5.3.2017

Aalto University
School of Science

MS-E2177 Operaatiotutkimuksen projektityöseminaari

Projektisuunnitelma:

Asiakasarvon määrittäminen päivittäistavarakaupassa

Projektiryhmä	Joonas Laihanen (projektipäällikkö) Aleksi Pasanen Eero Rantala Samuli Turunen
Aiheen asettaja	Kesko

Työn aihe ja taustat

Yritys voi hyödyntää parempaa ymmärrystä asiakkaiden käyttäytymisestä monin tavoin. Päivittäistavarakaupassa tämä voi tarkoittaa esimerkiksi markkinoinnin tehokkaampaa kohdentamista ja yksilöidympää palvelua. Toisaalta on mahdollista ennustaa asiakkaiden käyttäytymistä kerätyn tiedon perusteella toiminnan suunnittelussa. Päivittäistavaralla tarkoitetaan yleisesti elintarvikkeita ja ruokaostosten yhteydessä hankittavia kulutustavaroita.

Aiheen asettajayritys Kesko (myöhemmin myös “yritys”) omistaa kotimaiset K-kauppaketjut. Tämä projektityö keskittyy K-kaupoissa tehdyistä päivittäistavaraostoksista saatuun dataan. Päivittäistavarakaupassa tavallinen tapa kerätä tietoa asiakkaan käyttäytymisestä on hyödyntää hänen kanta-asiakkuuttaan. Bonuskortin lukeminen ostosten maksun yhteydessä tallentaa hyödyllistä tietoa kauppiaan tietojärjestelmiin. Keskon Plussa-asiakkuus on yrityksen oma bonusjärjestelmä. Plussa-asiakkaat saavat esimerkiksi ajoittain alennusta tietyistä tuotteista, kohdennettuja tarjouksia, sekä ostohyvityksiä niin sanottujen Plussa-pisteiden muodossa. Noin 2,2 miljoonalla taloudella on Plussa-kortti.

CLV (asiakasarvo, eng. customer lifetime value) on tapa mitata asiakkaasta tulevaisuudessa saatavaa tuottoa. Usein CLV määritellään asiakkaasta saatavana voittona, mutta tässä projektissa CLV:tä käsitellään asiakkaan tekeminä ostoina. Myös CLV:n käsittämä aikahorisontti voi vaihdella.

Yrityksen kanta-asiakkuuksien suuri määrä antaa hyvän pohjan tilastolliselle mallintamiselle. Projektissa kehitetään malli asiakastalouden CLV:n ennustamiselle seuraavan 5-10 vuoden ajalle. Projektin laajuus ei sisällä esimerkiksi mallin liiketoimintamahdollisuuksien määrittämistä, asiakastalouksien elämäntilanteen mallintamista, päivittäistavarakaupan tai väestörakenteen kehityksen arviointia tai graafisella käyttöliittymällä varustetun valmiin työkalun rakentamista. Mallin on tarkoitus tarjota odotusarvoon perustuva ennuste, joka saattaa poiketa toteutuneesta ostosten määrästä. Asiakasyritykseltä saatava data koostuu talouskohtaisesta ostosten loppusummista kuukausitasolla; tiedosta asiakastalouden kortinhaltijan iästä, asuinalueesta, sekä yrityksen itse mallintamista ominaisuuksista kuten elämänvaihe ja niin sanottu SoW (asiakasosuus, eng. Share of Wallet). Myös tarkempaa tuotekohtaista ostodataa on olemassa, mutta sitä ei oletusarvoisesti käytetä tässä työssä, ellei sitä koeta tarpeelliseksi jo tehdyn mallin jatkokehittämisessä.

Tavoitteet

Projektin aihe on määritelty varsin tarkasti asiakasyhtiön toimesta. Työn tavoitteena on kehittää Markov-ketjun simulointiin perustuva malli CLV:n määrittämiseksi, hyödyntäen yritykseltä saatua asiakasdataa. Asiakasyhtiö käyttää sitten saatua mallia liiketoimintansa kehittämiseen näkemällään tavalla tämän projektin valmistuttua. Työ mallin luomiseksi voidaan jakaa karkeasti kolmeen välitavoitteeseen, jotka ovat:

1. Ostoihin vaikuttavien tekijöiden tutkiminen ja määrittely

2. Markov-ketjun tilojen ja siirtymätodennäköisyyksien määrittely
3. Ennustavan mallin rakentaminen ja CLV:n laskeminen

CLV:n määrittäminen päivittäistavaramyynnissä aloitetaan tutustumalla siihen vaikuttaviin tekijöihin datan ja kirjallisuuden avulla. CLV:tä voidaan mallintaa useilla tavoilla vaihtelevin selittävin tekijöin. Kirjallisuuskatsauksen avulla etsitään tukea asiakkaan dataan perustuvan mallin kehittämiseen ja mahdollisia datan ulkoisia selittäjiä. Esimerkiksi varsinaisen datan ulkopuolelta ihmisten tulotasoista on avointa dataa [1] postinumeroihin perustuen, jota voidaan käyttään Keskon tarjoaman datan ohella tarvittaessa.

