

Hur man utnyttjar maskininlärning för målinriktad marknadsföring

Amos Weckström

Kandidatavhandling i datateknik

Handledare: Luigia Petre

Fakulteten för naturvetenskaper och teknik

Åbo Akademi

2023

Table of Contents

[Introduktion 3](#_Toc131433295)

[Teoretisk Bakgrund 0](#_Toc131433296)

[**2.1 Maskininlärning** 0](#_Toc131433297)

[**2.1.1 Övervakad och Oövervakad inlärning** 1](#_Toc131433298)

[**Övervakat inlärning** 1](#_Toc131433299)

[**Oövervakad inlärning** 3](#_Toc131433300)

[Metod 9](#_Toc131433301)

[**3.1 Dataförberedelse** 9](#_Toc131433302)

[**3.1.1 Datainsamling** 9](#_Toc131433303)

[**3.1.2 Datarengöring** 9](#_Toc131433304)

[**3.1.3 Datatransformering** 10](#_Toc131433305)

[**3.1.3 Datavisualisering** 10](#_Toc131433306)

[**3.1.3 Skalning** 12](#_Toc131433307)

[**3.1.4 Översampling** 12](#_Toc131433308)

[Empirisk Studie 13](#_Toc131433309)

**1. Introduktion**

2. **Maskininlärning**

2.1.1 Övervakad och Oövervakad inlärning

2.1.2 Klassificering och regression

2.1.3 ML Modeller

3. **Metod**

3.1 Dataförberedelse

3.1.1 Datainsamling

3.1.2 Datarengöring

3.1.3 Datamärkning

3.1.4 Datavisualisering

3.1.5 Scaling

3.1.6 Oversampling

3.3 **Modellering**

3.3.1 Val av modell

3.3.2 Optimering

3.4 Utvärdering

3.4.1 Prestandamätning

3.4.2 Resultatinterpretation

**Empirisk studie**

4.1 Beskrivning av datamängden

4.2 Analyser och resultat

4.2.1 Målgruppsanalys

4.2.2 Segmentering

4.3 Diskussion

**Slutsats**

5.1 Sammanfattning av resultat

5.2 Bidrag och implikationer

5.3 Begränsningar och framtida forskning

**Referenser**

**Bilagor**

7.1 Dataförteckning

7.2 Kodförteckning

# **Introduktion**

Internet samt den breda tillgängligheten till teknik har ändrat hur människor konsumerar. Konsumentbeteenden har ändrats drastiskt jämfört med för 20 år sedan. Den ökande användningen av enheter som är kopplade till internet har möjliggjort för användare att snabbt få information om produkter och tjänster [[]](https://www.sciencedirect.com/science/article/pii/S0969698920300771#sec4). I takt med den ökande tillgängligheten till internet måste företag engagera sig med konsumenterna i den digitala miljön för att uppnå effektiv marknadsföring. Detta innebär att den digitala marknadsföringsaspekten har blivit avgörande för företag och utgör en nödvändighet för att de ska kunna konkurrera och nå ut till kunderna. Digital marknadsföring definieras som olika sätt att övertyga konsumenter på internet att köpa produkter eller tjänster. Med hjälp av moderna teknologier samt integrering av datavetenskapliga metoder inom marknadsföringsbranschen kan företag öka sin effektivitet och lönsamhet [[1].](https://reader.elsevier.com/reader/sd/pii/S2444569X20300329?token=FF9B419B38C1D08CA90B7EE82424DC8677FF12BF7369EFA7BF02347800C67BE25D963E40910D3CD438EAAD626E194B26&originRegion=eu-west-1&originCreation=20230219125858)

För att kunna utföra effektiv marknadsföring behövs det data. Dagligen genereras stora mängder data som företag utnyttjar för att genomföra analyser, formulera hypoteser samt identifiera och lösa problem. Den ökande mängden data som används för analytiska ändamål är så omfattande att dess bearbetning kräver stöd från datorbaserade metoder, eftersom det inte längre är möjligt för människor att utföra analyserna manuellt. Stora mängder av data kallas stordata (på engelska Big Data) och de flesta företagen måste hantera sådana data. Stordata definieras med tre ord: variation, mängd och hastighet d.v.s. data som anländer snabbt, i stor variation och i stor mängd [[]](https://www.oracle.com/big-data/what-is-big-data/).

