Kandidatuppsats i Statistik

Mixad mediamodell för marknadsföring

Sambandet mellan konvertering i antal nya konton och spendering på marknadsföringskanaler

Johannes Hedström
Mikael Montén

Avdelningen för Statistik och maskininlärning
Institutionen för datavetenskap
Linköpings universitet

2023-05-25

Handledare: Josef Wilzén
Examinator: Linda Wänström

Sammanfattning

Uppsatsen är skriven för CTRL Digital som är en digital marknadsföringsbyrå, som i sin tur har kontaktat ett ej namngivet företag inom finans & lån. Den utforskar spenderingen på olika marknadsföringskanaler och deras påverkan på antalet nya konton för ett här icke namngivet företag inom finans- och lånetjänster. Syftet är att statistiskt modellera sambandet mellan spendering och resultat för att hjälpa företaget att utvärdera effektiviteten av deras marknadsbudget och undersöka om en budgetförändring skulle resultera i ökad respons och fler nya konton. Dessutom är detta av intresse för CTRL Digital då en godtycklig modell som kan tillämpas för olika företag skulle utöka deras produktutbud.

Data samlades in från olika datamaterial genom BigQuery och aggregerades på daglig nivå. Gemensamma variabler såsom spendering och antal nya konton kombinerades i ett datamaterial. Tre typer av transformationer utfördes på spenderingsvariablerna för att fånga eventuella kvarvarande effekter av marknadsföring på antalet nya konton. Externa variabler som ansågs påverka frekvensen av nya konton inkluderades också. Därefter anpassades 11 olika XGBoost-modeller med de tillämpade transformationerna, och de tre bästa modellerna valdes ut. De valda modellerna inkluderade en modell utan transformation, en med en sönderfallshastighetstransformation och en med en andelstransformation av variablerna.

Resultaten visade icke-linjära samband mellan spendering på marknadsföring och antalet nya konton för företaget. Modellerna överanpassade på träningsdata, och för vissa marknadsföringskanaler fanns det för få observationer för att dra pålitliga slutsatser.

Dock visade modellerna antingen överanpassning till träningsdata, dessutom så fanns det för få observationer för att dra pålitliga slutsatser för alla marknadsföringskanaler.

För företagets tre största kanaler visade resultaten att Google ger mest effekt när det utgör cirka 10 procent av budgeten, medan Meta och Affiliate ger mest effekt vid ungefär 25 procent vardera, förutsatt att den totala marknadsbudgeten är relativt hög. Vid en andel på cirka 60 procent av dagens budget visade Affiliate också en betydande ökning av antalet nya konton, även om den totala spenderingen då var låg.

Abstract

This academic thesis, conducted for CTRL Digital, a digital marketing agency, explores the spending on various marketing channels and their impact on the number of new accounts for an unnamed company in the finance and loan services industry. The objective is to statistically model the relationship between spending and outcomes to assist the company in evaluating the effectiveness of their marketing budget and investigate whether a budget change would result in increased response and more new accounts. Additionally, this research is of interest to CTRL Digital, as a generalizable model applicable to different companies would expand their product offerings.

Data was collected from various data sources using BigQuery and aggregated at a daily level. Common variables such as spending and the number of new accounts were combined into a dataset. Three types of transformations were applied to the spending variables to capture any carryover effects of marketing on the number of new accounts. External variables considered to affect the frequency of new accounts were also included. Subsequently, 11 different XGBoost models were fitted with the applied transformations, and the top three models were selected. The chosen models included one without any transformation, one with a decay rate transformation, and one with a proportion transformation of the variables.

The results revealed non-linear relationships between marketing spending and the number of new accounts for the company. However, the models either exhibited overfitting to the training data and insufficient observations which hindered the ability to draw reliable conclusions for all marketing channels.

For the company's three largest channels, the results showed that Google provides the most effect when it accounts for approximately 10 percent of the budget, while Meta and Affiliate yield optimal results at around 25 percent each, assuming the overall marketing budget is relatively high. At a share of approximately 60 percent of the current budget, Affiliate also demonstrated a significant increase in the number of new accounts, even though the total expenditure was low.

Förord

Vi vill börja med att tacka Ted Solomon och övriga kollegor på CTRL Digital som vi har haft kontakt med för att de givit oss problemet samt att de försett oss med domänkunskap inom digital marknadsföring.

Tack till företaget som anförsett oss med datamaterialet.

Till sist ett stort tack till Josef Wilzén som handlett oss genom projektets gång, bidragit med insikt och värdefulla kommentarer.

Innehållsförteckning

Figurer

Tabeller

1 Inledning

Bakgrunden behandlar ämnet marknadsföring och hur det sett ut tidigare till dagens digitala industri. Ämnesbegrepp tas upp för att lättare hänga med i uppsatsen samt problem och information från tidigare studier nämns för att ge en överblick om hur hantering av liknande problem utförts.

1.1 Bakgrund

Marknadsföring är en viktig stapel för ett företags fortsätta växande (Kotler 2012), vilket globalt innebär stora investeringar. Traditionell marknadsföring, eller offline marknadsföring, är all marknadsföring som inte sker genom internet. Exempel på medium för detta är radio, TV och tidningar (Kotler 2012). Denna uppsats omfattar snarare den digitala marknadsföringen, eller online marknadsföring, som i modern tid växt sig väldigt stark. Online marknadsföring innebär de olika marknadsföringskanaler som är internetbaserade, exempelvis sociala medier, sökmotorer och bloggar. Under 2022 lades globalt 567 miljarder USD (Lebow 2023) på digital marknadsföring, med en så pass stor industri är det föga förvånande att kartläggning och optimering av hur pengarnas spenderas är en stor faktor av intresse för företagen.

En marknadsföringskanal är alltså det särskilda medium som en viss kampanj eller annons ska visas i och företag betalar annonsföretag - exempelvis Meta och Google - en summa för att deras reklam ska visas på dess användares skärmar. Kartläggning och optimering av hur dessa pengar spenderas grundar sig i att utvärdera effektivitet i marknadsföringen genom statistik. Digitaliseringen har gjort att detta är mer tillgängligt än någonsin då de stora annonseringsplattformarna för statistik över metriker som är intressanta för utvärdering.

Uppdragsgivaren till studien är CTRL Digital, vilket är en digital marknadsföringsbyrå och en cloud marketingbyrå som jobbar med att effektivisera och optimera deras kunders marknadsföring med hjälp av Google Cloud Platform och maskininlärning med användarnas integritet i åtanke. Deras team består av analytics engineers, digital analysts, technical marketers, och data scientists. De har kontakt med ett företag inom finans & lån som vill undersöka möjligheten att optimera fördelningen av budgeten de spenderar på sina marknadsföringskanaler och det är detta företag som bidragit med datamaterial till uppsatsen. Företaget som tillhandahåller datamaterialet till studien kommer inte att nämnas ur konkurrenssyfte.

Den data som tillhandahålls är tidsseriedata på dagsnivå över en period om cirka 2 år, med variabler som spendering per marknadsföringskanal för 7 kanaler och antal nya konton hos företaget. Marknadsföringskanalerna kommer att användas som förklarande variabler och responsvariabeln som definierats av projektgruppen i samråd med CTRL Digital är då antal nya konton skapade hos företaget. För att kunna analysera effekten från marknadsföring existerar en vedertagen metod som kallas mixed media modelling (även kallat marketing mix modelling, förkortas MMM), vilket är en regressionsmodell som använder sig av marknadsföringsdata för att förstå sambandet mot en försäljningsmetrik (responsvariabel) och som används för att optimera den blandning (mix) av marknadsföring med anseende mot försäljning som presterar bäst (Chan 2017). En mix av marknadsföring kan innebära olika nivåer av budgetar eller olika former av kampanjer som ett företag vill använda sig av. Genom MMM kan man analysera vad förväntade försäljningssiffror är om alla kanaler sätts till 0, d.v.s är inaktiva, och vad den marginella skillnaden en inkludering eller budgetjustering i dessa kanaler medför. Detta gör det möjligt för marknadsförare att beräkna hur företagets nyckeltal (Key Performance Indicators, KPI) påverkas och således ger mer insyn i påverkan från spenderingen, vilket är av intresse då marknadsföring tenderar vara svåranalyserat i vad man faktiskt får ut av pengarna man spenderar.

Effektivitet från marknadsföring kan mätas på flera sätt genom användning av KPI. Dessa kan skifta beroende på vad ett företag vill mäta, men exempel är Cost of Goods Sold (COGS) (CFI 2022a) som är avkastning genom totala kostnaden för varje såld vara, Cost per Acquired Customer (CAC) (CFI 2022b) som är antalet nya kunder genom totala kostnaden för marknadsföring eller Return on Ad Spend (ROAS) (CFI 2022c) som är avkastningen per spenderad krona i marknadsföring.

För att kunna kartlägga och optimera hur pengar spenderas är det av högt intresse att estimera hur mycket av respektive marknadsföringskanal bidrar till förändring i de olika KPI:erna. Att kunna attribuera en konvertering (försäljning av en vara) till en särskild marknadsföringskanal är vitalt för att optimera marknadsföringsbudgeten mellan de olika kanalerna.

1.2 Tidigare studier

Marknadsföringsdata samlas ofta in över tid och det är välkänt att tidsserier kan resultera i modeller där residualerna har hög autokorrelation och att det då är svårt att dra slutsaser och att tolka parametrar och samband då många metoder antar oberoende bland residualerna (Newbold. Paul 1974). Detta måste tas i åtanke vid modellutvärdering för att säkerställa att tidsberoende inte finns för modellens residualer.

Det finns forskning som pekar på att det är rimligt att tro att det finns en effekt för marknadsföringen mellan kanaler (Dinner 2014), vilket är att kanalerna samverkar och möjligtvis har interaktionseffekter mellan varandra. Detta har skapat ett intresse hos företag för att förstå hur de ska fördela budgeten mellan kanalerna för att få maximal avkastning av marknadsföringen.

Utmaningar som tidigare studier identifierat vad gäller mixed media-modeller är till exempel att inte tillräckligt med data finns tillgängligt, att variabler som driver försäljning lämnas utanför modellen, reklam som riktas mot människor som redan visat intresse för en tjänst/produkt och säsonger som modellen inte pricksäkert fångar upp (Chan 2017).

Studier tar upp att det inom marknadsföring är välkänt att det finns en lagg-/carryover-effekt, det vill säga att köpet inte nödvändigtvis sker samma dag som kunden sett en annons utan det kan ta allt från dagar till månader innan konvertering och olika variabler kan tänkas ligga bakom detta och att det ska finnas i åtanke vid modellskapandet (Jin 2017) . Detta fenomen tas även upp i (Pandey 2021) som beskriver att det oftast är en effekt som under en kortare period och att den långa effekten som vissa studier visar på snarare är att datamaterialet som då använts är vecko- eller månadsnivå.

För att hantera den icke-linjära effekten använder sig (Jin 2017)av Hills-funktionen för att transformera förklaringsvariabler för att ge parametrarna i modellen en mättad effekt, ett annat sätt är att använda sig av modeller som hanterar icke-linjäritet likt XGBoost för att låta modellen hitta icke-linjära mönster, både med eller utan transformering av variabler (Wigren R 2019).

En ytterligare lämplig icke-linjär transformation är att testa flera olika sorters sönderfallsmönster för spenderat per kanal för att hitta det mönstret som skapar bästa modellen för konverteringar (J. C. Wolfe Michael och Crotts 2011). Detta då spendering på marknadsföring ofta sker över en längre tidsperiod samt att det ska kunna beskriva en eventuell carryover-effekt.

"Shape effect" är ett annat marknadsföringsfenomen, vilket är skillnaden i konverteringar när det spenderas mer på marknadsföring (Pandey 2021) . Hur detta samband ser ut är det som är intressant när det kommer till att optimera budgeten och att anta att den är linjär är logiskt omöjlig, då det säger att försäljningen går till oändligheten om spenderingen också gör det. Det är därav rimligt att tro att det är en icke-linjär form på sambandet, exempelvis en konkav eller S-formad kurva är mer trolig, speciellt den S-formade kurvan då det förväntas att lågt spenderade summor ger en väldigt liten effekt på försäljning och "äts upp" av marknaden samt en låg ökad effekt på försäljning när spenderingen är väldigt hög (Pandey 2021).

Ett problem med maskininlärningsmetoder är att det ofta är svårt att tolka effekten av varje enskild variabel på prediktioner, XGBoost kan attribuera feature importance till variabler i modellen men dessa är ofta inkonsekventa och inte riktigt pålitliga, istället kan SHAP-värden användas för att få ett mer pålitligt resultat (S. M. Lundberg and Lee 2017). En annan fördel med SHAP-värden är att de tar hänsyn till interaktionseffekter mellan förklarande variabler för att se hur olika värden på variablerna interagerar och påverkar SHAP-värdet (S. M. Lundberg and Lee 2017).

1.3 Syfte

Rapportens syfte är att skapa en robust modell för att förstå vilket samband spendering på marknadsföring har på antal nya konton hos företaget. Samtidigt är det av intresse att ta fram en metod som CTRL Digital kan använda sig av till andra kunder för att utöka deras arbetsätt.

Eftersom så mycket pengar spenderas på marknadsföring är det av intresse att förstå vad det ger för effekt på konverteringar, företaget har funderingar som lyder: "Bör budgeten förändras? Ska vi spendera mer eller mindre pengar?" och "Bör vi förändra budgetfördelningen mellan kanalerna?".

Förhoppningsvis kan rapportens resultat ge företaget en bättre bild över deras marknadsföring och samtidigt ge dem en riktning om eventuella förändringar av budgeten som enligt modellen ger en bättre avkastning mellan spenderat och antal nya konton.

Frågeställning

- För vilken spenderingsnivå påvisar kanaler en positiv effekt på antal nya konton?
- Kan transformering av spenderingen användas i modeller för att förklara den långvariga effekten från marknadsföring?
- Vilken andel bör varje enskild kanal ha av totalbudgeten för att ge en hög avkastning?

1.4 Avgränsningar

Vissa avgränsningar har verkställts för att arbetet inte ska bli för stort eller komplext och efter uppdragsgivarens instruktioner, denna studie kommer att fokusera på data endast från företagets norska avdelning då det är där det finns mest data tillgängligt. Det blir då också ingen effekt för skillnader i marknadsföring mellan länder som kan påverka resultatet och modellstrukturen, samt att ingen hänsyn till olika valutor behöver tas i åtanke på de spenderade summorna.

Marknadsföringskanalerna har dels inte nödvändigtvis samma definition för de insamlingsmetriker de använder sig av för analys, samt för en stor del av kanalerna saknas sådana metriker till fullo i datamaterialet studien har tillhandahållit vilket gör att vissa kommer att lämnas utanför modellskapandet.

Responsvariabeln valdes till antal nya konton hos företaget och inte en direkt försäljning då det ansågs vara den bästa variabeln för att beskriva ett resultat av marknadsföringen då andra icke-kontrollerbara faktorer anses påverka en försäljning av företaget.

1.5 Etiska och samhälleliga aspekter

Den data som används är hämtat från Googles molnbaserade data warehouse BigQuery. Ett data warehouse är ett datalager för lagring av data från olika källor, som sedan kan användas för att hämta och behandla data som är av intresse. BigQuery är ett väldigt vedertaget warehouse och har säkerhetsfunktioner för att data hanteras på rätt sätt. Data rapporteras direkt till BigQuery utan manuell hantering vilket innebär att det mänskliga felet reduceras.

Det finns etiska aspekter som måste övervägas vid arbete med marknadsföring och särskilt digital marknadsföring. Som koncept innebär digital marknadsföring att man samlar och kombinerar data från både online- och offlinekällor vilket sedan analyseras för att öka förståendet för sina kundmålgrupper (Shamsuzzoha 2021). Den data som används för uppsatsen gäller inte data insamlad på individuella kunder, så det finns ingen egentlig risk att hänga ut enskilda individer. Samtidigt går teknologin framåt och kunder vet inte hur data samlas

in och används vilket gör dem sårbara. Det finns en balansgång mellan att kunder vill ha mer personlig och individualiserad marknadsföring mot att bibehålla individernas integritet, vilket är en av de största aspekterna inom etik för marknadsföring tillsammans med att data hanteras på rätt sätt (Shamsuzzoha 2021).

Även om insamlingsprocessen till stor del är automatiserad och processad genom vedertagen teknologi finns det ytterligare problem. Amnesty har riktat hård kritik mot Google och META att deras företagsmodeller är ett hot mot integritet och mänskliga rättigheter. Google och META är två stora annonsplattformar som används i detta arbete, och inte minst då data lagras och hanteras av BigQuery som också är en Google produkt (Shamsuzzoha 2021).

I syfte att motverka integritets- och dataintrång lanserade i maj 2018 General Data Protection Regulation (GDPR), som reglerar användning och hantering av personlig data för invånare inom EU. Den gäller för alla företag som på något sätt hanterar EU-invånares data, oavsett var företaget är lokaliserat eller om det hanterar data indirekt åt ett annat företag och överträdelser av GDPR innebär tunga böter för företagen (Shamsuzzoha 2021).

2 Data

Datamaterialet som används tillhandahålls av företaget genom Google Clouds datawarehouse BigQuery. Totalt är det sju olika dataset som existerar i det BigQuery-projekt som används, vilka är olika data från olika datainsamlingsprocesser som företaget använder sig av. De sju aktuella dataseten är;

- google_ads
  - Ett stort dataset som innehåller över 100 tabeller med olika variabler och metriker från Google som analyseras för att marknadsförare ska kunna få insikt i sin data.
- facebook_ads
  - Innehåller tabell med ett 30-tal variabler och metriker som är intressanta för marknadsförare från META.
- aggregated
  - Innehåller en tabell med information kring var trafik till företagets hemsida kommer ifrån, samt olika metriker för exempelvis annonskostnad per annonsplattform och variabler för hur långt i låneansökningsprocessen en kund kommit.
- analytics
  - Innehåller olika metriker gällande vad en kund gjort på hemsidan och även om en låneansökningsprocess har startats.
- snapchat_ads
  - Tabell innehållande ca 40 variabler med metriker relevanta för digital marknadsföring från Snapchat.
- tiktok_ads
  - Tabell innehållande ca 10 variabler som mäter intressant data i utvärderingssyfte från plattformen Tiktok.
- Funnel
  - Innehåller data från andra kanaler som inte mäts med samma typ av metriker som de föregående strikt digitala kanalerna. Här finns data från de så kallade Funnel-kanalerna influencers, Programmatic och Affiliate. Programmatic, eller programmerad annonsering, är en automatiserad form av annonsering där beställning, betalning och placering av annons sköts automatiskt och förutbestämt. Influencers och Affiliate är olika former av samarbetspartners.

Dessa dataset har alla data med olika tidsdetaljnivå, men generellt finns det flera observationer för varje dag av olika variabler. För att kunna utföra säkerställd analys har arbete utförts i att sammanställa information från dessa olika dataset som är tillförlitlig i form av att definiera en tidsperiod som existerar i alla tabeller över en rimlig tidsdetaljnivå, samt matcha annonsplattformarnas gemensamma variabler. För att göra detta har data aggregerats till dagsnivå, vilket innebär att de olika numeriska variablerna har summerats till per dag. De variabler som anses vara relevanta som genomgående finns i datamaterialen är spenderingen varje dag per kanal. Utöver spenderingsvariabler inkluderas också ett antal externa variabler i syfte att hjälpa modellen förklara sambandet mer. Två dummyvariabler inkluderas för veckodag och månad, och en variabel som är responsvariabeln förskjuten med en dag (laggad), syftet med dessa tre är att hjälpa modellen hantera det tidsberoendet som uppstår i observationerna. Två ytterligare externa variabler används också: inflation, som redovisar vilken procent inflationen låg på för respektive månad i Norge under perioden, och totalspendering som summerar dagsspenderingen för alla kanaler. Det slutgiltiga datasetet innehåller data på dagsnivå under perioden 7 juni 2021 - 31 januari 2023. Totalt är det 604 observationer över 29 variabler.

Tabell 1: Beskrivning av datamaterialets variabler

	Variabler	Beskrivning
1	Antal nya konton	Responsvariabel Y. Antalet nya registrerade konton hos företaget
2	Spenderat	Dagsspendering för de 7 kanalerna. Kanal noteras med prefix.
3	Inflation	Inflationstakten för Norge per månad under tidsperioden
4	Veckodag	7 dummyvariabler för individuell veckodag. Aktuell veckodag noteras som 1, annars 0.
5	Månad	12 månadsvariabler för enskild månad. Aktuell månad noteras som 1, annars 0.
6	Lag Y	Responsvariabel Y förskjuten en dag i tiden. Värdet på variabeln är vad Y var dagen innan.

Tabell 1 visar de variabler som används i modellen.

2.1 Transformationer och bearbetning

2.1.1 Rullande kumulativ summa

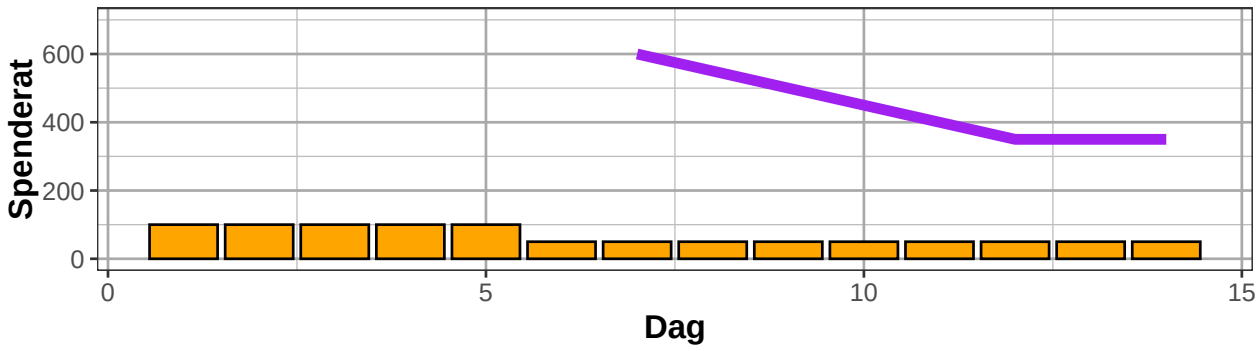
Eftersom spenderingen på marknadsföringen i teorin anses kunna ha en uppbyggande effekt över flera dagar så kommer rullande kumulativa summor beräknas för olika tidsintervall för att skapa en mer realistisk effekt på det som spenderats.

Beräkningen kommer ske enligt:

$$\hat{x}_{t,i} = \sum_{j=1}^{L} x_{t,i} + x_{t-1,i} + ... + x_{t-L,i} \tag{1}$$

Där $x_{t,i}$ är spenderingen för kanal i vid tidpunkt t och L är längden på tidsintervallet som summeras.

4 olika typer av intervall kommer att testas för att se vilken som förklarar spenderingen bäst och dessa är 7, 14, 21 och 28 dagar.



Figur 1: Staplarna är sann spendering per dag, det lila strecket är rullande kumulativa summan för de senaste 7 dagarna.

I figur 1 visas ett exempel på hur en rullande kumulativ summa på 7 dagar fungerar, de första observationerna försvinner från tidslinjen då de inte har en tillräckligt lång period att summeras över innan tiden det spenderats. Istället för ett rakt hopp från 100 till 50 så avtar den kummulativa summan sakta och planar ut efter 7 observationer med samma spendering.

2.1.2 Enkel sönderfallshastighetsfunktion

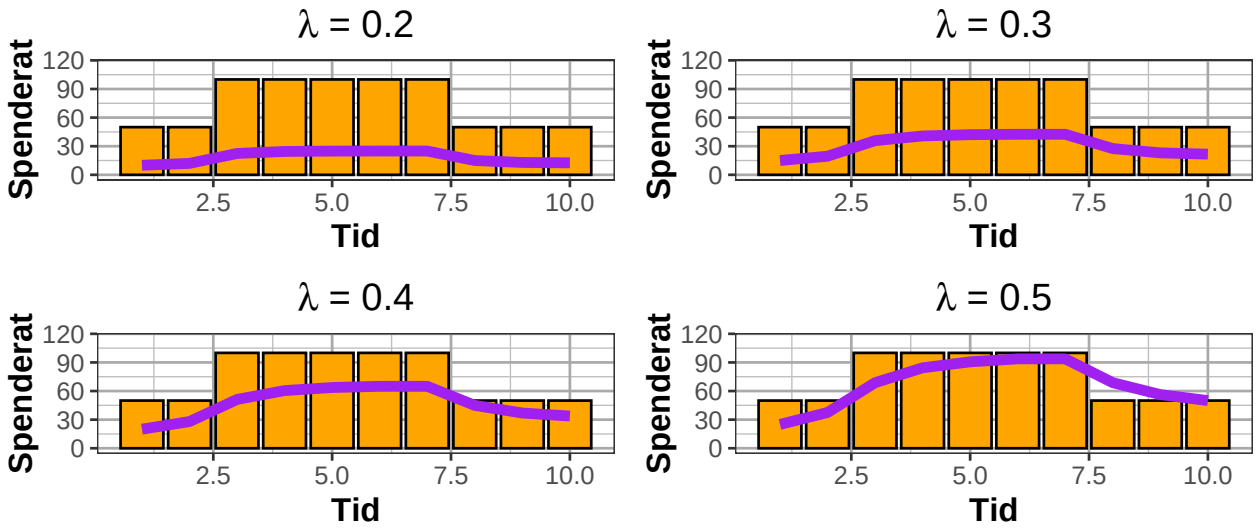
Ett annat sätt att transformera spenderingen för att realistiskt förklara kvarvarande effekten av marknadsföringen är att göra en sönderfallshastighetfunktion (M. Wolfe 2011). Detta innebär att effekten från dagens spendering kommer att påverka de kommande dagarna, men med en lägre effekt för varje dag som går. Vidare kommer denna funktion kallas för Decay.

