Stockholms universitet

Logistisk regression för att prediktera om en låneansökan ska beviljas eller avslås

Viktor Ellerud

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2023:13
Matematisk statistik
Juni 2023

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2023:13**
http://www.math.su.se/matstat

Logistisk regression för att prediktera om en låneansökan ska beviljas eller avslås

Viktor Ellerud*

Juni 2023

Sammanfattning

Syftet med detta arbete är att kunna prediktera samt avgöra vilka faktorer som är viktigast för att avgöra om en låneansökan ska beviljas eller avslås givet en viss information om låntagaren. Eftersom responsvariabeln är binär kommer logistisk regression att appliceras. De fyra framtagna modellernas prediktionsförmåga jämförs och tolkas utifrån ett flertal jämförelsemått där samtliga modeller presterar förhållandevis likvärdigt.

Slutsatsen som kan dras är att samtliga framtagna modeller skulle kunna användas i prediktivt syfte. Det visar sig att det nödvändigtvis inte behövs särskilt många förklarande variabler för att kunna förklara samt prediktera om en låneansökan ska beviljas eller avslås. Den slutliga modellen är en modell med endast en förklaringsvariabel, den mest inflytelserika faktorn Credit History, vilket motsvarar om den lånesökandes kredithistorik är bra eller dålig. Slutsatsen som kan dras från detta är att det är viktigt för privatpersoner att ha god kredithistorik för att låneansökan ska godkännas.

Dock kan slutresultatet ifrågasättas eftersom datasetet består av en relativt stor andel bortfall. Däremot har olika metoder för behandling av bortfallet jämförts och de resulterade i likvärdiga resultat. Detta talar för att bortfallet ej bör ha en betydande påverkan på tillförlitligheten i resultaten.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: Viktor@ellerud.se. Handledare: Ola Hössjer och Mohamed El Khalifi.

Abstract

The purpose of this work is to predict and determine the most important factors for determining whether a loan application should be approved or denied given specific information about the borrower. Since the response variable is binary, logistic regression will be applied. The predictive power of the four developed models is interpreted and compared based on several measures of comparison, where all models perform relatively similarly.

The conclusion that can be made is that all developed models could be used for predictive purposes. It turns out that it is not necessarily required to have a large number of explanatory variables to explain and predict whether a loan application should be approved or denied. The final model is a model with only one explanatory variable, Credit_History, which corresponds to whether the loan applicant's Credit_History is good or bad. The conclusion that can be drawn from this is that it is important for individuals to have a good Credit_History in order for their loan application to be approved.

However, the final result can be questioned since the dataset contains a relatively large proportion of missing data. Nevertheless, the different methods for handling missing data have been compared and resulted in very similar results. This indicates that the missing data should not have a substantial impact on the reliability of the results.

Förord

Till att börja med vill jag tacka mina två handledare Ola Hössjer och Mohamed El Khalifi som har bidragit med goda råd och stort engagemang i mitt arbete. Tack för att ni alltid har hjälpt till och funnits där när jag behövt hjälp. Ett stort tack är riktat till er båda! Dessutom ett extra stort tack till Mohamed El Khalifi som hjälpt mig att hitta samt välja datasetet. Jag vill även tacka Leo Levenius och Esma Özdolap för korrekturläsning och som hjälpsamma bollplank.

Detta är ett examensarbete i matematisk statistik motsvarande 15 högskolepoäng. Arbetet är skrivet vid den Matematiska institutionen på Stockholms universitet.

Innehållsförteckning

1 Introduktion **1**
1.1 Syfte och metodik 1
1.2 Tidigare forskning 1

2 Teori **3**
2.1 Lånförutsägelse samt fallissemang 3
2.2 Logistisk regression 3
2.3 Akaikes informationskriterium (AIC) 4
2.4 Bayesianska informatioskriteriet (BIC) 5
2.5 Bakåtelimination och framåtselektion 5
2.6 Hosmer–Lemeshow-testet 6
2.7 Förklaringsgrad (R^2) 6
2.7.1 McFadden R^2 7
2.7.2 Cox och Snell R^2 7
2.8 Sammanblandningsmatris 7
2.9 Receiver operating characteristic (ROC-kurva) 8
2.10 Arean under kurvan (AUC) 9
2.11 Residualavvikelse 9

3 Data **10**
3.1 Beskrivning av data 10
3.2 Bortfall 11
3.3 Explorativ dataanalys (EDA) 12

4 Resultat och analys **15**
4.1 Modellframtagning 16
4.2 Modelljämförelse 17

5 Slutsats **22**

6 Diskussion **23**

Referenser **26**

Appendix **28**

1 Introduktion

I denna rapport kommer fokus att ligga på lånförutsägelse. Det vill säga prediktera om en låneansökan kommer godkännas eller nekas. Detta grundar sig på kreditvärderingsmodeller som allmänt används inom finansbranschen för att bedöma betalningsförmågan hos den lånesökande. Kreditvärderingsmodeller är en väsentlig del av utlåningsprocessen. Därav kommer även fallissemang (inställande av betalningar, exempelvis beroende på personlig konkurs) att diskuteras eftersom sannolikheten för fallissemang ligger till grund för att avgöra om en given låneansökan ska beviljas eller avslås.

1.1 Syfte och metodik

Syfte med denna rapport är att kunna prediktera om en låneansökan kommer beviljas eller avslås givet viss information om låntagaren. Detta genom att skapa en modell som predikterar huruvida den lånesökande uppfyller bankens krav eller inte för ett givet lån.

Med en sådan modell skulle banker kunna reducera sin kreditrisk. Det vill säga att inte bevilja lån som man bedömer inte kan betalas tillbaka, i syfte att öka vinsten. Därav kan det finnas ett värde för banker och kreditinstitut att använda denna information för att kunna utveckla mer träffsäkra kreditbedömningar och minska risken för kreditförluster. Modellen skulle även kunna hjälpa privatpersoner att få en bättre förståelse för hur kreditgivare fattar beslut och identifiera de viktigaste faktorerna som bidrar till om en låneansökan accepteras eller avslås.

Det finns flera olika möjliga metoder för att förutspå om en låneansökan beviljas eller avslås men i denna rapport kommer logistisk regression att användas. Fyra olika modeller kommer att skapas. En modell kommer tas fram genom minimering av AIC samt den andra genom minimering av BIC. Den tredje modellen är en jämförelsemodell och den sista modellen tas fram utifrån en undersökning som ger en modell med de mest inflytelserika parametrarna.

1.2 Tidigare forskning

I artikeln av Louzada et al. (2016) [2, ss. 119–132] jämförs nio olika metoder för prediktion av låneförfall. Dessa metoder är neurala nätverk, stödvektormaskin, linjär regression, beslutsträd, logistisk regression, suddig logik, genetisk programmering, diskriminantanalys och bayesiska nätverk. Logistisk regression, beslutsträd och neurala nätverk är främst använda som jämförelsemetoder som måste överträffas. Stödvektormaskiner bygger på att

hitta det hyperplan som på ett optimalt sätt delar upp data i två delar. Detta gör metoden till en icke-probabilistisk binär linjär klassificeringsalgoritm. Vidare är suddig logik en metod som används för att lösa problem där lösningen ofta innefattar antaganden, förenklingar eller regler som inte nödvändigtvis garanterar en optimal lösning. Det är däremot en metod som på ett effektivt sätt kan ge tillräckligt bra resultat. Vidare är enligt Louzada et al. (2016) suddig logik och stödvektormaskin de bästa metoderna för prediktion av fallissemang (när en låntagare inte utför de nödvändiga betalningarna av en skuld). Vidare varierar prestanda mellan metoderna beroende på graden av obalans i träningsdata mellan antalet som lyckas samt misslyckas med att återbetala sina lån enligt avtal. Exempelvis har metoderna beslutsträd och logistisk regression varierande prestanda. Däremot har metoderna neurala nätverk, genetisk programmering samt linjär regression sämre prestanda vid obalans i träningsdata.

Vidare har ovannämnda metoder även undersökts i Hand och Henleys (1997) [3, ss. 534–537] artikel för att prediktera om en låneansökan resulterar i fallissemang eller inte. Enligt Hand och Henleys (1997) artikel blev analysen att metoder som är lätta att förstå är mycket mer tilltalande. Exempel på sådana metoder är logistisk regression, nearest neighbour och beslutsträd. Dessa metoder är mer tilltalande än metoder som i princip är "black boxes" där neurala nätverk utgör ett exempel. Denna metod är mest lämpad i fall då förståelsen för datans struktur är bristfällig. Författarnas slutsats är att med god förståelse för datamaterialet och problemet förväntas metoder som utnyttjar denna förståelse prestera bättre. Inom kreditvärdering finns en solid förståelse för problemet. Det är därför osannolikt att nya metoder kommer att leda till betydande förbättringar i prediktionsnoggrannheten. Detta är anledningen som Hand och Henleys (1997) ger till varför neurala nätverk inte har fått större genomslagskraft inom denna sektor trots att banker har experimenterat med metoden i flera år.

Om responsvariabeln istället är Loan_Status det vill säga om en låneansökan beviljas eller inte så ändras förutsättningarna något. Utifrån Tumuluru et al. (2022) [4, ss. 349–352] kan det finnas både likheter och skillnader jämfört med den tidigare nämnda forskningen kopplad till fallissemang. I denna artikel blir slutsatsen likt Hand och Henleys (1997) artikel, att enklare metoder presterar väl. Utifrån Tumuluru et al. (2022) framhävs random forest (ett flertal beslutsträd) samt logistisk regression som de två bästa metoderna för prediktion av om en låneansökan skall accepteras eller nekas. Detta ligger i linje med Hand och Henleys (1997) artikel eftersom båda artiklarna jämför samma metoder med några undantag. Dessutom överensstämmer resultatet från Kansal et al. (2021) [5, s. 3447] väl med Tumuluru et al. (2022). Liknande metoder jämförs i Kansal et al. (2021) för att förutspå om en låneansökan accepteras eller nekas. Resultatet blir återigen att random forest samt logis-

tisk regression är de bästa metoderna för att förutspå om en låneansökan ska beviljas eller avslås.

