insikten att värdefull dataanalys inte kräver ”big data” – och därmed heller inte tekniskt avancerad och kostsam infrastruktur och mjukvara. Inom ramen för detta tema redovisar vi därför hur en stor kunddatabas modelleras med ett publikt verktyg och på en vanlig laptop. Utöver detta visar vi en utökning av en tidigare föreslagen teknik för specialanpassade prediktioner, samt en undersökning av hur valet av *score function* påverkar modellernas egenskaper. Slutligen presenterar vi ytterligare ett egenutvecklat verktyg, vilket trots att det är enkelt att använda, möjliggör mycket avancerade simuleringar och ”what-if analyser”.

- **Kampanjer och personifiering:** Vi har i projektet studerat olika former av kampanjer där vi undersökt om kampanjverktyget (tryckt reklam eller digitala verktyg för reklam) påverkar responsen.

- **Datakvalitet och integritet:** Grunden i insamling av data har länge varit att samla in information med minimal påverkan på kund, men företagen har kommit till en nivå där det inte längre handlar om mer data utan om mer kvalitet i datan. För att kunna

komma förbi den paradox där kunder å ena sidan lämnar ifrån sig mängder av data i sociala medier men å andra sidan är restriktiva med vilken information de lämnar ut till företagen, har vi i denna delstudie i fokusgrupper låtit kunder diskutera hur högre involvering i insamling, analys och användande av företagens data skulle kunna få dem att lämna ifrån sig mer data av hög kvalitet.

- **Säkra prediktioner:** Inte sällan används modeller och prediktioner i nästa läge som beslutsunderlag, exempelvis för kampanjplanering eller personifierade erbjudanden. Naturligtvis vill beslutsfattare då ha möjlighet att kunna jämföra olika alternativ utifrån förväntad vinst. Tyvärr blir detta ofta vanskligt då man inte kan kvantifiera säkerheten i olika prediktioner. Besluten fattas därför ofta på ett underlag där säkerheten inte bara är otillräcklig, utan där osäkerheten i sig är omöjlig att uppskatta. *Conformal prediction*, introducerat av Vovk, Gammerman & Shafer (2005), är ett relativt nytt matematiskt ramverk som motverkar exakt det här problemet. I detta tema redovisas med två exempel grunderna för conformal prediction.

Varje tema inleds med en övergripande sammanfattning och en kort introduktion, medan de viktigaste slutsatserna samlas upp och konkretiseras i rekommendationer, sist i respektive tema.

Bakgrund

2

För att öka läsarens helhetsförståelse av rapporten har vi valt att ge en bakgrund som förhåller sig till ett viktigt begrepp, affärsintelligens. Vi presenterar också vad dataanalys som verktyg betyder samt identifierar de utmaningar som ligger inom ramen för vetenskapen och näringslivet.

2.1 Business Intelligence

Det vetenskapliga målet för detta projekt har varit att utveckla effektiva tekniker för att finna dold men värdefull information i data – det som traditionellt kallas data mining. För att kunna utveckla ett teoretiskt resonemang kring detta har det varit nödvändigt för oss forskare att hitta en gemensam nämnare som kan föra de vetenskapliga disciplinerna dataanalys och marknadsföring samman, där vi valde begreppet affärsintelligens (eng. business intelligence, BI). System för BI kombinerar verksamhetsdata med analytiska verktyg som gör det möjligt för användaren att ta bättre beslut och utforma strategiska och taktiska planer. Således omfattas begreppet BI både av tekniska aspekter (maskin) och beslutsfattande aspekter (människa) (Negash, 2004). Ur ett marknadsförings-perspektiv utmanar begreppet affärsintelligens organisationen, eftersom valet av verktyg styrs av företagets struktur. Det kan till exempel handla om företagets kärnvärderingar, organisering, kontrollsystem och maktförhållande (Audzeyeva & Hudson, 2015). Det innebär att det inte är framgångsrikt att skaffa ett system för affärsintelligens om inte organisationen fullt ut kan implementera affärsintelligensen i löpande verksamhet.

2.2 Dataanalys som verktyg

Traditionell databasmarknadsföring syftar till att identifiera och analysera kundbeteenden utifrån olika datakällor, exempelvis transaktionsdatabaser eller kundregister. Med hjälp av statistiska verktyg nås kunskaper om vilka typer av kunder som beter sig på ett visst sätt, givet en viss åtgärd (en kampanj, en prisförändring eller liknande). Mer moderna analysverktyg har dock oftast sitt ursprung inom området maskininlärning, och tillåter då en uttalat datadriven ansats, på ett helt annat sätt än de mer rigida statistiska teknikerna. Den mest typiska uppgiften inom data mining är prediktiv modellering, vilket är ett samlingsnamn för en stor mängd scenarier där historisk data används för att bygga en modell över något fenomen i syfte att senare använda modellen för förutsägelser och/eller förklaring. För handelns del används sådana modeller ofta för att kunna planera inköpsvolymer och för att kunna ta hänsyn till förändrad efterfråga beroende på vilka kampanjer som genomförs. Försäljningsprognostisering är särskilt viktigt för dagligvaru-branschen som hanterar stora volymer med färskvaror, men på grund av för stora lager, säsongsvariationer och reaproblematik har det också blivit centralt för andra handels-branscher. Än viktigare är dock mer moderna och kunskapsintensiva tillämpningar som churn-predicering och responsmodellering.

Prediktiv modellering handlar alltså om att utifrån tillgänglig data och med hjälp av en algoritm skapa en *modell* (funktion) mellan en uppsättning förklarande variabler och en målvariabel. Algoritmen utnyttjar då en uppsättning *instanser* (observationer) av det modellerade fenomenet, exempelvis försäljning, där man vet det korrekta värdet för målvariabeln. Algoritmen producerar en modell som försöker minimera en *score function* (ett felmått), som beskriver skillnaden mellan modellens *prediktion* (det uppskattade värdet) och målvariabelns faktiska värde, över en mängd instanser. Syftet är dock förstås att modellen senare ska kunna göra prediktioner på nya instanser där målvariabelns värde inte är känt. Därför är ett grundläggande antagande för prediktiv modellering att *träningsmängden*, (instanserna som modellen optimeras för och där målvariabelns korrekta värde är tillgängligt) verkligen innehåller det underliggande samband som man letar efter. Modeller med tillräckligt hög *träffsäkerhet*, alltså modeller som minimerar vald score function tillräckligt bra, antas fånga det underliggande sambandet och kan därmed användas för prediktion. Vid prediktiva problem där målvariabeln har en diskret mängd tillåtna värden, exempelvis {Churn, NoChurn}, kallas uppgiften för *klassifikation* och när målvariabeln är kontinuerlig, till exempel försäljningsvolym, benämns den som *regression*.

En *prediktiv teknik* består mer tekniskt av en *modellrepresentation* samt en *algoritm* som kan optimera modellens parametrar så att vald score function minimeras för träningsmängden. Prediktiva tekniker skiljer sig i den underliggande algoritmen och i modellrepresentationen. Modellrepresentationen avgör vilka samband som teoretiskt kan modelleras samt hur lätt modellen kan tolkas. Tolkningen av en modell kan ge nya insikter om hur de förklarande variablerna påverkar målvariabeln. Tolkningsbarhet kan även vara avgörande för säkerhetskritiska system eller om en prediktion behöver justeras manuellt. Modeller som uppfattas som tolkningsbara, till exempel *beslutsträd* eller *multipel linjär regression*, beskriver enklare samband som är lätta att överblicka och förstå för mänskliga beslutsfattare. I realiteten kan dock de verkliga sambanden vara så pass komplexa att tolkningsbara modeller inte kan beskriva dem med tillfredställande träffsäkerhet. Mer kraftfulla tekniker som *artificiella neuronnät*, eller *random forest* (Breiman, 2001) kan då ge en högre träffsäkerhet, men är samtidigt så pass komplexa att all tolkningsbarhet förloras. Detta dilemma mellan träffsäkerhet och tolkningsbarhet är välkänt och bör alltid tas i beaktande vid val av prediktiv teknik.

Det är dock inte säkert att modellrepresentationen tillåter modellering av det verkliga sambandet, eller att träningsmängden faktiskt innehåller de variabler och instanser som krävs för att hitta det. Därför behöver generaliteten hos prediktiva modeller alltid utvärderas rigoröst. Detta kräver att den färdiga modellen prövas på data som inte använts för att skapa den. Skulle man utvärdera modellen på datan som den skapats med kommer den utvärderingen att vara alltför optimistisk eftersom modellen är anpassad för exakt de instanserna. Standardsättet för att utvärdera hur generell en modell är, det vill säga vilken prediktiv prestanda man kan förvänta sig då den används skarpt, är därför att inte använda all tillgänglig data för modellbyggandet utan spara en del av den för utvärdering, Man ”låtsas” helt enkelt att denna data (kallat *testmängd*), som inte modellen anpassats

för, representerar ny data – och därmed motsvarar den träffsäkerhet som modellen har på testdatan vad man kan förvänta sig vid skarp användning.

Man måste dock vara medveten om att det alltid är vanskligt att försöka uppskatta prediktiv prestanda i förväg. Dessutom säger den typen av analyser ingenting om hur säker man kan vara på en specifik prediktion. Syftet med ramverket conformal prediction är därför att motverka problemet med prognosers osäkerhet. Mer konkret kan man med conformal prediction, under mycket generösa antaganden, välja en acceptabel nivå för prediktionsfelet, och ramverket garanterar sedan matematiskt att det faktiska felet kommer att närma sig denna nivå asymptotiskt.

2.3 Företagsutmaningar

En kartläggning av forskning publicerad i *Journal of Retailing* visar att det saknas svar på frågor som är relaterade till kundinsikt och som bygger på kvantitativa data om kunders beteenden och företagens transaktioner (Grewal & Levy, 2007). Vi menar att många begräsningar i dagens BI-lösningar kan avhjälpas med data mining, en ofta bortglömd del av BI. Det överordnade målet för den generiska aktiviteten data mining är att utnyttja lagrad data, i syfte att finna meningsfull och handlingsbar information (Berry & Linoff, 2000). Med olika data mining tekniker utökas BI-verktyglådan med verktyg för prediktiv och deskriptiv modellering, prognostisering, simulering och optimering. De sammantagna företagsutmaningarna som detta projekt har tagit sig an handlar om följande:

- **Lära sig att balansera effektivitet i dataanalys mot konsumenters eventuella negativa uppfattningar om kampanjer baserade på skräddarsydd kommunikation:** De alltmer kundanpassade erbjudande som dataanalys möjliggör är inte helt okontroversiella från ett kundperspektiv. Trots att de erbjudande som kunder får stämmer bättre med önskemål kan det skapa viss oro när privat information kopplas till kampanjer och säljfrämjande åtgärder. En del av denna oro kan grunda sig i bristen på transparens, det vill säga kunden har själv varken insyn i eller kontroll över den information som dagligen lagras om dem. En praktisk utmaning är därför att veta hur företaget ska balansera effektiviteten från dataanalys mot de eventuella negativa uppfattningar som kampanjer baserade på data mining kan åstadkomma.

- **Avgöra i vilka situationer så kallade big data-lösningar krävs:** Kostnaden för, och komplexiteten hos, system för analys av big data kan för många företag kännas överväldigande. Det är därför en viktig uppgift att identifiera för vilka uppgifter och datamängder den typen av verktyg verkligen är nödvändiga. Extra viktigt är att upptäcka situationer där dataanalys, utförd med betydligt enklare och billigare (”lättviktiga”) lösningar, ändå kan bidra med värdefull kunskap.

- **Identifiera vilka centrala processer dataanalys kan stödja:** Dataanalys är i sin utformning generisk, prediktiv modellering handlar tekniskt om klassificering och

regression. För ett företag är därför ofta utmaningen att förmå överföra dessa generella uppgifter till den egna verksamheten och då konkret identifiera vilka processer som dataanalysen kan stödja.

- **Utveckla kunskap kring dataanalys som aktivitet och om tillgängliga metoder, verktyg och algoritmer:** Många företag samlar i dag in data utan att riktigt veta vad de ska använda den till. Även då dataanalys sker, görs det naivt eller utgår helt från tillgänglig kompetens och programvara. Specifikt finns risken att analysen är en separat del av verksamheten, typiskt på it-avdelningen, istället för att vara naturligt integrerad med marknadsavdelningen. Det är därför en viktig utmaning för företagen att utveckla och sprida kunskap om dataanalys i organisationen. Först när data mining är en naturlig del av verksamheten kan man närma sig dess fulla potential.

Först när data mining är en naturlig del av verksamheten kan man närma sig dess fulla potential.

2.4 Vetenskapliga utmaningar

Många menar att det är dags att bredda handelsforskningen genom att bjuda in fler vetenskapliga discipliner än bara företagsekonomer (Ingene, 2009) och göra forskningen flervetenskaplig. Det finns också tydliga incitament för att vissa branscher inom handeln har ett ännu större behov av en bredare forskningsansats. Det rör sig särskilt om branscher som agerar på en marknad där behovet av innovation och utveckling är tydligt, till exempel inom dagligvaruhandeln. Innovationer har alltid varit efterfrågade men innovationer kommer inte av sig självt – ofta behöver utveckling och idéer födas via specialkunskap (Shankar & Yadav, 2011). Den vetenskapliga utmaningen har varit att använda kunskaper inom dataanalys med ett företagsekonomiskt perspektiv mot handel. En av de vetenskapliga utmaningarna har därför handlat om:

- **Metodfrågor:** Hur kan vi använda tekniska mätningar på respondenter som till exempel eyetracking med kvalitativa intervjuer för att koppla resultaten till tolkningsbara försäljningsprognoser? Hur kan vi fånga respondenters uppfattningar om vilken typ av data som uppfattas som personlig och mycket känslig samtidigt som vi vill ta reda på hur sådan information skulle kunna översättas i ett kommersiellt erbjudande?

Det övergripande vetenskapliga målet för projektets tekniska del har varit att skapa *förbättrade data mining algoritmer*. Mer konkret har följande utmaningar fokuserats:

- **Modellering av snabbt växande datamängder:** Ett vetenskapligt och tekniskt intressant problem för all dataanalys, speciellt då datamängderna blir större, är modellering av strömmande data, det vill säga att kontinuerligt, och i nära-realtid, kunna uppdatera modeller efterhand som mer data blir tillgängligt.

- **Träffsäkra tolkningsbara modeller:** En etablerad sanning inom data mining är att de mest träffsäkra modellerna är ogenomskinliga, det vill säga de kan inte tolkas och analyseras av människor. I många sammanhang är det här oacceptabelt, och då tvingas analytiker i stället välja en mindre kraftfull teknik, som dock genererar tolkningsbara modeller. Att minska konsekvenserna av detta accuracy vs. comprehensibility tradeoff är ett vetenskapligt intressant problem, vilket i praktiken dyker upp i många olika domäner.

- **Effektiv associationsregelanalys (market basket):** En viktig uppgift är att från kvittodata extrahera regler som beskriver vanliga köpbeteende med avseende på vilka varor som köps tillsammans. Det finns många algoritmer, varav flera har blivit standard i större system för dataanalys, som fokuserar på att ta fram dessa vanliga varukombinationer på så kort tid som möjligt, men problemet är erkänt svårt och har kraftigt ökande komplexitet och tidsåtgång som funktion av ökande antal artiklar, varukorgsstorlek och antal transaktioner. Givet detta är det en uppenbar utmaning att anpassa och optimera algoritmer för associationsregelanalys så att de kan användas på realistiska problem.

- **Prediktioner med garantier:** Ett viktigt område, med för närvarande hög forskningsaktivitet, handlar om prediktiv modellering där användaren förutom en prediktion även får ett mått på hur säker denna prediktion är. Det mest tydliga exemplet är ramverket conformal prediction, där varje prediktion kompletteras med en välkalibrerad sannolikhet för att den är korrekt. Ramverket är dock extremt generellt, och i huvudsak presenterat i väldigt matematiska beskrivningar, varför en viktig uppgift blir att konkretisera dessa för att kunna skapa metoder och tekniker som i form av algoritmer kan tillämpas i skarpa analyssituationer.

- **Domänanpassning:** De flesta metoder, tekniker och algoritmer för dataanalys är generiska, vilket är en av deras största fördelar. Dock kräver detta samtidigt ofta att mindre anpassningar sker för att de ska kunna användas optimalt i skarpa projekt. Dylika anpassningar, vilka alltså normalt görs för att lösa ett visst problem i en speciell situation, visar sig förvånansvärt ofta ha bäring utanför den aktuella situationen och domänen. Inom ramen för det här projektet har det därför vetenskapligt värde att identifiera och utvärdera de anpassningar som krävs för handelsdomänen, då dessa lösningar mycket väl kan utgöra generellt intressant kunskap.

2.5 Metod och kommunikation

Inom ramen för projektet har ett stort antal problem och frågor tacklats. De flesta frågeställningarna av teknisk natur har studerats med hjälp av kontrollerade experiment, vilket i det här fallet betyder genom framtagande av nya algoritmer och metoder för dataanalys, vilka sedan har prövats och utvärderats på olika datamängder. I många fall har vi följt praxis inom maskininlärning och data mining genom att använda publika benchmarking-datamängder. Samtidigt har vi hela tiden velat påvisa nyttan

av de föreslagna teknikerna för handelsdomänen, varför vi så ofta som möjligt även genomfört utvärderingen på skarp data från branschen. De frågeställningar som varit av samhällsvetenskaplig karaktär har studerats med hjälp av fokusgrupper, intervjuer, enkätstudier, eyetracking-experiment samt experiment med kroppsskanning och 3D-teknik. Konkreta metodval redovisas i detalj för respektive studie

Projektet har förstås genererat ett antal vetenskapliga publikationer, vilket framgår av den här rapporten. Samtidigt är det viktigt att vara medveten om att i ett så kort projekt (två år) så finns det naturligt då projekttiden löper ut, en hel del resultat som ännu inte granskats av samfundet eller publicerats. Vi har ändå valt att i rapporten här ta med en del ännu icke-publicerade resultat, men vi är förstås tydliga med när så är fallet. Omvänt anser vi att en del av de mest tekniska resultaten inte kan presenteras speciellt väl i den här typen av rapport – och de är antagligen heller inte speciellt relevanta för den typiske läsaren. För de studierna ges därför bara korta sammanfattningar här.

Projektet har engagerat personer från näringslivet i en styrgrupp, där deltagarna har varit med och påverkat de olika studiernas frågeställningar. Preliminära resultat har också presenterats för näringslivet, bland annat i form av workshops och seminarier. Inom ramen för projektet har vi erbjudit en mängd företag möjligheten att besöka SIIRs forskningsmiljö, Handelslabbet, för att se IT-piloter som kopplar samman information från databaser med produkter och konsumenter. ”Skarp data” i form av kassakvitton från en av de största dagligvarubutikerna i Sverige (Ica City) har erhållits och utgör en viktig grund för fortsatt arbete.

Nedan följer en lista över vad vi, ihop med slutsatser och rekommendationer i rapportens sammanfattning, anser utgör de för branschen viktigaste bidragen från projektet Framtidens Business Intelligence.

- Dataanalys är tillgängligt för en stor mängd företag och kräver varken exceptionell know-how eller stora investeringar i IT-infrastruktur, programvaror och konsulter. Av detta följer att det inte är ”big data” utan ”smart data” som är nyckeln till kundinsikt.

- Hänsyn måste tas till typen av media när kampanjanalyser genomförs för att förutsäga inköpsvolymer, där vi särskilt poängterar vikten av att förstå genomslagskraften i digitala medier och verktyg.

- Projektet har nyutvecklat ett program för att skapa associationsregler (market-basket analysis), vilket genom att utnyttja avancerade tekniker för att effektivisera beräkningar kan köras på en standardmaskin.

- Vi har även tagit fram en helt ny typ av datorstöd för avancerade simuleringar och ”what-if analyser”, vilket utnyttjar dataanalys och prediktiv modellering som byggstenar.

Empiri

3

I det här avsnittet redogör vi för de olika studierna i temaform där varje tema inleds med en sammanfattning, redovisning av resultaten samt en analys och efterföljande rekommendationer. I några fall presenteras också metodvalet mer ingående där det är av särskilt intresse.

3.1 Big data

Inom handelsnäringen har de senaste årens ständigt växande datamängder inneburit ett paradigmskifte. Handelsföretag som nu vill utveckla sin BI ytterligare kan inte nöja sig med system där standardmoduler kopplas till de klassiska affärssystemen, utan måste istället söka mer skräddarsydda lösningar, oftast baserade på de absolut senaste teknikerna för big data analytics. Inom detta tema redovisar vi tre studier med utgångspunkten big data.

