Kandidatuppsats i Statistik

Inkrementell responsanalys av Scandinavian Airlines medlemmar

Vilka kunder ska väljas vid riktad marknadsföring?

Erika Anderskär
Frida Thomasson

Avdelningen för Statistik och maskininlärning
Institutionen för datavetenskap
Linköpings universitet

Vårterminen, 2017

ISRN: LIU - IDA/STAT - G - - 17/005 - SE

Handledare: Måns Magnusson

Examinator: Annika Tillander

Abstract

Scandinavian Airlines has a large database containing their Eurobonus members. In order to analyze which customers they should target with direct marketing, such as emails, uplift models have been used. With a binary response variable that indicate whether the costumer has bought or not, and a binary dummy variable that indicates if the customer has received the campaign or not conclusions can be drawn about which customers are persuadable. That means that the customers that buy when they receive a campaign and not if they don't are spotted. Analysis have been done with one campaign for Sweden and Scandinavia. The methods that have been used are logistic regression with Lasso and logistic regression with Penalized Net Information Value. The best method for predicting purchases is Lasso regression when comparing with a confusion matrix. The variable that best describes persuadable customers in logistic regression with PNIV is `Flown` (customers that have flown with SAS within the last six months). In Lasso regression the variable that describes a persuadable customer in Sweden is membership level 1 (the first level of membership) and in Scandinavia customers that receive campaigns with delivery code 13 are persuadable, which is a form of dispatch.

Sammanfattning

Scandinavian Airlines har en stor kunddatabas som de vill använda för att lokalisera vilka kunder som ger högst lönsamhet vid marknadsföring. Detta sker genom riktade marknadsföringskampanjer till eurobonusmedlemmar i Scandinavian Airlines kundsystem. Responsvariabeln i datamaterialet är binär och indikerar på om kunden har genomfört ett köp eller inte efter utskicket. Det finns även en binär variabel som indikerar på om kunden ingår i kontrollgruppen eller kampanjgruppen. Detta för att kunna dra slutsatser om vilka kunder som är påverkningsbara, det vill säga köper när de får en kampanj, men inte när de inte får en kampanj. Analyser har gjorts på en kampanj för Sverige och Skandinavien. Metoden som används i uppsatsen är logistisk regression med två olika typer av variabelselektion, Lasso och Penelized Net Information Value. Den metod som predikterar köp bäst i datamaterialet är Lassoregression vid jämförelse med hjälp av en förväxlingsmatris. Den variabel som bäst förklarar vilka kunder som är påverkningsbara av marknadsföring i både Sverige och Skandinavien enligt logistisk regression med PNIV är `Flugit` (kunder som flugit med Scandinavian Airlines under de senaste sex månaderna). Enligt Lassoregressionen är det medlemmar som har medlemsnivå 1 (första nivån av medlemskap) som är mest påverkningsbara av marknadsföring i Sverige och i Skandinavien är påverkningsbara medlemmar de som har fått kampanjer med leveranskod 13, som är en form av utskicket som gjorts.

Förord

Uppsatsen har gjorts som en kandidatuppsats på programmet Statistik och dataanalys vid Linköpings universitet. Uppdragsgivaren är flygbolaget Scandinavian Airlines. Vi vill rikta ett stort tack till Scandinavian Airlines för förtroendet och möjligheten till att ha fått göra detta arbete på deras kunddata, och ett extra stort tack till Mattias Andersson, analyschef för CRM avdelningen på SAS och Botan Calli, Analytiker, för att de tagit sig tid att handleda detta arbete med stort engagemang och kunskap inom området.

Utöver dessa vill vi även rikta ett tack till Mathias Lanner analytiker på SAS Institute för givande samtal och information kring modeller och problemhantering.

Vi vill även tacka vår handledare vid Linköpings Universitet, Måns Magnusson, för hjälpsamma möten, kommentarer och stöd. Sist men inte minst vill vi rikta ett tack till Josefin Enoksson och Sofia Olausson för genomgång av uppsatsen samt goda insikter kring förbättringsområden.

Erika Anderskär och Frida Thomasson Linköpings universitet, Juni 2017

Innehåll

Figurer

Tabeller

Centrala begrepp

Kampanjgrupp/Exprimentgrupp	Individer som har fått kampanjen
Kontrollgrupp	Individer som inte har fått kampanjen
True positive/Sann positiv (TP)	Observationer som klassificeraren har klassat till 1 och som är 1
True negative/Sann negativ (TN)	Observationer som klassificeraren har klassat till 0 och som är 0
False positive/Falsk positiv (FP)	Observationer som klassificeraren har klassat till 1, men som egentligen är 0
False negative/Falsk negativ (FN)	Observationer som klassificeraren har klassat till 0, men som egentligen är 1

1. Introduktion

Scandinavian Airlines är det ledande flygbolaget i Skandinavien. Bolaget skapades 1946 genom en sammanslagning av flera flygbolag i Danmark, Norge och Sverige. Verksamhetsåret 2014/2015 flög 28,1 miljoner passagerare med Scandinavian Airlines till 119 olika destinationer i Europa, USA och Asien. Medlemskapet i Star Alliance ger kunderna bra förbindelser världen över då Star Alliance har totalt mer än 18500 avgångar varje dag till 1300 destinationer i 192 länder världen över. (SAS, 2017)

Scandinavian Airlines har ett medlemsskapssystem, Eurobonus, som ger kunder möjlighet att samla poäng på sina köp och få speciella erbjudanden. Många av dessa erbjudanden skickas ut per e-post till eurobonusmedlemmarna. Värt att notera är att Scandinavian Airlines även publicerar erbjudanden från utskicken på deras hemsida, vilket gör att kampanjerna inte är unika utan går att tillgå även för personer som inte är medlemmar i Eurobonus.

Erbjudanden, eller kampanjer, kan få en positiv effekt på försäljningen, men det kan också få en negativ effekt, om kunder upplever e-posten som ett störande moment och väljer att avsluta sina utskick från Scandinavian Airlines. För att få en så hög lönsamhet på marknadsföringen som möjligt är det viktigt att rikta sig till rätt kunder. Utskicken ska helst bara gå till påverkningsbara kunder, det vill säga kunder som genomför ett köp om de får ett utskick, men inte annars. Dessa kunder ger en ökad respons och lönsamhet av en kampanj. För att hitta dessa kunder kan inkrementella responsmodeller användas.

Inkrementella responsmodeller beräknar uplift, detta definieras som skillnaden mellan gruppen som får ett utskick och en kontrollgrupp som inte får ett utskick.

Tabell 1.1: Förklaring till en generell beräkning av uplift

Grupp	Antal kunder	Antal köp	Köp i procent
Kampanj	10000	2000	20%
Kontroll	10000	1200	12%

Svarar 20 procent av kampanjgruppen på erbjudandet i utskicket och 12 procent av kontrollgruppen så är det en uplift på 8 procentenheter. Andra ord för uplift är inkrementell respons, true lift, netlift, och incremental value.

Datamaterialet består av en binär responsvariabler som indikerar om kunder har genomfört köp eller inte. En indikatorvariabel som visar om kunden ingår i kampanj- eller kontrollgrupp, samt ett antal förklarande variabler. Bland de förklarande variablerna finns mosaikdata som Scandinavian Airlines har köpt in för att utöka sin databas, dock finns de inte att tillgå för hela Skandinavien. Därför delas datamaterialet in i två delar, Sverige och Skandinavien (Sverige, Norge och Danmark), för att kunna undersöka så stor mängd data som möjligt.

1.1 Tidigare studier

För att kunna karaktärisera kunder som köper på grund av en kampanj används ofta responsmodeller (*response modeling*). Responsmodeller försöker beräkna sannolikheten att en kund genomför ett köp om kunden får ett utskick. (Surry et al., 2011).

Inkrementella responsmodeller försöker istället att beräkna ökningen i sannolikhet för köp om en kund får ett utskick jämfört med om kunden inte får ett utskick. För att kunna mäta detta används en kontrollgrupp. Då jämförs effekterna av kampanjgruppen mot kontrollgruppen. En indikatorvariabel skapas som anger om kunden tillhört kontrollgrupp eller kampanjgrupp. (Surry et al., 2011).

Vid anpassandet av en inkrementell responsmodell kan två olika typer av modeller användas: separata modeller och gemensamma modeller. En separat modell anpassas genom att två olika modeller skattas, en med kampanjgruppen och en med kontrollgruppen. Därefter subtraheras de predikterade värdena från de två modellerna för att prediktera uplift. Fördelarna med denna modell är att den bland annat är enkel och inte kräver nya metoder eller mjukvara. Nackdelen är att det sällan fungerar speciellt bra i praktiken, exempelvis för att två välanpassade modeller inte garanterar att differensen mellan dem också är välanpassad. (Surry et al., 2011).

Den gemensamma modellen använder indikatorvariabler och anpassar en modell på hela datamängden. En gemensam modell presterar ofta bättre, men problemet med dessa är att det är svårt att mäta skillnaden, eftersom det inte går att mäta på individnivå (en individ kan inte vara med i både kampanjgruppen och kontrollgruppen samtidigt). (Surry et al., 2011)

Ett tillvägagångssätt för att skatta den inkrementella responsen är att använda logistisk regression som klassificerare (Lo, 2002). Den logistiska regressionen analyserar olika förklaringsvariablers inverkan på en binär responsvariabel (köp eller inte köp).

Karlsson et al. (2013) testar flera olika modeller för att skatta uplift. De testar att anpassa en separat modell med beslutsträd, en gemensam modell med logistisk regression (både frekventistisk och bayesiansk) och en gemensam modell med lassoregression. Samtliga modeller misslyckas med att klassificera responsvariabeln sett till utförda tester, då den prediktiva förmågan hos modellerna är inte speciellt bra. Dock är lassoregressionen den modell som anses vara mest lämpad. Det stora datamaterialet, som innehåller observationer från en period på 9 år, diskuteras som en eventuell orsak till felet. Innehåller ett datamaterial många olika kampanjer är det svårt att hitta något intressant. Det är lättare att få fram ett bra resultat om de istället koncentrerar sig på en specifik kampanj och försöker anpassa en prediktiv modell, som i sin tur kan användas vid framtida, liknande kampanjer.

De test Karlsson et al. (2013) använder är felkvot, Area Under Curve, Root Mean Squared Error etc. Testen utvärderar modellernas förmåga att prediktera köp, inte uplift, då det inte finns något bra, objektivt mått för att jämföra uplift i olika modeller.

Larsen (2010) har redovisat noden (incremental response) i Sas Enterprise Miner och presenterat Penalized Net Information Value (PNIV) som är ett informationsmått som kan användas för variabelselektion. PNIV mäter korrelationen mellan förklaringsvariabeln och den binära responsvariabel. Med PNIV rankas variablerna och de som bidrar mest till uplift väljs ut.

1.2 Syfte

Uppsatsen avser att anpassa modeller på medlemsdata från Scandinavian Airlines som kan identifiera medlemmar som köper när de får ett utskick, men inte annars. Detta för att få en så hög lönsamhet av marknadsföringen som möjligt. Inkrementella responsmodeller anpassas på datamaterialet för att undersöka vilka egenskaper medlemmar som genererar hög uplift besitter.

1.2.1 Frågeställningar

Rapportens mål under arbetets gång är att besvara frågeställningarna:

- Vilka variabler är lämpade för att förklara uplift hos Scandinavian Airlines kunder i Sverige och i Skandinavien?
- Vilken metod är mest lämpad för att analysera köp av medlemmar i Sverige och Skandinavien i Scandinavian Airlines datamaterial?
- Finns det likheter mellan Skandinavien och Sverige när det gäller variabler som förklarar uplift?

1.3 Etiska aspekter och samhällsnytta

Uppsatsens författare ansvarar för en god kvalité på arbetet och att forskningen är moraliskt korrekt. Uppsatsen får inte skada någon fysiskt eller psykiskt eller på annat sätt kränka samhällets människor. (Codex, 2016)

Kunder som blir medlemmar i Eurobonus behöver inte ange uppgifter specifikt knutna till dem, såsom personnummer. Detta medför att det inte går att identifiera kunderna i datamaterialet till en fysisk person. Medlemsnumret, som är unikt för varje kund, har kodats om i datamaterialet så att enbart Scandinavian Airlines kan koppla numret till en e-postadress.