Denna effekt kommer beräknas enligt:

$$a_{i,t} = \sum_{j=max(1,t-L)}^{t} \lambda^{t+j-1} \cdot x_{i,j} \tag{2}$$

Där λ vanligtvis väljs enligt industristandard för den aktuella branschen om det finns en sådan standard, L är antal dagar som effekten anses gå över, T är längden på tidsserien, j är dagen för observationen, i är vilken kanal och t är tidpunkt för effekten.

Den här studien kommer att testa olika sönderfallshastigheter på $\lambda = 0.5, 0.4, 0.3, 0.2$, då ingen standard för branschen har hittats. Tidsperioden L som används är 14 dagar, detta anses vara en tillräckligt lång period för att effekten ska ha mynnat ut och kvarvarande beloppet från spenderingen är nära noll.



Figur 2: Exempel på sönderfallshastighet för alla λ som kommer användas från ekvation 2. Oranga staplar är spendering, lila strecket är den transformerade spenderingen för de olika λ-värdena. L = 4.

Effekten av spenderingen planas ut och ger en mjukare kurva över tid när den transformeras för sönderfall. När λ är 0.5 tillåts spenderingen byggas upp mer, men när λ minskar resulterar det i en mer restriktiv effekt.

2.1.3 Andelar

Det är också av intresse att testa om andelen spenderat för varje kanal per dag förklarar antal nya konton bättre, dessutom så skulle detta kunna förklara en budgetfördelning för kanalerna. För att beräkna varje andel så beräknas radsumman för all spendering på kanalerna för varje dag till en total dagssumma för spenderat på marknadsföring.

$$\text{Total spendering}_t = \sum_{i=1}^{k} X_{it} \tag{3}$$

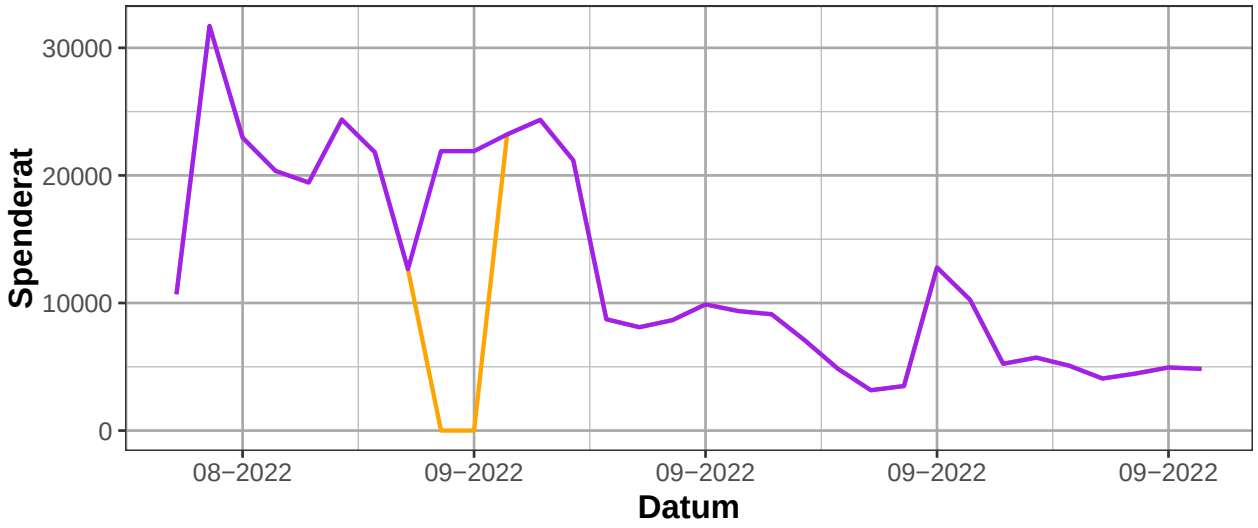
Där k är antal kanaler och t är dag i datamaterialet.

Sedan divideras varje enskild spendering per dag med totalsumman för den dagen enligt:

$$\text{Andel}_{it} = \frac{X_{it}}{\text{Total spendering}_t} \tag{4}$$

2.2 Imputering

I datamaterialet för Google-annonser saknades data under två dagar, 2022-09-04 samt 2022-09-05. Efter kontakt med representant från företaget är det sannolikt att Googles datasystem har legat nere varav inga observationer från deras sida finns. Spenderingen på Google var mycket högre under den veckan och därav imputeras de saknade värdena med ett genomsnitt för den veckan istället för att kolla på längre säsongsmönster. Nedan följer en graf för att illustrera imputeringen.



Figur 3: Imputering för saknade värden hos Google. Lila linjen är hur datamaterialet ser ut för närvarande, orangea linjen är hur tidsserien såg ut innan imputering.

Här syns att imputeringen som gjorts ser lämplig ut då den inte påverkar hur tidsserien beter sig åt något håll. Båda dagarna har fått samma värde.

2.3 Beskrivande statistik

Nedan följer beskrivande statistik för de olika spenderingarna.

Tabell 2: Variabelbeskrivning för spendering per kanal

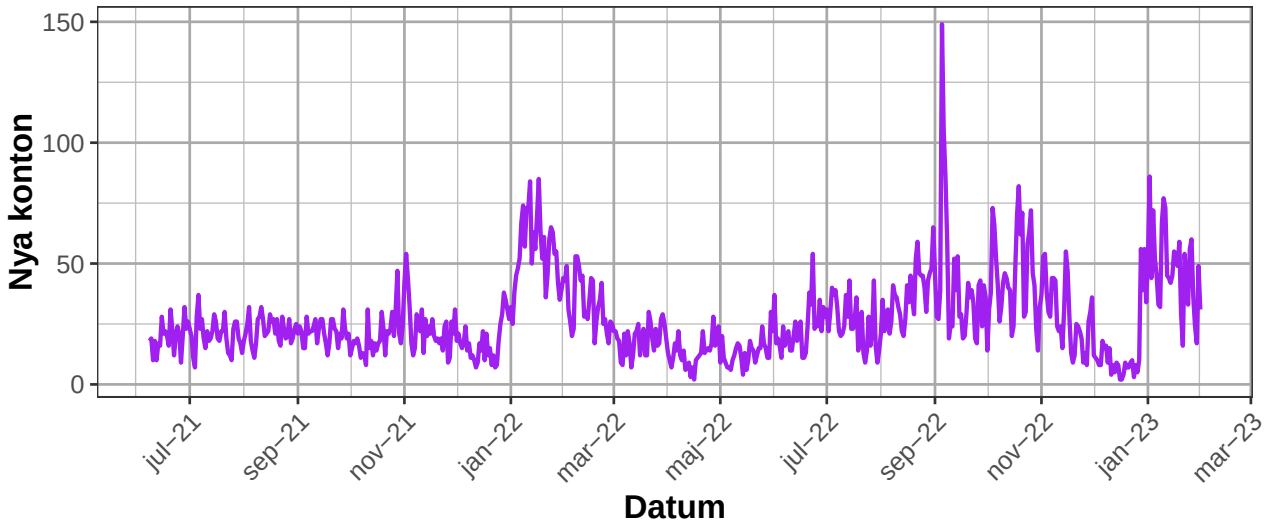
	Variabel	Median	Medelvärde	Std.Avv	Min	Max
1	Meta	6780.79	7499.75	5113.05	354.450	26826.130
2	Google	1004.59	3099.36	4354.06	56.843	31713.919
3	Tiktok	0.00	428.71	1235.79	0.000	5500.000
4	Snapchat	0.00	14.04	67.12	0.000	400.000
5	Affiliate	99.19	3807.76	5410.27	0.000	26256.531
6	Influencer	0.00	205.33	847.84	0.000	7222.668
7	Programmatic	0.00	1347.29	1732.49	0.000	17373.624

Meta har klart högst spendering vad gäller median och medelvärde. Affiliate har ett högre medelvärde än till exempel Google, men en väldigt låg median, vilket beror på att de haft långa perioder utan spendering och när de väl använder kanalen spenderar de mycket. TikTok, Snapchat, Influencer och Programmatic har alla en median som är lika med noll vilket också innebär att det finns få observationer som är nollskilda för de kanalerna. Särskilt syns att Snapchat har spenderats på så låga nivåer att kanalen troligtvis inte bidrar med någon effekt på antal nya konton.

2.4 Visualisering av data

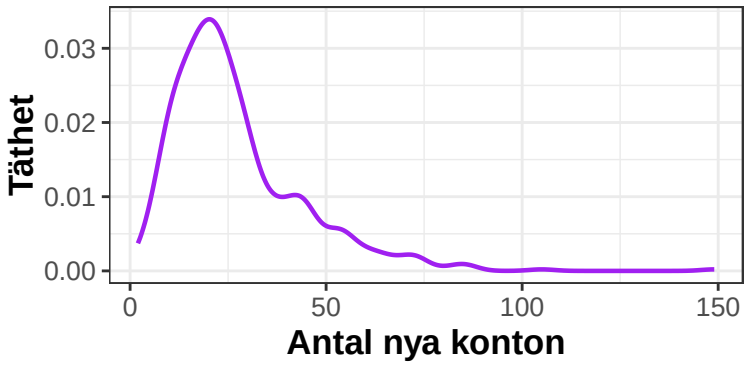
Generellt när det gäller tidsserier är det vissa aspekter som trend, säsong och cykliskt mönster som undersöks för att bättre förstå datamaterialet och hur eventuella problem ska angripas (Hyndman 2021).

2.4.1 Antal nya konton



Figur 4: Antal nya konton per dag

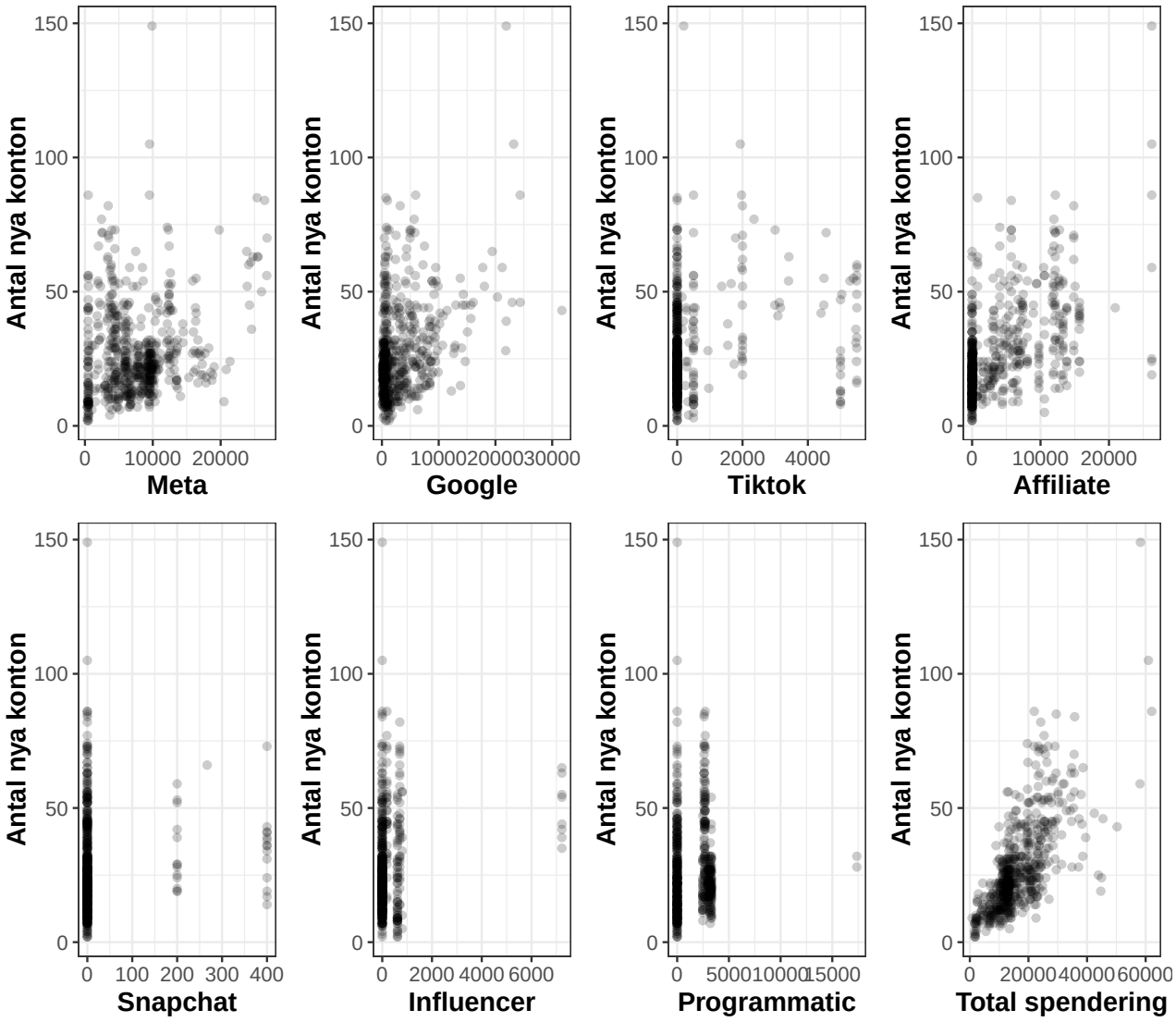
I figur 4 visualiseras antal nya konton som skapats per dag under perioden 7 juni 2021 till 31 januari 2023. Det högsta antalet skapade konton på en dag är 149 stycken vilket var den 5 september 2022. En väldigt svag positiv trend tycks finnas över perioden. För januari ökar antalet nya konton jämfört med resterande månader, men det är svårt att dra slutsatser kring säsongsmönster över månader då ej tillräckligt med observationer finns.



Figur 5: Fördelningen för antal nya konton

För att undersöka om tidigare noteringar stämmer så plottas fördelningen för antal nya konton, det syns tydligt att den är högerskev och extremvärdet skiljer mycket mot den resterande värden för variabeln.

2.4.2 Kanalspendering mot antal nya konton



Figur 6: Samband mellan antal nya konton och kanalspenderingar. Punkters nyans bestäms av densiteten, där mörkare innebär högre densitet.

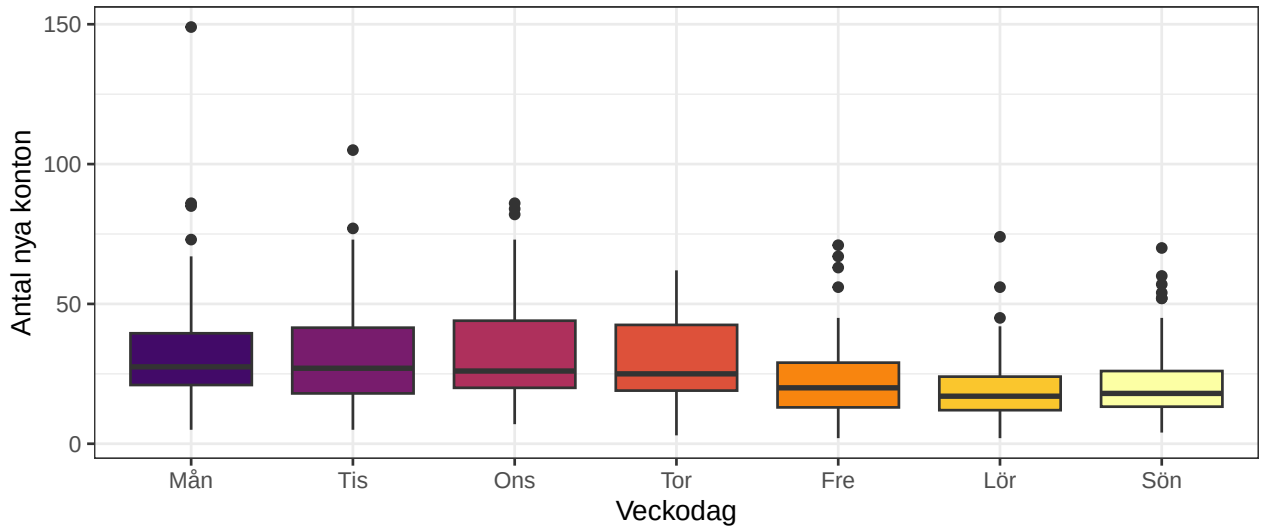
Det är svårt att tyda några starka linjära samband mellan antal nya konton och marknadsföringskanalerna vilket syns i figur 6. Detta motiverar undersökning av transformationer på variabler samt använda metoder som kan hitta icke-linjära samband. Eftersom Influencer, Programmatic och Snapchat spenderats i så få nivåer kommer det vara svårt att identifiera säkerställda samband, det enda som syns är att variansen skiljer sig markant för samma nivåer på spendering. Kanalerna med fler observationer påvisar något form av samband, dock inte entydigt linjärt och svårtolkat.

Tabell 3: Korrelationer för alla spenderingar samt nya konton

	Nya konton	Meta	Google	Tiktok	Snapchat	Affiliate	Influencer	Programmatic
Nya konton	1.00	0.21	0.38	0.24	0.10	0.54	0.18	0.09
Meta	0.21	1.00	-0.21	-0.32	-0.10	-0.33	0.09	0.41
Google	0.38	-0.21	1.00	0.18	0.15	0.52	-0.08	-0.30
Tiktok	0.24	-0.32	0.18	1.00	-0.02	0.37	0.07	-0.02
Snapchat	0.10	-0.10	0.15	-0.02	1.00	0.42	-0.01	-0.16
Affiliate	0.54	-0.33	0.52	0.37	0.42	1.00	0.05	-0.29
Influencer	0.18	0.09	-0.08	0.07	-0.01	0.05	1.00	0.00
Programmatic	0.09	0.41	-0.30	-0.02	-0.16	-0.29	0.00	1.00

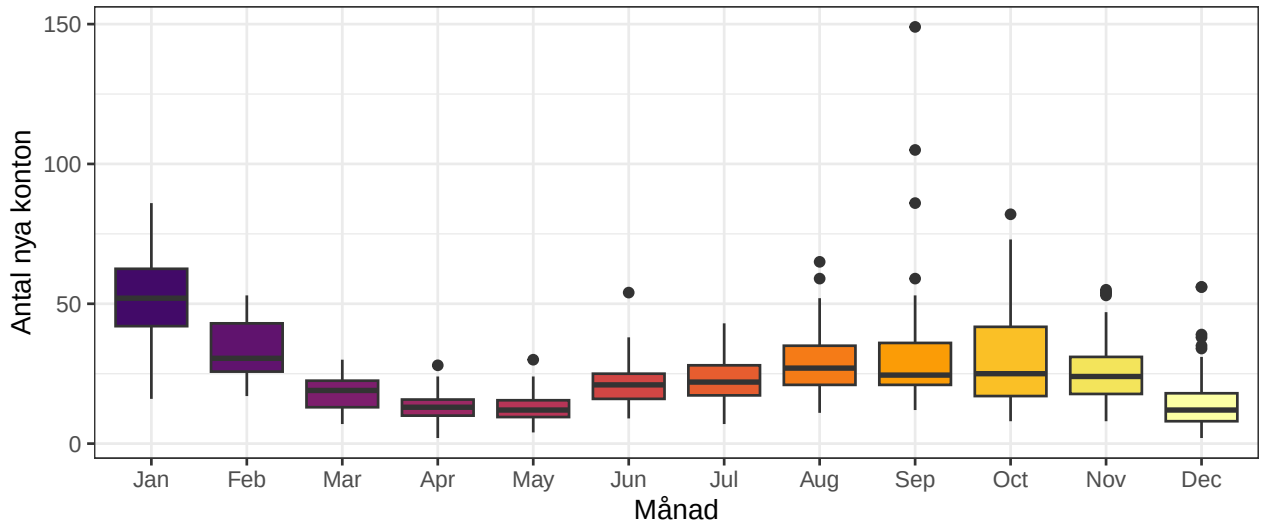
Korrelation Tabell 3 visar att beloppet spenderat på Affiliate är det som starkast korrelerar med antal nya konton med en korrelation på 0.54, de förklarande tidsserierna som starkast korrelerar med varandra är Affiliate och Google. Affiliate har relativt stark korrelation med flera förklarande variabler men ingen korrelation överstiger 0.52. Programmatic och Snapchat är de två kanalerna som har svagast korrelation med antal nya konton med en korrelation på 0.09 respektive 0.1. Detta stärker slutsatsen som tagits från figur 6 angående att en metod som kan hitta icke-linjära samband är intressant att undersöka då ingen variabel har jättestark korrelation mot antal nya konton.

2.4.3 Säsongsmönster



Figur 7: Lådagram över antal nya konton per veckodag.

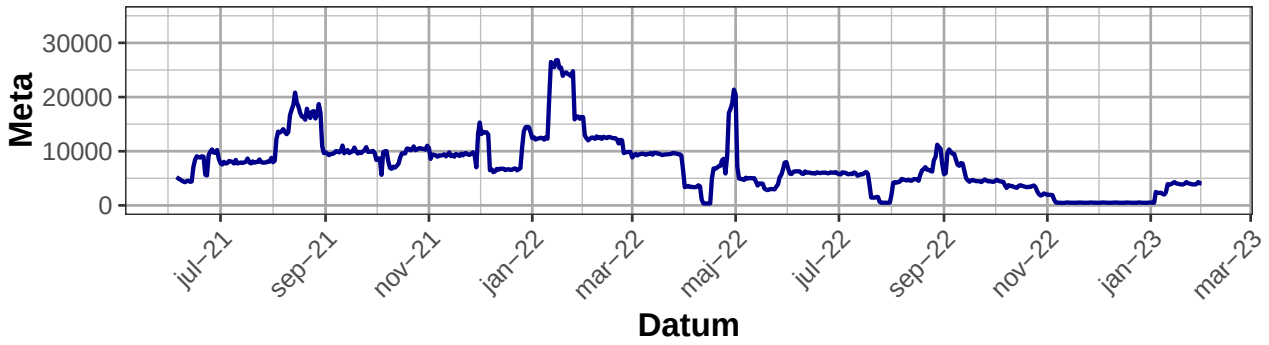
Generellt syns en mindre ökning under början av veckan fram till torsdag, varpå antal nya konton sjunker markant under fredag, lördag och söndag. Vissa extremvärden finns för helgdagarna men generellt så är magnituden lägre. Måndag och tisdag är de dagar när stora extremvärden finns med värden över 100.



Figur 8: Lådagram över antal nya konton per månad.

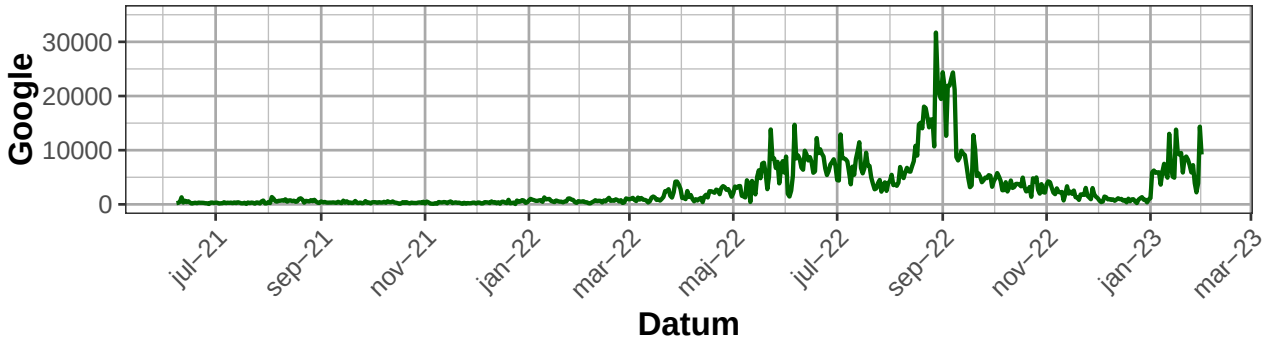
Ur figuren syns att januari är den månad med klart högst genomsnittligt antal skapade nya konton. Snittet sjunker under vårmånaderna och ökar sedan något under sommar och höst, men sjunker återigen under slutet av året. September är den månad med de högsta värdena. Noteras dock att datamaterialet inte går under 2 kompletta år så februari till maj är endast med en gång vilket bidrar med osäkerhet.

2.4.4 Spenderat per kanal



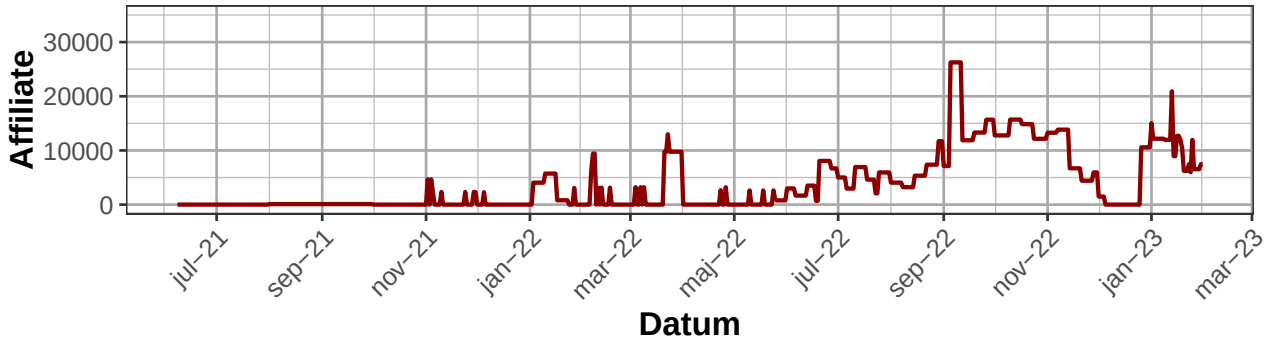
Figur 9: Spenderat på Meta per dag

Meta Figur 9 visualiserar beloppet som spenderats på Metas marknadsföringsplattform för tidsperioden i datamaterialet. En nedåtgående trend ser ut att finnas, samt att toppar uppstått i oregelbundna månadsintervall. Den högsta spenderade summan är den sextonde januari 2022, då 26138 NOK spenderats.