2 Teori

2.1 Lånförutsägelse samt fallissemang

Lånförutsägelse används av banker och kreditinstitut för att avgöra om en individs låneansökan ska beviljas eller inte. Om en låneansökan beviljas eller avslås baseras på en mängd olika faktorer kopplade till låntagaren samt dess förmåga att återbetala lånet. Några exempel på faktorer kan vara inkomst, kön och kredithistoria.

Utlåning medför en risk för långivarna som kan benämnas låneförfall eller fallissemang. Fallissemang inträffar när en låntagare inte genomför de betalningar som är överenskomna enligt lånevillkoren. Däremot kan definitionen för fallissemang variera. Det kan innebära att en betalning inte inträffat inom en viss tid (exempelvis tre månader) efter betalningen skulle varit utförd. Det kan även innebära att en individ helt och hållet misslyckas utföra de avtalade betalningarna (exempelvis på grund av personlig konkurs). Kreditgivaren ger lån till låntagare i utbyte mot löftet om återbetalning med ränta. Det är räntan som ger en vinst för långivaren som kompensation för den tagna risken. Eftersom det är viktigt för långivare att minimera antalet lån som inte betalas tillbaka behöver banker kunna förutspå om den lånesökande ska beviljas lånet eller inte givet en viss information. Lånförutsägelse grundar sig på om den lånesökande uppfyller bankens kriterier. Vidare baserar sig detta på sannolikheten för fallissemang. En låneansökan beviljas om sannolikheten för fallissemang är mindre än någon gräns som ligger strikt mellan noll och ett. Sannolikheten för kreditförlust eller fallissemang baseras på att långivaren anpassar en kreditvärderingsmodell som i sin tur avgör utfallet huruvida en lånesökande beviljas lånet eller inte [8, ss. 1–4].

2.2 Logistisk regression

Låt Y vara en binär responsvariabel som antingen antar värdet 0 eller 1. Vidare definieras $\pi(x)$ som $\mathbb{P}(Y = 1 \mid x)$ beroende på värdena $x = (x_1, x_2, ..., x_q)^T$ från q prediktorer.

Den logistiska regressionsmodellen är

$$\pi(x) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}},$$

där α är interecptet och där $\beta = (\beta_1, \ldots, \beta_q)$ innehåller effektparametrar för var och en av de q kovariaterna. Sannolikheten för $Y = 1$ givet x har egenskapen att då $\beta x \to -\infty$ så gäller att $\pi(x) \to 0$ samt när $\beta x \to \infty$ så gäller att $\pi(x) \to 1$.

Oddsen definieras som

$$\frac{\pi(x)}{1 - \pi(x)} = e^{\alpha + \beta x}.$$

Ekvivalent har logaritmen av oddsen (logit) ett linjärt samband

$$\text{logit}[\pi(x)] = \log \frac{\pi(x)}{1 - \pi(x)} = \alpha + \beta x.$$

Den logistiska regressionsmodellen besitter den viktiga egenskapen att sannolikheten $\pi(x)$ alltid är strikt mellan 0 och 1. Detta med anledning av den sökta länkfunktionen logit har egenskapen att logiten kan anta vilket reellt värde som helst, men $\pi(x)$ kommer alltid strikt vara mellan noll och ett [1, ss. 119–120].

2.3 Akaikes informationskriterium (AIC)

Akaikes informationskriterium (AIC) är mått på hur väl modellen passar ihop med datan (goodness of fit). AIC gör en avvägning mellan hur väl modellen passar till träningsdata och modellens komplexitet. AIC är en metod för modellval som bestraffar stora och komplicerade modeller.

Definitionen av AIC är

$$\text{AIC} = 2p - 2\log(L). \tag{1}$$

Här betecknar p antalet parametrar i en given modell och vidare är L likelihood-funktionens maximala värde för modellen. Den modell som enligt AIC är bäst är den modell som har det lägsta värdet på AIC. Däremot säger AIC i sig själv inte hur bra modellen är utan används för att jämföra olika modeller baserade på samma dataset. Notera att den logistiska regressionsmodellen ovan har $p = q + 1$ parametrar, interceptet och de q effektparametrarna [9, s. 284].

2.4 Bayesianska informatioskriteriet (BIC)

Det Bayesianska informatioskriteriet förkortat BIC, liknar AIC fast skillnaden är att BIC för stora datamaterial bestraffar modeller med fler antal parametrar hårdare.

BIC definieras enligt följande:

$$\text{BIC} = \log(n)p - 2\log(L). \tag{2}$$

Här svarar n mot antalet observationer i datamaterialet. Vidare är L likelihood-funktionens maximala värde för modellen, vilket gör att BIC uppmuntrar till mindre och enklare modeller än AIC, i synnerhet då $n > 7$. Detta med anledning av att 2 från definitionen av AIC (se ekvation (1)) byts ut mot $\log(n)$ (ekvation (3)) som faktor framför antalet parametrar. Det innebär att den valda modellen med BIC i allmänhet innehåller färre parametrar då $n > 7$ och därför är enklare än den modell som valts med AIC. Precis som för AIC är det modellen med lägst BIC som väljs [1, ss. 212–213].

2.5 Bakåtelimination och framåtselektion

Bakåtelimination (backward elemination) är en typ av stegvis variabelselektion. Bakåtelimination är en metod för modellval som grundar sig i utgångsantagandet att samtliga förklarande variabler tas med för att sedan successivt eliminera en variabel i taget tills processen avslutas, det vill säga då alla kvarvarande förklarande variabler är signifikanta. I de steg då flera variabler inte är signifikanta utesluts variabeln med det högsta p-värdet. Det innebär att i varje steg elimineras den variabel som bidrar minst till modellens anpassning till data.

Framåtselektion (forward selection) är även den en typ av stegvis variabelselektion som används för att hitta den bästa uppsättningen av prediktorer för att förutsäga responsvariabeln. Framåtselektion fungerar däremot på motsatt vis jämfört med bakåtelimination. Det vill säga att metoden startar med en modell som endast innehåller intercept och sedan läggs successivt en variabel i taget till, baserat på deras bidrag till modellens förmåga att förklara responsvariabeln. I varje steg inkluderas variabeln som är den mest signifikanta variabeln utanför modellen alltså den som ger den största förbättringen av modellanpassningen till data. Processen avslutas när ingen variabel utanför modellen ger ett signifikant bidrag till att prediktera responsvariabeln [9, ss. 93–94].

2.6 Hosmer–Lemeshow-testet

Hosmer–Lemeshow-testet är en metod för att bedöma hur väl en modellen passar med data. För ett givet antal skattade sannolikheter för olika händelser så rangordnas de skattade sannolikheterna i storleksordning. Hosmer-Lemeshow är en metod där man först anpassar modellen till data och sedan delar upp data i g grupper enligt deras skattade sannolikheter när de är storleksordnade. Vanligtvis delas de upp i $g = 10$ grupper baserat på de skattade sannolikheterna $P(Y = 1|x)$. Pearsons χ^2-statistika tillämpas för jämförelse mellan observerade och skattade värden. Utifrån statistikan

$$\sum_{i=1}^{g} \frac{\left(\sum_k y_{ik} - \sum_k \widehat{\pi}_{ik}\right)^2}{\sum_k \widehat{\pi}_{ik} \left(1 - \left(\sum_k \widehat{\pi}_{ik}\right)/n_i\right)},$$

där y_{ik} är utfallet för observation k i grupp i och där n_i är antalet observationer i grupp i. Under nollhypotesen att den anpassade logistiska regressionsmodellen är korrekt är detta en χ^2-fördelad statistika med $g-2$ frihetsgrader där g är antalet grupper [1, ss. 173–174].

Resultatet från detta approximativa χ^2-test kan användas för att avgöra om en viss modell är lämplig för att beskriva observerade data. Tolkningen av Hosmer-Lemeshow är att för p-värden under 0.05 anses modellen vara dålig. Däremot tyder högre p-värden endast på att det är mer troligt att det inte är en dålig modell. Det säger däremot inte nödvändigtvis att det behöver vara en bra modell.

Styrkan i ett test påverkas av storleken på det undersökta urvalet och kan ibland dölja eventuella brister som inte upptäcks. När antalet observationer ökar så förhöjs även risken för låga p-värdet trots att modellen ger en god anpassning till data. Detta kan leda till att alla modeller förkastas om datamaterialet är tillräckligt stort [10, ss. 147–152].

2.7 Förklaringsgrad (R^2)

Förklaringsgraden även benämnt R^2 används för att jämföra olika modeller. Man använder R^2 för att förklara hur stor del av variansen i responsvariabeln som förklaras av de oberoende variablerna. Detta gäller för linjär regression men för logistisk regression är däremot R^2 inte användbart. Istället för traditionella R^2 kommer därför två andra varianter av R^2 att användas. De två olika varianterna är McFadden R^2 samt Cox och Snell R^2 som båda grundar sig på maximum likelihood-anpassningar [9, ss. 90–91].

2.7.1 McFadden R^2

McFadden R^2 är ett väl använt mått på hur bra modellen överensstämmer med faktiska data för logistisk regression. Detta mått mäter hur väl en modell anpassar sig till data jämfört med en nollhypotesmodell. Nollhypotesmodellen är en modell som endast innehåller intercept. McFadden R^2 kan även benämnas pseudo R^2 och kan matematiskt formuleras som

$$R^2_{McF} = 1 - \frac{\log\left(L_{\text{modell}}\right)}{\log\left(L_{\text{null}}\right)}.$$

Här är L_{modell} likehooden för den anpassade modellen medan L_{null} är likelihooden för modellen med endast intercept. Modellen anses vara bättre desto större värdet är på McFadden R^2. Ett värde mellan 0.2–0.4 anses som en utmärkt modellanpassning [11, ss. 1–2].