3.1.1 Sammanfattning

Big data förknippas alltså normalt med problem relaterat till enorma datavolymer, alternativt att datamängderna växer snabbt eller innehåller ostrukturerad data som behöver processas innan den kan analyseras. Big data kan dock lika gärna innebära att man har ett väldigt stort antal mindre datamängder, till exempel försäljningsdata för ett helt sortiment av produkter, vilka alla behöver analyseras inom en viss tid. Om Ica skulle göra en tioveckorsprognos för försäljningen av alla produkter i varje butik, skulle detta kräva att närmare 160 miljoner prediktiva modeller skapades. Vill man, vilket är ett realistiskt scenario, genomföra denna modellering under en natt, måste tillgänglig processorkraft användas så effektivt som möjligt. I detta fall, som vi kallar *horisontell big data*, är själva uppdelningen och distributionen av datamängderna inte ett problem, varför målet blir att uppnå så hög träffsäkerhet som möjligt inom givna tidsramar.
En studie genomförd på 1 001 produkter visar att ensembletekniker är mer robusta och träffsäkra, men att de samtidigt kräver mer än 100 gånger så mycket processorkraft som den näst mest träffsäkra tekniken. Normalt är detta inte ett problem då ensembletekniker är naturligt parallelliserbara, men för horisontell big data, då det inte är datans storlek utan antalalet uppgifter som är problemet, kan de bli alltför beräkningstunga. Vår studie visade att enklare algoritmer ofta kan ge nästan lika hög prestanda som de mer beräkningstunga ensembleteknikerna. Vid en sammanvägning av träffsäkerhet och beräkningstid gav en beslutsträdteknik kallad *RepTree* bäst resultat.

En annan genomförd studie fokuserade på utvecklandet av ett nytt verktyg för associationsregler (market basket analysis), det vill säga regler som uttrycker vilka varor som ofta köps tillsammans. Arbetet utgick från existerande algoritmer, och det övergripande syftet var att genom olika optimeringar tillåta analyser av realistiska datamängder. Studien genomfördes på en datamängd bestående av nästan 60 000

produkter, uppdelade i över 400 produktergrupper. Söker man regler omfattande som mest fem produkter eller produktgrupper ger detta upphov till svindlande 10^{23} regler vilka behöver analyseras. Resultatet blev en optimerad algoritm samt en nyutvecklad prototyp till verktyg, vilken effektivt utnyttjar tillgänglig parallellism i hårdvara. Verktyget kör den underliggande analysen på i storleksordningen någon timme, även på standardmaskiner, varefter användaren interaktivt kan analysera regler med upp till fyra produkter i realtid. Den resulterande programvaran är därmed ett fullt realistiskt verktyg, även för enstaka butiker. En uppenbar användning kan vara att utnyttja den här typen av analyser för att optimera butikens layout utifrån kundernas köphistorik.

Ett alternativt sätt att hantera big data är att inte bygga helt nya prediktiva modeller när ny data erhålls utan att endast uppdatera befintliga modeller. Ofta utgör detta skillnaden mellan att teknikerna tillåter analys i realtid eller inte. Vi har inom projektet utvecklat en variant av en beslutsträdsteknik för detta ändamål. Algoritmen innehåller två nya viktiga aspekter: Först och främst uppdateras träden alltså efterhand, vilket är en betydligt enklare och mindre kostnadskrävande operation än att träna om träden från början varje gång ny data tillförs. Den andra innovativa aspekten innebär att trädens prognoser kompletterades med en konfidens, det vill säga en uppskattning av hur säker man kan vara på varje prediktion. Då ramverket conformal prediktion användes som underlag för beräkning av denna konfidens, kommer den att vara välkalibrerad, det vill säga väldigt väl motsvara den faktiska andelen felaktiga prognoser. Den föreslagna tekniken klarar därmed extrema realtidskrav samtidigt som dess prognoser kommer med starka garantier.

3.1.2 Introduktion

Som en konsekvens av behovet att analysera gigantiska och ofta även snabbt växande datamängder, har forskning om metoder och algoritmer för data mining etablerat sig som ett av de viktigaste områdena inom datavetenskapen. Specifikt, då mängden data blir så stor att standardsystem inte klarar av att samla in, bearbeta och behandla den inom en rimlig tid, benämns dessa datamängder för big data. Att sedan verkligen analysera och utnyttja dylika datamängder för att hitta intressanta mönster och utvinna värdefull information kallas följaktligen big data analytics. Exakt vad som utgör big data varierar förstås främst utifrån de tänkta uppgifterna, men beror även på faktorer som infrastrukturen och den ackumulerade erfarenheten hos företaget som äger datamängderna. I vilket fall som helst så kräver all big data analytics exceptionell teknik, för att på ett effektivt sätt kunna behandla data, inom givna tidsramar. Ett specifikt, och för big data analytics ofta förekommande problem, är det faktum att de datamängder som analyseras även växer snabbt. Analysmetoderna måste därmed hantera detta svåra specialfall, typiskt genom att kunna uppdatera modeller och beslutsunderlag efter hand och i nära-realtid. Därmed finns det idag generella tekniker, som map reduce, vilka kan hantera de flesta big data problem genom att portionera ut arbetet över en mängd datorer.

3.1.3 Redovisning – Horisontell big data

Om till exempel Ica skulle vilja göra en separat försäljningsprognos för varje enskild produkt i alla butiker och för var och en av de kommande tio veckorna skulle detta kräva att nästan 160 miljoner prediktiva modeller behövde byggas (1 300 butiker med i snitt 12 000 produkter och tio prognoser för varje).

I detta fall, som vi kallar *horisontell big data*, är alltså själva uppdelningen och distributionen av datamängderna inte ett problem, utan utmaningen är i stället att använda tillgänglig datakraft (antal processorer) så effektivt som möjligt. Därmed blir effektivitetsmålet att uppnå så hög träffsäkerhet som möjligt inom givna tidsramar.

Det finns idag en uppsjö av prediktiva tekniker som är baserade på olika typer av algoritmer. Varje teknik har sina egna styrkor och svagheter och det är inte givet vilket teknik som är bäst för ett visst problem, framförallt när träffsäkerhet ställs mot beräkningskraft. Det finns viss tidigare forskning med liknande frågeställning, Ahmed et al. (2010) jämförde till exempel åtta olika tekniker för autoregression, det vill säga inga förklarande variabler förutom historiska värden för målvariabeln, på 3003 tidsserier. Studien visade att artificiella neuronnät var den mest träffsäkra tekniken av de som jämfördes, men också den som tog längst tid (1,6 minuter per tidsserie). Vår studie, vilken redovisas nedan, skiljer sig i det att datamängderna är något mindre men å andra sidan har ett trettiotal förklarande variabler, samt att vi inkluderar de två state-of-the-art teknikerna M5P och random forest i utvärderingen. Resultaten som presenteras är en delmängd av en pågående större studie vilken kommer skickas in för publicering under 2016.

Metod

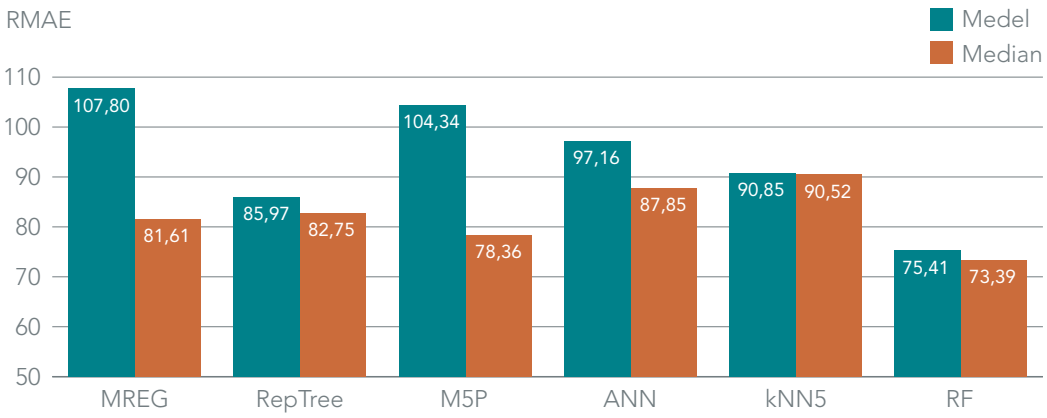
För att undersöka vilken teknik som når högst träffsäkerhet med så lite beräkningskraft som möjligt genomfördes ett experiment med 1001 produkter från Ica-Handlarna AB. Datamängderna valdes ut från frekvent kampanjade produkter inom sortimenten frys och kolonial och innehöll cirka tre års veckovis försäljningshistorik, samt pris och data om genomförda kampanjer, det vill säga kampanjmedia, rabatt, kampanjtyp etcetera. I båda experimenten utvärderas sex olika typer av prediktiva tekniker där de tre första normalt anses ge tolkningsbara modeller. *MREG* står för multipel linjär regression, *RepTree* skapar regressionsträd (beslutsträd med konstanter i löven), och *M5P* är en teknik som kombinerar de två föregående teknikerna, det vill säga beslutsträd med en multipel linjär regression i varje löv. De tre sista mer kraftfulla men icke-tolkningsbara teknikerna som utvärderas är: artificiella neuronnät (ANN), *k*-nearest neighbor med $k = 5$ (kNN) och ensembletekniken random forest med 100 träd (RF). Samtliga tekniker och experiment kördes i Weka (Hall et al., 2009). Utvärderingen görs för prognoser två veckor i framtiden (horisont 2) genom att använda de 25 procent senaste veckorna som testmängd.

Resultatet som redovisas är absolutfelet relativt det fel som en prediktion med medelvärdet skulle ge (RMAE). Därmed är ett lägre värde bättre, och ett värde på 65 procent innebär exempelvis att felet är 35 procent mindre än om prediktionen varit medelvärdet.

Ett högt RMAE behöver därför inte betyda att tekniken gör en dålig prognos, utan kan även innebära att medelvärdet ger en mycket bra prognos (det vill säga produkten uppvisar väldigt små variationer i försäljningsvolym). Därmed ska detta mått främst användas för att jämföra teknikerna mot varandra. RMAE används även för att kunna aggregera resultat över datamängderna trots att försäljningsvolym och volatilitet varierar kraftigt mellan produkterna. Tidsåtgången för träning av modellen och prediktion av samtliga testinstanser mäts i millisekunder. I experimentet användes ingen parallellisering eftersom det normalt sett är både enklare och mer effektivt att köra flera datamängder samtidigt istället för att parallellisera de enskilda teknikerna. Specifikt undviks förstås den overhead som alltid finns när ett problem bryts ner i mindre delar och dessa behöver synkroniseras.

Resultat

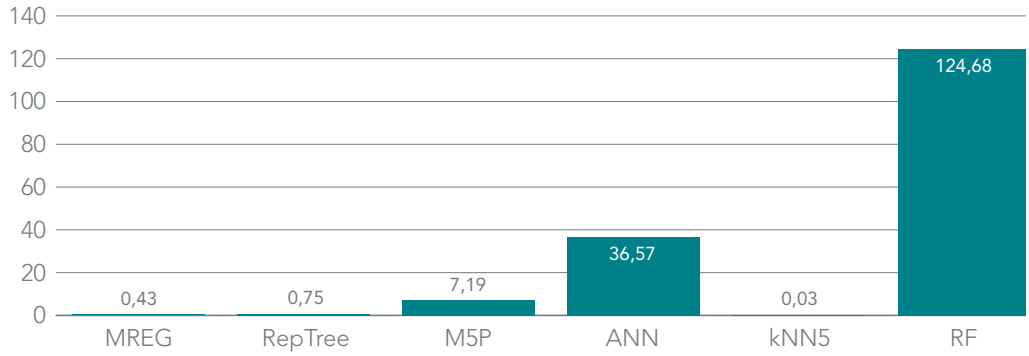
Figur 1 nedan visar medelfelet och medianfelet över 1 001 produkter. Ensembletekniken RF ger betydligt lägre fel jämfört med de andra teknikerna, och har exempelvis 10 procent lägre RMAE (i snitt) jämfört med den näst bästa tekniken RepTree. Ett annat intressant resultat är att både MREG och M5P har relativt höga medelfel, men samtidigt låga medianfel, det vill säga de har producerat ett antal riktigt dåliga prognoser vilka kraftigt ökar medelfelet.



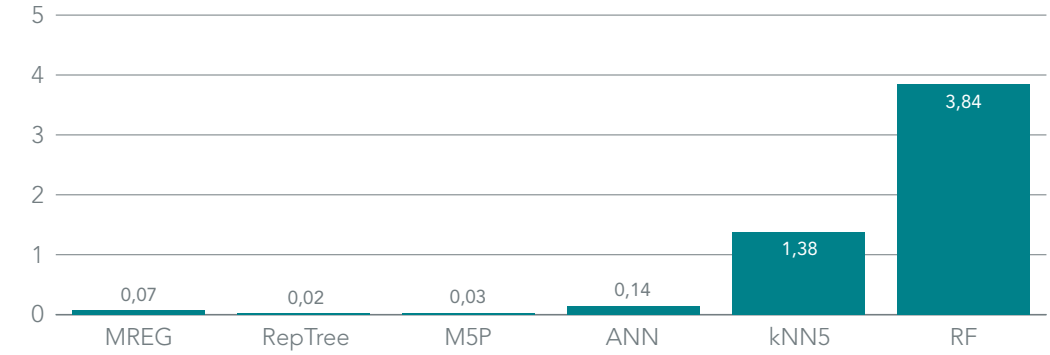
Figur 1. Medel och median RMAE för 1 001 Ica-produkter.

Figur 2 visar den tid det tar att träna modellerna som utvärderas ovan. Här framgår det tydligt att den ökade träffsäkerheten hos RF kommer till priset av längre träningstider. RF tar i snitt 165 gånger längre tid än RepTree och 290 gånger längre tid än MREG. Den absoluta skillnaden är dock inte mer än 124 millisekunder.

Tiden det tar att göra prediktionerna för testmängden (25 procent av instanserna, alltså ungefär 40 stycken per produkt) framgår av figur 3. Resultatet är snarlikt det för träningstiden, men här är dock de absoluta skillnaderna förstås betydligt mindre. Exempelvis tar det i snitt 3,8 millisekunder för random forest att genomföra de ungefär 40 prediktioner som krävs för varje produkt.

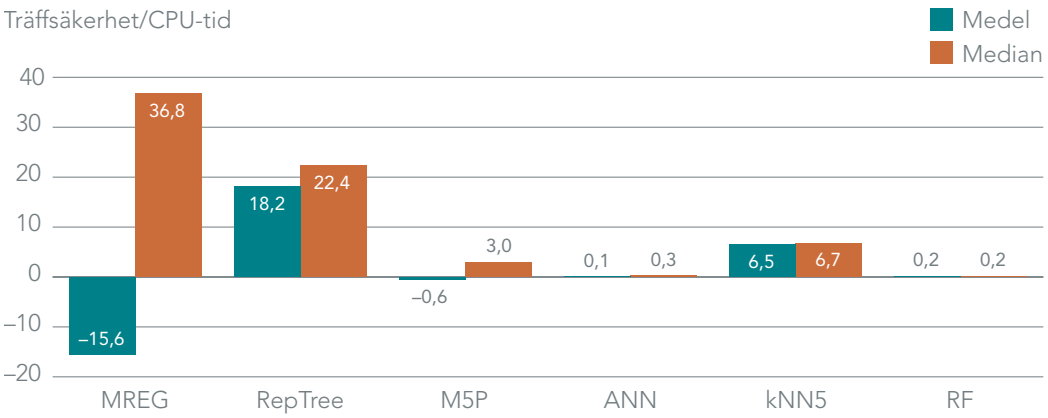


Figur 2. Träningstid i millisekunder.



Figur 3. Prediktionstid i millisekunder.

Till sist visar figur 4 den träffsäkerhet som uppnås (100-RMAE) per millisekund processortid som ett mått på teknikernas effektivitet. När medelvärdet för RMAE används är RepTree den mest effektiva tekniken som ger 18 procent/millisekund förbättring jämfört med att predicera medelvärdet. MREG och M5P får här negativa värden för de är i snitt faktiskt sämre än att predicera medelvärdet. För medianfelet är dock MREG den mest effektiva tekniken med en förbättring på 37 procent/millisekund följd av RepTree med 22 procent/ms. RF som var den mest träffsäkra tekniken får dock endast ett värde på 0,2 procent på grund av betydligt längre träningstid.



Figur 4. Träffsäkerhet per millisekunder.

Analys

Det är inte förvånande att ensembletekniken RF gav bäst prognoser då en stor mängd tidigare studier visat att ensembletekniker är mer träffsäkra och robusta, det vill säga fungerar väl på en stor mängd problem. Robustheten kan här till viss del åskådliggöras som skillnaden mellan medel- och medianfelet. Ett högt medelfel men lågt medianfel indikerar att tekniken inte är robust och att den ibland kan ge mycket höga fel. Ett tydligt exempel på icke-robusta tekniker, i den här studien är därmed MREG och M5P, vilka har nästan 20 procent lägre medianfel än medelfel. En orsak till detta är att både teknikerna ofta extrapolerar, och då predicerar värden mycket högre än tidigare observerade försäljningar. Uppenbarligen fungerar dessa tekniker bra för majoriteten av artiklar, men mycket dåligt för ett fåtal. Dessa tekniker kan dock fortfarande användas om tolkningsbarhet krävs, men behöver då kompletteras med en utvärdering av dess lämplighet för varje datamängd. Ett alternativ kan även vara att manuellt begränsa dess prognoser till högsta tidigare observerade värde.

Den ökade träffsäkerheten hos ensembletekniker kommer dock till priset av längre träningstider då ensemblen kan bestå av hundratals modeller, vilka alla behöver tränas. I normala fall är detta inte ett stort problem då ensembletekniker är naturligt parallelliserbara och varje ingående modell kan tränas separat. Vid horisontell big data hjälper dock inte detta då även icke-parallelliserbara tekniker kan köras parallellt på olika datamängder. Därmed kan ensembleteknikers lämplighet för horisontell big data problem ifrågasättas. I studiens exempel skulle det dock krävas 5 510 timmar processortid för att bygga de 160 miljoner modeller som krävs (124 millisekunder * 160 miljoner produkter = 5 510 timmar processortid 2,6 gigahertz). För att klara av beräkningarna inom fyra timmar, vilket är ett rimligt krav, skulle det krävas över 1 400 processorer och därmed ett större datacenter. Att sedan göra prognoser för de tio kommande veckorna skulle dock inte ta mer än cirka 20 timmar på en processor. I jämförelse skulle det med den näst bästa tekniken RepTree ta 0,75 millisekunder * 160 miljoner produkter = 33 timmar processortid 2,6 gigahertz för att bygga modellerna och mindre än en timme att göra prediktionerna.

Slutsatser

- Träningstiden och träffsäkerhet för olika prediktiva tekniker varierar kraftigt. I denna studie var RepTree den teknik som gav bäst prestanda i förhållande till förbrukad processorkraft.

- Random forest och andra ensembletekniker är robusta och mer träffsäkra än traditionella tekniker, men kräver mer processorkraft vilket gör dem mindre lämpliga för horisontell big data.

3.1.4 Redovisning – Market basket: verktyg för associationsregler

En intressant aspekt av studier i köpbeteende är att analysera vilka varor som brukar köpas tillsammans. Med denna kvantitativa kunskap som plattform kan man sedan gå vidare med kvalitativa studier för att undersöka vad som orsakar beteendet och även

finna tillämpningar av dessa insikter. Mer konkret så strävar vi i den här studien att dels optimera tidigare kända algoritmer för sökning av associationsregler men även försöka hitta metoder för en så effektiv tillämpning av dessa algoritmer som möjligt. Våra viktigaste utgångspunkter har varit:

- Tillämpbart och tillgängligt för små och medelstora företag.
- Kunna användas på konkret transaktionsdata som vi har tillgänglig.
- Minska det upplevda problemet relaterat till big data.

