LUNDS UNIVERSITET
Ekonomihögskolan

Nationalekonomiska institutionen

Klassificering av kreditkortskunder

En prestationsjämförelse mellan logistisk regressionsanalys och random forest.

Kandidatuppsats 15hp, kurs NEKH01 i Ekonometri

Författare: Carl Sandelius

Handledare: Peter Jochumzen

Framlagd: Maj 2022

Sammanfattning: Mängden data som genereras av olika aktörer ökar ständig och det brukar sägas att många organisationer är rika på data men samtidigt informationsfattiga. Behovet av varierande analysverktyg för att skapa insikt och beslutsunderlag har aldrig varit större, men för varierande uppgifter och data passar olika metoder mer eller mindre bra. Uppsatsen undersöker, beskriver och förklarar ett perspektiv på hur den ekonometriska modellen logistisk regressionsanalys presterar i förhållande till maskininlärningsmodellen random forest för binär klassificering av kreditkortskunders kreditvärdighet. Resultatet visar att random forest genererar en högre AUC än logistisk regression och presterar bättre i 4 av de 5 prestationsmåtten.

1 INNEHÅLLSFÖRTECKNING

EKVATIONER

FIGURER

TABELLER

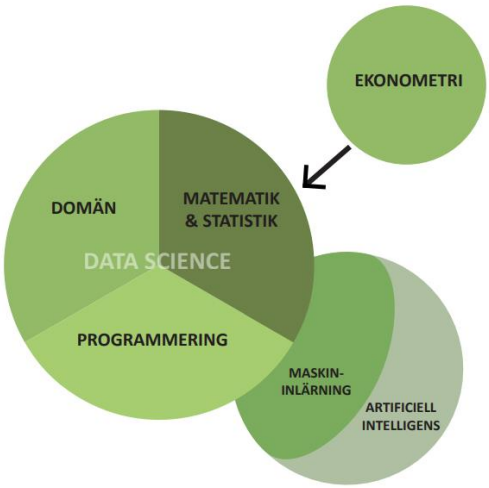
2 INTRODUKTION

Initialt presenterar kapitlet bakgrund och problemområde för uppsatsen, därefter följer frågeställning och syfte. Avslutande beskrivs de avgränsningar som gjorts och en begreppslista presenteras.

2.1 BAKGRUND

AI (artificiell intelligens), ML (maskininlärning) och de olika modeller som förekommer inom området har det senaste decenniet varit ett ämne som adresserats utifrån olika perspektiv av författare som till exempel Nick Bostrom (2014), Max Tegmark (2017) och David Sumpter (2019). Oaktat om det berört utvecklandet av dagens specifika AI mot morgondagens generella, de etiska aspekterna kopplat till detta, eller de underliggande framstegen som möjliggjort användandet av olika algoritmer, så beskriver författarna att användandet av modeller kopplat till klassificering, prediktion och slutligen beslutsfattande som områden som troligen kommer att fortsätta att integreras inom en rad olika delar av samhället.

Enligt Microsoft (2022) och Bell (2020, s. 3) så utgör ML ett centralt område inom AI och vidare beskriver Grus (2019, s. 153) att området är en väsentlig komponent inom data science. Grus (2019, s. xv-xvi) menar i sin tur att data science är ett tvärvetenskapligt ämne som återfinns i interceptet mellan statistik, matematik och programmering med ett stort mått av specifik domänvetande kring till exempel ett aktuellt affärsområde.



Figur 1 Områdesassociation, sammanställning baserad på Microsoft (2022), Grus (2019, s. xv-xvi), Grus (2019, s. 153) och Angrist (2021, *4:47)*

Dougherty (2016, s. 1) beskriver ekonometri som tillämpningen av statistiska metoder för kvantitativa utvärderingar av hypotetiska relationer baserat på data. Författaren betonar samtidigt att tillämpningen av ekonometri kan ske inom en brett spektra av vetenskapliga områden och inte är isolerat till den ekonomiska domänen även om dess namn kan antyda detta. Visserligen har merparten av den forskning och utveckling som historiskt bidragit till området sitt ursprung inom det ekonomiska fältet (Dougherty, 2016, s. 1) men som Kelleher, Mac Namee & D'Arcy (2020, s. 338–346) visar, nyttjas ekonometriska verktyg som till exempel logistisk regressionsanalys, som en möjlig maskininlärningsmodell. Angrist uttrycker den nära kopplingen mellan ekonometri, data science och i förlängningen maskininlärning i en intervju som:

*"Econometrics is the original data scien*6ce,*
before there where data science there where econometrics"
(Angrist, 2021, 4:47)

2.2 Problemområde

McAfee et al. (2012) argumenterar för att den allt större mängd data (big data) som genereras av olika aktörer i samhället bör gå hand i hand med ett datadrivet beslutsfattande för att säkerställa att den aktuella verksamheten bedrivs på ett konkurrenskraftigt sätt. Detta perspektiv stärks av Kubina, Varmus & Kubinova (2015) samtidigt som de betonar att det är essentiellt för verksamheter att fokusera på vad de processar för typ av data och när detta sker för att kunna säkerställa sin position på marknaden.

SAS Institute (2022) beskriver att hanteringen av big data utgör en komplex uppgift för organisationer när volym, variation och hastigheten för den genererade eller inhämtade datan är hög. Datavolymen härstammar från varierande källsystem och sensorer och kan vara av olika format och därigenom uppträda som strukturerad, semistrukturerad eller ostrukturerad (SAS Institute, 2022). IBM (2021) framställer strukturerad data genom den typ av atomisk data som normalt hanteras i en relationsdatabas (tabellbaserad) och ger samtidigt uttryck för att semistrukturerad data ofta har sitt ursprung i webservices (kommunikation mellan system) via filformat som JSON (JavaScript Object Notation) eller XML (Extensible Markup Language). Ostrukturerad data genereras till exempel via filer innehållande löpande text, ljud eller video IBM (2021).

Edwards (2018) menar att den enorma mängd data som genererats under de senaste decennierna i praktisk mening saknar värde utan rätt analysverktyg. IBM (2022) har en liknande syn och beskriver att bearbetning av big data sker via s.k. BDA (Big Data Analytics) där maskininlärning, via tillämpningen av olika modeller, utgör en viktig komponent för att möjliggöra analys av större datamängder. Angrist har gett uttryck för att maskininlärning gett upphov till nya metodologiska approacher inom ekonometri men att ML-området har lite att bidra med till ekonometriska studier där dataunderlaget är relativt litet (Angrist, 2019, 0:25). Detta stödjs även av Haldar (2015), som säger att datamängden vid maskininlärning bör vara av en viss storlek för att säkerställa användbar output, men betonar samtidigt att vad denna storlek exakt är beror på det aktuella problemet, kvaliteten på tillgänglig data och vilken modell som skall tillämpas.

Wolpert (1996) och Wolpert & Macready (2005) har via sitt ”No Free Lunch Theorem” framfört att det inte existerar någon inlärningsalgoritm som universellt fungerar optimalt för samtliga uppgifter. De beskriver att olika uppgifter eller problem kommer med sina specifika egenheter, resultatet blir att varierande inlärningsmodeller kan prestera olika bra på skilda problem. Edwards (2018) menar att baserat på den mängd modeller som kan användas för varierande eller lika syften i samband med maskininlärning, så blir det nödvändigt att jämföra prestationen för dessa för att kunna utvärdera vilken approach som är mest lämplig för ett givet problem. Detta perspektiv stärks av Kelleher, Mac Namee & D’Arcy (2020, s. 15–17) och Hastie, Tibshirani & Friedman (2009, s. 222) som har en liknande uppfattning.

2.3 Frågeställning

Hur presterar logistisk regressionsanalys som binär klassificeringsalgoritm jämfört med random forest?

2.4 Syfte och målsättning

Uppsatsen syftar till att undersöka, beskriva och förklara hur logistisk regressionsanalys presterar i förhållande till maskininlärningsmodellen random forest för binär klassificering av kreditkortskunders kreditvärdighet. Genom detta blir uppsatsens primära målsättning att dels bidra med ett perspektiv på hur två klassificeringsmodeller kan jämföras, men även att

undersöka hur maskininlärningsmodellen random forest skulle kunna fungera som ett kompletterande verktyg till den ekonometriska verktygslådan.

Vad som även kan tilläggas är att datasetet i sig avser att prediktera vilka kreditkortskunder som löper högre risk för förfall och i förlängningen hur potentiella förluster kan undvikas för olika kreditinstitutioner. Data från Federal Reserve (2022) visar att antalet förfallna kreditkortsräkningar i USA ökade under de år som föregick pandemin. Under pandemin stannade utvecklingen av, men av allt att döma verkar trenden nu återigen vara uppgående. Med andra ord verkar problematiken återkomma och vara korrelerad med ökad konsumtion. Därigenom blir en sekundär målsättning att uppsatsen även skall kunna bidra med ett möjligt tillvägagångssätt för vidare kunna arbeta med den här typen av prediktioner.

2.5 AVGRÄNSNINGAR

Eftersom uppsatsen rör sig i området mellan ekonometri, data science och maskininlärning där aktuella modeller (logistisk regressionsanalys och random forest) utifrån olika perspektiv kan klassificeras till en eller flera kategorier, kommer uppsatsen inte att fokusera på eller utreda AI (dit maskinlärning ofta räknas) utan använda befintliga definitioner. Detta för att området är för stort för att ignoreras men samtidigt inte ryms inom ramarna för uppsatsen.

Gällande området maskininlärning så kommer uppsatsen uteslutande att vara fokuserad mot övervakad maskininlärning via klassificeringsmodellerna logistisk regressionsanalys och random forest. Därmed kommer inte oövervakad, semiövervakad eller förstärkningsorienterad maskininlärning eller underliggande modeller att utredas. Inte heller kommer tekniker som djupinlärning, som kan tillämpas inom samtliga nämnda områden att avhandlas. Motiveringen till detta grundar sig i att de aktuella modellerna förutsätter att det existerar en målvariabel (beroende variabel) och att datan är strukturerad, vilket medför att avgränsningen kan göras till övervakad inlärning.

Eftersom uppsatsen undersöker två klassificeringsalgoritmer, kommer de utvärderingsmått som lyfts vara avgränsade till de som är aktuella för detta ändamål.

Uppsatsen kommer inte att i detalj beskriva hur underliggande algoritmer till inlärningsmodellerna fungerar eller konstruerats, utan begränsa sig till att ge en konceptuell bild över dess funktion. Detta då uppsatsen inte förutsätter kunskaper inom programmering,

pseudokod eller liknande. För den intresserad kan Iterative Dichotomizer 3 (ID3) gällande random forest och dokumentationen för sklearn.linear_model.LogisticRegression avseende logistisk regression ge ett fördjupat perspektiv.

Uppsatsen är även avgränsad till att nyttja ett strukturerat dataset. Datasetet är bestående av två delar om respektive 1 048 575 kreditansökningar och 438 557 kreditupplysningar som benchmark för de bägge klassificeringsmodellerna.

2.6 BEGREPPSLISTA

Tabell 1 Begreppslista

Begrepp	Beskrivning
Artificiell intelligens (specifik/smal AI)	I uppsatsen avser termen s.k. specifik (smal) AI och definieras enligt Agrawal, Gans & Golfrabs (2019) som en mjukvara med förmåga att fatta beslut baserat på dataanalys om predikterade utfall.
Data eller dataset	I uppsatsen avser begreppet uppmärkt strukturerad (rubricerad) data som följer normalformerna för databaser, såväl kategorisk som numerisk.
Maskininlärning eller data mining	Avser processen att identifiera mönster från data. Notera att i viss litteratur används begreppen maskininlärning och data mining som synonymer, medan data mining enligt andra källor ofta avser subprocesser för datainsamling (till exempel web scraping).
Modell eller maskininlärningsmodell	Avser i uppsatsen en inlärningsalgoritm.
Observation	Motsvarar en unik rad i en relationsdatabas (tabellbaserad databas).

2.7 ÖVRIGT

RMP-processen har medvetet inte lagts till som bilaga då kopiering av XML-formatet från pdf kan medföra att även dolda tecken från dokumentet inkluderas och korrumperar filen. Om processen/testet vill reproduceras så skickas nedladdningslänk vid förfrågan.

3 LITTERATURGENOMGÅNG

Kapitlet presenterar initialt en överblick över området oövervakad maskininlärning. Därefter följer ett praktiskt perspektiv på datakvalitet och datapreparering, samt ett stycke om två centrala problem kopplat till detta. Vidare redogörs för maskininlärning utifrån ett ekonometriskt perspektiv följt av varsitt stycke behandlande de bägge modellerna. Slutligen följer en redogörelse för en rad utvärderingsmått som möjliggör prestationsjämförelsen.

Avseende terminologin, så har den engelska benämningen nyttjats i de fall där det saknas en svensk motsvarighet eller där det bedömts underlätta för läsaren.

3.1 EN ÖVERBLICK PÅ ÖVERVAKAD MASKINSKININLÄRNING

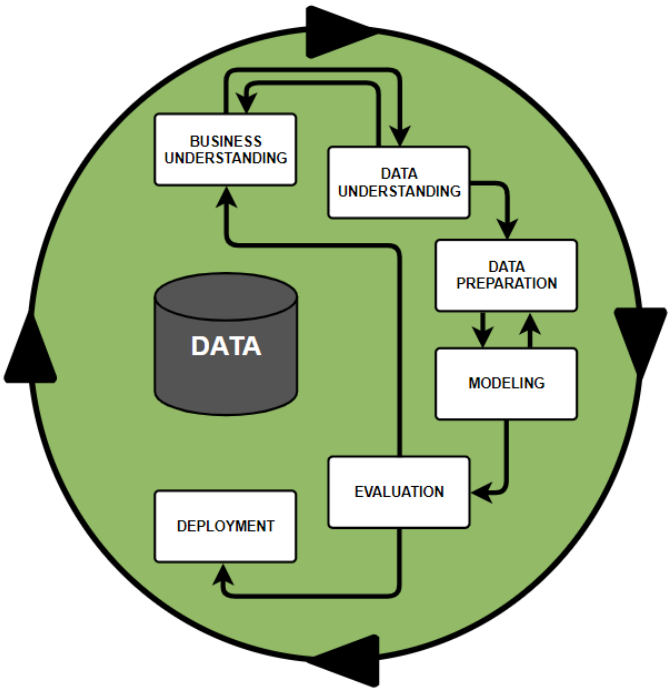
Kelleher, Mac Namee & D'Arcy (2020, s. 598) definierar övervakad maskininlärning (ML) som en automatiserad process för att extrahera mönster från data baserat på historiska observationer. Bell (2020, s. 3) nyttjar en liknande definition men tillägger att ML är en gren inom AI som möjliggör för att utveckla och träna modeller för prediktiva uppgifter. Annorlunda uttryckt, så kan maskininlärning betraktas som ett verktyg för att omvandla information till kunskap genom att identifiera annars svårupptäckta mönster och samband inom datan, vilket vidare kan nyttjas som beslutsunderlag (Edwards, 2018).