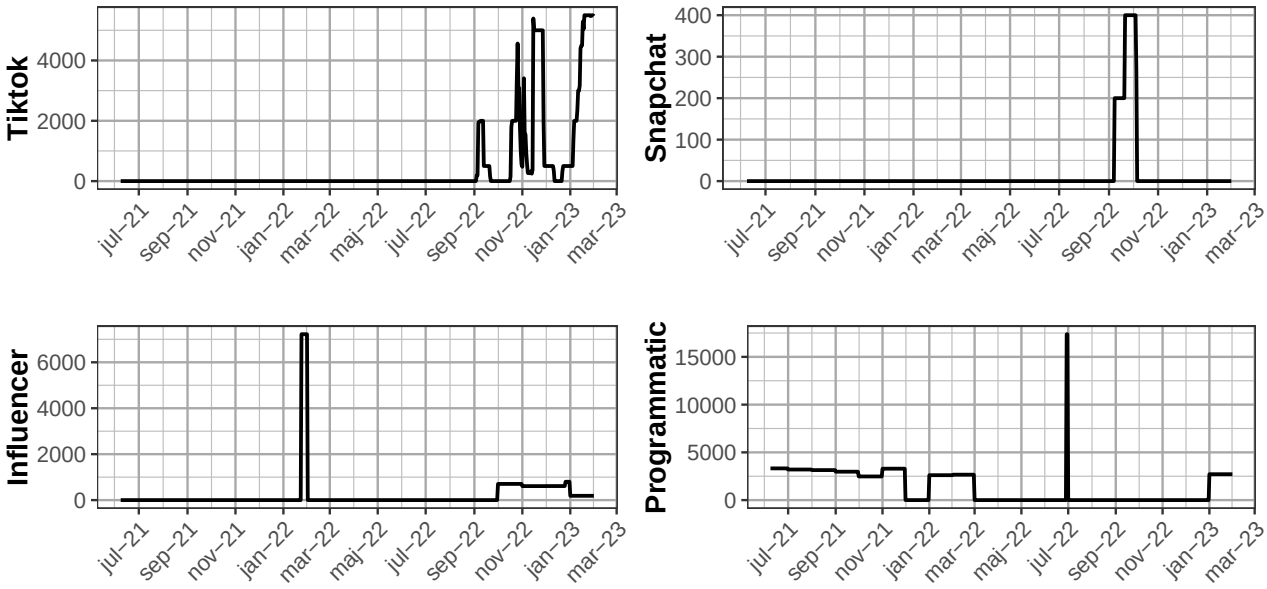
Figur 10: Spenderat på Google per dag

Google I figur 10 illustreras beloppet som spenderats per dag på annonsering via Google. Här syns att beloppet skiljer sig väldigt mycket mellan första och andra halvan av perioden. Den 28 augusti 2022 är dagen då mest pengar spenderats på Google, vilket är 31713 NOK.



Figur 11: Spenderat på Affiliates per dag

Affiliate Företagets spendering på Affiliates som visas i figur 11 ser ut att följa liknande trend som Google. Dagarna det spenderas som mest på Affiliates är under en vecka mellan 5 september 2022 och 11 september 2022, då spenderas cirka 26000 NOK.

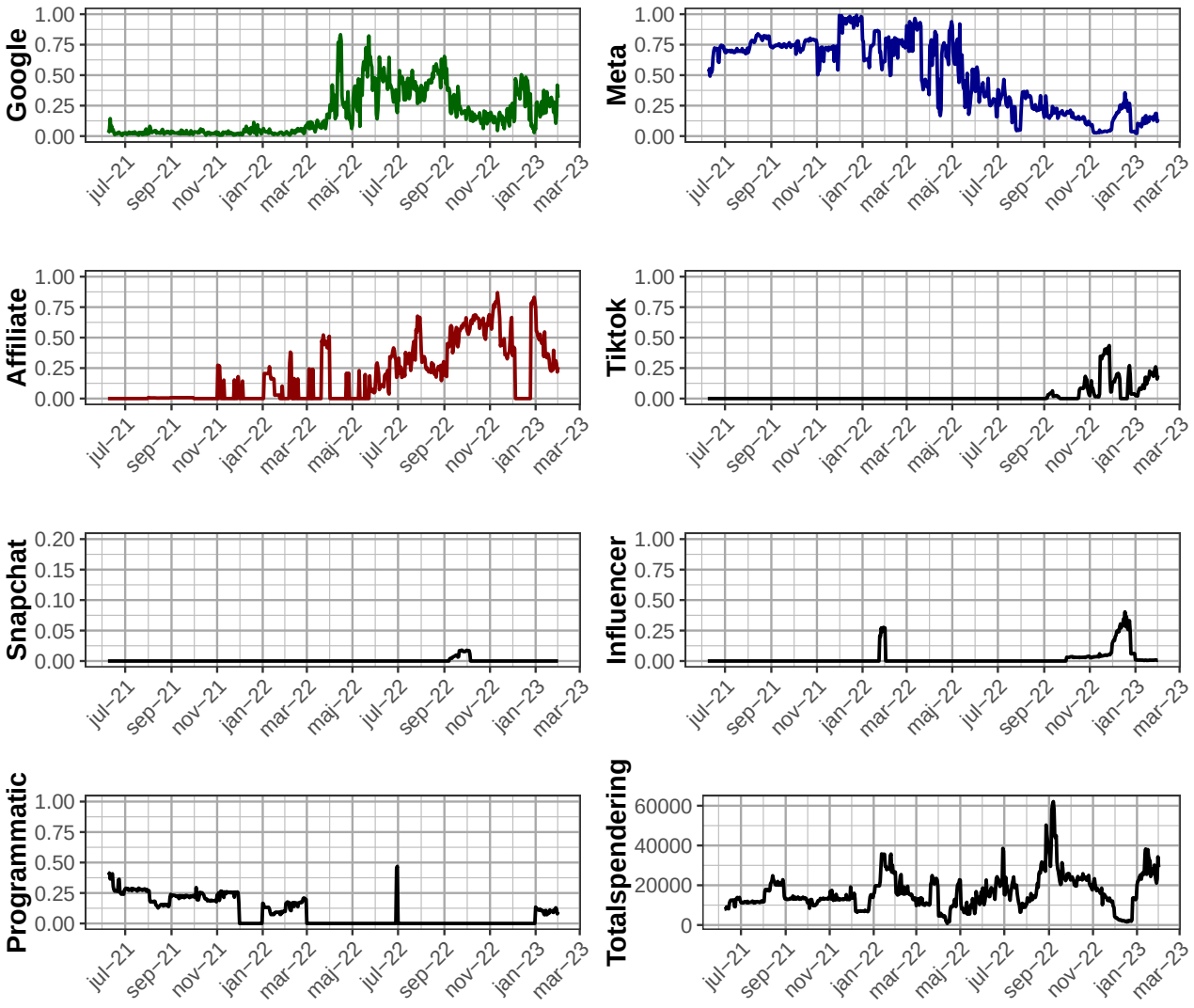


Figur 12: Övriga kanaler, notera skillnaderna i skalan på Y-axlarna

I figur 12 så ser vi de fyra övriga marknadsföringskanalerna som det spenderats pengar på sedan 7 juni 2021. De här kanalerna har mer sporadisk spendering än tidigare förutom TikTok som det har börjat spenderats på efter sommaren 2022. För dessa kanaler är spenderingen noll för en majoritet av dagar i datamaterialet, vilket gör det problematiskt att modellera samband mellan variabler. Notera att graferna har olika skalor på y-axlarna, högsta värdet på Tiktok är 5500 och högsta värdet på Snapchat är 400 NOK.

2.5 Andelar

Andelarna för respektive spendering under tidsperioden visualiseras för att ge en förståelse över hur marknadsföringsbudgeten förändrats över tid.



Figur 13: Kanalernas spenderade andel och totala spenderingen över tid, notera att Snapchat och totalspendering har en annan y-skala

Figur 13 visar liknande mönster som tidigare tidseriediagram, Meta har en negativ trend i andelar medan Google och Affiliate har positiv trend där andelen ökar 2022 jämfört med 2021. Tiktok som endast används från slutet på sommaren 2022 tar upp en större del av budgeten under slutet av året och är som mest runt 40 procent av en dagsspendering. Influencer har som Tiktok liknande mönster och får en ökad andel i slutet på 2022 och in i 2023.

Snapchats Y-axel har en mindre skala än alla andra grafer för att en skillnad av andelen ska synas, den rör sig aldrig över 2 procent av en dagsspendering. Programmatic bestod under 2021 till runt 25 procent av marknadsföringsbudgeten, men andelen av dess budget dras ner till 0 under 2022 förutom en kort period i juli.

Om totala spenderingen tas i hänsyn med andelen så syns det att när Tiktok och influencer har en hög andel så är den totala spenderingen relativt låg jämfört med hur det sett ut innan och efter deras toppar. Budgetfördelningen skiljer sig mycket över tid mellan kanalerna och den totala spenderingen gör också det.

Genomsnittliga andelar 7 september 2022 till 31 januari 2023

I syfte att förstå hur budgetfördelningen sett ut under en period där spendering har skett på alla kanaler så väljs alla dagar mellan 7 september 2022 till 31 januari 2023 ut. Denna period har valts dels då det är det mest aktuella segmentet av datamaterialet, men också då 7 september är första dagen som företaget spenderar pengar på TikTok. Detta gör perioden intressant för att kunna utvärdera budgeten och jämföra kanalspenderingar godtyckligt.

För att förstå hur budgeten sett ut i slutet på tidsperioden så väljs första dagen då TikTok spenderas ut som starten på perioden. Detta för att kunna utvärdera om budgeten går att optimera och för att bättre kunna jämföra vissa spenderingar.

Tabell 4: Medelvärde för spendering för respektive kanal under perioden

	Andel	NOK
Meta	0.136	2709.938
Google	0.221	4490.528
Tiktok	0.099	1723.434
Snapchat	0.002	56.445
Affiliate	0.457	10220.079
Influencer	0.063	440.205
Programmatic	0.021	560.677
Total	1.000	20201.306

Medelvärdet för andelen i budgeten skiljer sig mycket mellan kanalerna, Affiliate har störst andel av budgeten i perioden följt av Google och Meta, medelvärdet för dagstotalen är runt 20 000 NOK.

2.6 Slutgiltigt datamaterial

Alla spenderingar på kanaler transformeras och 10 nya datamaterial skapas.

Tabell 5: Datamaterialen

	Datamaterial	Antal variabler
1	Base	28
2	Kumsum 1 vecka	28
3	Kumsum 2 veckor	28
4	Kumsum 3 veckor	28
5	Kumsum 4 veckor	28
6	Decay 0.5	28
7	Decay 0.4	28
8	Decay 0.3	28
9	Decay 0.2	28
10	Komplett	84
11	Andelar	29

Gemensamt innehåller alla datamaterial antal nya konton, transformerad/icke-transformerad spendering per kanal, inflation, förskjuten Y-variabel, samt dummies för månader och veckodagar. Utöver detta innehåller datamaterial 11 även en variabel för total spendering då kanal-variablerna är relativa mått. Det kompletta datamaterialet innehåller alla transformerade variabler förutom andelstransformationen. Kanalvariabeln för Snapchat har exkluderats ur datamaterialen då spenderingen varit väldigt inkonsekvent och på mycket låga nivåer vilket skapar problem vid modellering.

3 Metod

I detta kapitel kommer metoderna som uppsatsen använder för att besvara frågeställningarna till datamaterialet som tillhandahållits att beskrivas.

3.1 Mixade media-modeller

Mixade media-modeller är en metod som används för att analysera sambanden mellan spendering på marknadsföring och konverteringar, modelleringen har funnits i olika former sedan 60-talet (Borden 1964). Syftet med metoden är framförallt att förstå effektiviteten av respektive marknadsföringskanal för att optimera spenderingsbudgeten (Chan 2017). Modellerna skapas genom statistiska regressionsmodeller för att analysera hur olika marknadsföringsaktiviteter påverkar konverteringarna över tid.

Generellt använder sig mixade mediamodeller av historisk tidsseriedata för konverteringar för ett företag och dess spendering på de olika marknadsföringskanalerna för att modellera sambanden över tid. Oftast så skapas modellerna på observationsdata vilket innebär att sambandsresultaten från modellerna anses som korrelationer snarare än kausala, vilket egentligen är målet med modelleringen. Metoden ger då en insikt på hur spenderingen påverkar konverteringar, men fastställda slutsatser är svårt dra då många andra bakomliggande faktorer som inte är med i modellen kan påverka resultatet.

Vanligt är att från en färdig modell använda prediktioner för att beräkna avkastningen av spendering (ROAS) per marknadsföringskanal för att ge insikt när avkastningen är som störst på investeringen.

3.2 Icke-linjäritet

Problemet som ska modelleras är ett icke-linjärt samband, anledningen för detta är att en ökning av budgeten inom respektive marknadskanal inte nödvändigtvis medför en enhetlig ökning i responsvariabeln, samt att det finns en så kallad avtagande avkastning (dimnishing returns) vad gäller spenderingen på kanalerna. Med avtagande avkastning menas att man inte kan förvänta sig att en predicerad ökning ur modellen håller sig konstant oavsett spenderat belopp, utan efter en viss gräns börjar avkastningen avta.

För att handskas med detta i linjära modeller tillämpas olika former av transformering av variablerna vilka nämndes i Bakgrundskapitlet, särskilt Hill-funktionen som utvecklades med grund i biokemin, samt enklare transformationer som sigmoid funktionen. Dessa transformationer kräver att man manuellt undersöker och definierar vikter för hur reaktionen på marknaden är i förhållande till mediaspenderingen vilket bidrar till att det kräver viss domänkunskap och öppnar för möjligheten att introducera bias eller dylikt genom mänsklig oaktsamhet.

Ett sätt att komma runt kravet på att transformera data och dessutom utveckla modeller som är väldigt kraftfulla är att använda sig av icke-linjära modeller.

Icke-linjära modeller anpassar funktionen av de förklarande variablerna mot beroende variabeln så väl som möjligt, snarare än att som för linjära modeller där antagandet om linjäritet ofta begränsar modellens prediktiva förmåga (G. James et al. 2013). Med anledning av detta kräver icke-linjära modeller fler observationer än linjära modeller, och ökar möjligheten att överanpassa sin modell efter de ofta precisa hyperparametrara en icke-linjär modell använder sig av i träningsfasen (G. James et al. 2013).

3.2.1 Hyperparametrar

Hyperparametrar har en omfattande tillämpning inom maskininlärning och har i denna uppsats utnyttjats för att optimera prestanda och reglering av XGBoost-modellerna. De representeras ofta i formler genom sym-

boler såsom α, θ, γ och optimeras numeriskt likt vikter som bestämmer graden av påverkan hos respektive hyperparameter i modellen (tune, tuning) (Brownlee 2019).

3.2.2 Korsvalidering

Korsvalidering är en metod för att utvärdera hur en metod presterar när man inte har ett särskilt stort distinkt datamaterial att dela upp i tränings- och valideringsdata. Det går ut på att man delar upp alla observationer i ungefärligt lika stora k antal grupper (folds) (G. James et al. 2013). Av de k grupperna blir en grupp vad man testar modellen mot, och $k-1$ är de folds man tränar modellen på. Utvärderingsmått beräknas på den ensamma gruppen, sedan itereras utvärderingen så en ny fold används som valideringsmängd varje gång (G. James et al. 2013). Formeln lyder,

$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^{k} MAE_i \tag{5}$$

Där k är antalet folds och ett godtyckligt utvärderingsmått, som till exempel MAE, beräknas för varje i:te fold.

Hyperparametrar som används för att optimera modeller till det bättre behöver ofta korsvalideras för att hitta ett värde som genererar en bra modell. Denna metod tillämpas så att en grid av värden för hyperparametern och en modell per värde i griden skapas för att sedan beräkna felet genom korsvalidering. Det värde på hyperparametern som genererar lägst fel väljs ut som det bästa värdet och används i det slutgiltiga modellskapandet (G. James et al. 2013).

3.3 Regularisering

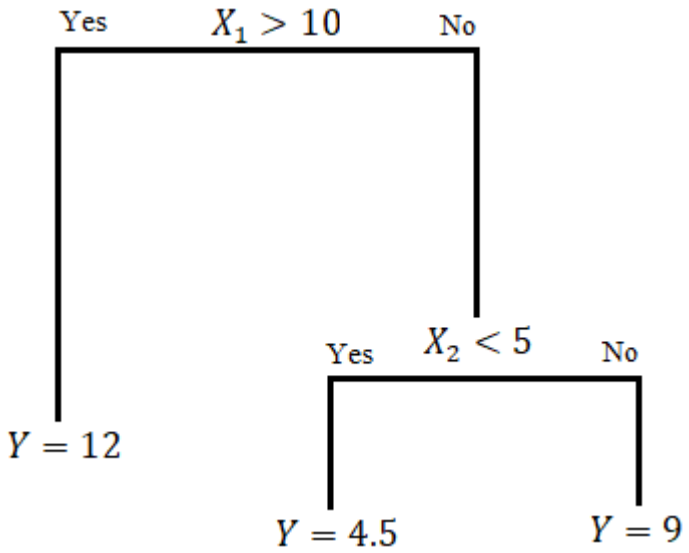
Regularisering används för att minska överanpassning på det datamaterial som modellen tränats på, ett önskat resultat av denna minskning av överanpassning är att prediktionernas pricksäkerhet på ny data ska förbättras (D. och H. James Gareth och Witten 2013).

3.3.1 Träning och valideringsmängd

Genom att dela upp datamaterialet i 2 delar, träning och valideringsmängd, så kan modellen tränas på en del av datamaterialet och sedan valideras genom att testa prediktionerna mot valideringsdata för att utvärdera vilken modell som är bäst. Ett utvalt utvärderingsmått likt MAE kan användas för att se vilken modell som presterar bäst på valideringsmängden och denna modell är potentiellt den slutgiltiga, givet att andra delar av analysen håller. Det är viktigt att inte ändra modellerna efter att de jämförts mot valideringsmängden för att förbättra precisionen, för då försvinner konceptet med uppdelningen av tränings- och valideringsmängd. En ändring efter validering för att sedan valideras igen blir att specifikt försöka anpassa modellen mot valideringsmängden och den önskade regulariseringseffekten försämras (D. och H. James Gareth och Witten 2013).

3.4 Beslutsträd

Beslutsträd är en väldigt vanligt förekommande icke-parametrisk modell som kan användas för både regression och klassificering. Metoden bygger på att man segmenterar värden på de variabler man vill använda för prediktion och sedermera skapar ett booleanskt alternativ där antingen det segmenterade värdet är uppfyllt eller inte. I nedåtgående led skapas då en "gren" som indikerar att antingen är värdet uppfyllt eller inte, och sedan kan denna gren utvecklas med ytterligare en segmenterad variabel som kan gälla eller ej gälla, så skapar man



Figur 14: Exempel på hur ett träddiagram kan se ut. Här är första regionella avgränsningen huruvida X_1 överstiger 10 eller inte. För $X_1 > 10$ har Y ett medelvärde på 12. För $X_1 \leq 10$ skapas också regioner för skillnader i X_2 som producerar olika resultat för Y beroende på dess värde.

en gren utifrån det s.k lövet också och fortsätter till man har en prediktion utifrån de interagerande variablerna man har tagit hänsyn till (G. James et al. 2013).

Att skapa ett regressionsträd innefattar två steg:

Algoritm för att skapa regressionsträd (G. James et al. 2013):

1. Dela upp de förklarande variablernas $X_1, X_2, ..., X_p$ i J olika distinkta området $R_1, R_2, ..., R_J$. Dessa områden R är strikt skilda från varandra.

2. Respektive observation av X_p som faller inom en viss region R_j erhåller samma prediktion, vilket är medelvärdet för responsvariabeln för den regionen.

Målet med skapande av dessa regioner R_J är att minimera RSS (residualkvadratsumman). Detta görs genom en metod som kallas recursive binary splitting (G. James et al. 2013). Det är en metod som börjar med en rotnod (toppen på trädet), delar datamängden i två delar vid en tröskel för den variabel som minskar residualsumman mest av alla variabler och sedan fortsätter dela respektive nod binärt med avseende att minska residualsumman så mycket som möjligt för varje steg. Att alltid göra splits efter att uppifrån maximera informationsvinsten utan att ta hänsyn med hur det påverkar noder längre ned i trädet anses vara en så kallad girig (greedy) strategi. Denna process som kan anses omfattande har en tendens att överanpassa data, just för att den egentligen kan iterera hur länge som helst. Detta problem löser man genom stoppkriterier för att begränsa modellen, ofta används "tree pruning", vilket innebär att man begränsar trädet genom att sätta en gräns för hur mycket bättre RSS måste bli vid nästa nod, om det inte förbättras tillräckligt så stannar processen. Detta resulterar i

mindre träd som ger lägre varians och bättre möjligheter för tolkning, mot att en bias introduceras då risk för att inte all information till handa används i prediktionen (G. James et al. 2013).

Optimering genom tree pruning (beskärning) sker genom att man skapar ett träd som är ett väldigt djupt träd T_0 och från det itererar man nedifrån och tar bort noder som inte presterar tillräckligt väl, vilket resulterar i ett antal delträd (sub-trees). En metod för att hantera detta och samtidigt slippa beräkna alla möjliga delträd är cost complexity pruning, där sekvenser av träd studeras indexerat efter en positiv hyperparameter α (G. James et al. 2013). Respektive värde på α tillhör ett delträd av T_0 så att följande formel minimeras. För att hitta ett bra värde på α så korsvalideras det för att hitta det värde som ger lägst valideringsfel.

$$\sum_{m=1}^{|T|} \sum_{i:} \left(y_i \in R_m - \hat{y}_{R_m} \right)^2 + \alpha |T| \tag{6}$$

Detta är RSS adderat med α-parametern, R_m är respektive region för en nod och $|T|$ som indikerar antal noder underträdet har. Parametern balanserar trädets bias mot varians (G. James et al. 2013).

3.4.1 Ensemblemodellering, Boosting

Ensemblemodeller används inom maskininlärning och syftar förbättra prestandan hos individuella modeller, i detta fall trädmodeller, genom att kombinera dem. Det finns olika varianter av ensemblemodeller, men i denna studie kommer "boosting" användas. (G. James et al. 2013).

Boosting är en algoritmisk metod för optimering av beslutsträd, och principen är att man skapar ett stort antal beslutsträd, $\hat{f}^1, ..., \hat{f}^B$. Algoritmen itererar över detta stora antal träd och minskar sekventiellt felmarginalen genom att för varje träd anpassa ett nytt träningsträd på ursrungliga trädets nod-residualer. Detta nya träd inkluderas sedan i modellen för att långsamt förbättra $\hat{f}$ stegvis (G. James et al. 2013).

Användning av boosting grundar sig i optimering av tre parametrar och nedan visas algoritmiskt en variant av boosting:

Algoritm för boosting av regressionsträd (G. James et al. 2013):

1. Sätt $\hat{f}_x = 0$ och $r_i = y_i$ för alla observationer i träningsdata.

2. För $b = 1, 2, ..., B$, iterera:

- (a) Anpassa ett träd $\hat{f}^b$ med d splits (dvs $d + 1$ avslutande löv) efter träningsdata (X, r).

- (b) Uppdatera $\hat{f}$ genom att addera en kortare version av det nya trädet

$$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x) \tag{7}$$

- (c) Uppdatera residualerna

$$r_i \leftarrow r_i - \lambda \hat{f}^b(x_i) \tag{8}$$

3. Visa den nya boostade modellen,

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x) \tag{9}$$

3.4.2 Gradient Boosting & XGBoost

Ett effektivt sätt att optimera algoritmen ovan är genom gradient boosting, vilket är en utveckling på boosting. De delar konceptet att svagare beslutsträd adderas ihop för att skapa en förbättrad ensemblemodell. Skillnaden är att gradient boosting använder sig av, som tyds ur namnet, en förlustfunktion och dess gradient - som kan liknas vid en multivariat derivata - i optimeringsfunktionen för att bestämma riktning på de residualbaserade träningsträden som läggs till i modellen. Modellen uppdateras iterativt för att minimera förlustfunktionen, som i detta fall för regression är det genomsnittliga absoluta felet (MAE). Gradient boosting innebär att man beräknar förlustfunktionens gradient och anpassar parametrarna efter störst negativ ökning i denna. Detta repeteras tills algoritmen konvergerat i ett godtyckligt minimum hos förlustfunktionen. Detta resulterar i en kraftfull modell som kan navigera stora mängder träd relativt effektivt, men det kräver optimering av hyperparametrar, mer datorkraft samt varsam regularisering för att inte överanpassa (Hastie, Tibshirani, and Friedman 2001).

eXtreme Gradient Boosting (XGBoost) är den metod som används för att skapa den icke-parametriska modellen för detta arbete. Algoritmen utvecklades av Tianqi Chen och presenterades i hans artikel "XGBoost: A Scalable Tree Boosting System" från 2016. Det främsta syftet med algoritmen var att skapa en metod för att effektivt använda modern datorkraft samt kunna producera godtyckligt bra modeller. XGBoost är en väldigt komplett metod och utöver att vara väldigt effektiv så har den möjlighet att tillämpa L2-regularisering för variabelselektion (T. Chen and Guestrin 2016). L2-regularisering fungerar i XGBoost på så sätt att modellens vikter straffas mot noll för att minska överanpassning och komplexitet. Ett större värde på L2-parametern ger regulariseringen ett större inflytande och modellen blir mer konservativ.