Metod

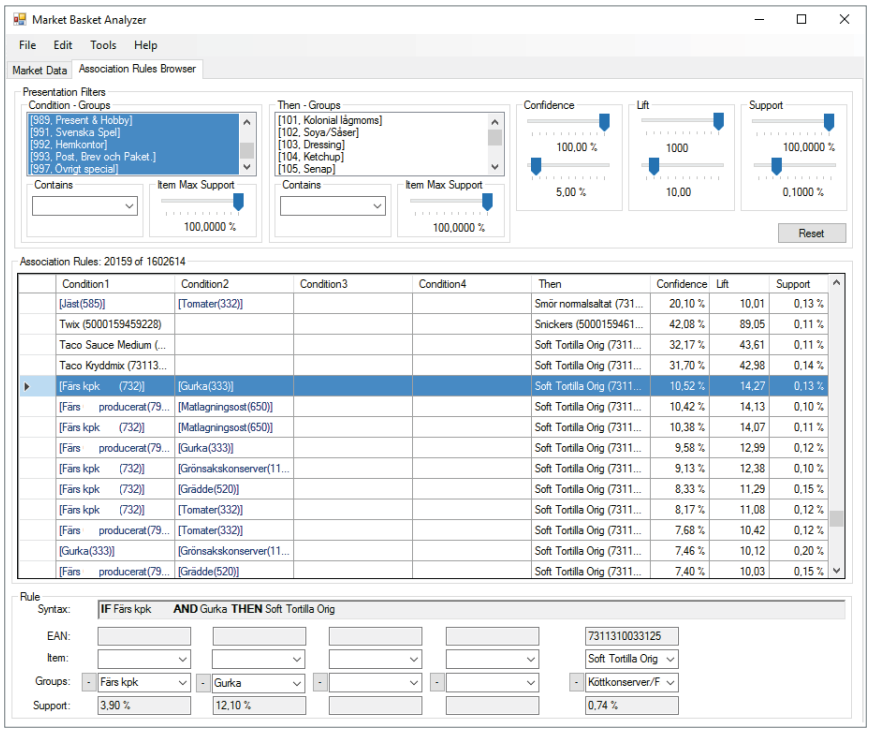
Vår ansats är att utgå från tillgänglig data med relaterad domänkunskap för att därifrån göra analyser och vidareutveckla specifika metoder för att forska fram relevanta resultat. Konkret har vi tillgång till en relativt stor datamängd i form av väl dokumenterad kvittodata från en närliggande stormarknad under sommaren 2014. Som en detaljhandel räknat utgör denna data en relativt stor utmaning, bland annat består produktsortimentet av hela 59 600 artiklar och 438 varugrupper. Det är därmed uppenbart att det går att kombinera produkter på ett stort antal sätt, exempelvis blir det teoretiska antalet möjliga kombinationer av fem varor hela 752 023 303 669 760 000 000 000 (det vill säga biljoner biljoner) stycken. Ska man dessutom leta efter kombinationer av artiklar kombinerat med varugrupper blir antalet möjligheter snabbt ännu större. Utifrån en kortare förstudie av befintliga algoritmer inom forskningsområdet stod det klart att en lämplig metod för kunna gå vidare med analysen av kvittodatan är optimering av existerande algoritmer utifrån våra identifierade domänbehov.

Vi identifierade också i vår förstudie att det är en uppenbar brist på tillgängliga verktyg, varför en viktig kompletterande forskningsmetod blev nyutveckling av verktyg. Detta verktyg bör kunna användas på för domänen tillgänglig hårdvara och därmed kunna dra nytta av aktuella egenskaper i samtida datorarkitektur, i exempelvis lite bättre arbetsstationer – samtidigt som det absolut inte ska kräva större investeringar i kostsam hårdvara. En viktig aspekt blir därmed att försöka utnyttja potentialen i parallellisering av algoritmer och direkt tillämpning i programutveckling.

Det tredje metodiska steget efter algoritm- och programutveckling blir naturligt att studera och mäta på tillämplig data genom experiment och praktiska studier. Inom detta steg ingår även att arbeta med själva kvittodatan så att den effektivt går att använda med utvecklade algoritmer och verktyg.

Då dessa tre metodiska steg inte är helt uttömmande i sig, och dessutom är delvis interberoende, är en naturlig övergripande metod att tillämpa en iterativ process av dessa tre steg, där resulterande insikter från tidiga experiment återkopplas till både algoritm- och programutveckling.

För både utvärdering och hjälp i utvecklandet av algoritm och verktyg finns ett antal viktiga begrepp och mätvärden att beakta. Ett av de mest grundläggande är *stöd* (eng. support). Stöd anger hur starkt underbyggd en regel (eller även en enstaka vara) är, med andra ord hur många transaktioner som regeln (eller varan) har förekommit i. Exempelvis 50 procent i stöd innebär att kombinationen av varor förekommer i hälften av alla varukorgar och 0,01 procent innebär att kombinationen av varor förekommer i 1/10 000 av alla varukorgar.



Figur 5. En funktionell prototyp på ett interaktivt verktyg för djupgående varukorgsanalyser.

Resultat

Förutom nya insikter och kunskap om både domänen och forskningsområdet, är de två viktigaste och konkreta bidraget från studien:

- En optimerad algoritm för varukorgsanalys, både i fråga om prestanda och möjlighet till uttryckande av regler i en högre nivå av abstraktion.

- En nyutvecklad prototyp till verktyg, vilken effektivt utnyttjar tillgänglig parallellism i hårdvara och låter användaren effektivt interagera för en användbar arbetsprocess.

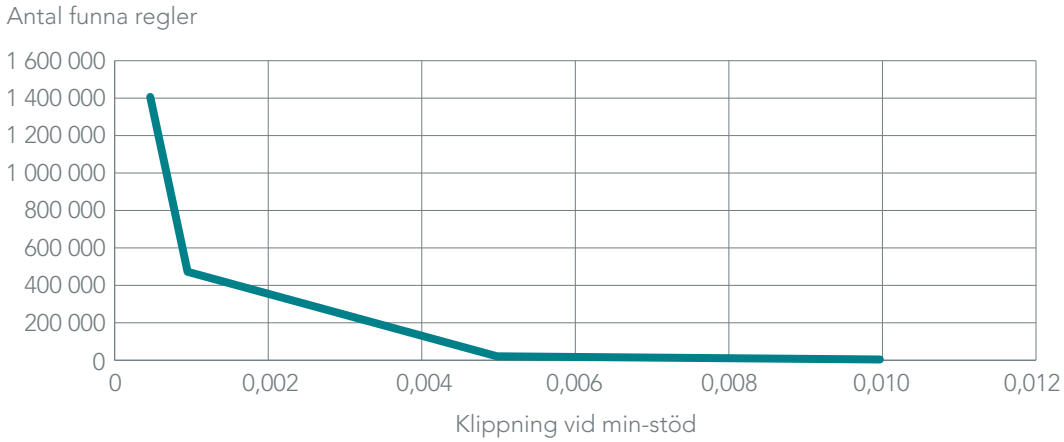
Vi valde att utgå ifrån den vanligaste och mest kända algoritmen för varukorgsanalys, den så kallade Apriori-algoritmen (Agrawal & Srikant, 1994). Då denna algoritm uppenbarligen inte är tillämplig (prestandamässigt) för data i den storleksordning av antalet artiklar (det vill säga tiotusentals) som vi har i vår domän, har optimeringen i huvudsak bestått av att identifiera och slå ihop två delsteg i algoritmen till en helhet för att på så sätt

undvika denna annars extremt höga komplexitet. Vidare har algoritmen utvidgats till att även kunna hitta och skapa regler som består av varukorgskombinationer som spänner över flera nivåer i produkthierarkin, exempelvis mellan både varugrupper och artiklar (observera att varje artikel dessutom alltid ingår i en viss varugrupp).

Vi har i projektet nyutvecklat ett verktyg med utgångspunkt från våra identifierade behov. Av detta verktyg finns en funktionell prototyp, vilken går att använda på vanliga eller kraftigare arbetsstationer för kontorsbruk. En illustrativ bild från verktyget visas i figur 5. En stor del av arbetet i utvecklandet av verktyget har lagts på användbarheten och att verktyget ska vara anpassat för det praktiska arbetet med att utvinna information inom domänen. Därmed har vi utvecklat metoder och funktioner i verktyget för att på bästa och effektivaste sätt kunna hitta regler av intresse, utifrån den enorma mängd kandidatregler som finns att tillgå (med andra ord adresserar vi problemet med att hitta en nål i en höstack). En detaljerad beskrivning av algoritmen finns i (Sundell, König & Johansson, 2015).

Analys

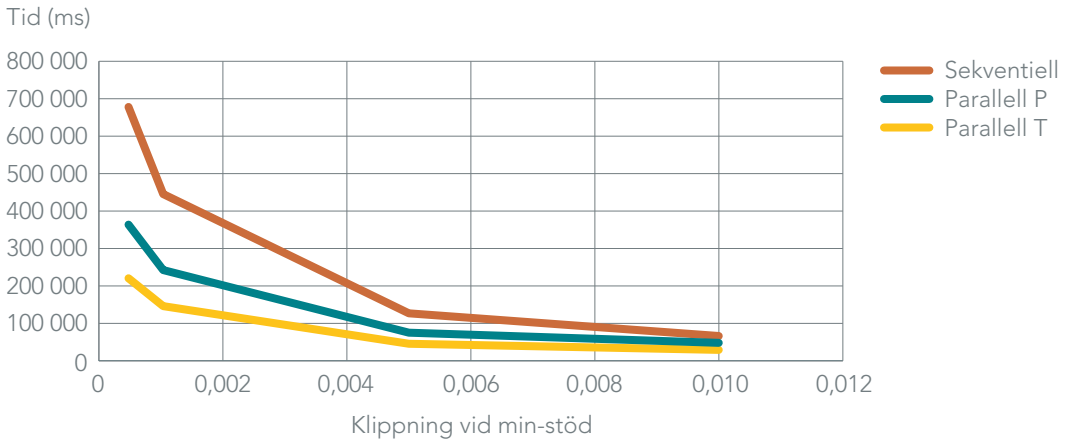
Resultatet har kontinuerligt utvärderats experimentellt, främst inom den iterativa processen med återkoppling till algoritm- och verktygsutvecklingen, men även som en mer omfattande analysdel i slutet av studien. En viktig analys är att studera hur inställningen av olika parametrar påverkar antalet funna regler. I figur 6 visas resultatet från experiment med olika inställningar av parametern som bestämmer minsta stöd av intresse för algoritmen. Förenklat kan man förklara ett minskat min-stöd som en ökning av sökdjupet, det vill säga man letar efter mer och mer sällsynt förekommande varukombinationer i varukorgen. Det är uppenbart att inställningar i sökdjupet har mycket stor påverkan på antalet funna regler, vilka växer exponentiellt med minskande inställning av min-stöd.



Figur 6. Min-stöd vs. antalet funna regler.

Eftersom verktyget är utvecklat för att effektivt kunna utnyttja parallellism i hårdvara, har ett antal experiment för att mäta prestanda utförts. I figur 7 visas hur prestanda för olika implementationer av algoritmen i verktyget beror på variationer i

inställningen av min-stöd. Tydligt är att prestanda blir avsevärt förbättrad på parallella hårdvaruplattformar.



Figur 7. Prestanda vs. nivå av stöd och implementation.

Även mer generella användningsstudier har genomförts. Bland annat har vi undersökt var gränserna för möjligheten att använda verktyget går. Om man applicerar verktyget på hela underlaget av kvittodata bestående av tre månaders transaktioner, och en mycket låg inställning av min-stöd, kan mer än 40 miljoner regler fås fram på cirka en timme. Onekligen är det anmärkningsvärt att domänspecifikt intressanta regler faktiskt kan dölja sig på denna nivå av sökdjup, varför det är viktigt att verktyget klarar av att söka i den stora mängden regler, men även att användaren effektivt kan kombinera verktygets funktionalitet med sin egen domänkunskap.

Slutsatser

Vi har tagit fram ett nytt verktyg för associationsregelinlärning avsedd för användning inom detaljhandeln, med följande egenskaper:

- Hittar regler som spänner över flera nivåer i produkthierarkin.
- Baserat på den väletablerade Apriori-algoritmen.
- Optimerad för praktisk användning i en realistisk miljö.
- Parallelliserad med hjälp av flera strategier inom data parallellism.
- Utvecklat med hjälp av det moderna ramverket Microsoft.net.
- Utformat med ett mångsidigt och användbart grafiskt gränssnitt.

Vi tror att våra resultat kommer att ha intresse i forskarsamhället, men kanske framförallt erbjuda direkta praktiska fördelar för detaljhandeln.

3.1.5 Redovisning – Strömmande data

Ett vetenskapligt och tekniskt intressant problem för all dataanalys, speciellt då datamängderna blir större *(big data analytics)*, är modellering av strömmande data, det vill säga att kontinuerligt kunna uppdatera modeller efterhand som mer data blir tillgängligt. Att finna och anpassa existerande lösningar till dylik ”online-analys” var också ett uttalat mål för projektet. Inom FBI har vi utvecklat och presenterat en variant av beslutsträd för strömmande data (Johansson, Sönströd & Linusson, 2014). Algoritmen innehåller två nya viktiga aspekter: Först och främst uppdateras träden efter hand, vilket är en betydligt enklare och mindre kostnadskrävande operation än att träna om träden från början varje gång ny data tillförs. Detta innebär att den nya tekniken har betydligt större möjligheter än traditionella beslutsträdsalgoritmer att klara av de extrema krav på effektivitet som blir avgörande för att klara de realtidskrav som ställs vid online-analys av växande data. Den andra innovativa aspekten innebär att de resulterande träden, under väldigt generösa antaganden, kan garanteras (matematiskt) uppfylla en av användaren bestämd noggrannhet. Den här mycket attraktiva egenskapen hos en prediktiv modell är en konsekvens av utnyttjande av ramverket conformal prediction, se avsnitt 3.5 – Säkra prediktioner.

I den här studien användes bara publika benchmark-datamängder, och studiens bidrag var uteslutande ”tekniska”, det vill säga riktade till forskare inom maskininlärning och data mining. Vi väljer därför att i denna rapport inte presentera studien i detalj, utan hänvisar den intresserade läsaren till artikeln (Johansson, Sönströd & Linusson, 2014).

3.1.6 Rekommendationer

- Börja alltid med enkla prediktiva tekniker då dessa i praktiken ofta är nästan lika bra som mer avancerade. Om dessa inte ger tillfredställande prestanda kan mer avancerade tekniker utvärderas.

- Ensembletekniker är alltid ett bra val för att nå hög träffsäkerhet, men de är inte alltid lämpade för horisontell big data då de är mer beräkningsintensiva än enklare tekniker.

- Fundera noga över vilken del av analysen som egentligen ställer realtidskrav, om själva modellerandet kan utföras offline öppnas många möjligheter till enklare och effektivare lösningar. Ett tydligt exempel på detta är vårt verktyg för associationsregler där användaren ges möjlighet att utan tidsfördröjning undersöka och analysera en enorm mängd regler skapade från big data – men där detta blir möjligt eftersom det mest tidskrävande momentet bara behöver genomföras en gång.

3.2 Smart data

Datavetenskapliga forskare lyfter nu allt oftare fram en fara med begreppet ”big data” – nämligen att företag tenderar att fokusera på begreppet ”big”, det vill säga det faktum att datamängderna är stora eller snabbt växande. Naturligtvis kräver verkligt stora datamängder exceptionella tekniker, men de företag som verkligen har den typen av

datamängder är betydligt färre än man kan tro. I själva verket är vår uppfattning att inte ens de största svenska e-handelsföretagens kunddatabaser utgör ”big data” i betydelsen att typiska ”big data”-lösningar i form av olika MapReduce verktyg, exempelvis Hadoop, krävs. Detta innebär sammantaget att det är andra faktorer än förmågan att snabbt samla in, lagra och bearbeta stora mängder data som avgör om svenska handelsföretag ska vara framgångsrika i sin dataanalys. Specifikt pratar man då ofta om mer fundamentala analyser, vilket leder till ”smart data”. Därmed blir också dataanalys en aktivitet som är möjlig även för mindre företag eftersom det inte måste utgå från gigantiska investeringar i hårdvara, analysverktyg och konsulttjänster. Temat innehåller fyra olika studier.

3.2.1 Sammanfattning

I detta tema har vi utforskat var gränsen för big data går och sett att det antagligen inte finns något svenskt e-handelsföretag som verkligen har big data problem, i alla fall inte så länge datan som analyseras utgörs av kunddatabasen eller liknande. I en studie som gjordes i samarbete med en ledande e-handlare, med syfte att predicera churn, analyserades nästan 250 000 kunder beskrivna med över 250 variabler. Trots att denna datamängd vid första anblick kan tyckas stor, tog analysen inte längre tid än några minuter med hjälp av ett open-source verktyg och en standard laptop. Eventuella konkurrensfördelar från dataanalys kommer därför inte från vem som kan hantera mest data utan vem som kan utnyttja tillgänglig data på bästa sätt.

Ett sätt att öka kvaliteten på sina prognoser är att säkerställa att man faktiskt minimerar rätt felmått. Det låter kanske självklart att en prediktiv teknik ska minimera prediktionsfelet men felet går att beräkna på många olika sätt. Ett mått kan ge större vikt till stora fel och ett annat kan tillåta några enstaka stora fel om det blir mindre fel i övrigt. Självklart får valet av mått konsekvenser, men då prediktiva tekniker oftast implicit bestämmer vilket mått som minimeras, glöms denna möjlighet bort. Vi har därför presenterat och utvärderat en teknik som kan optimera godtyckligt felmått, och utvärderingen syns stora skillnader i hur modeller som optimerats för olika felmått predicerade samma instanser.

Ett annat sätt att förbättra sina prediktioner på är att verkligen utnyttja all tillgänglig data. Traditionellt tränas en modell på en träningsmängd bestående av de instanser för vilka målvariabeln är känd. Vid prediktionsögonblicket har företag dock ofta tillgång till mer data då de själva vet vad de tänker göra i framtiden, till exempel vilka kampanjer som kommer genomföras med stöd av den planerade dataanalysen. Vi har tidigare föreslagit en teknik kallad *orcale coaching* för detta ändamål. Här har vi vidareutvecklat tekniken och anpassat den för regression, och resultaten visar entydigt att den föreslagna metoden förmår utnyttja den extra tillgängliga informationen och därmed skapa mer träffsäkra modeller.

Till sist har vi även, som komplement till den rena forskningen, skapat en plattformsoberoende lättviktsapplikation med en motor för dataanalys. Motorn stödjer prediktion, simulering av alternativa scenarion och känslighetsanalys. Applikationen är ett sätt att

exponera vår forskning, men utgör också ett konkret exempel på de komplexa uppgifter som kan tacklas även med ett mindre fristående verktyg för dataanalys. Applikationen är tillgänglig för utvärdering från www.tiplersoftware.com.

3.2.2 Introduktion

Dataanalys har på senare år blivit allt mer förknippad med big data, det vill säga datamängder som är så pass stora eller snabbt växande att de kräver stora dedikerade serverar för bearbetning och analys. Ledande leverantörer som SAS Instruments, IBM och Oracle har helhetslösningar vilka kan integreras med befintliga affärssystem. Naturligtvis är dessa lösningar avancerade, flexibla och kraftfulla men kostar därefter. Kostnader för licenser, konsulter och hårdvara brukar ofta innebära en avskräckande prislapp, speciellt för mindre och medelstora företag.

Prediktiv analys behöver dock inte vara dyrt eller omständligt. En scoring modell för ett lojalitetsprogram behöver till exempel inte uppdateras online utan kan uppdateras årligen. Om analyserna inte behöver göras online måste inte heller lösningen integreras med befintliga system. Tills sist är det extremt få företag som verkligen har big data vilket i kombination med den snabbt ökande prestandan på dagens arbetsstationer, innebär att så gott som alla analyser kan göras på befintlig hårdvara. Det finns alltså alternativ till de stora helhetslösningarna och det behöver därmed varken bli dyrt eller komplicerat att skaffa sig konkurrensfördelar med hjälp av dataanalys. Följande avsnitt redovisar resultat som visar hur vanliga arbetsstationer och fristående applikationer kan användas för avancerade analyser och att inte ens databaser med hundratusentals kunder behöver betraktas som big data.

3.2.3 Redovisning – Predicering av churn

Att med hjälp av prediktiv modellering förutsäga vilka kunder som kommer lämna (”churn”) har alltid varit en viktig och typisk uppgift för dataanalys. Liksom all prediktiv modellering används historisk data, vilken beskriver både kunder som har lämnat och inte, för att skapa modellen. Därefter kan modellen användas på nya kundprofiler – vilka inte utnyttjades för byggandet av modellen – för att förutsäga om de kommer lämna eller inte. Det är viktigt att inse att även om churn-modellering är betydelsefullt i sig, så delar det också många egenskaper med andra centrala uppgifter, exempelvis responsmodellering. Framförallt är det rimligt att anta att storleken på de datamängder som analyseras i andra scenarier inom handeln sällan är större, helt enkelt eftersom churn-modellering typiskt appliceras på hela kunddatabasen.

Studien använder skarp data från en ledande e-handelsaktör och genomförs med hjälp av fritt tillgänglig mjukvara på en standardmaskin. Ett uttalat syfte med denna fallstudie är därför att visa att avancerad dataanalys inte nödvändigtvis kräver speciell och kostsam hårdvara eller dyr och komplicerad programvara. Ett viktigt resultat är därmed att visa på möjligheten att utnyttja data mining för alla de handelsföretag som i dag har insamlad data men inte analyserar den på grund av bristande kompetens eller av kostnadsskäl.