En lyckad inkrementell responsmodell skulle kunna innebära för företaget ett effektivare arbete i form av att inte behöva lägga tid på kunder som inte är påverkningsbara av utskicken. Detta skulle i sin tur kunna generera en högre avkastning för företaget om endast de kunder som tenderar till att handla på kampanjer lokaliseras. Detta kan resultera i att företaget slipper lägga tid på att skicka marknadsföring till de kunder som ändå inte kommer att generera någon avkastning. Från kundernas synvinkel så återfinns nyttan i att endast de kunder som vill ha utskicken får dessa.

2. Data

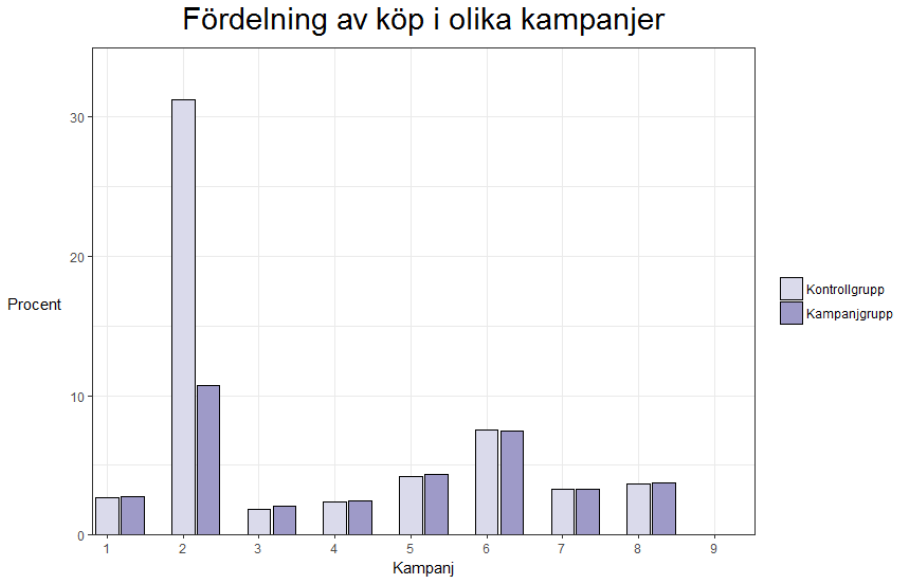
Datamaterialet som används i uppsatsen kommer från Scandinavian Airlines kunddatabas och innehåller enbart eurobonusmedlemmar. De förklarande variablerna omfattar exempelvis vilka

kampanjer som skickats, vart kunderna bor och om de tillhör kontrollgrupp eller kampanjgrupp. Samtliga variabler med förklaring redovisas i bilaga A. Datamaterialet omfattar även demografi om kunden i form av bland annat kön och ålder. Det är även utvidgat med så kallad mosaikdata (bakgrundinfo) som Scandinavian Airlines köpt in för att utvidga och förädla kunskapen kring kunderna. Förklaring av mosaik återfinns i bilaga C.

När det kommer till inkrementella responsmodeller behövs en responsvariabel som är binär, om kunden köpt eller inte köpt.

Kampanjerna som data är uppbyggt utefter är så kallade priskampanjer som ger kunderna erbjudanden, som skickas ut vid olika tidpunkter. Där det beteende som kunden påvisar vid utskicken sammanställs. Data är inhämtat under en 3-månadersperiod mellan 4:e oktober och 15:e december år 2016, som genererat 9 olika kampanjutskick.

Datamaterialet innehåller ungefär 1,85 miljoner unika kunder. Totalt handlar det om ungefär 8,7 miljoner observationer där varje rad innehåller en unik kombination av medlem och kampanjnummer. Vid varje kampanjutskick väljs en delmängd av medlemmar ut. Dessa delas sedan in i kampanj- och kontrollgrupp. Detta innebär att vid olika kampanjer har data samlats in om olika stora mängder kunder. Enligt Karlsson et al. (2013) i tidigare studier har det framkommit att användandet av flera olika kampanjer kan resultera i missvisande resultat. I detta datamaterial återfinns samma kund flera gånger vilket skulle orsaka korrelerade feltermer om samtliga kampanjer analyserades. Därför kommer endast en kampanj att väljas ut. Nedan visas data över de olika kampanjerna.

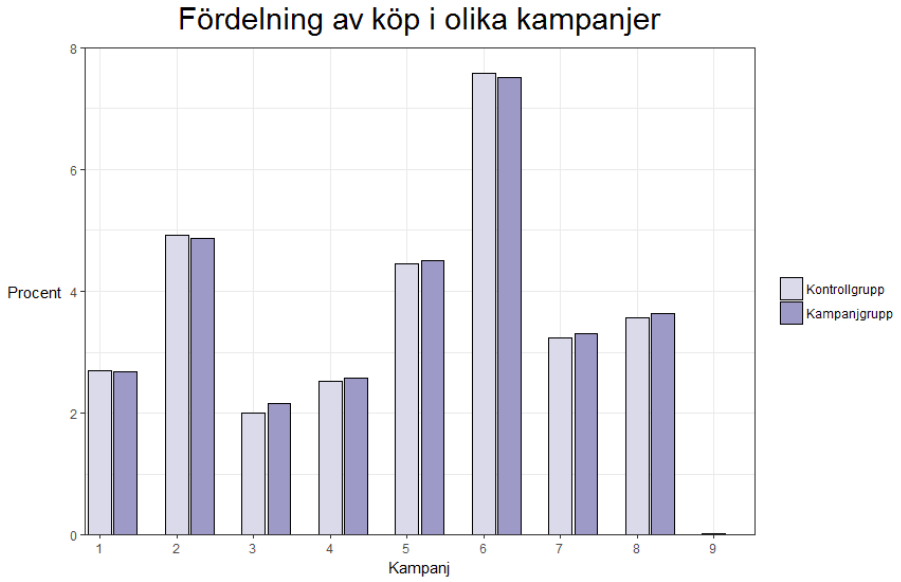


Figur 2.1: Fördelning av köp i olika kampanjer i Sverige

Tabell 2.1: Kampanjer i Sverige

	Kontrollgrupp			Kampanjgrupp			
Kampanj	Antal	Köp (%)	Icke-köp(%)	Antal	Köp(%)	Icke-Köp(%)	Uplift
1	47109	2,68	97,32	414796	2,73	97,27	0,05
2	16	31,25	68,75	84	10,71	89,29	-20,54
3	56406	1,88	98,12	496956	2,11	97,89	0,23
4	47761	2,4	97,6	416750	2,42	97,58	0,02
5	54705	4,23	95,77	480586	4,32	95,68	0,09
6	7277	7,54	92,46	64173	7,49	92,51	-0,05
7	48528	3,32	96,68	428157	3,32	96,68	0
8	53286	3,65	96,35	465798	3,72	96,28	0,07
9	6847	0	100	60604	0	100	0

I tabell 2.1 går det att se att det råder stora skillnader mellan antalet kunder som finns med i datamaterialet för de olika kampanjerna. Kampanj 3 har flest observationer i datamaterialet, där lite över 550 000 medlemmar tillhör kontroll- och kampanjgrupp. Enligt figur 2.1 finns det en liten skillnad i procentuella köp mellan kampanj- och kontrollgrupp för kampanj 3.



Figur 2.2: Fördelning av köp i olika kampanjer i Skandinavien

Tabell 2.2: Kampanjer i Skandinavien

	Kontrollgrupp			Kampanjgrupp			
Kampanj	Antal	Köp(%)	Icke-köp(%)	Antal	Köp(%)	Icke-Köp(%)	Uplift
1	109677	2,7	97,3	968349	2,68	97,32	-0,02
2	76415	4,92	95,08	676471	4,87	95,13	-0,05
3	133335	2	98	1177781	2,16	97,84	0,16
4	110919	2,52	97,48	974206	2,58	97,42	0,06
5	12925	4,45	95,55	1138490	4,51	94,49	0,06
6	18153	7,57	92,43	160402	7,5	92,5	-0,07
7	111776	3,23	96,77	987662	3,3	96,7	0,07
8	127040	3,57	96,43	1120727	3,63	96,37	0,06
9	17332	0,02	99,98	153945	0	100	-0,02

I tabell 2.2 går det att se att det råder stora skillnader mellan antalet kunder som finns med i datamaterialet för de olika kampanjerna. Kampanj 3 har flest observationer i datamaterialet, där lite över 1,3 miljoner medlemmar tillhör kontroll- och kampanjgrupp. Enligt figur 2.2 råder det en liten procentuell skillnad mellan köp i kampanj- och kontrollgrupp för kampanj 3.

Med tanke på att kampanj 3 har flest observationer i datamaterialet för både Skandinavien och Sverige används denna i uppsatsen för att skapa en modell där det inte finns problem med korrelerade feltermer.

2.1 Bearbetning

Genom undersökning av datamaterialet har flera saknade värden och tveksam information uppkommit. För att kunna göra en så bra modellering av datamaterialet som möjligt har data bearbetats på ett antal sätt. Detta skedde innan kampanj 3 valdes ut för uppsatsen.

- Datamaterialet innehöll observationer från flera olika länder utanför Skandinavien. Med tanke på att uppsatsen ska behandla erbjudanden som skickas inom Skandinavien har dessa observationer eliminerats.
- Variabeln som anger första postnummersiffran innefattade ett antal observationer som hade bokstäver istället för siffror. I Skandinavien är postnummer enbart nummerbaserade, vilket resulterat i en bearbetning på så sätt att endast siffror 0-9 sparades i datamaterialet.
- De kontinuerliga variablerna som finns i datamaterialet, men som även har en tillhörande grupperingsvariabel har eliminerats till förmån för gruppindelningarna. Där exempelvis åldersgrupp valts att användas istället för ålder.
- I datamaterialet finns det en indikatorvariabel som säger om medlemmen tillhör kontrollgrupp eller inte. Med tanke på att uppsatsens mål är att lokalisera de kunder som tenderar att köpa på kampanj har en ny variabel kampanj skapats, som även den är binär som säger om medlemmen istället tillhör kampanjgrupp eller inte.
- Om en medlem inte angivit födelseår vid skapandet av medlemskapet sätts förhandsvärdet 1900-01-01 in. Detta tenderar till att vissa medlemmar har en ålder på över 100 år. För att inte riskera att den äldre åldersgruppen ger felaktiga skattningar, har samtliga observationer med en ålder på över 90 år tagits bort ur datamaterialet.
- Mosaikdata finns enbart för Sverige och Norge, vilket medför att den inte kan användas för hela Skandinavien. Därför tas dessa förklaringsvariabler bort från data för Skandinavien. För att fortfarande kunna ta del av information från mosaikdata görs en modellering enbart över Sverige. Mer om hur mosaikdata är insamlat och uppbyggt återfinns i bilaga C.

3. Metod

I detta kapitel presenteras den inkrementella responsmodellen samt metoder för att hantera variabelselektion för inkrementella responsmodeller.

3.1 Beräkna uplift

Responsmodellen (response model) är den traditionella modellen som används för att beräkna *sannolikheten* att en kund genomför ett köp om kunden får ett utskick (Surry et al., 2011).

$$\max \sum_{S \in W} E(y_i \mid X_i; T_i = 1) \tag{3.1}$$

där S är mängden av kunder som har den högsta predikterade y_i givet kampanjgrupp. För att maximera sannolikheten för köp väljs mängden S från mängden W som är mängden av samtliga kunder.

Lo (2002) beräknar den *ökningen* i sannolikheten för att genomföra köp om kunden får utskick jämfört med om kunden inte får utskick. Detta görs genom att maximera skillnaden i respons mellan kampanjgrupp och kontrollgrupp enligt följande formel:

$$\max \sum_{S \in W} (E(y_i|X_i; T_i = 1) - E(y_i|X_i; T_i = 0)) \tag{3.2}$$

Skillnaden i väntevärdena maximeras genom att välja ut den subgrupp, S av kunder ur W som har störst skillnad i sannolikhet för köp mellan kontroll och kampanj.

När $E(y_i \mid X_i; T_i = 0)$ är nära 0 reduceras den inkrementella responsmodellen 3.2 till responsmodellen 3.1.

3.2 Separata och gemensamma modeller

Enligt Larsen (2010) är *difference score model* modeller som mäter differensen mellan de predikterade värdena i kampanjgruppen och kontrollgruppen, en vanlig metod inom inkrementell responsmodellering. Värdena kallas *difference scores* och rankas i fallande ordning för att de medlemmar som genererar högst uplift ska väljas ut. Modellen kan byggas på två olika sätt. Dels kan två, separata modeller från kampanj- och kontrollgruppen användas eller så används en enskild, kombinerad modell. En separat modell anpassar två modeller, en med kampanjgrupp och en med kontrollgrupp. Därefter beräknas *difference score* som differensen mellan de predik-

terade värdena från de två modellerna. Då detta tillvägagångssätt inte enligt Larsen (2010) i tidigare studier är att föredra, så kommer uppsatsen att tillämpa den gemensamma modellen.