För att konstruera prediktiva modeller beskriver Kelleher, Mac Namee & D'Arcy (2020, s. 5–7) att nyttjandet av deskriptiva variabler för att predicera en målvariabel är en central del. I praktisk mening kan det jämföras med terminologin inom ekonometrin, dvs. hur oberoende variabler används för att bestämma en beroende variabel inom regressionsanalys (Dougherty, 2016, s. 85).

Enligt Chollet (2021, s. 13–18) så har de mjukvaruverktyg som traditionellt använts för dataanalys varit uteslutande regelbaserade, till exempel om (a == b) returnera c, till skillnad mot mjukvarusystem som idag nyttjar maskininlärningsmodeller där inlärningsalgoritmen medför att de underliggande sambanden snarare identifieras och därigenom kan ligga till grund för en dynamisk approach gentemot det aktuella problemet, men också anpassas när dataunderlaget gör så (Chollet, 2021, s. 20–24). Edwards (2018) nämner att det är just mjukvarans förmåga till utvärdering och optimering som givit upphov till benämningen maskininlärning.

3.1.1 CRISP-DM som ramverk

Kelleher, Mac Namee & D'Arcy (2020, s. 15–17) lyfter att ett generellt ramverk för att använda maskininlärning som metod är via processchemat CRISP-DM (Cross Industry Process for Data Mining). Författarna beskriver att processen är strukturerad i sex underliggande steg där det första avser förståelse för den aktuella verksamheten och analysen som modellen skall generera (se figur 2). Kelleher, Mac Namee & D'Arcy (2020, s. 15–17) och Saltz (2021) betonar att CRISP-DM utgör det mest populära och vida nyttjade ramverket för data science (maskininlärning).



Figur 2 CRISP-DM, efter Kelleher, Mac Namee & D'Arcy (2020, s. 16)

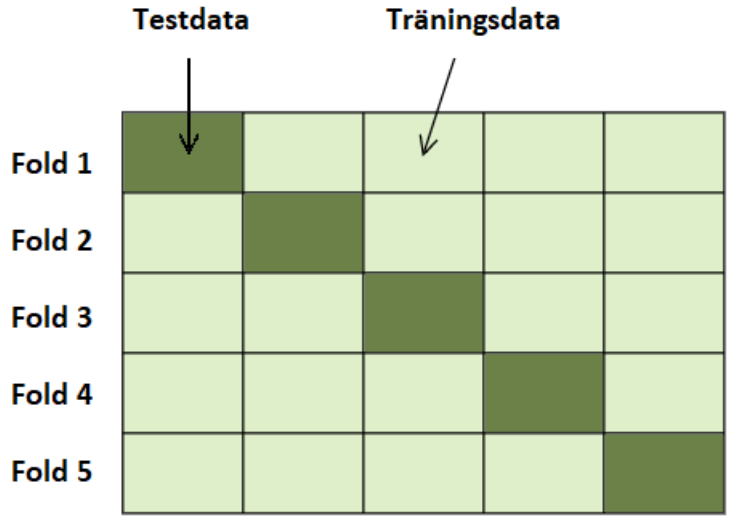
Därefter beskriver Kelleher, Mac Namee & D'Arcy (2020, s. 15–17) att förståelse för den tillgängliga datan, såväl som vad som eventuellt kan behövas inhämtas för att uppnå det föregående följer som steg två i processen.

Det tredje steget utgörs av preparering av datan. Här nämns EDA (Exploratory Data Analysis) för visualisering över till exempel distributioner hos variabler och deskriptiv statistik över desamma som en komponent som kommer ligga till grund för en datakvalitetsrapport. Denna rapport bör delvis innehålla informationen från EDA:n men även lista specifika kvalitetsproblem som noterats hos en deskriptiv variabel med potentiella åtgärdsstrategier

(Kelleher, Mac Namee & D'Arcy, 2020, s. 53–54). Ett exempel som nämns är när en binär deskriptiv variabel uppvisar en kardinalitet som avviker från 2 (kvalitetsproblem). Är den lägre, är det troligen klokt att ta bort variabeln helt (åtgärdsstrategi) då den inte bidrar med någon information. Är kardinaliteten däremot högre, bör de avvikande värdena kontrolleras, kanske har olika beteckningar använts för samma input vid datainsamlingen (till exempel kvinna eller k för det kvinnliga könet osv.). Här kan en värdeomvandling från k till kvinna hos berörda observationer fungera som en åtgärdsstrategi. Kelleher, Mac Namee & D'Arcy (2020, s. 15–17) beskriver att den färdiga datapreparereingen utmynnar i en s.k. ABT (Analytical Base Table) som innehåller det färdiga datasetet som sedan kan användas för inlärningsalgoritmen.

Det fjärde steget i processen beskrivs av Kelleher, Mac Namee & D'Arcy (2020, s. 15–17) som det som avser att identifiera ett antal alternativa maskininlärningsmodeller. Här kan uppsatsen fungera som ett exempel. Eftersom ett klassificeringsproblem föreligger, datasetet är av strukturerad typ och det finns möjlighet att konstruera (eller urskilja en beroende variabel) blir det möjligt att nyttja logistisk regressionsanalys och random forest som modeller.

Kelleher, Mac Namee & D'Arcy (2020, s. 15–17) framhåller att det femte steget avser utvärdering av de alternativa modellerna och slutligen ett modellval. För att möjliggöra jämförelse, men även konstruktionen av en maskininlärningsmodell som generaliserar bra på nya data, utvecklas normalt modellen på befintliga data och avhängande till storleken på denna delas den ofta i proportionerna 70/30 eller 80/20 för träning och testning (Kelleher, Mac Namee & D'Arcy, 2020, s. 535–540; Bronshtein, 2017; Hastie, Tibshirani & Friedman, 2009, s. 222–223). Detta görs för att möjliggöra en utvärdering av modellens prestation där utvärderingsunderlaget inte utgår från träningsdatan (Kelleher, Mac Namee & D'Arcy 2020, s. 535–540). Vidare beskrivs hur k-folded cross validation (där till exempel k=5) kan användas för att säkerställa utfallet från modellen (se figur 3). I praktisk mening separeras datan i fem grupper och fem iterationer görs. För varje iteration varieras vilken grupp som utgör testdata och viken som fungerar som träningsdatan. Efter korsvalideringen kan utfallet erhållas som varierande output och som ett medelvärde över iterationerna (Kelleher, Mac Namee & D'Arcy 2020, s. 543–545).



Figur 3 5-folded cross validation, efter Kelleher, Mac Namee & D'Arcy (2020, s. 543)

Det sjätte och avslutande steget avser introducerandet av modellen inom verksamhetens processer. I likhet med andra mjukvarusystem så inträder det nu i en typisk drift- och underhållscykel med eventuell vidareutveckling och integrering (Kelleher, Mac Namee & D'Arcy 2020, s. 15–17).

3.2 En praktisk vinkel på datakvalitet och datapreparering

För att generera datakvalitetsrapporten, som ligger till grund för datapreparereingen och slutligen det färdiga analytiska bastabellen (datasetet), bör två aspekter täckas in (Kelleher, Mac Namee & D'Arcy, 2020, s. 53–54). Först behövs egenskaperna för samtliga variabler i den obearbetade datan undersökas och dokumenteras på ett strukturerat sätt. Varje värdetyp som en variabel kan hålla, dess intervall och distribution blir centrala att utreda. Den andra aspekten som lyfts är identifieringen av eventuella kvalitetsproblem inom datan som kan komma att påverka prestationen för den färdiga modellen. Författarna beskriver typiska problem som saknad data, avvikande kardinalitet och förekomsten av extremvärden som vanligt förekommande kvalitetsproblem (Kelleher, Mac Namee & D'Arcy, 2020, s. 63). Avseende identifierade kvalitetsproblemen, kan dessa i sin tur kategoriseras som grundande på antingen icke-valid eller valida data. Problem kopplade till icke-valida data härstammar ofta från felaktig datainsamling eller notering av denna och beroende på omfattningen så kan olika hanteringsstrategier nyttjas skriver författarna. Om det, till exempel, noteras förekomma

saknade värden för en given deskriptiv variabel, skulle denna problematik kunna hanteras genom att de observationer (rader) innehållande de saknande värdena helt enkelt raderas. Kelleher, Mac Namee & D'Arcy (2020, s. 69–72) beskriver att detta kan vara en godtagbar lösning om antalet berörda observationer är förhållandevis få. Samtidigt understryker de att exkludering leder till en större informationsförlust och kan introducera bias i datan under antagandet att de saknade värdena inte uppträder helt randomiserat. En alternativ lösning på kvalitetsproblemet är att substituera in ett estimat för det saknade värdet för varje observation. För kontinuerliga värden kan till exempel median eller medelvärde nyttjas och för kategoriska data kan den mest frekvent förekommande värdet (mode) för variabeln användas (Kelleher, Mac Namee & D'Arcy 2020, s. 69–72). Den här typen av lösning kan dock påverka datasetets centrala tendens om en allt för hög andel saknade värden hanteras på detta sätt. Författarna nämner att då den totala andelen saknade värden för en deskriptiv variabel är >60% bör variabeln helt exkluderas från datasetet eftersom informationen som erhålls är av tveksam betydelse, men även att substituering inte bör användas för en större andel än maximalt 30% (Kelleher, Mac Namee & D'Arcy, 2020, s. 69–72).

Gällande kvalitetsproblem kopplat till valida data så kan hanteringen av extremvärden tas som exempel genom användandet av s.k. clamp transformation, vilket avser definitionen av övre- och lägre tröskel för vilka observationer som skall inkluderas (Kelleher, Mac Namee & D'Arcy 2020, s. 69–72). Definitionen för dessa är ofta baserad på 1.5 * kvartilavståndet från den första och tredje kvartilen för den aktuella variabeln. Värden som faller inom intervallet inkluderas och de utanför exkluderas, genom detta kvarstår inte längre problematiken med extremvärdena (Kelleher, Mac Namee & D'Arcy, 2020, s. 69–72). Alternativt så kan tröskeln baseras på z-score (Mishra, 2021) där ett värde på ca ± 2.698 skulle motsvara samma intervall som ovan för en normalfördelning.

I de fall där det aktuella datasetet innehåller kontinuerliga data med stor diversitet inom variablernas respektive intervall kan normalisering av dessa bidra till att underlätta för olika inlärningsalgoritmer (Kelleher, Mac Namee & D'Arcy, 2020, s. 87–88; Bell 2020, s. 191). Under antagandet om att eventuella extremvärden har hanterats, kan range normalization nämnas som en möjlig metod. Visserligen är metoden känslig för närvaron av extremvärden, men i kombination med en föregående clamp transformation kan den generera en god normalisering. Kelleher, Mac Namee & D'Arcy (2020, s. 87–88) beskriver att ett alternativt

tillvägagångssätt för att normalisera datan är via z-transformation, vilket avser genererandet av ett värde som representerar antalet standardavvikelser som observationen befinner sig från medelvärdet för den givna variabeln (z-score).

Ytterligare ett problemområde som kan uppträda i samband med datapreppareringen är behovet av att konvertera kontinuerliga data till kategorisk eller vice versa. Kelleher, Mac Namee & D'Arcy (2020, s. 89) och Kumar (2020) beskriver *binning* som en potentiell metod för att hantera omvandlingen från kontinuerliga till kategoriska data. Detta sker genom att intervallet delas upp i sub-intervall, s.k. bins. Dessa kan antingen vara baserade på lika bredd eller på samma frekvens. Lika bredd beskrivs helt enkelt som att det ursprungliga intervallet delas i ett förbestämt antal, med lika intervallbredd (Kelleher, Mac Namee & D'Arcy, 2020, s. 89; Kumar, 2020). Samma frekvens syftar till att antalet observationer i varje ny bin är den dikterande faktorn för avgränsningen för de nya sub-intervallen. Popov (2019) lyfter fram One-hot-incoding som en av de mest använda metoderna för att konvertera kategoriska data till numeriska. I praktisk mening fungerar metoden genom att introducera en s.k. dummy variabel som tar värdet 1 eller 0 beroende på huruvida villkoret är uppfyllt eller ej (Dougherty, 2016, s. 230; Popov, 2019). Alternativt kan varje kategorisk variabel även representeras av ett numeriskt värde för att möjliggöra kategoriseringen.

Kelleher, Mac Namee & D'Arcy (2020, s. 91–94) beskriver att det beroende på storleken på datasetet (ABT), i kombination med tillgänglig processorkapacitet, kan det bli aktuellt att nyttja sampling för att hantera annars mycket krävande uppgifter. Samtidigt understryker de att detta måste ske på ett sådant sätt att datan fortfarande är representativ gentemot den ursprungliga och att ingen bias introduceras. Författarna föreslår användandet av s.k. random sampling, vilket de menar troligen inte medför denna problematik.

