Hanna Hovd
Rikke Olsen Aase

Mislighold av kredittkortgjeld

En studie av ulike karakteristikker ved lånsøkere som påvirker sannsynligheten for mislighold av kredittkortgjeld

Masteroppgave i Økonomi og administrasjon
Veileder: Ranik Raaen Wahlstrøm
Mai 2022

Masteroppgave

NTNU
Norges teknisk-naturvitenskapelige universitet
Fakultet for økonomi
NTNU Handelshøyskolen

NTNU
Kunnskap for en bedre verden

Hanna Hovd
Rikke Olsen Aase

Mislighold av kredittkortgjeld

En studie av ulike karakteristikker ved lånsøkere som påvirker sannsynligheten for mislighold av kredittkortgjeld

Masteroppgave i Økonomi og administrasjon
Veileder: Ranik Raaen Wahlstrøm
Mai 2022

Norges teknisk-naturvitenskapelige universitet
Fakultet for økonomi
NTNU Handelshøyskolen

NTNU
Kunnskap for en bedre verden

Forord

Denne oppgaven er vårt avsluttende arbeid på masterstudiet i økonomi og administrasjon med hovedprofil Business Analytics, ved NTNU Handelshøyskolen. Et studie som har vært både utfordrende og spennende, gitt mye dårlig samvittighet for alt arbeid som skulle vært gjort og alle timer med digitale forelesninger som skulle vært sett, men samtidig også mestring og uforglemmelige øyeblikk. Rett og slett en liten berg-og-dal-bane, akkurat som skrivingen av denne masteroppgaven også ble.

Vi ønsker derfor å rette en stor takk til vår veileder Ranik Raaen Wahlstrøm for upåklagelig hjelp ved skriving av denne masteroppgaven. Uten hans bistand med kompetanse innen maskinlæring ville definitivt denne masteroppgaven vært noe annet. Vi vil også takke Sparebank 1 Kreditt for tilgang på data, muligheten til å skrive denne oppgaven og god hjelp underveis i prosessen. Ellers vil vi takke hverandre for nydelig samarbeid, morsomme stunder, flere nesten-sammenbrudd og mange gode timer på kontoret.

Innholdet i denne oppgaven står for forfatterenes regning.

Sammendrag

Tema i denne masteroppgaven er mislighold av kredittkortgjeld. Formålet er å gi långivere av kredittkortgjeld en indikasjon på hvilke variabler som er viktige å hensynta i vurderingsprosessen av lånsøkere. Problemstillingen som vil studeres i denne masteroppgaven er dermed som følger: *«Hvilke karakteristikker ved lånsøkere påvirker sannsynligheten for mislighold av kredittkortgjeld?»*. For å besvare problemstillingen utvikler vi prediksjonsmodeller, for deretter å studere variabelviktigheten. Prediksjonsmodellene utvikles ved bruk av logistisk regresjon og Extreme Gradient Boosting, mens variabelviktigheten studeres med henholdsvis Least Absolute Shrinkage and Selection Operator og SHapley Additive exPlanations.

Videre vil vi undersøke hvilken av de to estimeringsteknikkene som oppnår best prediksjonsevne ved utvikling av modeller for kredittvurdering. Da Extreme Gradient Boosting er en svart boks, vil vi også se på hvordan slike estimeringsteknikker kan gjøres tolkbare for begrunnelse av kredittvurderinger.

Studien er avgrenset til å evaluere innvilgede kredittkortsøknader i det norske privatmarkedet. Datasettet vi benytter er levert av Sparebank 1 Kreditt, og innholder data fra perioden november 2019 til desember 2021.

Vår studie konkluderer med at en lånsøker sin alder og årlige nettoinntekt er sentrale for prediksjon av mislighold av kredittkortgjeld. Videre finner vi også at informasjon om hvorvidt en kredittkortsøknad er sendt inn på natten, inngående beløp på debetkonto hos dataleverandør de siste 6 månedene, samt forholdet mellom søkt og innvilget kredittkortgrense er viktige variabler i kredittvurderingsmodeller. Dette er et unikt bidrag til litteraturen, da variablene tidligere ikke er nevnt i litteraturen innen kredittvurdering. Videre konkluderer oppgaven med at prediksjonsmodeller utviklet ved bruk av logistisk regresjon presterer noe bedre enn modeller estimert med Extreme Gradient Boosting. I tillegg finner vi at SHapley Additive exPlanations er godt egnet for å tolke svarte bokser.

Abstract

The theme of this master's thesis is credit card default. The aim of this thesis is to give lenders an indication of which variables are important in the process of evaluating credit card applicants. In extension of this, we will look at the following question: *«What characteristics of loan applicants affect the probability of default on credit card debt?»*. This will be examined by developing credit scoring models, and then by studying the variable importance. The credit scoring models are developed using logistic regression and Extreme Gradient Boosting, while the variable importance is examined with Least Absolute Shrinkage and Selection Operator and SHapley Additive exPlanations, respectively.

Furthermore, we will investigate which of the two estimation techniques achieves the best predictive accuracy when developing models for credit scoring. As Extreme Gradient Boosting is a black box, we will also look at how such estimation techniques can be made interpretable to justify credit ratings.

This study is limited to evaluate approved credit card applications in the Norwegian private market. The data set is provided by Sparebank 1 Kreditt, and includes data from the period November 2019 to December 2021.

Our study concludes that a credit card applicant's age and annual net income are key variables when predicting credit card default. Furthermore, we also find that whether a credit card application has been submitted at night, the total incoming amount on the applicant's debit account maintained by the provider of data in the last 6 months, and the relation between applied and granted credit card limit are important variables in credit scoring models. This is a unique contribution to the literature, as prior literature within credit scoring do not include these variables. Furthermore, the thesis concludes that credit scoring models developed using logistic regression perform somewhat better than models estimated with Extreme Gradient Boosting. In addition, we find that SHapley Additive exPlanations are well suited for interpreting black boxes.

Innhold

1 Introduksjon **1**

1.1 Tema og aktualisering 1

1.1.1 Kredittvurdering 2

1.2 Problemstilling 3

1.2.1 Forskningsspørsmål 4

1.3 Oppgavens struktur 4

2 Litteraturgjennomgang **6**

2.1 Estimeringsteknikker 6

2.2 Metoder for analyse av variabler 10

2.2.1 Metoder for variabelseleksjon 10

2.2.2 Metoder for variabelviktighet 12

2.3 Sentrale variabler 14

3 Data **17**

3.1 Variabler 17

3.1.1 Variabeltransformasjon 17

3.1.2 Utelatelse av variabler 19

3.2 Håndtering av manglende verdier 19

3.3 Håndtering av ubalansert data 20

3.4 Endelig datasett 21

4 Metode **22**

4.1 Datainndeling og kryssvalidering 22

4.2 Estimeringsteknikker 23

4.2.1 Logistisk regresjon 23

4.2.2 XGBoost 24

4.3 Metoder for analyse av variabler 25

4.3.1 LASSO 26

4.3.2 SHAP 26

4.4 Evalueringsmål . . . 28
4.4.1 AUC . . . 28
4.4.2 Brier Score . . . 29
4.5 Optimalisering av hyperparametere . . . 29
4.5.1 Optimalisering i LASSO . . . 29
4.5.2 Optimalisering i XGBoost . . . 30

5 Resultater **32**
5.1 Variabelviktighet . . . 32
5.1.1 LASSO . . . 32
5.1.2 SHAP . . . 33
5.2 Modellprediksjon . . . 35
5.2.1 Logistisk regresjon . . . 36
5.2.2 XGBoost . . . 36
5.3 Tolkning av svarte bokser . . . 36

6 Diskusjon **39**
6.1 Variabelviktighet . . . 39
6.2 Modellprediksjon . . . 45
6.3 Tolkning av svarte bokser . . . 46

7 Konklusjon **47**
7.1 Våre funn og implikasjoner . . . 47
7.2 Svakheter . . . 48
7.3 Forslag til videre forskning . . . 49

Referanser **50**

A Appendiks **62**
A.1 Variabler i endelig datasett . . . 62
A.2 Trening og test . . . 65
A.3 AUC og LASSO . . . 66
A.4 Logistisk regresjon . . . 76

A.5 Optimalisering av hyperparametere i XGBoost 80
A.6 SHAP 81
A.7 Beskrivende statistikk 91
A.8 Utelatte variabler 94

Figurer

1 Utvikling i forbrukslån og mislighold (Finanstilsynet, 2022) 1
2 Kryssvalidering for å finne optimal λ (scikit-learn, 2017) 30
3 Standardiserte SHAP-verdier 34
4 Låntaker A (Y=1) 37
5 Låntaker B (Y=0) 37
6 Prosessen for å trene og evaluere modeller 65
7 AUC-plott og LASSO-stiplott for periode 1 66
8 AUC-plott og LASSO-stiplott for periode 2 67
9 AUC-plott og LASSO-stiplott for periode 3 68
10 AUC-plott og LASSO-stiplott for periode 4 69
11 AUC-plott og LASSO-stiplott for periode 5 70
12 AUC-plott og LASSO-stiplott for periode 6 71
13 AUC-plott og LASSO-stiplott for periode 7 72
14 AUC-plott og LASSO-stiplott for periode 8 73
15 AUC-plott og LASSO-stiplott for periode 9 74
16 AUC-plott og LASSO-stiplott for periode 10 75
17 SHAP-plott for periode 1 81
18 SHAP-plott for periode 2 82
19 SHAP-plott for periode 3 83
20 SHAP-plott for periode 4 84
21 SHAP-plott for periode 5 85
22 SHAP-plott for periode 6 86
23 SHAP-plott for periode 7 87
24 SHAP-plott for periode 8 88
25 SHAP-plott for periode 9 89
26 SHAP-plott for periode 10 90

Tabeller

1 Hyperparametere i XGBoost 31

2 Evaluering av modellenes prediksjonsevne på trenings- og testdata 35

3 Beskrivelse av variabler 62

4 Estimerte LR-koeffisienter for variabler valgt av LASSO 76

5 Optimaliserte hyperparametere i XGBoost 80

6 Beskrivende statistikk for variabler i endelig datasett 91

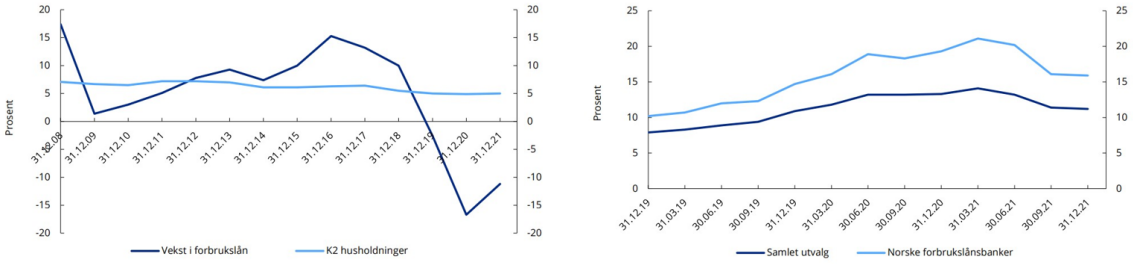
7 Beskrivelse av utelatte variabler 94

1 Introduksjon

1.1 Tema og aktualisering

Tema i denne masteroppgaven er mislighold av kredittkortgjeld, hvor mislighold defineres som et kontraktsbrudd som låntaker er ansvarlig for (Norges Bank, 2021, s. 112). Dette er et tema som har økt i aktualitet de seneste årene. Som grafen til venstre i figur 1 viser har det tidligere vært en økende trend av forbrukslån i Norge, altså en økning av både kredittkortgjeld og annen usikret gjeld. Denne veksten falt derimot fra 2017. I utgangen av 2018 var andelen misligholdte forbrukslån på 7,9%, mens andelen i utgangen av 2020 var på 13,3%. Selv om andelen misligholdte forbrukslån gikk ned i 2020, slik som grafen til høyre i figur 1 viser, er andelen fremdeles høy (Finanstilsynet, 2022, s. 4-8).

Figur 1: Utvikling i forbrukslån og mislighold (Finanstilsynet, 2022)



Grafen til venstre viser tolvmånedersvekst i det norske forbrukslånsmarkedet og husholdningenes innenlandsgjeld. Grafen til høyre viser misligholdt forbrukslån over 90 dager i prosent av totalt forbrukslån hos norske forbrukere fra henholdsvis norske forbrukslånsbanker (lyseblå) samt både norske og utenlandske banker og finansieringsforetak (mørkeblå).

Regjeringen (2017) har uttrykt bekymring over gjeldsbelastningen i private husholdninger. Spesielt er det uttrykt en bekymring for omfanget av den kostbare forbrukskreditten (Regjeringen, 2017). Mislighold av kredittkortgjeld er ikke bare et problem for finansinstutisjonene som påtar seg risiko ved utlån. Det er også et kostbart problem for hele samfunnet. Det offentlige bruker årlig rundt 120 millioner kroner på å avhjelpe gjeldsproblemer, mens gjeldsinstutisjonene har årlige tap på om lag én milliard kroner. Gjeldsproblemene

er det til slutt forbrukerne selv som må betale for, via økte renter (Regjeringen, 2019).

Den nevnte gjeldsproblematikken i Norge er årsaken til at myndighetene har innført flere tiltak for å regulere forbrukslånsmarkedet i løpet av de siste 5 årene (Norges Bank, 2021, s. 49). Gjeldsinformasjonsloven ble iverksatt juli 2019. Formålet med loven er å gi finansinstitusjonene et bedre grunnlag for kredittvurderingene som gjøres og dermed bidra til en mer forsvarlig utlånspraksis. Dette oppnås ved at långiverne får innsikt i hvor mye forbruksgjeld en kunde har fra før (Regjeringen, 2019). I 2021 ble utlånsforskriften iverksatt. Forskriften regulerer bankene ved å sette krav til utlånspraksisen av boliglån og forbrukslån. Det stilles blant annet krav til kundens betjeningsevne og gjeldsgrad (Finansdepartementet, 2021). Ifølge Finansdepartementet (2021) skal dette bidra til en mer bærekraftig utvikling av gjeld i husholdningene. Norske myndigheter har i tillegg innført flere andre reguleringer for kredittmarkedet. I 2017 innførte Finansdepartementet en forskrift om fakturering av kredittkortgjeld. Forskriften setter krav til finansforetakene om at fakturaen må inneholde kundens samlede utestående gjeld (Finansdepartementet, 2017). En forskrift om markedsføring av kreditt ble også innført i 2017. Hensikten er å hindre påtrengende markedsføring som villeder forbrukere vekk fra potensielt negative konsekvenser ved kredittbruk (Regjeringen, 2017).

1.1.1 Kredittvurdering

Et kredittkort innebærer at eieren av kortet har tilgang til kreditt. I stedet for at eierens konto blir debitert direkte, samler lånutgiveren kjøpene i en gitt periode og fakturerer for det samlede beløpet. Eieren av kortet kan velge å betale hele, deler eller ingenting av det utestående beløpet. Dersom eieren velger en av de to siste alternativene, vil det gjenstående beløpet inkluderes til neste periode og renter vil påløpe (Norges Bank, 2021, s. 79).

Lånutgivere tar en kredittrisiko ved å innvilge kredittkortlån. Begrepet kredittrisiko defineres som: «risikoen for tap knyttet til at en motpart ikke gjør opp sin forpliktelse verken til rett tid eller på et senere tidspunkt» (Gerdrup & Bakke, 2002, s. 196). I denne sammenheng utvikles modeller for kredittvurdering som estimerer sannsynligheten for mislighold hos låntaker eller lånsøker basert på flere karakteristikker ved personen. Slike modeller

skal bidra til å forbedre vurderingsprosessen for innvilgelse av lån (Yap mfl., 2011).

I dag er modeller for kredittvurdering svært utbredt, og nye teknikker blir stadig forsket på og tatt i bruk (Louzada mfl., 2016). Kredittvurderingsmodeller har lenge vært en del av vurderingsprosessen for innvilgelse av lån. Da kredittkortene kom for fullt på 1960-tallet, erfarte långiverne raskt at en manuell håndtering av alle søknadene var umulig. Det ble tydelig at prosessen måtte automatiseres. Automatiseringen ble en suksess og bidro til at modeller for kredittvurdering senere også ble benyttet til andre produkter slik som boliglån og små bedriftslån (Thomas, 2000). I 2018 trådte en ny lov om behandling av personopplysninger i kraft. Personopplysningsloven regulerer blant annet bruken av automatiserte individuelle avgjørelser. Enkeltpersoner har dermed rett til ikke kun å vurderes ved en automatisert prosess. Videre regulerer loven kravene til gyldig samtykke til behandling av personopplysninger, samt informasjonsrettigheter som rett til innsyn (Kommunal- og distriktsdepartementet, 2019). Ved anvendelse av kredittvurderingsmodeller for automatiserte beslutninger, må det dermed innlemmes tiltak for å opprettholde en søkers rettigheter til innsyn i beslutningen som blir tatt. I tillegg må kravet om at et menneske tar den endelige beslutningen, kunne etterfølges (Datatilsynet, 2018, s. 19).

Som statistikken og aktualiseringen viser, er temaet i vår masteroppgave dagsaktuelt og av interesse for både forbrukere, finansinstitusjoner og samfunnet som helhet. En av skribentene fikk selv innblikk i gjeldsproblematikken, da hun jobbet i Sparebank 1 Kreditt sommeren 2021. Dette ga spennende innsikt og videre motivasjon for denne masteroppgaven. Masteroppgaven er skrevet i samarbeid med Sparebank 1 Kreditt, som etter ønske vil bli kalt dataleverandør videre i oppgaven.

1.2 Problemstilling

Med bakgrunn i introduksjonen, har vår masteroppgave som formål å gi långivere av kredittkortgjeld en indikasjon på hvilke variabler som er viktige å hensynta i vurderingsprosessen av lånsøkere. Problemstillingen vi ønsker å belyse er dermed som følger:

«Hvilke karakteristikker ved lånsøkere påvirker sannsynligheten for mislighold av kredittkortgjeld?»

For å besvare problemstillingen utvikler vi prediksjonsmodeller ved bruk av to sett med teknikker. Logistisk regresjon (LR) benyttes for å utvikle den ene prediksjonsmodellen, sammen med Least Absolute Shrinkage and Selection Operator (LASSO) for å velge de viktigste variablene som skal inkluderes i LR-modellen. Den andre prediksjonsmodellen utvikles ved bruk av Extreme Gradient Boosting (XGBoost). For denne prediksjonsmodellen benyttes SHapley Additive exPlanations (SHAP) for å vurdere variabelviktigheten.

1.2.1 Forskningsspørsmål

For å underbygge vår problemstilling vil vi i tillegg se nærmere på følgende to forskningsspørsmål:

- *«Hvilken estimeringsteknikk gir best prediksjonsevne ved utvikling av modeller for kredittvurdering?»*
- *«Hvordan kan svarte bokser gjøres tolkbare for begrunnelse av kredittvurderinger?»*

Vår studie er avgrenset til å evaluere kredittkortsøknader innvilget av vår dataleverandør. Søknadene er begrenset til privatpersoner. Videre er modellene utviklet basert på data fra 47 625 lånsøkere i perioden november 2019 til desember 2021.

En nyvinning ved vår studie er bruken av et nytt og unikt datasett. Videre er studiets fokus på variabler og variabelviktighet et bidrag til litteraturen innen kredittvurdering, da tidligere forskning i liten grad diskuterer dette. I tillegg gjør vår studie et unikt funn da tre variabler, som tidligere ikke er nevnt i litteraturen, anses å være sentrale for prediksjon av mislighold av kredittkortgjeld. Dette er variabler som inneholder informasjon om hvorvidt en kredittkortsøknad er sendt inn på natten, sum av inngående beløp på debetkonto hos dataleverandør de siste 6 månedene, samt forholdet mellom søkt og innvilget kredittkortgrense.

1.3 Oppgavens struktur

I denne masteroppgaven vil relevant litteratur innen kredittvurdering gjennomgås i kapittel 2. Kapittelet vil først presentere ulike estimeringsteknikker benyttet ved kredittvurdering, før ulike metoder for variabelseleksjon og variabelviktighet, samt sentrale variabler

deretter gjennomgås. I kapittel 3 vil dataen som benyttes for å utvikle prediksjonsmodeller i vår studie beskrives. Videre vil kapittel 4 introdusere forskningsmetodene som er anvendt i masteroppgaven. I kapittel 5 presenteres resultatene, før resultatene drøftes videre i kapittel 6. Avslutningsvis følger vår konklusjon i kapittel 7, deretter referanseliste og appendiks.

2 Litteraturgjennomgang

Vi vil i dette kapittelet gjennomgå tidligere litteratur innen kredittvurdering. I delkapittel 2.1 presenteres forskjellige estimeringsteknikker benyttet til kredittvurdering, i delkapittel 2.2 gjennomgås ulike metoder for variabelseleksjon og variabelviktighet, mens i delkapittel 2.3 presenteres sentrale variabler. Litteraturen som benyttes omhandler i hovedsak kredittvurderinger av privatpersoner. Selv om vår studie kun ser på kredittkortgjeld, omhandler noe av litteraturen også andre lån. Enkelte studier innenfor konkursprediksjon for selskaper er også benyttet.

2.1 Estimeringsteknikker

Det er ikke et entydig svar på hvilken estimeringsteknikk som er best egnet til å utforme kredittvurderingsmodeller for best tolkbarhet og prediksjonsevne (Q. Wang mfl., 2018). Hvilken teknikk som er fordelaktig vil avhenge både av datastrukturen, variablene som brukes, samt formålet med klassifiseringen (Hand & Henley, 1997). Vi vil derfor i denne delen presentere noen av de mest brukte estimeringsteknikkene innenfor litteraturen om kredittvurdering.

Lineær diskriminantanalyse (LDA) ble først innført av Fisher (1936) som en klassifikasjonsteknikk (T.-S. Lee mfl., 2006). Durand (1941) anses gjerne for å være den første som benyttet LDA til å utforme en kredittvurderingsmodell og beviste med sin studie at metoden gir gode prediksjonsresultater for mislighold av kreditt (Hand & Henley, 1997). Altman (1968) var den første som benyttet multippel diskriminantanalyse, altså LDA med flere enn én variabel, for kredittvurdering. Han introduserte Z-score-modellen, som siden har blitt en av de mest kjente modellene for kredittvurdering (K. Li mfl., 2016). LDA har blitt benyttet i mange studier for å utforme kredittvurderingsmodeller, eksempelvis av Hand mfl. (1998), H. A. Abdou og Pointon (2011), samt Mahmoudi og Duman (2015). Bruken av LDA for kredittvurdering har likevel hatt en markant nedgang i løpet av 2000-tallet (Louzada mfl., 2016). Flere har stilt spørsmål ved hvorvidt LDA faktisk egner seg til å utvikle kredittvurderingsmodeller, blant annet på grunn av den kategoriske strukturen til kredittdata. I tillegg forutsetter LDA samme kovariansmatrise for de ulike klassene for

predikering, noe som sjelden vil være tilfellet (West, 2000). Videre forutsetter metoden at de uavhengige variablene er multivariat normalfordelt (Louzada mfl., 2016). Reichert mfl. (1983) hevdet likevel at et eventuelt brudd på antakelsen om normalfordeling ikke trenger å være en kritisk begrensning for å benytte metoden til kredittvurdering.

Lineær regresjon (LRE) ble først benyttet til kredittvurderinger av Myers og Forgy (1963), etterfulgt av Orgler (1970) (K. Li mfl., 2016). Metoden antar en lineær sammenheng mellom forklaringsvariablene og responsvariabelen, men har likevel blitt benyttet i kredittvurderingsmodeller hvor responsvariabelen er binær. Det er flere problemer knyttet til å bruke lineær regresjon på klassifikasjonsproblemer. Blant annet kan metoden gi verdier utover responsvariabelens verdier, noe som gjør resultatet mindre tolkbart. Resultatene kan dermed ikke direkte tolkes som sannsynligheten for det gitte utfallet (Louzada mfl., 2016). På tross av sine svakheter har metoden blitt brukt i flere studier for kredittvurderinger (Louzada mfl., 2016), eksempelvis av Hand (2002), Karlis og Rahmouni (2006), samt Efromovich (2010).

Logistisk regresjon (LR) er en statistisk teknikk som ofte benyttes i analyser av klassifikasjonsproblemer, hvor den avhengige variabelen er kvalitativ med distinkte verdier (James mfl., 2013, s. 130-137). LR benyttes særlig i tilfeller hvor den avhengige variabelen er binær (James mfl., 2013, s. 130-137). Teknikken er dermed også mye brukt ved utvikling av kredittvurderingsmodeller (H. A. Abdou mfl., 2016). LR ble først introdusert som en estimeringsteknikk for kredittrisiko av Wiginton (1980). I Louzada mfl. (2016) sin litteraturgjennomgang fremgår det at LR overordnet er den 5. mest brukte teknikken i litteraturen innen kredittvurdering i perioden mellom 1992 og 2015, men at det fra og med 2012 var en økning i bruken. LR er bevist å være en tilsvarende effektiv og treffsikker estimeringsteknikk som LDA (Harrell & Lee, 1985). Videre vil det være fordelaktig å benytte LR fremfor LRE ved klassifikasjon, da LR gir verdier innenfor intervallet til responsvariabelen. Selv om enkelte nyere estimeringsteknikker har vist seg å ha bedre prediksjonsevne, blir LR fremdeles i stor grad benyttet i litteraturen innenfor kredittvurdering hovedsakelig på grunn av sin tolkbarhet (Dong mfl., 2010). Tolkbarheten, samt metodens robusthet, er årsakene til at LR er den mest brukte estimeringsteknikken innen bankindustrien (Dong mfl., 2010). Bruken av LR har likevel blitt kritisert da også denne

metoden gjør noen strenge modellforutsetninger om blant annet homoskedastisitet, som kan redusere egnetheten for bruk i kredittvurderingsmodeller (T.-S. Lee mfl., 2006). Metoden anses likevel å ha mindre strenge forutsetninger enn både LDA og LRE, da LR verken antar at de uavhengige variablene må være normalfordelte, lineært relatert med responsvariabelen eller at det må være samme kovariansmatrise for de ulike klassene for predikering (Tabachnick & Fidell, 2019). Det er gjennomført mange studier hvor LR er benyttet for å estimere mislighold av kredittkortgjeld, blant annet av T. H. Lee og Jung (1999), H. Abdou mfl. (2007) og Lessmann mfl. (2015). LR har i mange studier også blitt benyttet som et sammenligningsgrunnlag mot andre estimeringsteknikker, eksempelvis av Yap mfl. (2011) og Pavlidis mfl. (2012).

Desisjonstrær er en populær klassifikasjonsmetode som predikerer basert på beslutningsregler organisert med en trebasert arkitektur. Desisjonstrær kan benyttes til både regresjon og klassifikasjon, avhengig av om responsvariabelen har henholdsvis en kontinuerlig eller kategorisk verdi (James mfl., 2013, s. 302). Klassifikasjons- og regresjonstrær (CART) ble først introdusert av Breiman mfl. (1984) og har vist seg å inneha flere fordeler ved utvikling av modeller for kredittvurdering (T.-S. Lee mfl., 2006). I motsetning til LDA, LRE og LR, gjør CART ingen antakelser om underliggende fordelinger (Breiman mfl., 1984). I tillegg kan metoden identifisere viktige forklaringsvariabler når flere variabler og potensielle trær vurderes. En annen sentral fordel med CART er tolkbarheten (T.-S. Lee mfl., 2006). Toklbarheten går derimot ofte på bekostning av modellenes prediksjonsevne. Prediksjonsevnen kan derimot vesentlig forbedres ved å sammenstille mange trær, eksempelvis ved bruk av metoden Random forest (James mfl., 2013, s. 316). Se Breiman (2001) for mer detaljert utledelse av metoden. Flere studier har benyttet samt bevist at metoden fungerer godt for utvikling av modeller for kredittvurdering, blant annet Brown og Mues (2012) og Lessmann mfl. (2015). Studien til Lessmann mfl. (2015) er en sammenligning av 41 ulike estimeringsteknikker, basert på 6 ulike evalueringsmål. Studien konkluderte med at Random forest med fordel bør benyttes, fremfor LR, ved sammenligning av ulike metoder for utvikling av kredittvurderingsmodeller (Lessmann mfl., 2015). Siden studien til Lessmann mfl. (2015) har det kommet ytterligere estimeringsteknikker som har vist gode prediksjonsevner, deriblant den trebaserte estimeringsteknikken XGBoost (Gunnarsson

mfl., 2021). Metoden ble utviklet av T. Chen og Guestrin (2016) og utnytter Gradient Boosting for å trene flere desisjonstrær (Gunnarsson mfl., 2021). XGBoost har oppnådd gode prediksjonsresultater ved estimering av kredittvurderingsmodeller, blant annet i studiene til Xia mfl. (2017), M. Wang mfl. (2018) samt Mushava og Murray (2022).

Multivariate adaptive regression splines (MARS) er en klassifikasjonsteknikk introdusert av Friedman (1991). Metoden ble presentert som en bedre løsning for fleksibel ikke-parametrisk regresjonsmodellering, spesielt ved høydimensjonalitetsdata. MARS bygger på en blanding av CART og generaliserte additive modeller (Friedman, 1991). Da MARS er en ikke-parametrisk metode, kan den benyttes for å modellere komplekse sammenhenger uten å gjøre noen eksplisitte forutsetninger om funksjonsformen til modellen (T.-S. Lee mfl., 2006). Videre utmerker metoden seg i evnen til å identifisere sentrale forklaringsvariabler og interaksjoner, samt den underliggende datastrukturen i høydimensjonale data (T. Lee & Chen, 2005). Metoden trenger ikke lang treningstid og gir resultater som er enkle å tolke (T.-S. Lee mfl., 2006). Det er videre bevist at MARS kan være nyttig i kombinasjon med nevrale nettverk, da fordelene ved metoden kan overkomme svakhetene ved nevrale nettverk (T. Lee & Chen, 2005). Basert på sine fordeler har metoden vist seg å være et nyttig verktøy for prediksjon og dermed blitt brukt i flere studier for vurdering av kredittrisiko (T.-S. Lee mfl., 2006), eksempelvis av Xiao mfl. (2006) og Afrilia mfl. (2021).

Nevrale nettverk (NN) er et nyttig alternativ til flere av de ovennevnte teknikkene, da NN har vist seg å gi mer treffsikre resultater (Chuang & Lin, 2009). NN er først og fremst fordelaktig i tilfeller hvor det er en ikke-lineær relasjon mellom den avhengige variabelen og forklaringsvariablene i modellen (T.-S. Lee mfl., 2006). Flere studier har sett på bruken av NN i modeller for kredittvurdering, blant annet West (2000), Marques mfl. (2012), samt Abellán og Mantas (2014). Det er likevel svakheter også ved denne metoden. Spesielt krever det mye tid å trene modeller med bruk av NN. Videre er det vanskelig å identifisere den relative betydningen av potensielle variabler. Disse egenskapene gjør metoden mindre egnet for utforming av kredittvurderingsmodeller, da årsaken til et eventuelt avslag på en kredittsøknad gjerne må begrunnes (Piramuthu, 1999; West, 2000; Akkoç, 2012).

Støttevektormaskiner (SVM) benyttes for klassifisering og ble introdusert av Vapnik (1995). Metoden er utviklet for klassifikasjon av binære responsvariabler (James mfl.,

2013, s. 337) og klassifiserer ved hjelp av et hyperplan (Louzada mfl., 2016). SVM er robust mot overtilpasning (S. Li mfl., 2006), i tillegg til å fungere bra ved små utvalg (Ravi Kumar & Ravi, 2007). Algoritmen er derimot kompleks og krever derfor mye beregningstid (Ravi Kumar & Ravi, 2007). SVM har suksessfullt blitt brukt til prediksjon innenfor flere fagfelt, før metoden senere også har blitt brukt til å utforme kredittvurderingsmodeller (Akkoç, 2012). Flere studier har benyttet estimeringsteknikken til å utvikle kredittvurderingsmodeller, blant annet S. Li mfl. (2006), Bellotti og Crook (2009) og Zhou mfl. (2010). Litteraturgjennomgangen til Louzada mfl. (2016) viser at SVM var den mest brukte estimeringsteknikken innenfor kredittvurderingsmodeller i perioden mellom 2006 og 2010, i tillegg til at metoden overordnet var den 4. mest brukte teknikken i litteraturen i perioden mellom 1991 og 2015.

Hybrider og **ensemblemetoder** har blitt mer vanlige estimeringsteknikker for utvikling av kredittvurderingsmodeller (Louzada mfl., 2016). Hybride metoder kombinerer ulike tradisjonelle metoder for å forbedre modellenes prediksjonsevne. I litteraturen finner vi eksempler på kredittvurderingsmodeller utviklet som en hybrid mellom diskriminantanalyse og NN (T.-S. Lee mfl., 2002), NN og MARS (T. Lee & Chen, 2005), samt genetisk algoritme og SVM (Huang mfl., 2007). Ensemblemetoder kombinerer ulike klassifikasjonsmetoder for å forbedre ytelsen til kredittvurderingsmodellene. Bagging, boosting og stacking er tre populære grupper av ensemblemetoder (Louzada mfl., 2016). Eksempler på bruken av slike metoder for kredittvurdering finner vi blant annet i studiene til Hsieh og Hung (2010), Marqués mfl. (2012) og Xu mfl. (2019).

2.2 Metoder for analyse av variabler

2.2.1 Metoder for variabelseleksjon

Ved å benytte metoder for variabelseleksjon (feature selection) kan det utvikles mer treffsikre og tolkbare kredittvurderingsmodeller. Disse metodene reduserer eventuelle problemer som følge av høy dimensjonalitet i data, i tillegg til å ekskludere irrelevante variabler fra den endelige modellen, gi en bedre forståelse av datastrukturen og bidra til å unngå overtilpasning (Tripathi mfl., 2020; Paraschiv mfl., 2021). Metodene for variabelseleksjon deles tradisjonelt inn i tre ulike metoder: filtermetoder, wrapper-metoder og embedded

metoder (Chandrashekar & Sahin, 2014). Det er mye tilgjengelig litteratur hvor metoder for variabelseleksjon benyttes (J. Li mfl., 2018; Q. Wang mfl., 2018), også innenfor studiene om kredittvurderingsmodeller (Q. Wang mfl., 2018). Ara, Fernandes og Louzada (2016) sin litteraturgjennomgang viser at 51% av de gjennomgåtte studiene benyttet metoder for variabelseleksjon, selv om verken studienes formål eller hovedfokus nødvendigvis var på bruken av disse metodene.