En gemensam modell beskrivs i Lo (2002). Lo använder sig av en indikatorvariabel, T_i där $T_i = 1$ om person i är med i experimentgruppen och 0 om person i är med i kontrollgruppen.

$$y_i = \beta_0 + \mathbf{x_i'}\beta_1 + T_i\beta_2 + (\mathbf{x_i'}T_i)\beta_3 + \epsilon_i$$

där y är responsvektorn för varje individ i datamaterialet, β_0 är interceptet, x är designmatrisen med mätvärdena för varje individ, β_1 är parameterskattningar för variablerna, T är indikatorvariabeln för experimentgrupp och kontrollgrupp, β_2 är parameterskattningar för de huvudsakliga effekterna i kampanjgruppen och β_3 mäter interaktionseffekter för de förklarande variablerna hos medlemmar som har fått kampanjen.

Modellen som rör experimentgruppen innehåller samtliga parametrar. Interaktionstermerna kan då tolkas som skillnaden i sannolikhet för köp i kontrollgrupp och kampanjgrupp, det vill säga uplift.

$$\hat{y}_{Ei} = \hat{\beta}_0 + \mathbf{x_i'}\hat{\beta}_1 + \hat{\beta}_2 + (\mathbf{x_i'}T_i)\hat{\beta}_3$$

Modellen för kontrollgruppen innehåller inga interaktionstermer eftersom denna är 0 för kontrollgruppen.

$$\hat{y}_{Ki} = \hat{\beta}_0 + \mathbf{x_i'}\hat{\beta}_1$$

Då erhålls *difference scores* med ekvationen:

$$\hat{DS}_i = \hat{y}_{Ei} - \hat{y}_{Ki} = \hat{\beta}_2 + \mathbf{x_i}\hat{\beta}_3$$

där i=1,2,...,n

Detta kan också skrivs som:

$$\hat{DS}_i = \hat{y}_{Ei} - \hat{y}_{Ki} = E(y_i|x_i, T_i = 1) - E(y_i|x_i, T_i = 0)$$

där i=1,2,...,n

3.3 Tränings-, validerings- och testmängd

Innan analyserna påbörjas sparas 10 procent av datamaterialet i en testmängd. Den resterande mängden delas in i 70 procent träningsmängd och 30 procent valideringsmängd. Indelningen är randomiserad. Träningsmängden används för att anpassa modellerna och ta fram parameterskattningar. Därefter används valideringsmängden för att studera fördelningen av sannolikheterna modellen skattar, och för att ta fram en förväxlingsmatris. I förväxlingsmatrisen undersöks modellens prediktionsförmåga när det kommer till att klassificerar köp.

Testmängden är till för att få en slutgiltig skattning av hur väl modellen klassificerar köp. Detta görs genom att skapa en ny förväxlingsmatris över testmängden.

3.4 Logistisk regression

Då responsvariabeln i data är binär, om kunden köpt på en kampanj eller inte köpt på en kampanj. Så är den logistiska regressionen en bra modell att använda, för att finna hur olika bakgrundvariabler förklarar responsvariabeln. För att modellera andelen genomförda köp genom logistisk regression används modellen Lo (2002):

$$p_i = E(\mathbf{Y}_i \mid \mathbf{X}_i) = \frac{e^{\beta_0+\beta_1'\mathbf{X}_i+\beta_2 T_i+\beta_3 \mathbf{X}_i T_i}}{1+e^{\beta_0+\beta_1\mathbf{X}_i+\beta_2 T_i+\beta_3 \mathbf{X}_i T_i}}$$

Där X_i och T_i är oberoende variabler, där $T_i = 1$ om person i är med i experimentgruppen och 0 om person i är med i kontrollgruppen. Och där β_0, β_1, β_2 och β_3 är de parametrar som uppsatsen är intresserade av att skatta. Där β_0 är interceptet för regressionen. β_1, β_2 och β_3 är huvudsakliga effekterna hos de oberoende variablerna. Detta beräknas för både experimentgrupp och kontrollgrupp för att sedan ta fram differensen mellan dessa, för att se om det finns någon uplift i att skicka ut kampanjen. Differensen beräknas enligt formeln nedan:

$$p_{i,T_i=1} = \frac{e^{\beta_0+\beta_1'\mathbf{X}_i+\beta_2 T_i+\beta_3 \mathbf{X}_i T_i}}{1+e^{\beta_0+\beta_1'\mathbf{X}_i+\beta_2 T_i+\beta_3 \mathbf{X}_i T_i}} - \frac{e^{\beta_0+\beta_1\mathbf{X}_i}}{1+e^{\beta_0+\beta_1\mathbf{X}_i}}$$

Om differensen blir större än 0 finns det en uplift i att skicka ut kampanjen. Genom att kolla på ekvationen visar den på att den uplift som råder ges utav $\beta_2 + \beta_3 X_i$. β_2 kan i detta fall tolkas som den ökning i uplift som råder för en slumpmässig kund och $\beta_2 + \beta_3 X_i$ kan ses som den ökning i uplift som råder givet faktorerna X för kund i.

3.5 Variabelselektion

Med hjälp av logistisk regression vill uppsatsen undersöka sambandet mellan responsvariabeln och de förklarande variablerna. För att kunna avgöra vilka variabler som ska bara med för att förklara responsvariabeln i modellen används variabelselektion. Det finns flera olika variabelselektioner som utgår ifrån p antal förklarade variabler. De som kommer användas med logistisk regression är lassoregression och Penalized Net Information Value. Dessa variabelselektioner har valts då tidigare undersökningar från Larsen (2010) och Karlsson et al. (2013) visar på goda resultat i inkrementella responsmodeller med dessa.

3.5.1 Net Weight of Evidence och Net Information Value

Larsen (2010) skriver om måtten *Weight of evidence* (WOE) och *Information value* (IV) och modifierar sedan dessa till *Net weight of evidence* ($NWOE$) och *Net Weight of evidence* (NIV) för att dessa ska kunna användas som variabelselektionsmetod vid anpassning av inkrementella responsmodeller. Vid användning av både tränings- och valideringsmängd kan *Penalized Net weight of evidence* ($PNIV$) användas.

WOE kan användas för att se var en prediktiv variabel har sin styrka. Först grupperas den prediktiva variabeln i B antal lika stora boxar (Bins). Sedan räknas WOE ut med följande formel:

$$WOE_b = \frac{Pr(X = x_i|Y = 1)}{Pr(X = x_i|Y = 0)}$$

där b=1,2 ... B

Då kan man analysera WOE i de ordnade boxarna för att se var styrkan i den predikterande variabeln är.

IV räknas ut på följande sätt:

$$IV = \sum (Pr(X = x_i|Y = 1) - Pr(X = x_i|Y = 0)) \cdot WOE_i$$

IV används för att mäta styrkan i korrelation mellan den förklarande variabeln och responsvariabeln.

Dessa koncept kan användas i inkrementella responsmodeller med en modifikation efter experiment- och kontrollgrupp.

$NWOE$ beräknas med:

$$NOWE = \log \left[\frac{\left(\frac{Pr_E(X = x_i|Y = 1)}{Pr_E(X = x_i|Y = 0)} \right)}{\left(\frac{Pr_K(X = x_i|Y = 1)}{Pr_K(X = x_i|Y = 0)} \right)} \right]$$

NWOE är log-odds kvoten som jämför oddsen för genomfört köp i kontrollgruppen och kampanjgruppen.

NIV beräknas med:

$$NIV = \sum (\gamma - \theta) \cdot NOWE_i$$

där:

$$\gamma = (Pr_E(X = x_i|Y = 1) Pr_K(X = x_i|Y = 0))$$
$$\theta = (Pr_E(X = x_i|Y = 0) Pr_K(X = x_i|Y = 1))$$

NIV får ett högt värde om andelen genomförda köp skiljer sig mycket mellan kampanjgrupp och kontrollgrupp. Med NIV-värdet rankas variablerna och de som genererar uplift identifieras. För att kunna utföra rankningen behöver kontinuerliga variabler transformeras till diskreta. Detta kan göras genom att de delas in i exempelvis 20 lika stora grupper.

När en valideringsmängd används kan NWOE variera mycket mellan träning- och valideringsmängden. För att undvika att detta påverkar modellen negativt kan NIV justeras med en *Penalty term*(ω) som beskriver skillnaden i *NWOE* i tränings- och valideringsmängden. Då bildas *Penelized Net Information Value*(PNIV).

$$Penalty = \sum [Pr_E(X = x_i|Y = 1) - Pr_K(X = x_i|Y = 0) - Pr_E(X = x_i|Y = 0) - Pr_K(X = x_i|Y = 1)] \cdot \omega$$

där

$$\omega = \mid NWOE_{träning} - NWOE_{validering} \mid$$

Då räknas PNIV ut som

$$PNIV = NIV - Penalty$$

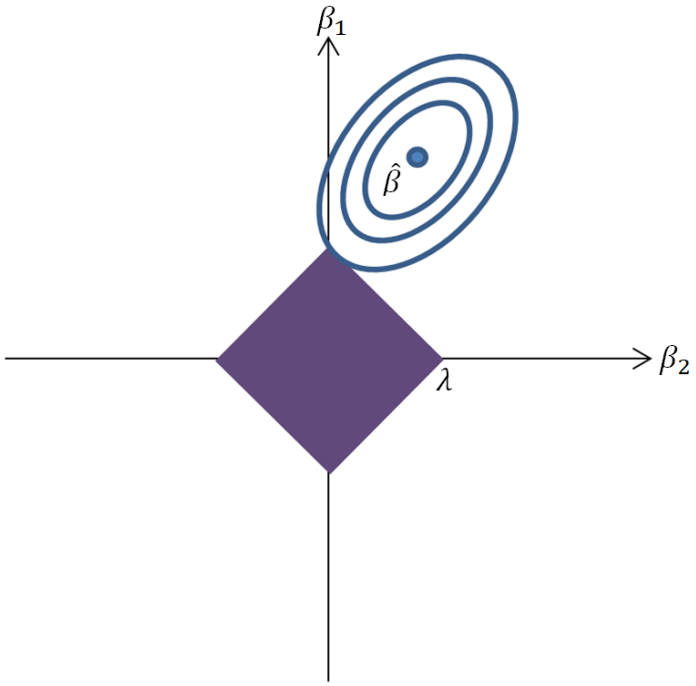
3.5.2 Lassoregression

Lassoregression (Least Absolute Shrinkage and Selection Operator) föreslås av Tibshirani (1996) som ett alternativ för att förbättra skattningarna med *Ordinary least squares*.

Många variabelselektioner behåller eller kastar koefficienterna, vilket är ett binärt tillvägagångssätt. Ett annat alternativ är Ridgeregression som förminskar koefficienterna, men sätter inga koefficienter till noll. Alla koefficienter behålls i modellen vilket gör modellen väldigt svårtolkad. LASSO är ett mellanting mellan dessa två tillvägagångssätt. LASSO förminskar vissa koefficienter och sätter andra till 0. Detta gör att både fördelarna från variabelselektion och Ridgeregression tas med i modellen.

En lassoregression anpassas med följande formel:

$$\min \hat{\beta} = \sum_{i=1}^{N} (y_i - \sum_j \beta_j x_{ij}) + \lambda \sum_j (|\beta_j|)$$



Figur 3.1: Parametervektor med två förklarande variabler anpassas med Lassoregression

Figur 3.1 illustrerar hur parametervektorn $\hat{\beta} = (\beta_1 \quad \beta_2)$ skattas med Lasso. De blå ringarna visar hur parameterskattningarna krymps med minsta kvadratmetoden. Diamanten visar begränsningsregionen för lassoregression ($\lambda \sum_j (\mid \beta_j \mid)$ som alltså är summan av absolutvärdena av parameterskattningarna multiplicerat med lassostraffet (λ). Parametervektorn kan inte skattas till mindre än detta område. När skattningen slår i begränsningsregionen har det minsta möjligt värdet uppnåtts och modellanpassningen med Lasso slutförs. Eftersom begränsningsregionen bildar en diamant med tydliga hörn skattas ofta lassoparametrarna vid dessa vilket leder till att vissa parametrar skattas till 0. I figur 3.1 visas ett exempel med två förklarande variabler där $\hat{\beta}$ har krympts så att β_2 skattats till 0. På detta sätt fungerar lassoregression som en variabelselektion då den har valt ut β_1, men inte β_2, det vill säga β_1 är mer relevant att ha med i modellen i detta fall.