Resultatet från ett gradient boostat träd noteras som

$$\hat{y}_i = \sum_{k=1}^{K} f_k(\mathbf{x}_i), \quad f_k \in \mathcal{F} \tag{10}$$

Där, K är antalet adderade träd i modellen

$\mathcal{F} = \left\{ f(\mathrm{x}) = w_{q(\mathbf{x})} \right\} (q : \mathbb{R}^m \to T, w \in \mathbb{R}^T)$ är rummet med regressionsträd

w_i representerar den kontinuerliga "score" som erhålls för var i:de löv

q noterar varje träds unika struktur ner till avslutande noder

T är antalet löv i trädet

w noterar vikter för varje löv

Varje f_k representerar ett individuellt träd

Praktiskt beräknas ett träd q med unik struktur vars resultat klassas in som vikter w vilka sedan summeras ihop till en slutgiltig prediktion (T. Chen and Guestrin 2016). För att optimera de individuella träd f_k som används i modellen, minimeras följande funktion,

$$\mathcal{L}(\phi) = \sum_i l\left(\hat{y}_i, y_i\right) + \sum_k \Omega\left(f_k\right) \tag{11}$$

där den första delen, $\mathcal{L}(\phi) = \sum_i l\left(\hat{y}_i, y_i\right)$, representerar förlustfunktionen och andra delen, $\sum_k \Omega\left(f_k\right)$, är regulariseringstermen för funktionen.

Här är l en deriverbar konvex förlustfunktion för att mäta skillnaden mellan skattat värde och sanna värdet från datamaterialet, exempel på en sådan funktion är MAE vilken förklaras nedanför. Regulariseringen definieras enligt:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2$$

Där λ och γ är regulariseringsvikter för L1- samt L2-regularisering. Denna term används för att undvika överanpassning när de slutgiltiga vikterna w sätts (T. Chen and Guestrin 2016).

Tillsammans bildar den ovanstående funktionerna den slutgiltiga formeln som XGBoost följer vid modellering,

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t\left(\mathrm{x}_i\right)\right) + \Omega\left(f_t\right) \tag{12}$$

Funktionen optimeras genom att addera unika träd f_t iterativt för varje prediktion $\hat{y}_i^t$. Detta innebär att förlustfunktionen minimeras då respektive träd i boosting minimerar residualsumman.

3.4.2.1 Tillämpning av XGBoost för uppsatsens datamaterial

Förlustfunktion & Modellutvärdering Trots att responsvariabeln är av antal så hanteras den som kontinuerlig då en stor del av alla observationer är större än 15 och ingen är 0. Eftersom det högsta värdet på responsvariabeln sticker ut från resterande observationer så kommer en förlustfunktion behöva hantera detta så att det avvikande värdet inte ska påverka modellen negativt.

MAE Absolut medelfel, används ofta som utvärderingsmått på regressionsproblem, men är inte lika känslig mot extremvärden i samma utsträckning som andra vanliga mått, ett lågt värde innebär att modellen predikterar nära de sanna värdena. (Géron 2019).

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} \tag{13}$$

Generella hyperparametrar

- booster, väljer vilken typ av modell som ska byggas, till exempel träd eller linjär regression

Trädens hyperparametrar

- eta, steglängden för varje iteration, definierad mellan 0 och 1, avgör inlärningstakten.
- colsample_bytree, colsample_bylevel, colsample_bynode, dessa tre väljer ut andel av variabler för varje träd, nivå eller nod som ska användas vid trädskapandet.
- max_depth, maxdjupet på de skapade träden.
- n_rounds, antal träd som skapas.

Regulariserande hyperparametrar

- gamma, minsta informationsvinsten för att fortsätta dela upp ett träd i fler noder.
- lambda, Ridge regression på vikterna, högre värden gör modellen mer restriktiv.

Inlärningsparametrar

- eval_metric, valideringsmått som till exempel MAE.
- objective, förlustfunktionen som modellen ska lära sig att minimera.

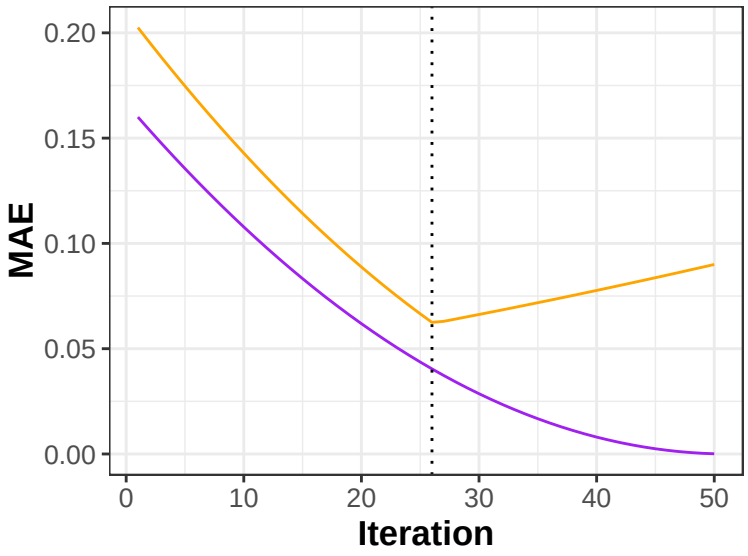
Randomized Search korsvalidering XGBoost har väldigt många hyperparametrar och att hitta den bästa modellen som skapas av en kombination av dessa är både tid- och beräkningskrävande. Ett sätt att korta ner tiden för beräkningarna och samtidigt lyckas hitta en bra modell är genom slumpmässig sökning med korsvalidering eller på engelska randomized search CV (Géron 2019).

Istället för att skapa ett rutnät (grid) och testa varje möjlig kombination av värden inom varje hyperparameters gridintervall så väljer metoden ut ett givet antal kombinationer genom att slumpmässigt dra ett nytt värde för varje hyperparameter inom ett bestämt intervall för varje iteration. Detta ger användaren en större kontroll över hur krävande beräkningarna ska vara då antalet iterationer väljs av användaren.

Korsvalideringen används för att testa vilken kombination av hyperparametrar som genererar den bästa modellen.

Early stopping Vid modellskapandet av XGBoost kan ett stort antal iterationer köras och modellen kommer tillslut oftast få en bra prediktionsförmåga på datamaterialet den tränats på. Om då modellen för varje iteration också testas på valideringsdatat och utvärderingsmått beräknas går det att se när modellen börjar att överanpassas. Early stopping kan då sättas in i modellskapandet, det detta gör är när utvärderingsmåttet för valideringsdatat inte längre sjunker kommer den att låta träningen fortgå i x antal iterationer för att se om en förbättring sker, om detta inte sker så avbryts träningen (Géron 2019).

Exemplet i figur 15 visar hur en överanpassning kan upptäckas genom att testa modellen för varje iteration i träningen, vid iteration 27 hade early stopping sett att en förbättring inte hade hänt och låtit modellen gått ett valt antal iterationer(kan vara 0) till och avbrutit modellträningen om inget förbättrat värde hade fåtts.



Figur 15: Exempel på överanpassning där early stopping hade avbrutit. Lila linjen är MAE för träningsmängd, orangea är för valideringsmängd.

3.5 Residualanalys

För att testa huruvida modellantaganden är uppfyllda analyseras modellens residualer genom att studera figurer av residualerna mot modellens skattade värden. Förlustfunktionen som används i uppsatsen antar att residualerna ska vara symmetriska kring medelvärdet, oberoende och ha en konstant varians. Figurer som studeras är residualernas fördelning för att se att symmetriantaganet är uppfyllt, residualerna plottade i observationsordning för att om möjligt se ett tidsberoende, en figur över skattad autokorrelationfunktion för att se om tidsberoende finns bland residualerna och skattade värden mot residualerna för att se om variansen är konstant. En modell kan inte anses trovärdig om antaganden inte är uppfyllda, vilket kan ge stora konsekvenser om den producerar koefficienter som inte är korrekt beräknade. Är något av antagandena inte uppfyllt är modellen formellt obrukbar.

3.5.1 Tidsberoende

Då det datamaterialet som behandlas är en tidsserie är det stor sannolikhet att antagandet om oberoende observationer inte kommer vara uppfyllt och att autokorrelation förekommer i residualerna, vilket innebär en korrelation i tiden (Bowerman 2004). Autokorrelation kan delas in i två typer, positiv och negativ autokorrelation. Positiv autokorrelation innebär att över tid så följs en positiv residual av en positiv residual eller att en negativ residual följs av en negativ residual. Negativ autokorrelation innebär att en positiv residual över tid följs av en negativ residual, eller att en negativ residual följs av positiv residual. Positiv autokorrelation medför ofta ett cykliskt mönster hos residualerna, medan negativ autokorrelation snarare följer ett skiftande mönster. Modellantagande gällande oberoende innebär att residualerna sorterade efter observationsordning inte följer något av dessa mönster utan faller slumpmässigt i tiden (Bowerman 2004). Autokorrelation förekommer ofta som en förstagradsautokorrelation, det vill säga att feltermen vid tidpunkt T är relaterad till feltermen vid tidpunkt T-1. Förstagradskorrelation är ett första steg i att analysera huruvida autokorrelation förekommer överhuvudtaget och därmed ett första steg att ta i modellering vid tidsserieproblem för att bekräfta om antagandet kring tidsoberoende inte håller.

I analys av tidsberoendet för att kunna dra slutsatser om vilka åtgärder som kan krävas för att behandla det används framför allt skattad autokorrelationsfunktion (förkortas SAC eller S-ACF). Funktionen mäter det linjära sambandet mellan observationer i en tidsserie för olika tidpunkter genom att man skapar så kallade laggade variabler. En laggad variabel är en variabel som har förskjutits ett visst antal tidsenheter (lags), vilken sedan kan jämföras med varandra för att avläsa hur korrelationen beter sig mellan olika lags (Bowerman 2004).

Formel för SAC:

$$r_k = \frac{\sum_{t=b}^{n-k}(z_t - \bar{z})(z_{t+k} - \bar{z})}{\sum_{t=b}^{n}(z_t - \bar{z})^2} \tag{14}$$

Där,

$z_b, z_{b+1}, ..., z_n$ är en stationär tidsserie (antingen ursprungliga värden eller exempelvis differentierade)

$$\bar{z} = \frac{\sum_{t=b}^{n} z_t}{(n - b + 1)} \tag{15}$$

n = antalet observationer

b = minsta antalet observationer som krävs för att beräkna SAC vid lag k

k = antalet lags

t = observation i tiden

r_k kommer alltid vara mellan -1 och 1 där ett högre värde visar på att observationer med som skiljer sig med ett lag på k visar en stark positiv tendens och vice versa för -1 fast stark negativ tendens (Bowerman 2004).

3.5.2 Additive feature attribution method

Maskininlärningsmodeller är ofta komplexa och svårtolkade, särskilt vid analys av hur mycket varje enskild variabel bidrar till resultatet. Ett vanligt sätt att hantera detta problem är att använda "additive feature attribution"-metoder, som används för att tilldela variabler värden efter vad de bidrar till modellens outputs. Om vi låter $z^{'} \in (0,1)^M$ vara en mängd av binära variabler som säger om $x^{'}$ är observerat eller ej kan metoden anses vara följande:

$$g(z^{'}) = \phi_0 \sum_{i=1} \phi_i z_i^{'} \tag{16}$$

Där M är antalet variabler och $\phi_i \in R$ och är tillskrivningsvärdet för respektive variabel. En viktig egenskap som krävs av en 'additive feature attribution'-metod är att det finns en unik lösning som eftersträvar tre egenskaper: lokal pricksäkerhet, saknad och följdriktighet (S. M. Lundberg and Lee 2017).

Lokal pricksäkerhet: Summan av variabel-tillskrivningen är densamma som outputen som den försöker beskriva.

Saknad: Saknade värden ska tilldelas 0 importance, alltså när $z_i^{'} = 0$.

Följdriktighet: En förändring av modellen så att en variabel får en större påverkan ska aldrig sänka den variabelns tilldelade importance.

Importance XGBoost beräknar importance scores av de variabler som används i modellen och de variabler som används mest vid viktiga beslut/splits i ensembleträden får högst score. Importance beräknas ofta med gain, vikt eller cover, men här studien kommer använda sig av genomsnittliga absoluta Shapley-värdet då det anses mer robust (S. M. Lundberg and Lee 2017).

3.5.3 Shapley

Shapley-värden härstammar ur spelteori och introducerades 1951 av Lloyd Shapley, som sedermera vann Nobelpris 2012 för upptäckten. Inom samarbetsspel används Shapley-värden för att rättvist fördela ut poäng eller pengar efter hur varje enskild spelare bidragit i spelet. Detta kan appliceras på XGBoost för att lättare förstå bidraget från varje variabel. Shapley-värden tolkas som det marginella bidraget från respektive variabel i modellens prediktioner för träningsmängden. Värdet beräknas enligt,

$$Shapley(X_j) = \phi_j = \sum_{S \subseteq N/j} \frac{k!(p-k-1)!}{p!}(f(S \cup j) - f(S)) \tag{17}$$

Där,

k är antalet permutaioner som kan skapas av variablerna i delmängd S.

p är totala antalet variabler

N/j är alla möjliga kombinationer av variabler uteslutande av X_j

S är en delmängd av variabler i N/j

$f(S)$ är modellens prediktioner med de variabler som är i S

$f(S \cup j)$ är modellens prediktion med de variabler i S plus X_j.

Exempel på beräkning av Shapley-värden Ett företag har testat olika kombinationer av 3 olika marknadsföringskanaler, A, B, och C, och vill testa hur mycket varje kanal bidrar med till försäljningen för en specifik månad. Detta resulterar i 8 olika kombinationer där varje kombination har en predikterad inkomst för månaden enligt följande:

Tabell 6: Exempel på inkomstprediktion för alla möjliga kombinationer av kanalerna A, B, C.

	Kombination	Inkomstprediktion
1	Ingen av A, B, C (tomma)	100
2	A	200
3	B	250
4	C	150
5	AB	350
6	BC	250
7	AC	300
8	ABC	400

För att beräkna Shapley-värdet för kanal A och således hur mycket kanalen bidrar med till försäljning tittar vi på de olika delarna i formel 17.

Totalt finns 4 kombinationer av kanaler som inte innehåller A, dvs “Ingen av A, B, C”, “B”, “C”, “BC”. Dessa används för att beräkna påverkan kanal A har genom att subtrahera deras inkomstprediktioner från vad respektive kombinations inkomstprediktion är när A är inkluderad, exempelvis AB-B och ABC-BC.

Då det är tre kanalvariabler som används i exemplet är $p = 3$ och således är nämnaren i formeln $p! = 3 \cdot 2 = 6$ genomgående.

Först ut används Shapley-formeln med den "tomma" kombinationen, dvs Shapley-värdet då ingen av kanalerna A, B, C används. Detta värde anses vara den predikterade inkomsten företaget kan förvänta sig utan användandet av någon kanal och således bidrar potentiellt andra faktorer utanför just marknadsföring till denna försäljning.

Shapley-värdet för den tomma kombinationen beräknas enligt:

Antalet möjliga permutationer $k = 0$ eftersom ingen variabel ingår.

Inkomstprediktionen för den aktuella tomma kombinationen $f(S) = 100$.

Inkomstprediktionen för den kombinationen A som undersöks $f(S \cup j) = 200$.

Sätt in dessa i formel 17,

$$\frac{k!(p-k-1)!}{p!}(f(S \cup j) - f(S)) = \frac{0!(3-0-1)!}{6}(200-100) = \frac{2}{6} \cdot 100 \tag{18}$$

Detta resultat används senare vid summering över N/j.

Shapley-värdet för kombination med B beräknas sedan på samma sätt:

Antalet permutationer $k = 1$ eftersom en variabel ingår.

Inkomstprediktionen för den aktuella kombinationen B är $f(S) = 250$.

Inkomstprediktionen för kombinationen innehållande A, dvs AB är $f(S \cup j) = 350$.

$$\frac{1!(3-1-1)!}{6}(350-250) = \frac{1}{6} \cdot 100 \tag{19}$$

Shapley-värdet för kombination med C:

Antalet permutationer återigen $k = 1$ eftersom en variabel ingår.

Inkomstprediktionen för den aktuella kombinationen C är $f(S) = 150$.

Inkomstprediktionen för kombinationen innehållande A, dvs AC är $f(S \cup j) = 300$.

$$\frac{1!(3-1-1)!}{6}(300-150) = \frac{1}{6} \cdot 150 \tag{20}$$

Shapley-värdet för kombination BC:

Antalet permutationer $k = 2$ eftersom BC även kan kombineras som CB.

Inkomstprediktionen för den aktuella kombinationen BC är $f(S) = 250$.

Inkomstprediktionen för kombinationen innehållande A, dvs ABC är $f(S \cup j) = 400$.

$$\frac{2!(3-2-1)!}{6}(400-250) = \frac{2}{6} \cdot 150 \tag{21}$$

Sedan summeras alla dessa beräkningar över N/j för att få Shapley-värdet för kanal A enligt:

$$\phi_A = \frac{2}{6} \cdot 100 + \frac{1}{6} \cdot 100 + \frac{1}{6} \cdot 150 + \frac{2}{6} \cdot 150 = 125 \tag{22}$$

Enligt Shapley-värdet bidrar marknadsföringskanal A med 125 enheter i den marginella förändringen för modellens prediktioner på inkomsten för den specifika månaden när A inkluderas i modellen. Det är viktigt att notera att Shapley-värden är baserade på modellens prediktioner och inte på den verkliga försäljningen, det ger en kvantitativ indikation på bidraget från varje variabel i modellen.

Tree-SHAP Ett problem med Shapley-värden är att det är beräkningstungt och därav har Tree-SHAP tagits fram som går mycket snabbare att beräkna än traditionella Shapleyvärden. Tree-SHAP är en algoritm som utnyttjar trädens struktur och sänker beräkningstiden från exponentiellt till polynom, exempelvis $T^M \Rightarrow T^2$, där T = tiden och M = antalet variabler i modellen.

Algoritmen använder repeterbarhet för att beräkna andelen av alla möjliga delmängder av variabler som flödar ner i varje löv i trädet istället för att beräkna alla delmängder. Algoritmen ökar storleken på delmängden när metoden rör sig nedåt i trädet och minskar storleken på delmängden när metoden rör sig uppåt i trädet. När metoden rör sig nedåt i trädet uppdateras andelarna för delmängden och på vägen upp beräknas SHAP-värdet. Med en ökad delmängd så ges en mer heltäckande bild över alla variabler som är relevanta medan den minskade delmängden fokuserar på de variabler som är viktigast genom minskningen av storleken. Genom att dynamiskt kunna ändra storleken på delmängderna beroende på riktningen som metoden rör sig i trädet så möjliggörs en effektivare beräkningsprocess och få en bättre bild över variablernas inverkan på resultatet.

För XGBoost så tas medelvärdet över alla träds SHAP-värde för varje variabel för att få ett sammanlagt värde för hela modellen.

SHAP interaktionsvärden Ett sätt att mäta hur variabler samverkar för att påverka modellens resultat är att beräkna SHAP-interaktioner och det kan beräknas genom en utökning av formel 17 enligt:

$$\Phi_{i,j} = \sum_{S \subseteq N \setminus \{i,j\}} \frac{k!(p-k-2)!}{2(p-1)!} \nabla_{ij}(S) \tag{23}$$

Där $i \neq j$ och $\nabla_{ij}(S)$ beräknas enligt:

$$\nabla_{ij}(S) = f(S \cup \{i,j\}) - f(S \cup \{i\}) - f(S \cup \{j\}) + f(S) \tag{24}$$

Formel 23 delar annars notationer med formel 17.

Interaktionsvärdet delas upp så den blir lika för de variabler (i och j) i interaktionen, den totala interaktionen blir då följande:

$$\phi_{i,j} + \phi_{j,i} \tag{25}$$

En interaktion kan tolkas som skillnaden i SHAP-värdet för variabel i när variabel j är närvarande och när den är frånvarande i en modell.

Effekten för en prediktion kan då beräknas genom att ta SHAP-värdet för en variabel och subtrahera med summan över dess interaktionsvärden.

$$\Phi_{i,i} = \phi_i - \sum_{j \neq i} \Phi_{i,j} \tag{26}$$

Algoritmen för tree SHAP kan appliceras på SHAP-interaktioner för att reducera beräkningstiden(S. M. Lundberg and Lee 2017).

SHAP-värden kommer att användas för att se vad varje spenderingsvariabel bidrar med till prediktionerna på antal nya konton, vilket naivt kan tolkas som ROAS vilket nämndes i bakgrunden. Beroende-diagram (dependence plots) kommer att användas för att se de marginella effekterna på prediktioner på träningsdata när respektive variabel inkluderas i modellen.

Beeswarm-diagram För att se en variabels SHAP-värdefördelning för olika nivåer av spendering kan beeswarm-diagram användas (S. Lundberg 2018). Det fungerar så att variablernas SHAP-värde ger en position på x-axeln för varje observation, färgen på punkterna beror på variabelns individuella värdefördelning där minimum och maximum har olika färger och värden mellan dem blir en färgblandning. Om punkterna för en variabel påvisar en svärm utmed y-axeln betyder det att antalet observationer runt det SHAP-värdet är fler jämfört med andra SHAP-värden för den variabeln. Från diagrammet går det alltså att se vilket SHAP-värde en stor del av observationerna för en variabel får och samtidigt se om SHAP-värdet skiljer sig beroende nivån av spendering för variabeln.

3.5.3.1 LOWESS-kurvor

LOWESS (Locally Estimated Weighted Scatterplot Smoothing) är en form av lokal (local) regression. Local regression är en metod för att anpassa icke-linjära funktioner och innebär att man beräknar flera olika regressioner för ett datamaterial med endast en delmängd av träningsobservationerna inkluderade för varje regression vilka repeteras för hela datarummet (G. James et al. 2013). Detta innebär att man kan producera regressionslinjer som båda är negativa och positiva snarare än entydigt bestämda. Med hjälp av LOWESS kan dessa regressionslinjer kopplas ihop och "smoothas" till en kurva. LOWESS-kurvor används i uppsatsen för att visualisera sambanden mellan kanalspendering och Shapley-värden. För de kanaler som har få observationer kommer LOWESS-kurvan ej kunna produceras då den kräver ett visst antal observationer per delmängd vilket inte håller för de specifika kanalerna.

3.6 Programvaror

All programmering, från datahantering till visualisering av modellernas resultat är skrivet i programmeringsspråket R. Olika paket har använts där de viktigaste för datahanteringen är 'dplyr' (Wickham et al. 2023) och 'lubridate' (Grolemund and Wickham 2011), visualisering av data är framförallt gjort med 'ggplot2' (Wickham 2016), för modelleringen används 'MLR' (Bischl et al. 2016) för tuneing av hyperparametrar och 'xgboost' (H. Chen T. 2022) för att skapa slutgiltiga modellen och till sist används 'shapviz' (Mayer 2023) för figurer med SHAP-värden.

4 Resultat

4.1 Modellskapande

Det första steget som gjordes i modellskapandet var att dela upp datamaterialet i tränings- och valideringsmängd, de två sista månaderna av datamaterialet (december 2022, januari 2023) valdes till valideringsmängd och all data från 7 juni 2021 till december 2022 används då som träningsdata. Det blir alltså en uppdelning på cirka 90/10 i träning/validering, eftersom spenderingen på vissa kanaler började efter sommaren 2022 och för att modellen ska kunna se effekten av dem så behövs så många observationer av dem som möjligt i träningsmängden. Nästa steg var att testa XGBoost-modeller för våra datamaterial för att sedan utföra en residualanalys där det undersöks om förlustfunktionens antaganen är uppfyllda och analysen visade följande:

Residualerna var ej normalfördelade utan positivt skeva, variansen var inte konstant utan ökade när amplituden ökade och ett tydligt tidsberoendesyndes.

Dessa problem med modelleringen var först och främst på grund av två dagar då antal nya konton som skapades var extremt höga med en väldigt markant ökning i magnitud sett emot alla andra dagar i datamaterialet trots att alla tillgängliga variabler är kring nivåer man kan förvänta sig. Vad som ytterligare försvårar problemet är att det är just under dessa två dagar som Google har saknade data vilket imputerades för i avsnitt 2.2. Det medför problem inte bara för modelleringen utan även för tolkning av samband och för imputeringen att extremvärden för responsvariabeln sker samtidigt som en av de största kanalerna har saknade data. Med grund i detta beslutades att exkludera dagarna ur analysen då ökningen kan bero på antingen Google-spendering eller yttre faktorer som ej är relevanta för modelleringen i denna uppsatsen. Dummyvariabler för veckodagar och månader samt en endags-lag på antal nya konton läggs in i modellen för att reducera tidsberoendet.