Tässä Markov-ketjulla mallinnetaan asiakkaan käyttäytymistä estimoimalla todennäköisyyttä siirtyä tilasta toiseen diskreetillä aikavälillä, kun oletetaan tilan muutoksen todennäköisyyden riippuvan ainoastaan tilasta nykyhetkellä [2]. Alkutilajakaumana käytetään tilojen nykyistä jakaumaa ja simuloimalla Markov-ketjua eteenpäin halutun ajanjakson 5-10 vuotta verran saadaan ennuste tulevaisuuden tilajakaumalle, josta voidaan CLV laskea suoraviivaisesti. Markov-ketjun tiloja voivat olla esimerkiksi talouksien elämäntilanteet ja kortinhaltijan ikä. Projektin tavoitteena on kuitenkin löytää jokin parempia ennusteita tuottava tarkempi luokittelutapa Markov-ketjun käyttöön.

Mallin rakentamisessa pyritään ennustamiskykyiseen algoritmiin, eli mallin halutaan toimivan mahdollisimman hyvin myös uudelle datalle. Sen kannalta asiakkaiden luokittelu optimaalisella tavalla on erittäin tärkeää. Se, että jokainen nykyinen asiakas olisi oma luokkansa ei ole kovin mielekäs luokittelutapa, koska mallin tulee antaa toimivia ennusteita uusillekin asiakkaille. Toisaalta laskennalliset rajoitteetkin kannustavat asiakkaiden luokitteluun, eli toisin sanoen pienentämään laskennallisen ongelman dimensioita.

Luokittelu on mahdollista tehdä pääpiirteissään kahdella periaatteella: algoritmisesti johonkin sopivaan menetelmään tukeutuen ilman manuaalista päättelyä, tai luokittelu rakennetaan päätellen datasta merkittävimmät CLV:n tekijät ja niiden riippuvuussuhteet. Algoritmiset menetelmät ovat tehokkaita, mutta niiden tuloksena olevat luokittelut voivat olla hyvin vaikeita ymmärtää. Toisaalta ihmisen päättelykyky ei välttämättä löydä parhaita mahdollisia luokkia. Työssä rakennetaan niin sanottu baseline-malli perustuen yksinkertaiseen päättelyyn, jonka jälkeen pyritään parempiin tuloksiin helposti tulkittavien algoritmisten luokittelumenetelmien avulla. Vahvana ehdokkaana luokittelun apuvälineeksi pidetään päätöspuita ja klusterointialgoritmeja tai muita koneoppimisen menetelmiä [3]. Mallin tarvitsema Markov-tilansiirtomatriisi on laskettavissa luokitellusta datasta, kuten odotusarvoinen CLV:kin [2]. Voidaan siis todeta tavoitteiden 2 ja 3 olevan suoritettavissa ainakin osittain samanaikaisesti. Tavoitteen 2 saavuttaminen parhaalla mahdollisella tavalla tukee vahvasti tavoitteen 3 saavuttamista.

Lisäksi tavoitteisiin edetään askel kerrallaan rakentamalla ensin toimiva baseline-malli eli yksinkertainen ja suhteellisen nopeasti toteutettavissa oleva malli, jota kehitetään ajan puitteissa niin, että oletusarvoisesti luokkien määrä kasvaa, ennustettavuus paranee ja myös ulkoisia tekijöitä varsinaisen datan ulkopuolelta otetaan mukaan mallinnukseen. Mallin