En annan aspekt av sampling som Alto (2021) beskriver är att den kan nyttjas för att lösa problematik kopplat till obalanserade data för den beroende variabeln genom s.k. under- eller oversampling. Vidare framför densamma att det i samband med klassificering eller prediktion av fenomen som är mindre frekvent förekommande (eller rent av sällsynta) så är ofta fallet att datan som ligger till grund för datasetet har en inbyggd skevhet i sin distribution genom den låga förekomsten av utfallet. Vidare skulle nyttjandet av datan medföra att inlärningsmodellen i sig självt kommer att lära sig att prestera i enlighet med denna bias, för att därigenom optimera sitt resultat (Alto, 2021). Genom undersampling av den mest frekvent

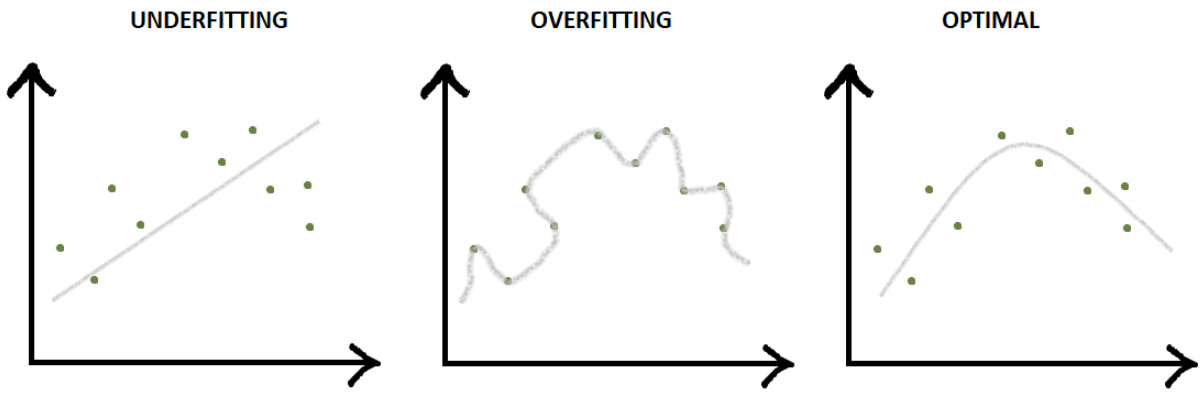
förekommande utfallet, så nyttjas en mot den minst frekvent förekommande utfallet proportionell mängd observationer. Med andra ord tas enbart en viss randomiserad mängd av dessa observationer med. Därigenom menar Alto (2021) att en modell kan tränas utan att introducera biasen från den ursprungliga datan, detta perspektiv stödjs även av Kelleher, Mac Namee & D'Arcy (2020, s. 91–94). Nackdelen med tillvägagångssättet som beskrivs är att den resterande datamängden blir avsevärt mindre (Alto, 2021). Alternativt, så kan oversampling av minoritetsklassen göras för att adressera problemet (Alto, 2021). Detta kan ske genom att använda random sampling med ersättning inom den minst frekvent förekommande utfallsgruppen till dess att önskad proportion har erhållits (Alto, 2021; Kelleher, Mac Namee & D'Arcy, 2020, s. 91–94). Nackdelen som lyfts är att duplicerad data kommer förekomma och inkluderas inom modellutvecklingen (Alto, 2021).

3.3 Två eventuella problem att beakta

Som Wolpert (1996) och Wolpert & Macready (2005) beskrivit så förekommer det inte någon modell som generellt kan tillämpas och fungera optimalt för olika uppgifter, utan specifika egenheter hos datan i kombination med de induktiva bias som varje modell inkodar kommer medföra att skilda inlärningsmodeller prestera olika bra. Kelleher, Mac Namee & D'Arcy (2020, s. 11–12) beskriver en modells induktiva bias som de antaganden som inlärningsalgoritmen gör beträffande den underliggande datan. Genom detta menar de att det inte initialt går att veta vilken eventuell inlärningsmodell som matchar problemet bäst utan en jämförelse är nödvändig.

Kelleher, Mac Namee & D'Arcy (2020, s. 13–15) skriver att det förekommer två typer av fel som ett resultat av induktiv bias, d.v.s. att en maskininlärningsmodell presterar på ett dysfunktionellt sätt. Antingen kan felet beskrivas som underfitting eller overfitting. Grus (2019, s. 155–157) förklarar overfitting som en modell som presterar bra på träningsdatan men generaliserar dåligt på testdata (se figur 4). Kelleher, Mac Namee & D'Arcy (2020, s. 13–15) beskriver att overfitting uppträder då inlärningsalgoritmen modellerar det verkliga sambandet allt för komplext och matchar datan perfekt (se figur 4). Detta innebär att modellen kommer att bli känslig för avvikande data och kommer få problem med att hantera ny. Detta perspektiv stärks även av von Luxburg & Schoelkopf (2008). Vidare beskriver Grus (2019, s. 155–157) att underfitting avser en modell som varken presterar bra på tränings- eller

testdata. Underfitting uppträder som ett resultat av att den nyttjade inlärningsmodellen porträtterar det underliggande sambandet inom datan på ett allt för förenklat sätt (Kelleher, Mac Namee & D'Arcy, 2020, s. 13–15; von Luxburg & Schoelkopf, 2008).



Figur 4 Underfitting, overfitting & optimal, efter Kelleher, Mac Namee & D'Arcy (2020, s. 15)

Al-Masri (2019) skriver att en modell som generaliserar bra, d.v.s. dess förmåga att generera rimliga outputdata baserat på input som den inte tidigare har hanterat, är en modell som varken lider av underfitting eller overfitting.

Problematiken med underfitting och overfitting beskriver von Luxburg & Schoelkopf (2008) avhänger till bias/variansdilemmat. En modells oförmåga att uttrycka det underliggande sambandet definierar von Luxburg & Schoelkopf (2008) som dess bias och variansen härleds utifrån skillnaden mellan hur väl modellen presterar på tränings- och testdata. Gällande overfitting beskrivs problematiken medföra låg bias då den kan matchar träningsdatan mycket bra, men med en hög varians genom att generalisera dåligt på testdatan. Samtidigt innebär underfitting en hög bias genom att inte kunna beskriva den underliggande träningsdatan bra, men med låg varians då avvikelserna gentemot testdatan består (von Luxburg & Schoelkopf, 2008). Al-Masri (2019) uttrycker att den ideala modellen är en med låg bias och låg varians, men att utformningen för att konstruera en modell som presterar önskvärt blir en trade-off mellan bias och varians. Denna uppfattning delas även av Grus (2021, s. 160).

3.4 Ett ekonometriskt perspektiv på maskininlärning

Geweke, Horowitz & Pesaran (2008) beskriver, i likhet med Dougherty (2016), ekonometri som tillämpningen av statistiska och matematiska metoder för att ge ett empiriskt stöd för ekonomiska samband och Sunil (2019) sammanfattar ekonometri som en samling verktyg för att åstadkomma detta. Alam (2020) uttrycker att många av de statistiska och matematiska tekniker som nyttjas inom ekonometrin för att modellera ekonomiska system, också är verktyg som är tillämpningsbara på problem inom data science och maskininlärning (till exempel linjär- och logistisk regression, K-means, ARIMA etc.).

I likhet med maskininlärning så nyttjar bägge områdena data, matematik och statistisk för att komma fram till sina slutsatser men en tydlig skillnad som Wrg (2019) lyfter är att det inom ekonometrin ofta står ett orsakssamband i centrum där utvärderingen är baserad på ett hypotestest. Detta är en aspekt som maskininlärning inte täcker in eftersom inlärningsalgoritmer i sitt utformande, tenderar att utvärderas mot dess prediktiva prestation (Kelleher, Mac Namee & D'Arcy 2020, s. 585).

Shmueli (2010) föreslår i sin artikel en distinktion mellan prediktiv- och förklaringsorienterad modellering. Författaren menar att prediktiv modellering utgörs av en process där data i kombination med en statistisk modell nyttjas för att till exempel prediktera eller klassificera. Förklarande modellering beskrivs som tillämpningen av en statistisk process för att bestämma ett orsakssamband (Shmueli, 2010). Varian (2014) framhåller att dessa tangerande områden (ekonometri och maskininlärning) erbjuder varandra aspekter som kan inkorporeras i respektive verktygslådor. Utifrån ett ekonometriskt perspektiv nämner författaren bland annat korsvalidering, användandet av modeller som till exempel random forest och neurala nätverk, bagging och boosting mm. Från ett maskininlärningsperspektiv lyfter Varian (2014) bland annat orsakssamband, naturliga- och explicita experiment mm.

3.5 Hur fungerar logistisk regressionsanalys som maskininlärningsmodell?

Följande stycke avser att förklara hur logistisk regressionsanalys fungerar och hur den kan nyttjas som maskininlärningsmodell. Initialt ges en kort överblick, därefter undersöks hur koefficienter kan tas fram och hur maximum likelihood används inom modellen och slutligen hur en variant på R^2 kan erhållas.

3.5.1 En överblick på logiskregressionsanalys

Kelleher, Mac Namee & D'Arcy (2020, s. 311–312) nämner att inom ML-området klassificeras logistisk regressionsanalys, likt andra regressionsmodeller till gruppen felbaserade inlärningsmodeller. Dougherty (2016, s. 369–377) beskriver att logistisk regressionsanalys (även kallad binary choice models) bygger på samma princip som den vanliga regressionsanalysen men med främst två grundläggande skillnader. Den första avser nyttjandet av maximum likelihood i stället för least squares för att minimera the cost function. Den andra skillnaden är introducerandet och användandet av en sigmoid-funktion (S-formad och även kallad logistiks funktion) i stället för till exempel en linjär funktion. Annorlunda uttryckt så predikterar logistisk regression en input till en av två alternativa output (till exempel 1 eller 0), till skillnad från linjär regression som predikterar en kontinuerlig output Dougherty (2016, s. 369–377). Vad som även bör nämnas är att när det förekommer mer än en oberoende variabel så erhålls en yta snarare än en linje för sigmoid-funktionen (Starmer, 2022, s. 108–118).

Starmer (2022, s. 108–118) beskriver att själva proceduren för att finna the line of best fit (eller the best fit) gentemot observationerna är skilda mellan linjär- och logistisk regression. För linjär regression nyttjas least squares*,* d.v.s. den slutliga trendlinjen korresponderar mot de koefficienter som minimerar summan av residualerna i kvadrat. Samtidigt används residualerna för att kalkylera R^2 och därigenom avgöra hur väl de oberoende variablerna kan beskriva den beroende (Starmer, 2022, s. 108–118; Grus, 2021, s. 185–188). Vidare nämner Starmer (2022, s. 108–118) att det beräknade p-värdet avgör om de oberoende variablerna är statistiskt signifikanta eller ej och via trenden är det möjligt att prediktera ett utfall för given input för den linjära regressionen.

Logistisk regression beskrivs sakna konceptet residual och kan därigenom inte nyttja least squares eller beräkna R^2 på samma sätt som linjär regression, i stället används maximum likelihood för att finna the best fit för sigmoid-funktionen gentemot datan (Starmer, 2022 s. 108–118; Grus, 2021, s. 203–214).

3.5.2 Logistisk regressionsanalys, maximum likelihood och koefficienter

Dougherty (2016, s. 369–377) beskriver att för logistisk regression så är den beroende variabelns intervall begränsat till sannolikhetsvärden mellan [0, 1] men för att möjliggöra för att undersöka de underliggande koefficienterna behöver modellen transformeras så att den

likt en linjär regressionsanalys kan sträcka sig mellan (-∞, +∞). Detta görs via logit-funktionen enligt Starmer (2022, s. 108–118).

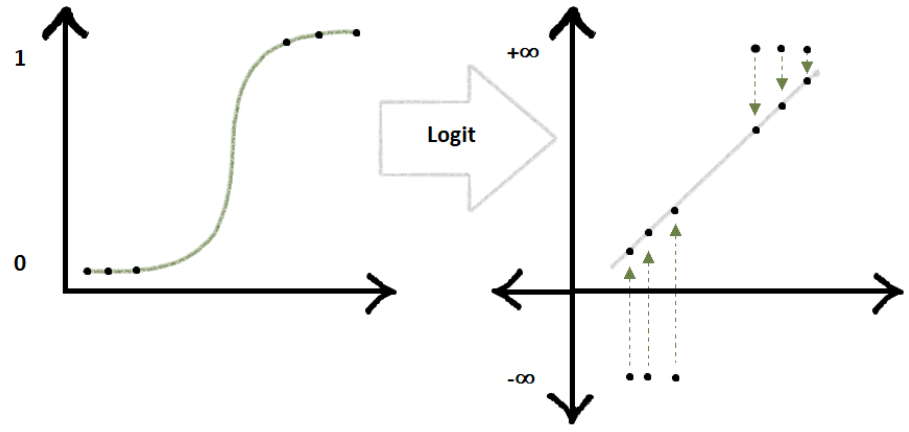
$$Logit(probability) = \log\left(\frac{probability}{1 - probability}\right)$$

Ekvation 1 Logit (Hastie, Tibshirani & Friedman, 2009, s. 121)

Starmer (2022, s. 108–118) lyfter att ett problem som uppstår i samband med detta är att observationerna (som tidigare binärt klassificerats till antingen 1 eller 0) nu blir lika med antingen (-∞) eller (+∞). Genom detta menar författaren att det inte längre är möjligt att nyttja *least squares* för att finna the best fit eftersom residualerna, till vilken slumpmässig linje som helst, också blir lika med (-∞) eller (+∞). Detta beskriver han kan hanteras genom att tillämpa maximum likelihood och därigenom passa in sigmoid-funktionens placering. I praktisk mening så innebär detta att observationerna först tillfälligt måste överföras till en slumpmässig kandidatlinje i ett nytt koordinatsystem för att vidare möjliggöra att omvandla dem tillbaka till probabilities (Starmer, 2022, s. 108–118). En tydlig definition av detta ger Bhattacharyya (2018):

$$Probability = \left(\frac{e^{Logit(probability)}}{1 + e^{Logit(probability)}}\right)$$

Ekvation 2 Probability (Bhattacharyya, 2018)



Figur 5 Sigmoid-kurva med illustration över logit-funktions transformering enligt Starmer (2022, s. 108–118)

Detta korresponderar mot en viss y-koordinat på sigmoid-funktionen (se figur 5), vilket kommer medföra att en observations *likelihood* kan beräknas enligt (Starmer, 2022, s. 108–118):

$$Likelihood = 1 - Probability$$

Ekvation 3 Likelihood (Starmer 2022, s. 116)

Starmer (2022, s. 108–118) beskriver att därefter beräknas produkten av samtliga observationers likelihood och kandidatlinjen kan på nytt slumpas (rotera lutningen). Detta genomförs iterativt till dess att en sigmoid-funktion som maximerar produkten av samtliga observationers likelihood identifierats (alternativt den som maximerar summan av log(likelihood)). Denna motsvarar maximum likelihood och därigenom har sigmoid-funktionen passats in till att motsvara the line of best fit för modellen (Starmer, 2022, s. 108–118).