Filtermetodene velger ut de viktigste forklaringsvariablene basert på et forhåndsdefinert kriterium som måler sammenhengen mellom hver enkelt uavhengig variabel og responsvariabelen (Paraschiv mfl., 2021). Eksempler på kriterium som kan benyttes i filtermetoder er kjikvadrattesten (Laborda & Ryoo, 2021) og Pearsons korrelasjonskoeffisient (Paraschiv mfl., 2021). Filtermetoder er uavhengig av estimeringsteknikken som benyttes (J. Li mfl., 2018), noe som er fordelaktig om estimeringsteknikken er påvirket av systematiske skjevheter (Paraschiv mfl., 2021). Videre krever filtermetodene svært liten beregningstid og unngår problemer med overtilpasning (Chandrashekar & Sahin, 2014). Filtermetoder har blitt benyttet i flere studier for å utvikle kredittvurderingsmodeller, blant annet i studiene til Liu og Schumann (2005), Somol mfl. (2005), samt F.-L. Chen og Li (2010). Ingen av de nevnte studiene finner filtermetodene som mest fordelaktige ved sammenligning med andre metoder for variabelseleksjon, noe som kan skyldes svakhetene ved metoden. For det første tar ikke filtermetoder for seg relasjonene mellom variablene. Videre, i de tilfeller hvor korrelasjon brukes som kriterium, tar filtermetoder kun hensyn til lineære sammenhenger mellom de uavhengige variablene og responsvariabelen. At metodene er uavhengig av estimeringsteknikken som brukes, kan også være en ulempe. Uavhengigheten kan føre til suboptimal tilpasning og prediksjon, dersom skjevhetene ved estimeringsteknikken overses (Paraschiv mfl., 2021).

Wrapper-metoder er iterative metoder som, i motsetning til filtermetoder, forsøker å identifisere de viktigste forklaringsvariablene ved å finne den beste sammensetningen av variabler som gir høyest treffsikkerhet. Recursive feature elimination method, exhaustive search og greedy forward search er eksempler på wrapper-metoder som er mye brukt (Ahmad & Starkey, 2018). Liu og Schumann (2005), Somol mfl. (2005), samt F.-L. Chen og Li (2010) finner at wrapper-metoder er bedre enn filtermetoder. Wrapper-metodene

har likevel også noen klare svakheter. Metodene vil ofte føre til overtilpasning, i tillegg til at de krever mye beregningstid (Q. Wang mfl., 2018). Dette gjør at metodene blir mindre brukt i praksis (J. Li mfl., 2018).

Embeddede metoder kan løse flere av svakhetene ved både filtermetoder og wrapper-metoder (Paraschiv mfl., 2021). Embedded metodene inkluderer interaksjonen med læringsalgoritmene, i tillegg til å være mye mer effektive enn wrapper-metodene da de ikke er avhengige av å evaluere variabler iterativt. De mest brukte embedded metodene er de som forsøker å minimere modellenes tilpasningsfeil, i tillegg til å straffe modeller ved for mange inkluderte variabler (J. Li mfl., 2018). LASSO, ridge regression og elastiske nett er de mest populære embedded metodene (Laborda & Ryoo, 2021). Flere studier innen litteraturen for kredittvurdering har benyttet ulike embedded metoder for å velge sentrale variabler, eksempelvis H. Chen og Xiang (2017), J. Li mfl. (2020) og Maldonado mfl. (2017).

2.2.2 Metoder for variabelviktighet

Flere maskinlæringsteknikker fungerer som svarte bokser, da metodene ikke gir innsikt i de interne prosessene og samspillene i de estimerte modellene. Slike metoder gir ofte mer treffsikre prediksjonsmodeller, men da på bekostning av modellenes tolkbarhet (James mfl., 2013, s. 8). Å kunne tolke slike modeller er av avgjørende betydning innenfor flere fagfelt, også innenfor kredittvurdering (Kvamme mfl., 2018). Å forstå hvordan modeller fungerer og hvorfor de fungerer slik de gjør, vil bidra til å styrke prediksjonene samt menneskers tillit til modellene (Ribeiro mfl., 2016b).

Når maskinlæringsteknikker fungerer som svarte bokser, er en mulig løsning å benytte såkalte modellagnostiske metoder. Dette er modelluavhengige metoder, som benyttes for å tolke de estimerte svarte boksene (Ribeiro mfl., 2016b). Metodene kan dermed benyttes til å forklare hvordan og i hvilken grad forklaringsvariablene i en svart boks bidrar til modellens prediksjoner (Casalicchio mfl., 2019, s. 656). Slike metoder kan med andre ord benyttes til å analysere variabelviktigheten, både på et globalt og lokalt nivå (Aas mfl., 2021).

Globale metoder benyttes for å beskrive modeller i sin helhet, ved å si noe om hvilke

forklaringsvariabler som overordnet påvirker modellene mest (Aas mfl., 2021). De globale metodene kan dermed benyttes til å si noe om forklaringsvariablenes gjennomsnittlige bidrag til modellers prediksjoner (Molnar, 2022, kap. 8). En mye brukt metode er Partial Dependence Plot (PDP) (Aas mfl., 2021). PDP viser den marginale effekten en eller to uavhengige variabler har på det predikerte utfallet (Friedman, 2001). Metoden er svært intuitiv, men er samtidig begrenset til data med lav dimensjonalitet (Friedman, 2001). I tillegg forutsettes det at ingen forklaringsvariabler er korrelerte, noe som sjelden vil være tilfellet (Molnar, 2022, kap. 8.1). En annen global metode, Permutation Feature Importance (PFI), måler endringen i modellenes prediksjonsfeil etter beregning av permuterte verdier for de uavhengige variablene (Molnar, 2022, kap. 8.5). En variabel vil være viktig dersom permutasjonen fører til en økning i prediksjonsfeilen, da økningen indikerer at variabelen påvirker modellens prediksjonsutfall (Kamath & Liu, 2021, s. 187). PFI er i likhet med PDP enkel å tolke, men kan i likhet også gi skjeve estimater dersom de uavhengige variablene er korrelerte. Ettersom PFI ikke er avhengig av at modellene trenes flere ganger, krever metoden liten beregningstid (Molnar, 2022, kap. 8.5). For å unngå feilaktige konklusjoner om hvilke forklaringsvariabler som er viktige, bør PFI med fordel benyttes på testdata (Kamath & Liu, 2021, s. 188).

Lokale metoder forsøker å forklare individuelle prediksjoner (Molnar, 2022, kap. 9). Slike metoder er spesielt nyttige i tilfeller hvor modeller gir ulike prediksjoner for ulike variabelsammensetninger, altså i tilfeller hvor de globale metodene ikke er representative for spesifikke tilfeller (Aas mfl., 2021). Local Interpretable Model-agnostic Explanations (LIME) og SHAP er to mye brukte metoder innenfor kredittvurdering (Provenzano mfl., 2020). LIME ble utledet av Ribeiro mfl. (2016a). Målet er å finne en tolkbar og troverdig modell lokalt rundt prediksjonen. En fordel med LIME er at den lokale tolkbare modellen ikke trenger å estimeres ved bruk av samme teknikk som ble benyttet for prediksjonen. Dette muliggjør at den teknikken som gir best tolkbarhet kan benyttes (Molnar, 2022, kap. 9.2). LIME må likevel benyttes med forsiktighet, blant annet fordi det ikke finnes en god fremgangsmåte for å definere maksimalt antall observasjoner som skal benyttes for å forklare den individuelle prediksjonen (Molnar, 2022, kap. 9.2). SHAP ble utledet av Lundberg og Lee (2017). SHAP-verdier blir beregnet ved å se hvor mye prediksjonen

endres når en forklaringsvariabel er ekskludert fra modellen. Dette gjøres for enhver prediksjon og forklaringsvariabel, noe som bidrar til at både konsistens og lokal nøyaktighet oppnås (Provenzano mfl., 2020). Dette skiller SHAP fra LIME, da sistnevnte kun inkluderer en andel av de nærmeste observasjonene til den prediksjonen som metoden er ute etter å forklare (Molnar, 2022, kap. 9.6). Videre viste studien til Lundberg og Lee (2017) at SHAP-verdier samsvarer bedre med menneskelige forklaringer, sammenlignet med LIME (Aas mfl., 2021). SHAP kan også benyttes til globale forklaringer da teknikken i tillegg beregner gjennomsnittlige SHAP-verdier for hver forklaringsvariabel (Molnar, 2022, kap. 9.6). Den største begrensningen ved SHAP er antallet variabler, da beregningstiden vokser eksponensielt (Aas mfl., 2021).

2.3 Sentrale variabler

Vår masteroppgave har som mål å identifisere karakteristikker ved lånsøkere som er av betydning for mislighold av kredittkortgjeld. Det er derfor relevant å se på hvilke variabler som brukes i kredittvurderingsmodeller i tidligere litteratur. En vanlig fremgangsmåte ved utvikling av prediksjonsmodeller er å utforske et stort antall forklaringsvariabler, for deretter å identifisere et mindre utvalg av karakteristikker som brukes i praksis. I kredittvurderingsmodeller er det vanlig å starte med 50 eller flere variabler, for deretter å velge ut 10-12 variabler i det endelige utvalget (Hand & Henley, 1997).

Steenackers og Goovaerts (1989) utviklet en LR-modell for kredittvurdering og fant blant annet at en låntakers alder, lengde på nåværende arbeidsforhold, yrke, månedlig inntekt og hvorvidt personen eier bolig er signifikante variabler. Hand og Henley (1997) presenterte i sin studie en oversikt over 14 variabler som ofte inngår i kredittvurderingsmodeller. I oversikten inngår flere av de samme forklaringsvariablene som nevnt i den foregående studien, i tillegg til en låntakers sivilstatus, bosituasjon, årlig inntekt, samt lengde på nåværende boforhold og kundeforhold til banken. Somol mfl. (2005) beviste at en kredittvurderingsmodell bestående av et utvalg på 5 til 10 forklaringsvariabler, gir signifikant bedre resultater enn en modell utviklet med hele det originale datasett. Videre fant de, i likhet med de ovennevnte studiene, at alder, bosituasjon, sivilstatus, samt lengde på nåværende arbeidsforhold er viktige variabler. I tillegg ble kredittgrense og tidligere be-

talingshistorikk valgt som sentrale variabler av metodene for variabelseleksjon brukt i studien. Videre fant Bellotti og Crook (2009) at blant annet låntakers alder, lengde på kundeforholdet til banken, eierskap av bolig og type kredittkort er av signifikant betydning for å avgjøre risikoen for mislighold. Akkoç (2012) benyttet LDA og LR som metoder for variabelseleksjon. Sivilstatus, utdanningsnivå, total arbeidsvarighet og lengde på nåværende arbeidsforhold ble funnet å være signifikante variabler i alle perioder for de to metodene.

Chiang mfl. (2002) påpekte at individuelle karakteristikker ved låntakere bør benyttes for å predikere risikoen for mislighold. Crook og Banasik (2004) vektla, i tillegg til de demografiske variablene, viktigheten av at variablene i en prediksjonsmodell ikke kun baseres på historiske lånedata om kunder. Tsai mfl. (2009) viste i sin studie at treffsikkerheten til prediksjonsmodeller for mislighold øker, dersom en låntakers holdning til penger inngår i modellen. Tsai mfl. (2009) hevdet derfor at sanntidsinformasjon om låntakerne, ikke kun historiske data, burde inkluderes i prediksjonsmodellen. Deres resultater viste at modellene predikerer bedre ved å inkludere sanntidsvariabler som sier noe om låntakerens holdning til penger. Informasjonen ble innhentet ved bruk av et spørreskjema utviklet av Kent T. Yamauchi og Donald J. Templer i 1982, kalt Money and Attitude Scale (MAS) (Tsai mfl., 2009). Skjemaet inneholder spørsmål som skal kartlegge 5 faktorer som beskriver kundens forhold til penger. Disse inkluderer kundens opplevelse av penger som kilde til prestisje, angst eller frykt, samt kundens tendens til nøye økonomisk planlegging og å ha en nølende holdning til situasjoner som inkluderer penger (Lejoyeux mfl., 2011).

T.-S. Lee mfl. (2006) utviklet flere kredittvurderingsmodeller ved bruk av ulike estimeringsteknikker hvor CART og MARS presterte best. Modellene ble utviklet på et datasett basert på bankinformasjon fra Taiwan og inkluderte 9 forklaringsvariabler for hver låntaker: kjønn, alder, sivilstatus, utdanningsnivå, yrke, jobbposisjon, årlig inntekt, bosituasjon og kredittgrense. Modellen utviklet med CART fant at kundens yrke og kredittgrense er de viktigste variablene. De samme variablene, i tillegg til bosituasjon, ble valgt av modellen utviklet med MARS. Sariannidis mfl. (2019) identifiserte nøkkelfaktorer ved kunder som misligholder. Dette ble gjort ved hjelp av et datasett som også er basert på bankinformasjon fra Taiwan, i perioden mellom april og september 2005. Datasettet bestod

av 23 forklaringsvariabler basert på demografi, samt regnskap- og kredittinformasjon om hver enkelt kunde. Blant annet inngikk variabler som alder, kjønn, sivilstatus, utdanning, innvilget kredittgrense og tilbakebetalingsstatus på kredittkort per måned. På tvers av de syv metodene som ble brukt i studien, er tilbakebetalingsstatusen på kredittkortet den viktigste forklaringsvariabelen. Variabelen inneholder informasjon om en kundes gjeldsstatus 1 måned før kredittvurdering og kan dermed sies å fungere som en indikator på mulig mislighold. Tilsvarende funn har også blitt gjort i studier av Dimitras mfl. (2017) og Hamori mfl. (2018). Videre fant Sariannidis mfl. (2019), noe overraskende, at fakturabeløp og summen av tidligere betalinger er mindre viktige.

Selv om det varierer hvilken informasjon som er tilgjengelig for forskere ved utvikling av kredittvurderingsmodeller (Lessmann mfl., 2015), er det fremdeles likheter i forklaringsvariablene som blir brukt i modellene (Kvamme mfl., 2018). I litteraturen benyttes ofte informasjon hentet fra blant annet kredittbyråer. Dette er variabler som inneholder informasjon om antall utestående og forfalte kontoer, i tillegg til saldo på andre lån. Videre benyttes gjerne informasjon om enkeltindividers månedlige inntekt, saldo, samt demografiske variabler som alder og sivilstatus (Kvamme mfl., 2018).

3 Data

Datasettet vi benytter for å utvikle prediksjonsmodeller er levert av en stor norsk leverandør av kredittkort til det norske privatmarkedet. Datasettet består opprinnelig av 47 712 observasjoner, hvor hver observasjon er en person som har fått innvilget kredittkort i perioden november 2019 til februar 2021.

3.1 Variabler

For hver observasjon er det oppgitt statistisk informasjon om lånsøkeren på søknadstidspunktet, gitt ved datasettets 72 forklaringsvariabler. Noen av forklaringsvariablene baserer seg på data hentet fra eksterne registre, som kjøretøyregisteret, folkeregisteret, skatteregistret, samt gjeldsregisteret. Dersom lånsøker var kunde hos dataleverandør ved søknadstidspunktet, er det i tillegg innhentet allerede eksisterende kundeinformasjon. Videre er resterende forklaringsvariabler basert på informasjon lånsøkerne selv har fylt inn, data fra kredittbyrå og beregninger gjort både i løpet av søknadsprosessen samt i ettertid. Dataen er anonymisert og den eneste personlige karakteristikken ved lånsøkeren er fødselsår. Av hensyn til dataleverandør er også enkelte variabelnavn anonymisert.

Responsvariabelen i datasettet er en dummyvariabel for hvorvidt en person har misligholdt sin kredittkortgjeld innen ett år etter godkjent kredittkortsøknad eller ikke. I vår studie defineres mislighold av kredittkortgjeld som at kravet er sendt til inkasso, da dette er definisjonen vår dataleverandør selv benytter. Vi setter verdien til 1 dersom kredittkortgjelden er misligholdt innen ett år og 0 hvis ikke. Variabelen er basert på data fra og med januar 2020, til og med desember 2021. Begrunnelsen for den korte utfallsperioden for mislighold, baserer seg på et ønske fra vår dataleverandør. Deres data tilsier at de fleste som misligholder, gjør det i løpet av det første året. I det opprinnelige datasettet er 1 182 (2,48%) av de 47 712 observasjonene, klassifisert som misligholdt.

3.1.1 Variabeltransformasjon

For at dataen skal gi gyldige analyseresultater er det nødvendig å behandle datasettet for å fjerne eventuelle feil, duplikater, mangler og potensielt villedende informasjon. Alle

variabler er nøye gjennomgått og evaluert, før eventuelle endringer er gjennomført.

Basert på enkelte variabler i det opprinnelige 2 dummyvariabler basert på søknadstidspunkt, for å innhente mer informasjon om når på døgnet og når i uken søknadene ble sendt inn. Dummyvariablene lages etter forslag fra vår dataleverandør, da de er interessert i å vite om søknadstidspunkt er av betydning ved prediksjon av mislighold av kredittkortgjeld. Vi lager dummyvariabelen *«App_ Night»* som får verdien 1 dersom søknaden er sendt inn på natten, noe vi definerer til å være i tidsrommet mellom klokken 22.00 og 07.00. Videre lager vi *«App_ Weekend»* som gir informasjon om kunden har sendt inn søknaden på en ukedag eller helg. Vi setter verdien til 1 dersom søknaden er sendt inn en helg, og 0 hvis ikke.

I tillegg foreslo vår dataleverandør å undersøke om antall timer fra en søknad ble mottatt til den ble godkjent, er av relevant betydning. Vi lager derfor variabelen «*Diff_ Hour*». Variabelverdien beregnes ved differansen mellom søknadstidspunkt og tidspunktet da kunden signerte søknaden.

Videre har vi en variabel i det opprinnelige datasettet som angir datoen for når en lånsøker ble kunde hos dataleverandør. Datoen er på formatet yy-mm-dd, og varierer i stor grad. På dette formatet vil variabelen dermed gi lite informasjon i analysen. Vi lager derfor en ny variabel, «Year_Customer», som angir antall år lånsøkeren har vært kunde hos dataleverandør. Verdiene beregnes ved å finne differansen mellom året for det eldste registrerte kundeforholdet og året da kredittkortsøknaden ble opprettet.

«*Age_ log*» angir logaritmen av lånsøkerens alder. Alderen beregnes ut fra lånsøkerens fødselsår, definert som datoen for innvilget lån minus fødselsdatoen til personen. Deretter gjennomfører vi en logaritmisk transformasjon av alderen. Dette gjøres for å fjerne ekstremverdier.

Til slutt har vi «*Car*», «*MC*» og «*Van*». I det opprinnelige datasettet er det tre variabler som angir henholdsvis antall biler, motorsykler og varebiler som en lånsøker har registrert i bilregisteret. Denne registerdataen oppdateres ikke ved salg av kjøretøy, noe som fører til at det registrerte antallet sjelden stemmer. Antallet registrerte kjøretøy vil derfor ofte være kunstig høy og dermed ikke pålitelig ved gjennomførelse av analyser. Vi

velger derfor å benytte dummyvariabler som en løsning på problemet. Dummyvariablene angir om lånsøkeren har, eller tidligere har hatt, en eller flere biler/motorsykler/varebiler i bilregisteret.

For redusere effekten av ekstremverdier velger vi å gjennomføre winsorizing, i likhet med Shumway (2001). Da winsorizing på 1. og 99. persentil fremdeles gir mange ekstremverdier i vår data, velger vi heller å winsorize på 5. og 95. persentil slik som Nyitrai og Virág (2019). Videre benyttes One-Hot-Encoding på alle kategorivariabler da estimeringsteknikkene vi bruker krever at forklaringsvariablenes verdier er nummeriske. Dette innebærer at hver kategori innenfor en kategorivariabel, transformeres til en ny variabel med verdier 0 og 1.

3.1.2 Utelatelse av variabler

Fra det opprinnelige datasettet velger vi å utelate to variabler med ID-informasjon, da disse er unike for hver lånsøker og dermed ikke kan benyttes til prediksjon. Da variablene er unike, vil de ikke være nyttige for å undersøke hvilke karakteristikker ved lånsøkere som er sentrale for prediksjon av mislighold av kredittkortgjeld. Videre er det i det opprinnelige datasettet inkludert datovariabler for når søknadene ble opprettet og godkjent. For begge tilfeller er det inkludert flere variabler, men på ulikt datoformat. Vi velger derfor kun å inkludere to datovariabler basert på godkjenningstidspunktet. Videre blir variabelen «*Sum_Pledge_Remarks_AMT*», altså sum på heftelser, utelatt da tilnærmet 80% av observasjonene er manglende for denne variabelen. Enkelte variabler er utelatt da informasjonen de gir, også fremgår av andre variabler i datasettet. I tillegg blir alle observasjoner fra 2021 utelatt fra det endelige datasettet, da det er få observasjoner for dette året. Totalt slettes 86 observasjoner av denne årsaken. En fullstendig oversikt over utelatte variabler finnes i appendiks A.8.

3.2 Håndtering av manglende verdier

I det opprinnelige datasettet er det mange manglende verdier. En vanlig måte å håndtere dette problemet på, er å fjerne alle observasjoner som inneholder manglende verdier (Studenmund & Johnson, 2017, s. 364). Ved å gjennomføre en slik sletting i det opprinnelige datasettet ville vår data kun bestått av resterende 2 841 observasjoner. En slik håndte-

ring av manglende verdier er dermed ikke aktuell i vårt tilfelle. I tillegg vil utelatelse av observasjoner med manglende verdier kunne føre til bias, da analyseutvalget blir redusert. Det vil dermed ikke lengre være selvsagt at det resterende utvalget er representativt for populasjonen (Ringdal, 2018, s. 280).

Som en løsning på de nevnte problemene knyttet til de manglende verdiene, benytter vi matching som tilbøyelighetsanalyse (Propensity Score Analysis). Metoden setter alle manglende verdier for en observasjon, lik gjennomsnittet av alle liknende observasjoner. Hva som er liknende observasjoner avgjøres basert på andre forklaringsvariabler som ikke har manglende verdier for de gjeldende observasjonene og som anses å utgjøre et rimelig sammenligningsgrunnlag. Vi velger å inkludere 7 forklaringsvariabler som sammenligningsgrunnlag. Variablene inneholder informasjon om en lånsøkers kredittgrense, månedlig nettoinntekt, antall innslag i gjeldsregisteret, alder, sivilstatus, ansettelsestype og bosituasjon, hvor de tre sistnevnte er one-hot-encodet.

Algoritmen vi benytter er den ikke-parametriske metoden k-nærmeste naboer (k-NN). k-NN plotter først alle observasjonene i et utfallsrom (James mfl., 2013, s. 39). Forklaringsvariablene i vårt tilfelle er de 7 ovennevnte variablene. Gitt en observasjon x_i, vil algoritmen først identifisere de K nærmeste observasjonene til x_i (James mfl., 2013, s. 39). Avstanden fra x_i til de K nærmeste naboene beregnes som regel ved bruk av euklidsk avstand. Basert på de K nærmeste observasjonene, beregnes gjennomsnittsverdien som skal erstatte den manglende verdien. Valget av K vil dermed ha stor betydning for hvilken verdi som beregnes. Hvilken K som er optimal, vil i stor grad avhenge av problemet som skal løses og dermed være en avveining mellom bias og varians (James mfl., 2013, s. 42). Vi velger K lik 5 for kontinuerlige variabler. I tillegg har vi én dummyvariabel hvor k-NN benyttes. Her benytter vi K lik 1, for å unngå gjennomsnittsverdier mellom de to kategoriene.

3.3 Håndtering av ubalansert data

Datasettet vi benytter er ubalansert da andelen misligholdt er mindre enn andelen ikke misligholdt. Dette kan ha en negativ effekt på prediksjonsevnen til maskinlæringsmodeller. En måte å håndtere den ubalanserte dataen på, er å benytte såkalte resampling-

teknikker (Lessmann mfl., 2015). Dette er teknikker som endrer datasettet slik at andelen misligholdt kredittkortgjeld blir større (over-sampling) og/eller andelen ikke misligholdt kredittkortgjeld blir mindre (under-sampling). Slike teknikker kan derimot skape bias i prediksjonsmodellene siden de trenes på et datasett med et annet forholdstall mellom misligholdt og ikke misligholdt, sammenlignet med den virkelige verden som modellene skal brukes på. Vi velger derfor ikke å benytte resampling-teknikker for å håndtere det ubalanserte datasettet. Dette er i likhet med flere andre studier innenfor økonomisk litteratur som predikerer mislighold eller konkurs: Shumway (2001), Chava og Jarrow (2004), Tian mfl. (2015), Tian og Yu (2017), Provenzano mfl. (2020) og Paraschiv mfl. (2021).

3.4 Endelig datasett

Det endelige datasettet som blir brukt til å utvikle prediksjonsmodeller i denne masteroppgaven består dermed av 47 625 observasjoner og 106 forklaringsvariabler. Det er totalt 1 179 observasjoner som er klassifisert som misligholdt, tilsvarende en andel på 2,48%. Andelen misligholdt er dermed ikke redusert fra det opprinnelige datasettet. En oversikt over alle variablene som inngår i det endelige datasettet finnes i appendiks A.1.

4 Metode

I dette kapittelet vil vår forskningsmetode presenteres. Først vil metoden som benyttes for kryssvalidering gjennomgås i delkapittel 4.1. Delkapittel 4.2 presenterer LR og XGBoost, som brukes til å estimere kredittvurderingsmodeller i denne oppgaven. Videre presenteres metodene som benyttes for å analysere variablenes viktighet, hvor LASSO benyttes med LR og SHAP med XGBoost. Evalueringsmålene som benyttes i studien beskrives i 4.4, før vi avslultningsvis gjennomgår optimaliseringen av hyperparametere i 4.5.

4.1 Datainndeling og kryssvalidering

Ved modellestimering er det en risiko for overtilpasning. Dette innebærer at modellen tilpasses for godt med treningsdata, og det er dermed fare for at modellen lærer mønster i data som er urelevant. En slik overtilpasset modell vil gi unøyaktige estimater dersom den benyttes på nye data, altså data som ikke er en del av treningssettet (James mfl., 2013, s. 32). For å unngå overtilpasning bruker vi k-fold kryssvalidering (James mfl., 2013, s. 248) og forward validation (Ladyzynski mfl., 2013, s. 447) til å trene modellene.

Ved forward validation deles dataen i trening- og testsett på en slik måte at førstnevnte kun inneholder data som har forekommet før dataen i testsettet. Slik sørger vi for at modellene tilpasses data fra fortiden, for å predikere fremtiden. Videre kan vi enten velge å trene modellene ved å benytte all data som er tilgjengelig eller kun den nyeste dataen. Disse teknikkene kalles henholdsvis Expanding Window og Rolling Window (Brownlee, 2016b). Slik som Paraschiv mfl. (2021), velger vi å benytte Rolling Window da vi ønsker at modellene skal være sammenlignbare for de ulike tidsperiodene. Dette illustreres i appendiks A.2.

I likhet med West (2000) og Akkoç (2012) benytter vi 10-fold kryssvalidering. Vi deler derfor det endelige datasettet i 10 perioder. Hver periode består av 5 måneder, hvor de 4 første utgjør treningssettet og den siste utgjør testsettet. I den første perioden består treningssettet av observasjoner fra 1.november 2019 til 29.februar 2020, mens testsettet består av observasjoner fra mars 2020. For hver periode beveger vi oss én måned frem. I periode 2 består dermed treningssettet av observasjoner fra 1.desember 2020 til 31.mars

2020, mens testsettet inneholder observasjoner fra april 2020. For hver periode trener vi en modell ved bruk av treningssettet og evaluerer modellens prediksjonsevne på treningssettet samt testsettet.

4.2 Estimeringsteknikker

4.2.1 Logistisk regresjon

LR produserer sannsynligheten for at den avhengige variabelen Y tilhører en gitt kategori (James mfl., 2013, s. 132), som i vår studie er mislighold (Y=1) og ikke mislighold (Y=0). Vektoren for de predikerte sannsynlighetene for mislighold $\vec{\hat{y}} = \{\hat{y}_n\}_{n=1,\dots,N} \in [0,1]^N$ er gitt ved (Paraschiv mfl., 2021):

$$\vec{\hat{y}} = \vec{\iota} \oslash \left(\vec{\iota} + \exp\left(-\mathbf{X}\vec{\beta} - \vec{\iota}\beta_0\right)\right) \tag{1}$$

Vi har at $\mathbf{X} = \{x_{(n,i)}\}_{n=1,\dots,N,i=1,\dots,I}$ er en matrise av verdier for forklaringsvariablene i fra datasettet n, $\vec{\beta} = \{\beta_i\}_{i=1,\dots,I}$ er en vektor av koeffisienter, β_0 er konstantleddet, $\vec{\iota}$ er en $N \times 1$ vektor av enere, og $\oslash$ står for Hadamard-divisjon. Vi ser bort fra tidsindeksene for å gjøre tolkningen av likningen enklere (Paraschiv mfl., 2021).

Videre benyttes sannsynlighetsmaksimeringsprinsippet for å estimere koeffisientene til de ukjente forklaringsvariablene (James mfl., 2013, s. 133). Koeffisientene $\vec{\beta}$ og β_0 estimeres ved å minimere den negative av logit-funksjonen $\ell(\vec{\beta}, \beta_0)$ som er gitt ved (Paraschiv mfl., 2021):

$$\ell(\vec{\beta}, \beta_0) = \sum_{n=1}^{N} \left[\vec{y} \odot \left(\mathbf{X}\vec{\beta} + \vec{\iota}\beta_0\right) - \log\left(\vec{\iota} + \exp\left(\mathbf{X}\vec{\beta} + \vec{\iota}\beta_0\right)\right)\right] \tag{2}$$

Hvor $\vec{y} = \{y_n\}_{n=1,\dots,N} \in \{0,1\}^N$ er vektoren med de faktiske klassifikasjonene av mislighold eller ikke-mislighold, og $\odot$ står for Hadamard-produktet. Videre benyttes Wald-statistikk til å bestemme koeffisientenes z-verdier som avgjør om variablene er signifikante eller ikke (Paraschiv mfl., 2021).

Formålet er å velge koeffisienter for de uavhengige variablene slik at den predikerte sannsynligheten for mislighold for hver enkelt observasjon, blir så nær kundens faktiske misligholdsstatus som mulig. Når vi setter de estimerte koeffisientene inn i LR-modellene, skal utfallsverdien dermed gi et tall nær 1 for kunder som har misligholdt og nær 0 for kunder som ikke har misligholdt (James mfl., 2013, s. 133). Dette betyr at vi kan benytte regresjonskoeffisientene fra LR-modellene til direkte å tolke hvor stort bidrag hver enkelt variabel har på prediksjonen.

4.2.2 XGBoost

XGBoost er et skalerbart maskinlæringssystem for tre-boosting, utviklet av T. Chen og Guestrin (2016). Metoden baseres på desisjonstrær og implementerer rammeverket til Gradient Boosting. Teknikken er benyttet av mange dataanalytikere i diverse utfordringer knyttet til maskinlæring og datareduksjon (T. Chen & Guestrin, 2016).

Enkelt forklart innebærer bygging av desisjonstrær to steg. Først deles utfallsrommet, som er et sett av mulige verdier for alle de I uavhengige variablene X_1,..., X_i, i T distinkte og ikke-overlappende regioner, R_1,..., R_t. Inndelingen av regioner skjer basert på regler som angir hvordan inndelingen skal foregå. Reglene kan summeres i en trestruktur, derav navnet desisjonstrær (James mfl., 2013, s. 303). Etter at regionene R_1,..., R_t er laget, kan prediksjoner gjennomføres for observasjonene i testsettet. Dette vil være steg to i prosessen. Steget innebærer at hver observasjon som faller i region R_t, får tildelt den samme predikerte verdien (James mfl., 2013, s. 306). Ved klassifikasjon predikeres hver observasjon i samme region R_t, til den mest forekommende klassen av observasjonene i treningssettet (James mfl., 2013, s. 311). Slike desisjonstrær benyttes som svake modeller i Gradient Boosting (Brownlee, 2016a).

Gradient Boosting ble først introdusert av Friedman (2001), og er en av de kraftigste maskinlæringsteknikkene for å lage prediksjonsmodeller (Nguyen mfl., 2021). Gradient Boosting består av tre elementer: en forhåndsbestemt tapsfunksjon, en svak modell og en additiv modell. Tapsfunksjonen er et mål på hvor gode modellens koeffisienter er tilpasset dataen (Nguyen mfl., 2021). I vårt tilfelle vil tapsfunksjonen være et mål på hvor god modellen er på å klassifisere mislighold. Videre benyttes svake modeller for å gjennomføre

prediksjoner og en additiv modell for å legge til de svake modellene for å minimere tapsfunksjonen (Brownlee, 2016a). Disse svake modellene trenes sekvensielt, hvor hver modell trenes ved bruk av informasjon fra modellen ved forrige iterasjon. For hver iterasjon benyttes de svake modellene til å minimere tapsfunksjonen. Dette innebærer at det inkluderes et feilledd som den svake modellen ikke klarer å forklare. Ved neste iterasjon legges det til en ny modell som reduserer dette feilleddet. Videre blir den nye svake modellen lagt til i tapsfunksjonen for å oppdatere feilleddene. På denne måten fortsetter modellen å forbedre seg gradvis for hver iterasjon. Til slutt summeres de estimerte prediksjonene over alle de trente svake modellene, som gir den endelige modellen (James mfl., 2013, s. 321).