Företag samlar flitigt på kunddata. Det finns många kanaler genom vilka företag kan samla data på individer. De vanligaste sätten är med transaktioner, frågeformulär, kakor, e-postspårning, applikationer med integrerade SDKar (på eng. Software Development Kit) och genom att köpa data från tredjepartsföretag. Exempelvis används kakor för att identifiera användare på olika webbsidor för att bland annat spåra vilka webbsidor användaren besöker. [[]](https://pages.cs.wisc.edu/~pb/www16_final.pdf)

Utöver dessa kanaler genererar företag data kontinuerligt under verksamhetens gång, såsom information om transaktioner eller marknadsföringskampanjer.

Med den mängd kanaler som finns tillgängliga för datainsamling är det lätt att förstå förekomsten av stordata. Trots att stordata erbjuder många fördelar för verksamheter, finns det även en del utmaningar som stordata medför. Dessa utmaningar inkluderar att skapa tillräckligt med utrymme för all data som genereras samt att rensa datat, det vill säga att bestämma vilket data som är relevant för verksamhetens målsättningar. Mycket av data är även ostrukturerat. Detta innebär främst att det är data som inte är lagrat i en konventionell databasstruktur. Detta kan vara saker som PDF-filer, sensordata, bilder o.s.v. Detta innebär att det inte alltid är så lätt att utföra analyser på stordata. För att enklare hantera stordata samt finna nya insikter, har flera industrier börjat använda maskininlärning för att främja deras verksamhet. En av dessa branscher är marknadsföringsbranschen men även branscher som medicin, agrikultur och finans har blivit mer data-drivna [[]](https://www.kdnuggets.com/2022/08/6-ways-businesses-benefit-machine-learning.html). I marknadsföring används maskininlärning på flera olika sätt. Det används för att skapa rekommendationssystem, utföra kundsegmentering och för att förutspå enskilda kunders köp m.m.

Maskininlärning innebär att vi kan skapa modeller som kan förutsäga olika information när de förses med nya data. Till exempel kan vi skapa en modell som, givet en kunds köphistorik, kommer att förutsäga om denna kund kommer att delta i en kommande marknadsföringskampanj eller inte. Att skapa sådana förutsägningsmodeller kallas träning och det går ut på att ge en modell ett stort antal tränings exempel (t.ex. köphistorik, tidigare marknadsföringskampanj deltagande) från vilka vissa mönster härleds. Dessa mönster kan vara betydelsefulla för analytiker eller inte. Ändå visar de sig vara mycket effektiva inom vetenskapen, teknologin, politiken, för att bara nämna några tillämpliga områden.

I denna avhandling kommer jag att utföra en analys på en datamängd av kunder och olika egenskaper som de har. Jag kommer att försöka svara på följande frågor genom att utforska datat med data visualisering och andra metoder, sedan utvärdera olika maskininlärningsmodeller baserat på hur bra de presterar i enlighet med syftet. Första uppgiften är att försöka förutspå vilka kunder som kommer att reagera positivt till en kampanj. Andra uppgiften kommer att vara kundsegmentering. Syftet med detta är att erbjuda en inblick i hur maskininlärning kan se ut inom marknadsförings branschen, samt att försöka ta reda på vilka maskininlärnings modeller kan användas för marknadsföringssyften.

I denna avhandling kommer vi att behandla följande frågor:

* Hur ser arbetsflödet för ett mindre maskininlärningsprojekt ut?
* Vilka modeller fungerar bäst för att förutspå kundrespons?
* Hur kan vi förbättra framtida kampanjer?
* Hur skiljer sig de olika kundgrupperna?