Genom den optimerade kandidatlinjen kan nu koefficienter, intercept, fel osv. återfinnas för den underliggande datan till modellen (Starmer, 2022, s. 108–118). Vidare menar Starmer (2022, s. 108–118) att samma koncept som till exempel t-test för hypotestestning nu kan tillämpas på den omformaformaterade modellen. Vad som bör noteras är att koefficienterna är desamma som för en linjär regression, det som skiljer dem är att y-axeln för det nya koordinatsystemet är skalad enligt logit-funktionen (Starmer, 2022, s. 108–118).

3.5.3 Logistisk regression och dess R^2

Precis som i fallet med linjär regression så behövs ett sätt för att bestämma huruvida en viss modell är användbar eller inte, med andra ord behövs ett mått på hur väl vi kan beskriva en beroende variabel via de oberoende. Som Starmer (2022, s. 108–118) lyfter fram så förekommer ingen konsensus kring exakt hur detta skall göras för logistisk regression. Ett av de mer frekvent använda sätten är via McFaddens R^2 och beskrivs enligt (Bartlett, 2014):

$$R^2_{McFadden} = 1 - \left(\frac{\log\left(maximum\ likelihood\right)}{\log\left(null\ model\right)}\right)$$

Ekvation 4 McFaddens R^2 (Bartlett, 2014)

Bartlett (2014) beskriver the null model som en modell innehållande inga oberoende variabler utan enbart ett intercept.

3.6 VAD ÄR RANDOM FOREST OCH HUR FUNGERAR MODELLEN?

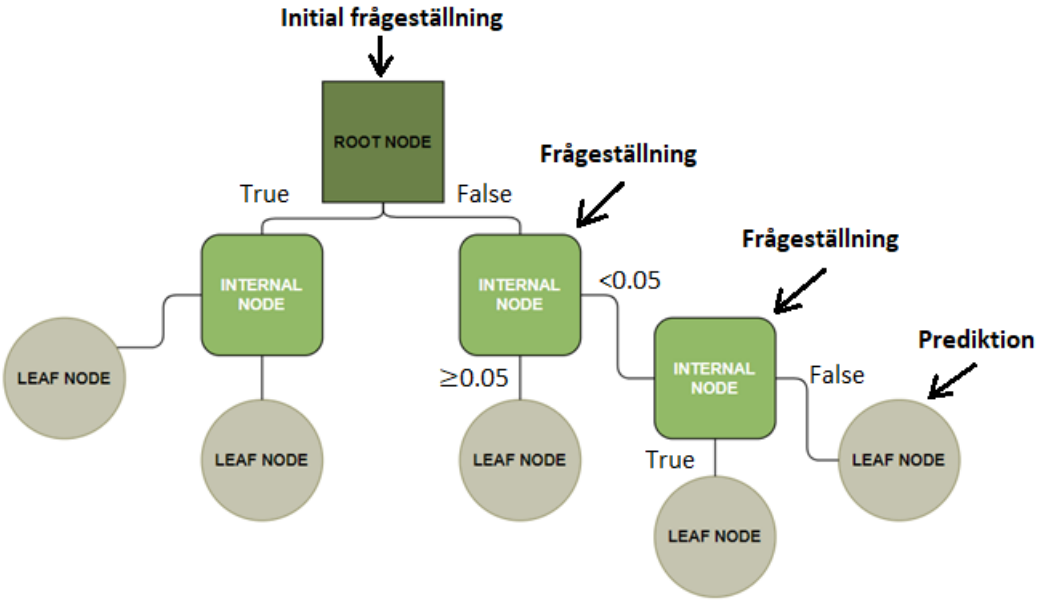
Följande stycke kommer att förmedla en samlad bild kring random forest. Inledande ges en kort överblick av maskininlärningsmodellen. Därefter beskrivs decision trees, följt av entropi och informationsvinst. Slutligen ges ett mer fördjupat perspektiv på random forest och hur dess olika komponenter samverkar.

3.6.1 En kort sammanfattning av random forest

Kelleher, Mac Namee & D'Arcy (2020, s. 158–162) beskriver random forest som en ensemblemodell som i praktiken består av ett antal underliggande modeller i form av decision trees och mot aggregationen av deras output genererar modellen sin prediktion. Med andra ord så bildar flera decision trees en forest och dess gemensamma prediktion kan till exempel baseras på ett majoritetsbeslut. Författarna förklarar att inom informationsbaserad inlärning så utgör decision trees en fundamental struktur för att beräkna entropi och informations gain och därigenom möjliggöra förgreningen inom ett decision tree på ett optimalt sätt (Kelleher, Mac Namee & D'Arcy, 2020, s. 117).

3.6.2 Decision trees

Kelleher, Mac Namee & D'Arcy (2020, s. 120–123) skriver att den grundläggande idén bakom decision trees är att det via noder (frågeställning) och en grenar (villkor) styr informationsflödet (d.v.s. den väg algoritmen tar) på ett optimerat sätt för att nå en slutsats. Ett decision tree kan liknas vid ett upp-och-nervänt träd där den översta noden (root node) förmedlar den initiala frågeställningen (se figur 6). Därefter förgrenas den vidare till nya noder (internal nodes) för att efter ett antal nya noder nå en slutlig punkt (leaf node) och genom denna har klassificeringen skett (Kelleher, Mac Namee & D'Arcy, 2020, s. 120–123).



Figur 6 Exemplifiering av ett Decision tree, efter Grus (2020, s. 216)

Starmer (2022, s. 185) skriver att beroende på vilken typ av information som en visst *decision tree* behandlar benämns de olika. Om datasetet innehåller en eller flera kategoriska variabler kallas de classification tree och ifall samtliga variabler är av numerisk typ beskrivs den som ett regression tree (Starmer, 2022, s. 185). Annorlunda uttryckt så behandlar classification trees både kategoriska- och numeriska data till skillnad från regression trees som enbart håller frågeställningar kopplat till numeriska data.

3.6.3 Entropi och informationsvinst

Kelleher, Mac Namee & D'Arcy (2020, s. 123–127) förklarar Shannon's entropy model genom att likna entropin vid den osäkerhet som finns associerad till att gissa utfallet för ett slumpmässigt val från en viss mängd möjliga utfall. Om mängden består av samma utfall blir den korresponderande entropin obefintlig. Det råder ingen osäkerhet kopplat till vilket utfall som blir valt, givet att samtliga utfall har samma sannolikhet att bli vald. Ifall mängden består av unika utfall blir däremot den associerade entropin mycket hög. Detta kan exemplifieras genom att vi tänker oss en påse med stenar. Vi vet att det finns både vita och grå stenar i påsen, men vi vet också att antalet vita stenar är betydligt fler till antalet än de grå. Om vi sedan slumpmässigt tar upp en sten ur påsen kan vi således anta att vi troligen kommer att ta

upp en vit sten. Det rådde nästan ingen osäkerhet kring utfallet (låg entropi). Om alla stenar i stället hade haft olika färg (givet att antalet stenar är >2) hade entropin, alltså osäkerheten kring utfallet, istället varit hög.

Vidare beskriver författarna att då detta kombineras med sannolikheten för ett visst utfall så möjliggör det beräkningen för ett värde för entropin inom en viss informationsmängd. Hög sannolikhet medför låg entropi och vice versa (Kelleher, Mac Namee & D'Arcy, 2020, s. 123–127). Shannon's entropy model beskriver en viktad summa av logaritmen för sannolikheten kopplat till ett visst utfall enligt (Kelleher, Mac Namee & D'Arcy, 2020, s. 125):

$$Entropy = -\sum_{i=1}^{l}(P(t=i) \times log_2(P(t=i)))$$

Ekvation 5 Entropy (Kelleher, Mac Namee & D'Arcy, 2020, s. 125)

P(t=i) avser sannolikheten för utfallet t är av en viss typ i, l motsvarar antalet olika typer av utfall och $\log_2$ syftar till att erhålla binär output i bits (Kelleher, Mac Namee & D'Arcy, 2020, s. 125–126). Samtidigt uttrycker de att outputen representerar heterogeniteten eller den orenhet som finns inom den aktuella datan.

I samband med att ett decision tree skall konstrueras, uppträder frågeställningar rörande vilken data som en viss nod ska hålla och hur många förgreningar trädet skall bära (Grus 2021, s. 215–225). Svaret avhänger till hur information gain kan maximeras vid varje nod (delning av datan) (Kelleher, Mac Namee & D'Arcy, 2020, s. 127–132). Delningen är baserad på en sekvens av tester där samtliga potentiella leaf nodes heterogenitet undersöks. Informationsvinsten fungerar som ett mått på minskningen av den samlade entropin för datasetet genom att testa för en viss oberoende variabel (Kelleher, Mac Namee & D'Arcy 2020, s. 127–132). Syftet är att erhålla leaf nodes som perfekt diskriminerar mellan de alternativa utfallen för en viss frågeställning. Annorlunda uttryckt, så innebär en högre heterogenitet att vi har en lägre osäkerhet (entropi) då vi har fler vita stenar och helst enbart vita stenar i en viss påse (leaf node). Därför kommer algoritmen initialt, för den första noden (root node), att utvärdera informationsvinsten för varje oberoende variabel som delningskriterium. Den som medför högst informationsvinst kommer därefter att väljas. Därefter, i nästa nod (internal node), kommer algoritmen på nytt beräkna informationsvisten

för samtliga kvarvarande oberoende variabler och en ny delning sker. Detta fortgår till dess att vi förhoppningsvis erhållit hög heterogenitet och kan särskilja mellan olikfärgade stenar med hög säkerhet.

Vad som avslutande bör nämnas är att det förekommer andra alternativ för att åstadkomma delningen men principen är ungefär densamma, till exempel kan gini-impurity användas för samma syfte (Starmer, 2020 s. 190).

3.6.4 Mer om random forest

Nyttjandet av random forest och inte uteslutande ett decision tree grundar sig i att den sistnämnda kan uppvisa en tendens till felaktiga prediktioner gentemot testdata men mycket goda prediktioner på träningsdatan (låg bias men hög varians, s.k. overfitting), vilket man kommer runt genom aggregerandet av flera decision trees till en random forest (Hastie, Tibshirani & Friedman, 2009, s. 587–589). Konstruktionen av en random forest är baserad mot bagging vilket avser att ett nytt slumpmässigt dataset, utifrån den ursprungliga datan, genereras för varje decision tree inom ensemblen (Hastie, Tibshirani & Friedman, 2009, s. 282–283). Dessa slumpmässiga dataset är av samma storlek som det ursprungliga och skapas genom ett randomiserat urval där varje vald observation ersätts med en kopia av sig själv innan nästa iteration sker. Detta medför att det slumpmässiga datasetet kan innehålla flera identiska observationer medan det ursprungliga datasetet förblir intakt (Hastie, Tibshirani & Friedman, 2009, s. 282–283).

Vidare kan s.k. subspace sampling tillämpas, vilket innebär att ett randomiserat urval av oberoende variabler inkluderas i datasetet för varje decision tree (Ho, 1998). Syftet med detta är att nyttja den känslighet som decision trees har för små variationer i den underliggande datan. Därigenom kommer olika decision trees separera datan vid olika noder och förgreningen inom de olika träden kommer ske på ett varierande sätt (Ho, 1998). Dess prediktioner vägs samman inom ensemblen genom en röstningsmekanism, till exempel genom majoritet eller median beroende på vilken informationstyp som behandlas (Yiu, 2019).

Den diversitet som erhålls genom tillvägagångssättet medför att ensemblen generaliserar ny input effektivt genom att de olika underliggande modellerna är relativt okorrelerade (Ho, 1998; Yiu, 2019). Yiu (2019) liknar detta med hur en portfölj innehållande lågt korrelerade tillgångar erhåller en lägre risk än summan av respektive värdepapper. I det här fallet erhåller

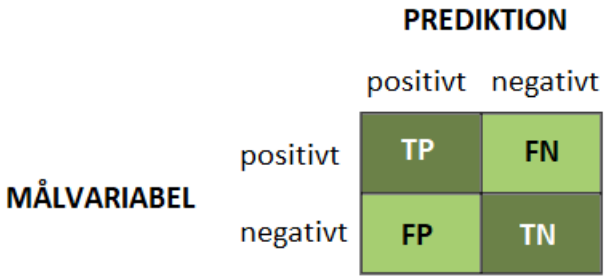
ensemblen en betydligt lägre varians än summan av dess modeller. Den övergripande idén kallas wisdom of the crowd och syftar till att minska variansen mellan tränings- och testdata genom antagandet att diversiteten mellan de olika träden bättre fångar in och klassificerar ny data samtidigt som biasen ökar då samtliga är baserade på en variant av samma underliggande data. Gruppens samlade bedömning bli helt enkelt mer precis på okänd data än en enskild experts.

Gällande hyperparametrarna för random forest, d.v.s. till exempel antalet decision trees och antalet oberoende variabler som inkluderas vid subspace sampling så förekommer det enligt Yiu (2019) inget givet tillvägagångssätt, utan den slutliga finjusteringen sker genom try-and-error approach där olika konfigureringar testas.