XGBoost er en avansert versjon av Gradient Boosting. En av fordelene med XGBoost er at metoden inkluderer et ledd for regularisering i modellformuleringen, noe som kontrollerer overtilpasning bedre enn hva Gradient Boosting gjør. Dette medfører at modellenes prediksjonsevne blir bedre (T. Chen, 2015). Videre er skalerbarheten til XGBoost en viktig faktor for hvorfor denne metoden er veldig suksessfull. Sammenlignet med andre populære metoder, kjøres XGBoost mer enn ti ganger raskere. Dette skyldes flere viktige systemer og algoritmiske optimaliseringer (T. Chen & Guestrin, 2016). Blant annet benytter XGBoost alle kjernene til datamaskinens prosessor for å trene modellene. Slik bruker metoden mindre tid og dataressurser til å trene modeller og gjennomføre prediksjoner (Jorly, 2021). En klar svakhet ved metoden er derimot vankeligheten med å tolke variablenes bidrag på prediksjonen, da modeller estimert med XGBoost er svarte bokser. Det er derfor nyttig å benytte en metode for analyse av variabelviktighet sammen med XGBoost, for å gjøre resultatet tolkbart (Sagi & Rokach, 2021). For en mer detaljert utledning av XGBoost, se T. Chen og Guestrin (2016).

4.3 Metoder for analyse av variabler

Ved prediksjon av mislighold av kredittkortgjeld er det ikke bare modellenes prediksjonsevne som er av betydning. Det er i tillegg ønskelig med modeller som forklarer resultatet og som dermed er tolkbare (James mfl., 2013, s. 17-20). I mange tilfeller presterer komplekse, gjerne lite tolkbare maskinlæringsmodeller, bedre enn de mer tolkbare og tradisjonelle LR-modellene. Dette har ført til en avveining mellom treffsikkerhet og tolkbarhet. Som

en løsning på dette problemet har det blitt et økt fokus på metoder for variabelviktighet, som indikerer hvilke variabler som er av betydning og ikke. Slike metoder kan benyttes for bedre å forstå og tolke prediksjonene fra avanserte maskinlæringsmodeller (Aas mfl., 2021).

4.3.1 LASSO

LASSO ble gjort populær som metode av Tibshirani (1996). Teknikken innlemmer variabelseleksjon som en del av modelltreningen (Paraschiv mfl., 2021). LASSO trener en modell som inkluderer alle uavhengige variabler, men med et straffeledd proposjonalt med koeffisientverdiene som gjør at enkelte koeffisienter krympes til null. På denne måten eliminerer teknikken de forklaringsvariablene som ikke har noen betydning (Tibshirani, 1996). Vi benytter derfor LASSO for å ekskludere irrelevante forklaringsvariabler før vi estimerer de endelige LR-modellene.

Ved bruk av LASSO, estimeres koeffisientene β og β_0 ved å minimere følgende i treningssettet (Paraschiv mfl., 2021):

$$-\ell(\vec{\beta}, \beta_0) + \lambda||\vec{\beta}||_1 \tag{3}$$

$\ell(\vec{\beta}, \beta_0)$ er den samme logit-funksjonen som er gitt i likning 2. Videre er λ en positiv hyperparameter og $||\vec{\beta}||_1$ er l_1-regulariseringen av $\vec{\beta}$, tilsvarende $\sum_{i=1}^{I} \beta_i$. Det er hyperparameteren λ som avgjør hvor mye det straffer seg å inkludere flere uavhengige variabler i modellen (Paraschiv mfl., 2021). Dersom λ er tilstrekkelig stor, vil straffeleddet bli så dominerende at alle de estimerte koeffisientene settes til 0. Valget av λ-verdien vil dermed være av avgjørende betydning for resultatene fra LASSO. For å få best mulige estimater er det dermed vanlig å benytte kryssvalidering for å finne den optimale λ-verdien (Tibshirani, 1996). Vi vil forklare hvordan dette gjøres i vår studie i delkapittel 4.5.

4.3.2 SHAP

SHAP ble utledet av Lundberg og Lee (2017). Metoden benyttes for å forklare hvordan en svart boks har kommet frem til individuelle prediksjoner, ved å beregne hver enkelt variabel sitt bidrag på prediksjonen (Molnar, 2022, kap. 9.6.1). SHAP er dermed ikke en

metode for variabelseleksjon, men en metode som kan benyttes for å analysere variabelviktighet. En klar fordel med SHAP er at metoden kan benyttes til å tolke både lokal og global variabelviktighet (Aas mfl., 2021). Videre utfører SHAP raske beregninger, spesielt for trebaserte modeller (Molnar, 2022, kap. 9.6.10). Vi velger dermed å benytte SHAP for å tolke XGBoost-modellene, da XGBoost fungerer som en svart boks.

SHAP bygger på spillteori og såkalte Shapley-verdier, introdusert av Shapley (1953). Metoden ble opprinnelig brukt for å tildele utbytte til hver enkelt spiller, basert på deres bidrag til det totale utbyttet (Molnar, 2022, kap. 9.6.1). Shapley-verdiene bygger på en antakelse om at spillerne samarbeider og danner koalisjoner (Aas mfl., 2021). En spiller j kan dermed forvente å motta et utbytte gitt ved:

$$\phi_j(v) = \phi_j = \sum_{S \subseteq M \setminus \{j\}} \frac{|S|!(M - |S| - 1)!}{M!}(v(S \cup \{j\}) - v(S)), \quad j = 1, ..., M \tag{4}$$

Dette tilsvarer det vektede gjennomsnittet over alle grupper S av spillere, hvor spiller j ikke er inkludert (Aas mfl., 2021). Shapley-verdier vil dermed være det gjennomsnittlige marginale bidraget for en bestemt variabel, på tvers av alle mulige koalisjoner. For den gitte forklaringsvariabelen som ønskes undersøkt gjennomføres det en prediksjon, først med og deretter uten denne forklaringsvariabelen. Dette gjøres for alle mulige koalisjoner, for å finne det gjennomsnittlige marginale bidraget (Molnar, 2022, kap. 9.5.1).

Logikken fra spillteorien og Shapley-verdier legger grunnlaget for maskinlæringsmetoden SHAP (Lundberg & Lee, 2017). Arbeidsoppgaven maskinlæringsalgoritmen skal løse, anses her som spillet. Videre defineres den enkelte prediksjonen som utbytte, mens forklaringsvariablene utgjør spillerne (Aas mfl., 2021). SHAP kan formuleres ved følgende formel (Lundberg & Lee, 2017):

$$g\left(z'\right) = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j \tag{5}$$

Hvor g er forklaringsmodellen, definert som enhver tolkbar tilnærming til den originale og gjerne komplekse modellen. Videre er $z' \in \{0,1\}^M$ koalisjonsvektoren, M er antall forenklede forklaringsvariabler og $\phi_i \in R$ er Shapley-verdiene (Lundberg & Lee, 2017).

For å vurdere variablenes globale variabelviktighet benytter vi såkalte SHAP-plott for hver periode, vedlagt i appendiks A.6. Y-aksen viser de 40 viktigste variablene i den gitte perioden, sortert etter variabelviktighet. Ved siden av variabelnavnene angis variablenes SHAP-verdi, som tilsvarer det gjennomsnittlige bidraget en gitt variabel har på prediksjonen av mislighold av kredittkortgjeld. En høy verdi indikerer at en variabel bidrar mye i prediksjonen av mislighold, mens en lav verdi indikerer at variabelen bidrar lite i prediksjonen. Videre representerer hver prikk i plottet, én observasjon fra datasettet. Fargene i plottet angir observasjonenes opprinnelige verdi for den enkelte variabelen, mens x-aksen viser observasjonenes SHAP-verdier (Lundberg, 2018). Basert på dette, kan vi tolke om variabelen bidrar mest i retning misligholdt eller ikke misligholdt.

Videre vurderes variablenes lokale variabelviktighet av plottene vedlagt i delkapittel 5.3. Plottene angir SHAP-verdier for de 10 forklaringsvariablene som har størst bidrag på prediksjonen av mislighold for en gitte observasjon. Plottet gir dermed en forklaring på hvorfor en observasjon er predikert til misligholdt eller ikke misligholdt.

4.4 Evalueringsmål

4.4.1 AUC

For å evaluere prediksjonsevnen til de estimerte modellene, benytter vi Area Under the Receiver Operating Characteristic Curve (AUC) både på trenings- og testsettet. I tillegg benytter vi AUC som evalueringsmål i optimaliseringen av hyperparametere. AUC er et mye brukt evalueringsmål (Kvamme mfl., 2018). Spesielt har AUC vist seg å være et nyttig evalueringsmål når det er stor ubalanse mellom klassene i datasettet, slik det gjerne er ved prediksjon av mislighold av kredittkortgjeld. AUC er i slike tilfeller å foretrekke fremfor enklere evalueringsmål som treffsikkerhet, da AUC ikke vil gi skjeve estimater (Brzezinski & Stefanowski, 2017). AUC representerer arealet under grafen når ekte positivrate plottes mot falsk positivrate på tvers av alle observasjoner, summert over alle mulige terskelverdier (James mfl., 2013, s. 147). AUC $\in [0.5, 1]$, hvor høyere verdi indikerer at modellen har bedre prediksjonsevne (Paraschiv mfl., 2021). Basert på Hosmer mfl. (2013) anser vi AUC $\in [0.7, 0.8)$ som en akseptabel modell, AUC $\in [0.8, 0.9)$ som utmerket og AUC $\geq$ 0.9 som fremragende.

4.4.2 Brier Score

I tillegg til AUC velger vi å benytte Brier (1950) Score for å evaluere de endelige prediksjonsmodellene, både på trening- og testdata. Vi velger å benytte to evalueringsmål av ulik type, da ulike typer evalueringsmål kan gi ulik fremstilling av modellers prediksjonsevne (Lessmann mfl., 2015). Brier Score er et mål på den gjennomsnittlige kvadrerte avstanden mellom den predikerte sannsynligheten knyttet til mulige utfall for et sett med forklaringsvariabler og det faktiske utfallet (Paraschiv mfl., 2021). Brier Score er dermed et evalueringsmål på modellers nøyaktighet (Lessmann mfl., 2015) og ønskes dermed så lav som mulig (Brier, 1950). Brier Score er ifølge Lessmann mfl. (2015) sin litteraturgjennomgang mindre brukt innen kredittvurdering, enn AUC. Evalueringsmålet er likevel blitt benyttet i blant annet Kvamme mfl. (2018) sin studie om kredittvurdering, samt Paraschiv mfl. (2021) sin studie om konkursprediksjon.

4.5 Optimalisering av hyperparametere

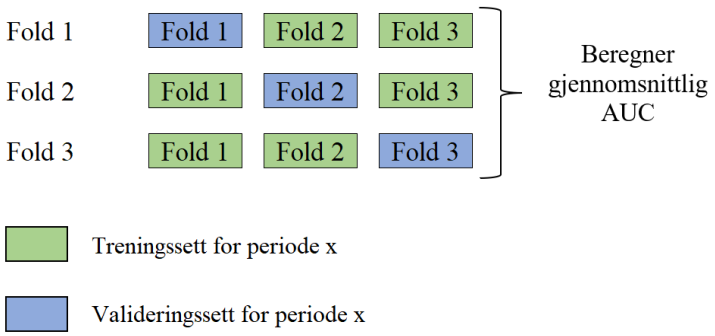
Før vi kan trene modellene, må vi optimalisere modellenes hyperparametere. Dette gjøres på treningsdata da beregning av prediksjonsevne på samme data som modellene trenes på, vil gi et unøyaktig mål på hvor bra modellene faktisk er. Optimalisering innebærer å finne den kombinasjonen av verdier for hyperparametere som gir den modellen med best prediksjonsevne på testsettet. Hyperparameterne må optimaliseres for hver enkelt modell, da de kontrollerer kompleksiteten ved modellene. Dette betyr at hyperparameterne for den beste modellen i et datasett, ikke vil gjelde i et annet datasett (Koehrsen, 2018). I dette kapittelet beskrives hvordan vi optimaliserer hyperparametere i LASSO og XGBoost.

4.5.1 Optimalisering i LASSO

Ved gjennomførelse av LASSO optimaliserer vi λ, hvor vi benytter gridsøk som metode. Dette innebærer at det først blir laget et grid (rutenett) med ulike verdier for λ. Lengden på gridet er satt til 100 verdier, hvor den laveste verdien er definert som 10^{-2} og høyeste verdi er 10^{2}. Videre benyttes k-fold kryssvalidering, med $k = 3$, til å vurdere de ulike λ-verdiene i gridet. Én fold benyttes som valideringssett, mens de 2 resterende benyttes for å trene modellen med valgt λ-verdi (James mfl., 2013, s. 181). Den trente model-

len blir evaluert på valideringssettet ved bruk av AUC. Dette gjøres for hver fold, hvor ulike valideringssett benyttes hver gang. For hver λ-verdi i gridet, beregnes modellenes gjennomsnittlige AUC for alle de tre valideringssettene. Prosessen er illustrert i figur 2.

Figur 2: Kryssvalidering for å finne optimal λ (scikit-learn, 2017)



Ettersom straffeleddet i LASSO øker med økende λ, vil antallet variabler i modellene reduseres ved en høyere λ-verdi. Vi velger derfor å sammenligne tre ulike verdier for λ med tilhørende AUC-verdi, for å avgjøre hvilken λ-verdi vi videre benytter i estimeringen av LR-modellene. Vi benytter AUC-plott, vedlagt i appendiks A.3, for å vurdere verdiene. Plottene viser AUC mot $\log(\lambda)$, gitt ved kryssvalideringen for de ulike periodene. Streken lengst til venstre i plottene representerer den λ-verdien som gir høyest gjennomsnittlig AUC-verdi på valideringssettet. Videre angir den midterste streken den største verdien av λ hvor den gjennomsnittlige AUC-verdien er innenfor 1 standardfeil fra AUC-verdien til den førstnevnte λ-verdien (S.-H. Lee, 2021). λ-verdien som representeres av streken til høyre beregnes likt som sistnevnte, men med 1,5 standardfeil. AUC-plottene gir i tillegg en indikasjon på hvor mange forklaringsvariabler som blir valgt av LASSO for de ulike λ-verdiene. Videre lager vi LASSO-stiplott, vedlagt i appendiks A.3, basert på den valgte λ-verdien. Disse illustrerer hvilke forklaringsvariabler som ikke blir satt til 0 av LASSO i de ulike periodene.

4.5.2 Optimalisering i XGBoost

For XGBoost-modeller anbefales det, ifølge Saraswat (2016), å optimalisere hyperparameterne i tabell 1. Det anbefales å optimalisere alpha dersom det er behov for at algoritmen skal kjøre raskere på grunn av høydimensjonalitetsdata (Saraswat, 2016). Dette er derimot

ikke tilfellet i vår data. Videre velger vi ikke å optimalisere lambda da dette som oftest ikke gjennomføres i XGBoost-modeller (Jain, 2016). Vi velger derfor å optimalisere alle hyperparameterne oppgitt i tabellen, utenom alpha og lambda.

Tabell 1: Hyperparametere i XGBoost

Hyperparameter	**Beskrivelse**
Alpha	Kontrollerer L1-regularisering. Benyttes til variabelseleksjon.
Eta	Kontrollerer læringshastigheten, dvs. hvor raskt modellen lærer seg mønster i data.
Gamma	Kontrollerer regularisering. Benyttes til å unngå overtilpasning.
Lambda	Kontrollerer L2-regularisering. Benyttes til å unngå overtilpasning.
Max depth	Kontrollerer dybden av modellen. Større dybde gir mer komplekse modeller og høyere sannsynlighet for overtilpasning.
Min child weight	Kontrollerer overtilpasning ved å blokkere for potensielle interaksjoner mellom forklaringsvariabler.
Nrounds	Maksimalt antall iterasjoner, dvs. maks antall desisjonstrær som skal lages.
Subsample	Kontrollerer antall observasjoner som knyttes til modellen.

For å optimalisere de resterende hyperparameterne, velger vi å benytte Bayesiansk Optimalisering (BO). Flere studier har vist at dette er en bedre metode for optimalisering av hyperparametere i XGBoost-modeller sammenlignet med teknikker som gridsøk, tilfeldig og manuell søk (Bergstra mfl., 2015; Xia mfl., 2017; Mushava & Murray, 2022). Dette innebærer at vi trener flere modeller, før den beste kombinasjonen av hyperparametere velges. Den beste kombinasjonen tilsvarer antall iterasjoner der det ikke er noen signifikant økning i AUC (Mushava & Murray, 2022). Vi velger å sette maksimalt antall iterasjoner lik 10 000, i tillegg til et stoppkriterie lik 50. Metoden vil dermed trene opp til 10 000 modeller, gitt en forbedring i AUC. Dersom AUC ikke forbedres på 50 modeller, vil treningen av modellene stoppes (Nugent, 2018). En oversikt over de optimaliserte hyperparameterne finnes i appendiks A.5.

5 Resultater

5.1 Variabelviktighet

5.1.1 LASSO

Den relative betydningen av forklaringsvariablene valgt av LASSO i de 10 periodene, kan analyseres ved LASSO-stiplottene og de tilhørende AUC-plottene vedlagt i appendiks A.3. Samtlige AUC-plott indikerer at LASSO velger flest forklaringsvariabler for λ-verdien som maksimerer AUC, gitt ved de stiplede linjene lengst til venstre i plottene. Da vi ønsker å analysere variabelviktigheten, er det fordelaktig å redusere antall forklaringsvariabler uten å svekke modellenes prediksjonsevne i vesentlig grad. Av plottene ser vi at antall forklaringsvariabler valgt av LASSO reduseres dersom λ-verdiene gitt ved streken i midten og til høyre heller benyttes, uten at AUC svekkes i noen betydelig grad. Ettersom vi ønsker færrest mulig forklaringsvariabler, velger vi å benytte λ-verdien gitt ved streken til høyre.

LASSO-stiplottene lages basert på den valgte λ-verdien i de enkelte periodene. For å gjøre plottene tolkbare velger vi kun å inkludere forklaringsvariablene som *ikke* blir satt til 0 av LASSO for den valgte λ-verdien i de ulike periodene. Variabler som inkluderes i LASSO-stiplottene ved høyere verdier av λ har sterkere forklaringskraft. Variabler som inkluderes lengre mot høyre i stiplottene indikerer dermed en høyere relativ variabelviktighet. Figurene viser at *«RegisteredDebtByIncome»* er den klart viktigste variabelen i samtlige perioder.

Tabell 4 i appendiks A.4 viser en oversikt over forklaringsvariablene som blir valgt av LASSO i hver av de 10 periodene. Tabellen er sortert slik at variablene som blir valgt i alle de 10 periodene kommer først. Deretter følger resterende forklaringsvariabler i synkende rekkefølge basert på antall perioder de blir valgt av LASSO. Variablene LASSO velger, benyttes videre til å trene en LR-modell for hver periode. Overordnet velger LASSO totalt ut 55 variabler, på tvers av de 10 periodene. 14 av disse forklaringsvariablene blir kun valgt i én periode, mens 13 blir valgt i kun 2 perioder. Det er likevel også 7 variabler som blir valgt i alle de 10 periodene, samt 2 variabler som blir valgt i alle perioder utenom én. I den ene perioden hvor *«Homeowner_NoHouseOrCoop»* ikke blir valgt av LASSO,

blir *«Homeowner_HouseOrCoop»* valgt. Begge disse variablene representerer hvorvidt en lånsøker eier hus eller eiendom i borettslag med boliglån hos dataleverandør. Dette indikerer at LASSO velger variabler som måler det samme aspektet ved lånsøkeren.

I tabell 4 er de estimerte koeffisientverdiene fra LR-modellene oppgitt, samt de tilhørende z-verdiene. Vi definerer alle variabler med signifkansnivå på 5% eller lavere, som signfikante variabler. Av de totalt 7 forklaringsvariablene som blir valgt i alle perioder, er det *«Student_Loan_AMT»*, *«RegisteredDebtByIncome»* og *«Habitation_Renter»* som er signifkante i samtlige. Videre er *«AppliedByGranted»* og *«Age_log»* signifkante i totalt 9 av 10 perioder, mens *«Net_Income_AMT»* og *«App_Night»* er signifikante i 8 perioder. Det er ikke et entydig mønster når det gjelder hvilke perioder forklaringsvariablene er signifkante og ikke. Videre er det verdt å merke seg at *«Homeowner_NoHouseOrCoop»*, som blir valgt ut i 9 perioder, kun er signifikant i én av disse periodene. *«Habitation_Homeowner»*, som blir valgt av LASSO i 6 av 10 perioder, er ikke signifikant i noen av de utvalgte periodene. *«Employment_SocialSecurity»*, som blir valgt i 3 perioder og som også har høye standardiserte koeffisientverdier, er heller ikke signifkant i noen av de utvalgte periodene. Det er i tillegg enkelte andre variabler som ikke er signifkante i noen av de utvalgte periodene. Flesteparten av disse ikke-signifkante variablene er variabler som kun blir valgt av LASSO i én periode. Resterende forklaringsvariabler er signifkante i alle utvalgte perioder eller, gjeldende for de fleste, signifkante i utvalgte perioder med unntak av 1 til 2 perioder.

5.1.2 SHAP

I appendiks A.6 har vi vedlagt SHAP-plott for hver periode. Disse gir informasjon om variablenes gjennomsnittlige SHAP-verdier i hver periode, og hvorvidt de bidrar til prediksjon av mislighold eller ikke mislighold. Videre viser figur 3 alle variabler med tilhørende standardiserte SHAP-verdier for de ulike periodene. Verdiene er standardisert mellom 0 og 100. Verdien 100 er tildelt de variablene som ifølge SHAP har høyest variabelviktighet i de enkelte periodene. De resterende variablene er satt til en verdi mellom 0 og 100, tilsvarende den prosentvise viktigheten i forhold til den viktigste variabelen i den gitte perioden. Dette illustreres også ved bruk av farger, der høyere verdi gir mørkere blåfarge. Variabler med verdi lik 0 i alle perioder er ekskludert fra tabellen.

	Periode_1	Periode_2	Periode_3	Periode_4	Periode_5	Periode_6	Periode_7	Periode_8	Periode_9	Periode_10
Age_log	100,000	82,177	61,456	37,000	61,727	48,550	74,848	65,767	62,674	100,000
App_Night	0,000	0,000	0,000	0,000	4,274	1,415	0,000	0,000	0,000	0,000
App_Weekend	9,811	12,046	0,000	7,169	13,064	14,781	18,134	12,113	5,720	5,561
AppliedByGranted	58,756	62,245	45,277	15,467	23,832	20,141	30,885	15,619	22,235	35,942
Channel_OnlineBank	12,473	0,000	0,000	6,826	12,064	18,222	33,273	20,694	14,371	21,051
Channel_Operator	14,696	41,944	15,765	19,812	33,321	21,420	38,884	31,602	38,315	58,111
Channel_PhoneBank	13,645	16,337	7,145	7,254	10,600	22,531	32,304	17,991	30,618	28,564
Channel_Web	0,000	0,000	0,000	0,000	3,497	5,160	0,000	0,000	0,000	0,000
Debit_Card_IND	12,658	14,986	15,835	10,972	3,617	3,734	9,324	2,224	8,214	19,611
Debt_Ratio_AMT	57,334	47,246	35,681	45,432	100,000	83,905	56,259	24,429	26,363	54,869
DebtRegisterCreditFacilityNum	12,710	37,802	14,385	29,352	47,255	32,013	38,052	21,917	9,070	34,123
DebtRegisterCreditLimit	24,671	53,849	22,342	13,267	27,281	30,271	24,392	23,446	42,347	80,658
DebtRegisterIELA	40,765	76,060	36,541	65,004	47,804	78,435	43,530	100,000	99,309	88,053
DebtRegisterNonIELA	22,634	15,567	19,972	27,045	26,482	14,948	6,148	11,991	16,560	28,475
DebtRegisterNum	49,913	26,780	43,203	39,539	30,325	50,111	54,034	56,841	26,298	31,216
DebtRegisterRepaymentLoanBalance	10,817	27,541	7,055	9,524	18,039	25,010	14,366	17,782	5,723	10,433
DebtRegisterRepaymentLoanNum	0,000	10,777	0,000	8,791	0,000	0,000	0,000	0,000	0,000	0,000
Diff_Hour	74,393	62,655	47,192	43,364	57,056	65,610	52,820	36,751	31,420	44,757
Distributor_4G	0,000	10,347	0,000	0,000	8,881	13,883	17,031	8,133	12,206	0,000
Distributor_4J	0,000	15,467	3,199	9,622	8,137	6,216	18,512	14,431	17,210	29,776
Distributor_4K	0,000	16,932	0,000	0,000	19,636	5,107	8,074	6,336	0,000	5,878
Distributor_4O	14,806	0,000	0,000	3,810	10,557	4,103	9,979	5,460	3,156	10,895
Employment_Duration_1To3	11,662	0,000	9,271	15,600	5,362	11,723	10,323	0,000	4,654	4,478
Employment_Duration_Less1	0,000	14,555	0,000	0,000	2,030	2,302	6,508	0,000	0,000	6,896
Employment_Duration_More3	0,000	19,658	0,000	0,000	2,994	6,122	0,000	14,061	4,503	10,190
Employment_Duration_NoJoB	11,585	0,000	0,000	9,145	26,354	3,444	19,077	5,125	0,000	0,000
Employment_Employee	9,374	15,781	6,812	8,542	8,373	17,468	8,458	7,127	4,504	0,000
Employment_Student	7,106	24,208	10,971	5,762	4,508	5,204	0,000	0,000	0,000	0,000
FLI_AMT	35,150	50,115	38,571	27,484	55,590	75,262	50,716	28,671	28,287	71,497
Habitation_Homeowner	25,869	37,274	15,717	9,054	33,681	20,669	28,949	16,368	6,710	10,445
Habitation_Parents	0,000	0,000	0,000	0,000	6,499	10,589	0,000	6,593	3,927	0,000
Habitation_Renter	31,128	35,839	35,467	16,643	29,405	25,286	32,701	40,571	16,265	43,805
Homeowner_HouseOrCoop	6,159	24,296	16,461	10,509	6,464	5,307	14,579	7,618	0,000	0,000
Homeowner_NoHouseOrCoop	0,000	29,242	0,000	13,093	14,776	12,883	0,000	0,000	14,116	22,594
HourCompleted	38,007	41,806	23,332	27,878	32,594	31,240	24,166	27,458	15,786	24,815
Logins_Num	57,789	55,850	37,975	30,537	41,381	42,251	28,910	27,707	21,197	45,120
Manual_Handling_IND	11,214	40,168	12,629	20,566	11,449	7,629	0,000	5,414	2,202	6,176
MC_2	41,961	31,420	34,882	33,318	26,875	32,334	33,594	26,823	13,734	9,790
MC_4	0,000	41,630	12,386	18,961	25,650	14,050	15,447	9,677	14,667	36,773
MC_4_Young	0,000	0,000	0,000	7,356	15,509	28,423	27,073	13,824	21,599	18,445
Method_Ordinary	38,263	16,197	30,827	32,792	27,359	45,179	45,967	32,170	14,689	24,918
Method_Predefined	0,000	34,483	12,663	0,000	7,328	0,000	0,000	0,000	0,000	0,000
Method_Young	0,000	11,219	0,000	0,000	7,326	0,000	13,928	4,802	16,216	13,858
Mortgage_IND	20,181	36,972	13,812	11,655	6,869	9,458	11,870	5,156	19,679	22,361
Mortgages_AMT	32,731	28,241	8,612	20,521	39,938	42,852	22,672	11,841	15,494	30,784
Net_Income_AMT	50,402	82,534	70,418	100,000	77,865	69,725	57,035	45,007	100,000	49,461
Net_Monthly_Income_AMT	33,372	42,166	40,276	30,442	36,600	34,448	36,308	22,428	35,266	54,400
NoOfAdressChanges_CNT	17,951	20,138	22,629	10,801	24,341	32,431	22,144	12,712	18,733	21,287
NoOfChildren	37,001	8,914	11,463	0,000	14,332	10,959	24,440	13,576	18,990	26,197
RegisteredDebtByIncome	89,780	100,000	88,956	55,802	33,992	48,281	57,106	21,705	26,594	28,317
SFLI_AMT	76,307	35,168	36,309	35,049	40,619	39,445	38,010	29,247	34,311	55,228
Status_Cohabiting	24,489	0,000	0,000	9,105	10,058	12,380	11,700	3,310	11,259	11,315
Status_Married	48,193	19,542	0,000	10,431	19,928	18,399	48,831	21,543	40,779	64,505
Status_Single	9,143	12,529	34,825	8,707	17,997	45,485	27,048	19,871	14,083	18,824
Student_Loan_AMT	59,176	92,425	55,355	30,861	49,655	60,407	62,985	41,033	37,472	53,831
SumAvailable	51,906	51,876	30,592	33,336	45,476	83,054	54,380	39,864	29,564	42,488
Total_Incoming_AMT	89,255	96,727	94,160	54,419	86,664	100,000	100,000	50,847	61,463	74,876
Total_Tax_AMT	21,793	44,943	100,000	95,881	85,023	73,586	40,117	52,717	76,831	62,610

Figur 3: Standardiserte SHAP-verdier

Vi ser at *«Total_Incoming_AMT»* anses som en viktig variabel da den stort sett har høye standardiserte SHAP-verdier i alle periodene, og hvor den laveste verdien er på 50,847. Videre ser vi også at *« Age_log»*, *«Net_Income_AMT»*, *«Total_Tax_AMT»*, *«Transactions_Num»* og *«Wealth_AMT»* anses som stabilt viktige variabler da de har

høye standardiserte SHAP-verdier i de fleste periodene. *«DebtRegisterIELA»*, *«RegisteredDebtByIncome»* og *«Debt_Ratio_AMT»* er derimot kun blant de viktigste variablene i enkelte perioder, da de har SHAP-verdier mellom 70 og 100 for henholdsvis 5, 3 og 2 perioder. Videre ser vi at *«Student_Loan_AMT»* og *«Diff_Hour»* er relativt stabile, da de har middels høye SHAP-verdier for alle perioder. Enkelte variabler har SHAP-verdier over 0 i kun noen få perioder. Dette gjelder blant annet *«App_Night»*, *«Channel_Web»*, *«DebtRegisterRepaymentLoanNum»* og *«Method_Predefined»*. Disse har kun verdier over 0 for 2-3 perioder, hvor verdiene er lave.

5.2 Modellprediksjon

Tabell 2 viser modellenes prediksjonsevne på trenings- og testsettet målt med AUC og Brier Score, for hver enkelt periode. I tillegg er den prosentvise forskjellen i prediksjonsevne mellom LR- og XGBoost-modellene for hver periode[1] inkludert. Forskjellene er større på treningssettene, sammenlignet med testsettene. Videre indikerer evalueringsmålene at LR-modellene stort sett presterer bedre enn XGBoost-modellene.

Tabell 2: Evaluering av modellenes prediksjonsevne på trenings- og testdata

	AUC - Trening			AUC - Test			Brier - Trening			Brier - Test		
	Logistisk	**XGboost**	**Endring i %**	**Logistisk**	**XGboost**	**Endring i %**	**Logistisk**	**XGboost**	**Endring i %**	**Logistisk**	**XGboost**	**Endring i %**
Periode 1	0,8030	0,9580	19,30 %	0,7987	0,7914	-0,91 %	0,0195	0,0144	-26,15 %	0,0264	0,0262	-0,76 %
Periode 2	0,8128	0,9513	17,04 %	0,8521	0,8513	-0,09 %	0,0212	0,0155	-26,89 %	0,0253	0,0253	0,00 %
Periode 3	0,8181	0,9566	16,93 %	0,8253	0,8086	-2,02 %	0,0219	0,0156	-28,77 %	0,0197	0,0206	4,57 %
Periode 4	0,8222	0,9571	16,41 %	0,753	0,8235	9,36 %	0,0225	0,0161	-28,44 %	0,0211	0,0218	3,32 %
Periode 5	0,8170	1,0000	22,40 %	0,7946	0,7634	-3,93 %	0,0226	0,0031	-86,28 %	0,0317	0,0336	5,99 %
Periode 6	0,8360	1,0000	19,62 %	0,8014	0,7485	-6,60 %	0,0235	0,002	-91,49 %	0,0281	0,0321	14,23 %
Periode 7	0,8329	0,9580	15,02 %	0,7284	0,7604	4,39 %	0,0241	0,0173	-28,22 %	0,0219	0,0228	4,11 %
Periode 8	0,8197	0,9659	17,84 %	0,8174	0,8156	-0,22 %	0,0242	0,0164	-32,23 %	0,0221	0,0239	8,14 %
Periode 9	0,8357	0,9683	15,87 %	0,8190	0,8145	-0,55 %	0,0245	0,0163	-33,47 %	0,0225	0,0237	5,33 %
Periode 10	0,8144	0,9663	18,65 %	0,7850	0,7892	0,54 %	0,0228	0,0159	-30,26 %	0,0308	0,0331	7,47 %

[1]De prosentvise forskjellene viser LR-modellenes verdier i forhold til XGBoost-modellenes verdier

5.2.1 Logistisk regresjon

For LR-modellene varierer AUC fra 0,803 til 0,836 på treningssettet, og fra 0,7284 til 0,8521 på testsettet. Sammenlignet med treningssettet, reduseres AUC på testsettet for enkelte perioder, mens den øker for andre. Modellenes prediksjonsevne er akseptable til utmerket da AUC-verdiene på testsettet ligger innenfor [0.7, 0.9).

Modellenes prediksjonsevne målt med Brier Score varierer fra 0,0195 til 0,0245 på treningssettet, og fra 0,0197 til 0,0317 på testsettet. I likhet med AUC, varierer det om verdiene for Brier Score på testsettet er større eller mindre enn verdiene på treningssettet. Resultatene gitt ved Brier Score samsvarer med resultatene gitt ved AUC.