2.7.2 Cox och Snell R^2

Cox och Snell R^2 kan även benämnas som generaliserad R^2 och är ett mått för att bedöma hur väl modellen passar till data. Även för detta mått tyder ett större värde på en bättre anpassad modell. Cox och Snell R^2 formuleras enligt

$$R^2_{CS} = 1 - \left(\frac{L_{\text{null}}}{L_{\text{modell}}}\right)^{\frac{2}{n}},$$

där n är antalet observationer i träningsdata [11, ss. 1–2].

2.8 Sammanblandningsmatris

Sammanblandningsmatris är ett verktyg som används för att på ett tydligt och rent visuellt vis utvärdera en modells klassificeringsförmåga (accuracy). I sammanblandningsmatrisen skapas en översikt av antalet fall som har klassificerats korrekt eller felaktigt för testdata. Prediktionen benämns lyckad (eller positiv) om den är 1 eller misslyckad (eller negativ) om den är 0. Utfallet av prediktionen beror på om den skattade sannolikheten av utfallet är större eller mindre än ett visst tröskelvärde c. I fallet då observation i i testdata med förklarande variabler x_i har skattad sannolikhet $\hat{\pi}_i = \hat{\pi}(x_i) > c$ ansätts $\widehat{y}_i = 1$ samt omvänt ansätts $\widehat{\pi}_i < c$ då $\widehat{y}_i = 0$ [1, ss. 223–224].

Sammanblandningsmatris är en tabell med de fyra möjliga elementen:

- True Positives (TP): Positiva fall korrekt klassificerade ($y_i = \hat{y}_i = 1$)
- False Positives (FP): Negativa fall felaktigt klassificerade ($y_i = 0$, $\hat{y}_i = 1$)
- True Negatives (TN): Negativa fall korrekt klassificerade ($y_i = \hat{y}_i = 0$)
- False Negatives (FN): Positiva fall felaktigt klassificerade ($y_i = 1$, $\hat{y}_i = 0$).

Utifrån detta är det möjligt att beräkna både specificiteten samt sensitiviteten. Det är även möjligt att utifrån informationen i tabellen beräkna proportionen av de korrekt klassificerade händelserna. Klassificieringsförmågan formuleras formellt enligt följande

$$\mathbb{P}(\text{Korrekt prediktion}) = \frac{TN + TP}{FN + FP + TN + TP}. \tag{3}$$

Detta värde föredras vara så högt som möjligt.

Vidare formuleras specificiteten samt sensitiviteten

$$\text{Specificiteten} = \frac{TN}{FP + TN}$$

$$\text{Sensitivitet} = \frac{TP}{FN + TP}.$$

Måttet sensitivitet bedömer prediktionsförmågan hos en modell eftersom måttet mäter förmågan att korrekt identifiera verkliga positiva resultat. Ju högre sensitivitet en modell har desto bättre är dess förmåga att identifiera verkliga positiva fall, vilket ger modellen bättre prediktionsförmåga.

Måttet specificitet är istället modellens förmåga att identifiera negativa resultat. Ju högre specificitet desto bättre är modellen på att identifiera verkliga negativa resultat vilket talar för en bättre modell [12, ss. 861–863].

2.9 Receiver operating characteristic (ROC-kurva)

ROC är en förkortning för receiver operating characteristic vilket är en kurva som används för utvärdering av hur precist en modell predikterar. ROC-kurvan är en graf där sensitiviteten plottas mot ett minus specificiteten för alla möjliga tröskelvärden c. ROC-kurvan kan anses vara mer informativ än sammanblandningsmatrisen i avsnitt 2.8 eftersom ROC-kurvan är baserad på varierande tröskelvärden. Om den estimerade sannolikheten för observation i i testdata är större än tröskelvärdet så predikteras fallet som positivt (händelsen $Y_i = 1$ inträffade). Om den estimerade sannolikheten istället är mindre än tröskelvärdet predikteras responsvariabeln istället som ett negativt fall (händelsen inträffade ej). Kurvan har vanligtvis en konkav form. Det optimala utfallet i grafen är att den går igenom punkten $(0, 1)$ eftersom modellen i detta fall predikterar korrekt för samtliga observationer i testdata [1, s. 224].

2.10 Arean under kurvan (AUC)

AUC betyder arean under ROC-kurvan och för att ROC-kurvan ska kunna användas som ett mått på hur precist en modell predikterar behövs AUC. Det är ett mått som beräknar arean under ROC-kurvan vilket är ett värde som ligger strikt i intervallet $[0, 1]$. Ett högt värde på AUC indikerar att modellen har god prediktionsförmåga. Det motsatta gäller även för låga värden på AUC. Då AUC ≤ 0.5 anses modellen vara dålig på att prediktera eftersom modellen således inte ger bättre resultat än att gissa slumpmässigt värdet på responsvariabeln för testdata. Detta visas i ROC-grafen med en referenslinje vilket är en rät linje och det motsvarar ett AUC-värde på 0.5. Enligt Hosmer och Lemeshow (2000) ges följande riktlinjer för hur olika AUC-värden ska tolkas: [10, ss. 160–163]

$$\begin{cases} \text{Om AUC} = 0.5, & \text{Modellen inte bättre än att gissa.} \\ \text{Om } 0.7 < \text{AUC} < 0.8, & \text{Modellen anses acceptabel.} \\ \text{Om } 0.8 < \text{AUC} < 0.9, & \text{Modellen anses mycket bra.} \\ \text{Om AUC} \geq 0.9, & \text{Modellen anses utmärkt.} \end{cases} \tag{4}$$

2.11 Residualavvikelse

Residualavvikelse (residual deviance) är ett mått på residualernas spridning som används för att bedöma hur väl modellen passar till träningsdata. För varje observation i från träningsdata, så definieras avvikelseresidualerna (deviance residual) som

$$r_i^D = \sqrt{d_i} \cdot \text{sign}(y_i - n_i\widehat{\pi}_i), \tag{5}$$

där

$$d_i = 2y_i \cdot \log\frac{y_i}{n_i\widehat{\pi}_i} + 2(n_i - y_i)\log\frac{n_i - y_i}{n_i - n_i\widehat{\pi}_i},$$

Här betecknar y_i binomialutfallet för utfall i av de förklarande variablerna där $i = 1, \ldots, N$, medan n_i är antalet observationer med värde x_i på kovariaterna. Vidare betecknar $\hat{\pi}_i = \hat{\pi}(x_i)$ det skattade värdet på sannolikheten $\pi(x_i) = P(Y = 1|x_i)$ för utfall x_i av de förklarande variablerna [1, ss. 215–216].

Vidare med hjälp av ekvation (5) fås den slutgiltiga definitionen av residualavvikelse. Den definieras som

$$\text{Residualavvikelse} = \sum_i (r_i^D)^2.$$

Residualavvikelse svarar approximativt mot en kvadratsumma av normaliserade avvikelser mellan responsvariablerna Y_i och de skattade medelvärdena

$n_i\hat{\pi}_i$, det vill säga ett mått på hur väl en modell anpassats till träningsdata. Vilket innebär att ett lägre värde på residualavvikelsen tyder på att modellen passar datan bättre samt det motsatta för ett högre värde [19, s. 205].

3 Data

3.1 Beskrivning av data

Datamaterialet som används är nedladdat som en CSV-fil från "Kaggle" där ägarna av datan ursprungligen är "Analytics Vidhya". Analytics Vidhya är en digital plattform för de som vill lära sig och utveckla sina färdigheter inom datavetenskap och maskininlärning. Plattformen grundades år 2013 och erbjuder en rad resurser för att främja lärande inom datavetenskap såsom artiklar, kurser, tävlingar och community-baserade evenemang. Idag är Analytics Vidhya en av de ledande plattformarna för lärande inom datavetenskap och maskininlärning i Indien [6].

Datamaterialet som används för att skapa modellen är baserad på totalt 614 olika individer. Vidare består datamaterialet av 11 förklarande variabler samt responsvariabeln Loan_Status. Dessutom finns ett testdataset för att mäta modellernas prediktionsförmåga på ny data, för 367 andra individer. Informationen har samlats in om individer i syfte att bedöma om en individ ska beviljas lån eller inte.

De variabler som är med i urvalet för de framtida modellerna ses i tabell 1 med en kort presentation av variablernas betydelse [7].

Tabell 1: Förklaring och översikt över variablerna som är inkluderade i datamaterialet för att ta fram modeller.

Beskrivning	**Variabler**	**Typ**	**Nivåer**
Lånesökandes kön	Gender	Nominal	0 = Kvinna 1 = Man
Om den lånesökande är gift	Married	Nominal	0 = Nej 1 = Ja
Antalet anhöriga till den lånesökande	Dependents	Diskret	0, 1, 2, ≥ 3
Om den lånesökande har en gymnasieexamen	Education	Nominal	0 = Nej 1 = Ja
Om den lånesökande är egenföretagare	Self_Employed	Nominal	0 = Nej 1 = Ja
Lånesökandes inkomst	ApplicantIncome	Kontinuerlig	150 – 81 000
Medsökandes inkomst	CoapplicantIncome	Kontinuerlig	0 – 41 700
Lånebelopp	LoanAmount	Kontinuerlig	9 – 700
Antalet månader till lånet ska vara helt avbetalat	Loan_Amount_Term	Diskret	12, 36, 60, 84, 120, 180, 240, 300, 360, 480
Lånesökandes kredit-historia	Credit_History	Nominal	0 = Dålig 1 = Bra
Lånesökandes boendeområde	Property_Area	Nominal	Urban, Semiurban, Rural
Om den lånesökande beviljas lånet	Loan_Status	Nominal	0 = Nej 1 = Ja