3.7 HUR KAN MODELLERS PRESTATION JÄMFÖRAS?

Kelleher, Mac Namee & D'Arcy (2020, s. 534–535) uttrycker att den främsta komponenten vid design av inlärningsmodell är hur dess prestation skall mätas och utvärderas. Samtidigt definieras den centrala frågeställningen i samband med detta som *"huruvida den aktuella modellen faktisk kan utföra det jobb som den utvecklats för?"* (Kelleher, Mac Namee & D'Arcy, 2020, s. 534). Författarna beskriver vidare syftet med utvärderingen som trefaldig, där man initialt jämför de aktuella modellerna gentemot varandra och avgör vilken som kan prestera bäst. Därefter bör man estimera hur väl modellen kommer att prestera när den implementeras inom en process och slutligen skall detta presenteras och motiveras mot den implementerande parten inom verksamheten (Kelleher, Mac Namee & D'Arcy, 2020, s. 534–535).

Fawcett (2006) beskriver användandet av confusion matrix (även kallat error matrix) som ett grundläggande verktyg för att förenkla visualiseringen av den faktiska klassificeringen mot den predikterade klassen för beroende variabel (se figur 7).



Figur 7 Confusion matrix efter Kelleher, Mac Namee & D'Arcy (2020, s. 537)

Fawcett (2006) understryker samtidigt att en confusion matrix utgör den grund som flertalet andra prediktionsmått är baserade på. Detta perspektiv delar även Minaee (2019) som förklarar att de diagonala elementen true positives (TP) och true negatives (TN), sett från TP till TN, utgör de korrekta prediktionerna. De resterande två, false positives (FP) och false negatives (FN), svarar för felaktigheter. Dougherty (2016, s. 39) beskriver FP som type I errors, vilket innebär att modellen felaktigt predikterar positivt medan det faktiska utfallet är negativt. Vidare beskrivs FN som type II errors och medför att inlärningsmodellen felaktigt predikterar negativt medan det faktiska datan visar positivt (Dougherty, 2016, s. 39).

Minaee (2019) beskriver accuracy som ett enkelt men mycket användbart mått, det definieras av Fawcett (2006) som antalet korrekta prediktioner dividerat med det totala antalet prediktioner enligt:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

Ekvation 6 Accuracy (Fawcett, 2006)

I scenarion där vi har obalanserade data för den oberoende variabeln och detta inte har adresserats menar Minaee (2019) att accuracy som mått blir missvisande. Vidare lyfter författaren att om så är fallet så kan måttet precision nyttjas då det är ett klasspecifikt mått. Kelleher, Mac Namee & D'Arcy (2020, s. 548–552) uttrycker detta som att precision specificerar hur ofta en modell gör en positiv prediktion. Fawcett (2006) definierar precision som:

$$Precision = \frac{TP}{TP + FP}$$

Ekvation 7 Precision (Kelleher, Mac Namee & D'Arcy 2020, s. 549)

Kelleher, Mac Namee & D'Arcy (2020, s. 548–552) beskriver recall som ett mått på den säkerhet som kan tillskrivas att en modell identifierat samtliga positiva utfall. Både precision och recall antar ett värde i intervallet [0, 1], högre värde indikerar en bättre prestation från modellen (Kelleher, Mac Namee & D'Arcy, 2020, s. 548–552).

$$Recall\ = \frac{TP}{TP + FN}$$

Ekvation 8 Recall (Kelleher, Mac Namee & D'Arcy 2020, s. 549)

Vidare beskriver Kelleher, Mac Namee & D'Arcy (2020, s. 548–552) att precision and recall kan inkorporeras till ett gemensamt prestationsmått, nämligen F1 score (även kallat F-measure eller F-score). Detta medför ett möjligt sätt att uttrycka andelen missklassificeringar som en modell ger upphov till. F1 score kan uttryckas som (Kelleher, Mac Namee & D'Arcy 2020, s. 550):

$$F1\ score = 2\ \times \frac{precision \times recall}{precision + recall}$$

Ekvation 9 F1 score (Kelleher, Mac Namee & D'Arcy, 2020, s. 550)

Minaee (2019) uttrycker att det alltid förekommer en trade-off mellan precision och recall för en modell. Med andra ord kommer det alltid förekomma en differens mellan hur ofta en modell gör en positiv prediktion kontra den säkerhet modellen erbjuder för att samtliga positiva utfall identifierats (Kelleher, Mac Namee & D'Arcy, 2020, s. 548–552).

ROC-kurva (Reciver Operating Characteristics curve) avser en tvådimensionell graf som förmedlar prestationen hos en eller flera binära klassificeringsalgoritmer (Fawcett, 2006). ROC-kurvan förmedlar förhållandet mellan true positive rate (TPR) och false positive rate (FPR), alternativt mellan sensitivity och specificity, vilka definieras enligt följande (Fawcett, 2006):

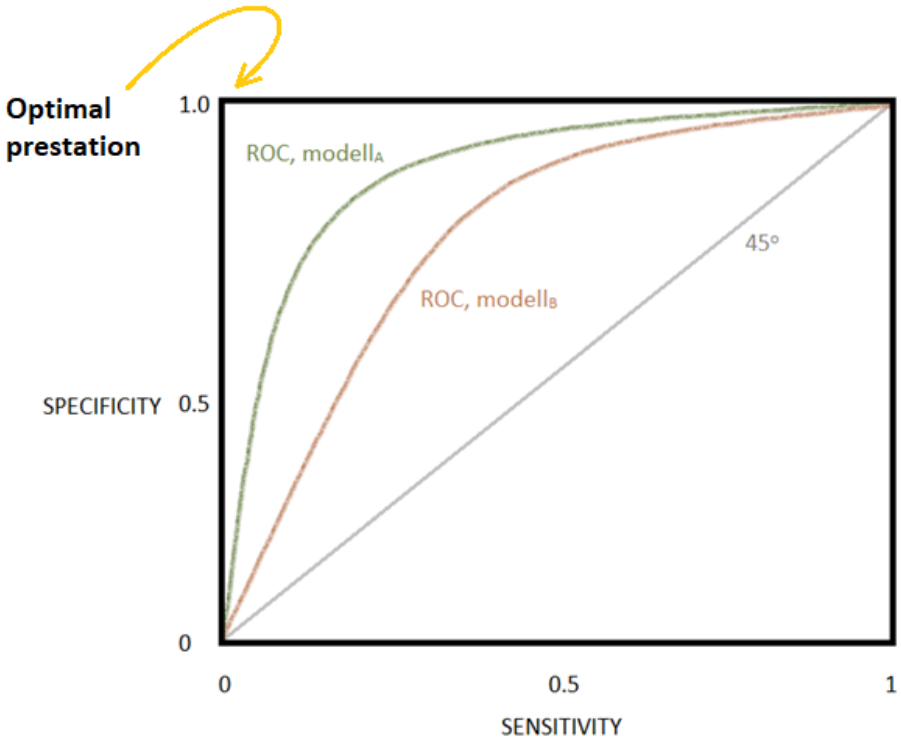
$$Sensitivity = \ Recall = TPR \ = \frac{TP}{TP + FN}$$

Ekvation 10 Sensitivity (Fawcett, 2006)

$$Specificity = TNR \ = \frac{TN}{TN + FP}$$

Ekvation 11 Specificity (Fawcett, 2006)

TRP placeras på x-axeln och TRP plottas på y-axeln (Fawcett, 2006). Minaee (2019) framhåller att flertalet klassificeringsalgoritmer är baserade på sannolikhetslära och i praktiken så jämförs den aktuella outputen gentemot tröskel (ett givet värde) för att åstadkomma själva prediktionen. Författaren beskriver att ROC-kurvan undersöker och uttrycker TRP och FPR för olika trösklar och presenterar detta grafiskt (se figur 8). Fawcett (2006) skriver att olika punkter inom ROC-kurvan representerar olika utfall. Till exempel lyfter författaren att punkten korresponderande mot koordinaterna (0, 0) som representerande av att en tröskel som medför att en modell aldrig klassificerar positivt utfall och då verken TP eller FP. Vad som bör noteras gällande ROC-kurvan, som Fawcett (2006) understryker, är att punkten i det övre vänstra hörnet (0,1) korresponderar mot en perfekt klassificering av inlärningsmodellen och att samtliga punkter till nordväst är att betrakta som bättre än till sydöst. Detta grundar sig i att en abstrakt 45^{o} diagonal linje från origo representerar ett slumpmässigt utfall, d.v.s. en gissning om 50/50 (Fawcett, 2006).



Figur 8 ROC, efter Kelleher, Mac Namee & D'Arcy (2020, s. 562)

AUC (Area Under the Curve) eller ROC-index avser den yta, i intervallet (0, 1), som uppstår under en viss ROC-kurva (Kelleher, Mac Namee & D'Arcy, 2020, s. 558–564). Storleken på denna medför att varje modells prestation kan jämföras och eftersom det översta vänstra hörnet i grafen motsvarar en optimal klassificering, går det därigenom att avgöra vilken modell som presterar förhållandevis bäst och bör tillämpas (Kelleher, Mac Namee & D'Arcy, 2020, s. 558–564; Fawcett 2006). Annorlunda uttryckt, desto närmre en yta är arean = 1, desto bättre (Minaee, 2019). Kelleher, Mac Namee & D'Arcy (2020, s. 562) föreslår att trapezoidal metoden kan nyttjas för att enkelt beräkna ROC-kurvans integraler enligt:

$$AUC = \sum_{i=2}^{|T|} \frac{FPR(T[i]) - FPR(T[i-1]) \times TPR(T[i]) - TPR(T[i-1])}{2}$$

Ekvation 12 AUC (Kelleher, Mac Namee & D'Arcy, 2020, s. 562)

4 METOD

Följande kapitel kommer initialt förmedla en motivering till metodval, därefter kommer arbetsprocessens olika deluppgifter som lett fram till resultatet beskrivas. Deluppgifterna är; dataförståelse, preparering av datan, konstruktion av modellerna samt utvärderingsmått. Slutligen följer två stycken, ett om etik kopplat till datan som nyttjats och ett gällande reliabilitet och replikerbarhet för uppsatsens resultat.

4.1 METODVAL

Metoden för uppsatsen har baserats på Kelleher, Mac Namee & D'Arcy (2020, s. 15–17) beskrivning av CRISP-DM ramverket (Cross Industry Process for Data Mining). I den här kontexten syftar inte data mining till processen för att extrahera rådata från varierande källsystem, utan till maskininlärning som arbetssätt. Som det har nämnts i litteraturgenomgången (se sid. 12) så är ramverket uppbyggt av sex underliggande komponenter, den första (förståelse för verksamheten och den färdiga modellens funktion inom denna) och sista (implementering av modellen i verksamhetsprocesser) blir utifrån uppsatsens sammanhang inte aktuella att gå djupare in i, men kan nämnas för att tydliggöra funktionssättet gentemot de övriga delarna. Det kan dock förtydligas att det genom data från Federal Reserve (2022) finns en idé om i vilket sammanhang som den här typen av klassificeringsproblem kan tänkas vara intressant, nämligen för till exempel olika kreditinstitutioner som hypotetiskt kan antas ha ett behov av ett helt eller delvis automatiserat beslutsstöd vid kreditgivning.

Gällande den avslutande komponenten inom CRISP-DM ramverket, så förekommer det inte någon implementeringsfas för de modeller som konstruerats. Förvisso är modellerna, som skript, färdigt för att tillämpas på ny data men detta ligger utanför uppsatsen.

Med andra ord så grundar sig arbetsprocessen på en modifierad version av CRISP-DM, där dataförståelse, preparering av data, modellkonstruktion och utvärdering utgör kärnan för att nå resultatet. Valet föll på ramverket eftersom det är ett populärt och beprövat arbetssätt för projekt inom området och att det samtidigt passar bra för såväl en ensam utvecklare som för en grupp (Kelleher, Mac Namee & D'Arcy, 2020, s. 15–17; Saltz, 2021). Perspektivet för uppsatsen, d.v.s. utifrån området maskininlärning, har valts då det möjliggör för jämförelse av

bägge klassificeringsmodellerna under ett gemensamt paraply, även om de respektive modellerna har sitt ursprung inom närliggande områden (se sid. 6). Ett sätt att betrakta den nära kopplingen, är genom att maskininlärning utgör ett verktyg inom data science och att denna i sin tur är baserad på ekonometrin (se sid. 6).

4.2 DATAFÖRSTÅELSE

Datan som nyttjats i uppsatsen består av två underliggande dataset, ett behandlar kreditansökningar och ett kreditupplysningar. I sin initiala, obearbetade form håller de 1 048 575 respektive 438 557 observationer. De är bägge av strukturerad typ vilket beskrivits som en förutsättning för att tillämpa övervakad maskininlärning (IBM, 2021; Kelleher, Mac Namee & D'Arcy, 2020, s. 5) och avser att fungera som benchmark för inlärningsmodellerna. Nedan (se tabell 2 och tabell 3) presenteras variablerna från rådatan och en kort beskrivning.

Tabell 2 Kreditansökan (rådata)

Variabel	Beskrivning
ID	Unikt identifikationsnummer för en person
CNT_CHILDREN	Antal barn
AMT_INCOME_TOTAL	Årlig inkomst
DAYS_BIRTH	Dagar sedan födsel, där 0 avser född idag och (– 365) för ett år sedan
DAYS_EMPLOYED	Dagar sedan anställningen påbörjades, där 0 avser anställd idag och (– 365) för ett år sedan. Där +365 avser arbetslöshet i ett år
FLAG_MOBIL	Finns det en mobiltelefon
FLAG_WORK_PHONE	Finns det en arbetstelefon
FLAG_PHONE	Finns det en stationär telefon
FLAG_EMAIL	Finns det en email
CNT_FAM_MEMBERS	Antalet familjemedlemmar
CODE_GENDER	Kön
FLAG_OWN_CAR	Finns det en bil
FLAG_OWN_REALTY	Ägs boendet
NAME_INCOME_TYPE	Personens inkomsttyp
NAME_EDUCATION_TYPE	Personens utbildningsnivå
NAME_FAMILY_STATUS	Personens civilstånd
NAME_HOUSING_TYPE	Boendetyp
OCCUPATION_TYPE	Yrke

Tabell 3 Kreditupplysning (rådata)

Variabel	Beskrivning	
ID	Unikt identifikationsnummer för en person	
MONTH_BALANCE	Tidpunkten för upplysningen i förhållande till aktuell tidpunkt. Där 0 avser befintlig månad och (-1) föregående månad osv.	
STATUS	Avser personens kreditbalans där:	
	0	1 – 29 dagar efter med betalningen
	1	30 – 59 dagar efter med betalningen
	2	60 – 89 dagar efter med betalningen
	3	90 – 119 dagar efter med betalningen
	4	120 – 149 dagar efter med betalningen
	5	Övertrassering eller förfallna kreditskulder eller >149 dagar efter med betalningen
	C	Avbetalad kreditskuld den aktuella månaden
	X	Ingen kreditskuld

4.3 PREPARERING AV DATAN

Detta steg i arbetsmetoden har utgjorts av en iterativ process dels avseende EDA och datakvalitetsrapport, dels angående det praktiska genomförandet av de olika hanteringsstrategierna för de underliggande kvalitetsproblemen som identifierats (se sid. 15). Som det har beskrivits så är det övergripande syftet att dyka djupt in i den tillgängliga datan och dokumentera centrala tendenser och eventuella kvalitetsproblem (se sid. 15). Nedan redovisas de två separata datakvalitetsrapporterna som initialt utformades för att svara mot detta behov för rådatan. Respektive tabell innehåller en sektion för numerisk- och en för

kategorisk deskriptiv statistik, dessa har utformats i enlighet med den beskrivning som getts av Kelleher, Mac Namee & D'Arcy (2020, s. 53–54).