4.1.1 Randomized search korsvalidering

Innan modellerna skapas måste hyperparametrarna sättas eller tilldelas intervall där värden kommer testas för att hitta bra modeller för datamaterialen och varje enskilt datamaterial kommer att tilldelas hyperparametrar som korsvaliderats fram som den bästa för dem. Efter att manuellt testat vilka hyperparametrar som påverkar modellerna mest och inom vilka ramar som är rimliga så väljs följande intervall eller värden till parametrarna i tabell 6. Efter att varje modell har fått en hyperparameterkombination som anses vara bäst av alla kombinationer som testats så plockas de ut för att köra om modellen med earlystopping för att motverka eventuell överanpassning.

Tabell 7: Intervall för randomized search CV

Parameter	Intervall/värde
booster	gbtree
eta	[0.1-0.4]
colsample_bytree	[0.5-1]
max_depth	[3-10]
n_rounds	2000
gamma	[0.5-1]
lambda	[0-10]
eval_metric	MAE
objective	Absolute error

Den slumpmässiga sökningen körs i 1000 iterationer för varje respektive datamaterial, vilket ger 1000 slumpmässiga kombinationer av värden från intervallen på alla hyperparametrar. Värdena på intervallen är satt så att

kombinationer kan regulariserar till viss del, men samtidigt tillåta kombinationer som ger komplexa träd med högt maxdjup. Korsvalidering sker för utvärdering av varje iteration så att kombinationer där regulariseringen inte är lika stark och mycket överanpassning sker troligtvis kommer att resultera i att en mer regulariserad modell får lägre korsvalideringsfel och anses vara bättre. Alla intervall utom maxdjupet är flyttal och kan anta alla värden inom intervallen, maxdjupet kan endast anta heltal.

4.1.2 Bästa hyperparametrarna

Att studera vilka hyperparametrar som ger den bästa modellen enligt den slumpmässiga sökningen ger en bättre förståelse för hur parametrarna regulariserar modellerna samt om intervallen är rimliga eller om en stor del av alla värden ligger nära ett gränsvärde.

Tabell 8: Hyperparametrar som validerats efter MAE

	Modell	eta	gamma	max_depth	colsample_bytree	lambda
1	Base	0.12	1.90	4	0.53	8.81
2	Kumsum 1v	0.13	1.38	3	0.65	8.37
3	Kumsum 2v	0.18	2.94	4	0.58	7.03
4	Kumsum 3v	0.12	1.75	4	0.50	8.72
5	Kumsum 4v	0.22	2.69	7	0.78	9.58
6	Decay 0.5	0.35	4.74	10	0.57	9.86
7	Decay 0.4	0.14	3.76	8	0.62	8.77
8	Decay 0.3	0.12	2.51	4	0.50	9.35
9	Decay 0.2	0.15	2.90	5	0.68	8.36
10	Komplett	0.13	3.96	8	0.52	9.41
11	Andelar	0.37	3.66	4	0.63	9.19

Den hyperparameterkombination som får lägst MAE för varje modell presenteras i tabell 7. För modellerna så är lambda och colsample_bytree de hyperparametrarna som är mest lika för alla modeller, detta tyder på att dem viktiga parametrar för att regularisera modellen. Gamma och maxdjupet skiljer sig mellan modellerna men det går att se ett mönster över hur dem kombineras, ett lägre gamma är ofta följt av ett lågt maxdjup och ett högre gamma är följt av ett större maxdjup. Detta tyder på att maxdjupet och gamma kan kompensera för varandra i den mån att när minsta informationsvinsten för en split är hög så spelar det inte lika står roll att maxdjupet är stort då de skapade träden troligtvis inte blir så djupa.

4.1.3 Utvärdering av modeller

Första steget för att utvärdera modellerna är att studera utvärderingsmåttet MAE, där ett lågt värde tyder på en bättre modell.

Tabell 9: Utvärderingsmått för modellerna. Kvoten beräknas som validerings-MAE dividerat med tränings-MAE.

	Modell	Träning	Validering	Kvot
1	Base	1.98	8.46	4.27
2	Kumsum 1v	2.58	8.98	3.48
3	Kumsum 2v	2.99	10.09	3.37
4	Kumsum 3v	2.17	11.81	5.45
5	Kumsum 4v	0.22	11.08	50.60
6	Decay 0.5	2.08	9.93	4.79
7	Decay 0.4	1.45	9.32	6.41
8	Decay 0.3	2.37	9.71	4.10
9	Decay 0.2	1.14	10.07	8.86
10	Komplett	0.33	9.59	28.71
11	Andelar	4.02	9.43	2.35

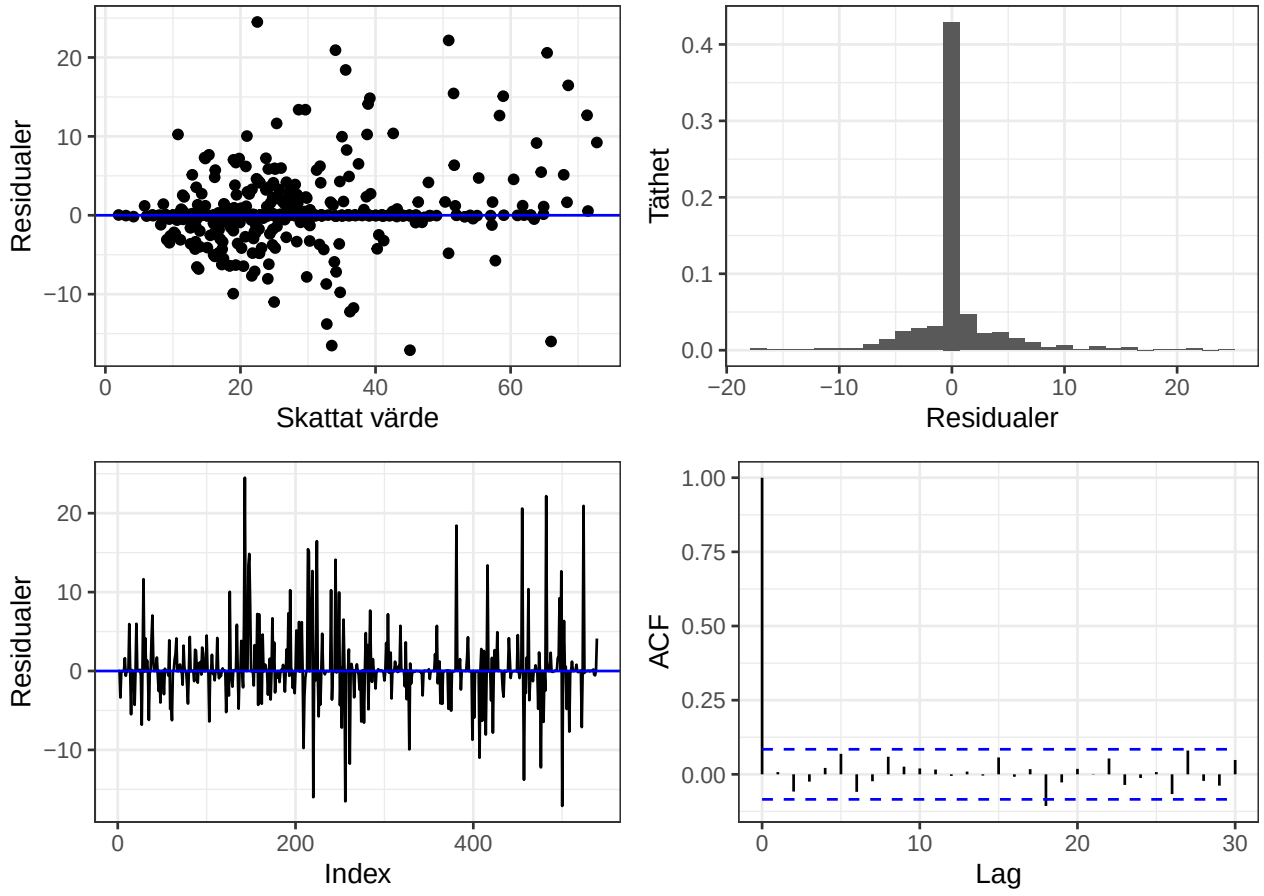
Modeller för de bästa hyperparameterkombinationerna har skapats och i tabell 8 så presenteras utvärderingsmåtten för respektive modell. Modellerna med lägst validerings-MAE är base-modellen utan transformering, kumulativ summa på 1 vecka och sönderfallshastighet på 0.4. Kvoten visar att vissa modeller har överanpassats trots att regulariserande hyperparametrarna används vid modellskapandet. Den modellen som är minst överanpassad är andelsmodellen, detta syns då kvoten på MAE mellan träning- och valideringsmängden är minst av alla modeller.

Numreringen på raderna från tabell 8 kommer att användas för att hänvisa till modellerna i senare delkapitel, där basemodellen med radnummer 1 refereras som modell 1.

4.2 Modellanalys

För alla skapade modeller har residualerna från träningsmängden analyserats för att se vilka som uppfyller antaganden för modellen, resultatet som presenteras i detta delkapitel är de modeller som har lägst MAE för respektive transformation och samtidigt någorlunda uppfyller antaganden för MAE.

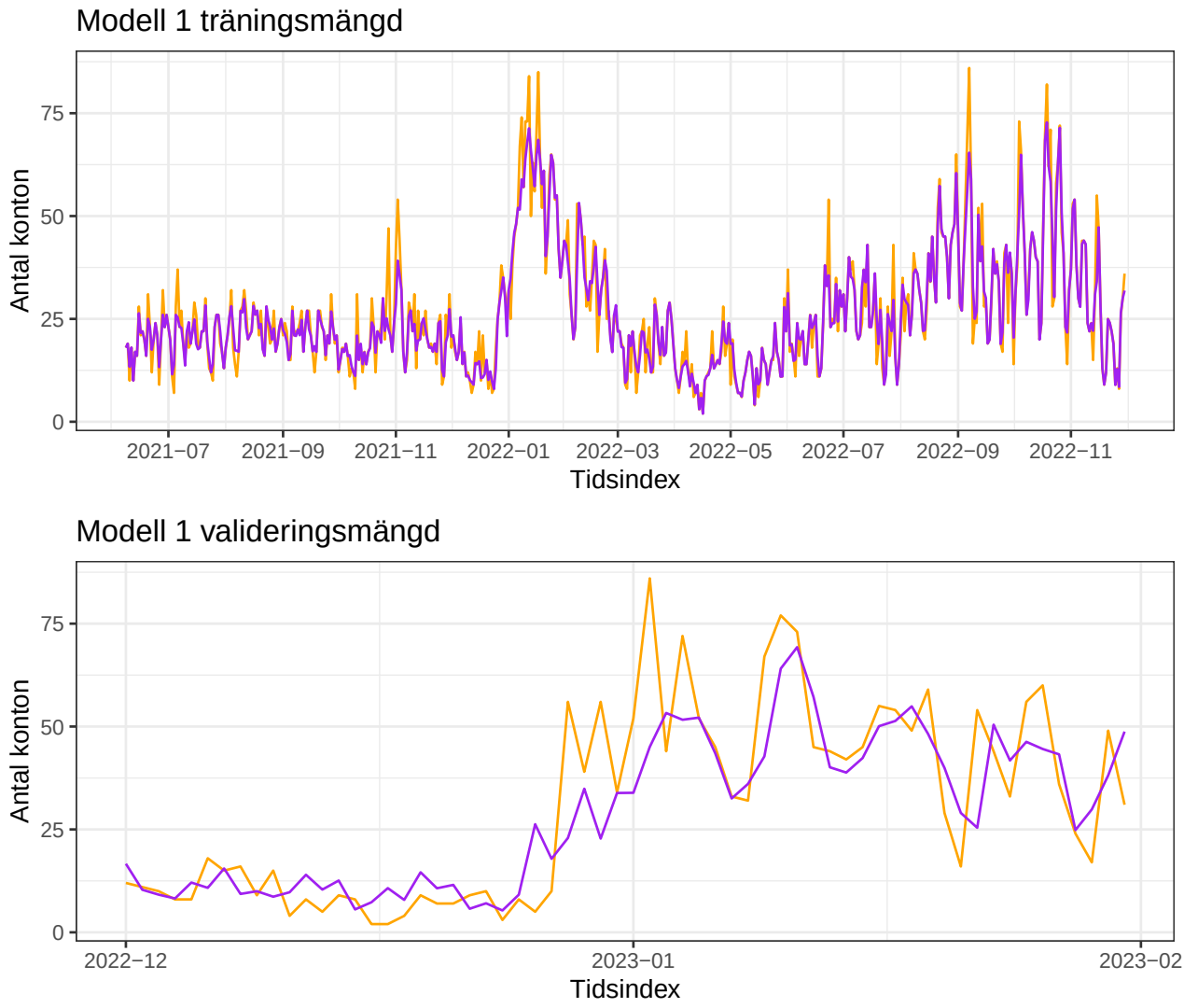
4.2.1 Modell 1



Figur 16: Residualanalys Modell 1 MAE

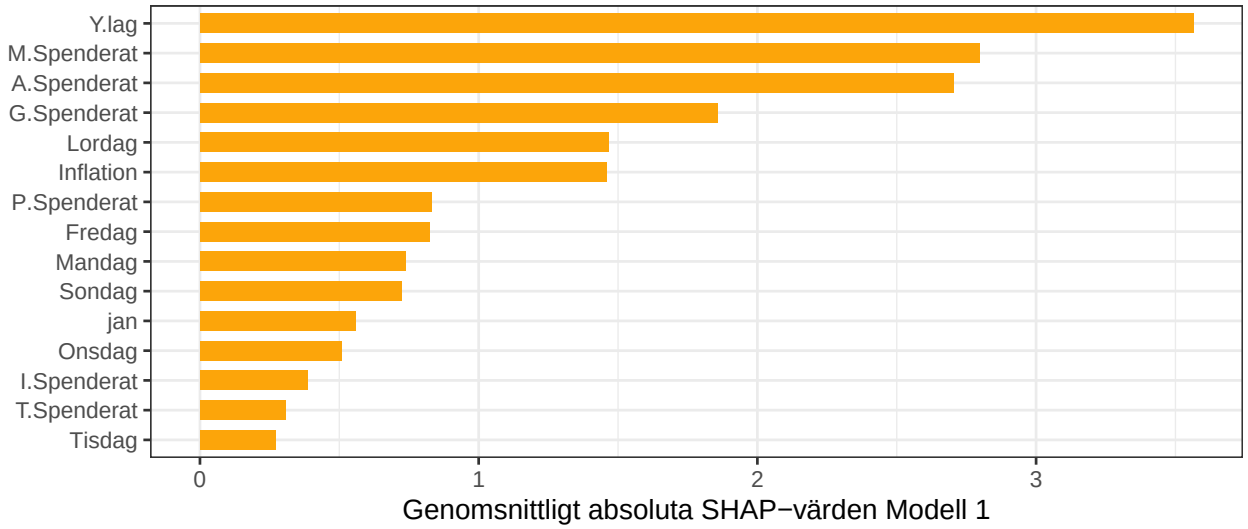
Residualanalysen ovan används för att bekräfta modellens antaganden och således avgöra att problemet modellerats korrekt för. Det som gäller för MAE som förlustfunktion är symmetriska residualer, vilket tyds enklast genom att titta på histogrammet i figur 16, fördelningen är symmetrisk kring medelvärdet då enskilda värden vid högre varians syns åt båda hållen. Stor majoritet av modellens skattade värden har ett residualvärde på 0, med de högsta residualerna kring |20|. Konstant varians är viktigt att man kan anse uppfyllt så modellen skattar robust för alla magnituder av responsvariabeln, vilket kan tydas ur det övre punktdiagrammet till vänster. En någorlunda ”tratt”-form går att tyda vilket indikerar en icke-konstant varians där den ökar för högre anpassade värden, däremot går det att argumentera för att variansen ändå är någorlunda konstant då de är relativt jämna från 0-30 och 30-70 utmed x-axeln.

För att analysera huruvida residualerna kan anses vara tidsberoende studeras dels indexfiguren i det nedre vänstra hörnet men också autokorrelationsfiguren till höger, som visar genomsnittlig korrelation mellan respektive residual och dess omkringliggande värden vilket representerar lags i tiden. Ur indexfiguren i residualanalysen syns hur residualerna är fördelade i tiden, och då man inte vill ha ett beroende så gäller antagandet att residualerna till synes ska vara slumpmässiga och inte påvisa några särskilda mönster. Detta kan anses uppfyllt då det fluktuerar mycket mellan både negativt och positivt värde samt genomgående höga och låga magnituder. Figuren över autokorrelationen visar allmänt väldigt låga ”spikar”. Man tittar efter periodicitet eller särskilda mönster vilket skulle kunna förklara att residualerna är beroende varandra, samt efter höga spikar. För modellen i fråga syns att inga uppenbara mönster i spikarnas position går att tydas, samt endast en korsar signifikanslinjen som är streckad i blått. Signifikanslinjen anger en gräns för vad som kan anses vara vitt brus i tidsserien och ligger vid |0.1| eller 10 procent korrelation. Med grund i dessa figurer kan antaganden om oberoende i tiden vara uppfyllt.



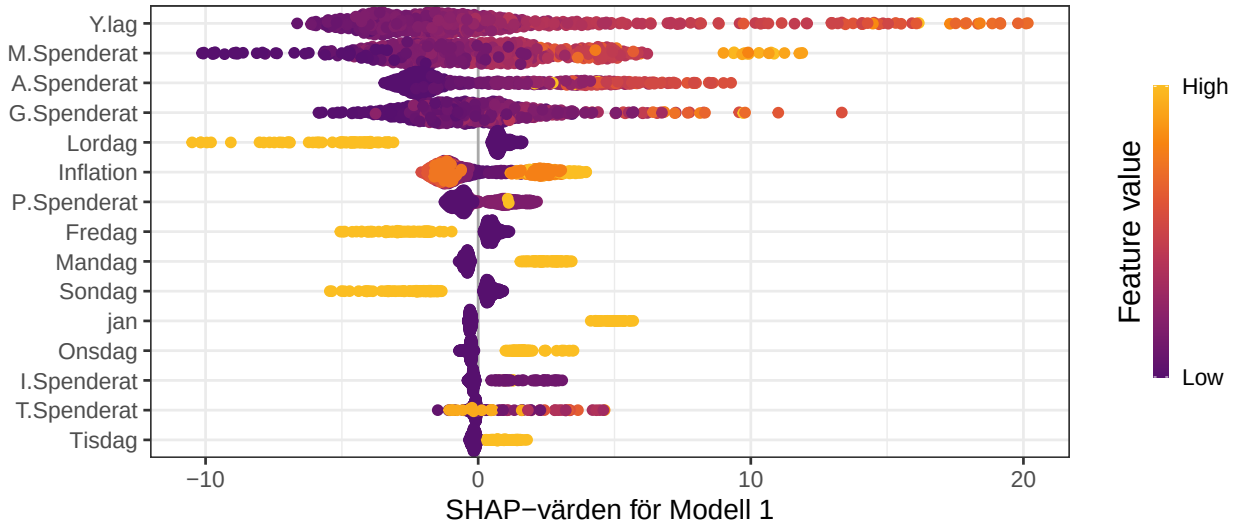
Figur 17: Linjediagram över sanna värden samt prediktioner på tränings- och valideringsdata för modell 1, orange är sanna värden på antal nya konton och lila är modellens prediktioner

I figur 17 syns sanna antal konton i orange och skattade värden i lila för tränings- och valideringsdata. I träningsdata går att tyda att modellen är ganska restriktiv och inte plockar upp de högsta magnituderna av responsvariabeln, det verkar vara så att modellens variabler inte kan förklara de höga topparna. I båda figurerna syns det att modellen följer generella mönster i responsvariabel, men att modellen har problem att fånga upp den plötsliga ökningen av antal nya konton på valideringsmängden.



Figur 18: Stapeldiagram över genomsnittligt absoluta SHAP-värden för variablerna i modell 1.

Figur 18 visar genomsnittligt absoluta SHAP-värden, vilket är ett mått för att utvärdera vilka variabler som bidrar mest till modellens prediktioner. Ett högre värde indikerar att det är en viktigare variabel för modellen. Då trädmodeller är icke-parametriska är det ett sätt att attribuera ett värde för genomsnittliga effekten på responsvariabeln respektive förklarande variabel bidrar med, motsvarande en parameter i linjär regression. Enligt figuren är ”Y.lag” den variabeln som bidrar mest till förklaring i responsvariabeln, däremot är detta Y-variabeln laggad en dag och används endast för att behandla tidsberoendet som uppstått annars och således inte lika intressant för vidare analys som gör sig synes. Det mest intressanta är egentligen att titta på de olika kanalernas variabler, även om externa variabler som inflation eller veckodagar bidrar till förklaring så är de just externa och endast inkluderade i syfte att hjälpa modellen hitta mönster med hjälp av dessa. Högst genomsnittligt absoluta SHAP-värde har generellt Meta och Affiliate, vilket alltså innebär att de har en högre inverkan på skattningen av Y än andra spenderingskanaler, däremot visar inte figuren detaljerna kring påverkan som åt vilket håll det påverkas. Google är den tredje viktigaste kanalen i påverkan på responsvariabeln, sedan följer Programmatic, Tiktok och Influencer som kanaler med lägst SHAP-värden.



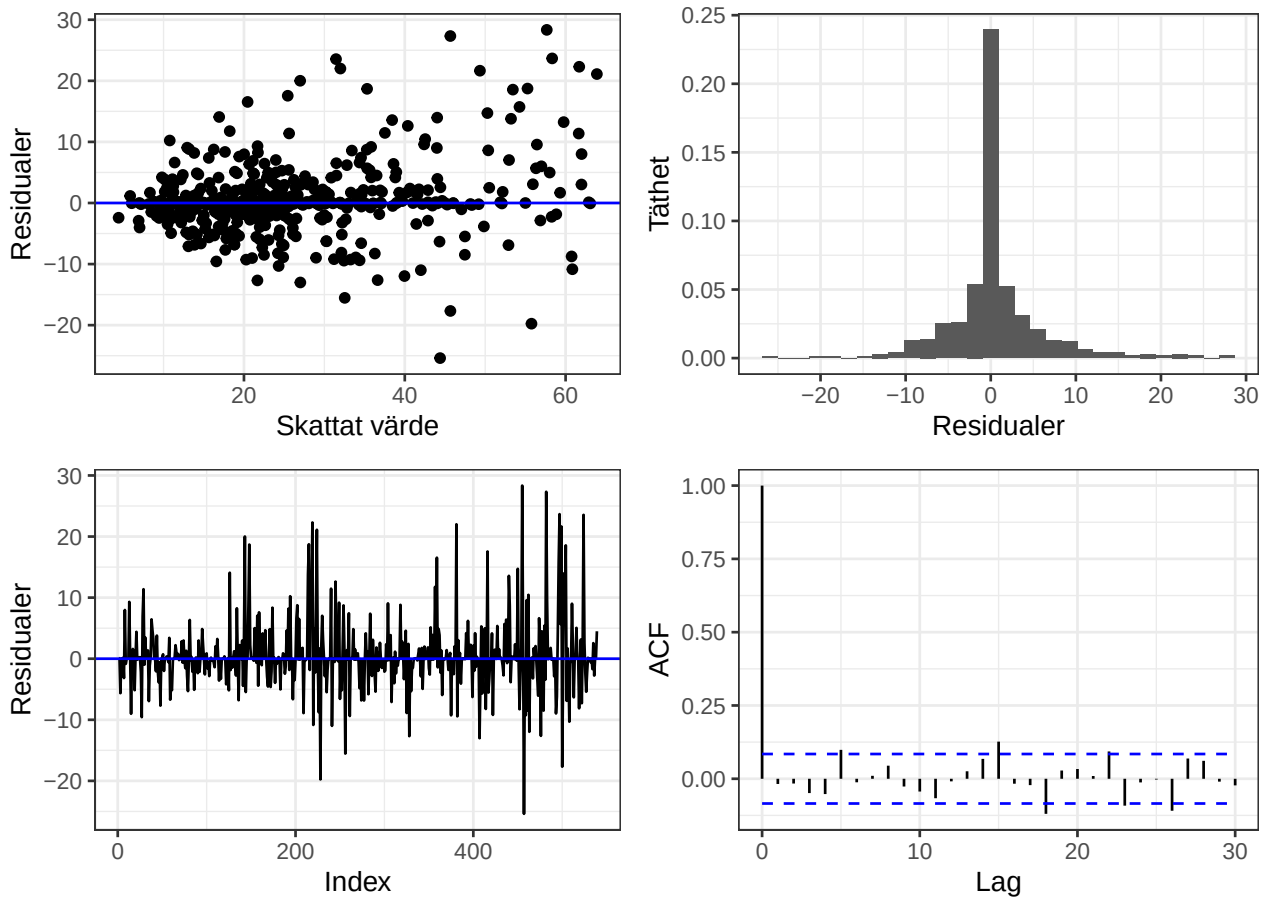
Figur 19: Beeswarm-diagram över varje variabels SHAP-värde för varje observation i modell 1, en ljusare färg indikerar på högre variabelvärde och mörk färg lägre.

SHAP-värden från modell 1 visas i figur 19 som en beeswarm-figur. Figuren indikerar SHAP-värde på x-axeln för respektive observation och högre densitet kring ett SHAP-värde medför ett bredare kluster. Laggade Y har högst SHAP-värde vilket är väntat men ej relevant för tolkning av modellen. Detta gäller egentligen för alla externa variabler som månads- och veckodagsvariabler också, även om man exempelvis kan bekräfta det samband som presenterades i Data-kapitlet med lägre genomsnittligt antal skapade nya konton på lördagar, som här visas att ett högt värde på Lördag-dummyvariabeln (dvs 1) innebär ett negativt SHAP-värde.