Metod

Datamängden består av mer än 250 000 kunder, uppdelade i en träningsmängd (cirka 100 000 kunder) och en testmängd (cirka 150 000 kunder). Målvariabeln är huruvida en viss kund ska lämna eller inte, vilket i den här studien definieras som att hen inte har gjort något nytt köp inom ett år från senaste köp. Varje kund beskrivs med 276 attribut. Vi är inte tillåtna att ge en detaljerad beskrivning av attributen, men de inkluderar statistik som antalet gjorda köp, antalet besök i e-handelsbutiken och om kunden öppnat mail innehållande erbjudanden från företaget.

Från tidigare analys är företaget väl medvetet om att en mycket viktig variabel för att förutsäga churn är antalet tidigare inköp. I själva verket är det extremt stor skillnad på andelen av kunder som lämnar om de genomfört bara ett köp (ungefär 80 procent), jämfört med till exempel fyra eller fler köp (strax över 30 procent). När datamängden innehåller ett så tydligt mönster kommer det med nödvändighet att upptäckas vid modelleringen, oavsett vilken teknik som används. Det finns dock i alla fall minst två problem med att direkt modellera datamängder med redan kända mönster: algoritmerna kommer att lägga kraft på att upptäcka saker vi redan vet, och det är en uppenbar risk att mer intressanta upptäckter missas. Om vi exempelvis predicerar att alla kunder med bara en order kommer lämna, så har den modellen en träffsäkerhet på 80 procent. Eftersom de flesta modelleringsalgoritmer innehåller element som strävar mot enklare modeller, är det inte säkert att den delen av kunddatabasen (kunder med bara ett köp) skulle analyseras ytterligare, det vill säga modellen skulle för det segmentet vara mer eller mindre värdelös. Utifrån detta och liknande resonemang är den generella rekommendationen att i situationer när man vet något om det samband som ska modelleras, så bör man på något sätt ta hänsyn till det i stället för att låta algoritmerna lära sig det redan uppenbara. I det aktuella fallet är det rimligt, både utifrån de tekniska aspekterna och affärslogiken, att i alla fall utvärdera effekten av att först manuellt dela in datamängden utifrån hur många inköp kunderna gjort, och sedan analysera dessa delar separat. I experimenten prövas därför både en dylik manuell uppdelning och att direkt modellera hela datamängden.

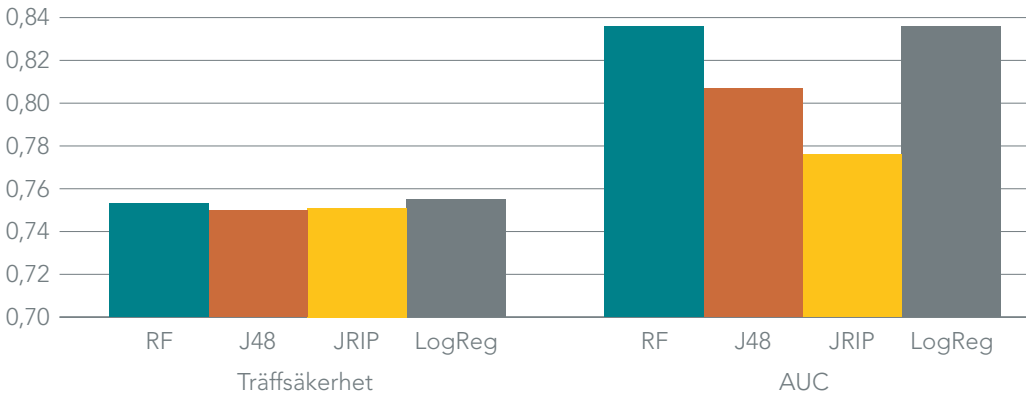
Även om träffsäkerhet, mätt som andelen korrekta prediktioner, är det mest naturliga måttet vid klassificering, så kan det vara för trubbigt i många situationer, framförallt om klasserna är olika stora. I den aktuella studien är datamängden som sådan väl balanserad med 51 procent churn, medan grupper indelande efter antalet order blir tydligt obalanserade. Utifrån detta inkluderar vi även AUC (”area under the ROC”) som mått. Vid beräkning av AUC rangordnas först alla instanser utifrån den predicerade sannolikheten att de tillhör den positiva klassen (churn) varefter AUC mäter, enkelt uttryckt, sannolikheten för att en instans tillhörande den negativa klassen rankas före en instans tillhörande den positiva klassen. AUC är därmed ett mer informerat mått då det inte bara tar hänsyn till den faktiska klassificeringen (churn eller inte) utan modellens sannolikhetsuppskattningar. AUC blir naturligtvis extra viktigt i analyser där syftet är just att rangordna kunderna, exempelvis då mottagarna för en viss kampanj ska utses, och man väljer kunder ”från toppen”, det vill säga de modeller bedömer har störst chans att svara positivt.

I studien användes det publika verktyget Weka (Hall et al., 2009) och alla körningar genomfördes på en standard laptop (Intel i7 4710MQ CPU med 16 GB RAM). Ett uttalat syfte är därmed, enligt tidigare, att påvisa möjligheten till avancerad analys av relativt stora datamängder utan tillgång till anpassad hårdvara och dyrbara system. I experimenten utvärderades totalt fyra olika algoritmer med väldigt olika egenskaper. Random forest (Breiman, 2005) är en ensembleteknik, det vill säga de producerade modellerna är inte tolkningsbara, men erfarenhetsmässigt brukar Random forest ofta vara bland de algoritmer som skapar mest träffsäkra modeller, oavsett domän och data. I dessa experiment består en Random forest av 100 träd, vilket är default-inställning i Weka. JRip är en implementering av algoritmen RIPPER (Cohen, 1995) vilken skapar ordnade regelmängder och J48 är Wekas variant av beslutsträdsalgoritmen C4.5 (Quinlan, 1993). JRip och J48 ger därför båda tolkningsbara modeller som gör det möjligt för en beslutsfattare att manuellt analysera de funna sambanden. Den fjärde algoritmen som prövas är logistisk regression, en enkel teknik som dock ofta ger goda resultat för den här typen av modellering. Modeller skapade med logistisk regression är i princip möjliga att tolka, men långt ifrån lika tydliga som regelmängder och beslutsträd.

Innan den faktiska modelleringen genomfördes en automatisk attributselektion i Weka. Detta är ett standardsteg i prediktiv modellering av större datamängder, där attribut som inte bidrar elimineras, det vill säga syftet är inte bara att minska körningstider utan att förbättra träffsäkerhet och tolkningsbarhet. Rent konkret mäts korrelationen mellan alla olika attribut och de attribut som inte korrelerar med målvariabeln, alternativt har stark korrelation med andra inputattribut, tas bort. I denna studie resulterade attributselektionen i att antalet attribut minskades från 276 till färre än 50.

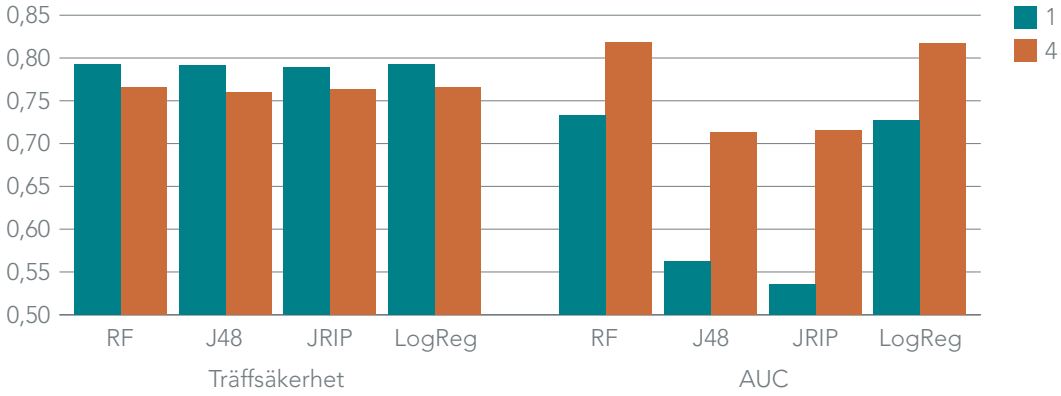
Resultat och analys

Den prediktiva prestandan hos de olika metoderna, då all data används för modelleringen, framgår av figur 8 nedan. Det är naturligtvis svårt att avgöra vad som motsvarar en värdefull (eller ens acceptabel) nivå, men modellerna är i alla fall mycket mer informativa än en gissning på majoritetsklassen. En intressant iakttagelse är att i detta experiment är den enkla tekniken logistisk regression minst lika bra som random forest, medan J48 och JRip har sämre prestanda, särskilt avseende AUC.



Figur 8. Prestanda då datamängden analyserades i sin helhet.

I figur 9 nedan visas den prediktiva prestandan för kunder med en order respektive med minst fyra order. I segmentet med bara en order ser vi att alla modeller är ungefär lika bra om man bara beaktar träffsäkerhet (accuracy). Tyvärr ligger nivåerna väldigt nära den naiva gissningen på att alla lämnar. Tittar man däremot på AUC så har random forest och logistisk regression mycket högre värden än JRip och J48, vilka båda närmar sig nivån 0,5 som motsvarar ingen information alls. Samma mönster återfinns i segmentet med kunder med minst fyra order, det vill säga ganska lika prestanda mellan teknikerna avseende träffsäkerhet, medan random forest och logistisk regression är överlägsna på AUC.

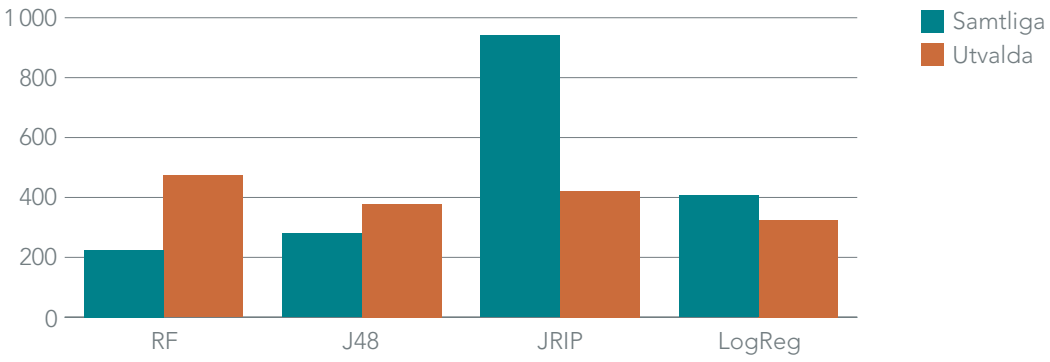


Figur 9. Prestanda för kunder med exakt en order respektive fyra eller flera.

Det är här viktigt att kunna tolka resultaten rätt, specifikt att inse att AUC är det viktigaste kvalitetsmåttet. Tittar man bara på träffsäkerhet blir bilden att modellerna är bättre på att predicera kunder med bara en order än kunder med fyra eller fler, och att valet av teknik inte spelar någon roll. Detta må vara sant rent siffermässigt, men anledningen är den kraftigt obalanserade fördelningen mellan klasserna. Vill man däremot hitta icke-triviala samband, alternativt rangordna kunderna efter uppskattad risk att de lämnar, så bör tekniken väljas utifrån AUC, och då är det tydligt att random forest och logistisk regression har mycket bättre prestanda. Specifikt, på segmentet med bara en order, ger de viss information, medan JRip och J48 har väldigt liten extra förmåga jämfört med den naiva gissningen på att alla lämnar. För segmentet med fyra eller fler order, vilket man kan tänka sig är väldigt viktigt för företaget, indikerar relativt höga AUC-värden att modellerna borde kunna användas för att välja ut de kunder som har högst risk för churn.

Figur 10 visar körningstiderna för de olika algoritmerna, med och utan attributselektion, då hela datamängden analyseras. Tiderna inkluderar både byggandet av modellerna och själva prediktionerna – där byggandet normalt står för mer än 90 procent av tiden. När attributselektion används ingår även tiden för detta steg. Det klart viktigaste resultatet är att alla körningstider är acceptabla med tanke på att den här typen av analyser typiskt görs väldigt sällan. Tiderna varierar mellan strax över tre minuter upp till en kvart. Detta gör att en direkt jämförelse mellan de olika teknikerna blir relativt ointressant, men

det är ändå möjligen intressant att den mest kraftfulla tekniken random forest (och då utan attributselektion) tar minst tid av alla. Utöver det kan också vara värt att notera att för två tekniker (random forest och J48) så tar själva attributselektionen så lång tid att totaltiderna ökar. JRip har mest nytta av attributselektionen medan den för logistisk regression har ganska liten betydelse.



Figur 10. Tidsåtgång i sekunder för de olika algoritmerna med och utan attributselektion – träning och test.

Sammanfattningsvis framstår random forest och logistisk regression som de bästa alternativen då de ger mest information, utan att körningstiderna för den skull ökar. J48 och JRip ger visserligen tolkningsbara modeller, men det sker här tydligt på bekostnad av rangordningsförmågan, vilket ofta är det mest centrala i den här typen av analyser. Detta mönster är här särskilt tydligt då specifika segment analyseras separat.

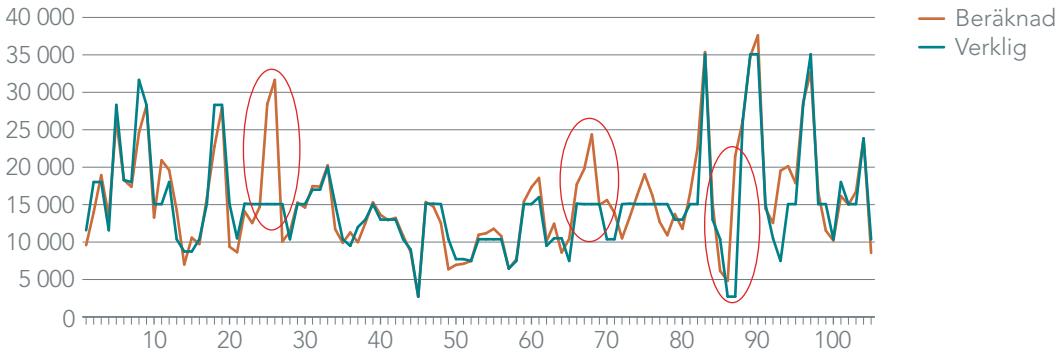
Slutsatser

Den här studien visar tydligt att prediktiv modellering som utnyttjar relativt stora datamängder kan genomföras med fritt tillgänglig programvara och på en standarddator, det vill säga utan krav på den typ av lösningar som normalt förknippas med big data. Vår bedömning är dessutom att de flesta liknande datamängder inom handeln troligtvis är mindre än den här analyserade, eftersom den baseras på en ledande e-handlares kundregister omfattande fler än 250 000 kunder. Vår slutsats är därför att många aktörer med liknande problem, och motsvarade tillgång till insamlad data, relativt enkelt kan utöka sin verktygslåda med dataanalys – och då utan att det kräver stora investeringar i hårdvara eller avancerade analysprogram.

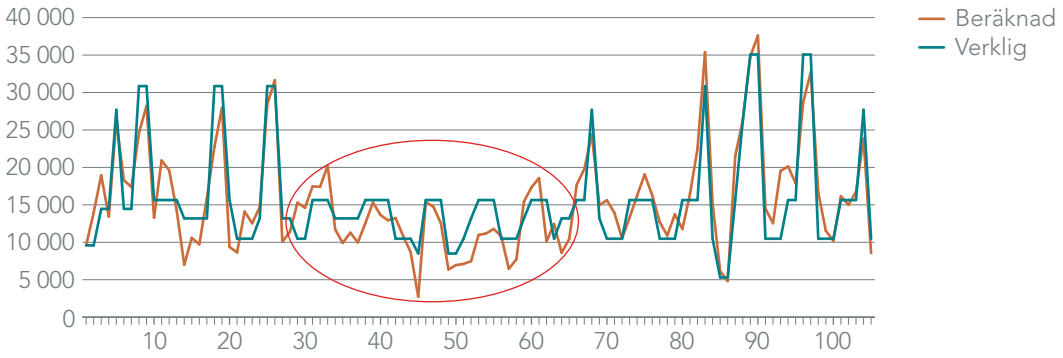
3.2.4 Redovisning – Alternativa optimeringsfunktioner

Det går att mäta prediktionsfel på många olika sätt, och varje mått beskriver en viss egenskap hos felet. Skillnaderna blir särskilt intressanta om måtten används som del av en optimeringsfunktion vid prediktiv modellering, eftersom valet av mått då direkt kommer styra modellen. Figur 11 visar en försäljningsprognos för en modell som är optimerad för att minimera *mean absolute error (MAE)*. Som synes finns det tre riktigt stora fel medan prediktionerna runt medelvärdet är förhållandevis träffsäkra. MAE är

här 15 procent av den genomsnittliga försäljningen och *root mean square error (RMSE)* 26 procent. Figur 12 visar istället en prognos från en modell optimerad, på exakt samma data, för att minimera RMSE. I detta fall finns inga riktigt stora fel men prediktionerna runt medelvärdet är betydligt sämre. Det syns även i att MAE här är 19 procent av medelförsäljningen men RMSE 22 procent.



Figur 11. Försäljningsprognos från en modell optimerad för att minimera MAE.



Figur 12. Försäljningsprognos från en modell optimerad för att minimera RMSE.

Exemplet visar att det felmått som minimeras av den prediktiva tekniken ger konsekvenser för modellen, typiskt att modeller optimerade för MAE ger ett lågt MAE men högre RMSE och tvärtom. Praktiska konsekvenser vid försäljningsprognostisering kan till exempel vara:

- *RMSE* är bättre att optimera om man vill undvika stora fel, till exempel för att undvika att för få produkter beställs i samband med en större kampanj.
- *MAE* kan med fördel optimeras om fokus ligger på den normala försäljningen och kampanjer endast förekommer vid enstaka tillfällen och då hanteras manuellt.
- *Korrelation:* beskriver hur mycket av målvariabelns variation som fångas av modellen och kan därför vara ett bra val om modellen ska användas för att hitta nya kundinsikter.

- *Mean Absolute Percentage Error:* lägger större vikt vid fel på små volymer och tenderar därför att ge mer konservativa prognoser, kan vara fördelaktigt för att undvika för stora inköp.

I artikeln (König & Johansson, 2014) presenterades en ny teknik som kan optimera godtyckligt felmått. Effekten av att optimera fyra olika felmått utvärderades på ett urval av 179 frekvent kampanjade Ica-produkter (en delmängd av de artiklar som utvärderades i avsnitt 3.1.3). Experiment visade att det gick att optimera felmåtten och att det gav uppenbara konsekvenser för modellerna och dess prediktioner.

Slutsatser

- Det felmått som minimeras vid skapande av en prediktiv modell har stor påverkan på modellens prediktioner. Olika mått belyser olika egenskaper hos felet och modellen kommer att bli bra på just detta.

- Då felmåttet är inbyggt i det flesta traditionella prediktiva tekniker blir valet av teknik automatiskt även ett val av felmått.

3.2.5 Redovisning – Situationsanpassade prediktiva modeller

Vid traditionell prediktiv modellering används alltså historisk data för att bygga en generell modell vilken sedan används i de faktiska prediktionerna. Detta är ett så vanligt sätt att arbeta att det sällan eller aldrig ifrågasätts. Det är dock viktigt att inse att man med detta arbetssätt i själva verket löser ett svårare problem än det man egentligen står inför. När man bygger en generell prediktiv modell utifrån historisk data så blir konsekvensen att modellen kommer fungera för vilken testdata som helst – så länge den kommer från, lite slarvigt uttryckt, samma fördelning. Det här är alltså själva grundtesen i all prediktiv modellering, vi bygger modellen och den kan i ett senare skede användas för prediktionerna. Dock finns det ett otal situationer där man redan vid modelleringssituationen vet på vilka instanser som modellen senare ska användas, och då öppnar sig, i princip, möjligheten att på något sätt utnyttja denna kunskap. Låt oss ta ett exempel från handelsdomänen för att illustrera fenomenet: vid prediceringen av churn (se avsnitt 3.2.3) så byggdes alltså modellen från historisk data där man för varje kund vet huruvida denne lämnat eller inte, utifrån den valda definitionen. Modellen användes sedan på testdatan, vilken då förstås bestod av ett antal kunder där vi i princip inte hade tillgång till rätt värde på målvariabeln, alltså om kunden lämnat eller inte. Men, och detta är en viktig poäng, i det läget att ett företag bygger en modell för att avgöra vilka kunder som kommer lämna eller inte, så är normalfallet rimligen att man redan har de kundprofiler som den skarpa prediktionen (om de ska lämna eller inte) ska göras på tillgängliga. Samma argument kan enkelt föras vid responsmodellering – vi har redan innan modelleringstillfället bestämt från vilken mängd kunder mottagarna ska väljas.