# **Teoretisk Bakgrund**

## **2.1 Maskininlärning**

För att kunna tolka och analysera de stora mängder data som skapas dagligen har man tvingats hitta metoder som inte längre kräver mänskliga omdömen. Maskininlärning (ML) är en gren av artificiell intelligens (AI) och ett verktyg som man kan använda för att hitta mönster i data och förutspå händelser baserat på historiska data med hjälp av iterativa processer och statistiska algoritmer, även kallade maskininlärningsmodeller.[[]](https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-ml/)

ML skiljer sig från traditionell statistisk analys eftersom ML-modeller kan hitta mönster på egen hand utan att en människa måste moderera mellanstegen. Traditionella metoder utgår från att man ger en direkt indata och får fram en direkt utdata som baserar sig på färdigprogrammerad logik i datorn. ML introducerar dock ett steg mellan indata och utdata, vilket är inlärning. ML använder data för att skapa en modell som är kapabel att fatta beslut (Machine Learning for Absolute Beginners).

Data som man matar in i modellen kommer att påverka modellens användbarhet. Detta innebär att en av de viktigaste aspekterna i maskininlärning är data och dess behandling. Innan man matar in data i modellen måste man utföra datatvätt samt funktionskonstruktion. Detta innebär alltså att man ser till att data är i ett sådant format som en dator kan förstå. Oftast innebär detta att man ändrar alla datapunkter till ett numeriskt format. Funktionskonstruktion är sedan att välja/skapa de data som är relevanta för det man vill åstadkomma. Detta kan åstadkommas med logiska slutsatser eller erfarenhet inom ämnet. T.ex. Om man försöker uppskatta någons ålder genom bildklassificering så är data om personens favoritfärg högst antagligen irrelevant för uppgiften. Om man skulle inkludera favoritfärgsdata i modellens träning skulle det kunna hitta korrelationer mellan ålder och favoritfärg som inte finns i den verkliga världen.

Maskininlärning går ut på att en modell finner mönster och relationer i data genom statistiska metoder, upprepade försök och misstag. Detta kallas att träna maskininlärningsmodellen. Efter träningsskedet ska man testa modellen för att utvärdera prestandan. Detta görs med data som inte används i träningsskedet (valideringsdata) för att justera hyperparametrar och utvärdera modellens prestanda under träningen. Eftersom man använder valideringsdata för att justera hyperparametrarna så måste man även ha en sista testdata som sedan beskriver hur bra modellen presterar på helt ny data.

ML är ett starkt verktyg och används i flera branscher för en mängd olika uppgifter. Några exempel:

* Identifiering av sjukdomar med hjälp av bild-klassificering metoder
* Spam identifiering i e-post
* Rekommendationsmotorer

#### **2.1.1 Övervakad och Oövervakad inlärning**

ML kan delas in i tre breda grenar: Övervakat, Oövervakat och Förstärkningsinlärning. [[]](https://course.elementsofai.com/se/4/1) Jag kommer snabbt att förklara vad Övervakad samt Oövervakad inlärning innebär.

##### **Övervakat inlärning**

Övervakad inlärning är en av de vanligaste typerna av maskininlärning. [[]](https://www.coursera.org/articles/types-of-machine-learning) Modellen tränas med hjälp av ett träningsset av data som redan är klassificerat eller etiketterat. Det innebär att varje datapunkt i träningssetet associeras med en etikett. Till exempel kan en bild klassificeras som antingen en katt eller en hund. Under träningsskedet matas träningsdatan in i modellen, som innehåller både inputdata och de rätta svaren för varje input. På så sätt får modellen informationen den behöver för att kunna göra en förutsägelse och jämföra den med det faktiska svaret. Modellen mäter sedan sin prestanda och justerar sig själv tills felen minimeras. [[]](https://www.ibm.com/topics/supervised-learning#:~:text=the%20next%20step-,What%20is%20supervised%20learning%3F,data%20or%20predict%20outcomes%20accurately.) Modellen skapar en matematisk funktion som representerar relationerna mellan input- och output-data. Denna funktion kan sedan användas på liknande input-data för att göra högprecisionförutsägelser. De två vanligaste metoderna inom övervakad inlärning är klassificering och regression.