3.2 Bortfall

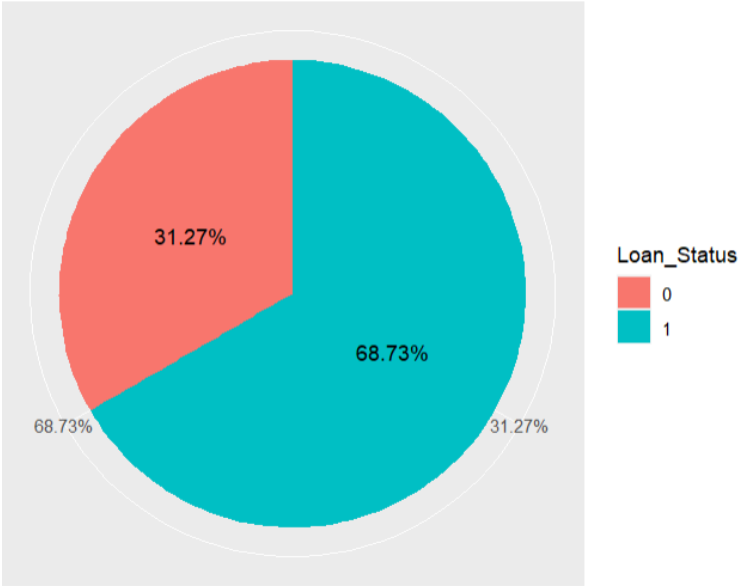
Bortfall eller avsaknad av data är ett vanligt problem inom dataanalys. Avsaknad av data kan bero på ett flertal olika anledningar. Det är viktigt att behandla bortfallet av data på ett systematiskt och lämpligt vis för att förhindra att resultaten blir snedvridna (biased). Bortfall kan resultera i en osäkerhet i de slutsatser som dras från analysen av data eftersom det är lätt att hantering av bortfall leder till bias det vill säga ett systematiskt fel. Detta gäller speciellt om andelen avsaknade data är stor i förhållande till den totala datastorleken.

Det finns ett stort antal olika metoder för att behandla avsaknad data på. De två metoderna som kommer att användas i denna dataanalys är uteslutning samt medelvärde- eller typvärdesimputation. Den första metoden är ett vanligt tillvägagångssätt nämligen att utesluta observationer med avsaknad data helt och hållet. Att endast använda observationer med komplett data kan däremot leda till missvisande resultat om det finns specifika samband inom bortfallet. För att denna metod ska kunna appliceras utan att leda till en bias skall bortfallet vara fullkomligt slumpmässigt det vill säga MCAR (missing completely at random) [15, ss. 173–177].

Den andra metoden för att hantera bortfall är medelvärde- eller typvärdesimputation. Metoden innebär att variabeln med bortfall ersätts av medelvärdet eller typvärdet för variabeln. För variablerna som är kontinuerliga ersätts avsaknad data med medelvärdet. Medelvärdet för variabeln med bortfall beräknas för samtliga de resterande observationer där det ej föreligger bortfall [15, ss. 173–177]. Vidare används typvärdesimputation för kategoriska variabler vilket innebär att bortfall av data ersätts med typvärdet för variabeln. Även imputation kan leda till missvisande resultat. För att denna behandlingsmetod av bortfall ej ska leda till bias ska bortfallet vara MCAR. Det betyder att orsakerna till bortfallet inte ska innehålla någon specifik relation eller systematik[16, ss. 1–3].

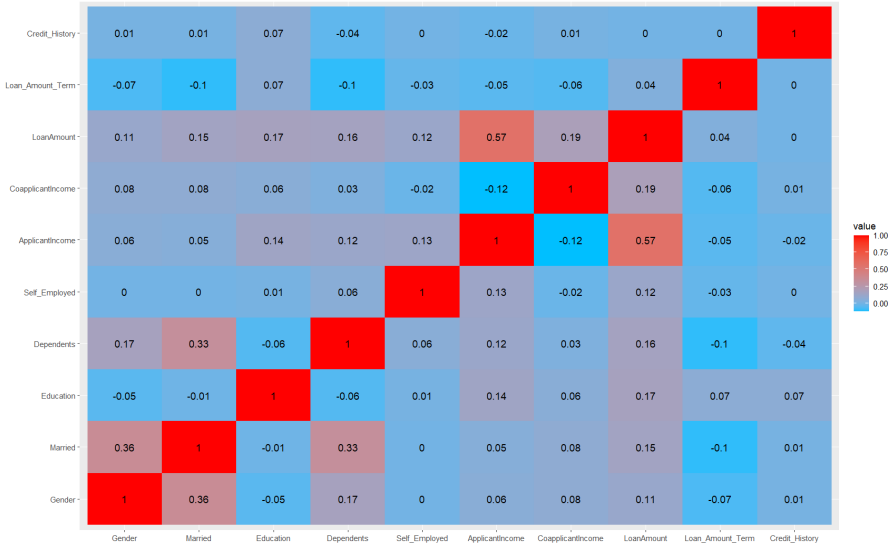
3.3 Explorativ dataanalys (EDA)

Explorativ dataanalys (exploratory data analysis) syftar till att ge en bättre förståelse av datamaterialet och avslöja insikter som kan användas för att fatta informerade beslut. Till att börja med presenteras andelen av antalet låneansökningar som beviljas samt avslås. Om Loan_Status är lika med 1 syftar det till att den lånesökande uppfyller bankens krav samt 0 till att den lånesökande inte uppfyller bankens krav.



Figur 1: Proportionen av låneansökningarna som möter bankens krav då bortfallet för Loan_Status är imputerat med medelvärde eller typvärde.

I figur 1 ses att en majoritet av de lånesökande mötte bankens krav. Totalt var det 68.73 procent av de lånesökande i detta datamaterial som uppfyllde bankens krav och därav fick sin låneansökan godkänd. Motsatt ser vi att det var 31.27 procent som ej mötte bankens kriterium och därav godkändes inte deras låneansökan.



Figur 2: Korrelationen mellan de förklarande variablerna då bortfallet är imputerat med medelvärde eller typvärde.

Från korrelationsmatrisen i figur 2 kan det fastställas att det inte finns någon anmärkningsvärt hög korrelation mellan någon av de förklarande variablerna då bortfallet är imputerat. Detta innebär att det inte råder någon anmärkningsvärd multikollinearitet eftersom variablerna är relativt oberoende av varandra. Detta möjligtvis med undantag av korrelationen 0.57 mellan ApplicantIncome och LoanAmount, men denna anses inte vara tillräckligt hög för att någon av dessa variabler ska uteslutas. Fördelen med att det inte föreligger någon anmärkningsvärd multikollinearitet är att risken för överanpassning minskar samt att precisionen i modellens parameterskattningar ökar. Dessutom är korrelationsmatrisen när bortfallet är uteslutet mycket lika. Med andra ord är det inga stora skillnader mellan korrelationsmatrisen då bortfallet är imputerat jämfört med uteslutet.

För senare modellering har Education, Married, Self_Employed och Loan_-Status gjorts om till binära variabler $\{0, 1\}$ då de endast antar två värden. Vidare behandlas variabeln Loan_Amount_Term som diskret och inte en kontinuerlig variabel. Detta eftersom variabeln inte antar ett stort antal olika värden då man vanligtvis får lån med en bestämd löptid. Denna variabel kommer därför fortsättningsvis att behandlas som en diskret variabel. Även Dependents kommer att behandlas som en diskret variabel med samma motivering som ovan.

Som nämnts ovan råder det bortfall av observationer i detta dataset. Det finns totalt sett 134 olika observationer med "NA", " " eller dylikt vilket motsvarar 21.82 procent av det totala antalet observationer i träningsdata. Enligt avsnitt 3.2 kommer vi att hantera bortfallet med två olika metoder. Först genom att utesluta observationer med avsaknad data samt sedan med medelvärde- eller typvärdes-imputation. För att applicera denna metod kommer cellen där datan är avsaknad att ersättas med typvärdet eller medelvärdet. Det vill säga bortfallet ersätts med värdet på den nivå av variabeln som förekommer oftast om det är en kategorisk variabel och medelvärdet om det är en kontinuerlig variabel. Även Loan_Amount_Term och Dependents kommer att behandlas som kategoriska variabler och ersättas med typvärdet, vilket för Loan_Amount_Term är 360 för både testdata och träningsdata. För samtliga kategoriska variabler är det en klar övervikt för typvärdet inom varje variabel vilket motiverar att det är rimligt att ersätta det avsaknade värdet med typvärdet. Om det är en kontinuerlig variabel ersätter vi istället bortfallet med medelvärdet. I testdatasetet är andelen observationer med någon avsaknad variabel 21.25 procent vilket motsvarar att bortfallet totalt påverkar 78 olika observationer. Även i detta datamaterial kommer avsaknade data ersättas på liknade vis enligt någon av dessa två olika metoder. Även för testdatasetet gäller att det är en klar övervikt för typvärdet för kategoriska variabler. Vilket återigen innebär att det är rimligt att ersätta bortfallet med typvärdet för den variabel som saknas.

4 Resultat och analys

Följande modellering tillämpar logistisk regression i programmet `R` med hjälp av den generaliserade linjära modellen (glm). Utifrån de givna variablerna i tabell 1 kommer modeller skapas för att prediktera om en låneansökan beviljas eller inte. Målet är att skapa en modell som inte är allt för komplicerad och som dessutom predikterar väl.

4.1 Modellframtagning

Fyra olika modeller kommer att tas fram, först och främst en modell med lägst AIC samt en med lägst BIC. Dessutom tas en modell fram med endast Credit_History som förklarande variabel eftersom tidigare analys gett insikten att Credit_History är den mest betydelsefulla förklarande variabeln, oavsett antalet parametrar i modellerna. Även en jämförelsemodell med samtliga förklarande variabler inkluderade, exklusive interaktionstermer kommer tas fram. Eftersom bortfallet i data är förhållandevis stort kommer avsaknaden av data att behandlas på två olika vis. Därav presenteras först modeller där bortfallet är imputerat med medelvärde eller typvärdet av variabeln beroende på om det är en kontinuerlig eller kategorisk variabel. Därefter presenteras modeller med bortfallet uteslutet.