Wu et al. (2009) använder lasso generellt vid logistisk regression. Sannolikheten för köp, givet värden på x_i skrivs som:

$$p_i = Pr(y_i = 1) = \frac{\exp^{\beta_0 + x_i^t \beta}}{1 + \exp^{\beta_0 + x_i^t \beta}}$$

Parametervektorn β skattas vanligtvis genom att maximera loglikelihoodfunktionen:

$$L(\beta) = \sum_{i=1}^{n} [y_i \log p_i + (1 - y_i) \log(1 - p_i)]$$

Denna skattning adderas sedan med *lassostraffet* $\lambda \sum_j (\mid \beta_j \mid)$:

$$LASSO(\beta) = \sum_{i=1}^{n}[y_i \log p_i + (1 - y_i) \log(1 - p_i)] + \lambda \sum_j (\mid \beta_j \mid)$$

För att tillämpa LASSO i en inkrementell responsmodell används interaktionstermer enligt (Lo, 2002).

$$p_i = \frac{e^{\beta_0 + \beta_1 \mathbf{X}_i + \beta_2 T_i + \beta_3 \mathbf{X}_i T_i}}{1 + e^{\beta_0 + \beta_! \mathbf{X}_i + \beta_2 T_i + \beta_3 \mathbf{X}_i T_i}}$$

Då används dessa sannolikheter för köp i maximeringen av loglikelihoodfunktionen adderat med lassostraffet.

$$LASSO(\beta) = \sum_{i=1}^{n}[y_i \log \frac{e^{\beta_0 + \beta_1 \mathbf{X}_i + \beta_2 T_i + \beta_3 \mathbf{X}_i T_i}}{1 + e^{\beta_0 + \beta_! \mathbf{X}_i + \beta_2 T_i + \beta_3 \mathbf{X}_i T_i}} + (1 - y_i) \log(1 - \frac{e^{\beta_0 + \beta_1 \mathbf{X}_i + \beta_2 T_i + \beta_3 \mathbf{X}_i T_i}}{1 + e^{\beta_0 + \beta_! \mathbf{X}_i + \beta_2 T_i + \beta_3 \mathbf{X}_i T_i}})] + \lambda \sum_j (\mid \beta_j \mid)$$

Lassoregressionen anpassas genom att ett flertal modeller med olika värden på λ (lassostraffet) genereras, där den modell med lägst Average Square Error (ASE) i valideringsmängden väljs.

$$ASE = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{N}$$

3.6 Modellutvärderingar

För att kunna svara på frågeställningen om vilken metod som är mest lämpad att prediktera köp i Scandinavian Airlines datamaterial utvärderas modellerna mot varandra med hjälp av mått för modellutvärdering. Anledningen till att modellerna jämförs med avseende på köp och inte uplift är för att det är svårt att finna något bra mått för att kunna jämföra uplift mellan olika modeller på ett objektivt sätt, enligt Karlsson et al. (2013). Därför analyseras istället modellernas förmåga att prediktera köp. Modellerna innehåller variabler som förklarar uplift (interaktionstermer) och om modellen i sin helhet är välanpassad är detta ett argument för att interaktionstermerna är relevanta och förklarar uplift. Eftersom uplift är inkluderad i modellerna tas det hänsyn till skillnaden mellan kampanj- och kontrollgrupp när resultatet tas fram.

3.6.1 Förväxlingsmatris

Ett vanligt sätt att utvärdera en modell är med hjälp av en förväxlingsmatris. Detta är en lämplig metod när responsvariabeln är binär, det vill säga att modellen klassificerar 1 eller 0.

Tabell 3.1: Förväxlingsmatris

		Predikterad klass	
		1	0
Faktisk klass	1	TP	FN
	0	FP	TN

Felkvoten anger hur stor andel av modellen som är fel. Detta beräknas med:

$$Felkvot = \frac{FN + FP}{TP + FN + FP + TN}$$

Precision och recall är två mått som berättar vilket typ av fel modellen gör. Precision visar hur stor andel av de som modellen väljer ut som positiva som faktiskt är positiva. Det vill säga om klassificeraren har identifierat 100 observationer som positiva, men bara 75 av dessa är positiva egentligen så är precisionen för klassificeraren 75 procent.

Tabell 3.2: Förväxlingsmatris som visar 75 procents precision

		Predikterad klass	
		1	0
Faktisk klass	1	75	0
	0	25	0

Precision räknas ut på följande sätt (de celler som är rosamarkerade i tabell 3.2 används i beräkningen):

$$Precision = \frac{TP}{TP + FP} = \frac{75}{75 + 25} = 75\%$$

Recall är ett mått som anger hur stor andel av de som faktiskt är positiva som modellen väljer ut.

Om det finns 100 positiva observationer och modellen bara identifierar 20 av dessa har modellen 20 procents recall.

Tabell 3.3: Förväxlingsmatris som visar 20 procents recall

		Predikterad klass	
		1	0
Faktisk klass	1	20	80
	0	0	0

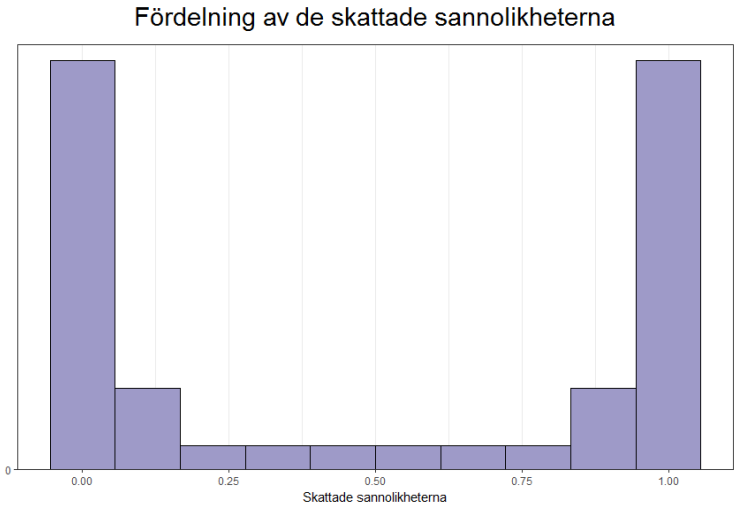
Recall beräknas på följande sätt (de celler som är rosamarkerade i tabell 3.3 används i beräkningen

$$Recall = \frac{TP}{TP + FN} = \frac{20}{20 + 80} = 20\%$$

Måtten Precision och Recall är sannolikheter med värden mellan 0 och 1, där så höga värden som möjligt eftersträvas.

Eftersom resultaten från logistisk regression består av skattade sannolikheter, $\hat{p}_i$, behöver en gräns bestämmas för när dessa ska klassas som 0 eller 1, det vill säga köp eller inte. Vid vanlig avrundning dras gränsen vid $\hat{p}_i \geq 0,5 = 1$ och $\hat{p}_i < 0,5 = 0$. Problem med den gränsen i detta fall är att väldigt få eller väldigt många sannolikheter klassas som köp. Ett alternativ är att dra gränsen vid den 1−(andel köp i datamaterialet) percentilen av ordnade $\hat{p}_i$. Detta leder till andelen köp som modellen klassificerar är lika stor som den faktiska andelen köp i datamaterialet. Anledningen till den framtagna gränsen är att uppsatsen vill att modellen skattar lika många köp som faktiskt är köp. Med detta för att undvika att modellen predikterar för många eller för få köp.

För att studera hur modellen har skattat kan man visualisera $\hat{p}_i$ i ett histogram. Då ser man om det finns en tydlig uppdelning mellan köp och icke-köp i modellens skattade sannolikheter.



Figur 3.2: Den ultimata fördelningen av $\hat{p}$

Figur 3.2 visar en modell som har många skattningar runt 0 och 1 och färre skattningar däremellan. Om modellens skattningar stämmer vore detta en bra fördelning av $\hat{p}$.

3.7 Programvaror

Logistisk regression för PNIV utförs med noden incremental response i SAS Enterprise Miner och LASSO utförst med noden LARs. Förväxlingsmatriser för de två modellerna kodas i R-studio. För kod till förväxlingsmatris se bilaga D.

4. Resultat och Analys

Analyserna har gjorts för kampanj 3. Resultaten redovisas metodvis för Sverige och Skandinavien. Detta för att kunna jämföra metoderna och länderna mot varandra.

Då de skattade parametrarna är många görs ett urval för att underlätta tolkningen. I den logistiska regressionen med PNIV visas endast de signifikanta variablerna på 10 procents signifikansnivå och i lassoregressionen visas de 20 största parametrarna. Dessa värden har valts för att kunna studera interaktionstermer, då lägre gränser inte resulterar i några sådana.

4.1 Logistisk regression med PNIV

Parametrarna skattas med träningsmängden och modellens styrka mäts med valideringsmängden.

4.1.1 Sverige

Tabell 4.1: Variabler valda av PNIV för Sverige

Variabel	PNIV	NIV	Vald
Flugit	103.2227	127.8048	Ja
År	10.8667	29.8021	Ja
Åldersgrupp	9.973	24.0354	Ja
Boendeform	9.5937	32.1353	Ja
Nordisk gruppkod	3.5713	25.1254	Ja
Postnummersiffra	3.2461	18.5672	Ja
Mobil	0.1349	0.1672	Ja
Medlemsnivå	0.0245	10.9333	Nej
Leveranskod	0.0228	0.0242	Nej
Mosaikgruppkod	-1.0592	16.5209	Nej
Mosaiktypkod	-2.3372	27.602	Nej
Kön	-7.715	7.0135	Nej
Nordisk typkod	-10.6748	17.1091	Nej
Poänggrupp	-40.7694	13.1479	Nej

I tabell 4.1 visas variablerna. Kolumnen Vald visar vilka variabler som valts ut av algoritmen med variabelselektionen PNIV. De variabler som är viktigast för att förklara uplift är om medlemmen flugit de senaste 6 månaderna, hur många år medlemmen varit med i Eurobonus, åldersgrupp,

om medlemmen bor i hus eller lägenhet, den nordiska gruppkoden, första siffran i postnumret och huruvida medlemmen angett mobilnummer till Scandinavian Airlines eller inte.

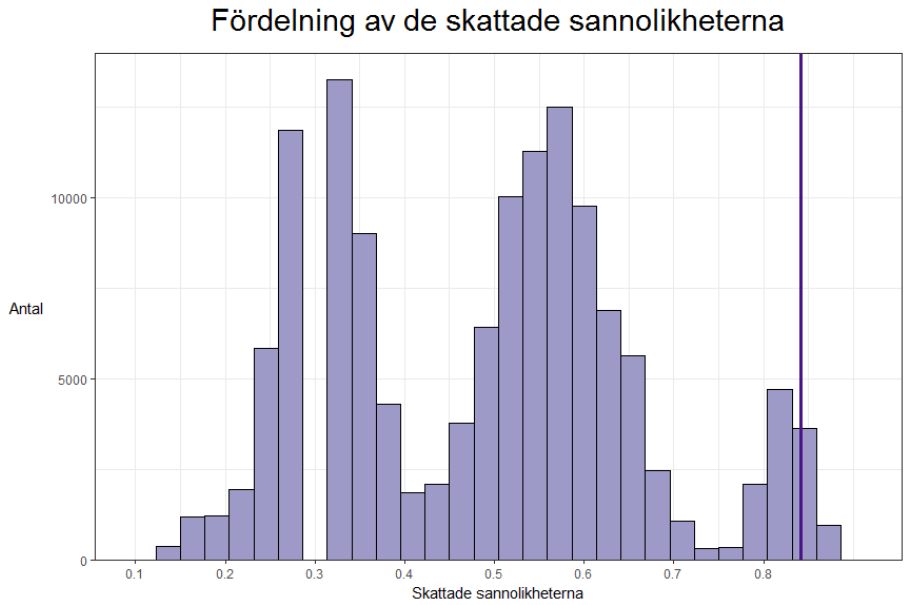
Tabell 4.2: Parameterskattningar på 10 procent signifikansnivå för Sverige (PNIV)

Variabel	Parameterskattning	P-värde
Flugit 0	-0.787	0
Åldersgrupp 18-29	1.1616	0
År	0.041	0
Nordisk gruppkod E	0.2688	0
Flugit 0 * Kampanj 0	-0.1102	2e-04
Nordisk gruppkod F	-0.3978	0.0016
Mobil 0	-0.3885	0.0034
Postnummersiffra 8 * Kampanj 0	-0.2524	0.0049
Postnummersiffra 5	-0.2942	0.0052
Åldergrupp 70-79	-0.3037	0.0092
Postnummersiffra 4	-0.1802	0.0107
Postnummersiffra 1	0.0988	0.0277
Nordisk typkod D	0.1224	0.029
Åldergrupp 40-49	-0.1918	0.0292
Postnummersiffra 5 * Kampanj 0	0.2292	0.0294
Nordisk gruppkod F * Kampanj 0	-0.2736	0.0302
Åldersgrupp 18-29 * Kampanj 0	0.1687	0.0526
Boendeform Lägenhet * Kampanj 0	0.0517	0.0596

I tabell 4.2 visas alla signifikanta parameterskattningar för PNIV. Det som är mest intressant är interaktionstermerna som förklarar skillnaden mellan kontrollgrupp och kampanjgrupp. Kampanjgrupp används som referens, vilket innebär att interaktionstermerna tolkas som skillnaden i sannolikhet för de som inte har fått kampanj 3. De som inte har flugit och fått kampanjen köper mer. Detta kan bero på att för medlemmar som reser ofta är det ett litet steg att boka en resa när en bra kampanj kommer. De som bor i postnummerområdet som börjar med postsiffra 8, vilket är Mittsverige, och får en kampanj köper mer, medan medlemmar som bor i postnummerområde som börjar med postnummersiffra 5, som är området runt Jönköping, köper mindre om de får en kampanj. Genom analys av datamaterialet framkommer det att i postnummerområde 5 finns det procentuellt sett färre medlemmar som har flugit de senaste 6 månaderna jämfört med de andra postnummerområdena. Detta kan ha påverkat resultatet.