5.2.2 XGBoost

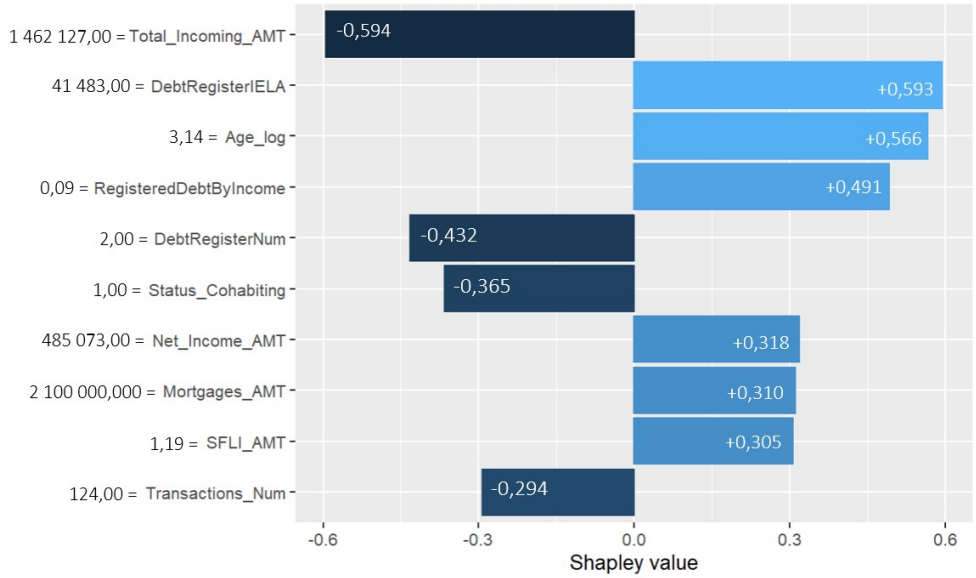
XGBoost-modellenes prediksjonsevne målt med AUC varierer fra 0,9513 til 1,000 på treningssettet, og fra 0,7485 til 0,8513 på testsettet. AUC-verdiene på treningssettet er spesielt høye. Resultatene er å forvente da XGBoost er en svært fleksibel estimeringsteknikk. AUC-verdiene på testsettet viser at XGBoost-modellenes prediksjonsevne er akseptable til utmerket.

Verdiene for Brier Score varierer fra 0,002 til 0,0173 på treningssettet, og fra 0,0206 til 0,0336 på testsettet. Dette stemmer overens med resultatene gitt med AUC.

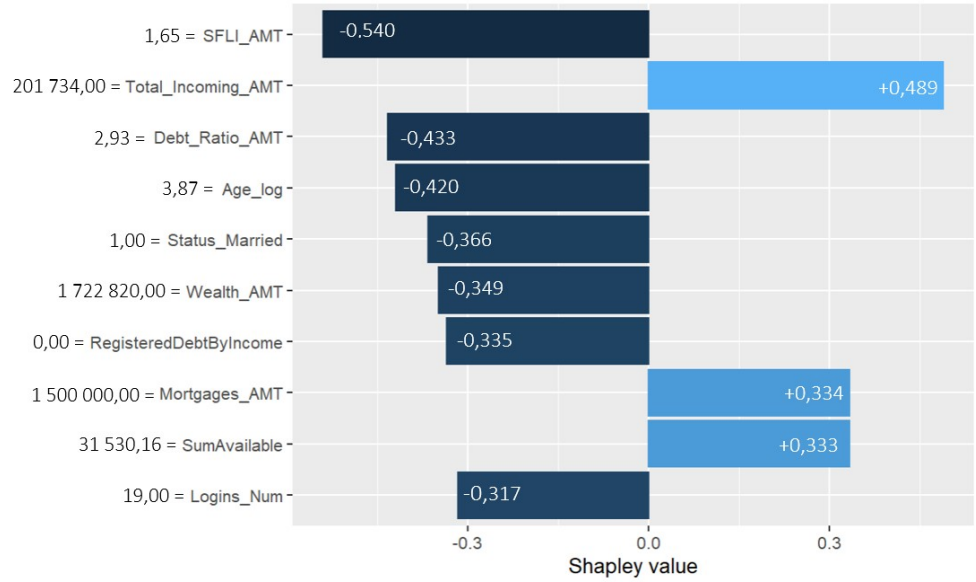
5.3 Tolkning av svarte bokser

For at prediksjonsmodeller estimert ved bruk av svarte bokser skal kunne benyttes til kredittvurdering, må de møte kravene som stilles i personopplysningsloven. Vi benytter SHAP for lokal tolkning av prediksjonene gjort med XGBoost-modellen i periode 1, for to observasjoner. Dette illustreres i figur 4 og 5. Vi velger å inkludere de 10 variablene med høyest variabelviktighet for hver av de individuelle prediksjonene. Figurene viser i tillegg de sanne variabelverdiene.

Figur 4: Låntaker A (Y=1)



Figur 5: Låntaker B (Y=0)



Den øverste figuren illustrerer variablene som har størst bidrag på det predikerte utfallet for låntaker A, altså mislighold. De sanne variabelverdiene til de 3 viktigste variablene tilsier at dette er en 23 år gammel låntaker med et rentebærende beløp på kr 41 483 registrert i gjeldsregisteret og med totalt kr 1 462 127 som innkommende beløp på konto de siste 6 månedene. Av de 10 viktigste variablene, er det totalt 6 variabler som

bidrar i den predikerte retningen. «*DebtRegisterIELA*» er variabelen som bidrar mest i retning mislighold, etterfulgt av «*Age_log*» og «*RegisterDebtByIncome*». I tillegg bidrar også «*Net_Income_AMT*», «*Mortgages_AMT*» og «*SFLI_AMT*» i retning mislighold, alle med SHAP-verdier rundt +0,3. Variabelen med høyest variabelviktighet, «*Total_Incoming_AMT*», bidrar derimot i retning ikke mislighold. «*Transactions_Num*» bidrar minst på det predikerte utfallet av de topp 10 viktigste variablene.

Figur 5 viser hvilke variabler som bidrar mest i prediksjonen for en gitt låntaker som ikke har misligholdt. Dette er en 48 år gammel låntaker som blant annet har SFLI lik 1,65, totalt kr 201 734 som innkommende beløp på konto de siste 6 månedene, samt en gjeldsgrad på 2,93. Vi ser at 7 variabler bidrar i retning ikke mislighold, mens kun 3 variabler bidrar i retning mislighold. For denne låntakeren er det *«SFLI_AMT»* som har størst bidrag på prediksjonen, i retning ikke mislighold. Den nest viktigste variabelen, *«Total_Incoming_AMT»*, bidrar derimot i motsatt retning. Videre bidrar *«Debt_Ratio_AMT»*, *«Age_log»*, *«Status_Married»*, *«Wealth»* og *«RegisteredDebtByIncome»* i retning ikke mislighold. Til slutt bidrar *«Mortgages_AMT»* og *«SumAvailable»* i retning mislighold, mens *«Logins_Num»* bidrar i retning ikke mislighold.

6 Diskusjon

6.1 Variabelviktighet

Resultatene indikerer flere klare likheter når det gjelder variabelviktighet basert på LASSO og SHAP. For det første anses *«Age_log»* som en sentral variabel av begge metodene. Dette samsvarer med både Steenackers og Goovaerts (1989), Somol mfl. (2005) og Bellotti og Crook (2009) som alle fant at alder er av signifikant betydning i modeller for kredittvurdering. Regresjonskoeffisientene i LR-modellene er negative for samtlige perioder og indikerer dermed at sannsynligheten for mislighold reduseres med økende alder. SHAP-plottene indikerer tilsvarende sammenheng, da høyere alder er assosisert med høye og negative bidrag på hvorvidt en lånsøker vil misligholde. Resultatene virker rimelige, da en persons økonomiske situasjon typisk stabiliseres ved økende alder. Det er gjerne i en etableringsfase som ung, hvor inntektene nødvendigvis ikke er store, at behovet for usikret gjeld er størst. Samtidig er det ofte i en slik livssituasjon at en person høyst sannsynlig har større problemer med å tilbakebetale egen gjeld.

Vi inkluderer både månedlig og årlig nettoinntekt i vår data da *«Net_Income_AMT»* er hentet fra skattedata, mens kunden selv har fylt inn *«Net_Monthly_Income_AMT»*. Selv om de to variablene i utgangspunktet kan anses å være perfekt korrelert, vil det i følge en representant fra dataleverandør ofte være et avvik mellom de to variablene i vår data. Våre resultater viser at *«Net_Income_AMT»* anses som en viktig variabel av både LASSO og SHAP. Regresjonskoeffisientene i LR-modellene indikerer mindre sannsynlighet for mislighold, ved større verdier for *«Net_Income_AMT»*. Dette ser vi også av SHAP-plottene, med unntak av periode 1 hvor det indikeres at en høyere årlig nettoinntekt er assosiert med høyere sannsynlighet for mislighold. Med unntak av denne ene perioden, kan resultatene fra både LASSO og SHAP sies å være som forventet ut fra økonomisk teori. Det er rimelig å anta at høyere årlig nettoinntekt gir en økonomisk stabilitet som reduserer sannsynligheten for at en lånsøker vil misligholde. Det kan videre tenkes at flere med lav inntekt tar opp kredittkortgjeld, som kompensasjon for den lave inntekten. Personer med høyere nettoinntekt vil dermed være bedre stilt enn personer med lav inntekt, til å håndtere den pådratte kredittkortgjelden. I den tidligere litteraturen blir inntekt ansett

som en sentral forklaringsvariabel i kredittvurderingsmodeller (Hand & Henley, 1997; Steenackers & Goovaerts, 1989; Kvamme mfl., 2018). Det varierer derimot noe hvorvidt årlig eller månedlig inntekt vektlegges som viktig, men dette skyldes trolig at litteraturen anser de to variablene som det samme da de i utgangspunktet skal være perfekt korrelert. I vår studie blir ikke månedlig inntekt valgt av LASSO i noen perioder. SHAP anser derimot *«Net_Monthly_Income_AMT»* å være av betydning for det predikerte utfallet, selv om variabelens bidrag derimot er i det midtre eller lavere sjiktet. Noe overraskende indikerer SHAP-plottene at høyere verdier for *«Net_Monthly_Income_AMT»* er assosiert med mislighold. Dette er motstridene resultater sammenlignet med årlig nettoinntekt og hva som er å forvente ut fra økonomisk teori. Resultatene bekrefter dermed at det er et avvik mellom variablene i vår data, og at årlig nettoinntekt hentet fra skattedata anses som en mer viktig og troverdig variabel.

SHAP anser *«Total_Incoming_AMT»* som en av de viktigste variablene for prediksjon av mislighold. SHAP-plottene viser at høye variabelverdier er assosiert med middels høye og negative bidrag på mislighold, tilsvarende som for *«Net_Income_AMT»* og som forventet ut fra økonomisk teori. Variabelen blir derimot kun valgt i 2 perioder av LASSO. *«Total_Incoming_AMT»* angir summen av inngående beløp på debetkonto hos dataleverandør de siste 6 månedene. Det kan tenkes at inntekt utgjør en stor del av en lånsøker sin inngående beløp og at variabelen dermed fanger opp mye av den samme informasjonen som *«Net_Income_AMT»*. *«Total_Incoming_AMT»* blir ikke nevnt i den tidligere litteraturen, noe som muligens skyldes at inntektsvariabler heller inkluderes. En annen mulig forklaring kan være at tidligere litteratur ikke har hatt en tilsvarende variabel tilgjengelig. Resultatene fra SHAP indikerer dermed et unikt bidrag til litteraturen innen kredittvurdering. To andre variabler som anses viktig av SHAP er *«Total_Tax_AMT»* og *«Wealth_AMT»*. Det kan tenkes at også disse variablene gir mye av den samme informasjonen som *«Net_Income_AMT»* da begge sier noe om en persons økonomiske situasjon.

En variabel som ikke eksplisitt blir nevnt som viktig i litteraturen om kredittvurdering, men som i vår studie er av sentral betydning, er hvorvidt en person har studielån. Særlig i modellene estimert med LR er *«Student_Loan_AMT»* viktig. Regresjonskoeffisientene indikerer en negativ sammenheng mellom variabelen og mislighold, altså at større studielån

reduserer sannsynligheten for mislighold. Tilsvarende sammenheng vises i SHAP-plottene. Dette er i utgangspunktet et overraskende funn, da det virker fornuftig at økt gjeld heller øker sannsynligheten for mislighold. Det kan likevel diskuteres hvorvidt høyere studielån er en indikasjon på høyere utdanningsnivå, som av Akkoç (2012) anses å være av signifikant betydning for kredittvurdering. Denne eventuelle sammenhengen vil derimot kun stemme til en viss grad, da en person sitt utdanningsnivå ikke nødvendigvis gjenspeiles i størrelsen på studielånet. Store deler av studielånet kan være nedbetalt og dermed gi indikasjoner på lavere utdanningsnivå, selv om en person potensielt har høy utdanning. Dersom det legges til grunn en sammenheng mellom *«Student_Loan_AMT»* og utdanningsnivå virker det rimelig at økt studiegjeld reduserer sannsynligheten for mislighold, slik våre resultater viser. Personer med høyere utdanningsnivå har som regel høyere inntekt enn personer med lavere utdanningsnivå og dermed høyst sannsynlig mindre problemer med å håndtere kredittkortgjeld.

«RegisteredDebtByIncome» anses som en viktig variabel av LASSO. Variabelen angir hvor stor andel av en lånsøker sin årlige nettoinntekt som består av rentebærende gjeld registrert i gjeldsregisteret. Fortegnene til LR-regresjonskoeffisientene er positive. En høyere andel av nettoinntekten som består av rentebærende gjeld registrert i gjeldsregisteret, vil dermed øke sannsynligheten for mislighold. Dette virker som en fornuftig antakelse. Videre kan variabelviktigheten ses i sammenheng med kravet i utlånsforskriften om at lånsøkeren sin gjeldsgrad må være under 5. I datasettet angis gjeldsgraden av *«Debt_Ratio_AMT»*. Variabelen blir valgt i færre perioder enn *«RegisteredDebtByIncome»*. Dette indikerer at andelen usikret gjeld av lånsøker sin inntekt er viktigere enn andelen samlet gjeld ved prediksjon av mislighold av kredittkortgjeld. De standardiserte SHAP-verdiene viser at variabelviktigheten til *«RegisteredDebtByIncome»* er varierende. De fleste periodene viser likevel at sammenhengen mellom variabelen og mislighold er tilsvarende som for LASSO.

«Transactions_Num» anses av SHAP å være blant de viktigste variablene for å predikere mislighold av kredittkortgjeld. Variabelen blir noe motstridene kun valgt i 4 perioder av LASSO, men er da signifkant i samtlige. Variabelen inneholder informasjon om antall transaksjoner ut og inn fra debetkonto hos dataleverandør den siste avsluttede måneden. SHAP-plottene indikerer at høyere antall transaksjoner bidrar til større sannsynlighet for

mislighold. Funnet er noe ulikt den tidligere litteraturen om kredittvurdering. I litteraturen blir tidligere nedbetalingshistorikk av usikret gjeld (Somol mfl., 2005) og tilbakebetalingsstatus på kredittkortet (Sariannidis mfl., 2019) fremhevet som sentrale variabler. Selv om litteraturen viser til betalingshistorikk for usikret gjeld, mens vår variabel viser til betalingshistorikk på debetkort, kan alle tre variabler sies å omhandle en lånsøker sin betalingsmønster.

Videre anser LASSO *«Habitation_Renter»* som en viktig variabel for prediksjon av mislighold. Dette er i likhet med Hand og Henley (1997), Somol mfl. (2005) og T.-S. Lee mfl. (2006) som fant at bosituasjon er viktig. Regresjonskoeffisientene fra LR-modellene er positive, noe som tilsier at det er større sannsynlighet for mislighold dersom lånsøker leier fremfor eier bolig. Tilsvarende sammenheng indikerer SHAP-plottene. En annen variabel som omhandler lånsøker sin bosituasjon, og som LASSO anser som viktig, er *«Homeowner_NoHouseOrCoop»*. Variabelen angir om lånsøker eier hus eller er boligeier i et borettslag, gitt at lånsøker har boliglån hos dataleverandør. Dette samsvarer med resultatene fra Bellotti og Crook (2009) som fant at hvorvidt en lånsøker er boligeier er signifikant. Regresjonskoeffisientene til *«Homeowner_NoHouseOrCoop»* er lavere enn *«Habitation_Renter»*. Videre varierer fortegnene i de ulike periodene, i tillegg til at variabelen kun er signifikant i én periode. Dette kan skyldes at det er en sterk positiv korrelasjon mellom de to variablene (ρ=0.564). Videre kan dette påvirke hvilken variabel som blir valgt av metodene da de i enkelte tilfeller kan erstatte hverandre. I tillegg inngår variabelen *«Habitation_Homeowner»* i datasettet. Variabelen baserer seg på hvorvidt lånsøker selv oppgir å være boligeier. Variabelen er ikke signifikant i noen perioder, noe som kan skyldes at også denne variabelen er sterkt korrelert med *«Habitation_Renter»* (ρ=-0.606). Denne sammenhengen er også gjeldende i SHAP, hvor både *«Homeowner_NoHouseOrCoop»* og *«Habitation_Homeowner»* generelt har lavere variabelviktighet enn *«Habitation_Renter»*.

Bellotti og Crook (2009) fant at type kredittkort er av signifikant betydning for prediksjon av mislighold. I vår studie anser både LASSO og SHAP to ulike typer kredittkort, *«MC_2»* og *«MC_4»*, å påvirke prediksjonen av mislighold. I tillegg blir *«MC_2_Young»* valgt av LASSO i enkelte perioder, mens *«MC_4_Young»* har lave standardiserte SHAP-

verdier i de fleste periodene. Resultatene indikerer mindre sannsynlighet for mislighold dersom en lånsøker har produktene *«MC_2»* og *«MC_2_Young»*, samt større sannsynlighet for mislighold for produktypene *«MC_4»* og *«MC_4_Young»*. Våre resultater indikerer dermed, i likhet med litteraturen, at produkttype kan være av betydning ved prediksjon av mislighold av kredittkortgjeld. Resultatene tyder på at kundesegmentet som benytter *«MC_2»*-produkter har en lavere kredittrisiko i forhold til de som benytter *«MC_4»*-produkter.

SHAP anser *«DebtRegisterIELA»* som en av de viktigste variablene i mange av periodene. Variabelen angir lånsøker sin samlede rentebærende gjeld i gjeldsregisteret. Dette er i likhet med Kvamme mfl. (2018), som fant at nettopp saldo på andre lån ofte benyttes av kredittbyråer. Videre kan vi se en sammenheng med Dimitras mfl. (2017), Sariannidis mfl. (2019) og Hamori mfl. (2018), som fant at tilbakebetalingsstatus på kredittkortet per måned er den viktigste variabelen. Dette sier, i likhet med *«DebtRegisterIELA»*, noe om hvor mye penger lånsøker har i usikret gjeld. Videre er sammenhengen mellom mislighold og *«DebtRegisterIELA»* som forventet, da SHAP-plottene viser at høye verdier på variabelen assosieres med større sannsynlighet for mislighold. LASSO anser derimot ikke denne variabelen som viktig, da den ikke velges i noen av periodene. Likevel betrakter LASSO *«RegisteredDebtByIncome»*, som også inneholder informasjon om en lånsøker sin usikrede gjeld, som en viktig variabel.

Flere av våre resultater stemmer godt overens med den tidligere litteraturen innen kredittvurdering. Det er likevel også flere variabler som anses som viktige for kredittvurdering i litteraturen, men som ikke blir ansett som viktig i vår studie. Lengde på nåværende arbeidsforhold (Steenackers & Goovaerts, 1989; Hand & Henley, 1997) og total arbeidsvarighet (Akkoç, 2012) blir i litteraturen ansett som viktige variabler for prediksjon av mislighold av kredittkortgjeld. I vårt datasett har vi ingen variabel for total arbeidsvarighet, så det er derfor ikke overraskende at vi ikke har samme funn. Det er likevel rimelig å anta at alder er korrelert med total arbeidsvarighet, da eldre stort sett har jobbet lengre enn yngre. Vår studie finner at alder er av betydning for prediksjon av mislighold og funnet kan dermed sies å samstemme med litteraturen. Lengde på nåværende arbeidsforhold er inkludert i vår data, men anses derimot ikke å ha høy variabelviktighet av verken

LASSO eller SHAP. Det kan argumenteres for at gjelds- og inntektsdataen inkludert i vår data gir mer presis og korrekt informasjon, sammenlignet med variabelen for lengde på nåværende arbeidsforhold. At en person er i jobb trenger ikke nødvendigvis å bety at personen har mye penger. Det kan dermed tenkes at *«Net_Income_AMT»*, som anses å ha høy variabelviktighet i våre studier, fanger opp mye av den samme informasjonen. Dette kan være en mulig årsak til at variabelviktigheten til lengden på nåværende arbeidsforhold reduseres. Tilsvarende argument kan også gjøres gjeldende for yrke, som av Steenackers og Goovaerts (1989) og T.-S. Lee mfl. (2006) anses å være av sentral betydning. Avslutningsvis blir også holdning til penger ansett som viktig for prediksjon av mislighold i litteraturen (Tsai mfl., 2009). Igjen er dette informasjon som ikke er tilgjengelig i vår data og det er dermed ikke mulig å si noe om variabelviktigheten i vår studie.

To av våre resultater fra LASSO indikerer unike funn, sammenlignet med den tidligere litteraturen innen kredittvurdering. For det første anses *«AppliedByGranted»* å være viktig, da den blir valgt av LASSO i samtlige perioder. Variabelen er en beregning av kredittkortgrensen det er søkt om, delt på innvilget kredittkortgrense. Variabelverdien vil dermed være større eller lik 1, hvor 1 indikerer at en lånsøker har fått innvilget kredittkortgrensen det er søkt om. Våre resultater viser en positiv sammenheng, altså at økt variabelverdi øker sannsynligheten for mislighold. Dette virker fornuftig da personer som søker om høyere kredittkortgrense enn det de får innvilget, ønsker tilgang på mer penger enn hva deres betalingsevne tilsier at de kan betjene. I den tidligere litteraturen er kredittgrense (T.-S. Lee mfl., 2006) og innvilget kredittgrense (Sariannidis mfl., 2019) inkludert, men ikke en tilsvarende variabel som *«AppliedByGranted»*. Resultatene fra LASSO indikerer dermed at en slik variabel med fordel kan inkluderes i modeller for kredittvurdering. Videre blir også variabelen *«App_Night»* valgt ut i samtlige perioder av LASSO, i tillegg til å være signifikant i 8 av disse. De estimerte koeffisientverdiene indikerer høyere sannsynlighet for mislighold dersom en person søker kredittkort om natten, fremfor om dagen. Vi har ikke sett at søknadstidspunkt har vært inkludert i annen tidligere litteratur innenfor kredittvurdering, og funnet kan derfor sies å være unikt. Det er likevel viktig å påpeke at resultatene fra SHAP indikerer lavere variabelviktighet for både *«AppliedByGranted»* og *«App_Night»*. De standardiserte SHAP-verdiene for *«AppliedByGranted»* kan sies å være

middels til lave, mens verdiene for *«App_Night»* er lave med flere verdier lik 0. Resultatene fra SHAP svekker dermed funnene gjort med LASSO. De motstridene resultatene indikerer likevel at begge variablene kan være sentrale i modeller for kredittvurdering og at det bør undersøkes nærmere hvorvidt variablene faktisk er av signifikant betydning for prediksjon av mislighold av kredittkortgjeld.

Datasettet inneholder flere variabler med informasjon fra gjeldsregisteret. Vår dataleverandør var interessert i å se hvilken effekt disse variablene har for prediksjon av mislighold, da gjeldsregisteret ble innført for å redusere misligholdsandelen i Norge. SHAP og LASSO anser *«RegisteredDebtByIncome»* og *«DebtRegisterIELA»* å være av betydning for prediksjon av mislighold. En mulig årsak for at ikke flere variabler blir ansett som viktige er at datasettet vårt kun inneholder innvilgede søknader. Søknader hvor informasjon fra gjeldsregisteret er avgjørende ved vurderingen, vil ofte ekskluderes allerede før innvilgelse av kredittkort. Likevel indikerer våre resultater at gjeldsregisteret har en betydning for prediksjon av mislighold av kredittkortgjeld.

6.2 Modellprediksjon

Våre resultater støtter bransjestandarden om bruk av LR-modeller for kredittvurdering (Dong mfl., 2010). Prediksjonsevnen til LR-modellene er stort sett noe bedre enn for XGBoost-modellene. AUC-verdiene for periodene 4, 7 og 10, samt Brier Scoren for periode 1, indikerer derimot at XGBoost-modellene presterer best i disse periodene. Våre resultater impliserer likevel at LR-modellene gir tilsvarende gode eller bedre prediksjonsresultater på testsettet enn XGBoost-modellene. I tillegg gir LR-modellene mer tolkbare resultater, noe som også er sentralt ved kredittvurdering (Kvamme mfl., 2018).

En mulig årsak til at LR-modellene overordnet oppnår noe bedre prediksjonevne på testsettet, kan skyldes at LASSO velger ut de viktigste variablene før LR-modellene estimeres. LR-modellene blir dermed estimert med kun de forklaringsvariablene LASSO anser å være viktigst, i motsetning til XGBoost-modellene som estimeres med alle forklaringsvariablene i det endelige datasettet. At irrelevante variabler ekskluderes før modellestimering kan dermed være av sentral betydning, slik det også fremgår i litteraturen (Paraschiv mfl., 2021; Tripathi mfl., 2020).

Optimaliseringen av hyperparameterne i XGBoost-modellene, kan være en annen mulig årsak til at LR-modellene generelt presterer noe bedre. Algoritmen til XGBoost krever optimalisering av flere hyperparametere, i motsetning til LR som i vårt tilfelle kun krever optimalisering av λ-verdien som benyttes i LASSO. XGBoost-modellene er dermed i større grad påvirket av våre valg av hyperparametere, sammenlignet med LR-modellene. Optimaliseringen vil i stor grad påvirke resultatene, noe studien til Xia mfl. (2017) tydelig viser.

6.3 Tolkning av svarte bokser

Som resultatene viser, kan tilsvarende plott som figur 4 og figur 5 med fordel benyttes til å gjøre svarte bokser tolkbare for begrunnelse for kredittvurderinger. For det første kan plottene gis til en lånsøker som forklaring på hvorfor kredittkortsøknaden ble innvilget eller avslått. Lånsøker får da innsikt i hvilke variabler som påvirker avgjørelsen mest, samt i hvilken retning de ulike variablene bidrar. Slike plott øker dermed gjennomsiktigheten i beslutningene som tas. Videre vil kredittinstitusjoner som benytter slike plott for begrunnelse, oppfylle personopplysningslovens krav om en lånsøkers rett på innsyn. I tillegg vil det være enklere for kredittinstitusjoner å etterse de automatiserte beslutningene.

Fremstillingen av den lokale variabelviktigheten i figur 4 og figur 5 er kun én mulig løsning på hvordan prediksjoner gjort av svarte bokser kan gjøres mer tolkbare. Vi har kun inkludert de 10 forklaringsvariablene som har størst bidrag på prediksjonen, men antallet kan videre utvides for mer innsikt. Det er likevel ikke en selvfølge at slike plott gjør tolkningen av en gitt prediksjon enkel. Figur 4 gir ikke like klare indikasjoner på at det predikerte utfallet er mislighold, slik som figur 5 gir indikasjoner på at det ikke er misligholdt. For låntaker A er det dermed ikke umiddelbart logisk hva prediksjonsutfallet er. Dette er problematisk da det med stor sannsynlighet er nettopp kundene som får avslag på en søknad som eventuelt vil kreve en begrunnelse. At begrunnelsen i enkelte tilfeller er noe tvetydig, svekker plottets funksjon.

7 Konklusjon

Denne masteroppgaven har undersøkt hvorvidt enkelte variabler ved lånsøkere er sentrale å hensynta i vurderingsprosessen før innvilgelse av kredittkortgjeld. Videre har vi også forsøkt å besvare hvilken av estimeringsteknikkene LR og XGBoost som gir best prediksjonsevne for kredittvurdering, samt vise hvordan svarte bokser kan gjøres tolkbare for begrunnelse av kredittvurderinger. Vi vil i dette kapittelet oppsummere våre funn, samt implikasjonene de kan ha. Videre vil vi presentere svakheter ved vår studie, før vi avslutningsvis vil komme med forslag til videre forskning.

7.1 Våre funn og implikasjoner

Denne masteroppgaven har forsøkt å besvare problemstillingen presentert innledningsvis: *«Hvilke karakteristikker ved lånsøkere påvirker sannsynligheten for mislighold av kredittkortgjeld?»*. Vårt hovedfunn viser at en lånsøker sin alder og årlige nettoinntekt er av sentral betydning ved prediksjon av mislighold av kredittkortgjeld. Dette funnet er i vår studie gjeldende uavhengig av settet med teknikker som er benyttet, altså uavhengig av om prediksjonsmodellene er estimert ved bruk av LR og LASSO eller XGBoost og SHAP. I tillegg til de 2 nevnte forklaringsvariablene, indikerer LR og LASSO videre at 5 andre karakteristikker ved en lånsøker er av vesentlig betydning for prediksjon av mislighold av kredittkortgjeld. En lånsøkers studielån, hvor stor andel av inntekten som består av usikret gjeld registrert i gjeldsregisteret, søkt kredittkortgrense delt på innvilget kredittkortgrense, hvorvidt en lånsøker leier bolig, samt om søknaden ble sendt om natten er også signifikante variabler for kredittvurdering. I tillegg til hovedfunnet indikerer modellene estimert ved bruk at XGBoost og SHAP at en lånsøker sin formue, totalt betalt skattebeløp, antall transaksjoner og sum av inngående beløp på debetkonto hos dataleverandør de siste 6 månedene påvirker sannsynligheten for mislighold. Videre viser våre resultater at variablene basert på gjeldsregisteret er av betydning da kundens samlede rentebærende gjeld i gjeldsregisteret og hvor stor andel av inntekten som består av rentebærende gjeld registrert i gjeldsregisteret anses som viktig. Vi finner at tre variabler, som tidligere ikke er nevnt i litteraturen, er avgjørende for prediksjon av mislighold av kredittkortgjeld. Dette

er variabler som inneholder informasjon om hvorvidt en kredittkortsøknad er sendt inn på natten, sum av inngående beløp på debetkonto hos dataleverandør de siste 6 månedene, samt forholdet mellom søkt og innvilget kredittkortgrense. Dette er dermed et unikt bidrag til litteraturen innen kredittvurdering.

Videre har vi forsøkt å besvare våre to forskningsspørsmål: «*Hvilken estimeringsteknikk gir best prediksjonsevne ved utvikling av modeller for kredittvurdering?*» og «*Hvordan kan svarte bokser gjøres tolkbare for begrunnelse av kredittvurderinger?*». Vår studie viser overordnet at LR gir noe bedre prediksjonsevne for kredittvurdering enn XGBoost, selv om forskjellen kan sies å være minimal. Vårt resultat støtter dermed opp under bankindustriens bruk av LR i kredittvurderingsmodeller. Videre har vi bevist at svarte bokser kan gjøres tolkbare for begrunnelse av kredittvurderinger ved å benytte plott som viser den lokale variabelviktigheten for de 10 variablene som har størst bidrag på en gitt prediksjon.

Vår studie gir implikasjoner for bankindustrien om hvilke forklaringsvariabler som bør inngå i kredittvurderingsmodeller for best mulig prediksjonsevne. I tillegg impliserer studien hvilke forklaringsvariabler som med fordel kan utelates fra kredittvurderingsmodeller, da våre to benyttede sett med teknikker konkluderer med at enkelte variabler ikke har noen betydning på prediksjonen av mislighold av kredittkortgjeld. Vår studie bidrar dermed med innsikt som videre kan benyttes av banker og andre finansinstitusjoner til å forbedre eksisterende kredittvurderingsmodeller. Det er rimelig å anta at en forbedring av eksisterende kredittvurderingsmodeller kan bidra til mer treffsikre og rettferdige kredittvurderinger. Dette vil videre kunne føre til en mer nøyaktig utskillelse av kunder som ikke bør få innvilget kredittkort, og dermed være med å redusere gjeldsproblematikken i Norge. I tillegg viser vår studie hvordan plott kan benyttes for å gjøre svarte bokser tolkbare, og dermed oppfylle kravet om en lånsøkers rett på innsikt.

7.2 Svakheter

Det er viktig å påpeke at vår studie har benyttet data basert på innvilgede kredittkortsøknader, noe som har ført til en bias i våre modeller. Dette skyldes at mange av kredittkortsøknadene som ikke anses tilfredsstillende, allerede er ekskludert fra vurderingsprosessen og dermed ikke inngår i vår data. Dette kan bety at våre modeller i praksis ville ha inn-

vilget mange av kredittkortsøknadene som i dag blir avslått av dagens vurderingssystem hos dataleverandør.

En annen svakhet ved vår studie er at utfallsperioden for mislighold er relativt kort, da våre modeller predikerer hvorvidt en lånsøker vil misligholde innen ett år. Dersom vi hadde benyttet en lengre utfallsperiode, kan det tenkes at andelen misligholdt hadde vært større.

Da vi benyttet k-NN til å håndtere manglende verdier, er det rimelig å anta at enkelte verdier ikke nødvendigvis var representable og dermed feilaktig kan ha påvirket resultatene. I tillegg er resultatene avhengig av variablene vi valgte å benytte som sammenligningsgrunnlag i k-NN. Videre har vi ikke vurdert kostnadene ved å innføre, samt opprettholde, en kredittvurderingsmodell i praksis.

7.3 Forslag til videre forskning

Da vår studie er gjennomført med et unikt datasett, hadde det vært interessant å gjennomføre tilsvarende analyser med andre datasett. Dette kunne i større grad vært med på å generalisere resultatene for det norske markedet. I tillegg vil det være aktuelt å se mer på denne studiens nyvinninger, altså betydningen av hvorvidt en søknad er sendt inn på natten, sum av inngående beløp på debetkonto hos dataleverandør de siste 6 månedene samt forholdet mellom søkt og innvilget kredittkortgrense. Videre ville det også vært interessant å kjøre analysene på nytt med akkurat samme variabelsett i LR- og XGBoost-modellene, da LR-modellene i vår studie er estimert med færre variabler ettersom det først ble gjennomført en variabelseleksjon. Da den korte utfallsperioden for mislighold kan anses som en svakhet ved vår studie, hadde det også vært aktuelt å estimere modeller med lengre utfallsperiode for deretter å sammenligne resultatene. Avslutningsvis hadde det vært interessant å estimere kredittvurderingsmodeller ved bruk av andre estimeringsteknikker på tilsvarende datasett som benyttet i denne studien.

Referanser

Abdou, H., El-Masry, A. & Pointon, J. (2007). On the applicability of credit scoring models in Egyptian banks. *Banks and bank systems*, *2*(1), 4–20.