###### **Regression**

Regression är en metod inom övervakad inlärning som används för att förutsäga kontinuerliga numeriska värden. Till exempel kan en regression användas för att förutsäga priset på en bostad baserat på dess egenskaper som antal sovrum, kvadratmeter, läge och liknande. En vanlig typ av regression är linjär regression, där modellen försöker hitta en linje som bäst passar givna data genom att minimera avståndet mellan linjen och datapunkterna. När modellen är tränad kan man använda den skapade funktionen för att göra förutsägelser (se figur 1). [[]](https://www.javatpoint.com/regression-vs-classification-in-machine-learning#:~:text=The%20main%20difference%20between%20Regression,Spam%20or%20Not%20Spam%2C%20etc.)

###### **Klassificering**

Klassificering är en metod inom övervakad inlärning som används för att kategorisera input-data i fördefinierade klasser baserat på dess egenskaper. Det är en kraftfull teknik som kan exempelvis användas inom marknadsföring för att klassificera kunder baserat på deras köpbeteende och preferenser. Till exempel kan en kund klassificeras som "tillfällig" eller "lojal" baserat på deras historiska köpmönster. I klassificering måste output-variabeln vara diskret. I nedanstående figur ser man hur en klassificerings beslutsgräns skulle kunna se ut. Exempelvis skulle en kund klassificeras som lojal om hen hamnar på vänstra sidan av gränsen och tillfällig om hen hamnar på högra. [[]](https://www.javatpoint.com/regression-vs-classification-in-machine-learning#:~:text=The%20main%20difference%20between%20Regression,Spam%20or%20Not%20Spam%2C%20etc.)



Figur 1

##### **Oövervakad inlärning**

Oövervakad inlärning är en typ av ML där algoritmer tränas att hitta mönster, strukturer och relationer i data utan förutbestämda etiketter eller svar. I denna inlärningsmetod använder algoritmen endast insamlade data och försöker självständigt identifiera och organisera information, till exempel genom att gruppera liknande datapunkter (klustring) eller reducera data till lägre dimensioner (dimensionalitetsreduktion). Målet med oövervakad inlärning är att hitta underliggande strukturer i data och sedan gruppera dem. Oövervakad inlärning kan delas in i två typers problem: Klustring och Association. [[]](https://www.javatpoint.com/unsupervised-machine-learning)

Jag kommer bara att ta upp klustring eftersom association inte är relevant för avhandlingen.

**Klustring**

Klustring är en ML metod som används flitigt inom marknadsföring. Den används främst för kundsegmentering. Klustring innebär att en modell finner strukturer och relationer mellan datapunkterna, och sedan grupperar dem i en förbestämd mängd grupper (se fig 2).



Figur 2

Klustrings modeller kan delas in i exklusiva, hierarkiska, överlappande och probabilistiska.

*Exklusiv klustring* innebär att datapunkter delas in i disjunkta kluster, där varje datapunkt tillhör ett enda kluster. En vanlig metod är k-means klustering, där antalet kluster bestäms på förhand och datapunkterna placeras i det närmaste klustret. [[]](https://www.ibm.com/topics/unsupervised-learning)

*Hierarkisk klustring* kan vidare delas in två metoder: *Agglomerativ* och *Splittrande*. I analysen kommer jag att använda *Agglomerativ hierarkisk klustring*.

I *Agglomerativ hierarkisk klustring* börjar man med att placera alla datapunkter i separata kluster och sedan successivt sammanfogar de mest liknande klustren tills alla datapunkter ingår i ett eller k-antal kluster (k är förbestämd mängd av kluster). Likheten mellan kluster beräknas med hjälp av distansen mellan datapunkterna.

*Splittrande klustring* är i grund och botten motsatsen till *Agglomerativ klustring*.