I dessa situationer – då vi i modelleringsögonblicket har tillgång till de instanser för vilken prediktionen ska göras – har vi alltså extra information som inte används vid traditionell prediktiv modellering. Vi har tidigare föreslagit en metod för att utnyttja

detta kallad oracle coaching (Johansson & Niklasson, 2009). Vid oracle coaching är målet en tolkningsbar modell som är speciellt anpassad för de specifika instanser för vilken prediktionen ska göras. I tidigare studier exempelvis (Johansson et al., 2012) har metoden använts för klassificering, men inom ramen för FBI-projektet har vi nu utökat den till regression, se (Johansson, Sönströd & König, 2014). Eftersom vi använde publika benchmark-datamängder, och studiens bidrag i huvudsak var på algoritm- och metodnivå, det vill säga riktade till forskare inom maskininlärning och data mining, presenteras den här inte i detalj. Värt att notera är dock likheten med metoden (beskriven i avsnitt 3.2.4) som utnyttjar regelextrahering som ett sätt att ”filtrera” data innan modelleringsfasen.

Resultaten i både denna nya studie och tidigare studier visar entydigt att den föreslagna metoden förmår utnyttja den extra tillgängliga informationen för att genom situationsanpassning skapa mer träffsäkra modeller. Slutsatsen är därmed att för den specifika situationen då man vid modelleringsögonblicket redan har den data på vilken modellen ska användas tillgänglig, och tolkningsbara modeller är önskvärt, så fungerar oracle coaching för både klassificering och regression.

3.2.6 Redovisning – Verktyg för prediktiv modellering och känslighetsanalys

Ett flertal nya och avancerade algoritmer och tekniker har utvecklats under detta projekt. Forskningsresultat av denna typ är dock ofta svåra förmedla till branschen. För att bättre tillgängliggöra resultaten har vi därför, vid sidan av projektet, även implementerat några av de mest lovande resultaten i en fristående applikation som finns tillgänglig för utvärdering och kommersiell användning via www.tiplersoftware.com.

Mer specifikt är programmet en plattformsoberoende lättviktsapplikation (160 kb) med en motor för dataanalys. Motorn stödjer prediktion, simulering av alternativa scenarion och känslighetsanalys, det vill säga värdering av variabeleffekt (som ibland även benämns attribution). Applikationen har även försetts med ett intuitivt gränssnitt och designats för att vara ”one click”, det vill säga all konfiguration av underliggande algoritmer görs automatiskt och ”under huven”. Därmed är verktyget enkelt att använda och kräver ingen specialistkompetens. I följande avsnitt beskrivs den grundläggande funktionaliteten och hur vi anser att denna kan gagna företag i handeln. Applikationen är dock generell och det finns ett otal möjligheter att utnyttja och vidareutveckla inbyggd funktionalitet i samarbete med företag intresserade av prediktiv analys.

Prediktion

Den viktigaste funktionalitet för denna typ av analysverktyg är självfallet den prediktiva modelleringen, då den ska säkerställa träffsäkra och generella modeller. Då både tidigare forskning till exempel (Meyer, Leisch & Hornik 2003) och experiment redovisade i avsnitt 3.1.3 visat att random forest oftast är både mer robust och mer träffsäker jämfört med andra prediktiva tekniker, blev detta kärnan för den prediktiva motorn.

För att träna en random forest på en vald datamängd krävs endast ett klick på knappen ”Forecast”. Modellen utvärderas därmed på avsatt testmängd och resultaten visas med lämpliga felmått. Röda staplar visar felet på träningsmängden och orange på testmängden. Exemplet som visas i figur 13 visar prediktion av den dagliga försäljningen för en ”food truck”. Som förklarande variabler används försäljning föregående dag samt samma dag föregående vecka, millimeter regn den aktuella dagen, dag i veckan, lönedag och event som beskriver om något större event genomförts i närheten av aktuell food truck. Exemplet innehåller endast data för cirka tre månaders försäljning men som redovisats i avsnitt 3.2.3, kan en vanlig arbetsstation analysera datamängder med upp till ett par hundra tusen instanser utan problem.

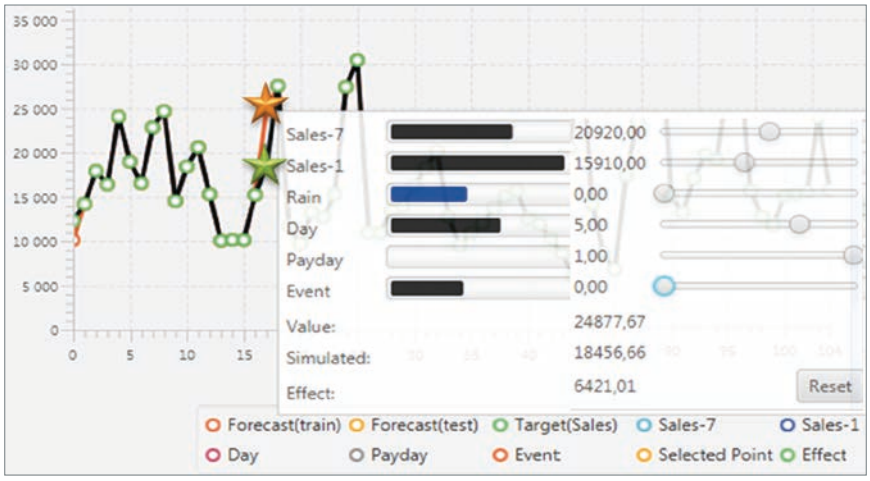
Figur 13. Daglig försäljningsprognos för en food truck.

Simulering av alternativa scenarion

Det grundläggande antagandet för prediktiv modellering är att det studerade fenomenet, här försäljningen, kan förklaras utifrån en uppsättning variabler. Den prediktiva modellen blir därmed en beskrivning av det underliggande sambandet och kan användas för att förklara observerade exempel och för prediktion av framtida instanser. Förklaring och prediktion är den typiska användningen av prediktiva modeller men då modellen antas beskriva ett faktiskt underliggande samband kan den även användas för simulering av alternativa scenarion.

Ett alternativt scenario definieras rättframt genom att skapa en kopia av en instans och ändra värdet för en eller flera variabler. Därefter används den prediktiva modellen för att göra en prediktion för den förändrade instansen och därmed för det alternativa scenariot. Då det är modellens prediktion som används för simuleringen kan tekniken användas både för redan observerade värden och för framtida instanser. En mängd alternativa

scenarion kan till exempel skapas för att utvärdera utfallet av olika framtida kampanjer och därmed för att optimera exempelvis mediainvesteringar.



Figur 14. Simulering alternativt scenario.

Att skapa alternativa scenarion genom att definiera nya datapunkter kan vara krångligt och onödigt komplext för en normal användare. I applikationen har därför denna uppgift förenklats och integrerats i det grafiska gränssnittet. För att skapa ett alternativt scenario klickar man på en datapunkt vilket visar ett ”popup-fönster” med ”slidebars”-kontroller för varje variabel, se figur 14. Från början visas värden för den ursprungliga instansen vilka dock enkelt kan manipuleras, med respektive slidebar, för att skapa ett alternativt scenario. Varje förändring ger direkt upphov till en ny prediktion och resultatet kan observeras interaktivt i diagrammet. Därmed kan en användare på några sekunder interaktivt laborera med en mängd olika scenarion. I figur 14 visar den orangefärgade stjärnan den ursprungliga prediktionen för en träningsinstans medan den gröna stjärna visar vad försäljningen skulle varit samma dag om ett specifikt event inte genomförts.

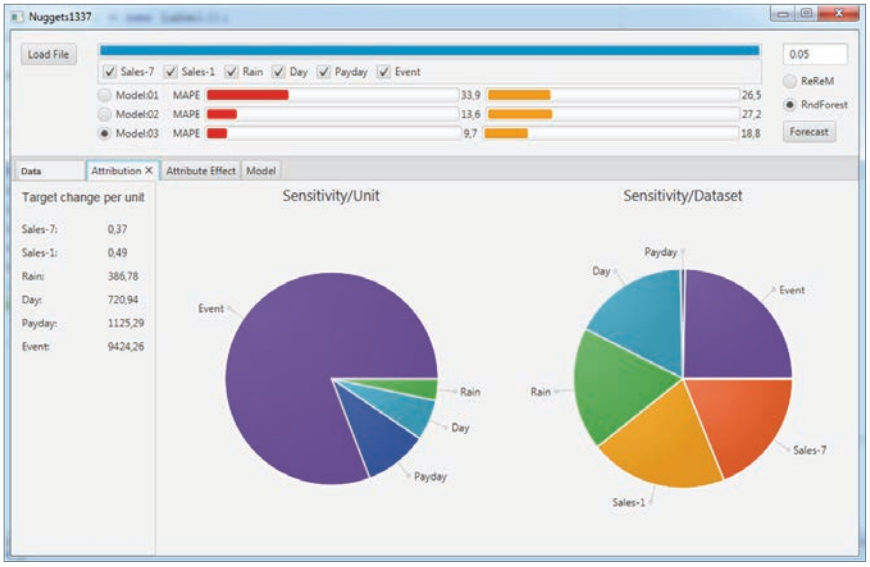
Känslighetsanalys

En nackdel med random forest och andra ensembletekniker är, enligt tidigare, att de inte är tolkningsbara, det vill säga det går inte att få en enkel förklaring av en enskild prognos på grund av en alltför komplex modell. Detta innebär dock inte att det är omöjligt att få en inblick i det samband som modellen har hittat. Ett vanligt förekommande sätt är känslighetsanalys, vilket innebär att utvärdera vilken påverkan förändringar i de ingående variablerna har på modellens prediktion. Om en förändring av värdet för en variabel ger en stor påverkan på prediktionen innebär detta också att variabeln är viktig för modellen, det vill säga har en stor påverkan på målvariabelns värde. Saltelli, Chan och Scott (2000) ger en bra introduktion och en mycket omfattande survey över de vanligast förekommande teknikerna. Isukapalli (1999) identifierar vidare tre olika typer av tekniker för känslighetsanalys; *variabel(parameter)-variation* samt *global(domain)* och *lokal* känslighetsanalys. Variabelvariation används för att generellt analysera effekten av att inkludera eller exkludera en viss variabel i modellen. Global och lokal

känslighetsanalys analyserar hur förändringar i ingående variabler påverkar prediktionen. Den globala känslighetsanalysen görs för hela domänen, alltså hela datamängden och alla möjliga variabelvärden till skillnad för den lokala känslighetsanalysen som görs för en enskild instans/datapunkt.

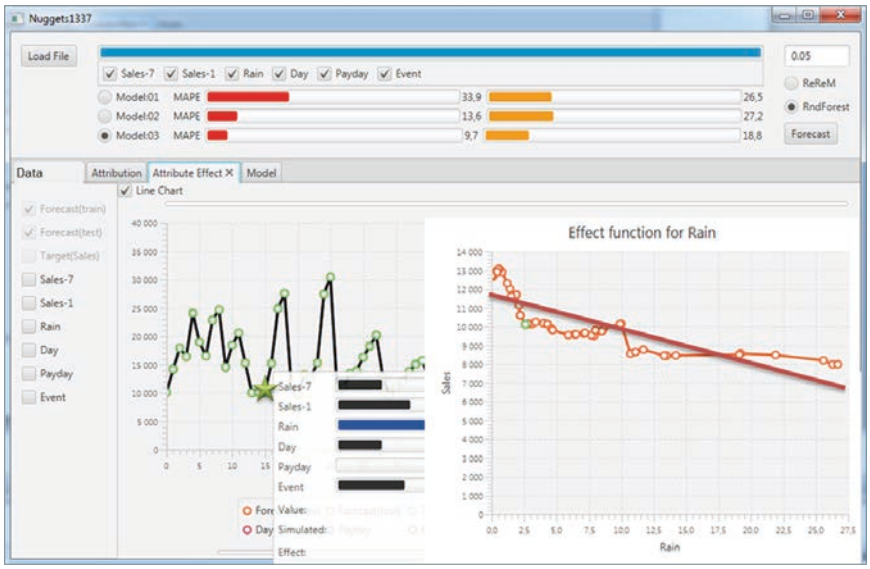
Följande avsnitt beskriver en approach till global och lokal känslighetanaslys (som används i applikationen) och hur resultaten kan gagna företag inom handeln. Den grundläggande idén är att utvärdera hur en *perturbation*, (en slumpmässig förändring) påverkar modellens prediktioner. Mer specifikt noteras först modellens prediktion för den ursprungliga instansen. Därefter ersätts värdet för en variabel i taget med värdet för samma variabel från en annan slumpmässigt vald instans. Modellen får predicera den nya instansen och skillnaden mot den ursprungliga prediktionen noteras. För att få en mer tillförlitlig uppskattning upprepas processen ett stort antal gånger. När samtliga instanser och variabler bearbetats summeras den totala skillnaden som perturbationerna gett upphov till för de olika variablerna och variablernas inbördes betydelse kan sedan beräknas utifrån dessa värden.

Figur 15 nedan visar en känslighetsanalys av den random forest som beskrevs i föregående exempel. Det högra diagrammet visar resultatet av en global känslighets-analys, alltså den genomsnittliga effekten av variablerna sett över hela datamängden. Event, tidigare försäljning, regn och dag i veckan har alla stor påverkan. Denna typ av global känslighetsanalys kan användas för att ge en överblick över vilka de viktigaste variablerna är för den totala omsättningen. Analysen säger dock inte hur viktig en förändring i en variabel är, då antalet observationer och skalan för variablerna inte tas med i beräkningarna. Lönedag har i realiteten en markant påverkan på försäljningen men har här bara förknippats med en minimal effekt, då datamängdens 104 dagar bara innehåller tre lönehelger. Diagrammet visar dock att regn, som tyvärr inträffar mycket oftare än löneutbetalning, har en betydligt större påverkan på den totala omsättningen.



Figur 15. Global och normaliserad känslighetsanalys.

För att ge en mer detaljerad förklaring av variablernas betydelse kan en normaliserad känslighetsanalys göras. Skillnaden är att den förändring som perturbationen ger upphov till divideras med perturbationens värde. Därmed blir förändringen normaliserad till den enhet som används för variabeln. Diagrammet till vänster i figuren ovan visar en normaliserad känslighetsanalys för samma prediktiva modell. I detta diagram har lönedag en betydligt större inverkan och regn en mindre inverkan. Tabellen längs till vänster ger den faktiska effekten av respektive variabel och enhet. Här bedöms exempelvis varje millimeter regn ge en minskad försäljning på cirka 380 kronor och lönedag en ökad försäljning på ungefär 1 000 kronor.



Figur 16. Lokal känslighetsanalys.

En lokal känslighetsanalys tillför en ytterligare dimension då analysen där görs för en enskild instans (här dag). Känslighetsanalysen sker på samma sätt som i det globala fallet men endast för den aktuella instansen. Resultatet kan sedan presenteras relativt till övriga variabler vilket visas med de horisontella staplarna i figur 14 och figur 16. I figur 16 har regn störst inverkan på försäljningen den aktuella dagen (markerad med en stjärna) till skillnad från dagen i figur 14 där regn är den minst påverkande variabeln. Skillnaden mot den normaliserade känslighetsanalysen, som visar ett medelvärde, är just att relationerna mellan variablerna kan skifta från dag till dag. Regn kan till exempel ha en stor inverkan en normal måndag men mindre inverkan en dag då ett stort event drar folk till staden oavsett väder. Därmed kan den lokala känslighetsanalysen användas för att snabbt och enkelt analysera orsakerna till en på något sätt avvikande försäljning.

Ett annat sätt att presentera resultatet, av en lokal känslighetsanalys, är som en funktion vilken beskriver hur en variabel påverkar försäljningen en specifik dag. Grafen till höger i figuren visar exempelvis hur, enligt modellen, olika mängder regn påverkar försäljningen den aktuella dagen. Då random forest är en icke-linjär teknik är kurvan inte en rät linje. Diagrammet visar i stället en stor negativ effekt mellan 0,1–2,5 millimeter regn och

därefter en ganska liten effekt upp till 10 millimeter regn, varefter försäljningen åter minskar markant. Traditionella tekniker som multipel linjär regression skulle istället ge den räta röda linje som visas i diagrammet och därmed inte ge lika mycket information. En viktig detalj är att denna typ av analys är mer tillförlitlig för mindre perturbationer, eftersom stora skillnader mellan perturbationen och det verkliga värdet kan generera ett mycket osannolikt exempel.

Till sist bör det noteras att korrektheten hos alla känslighetsanalyser inte helt kan garanteras. Korrektheten är visserligen kopplad till den prediktiva modellens träffsäkerhet men även en modell med perfekt träffsäkerhet kan beskriva ett felaktigt samband beroende på tvetydigheter i datamängden eller att det underliggande sambandet förändrats. För en djupare diskussion om detta se (König, Johansson & Niklasson, 2010). Därför bör alltid en känslighetsanalys tolkas av domänexperter vilka kan avgöra om den ger en rimlig förklaring.

Slutsatser

- Prediktiv modellering möjliggör simulering av alternativa scenarion vilket kan vara ett kraftfullt planeringsverktyg.

- Global känslighetsanalys ger en generell bild över hur ingående variabler påverkar modellen för en mängd exempel eller över en längre tidsperiod. Global känslighetsanalys är därför ett bra planeringsverktyg och kan därmed med fördel användas för strategiska beslut som optimering av kampanjstrategi.

- Icke-linjära tekniker kan ge mer detaljerade insikter än traditionella linjära tekniker.

Dataanalys behöver inte vara kostsamt.

3.2.7 Rekommendationer

Dataanalys behöver inte vara kostsamt. Väldigt få svenska handelsföretag står i dag inför verkliga big data-problem, varför de flesta kan utföra avancerad dataanalys med vanliga arbetsstationer och open-scource verktyg. Företag måste därmed inte köpa dyra helhetslösningar utan kan använda lättviktiga lösningar som antingen används parallellt med övriga system, för exempelvis årliga punktinsatser, eller som integreras direkt i existerande affärssystem. Fokus måste konkret flytta från insamlande och processerande av data till hur tillgänglig data ska utnyttjas på bästa sätt som beslutsstöd i kritiska processer. Detta kräver möjligen att kunskapen i organisationen om tillgängliga verktyg ökar, men framförallt att man klarar av att identifiera vilka processer som dataanalysen kan stödja. Två mikrotekniker för förbättrad prediktiv modellering som utvärderats i projektet är:

- Möjligheten att definiera ett optimeringsmått som passar problemet, och sedan välja en teknik som kan optimera detta. Oftast väljer man teknik först och får då ett felmått på köpet, vilket kanske inte alls är det bästa valet för problemet, och därför resulterar i suboptimala prediktioner.

- Möjligheten att även inkludera information om den mängd instanser (typiskt kunder) som man vill genomföra prediktionen på, vid själva modellerandet.

3.3 Kampanjer och personifiering

Strävan efter att utveckla strukturer och system som kan svara upp mot efterfrågan i ”realtid”, kombinerat med djupare kundinsikt (Sundström & Ericsson, 2012; 2015) om hur individer reagerar på olika kampanjer kan leda till ökad träffsäkerhet i prognoser. Av det skälet är det relevant att undersöka hur kampanjer påverkar försäljningssiffrorna samt vilka effekter olika kampanjverktyg har.

3.3.1 Sammanfattning

Djupare kundinsikt handlar i detta tema om personifierade kampanjer som vänder sig till slutkonsument. Tack vare ökad konkurrens och lägre marginaler har handeln inte råd att göra misstag och måste bli duktigare på att beräkna volymer baserat på efterfrågan (Sandberg & Abrahamsson, 2011). I framtiden kommer det inte att räcka att samla in data utan affärsintelligensen måste veta vilka kampanjvariabler som ger hög relevans. När denna studie gjordes var det kampanjverktyg med möjlighet att individualisera som hade hög relevans. Vi visar att prognostiseringsmodeller som tar hänsyn till kampanjvariabler måste uppdateras och innebära sådana aktiviteter som för stunden har hög kundrelevans, till exempel QR-koder och app-användning. Det våra studier visar är:

- Att tryckt massreklam inte har någon hög personlig kundrelevans och att det inte finns något samband mellan ökad försäljning och den tid individer tittar på ett reklambudskap.

- Att det går att rikta mottagarens blick mot särskilda budskap i tryckt reklam och därigenom påverka deras uppfattning om erbjudandet.