Tabell 4 Datakvalitetsrapport för kreditansökningar (rådata)

Kreditansökan

Variabel	Antal	Saknade %	Kardinalitet	Min	1 Kvartil	Medelvärde	Median	3 Kvartil	Max	Std. Dev.	Kvalitetsproblem	Hanteringsstrategi
ID	438557	0	438510	5008804	5609374.5	6022176.270	6047745	6456971.5	7999952	571637.023		
CNT_CHILDREN	438557	0	12	0	0	0.427	0	1	19	0.725	Hög korrelation med CNT_FAM_Members	Borttagning
AMT_INCOME_TOTAL	438557	0	866	26100	12500	187524.286	160780.5	225000	6750000	187524.286		
DAYS_BIRTH	438557	0	16379	-25201	-19483.5	-15997.905	-15630	-12514	-7489	4185.030	Avvikande format	Omformattering
DAYS_EMPLOYED	438557	0	9046	-17531	-3103	60563.675	-1467	-371	365243	138767.800	Avvikande format	Omformattering
FLAG_MOBIL	438557	0	1	1	1	1	1	1	1	0	Kardinalitet = 1	Borttagning
FLAG_WORK_PHONE	438557	0	2	0	0	0.206	0	0	1	0.405		
FLAG_PHONE	438557	0	2	0	0	0.288	0	1	1	0.453		
FLAG_EMAIL	438557	0	2	0	0	0.108	0	0	1	0.311		
CNT_FAM_MEMBERS	438557	0		1	2	2.194	2	2	20	0.897		

Variabel	Antal	Saknade %	Kardinalitet	Mode	Mode frekv.	Mode %	2 Mode	2 Mode frekv.	2 Mode %	Kvalitetsproblem	Hanteringsstrategier
CODE_GENDER	438557	0	2	F	294440	67.14	M	144117	62.86	Ej numerisk	Typomvandling
FLAG_OWN_CAR	438557	0	2	N	275459	62.81%	Y	163098	37.19	Ej numerisk	Typomvandling
FLAG_OWN_REALTY	438557	0	2	Y	304074	69.34	N	134483	30.66	Ej numerisk	Typomvandling
NAME_INCOME_TYPE	438557	0	5	Working	226104	51.56	Commercial associate	100757	17.21	Ej numerisk	Typomvandling
NAME_EDUCATION_TYPE	438557	0	5	Secondary/Secondary special	301821	68.82	Higer education	117522	26.80	Ej numerisk	Typomvandling
NAME_FAMILY_STATUS	438557	0	5	Married	299828	68.37	Singel/not married	55271	12.60	Ej numerisk	Typomvandling
NAME_HOUSING_TYPE	438557	0	6	House/apartment	393831	89.80	With parents	19077	4.35	Ej numerisk	Typomvandling
OCCUPATION_TYPE	304359	30.60	18	Laborers	78240	25.71	Core staff	43007	14.13	1. Saknade (134203), 2. Ej numerisk	1. Behåll icke-saknade, 2. Typomvandling

Tabell 5 Datakvalitetsrapport för kreditupplysningar (rådata)

Kreditupplysning

Variabel	Antal	Saknade %	Kardinalitet	Min	1 Kvartil	Medelvärde	Median	3 Kvartil	Max	Std. Dev.	Kvalitetsproblem	Hanteringsstrategi
ID	1048575	0	45985	5001711	5023644	5068286.425	5062104	5113856	5150487	46150.579	Dubletter	Bortagning dubletter
MONTH_BALANCE	1048575	0	61	-60	-29	-19.137	-17	-7	0	14.023	Redundant	Bortagning

Variabel	Antal	Saknade %	Kardinalitet	Mode	Mode frekv.	Mode %	2 Mode	2 Mode frekv.	2 Mode %	Kvalitetsproblem	Hanteringsstrategier
STATUS	1048575	0	8	C	442031	42.16	0	383120	36.54	Omvandling	Använd för att generera målvariabel samt historisk kvot

Som det framgår av kreditansökningstabellens första del (se tabell 4), så uppvisar variabeln CNT_CHILDREN hög korrelation (>0.7) med CNT_FAM_MEMBERS, detta listas som ett kvalitetsproblem då det kan medföra multikollinearitet för den logistiska regressionsmodellen. Hanteringsstrategin som listas för att adressera detta blir helt enkelt borttagning av CNT_CHILDREN och bygger på Dougherty (2016, s. 178) beskrivning. Ytterligare två kvalitetsproblem som listas är avvikande formatering för DAYS_BIRTH och DAYS_EMPLOYED, visserligen är detta i sig självt inte ett problem för inlärningsalgoritmerna men att till exempel ange ålder som ett negativ mått på antalet dagar från födseln kan vara

logiskt korrekt programmeringsmässigt men svårtytt för en människa. Därför föreslås omformatering som hanteringsstrategi. Variabeln FLAG_MOBIL har samma minsta och maximala värde, samt en kardinalitet lika med 1. Detta antyder att samtliga observationer enbart håller ett värde. Med andra ord så har samtliga kunder en mobiltelefon och variabeln kan i enlighet med den beskrivning som givits i litteraturgenomgången (se sid. 12) tas bort då den inte tillför någon ytterligare information.

I den andra delen av tabellen, framträder främst kvalitetsproblematik för OCCUPATION_TYPE som saknar input för en stor andel av observationerna (>30%). Eftersom den totala andelen saknade värden är <60% bör inte variabeln helt exkluderas enligt Kelleher, Mac Namee & D'Arcy (2020, s. 69–72) eftersom det fortfarande tillför information till datasetet. Men då andelen uppgår till strax över 30% så borde inte heller substituering via till exempel dess mode (yrkesgruppen laborers) att nyttjas som approach då det kan komma att påverka variabelns centrala tendens allt för mycket (se sid. 15). I stället har tillvägagångssättet med radering av dessa observationer använts. Visserligen medför detta att en viss informationsmängd går förlorad men eftersom datasetet initialt är mycket stort så förblir det totala antalet observationer fortfarande påtagligt (>304 000 observationer).

Slutligen gällande kreditansökningstabellen för rådatan, så behöver samtliga kategoriska variabler typomvandlas till numeriska värden för att kunna fungera som input till den logistiska regressionen (se sid. 15). För variablerna med en kardinalitet lika med 2 så nyttjas dummyvariabler och för de med högre kardinalitet översätts kategorin direkt till ett numeriskt värde, till exempel NAME_EDUCATION_TYPE har en kardinalitet om 5 och dessa kategorier kommer i stället representeras av värdena 0–4 (se sid. 15).

Gällande kreditupplysningstabellen, som innehåller betydligt färre variabler men nästan dubbelt så många observationer (se tabell 5) så är ett mindre kvalitetsproblem att tabellen innehåller dubbletter för vissa av observationerna och detta kunde enkelt hanteras genom radering. Gemensamt för de respektive tabellerna är att ingen av den innehåller någon definierad målvariabel (beroende variabel) utan detta är en komponent som behöver konstrueras.

Genom STATUS åtta kategorier i kreditupplysningstabellen (se tabell 3), kan dessa nyttjas för att definiera vad som i uppsatsen betraktas som en positiv eller ett negativ indikator för en

specifik individs kreditvärdighet. X (ingen kreditskuld) och C (avbetalad kreditskuld) ses som positiva indikatorer vid ansökan om ny kredit och de andra (övertrasserad, förfallna eller X dagar efter med betalningen) som negativa indikator, där positivt = 0 och negativt = 1. Genom detta har en binär målvariabel definierats. Samtidigt kan en ny variabel sammanfattande en individs kredithistorik genereras där antalet positiva förekomster för varje individ summeras och divideras med de negativa. I praktisk mening innebär detta att om en individ varit kund under en längre tid och skött sin kredit erhålls ett högre värde, och om en kund som varit kund länge missar en betalning, så påverkar detta kreditvärdigheten men kvoten förblir ändå betydligt högre än för en helt ny kund. För en ny kund ökar kreditvärdigheten snabbt efter ett lägre antal perioder, förutsatt att betalningarna sköts. Desto närmare eller lika med 0 avser en sämre kreditvärdighet.

MONTH_BALANCE har bedömts som redundant i sammanhanget då uppsatsen inte justerar för tidpunkten för senaste upplysningen utan snarare undersöker de befintliga uppgifter som finns kring en individs kredithistorik. Därför tas variabeln inte med och hanteringsstrategin blir borttagning.

I samband med att de sista kvalitetsproblemen hanterats kunde följande kvalitetsrapport genereras. Denna kommer samtidigt att utgöra ABT:n med tillägget av den beroende variabeln.

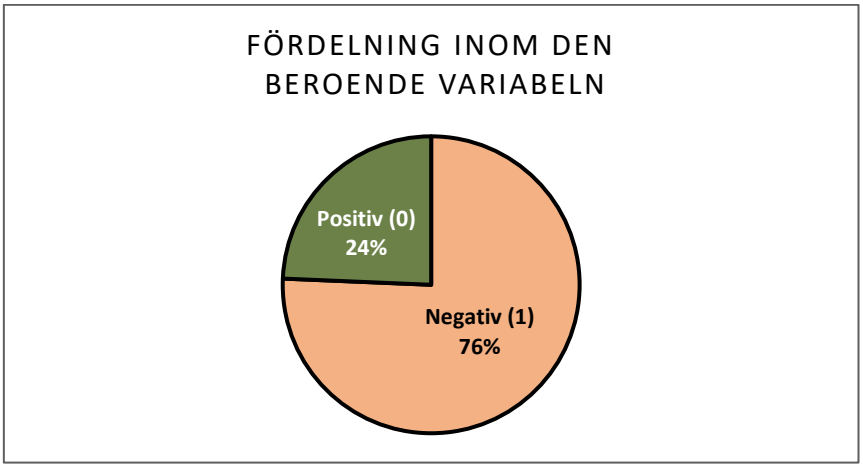
Tabell 6 Datakvalitetsrapport kombinerad tabell (färdigställd efter bearbetning av rådata)

Kombinerad tabell

Variabel	Antal	Saknade %	Kardinalitet	Min	1 Kvartil	Medelvärde	Median	3 Kvartil	Max	Std. Dev.	Kvalitetsproblem	Hanteringsstrategi
Gender	10 000	0	2	0	0	0.618	1	1	1	0.486		
Car	10 000	0	2	0	0	0.580	1	1	1	0.494		
Income	10 000	0	168	27000	135000	197034.204	180000	225000	1575000	109867.979		
Realty	10 000	0	2	0	0	0.344	0	1	1	0.475		
Income source	10 000	0	5	0	0	0.564	0	1	4	0.900		
Education	10 000	0	5	0	0	0.781	1	1	4	0.550		
Historical ratio	10 000	0	1226	0	0.180	27.995	1.192	4.884	610	86.511		
Civil status	10 000	0	5	0	1	1.243	1	1	4	0.777		
Housing	10 000	0	6	0	1	1.172	1	1	5	0.650		
Age	10 000	0	3965	21.159	32.805	40.636	40.104	67.427	67.427	9.596		
Employment time	10 000	0	2724	0.047	2.625	7.069	5.123	43.049	43.049	6.414		
Work phone	10 000	0	2	0	0	0.266	0	1	1	0.422		
Phone	10 000	0	2	0	0	0.289	0	1	1	0.453		
Email	10 000	0	2	0	0	0.102	0	0	1	0.303		
Occupation	10 000	0	18	0	3	4.656	4	6	17	3.235		
Family members	10 000	0	19	1	2	2.295	2	3	20	0.973		

När en kombinerad kvalitetstabell erhållits justeras datan för eventuella extremvärden. Den approach som användes baserades mot clamp transformation via den beskrivning som givits tidigare i litteraturgenomgången (se sid. 15). De variabler som undersöktes och därigenom utgjorde selektionsgrund var de som håller en kardinalitet >2 (se tabell 6).

Slutligen, när ABT:n erhållits noterades en obalans i den beroende variabeln (se figur 9). För att förhindra introducerandet av bias och potentiella problem med underfitting (se sid. 18) så nyttjades undersampling av den mest frekvent förekommande utfallet (1) efter den beskrivning som Alto (2021) givit. Tillvägagångssättet medför att datasetet blir betydligt mindre, men eftersom antalet observationer är stort så har metoden ändå valts då den jämfört med oversampling inte introducerar duplicerade observationer (se sid. 15).