Medlemmar som ingår i åldersgruppen 18-29 år köper mindre om de får en kampanj. Detta kan bero på att medlemmar i denna åldersgrupp har lägre inkomst och därför inte kan vara lika spontana och köpa när en kampanj kommer, utan istället planerar en resa när de har råd.

Medlemmar som bor i lägenhet tenderar också till att köpa mindre om de får en kampanj. Medlemmar som tillhör Nordisk gruppkod F köper mer om de får en kampanj.



Figur 4.1: Fördelning av $\hat{p}$ i valideringsmängden
Logistisk regression med PNIV för Sverige

I figur 4.1 visas det hur skattningarna av $\hat{p}$ fördelas i valideringsmängden. Den vertikala linjen visar gränsen (0,8421) för klassificeringen av sannolikheter för köp. Många observationer ligger runt 0,3-0,4 och många runt 0,55-0,65. Det är få skattningar runt 0,75 men därefter ökar det lite igen. Andelen som klassificeras som köp är ungefär lika stor som de faktiska köp som ingår i valideringsmängden, vilket är 2,1 procent.

Tabell 4.3: Förväxlingsmatris för PNIV (Sverige, valideringsmängd)

		Predikterad klass	
		1	0
Faktisk klass	1	254	2859
	0	2806	143488

Med förväxlingsmatrisen beräknas följande mått:

Tabell 4.4: Jämförande mått för PNIV (Sverige, valideringsmängd)

	PNIV
Felkvot	0,0379
Precision	0,0830
Recall	0,0859

Felkvoten är låg, där modellen klassificerar endast 3,79 procent fel. Dock är recall och precision väldigt låga. Modellen hittar ungefär 8,3 procent av de köp som finns i datamaterialet och av de observationer som modellen väljer ut som köp är det 8,6 procent som är faktiska köp.

4.1.2 Skandinavien

Tabell 4.5: Variabler valda av PNIV för Skandinavien

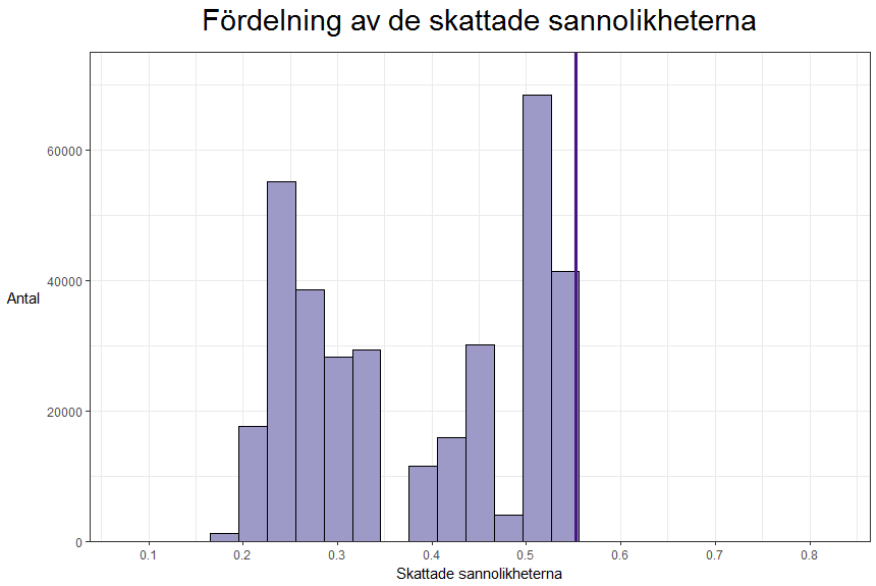
Variable	PNIV	NIV	Vald
Flugit	8.3793	14.0964	Ja
Mobil	0.0170	0.032	Ja
Kön	-0.1902	0.0597	Ja
Leveranskod	-0.2684	0.4314	Ja
Landskod	-0.2853	0.4992	Ja
Åldersgrupp	-0.518	4.1096	Nej
Postnummersiffra	-0.559	2.8105	Nej
Poänggrupp	-0.6204	3.146	Nej
År	-1.3 78	3.2943	Nej
Boendeform	-1.5538	0.5067	Nej

I tabell 4.5 visas de variabler som valts med variabelselektionen PNIV. De variabler som är viktigast när det kommer till att förklara upliften i data är om medlemmen flugit under de senaste sex månaderna eller inte. Om medlemmen har angett mobilnummer, vilket kön personen har, leveranskoden på kampanjen och Landskoden.

Tabell 4.6: Parameterskattningar på 10 procent signifikansnivå för Skandinavien (PNIV)

Namn	Parameterskattningar	P-värde
Flugit 0	-0.895	0
Kön kvinna	-0.1826	0
Mobil 0	-0.2169	0
Flugit 0 * Kampanj 0	-0.0356	0.0052
Leveranskod 13 *Kampanj 0	-0.7291	0.0746
Landskod NO *Kampanj 0	0.7232	0.077
Landskod NO	0.7094	0.0828

I tabell 4.6 visas de signifikanta parameterskattningarna på 10 procents signifikansnivå. De skattningar som är av störst intresse för uppsatsens syfte är interaktionstermerna, som antyder skillnaden mellan kontrollgrupp och kampanjgrupp. De medlemmar som inte har flugit och inte heller fått kampanjen tenderar till att köpa mindre. Även de medlemmar som har leveranskod 13 och inte fått kampanjen köper mindre. Medlemmar i Norge som inte får kampanjen tenderar dock istället till att köpa mer.



Figur 4.2: Fördelning av $\hat{p}$ i valideringsmängden
Logistisk regression med PNIV för Skandinavien

I figur 4.2 visas det hur skattningarna av $\hat{p}$ fördelas i valideringsmängden. Den vertikala linjen visar gränsen (0,5527) för klassificeringen av sannolikheter för köp. Många observationer har en skattad sannolikhet på mellan 0,2-0,3 eller runt 0,5. Andelen som klassificeras som köp är väldigt låg, på 1 procent, jämfört med de faktiska köpen som uppgår till 4,5 procent.

Tabell 4.7: Förväxlingsmatris för PNIV (Skandinavien, valideringsmängd)

		Predikterad klass	
		1	0
Faktisk klass	1	470	14943
	0	3237	323639

Gränsen enligt ordnad percentil predikterar få observationer till köp. Vid dragning av gränsen framkommer det att många observationer får samma skattningar, vilket kan bero på att det enbart är kategoriska variabler. Detta innebär att om gränsen dras aningen lägre predikterar modellen mer än dubbelt så många köp jämfört med hur många köp det finns i datamaterialet. Således görs en avvägning mellan för få eller för många skattade köp.

Med förväxlingsmatrisen beräknas följande mått:

Tabell 4.8: Jämförande mått för PNIV (Skandinavien, valideringsmängd)

	PNIV
Felkvot	0,053
Precision	0,03
Recall	0,127

Felkvoten är låg, där modellen klassificerar 5,3 procent fel. Recall och precision väldigt låga. Modellen hittar ungefär 3 procent av de köp som finns i datamaterialet och av de observationer som modellen väljer ut som köp är det 12,7 procent som är faktiska köp.

4.2 Logistisk regression med LASSO

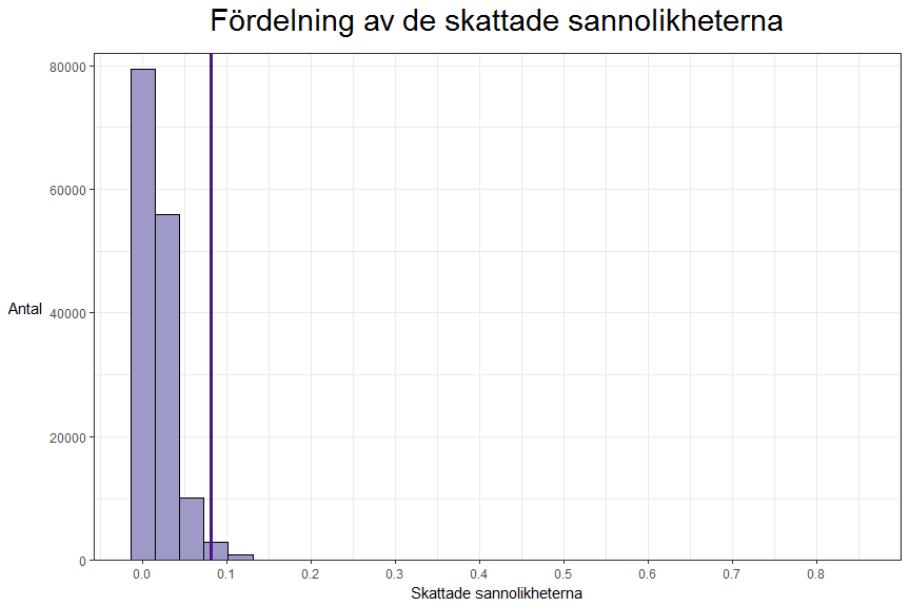
Parametrarna för lassoregressionen tas fram med träningsmängden. Därefter används valideringsmängden för att mäta modellens styrka.

4.2.1 Sverige

Tabell 4.9: De 20 största parameterskattningarna för Sverige (LASSO)

Variabel	Parameterskattning
Medlemsnivå 1 * Kampanj	0.1057
Leveranskod 6 * Kampanj	0.0955
Leveranskod 7 * Kampanj	0.038
Intercept	0.033
Medlemsnivå 1	0.0272
Medlemsnivå 2	0.026
Mosaiktypkod D16	0.0257
Åldersgrupp 18-29	0.0244
Medlemsnivå 3	0.0236
Flugit 0	-0.0145
Poänggrupp 0	-0.0142
Poänggrupp 1	-0.0133
Poänggrupp 8	-0.0121
Poänggrupp 7	-0.0107
Postnummersiffra 4	-0.0085
Postnummersiffra 2	-0.0073
Mosaiktypkod B08 * Kampanj	0.0071
Postnummersiffra 5	-0.0069
Poänggrupp 2	-0.0067
Postnummersiffra 6	-0.0065

I tabell 4.9 visas de största parameterskattningarna, det vill säga de variabler som har högst inverkan på skattningen av sannolikheten för köp. De interaktionstermer som kommer med är medlemsnivå 1, leveranskod 6 och 7 och mosaiktypkod B08. Parameterskattningarna är positiva vilket innebär att de som har fått kampanjen och har medlemsnivå 1, leveranskod 6 och 7 eller mosaiktypkod B08 tenderar till att köpa mer vid kampanj. Medlemsnivå 1 innebär den lägsta medlemsnivån (Medlem), samtliga medlemsnivåer återfinns i bilaga B. Mosaiktypkod B08 innebär kulturkapitalister som är namnet på en konsumentgrupp i mosaikdata.



Figur 4.3: Fördelning av $\hat{p}$ i valideringsmängden
Logistisk regression med LASSO för Sverige

I figur 4.3 visas fördelningen av de skattade sannolikheterna för köp i valideringsmängden. Modellen har predikterat låga sannolikheter för köp. Vid ordnad percentil används gränsen 0,0806 eftersom metoden beräknar väldigt låga sannolikheter, se figur 4.3. Andelen som klassas som köp är lika stor som de faktiska köp som ingår i valideringsmängden i datamaterialet, vilket är 2,1 procent.