I *splittrande klustring* börjar man med att placera alla datapunkter i ett kluster, och sedan successivt splittra på klustret. [[]](https://www.ibm.com/topics/unsupervised-learning)



Agglomerativ och Splittrande klustring (Figur 3

**2.2 ML modeller**

2.2.1 Klassificerings modeller

**Logistisk regression**

Logistisk regression är en övervakad inlärnings ML modell som strävar till att förutsäga en ändlig mängd av resultat, exempelvis 0 eller 1. Modellen använder sig av logistiska funktionen för att kartlägga y som en sigmoid funktion av x. Där y är beroende variabeln och x är oberoende variabeln. En beroende variabel är den variabel som man försöker förutspå, medan oberoende variabler används för att förutspå beroende variabeln. Funktionen producerar en S-formad kurva som kan förvandla alla reella tal till ett värde mellan 0 < y < 1. Detta innebär att modellen transformerar alla oberoende värden till en sannolikhet 0 < y < ,1 där 1 == 100% sannolikhet och 0 == 0% sannolikhet. I binär klassificering kollar man sedan hur nära y är antingen 0 eller 1, där y < 0.5 klassificeras som 0, och y > 0.5 klassificeras som 1. (Machine Learning for absolute beginners)





**Decision Tree Classifier**

Beslutssträd klassificerare är en övervakad klassificerings modell. Ett beslutsträd representerar en sekvens av beslut och deras möjliga utfall i en trädstruktur. Trädet börjar från rot-noden (på eng. Root Node) och där splittras all data i två eller flera grenar. Varje beslutsnod (Decision Node) i trädet representerar ett beslut eller en test på en egenskap som vidare splittrar data, och varje gren representerar ett utfall av det beslutet. Löv-noden (på eng. Leaf Node) i trädet representerar de slutliga besluten eller klasserna.[[]](https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm)



Under träningfasen tittar beslutsträdsmodellen på indata och försöker hitta de viktigaste egenskaperna eller attributen som kan användas för att dela in datan. Detta kan uppnås genom att kolla på vilka variabler som kan splittras för att minska på entropin i nästa nod. Entropi beskriver variansen i data. [ML for absolute beginners]

**Exempel 1:**

|  |  |  |  |
| --- | --- | --- | --- |
| Kunder | Är gift? | Ålder < 40 | Köpt produkt? |
| 4 | 4 | 2 | Ja |
| 2 | 0 | 2 | Nej |

Här har vi 4 datapunkter o vi försöker förutspå beroende variabeln: ”Köpt produkt?” med hjälp av de oberoende variablerna: ”Kunder, Är gift?, Ålder<40”

I diagrammet nedan ser man hur entropin ökar, d.v.s. data blir mindre homogen om man splittrar data med variabeln ”Ålder < 40” jämfört med om man splittrar det med variabeln ”Är gift”:



När modellen är tränad kan den användas för att förutsäga klass för ny indata genom att följa beslutsserien från roten till en lövnod. Den förutsagda klassen är den som är associerat med den lövnod som nås av indata.

**Random Forest Classifier**

Random Forest Classifier är en övervakad ensamble modell. En ensemble-modell är en modell som använder sig av flera svagare modeller, för att sedan skapa en stark modell genom att aggregera resultaten från de svagare modellerna. Modellen fungerar genom att kombinera flera beslutsträd. Modellen börjar med att slumpmässigt välja en delmängd av oberoende variabler från den ursprungliga datamängden. Därefter skapas ett beslutsträd baserat på dessa variabler. Beslutsträdet delar upp datamängden i två eller flera homogena uppsättningar.

Modellen upprepar sedan denna process för flera beslutsträd, varje gång väljer den en annan delmängd av oberoende variabler slumpmässigt. När alla beslutsträden har skapats, gör modellen en förutsägelse genom att ta majoritetsomröstningen från alla träden. Om 80 av 100 träd förutsäger en datapunkt som "Klass A", kommer den slutliga förutsägelsen att vara "Klass A". [[]](https://towardsdatascience.com/understanding-random-forest-58381e0602d2)



**Gradient Boosting Klassificerare**

Gradient Boosting Klassificerare är också en övervakade ensemble ML modell. Modellen använder sig av Boosting. Boosting innebär att modellen initialiseras med endast en modell som gör förutsägelser på data. Denna första modell kommer inte att vara noggrann med förutsägelserna, men det är för att få en start på algoritmen. Sedan tränas det en ny modell som fokuserar på att minimera felet i den förra modellen o.s.v. tills man har en modell som är bra.