- Att produktinformation som ger kunden konkret nytta och/eller underhållning har hög relevans för unga kunder och bör vara aktuella att väga in i framtidens prognostiseringsmodeller.

- Att kampanjer som erbjuder eller innehåller digitala inslag i högre grad tilltalar unga kunder och bör vara aktuella att väga in i prognostiseringsmodeller.

3.3.2 Introduktion

Data mining har blivit ett verktyg som särskilt de stora företagen lärt sig att använda och som bidrar till minskade lagerkostnader och minskat svinn. Data mining har också

inneburit att handelsföretag kan agera snabbare på efterfrågesvängningar, både lokalt och globalt (Ngai, Xiu & Chau, 2009). Kampanjer har däremot ofta utgjort ett inslag som gör prognoserna osäkra eftersom det är svårt att veta hur väl de når fram. Sådan osäkerhet har hanterats med hjälp av erfarenhet och historisk data och ofta resulterat i *relativt* korrekta prognoser, men blivit bättre tack vare utvecklingen av kundkorgsanalyser (Richards, Hamilton & Yonezawa, 2015) som i sin tur bidragit med identifiering av finare och mer träffsäkra kundsegment (Ismail et al., 2015). Data mining verktyg har gjort att bland annat dagligvarubranschen blivit duktigare på att analysera kunders köphistorik och förutse beteendet kopplat till kampanjen. Optiska lösningar har bidragit till högre precision vad gäller kampanjutfall. Särskilt dagligvarukedjorna har blivit duktiga på att mäta utfallet av tryckt reklam med hjälp av streckkoder på kuponger. En föregångare vad gäller sådan precision i kampanjarbetet har varit Ica som använt data mining verktyg för att analysera varje kunds unika köphistorik och basera framtida person anpassade erbjudanden ”Mina varor” på detta.

Aldrig tidigare har företag haft så mycket potentiellt värdefull information om sina kunder som nu och det gäller särskilt de företag som har tillgång till någon form av CRM-system. År 2008 när Ica för första gången testade individuella erbjudanden baserade på kundens inköpshistorik blev det ifrågasatt. Exempelvis Sveriges Konsumenter protesterade och menade att konsekvenserna av sådan data mining skulle påverka den personliga integriteten negativt (Andersson, 2008). Men nu har konsumenterna börjat vänja sig vid riktade erbjudanden och utskickens höga relevans bidrar till hög upplevd kundnytta.

De enskilda Ica-handlarna kan dock inte använda sådana kampanjer utan ger istället ut massreklam i form av tryckta reklamblad. De måste således också ha förmågan att kunna förutsäga efterfrågan av ett erbjudande, för ingen kund är så missnöjd som den kund som fått ett erbjudande men som sedan inte hittar varan i butiken. Det handlar om att kunna beräkna relevansen i ett erbjudande redan innan erbjudandet går ut. Det är också relevansen som måste identifieras och omvandlas till hanterbar data i prognostiseringsmodellerna. Eftersom handelns digitalisering är ett pågående fenomen (Hagberg, Sundström & Egels-Zandén, 2016) var det därför relevant att undersöka något media som förenade de virtuella och analoga världarna, som kan användas i reklamsammanhang och som kan användas när behovet är högt av en viss produkt. Valet föll på QR-koder och hur dessa kan påverka konsumentens agerande i en butikskampanj. Tidigare forskning om QR-koder har visat att unga konsumenter i högre grad än äldre använder QR-koder och en amerikansk studie visar att det kundupplevda värdet ökar om unga individer har möjlighet att utöka sin butiks- och produktupplevelse med hjälp av mobilen (Sago, 2011). Även andra studier visar att QR-koder är uppskattade verktyg eftersom det integrerar den virtuella och verkliga världen och kan påverka kundupplevt värde av en produkt eller tjänst (Fine &

Clark, 2015). QR-koder är på det sättet också ett samtida verktyg som bör undersökas vad gäller relevans.

Eftersom vi i denna studie bland annat ville ta reda på hur mindre företag utan CRM-program för kundvård kan nå en tydlig kundförståelse, intresserade vi oss för tryckt massreklam då det är ett vanligt kampanjmedia. Företaget som vi samarbetade med i studien var Ica City-gruppen i Borås, där det fanns en uppfattning om att de lokala reklambladen är effektiva. Uppfattningen baseras på att utskicken leder till ökad försäljning samt att kunder ofta frågar efter veckobladet i butiken om de inte fått hem det i brevlådan. Men det fanns lägre kunskap kring om det är reklambladen i sig som leder till ökad försäljning eller om det är skyltning av kampanjvaror i butik som leder till ett positivt samband. Det är fullt möjligt att kampanjvarorna säljs lika mycket tack vare exponering i butik som via den tryckta reklamen. Det vi ville ta reda på var om kundernas uppmärksamhet på ett tryckt reklamblad leder till att de köper fler produkter, samt om det gick att rikta deras uppmärksamhet och påverka deras uppfattning av ett erbjudande.

I den andra studien ville vi undersöka vikten av att få rätt typ av information vid rätt tillfälle, för att på så sätt avgöra om sådana variabler bör läggas till i prognostiseringen. Vi vände oss i denna studie till medelstora företags problematik med att ha svårt att förutsäga effekterna av kampanjer i butik, där vi gjorde tester tillsammans med Hööks och Hemtex.

3.3.3 Redovisning

Metodmässigt har de två olika studierna skiljt sig åt. Studie 1 inleddes med fokus-grupper för att få en kunskap om hur konsumenter uppfattar att de påverkas av kampanjer i butik. Därefter genomfördes eyetracking-tester för att mäta ögonpositioner och pupillernas rörelse över ett fast objekt, ett reklamblad. Urvalet av respondenter gjordes slumpmässigt och i avsikt att representera ”vem som helst” som också skulle kunna vara en Ica City-kund. Respondenterna delades in i två grupper där grupp A fick se ett reklamblad från Ica med olika produkter och en prisangivelse vid varje produkt. Grupp B fick se ett identiskt reklamblad med samma produkter och prisangivelser fast med ett tillägg vid fyra av produkterna där vi hade skrivit ”under halva priset”. Reklambladet som användes i testet var ett reklamblad som getts ut ett år tidigare där vi också hade tillgång till försäljningssiffrorna före och efter kampanjen. Uppdraget till respondenterna var att läsa reklambladet som man skulle gjort om man varit hemma. När man kände sig klar med att titta på bladet fick man säga till och då avbröts testet och bilden släcktes ner. Efteråt ombads respondenterna i grupp A och B att försöka minnas vilka varor man sett på bladet och vad de kostade. Den sista frågan var hur de värderade det sammantagna erbjudandet samt hur mycket pengar de upplevde att de kunde spara på erbjudandet. Det avslutande momentet i testet var att respondenten fick se de båda reklambladen bredvid varandra och ombads studera dem noga och tala om ifall de kunde se någon skillnad på bladen.

I Studie 2 användes experiment i butik, där 150 enkätsvar analyserades. Enkätstudien genomfördes i två butiker, Hemtex och Hööks, där undersökningen startade genom att be respondenten att i butik skanna en QR-kod som var kopplad till en viss produkt i butiken. Respondenterna informerades om experimentet med hjälp av en skylt enligt följande: ”Skanna QR koden och få mer information om den här produkten”. Varje förbipasserande kund tillfrågades om de visste vad en QR-kod var och ombads sedan använda den (Sundström et al., 2016). Resultaten från fokusgrupperna visar att konsumenter inte uppfattar att de påverkas av kampanjer i butik. Det gäller särskilt analog reklam som till exempel skyltning och exponering. De anser att de är väl förberedda när de besöker butiken och ofta har planerat sina inköp med hjälp av en inköpslista. Det gällde särskilt personer som var 35 år eller äldre. Däremot har de ofta läst ett reklamblad om veckans erbjudande innan de besöker butiken. Analys av eyetracking-datan visade att det fanns produkter som hade mer uppmärksamhet och längre fixations-tid än de andra. Men det fanns ingen ökad försäljning av dessa varor i butiken när vi gick igenom försäljningssiffrorna, utan samtliga produkter hade likvärdig försäljningsökning. Vår hypotes att längre fixationstid skulle medföra ökad försäljning kunde därigenom förkastas.

Vi gjorde fler eyetracking-tester med nya reklamblad och jämförde även där försäljningsdata från butiken men kunde återigen inte hitta något samband mellan ökad fixationstid och/eller uppmärksamhet och försäljning. Däremot visade resultaten från det ursprungliga reklambladet med tilläggstext att uppfattningen av reklambladet som bra eller dåligt påverkades i de fall då texten ”under halva priset” fanns med (Grupp B – med tilläggstext). Fler respondenter uppfattade erbjudandet som mycket bra när de tillfrågades. Resultaten blev också ännu tydligare när vi jämförde respondenternas uppfattning om hur mycket de kunde spara på erbjudandet. Grupp B (med tilläggstext) ansåg att de sparade mer pengar jämfört med Grupp A (utan tilläggstext). När respondenterna i efterhand fick titta på båda reklambladen (med tilläggstext och utan) var det endast sju respondenter som kunde notera tilläggstexten. Övriga upplevde reklambladen som identiska. I samband med testerna frågade vi också deltagarna om de brukade läsa reklamannonser. I båda grupperna var det mer vanligt att personer över 35 år och uppåt läste reklamannonser jämfört med personer under 35 år.

Resultaten från butiksexperimenten visar att konsumenters uppfattning om värdet av digitala inslag i butikskampanjer skiljer sig åt beroende på ålder. Konsumenter i åldern 19–40 år använder gärna QR-koder när de inte kan hitta någon i butiken att fråga eller när de har bråttom. Konsumenter i åldern 41–51 använder QR-koden om de får ett lägre pris.

Studiernas slutsatser är att det finns ett positivt samband mellan tryckt reklam och ökad försäljning men att det *inte* finns någon korrelation mellan hur länge en individ tittar på ett reklamblad och ökad försäljning. Således talar resultaten för att det inte är värt mödan att lägga in sådana variabler i en prognostiseringsmodell och ett företag kan nöja sig med att ange media som variabel. Däremot visar eyetracking-studien att individer ser texter som påverkar deras *uppfattning* om erbjudandet, utan att de själva är medvetna om det.

På en praktisk nivå innebär det en förenkling av prognostisering av inköp för en butik eftersom de kan lita på sina historiska kampanjdata och inte ta hänsyn till hur länge mottagaren tittar på olika produkter. Det gör också arbetet med data mining något enklare för en enskild handlare eftersom vi uteslutit en ytterligare dimension med hjälp av denna studie. Det innebär också att det troligen går att *styra* individens uppfattning men inte beteende, med hjälp av tilläggstexter. Respondenternas åsikt om reklamblad i brevlådan skiljer sig åt och även om vi i denna studie inte kan dra några generella slutsatser eftersom stickprovet är litet, så finns det skäl att misstänka (och fortsätta forska om) att yngre (under 30 år) människor har en mer negativ uppfattning till tryckt reklam, eftersom de inte läser den i lika hög grad som äldre gör. På en teoretisk nivå innebär resultaten att massreklam kan ”styra” mottagarens uppfattning, vilket stödjer klassiska teorier om att reklam är kraftfullt och påverkar attityder (Packard, 1957 & 1960; Nelson, 2008).

Vi har sammantaget visat att det *inte* finns något samband mellan ökad försäljning och längden då individer tittar på ett reklambudskap. Vi visar också att det går att rikta mottagarens blick mot särskilda budskap i tryckt reklam och därigenom påverka deras uppfattning om erbjudandet. Vi har också visat att produktinformation i en kampanj som ger kunden konkret nytta och/eller underhållning är värdefull för unga kunder och bör vägas in som en variabel i prognostiseringsmodellerna.

3.3.4 Rekommendationer

För de företag som överväger att förfina sina prognostiseringsmodeller visar våra studier att även om möjligheten finns att använda massvis med variabler, så är det inte alltid värt mödan att lägga in många kampanjvariabler när det gäller tryckt reklam. De prognostiseringsmodeller som finns på marknaden idag är relativt träffsäkra och tar tillräcklig hänsyn till kampanjvariabler vad gäller tryckt reklam. Det finns således inget behov av att lägga till fler variabler som handlar om hur individer tittar på tryckt reklam. Det som däremot bör studeras är konsumenters inställning och attityd till relevanta budskap (one-to-one) istället för massreklam (one-to-many) utan unik relevans. I en mer digital värld där människor ständigt omges av kommersiella budskap har tekniken med databasmarknadsföring och data mining gjort reklamen mer effektiv, för både företagen och konsumenterna. Digitaliseringen har också medfört nya möjligheter att kommunicera med kunden direkt i butik och det är viktigt att företag ger plats för nya variabler att ta hänsyn till i sin kampanjplanering. Kampanjvariabler som uppfattas ha hög kundrelevans bör vägas in i framtidens prognostiseringsmodeller.

I en mer digital värld där människor ständigt omges av kommersiella budskap har tekniken med databasmarknadsföring och data mining gjort reklamen mer effektiv, för både företagen och konsumenterna.

Rekommendationer till företag ligger i linje med studier inom temaområdet smart data där vi säger:

- Företag måste inte köpa dyra helhetslösningar utan kan använda enklare lösningar som kan integreras i existerande affärssystem.

- Företag måste vara observanta på vad som kunden uppfattar som hög relevans i erbjudandet, men även vad som är hög relevans i medieverktyget och använda den kunskapen när prognostisering sker.

- Företag måste följa med i konsumentens digitaliseringsresa och förbereda kampanjverksamheter som tilltalar en mer digital kund i framtiden.

3.4 Datakvalitet och integritet

Att data om individer samlas in och används för kommersiella syften ifrågasätts naturligtvis då och då. Även lagstiftarna har uppmärksammat fenomenet och snart skärps lagstiftningen i EU om hur företag får hantera och samla in data. Denna trend har delvis drivits fram av konsumentens oro för företagens insamling och hantering av data vilket skapar en paradoxal syn på personifierade erbjudanden. I denna delstudie diskuterar vi alltså data mining från ett kundperspektiv.

3.4.1 Sammanfattning

Tänk dig en mer transparent relation mellan dig och olika företag avseende den information företagen har om dig. Vilken kunskap om dig skulle kunna hjälpa företagen att förbättra dina inköp? Ett uppenbart exempel är information om vilka av varorna i butiken som innehåller ämnen som du eller någon i din familj är allergisk mot. På en annan nivå skulle den ovane köparen kunna få reda på vilket varumärke på tvättmedel som familjen vanligen använder eller vilken olja som senast köptes till bilen. Man skulle även kunna guidas med storlekar, passform eller smak vid klädinköp. I förlängningen skulle företagen kunna förutspå förändringar i köpbeteende och därmed ge informerade förslag på kommande inköp. Allt det här skulle rimligen vara förmånligt för såväl kunden som företagen, men det kräver tillgång till betydligt mer personlig data än vad som i dag normalt finns i systemen. En övergripande fråga blir därför hur kunder skulle reagera på att känna till den personifierade information som företagen redan i dag analyseras, och i förlängningen om man som kund skulle vara beredd att dela med sig av än mer personlig och specifik data, givet att det skulle leda till olika fördelar.

I denna delstudie har vi tittat på kundens syn på företagens ackumulerade kunskap om dem. Fokus har varit på diskussionen kring viljan att dela med sig av personlig och potentiellt känslig information, till exempel i form av en 3D-scannad bild (en ”avatar”) av sig själv, för att i utbyte få hjälp och stöd i sin inköpsprocess. Genom att använda två enkla modeller – *The Johari window* (Luft, 1982) och *The Uncanny Valley* (Mori, 2012) i en omarbetad version från Strong (2015) diskuterar vi i denna delstudie hur öppenhet (transparens) i kunddata skulle kunna vara en möjlig väg för företagen att få tillgång till mer data samtidigt som kunden får underlag för att fatta bättre beslut. Resultatet från fokusgrupperna indikerar att kunder är villiga att dela med sig av till och med så känslig

information som sin egen virtuella avatar om de känner att informationen kan hjälpa dem att fatta ett bättre beslut men att det finns ett antal faktorer som påverkar hur känsliga de är för företagens användande av personlig information. Möjlighet att ta del av andras information (vänner, bekanta men även okända anonymt) skulle kunna hjälpa till att värdera redan befintliga kommentarer kring storlekar och färg framförallt vid inköp via internet. Att själv kunna förfoga över sin kunddata uppfattades generellt som positivt.

3.4.2 Introduktion – Mellan *cute* och *creepy*

2012 skapade Target stora rubriker när de lyckades avslöja en tonårsgraviditet innan flickans egen far gjorde det (Quirk, 2012). Efter att ha identifierat 25 produkter som indikerade graviditet hade företaget börjat skicka ut direktreklam med kuponger för bland annat mammakläder och barnmöbler för att fånga in ett vad man ansåg vara relativt lättpåverkat segment (Lubin, 2012; Hill 2012; Duhigg, 2012). Fallet är ett av många som gav nytt bränsle till debatten kring hur mycket företagen egentligen vet om oss konsumenter, och etiken kring personifierade erbjudanden (Strong, 2015). Hur bekväma är vi egentligen med att erbjudanden blir uttalat personliga? Eller, annorlunda uttryckt, finns det en gräns för hur personliga erbjudanden får bli? I en undersökning som gjordes i samarbete mellan Guardian Media Network och GfK 2013 upplevde 69 procent av de brittiska konsumenterna företagens hantering av personlig data som *creepy* (Strong, 2015; Coll, 2013). Begreppen *cute* och *creepy* är lånade från en modell framtagen under 1970-talet av Masahori Mori benämnd som *the Uncanny Valley*. I originalartikeln som översattes 2012 (Mori, 2012) användes modellen för att förklara att ju mer lik till exempel en handprotes är en riktig hand, desto mer positivt uppfattas detta *(cute)* fram till en viss gräns där likheten med originalet är så stor att de få onaturliga dragen istället skapar en motsatt reaktion *(creepy)*. Ett liknande exempel är den japanska androiden Erica, som trots sin stora likhet med en människa, av många upplevs som *creepy*. Modellen har även använts av forskare för att förklara hur en viss grad av igenkännande kan upplevas positivt vad gäller personifierade marknadsföringsinsatser eller sociala medier, men endast till en viss gräns (Strong, 2015; van den Berg, 2011).

Så vad är det då som gör att vi upplever viss personifierad reklam som positiv *(cute)* medan annan reklam uppfattas negativ *(creepy)*. Strong (2015) föreslår att en del av problematiken ligger i att företagens kunskap överstiger kundens egen kunskap om sig själv. Beakta matrisen i figur 17 nedan som illustrerar *The Johari window* framtagen av Joeseph Luft och Harry Ingham (Luft, 1982).

Figur 17. Johari window (Luft, 1982).

Modellen togs fram som ett verktyg för att diskutera interaktion mellan människor och utgör en fyrfältare. Det första (övre vänstra) fältet innehåller beteenden och motiv som är kända av såväl dig själv som andra. Till höger om detta fält är beteenden och motiv som är kända för andra men som du själv inte känner till. Under det första fältet finns beteenden och motiv som är känt för dig själv men okänt för andra. Det fjärde fältet, slutligen, innehåller beteenden och motiv som är okänt för såväl dig själv som andra. I den omarbetade versionen av Strong (2015) omvandlas detta till företagsinformation om kunder; där första fältet innehåller information om kunden som både företaget och kunden själv känner till, det andra fältet innehåller information som kunden inte själv är medveten om (till exempel vilket kundsegment man tillhör) och det tredje fältet kunskap som företaget inte har om kunden (exempelvis inköp hos konkurrenter eller starkt personlig information). Strong (2015) menar då att det är i det andra fönstret – när företagen skapar erbjudanden utifrån analyser av data som kunden inte är medveten om att företaget har tillgång till – som kunden upplever situationen som *creepy*.

Men det är inte enbart kundens ambivalens gentemot personifierad reklam som skapar en utmaning för företagen i denna diskussion. Vi har i den här rapporten argumenterat för att dataanalys är centralt för handelns konkurrenskraft och specifikt att behovet av smart data är större än behovet av big data. Samtidigt finns det en klar paradox i att även om kunderna förefaller mer än villiga att dela information i sociala medier så är de betydligt mer restriktiva när det gäller att dela med sig av information till företagen (Strong, 2015). Att för företag samla in smart data är inte helt oproblematiskt vilket ledde till en diskussion under den workshop som hölls inom ramen för projektet. Under presentationerna återkom företagen till faktorer som de ansåg saknades i den kunddata som företagen förfogar över. Samtidigt ställde de sig frågande till att denna typ av information skulle vara möjlig att samla in. ”Hur samlar vi in mer data om kunden utan att de vet om det?” För en av de stora fördelarna med big data är att den vanligen samlas in utan att kunden märker det till exempel via aktiviteter online, tidigare köp eller medlemsregister (Strong, 2015).