Den stegvisa bakåtelimination startar med att samtliga förklarande variabler är inkluderade samt deras tvåvägs-interaktioner. Därefter tas i varje steg enligt avsnitt 2.5 någon variabel eller samspel bort från den ursprungliga mättade modellen. För att sedan fortsätta reducera modellen för att till sist välja modellen med lägst AIC respektive BIC. För att finna modellen med lägst AIC respektive BIC används `step` för att utföra både framåtselektion och bakåtelimination. Vi väljer att i varje steg gå vidare med den modell som har lägst AIC respektive BIC av framåtselektion och bakåtelimination, där bortfallet har ersatts med medelvärde eller typvärde för den saknade variabeln enligt avsnitt 3.2. Nedan presenteras ModellBIC i tabell 2 och ModellAIC i tabell 8 i appendix.

Tabell 2: ModellBIC är modellen som har lägst BIC då bortfallet är imputerat med antingen medelvärdet eller typvärdet för den saknade variabeln.

ModellBIC	**Skattning**	**Std.fel**	***p*-värde**
Intercept	-2.918	0.434	2e-11
Credit_History	3.857	0.414	<2e-16
Property_AreaSemiurban	0.912	0.263	0.001
Property_AreaUrban	0.247	0.249	0.322

Vidare, fortfarande med bortfallet imputerat, kommer en modell som endast innehåller Credit_History som förklarande variabel användas då det är den absolut mest inflytelserika variabeln. Denna modell benämns ModellCredit_History och dessutom tas modellen ModellFörstagradstermer med som en jämförelsemodell. ModellCredit_History presenteras nedan i tabell 3 medan ModellFörstagradstermer presenteras i tabell 9 i appendix.

Tabell 3: ModellCredit_History är modellen med endast Credit_History som förklarande variabel då bortfallet är imputerat med antingen medelvärdet eller typvärdet för den saknade variabeln.

ModellCredit_History	**Skattning**	**Std.fel**	***p*-värde**
Intercept	-2.461	0.394	4e-10
Credit_History	3.789	0.408	<2e-16

Både ModellCredit_History och ModellFörstagradstermer visas nu istället då bortfallet är uteslutet (se avsnitt 3.2). Modellerna med lägst AIC och BIC tas fram på liknande sätt som tidigare fast nu istället med denna metod för att behandla bortfallet. Där ModellCredit_History visas nedan i tabell 4 samt ModellBIC i tabell 5, båda med bortfallet uteslutet.

Tabell 4: ModellCredit_History är modellen med endast Credit_History som förklarande variabel då observationer med bortfall är uteslutna.

ModelCredit_History	**Skattning**	**Std.fel**	***p*-värde**
Intercept	-2.197	0.398	4e-8
Credit_History	3.538	0.417	<2e-16

Tabell 5: ModellBIC är modellen som har lägst BIC med observationer med bortfall är uteslutna.

ModelBIC	**Skattning**	**Std.fel**	***p*-värde**
Intercept	-2.692	0.449	2e-09
Credit_History	3.610	0.424	<2e-16
Property_AreaSemiurban	0.985	0.294	0.0008
Property_AreaUrban	0.249	0.2837	0.380

Vidare visas ModellFörstagradstermer i tabell 11 samt ModellAIC i tabell 10 i appendix då bortfallet är uteslutet.

4.2 Modelljämförelse

Nu jämförs de fyra olika framtagna modellerna med de två olika metoderna för att behandla bortfallet. De modeller som tagits fram är bland annat den enklaste modellen ModellCredit_History som endast har en förklarande variabel Credit_History. Sedan har vi ModellFörstagradstermer där samtliga förklarande variabler är inkluderade exklusive interaktionstermer. Vidare har ModellAIC samt ModellBIC tagits fram för att minimera AIC respektive BIC. I tabell 6 presenteras ett antal mått för att jämföra modellernas prediktiva förmåga på testdata samt hur väl de anpassar sig till träningsdata, där

bortfallet är imputerat med medelvärde eller typvärde för de kontinuerliga respektive kategoriska variablerna.

Tabell 6: Fyra modeller har anpassats till träningsdata. Bortfallet i träningsdata har ersatts med medelvärdet eller typvärdet för variabeln beroende på om variabeln är kontinuerlig eller kategorisk. Vidare är de avsaknade värdena ersatta på samma vis för testdata vilket används för att jämföra prediktionsförmågan hos de olika modellerna med kriterierna AUC, klassificeringsförmåga, sensitivitet och specificitet. Där tröskelvärdet $c = 0.5$ har använts för att beräkna klassificeringsförmåga, sensitivitet och specificitet. I tabellen jämförs de olika modellerna utifrån olika jämförelsemått, där de resterande icke-prediktiva kriterierna beräknas med hjälp av träningsdata.

Test	**ModellAIC**	**ModellBIC**	**Modell Första-ordningen**	**Modell Credit_ History**
AIC	**563.417**	582.711	587.420	592.024
BIC	673.917	**600.391**	644.880	600.864
CoxSnell R^2	**0.334**	0.264	0.280	0.248
MCFadden R^2	**0.327**	0.247	0.264	0.229
AUC	0.950	0.919	**0.969**	0.940
Hosmer-Lemeshow p-värde	0.127	**1.000**	0.301	**1.000**
Klassificierings-förmåga	0.929	**0.978**	**0.978**	**0.978**
Sensitivitet	0.866	**0.881**	**0.881**	**0.881**
Specificitet	0.943	**1.000**	**1.000**	**1.000**
Residual-avvikelse	**513.420**	574.710	561.420	588.020

Utifrån tabell 6 kan det avläsas att ModellAIC respektive ModellBIC har lägsta värdena på AIC respektive BIC. Vilket är som förväntat eftersom modellerna tagits fram med syftet att minimera AIC respektive BIC.

Vidare ser vi i tabell 6 ovan att ModellAIC är modellen som får de högsta värdena på både McFadden R^2 samt Cox och Snell R^2. Vilket indikerar en bättre modell enligt detta mått då de bättre anpassar sig till träningsdata. Dessa två olika varianter av R^2 antyder att ModellAIC är något bättre än de övriga för detta mått (även om detta delvis kan bero på att R^2 premierar större modeller som tenderar att överanpassa till träningsdata). Eftersom ett McFadden R^2 värde i intervallet 0.2–0.4 anses som en utmärkt modellanpassning kan slutsatsen dras att samtliga modeller anpassar sig väl till träningsdata (se avsnitt 2.7.1).

Det är ModellFörstagradstermer som har högst AUC vilket tyder på att det är modellen med bäst prediktionsförmåga. Skillnaderna mellan modellerna kan inte anses som speciellt stora utan samtliga modeller har en hög prediktiv förmåga. Detta med anledning av att samtliga modeller har AUC som ligger i intervallet 0.9–1 vilket enligt avsnitt 2.10 (ekvation (4)) anses som utmärkt. Detta kan avläsas från ROC-kurvorna samt AUC för de olika modellerna vilket kan även ses i figur 4 och figur 5 i appendix.

Från tabell 6 kan även Hosmer-Lemeshow p-värdena för de olika modellerna avläsas. Där ser vi att samtliga modeller har p-värden över 0.05 vilket tyder på att ingen av modellerna klassas som dåliga enligt Hosmer-Lemeshow. Både ModellBIC och ModellCredit_History har höga p-värden vilket kan anses vara fördelaktigt. Enligt avsnitt 2.6 tyder högre p-värden på att det är mindre troligt att det är en dålig modell. Det påvisar dock inte att det är en bra modell vilket gör att detta mått inte bidrar i lika hög grad till valet av modell.

		Prediktion	
		0	**1**
Utfall	**0**	58	17
	1	9	283

(a) ModellAIC

		Prediktion	
		0	**1**
Utfall	**0**	59	0
	1	8	300

(b) ModellBIC

		Prediktion	
		0	**1**
Utfall	**0**	59	0
	1	8	300

(c) ModellFörstagradstermer

		Prediktion	
		0	**1**
Utfall	**0**	59	0
	1	8	300

(d) ModellCredit_History

Figur 3: Sammanblandningsmatriser för samtliga modeller där bortfallet är imputerat

Klassificieringsförmågan från avsnitt 2.8 anger den korrekta andelen klassificeringar (ekvation (3)), vilket är en tydlig indikation på hur väl modellerna predikterar. Utifrån tabell 6 kan det utläsas att samtliga modeller förutom ModellAIC har en klassificieringsförmåga på 0.978, vilket motsvarar 2.2 procent felaktiga prediktioner. Sammanblandningsmatriserna för testdatan då bortfallet är imputerat med medelvärdet eller typvärdet ses i sammanblandningsmatrisen i figur 3 för samtliga modeller. För samtliga modeller förutom ModellAIC är det åtta felaktiga prediktioner som inte överensstämmer med det faktiska utfallet gällande om en lånesökande fick lånet beviljat eller inte. Vi ser därtill att samtliga modeller har både hög sensitivitet samt specificitet och de kan därför anses vara modeller med god klassificeringsförmåga.

Från tabell 6 kan det dessutom utläsas att ModellAIC har lägst residualavvikelse vilket enligt avsnitt 2.11 tyder på att det är den modell som passar träningsdata bäst. Detta eftersom ett lägre värde på residualavvikelsen visar på att modellen förklarar datan bättre. Problemet med residualavvikelse baserat på träningsdata är att det är ett mått som alltid har lägre värde för större modeller eftersom residualavvikelsen inte kan bli högre genom att lägga till förklarande variabler till en modell. Därav ser vi att det är den största modellen med flest parametrar som har lägst residualavvikelsen. För de resterande mindre modellerna är residualavvikelsen förhållandevis lika vilket tyder på att modellerna passar datan approximativt lika väl. De två R^2-kriterierna ger också högst värden för ModellAIC. Som nämnts ovan är detta inte förvånande, eftersom dessa kriterier också utgår från träningsdata och därmed är känsliga för överanpassning, så att den större ModellAIC premieras.