Abdou, H. A. & Pointon, J. (2011). Credit Scoring, Statistical Techniques and Evaluation Criteria: A Review of the Literature. *Intelligent Systems in Accounting, Finance and Management*, *18*(2-3), 59–88. https://doi.org/10.1002/isaf.325

Abdou, H. A., Tsafack, M. D. D., Ntim, C. G. & Baker, R. D. (2016). Predicting creditworthiness in retail banking with limited scoring data. *Knowledge-Based Systems*, *103*, 89–103. https://doi.org/10.1016/j.knosys.2016.03.023

Abellán, J. & Mantas, C. J. (2014). Improving experimental studies about ensembles of classifiers for bankruptcy prediction and credit scoring. *Expert systems with applications*, *41*(8), 3825–3830. https://doi.org/10.1016/j.eswa.2013.12.003

Afrilia, A., Joharudin, A., Zaky, M., Budiman, B. & Fauziah, M. (2021). Credit scoring model using MARS method to comply with FSA regulation. *Journal of Physics: Conference Series*, *1869*(1), 012135. https://doi.org/10.1088/1742-6596/1869/1/012135

Ahmad, A. U. & Starkey, A. (2018). Application of feature selection methods for automated clustering analysis: A review on synthetic datasets. *Neural Computing and Applications*, *29*(7), 317–328. https://doi.org/10.1007/s00521-017-3005-9

Akkoç, S. (2012). An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis: The case of Turkish credit card data. *European Journal of Operational Research*, *222*(1), 168–178. https://doi.org/10.1016/j.ejor.2012.04.009

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance*, *23*(4), 589–609. https://doi.org/10.2307/2978933

Bellotti, T. & Crook, J. (2009). Support vector machines for credit scoring and discovery of significant features. *Expert Systems with Applications*, *36*(2), 3302–3308. https://doi.org/10.1016/j.eswa.2008.01.005

Bergstra, J., Komer, B., Eliasmith, C., Yamins, D. & Cox, D. D. (2015). Hyperopt: A Python library for model selection and hyperparameter optimization. *Computational Science & Discovery*, *8*(1), 14008–14024. https://doi.org/10.1088/1749-4699/8/1/014008

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984). *Classification And Regression Trees* (1. utg.). Wadsworth International Group. https://doi.org/10.1201/9781315139470

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Brown, I. & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert systems with applications*, *39*(3), 3446–3453. https://doi.org/10.1016/j.eswa.2011.09.033

Brownlee, J. (2016a, 8. september). *A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning*. Machine Learning Mastery. https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/

Brownlee, J. (2016b, 18. desember). *How To Backtest Machine Learning Models for Time Series Forecasting*. Machine Learning Mastery. https://machinelearningmastery.com/backtest-machine-learning-models-time-series-forecasting/

Brzezinski, D. & Stefanowski, J. (2017). Prequential AUC: Properties of the area under the ROC curve for data streams with concept drift. *Knowledge and Information Systems*, *52*(2), 531–562. https://doi.org/10.1007/s10115-017-1022-8

Casalicchio, G., Molnar, C. & Bischl, B. (2019). Visualizing the Feature Importance for Black Box Models. I M. Berlingerio, F. Bonchi, T. Gärtner, N. Hurley & G. Ifrim (Red.). Springer International Publishing. https://doi.org/10.1007/978-3-030-10925-7_40

Chandrashekar, G. & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024

Chava, S. & Jarrow, R. A. (2004). Bankruptcy Prediction with Industry Effects*. *Review of Finance*, *8*(4), 537–569. https://doi.org/10.1093/rof/8.4.537

Chen, F.-L. & Li, F.-C. (2010). Combination of feature selection approaches with SVM in credit scoring. *Expert Systems with Applications*, *37*(7), 4902–4909. https://doi.org/10.1016/j.eswa.2009.12.025

Chen, H. & Xiang, Y. (2017). The Study of Credit Scoring Model Based on Group Lasso. *Procedia Computer Science*, *122*, 677–684. https://doi.org/10.1016/j.procs.2017.11.423

Chen, T. (2015, 4. september). *What is the difference between the R gbm (gradient boosting machine) and xgboost (extreme gradient boosting)?* Quora. https://www.quora.com/What-is-the-difference-between-the-R-gbm-gradient-boosting-machine-and-xgboost-extreme-gradient-boosting/answer/Tianqi-Chen-1

Chen, T. & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13–17*, 785–794. https://doi.org/10.1145/2939672.2939785

Chiang, R. C., Chow, Y.-F. & Liu, M. (2002). Residential Mortgage Lending and Borrower Risk: The Relationship Between Mortgage Spreads and Individual Characteristics. *The journal of real estate finance and economics*, *25*(1), 5–32. https://doi.org/10.1023/A:1015347516812

Chuang, C.-L. & Lin, R.-H. (2009). Constructing a reassigning credit scoring model. *Expert Systems with Applications*, *36*, 1685–1694. https://doi.org/10.1016/j.eswa.2007.11.067

Crook, J. & Banasik, J. (2004). Does reject inference really improve the performance of application scoring models? *Journal of Banking & Finance*, *28*(4), 857–874. https://doi.org/10.1016/S0378-4266(03)00203-6

Datatilsynet. (2018 januar). *Kunstig intelligens og personvern*. https://www.datatilsynet.no/regelverk-og-verktoy/rapporter-og-utredninger/kunstig-intelligens/

Dimitras, A. I., Papadakis, S. & Garefalakis, A. (2017). Evaluation of empirical attributes for credit risk forecasting from numerical data. *Investment Management and Financial Innovations*, *14*(1), 9–18. https://doi.org/10.21511/imfi.14(1).2017.01

Dong, G., Lai, K. K. & Yen, J. (2010). Credit scorecard based on logistic regression with random coefficients. *Procedia Computer Science*, *1*(1), 2463–2468. https://doi.org/10.1016/j.procs.2010.04.278

Durand, D. (1941). Review of Risk Elements in Consumer Instalment Financing. *Journal of Marketing*, *6*(4), 407–408. https://doi.org/10.2307/1246534

Efromovich, S. (2010). Oracle inequality for conditional density estimation and an actuarial example. *Annals of the Institute of Statistical Mathematics*, *62*(2), 249–275. https://doi.org/10.1007/s10463-008-0185-1

Finansdepartementet. (2017, 4. april). *Strengere krav til fakturering av kredittkortgjeld*. Regjeringen.no. https://www.regjeringen.no/no/dokumentarkiv/regjeringen-solberg/aktuelt-regjeringen-solberg/fin/nyheter/2017/kredittkortgjeld/id2547981/

Finansdepartementet. (2021, 4. oktober). *Utlånsforskriften*. Regjeringen.no. https://www.regjeringen.no/no/tema/okonomi-og-budsjett/finansmarkedene/utlansforskriften/id2791101/

Finanstilsynet. (2022). *Utviklingen i forbruksgjeld april 2022*. https://finanstilsynet.no/contentassets/e238d68d2f0d4cfb8e412683307ec20f/utviklingen-i-forbruksgjeld-april-2022.pdf

Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, *7*(2), 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics*, *19*(1), 1–67. https://doi.org/10.1214/aos/1176347963

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Gerdrup, K. R. & Bakke, B. (2002). Bankenes motpartsrisiko - resultater fra en kartlegging gjennomført av Norges Bank og Kredittilsynet. *Norges Bank*, 10. https://www.norges-bank.no/globalassets/upload/publikasjoner/penger_og_kreditt/2002-04/gerdrup.pdf

Gunnarsson, B. R., vanden Broucke, S., Baesens, B., Óskarsdóttir, M. & Lemahieu, W. (2021). Deep learning for credit scoring: Do or don't? *European Journal of Operational Research*, *295*(1), 292–305. https://doi.org/10.1016/j.ejor.2021.03.006

Hamori, S., Kawai, M., Kume, T., Murakami, Y. & Watanabe, C. (2018). Ensemble Learning or Deep Learning? Application to Default Risk Analysis. *Journal of Risk and Financial Management*, *11*(1), 12. https://doi.org/10.3390/jrfm11010012

Hand, D. J. (2002). Superscorecards. *IMA Journal of Management Mathematics*, *13*(4), 273–281. https://doi.org/10.1093/imaman/13.4.273

Hand, D. J. & Henley, W. E. (1997). Statistical Classification Methods in Consumer Credit Scoring: A Review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *160*(3), 523–541. https://doi.org/10.1111/j.1467-985X.1997.00078.x

Hand, D. J., Oliver, J. J. & Lunn, A. D. (1998). Discriminant analysis when the classes arise from a continuum. *Pattern Recognition*, *31*(5), 641–650. https://doi.org/10.1016/S0031-3203(97)00083-6

Harrell, F. E. & Lee, K. L. (1985). A Comparison of the Discrimination of Discriminant Analysis and Logistic Regression. *Statistics in Biomedical, Public Health and Environmental Sciences*, 333–343.

Hosmer, D. W., Lemeshow, S. & Sturdivant, R. X. (2013). *Applied logistic regression* (3. utg). John Wiley and Sons. https://doi.org/10.1002/9781118548387

Hsieh, N.-C. & Hung, L.-P. (2010). A data driven ensemble classifier for credit scoring analysis. *Expert Systems with Applications*, *37*(1), 534–545. https://doi.org/10.1016/j.eswa.2009.05.059

Huang, C.-L., Chen, M.-C. & Wang, C.-J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications*, *33*(4), 847–856. https://doi.org/10.1016/j.eswa.2006.07.007

Jain, A. (2016, 1. mars). *XGBoost Parameters | XGBoost Parameter Tuning*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Bd. 103). Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

Jorly, J. (2021, 5. juni). *XGBOOST — IN A NUTSHELL*. Medium. https://ai.plainenglish.io/xgboost-in-a-nutshell-211e170e8b48

Kamath, U. & Liu, J. (2021). *Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning*. Springer International Publishing. https://doi.org/10.1007/978-3-030-83356-5

Karlis, D. & Rahmouni, M. (2006). Analysis of defaulters' behaviour using the Poisson-mixture approach. *IMA Journal of Management Mathematics*, *18*(3), 297–311. https://doi.org/10.1093/imaman/dpm025

Koehrsen, W. (2018). *Intro to Model Tuning: Grid and Random Search*. Kaggle. https://kaggle.com/willkoehrsen/intro-to-model-tuning-grid-and-random-search

Kommunal- og distriktsdepartementet. (2019, 30. oktober). *Ny personopplysningslov*. Regjeringen.no. https://www.regjeringen.no/no/tema/statlig-forvaltning/personvern/ny-personopplysningslov/id2340094/

Kvamme, H., Sellereite, N., Aas, K. & Sjursen, S. (2018). Predicting mortgage default using convolutional neural networks. *Expert Systems with Applications*, *102*(17), 207–217. https://doi.org/10.1016/j.eswa.2018.02.029

Laborda, J. & Ryoo, S. (2021). Feature Selection in a Credit Scoring Model. *Mathematics*, *9*(7), 746. https://doi.org/10.3390/math9070746

Ladyzynski, P., Zbikowski, K. & Grzegorzewski, P. (2013). Stock Trading with Random Forests, Trend Detection Tests and Force Index Volume Indicators. I L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L. A. Zadeh & J. M. Zurada (Red.). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-38610-7

Lee, S.-H. (2021, 30. oktober). *R-bloggers*. lambda.min, lambda.1se and Cross Validation in Lasso : Binomial Response. https://www.r-bloggers.com/2021/10/lambda-min-lambda-1se-and-cross-validation-in-lasso-binomial-response/

Lee, T. & Chen, I. (2005). A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications*, *28*(4), 743–752. https://doi.org/10.1016/j.eswa.2004.12.031

Lee, T.-S., Chiu, C.-C., Chou, Y.-C. & Lu, C.-J. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics & Data Analysis*, *50*(4), 1113–1130. https://doi.org/10.1016/j.csda.2004.11.006

Lee, T.-S., Chiu, C.-C., Lu, C.-J. & Chen, I.-F. (2002). Credit scoring using the hybrid neural discriminant technique. *Expert Systems with Applications*, *23*(3), 245–254. https://doi.org/10.1016/S0957-4174(02)00044-1

Lee, T. H. & Jung, S.-C. (1999). Forecasting creditworthiness: Logistic vs. artificial neural net. *The Journal of Business Forecasting Methods & Systems*, *18*(4), 28–30.

Lejoyeux, M., Richoux-Benhaim, C., Löhnardt, H. & Lequen, V. (2011). Money Attitude, Self-esteem, and Compulsive Buying in a Population of Medical Students. *Frontiers in Psychiatry*, *2*(13). https://www.frontiersin.org/article/10.3389/fpsyt.2011.00013

Lessmann, S., Baesens, B., Seow, H.-V. & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, *247*(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J. & Liu, H. (2018). Feature Selection: A Data Perspective. *ACM Computing Surveys*, *50*(6), 1–45. https://doi.org/10.1145/3136625

Li, J., Chang, M., Tian, P., Chen, L. & Mu, X. (2020). Personal Credit Scoring via Logistic Regression with Elastic Net Penalty. I Y. Jia, J. Du & W. Zhang (Red.), *Proceedings of 2019 Chinese Intelligent Systems Conference* (s. 422–428). Springer Singapore. https://doi.org/10.1007/978-981-32-9682-4_44

Li, K., Niskanen, J., Kolehmainen, M. & Niskanen, M. (2016). Financial innovation: Credit default hybrid model for SME lending. *Expert Systems with Applications*, *61*(28), 343–355. https://doi.org/10.1016/j.eswa.2016.05.029

Li, S., Shiue, W. & Huang, M. (2006). The evaluation of consumer loans using support vector machines. *Expert Systems with Applications*, *30*(4), 772–782. https://doi.org/10.1016/j.eswa.2005.07.041

Liu, Y. & Schumann, M. (2005). Data mining feature selection for credit scoring models. *The Journal of the Operational Research Society*, *56*(9), 1099–1108. https://doi.org/10.1057/palgrave.jors.2601976

Louzada, F., Ara, A. & Fernandes, G. B. (2016). Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research*

and Management Science, *21*(2), 117–134. https://doi.org/10.1016/j.sorms.2016.10.001

Lundberg, S. (2018). *Bar Plot*. SHAP documentation. Hentet 15. mai 2022, fra https://shap.readthedocs.io/en/latest/example_notebooks/api_examples/plots/bar.html

Lundberg, S. & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. I U. von Luxburg, I. Guyon, S. Bengio, H. Wallach & R. Fergus (Red.). Curran Associates Inc. https://arxiv.org/abs/1705.07874

Mahmoudi, N. & Duman, E. (2015). Detecting credit card fraud by Modified Fisher Discriminant Analysis. *Expert Systems with Applications*, *42*(5), 2510–2516. https://doi.org/10.1016/j.eswa.2014.10.037

Maldonado, S., Pérez, J. & Bravo, C. (2017). Cost-based feature selection for Support Vector Machines: An application in credit scoring. *European Journal of Operational Research*, *261*(2), 656–665. https://doi.org/10.1016/j.ejor.2017.02.037

Marques, A. I., Garcia, V. & Sanchez, J. S. (2012). Exploring the behaviour of base classifiers in credit scoring ensembles. *Expert systems with applications*, *39*(11), 10244–10250. https://doi.org/10.1016/j.eswa.2012.02.092

Marqués, A., García, V. & Sánchez, J. (2012). Two-level classifier ensembles for credit risk assessment. *Expert Systems with Applications*, *39*(12), 10916–10922. https://doi.org/10.1016/j.eswa.2012.03.033

Molnar, C. (2022). *Interpretable Machine Learning: A Guide For Making Black Box Models Explainable* (2. utg.). https://christophm.github.io/interpretable-ml-book/shap.html

Mushava, J. & Murray, M. (2022). A novel XGBoost extension for credit scoring class-imbalanced data combining a generalized extreme value link and a modified focal loss function. *Expert Systems with Applications*, *202*, 117233. https://doi.org/10.1016/j.eswa.2022.117233

Myers, J. & Forgy, E. (1963). The Development of Numerical Credit Evaluation Systems. *Journal of the American Statistical Association*, *58*(303), 799–806. https://doi.org/10.1080/01621459.1963.10500889

Nguyen, H. D., Truong, G. T. & Shin, M. (2021). Development of extreme gradient boosting model for prediction of punching shear resistance of r/c interior slabs. *Engineering Structures*, *235*, 112067. https://doi.org/10.1016/j.engstruct.2021.112067

Norges Bank. (2021, 30. juni). *Det norske finansielle systemet 2021* (Nr. 2021). https://www.norges-bank.no/aktuelt/nyheter-og-hendelser/Publikasjoner/det-norske-finansielle-systemet/2021-dnfs/

Nugent, C. (2018, 17. januar). *Gradient Boosting and Parameter Tuning in R*. Kaggle. https://kaggle.com/camnugent/gradient-boosting-and-parameter-tuning-in-r

Nyitrai, T. & Virág, M. (2019). The effects of handling outliers on the performance of bankruptcy prediction models. *Socio-Economic Planning Sciences*, *67*, 34–42. https://doi.org/10.1016/j.seps.2018.08.004

Orgler, Y. E. (1970). A Credit Scoring Model for Commercial Loans. *Journal of Money, Credit and Banking*, *2*(4), 435–445. https://doi.org/10.2307/1991095

Paraschiv, F., Schmid, M. & Wahlstrøm, R. R. (2021). *Bankruptcy Prediction of Privately Held SMEs Using Feature Selection Methods* (SSRN Scholarly Paper). Social Science Research Network. https://doi.org/10.2139/ssrn.3911490

Pavlidis, N. G., Tasoulis, D. K., Adams, N. M. & Hand, D. J. (2012). Adaptive consumer credit classification. *The Journal of the Operational Research Society*, *63*(12), 1645–1654. https://doi.org/10.1057/jors.2012.15

Piramuthu, S. (1999). Financial credit-risk evaluation with neural and neurofuzzy systems. *European Journal of Operational Research*, *112*(2), 310–321. https://doi.org/10.1016/S0377-2217(97)00398-6

Provenzano, A. R., Trifirò, D., Datteo, A., Giada, L., Jean, N., Riciputi, A., Pera, G. L., Spadaccino, M., Massaron, L. & Nordio, C. (2020). *Machine Learning Approach for Credit Scoring*. http://arxiv.org/abs/2008.01687

Ravi Kumar, P. & Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review. *European Journal of Operational Research*, *180*(1), 1–28. https://doi.org/10.1016/j.ejor.2006.08.043

Regjeringen. (2017, 4. april). *Spørsmål og svar om endring av regler for markedsføring av kreditt*. Regjeringen.no. https://www.regjeringen.no/no/tema/okonomi-og-

budsjett/finansmarkedene/sporsmal-og-svar-om-endring-av-regler-for-markedsforing-av-kreditt/id2548055/

Regjeringen. (2019, 5. juli). *Gjeldsinformasjonsloven*. Regjeringen.no. https://www.regjeringen.no/no/tema/forbruker/gjeldsinformasjonsloven/id2510537/

Reichert, A. K., Cho, C.-C. & Wagner, G. M. (1983). An Examination of the Conceptual Issues Involved in Developing Credit-Scoring Models. *Journal of Business & Economic Statistics*, *1*(2), 101–114. https://doi.org/10.1080/07350015.1983.10509329

Ribeiro, M. T., Singh, S. & Guestrin, C. (2016a). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Ribeiro, M. T., Singh, S. & Guestrin, C. (2016b). Model-Agnostic Interpretability of Machine Learning. *ArXiv*. http://arxiv.org/abs/1606.05386

Ringdal, K. (2018). *Enhet og mangfold: Samfunnsvitenskapelig forskning og kvantitativ metode* (4. utg.). Fagbokforlaget.

Sagi, O. & Rokach, L. (2021). Approximating XGBoost with an interpretable decision tree. *Information Sciences*, *572*, 522–542. https://doi.org/10.1016/j.ins.2021.05.055

Saraswat, M. (2016). *Beginners Tutorial on XGBoost and Parameter Tuning in R*. Hackerearth. https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/beginners-tutorial-on-xgboost-parameter-tuning-r/tutorial/

Sariannidis, N., Papadakis, S., Garefalakis, A., Lemonakis, C. & Kyriaki-Argyro, T. (2019). Default avoidance on credit card portfolios using accounting, demographical and exploratory factors: Decision making based on machine learning (ML) techniques. *Annals of operations research*, *294*(1-2), 715–739. https://doi.org/10.1007/s10479-019-03188-0

scikit-learn. (2017). *Cross-validation: Evaluating estimator performance*. https://scikit-learn/stable/modules/cross_validation.html

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, *2*(28), 307–317. https://apps.dtic.mil/sti/citations/AD0604084

Shumway, T. (2001). Forecasting Bankruptcy More Accurately: A Simple Hazard Model. *The Journal of Business*, *74*(1), 101–124. https://doi.org/10.1086/209665

Somol, P., Baesens, B., Pudil, P. & Vanthienen, J. (2005). Filter- versus wrapper-based feature selection for credit scoring. *International Journal of Intelligent Systems*, *20*(10), 985–999. https://doi.org/10.1002/int.20103

Steenackers, A. & Goovaerts, M. J. (1989). A credit scoring model for personal loans. *Insurance, Mathematics & Economics*, *8*(1), 31–34. https://doi.org/10.1016/0167-6687(89)90044-9

Studenmund, A. H. & Johnson, B. K. (2017). *A Practical guide to using econometrics* (7. utg). Pearson.

Tabachnick, B. G. & Fidell, L. S. (2019). *Using multivariate statistics* (7. utg). Pearson.

Thomas, L. C. (2000). A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers. *International Journal of Forecasting*, *16*(2), 149–172. https://doi.org/10.1016/S0169-2070(00)00034-0

Tian, S. & Yu, Y. (2017). Financial ratios and bankruptcy predictions: An international evidence. *International Review of Economics & Finance*, *51*, 510–526. https://doi.org/10.1016/j.iref.2017.07.025

Tian, S., Yu, Y. & Guo, H. (2015). Variable selection and corporate bankruptcy forecasts. *Journal of Banking & Finance*, *52*, 89–100. https://doi.org/10.1016/j.jbankfin.2014.12.003

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288. https://doi.org/http://www.jstor.org/stable/2346178

Tripathi, D., Edla, D. R., Kuppili, V. & Dharavath, R. (2020). Binary BAT algorithm and RBFN based hybrid credit scoring model. *Multimedia Tools and Applications*, *79*(43-44), 31889–31912. https://doi.org/10.1007/s11042-020-09538-6

Tsai, M.-C., Lin, S.-P., Cheng, C.-C. & Lin, Y.-P. (2009). The consumer loan default predicting model – An application of DEA–DA and neural network. *Expert Systems with Applications*, *36*(9), 11682–11690. https://doi.org/10.1016/j.eswa.2009.03.009

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory* (1.utg.). Springer New York. https://doi.org/10.1007/978-1-4757-2440-0

Wang, M., Yu, J. & Ji, Z. (2018). Personal Credit Risk Assessment Based on Stacking Ensemble Model. I Z. Shi, E. Mercier-Laurent & J. Li (Red.). Springer, Cham. https://doi.org/10.1007/978-3-030-00828-4_33

Wang, Q., Hu, Y. & Li, J. (2018). Community-Based Feature Selection for Credit Card Default Prediction. I C. Cherifi, H. Cherifi, M. Karsai & M. Musolesi (Red.), *Complex Networks & Their Applications VI* (s. 153–165). Springer International Publishing. https://doi.org/10.1007/978-3-319-72150-7_13

West, D. (2000). Neural network credit scoring models. *Computers & Operations Research*, *27*(11), 1131–1152. https://doi.org/10.1016/S0305-0548(99)00149-5

Wiginton, J. C. (1980). A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior. *The Journal of Financial and Quantitative Analysis*, *15*(3), 757–770. https://doi.org/10.2307/2330408

Xia, Y., Liu, C., Li, Y. & Liu, N. (2017). A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications*, *78*, 225–241. https://doi.org/10.1016/j.eswa.2017.02.017

Xiao, W., Zhao, Q. & Fei, Q. (2006). A comparative study of data mining methods in consumer loans credit scoring management. *Journal of Systems Science and Systems Engineering*, *15*(4), 419–435. https://doi.org/10.1007/s11518-006-5023-5

Xu, D., Zhang, X. & Feng, H. (2019). Generalized fuzzy soft sets theory-based novel hybrid ensemble credit scoring model. *International Journal of Finance & Economics*, *24*(2), 903–921. https://doi.org/10.1002/ijfe.1698

Yap, B. W., Ong, S. H. & Husain, N. H. M. (2011). Using data mining to improve assessment of credit worthiness via credit scoring models. *Expert Systems with Applications*, *38*(10), 13274–13283. https://doi.org/10.1016/j.eswa.2011.04.147

Zhou, L., Lai, K. K. & Yu, L. (2010). Least squares support vector machines ensemble models for credit scoring. *Expert Systems with Applications*, *37*(1), 127–133. https://doi.org/10.1016/j.eswa.2009.05.024

Aas, K., Jullum, M. & Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial intelligence*, *298*(17), 103502–. https://doi.org/10.1016/j.artint.2021.103502

A Appendiks

A.1 Variabler i endelig datasett

Tabell 3: Beskrivelse av variabler

Variabelnavn	**Beskrivelse**
Age_log	Kontinuerlig; logartimen av lånsøker sin alder
App_Night	Dummy; 1 dersom søknad mottatt mellom kl. 22.00 og 07.00
App_Weekend	Dummy; 1 dersom søknad mottatt lørdag eller søndag
AppliedByGranted	Kontinuerlig; søkt kredittkortgrense/innvilget kredittkortgrense
Channel_OnlineBank	Dummy; 1 dersom lånsøker har søkt via nettbank
Channel_Operator	Dummy; 1 dersom lånsøker har søkt via rådgiver i bank
Channel_PhoneBank	Dummy; 1 dersom lånsøker har søkt via mobilbank
Channel_ResponsePage	Dummy; 1 dersom lånsøker har søkt via en kampanje
Channel_Web	Dummy; 1 dersom lånsøker har søkt via en annen digital løsning
Debit_Card_IND	Dummy; 1 dersom lånsøker har et aktivt debetkort hos dataleverandør
Debt_Ratio_AMT	Kontinuerlig; gjeldsgrad
DebtRegisterCreditFacilityNum	Kontinuerlig; antall kredittkort i gjeldsregisteret
DebtRegisterCreditLimit	Kontinuerlig: sum av beløp på åpne rammer i gjeldsregisteret
DebtRegisterIELA	Kontinuerlig; sum av rentebærende beløp for alle innslag i gjeldsegisteret
DebtRegisterNonIELA	Kontinuerlig; sum av ikke-rentebærende beløp for alle innslag i gjeldsregisteret
DebtRegisterNum	Kontinuerlig: antall innslag i gjeldsregisteret
DebtRegisterRepaymentLoanBalance	Kontinuerlig; sum av balanse på nedbetalingslån
DebtRegisterRepaymentLoanNum	Kontinuerlig; antall nedbetalingslån i gjeldsregisteret
Diff_Hour	Kontinuerlig; antall timer fra søknad ble mottatt til innvilget
Distributor_1	Dummy; 1 dersom lånsøker har søkt via distribrutør 1
Distributor_3	Dummy; 1 dersom lånsøker har søkt via distribrutør 3
Distributor_4A	Dummy; 1 dersom lånsøker har søkt via distribrutør 4A
Distributor_4B	Dummy; 1 dersom lånsøker har søkt via distribrutør 4B
Distributor_4C	Dummy; 1 dersom lånsøker har søkt via distribrutør 4C
Distributor_4D	Dummy; 1 dersom lånsøker har søkt via distribrutør 4D
Distributor_4E	Dummy; 1 dersom lånsøker har søkt via distribrutør 4E
Distributor_4F	Dummy; 1 dersom lånsøker har søkt via distribrutør 4F
Distributor_4G	Dummy; 1 dersom lånsøker har søkt via distribrutør 4G
Distributor_4H	Dummy; 1 dersom lånsøker har søkt via distribrutør 4H
Distributor_4I	Dummy; 1 dersom lånsøker har søkt via distribrutør 4I
Distributor_4J	Dummy; 1 dersom lånsøker har søkt via distribrutør 4J
Distributor_4K	Dummy; 1 dersom lånsøker har søkt via distribrutør 4K
Distributor_4L	Dummy; 1 dersom lånsøker har søkt via distribrutør 4L
Distributor_4M	Dummy; 1 dersom lånsøker har søkt via distribrutør 4M

Tabell 3: Beskrivelse av variabler

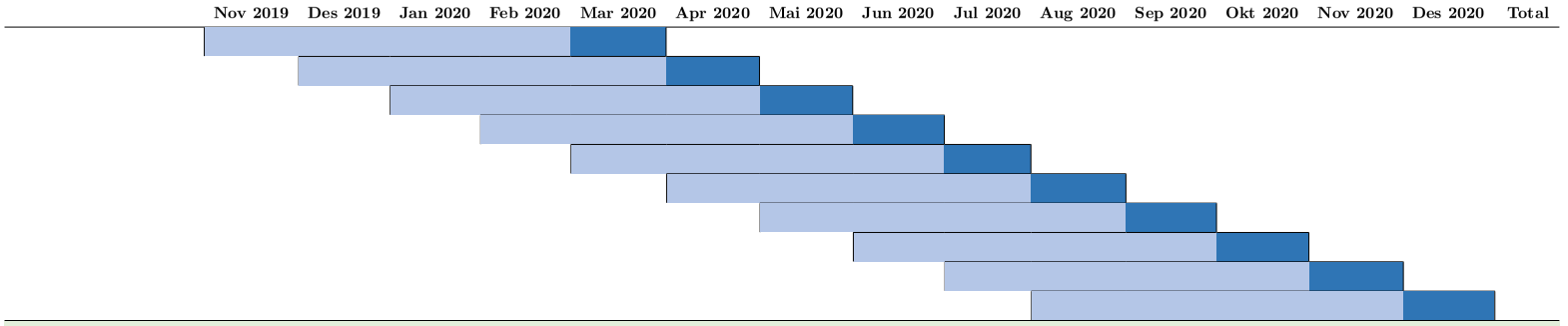
Variabelnavn	**Beskrivelse**
Distributor_4N	Dummy; 1 dersom lånsøker har søkt via distribrutør 4N
Distributor_4O	Dummy; 1 dersom lånsøker har søkt via distribrutør 4O
Employment_Duration_1To3	Dummy; 1 dersom lånsøker sin lengde på nåværende arbeidsforhold er mellom 1 til 3 år
Employment_Duration_Less1	Dummy; 1 dersom lånsøker sin lengde på nåværende arbeidsforhold er mindre enn 1 år
Employment_Duration_More3	Dummy; 1 dersom lånsøker sin lengde på nåværende arbeidsforhold er mer enn 3 år
Employment_Duration_NoJoB	Dummy; 1 dersom lånsøker ikke har jobb
Employment_DisabilityPensioner	Dummy; 1 dersom lånsøker er uførepensjonist
Employment_Employee	Dummy; 1 dersom lånsøker er heltidsansatt
Employment_Home	Dummy; 1 dersom lånsøker er hjemmeværende
Employment_Other	Dummy; 1 dersom lånsøker har valgt annet som ansettelsestype
Employment_Retiree	Dummy; 1 dersom lånsøker er pensjonist
Employment_SelfEmployed	Dummy; 1 dersom lånsøker er selvstendig næringsdrivende
Employment_SocialSecurity	Dummy; 1 dersom lånsøker mottar trygd
Employment_Student	Dummy; 1 dersom lånsøker er student
Employment_TempEmployee	Dummy; 1 dersom lånsøker er deltidsansatt
Employment_Unemployed	Dummy; 1 dersom lånsøker er arbeidsledig
FLI_AMT	Kontinuerlig; forenklet likviditetsindikator
Habitation_Apartment	Dummy; 1 dersom lånsøker bor i egen leilighet
Habitation_Homeowner	Dummy; 1 dersom lånsøker bor i eget hus
Habitation_Other	Dummy; 1 dersom lånsøker har valgt annet som bosituasjon
Habitation_Parents	Dummy; 1 dersom lånsøker bor hos foreldrene
Habitation_Renter	Dummy; 1 dersom lånsøker leier
Homeowner_HouseAndCoop	Dummy; 1 dersom lånsøker eier bolig og er boligeier i borettslag, med boliglån hos dataleverandør
Homeowner_HouseOrCoop	Dummy; 1 dersom lånsøker eier hus eller er boligeier i borettslag, med boliglån hos dataleverandør
Homeowner_NoHouseOrCoop	Dummy; 1 dersom lånsøker ikke eier hus eller er boligeier i borettslag
HourCompleted	Kontinuerlig; tidspunkt for når søknad ble signert
ID_Bank	Dummy; 1 dersom lånsøker har identifisert seg via bankID
ID_Physical	Dummy; 1 dersom lånsøker har identifisert seg fysisk i banken
ID_Pum	Dummy; 1 dersom lånsøker har identifisert seg via PUM
Logins_Num	Kontinuerlig; antall innlogginger i nettbank siste avsluttede måned
Manual_Handling_IND	Dummy; 1 dersom søknad er manuelt behandlet
MC_1	Dummy; 1 dersom lånsøker har søkt om kredittkorttype 1
MC_2	Dummy; 1 dersom lånsøker har søkt om kredittkorttype 2
MC_2_Young	Dummy; 1 dersom lånsøker har søkt om kredittkorttype 2 Ung
MC_3	Dummy; 1 dersom lånsøker har søkt om kredittkorttype 3
MC_3_Extra	Dummy; 1 dersom lånsøker har søkt om kredittkorttype 3 Ekstra