I företagens försök att berika befintlig data, använder de sig vanligen av traditionella sätt som kundenkäter, vilket dock knappast ger den kvalitet på datan som krävs för att bygga prediktiva modeller. Kundenkäter ger visserligen viss information om kundens preferenser och känslor, men är svåra att koppla till faktiskt agerande. Dessutom saknar kundenkäter den ”real-time effekt” som man får genom medlemsregister. Men ännu viktigare är den motvilja att dela med sig av information som företagen ofta bemöts med av konsumenterna och många studier pekar på att flertalet konsumenter fortfarande uppvisar en osäkerhet kring hur företagen samlar in och hanterar kunddata (Phelps, Nowak & Ferrell, 2000; Kshetri, 2014, Stewart & Segars, 2002). I korthet ställs företagen inför följande dilemma: för att kunna samla in ”smart data” behöver de engagera sina kunder i att longitudinellt och kontinuerligt släppa ifrån sig information om sig själva men motiven för att engagera sig i teknik skiljer sig markant åt, något som till exempel studien kring QR koder visade ovan (se även Sundström et al., 2016).

I denna delstudie ställer vi därför frågan: ”Hur ska vi kunna få kunden att vilja dela med sig av information om sig själv till företag?”

3.4.3 Redovisning

Denna delstudie bygger på sex fokusgruppsintervjuer med 3–5 deltagare där totalt 22 kvinnor i åldrar mellan 18–53 deltog. Varje fokusgrupp tog cirka tre timmar och för att få till stånd en bra dialog kring ämnen som kan uppfattas som känsliga gjordes urvalet genom att individer som anmält sitt intresse att delta själva satte ihop grupper med personer som de kände att de kunde prata fritt med. Förutom löpande anteckningar, videofilmades sessionerna.

Temat för fokusgrupperna utgick från en diskussion kring kroppstyper och måttagning för val av klädstorlek, samt respondenternas upplevelse av hur denna typ av information skulle kunna hjälpa dem vid val av kläder. Under 2015 gjordes en förstudie av två studenter vid Textilhögskolan (di Natali & Ivarsdottir, 2015) som fokuserade på just synen på kroppstyper. Empirin i denna förstudie byggde dock enbart på kroppsskanning och måttagning utan efterföljande gruppintervjuer. Resultaten visar bland annat att respondenterna hade mycket svårt att innan de såg avataren uppskatta sin egen kroppstyp och att även om de valde ”rätt” storlek online så skilde sig måtten framtagna med kroppsskanning väsentligt från de mått som togs med måttband. Detta ledde till ett antal frågor. Hur bra är de tvådimensionella kroppstypsdiagrammen för vår förståelse av vilken kroppstyp vi tillhör? Kan vi med hjälp av en 3D-bild av oss själva ändra uppfattning om vilka kläder som passar oss? Men framför allt – skulle vi vara villiga att använda oss av avataren för att prova och köpa kläder i framtiden? I så fall, skulle vi vara villiga att dela med oss av denna information till andra? En annan viktig iakttagelse från denna förstudie var hur känsligt det var för deltagarna att se sin avatar. Samtliga deltagare reagerade på olika sätt med olika grad av förvåning och ogillande på avatarens detaljrikedom, därav vikten att deltagarna i fokusgrupperna kände förtroende för varandra.

Innan gruppintervjuerna startade användes en 3D-scanner för att skapa en avatar av deltagarna. De fick dessutom svara på ett antal frågor kring vilka storlekar de brukar beställa på nätet samt vilken kroppstyp de uppfattar att de har. De efterföljande gruppintervjuerna var explorativa i sitt upplägg, men startade med att deltagarna fick se gruppens avatarer, en efter en, och med hjälp av skyltar tala om vilken kroppstyp de uppfattade att avataren hade. Utifrån detta diskuterade deltagarna kring hur de resonerat när de valde skylt och vad kroppstyper egentligen har för betydelse för dem. Denna dialog ledde in på nästa ämne – hur en avatar skulle kunna användas för att hjälpa kvinnorna i deras beslutsfattande vid inköp av kläder.

Resultatet i denna delstudie sammanfattas i tre teman: först den del av diskussionerna som berör användandet av avataren i eget beslutsfattande, sedan den del av diskussionerna som berör viljan att dela med sig av informationen till andra (bland annat företag) och slutligen diskussioner som berör potentiella reaktioner om företagen skulle

agera utifrån den kunskap de har för att interagera på ett personligt plan med sina konsumenter.

Trots att det var tydligt att kvinnorna uppfattade det som relativt känsligt att se sig själva i form av en gyllne avatar, så var de mycket kreativa i sina tankar kring hur de skulle kunna använda sig av avataren i den mån det gick att lösa tekniskt. Att ladda ner avataren i ett eget program eller app och sedan använda den för att prova kläder ansågs inte bara positivt utan även väldigt önskvärt. I varje fokusgrupp framkom olika varianter på hur detta skulle kunna ske – en del diskuterar fördelarna att ha det i mobiltelefonen och andra föredrar datorn. Vissa ser det som ett sätt att kunna hitta kombinationer, medan andra snarare ser det som ett spel eller förströelsemoment. Men det var tydligt att kvinnorna såg en klar möjlighet i att avataren, givet rätt teknik, skulle kunna hjälpa dem att hitta rätt storlek och bra passform vid köp på nätet. Tanken på att själv äga informationen gav upphov till tankar kring vilken övrig information de skulle kunna använda sig av. Att ha tillgång till befintlig garderob kom upp i samtliga fokusgrupper. Att se vilka storlekar som tidigare passat när de köpt ett visst varumärke ansåg många skulle kunna vara en fördel.

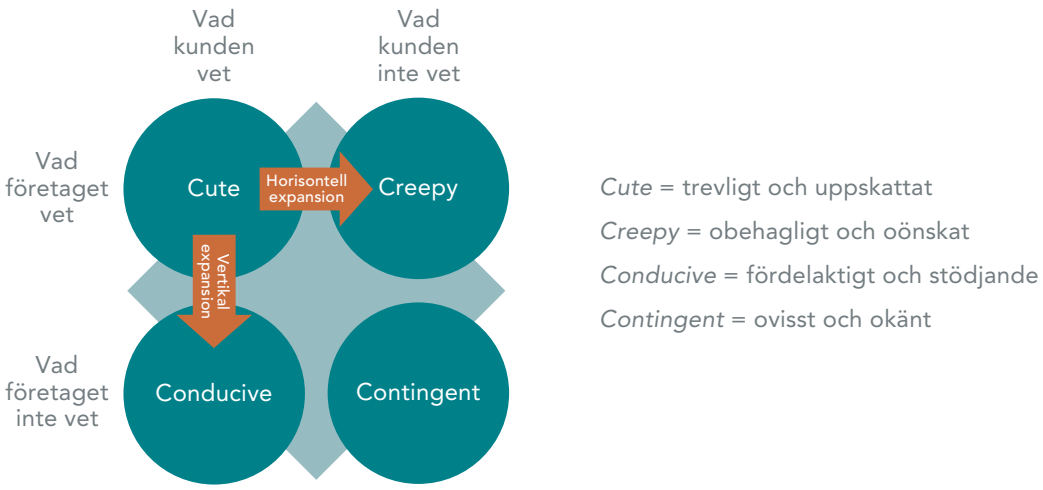
Att lämna ifrån sig sin avatar direkt till ett företag uppfattades som helt uteslutet. Att dela med vänner och familj sågs däremot som en möjlighet. Även annan information som diskuterats, som till exempel garderoben, kunde en del se fördelar med att dela med sig av. Motivet för att dela med sig av denna information till företagen skiljde sig inte oväntat åt mellan respondenterna (speciellt med tanke på det stora åldersspannet). De yngre hade betydligt lättare för att hitta motiv. Även om många här var villiga att dela med sig bara för att få tillgång till befintligt sortiment, så framförde speciellt en av de yngre respondenterna att en prenumeration på kläder från favoritvarumärket så att man var först med att kunna ”prova” online, kanske till och med innan kläderna kom ut till försäljning, skulle vara en mycket stark drivkraft för henne. De som hade som vana att läsa produktinformation innan de köper kläder online såg en klar fördel i att kunna se ”vem” som lämnat produktinformationen även om den var anonym. Informationen ”liten i storleken” blir naturligtvis lättare att relatera till om måtten på informanten finns tillgänglig.

Slutligen diskuterades även reaktioner på möjligheten att företagen utifrån egna analyser av den data som de då skulle kunna få tillgång till, börjar ge råd och tips till dem som konsumenter. Även här fanns stora skillnader mellan kvinnorna i studien. Vissa tyckte att det skulle kunna vara roligt att som en del i ett spel kunna ge tips till andra, men även få tips kring klädstil och passform, medan andra kände sig hemma i sin stil och inte såg det som aktuellt att byta bara för att de nu ser att de har en annan kroppstyp.

3.4.4 Analys

Om vi går tillbaka till the Johari window som introducerades i början på det här kapitlet och relaterar de resultat som vi fått via fokusgrupperna till den modellen, föreslår vi

följande modifikation som ett underlag för en vidare diskussion kring smart data, big data och annan typ av data (se figur 18).



Figur 18. Omarbetad Johari window.

I det första fältet ligger all den information som företagen har kunskap om och som kunden vet att den har delat med sig av. Precis som Strong (2015) föreslog så kan kunden förväntas känna sig trygg och uppskatta den hjälp och fördelar som personifierade erbjudanden ger så länge företagen använder sig av information som ligger inom detta fält. Till höger om fältet *cute* finns den information om kunden som företagen fått fram genom analyser men där kunden själv inte är medveten om varken vilken underliggande data som företaget har tillgång till eller vilka analyser som görs (se exemplet med Target ovan). Detta fält innefattar alltså typiskt resultatet av olika analyser och prediktiva modeller. När företagen använder sig av denna framtagna kunskap i sina riktade kampanjer finns en uppenbar risk att det upplevs som *creepy* av kunden, vilket alltså diskussionen i fokusgrupper bekräftar. I fältet under *(conducive)* ligger den information som företagen gärna skulle vilja ha tillgång till men som kunden själv äger och av olika anledningar valt att inte dela med sig av. För att få tillgång till denna uppenbart personifierade, och därmed extremt värdefulla information, måste företagen hitta nya strategier för att övertyga kunden om att det finns ett tydligt värde för hen att dela med sig av informationen. Slutligen, finns det naturligtvis information som företagen inte känner till om sina kunder och som kunden själv heller inte är medveten om *(contingent)*. Värdet av den informationen är därmed okänt innan den hittas, möjligen genom att företagen upptäcker den genom mer generell och explorativ analys av data.

Modellen understryker alltså att det finns två vägar för att expandera den data som är användbar för företagen – antingen genom att motivera kunden att dela med sig av mer information (vertikalt), alternativt att medvetandegöra kunden om den information som den inte känner till om sig själv genom transparens (horisontellt). Båda dessa vägar

bör rimligtvis innebära lösningar för att involvera kunden i datainsamlingen snarare än att hitta fler sätt att samla in information dolt. Detta tangerar den slutsats som även Phelps, Nowak och Ferrell (2000) drar i sin studie där de föreslår att företag genom att ge kunder mer kontroll över insamling och användande av sin egen information skulle kunna öka deras villighet att dela med sig av mer information. För att nå dit finns det ett antal faktorer som företagen måste beakta. Kan det finnas ett sätt att samspela med andra aktörer inom branschen? Vilket förtroende har vi idag hos våra kunder? Hur kan vi rikta ett erbjudande så att flera olika segment av kunder känner sig attraherade att bli involverade.

3.4.5 Rekommendationer

Utifrån de genomförda studierna och diskussionen ovan anser vi att företag bör överväga möjligheten att låta kunder få större insyn i vilken kunskap företagen har om kunderna. På så sätt kan förtroendet ökas, och kunden kan enklare uppfatta den egna nyttan av företagens dataanalys och de resulterande personifierade erbjudandena. Ny teknik möjliggör att kunden kan äga sin information själv, till exempel i mobilen, och givet rätt incitament visar studierna att kunder verkar vara beredda att dela med sig av även relativt känslig och starkt personlig information. På detta sätt skulle alltså företagen alltså kunna berika sin data med såväl real-time information kring faktiska beslutssituationer som personliga uppgifter. Företag bör därför redan nu överväga att i nästa generation av lojalitetsprogram låta kunden ha inte bara insyn i utan också makt över sin egen data. Rent konkret har kunderna då möjlighet att ”styra” hur tillgänglig informationen ska vara för företaget i olika situationer. Vi är medvetna om att detta är en revolutionerande tanke, och att det antagligen av de flesta företag skulle kännas som att ”ge upp” den kunskap om kunderna som till stora kostnader redan samlats in. Vår uppfattning är dock att det här paradigmskiftet ligger väldigt väl i tiden, då så många redan delar med sig av så mycket på sociala medier. Specifikt ser vi förstås att ett system där kunder äger sin egen data och väljer vad och till vem hen vill dela med sig av den, inte bara signifikant stärker relationen mellan kund och företag, utan även utgör verklig ”smart data”. Resultatet blir därmed mycket bättre analyser och verkligt personifierade erbjudanden av hög kvalitet.

Denna delstudie är bara en första explorativ studie som bygger på de diskussioner som kvinnorna hade kring sina avatarer och hur de skulle kunna användas. Planen är att utveckla experiment med hjälp av den teknik som Handelslabbet på Swedish Institute for Innovative Retailing vid Högskolan i Borås tillhandahåller i form av till exempel informationsdisk, RFID och ett medlemskort i form av en app för att skapa experiment där kunden kan välja vilken data de väljer att dela med sig utifrån olika erbjudanden. Samtidigt sker en studie kring kundens uppfattning av passform utifrån sin avatar av Nina Hernandez inom ramen för hennes doktorsavhandling. I denna studie får respondenterna se sina avatarer klädda i en skjorta i nio varianter – tre utformningar (färgglad, enfärgad och genomskinlig) i tre storlekar. Även denna forskning kommer att bidra till helheten. Förhoppningen är att vi på sikt kan ha data som visar hur kunden skulle reagera på olika scenarion.

3.5 Säkra prediktioner

Ett teoretiskt intressant och samtidigt praktiskt angeläget ämne är huruvida det går att bygga ett matematiskt ramverk där prediktiva metoder för dataanalys kan fås att garantera en viss träffsäkerhet. Av det skälet formulerades ett av projektens teman kring dessa frågeställningar.

3.5.1 Sammanfattning

Problemet med all prognosticering är att prognoser inte alltid blir rätt, och framför allt, att det inte på förhand går att avgöra hur mycket vi kan lita på en viss prognos. Om vår modell predicerar att en kund ska lämna oss, vad är risken (sannolikheten) att hen verkligen gör det? Om vår modell säger att vi kan förvänta oss att sälja 1 500 exemplar av varan X nästa månad – vad säger egentligen den prognosen, och hur säker är den? Vad är sannolikheten att vi säljer fler än 2 500? Färre än 1 000? Den bistra sanningen är att vi i normal prediktiv modellering har extremt svårt att i förväg uppskatta kvaliteten hos prediktionerna, vilket förstås gör det oerhört vanskligt att använda prediktioner som utgångspunkt för olika kalkyler.

Ramverket conformal prediction är ett sätt att få prediktioner med matematiska garantier. Ramverket garanterar – under mycket generösa antaganden – att den faktiska andelen felaktiga prediktioner konvergerar mot den valda signifikansnivån. Conformal prediction fungerar för både klassificering och regression, och det enda priset man betalar för garantin är att prediktionerna blir mängder istället för punktprediktioner. För regression innebär det att en prediktion beskrivs som ett intervall, medan det vid klassificering blir en mängd av klasser. Ett prediktionsfel inom conformal prediction är då det korrekta värdet inte finns med i intervallet/mängden.

En viktig egenskap hos ramverket är att användaren själv väljer signifikansnivån, det vill säga hur många fel som ska göras. Ramverket ser sedan till att de resulterande intervallen/mängderna blir lagom stora för att andelen fel ska gå (exakt) mot den valda signifikansnivån. Det finns alltså en direkt möjlighet för användaren att utifrån situationen välja hur säkra prediktionerna måste vara. Rent generellt är så klart mindre intervall/mängder (för en given signifikansnivå) mer informativa, varför strävan är att minimera dessa. Storleken hos de resulterande intervallen/mängderna beror på ett flertal olika parametrar, exempelvis svårigheten hos problemet som modelleras och kvaliteten på den underliggande modellen, men även ett antal interna faktorer i ramverket conformal prediction.

För branschen är det viktigt att vara medveten om att conformal prediction, trots sin matematiska elegans och unika egenskaper, är oerhört generellt och dessutom enkelt att använda. Mer konkret fungerar det för godtycklig underliggande modell, det vill säga man kan enkelt lägga till conformal prediction i sin modelleringsprocess, alternativt tillföra det direkt ovanpå redan existerande modeller. De faktiska algoritmerna är dessutom enkla, och det finns även några publika kodbibliotek publicerade (pypi.python.

org/pypi/nonconformist), varför tekniken är i princip fullt tillgänglig, trots att den ännu inte nått ut till kommersiell mjukvara.

3.5.2 Introduktion

Inom dataanalys så ligger forskningsfronten alltid före vad som finns tillgängligt i kommersiella produkter. För branschen är det dock omöjligt att överblicka den stora mängd nya metoder och algoritmer för dataanalys som publiceras varje år i tidskrifter eller på konferenser. Specifikt är det väldigt svårt att avgöra vilka nyheter som kan vara relevanta för den egna verksamheten, samt förstås bedöma vilka av alla dessa artiklar som är nydanande eller ens innebär en marginell förbättring mot existerande lösningar.

Samtidigt måste man vara medveten om att fördröjningen mellan att banbrytande algoritmer, metoder och tekniker publiceras, och det att de inkluderas i ledande kommersiella programvaror, typiskt är 5–10 år. Ett exempel på detta är algoritmen random forest (Breiman, 2001), vilken i dag är en standardalgoritm i alla verktyg för prediktiv modellering. Algoritmen publicerades 2001, och fanns då tillgänglig i mjukvara för forskare inom området, men det dröjde till 2007–2010 innan den dök upp i kommersiella analysverktyg.

På motsvarande sätt finns det i dag viktiga framsteg av karaktären ”etablerade sanningar” (för forskare inom dataanalys) som inte nått ut i kommersiella produkter. Ett sådant område är det matematiska ramverket conformal prediction (Vovk et al., 2004) där man, under väldigt generella antaganden, matematiskt kan garantera att en prediktion, med en vald sannolikhet, är korrekt. Vår forskningsgrupp arbetar intensivt, i ett flertal olika projekt, med att vidareutveckla conformal prediction. Arbetet bedrivs i samverkan med exempelvis Royal Holloway, Frederick University, Stockholms universitet, AstraZeneca och Scania, med en uttalad målsättning att även introducera ramverket i olika branscher, inklusive nödvändiga anpassningar. Vi tror att conformal prediction kommer vara ”allmänt accepterat” (och finnas tillgängligt i kommersiell programvara) inom fem år, men ser också möjligheter för företag att skapa sig fördelar genom att vara tidigt ute.

3.5.3 Bakgrund

Inte sällan används modeller och prediktioner i nästa läge som beslutsunderlag, exempelvis för kampanjplanering eller personifierade erbjudanden. Naturligtvis vill beslutsfattare då ha möjlighet att kunna jämföra olika alternativ, exempelvis utifrån förväntad vinst. Tyvärr blir detta ofta vanskligt då man inte kan kvantifiera säkerheten i olika prediktioner. Besluten fattas därför ofta på ett underlag där säkerheten inte bara är otillräcklig, utan där osäkerheten i sig är omöjlig att uppskatta. Conformal prediction introducerat av Vovk, Gammerman & Shafer (2005), är ett relativt nytt matematiskt ramverk som motverkar exakt det här problemet. Mer konkret kan man med conformal prediction, under mycket generösa antaganden, välja en acceptabel nivå för prediktionsfelet, och ramverket garanterar sedan att det faktiska felet kommer att närma sig denna nivå asymptotiskt. Priset som betalas för garantin är att prediktionerna blir multivärda

i stället för atomära – vilket i regression motsvaras av prediktionsintervall och i klassificering av prediktioner som består av en mängd klasser.

Conformal prediction har hittills oftast använts i säkerhetskritiska tillämpningar, men egenskapen att producera välkalibrerade sannolikheter ger förstås möjligheter i många områden. Målvariabeln i prediktion av för handeln typiska uppgifter, exempelvis churn, responsmodellering eller lifetime value kan direkt kopplas till antingen kostnader eller intäkter, vilket ger möjlighet att fatta informerade beslut. För att dessa kalkyler ska vara vettiga krävs dock en korrekt kvantifiering av sannolikheterna för att en prediktion är korrekt – eller omvänt uttryckt, utan denna egenskap kommer kalkylerna att vara inte bara osäkra utan i många fall direkt vilseledande.