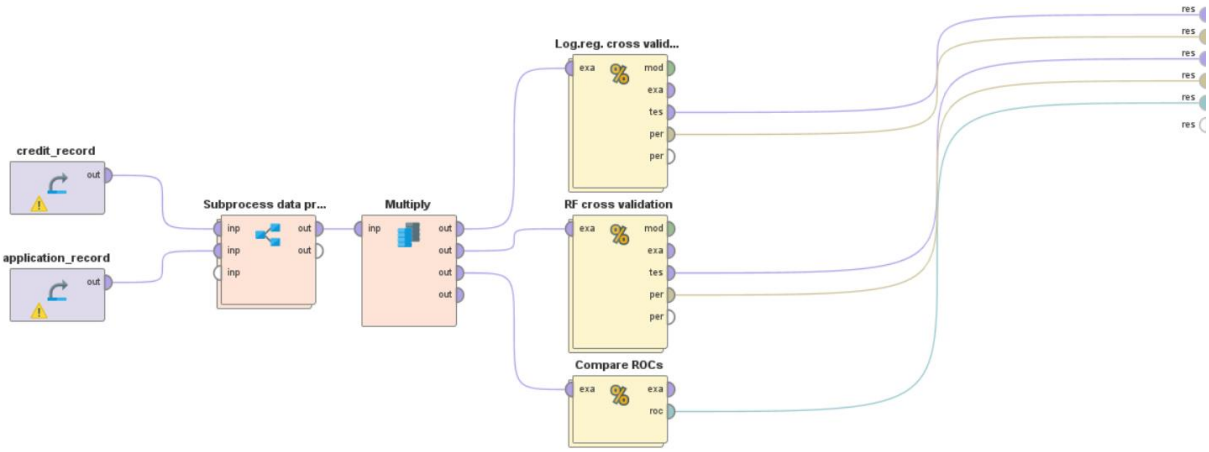
Figur 9 Fördelning i den beroende variabel (obalanserad)

Efter balanseringen av den beroende variabeln har den färdiga ABT:n erhållits och håller nu 10 000 observationer som kommer ligga till grund för träningen och testningen av modellerna.

4.4 PROCESSKONSTRUKTION, FRÅN DATA TILL RESULTAT

För att möjliggöra för såväl datapreppareringens olika delmoment som konstruktionen av modellerna, dess resultat och utvärderingen av desamma så har Rapidminer använts. Rapidminer är en mjukvara avsedd för maskininlärning eller andra analysorienterade projekt och tillhandahåller ett grafiskt interface som tillåter både implementeringen av egna skript eller design av analysprocessen genom olika moduler som kopplas samman i ett processflöde.

För de respektive modulerna definieras vilka underliggande metoder de skall nyttja och varierande hyperparametrar justeras för att uppnå önskad output. Figur 10 visar processflödet som använts i uppsatsen.



Figur 10 Process från Rapidminer i makroperspektiv

Initialt läses datan från de respektive dataseten in och kopplas samman med datapreparingen. Denna modul innehåller ett antal nästlade subprocesser (se beskrivning i föregående stycke) för att slutligen generera ABT:n som output. Denna kopieras därefter och passas över som input till subprocessen för korsvalidering av den logistiska regressionsanalysen, subprocessen för korsvalidering av random forest och även en operator för att plotta de respektive ROC-kurvorna för inlärningsmodellerna. Outputen från dessa tre moduler ger upphov till de prestationsmått som utgör uppsatsens resultat.

Gällande konfigureringen för den logistiska regressionen så ligger denna som en delkomponent inom en 10-folded cross validation. Detta för att träna och testa modellen i en upprepad sekvens för att säkerställa och möjliggöra för att utvärdera dess prestation (se sid. 12). I den följande träningsfasen nyttjas normalversionen av operatorn Logistic Regression och dess funktionssätt följer den som beskrivs i litteraturgenomgången (se sid. 20). När en generaliseringsbar modell erhållits så övergår cross valideringsmodulen till testning och listade prestationsmått erhålls som output.

Tillvägagångssättet för den andra modellen, random forest, är identisk och baserad på samma datainput som den föregående. Den enda skillnaden är att cross valideringen tillämpar random forest som inlärningsmodell i stället och dess funktion följer tidigare beskrivna (sid. 24), med tillägget att antalet decision trees satts till 125st med ett djup på 12 förgreningar och utan trimning. Motiveringen till detta grundar sig helt på try-and-error via optimeringsoperator som utvärderade varierande konfigureringar av modellens hyperparametrar (ej inkluderat i figur 10). Den ovan nämnda konfigureringen motsvarar högst AUC (se sid. 24).

4.5 UTVÄRDERING

Det slutliga steget som uppsatsens modifierade version av CRISP-DM ramverket tar är också kanske det kortaste - utvärderingen av de respektive modellerna i förhållande till varandra. Detta görs helt numeriskt genom de utvärderingsparametrar som beskrivits tidigare (se sid. 28). Som nämnts så är dessa frekvent nyttjade mått för klassificeringsproblem inom övervakad maskininlärning. De ger tillsammans en god överblick över de respektive modellernas prestation utifrån olika perspektiv men där AUC, via ROC-kurvan, utgör den i särklass mest betydelsefulla parametern (se sid. 28) för den här typen av jämförelse. Genom Rapidminer erhålls dessa mått i separata tabeller eller i den gemensamma grafen för ROC-kurvorna och kan därefter enkelt sammanställas och presenteras.

4.6 DATAETIK

Gällande de etiska aspekter som oftast finns kopplade till hantering av data rörande personuppgifter så skriver Walliman (2017) att material som kan bedömas vara av känslig natur skall hanteras och presenteras på ett sådant sätt att det värnar den personliga integriteten. Utifrån uppsatsens perspektiv så medför detta att den data som behandlas potentiellt kan antas vara av känslig karaktär då den innehåller ID-nummer (Integritetsmyndigheten, 2022). Visserligen så framgår det inte om detta är individuella ID-nummer, kundnummer, eller autogenererade och enbart förekommande i syftet att separera observationerna inom dataseten, men likväl kan det vara känslig information. Oaktat detta så har det inom arbetet inte funnits anledning att fokusera på något eller några enskilda observationer, eftersom datan enbart använts för träning och testning av klassificeringsalgoritmerna utifrån ett makroperspektiv. Samtidigt så kan kvantiteten i sig

antas medföra en viss anonymisering då det är oerhört svårt att särskilja en viss observation från mängden.

4.7 RELIABILITET OCH REPLIKERBARHET

Hammond (2012, s. 131) definierar reliabilitet för studier som ett mått på den tillförlitlighet som kan tillskrivas för att utfallet förblir detsamma under upprepade genomföranden. Med andra ord, att resultatet är stabilt. Samtidigt beskrivs en kvantitativ studies replikerbarhet som en central aspekt för att kunna validera resultatet.

Gällande reliabilitet och utifrån uppsatsens perspektiv, där 10-folded korsvalidering använts just med syftet att erhålla hög tillförlitlighet och utfallet är ett medelvärde av de iterationer mellan träning- och testning som gjorts (se sid. 39), så kan resultatet antas att utifrån de givna förutsättningarna och metodvalen som gjorts, svara mot en god reliabilitet då prestationsmåtten som redovisas i resultatdelen varierar inom ett relativt snävt intervall kring sitt medelvärde (se sid. 43).

Eftersom uppsatsens rpm-process och dataset finns tillgängliga som XML och CSV (skickas vid förfrågan), och i princip kan köras med vilken laptop som helst, med enbart ett par installationer, kan det även antas att replikerbarheten är god då den enkelt kan genomföras.

5 RESULTAT

Följande kapitel avser att presentera tabeller och en ROC-graf innehållande resultatet från modellerna. Initialt visas tabeller över confusion matrices för de bägge modellerna följt av en tabell över deras respektive prestation i förhållande till varandra. Slutligen porträtteras deras ROC-kurvor i samma graf för att underlätta jämförelsen.

5.1 CONFUSION MATRICES

Logistisk regression		Målvariabel	
		Positivt	Negativt
Prediktion	Positivt	4676	324
	Negativt	1547	3453

Tabell 7 Confusion matrix för logistisk regression

Random forest		Målvariabel	
		Positivt	Negativt
Prediktion	Positivt	4237	763
	Negativt	623	4377

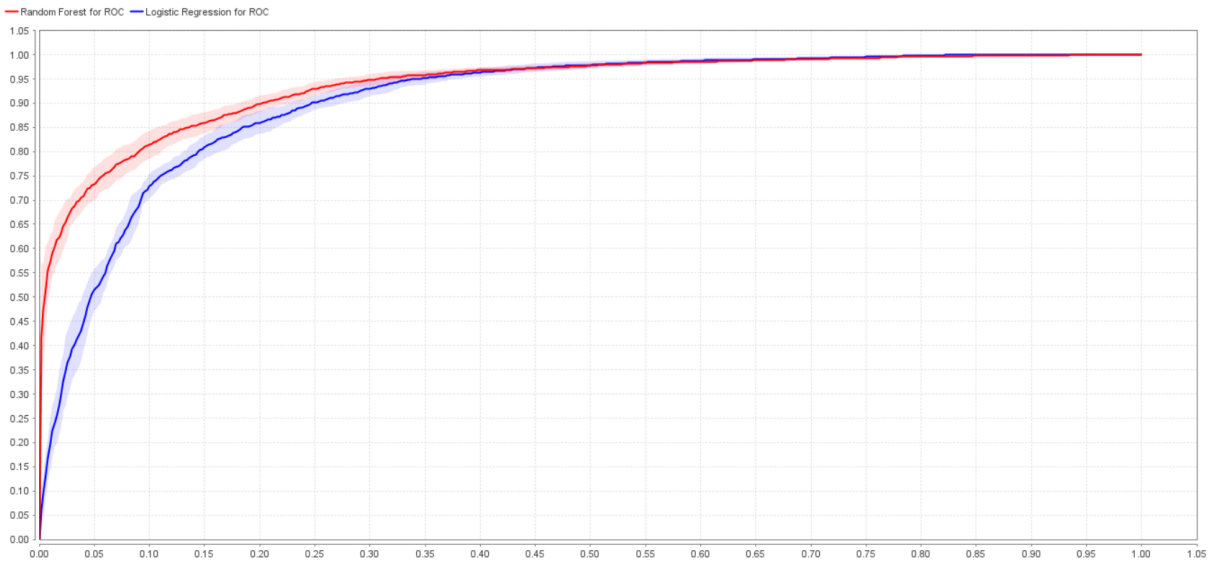
Tabell 8 Confusion matrix för random forest

5.2 PRESTATIONSMÅTT

	Logistisk regression	Random forest
Accuracy	**81.29%** ± 1.13%	**86.14%** ± 0.61%
AUC	**0.906** ± 0.008	**0.938** ± 0.003
Precision	**75.17%** ± 1.44%	**87.19%** ± 0.85%
Recall	**93.52%** ± 1.40%	**84.74%** ± 1.18%
F1-score	**83.33%** ± 0.88%	**85.94%** ± 0.66%

Tabell 9 Prestationsmått för modellerna

5.3 GRAF ÖVER ROC-KURVOR FÖR MODELLERNA



Figur 11 ROC-kurvorna för logistisk regression (blå) och random forest (röd)

6 DISKUSSION

I resultatet framgår det tydligt att random forest presterar bättre än logistisk regression i fyra av de fem prestationsmåtten som presenteras. I det centrala AUC-måttet så uppgår differensen till 0.032 mellan areorna. En liten skillnad kan tyckas, men ändå en skillnad. Vad som skall understrykas är att bägge inlärningsmodellerna presterar bra och på en relativt hög nivå sett utifrån en abstrakt 45^{o} linje, representerande en slumpad gissning. Uttryckt i accuracy, d.v.s. antalet korrekta prediktioner kontra det totala antalet (se sid. 28), så görs en korrekt klassificering i drygt 86% av fallen för random forest och 81% av gångerna för logistisk regression. Eftersom datan har balanserats (se sid. 35) så blir precision av mindre betydelse i uppsatsens fall då den är klasspecifik (håller enbart information om positiva prediktioner, se sid. 28). Samtidigt bör måttet inkluderas i de prestationsmått som undersöks då det utgör en förutsättning för att beräkna F1 score (se sid. 28).

I avseendet kring recall så framgår det att logistisk regression har en bättre säkerhet kring att prediktera samtliga positiva utfall i förhållande till random forest. Differensen här uppgår till 8.78 procentenheter. Men då både precision och recall inkorporeras i det gemensamma måttet F1 score (se sid. 28), för att undersöka andelen missklassificeringar, framträder random forest som det bättre presterande modellen med en differens mot logistisk regression om 2.61 procentenheter. En tydlig skillnad men på relativt höga nivåer. Genom detta blir det uppenbart att logistiska regressionens sigmoid-funktion har svårare att generalisera ny data än det icke-linjära tillvägagångssättet inom random forest och i det här fallet verkar den sistnämnda vara modellen att föredra utifrån ett prestationsorienterat perspektiv.

Logistisk regression är som tidigare nämnt en felbaserad inlärningsmodell som fungerar bra gentemot binär klassificering (se sid. 20). Samtidigt så erbjuder modellen en tydlig insikt i hur den härleder sina slutsatser genom de koefficienter som ligger till grund för att passa in sigmoid-funktionen efter maximum likelihood (se sid. 21). Till skillnad mot denna så räknas random forest som en informationsbaserad inlärningsalgoritm eftersom den orienterar strukturen på sina underliggande decision trees mot hur information gain kan maximeras (se sid. 25). Inom random forest, som namnet antyder och som tidigare beskrivits, så är randomisering en fundamental grundsten för att skapa diversiteten mellan de olika träden. Detta medför att den inte är lika transparent då den, som i uppsatsens fall, består av över

1500 noder som konstruerat utifrån en randomisering. Så även om principen är enkel, medför kvantiteten att det blir svårt för en människa att överblicka helheten för hur exakt en slutsats härleds. Det blir bokstavligt talat svårt att se skogen för alla träd.