Utifrån tabell 6 samt figur 4 och figur 5 i appendix kan vi se att olika jämförelsemått ger olika bästa modeller, med ModellAIC, ModellBIC och ModellCredit_History som vinnare för de flesta kriterierna. Med anledning av att ModellCredit_History är den enklaste modellen som ger tämligen lika värden på testerna som de andra modellerna, så överväger enkelheten hos ModellCredit_History vissa av de små fördelar som de komplexare modellerna besitter. Detta medför att vi kan använda oss av en princip från Sundberg [9, s. 89] ”Om en enkel och en mer komplicerad förklaringsmodell passar lika bra, så skall man nog hysa störts tilltro till den enklaste.” Av det följer att ModellCredit_History väljs som den bästa modellen med anledning av att den har god prediktiv förmåga samt passar datamaterialet väl, samtidigt som det är en enkel modell då den endast har med en förklarande variabel.

Nu jämförs de fyra olika framtagna modellerna då bortfallet är uteslutet, där ModellCredit_History, ModellFörstagradstermer och ModellBIC består av samma parametrar som tidigare med skillnaden att datan som nu används har observationer, med någon saknad variabel, borttagna. Nu används ute-

slutet bortfall för både testdata och träningsdata vilket ger andra skattade parametrar samt en annan prediktionsförmåga. Vidare skiljer sig ModellAIC en aning från tidigare då bortfallet var imputerat vilket kan ses i tabell 8 och tabell 10 i appendix. Tabell 7 presenteras nedan med ett antal mått för att jämföra de olika modellernas prediktiva förmåga på testdata samt hur väl de anpassar sig till träningsdata när bortfallet är uteslutet.

Tabell 7: Fyra modeller har anpassats till träningsdata där observationer med saknade förklarande variabler är uteslutna. Vidare är de avsaknade värdena uteslutna på samma vis för testdata vilket används för att jämföra prediktionsförmågan hos de olika modellerna med kriterierna AUC, klassificeringsförmåga, sensitivitet och specificitet. Där tröskelvärdet $c = 0.5$ har använts för att beräkna klassificeringsförmåga, sensitivitet och specificitet. I tabellen jämförs de olika modellerna utifrån olika jämförelsemått, där de resterande icke-prediktiva kriterierna beräknas med hjälp av träningsdata.

Test	**ModellAIC**	**ModellBIC**	**Modell Första-ordningen**	**Modell Credit_ History**
AIC	**445.424**	459.380	463.973	468.023
BIC	558.116	**476.075**	518.232	476.371
CoxSnell R^2	**0.343**	0.256	0.276	0.236
MCFadden R^2	**0.341**	0.239	0.262	0.218
AUC	0.905	0.935	**0.978**	0.951
Hosmer-Lemeshow p-value	0.771	**1.000**	0.154	**1.000**
Klassificierings-förmåga	0.893	**0.983**	0.976	**0.983**
Sensitivitet	0.863	**0.902**	**0.902**	**0.902**
Specificitet	0.899	**1.000**	0.992	**1.000**
Residual-avvikelse	**391.420**	451.380	437.970	464.020

Från tabell 7 samt figur 6 och figur 7 i appendix kan det utläsas att även med denna metod för behandling av bortfallet presterar modellerna bra för samma jämförelsemått som då saknade förklarande variabler imputerats. Slutsatsen som dras från tabellen är att båda metoderna för att hantera bortfallet ger samma bästa modell för de flesta av jämförelsemåtten. Därav väljs ModellCredit_History även då de avsaknade värdena är uteslutna med samma motivering som ovan. Nämligen att om flera modeller presterar likvärdigt så väljs den enklaste.

Vi har imputation som grundmetod för behandling av bortfallet. När vi tolkar resultaten ser vi att det inte är stora skillnader kopplat till hur väl modellerna predikterar för de två metoderna imputation samt uteslutning av bortfallet. Om det hade varit stor skillnad mellan hur väl modellerna predikterade testdata hade det inneburit att metoden för hur bortfallet hanterats har stor påverkan. Med anledning av att det är marginella skillnader mellan tabell 6 och tabell 7 leder det till att behandlingen av de avsaknade värdena inte har en påtaglig inverkan på resultatet. Däremot bör resultaten baserade på imputerade bortfall vara något mer exakta, även om det från tabell 6 och tabell 7 indikerar att båda analyserna ger liknande resultat. Eftersom imputation anses vara en bättre metod för att hantera avsaknade data så bör den användas för datamaterialet i detta arbete, speciellt när det är en förhållandevis stor andel variabelvärden som saknas. Fördelarna med imputation kontra uteslutning av observationer med avsaknade värden är att risken för snedvridning av resultaten minskar. Ytterligare en positiv faktor är att då inga observationer utesluts blir den statistiska styrkan högre. Det är därför ModellCredit_History med imputerat bortfall väljs som slutgiltig modell och som kommer att användas fortsättningsvis.

5 Slutsats

Slutsatserna som kan dras från resultatet av modellanpassningarna är att samtliga modeller som presenterats skulle kunna användas i prediktivt syfte. Det är även intresseväckande hur marginella skillnader det är mellan den mycket mer komplexa ModellAIC och den enkla ModellCredit_History. Precis som nämnts i resultatavsnittet är det ModellCredit_Historys enkelhet samt utmärkta prediktionsförmåga som är avgörande. Detta är anledningen till att den slutgiltigt valda modellen är just ModellCredit_History med bortfallet imputerat. Bortfallet imputerat blir alltså den slutgiltiga metoden för behandling av avsaknade data. Som tidigare nämnts beror det på att det är en mer robust och tillförlitlig metod. Detta kan motiveras med att imputation minskar risken för bias samt ger högre statistisk styrka jämfört med uteslutning av bortfall.

Från figur 3 ModellCredit_History kan det avläsas hur utomordentligt väl den slutgiltiga modellen predikterar testdata. Från sammanblandningsmatrisen ses att endast åtta fall klassificeras felaktigt av de totalt 367 fallen. Detta motsvarar 2.2 procent felaktigt predikterade låneansökningar vilket tyder på att modellen besitter en mycket god prediktiv förmåga för att avgöra om en låneansökan ska godkännas eller nekas. Detta kan dessutom ses från AUC samt ROC-kurvan i figur 5 som tyder på samma sak.

Analysen har dessutom bidragit till förståelsen att det är Credit_History som är den klart mest inflytelserika förklarande variabeln för att avgöra om en låneansökan blir beviljad eller inte. För alla testade modeller oavsett storlek var i samtliga fall Credit_History den mest signifikanta förklarande variabeln. Av detta följer att när ett kreditinstitut ska avgöra om en individ ska beviljas ett givet lån behövs i stort sett endast information om den lånesökandes kredithistoria. Om den lånesökande har en god kredithistoria blir individen accepterad lånet enligt ModellCredit_History samt i annat fall nekad, vilket onekligen är en mycket enkel modell. Därav lär inte användningen av denna modell vara speciellt stor hos kreditinstitut men utifrån datan som används i denna analys blir slutsatsen trots detta att modellen predikterar mycket väl om en låneansökan ska beviljas eller inte. Detta visades både genom ett högt AUC och en god klassificeringsförmåga vilket tyder på att det är en mycket träffsäker modell. Slutsatsen som kan dras från detta är att det viktigaste för privatpersoner är att ha en god kredithistorik för att deras låneansökan ska godkännas. Således är det viktigt att du betalar dina räkningar i tid och inte innehar för många samt för stora skulder för att förbättra dina chanser att få en låneansökan beviljad.

6 Diskussion

I allmänhet är modeller förenklingar av verkligheten och frågan är om den valda modellen är alltför förenklad då den endast består av en förklarande variabel. En förklaring till att modellen endast består av den förklarande variabeln kredithistoria kan bero på hur fallissemang är definierat. Eftersom sannolikheten för fallissemang ligger till grund för bedömning om en låneansökan ska beviljas eller avslås. Därav har den underliggande definitionen för hur man definierar fallissemang en påverkan på prediktionen om en låneansökan ska beviljas eller avslås. Om fallissemang är definierat som att en betalning inte inträffat inom en viss tid efter betalningen skulle varit utförd har det stora likheter med kredithistora. Om däremot definitionen är att låntagaren helt och hållet misslyckas utföra de avtalade betalningarna är det mer troligt att även andra ytterligare faktorer torde ha en större påverkan. Detta kan alltså vara en förklaring till att den slutgiltiga modellen endast består av en förklarade variabel.

Slutsatsen som drogs är att samtliga modeller visar på en utmärkt prediktionsförmåga. Däremot kan storleken på datasetet samt storleken på bortfallet vara något att ha i åtanke vid tolkning av resultaten. Det hade därav varit intressant att ha ett större träningsdataset för att möjligen kunna öka tillförlitligheten hos modellanpassningen, eftersom större dataset bidrar till en högre statistisk styrka för korrekta prediktioner för tesdata. Det är svårt

att få tag i bra dataset, men ett större dataset som innehåller en lägre andel avsaknade värden hade varit intressant att undersöka. Som nämnts i avsnitt 3.3 är bortfallet i träningsdata ungefär 20 procent, vilket är en stor andel. Därav kan bortfallet ha en påverkan på resultatet i denna undersökning. Risken för detta har vi dock försökt att minimera genom att använda olika metoder för behandling av bortfallet. Trots detta går det däremot inte att utesluta att bortfallet kan påverka resultatens tillförlitlighet. Detta med anledning av att bortfall kan leda till bias om bortfallet är systematiskt. Detta kan inträffa om det finns specifika samband eller mönster inom bortfallet som leder till att datan inte representerar hela populationen. Sådana situationer uppstår exempelvis om bortfallet inte är slumpmässigt eftersom det är svårt att veta anledningarna till varför bortfallen uppstod. Däremot är det också viktigt att notera att bortfallet inte nödvändigtvis behöver ha påverkat resultaten. Dock bör man ha bortfallet i åtanke när man tolkar resultaten.

Dessutom skulle det vara intressant att prova ytterligare metoder för att hantera de avsaknade värdena på. Detta för att se om det finns bättre metoder för att hantera bortfallet samt se hur resultaten förändras om dessa metoder används.