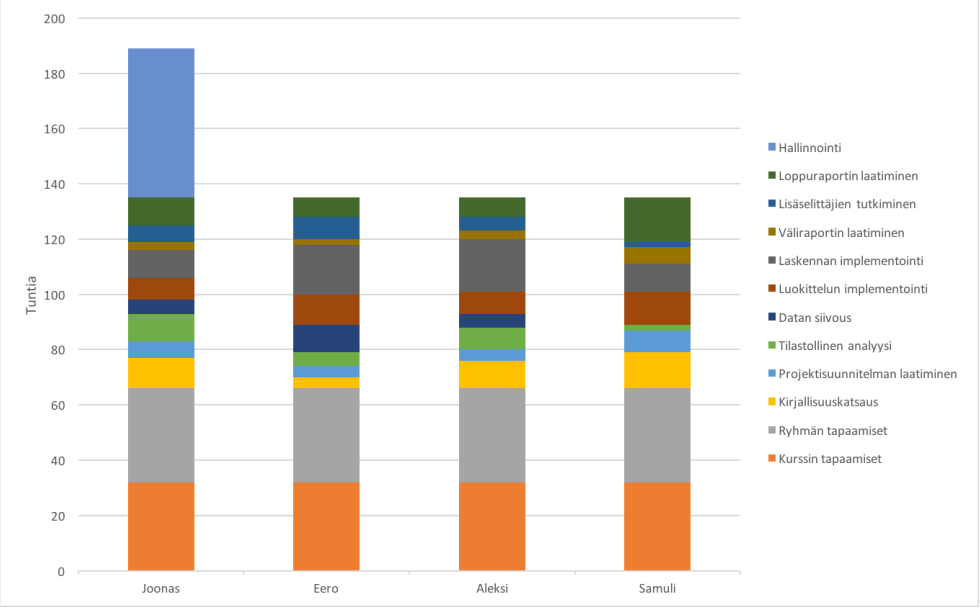
toimivuus validoidaan ja laskennan oikeellisuus verifioidaan historiaan perustuvalla testidatalla sekä erillaisilla luokitteluilla laskettujen ennusteiden vertailulla.

Tehtävät ja aikataulu

Kurssin ja projektin suorittaminen on jaettu yhteensä 12 tehtävään, joista osa liittyy kurssin pakollisiin osasuorituksiin (projektisuunnitelma, väliraportti, loppuraportti ja näihin liittyvät esitykset ja yritysekskursiot) ja osa puolestaan projektin tavoitteiden saavuttamiseen. Projektipäälliköllä on 54 tuntia projektin hallinnointia vastaten hänelle myönnettävää kahta ylimääräistä opintopistettä. Muiden ryhmän jäsenten kokonaistyömäärä on 135 tuntia, joka vastaa viittä opintopistettä. Kuvassa 1 on esitetty tehtäviin allokoidut tuntimäärät ryhmän jäsenten tasolla. Vastuualueet voidaan myös jakaa laajemmin taulukossa 1 esitetyllä tavalla.

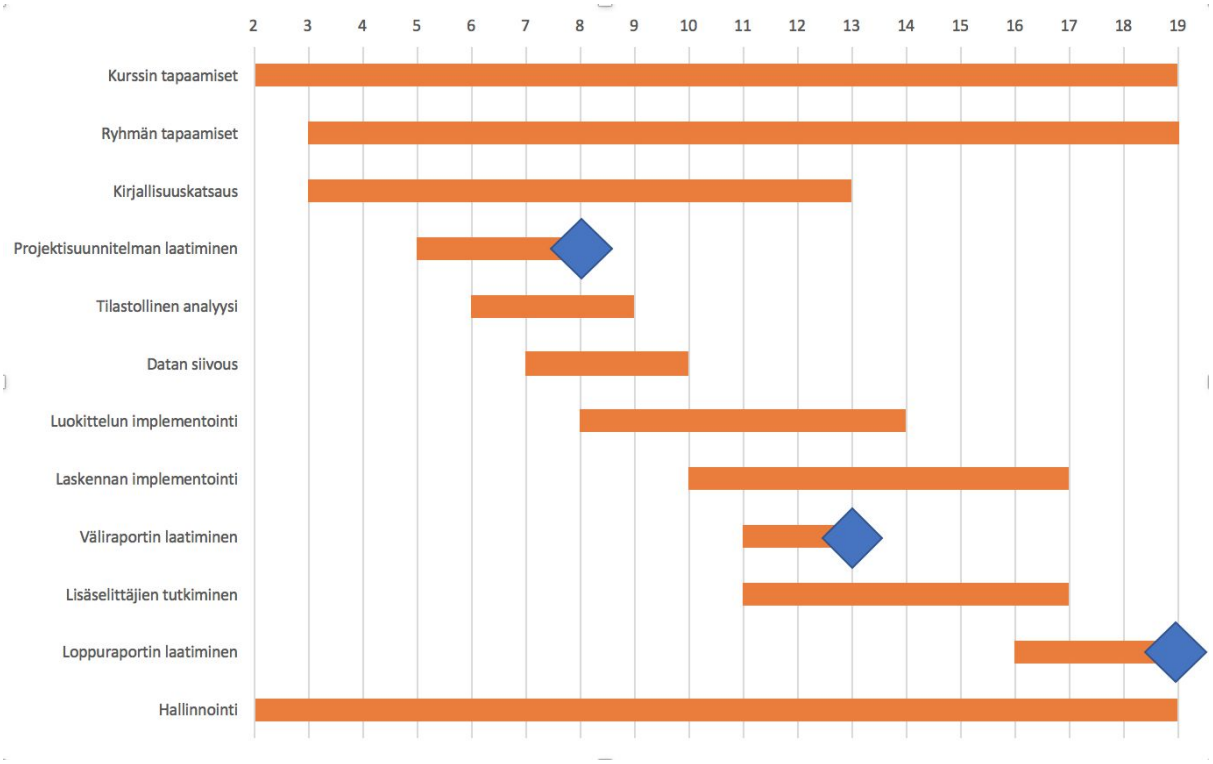
Taulukko 1. Ryhmän jäsenten erityisvastuualueet.