Tabell 3: Beskrivelse av variabler

Variabelnavn	Beskrivelse
MC_3_Plus	Dummy; 1 dersom lånsøker har søkt om kredittkorttype 3 Pluss
MC_3_Young	Dummy; 1 dersom lånsøker har søkt om kredittkorttype 3 Ung
MC_4	Dummy; 1 dersom lånsøker har søkt om kredittkorttype 4
MC_4_Extra	Dummy; 1 dersom lånsøker har søkt om kredittkorttype 4 Ekstra
MC_4_Unique	Dummy; 1 dersom lånsøker har søkt om kredittkorttype 4 Unik
MC_4_Young	Dummy; 1 dersom lånsøker har søkt om kredittkorttype 4 Ung
Method_Mortgage	Dummy; 1 dersom lånsøker er vurdert i segmentet «boliglån»
Method_Ordinary	Dummy; 1 dersom lånsøker er vurdert i segmentet «ordinær behandling»
Method_Predefined	Dummy; 1 dersom lånsøker vurdert i segmentet «forhåndsvurdert»
Method_Unique	Dummy; 1 dersom lånsøker er vurdert i segmentet «unik»
Method_Young	Dummy; 1 dersom lånsøker er vurdert i segmentet «ung»
Mortgage_IND	Dummy; 1 dersom lånsøker har boliglån hos dataleverandør
Mortgages_AMT	Kontinuerlig; beløp på boliglån
Net_Income_AMT	Kontinuerlig; årlig nettoinntekt
Net_Monthly_Income_AMT	Kontinuerlig; månedlig nettoinntekt
NoOfAdressChanges_CNT	Kontinuerlig; antall adresseendringer
NoOfChildren	Kontinuerlig; antall barn
RegisteredDebtByIncome	Kontinuerlig; DebtRegisterIELA/Net_Income_AMT
SFLI_AMT	Kontinuerlig; FLI dersom boliglånsrenten øker med 5%
Status_Cohabiting	Dummy; 1 dersom lånsøker sin sivilstatus er samboer
Status_Divorced	Dummy; 1 dersom lånsøker sin sivilstatus er skilt
Status_Married	Dummy; 1 dersom lånsøker sin sivilstatus er gift
Status_Single	Dummy; 1 dersom lånsøker sin sivilstatus er singel
Status_Widowed	Dummy; 1 dersom lånsøker sin sivilstatus er enke/enkemann
Student_Loan_AMT	Kontinuerlig; beløp på studielån
SumAvailable	Kontinuerlig; beløp lånsøker har igjen på slutten av måneden etter at utgifter er betalt
Total_Incoming_AMT	Kontinuerlig; sum av innkommende beløp på konto de siste 6 månedene
Total_Tax_AMT	Kontinuerlig; totalt betalt i skatt
Transactions_Num	Kontinuerlig; antall transaksjoner siste avsluttede måned
Vehicle_Loan_AMT	Kontinuerlig; beløp på billån
Wealth_AMT	Kontinuerlig; formue
Year_Completed	Kontinuerlig; året da søknaden ble signert
Year_Customer	Kontinuerlig; differansen mellom søknadstidspunkt og første kundeforhold, oppgitt i antall år
Van	Dummy; 1 dersom lånsøker har, eller har hatt, en eller flere varebiler registrert i bilregisteret
Car	Dummy; 1 dersom lånsøker har, eller har hatt, en eller flere biler registrert i bilregisteret
MC	Dummy; 1 dersom lånsøker har, eller har hatt, en eller flere motorsykler registrert i bilregisteret

A.2 Trening og test

Figur 6: Prosessen for å trene og evaluere modeller

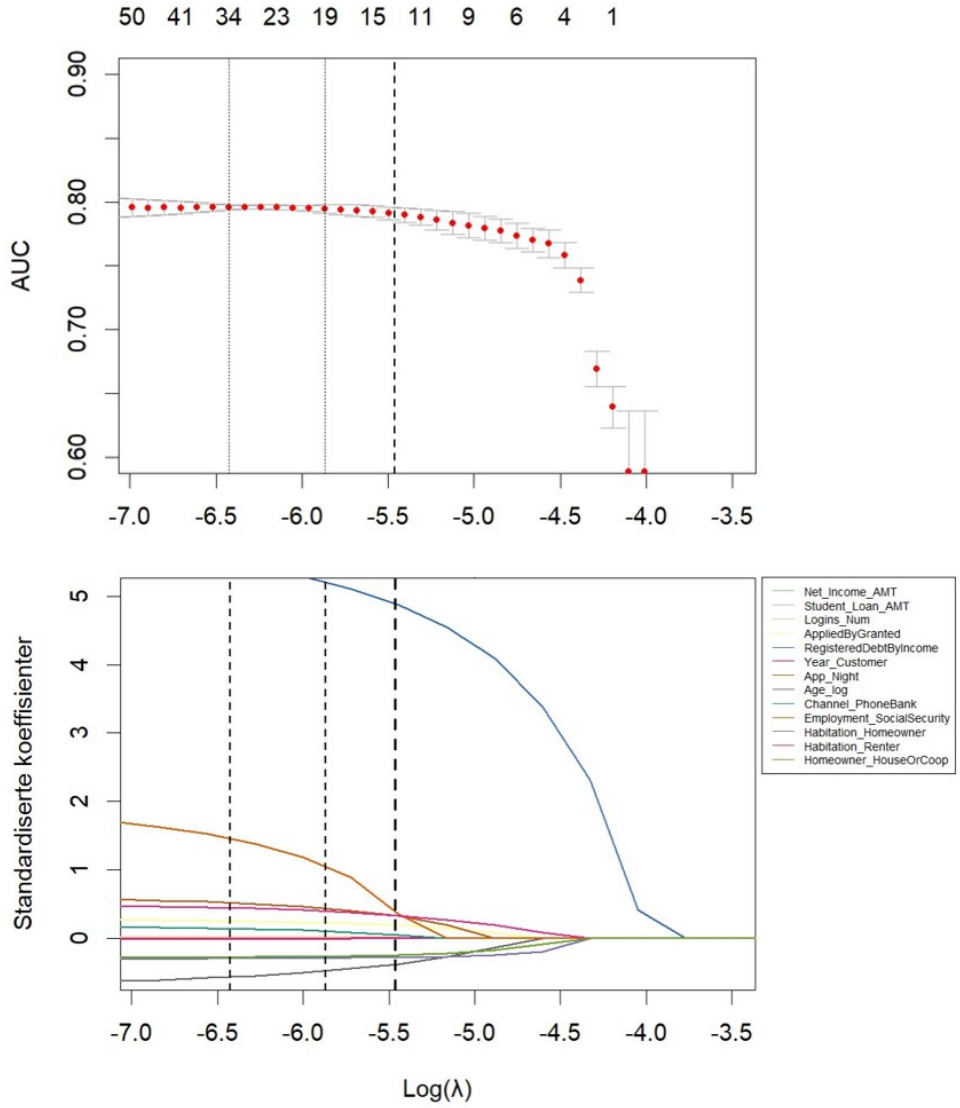


	Nov 2019	Des 2019	Jan 2020	Feb 2020	Mar 2020	Apr 2020	Mai 2020	Jun 2020	Jul 2020	Aug 2020	Sep 2020	Okt 2020	Nov 2020	Des 2020	Total
Obs (Misligholdt)	76	65	75	91	90	69	66	83	107	89	107	96	86	79	1 179
Obs (Ikke misligholdt)	3 713	2 681	3 958	3 910	2 996	2 323	3 089	3 745	2 992	2 836	4 543	3 920	3 494	2 246	46 446
Andel misligholdt	2.01%	2.37%	1.86%	2.27%	2.92%	2.88%	2.09%	2.17%	3.45%	3.04%	2.30%	2.39%	2.40%	3.40%	2.48%

Vi benytter rolling window og forward validation til å trene og evaluere modeller for prediksjon av mislighold av kredittkortgjeld. Vi inkluderer 4 måneder i treningssettet (lys blå) som benyttes til å trene og evaluere prediksjonsevnen til hver modell. For å evaluere modellenes prediksjonsevne på testsettet (mørk blå) benyttes måneden som følger etter siste måned i treningssettet. Av tabellen fremgår også antall observasjoner som har misligholdt eller ikke misligholdt sin kredittkortgjeld, i tillegg til andelen misligholdt hver måned.

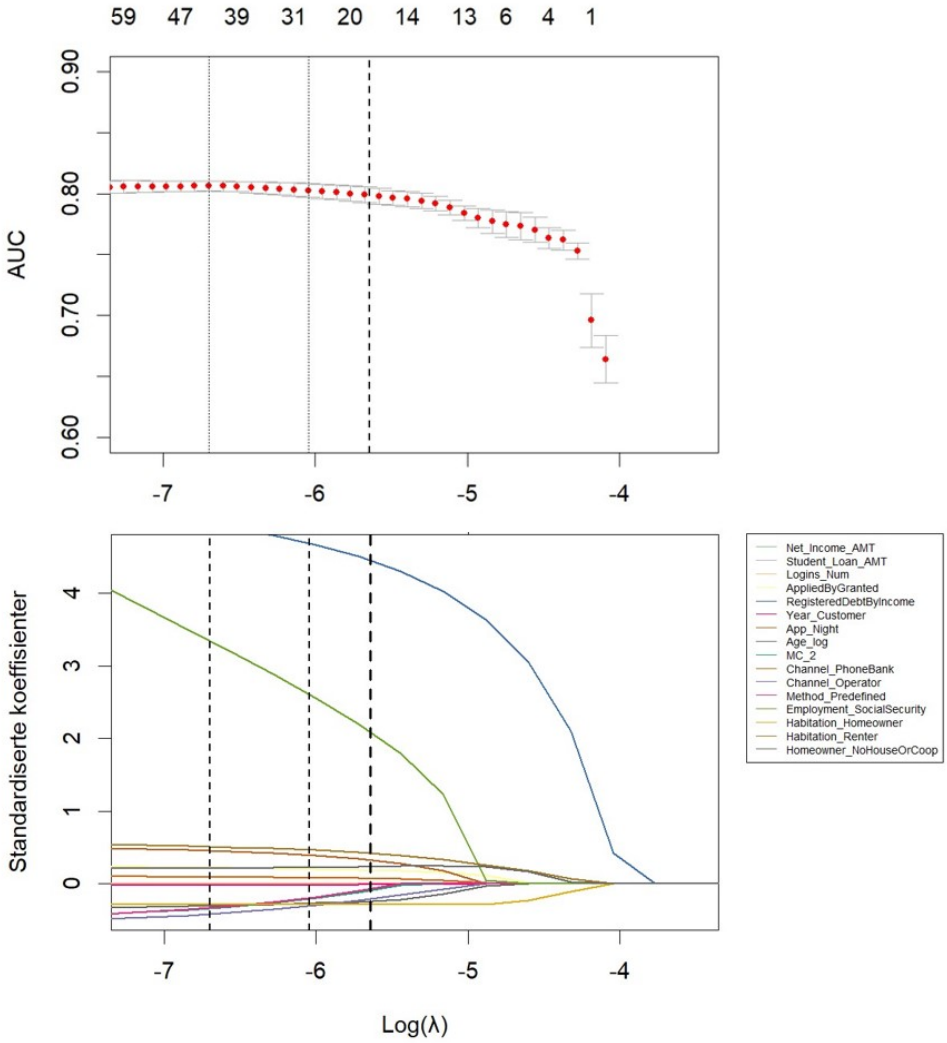
A.3 AUC og LASSO

Figur 7: AUC-plott og LASSO-stiplott for periode 1



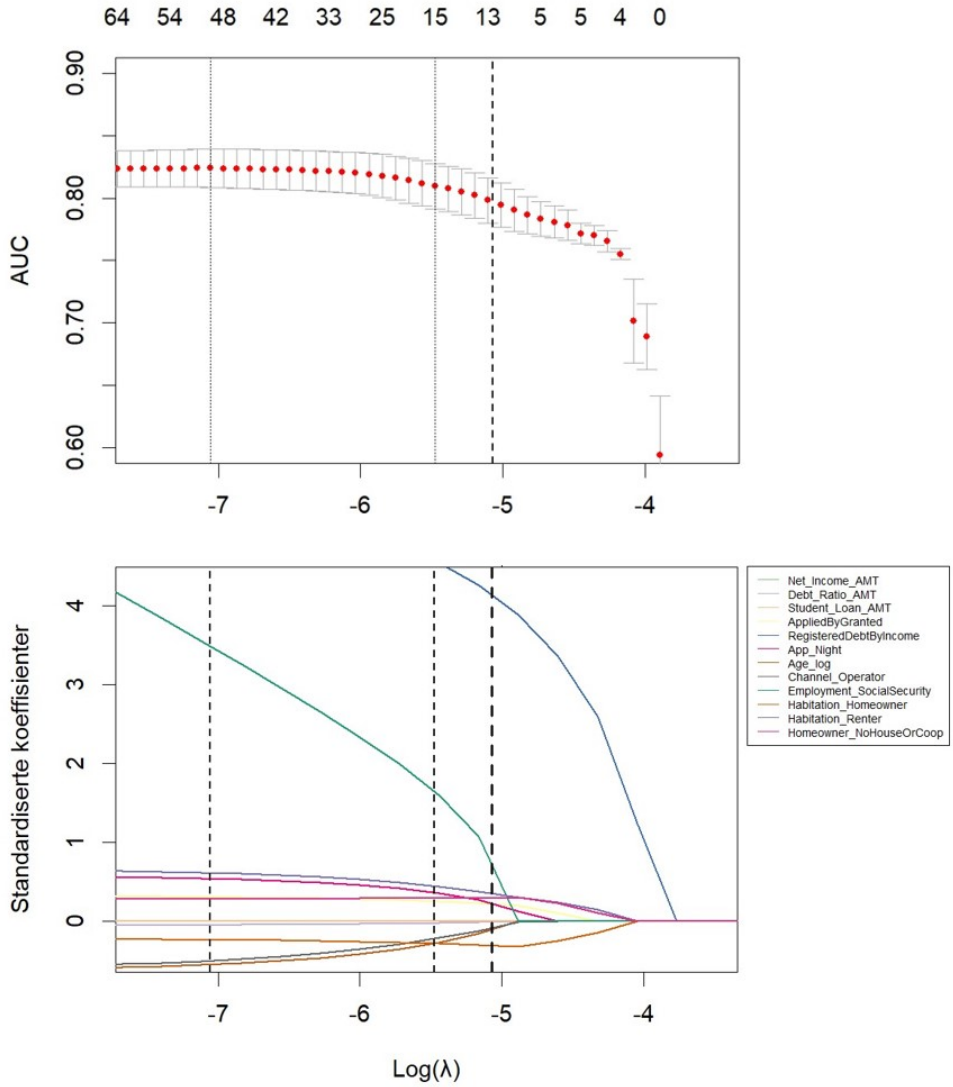
Det øverste plottet viser AUC mot log(λ), gitt ved kryssvalidering. De stiplede linjene representerer de 3 ulike λ-verdiene som ble vurdert. Linjen til høyre representerer λ-verdien som ble brukt til modellestimering, altså den største verdien av λ hvor den gjennomsnittlige AUC-verdien er innenfor 1,5 standardfeil fra AUC-verdien til linjen lengst til venstre i plottet. Nederst er LASSO-stiplottet, lagd ved bruk av den gitte λ-verdien. For å gjøre plottet lesbart velger vi kun å inkludere forklaringsvariablene som *ikke* blir satt til 0 av LASSO for den gitte λ-verdien.

Figur 8: AUC-plott og LASSO-stiplott for periode 2



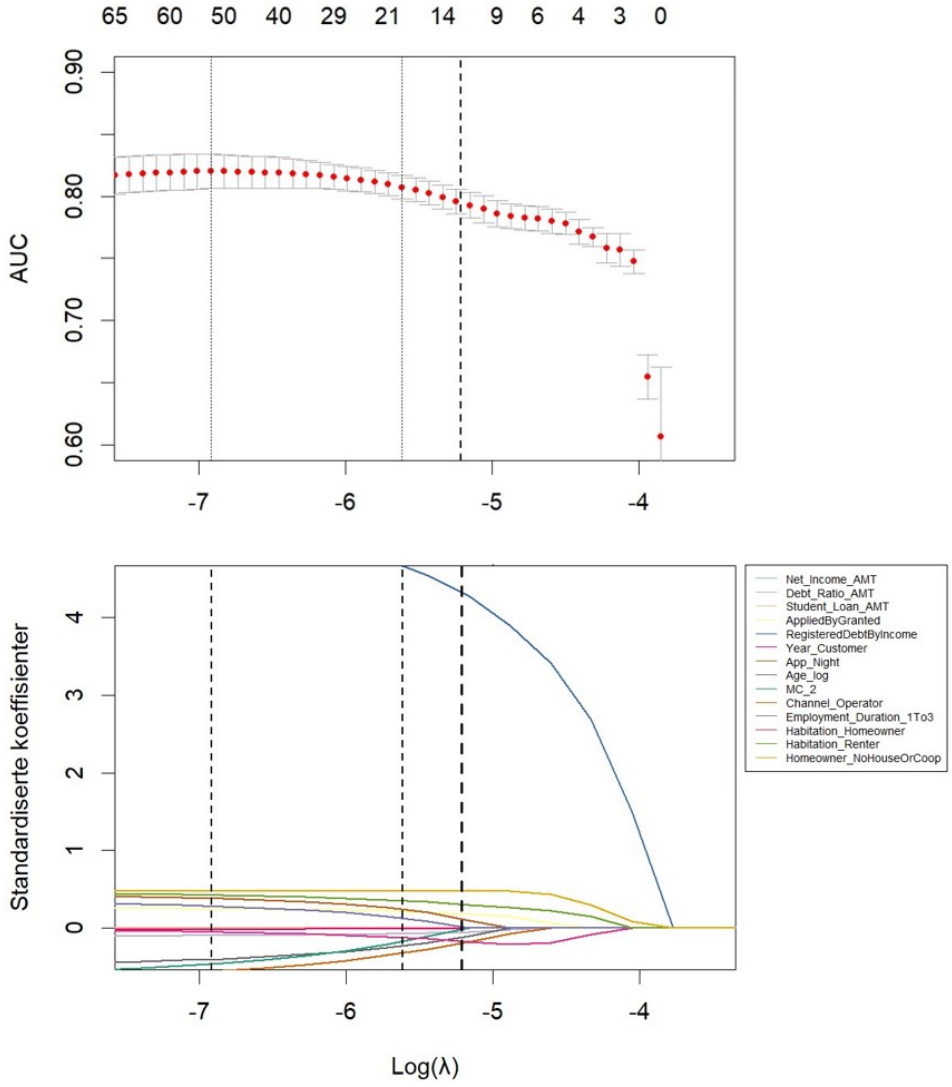
Det øverste plottet viser AUC mot log(λ), gitt ved kryssvalidering. De stiplede linjene representerer de 3 ulike λ-verdiene som ble vurdert. Linjen til høyre representerer λ-verdien som ble brukt til modellestimering, altså den største verdien av λ hvor den gjennomsnittlige AUC-verdien er innenfor 1,5 standardfeil fra AUC-verdien til linjen lengst til venstre i plottet. Nederst er LASSO-stiplottet, lagd ved bruk av den gitte λ-verdien. For å gjøre plottet lesbart velger vi kun å inkludere forklaringsvariablene som *ikke* blir satt til 0 av LASSO for den gitte λ-verdien.

Figur 9: AUC-plott og LASSO-stiplott for periode 3



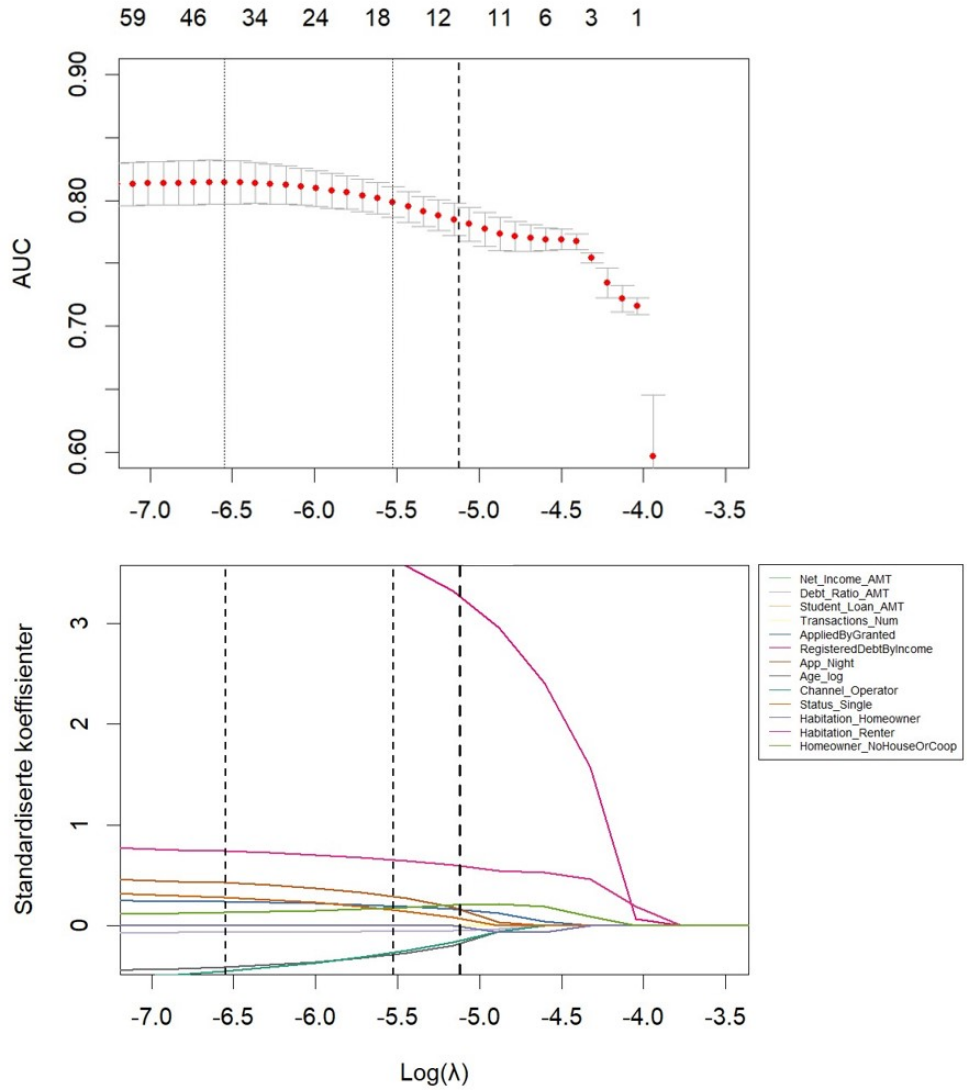
Det øverste plottet viser AUC mot log(λ), gitt ved kryssvalidering. De stiplede linjene representerer de 3 ulike λ-verdiene som ble vurdert. Linjen til høyre representerer λ-verdien som ble brukt til modellestimering, altså den største verdien av λ hvor den gjennomsnittlige AUC-verdien er innenfor 1,5 standardfeil fra AUC-verdien til linjen lengst til venstre i plottet. Nederst er LASSO-stiplottet, lagd ved bruk av den gitte λ-verdien. For å gjøre plottet lesbart velger vi kun å inkludere forklaringsvariablene som *ikke* blir satt til 0 av LASSO for den gitte λ-verdien.

Figur 10: AUC-plott og LASSO-stiplott for periode 4



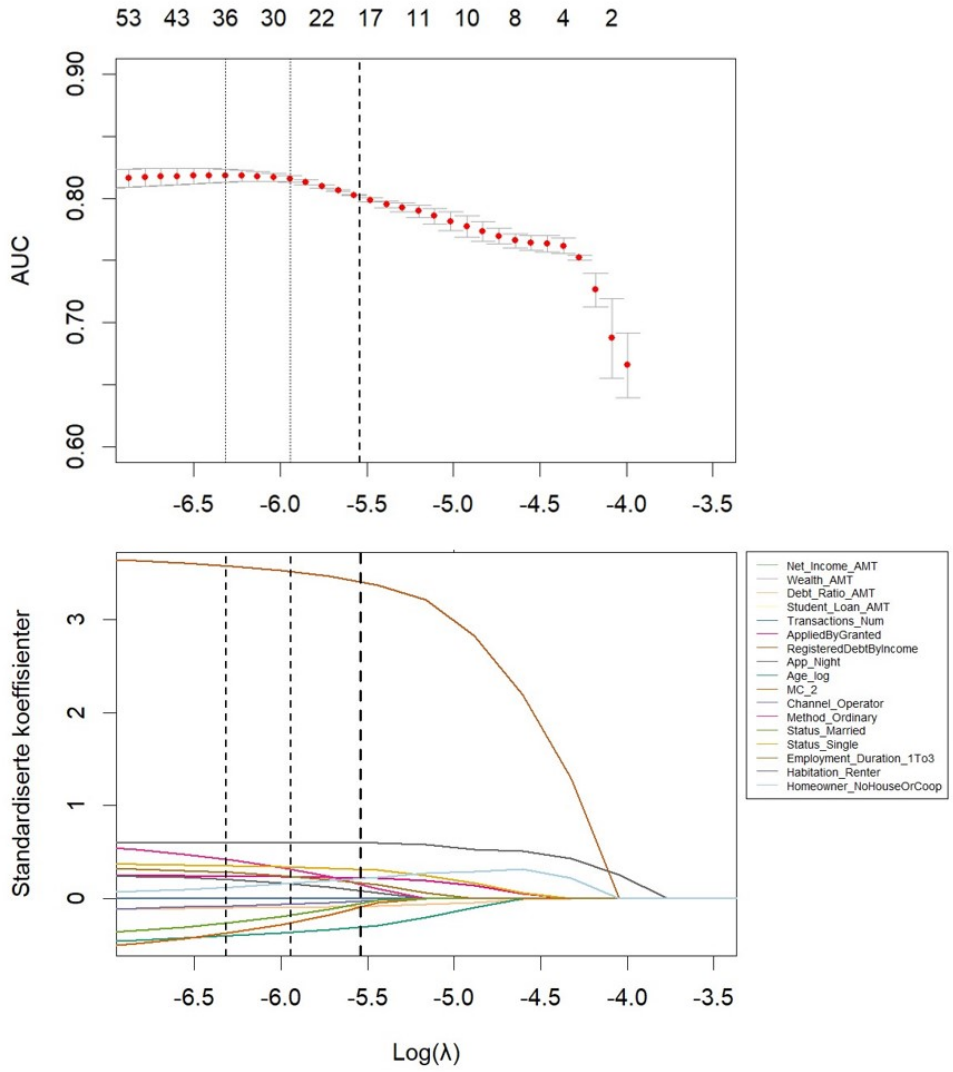
Det øverste plottet viser AUC mot log(λ), gitt ved kryssvalidering. De stiplede linjene representerer de 3 ulike λ-verdiene som ble vurdert. Linjen til høyre representerer λ-verdien som ble brukt til modellestimering, altså den største verdien av λ hvor den gjennomsnittlige AUC-verdien er innenfor 1,5 standardfeil fra AUC-verdien til linjen lengst til venstre i plottet. Nederst er LASSO-stiplottet, lagd ved bruk av den gitte λ-verdien. For å gjøre plottet lesbart velger vi kun å inkludere forklaringsvariablene som *ikke* blir satt til 0 av LASSO for den gitte λ-verdien.

Figur 11: AUC-plott og LASSO-stiplott for periode 5



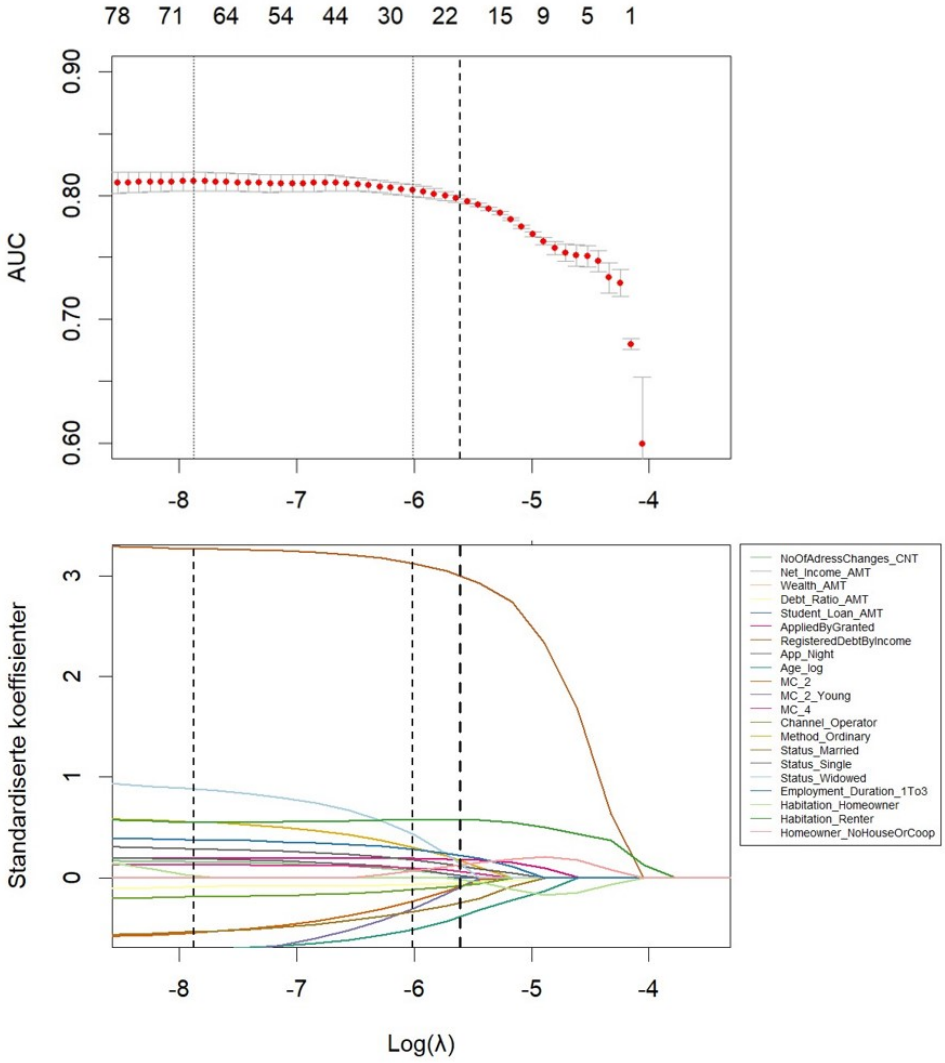
Det øverste plottet viser AUC mot log(λ), gitt ved kryssvalidering. De stiplede linjene representerer de 3 ulike λ-verdiene som ble vurdert. Linjen til høyre representerer λ-verdien som ble brukt til modellestimering, altså den største verdien av λ hvor den gjennomsnittlige AUC-verdien er innenfor 1,5 standardfeil fra AUC-verdien til linjen lengst til venstre i plottet. Nederst er LASSO-stiplottet, lagd ved bruk av den gitte λ-verdien. For å gjøre plottet lesbart velger vi kun å inkludere forklaringsvariablene som *ikke* blir satt til 0 av LASSO for den gitte λ-verdien.

Figur 12: AUC-plott og LASSO-stiplott for periode 6



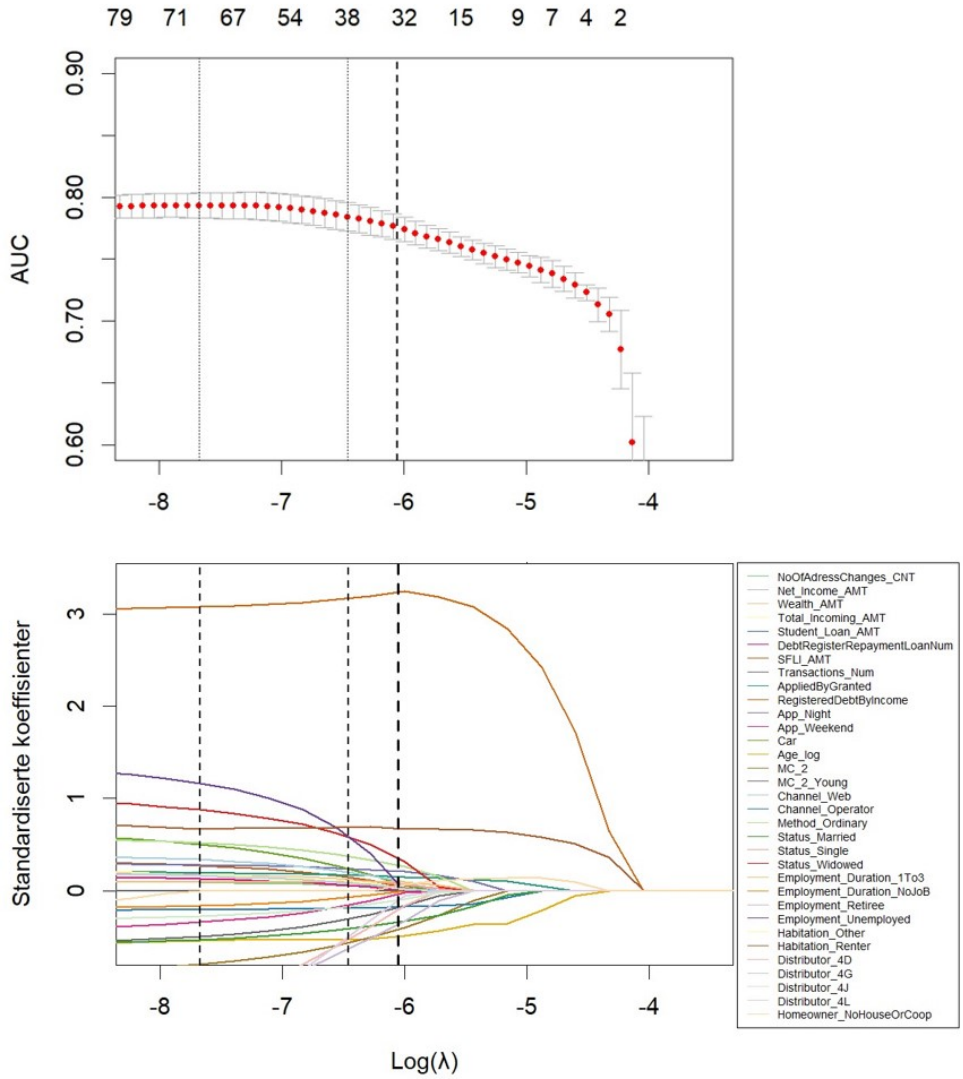
Det øverste plottet viser AUC mot log(λ), gitt ved kryssvalidering. De stiplede linjene representerer de 3 ulike λ-verdiene som ble vurdert. Linjen til høyre representerer λ-verdien som ble brukt til modellestimering, altså den største verdien av λ hvor den gjennomsnittlige AUC-verdien er innenfor 1,5 standardfeil fra AUC-verdien til linjen lengst til venstre i plottet. Nederst er LASSO-stiplottet, lagd ved bruk av den gitte λ-verdien. For å gjøre plottet lesbart velger vi kun å inkludere forklaringsvariablene som *ikke* blir satt til 0 av LASSO for den gitte λ-verdien.