Därtill skulle det vara intressant för framtida analyser att undersöka modeller med ytterligare förklarande variabler som inte ingick i detta dataset, för att se om det finns flera variabler som kan vara inflytelserika. Exempel på sådana variabler skulle kunna vara syftet bakom lånet samt ålder vilket även nämns i Ali (2015), där den bakomliggande orsaken till lånet ger ett signifikant bidrag till modellanpassningen. Det hade alltså varit intressant att se om dessa förklarande variabler skulle kunna hjälpa till och bidra till en ännu precisare modell som noggrannare kan förutspå om en individ får en låneansökan beviljad eller inte [18, s. 12–22].

För framtida forskning skulle det dessutom vara intressant att jämföra hur väl logistisk regression presterar jämfört med andra metoder som nämnts i avsnitt 1.2. Enligt Tumuluru et al. (2022) och Kansal et al. (2021) framhävs random forest som ytterligare en bra metod för att prediktera om en låneansökan ska beviljas eller avslås och det skulle vara intressant att jämföra denna metod med logistisk regression. Dessutom skulle de metoder som enligt Louzada et al. (2016) är de bästa för att prediktera låneförfall, nämligen stödvektormaskin och suddig logik, vara intressanta att applicera. Detta eftersom prediktion av låneförfall samt prediktion kopplat till om en låneansökan accepteras eller nekas är likartade problemställningar. Därför kan man förvänta sig att även dessa metoder bör kunna prestera väl. Att applicera flera olika metoder för att prediktera om en låneansökan beviljas eller inte är ett intressant område för vidare forskning för att se hur väl logistisk

regression står sig gentemot andra och mer avancerade metoder.

Liknande frågeställningar som behandlats i denna rapport för privatpersoner har även gjorts för företag. Ett sådant exempel är Carlehed och Petrov (2012) som tar upp hur nationalekonomiska faktorer påverkar sannolikheten för fallissemang. Detta är en mycket värdefull insikt eftersom de ekonomiska cyklerna har en stor påverkan på sannolikheten för fallissemang. Detta är relevant då sannolikheten för fallissemang ligger till grund för om en låneansökan beviljas eller inte. För framtida forskning skulle även detta vara spännande att applicera, inte bara för företag utan även för privatpersoners sannolikhet för fallissemang. Jag skulle tro att ekonomiska cykler har en påverkan även för privatpersoner, men det är något att undersöka noggrannare i framtida forskning [17, ss. 7–10].

Referenser

[1] Agresti, A. (2013) *Categorical Data Analysis* (Tredje upplagan), Wiley.

[2] Louzada, F., et al. (2016) "Classification methods applied to credit scoring: Systematic review and overall comparison", *Surveys in Operations Research and Management Science*, volym 21.

[3] Hand, D.J. och Henley, W.E. (1997) "Statistical Classification Methods in Consumer Credit Scoring: a Review", *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, volym 160.

[4] Tumuluru, P., et al. (2022) "Comparative Analysis of Customer Loan Approval Prediction using Machine Learning Algorithms," *Second International Conference on Artificial Intelligence and Smart Energy (ICAIS).*

[5] Kansal, V., et al. (2021) "Home Loan Prediction Using Machine Learning Models", *Elementary Education Online*, volym 20.

[6] Analytics vidhaya (2023) *About Analytics Vidhya.* Hämtad 2023-03-07 från `https://www.analyticsvidhya.com/about/`.

[7] Kaggle (2019) *Loan Prediction Problem Dataset.* Hämtad 2023-03-07 från `https://www.kaggle.com/datasets/altruistdelhite04/loan-prediction-problem-dataset`.

[8] Basel Committee on Banking Supervision (2000) *Principles for the management of credit risk.* Hämtad 2023-04-12 från `https://www.bis.org/publ/bcbs75.pdf`.

[9] Sundberg, R. (2022) *Lineära stokastiska modeller*, Stockholms Universitet.

[10] Hosmer, D.W. och Lemeshow, S. (2000) *Applied Logistic Regression* (Andra upplagan), Wiley.

[11] Allison, P.D. (2014) "Measures of fit for logistic regression", *Statistical Horizons LLC and the University of Pennsylvania.*

[12] Fawcett, T. (2006) "An introduction to ROC analysis", *Pattern Recognition Letters*, volym 27.

[13] James, G., et al. (2013) *An Introduction to Statistical Learning*, Springer.

[14] Hyndman, R.J. och Athanasopoulos, G. (2018) *Forecasting: principles and practice* (Andra upplagan), Otexts.

[15] Reuterberg, S.E. (2001) ”Hantering av bortfall i longitudinella studier: Ett exempel”, *Pedagogisk Forskning i Sverige*, volym 6.

[16] Sessa, J. och Syed, D. (2016) ”Techniques to deal with missing data”, *2016 5th International Conference on Electronic Devices.*

[17] Carlehed, M. och Petrov, A. (2012) ”A methodology for point-in-time–through-the-cycle probability of default decomposition in risk classification systems”, *Journal of Risk Model Validation*, volym 6.

[18] Ali, M. (2015) *Performance of Three Classification Techniques in Classifying Credit Applications Into Good Loans and Bad Loans: A Comparison*, Statistiska instutionen Uppsala universitet.

[19] Chambers, J.M. och Hastie, T.J. (1992) *Statistical Models in S*, Wadsworth & Brooks/Cole.

Appendix

Tabell 8: ModellAIC är modellen som har lägst AIC då bortfallet är imputerat, antingen med medelvärdet eller med typvärdet för den avsaknade förklarande variabeln.

ModellAIC	**Skattning**	**Std.fel**	***p*-värde**
Intercept	-9.185	3.173	0.004
Credit_History	3.857	0.483	<2e-16
Gender	3.246	2.105	0.123
Married	0.304	0.277	0.273
Dependents	-0.252	0.187	0.178
Education	4.659	1.659	0.005
Self_Employed	-25.783	520.067	0.961
ApplicantIncome	0.002	0.001	0.010
CoapplicantIncome	0.001	2e-4	0.010
LoanAmount	-0.048	0.017	0.005
Loan_Amount_Term	0.017	0.008	0.051
Property_AreaSemiurban	0.730	0.448	0.103
Property_AreaUrban	-0.640	0.482	0.184
Gender:Loan_Amount_Term	-0.009	0.006	0.125
Married:Self_Employed	1.122	0.729	0.124
Dependents:ApplicantIncome	6e-5	3e-5	0.021
Education:CoapplicantIncome	-6e-4	2e-4	0.005
Education:Loan_Amount_Term	-0.010	0.005	0.027
Self_Employed:ApplicantIncome	3e-4	1e-4	0.012
Self_Employed:Loan_Amount_Term	0.020	0.008	0.010
Self_Employed:Credit_History	16.585	520.054	0.975
ApplicantIncome:Loan_Amount_Term	-6e-6	2e-6	0.005
ApplicantIncome:Property_AreaSemiurban	5e-5	7e-5	0.433
ApplicantIncome:Property_AreaUrban	2e-4	9e-5	0.036
LoanAmount:Loan_Amount_Term	1e-4	5e-5	0.008

Tabell 9: ModellFörstagradstermer är modellen med samtliga förklarade variabler inkluderade exklusive samspel då bortfallet är imputerat med antigen medelvärdet eller typvärdet för den avsaknade förklarande variabeln.

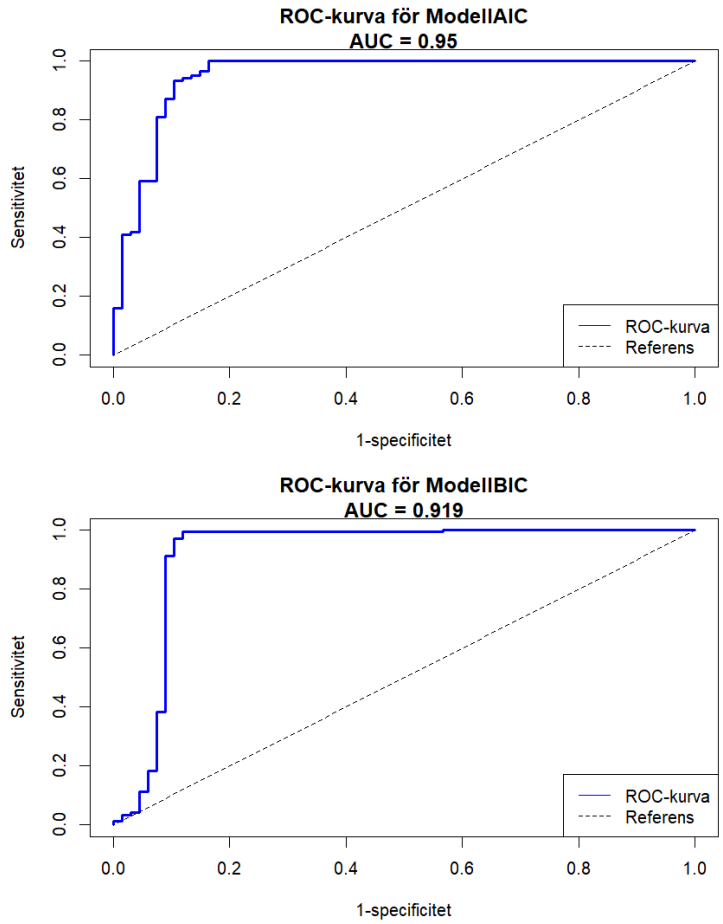
ModellFörstagradstermer	**Skattning**	**Std.fel**	***p*-värde**
Intercept	-2.971	0.837	4e-4
Credit_History	3.901	0.418	<2e-16
Gender	0.009	0.297	0.975
Married	0.556	0.248	0.025
Dependents	0.041	0.117	0.724
Education	0.417	0.259	0.107
Self_Employed	-0.030	0.315	0.924
ApplicantIncome	1e-5	2e-5	0.640
CoapplicantIncome	-5e-5	3e-5	0.161
LoanAmount	-0.002	0.002	0.200
Loan_Amount_Term	-9e-4	0.002	0.606
Property_AreaSemiurban	0.863	0.267	0.001
Property_AreaUrban	0.174	0.256	0.497

Tabell 10: ModellAIC är modellen som har lägst AIC, då observationer med avsaknade förklarande variabler är uteslutna.