3.5.4 Genomfört arbete

Det huvudsakliga arbetet med utveckling av conformal prediction har skett utanför FBI-projektet. Projektets forskare har dock de senaste åren publicerat ett stort antal artiklar inom området, exempelvis (Johansson, Boström & Löfström, 2013) och (Johansson et al., 2014b). Samtliga dessa studier har fokuserat på generella och tekniska aspekter av ramverket.

Vi är dock även under projektets sista månader applicerat conformal prediction på några typiska problem från handelsdomänen. Vi anser att resultaten är intressanta, och konceptet väckte också stort intresse på den publika workshop som projektet anordnade i Borås, i december 2015. Tyvärr är studien ännu inte inskickad för publicering. Trots detta väljer vi att inkludera delar från den studien i vår rapport, helt enkelt med motivet att vi anser att det ger en bra bild av möjligheterna med conformal prediction.

Metod

I studien appliceras conformal prediction på två olika problem; ett klassificeringsproblem och ett regressionsproblem. Klassificeringsproblemet är att predicera *churn*, alltså om en viss person kommer upphöra att vara kund. Tyvärr har vi i dagsläget inte vårt partnerföretags tillåtelse att redovisa vårt arbete med regressionproblemet, varför vi i denna rapport har ersatt det med ett välkänt benchmarking-problem kallat *Boston housing*, det vill säga att utifrån ett antal egenskaper hos fastigheter i Boston predicera dess värde.

Vid predicering av churn består datan av totalt instanser 136 000 instanser, det vill säga kunder. Dessa är uppdelade i en del för att bygga modellen (ungefär 55 000) och resten,används som testmängd. Varje kund är beskriven med 276 attribut. Samtliga kunder i denna studie har handlat minst fyra gånger hos vårt partnerföretag, en ledande e-handelsaktör. Churn definieras här som att kunden inte kommer att genomföra något köp inom ett år från senaste ordern. Tyvärr får vi inte beskriva attributen i detalj, men de inkluderar variabler som valt betalsätt, antal besök på hemsidan, antal erhållna mail från företaget som kunden öppnat etcetera. En väldigt viktig variabel är antalet dagar sedan senaste köp – vilket blir närmast självklart med den definition på churn som valts.

Boston housing är en mycket mindre datamängd med bara 506 instanser (fastigheter) där varje fastighet beskrivs med 13 attribut. I de tester som redovisas här användes 340 instanser för att bygga modellerna, och övriga 166 för testningen. Som underliggande modeller användes random forest för båda problem. Vid klassificering av churn prövades även logistisk regression, då det är en ofta använd enklare teknik. I experimentet testades även ett flertal olika signifikansnivåer, från 0,2 till 0,01.

Resultat

Nedan i tabell 1 visas 20 exempel på prediktioner från testmängden i Boston housing problemet. Den första kolumnen visar det korrekta värdet. Därefter följer fyra dubbelkolumner, en för vardera av de fyra signifikansnivåerna. De siffror som visas här är gränserna för prediktionsintervallet. Som synes blir förstås intervallen större ju säkrare vi kräver att vi ska vara. De felaktiga prediktionerna är markerade i rött, det vill säga för dessa tjugo exempelinstanser så görs fyra fel då signifikansnivån är 0,2, ett fel på nivån 0,1 och inget fel för 0,05 och 0,01.

Korrekt	ε = 0,2		ε = 0,1		ε = 0,05		ε = 0,01	
	Min.	Max.	Min.	Max.	Min.	Max.	Min.	Max.
10,8	6,7	23,2	2,7	27,3	0,0	31,0	0,0	40,7
14,9	9,9	26,4	5,8	30,4	2,1	34,1	0,0	43,8
12,6	10,4	26,3	6,6	30,1	3,0	33,7	0,0	43,0
14,9	16,8	30,2	13,5	33,5	10,5	36,5	2,6	44,4
12,7	8,7	22,1	5,4	25,4	2,4	28,4	0,0	36,3
20,0	11,8	28,2	7,8	32,2	4,1	35,8	0,0	45,5
16,4	15,6	32,1	11,5	36,1	7,8	39,8	0,0	49,6
20,2	14,0	28,2	10,5	31,7	7,3	34,9	0,0	43,3
19,1	9,2	25,6	5,2	29,6	1,5	33,3	0,0	43,0
20,1	11,7	28,1	7,7	32,1	4,1	35,8	0,0	45,4
19,9	10,2	26,5	6,2	30,5	2,5	34,2	0,0	43,9
23,0	12,9	29,2	8,9	33,2	5,2	36,9	0,0	46,6
23,7	20,5	36,4	16,7	40,2	13,1	43,8	3,8	53,1
21,8	13,1	28,5	9,4	32,2	6,0	35,7	0,0	44,7
20,6	13,0	29,4	9,0	33,4	5,3	37,1	0,0	46,7
19,1	11,1	27,4	7,1	31,4	3,4	35,1	0,0	44,8
15,2	10,3	26,8	6,3	30,8	2,6	34,5	0,0	44,3
7,0	7,7	24,2	3,6	28,2	0,0	31,9	0,0	41,6
24,5	18,0	23,4	16,6	24,8	15,4	26,0	12,2	29,2
11,9	17,8	24,1	16,3	25,6	14,9	27,1	11,1	30,8

Tabell 1. Exempelinstanser conformal regression.

En sammanfattning av resultaten för hela datamängden visas i tabell 2 nedan. Vi ser att även för så här få testinstanser så ligger de faktiska felnivåerna väldigt nära de valda signifikansnivåerna. Vi ser också att intervallens storlek ökar då en större säkerhet krävs.

	ε = 0,2	ε = 0,1	ε = 0,05	ε = 0,01
Andel fel	0,201	0,090	0,053	0,011
Medelintervall	10,1	16,0	19,4	32,8

Tabell 2. Sammanfattande resultat conformal regression.

Det är så klart omöjligt att avgöra hur informativa dessa prediktioner egentligen är utan att vara någorlunda insatt i boston housing problemet. Syftet med exemplet var därför bara att visa på hur conformal prediction fungerar för regression.

Tabell 3 visar 16 exempelprediktioner från churn-problemet. De möjliga alternativ som finns för conformal prediction vid klassificering med två klasser är att returnera båda klasserna, en av klasserna eller ingen av klasserna. Ett fel görs då prediktionsmängden inte innehåller den korrekta klassen. Som framgår i tabellen görs fler och fler dubbelprediktioner då kraven på säkerhet ökar – detta är analogt med större prediktionsintervall i regressionsfallet. Då vi tillåter 20 procent felaktiga prediktioner är nästan alla prediktioner enkla, det vill säga innehåller bara en klass, medan om vi bara tillåter 1 procent felaktiga prediktioner så är en stor majoritet av alla prediktioner dubbla. Detta är typiskt för hur man i conformal prediktion kan välja vilken säkerhet som krävs och därefter få anpassade prediktioner. Felaktiga prediktioner är markerade i rött – vi ser att det görs tre fel på signifikansnivån 0,2, två fel på nivån 0,1 och ett fel på nivån 0,05.

Tabell 4 sammanfattar resultaten för hela datamängden och de båda prövade teknikerna random forest och logistisk regression. För denna stora datamängd ligger de faktiska felnivåerna extremt nära de valda signifikansnivåerna. Vi ser även att andelen singelprediktioner förstås minskar för lägre signifikansnivåer. Om vi tillåter 20 procent fel så innehåller klart mer än 90 procent av alla prediktioner bara en klass, medan motsvarande siffra för signifikansnivån 0,01 är drygt 20 procent singleton-prediktioner. Vi noterar slutligen att för det här problemet är den enklare tekniken logistisk regression i stort sett lika effektiv som random forest.

Korrekt	ε = 0,2	ε = 0,1	ε = 0,05	ε = 0,01
Churn	{Churn}	{Churn}	{Churn}	{Churn}
No	{Churn}	{Churn}	{No, Churn}	{No, Churn}
No	{ }	{No}	{No}	{No}
Churn	{No, Churn}	{No, Churn}	{No, Churn}	{No, Churn}
Churn	{Churn}	{Churn}	{No, Churn}	{No, Churn}
Churn	{Churn}	{Churn}	{Churn}	{No, Churn}
No	{No}	{No}	{No, Churn}	{No, Churn}
Churn	{Churn}	{Churn}	{Churn}	{Churn}
No	{No}	{No, Churn}	{No, Churn}	{No, Churn}
No	{No}	{No}	{No}	{No, Churn}
Churn	{Churn}	{No, Churn}	{No, Churn}	{No, Churn}
Churn	{Churn}	{No, Churn}	{No, Churn}	{No, Churn}
No	{No}	{No}	{No}	{No}
Churn	{No}	{No}	{No}	{No, Churn}
No	{No, Churn}	{No, Churn}	{No, Churn}	{No, Churn}
No	{No}	{No}	{No}	{No, Churn}

Tabell 3. Exempelinstanser conformal klassificering.

	ε = 0,2	ε = 0,1	ε = 0,05	ε = 0,01
Random forest				
Andel singelprediktioner	0,939	0,666	0,481	0,209
Andel fel	0,202	0,100	0,052	0,010
Logistisk regression				
Andel singelprediktioner	0,925	0,653	0,475	0,210
Andel fel	0,199	0,096	0,050	0,011

Tabell 4. Sammanfattande resultat conformal klassificering.

Slutsatser

Prediktiv modellering ger möjligheter att ”räkna på”, och i förväg jämföra, värdet av olika beslut. Tekniken kan appliceras på många av de centrala processerna i handeln, och ger då ett väldigt bra beslutsunderlag – men det bygger på att modellerna är korrekta. Dock är prediktioner alltid osäkra, och specifikt kan den osäkerheten normalt inte kvantifieras – det vill säga beslutsunderlagen är egentligen mycket mer osäkra än vad de framstår. Ramverket conformal prediction löser precis det här problemet, då det matematiskt garanterar att andelen felaktiga prediktioner kommer att motsvara den valda signifikansnivån.

Vi har här demonstrerat hur conformal prediction kan användas för både prediktiv klassificering och regression. Resultaten visar att den uppmätta andelen fel mycket

riktigt ligger väldigt nära det förväntade. Det framgår också tydligt hur en användare kan få mindre eller större prediktioner (intervall eller mängder) genom att variera den acceptabla felnivån.

3.5.5 Rekommendationer

Vi rekommenderar alla de aktörer som i dag använder dataanalys för beslutsstöd att överväga införandet av conformal prediction. Vår uppfattning är att det i många fall signifikant skulle öka kvaliteten på beslutsunderlagen.

Referenser

Agrawal, R. och Srikant, R. (1994). Fast algorithms for mining association rules in large databases. VLDB '94 Proceedings of the 20th *International Conference on Very Large Data Bases*. Sid. 487–499.

Ahmed, S. R. (2004). Applications of data mining in retail business. *Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004. International Conference on.* Vol. 2, sid. 455–459.

Ahmed N. K., Atiya A. F., Gayar, N. E. och El-Shishiny, H. (2010). An Empirical Comparison of Machine Learning Models for Time Series Forecasting. *Econometric Reviews.* Vol. 29(5–6), sid. 594–621.

Andersson, E. (2008). Riktad Ica-reklam retar kunder. *Sydsvenskan.* 17 oktober, http://www.sydsvenskan.se/2008-10-17/riktad-ica-reklam-retar-kunder (2016-08-26).

Audzeyeva, A. och Hudson, R. (2015). How to get the most from a business intelligence application during the post implementation phase & quest; Deep structure transformation at a UK retail bank. *European Journal of Information Systems.* Vol. 24(1), sid. 1–18.

Berry, M. A. och Linoff, G. S. (2000). Mastering data mining: the art and science of customer relationship management. *Industrial Management & Data Systems*. Vol. 100(5), sid. 245–246.

Breiman, L. (2001). Random forests. *Machine Learning.* Vol. 45(1), sid. 5–32.

Cohen, W. (1995). Fast Effective Rule Induction. *Proceedings of the Twelfth International Conference on Machine Learning.* Sid. 115–123.

Coll, S. (2013). Consumption as biopower: Governing bodies with loyalty cards. *Journal of Consumer Culture.* Vol. 13(3), sid. 201–220.

DiNatali, N. och Ivarsdottir, M. (2015). *Perception meet Reality: A Pilot Study of the Self-congruence of Female Online Shoppers, with Regards to Fit, Size and Shape.* Högskolan i Borås. Master Thesis in Textile Management. Nr 2015.15.02.

Duhigg, C. (2012). How Companies Learn Your Secrets. *The New York Times Magazine*. 16 februari. http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html (2016-08-26).

Grewal, D. och Levy, M. (2007). Retailing research: Past, present, and future. *Journal of retailing.* Vol. 83(4), sid. 447–464.

Hagberg, J., Sundström, M. och Egels-Zandén, N. (2016). The digitalization of retailing: A review and framework. *International Journal of Retail & Distribution Management.* Accepted with revision.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. och Witten, I. H. (2009). The Weka Data Mining Software: An Update. *SIGKDD Explorations*. Vol. 11(1), sid. 10–18.

Hill, K. (2012). How Target Figures Out A Teen Girl Was Pregnant Before Her Father Did. *Forbes*. 16 februari. http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/ (2016-08-26).

Ingene, C. A. (2009) From Before the Beginning…to Beyond the Ending: Reflections on the Past, Present, and Future of the Journal of Retailing. *Journal of Retailing.* Vol. 85(4), sid. 510–518.

Ismail, M., Ibrahim, M. M., Sanusi, Z. M. och Nat, M. (2015). Data Mining in Electronic Commerce: Benefits and Challenges. *International Journal of Communications, Network and System Sciences*. Vol. 8(12), sid. 501.

Isukapalli, S. S. (1999). *Uncertainty Analysis of Transport Transformation Models*. PhD Thesis. Rutgers University.

Johansson, U. och Niklasson, L. (2009). Evolving Decision Trees Using Oracle Guides. *Computational Intelligence and Data Mining, 2009. CIDM '09. IEEE Symposium on,* Nashville, TN. Sid. 238–244.

Johansson, U., Sönströd, C., Löfström, T. och Boström, H. (2012). Obtaining accurate and comprehensible classifiers using oracle coaching. *Intelligent Data Analysis*. Vol. 16(2), sid. 247–263.

Johansson, U., Sönströd, C. och König, R. (2014). Accurate and Interpretable Regression Trees using Oracle Coaching. *Computational Intelligence and Data Mining (CIDM), 2014 IEEE Symposium on,* Orlando, FL. Sid. 194–201.

Johansson, U., Sönströd, C. och Linusson, H. (2014) Interpretable Streaming Regression Models with Local Performance Guarantees. *Big Data (Big Data), 2014 IEEE International Conference on,* Washington, DC. Sid. 461–470.

Kshetri, N. (2014). Big Data's Impact on Privacy, Security and Consumer Welfare. *Telecommunications Policy*. Vol. 38(11), sid. 1134–1145.

König, R. och Johansson, U. (2014). Rule extraction using genetic programming for accurate sales forecasting. *Computational Intelligence and Data Mining (CIDM), 2014 IEEE Symposium on,* Orlando, FL. Sid. 210–216.

Lubin, G. (2012). The Incredible Story Of How Target Exposed A Teen Girl's Pregnanacy. *Business Insider.* 16 februari. http://www.businessinsider.com/the-incredible-story-of-how-target-exposed-a-teen-girls-pregnancy-2012-2?IR=T (2016-08-26).

Luft, J. (1982). The Johari Window: A Graphic Model of Awareness in Interpersonal Relations. *NTL Reading Book for Human Relations Training.* NTL Insitute. http://www.convivendo.net/wp-content/uploads/2009/05/johari-window-articolo-originale.pdf (2018-08-16).

Meyer, D., Leisch, F. och Hornik, K. (2003). The support vector machine under test. *Neurocomputing.* Vol. 55(1–2), sid. 169–186.

Mori, M., MacDorman K. F. och Kageki, N. (2012). The Uncanny Valley [From the Field]. *IEEE Robotics & Automation Magazine*. Vol. 19(2), sid. 98–100.

Negash, S. (2004). Business intelligence. *The communications of the Association for Information Systems*. Vol. 13(1), sid. 177–195.

Nelson, M. R. (2008). The hidden persuaders: then and now. *Journal of Advertising.* Vol. 37(1), sid. 113–126.

Ngai, E. W., Xiu, L. och Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert systems with applications.* Vol. 36(2), sid. 2592–2602.

Packard, V. (1957). *The Hidden Persuaders*. New York: Pocket Books, 1958.

Packard, V. (1960). The Growing Power of Amen. I Sandage, C. H. och Fryburger, V. (red.) *The Role of Advertising in Society.* Homewood, IL: Richard D. Irwin. Sid. 266–274.

Phelps, J., Nowak, G. och Ferrell, E. (2000). Privacy Concerns and Consumer Willingness to Provide Personal Information. *Journal of Public Policy & Marketing.* Vol 19(1), sid. 27–41.

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

Quirk, M. B. (2012). Target Figures Out Teen Girl Is Pregnant Before Her Father Does, Sends Helpful Coupons. *Consumerism*. 17 februari. http://consumerist.com/2012/02/17/target-figures-out-teen-girl-is-pregnant-before-her-father-does-sends-helpful-coupons/ (2016-08-26).

Richards, T., Hamilton, S. och Yonezawa, K. (2015). *Retail Market Power in a Shopping Basket Model of Supermarket Competition.* Nr 1503. Working Papers from California Polytechnic State University. Department of Economics.

Saltelli, A., Chan, K. och Scott, E. M. (2000). *Sensitivity Analysis*. John Wiley & Sons.

Sandberg, E. och Abrahamsson, M. (2011). Logistics capabilities for sustainable competitive advantage. *International Journal of Logistics: research and applications.* Vol. 14(1), sid. 61–75.

Schankar, V. och Yadav, M. S. (2011). Innovations in Retailing. *Journal of Retailing.* Vol. 87(1), Editorial.

Stewart, K. A. och Segars, A. H. (2002). An Empirical Examination of the Concern for Information Privacy Instrument. *Information Systems Research*. Vol. 13(1), sid. 36–49.

Strong, C. (2015). *Humanizing Big Data. Marketing at the meeting of data, social science and consumer insight.* London: Kogan Page Ltd.

Sundell, H., Gidenstam, A., Papatriantafilou, M. och Tsigas, P. (2011). A lock-free algorithm for concurrent bags. *Proceedings of the Twenty-third Annual ACM Symposium on Parallelism in Algorithms and Architectures*. SPAA '11. Sid. 335–344.

Sundell, H., König, R. och Johansson, U. (2015). Pragmatic Approach to Association Rule Learning in Real-World Scenarios. *2015 International Conference on Computational Science and Computational Intelligence (CSCI)*, Las Vegas, NV. Sid. 356–361.

Sundström, M. och Radon, A. (2015). Don't Forget Consumer Value – Investigating consumer Attitudes toward QR-codes. *International Conference on Innovation and Management.* Singapore. 3–6 februari.

Sundström, M., Radon, A. och Wallström, S. (2016). Don't Forget Consumer Value – Investigating Consumer Attitudes towards QR-codes. *International Journal of Innovation in Management*. Vol. 3(2). Forthcoming.

Sundström, M. och Ericsson, D. (2012). Value Innovation and Demand Chain Management – keys to future success in the fashon industry. *The Nordic Textile Journal*. Special edition: Sustainability & Innovation in the Fashion Field. The Textile Research Centre, CTF. Sid. 82–90.

Sundström, M. och Ericsson, D. (2015). *Detaljhandel i förändring – Konsumentinsikt, värdenät och nya affärsmodeller*. I Solli, R. (red.) rapportserie i Styrning, Organisering och Ledning. Högskolan i Borås.

Van den Berg, B. (2010). The Uncanny Valley Everywhere? On Pricacy Perception and Expectation Management. I Fischer-Hübner et al. (red.) *Privacy and Identity Management för Life*. 6th IFIP WG 9.2, 9.6/11.7, 11.4, 11.6/PrimeLife International Summer School, Helsingborg, Sweden, August 2–6, 2010, Revised Selected Papers. Springer.

”Forskning för att stärka handelns konkurrenskraft och skapa goda villkor för branschens medarbetare.

Handelsrådet

Handelsrådet | 103 29 Stockholm
Besöksadress: Kungsgatan 24
Telefon växel 010-471 85 80
www.handelsradet.nu

www.fotoskrift.se | 1608