Annars syns att Meta, Affiliate och Google är de kanaler som generellt har störst inverkan på responsvariabeln, och generellt syns att ju högre värde på varje kanal ju mer positiv påverkan har kanalen på responsvariabeln. Detta syns särskilt för Meta, som har ett kluster på ett fåtal observationer där värdet varit i särklass högst och även SHAP-värdet väldigt högt. Affiliate har någon observationer där kanalspenderingen är hög även om Shap-värdet är högt, men generellt gäller ändå det positiva sambandet.

Programmatic har generellt en låg spendering, däremot syns att ett par observationer har hög spendering ändock utan särskilt bättre SHAP-värde. Influencer och Tiktok har SHAP-värden kring 0, och särskilt för TikTok syns att det är spridda skurar vad gäller magnitud på spenderingen kontra SHAP-värdet.

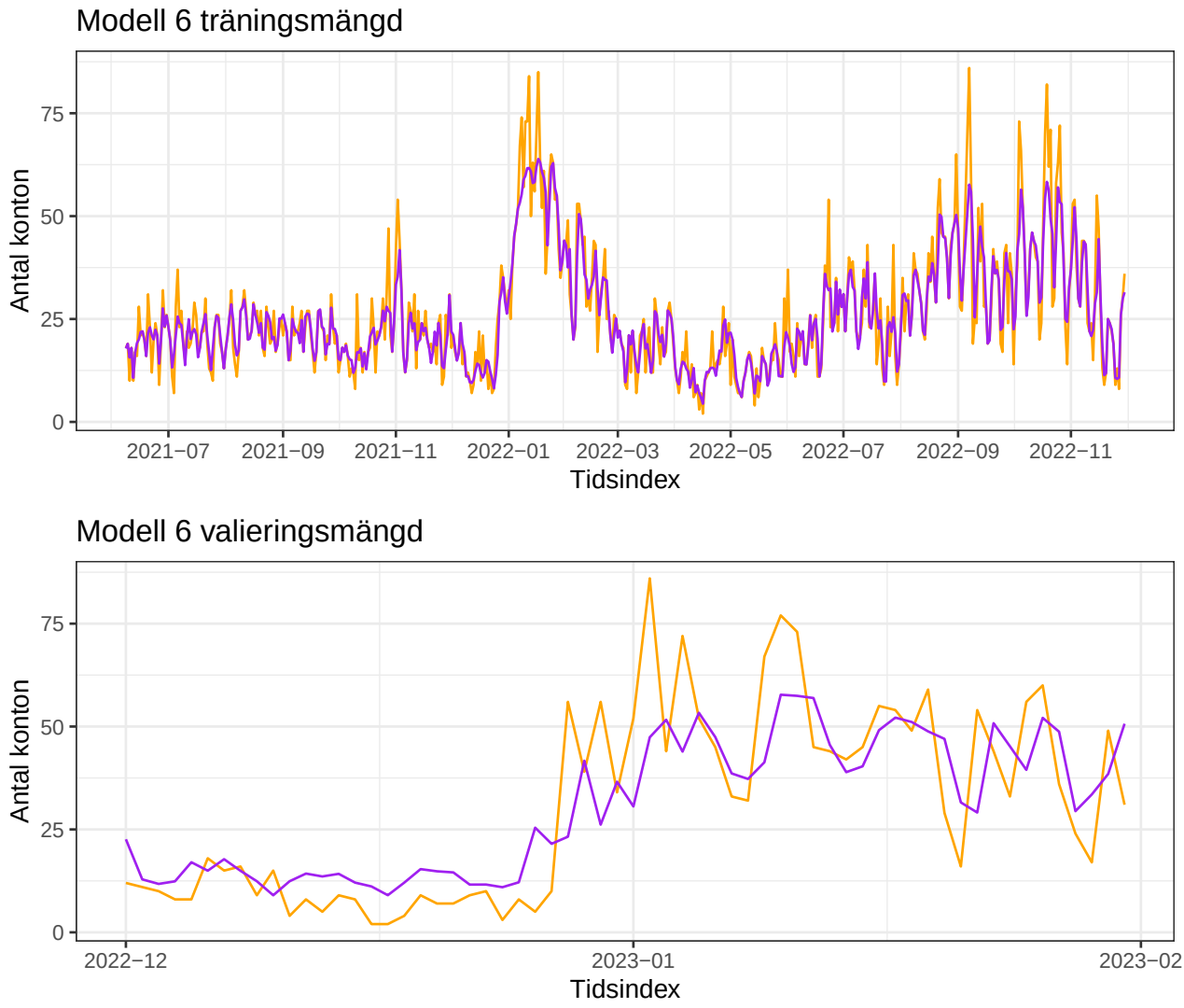
4.2.2 Modell 6



Figur 20: Residualanalys Modell 6 MAE

Modellen med variabler transformerade med en sönderfallshastighet på 0.5 får symmetriska residualer och det syns tydligt i histogrammet. Det finns en liten trattform i residualvariansen när de plottas mot skattade värden, men det går också att urskilja konstant varians om extrempunkter exkluderas.

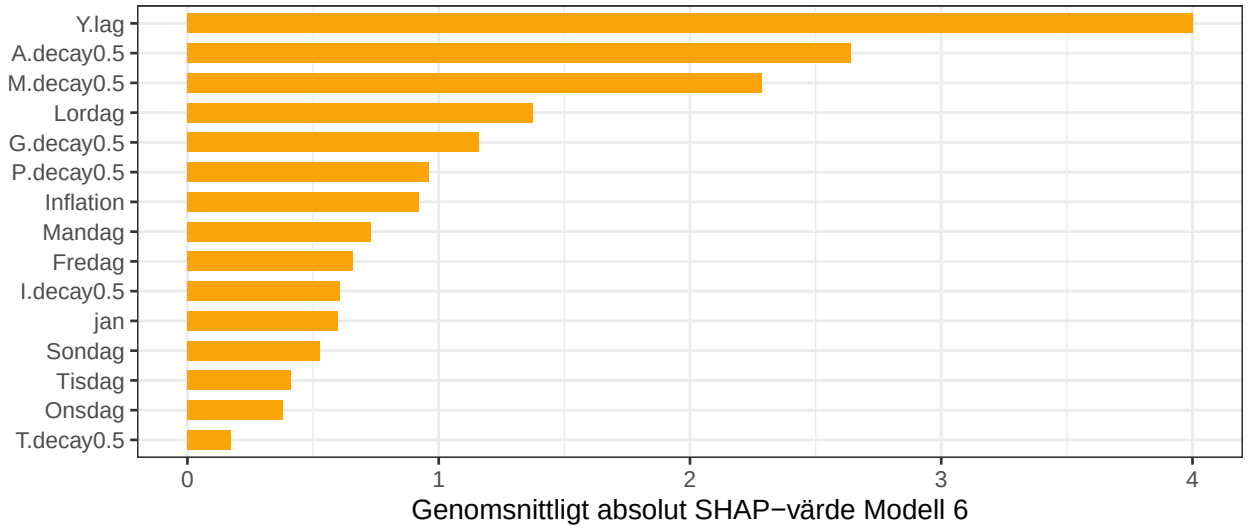
Det är svårt att tyda om eventuellt tidsberoende finns kvar bland residualerna genom att plotta dem mot tidsindex och autokorrelationsfunktionen för residualerna studeras istället. ACF-plotten visar en del spikar för vissa lags, men det går inte att se ett speciellt mönster över dessa lags då avståndet mellan dem är oregelbundna, därav kan detta anses som vitt brus snarare än ett tidsberoende och där av uppfylls antagandet om oberoende residualer.



Figur 21: Linjediagram över sanna värden samt prediktioner på tränings- och valideringsdata för modell 6

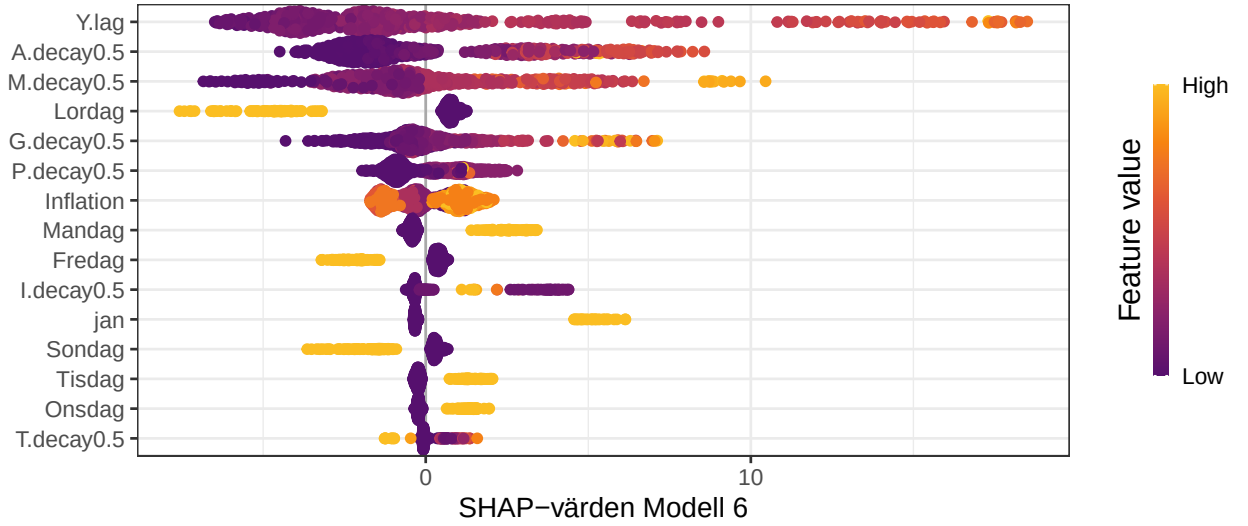
När modellens skattningar för träningsmängden plottas mot sanna värden ser vi att prediktionerna inte riktigt fångar upp de högsta topparna på responsvariabeln men att modellen följer de generella variationerna över tid väldigt bra. Modellen missar även en del av de djupaste dalarna av antal nya konton, men fångar upp dessa bättre än de högsta topparna.

När skattningar för valideringsmängden plottas mot valideringsmängdens sanna värden så ser vi även här att modellen missar de högsta topparna som den även gjorde på träningsmängden, men även att den överskattar de låga värdena lite. Modellen fångar upp den generella variationen för valideringsdata och hittar till exempel ökningen av Y som sker runt slutet på december 2022.



Figur 22: Stapeldiagram över genomsnittliga absoluta SHAP-värden för variablerna i modell 6

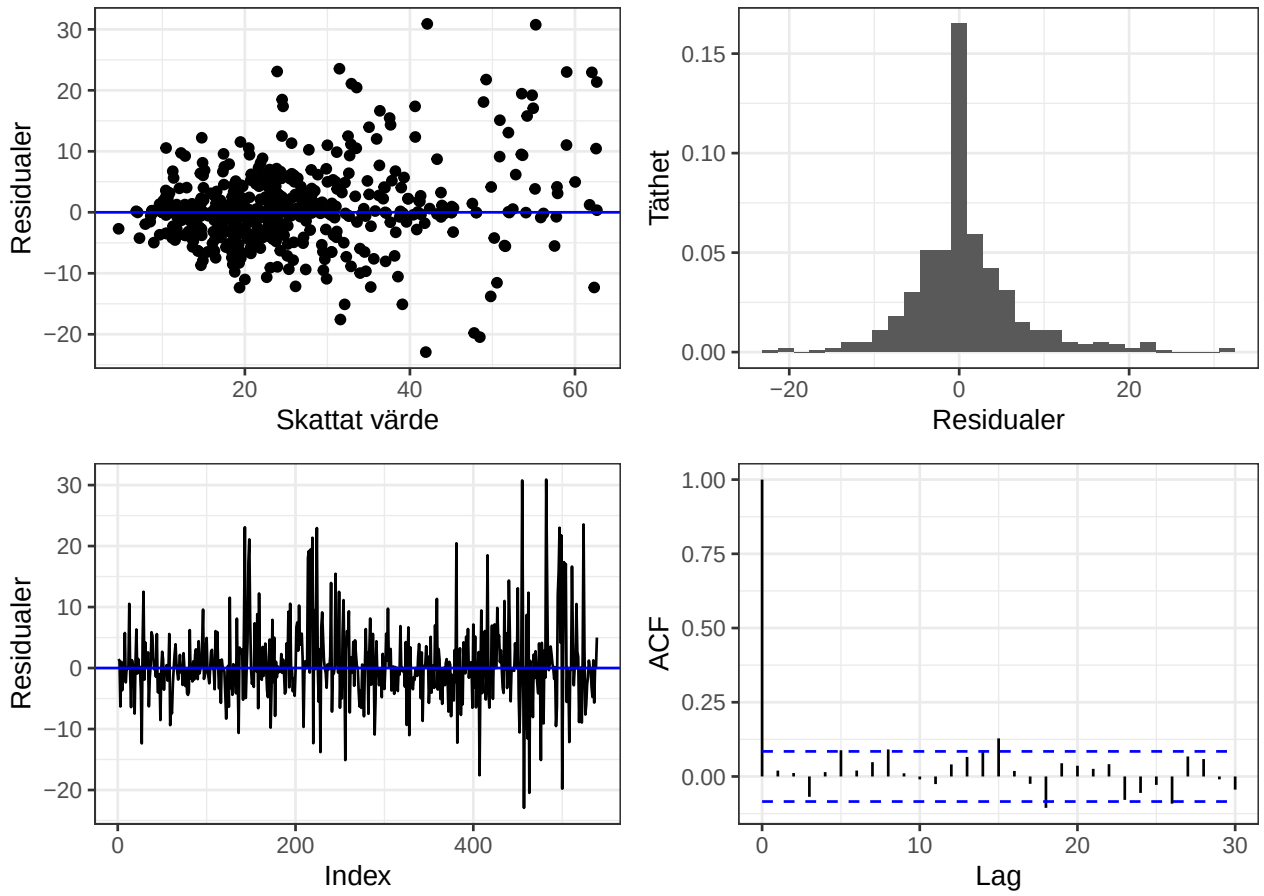
Figur 22 visar genomsnittliga absoluta SHAP-värden för modell 6. Affiliate och Meta är de två transformerade spenderingarna som i genomsnitt påverkar shapvärdena mest. Googles genomsnittliga SHAP-värde har sjunkit gentemot den icke transformerade spenderingen, vilket kan vara att den önskade transformerade effekten inte speglar den riktiga effekten från Google särskilt bra. Annars så följer de genomsnittliga shapvärdena på de transformerade spenderingsvariablerna samma ordning som i modell 1, se figur 18.



Figur 23: Beeswarm-diagram över varje variabels SHAP-värde för varje observation i modell 6

Variablerna i figur 23 är sorterade efter genomsnittliga absolut SHAP-värdet från figur 22, högre spendering på Affiliate, Meta, Google och Programmatic får generellt ett högre SHAP-värde och det syns då de har mörkare färg på observationer som har ett positivt SHAP-värde för deras rader.

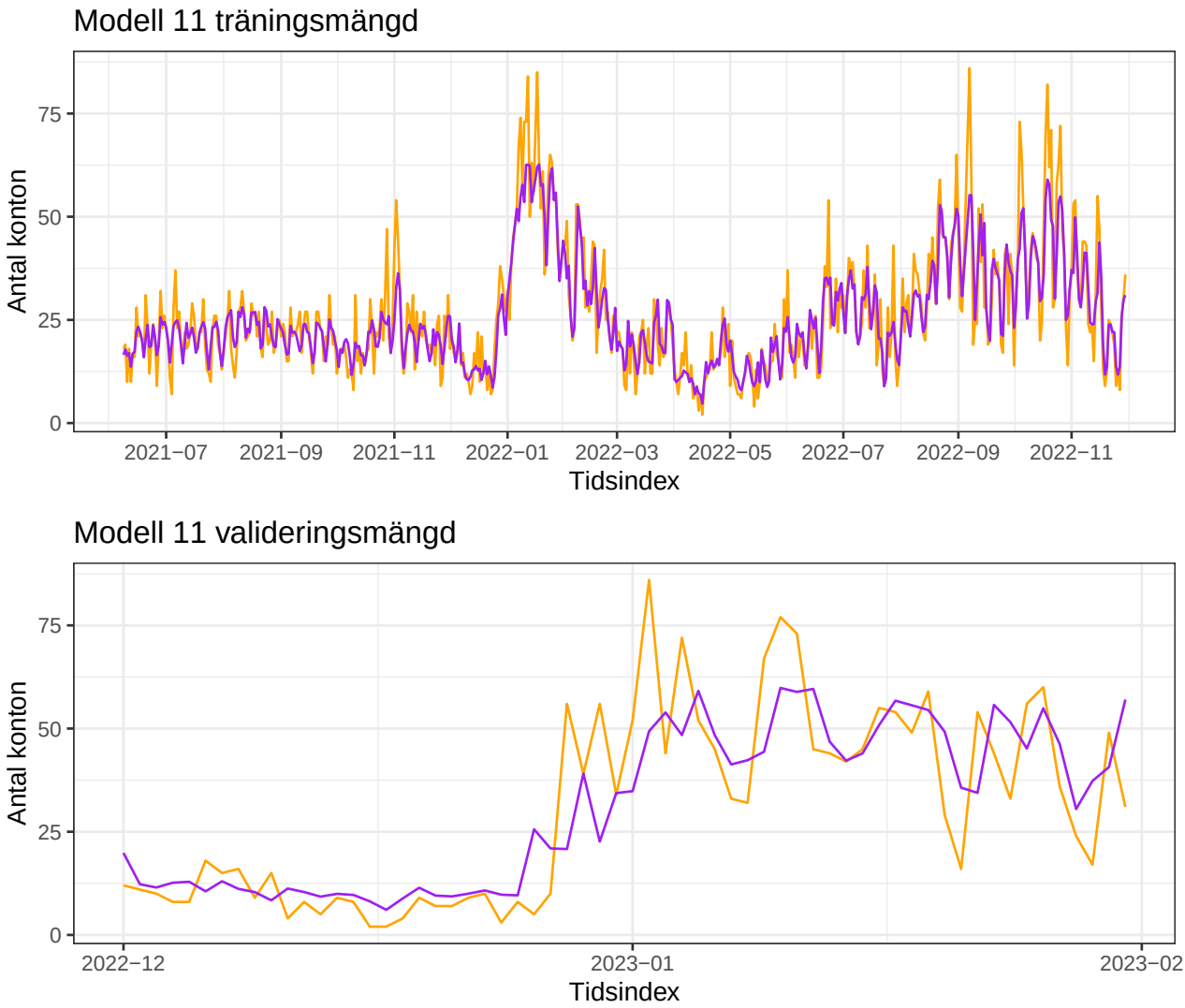
4.2.3 Modell 11



Figur 24: Residualanalys Modell 11 MAE

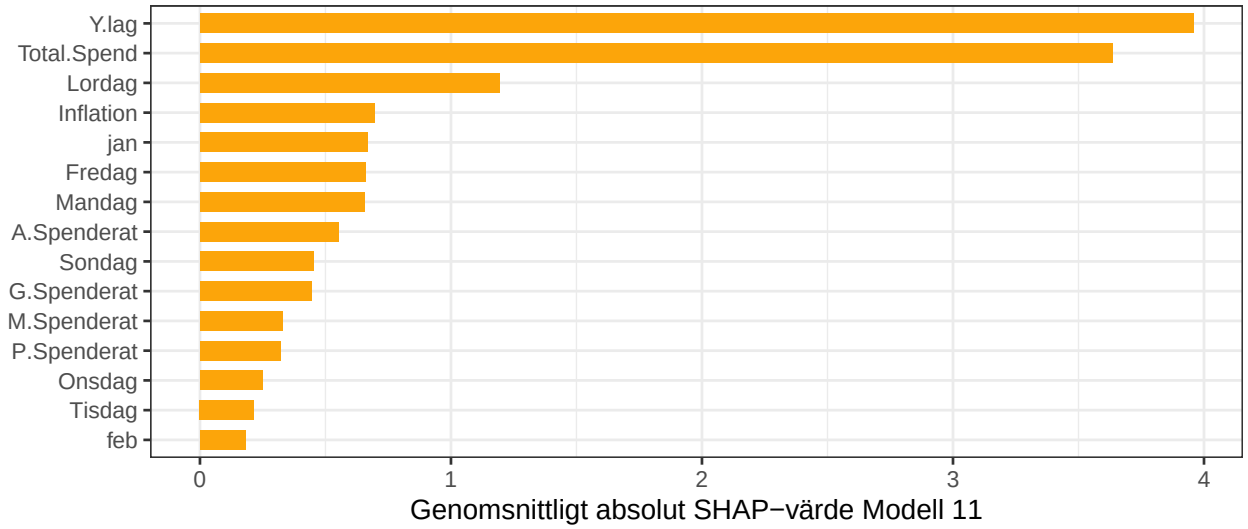
Utifrån histogrammet i figur 24 syns att antagandet om symmetriska residualer tycks hålla till viss del, däremot är det ett antal positiva observationer med högre värde än dess negativa motsvarighet vilket gör att fördelningen är något högerskev. Variansen påvisar trattmönster i residualerna när magnituden ökar och antagandet om konstant varians är inte definitivt uppfyllt.

Vad gäller tidsberoendet påvisar autokorrelationsfiguren ett visst mönster i spikarna på lag 5–8, 12–15, 23–26 och således kan ett svagt beroende i tiden finnas, men eftersom det är så små spikar kan det lika gärna vara vitt brus.



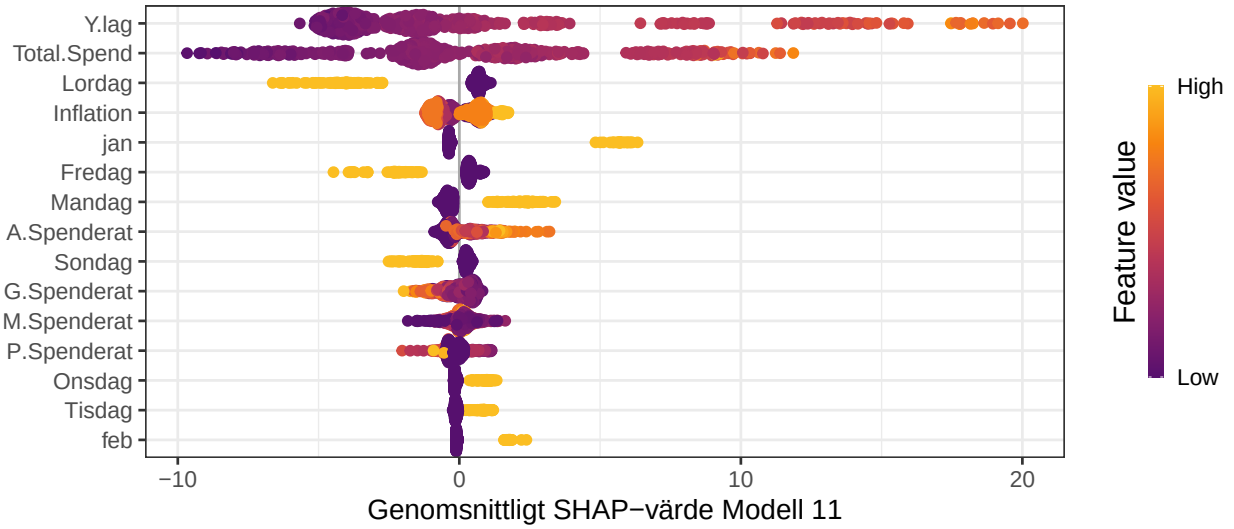
Figur 25: Linjediagram över sanna värden samt prediktioner på tränings- och valideringsdata för modell 11

I tabell 8 så har modell 11 ett MAE på 4.02 och 9.43 för tränings- respektive valideringsdata, således är MAE för träning cirka 1.93 gånger större. Detta är den bästa kvoten som erhållits och därav den modell som överanpassar minst på träningsdata men har ändå ett relativt bra mått i valideringen. Resultatet åskådliggörs i linjediagrammet i figur 25 som visar att modellen är restriktiv på träningsdata och inte fångar upp toppar särskilt bra, men för valideringsdata syns att den följer det generella mönstret väl.



Figur 26: Stapeldiagram över genomsnittliga absoluta SHAP-värden för variablerna i modell 11

Figur 26 visar genomsnittligt absolut SHAP-värde för modellen med andelstransformation. För andelar är som nämnts tidigare den absoluta spenderingen per dag med i modellen som komplement till de relativa andelarna över kanalerna. Denna har, tillsammans med laggade Y-variabeln, klart högst SHAP-värde vilket följer av att en högre spendering generellt kommer leda till en högre antal nya konton. Däremot är det endast så intressant den variabeln blir, det vill säga att mer pengar lagda på de olika kanalerna kommer leda till fler skapade konton. Vi ser att de olika kanalerna, som alltså är andelar definierade inom [0,1], generellt har låg importance. Affiliate, Google och Meta är de viktigaste av kanalandelarna för modellen vilket är förväntat efter resultatet från tidigare modeller. Influencer hamnar här inte ens bland de 15 variablerna som har högst genomsnittligt absolut SHAP-värde.