Gradient Boosting Består av tre huvudkomponenter:

**Förlustsfunktion**

Förlustsfunktionen är ett sätt att mäta hur mycket förutspådda värden skiljer sig från de egentliga värden i träningsskedet. Så i varje iteration räknas felet med hjälp av denna funktion

**Svaga inlärningsmodeller**

**Additiva Modeller**

2.2.2 Klustrings Modeller

# **Metod**

## **3.1 Dataförberedelse**

### **3.1.1 Datainsamling**

Datainsamling definieras som processen av att samla och mäta information på specifika variabler, som t.ex. inkomst av en person. Datainsamling för marknadsföring innebär att man samlar data på kampanjer, kunder, projekt, konkurrenter. För att man skall kunna arbeta med samlade data, måste data vara lagrad på en gemensam plats. I kontexten av ett företag finns det 4 olika typers data som de samlar.

1. Personlig data (IP Adress, Kön, cookies)
2. Beteende data (Hur ofta du köper, vad du köper)
3. Engagemangsdata (Hur du engagerar med företagets hemsida, e-post etc.)
4. Attityd data (Hur eftertraktad en produkt är, hur nöjda är kunder)[[]](https://research.aimultiple.com/data-collection-for-marketing/)

För maskininlärning så är optimala data balanserade, saknar inga datapunkter, har inga duplikat och är relevanta. I verkligheten så är data sällan perfekt och man måste utföra dataförberedelse. Men i datainsamlings skedet är det viktigt att man har pålitliga datakällor. [[]](https://www.simplilearn.com/tutorials/machine-learning-tutorial/machine-learning-steps)

### **3.1.2 Datarengöring**

Efter man har samlat data som man vill använda för att träna sin modell, måste man först se till att data är användbart för vidare analys. Datarengöring innebär främst att man korrigerar eller tar bort irrelevanta, inkorrekta, felformaterade, duplicerade datapunkter. Eftersom sådana datapunkter kan påverka modellens prestanda. Detta kan innebära att man formaterar strängar av datum till ett date-time format eller att man tar bort rader av data där det saknas datapunkter.

Det är viktigt att man tar i beaktande att rengörings metoderna bör anpassas till de data man har. [[]](https://www.tableau.com/learn/articles/what-is-data-cleaning#:~:text=Data%20cleaning%20is%20the%20process,to%20be%20duplicated%20or%20mislabeled.)

## **3.1.3 Datatransformering**

Datatransformering innebär att man transformerar data så att en modell kan förstå data. Modeller tar oftast bara in data som är numerisk. Exempelvis alla variabler som innehåller strängar måste man representera i ett numeriskt format. Om strängarna är kategorier kan man tilldela varje kategori en unik siffra. Man kan även använda One-hot encoding.

**One-hot encoding**

Denna metod innebär att man tilldelar varje kategori inom en kolumn en egen kolumn som kan enbart ha binära datapunkter. Exempelvis om vi har en kolumn ”Land” som kan ha sträng värden ”FI”,”SE”, ”NO”. Så kan vi tilldela varje land en egen kolumn där en 1 betonar att personen kommer från det landet medan en 0 betonar att en person inte kommer från det landet.

### **3.1.3 Datavisualisering**

Data visualisering är en central aspekt i ett ML projekt. Genom att visualisera data kan man lättare upptäcka mönster, trender och anomalier, vilket ger en bättre grund för att bygga och förbättra modellerna. I figurerna nedan kan vi se exempel på hur det skulle kunna se ut om det existerar extrempunkter i data.

 

**Korrelationsmatris**

En korrelationsmatris är en tabell som visar korrelationen mellan olika variabler i en datamängd. Korrelationen mäter hur starkt två variabler är relaterade till varandra, och kan variera från -1 (fullständig negativ korrelation) till +1 (fullständig positiv korrelation). En korrelationsmatris kan ge insikter i vilka oberoende variabler har en högre påverkan på den beroende variabeln. [[]](https://corporatefinanceinstitute.com/resources/excel/correlation-matrix/)



**Förvirringsmatris**

En förvirringsmatris (på eng. Confusion Matrix) är en tabell som används för att utvärdera prestandan hos en klassificeringsmodell genom att visa antalet korrekta och felaktiga förutsägelser av modellen. Förvirringsmatrisen består av fyra huvudkomponenter: True Positive (TP), False Positive (FP), False Negative (FN) och True Negative (TN). [[]](https://www.sciencedirect.com/topics/engineering/confusion-matrix) Där True och False beskriver om gissiningen var rätt, och positiv och negativ beskriver hurdan gissning det var.