Tabell 4.10: Förväxlingsmatris för LASSO (Sverige, valideringsmängd)

		Predikterad klass	
		1	0
Faktisk klass	1	363	2750
	0	2750	143544

Med förväxlingsmatrisen beräknas följande mått:

Tabell 4.11: Jämförande mått för LASSO (Sverige, valideringsmängd)

	LASSO
Felkvot	0,0368
Precision	0,1166
Recall	0,1166

Felkvoten är låg och säger att modellen klassar 3,68 procent fel. Recall och precision väldigt låga. Modellen hittar enbart 12 procent av köpen i datamaterialet och enbart 12 procent av de som modellen klassificerar som köp är faktiskt köp.

4.2.2 Skandinavien

Tabell 4.12: De 20 största parameterskattningarna för Skandinavien (LASSO)

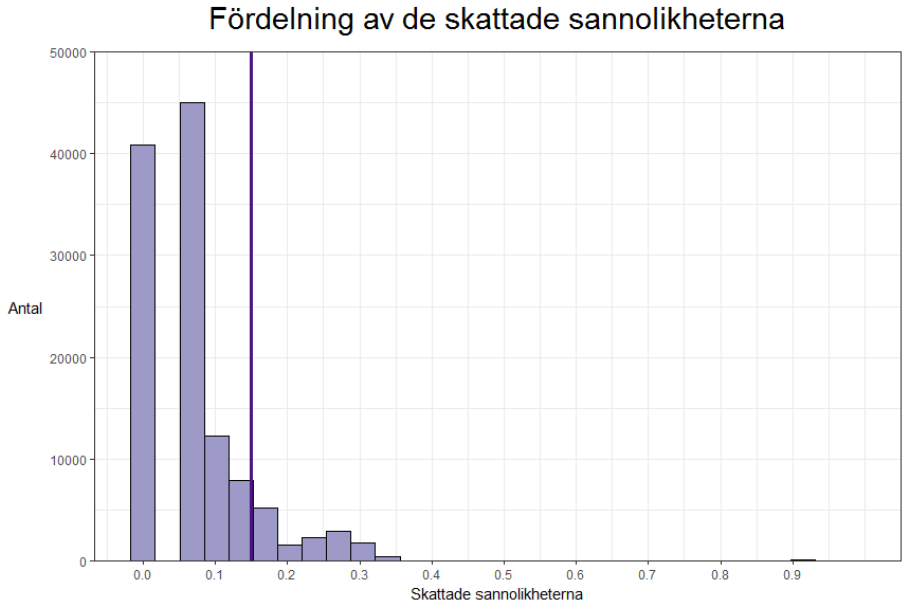
Variabel	Parameterskattning
Medlemsnivå 1	0.4282
Kampanj * Leveranskod 13	0.1781
Kampanj * Landskod NO	-0.1777
Landskod NO	0.1436
Kampanj * Medlemsnivå 1	-0.1411
Leveranskod 13	-0.1362
Kampanj * Leveranskod 12	0.1154
Kampanj * Landskod DK	-0.115
Medlemsnivå 2	0.0895
Leveranskod 12	-0.0688
Landskod DK	0.0605
Poänggrupp 0	-0.0566
Poänggrupp 1	-0.0563
Poänggrupp 2	-0.0526
Poänggrupp 3	-0.0516
Poänggrupp 4	-0.0456
Poänggrupp 5	-0.04
Poänggrupp 7	-0.038
Medlemsnivå 3	0.0375
Poänggrupp 8	-0.0352

I tabell 4.12 visas de största parameterskattningarna, det vill säga de variabler som har högst inverkan på skattningen av sannolikheten för köp. De interaktionstermer som kommer med är leveranskod 12 och 13, landskod NO och DK och medlemsnivå 1.

Parameterskattningarna för interaktionerna med leveranskod 12 och 13 är båda positiva, vilket indikerar på att de personer som fått kampanj och dessa leveranskoder tenderar till att köpa mer.

Norge och Danmark kommer med som variabler som förklarar en minskning av köp vid kampanj. Detta innebär att dessa länder köper mindre om de får en kampanj, jämfört med om de inte får en.

Medlemsnivå 1 innebär den lägsta medlemsnivån (Medlem). Dessa medlemmar köper mindre om de får en kampanj.



Figur 4.4: Fördelning av $\hat{p}$
Logistsik regression med LASSO för Skandinavien

I figur 4.4 visas fördelningen av de skattade sannolikheterna för köp. Modellen har predikterat låga sannolikheter för köp. Vid ordnad percentil används gränsen 0,149 för att prediktera sannolikheten för köp, detta eftersom metoden beräknar väldigt låga sannolikheter, se figur 4.4. Enligt den angivna gränsen klassificeras 4,5 procent till köp, vilket är lika stor andel som de faktiska köpen.

Tabell 4.13: Förväxlingsmatris för LASSO (Skandinavien, valideringsmängd)

		Predikterad klass	
		1	0
Faktisk klass	1	2651	12762
	0	12762	314114

Med förväxlingsmatrisen beräknas följande mått:

Tabell 4.14: Jämförande mått för LASSO (Skandinavien, valideringsmängd)

	LASSO
Felkvot	0,0746
Precision	0,172
Recall	0,172

Modellen har en låg felkvot som ligger på 7,5 procent. Precision och recall har låga värden som säger att modellen enbart hittar 17 procent utav köpen i datamaterialet och endast 17 procent av de som modellen klassificerar som köp är faktiskt köp.

4.3 Modellutvärdeing

För en slutgiltig analys av modellernas prediktiva förmåga används testmängden för att beräkna felkvot, precision och recall. Samma gräns används som när valideringsmängden klassificerades för att se hur gränsen fungerar på testmängden.

4.3.1 Sverige

Eftersom lassoregression är den metod som får bäst resultat för Sverige i valideringsmängden enligt måtten, presenteras här testmängden för lassoregressionen.

Tabell 4.15: Förväxlingsmatris för LASSO (Sverige, testmängd)

		Predikterad klass	
		1	0
Faktisk klass	1	133	1022
	0	1040	53143

Utvärderingsmåtten för testmängden blir då:

Tabell 4.16: Jämförande mått för LASSO (Sverige, testmängd)

	Lasso
Felkvot	0,0373
Precision	0,1133
Recall	0,1152

Felkvoten är låg och säger att modellen klassificerar fel i 3,73 procent av fallen. Precision säger att 11,3 procent av de observationer som modellen säger är positiva, faktiskt är positiva. Recall säger att modellen hittar endast 11,5 procent av de köp som finns i datamaterialet. Detta innebär att modellen klassificerar ungefär lika i testmängden som den gjorde i valideringsmängden.

4.3.2 Skandinavien

Även i valideringsmängden för Skandinavien är lassoregressionen den metod som enligt måtten får bäst resultat.

Tabell 4.17: Förväxlingsmatris för LASSO (Skandinavien, testmängd)

		Predikterad klass	
		1	0
Faktisk klass	1	948	4762
	0	4855	116212

Utvärderingsmåtten för testmängden blir då:

Tabell 4.18: Jämförande mått för LASSO (Skandinavien, testmängd)

	LASSO
Felkvot	0,0758
Precision	0,1634
Recall	0,166

Felkvoten för Skandinavien är låg och säger att modellen klassificerar fel i 7,58 procent av fallen. Precision säger att 16,34 procent av de observationer som modellen skattar som positiva, faktiskt är positiva. Recall säger att modellen hittar endast 16,6 procent av de köp som finns i datamaterialet. Modellen klassificerar ungefär lika i testmängden som i valideringsmängden.

5. Diskussion

Ett problem med inkrementella responsmodeller är att det är svårt att utvärdera hur väl modellen anpassar *ökningen* i sannolikhet för köp om kunder får kampanjen eller inte, det vill säga uplift. Detta eftersom det inte finns något test eller specifikt mått, i alla fall har det inte hittats något sådant i litteratursökningen. Detta leder till att testen som använts i uppsatsen främst har undersökt hur väl modellerna predikterar köp. Modellerna i sig är framtagna för att modellera uplift och ta hänsyn till en kampanj- och kontrollgrupp, men utvärderingen har alltså främst bestått av att undersöka hur väl modellen klassificerar köp. Eftersom modellerna innehåller interaktionstermer som förklarar uplift kan det antas att dessa är relevanta om modellen i sin helhet är välanpassad.

När lassoregressionen anpassas skapas det en rad olika modeller med olika värden på λ, där modellen med lägst ASE i valideringsmängden väljs ut. Detta görs i SAS Enterprise Miner och algoritmen skattar de olika modellerna i bakgrunden, vilket gör att vi inte vet vilket λ som har använts. Från start var tanken att kunna generera ett optimalt λ och skapa lassoregressionen i R. Detta för att få mer valfrihet och kontroll över det λ som valdes. På grund av problem med RAM-minnet och buggar i programmet kunde detta inte genomföras, varpå vi valde att göra det i SAS Enterprise Miner istället. Då Scandinavian Airlines själva använder SAS Enterprise Miner kan detta medföra att de i framtida undersökningar kan använda sig av resultatet. Detta gör det lämpligt att använda SAS Enterprise Miners metod LARs för lassoregression.

Ett annat problem som råder är att kontrollgruppen är väldigt liten vilket leder till att det är få som köper i denna grupp vilket gör att modellen har svårt att anpassa uplift. Förmodligen hade resultatet blivit mer utförligt om kontrollgruppen hade varit större och innehållit mer köp. Scandinavien Airlines hade förmodligen tjänat på att använda en större kontrollgrupp för att bättre kunna utvärdera kampanjernas lönsamhet.

När modellen anpassas med PNIV genereras interaktionstermer mellan responsvariabeln och de förklarande variablerna. Eftersom PNIV bygger på logistisk regression behöver responsvariabel använda $1 = köp$ och $0 = icke - köp$. När interaktionstermerna genereras sorteras variabeln fallande och interaktionstermerna skapas för kontrollgruppen med kampanjgruppen som referens. Detta innebär att interaktionstermerna behöver tolkas som hur kontrollgruppen agerar i förhållandet i kampanjgruppen. Positiva parameterskattningar indikerar då på att kontrollgruppen köper mer i förhållande till kampanjgruppen. Vi anser att det är tydligare att använda kontrollgruppen som referensens då då det är kampanjgruppen vi är intresserade av. Därför använder vi kampanjgruppen vid skapandet av interaktionstermer i Lasso. Positiva parameterskattningar innebär då här att kampanjgruppen köper mer än kontrollgruppen. Detta gör att parameterskattningarna är lättare att tolka i förhållande till våra frågeställningar som handlar om att identifiera ökningen i sannolikhet för köp i kampanjgruppen.

När förväxlingsmatrisen tas fram för logistisk regression med PNIV för Skandinavien skattas inte lika många köp som det finns köp i valideringsmängden. Detta beror på att många kunder har samma skattningar och många ligger precis där gränsen dras. Detta är ett resultat av att enbart kategoriska variabler har använts och att många kunder besitter samma egenskaper och

därför får samma skattning.

Kampanjerna som skickas ut finns även att tillgå för samtliga medlemmar på hemsidan. Detta innebär att även de som är med i kontrollgruppen kan ha sett kampanjerna. Detta gör att resultatet kan bli missvisande. Vi tror att det hade varit fördelaktigt för Scandinavian airlines att börja skicka mer riktade kampanjer till specifika kunder beroende på deras resehistorisk, och därför är vår rekommendation till Scandinavian Airlines att de samlar in mer resdata om kunderna, så att man skulle kunna anpassa mer skräddarsydda kampanjer.

Uppsatsens hade från början avsikten att testa ett flertal metoder, så som beslutsträd, neurala nätverk och randome forest. Tyvärr fanns det inte möjlighet att hinna med detta inom ramarna för denna kurs, men hade varit intressant att applicera datat för flera modeller, för att få en utförligare analys över lämpliga modellval för Skandinavian Airlines data.

I uppsatsen används variabelsekektionerna LASSO och PNIV, hade varit intressant att kolla på flera olika varableselektioner så som stegvis-, framåt- och bakåteliminering.

6. Slutsats

Nedan besvaras frågeställningarna till uppsatsen. Värt att inflika är att vid användande av detta resultat för kommande riktade kampanjer, bör kampanjen vara likvärdig kampanj 3. Detta då andra typer av kampanjer kan locka andra eurobonusmedlemmar.