Jäsen	**Vastuualue**
Joonas	Hallinnointi, kokonaisuus
Aleksi	Implementointi, tilastollinen analyysi
Eero	Datan esikäsittely ja ulkoiset selittäjät, implementointi
Samuli	Raportointi, luokittelu



Kuva 1. Tehtävien jako ja allokoidut työtunnit ryhmän jäsenille.

Kirjallisuuskatsaus sisältää aiheeseen tutustumisen sekä mallin kehittämisen aikana tehtävän lisäselvityksen tarvittaessa. Tilastollisessa analyysissä tuotetaan deskriptiivistä statistiikkaa käytössä olevasta datasta. Datan siivouksessa poistetaan selkeät poikkeamat ja virheelliset havainnot sopivalla tavalla. Implementoinnilla tarkoitetaan algoritmin kehittämistä. Luokittelun tulee toimia ennen Markov-ketjun ja CLV:n laskemista, johon viitataan laskennan implementointi -tehtävällä. Lisäselittäjien tutkiminen sisältää mahdollisten datan ulkoisten CLV:n selittäjien tutkimisen ja soveltamisen. Kurssin kesto on noin 17 viikkoa ja loppuraporttien esittäminen tapahtuu 19. toukokuuta, mikä on koko projektin viimeinen määräaika. Edellä mainitut tehtävät on aikataulutettu kurssin keston ajaksi siten, että projekti etenee sujuvasti ja jokaisen tehtävän alkaessa on saatu riittävään valmiuteen välttämättömät muut tehtävät. Kurssin aikataulu on esitetty kuvassa 2.



Kuva 2. Tehtävien aikataulutus. Vinoneliöt kuvaavat kurssin tehtävien palautuksia.

Resurssit ja riskit

Merkittävimpänä resurssina ovat projektiryhmäläisten edellä eritelty aika työskennellä projektin parissa ja heidän osaamisensa operaatiotutkimuksen ja data-analyysin saralla. Siten erityisesti teknisissä asioissa ryhmän jäsenten Aallossa käymien kurssien oppimateriaalit ovat hyödyllisiä. Lisäksi tiedonhankinnassa käytetään monipuolisesti alan kirjallisuutta ja tieteellisiä julkaisuja, ja projektissa voidaan hyödyntää sidosryhmien, kurssin vetäjän professori Salon, sekä asiakasyhtiö Keskon edustajien kokemusta, unohtamatta yhteistyötä muiden ryhmien, erityisesti opponointiryhmän (Fortum) antamaa palautetta.

Asiakasyritys antaa projektiryhmän käyttöön dataa erityyppisten asiakkaiden ostosten suuruudesta kahden vuoden ajalta kuukausitasolla, joka on projektin loppuun saattamisen

kannalta välttämätöntä. Myös muunlaisia datalähteitä on tarjolla, niin Keskon tarjoamia kuin myös täysin ulkopuolisia [1], mikäli ne koetaan kiinnostaviksi. Datan luottamuksellisuuden takia kyseiseen resurssiin on kiinnitettävä erityistä huomiota. Datan käsittely ja laskenta tehdään käyttäen ainakin R-ohjelmistoa ja laskenta suoritetaan Aallon tietokoneilla. Se, kuinka oleellisia eri ohjelmistojen ominaisuudet ja laskentaan käytettävien koneiden tehokkuus ovat, riippuu pitkälti mallin laskennallisten ratkaisujen valinnoista, jotka muovautuvat projektin edetessä lopulliseen muotoonsa.