### **3.1.3 Skalning**

I maskininlärning syftar skalning (på eng. Scaling) på att normalisera variabler i en dataset till en jämn skala. Detta görs för att undvika att en variabel dominerar en annan på grund av dess skala och påverkar modellens resultat. Vanliga skalningsmetoder inkluderar min-maxskalning, standardisering och normalisering.

**Min-maxskalning** är en skalningsmetod i maskininlärning som normaliserar variabler i en dataset till en fördefinierad intervall, vanligtvis mellan 0 och 1. Formeln för min-max skalning är:

Skalad\_värde = (värde - minvärde) / (maxvärde - minvärde)

Där "värde" är det ursprungliga värdet, "min\_värde" och "max\_value" är det minsta respektive största värdet för variabeln i datasetet, och "scaled\_value" är det nya normaliserade värdet.

### **3.1.4 Översampling**

Översampling är en metod inom maskininlärning som används för att hantera obalanserade dataset. Obalanserade dataset innebär att en klass är underrepresenterad i data. Översampling innebär att man skapar fler observationer i den underrepresenterade klassen genom att duplicera befintliga observationer eller genom att syntetisera nya observationer. Översampling kan hjälpa till att förbättra prestandan för maskininlärningsmodeller genom att minska risken för överanpassning och förbättra generaliseringsförmågan. [[]](https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/)

## **3.2 Träning**

## **3.3 Prestanda mätning**

#

# **Empirisk Studie**

I denna analys har jag försökt identifiera vilken ML modell som presterar bäst för att förutspå hur en kund kommer att reagera på en kampanj. Utöver detta kommer analysen att innehålla en beskrivning av data, analysresultat, samt möjliga förbättringar av processen. Hela projektet är gjort i Python

**Beskrivning av datamängden**

Denna studie baseras på en datamängd som hämtats från kaggle.com och innehåller information om kunder och deras köpbeteende. Datamängden består av flera attribut som beskriver varje kund och deras köpvanor t.ex. så finns det kolumner som beskriver hur mycket en kund har spenderat på köttprodukter eller hur de har reagerat på en viss kampanj. Utöver köpbeteende relaterade attribut, finns det även data om deras ålder, hur många barn de har, vilket land de kommer ifrån m.m. Totalt har databasen 28 kolumner och 2241 rader. Från dessa kolumner vill vi förutspå kolumnen ”Response”. Kolumnen ”Response” är binär, där en 1 representerar att en kund har accepterat den senast kampanjen, medan 0 innebär att en kund har ignorerat den. ”Response” kolumnen är alltså den beroende variabeln och de resterande kolumnerna är de oberoende variablerna. Datamängden var väldigt strukturerad redan från början, men jag måste ändå utföra vissa åtgärder för att preparera den för träningen av modellerna.

Några saker som åtgärdades:

* Null-värden togs bort
* ”Response” kolumnen var väldigt obalanserad, vilket krävde översampling.
* Det fanns kategoriskt data i vissa kolumner såsom i kolumnen ”Country” vilket hade länderna i strängar. Detta löstes med ”One-hot-encoding”.
* Extremvärden i kolumnerna ”Age”,”Income”,”Meat”,”NumWebPurchases” och ”NumCatalogPurchases” togs bort.
* Skalning av data med hjälp av min-maxskalning för att normalisera alla värden.

Utöver dessa åtgärder fanns det många små ändringar samt transformeringar av datamängden.

4.2 Analyser och resultat

4.2.1 Målgruppsanalys

4.2.2 Segmentering

4.3 Diskussion

# **Slutsats**