Vilka variabler är bäst för att beräkna uplift?

I Sverige är det följande variabler som blir signifikanta i logistisk regression med PNIV: `Flugit`, `Postnummersiffra 8`, `Postnummersiffra 5`, `Nordisk gruppkod F`, `Åldersgrupp 18-29` och `Boendeform Lägenhet`. `Postnummersiffra 8` innebär Mittsverige och `Postnummersiffra 5` innebär Jönköping. Medlemmar som flugit de senaste 6 månaderna, bor i Mittsverige eller har nordisk gruppkod F köper mer om de får en kampanj. Medlemmar som är mellan 18-29 år eller bor i Jönköping köper mindre om de får en kampanj.

Detta innebär att för att få ut så hög avkastning på marknadsföringen som möjligt bör Scandinavian Airlines fokusera på att skicka kampanjer till de eurobonusmedlemmarna som bor i Mittsverige, har flugit de senaste sex månaderna eller faller under nordisk gruppkod F. De bör undvika att skicka kampanjer till unga medlemmar (18-29 år) eller till medlemmar som bor i Jönköping. Detta då dessa medlemmar tenderar till att köpa mindre om de får en kampanj.

Av Lasso för Sverige väljs `Medlemsnivå 1`, `Leveranskod 6` och `Leveranskod 7` samt `Mosaiktypkod B08` ut som variabler som kan förklara uplift. `Mosaiktypkod B08` motsvaras av Kulturkapitalister. Medlemmar som faller i dessa kategorier tenderar att köpa mer om de får en kampanj jämfört med om de inte får en kampanj.

Resultatet från Lassoregressionen säger att Scandinavian Airlines bör skicka kampanjer till kunder som har den lägsta medlemsnivån, leveranskod B08 eller är kulturkapitaliser för att få ut så mycket av marknadsföringen som möjligt.

De variabler som blir signifikanta av PNIV för Skandinavien är `Flugit`, `Leveranskod 13` och `Landskod NO`. För Skandinavien i Lasso väljs variablerna `Leveranskod 13`, `Landskod NO`, `Medlemsnivå 1`, `Leveranskod 12` och `Landskod DK` ut för att förklara uplift. Där de medlemmar som Flugit och fått leveranskod 12 och 13 tenderar till att köpa mer när de får en kampanj, och de medlemmar som kommer från Danmark och Norge och de som har medlemsnivå 1 tenderar till att köpa mindre om de får en kampanj.

Detta innebär, för att Scandinavian Airlines ska få ut så hög avkastning som möjligt av marknadsföringen bör kampanjer skickas till kunder som flugit de senaste sex månaderna eller fått kampanjer med leveranskod 12 och 13. Vidare bör de lägga mindre fokus på att skicka kampanjer till kunder som har lägsta medlemsnivån eller som bor i Norge och Danmark.

Den variabel som får markant högst NIV och PNIV för Sverige och Skandinavien är `Flugit`. Detta innebär att personen som flugit med Scandinavian Airlines de senaste sex månaderna har en tydlig skillnad mellan kampanj- och kontrollgrupp.

Vilken metod är bäst för att klassificera köp av medlemmar?

För att kunna jämföra de olika metoderna och deras förmåga att prediktera köp bland medlemmarna i Sverige presenteras här de två förväxlingsmatriserna bredvid varandra.

Tabell 6.1: Jämförelse av förväxlingsmatriserna för Sverige

		Predikterad klass	
		1	0
Faktisk klass	1	363	2750
	0	2750	143544

(a) Lasso

		Predikterad klass	
		1	0
Faktisk klass	1	254	2859
	0	2806	143488

(b) PNIV

Tabell 6.1 visar att Lasso har predikterad fler korrekta köp. Falsk positiv och falsk negativ är lägre i Lasso och sann negativ är högre. Lasso verkar vara en bättre metod för att prediktera köp av Scandinavian Airlines kunder i Sverige.

Tabell 6.2: Jämförelse av jämförande mått för Sverige

	PNIV	Lasso
Felkvot	0,0379	0,0368
Precision	0,0830	0,1166
Recall	0,0859	0,1166

I tabell 6.2 visar det att Lasso har en lägre felkvot och högre precsion och recall. Detta innebär att Lassoregression är en bättre metod för att prediktera köp bland Scandinavian Airlines kunder i Sverige.

För att kunna jämföra de olika metoderna och deras förmåga att prediktera köp bland medlemmarna i Skandinavien presenteras även här de två förväxlingsmatriserna bredvid varandra.

Tabell 6.3: Jämförelse av förväxlingsmatriserna för Skandinavien

		Predikterad klass	
		1	0
Faktisk klass	1	2651	12762
	0	12762	314114

(a) Lasso

		Predikterad klass	
		1	0
Faktisk klass	1	470	14943
	0	3237	323639

(b) PNIV

Tabell 6.3 visar att Lasso har predikterat fler korrekta till köp. Falsk positiv är lägre för PNIV och falsk negativ är lägre i Lasso och sann negativ är högre i PNIV. Lasso verkar vara en bättre metod för att prediktera köp av Scandinavian Airlines kunder i Skandinavien.

Tabell 6.4: Jämförelse av jämförande mått för Skandinavien

	PNIV	Lasso
Felkvot	0,053	0,0746
Precision	0,03	0,172
Recall	0,127	0,172

I tabell 6.4 visar det att PNIV har en lägre felkvot medans Lasso har högre precision och recall. Detta innebär att Lassoregression hittar fler faktiska köp i datamaterialet då den har en högre precision.

Generellt är varken Lassoregression eller PNIV speciellt bra på att prediktera köp i datamaterialet, dock är Lassoregression något bättre. Detta beror förmodligen på en låg andel köp bland Eurobonusmedlemmar i kampanj 3. Datamaterialet verkar innehålla mycket information som inte genererar några tydliga mönster. Förmodligen hade resultatet blivit mer intressant med andra förklarande variabler, såsom utförligare resdata, och ett mer strukturerat datamaterial.

Finns det likheter mellan Skandinavien och Sverige när det gäller variabler som förklarar uplift?

De variabler som förklarar uplift i både Sverige och Skandinavien är Flugit och Medlemsnivå 1.

Om medlemmen flugit de senaste sex månaderna är den variabel som väljs ut av PNIV för både Sverige och Skandinavien. De som inte har flugit de senast 6 månaderna och inte fått kampanjen köper mindre med 90 procents säkerhet.

Medlemsnivå 1 är en viktig interaktionsterm för både Sverige och Skandinavien i och med att den kommer med bland de största parameterskattningarna i Lassoregressionen. Dock är interaktionstermen positiv för Sverige och negativ för Skandinavien. Detta innebär att de som har medlemsnivå 1 i Sverige tenderar att köpa mer om de får ett utskick, medan medlemmar i hela Skandinavien tenderar till att köpa mindre om de får ett utskick om de tillhör medlemsnivå 1.

6.1 Framtida arbeten

Enligt modellen logistisk regression med PNIV är `Flugit` en variabel som kan förklara uplift. Denna variabel indikerar om medlemmen har flugit med Scandinavian Airlines de senaste sex månaderna. Det hade varit intressant att i fortsatta studier undersöka detta närmare genom att titta på mer data över hur medlemmar handlat tidigare. Exempelvis skulle det vara spännande att undersöka hur det ser ut för medlemmar som flugit med Scandinavian Airlines det senaste året eller längre tid tillbaka. Det skulle också gå att kika på vilken typ av resor medlemmen köpt tidigare och hur ofta medlemmen reser.

I dagsläget finns det inte något kvalitativt och objektivt sätt att utvärdera och jämföra modeller med avseende på uplift. Därför vore det intressant att i framtida studier utveckla ett sådant mått.

Litteraturförteckning

Codex. Forskarens etik, 2016. URL `http://www.codex.vr.se/forskarensetik.shtml`. Hämtad den 1 juni 2017.

Insightone, 2017. URL `http://insightone.se/`. Hämtad den 3 mars 2017.

Jonas Karlsson et al. Inkrementell responsanalys. 2013.

K Larsen. Net lift models: Optimizing the impact of your marketing efforts. 2010.

Victor S Y Lo. The True Lift Model - A Novel Data Mining Approach to Response Modeling in Database Marketing. *ACM SIGKDD Explorations Newsletter*, 4(2), 2002.

SAS, 2017. URL `http://www.sasgroup.net/en/category/about-sas/`. Hämtad den 3 feb 2017.

Patrick D Surry et al. Real-World Uplift Modelling with Significance-Based Uplift Trees. *Stochastic Solutions White Paper*, (section 6):1–33, 2011.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statististical Society. Series B (Methodological)*, 58(1), 1996.

Wu et al. Genome-wide association analysis by lasso penalized logistic regression. *International Society for computational biology*, 25(6), 2009.

A. Datamaterial

Tabell A.1: Variabler i datamaterialet

Variabler	SAS variabel	Variabelbeskrivning
Medlemsnummer	Mbr_seq_no	Eurobonusmedlem
Kontrollgrupp	In_control_grp	1 om personen tillhör kontrollgrupp, 0 annars
Leveranskod	Delivery_code_seq_no	Benämningen på mailet som skickats.
Kampanjnummer	Campaign_seq_no	Vilken kampanj mailet behandlar
Datum	ContactDate	Datum kampanjen skickas ut
Medlemsnivå	Inc_lvl_cd	Nivåer på medlemskapet
Poänggrupp	Points_to_use_group	Poänggrupp
Mobil	has_mobile	1 om angett mobilnummer annars 0
År	year_in_program	Antal år i Eurobonus
Flugit	Flewlast6months	1 om flugit under de senaste 6 månaderna 0 annars
Köpt	Bought2	1 om köpt på utskicket 0 annars
Nordisk Gruppkod	Nord_groupcode	Nordisk gruppering
Nordisk typkod	Nord_typecode	Nordisk gruppering område
Postsiffra	first_zip_nr	Första siffran i postnumret
Boendeform	Typeofliving	Boendetyp
Åldersgrupp	age_group	Åldersgrupp
Kön	gender	Kundens kön
Landkod	S_country_code	Vart medlemmen kommer ifrån
Mosaikgruppkod	local_mosg5_group_code	Mosaikdata kring var kunden bor i för område
Mosaiktypkod	local_mosg5_group_typecode	Mosaikdata kring var kunden bor i området
Kampanj	-	1 om personen tillhör kampanjgrupp, 0 annars

Seqn0	Varje kund har ett unikt medlemsnummer. För att uppsatsen inte ska kunna knyta samman data till en person har dess nummer alternerats till andra nummer, som inte har någon betydelse för uppsatsen, utan endast för företaget.
Delivery_code	Då utskicket i data skickats till tre olika länder benämner denna variabel utöver utskickets innehåll vilket språk den är skriven på. Det vill säga till vilket land utskicket gjorts.
Inc_lvl_crm	Säger vilken nivå av medlemskap i Eurobonus som personen har. De fyra nivåer som finns är medlem (M), silver (S), guld (G) och diamant(D) och återfås beroende på hur aktiv medlemmen är. Mosaikdata är data som företag köper in som kompletterar deras data. Detta för att få tillgång till fler variabler kring kunderna som är registrerade i Eurobonus
Total_points_for_use	Som eurobonusmedlem samlar kunderna poäng, som i slutändan kan generera allt från billigare flygresor till val av mindre saker. Poängen beror av hur aktivt kunden brukar sitt medlemskap, där beroende på vad kunden köper eller använder sitt medlemskap till så genererar de olika poängsummor. Tål att nämnas är även att poäng kan återfås på många fler sätt än att endast köpa flygbiljetter hos SAS. Denna variabel förklarar hur många poäng de olika medlemmarna har till sitt förfogande.
Flewlastsixmonths:	Variabeln indikerar om personen de senaste sex månaderna innan kampanjen kom ut flygit med Scandinavian airlines. Här kodas 1 till de personer som flugit under 6-månadersperioden.
Bought2	Har kunden köpt på kampanjen eller inte. 1 menas med att kunden gjort ett köp och 0 annars. Detta är den binära responsvariabeln som råder i data.
Nord_group	Nordisk gruppering finns endast över Sverige och Norge, vilket medför saknade värden för samtliga utskick till Danmark. Den nordiska gruppkoden vairerar även inom länderna, vilket gör att variabeln är med i datat över Sverige