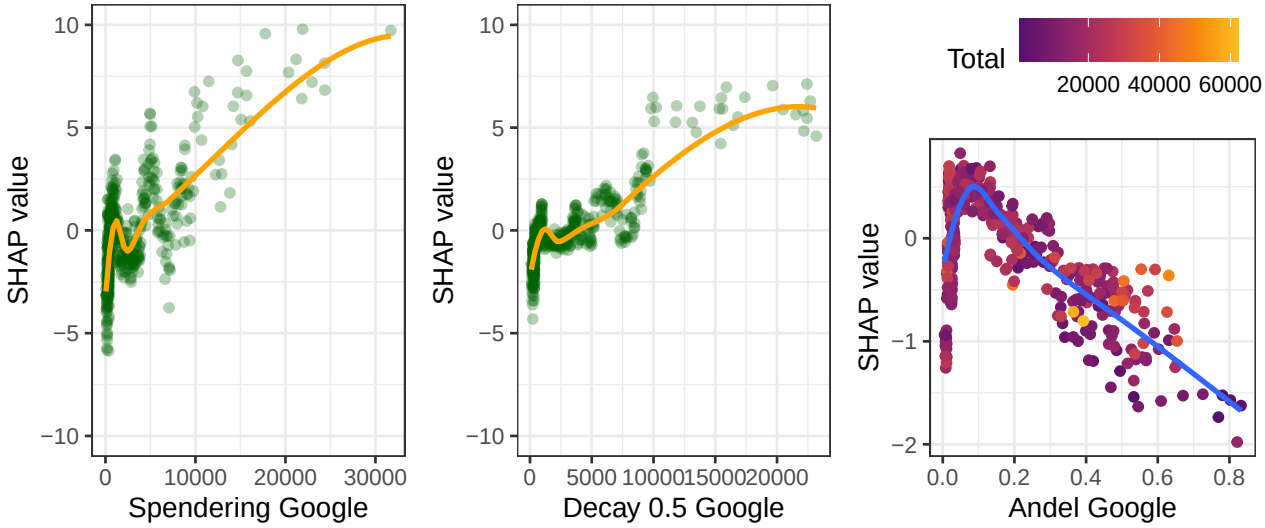
Figur 27: Beeswarm-diagram över varje variabels SHAP-värde för varje observation i modell 11

Figur 27 visar SHAP-värden för varje observation. Här syns som i föregående figur att kanalvariablerna har låg importance då mycket av informationsvinsten faller till laggad Y och total spendering. Affiliate har högst SHAP-värde, och visar även att SHAP-värdet ökar när andelen på kanalen ökar. Detta är inte nödvändigtvis fallet för de andra kanalerna, som exempelvis Google där högre andel är förknippat med ett lägre SHAP-värde. För Meta och Programmatic är det inte lika entydigt utan SHAP-värdet varierar mycket.

4.3 Marginella effekter per kanal

I detta delkapitel kommer beroende-figurer för kanalernas SHAP-värden studeras för att se de marginella effekterna på SHAP-värdet från när respektive kanal inkluderas i modellen. Notera att det är skillnad mellan skalor på y-axlar för andelsfigurerna. Detta för att illustrera sambandet mer detaljerat per modell.

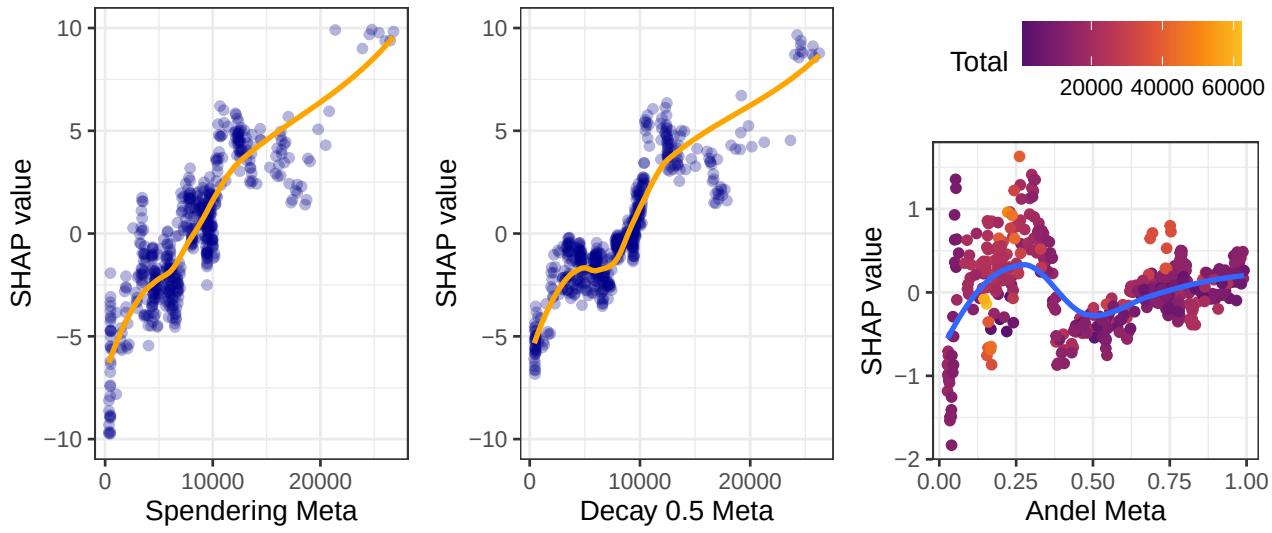
4.3.1 Google



Figur 28: Beroende-diagram för Google-kanalens påverkan på prediktioner för respektive modell. Figuren till höger är färgad efter magnitud på totalspendering.

Effekten av spenderingen på Google ser i figur 28 relativt liknande ut för direkt spendering och sönderfallstransformering med 0.5, men sönderfallstransformeringen har inte lika stor varians för observationer med liknande spenderingsvärde. Efter en spendering eller en sönderfallsspendering på 10 000 NOK så finns det inte lika många observationer så resultatet på dessa värden blir mindre pålitligt. Effekten från spenderingen på Google är som högst när Google har en andel på runt 10 procent av den dagliga spenderingen samtidigt som den totala spenderingen varit relativt låg.

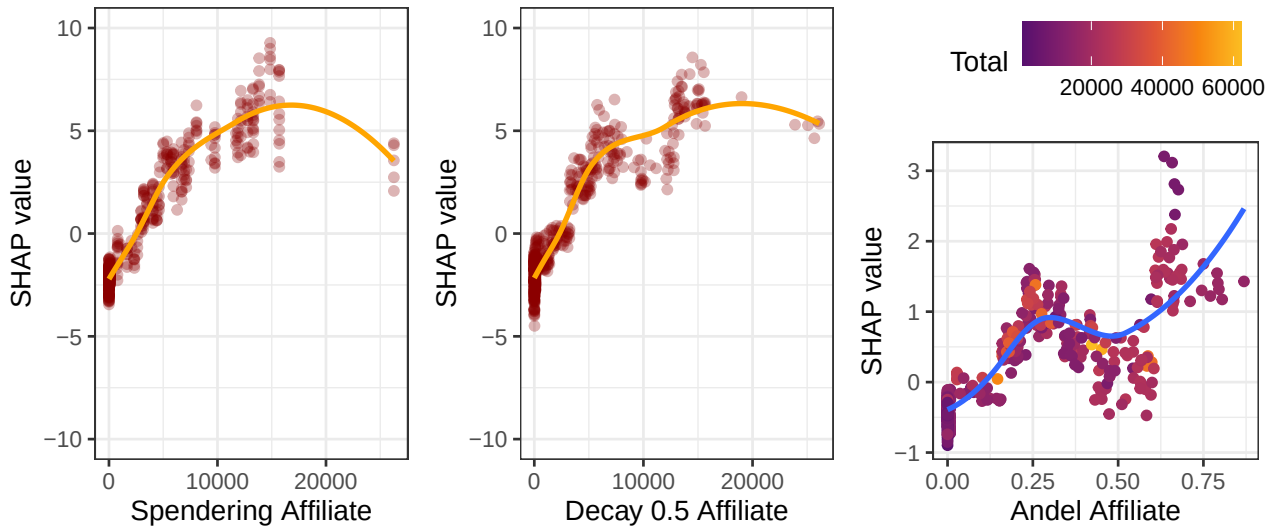
4.3.2 Meta



Figur 29: Beroende-diagram för Meta-kanalens påverkan på prediktioner för respektive modell

I figur 29 syns en positiv effekt på SHAP-värdet finns i spendering på Meta, liknande som för google så finns det färre observationer för de allra högsta spenderingarna så effekten från dem är inte lika pålitlig som de lägre spenderingarna. Att spendera lite på Meta ser inte ut att ge någon speciell effekt, men här är variationen för liknande värden hög och det kan vara interaktioner mot andra spenderingsvariabler som bör undersökas vidare. Effekten från sönderfallstransformeringen visar en stark ökning för värden som ligger mellan 10 000 och 15 000 NOK. Effekten från spendering på Meta är som högst när Metas andel av dagsspendering på marknadsföring är runt 25 procent och samtidigt som den totala spenderingen varierar runt denna andel.

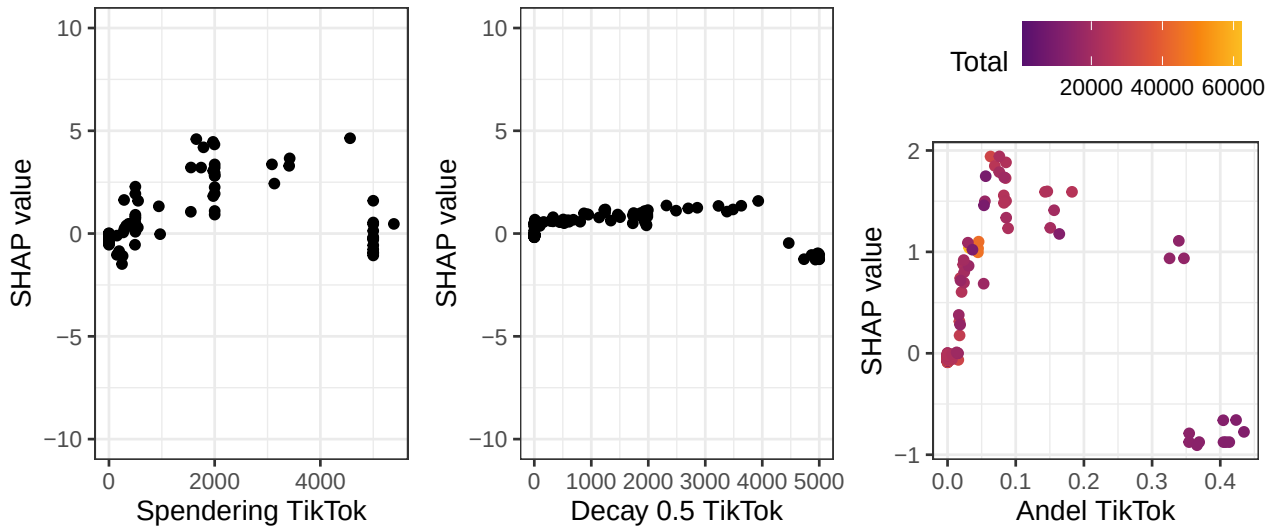
4.3.3 Affiliate



Figur 30: Beroende-diagram för Affiliate-kanalens påverkan på prediktioner för respektive modell

Figur 30 visar SHAP-värdet för respektive Affiliate-kanal. Beräknat från den ursprungliga spenderingen samt decay 0.5 syns att det generellt är ett positivt ökande SHAP-värde ju mer pengar som spenderas på Affiliate, men kring 20 000 NOK har effekten avtagit för båda transformeringarna. För ursprunglig spendering syns särskilt tydliga mönster att det är varierande magnitud på SHAP-värdet för samma spendering på kanalen, vilket tyder på att det kan finnas interaktioner med andra kanaler som också bidrar till effekten i responsvariabeln. Andelsfiguren visar att när totala spenderingen är ganska låg men Affiliate är en stor del av den så får man ut en stark effekt från kanalen. Affiliate ligger ofta kring 20-40% av andelen när totala spenderingen är hög, men påvisar då inte lika stark effekt i responsvariabeln. Detta innebär att Affiliate-kanalen är en bra kanal att lägga en större del av sin budget på när man inte har en särskilt stor budget, men att effekten verkar mattas av ganska fort när den ökar och marknaden mättas.

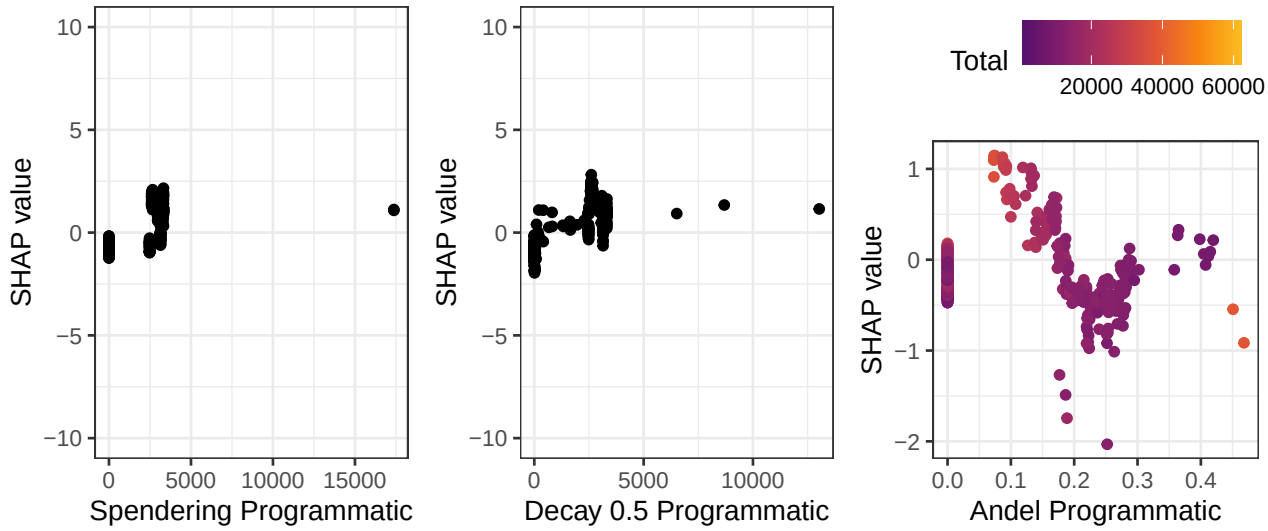
4.3.4 TikTok



Figur 31: Beroende-diagram för TikTok-kanalens påverkan på prediktioner för respektive modell

Figur 31 visar att ursprungliga TikTok-spenderingen är effektiv kring summor om 2000-4000 NOK, och generellt verkar kanalen prestera positivt. Däremot syns att när spenderingen når nivåer över 4000 NOK så mattas effekten från kanalen ut. Sönderfallstransformerade spenderingen är konsoliderad och visar ännu tydligare den nästan entydiga ökning av effekt när budgeten rör sig mot 4000 NOK, och avtappet som sker när det når 5000 NOK. I figuren för andelar syns att TikTok har generellt starkast positiva effekt när kanalen står för runt 5-15% av budgeten, oberoende av magnitud på den totala spenderingen. De få dagar när kanalen står för 30-40% av totala spenderingen erhålls ett negativt SHAP-värde vilket innebär att effekten på Y varit dålig. Värt att notera är att TikTok generellt har väldigt få observationer, och fler observationer hade kunnat utlysa fler och större mönster än detta.

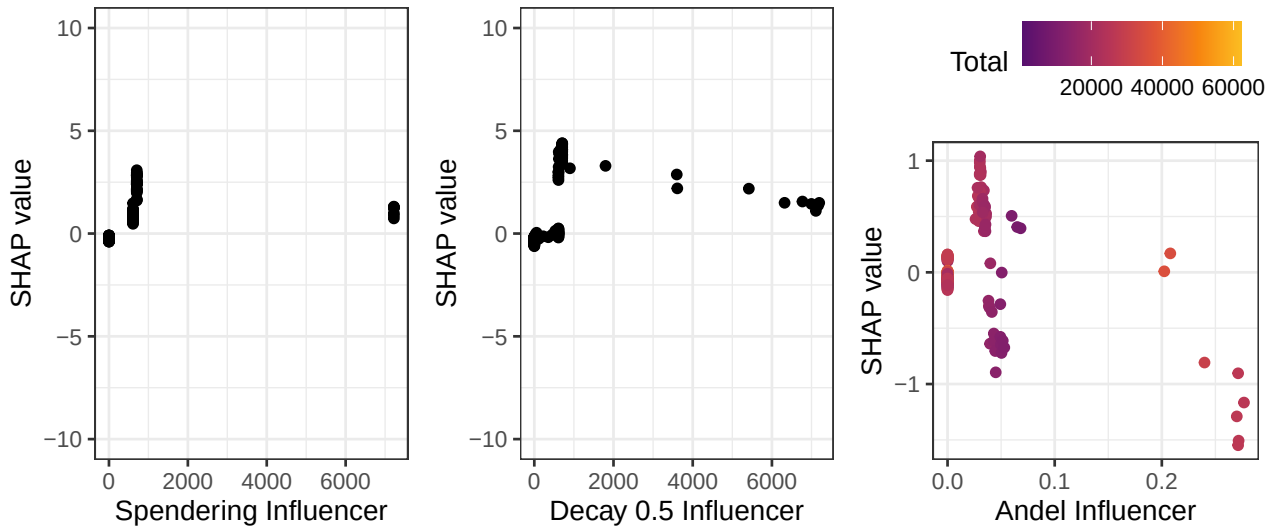
4.3.5 Programmatic



Figur 32: Beroende-diagram för Programmatic-kanalens påverkan på prediktioner för respektive modell

Figur 32 visar att för både ursprunglig spendering och sönderfallstransformering är det väldigt stor varians i effekten från Programmatic vilket gör det svårt att dra slutsatser kring spenderingen. Syns att det generellt aldrig lagts mer än runt 3000 NOK på kanalen och att denna summa bidragit med delvis positiv men även försumbar effekt i skapade konton. För andelar syns att positiv effekt från kanalen i majoritet av fallen kommer från när kanalen står för runt 10% av budgeten och när totala spenderingen är hög. Generellt påvisas en negativ trend när andelen ökar, men totala spenderingen har också sjunkit i takt med att andelen ökat.

4.3.6 Influencer



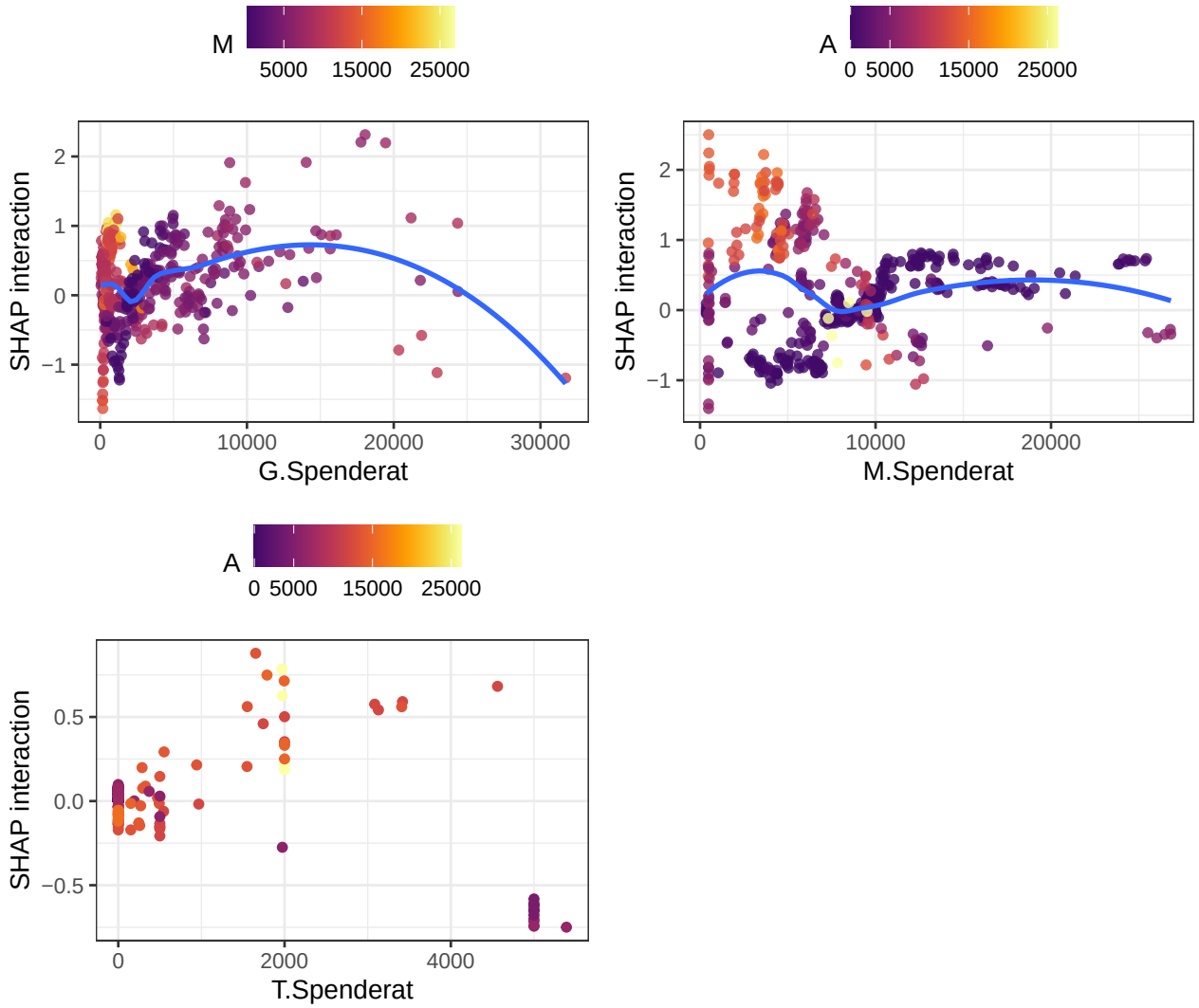
Figur 33: Beroende-diagram för Influencer-kanalens påverkan på prediktioner för respektive modell

Influencer har en relativt hög varians för de olika spenderingsnivåerna vilket syns i figur 33, men att spendera på influencers ser ut att ge högre SHAP-värden för de låga spenderingarna och för de lite högre finns det väldigt få observationer och resultatet från dem är inte lika pålitligt. När sönderfallstransformerade variabeln undersöks så visar den på liknande mönster som den icke transformerade spenderingen. Variansen för SHAP-värdena är relativt hög och det interaktioner kan vara intressanta att undersöka. När Influencer har en andel på cirka 5 procent är effekten av som högst, den totala dagsspenderingen har dock varit relativt låg för de andelsnivåerna.

4.4 Interaktionsdiagram

Eftersom det syntes tydlig varians mellan SHAP-värden för samma spenderingsnivå för flera kanaler indikerar det på att interaktioner med andra kanaler kan finnas. Interaktionsfigurerna som visas i delkapitlet är de som för varje modell visar någon form av interaktionseffekt mellan kanalerna för respektive modell.

4.4.1 Modell 1



Figur 34: Interaktionsdiagram mellan kanalerna (1) Google mot Meta (uppe till vänster), (2) Meta mot Affiliate (uppe till höger) och (3) TikTok mot Affiliate för modell 1 (nere till vänster).

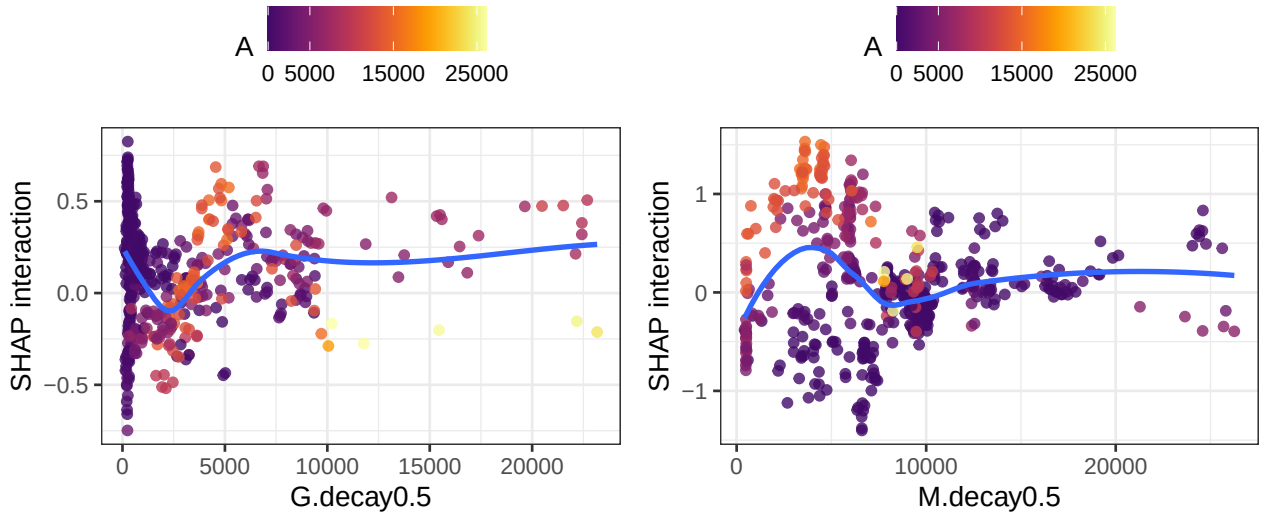
Figur 34 visar interaktionerna som anses intressanta i modell 1. Uppe till vänster syns Google mot Meta. Figuren påvisar en interaktionseffekt mellan kanalerna när man lägger pengar på Meta, spendering på båda kanalerna samtidigt bidrar till högre SHAP-värden än enskild spendering, däremot spenderas inte avsevärt

mycket på kanalerna samtidigt vilket gör att sambandet är svårt att bekräfta. Samtidigt finns det en ökande effekt även när Google ökar, och Meta är låg, vilket än mer tyder på att ett samband finns mellan kanalerna.