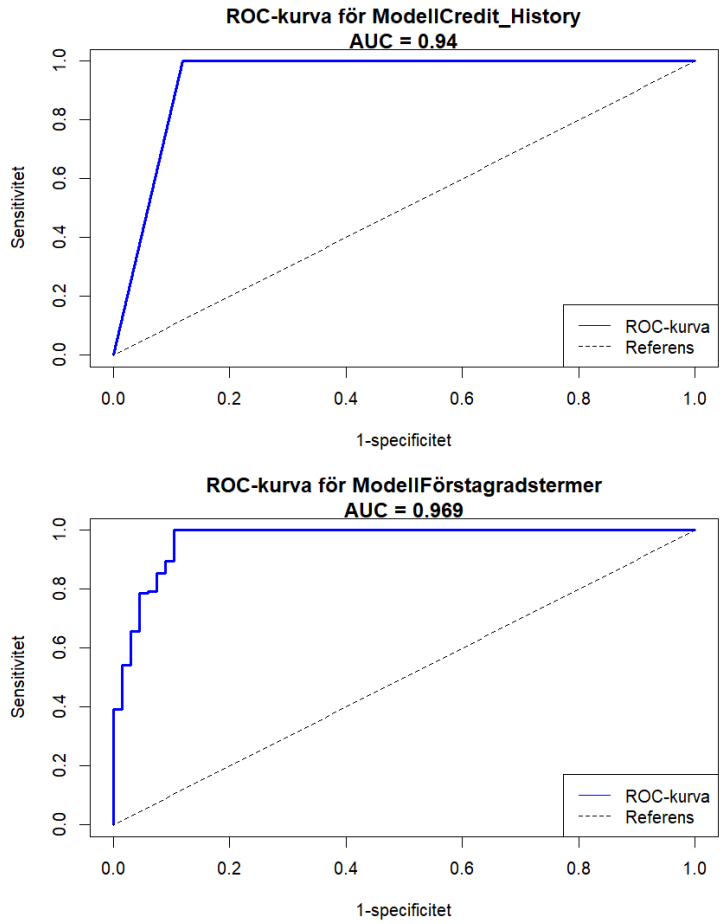
ModelAIC	**Skattning**	**Std.fel**	***p*-värde**
Intercept	-10.376	3.981	0.009
Credit_History	4.681	0.713	5.2e-11
Gender	-0.595	0.744	0.424
Married	5.555	2.526	0.028
Dependents	-2.780	1.182	0.019
Education	0.142	0.719	0.844
Self_Employed	-26.347	919.092	0.977
ApplicantIncome	0.002	8e-4	0.026
CoapplicantIncome	4e-5	5e-4	0.931
LoanAmount	0.002	0.004	0.581
Loan_Amount_Term	0.016	0.011	0.130
Property_AreaSemiurban	1.433	2.333	0.539
Property_AreaUrban	6.064	2.808	0.031
Gender:Education	1.339	0.823	0.104
Married:Loan_Amount_Term	-0.015	0.007	0.039
Dependents:Loan_Amount_Term	0.008	0.003	0.016
Education:CoapplicantIncome	-0.001	3e-4	0.008
Self_Employed:LoanAmount	0.016	0.006	0.005
Self_Employed:Loan_Amount_Term	0.019	0.009	0.035
Self_Employed:Credit_History	17.248	919.084	0.985
ApplicantIncome:LoanAmount	-7e-7	4e-7	0.051
ApplicantIncome:Loan_Amount_Term	-5e-6	2e-6	0.037
CoapplicantIncome:LoanAmount	-1e-6	7e-5	0.049
CoapplicantIncome:Loan_Amount_Term	3e-6	1e-6	0.015
CoapplicantIncome:Credit_History	-3e-4	2e-4	0.097
Loan_Amount_Term:Property_AreaSemiurban	-0.001	0.007	0.873
Loan_Amount_Term:Property_AreaUrban	-0.017	0.008	0.033

Tabell 11: ModellFörstagradstermer är modellen med samtliga förklarade variabler inkluderade exklusive samspel då observationer med avsaknade förklarande variabler är uteslutna.

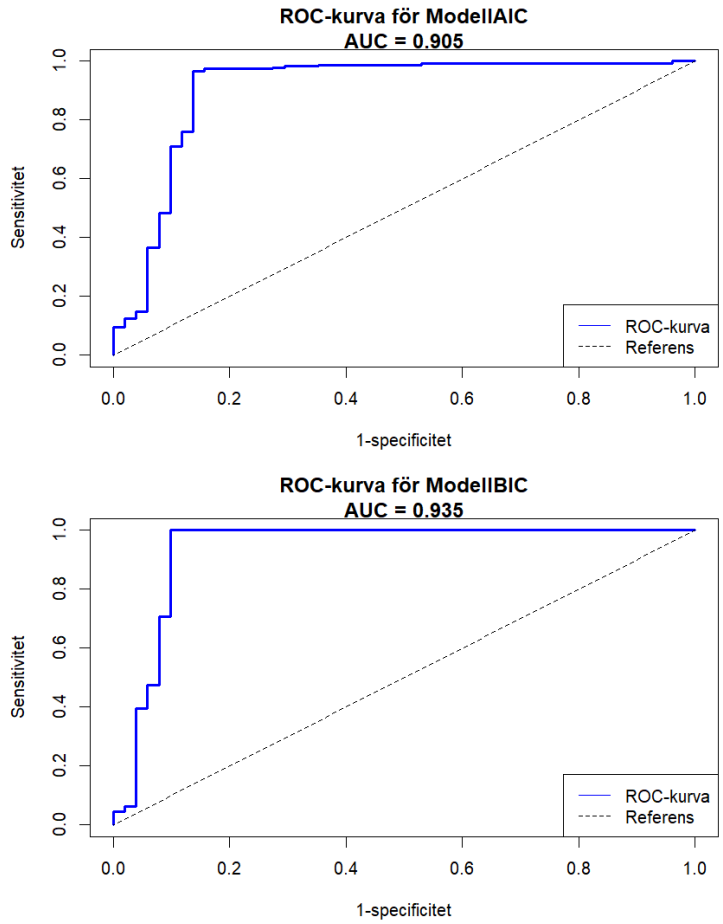
ModelFörstagradstermer	**Skattning**	**Std.fel**	***p*-värde**
Intercept	-2.963	0.918	0.001
Credit_History	3.632	0.431	<2e-16
Gender	0.371	0.329	0.259
Married	0.525	0.286	0.066
Dependents	0.083	0.134	0.539
Education	0.424	0.303	0.161
Self_Employed	-0.152	0.350	0.665
ApplicantIncome	5e-6	3e-5	0.853
CoapplicantIncome	-5e-5	4e-5	0.265
LoanAmount	-0.003	0.002	0.110
Loan_Amount_Term	-6e-4	0.002	0.757
Property_AreaSemiurban	0.949	0.301	0.002
Property_AreaUrban	0.099	0.295	0.738



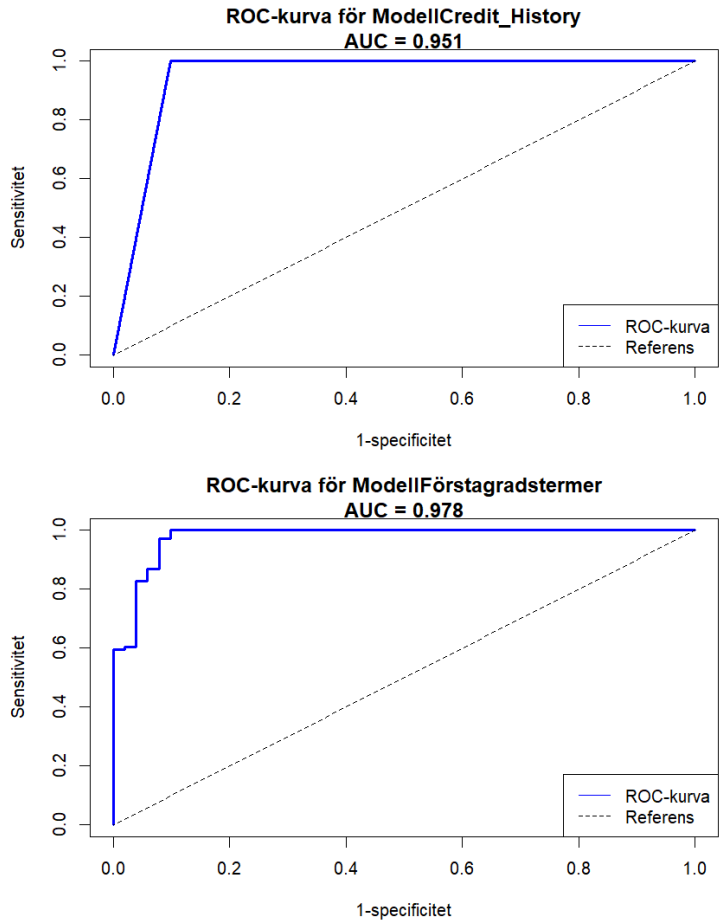
Figur 4: ROC-kurvan samt AUC för ModellAIC och ModellBIC med bortfallet imputerat.



Figur 5: ROC-kurvan samt AUC för ModellCredit_History och ModellFörstagradstermer med bortfallet imputerat.



Figur 6: ROC-kurvan samt AUC för ModellAIC och ModellBIC då observationer med avsaknade förklarande variabler är uteslutna.



Figur 7: ROC-kurvan samt AUC för ModellCredit_History och ModellFörstagradstermer då observationer med avsaknade förklarande variabler är uteslutna.