Figur 13: AUC-plott og LASSO-stiplott for periode 7



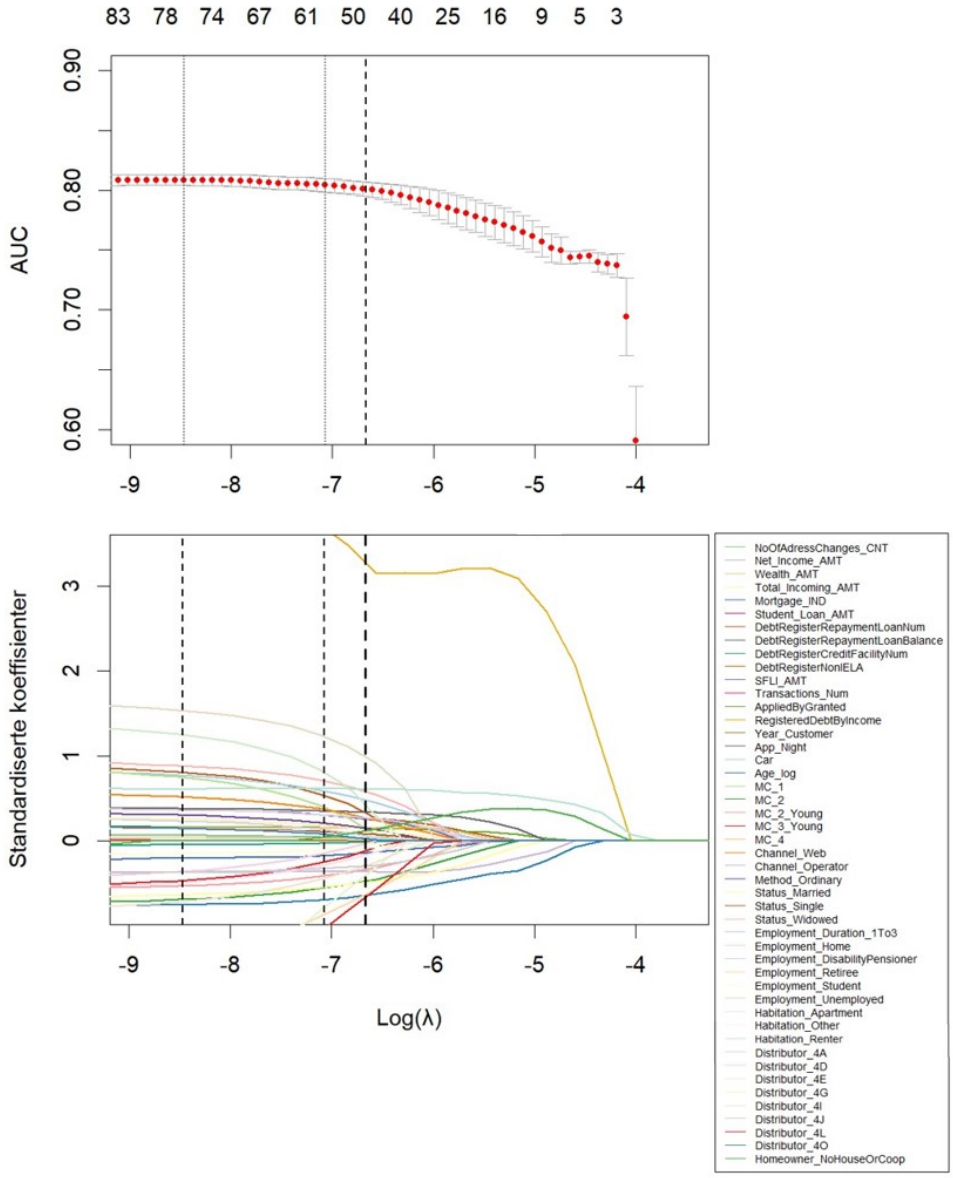
Det øverste plottet viser AUC mot log(λ), gitt ved kryssvalidering. De stiplede linjene representerer de 3 ulike λ-verdiene som ble vurdert. Linjen til høyre representerer λ-verdien som ble brukt til modellestimering, altså den største verdien av λ hvor den gjennomsnittlige AUC-verdien er innenfor 1,5 standardfeil fra AUC-verdien til linjen lengst til venstre i plottet. Nederst er LASSO-stiplottet, lagd ved bruk av den gitte λ-verdien. For å gjøre plottet lesbart velger vi kun å inkludere forklaringsvariablene som *ikke* blir satt til 0 av LASSO for den gitte λ-verdien.

Figur 14: AUC-plott og LASSO-stiplott for periode 8



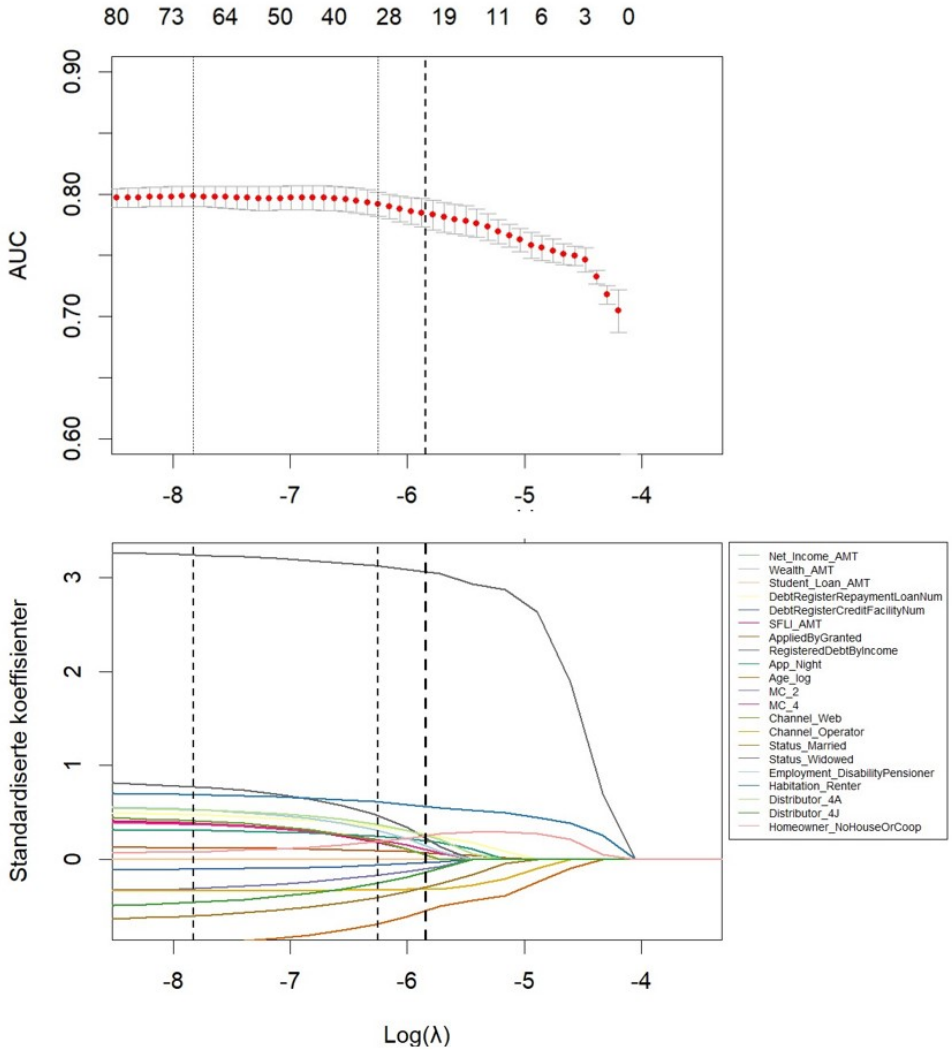
Det øverste plottet viser AUC mot log(λ), gitt ved kryssvalidering. De stiplede linjene representerer de 3 ulike λ-verdiene som ble vurdert. Linjen til høyre representerer λ-verdien som ble brukt til modellestimering, altså den største verdien av λ hvor den gjennomsnittlige AUC-verdien er innenfor 1,5 standardfeil fra AUC-verdien til linjen lengst til venstre i plottet. Nederst er LASSO-stiplottet, lagd ved bruk av den gitte λ-verdien. For å gjøre plottet lesbart velger vi kun å inkludere forklaringsvariablene som *ikke* blir satt til 0 av LASSO for den gitte λ-verdien.

Figur 15: AUC-plott og LASSO-stiplott for periode 9



Det øverste plottet viser AUC mot log(λ), gitt ved kryssvalidering. De stiplede linjene representerer de 3 ulike λ-verdiene som ble vurdert. Linjen til høyre representerer λ-verdien som ble brukt til modellestimering, altså den største verdien av λ hvor den gjennomsnittlige AUC-verdien er innenfor 1,5 standardfeil fra AUC-verdien til linjen lengst til venstre i plottet. Nederst er LASSO-stiplottet, lagd ved bruk av den gitte λ-verdien. For å gjøre plottet lesbart velger vi kun å inkludere forklaringsvariablene som *ikke* blir satt til 0 av LASSO for den gitte λ-verdien.

Figur 16: AUC-plott og LASSO-stiplott for periode 10



Det øverste plottet viser AUC mot log(λ), gitt ved kryssvalidering. De stiplede linjene representerer de 3 ulike λ-verdiene som ble vurdert. Linjen til høyre representerer λ-verdien som ble brukt til modellestimering, altså den største verdien av λ hvor den gjennomsnittlige AUC-verdien er innenfor 1,5 standardfeil fra AUC-verdien til linjen lengst til venstre i plottet. Nederst er LASSO-stiplottet, lagd ved bruk av den gitte λ-verdien. For å gjøre plottet lesbart velger vi kun å inkludere forklaringsvariablene som *ikke* blir satt til 0 av LASSO for den gitte λ-verdien.

A.4 Logistisk regresjon

Tabell 4: Estimerte LR-koeffisienter for variabler valgt av LASSO

Panel A-1: For månedene mars-juli

	Mars	April	Mai	Juni	Juli
Net_Income_AMT	-5.83e^{-7}(−1.32)	-1.10^{-6}(−2.43)	-1.28e^{-6}(−2.85)	-1.20e^{-6}(−2.69)	-1.91e^{-6}(−4.255)
Student_Loan_AMT	-4.71e^{-6}(−5.31)	-5.40e^{-6}(−5.95)	-5.61e^{-6}(−6.02)	-3.98e^{-6}(−4.65)	-5.64e*-6(-5.84)
AppliedByGranted	0.287(3.59)	0.25(3.12)	0.32(4.07)	0.27(3.27)	0.26(3.24)
RegisteredDebtByIncome	5.92(13.06)	5.23(11.71)	5.47(12.74)	5.46(12.73)	4.41(9.81)
App_Night	0.62(3.51)	0.52(2.96)	0.58(3.29)	0.43(2.26)	0.49(2.68)
Age_log	-0.70(-3.26)	-0.34(-1.57)	-0.62(-3.00)	-0.47(-2.23)	-0.51(-2.39)
Habitation_Renter	0.51(3.43)	0.57(3.83)	0.66(4.51)	0.46(2.98)	0.87(5.43)
Channel_Operator		-0.56(-3.13)	-0.60(-3.61)	-0.66(-3.99)	-0.60(-3.62)
Homeowner_NoHouseOrCoop		0.21(0.90)	0.28(1.20)	0.49(2.11)	0.20(0.89)
MC_2		-0.49(-2.92)		-0.59(-3.45)	
Habitation_Homeowner	-0.32(-1.42)	-0.29(-1.21)	-0.23(-0.94)	-0.01(-0.05)	0.31(1.22)
Wealth_AMT					
Debt_Ratio_AMT			-0.06(-0.96)	-0.11(-1.69)	-0.10(-1.58)
Status_Married					
Status_Single					0.39(2.79)
Employment_Duration_1To3				0.34(2.67)	
Transactions_Num					0.003(2.37)
Year_Customer	-0.02(-3.67)	-0.02(-3.13)		-0.02(-2.84)	
Method_Ordinary					
Status_Widowed					
NoOfAdressChanges_CNT					
DebtRegisterRepaymentLoanNum					
SFLI_AMT					
MC_2_Young					
MC_4					
Channel_Web					
Employment_SocialSecurity	1.95(1.28)	0.14(0.04)	13.80(0.04)		
Distributor_4J					
Total_Incoming_AMT					
DebtRegisterCreditFacilityNum					
Logins_Num	0.01(2.40)	0.01(2.37)			
Car					
Channel_PhoneBank	0.19(1.49)	0.10(0.76)			
Employment_DisabilityPensioner					
Employment_Retiree					
Employment_Unemployed					
Habitation_Other					
Distributor_4A					
Distributor_4D					
Distributor_4G					
Distributor_4L					

Tabell 4: Estimerte LR-koeffisienter for variabler valgt av LASSO

Panel A-2: For månedene mars-juli

	Mars	April	Mai	Juni	Juli
DebtRegisterRepaymentLoanBalance					
DebtRegisterNonIELA					
App_Weekend					
MC_1					
MC_3_Young					
Method_Predefined		-0.50(-2.62)			
Employment_Duration_NoJoB					
Employment_Home					
Employment_Student					
Habitation_Apartment					
Distributor_4E					
Distributor_4I					
Distributor_4O					
Homeowner_HouseOrCoop	-0.30(-1.35)				
Intercept	-1.77(-2.50)	-2.68(-3.44)	-2.00(-2.71)	-2.22(-2.98)	-2.48(-3.05)

Panel A-1 og A-2 viser estimerte regresjonskoeffisienter for variabler valgt av LASSO i månedene mars-juli og tilhørende z-verdier i parentes. Dette tilsvarer periode 1-5 i figur 6.

Tabell 4: Estimerte LR-koeffisienter for variabler valgt av LASSO

Panel B-1: For månedene august-desember

	August	September	Oktober	November	Desember
Net_Income_AMT	-1.65e^{-06}(−3.70)	-1.67e^{-06}(−3.81)	-1.65e^{-06}(−3.91)	-5.77e^{-07}(−1.36)	-1.18e^{-06}(−2.94)
Student_Loan_AMT	-5.49e^{-06}(−5.85)	-5.82e^{-06}(−6.43)	-5.60e^{-06}(−6.97)	-4.68e^{-06}(−6.10)	-5.24e^{-06}(−6.64)
AppliedByGranted	0.25(3.03)	0.19(2.41)	0.21(2.653)	0.16(2.07)	0.13(1.59)
RegisteredDebtByIncome	3.70(8.09)	3.31(7.08)	3.03(4.41)	4.50(5.77)	3.28(4.43)
App_Night	0.29(1.56)	0.20(1.09)	0.30(1.72)	0.39(2.30)	0.32(1.81)
Age_log	-0.53(-2.26)	-0.81(-3.47)	-0.61(-2.57)	-0.80(-3.04)	-0.95(-4.29)
Habitation_Renter	0.60(4.10)	0.61(4.05)	0.77(5.28)	0.64(4.37)	0.71(5.35)
Channel_Operator	-0.14(-0.87)	-0.21(-1.37)	-0.22(-1.50)	-0.38(-2.52)	-0.34(-2.45)
Homeowner_NoHouseOrCoop	0.01(0.05)	-0.01(-0.05)	-0.24(-1.33)	-0.12(-0.58)	0.06(0.35)
MC_2	-0.67(-3.61)	-0.62(-2.83)	-0.91(-4.69)	-0.75(-3.23)	-0.35(-1.67)
Habitation_Homeowner		0.28(1.17)			
Wealth_AMT	-2.00e^{-06}(−3.79)	-3.21e^{-06}(−4.67)	-1.33e^{-06}(−3.79)	-1.22e^{-06}(−3.64)	-1.48e^{-06}(−4.17)
Debt_Ratio_AMT	-0.13(-2.07)	-0.13(-2.10)			
Status_Married	-0.49(-2.50)	-0.60(-3.05)	-0.58(-3.27)	-0.67(-3.69)	-0.66(-3.88)
Status_Single	0.39(2.73)	0.33(2.28)	0.22(1.64)	0.15(1.14)	
Employment_Duration_1To3	0.37(2.92)	0.40(3.25)	0.10(0.79)	0.04(0.28)	
Transactions_Num	2.00e^{-3}(2.37)		3.00e^{-3}(2.32)	3.00e^{-3}(2.16)	
Year_Customer				-0.006(-0.980)	
Method_Ordinary	0.69(4.65)	0.61(3.60)	0.58(3.96)	0.33(1.74)	
Status_Widowed		0.98(3.32)	1.04(3.49)	0.95(3.25)	0.85(3.14)
NoOfAdressChanges_CNT		0.17(2.02)	0.11(1.29)	0.07(0.86)	
DebtRegisterRepaymentLoanNum			0.16(0.75)	0.89(3.70)	0.51(2.37)
SFLI_AMT			0.35(3.17)	0.40(3.65)	0.44(4.26)
MC_2_Young		-0.90(-2.78)	-0.59(-2.31)	-0.57(-2.38)	
MC_4		0.13(0.76)		0.15(0.89)	0.42(3.04)
Channel_Web			0.39(2.28)	0.57(2.85)	0.47(2.74)
Employment_SocialSecurity					
Distributor_4J			-0.33(-2.05)	-0.39(-2.17)	-0.51(-3.05)
Total_Incoming_AMT			-9.75e^{-07}(−3.66)	-9.66e^{-07}(−3.52)	
DebtRegisterCreditFacilityNum				-0.06(-0.96)	-0.12(-1.93)
Logins_Num					
Car			0.64(2.01)	0.84(2.43)	
Channel_PhoneBank					
Employment_DisabilityPensioner				0.38(1.65)	0.59(2.66)
Employment_Retiree			-1.95(-2.61)	-1.82(-2.44)	
Employment_Unemployed			1.38(1.76)	1.64(2.05)	
Habitation_Other			0.27(1.06)	0.17(0.66)	
Distributor_4A				0.27(1.06)	0.57(2.61)
Distributor_4D			-1.85(-1.83)	-0.45(-0.85)	
Distributor_4G			0.20(1.38)	0.28(1.70)	
Distributor_4L			-13.52(-0.05)	-13.50(-0.05)	

Tabell 4: Estimerte LR-koeffisienter for variabler valgt av LASSO

Panel B-2: For månedene august-desember

	August	September	Oktober	November	Desember
DebtRegisterRepaymentLoanBalance				-2.05e-05(-3.53)	
DebtRegisterNonIELA				-2.93e-05(-1.59)	
App_Weekend			-0.43(-2.46)		
MC_1				0.90(2.45)	
MC_3_Young				-0.55(-1.63)	
Method_Predefined					
Employment_Duration_NoJoB			-0.19(-1.19)		
Employment_Home				1.38(1.25)	
Employment_Student				-0.17(-0.89)	
Habitation_Apartment				-0.58(-1.74)	
Distributor_4E				-13.50(-0.04)	
Distributor_4I				-0.83(-1.57)	
Distributor_4O				0.18 (1.15)	
Homeowner_HouseOrCoop					
Intercept	-2.14(-2.51)	-1.14(-1.37)	-2.49(-2.80)	-2.09(-2.19)	-0.83(-1.13)

Panel B-1 og B-2 viser estimerte regresjonskoeffisienter for variabler valgt av LASSO i månedene august-desember og tilhørende z-verdier i parentes. Dette tilsvarer periode 6-10 i figur 6.

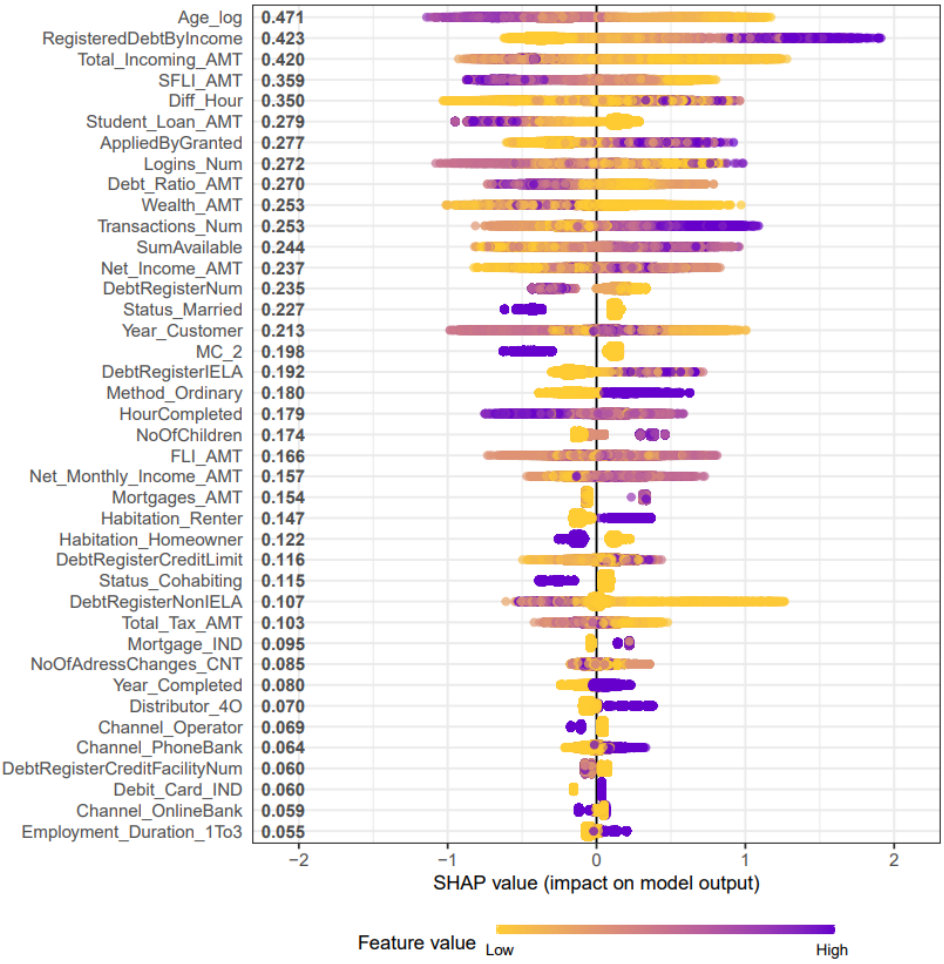
A.5 Optimalisering av hyperparametere i XGBoost

Tabell 5: Optimaliserte hyperparametere i XGBoost

	Eta	Gamma	max_depth	min_child_weight	subsample	nfold
Periode 1	0,388	5,830	3,000	23,637	0,611	0,611
Periode 2	0,388	5,830	3,000	23,637	0,611	0,611
Periode 3	0,388	5,830	3,000	23,637	0,611	0,611
Periode 4	0,388	5,830	3,000	23,637	0,611	0,611
Periode 5	0,474	0,000	3,000	19,749	1,000	1,000
Periode 6	0,333	0,000	3,000	25,000	1,000	1,000
Periode 7	0,388	5,830	3,000	23,637	0,611	0,611
Periode 8	0,388	5,830	3,000	23,637	0,611	0,611
Periode 9	0,388	5,830	3,000	23,637	0,611	0,611
Periode 10	0,388	5,830	3,000	23,637	0,611	0,611

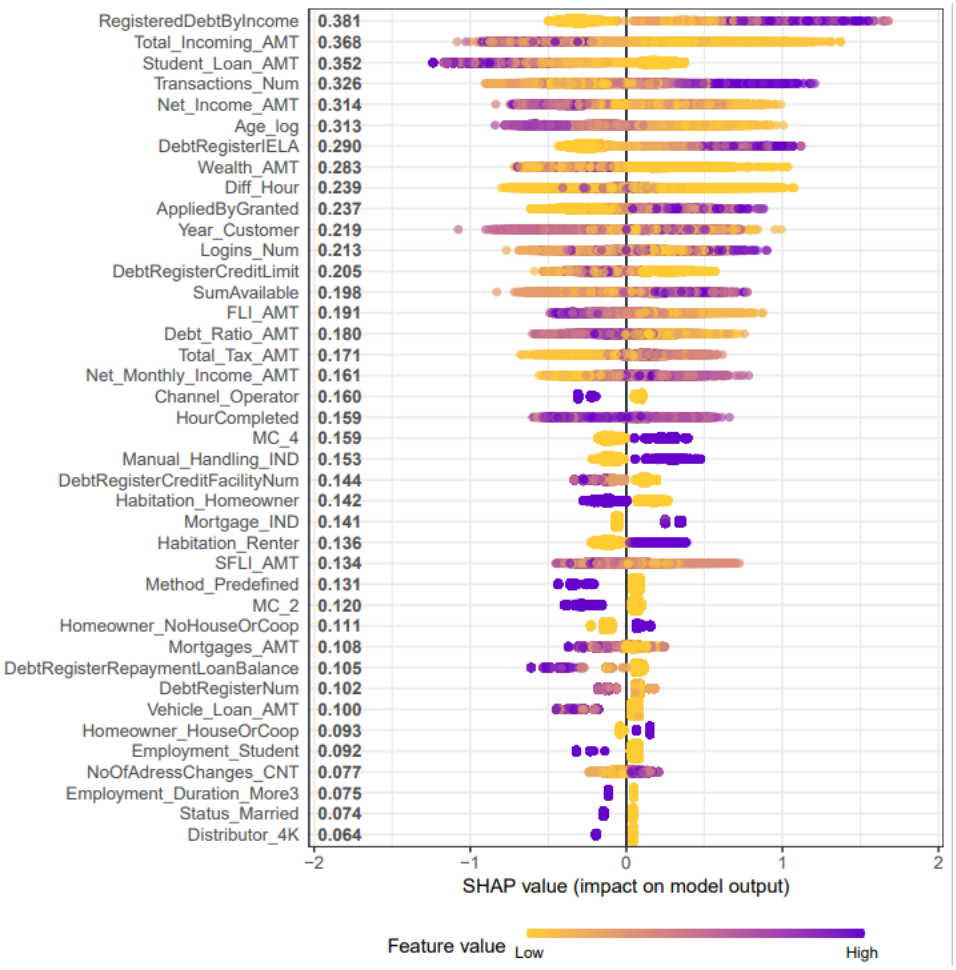
A.6 SHAP

Figur 17: SHAP-plott for periode 1



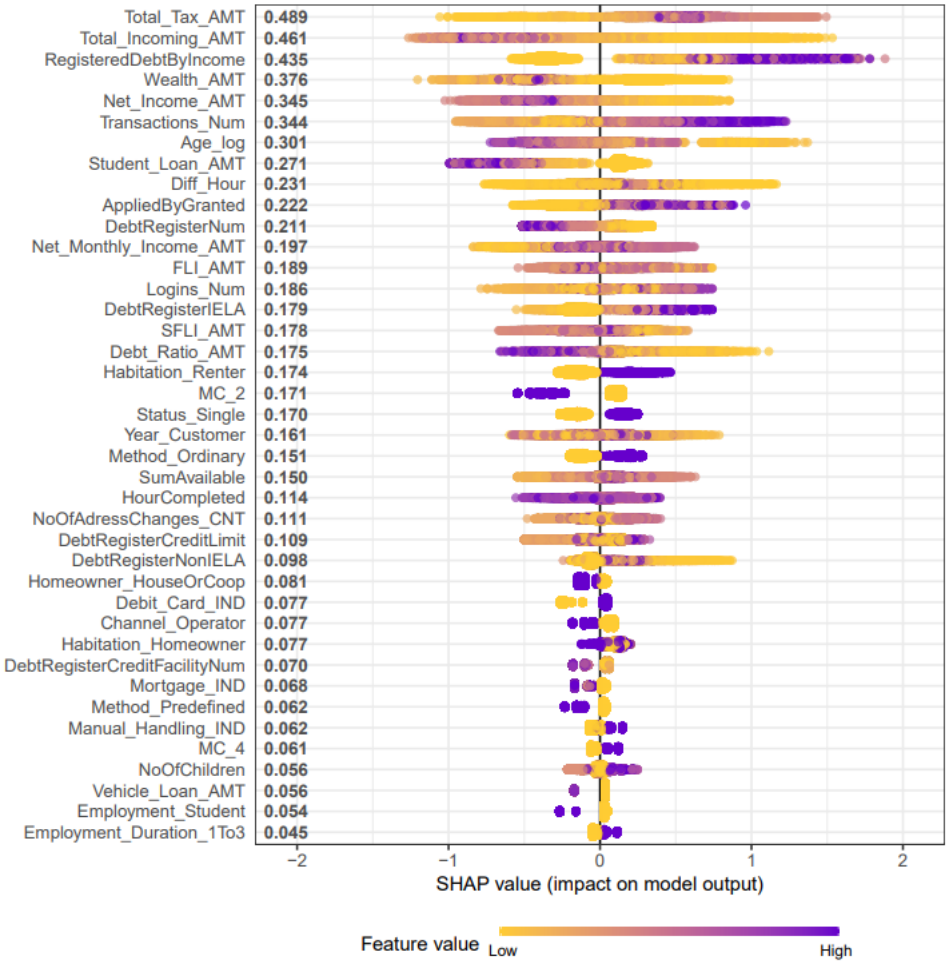
SHAP-plottet viser den globale variabelviktighet til de 40 viktigste variablene for den gitte perioden. Variabelen med høyest variabelviktighet er øverst, deretter følger de resterende variablene i synkende rekkefølge basert på viktigheten. Variablenes SHAP-verdi er gitt langs y-aksen. Videre representerer hver prikk én observasjon fra datasettet. Fargene i plottet angir observasjonenes opprinnelige verdi for den enkelte variabelen, mens x-aksen viser observasjonenes SHAP-verdier.

Figur 18: SHAP-plott for periode 2



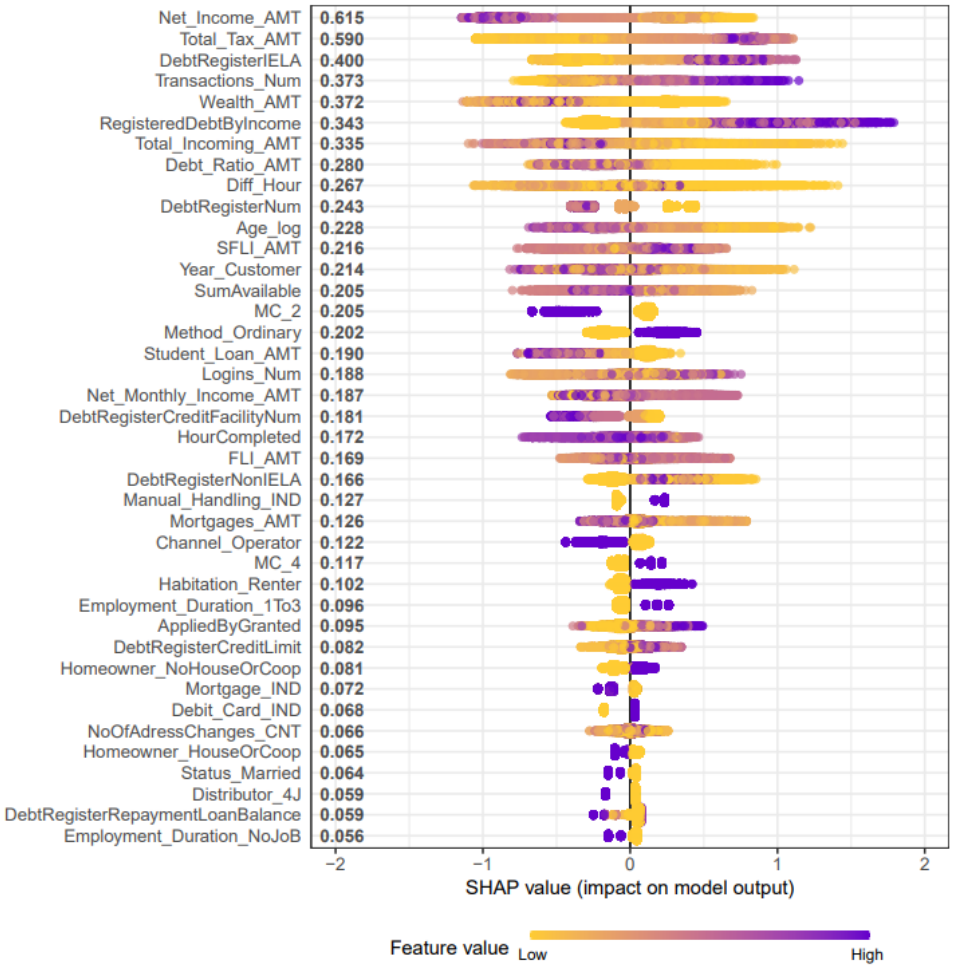
SHAP-plottet viser den globale variabelviktighet til de 40 viktigste variablene for den gitte perioden. Variabelen med høyest variabelviktighet er øverst, deretter følger de resterende variablene i synkende rekkefølge basert på viktigheten. Variablenes SHAP-verdi er gitt langs y-aksen. Videre representerer hver prikk én observasjon fra datasettet. Fargene i plottet angir observasjonenes opprinnelige verdi for den enkelte variabelen, mens x-aksen viser observasjonenes SHAP-verdier.