Merkittävin riski projektissa on, malli ei suoriudu riittävän hyvin. Aluksi rakennettavan baseline-mallin ei odotetakaan olevan kovin tarkka, joten sen odotetaan syntyvän suhteellisen vaivatta. Paranneltu malli voi suoriutua heikosti useista syistä, mutta merkittävimmät tekijät lienevät joko luokittelun epäonnistuminen tai siirtymämatriisin simuloinnin epäonnistuminen. Luokittelun odotetaan olevan haastava tehtävä, mutta riskiä hallitaan lähestymällä ongelmaa rinnakkain päättelemällä luokittelun parametrit kirjallisuuden avulla sekä algoritmisella menetelmällä. Simuloinnin ongelmat liittynevät kohinaan datassa, puuttuviin havaintoihin ja erittäin pieniin siirtymätodennäköisyyksiin. Laskennan riskejä hallitaan tavoittelemalla luokittelua, josta laskettava tilansiirtomatriisi on dimensioiltaan pieni. Täysin toimimaton malli odotetaan syntyvän vain äärimmäisessä tapauksessa, eli käytännössä työn jäävän kokonaan kesken.

Vaikutuksiltaan vakava riski on myös luottamuksellisen tiedon joutuminen väärään paikkaan tai ulkopuolisten tietoon. Riskin realisoitumista on jo alkuvaiheessa pyritty estetty useilla toimenpiteillä.

CLV:n lisäselittäjien etsintä on sekundäärinen tavoite, mutta samalla heikoiten määritelty. Riskinä on, ettemme löydä kirjallisuuteen perustuvia ja dataan soveltuvia lisäselittäjiä tai että niiden ennustuskyky on erittäin heikko.Tämän takia keskitytään malleihin, joissa selittäjien määrä on pieni. Riskejä on eritelty taulukossa 2.

Taulukko 2. Työn riskit.

Riski	Todennäköisyys	Vakavuus	Vaikutus	Toimenpiteet
Heikko kommunikointi asiakkaan kanssa	Matala	Keskitaso	Työn lopputulos ei ole halutunlainen tai työn eteneminen kärsii (kuitenkin tehtäväksianto on jo annettu tässä vaiheessa)	Kommunikoidaan asiakkaan kanssa säännöllisesti ja järjestetään myös tapaamisia
Heikko kommunikointi ryhmän kesken	Matala	Keskitaso	Työn eteneminen kärsii (kuitenkin tavoitteet on jo määritelty tässä vaiheessa)	Säännölliset tapaamiset, tiedon ja tulosten jakaminen aktiivisesti
Luokittelu osoittautuu erittäin vaikeaksi	Korkea	Korkea	CLV:tä ei saada laskettua mielekkäällä tarkkuudella	Sovelletaan aluksi ns. baseline-malli, jota laajennetaan ja lopuksi sovelletaan (osin) algoritmista menetelmää
Markov-ketjun laskenta osoittautuu erittäin vaikeaksi	Matala	Korkea	CLV:tä ei saada laskettua mielekkäällä tarkkuudella	Rajataan tilansiirtomatriisin tilojen määrä pieneksi, huomioidaan tilansiirtomatriisin laskennalliset rajoitteet luokittelussa
Projektin hallinnan ongelmat	Keskitaso	Keskitaso	Työn aikataulussa ei pysytä tai työn laajuus ei pysy hallinnassa	Tehdyn aikataulun ja suunnitelman noudattamista seurataan ja tuetaan säännöllisellä kommunikoinnilla ja tapaamisilla.
Luottamuksellisen tiedon vuoto	Hyvin matala	Hyvin korkea	Luottamuksellista tietoa joutuu väärään paikkaan tai ulkopuolisten tietoon	Luottamuksellinen data tallennetaan vain sitä varten perustetulle tallennusmedialle. Tapaamiset pidetään suljettujen ovien takana. Kirjallisissa tuotoksissa esitetään vain koonteja, eikä yksityiskohtaista dataa.

Lähteet

[1] Tilastokeskus: Paavo – Postinumeroalueittainen avoin tieto. Saatavissa: http://www.stat.fi/tup/paavo/index.html

[2] Cheng, C.-J., Chiu, S., Cheng, C.-B., Wu J. Customer lifetime value prediction by a Markov chain based data mining model: Application to an auto repair and maintenance company in Taiwan, 2011. Saatavissa: http://www.sciencedirect.com/science/article/pii/S1026309812000107/pdfft?md5=39b0c933456b38a137399deb37b9b7e1&pid=1-s2.0-S1026309812000107-main.pdf

[3] Hruschka, H., Natter,M., Comparing performance of feedforward neural nets and K-means for cluster-based market segmentation, 1999. Saatavissa: http://www.sciencedirect.com/science/article/pii/S0377221798001702