B. Poänggrupper och medlemsnivåer

Tabell B.1: Poängintervallen för Poänggrupperna

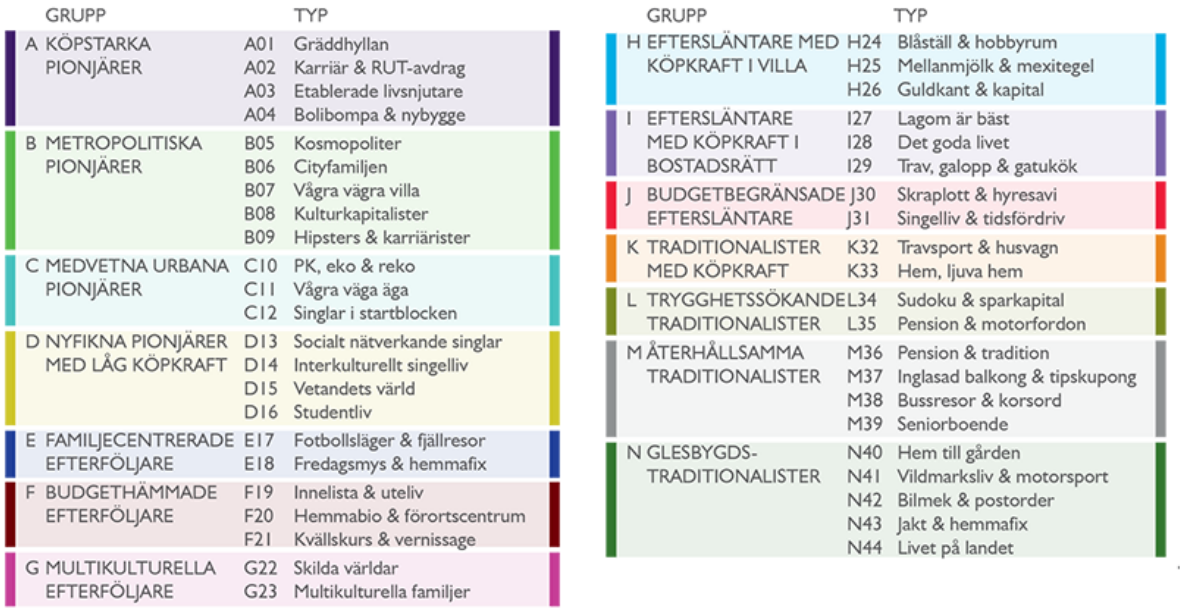
Poänggrupp	Poängintervall
0	0
1	1-1000
2	1-5 K
3	5-10 K
4	10-20 K
5	20-30 K
6	30-40 K
7	40-50 K
8	50-60 K
9	>60 K

Tabell B.2: Medlemsnivåer i Eurobonus

Medlemsnivå	Namn
1	Medlem
2	Silver
3	Guld
4	Diamant

C. Mosaikdata

Den mosaikdata som finns är framtagen av InsightOne. De klassificerar in grupper enligt figur C.1



GRUPP		TYP
A KÖPSTARKA PIONJÄRER	A01	Gräddhyllan
	A02	Karriär & RUT-avdrag
	A03	Etablerade livsnjutare
	A04	Bolibompa & nybygge
B METROPOLITISKA PIONJÄRER	B05	Kosmopoliter
	B06	Cityfamiljen
	B07	Vågra vägra villa
	B08	Kulturkapitalister
	B09	Hipsters & karriärister
C MEDVETNA URBANA PIONJÄRER	C10	PK, eko & reko
	C11	Vågra väga äga
	C12	Singlar i startblocken
D NYFIKNA PIONJÄRER MED LÅG KÖPKRAFT	D13	Socialt nätverkande singlar
	D14	Interkulturellt singelliv
	D15	Vetandets värld
	D16	Studentliv
E FAMILJECENTRERADE EFTERFÖLJARE	E17	Fotbollsläger & fjällresor
	E18	Fredagsmys & hemmafix
F BUDGETHÄMMADE EFTERFÖLJARE	F19	Innelista & uteliv
	F20	Hemmabio & förortscentrum
	F21	Kvällskurs & vernissage
G MULTIKULTURELLA EFTERFÖLJARE	G22	Skilda världar
	G23	Multikulturella familjer

GRUPP		TYP
H EFTERSLÄNTARE MED KÖPKRAFT I VILLA	H24	Blåställ & hobbyrum
	H25	Mellanmjölk & mexitegel
	H26	Guldkant & kapital
I EFTERSLÄNTARE MED KÖPKRAFT I BOSTADSRÄTT	I27	Lagom är bäst
	I28	Det goda livet
	I29	Trav, galopp & gatukök
J BUDGETBEGRÄNSADE EFTERSLÄNTARE	J30	Skraplott & hyresavi
	J31	Singelliv & tidsfördriv
K TRADITIONALISTER MED KÖPKRAFT	K32	Travsport & husvagn
	K33	Hem, ljuva hem
L TRYGGHETSSÖKANDE TRADITIONALISTER	L34	Sudoku & sparkapital
	L35	Pension & motorfordon
M ÅTERHÅLLSAMMA TRADITIONALISTER	M36	Pension & tradition
	M37	Inglasad balkong & tipskupong
	M38	Bussresor & korsord
	M39	Seniorboende
N GLESBYGDS-TRADITIONALISTER	N40	Hem till gården
	N41	Vildmarksliv & motorsport
	N42	Bilmek & postorder
	N43	Jakt & hemmafix
	N44	Livet på landet

Figur C.1: Mosaikdata

Mosaikdatat som finns till uppsatsens förfogande är producerat av InsightOne, och är ett datamaterial som kan hjälpa företag att få en ökad insikt och djupare förståelse kring deras konsumenters livsstil och beteende. InsightOne gör en klassificering Sveriges hushåll till 44 olika livsstilar och totalt 14 övergripande grupper. klassificeringer ger enligt dem en nyanserad och helhetstäckande bild över konsumenternas val, preferenser och vanor. Klassificeringen är på hushållsnivå och behandlar socialdemografi, livsstilar, köpbeteenden och värderingar. Mosaikdata är till för att förädla det datamaterial företagen själva innehar, och ska ge en helhetsbedömning av kunder (Insightone, 2017).

D. R-kod Förväxlingsmatris

Nedan visas R-koden för Skandinaviens förväxlingsmatris för testmängden. Samma procedur har använts för samtliga förväxlingsmatriser. Vid Lassoregression har $\hat{p}$ räknats ut med annan programvara och enbart klassificeringen $1 = köp$ och $0 = icke - köp$ har gjorts i R.

```
valT <- read.csv("Z:\\Kandidatuppsatsen\\data och beskrivande statistik\\
em_save_TEST.csv")

#Läser in tränings-/valideringsdata från SAS
#Gör om valideringsoutput till en dataframe
valT <- as.data.frame(valT)

#Läser in parameterskattningarna från INC-noden
skattINCT <- read.csv2("Z:\\Kandidatuppsatsen\\data och beskrivande statistik
\\nonePARAMETER.csv")



#Sparar namnen på parameterskattningarna (dessa namn innehåller både variabel-
namn
och grupp ex "age_group 18-29")
namnT <- as.character(skattINCT[,2])

#Delar upp Datamaterialet efter mellanslag så att variabelnamn och grupp hamnar
i olika kolumner

namn2T <- strsplit(namnT, " ")


#Gör en ny dataframe som innhåller parameterskattningar och namn på variabel
och grupp
inputT <- data.frame(nrow=20, var=3, group=2, int=2, var2=5, group2=6)

for (i in 1:length(namn2T)){

  inputT[i,] <- namn2T[[i]]

}

#Lägger till parameterskattningarna i den nya dataframen
inputT[,7] <- skattINCT[,3]
```

```
#Delar in i två dataframes med "vanliga" variabler och interaktionstermer
input_varT <- data.frame(var=3, group=2, skatt=3)
input_intT <- data.frame(nrow=20, var=3, group=2, int=2, var2=5)


inputT <- na.omit(inputT)
input_varT<-input[c(1,2,3,7,8,9,10,14),c(2,3,7)]
input_intT<-input[c(4,5,6,11,12,13),c(2,3,5,6,7)]


input_varT[,3] <- as.numeric(levels(input_varT[,3]))[input_varT[,3]]
input_intT[,5] <- as.numeric(levels(input_intT[,5]))[input_intT[,5]]

input_varT[7,1] <- "kampanj"

val_varT <- valT




#Sparar bara de variabler som finns i parameterskattningarna (dvs de variabler
som inte valts av NIV tas bort)
val_varT <- val_varT[,names(val_varT) %in% input_varT[,1]]

head(val_varT)

#Skapar tom dataframe att fylla resultat med

res2T <- data.frame(matrix(ncol = 20, nrow = 0))

val_var2T <- val_varT

#Skapar en loop som tar ut resultatet

for (j in  1:length(val_var2T[,1])){ #För varje kund j i valideringsmängden
  n <- 1                           #n skrivs till 1, n används för att  resultatet
  ska sparas på rätt ställe i resultatdataframen
  for (i in 1:length(val_var2T)){ #För varje variabel i valideringsmängden
    for (k in 1:length(input_varT[,1])){ #För varje variabel i parameterskatt-
    ningarna
      if (names(val_var2T)[i] == input_varT[k,1]){ #Om variabeln i input
      stämmer med
      variabeln i parameterskattningarna
        if (val_var2T[j,i]==input_varT[k,2]){ #Om gruppen i valideringsmängden
        stämmer med gruppen i parameterskattningarna
          res2T[j,n:(n+3)] <- input_varT[k,] #Då stoppas variabel, gruppen och
          parameterskattningen i resultaten, varje kund har en egen rad där dessa
          parameterskattningar skapas
          n <- n+3 #n ökar med 3 för varje steg
```

```
        }
      }
    }
  }
}
#Samma sak för interaktionstermer
val_intT <- valT
#Spara de som tillhör kampaj
kampanjT <- val_intT$kampanj
#spara de som köpt för att kunna jämföra senare
kopT <- val_intT$Bought2
#Tar bara med de parameterar som ingår i parameterskattningarna, alltså
de som valts av NIV
val_intT <- val_intT[,names(val_intT) %in% input_intT[,1]]
#Lägger in kampanj som en variabel
val_intT[length(val_intT)+1] <- kampanjT

#Skapar tom dataframe att fylla resultat med

res3T <- data.frame(matrix(ncol = 20, nrow = 0))

val_int2T <- val_intT


#Skapar en loop som tar ut res, denna fungerar på samma sätt som loopen ovan
for (j in  1:length(val_int2T[,1])){
  n <- 1
  for (i in 1:length(val_int2T)){
    for (k in 1:length(input_intT[,1])){
      if (names(val_int2T)[i] == input_intT[k,1]){
        if (val_int2T[j,i]==input_intT[k,2]){
          if (val_int2T[j,5]==1){
            res3T[j,n:(n+3)] <- input_intT[k,c(1,2,5)]
            n <- n+3


          }else {res3T[j,n:(n+3)]<-0}
        }
      }
    }
  }
}




#Summerar radvis de celler som innehåller parameterskattningar (dvs. var tredje)
Summerar också vanliga variabler och interaktionstermer
phatT <-rowSums(res3T[,c(3,6,9,12)],na.rm=TRUE)+rowSums(res2T[,c(3,6,9,12,15)],
na.rm=TRUE)

#Skattar p
```

```
phatT <- exp(phatT)/(1+exp(phatT))

kop2T <- kopT

#De slås samman för att kunna jämföra
resultatT <- cbind(phatT,kop2T)

resultatT <- as.data.frame(resultatT)
resultatT[,3] <- phatT

#Beräkning av egen gräns
table(kop2T)
5710/length(kopT) #Där 5710 är antal faktiska köp i datamaterialet


quantile(resultatT[,1],1-(5710/length(kop2T)))

#Loop över skattade phat, som klassas (1/0) efter den satta gränsen
for (i in 1:length(phatT)){
  if (resultatT[i,1] < 0.5527389 ){
    resultatT[i,3] <- 0
  } else{
    resultatT[i,3]  <- 1
  }
}

#Tar fram förväxlingsmatrisen
table(resultatT[,2:3])

#Beräknar utvärderingsmåtten
MissclassT <- (table(resultatT[,2],resultatT[,3])[2]+table(resultatT[,2],
resultatT[,3])[3])/sum(table(resultatT[,2],resultatT[,3]))

PrecisionT <- (table(resultatT[,2],resultatT[,3])[4])/(table(resultatT[,2],
resultatT[,3])[4]+table(resultatT[,2],resultatT[,3])[3])

RecallT <- (table(resultatT[,2],resultatT[,3])[4])/(table(resultatT[,2],
resultatT[,3])[4]+table(resultatT[,2],resultatT[,3])[2])
```