Figur 19: SHAP-plott for periode 3



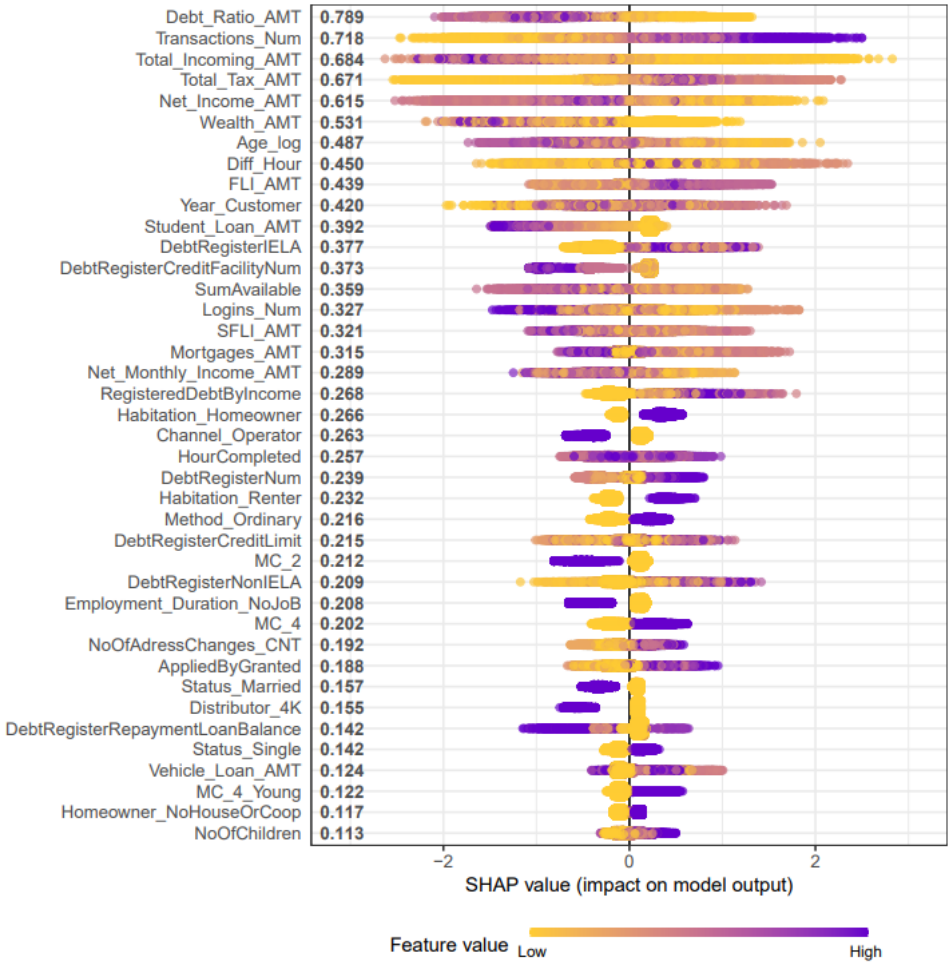
SHAP-plottet viser den globale variabelviktighet til de 40 viktigste variablene for den gitte perioden. Variabelen med høyest variabelviktighet er øverst, deretter følger de resterende variablene i synkende rekkefølge basert på viktigheten. Variablenes SHAP-verdi er gitt langs y-aksen. Videre representerer hver prikk én observasjon fra datasettet. Fargene i plottet angir observasjonenes opprinnelige verdi for den enkelte variabelen, mens x-aksen viser observasjonenes SHAP-verdier.

Figur 20: SHAP-plott for periode 4



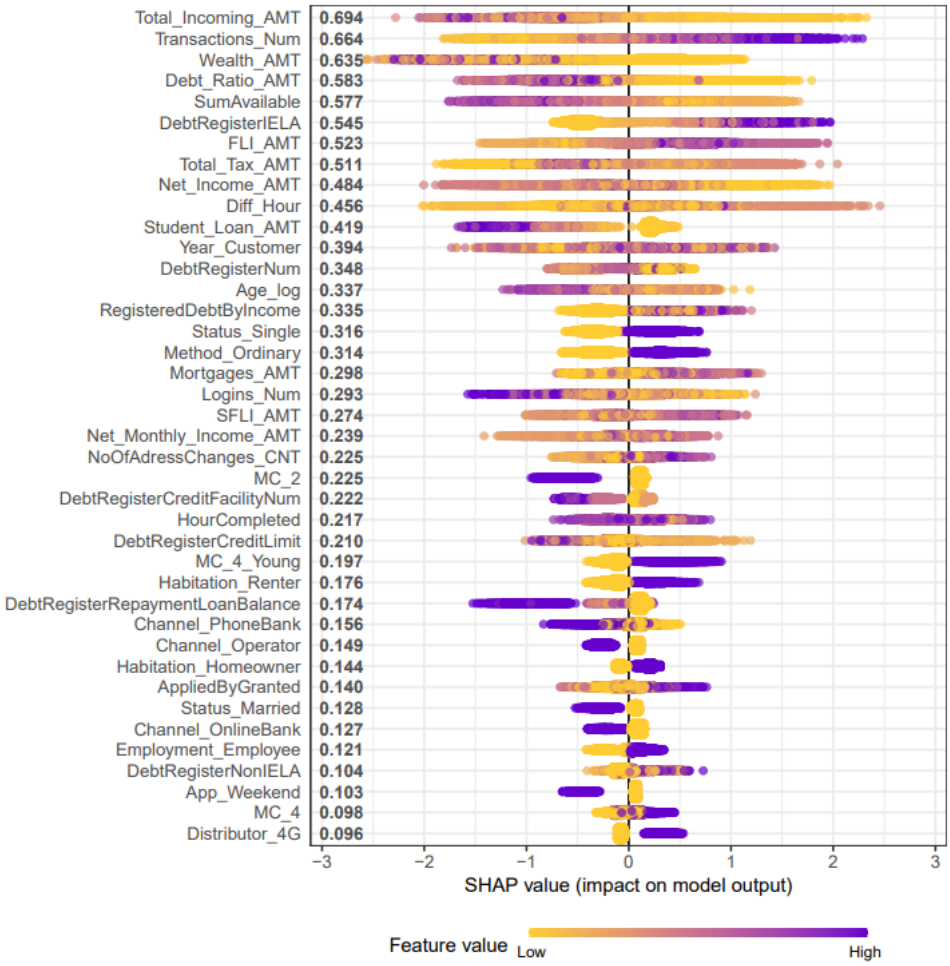
SHAP-plottet viser den globale variabelviktighet til de 40 viktigste variablene for den gitte perioden. Variabelen med høyest variabelviktighet er øverst, deretter følger de resterende variablene i synkende rekkefølge basert på viktigheten. Variablenes SHAP-verdi er gitt langs y-aksen. Videre representerer hver prikk én observasjon fra datasettet. Fargene i plottet angir observasjonenes opprinnelige verdi for den enkelte variabelen, mens x-aksen viser observasjonenes SHAP-verdier.

Figur 21: SHAP-plott for periode 5



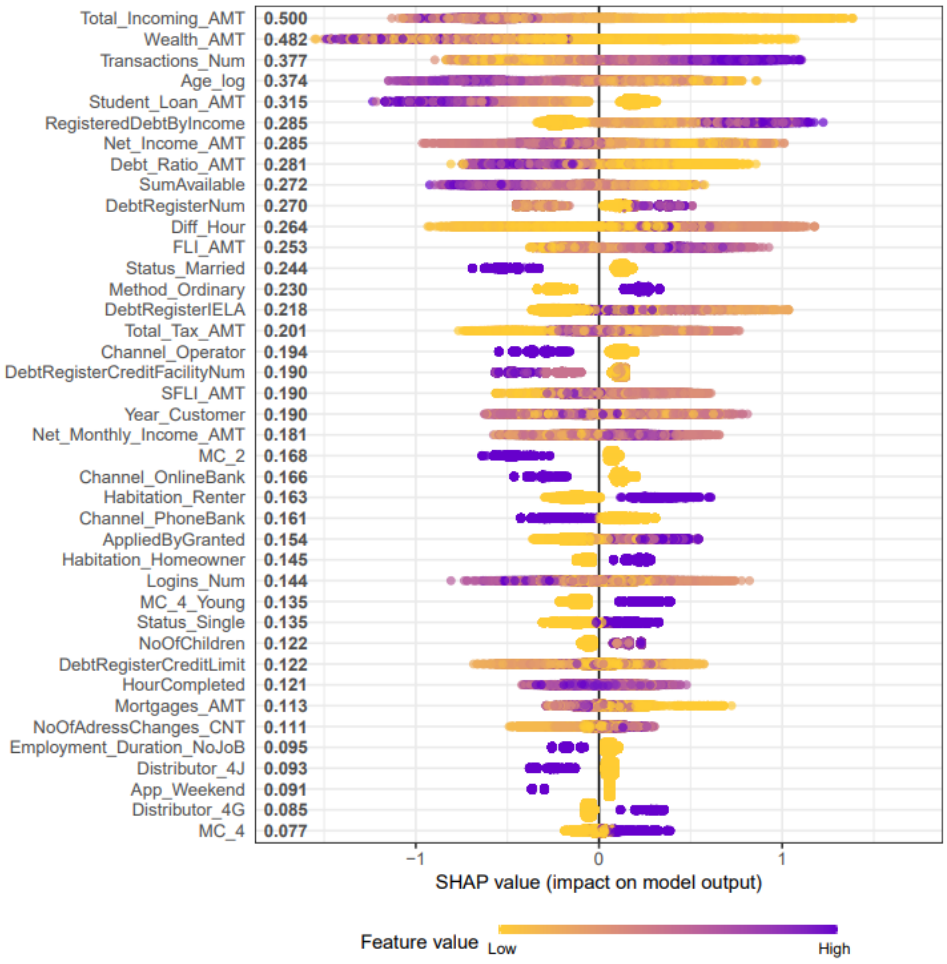
SHAP-plottet viser den globale variabelviktighet til de 40 viktigste variablene for den gitte perioden. Variabelen med høyest variabelviktighet er øverst, deretter følger de resterende variablene i synkende rekkefølge basert på viktigheten. Variablenes SHAP-verdi er gitt langs y-aksen. Videre representerer hver prikk én observasjon fra datasettet. Fargene i plottet angir observasjonenes opprinnelige verdi for den enkelte variabelen, mens x-aksen viser observasjonenes SHAP-verdier.

Figur 22: SHAP-plott for periode 6



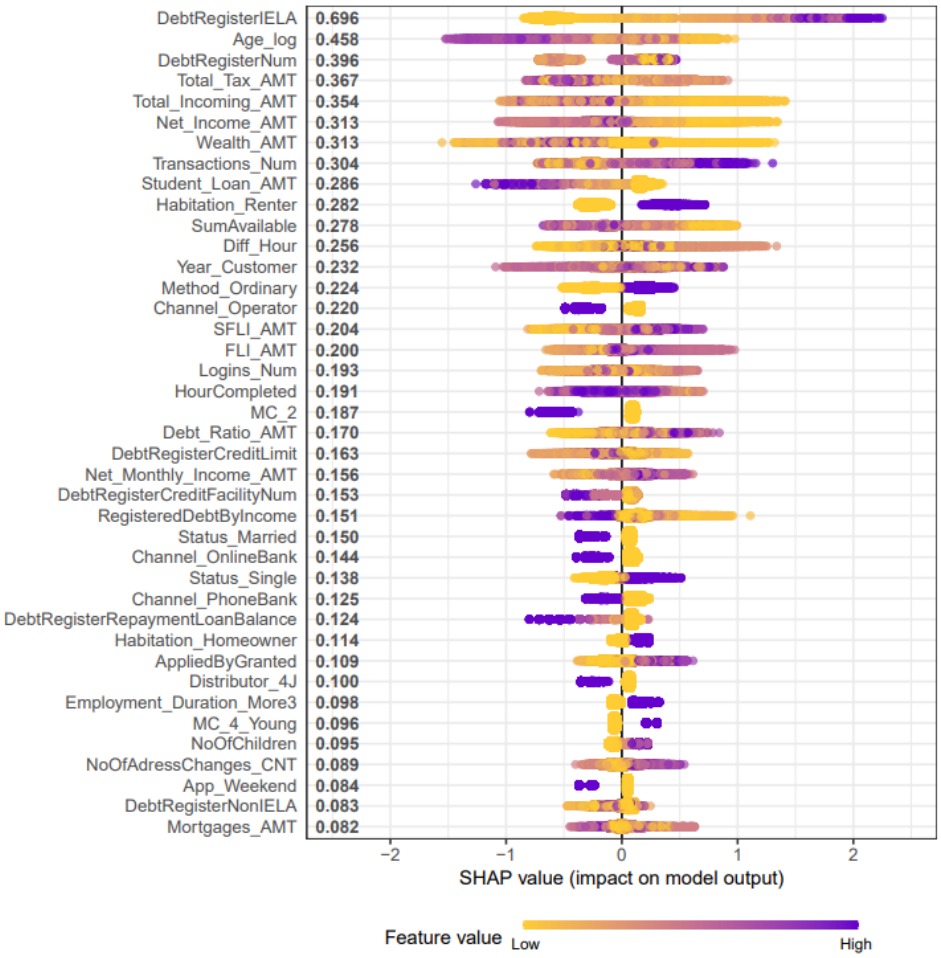
SHAP-plottet viser den globale variabelviktighet til de 40 viktigste variablene for den gitte perioden. Variabelen med høyest variabelviktighet er øverst, deretter følger de resterende variablene i synkende rekkefølge basert på viktigheten. Variablenes SHAP-verdi er gitt langs y-aksen. Videre representerer hver prikk én observasjon fra datasettet. Fargene i plottet angir observasjonenes opprinnelige verdi for den enkelte variabelen, mens x-aksen viser observasjonenes SHAP-verdier.

Figur 23: SHAP-plott for periode 7



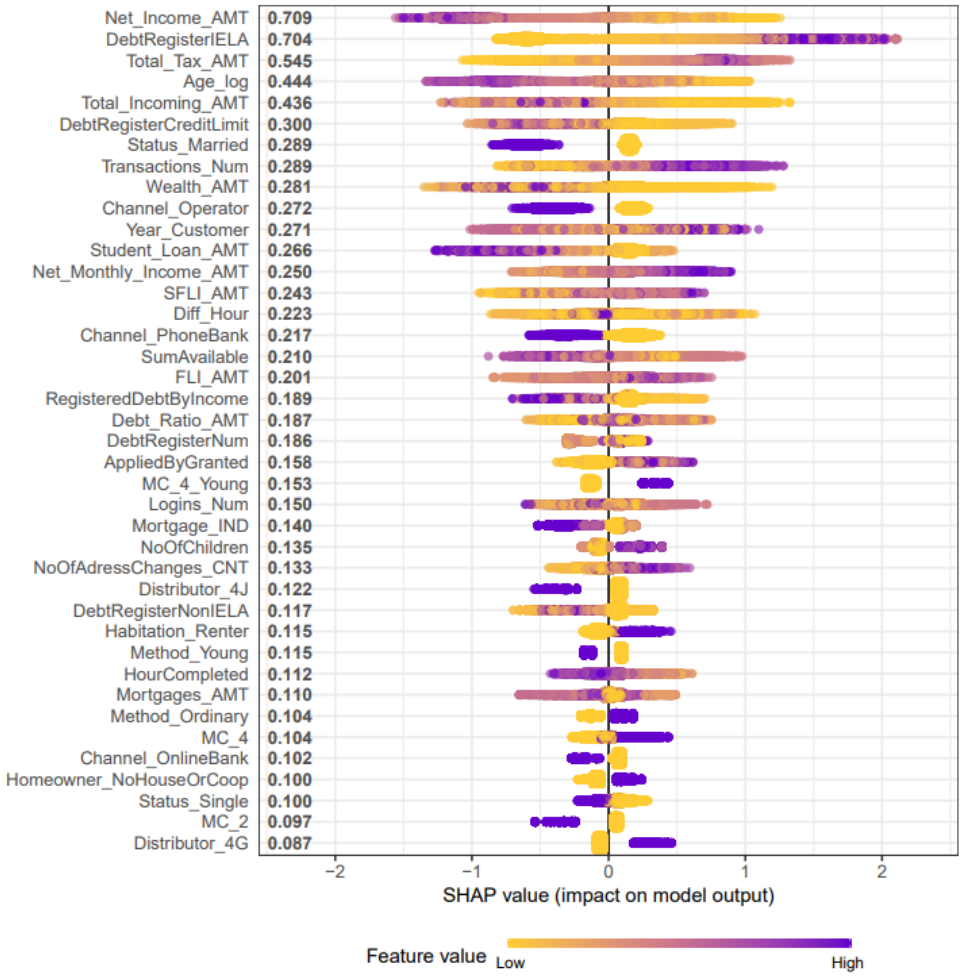
SHAP-plottet viser den globale variabelviktighet til de 40 viktigste variablene for den gitte perioden. Variabelen med høyest variabelviktighet er øverst, deretter følger de resterende variablene i synkende rekkefølge basert på viktigheten. Variablenes SHAP-verdi er gitt langs y-aksen. Videre representerer hver prikk én observasjon fra datasettet. Fargene i plottet angir observasjonenes opprinnelige verdi for den enkelte variabelen, mens x-aksen viser observasjonenes SHAP-verdier.

Figur 24: SHAP-plott for periode 8



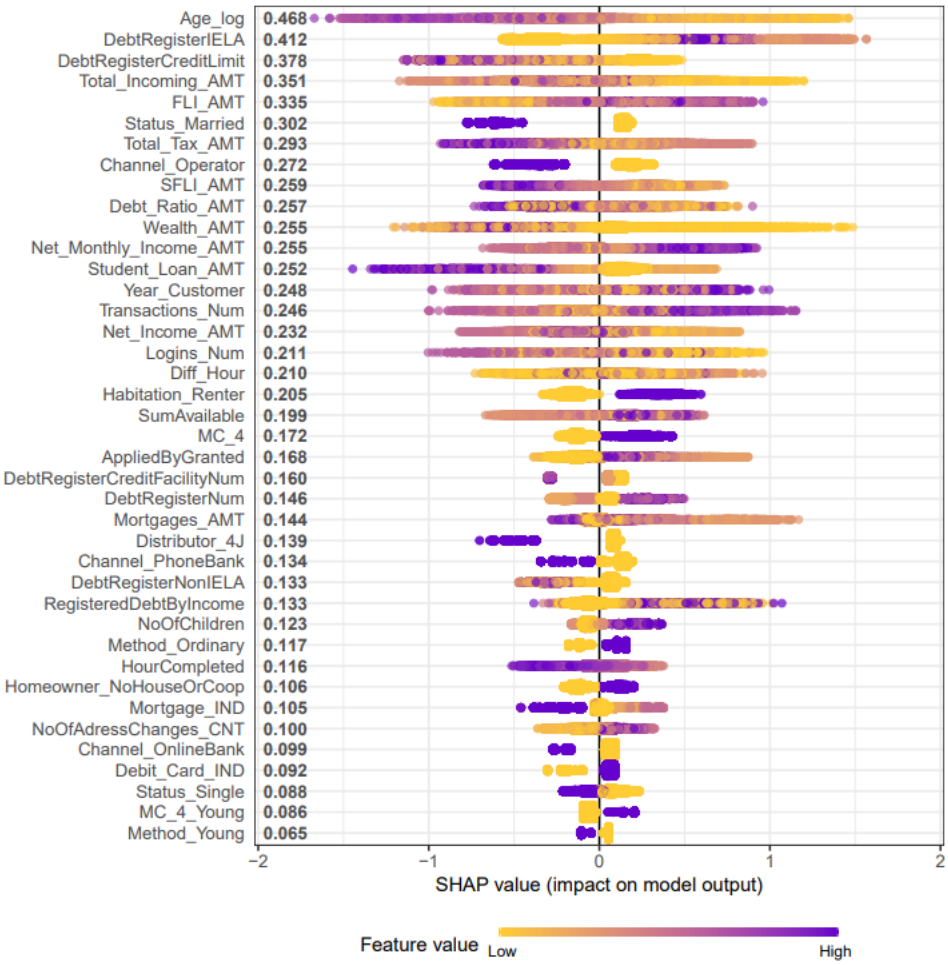
SHAP-plottet viser den globale variabelviktighet til de 40 viktigste variablene for den gitte perioden. Variabelen med høyest variabelviktighet er øverst, deretter følger de resterende variablene i synkende rekkefølge basert på viktigheten. Variablenes SHAP-verdi er gitt langs y-aksen. Videre representerer hver prikk én observasjon fra datasettet. Fargene i plottet angir observasjonenes opprinnelige verdi for den enkelte variabelen, mens x-aksen viser observasjonenes SHAP-verdier.

Figur 25: SHAP-plott for periode 9



SHAP-plottet viser den globale variabelviktighet til de 40 viktigste variablene for den gitte perioden. Variabelen med høyest variabelviktighet er øverst, deretter følger de resterende variablene i synkende rekkefølge basert på viktigheten. Variablenes SHAP-verdi er gitt langs y-aksen. Videre representerer hver prikk én observasjon fra datasettet. Fargene i plottet angir observasjonenes opprinnelige verdi for den enkelte variabelen, mens x-aksen viser observasjonenes SHAP-verdier.

Figur 26: SHAP-plott for periode 10



SHAP-plottet viser den globale variabelviktighet til de 40 viktigste variablene for den gitte perioden. Variabelen med høyest variabelviktighet er øverst, deretter følger de resterende variablene i synkende rekkefølge basert på viktigheten. Variablenes SHAP-verdi er gitt langs y-aksen. Videre representerer hver prikk én observasjon fra datasettet. Fargene i plottet angir observasjonenes opprinnelige verdi for den enkelte variabelen, mens x-aksen viser observasjonenes SHAP-verdier.

A.7 Beskrivende statistikk

Tabell 6: Beskrivende statistikk for variabler i endelig datasett

Variabel	Gj.snitt	Std	Min	1. kvartil	Median	3. kvartil	Maks
Age_log	3,60	0,41	2,89	3,26	3,58	3,93	4,56
App_Night	0,05	0,22	0,00	0,00	0,00	0,00	1,00
App_weekend	0,11	0,31	0,00	0,00	0,00	0,00	1,00
AppliedByGranted	1,32	0,60	1,00	1,00	1,00	1,32	3,00
Channel_OnlineBank	0,24	0,43	0,00	0,00	0,00	0,00	1,00
Channel_Operator	0,41	0,49	0,00	0,00	0,00	1,00	1,00
Channel_PhoneBank	0,25	0,43	0,00	0,00	0,00	0,00	1,00
Channel_ResponsePage	0,00	0,00	0,00	0,00	0,00	0,00	1,00
Channel_Web	0,10	0,30	0,00	0,00	0,00	0,00	1,00
Debit_Card_IND	0,77	0,42	0,00	1,00	1,00	1,00	1,00
Debt_Ratio_AMT	1,81	1,73	0,00	0,07	1,43	3,25	5,13
DebtRegisterCreditFacilityNum	1,21	1,26	0,00	0,00	1,00	2,00	4,00
DebtRegisterCreditLimit	33 615,00	44 245,00	0,00	0,00	15 000,00	50 000,00	151 000,00
DebtRegisterIELA	8 349,00	22 315,00	0,00	0,00	0,00	2,00	89 223,00
DebtRegisterNonIELA	1 811,00	4 080,00	0,00	0,00	0,00	679,00	15 197,00
DebtRegisterNum	1,38	1,44	0,00	0,00	1,00	2,00	5,00
DebtRegisterRepaymentLoanBalance	3 952,00	14 110,00	0,00	0,00	0,00	0,00	59 715,00
DebtRegisterRepaymentLoanNum	0,09	0,29	0,00	0,00	0,00	0,00	1,00
Diff_hour	24,64	45,57	0,03	0,09	1,15	24,45	166,58
Distributor_1	0,03	0,17	0,00	0,00	0,00	0,00	1,00
Distributor_3	0,13	0,34	0,00	0,00	0,00	0,00	1,00
Distributor_4A	0,04	0,20	0,00	0,00	0,00	0,00	1,00
Distributor_4B	0,01	0,10	0,00	0,00	0,00	0,00	1,00
Distributor_4C	0,01	0,11	0,00	0,00	0,00	0,00	1,00
Distributor_4D	0,01	0,10	0,00	0,00	0,00	0,00	1,00
Distributor_4E	0,01	0,09	0,00	0,00	0,00	0,00	1,00
Distributor_4F	0,01	0,09	0,00	0,00	0,00	0,00	1,00
Distributor_4G	0,13	0,34	0,00	0,00	0,00	0,00	1,00
Distributor_4H	0,02	0,14	0,00	0,00	0,00	0,00	1,00
Distributor_4I	0,02	0,14	0,00	0,00	0,00	0,00	1,00
Distributor_4J	0,17	0,38	0,00	0,00	0,00	0,00	1,00
Distributor_4K	0,13	0,33	0,00	0,00	0,00	0,00	1,00
Distributor_4L	0,02	0,12	0,00	0,00	0,00	0,00	1,00
Distributor_4M	0,03	0,17	0,00	0,00	0,00	0,00	1,00
Distributor_4N	0,03	0,17	0,00	0,00	0,00	0,00	1,00
Distributor_4O	0,20	0,40	0,00	0,00	0,00	0,00	1,00
Employment_Duration_1To3	0,20	0,40	0,00	0,00	0,00	0,00	1,00
Employment_Duration_Less1	0,12	0,33	0,00	0,00	0,00	0,00	1,00
Employment_Duration_More3	0,43	0,50	0,00	0,00	0,00	1,00	1,00
Employment_Duration_NoJoB	0,25	0,43	0,00	0,00	0,00	0,00	1,00

Tabell 6: Beskrivende statistikk for variabler i endelig datasett

Variabel	Gj.snitt	Std	Min	1. kvartil	Median	3. kvartil	Maks
Employment_DisabilityPensioner	0,05	0,21	0,00	0,00	0,00	0,00	1,00
Employment_Employee	0,69	0,46	0,00	0,00	1,00	1,00	1,00
Employment_Home	0,00	0,03	0,00	0,00	0,00	0,00	1,00
Employment_Other	0,01	0,10	0,00	0,00	0,00	0,00	1,00
Employment_Retiree	0,09	0,28	0,00	0,00	0,00	0,00	1,00
Employment_SelfEmployed	0,02	0,15	0,00	0,00	0,00	0,00	1,00
Employment_SocialSecurity	0,00	0,01	0,00	0,00	0,00	0,00	1,00
Employment_Student	0,10	0,30	0,00	0,00	0,00	0,00	1,00
Employment_TempEmployee	0,05	0,21	0,00	0,00	0,00	0,00	1,00
Employment_Unemployed	0,00	0,04	0,00	0,00	0,00	0,00	1,00
FLI_AMT	1,47	0,59	0,49	1,19	1,47	1,86	2,87
Habitation_Apartment	0,07	0,26	0,00	0,00	0,00	0,00	1,00
Habitation_Homeowner	0,57	0,50	0,00	0,00	1,00	1,00	1,00
Habitation_Other	0,04	0,19	0,00	0,00	0,00	0,00	1,00
Habitation_Parents	0,10	0,30	0,00	0,00	0,00	0,00	1,00
Habitation_Renter	0,22	0,41	0,00	0,00	0,00	0,00	1,00
Homeowner_HouseAndCoop	0,01	0,12	0,00	0,00	0,00	0,00	1,00
Homeowner_HouseOrCoop	0,55	0,50	0,00	0,00	1,00	1,00	1,00
Homeowner_NoHouseOrCoop	0,44	0,50	0,00	0,00	0,00	1,00	1,00
HourCompleted	13,97	4,58	0,00	11,00	14,00	17,00	23,00
ID_Bank	0,92	0,27	0,00	1,00	1,00	1,00	1,00
ID_Physical	0,08	0,27	0,00	0,00	0,00	0,00	1,00
ID_Pum	0,00	0,03	0,00	0,00	0,00	0,00	1,00
Logins_Num	22,37	23,23	0,00	4,00	15,00	22,37	83,00
Manual_Handling_IND	0,27	0,44	0,00	0,00	0,00	1,00	1,00
MC_1	0,03	0,17	0,00	0,00	0,00	0,00	1,00
MC_2	0,22	0,41	0,00	0,00	0,00	0,00	1,00
MC_2_Young	0,03	0,18	0,00	0,00	0,00	0,00	1,00
MC_3	0,08	0,28	0,00	0,00	0,00	0,00	1,00
MC_3_Extra	0,12	0,13	0,00	0,00	0,00	0,00	1,00
MC_3_Plus	0,00	0,06	0,00	0,00	0,00	0,00	1,00
MC_3_Young	0,03	0,16	0,00	0,00	0,00	0,00	1,00
MC_4	0,35	0,48	0,00	0,00	0,00	1,00	1,00
MC_4_Extra	0,03	0,18	0,00	0,00	0,00	0,00	1,00
MC_4_Unique	0,01	0,12	0,00	0,00	0,00	0,00	1,00
MC_4_Young	0,19	0,40	0,00	0,00	0,00	0,00	1,00
Method_Mortgage	0,14	0,35	0,00	0,00	0,00	0,00	1,00
Method_Ordinary	0,46	0,50	0,00	0,00	0,00	1,00	1,00
Method_Predefined	0,13	0,33	0,00	0,00	0,00	0,00	1,00
Method_Unique	0,02	0,13	0,00	0,00	0,00	0,00	1,00
Method_Young	0,25	0,44	0,00	0,00	0,00	1,00	1,00
Mortgage_IND	0,32	0,44	0,00	0,00	0,00	1,00	1,00
Mortgages_AMT	876 144,00	979 359,50	0,00	0,00	540 000,00	1 600 000,00	3 000 000,00

Tabell 6: Beskrivende statistikk for variabler i endelig datasett

Variabel	Gj.snitt	Std	Min	1. kvartil	Median	3. kvartil	Maks
Net_Income_AMT	303 275,00	209 539,00	8 746,00	128 533,00	288 564,00	428 256,00	773 624,00
Net_Monthly_Income_AMT	27 585,00	10 737,00	8 000,00	20 000,00	27 000 ,00	34 000,00	50 000,00
NoOfAdressChanges_CNT	1,55	0,61	1,00	1,00	1,40	2,00	3,00
NoOfChildren	0,92	1,10	0,00	0,00	0,00	2,00	3,00
RegisteredDebtByIncome	0,03	0,08	0,00	0,00	0,00	0,00	0,34
SFLI_AMT	1,43	0,56	0,49	1,08	1,43	1,72	2,73
Status_Cohabiting	0,24	0,43	0,00	0,00	0,00	0,00	1,00
Status_Divorced	0,04	0,20	0,00	0,00	0,00	0,00	1,00
Status_Married	0,32	0,47	0,00	0,00	0,00	1,00	1,00
Status_Single	0,37	0,48	0,00	0,00	0,00	1,00	1,00
Status_Widowed	0,03	0,16	0,00	0,00	0,00	0,00	1,00
Student_Loan_AMT	48 145,00	95 868,03	0,00	0,00	0,00	30 000,00	320 000,00
SumAvailable	9 809,00	9 403,72	-6 400,00	3 935,00	8 996,00	14757,00	31 530,00
Total_Incoming_AMT	323 172,00	372 119,20	0,00	49 825,00	219 101,00	413 731,00	1 462127,00
Total_Tax_AMT	100 456,00	80 490,58	0,00	31 701,00	91 944,00	143 000,00	293 317,00
Transactions_Num	63,32	56,94	0,00	7,00	54,40	101,00	188,00
Vehicle_Loan_AMT	24 352,00	60 351,07	0,00	0,00	0,00	0,00	220 000,00
Wealth_AMT	223 602,00	459 385,10	0,00	0,00	0,00	163 338,00	1 722 820,00
Year_Completed	19,86	0,34	19,00	20,00	20,00	20,00	20,00
Year_Customer	14,65	11,09	0,10	4,40	13,95	22,43	36,68
Van	0,05	0,23	0,00	0,00	0,00	0,00	1,00
Car	0,95	0,23	0,00	1,00	1,00	1,00	1,00
MC	0,05	0,22	0,00	0,00	0,00	0,00	1,00

A.8 Utelatte variabler

Tabell 7: Beskrivelse av utelatte variabler

Variabelnavn	Beskrivelse
BK_Account_ID	Kontinuerlig; unik ID for konto
BK_Application_ID	Kontinuerlig; unik ID for søknad
PeriodId	Kontinuerlig; dato for når søknaden ble opprettet, på format: YYYYMMDD
Created_DT	Kontinuerlig; tidspunkt for når søknaden ble opprettet, på format: YYYY-MM-DD HH:MM:SS.S
Completed_DT	Kontinuerlig; tidspunkt for når den godkjente søknaden ble signert, på format: YYYY-MM-DD HH:MM:SS.S
HourCompleted	Kontinuerlig; timen på døgnet, hentet ut fra Completed_DT
BirthYear	Kontinuerlig; fødselsår, på format: YY
Homeowner_IND	Dummy; 1 om personen eier hus, med boliglån hos dataleverandør
Housing_Cooperative_IND	Dummy; 1 om personen er eier i borettslag, med boliglån hos dataleverandør
NoOfCars	Kontinuerlig; antall biler
Debt_Ratio_AMT	Kontinuerlig; gjeldsgrad
NoOfCampers	Kontinuerlig; antall campingvogner
NoOfMC	Kontinuerlig: antall motorsykler
NoOfVan	Kontinuerlig; antall varebiler
NoOfOtherVehicle	Kontinuerlig; antall andre kjøretøy
TX_Gross_Income_AMT	Kontinuerlig: årlig bruttoinntekt, beregnet basert på skattedata
Total_Debt_AMT	Kontinuerlig; total gjeld i banken (dataleverandør)
Rental_Expenses_AMT	Kontinuerlig; beløp på husleieutgifter
Monthly_Rental_Income_AMT	Kontinuerlig; beløp på inntekter fra utleie
Sum_Pledge_Remarks_AMT	Kontinuerlig; sum på heftelser
Consent_IND	Dummy; 1 dersom lånsøker har gitt samtykke til å hente ut data fra dataleverandør
Customer_From_DT	Kontinuerlig; tidspunkt for det eldste registrerte kundeforhold hos dataleverandør, på format: YYYY-MM-DD HH:MM:SS.S
LmtByIncome	Kontinuerlig; Applied_Credit_Limit_AMT/Net_Income_AMT
RatioFLI	Kontinuerlig; SFLI_AMT/FLI_AMT
DebtByIncome	Kontinuerlig; Total_Debt_AMT/Net_Income_AMT
ChangeInIncomeABS	Kontinuerlig; absoluttverdien av %-vis endring i inntekt de to siste skatteårene (basert på de to siste årene sin Net_Income_AMT)
ChangeInIncome	Kontinuerlig; %-vis endring i inntekt de to siste skatteårene (basert på de to siste årene sin NET_INCOME_AMT)
TransactionsByIncome	Kontinuerlig; Total_Incoming_AMT/Net_Income_AMT

Hanna Hovd og Rikke Olsen Aase

Misligh old av kredittkortgjeld

NTNU
Kunnskap for en bedre verden