Uppe till höger syns Meta mot Affiliate, figuren visar att när Affiliate är hög och Meta är låg finns det en interaktionseffekt. När Meta ökar är antalet observationer där Affiliate är nollskild få vilket gör att det är svårt att se interaktionseffekter från högre nivåer på Meta.

Nere till vänster visas TikTok mot Affiliate. Få observationer gör resultatet mindre pålitligt, men de som finns visar på ett samband mellan kanalerna.

4.4.2 Modell 6

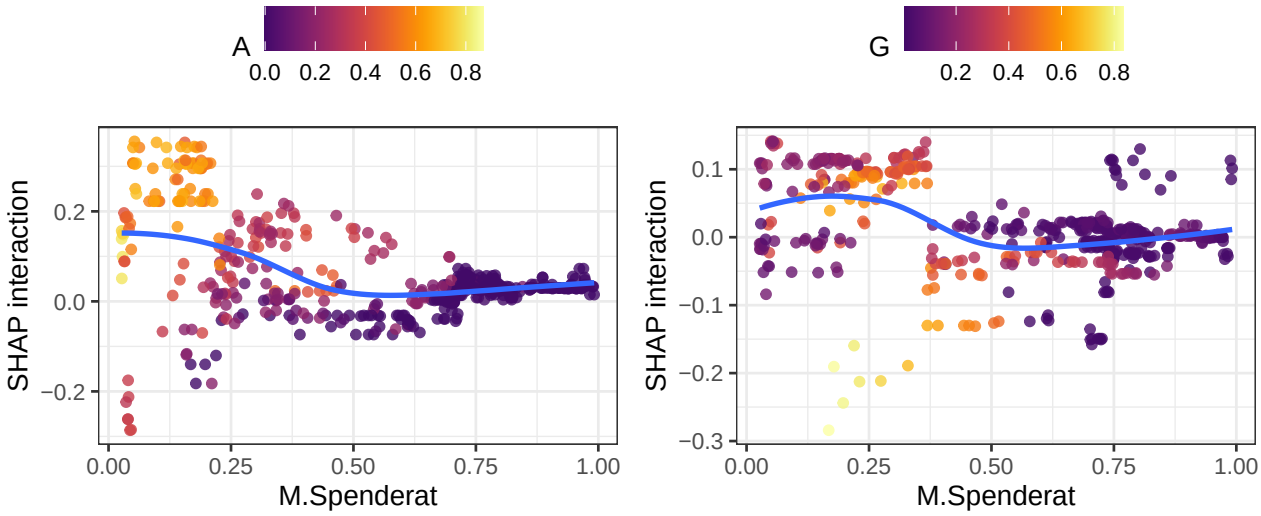


Figur 35: Interaktionsdiagram mellan kanalerna (1) Google mot Affiliate och (2) Meta mot Affiliate för modell 6.

I figuren för Google mot Affiliate syns att det verkar finnas en interaktionseffekt mellan kanalerna när sönderfallsspenderingen på Affiliate ökar och Google är låg. Däremot när Google ökar så verkar sambandet förändras, men här är observationerna väldigt få.

För Meta mot Affiliate syns en liknande effekt, när sönderfallsspenderingen på Meta ligger runt 5000 NOK och Affiliate runt 15000 NOK. För ökande värden på Meta tycks interaktionseffekten minska.

4.4.3 Modell 11



Figur 36: Interaktionsdiagram mellan kanalerna (1) Meta mot Affiliate och (2) Meta mot Google för modell 11.

Figur 36 visar hur andelarna samverkar med varandra. Meta mot Affiliate visar på en interaktionseffekt kanalerna mellan när Affiliate har runt 60%-80% av budgeten och Meta har runt 10-15%. Denna effekt börjar sedan avta när Meta taru pp en större del av budgeten och Affiliate minskar.

För Meta mot Google finns en svag effekt när Google har cirka 40-50% av budgeten och Meta har runt 25%, däremot när Googles andel ökar tycks sambandet avta. När Meta ökar verkar effekt generellt avta mot 0.

5 Diskussion

5.1 Resultatdiskussion

I detta kapitlet analyseras det resultat som erhållits för att kunna besvara frågeställningar i slutsatsen.

För att kunna analysera samband mellan magnituden av spenderingen på en marknadsföringskanal och effekten i responsvariabeln användes dependence plots, som genom spridningsdiagram visar SHAP-värdet respektive kanal har för varje observation. Ett högre SHAP-värde innebär en högre bidragande faktor till värdet på responsvariabeln för varje dag. Transformeringar genomfördes för att försöka hitta nya beteenden mellan förklarande variabler mot responsvariabeln.

Tre modeller med olika transformeringar användes för vidare analys då de presterade bra vad gällde MAE på validerings- och träningsmängd, hade residualer som ansågs bäst uppfylla de antaganden som modellen kräver och ej innehöll autokorrelation. Däremot finns det problem med ökande varians för ökade prediktionsvärden vilket gör att det antagandet inte är entydigt uppfyllt för modellerna.

De tre modellerna var modell med (1) ursprunglig spendering, (2) sönderfallshastighet om 0.5 och (3) andelar. I analys av SHAP-värden för dessa syntes att Meta, Google och Affiliate är de kanaler som överlag har högst inverkan på responsvariabeln för alla tre modeller, däremot är de inte i enhetlig ordning vad gäller påverkan mellan modellerna. Programmatic presterar genomgående strax utanför de tre viktigaste kanalerna, och Influencer med TikTok är i botten vad gäller SHAP.

Google och Meta har för båda modellerna med absoluta spenderingsvariabler väldigt positiva trender i sin data och högre spendering på kanalerna är förknippat med en stark positiv respons i antalet nya konton som skapas hos företaget oavsett magnitud på spenderingen, även om antalet observationer på högre nivåer är ganska få för båda. Affiliate påvisar en liknande positiv effekt, men mattas av ganska fort när budgeten ökar och ett tydligare "tak" där marknaden för Affiliate verkar mättas kan tänkas vara resultat av detta. Spendering på TikTok ger positivt resultat så länge spenderingen hålls ganska låg, men ökas budgeten uppkommer en tydlig avtagande effekt. Programmatic och Influencer har generellt sporadisk spendering med hög variation i SHAP-värde för lika magnitud av spendering, men även där tycks någon form av avtagande effekt när budgeten ökar kunna tydas.

Magnitud i spendering för att se positiv effekt i antal nya kanaler skiljer sig mellan kanaler. För Google syns generellt att spendering över 4000 NOK om dagen krävs för att få en enhetlig positiv respons. För Meta ligger gränsen runt 8000 NOK. Kanalen för Affiliate påvisar tidigt en positiv effekt i marknaden och antalet nya konton ökar vid en spendering om 3000 NOK. TikTok och Influencer har mer variation i när effekt kan tydas, men generellt kan man förvänta sig att spendera ungefär 1000 NOK per dag för att påverka mängden nya konton som skapas. Programmatic har spenderats i liknande nivåer genomgående vilket resulterat i både negativ och positiv påverkan på modellen, men en majoritet av dagar har antal nya konton påverkats positivt och således, utifrån den data som finns, bör en spendering vara runt 3000 kr om dagen för att kunna se positiva resultat.

Från studiens resultat är det svårt att med säkerhet säga vilken andel av budgeten för varje marknadsföringskanal som ger en hög avkastning eftersom att resultaten från andelsmodellen inte är helt pålitliga för alla kanaler. Figurerna som visar budgetandelar per dag för varje kanal syftar ge insyn i hur fördelningen skulle kunna optimeras. Google-kanalen har en positiv effekt när andelen är kring 10% av budgeten och totalspenderingen är ganska låg, däremot avtar effekten mot det negativa när andelen i Google ökar. Meta har generellt en stark positiv effekt när andelen är kring 15-20% av dagsbudgeten och totala spenderingen är ganska hög. Det finns även data som indikerar att Meta har en positiv effekt när andelen är hög och den totala spenderingen är låg. Affiliate-kanalen har positiv effekt när andelen är kring 25% och totala spenderingen är hög, samt när andelen är kring 70% och totala spenderingen är låg. Däremot är effekten från Affiliate svag kring 0 i SHAP-värde och i vissa fall negativ när budgeten är hög och andelen kring 50%. Detta verkar återigen peka på att Affiliate är en effektiv kanal för respons men att marknaden verkar mättas ganska fort. TikTok och Programmatic presterar

klart bäst när total spendering är ganska högt och andelen inte överstiger 15%, ökning av budget innebär att effekten sjunker mot det negativa. Programmatic påvisar även att kanalen generellt presterar dåligt när budgeten är låg. Influencer visar egentligen inga tydliga mönster, kanalen har endast stått för som mest 25 procent av budgeten och totalspendering var då hög men effekten från kanalen var negativ. Generellt verkar Influencer vara mest effektivt när inte mer än ca 5% av budgeten läggs på kanalen.

För att än mer analysera hur kanalerna påverkar antal konton som skapas har interaktioner studerats mellan de olika modellerna. Generellt är det få interaktioner mellan kanalerna som visar entydiga, meningsfulla resultat. Däremot finns det ett antal intressanta samband som har positiva interaktionseffekter mellan kanaler. Google och Meta verkar generellt ha en interaktionseffekt mellan sig, däremot endast när nivån på spenderingen är hög för den ena och låg för den andra. Kanalerna har inte spenderats högt samtidigt vilket gör att det är svårt att säga något om ett eventuellt sådant samband. Affiliate är den kanal som visar störst interaktionseffekter, särskilt mot Meta, men också TikTok och Google. Är Affiliate hög och Meta låg syns en stark positiv effekt, som sedan avtar ju mer budgeten på Meta ökar, vilket även gäller för andelarna. Trots få observationer verkar också Affiliate och TikTok ha en positiv effekt, om än ganska svag.

5.2 Diskussion av felkällor

Genom att titta på hur budgetfördelningen sett ut över tid har företagets spenderingsmönster kunnat urskiljas, särskilt syns stora skillnader under sista perioden när nya kanaler testats. Utifrån att det finns väldigt få observationer för vissa kanaler är det svårt att pålitligt kunna tyda deras effekt på antal nya konton och att eventuellt optimera andelen efter dessa samband blir inte heller pålitligt då färre observationer ökar ovissheten. Utöver det så överanpassar alla modeller lite vilket gör att sambanden och effekterna som syns från modellerna ska ses med en viss osäkerhet. Dessutom bidrar det med ytterligare osäkerhet när antagandet om konstant varians inte är uppfyllt då modellen inte presterar likvärdigt för alla predikterade värden.

När Tiktok inkluderas i budgeten så sänks inte bara spenderingen på Google, Meta och Affiliate utan den totala spenderingen minskar då också, detta gör att det är många förändringar under kort tid vilket gör det svårare för en modell att kunna urskilja säkerställda samband mellan kanalvariabler och antal nya konton. Fokuset med denna undersökning är att försöka skatta sambandet mellan användandet av kanaler och responsen som sker i marknaden. Då den data som samlats in är observationsdata och inte data från ett kontrollerat experiment är det svårare att dra definitiva slutsatser kring kausaliteten mellan variabler (Kleinbaum 2015). Kontrollerade experiment ger större kontrol över de variabler som används och kan enklare isolera effekten från enskild variabel på responsvariabeln, däremot kan observationsdata ge mer generaliserade resultat som inte är lika beroende av de exakta specifika observationerna som uppsatsen har tillgång till. När kanalspenderingar skiftar så pass mycket som de gör vid exempelvis inkludering av Tiktok blir det än svårare att konstatera slutsatser.

Ett problem som tidigare studier inom MMM tar upp och som nämndes i bakgrunden är att inte tillräckligt med data finns tillgängligt för att en robust och pålitlig modell ska kunna byggas, vilket även är fallet för denna studie. Detta påverkar möjligheten att pålitligt kunna se effekten som spendering på marknadsföring ger för vissa kanaler vad gäller konverteringar. Snapchat uteslöts ur analysen på grund av hur lågt belopp som lagts på kanalen samt under väldigt kort tid. Detta fenomen går även att tyda hos flera av de kanaler som behölls i analysen, dock i mindre utsträckning. För att analysera de sanna effekterna från dessa kanaler mot antal nya konton så hade mer data och högre belopp per dag behövts för att kunna dra säkerställda slutsatser.

Imputeringen på Google utförs under samma dagar som antal nya konton är som högst, alltså de värden som exkluderats för att få symmetriska residualer för modellerna, vilket betyder att modellens variabler inte kunde förklara de höga värdena på Y innan exkludering. Eftersom det är just de dagarna som det saknas data från Google så kan det vara så att Google egentligen förklarar de höga värdena och att imputeringen är långt ifrån de sanna värdena. Detta går endast att spekulera runt och då de andra höga spenderingarna på Google inte bidrar otroligt mycket mer till prediktioner enligt SHAP-värdet så är det från modellerna svårt att se att Googles saknade data ska kunna förklara de extremt höga Y-värdena som exkluderats, men finns ändå en osäkerhet kring det.

5.2.1 Andra metodval

Tidigare studier inom MMM har använt bayesiansk tidsserieregression som modellval vilket har resulterat i bra modeller, men att komplicerade transformationer behövs för att modellera de icke-linjära effekterna från spenderingen (Jin 2017). Det hade varit intressant att testa hur den här studiens modeller står sig mot en sådan modell och om effekterna på spenderingen från modellerna är lika.

(Wigren R 2019) testar flera olika modeller för att se vilken som presterar bäst på deras simulerade data och på data från ett företag, resultatet från den studien är att XGBoost är bäst för 2 av 3 utvärderingsmått i det simulerade datamaterialets testmängd, men att modellvalet inte presterar särskilt bra på företagets riktiga data. De bästa modellerna där är istället bayesiansk tidsserieregression, bayesiansk hierarkisk modellering och även Shapley-värde regression.

De icke-linjära effekterna finns beskrivet inom marknadsföring samt att inga starka linjära samband hittades mellan antal nya konton och spenderingsvariablerna gjorde att XGBoost ansågs som en bättre metod framför linjär och bayesiansk regression.

Att mäta effekten av spendering på marknadsföring genom antal nya konton är inte vanligt i andra studier som genomfört MMM utan snarare görs det på försäljningsiffror, då antal nya konton inte är en direkt inkomstkälla för företaget. Eftersom företaget har en lång period från när första steget för en eventuell försäljning sker till betalning samt potentiell påverkan av yttre faktorer som lönebeviljande utifrån ekonomisk förmåga så valdes antal nya konton som konverteringsvariabel. Detta anses vara en konvertering som är så pass nära marknadsföringen som möjligt men ändå noterar en bekräftat ny anskaffad kund som troligtvis leder till försäljningsvärde.

Det finns modeller som har antagande att variansen varierar över perioden vilket skulle kunna ha använts för att göra uppsatsen än mer djupgående, exempelvis Natural Gradient Boosting (NGBoost) (Duan et al. 2020) och XGBoostLSS (XGBoost Location, Scale, Shape) (März 2019). Det kan även finnas vinning i att utvärdera förlustfunktionen än mer, exempelvis Huber-Loss som fungerar som MAE vid stora residualer och RMSE vid små residualer (Hastie, Tibshirani, and Friedman 2001).

5.2.2 Optimering av XGBoost

Eftersom valet landade i att använda sig av random search korsvalidering med 1000 iterationer för optimering av hyperparametrarna i XGBoost så är det många möjliga kombinationer mellan alla hyperparametrar som inte testas, men att tid istället har sparats. Modellerna regulariseras inte tillräckligt med de hyperparameter-kombinationer som valts, utan hade kunnat undersökas mer för eventuell förbättring mot prestanda. Det är därav troligt att bättre modeller hade kunnat skapats vid fler antal iterationer eller att byta metod av tuning till bayesian optimization som är en mer beräkningstung metod men också mer "blackbox" än random search korsvalidering. Den lär sig om hyperparametrarnas inverkan på modellens skattningar och stegvis förbättrar värden på hyperparametrarna för att avslutningsvis få en väldigt bra modell, snarare än att testa ett givet antal kombinationer (Géron 2019).

Förutom en förändring i val av tuning så har XGBoost fler hyperparametrar än de studien fokuserat att använda sig av och det är därav också troligt att inkludering av andra hyperparametrar hade kunnat resultera i bättre modeller, men för varje tillagd hyperparameter så ökar antalet kombinationer exponentiellt.

5.2.3 Fortsatta arbeten till framtiden

Att dela upp datamaterialet i en till del, testmängd, skulle göra att eventuell överanpassning kan ses ännu tydligare eller stärka att den bästa modellen regulariserats på ett bra sätt för att plocka upp de generella sambanden (Géron (2019)).

Ytterligare steg som skulle kunna tas i en ny analys är att utforska variabeltransformeringar mer. Exempelvis hade det varit intressant att pröva en sönderfallsfunktion eller rullande kumulativ summa för andelar för att se om de har någon kvarvarande effekt i tiden. Även att optimera den hyperparameter som styr sönderfallsfunktionen enskilt per kanal för att hitta skillnader i hur effekten från marknadsföringen skiljer sig mellan olika kanaler och möjligt olika kundgrupper. Resultaten från modellen visade också att transformeringen för sönderfallshastighet inte nödvändigtvis förbättrar modellen, och att lägst MAE på valideringsdata erhölls av modellen utan variabeltransformeringar. Återigen kan detta bero på att kanalerna här har kodats med entydig takt på sönderfallet och att mer enskilt optimerad takt per kanal hade gett bättre resultat, men det kan också bero på att sönderfallshastighetsfunktionen inte representerar den långvariga effekten från marknadsföringen över tid för denna bransch.

Den enda externa variabel som testats för utöver tidsvariabler är inflation. Inflation är tydligt relaterat till lån i allmänhet vilket gör att den potentiellt påverkar folks inställning till att bli kunder hos företaget, men det kan också finnas andra externa variabler som kan ha påverkan. Däremot kräver det en viss domänkunskap för att kunna resonera väl kring vilka dessa externa faktorer kan vara och för studiens omfång ansågs det tillräckligt att endast inkludera denna.

Nästa steg när en pålitlig och robust modell har skapats, för att få ett fulländat resultat från en mixad mediamodell, är att göra en numerisk optimering av modellen. Det görs för att hitta den kombination av andelar för varje kanal som tillsammans med till exempel medelvärdet på totalspenderingen ger det högsta predikterade värdet på antalet nya konton. Detta för att få en optimering om när budgetfördelningen har varit som mest effektiv.

6 Slutsats

- För vilken spenderingsnivå påvisar kanaler en positiv effekt på antal nya konton?
  - För Google syns att en dagsspendering på 4000 NOK krävs för att se en positiv effekt i antal nya konton. För Meta krävs en spendering på 8000 NOK, för Affiliate krävs en spendering på 3000 NOK. TikTok, Influencer och Programmatic är de kanaler med resultat som anses mindre pålitliga, men en dagsspendering på 1000 NOK för de två förstnämnda och 3000 NOK för den sistnämnda krävs generellt för att en positiv effekt i antal nya konton ska kunna tydas.
- Kan transformering av spenderingen användas i modeller för att förklara den långvariga effekten från marknadsföring?
  - Transformering av spendering i kanalerna ger inte ett bättre resultat i modelleringen än de icke-transformerade. Detta innebär att en eventuell långvarig effekt inte förklaras av de transformeringar som gjorts.
- Vilken andel bör varje enskild kanal ha av totalbudgeten för att ge en hög avkastning?
  - Google bidrar som mest när andelen är kring 10 procent. Meta bidrar som mest när andelen är kring 25 procent. Affiliate bidrar som mest när andelen är antingen 25 procent vid hög spendering eller 60 procent vid låg spendering. TikTok och Programmatic bidrar som mest när andelen är kring 10 procent. Influencer bidrar som mest när andelen är kring 4 procent.

7 Referenser

Bischl, Bernd, Michel Lang, Lars Kotthoff, Julia Schiffner, Jakob Richter, Erich Studerus, Giuseppe Casalicchio, and Zachary M. Jones. 2016. "mlr: Machine Learning in r." *Journal of Machine Learning Research* 17 (170): 1–5. https://jmlr.org/papers/v17/15-066.html.

Borden, N. H. 1964. "The Concept of the Marketing Mix."

Bowerman, Richard T. och Koehler, Bruce L. och O'Connell. 2004. *Forecasting, Time Series, and Regression : An Applied Approach.* Duxbury Applied Series. Thomson Brooks/Cole.

Brownlee, Jason. 2019. "What Is the Difference Between a Parameter and a Hyperparameter." Machine Learning Mystery. https://machinelearningmastery.com/difference-between-a-parameter-and-a-hyperparameter/.

CFI. 2022a. "Cost of Goods Sold (COGS)." Corporate Finance Institute. https://corporatefinanceinstitute.com/resources/accounting/cost-of-goods-sold-cogs/.

———. 2022b. "CAC LTV Ratio." Corporate Finance Institute. https://corporatefinanceinstitute.com/resources/valuation/cac-ltv-ratio/.

———. 2022c. "ROAS (Return on Ad Spend)." Corporate Finance Institute. https://corporatefinanceinstitute.com/resources/accounting/roas-return-on-ad-spend/.

Chan, Mike, David och Perry. 2017. "Challenges and Opportunities in Media Mix Modeling." Google Inc.

Chen, He, T. 2022. "Package Xgboost." https://cran.r-project.org/web/packages/xgboost/xgboost.pdf.

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." *CoRR* abs/1603.02754.

Dinner, Harald och Neslin, Isaac och van Heerde. 2014. "Driving Online and Offline Sales: The Cross-Channel Effects of Traditional, Online Display, and Paid Search Advertising." *Journal of Marketing Research* 51 (October): 527–45. https://doi.org/10.1509/jmr.11.0466.

Duan, Tony, Anand Avati, Daisy Yi Ding, Khanh K. Thai, Sanjay Basu, Andrew Y. Ng, and Alejandro Schuler. 2020. "NGBoost: Natural Gradient Boosting for Probabilistic Prediction." https://doi.org/10.48550/arXiv.1910.03225.

Géron, Aurélien. 2019. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow : Concepts, Tools, and Techniques to Build Intelligent Systems / Aurélien Géron.* Second edition. Sebastopol, CA: O'Reilly Media.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2001. *The Elements of Statistical Learning.* Springer.

Hyndman, G., R. J. och Athanasopoulos. 2021. *Forecasting: Principles and Practice.* 3rd ed. OTexts: Melbourne, Australia. OTexts.com/fpp3.

James, Daniela och Hastie, Gareth och Witten. 2013. *An Introduction to Statistical Learning: With Applications in r.* 2nd ed. Springer. https://faculty.marshall.usc.edu/gareth-james/ISL/.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning: With Applications in r.* Springer.

Jin, Yueqing och Sun, Yuxue och Wang. 2017. "Bayesian Methods for Media Mix Modeling with Carryover and Shape Effects." Google Inc.

Kleinbaum, L och Nizam, David.G och Kupper. 2015. *Applied Regression Analysis and Other Multivariable Methods.* 5nd ed. Cengage.

Kotler, G., P. & Armstrong. 2012. *Principles of Marketing.* 14th ed. Pearson Education.

Lebow, Sara. 2023. Insider Intelligence. https://www.insiderintelligence.com/content/worldwide-digital-ad-spend-will-top-600-billion-this-year.

Lundberg, Scott. 2018. "Beeswarm Plot." https://shap.readthedocs.io/en/latest/example_notebooks/api_examples/plots/beeswarm.html.

Lundberg, Scott M., and Su-In Lee. 2017. "Consistent Feature Attribution for Tree -Ensembles." *CoRR.* http://arxiv.org/abs/1706.06060.

März, Alexander. 2019. "XGBoostLSS – an Extension of XGBoost to Probabilistic Forecasting." https://doi.org/10.48550/arXiv.1907.03178.

Mayer, Stando, M. 2023. "Package Shapviz." https://cran.r-project.org/web/packages/shapviz/shapviz.pdf.

Newbold. Paul, Granger. Clive och. 1974. “Spurious Regressions in Econometrics.” *Journal of Econometrics* 2 (2): 111–20. https://doi.org/https://doi.org/10.1016/0304-4076(74)90034-7.

Pandey, S och Chhajed, S och Gupta. 2021. “Marketing Mix Modeling (MMM) -Concepts and Model Interpretation.” *International Journal of Engineering and Technical Research* 10 (June): 784–93. https://doi.org/10.17577/IJERTV10IS060396.

Shamsuzzoha, Heli, Ahm och Raappana. 2021. “Perspectives of Business Process Ethics in Data-Driven Marketing Management.” *SECURITY AND PRIVACY* 4 (6). https://onlinelibrary.wiley.com/doi/abs/10.1002/spy2.177.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*.

Wigren R, Cornell F. 2019. “Marketing Mix Modelling: A Comparative Study of Statistical Models.” Linköpings Universitet. http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-160082.

Wolfe, John C., Michael och Crotts. 2011. “Marketing Mix Modeling for the Tourism Industry: A Best Practices Approach.” *International Journal of Tourism Sciences* 11 (January): 1–15. https://doi.org/10.1080/15980634.2011.11434633.

Wolfe, Michael. 2011. “Marketing Mix Modeling for the Tourism Industry: A Best Practices Approach.” *International Journal of Tourism Sciences* 11 (January): 1–15. https://doi.org/10.1080/15980634.2011.11434633.