Inom uppsatsens ramar exkluderades visserligen det avslutande steget inom CRISP-DM (implementeringen inom en verksamhet) men hypotetisk så är det min uppfattning att transparensen som de respektive inlärningsmodellerna för med sig potentiellt sett kan vara en viktig faktor att ta hänsyn till. Därigenom kanske inte valet per automatik landar på den bäst presterande algoritmen då differensen är tämligen liten och prediktionstypen tillåter det. Som i uppsatsens fall, där en tänkt kreditinstitution eventuellt skulle kunna nyttja en modell med något lägre prestation men högre transparens, är detta kanske ett bra alternativ för att kunna svara mot lagstiftning gällande datainsamling, hantering och utlämning som till exempel GDPR (Integritetsmyndigheten, 2022). Detta till skillnad mot exempelvis en medicinsk applikation där transparensen blir sekundär i förhållande till den aktuella prestationen och förmågan att till exempel ställa kritiska diagnoser gällande liv och hälsa.

Random forest som inlärningsmodell levererar ett mer träffsäkert resultat än logistisk regressionsanalys, men den är lite som en svart låda. Varje träd i skogen har en röst i valet om att bevilja en kreditansökning eller ej, och det går att förstå varför ett träd röstat så som det gjort, men ställer man sig frågan kring vad den grundar sin prediktion på, blir modellen genast mycket komplex att överblicka. Beslutsvägen algoritmen tar genom noderna, baserat på beräkning, särskiljning och randomisering av träningsdatan, leder fram till en så homogen grupp av observationer som möjligt, vilket ligger till grund för prediktionen. Det är därför oerhört svårt, för att inte säga omöjligt för den som använder sig av modellen att kunna härleda varför ett beslut fallit ut så som det gjort. Detta eftersom det skulle innebära att man samtidigt redogör för den exakta konstruktionen av inlärningsmodellen och denna är som vi vet konceptuellt överblickbar, men i praktiken påtagligt komplex i och med sin storlek i kombination med det återkommande användandet randomiseringen för att skapa diversitet inom ensemblen.

Resultatet visar att random forest är en mer träffsäker metod än logistisk regressionsanalys, trots detta verkar den senare vara en mer använd metod. Detta skulle eventuellt kunna förklaras med att dess koefficienter medför att en slutsats kan härledas mer exakt, något som är nödvändigt vid hantering av till exempel personuppgifter. I och med den europeiska

lagstiftningen om GDPR måste alla som samlar in, hanterar och använder sig av känslig data, till exempel personuppgifter som på ett eller annat sätt kan knytas till en specifik person, kunna specificera varför de gör det, vad datan används till, och på vilket sätt. Om till exempel ett kreditinstitut skulle använda sig av random forest som modell och få frågan varför beslutet föll ut som det gjorde skulle det mest troligt vara oerhört svårt att härleda beslutet – hur korrekt beslutet än må ha varit.

Även om en ökad träffsäkerhet, i detta fall en differens i AUC på 0.032 (se sid. 43), minskar risken för kreditinstituten vid utlåning skulle användningen av random forest också kunna innebära en kostnad, till exempel i form av viten för att ha brustit i personuppgiftshanteringen eller stämningar av kunder som anser sig diskriminerade och inte enkelt kan få en förklaring till utfallet. En högre felprediktion via logistisk regressionsanalys, ger kanske i detta fall trots allt en lägre kostnad i förlängningen i och med att modellen är mer transparent och att det därigenom går att härleda beslut och på vilket sätt datan som använts vid prediktionen har gjort det på rätt sätt och laglig grund. I det stora hela skulle man kunna kalla det för ett optimeringsproblem; eventuella förluster inom ramen för differensen i AUC om 0.032, kontra böter och stämningar till följd av en träffsäkrare metod. De etiska grunderna för införandet av dataskyddsförordningen kontra kreditinstitutens träffsäkerhet innebär sannolikt utmaningar för de som önskar använda sig av algoritmer likt i random forest.

Huruvida random forest användes som metod för att avgöra kreditvärdighet innan GDPR infördes vet jag inte, men det vore intressant att jämföra träffsäkerheten hos kreditinstituten före och efter införandet av en gemensam europeisk dataskyddsförordningen.

7 SLUTSATS

Logistisk regressionsanalys har en god prestation (AUC ≈ 0.906) som binär klassificeringsalgoritm men inte bättre än random forest (AUC ≈ 0.938) för det givna datasetet. I de mått som presenteras i uppsatsen så presterar random forest bättre i 4 av 5, men där logistisk regression uppvisar en högre säkerhet på att samtliga positiva utfall identifierats av modellen. Vad som även kan konstateras är att bägge modellerna presterar på höga nivåer relativt ett slumpmässigt utfall.

Prestationsmåtten som nyttjas har samtidigt medfört att klassificeringsmodellerna enkelt kunnat jämföras i både tabellformat och grafiskt. I litteraturgenomgången har flera potentiella aspekter rörande hur maskininlärningsmodellen random forest skulle kunna fungera som ett kompletterande verktyg till den ekonometriska verktygslådan beskrivits.

Utifrån ett kreditinstituts perspektiv, där man antas ha ett intresse av klassificering av kreditvärdighet, kan uppsatsens tillvägagångssätt eventuellt visa en möjlig approach till hur detta kan utformas. Om inte annat, så vittnar det om den grundläggande principen bakom de inlärningsmodeller som helt eller delvis fungerar som beslutsstöd vid kreditbedömning.

8 REFERENSER

Agrawal, A., Gans Joshua, S. & Goldfarb, A. (2019). Artificial Intelligence : The Ambiguous Labor Market Impact of Automating Prediction, *The Journal of Economic Perspectives,* vol. 33, no. 2**,** s. 31-50.

Al-Masri, A. (2019). What is Overfitting and Underfitting in Machine Learning? Towards Data Science, 22 Juni, Tillgänglig online: https://towardsdatascience.com/what-are-overfitting-and-underfitting-in-machine-learning-a96b30864690 [Hämtad: 2022-05-05]

Alam, M. (2021). Econometrics techniques for data science. Towards Data Science*,* 21 December, Tillgänglig online: https://towardsdatascience.com/econometrics-techniques-for-data-science-ef4a880415b4 [Hämtad: 2022-05-05]

Alto, V. (2021). How to deal with Unbalanced Dataset in Binary Classification – Part 1, Medium*,* 24 januari, Tillgänglig online: https://medium.com/dataseries/how-to-deal-with-unbalanced-dataset-in-binary-classification-part-1-2c25fae0e9e4 [Hämtad: 2022-05-05]

Angrist, J. (2019). *Joshua Angrist: Are Machine Learning and Big Data Changing Econometrics?* Youtube, [video online], Tillgänglig online: https://www.youtube.com/watch?v=Bm6CAjVtrIw&t=34s&ab_channel=MarginalRevolutionUniversity [Hämtad: 2022-05-08]

Angrist, J. (2021). *Joshua Angrist – Econometrics is the original data science,* Youtube, [video online], Tillgänglig online: https://www.youtube.com/watch?v=T24j8XTcpe0&list=PLPcq-fcJ-S6o4oky2cKfarQWEuhVyAb-l&index=20&ab_channel=RajkCollegeforAdvancedStudies [Hämtad: 2022-05-08]

Bartlett, J. (2014). *R squared in logistic regression.* Tillgänglig online: https://thestatsgeek.com/2014/02/08/r-squared-in-logistic-regression/ [Hämtad: 2022-05-05]

Bell, J. (2020) Machine Learning: Hands-On for Developers and Technical Professionals. 2. uppl. John Wiley & Sons, Incorporated.

Bhattacharyya, S. (2018). 'Logit' of Logistic Regression; Understanding the Fundamentals, Towards Data Science*, 21* Oktober*,* Tillgänglig online: https://towardsdatascience.com/logit-of-logistic-regression-understanding-the-fundamentals-f384152a33d1 [Hämtad: 2022-05-12]

Bronshtein, A. (2017). Train/Test Split and Cross Validation in Python, Towards Data Science*,* 17 Maj, Tillgänglig online: https://medium.com/towards-data-science/train-test-split-and-cross-validation-in-python-80b61beca4b6 [Hämtad: 2020-05-05]

Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies: Oxford University Press.

Chollet, Francois. (2021). *Deep Learning with Python*. 2. uppl. New York: Simon and Schuster.

Dougherty, C. (2016). *Introduction to Econometrics.* 5. uppl. Oxford: Oxford University Press.

Edwards, G. (2018). Machine Learning an Introduction, Towards Data Science, 18 November, Tillgänglig online: https://towardsdatascience.com/machine-learning-an-introduction-23b84d51e6d0 [Hämtad: 2022-05-05]

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters* 27 (2006) 861–874. Tillgänglig online: https://doi.org/10.1016/j.patrec.2005.10.010 [Hämtad: 2022-05-05]

Federal Reserve. (2022). *Delinquency Rate on Credit Card Loans, All Commercial Banks.* Tillgänglig online: https://fred.stlouisfed.org/graph/?g=Bod4 [Hämtad: 2022-05-05]

Geweke, J., Horowitz, J. L., & Pesaran, H. (2008). The New Palgrave Dictionary of Economics Online.

Grus, J. (2019). *Data Science from Scratch.* 2. uppl. Sebastopol: O'Reilly.

Haldar, M. (2015). How much tranining data do you need? Medium, 28 November, Tillgänglig online: https://malay-haldar.medium.com/how-much-training-data-do-you-need-da8ec091e956 [Hämtad: 2022-05-10]

Hammond, M., Wellington, J. (2013). Research Methods the Key Concepts. 1. uppl. New York: Taylor & Francis Books.

Hastie, T., Tibshirani, R., Friedman, J. (2009). The Elements of Statistical Learning. 2. uppl. Stanford: Springer.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 20(8), pp. 832–844. doi: 10.1109/34.709601.

IBM. (2021). Structured vs. Unstructured Data: What's the Difference? Tillgänglig online: https://www.ibm.com/cloud/blog/structured-vs-unstructured-data [Hämtad: 2022-05-11]

IBM. (2022) Big data analytics. Tillgänglig online: https://www.ibm.com/in-en/analytics/hadoop/big-data-analytics [Hämtad: 2022-05-04]

Integritetsmyndigheten. (2022). Introduktion till dataskyddsförordningen (GDPR). Tillgänglig online: https://www.imy.se/privatperson/dataskydd/introduktion-till-gdpr/ [Hämtad: 2022-05-12]

Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2020). *Fundamentals of Machine Learning for Predictive Data Analytics.* 2. uppl. London: The MIT Press.

Kubina, M., Varmus, M., & Kubinova, I. (2015). Use of big data for competitive advantage of company. *Procedia Economics and Finance*, *26*, 561-565.

Kumar, S. (2020). Feature Engineering – deep dive into Encoding and Binning techniques, Medium, 26 Augusti, Tillgänglig online: https://medium.com/towards-data-science/feature-engineering-deep-dive-into-encoding-and-binning-techniques-5618d55a6b38 [Hämtad: 2022-05-08]

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data: the management revolution. *Harvard business review*, *90*(10), 60–68.

Microsoft. (2022). *Djupinlärning jämfört med maskininlärning i Azure Machine Learning*. Tillgänglig online: https://docs.microsoft.com/sv-se/azure/machine-learning/concept-deep-learning-vs-machine-learning [Hämtad: 2022-05-05]

Minaee, S. (2019). 20 Popular Machine Learning Metrics. Part1: Classification & Regression Evaluation Metrics, Medium, 28 Oktober, Tillgänglig online: https://medium.com/p/1ca3e282a2ce [Hämtad: 2022-05-05]

Mishra, P. (2021). 5 Outlier Detection Techniques that every ”Data Enthusiast” Must Know, Towards Data Science, 12 Juni, Tillgänglig online: https://towardsdatascience.com/5-outlier-detection-methods-that-every-data-enthusiast-must-know-f917bf439210 [Hämtad: 2020-05-10]

Popov, V. (2019). Dealing with Categorical Data, Medium, *31 Augusti,* Tillgänglig online: https://medium.com/machine-learning-eli5/dealing-with-categorical-data-f4c8556cbda0 [Hämtad: 2022-05-07]

Saltz, J. S. (2021) ‘CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps’. *IEEE International Conference on Big Data*. pp. 2337–2344. doi: 10.1109/BigData52589.2021.9671634.

SAS Institute. (2022). Big Data. Tillgänglig online: https://www.sas.com/sv_se/insights/big-data/what-is-big-data.html [Hämtad 2022-05-04]

Shmueli, G. (2010). To explain or to predict? *Statistical science*, *25*(3), 289–310

Starmer, J. (2022). The StatQuest Illustrated Guide to Machine Learning. Independently published.

Sumpter, D. J. T. (2019). Uträknad: Sanningen Om Algoritmerna Som Styr Världen, Första utgåvan: Volante.

Sunil, K. S. (2019). Econometrics and Data Science: An Econometric Perspective. *COJ Technical & Scientific Research*. 2(2). COJTS.000531.2019.

Tegmark, M. (2017). Liv 3.0: Att Vara Människa I Den Artificiella Intelligensens Tid. Volante.

Tegmark, M. (n.d.). *Benefits & Risks of Artificial Intelligence*. Tillgänglig online: https://futureoflife.org/background/benefits-risks-of-artificial-intelligence [Hämtad 2022-05-16]

Varian, Hal R. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, 28 (2): 3-28. DOI: 10.1257/jep.28.2.3

von Luxburg, U. and Schoelkopf, B. (2008). Statistical Learning Theory: Models, Concepts, and Results. https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,cookie,url,shib&db=edsarx&AN=edsarx.0810.4752&lang=sv&site=eds-live&scope=site [Hämtad: 2022-05-08]

Walliman, N. (2017) Research Methods. 2. uppl. Florence: Taylor & Francis Books.

Wolpert, D. (1996). The Lack of A Priori Distinctions between Learning Algorithms, *Neural Computation*, pp. 1341–1390.

Wolpert, D.H., and Macready, W.G. (2005). Coevolutionary free lunches, *IEEE Transactions on Evolutionary Computation*, 9(6): 721–735

Wrg, A. (2019). From Econometrics to Machine Learning, Towards Data Science, 8 September, Tillgänglig online: https://towardsdatascience.com/from-econometrics-to-machine-learning-ee182f3a45d7 [Hämtad: 2022-05-05]

Yiu, T. (2019). Understanding Random Forest, Towards Data Science, 12 Juni, Tillgänglig online: https://towardsdatascience.com/understanding-random-forest-58381e0602d2 [Hämtad: 2022-05-11]