Stockholms universitet

Undersökning av metoder för att bestämma flyttantaganden för livförsäkring

Cecilia Söderberg

Masteruppsats i försäkringsmatematik
Master Thesis in Actuarial Mathematics

Masteruppsats 2022:12
Försäkringsmatematik
September 2022

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematisk statistik
Stockholms universitet
Masteruppsats **2022:12**
http://www.math.su.se/matstat

Prissättning av försäkringskontrakt med fokus på hur kontrakt i ett nytt område ska prissättas

Cecilia Söderberg*

September 2022

Sammanfattning

I det här arbetet har vi undersökt metoder för att ta fram flyttantaganden för livförsäkringar för ett svenskt livsförsäkringsbolag. Vi har undersökt tre olika metoder; Random forest och logistisk regression med GLM och GLMM.

Data som har använts är longitudinell data för åren 2010-2019. Det innebär att samma försäkringsnummer kan förekomma flera gånger, men för olika år.

Andelen kontrakt som flyttar är väldigt låg vilket innebär att vi har obalanserat data. För att förbättra förutsättningarna för prediktiv modellering har vi balanserat data. Först med metoden under sampling och sedan en variant av under sampling, men där vi tagit hänsyn till att ett kontrakt kan förekomma flera gånger. För båda metoderna såg vi en förbättring av våra modeller.

För att utvärdera modellerna har vi undersökt hur modellen presterar på ett år som är exkluderat från data som används för att bygga modellen. Vi har då jämfört andel flyttar och andel flyttat belopp. För att få säkrare skattningar har vi utifrån modellerna simulerat utfallen 10 000 gånger och tagit medelvärde och standardavvikelse från dessa samt åskådliggjort resultaten från simuleringarna i histogram.

När vi jämför andelen av antal flyttar är det inte någon av metoderna som är nära verkligt utfall. Däremot kommer vi närmare när vi använder flyttbart belopp. Då träffar vi verklig andel i simuleringarna i fyra fall. Bäst resultat fås av att utföra GLM på balanserad data med metoden under sampling. Dock så underskattar vi andelen flyttbart belopp i majoriteten av simuleringarna, vilket inte önskvärt.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige. E-post:cecilia.bk.soderberg@gmail.com. Handledare: Filip Lindskog.

Abstract

In this master thesis we examine methods to predict lapses for life insurances. We have used three different methods; Random Forest, logistic regression with GLM and GLMM.

We have used panel data for the years 2010-2019. Meaning that a specific contract can appear several times for different time periods. There is a very small proportion of the observation that lapses, which makes our data imbalanced. To correct this, we have balanced the data with the method under sampling. We have also used a variation of under sampling where we have taken into consideration that a contract might appear several times in the data. For both the new data sets we see an improvement in predictions when we use them to create our models.

To evaluate the performance of our models we have excluded one year from our data when we build the models. We use this data to compare how the share of number of lapses and the share of the lapsed amount differs from the predicted values from our models. To make our predictions more robust we simulate the results from our models 10 000 times. We take the mean value and the standard deviation and plot the results in histograms.

Of all the models, none of them does well predicting share of number of lapses. However, 4 of them predict correctly in some of the 10 000 simulations for the share of lapsed amount. The model that is most close to the correct amount is GLM using balanced data with the method under sampling. However, in most of the simulations the model underestimate the lapsed amount, which is not desired.

Förord och tack

Jag vill rikta ett stort tack till min handledare Filip Lindskog för ett stort tålamod och bra handledning. Jag har lärt mig oerhört mycket under arbetets gång och med din hjälp har jag lärt mig nya intressanta metoder och fått en ökad förståelse för metoder jag använt tidigare.

Jag vill även tacka mina kollegor som har stöttat med idéer och hjälp med data.

Slutligen vill jag även tacka Johan för ett stort stöd och mina vänner som har uppmuntrat mig under denna långa process.

Innehåll

Kapitel 1

Inledning

1.1 Bakgrund

Livförsäkringskontrakt har långa durationer och för att beräkningar, både avseende lönsamhetsanalys och till försäkringstekniska avsättningar (FTA), ska bli korrekta behövs lämpliga antaganden för annullationer. Är antagandena satta för låga kan det innebära likviditetsproblem i framtiden medan om dom är för höga ser lönsamheten sämre ut än vad den faktiskt är vilket kan ge följdeffekter som för låg återbäring till kunderna för ett ömsesidigt bolag. Då det är en stor konkurrens om kunderna kan detta innebära att kunder i sin tur väljer att flytta sin försäkring till ett försäkringsbolag med högre återbäring.

Det finns flera olika typer av annullationer. Vi menar i det här arbetet att säga upp sitt försäkringskontrakt i förtid. Det finns då två typer av annullationer: flyttar och återköp. En flytt innebär att ett försäkringskontrakt flyttar sitt försäkringskapital till en annan försäkring medan ett återköp innebär att försäkringstagaren tar ut sitt sparande innan den avtalade utbetalningstiden. Om ett kontrakt får flytta eller återköpa beror på vilken skatteklass kontraktet tillhör. K-skattade försäkringar (kapitalförsäkringar) får återköpa medan P-skattade (pensionsförsäkringar) får flytta. Även P-skattade försäkringar kan återköpas om beloppet understiger ett prisbasbelopp (konsumenternas.se 2015). I det här arbetet kommer vi att fokusera på att undersöka metoder för att bestämma antaganden för flyttar.

1.2 Syfte & metod

Att ha korrekta antaganden för annullationer är alltså väldigt viktigt ur flera perspektiv. I det här arbetet vill ett svenskt livförsäkringsbolag undersöka metoder för hur antaganden för annullationer sätts. Syftet är inte att förutsäga om ett specifikt kontrakt kommer att flytta, utan vi vill fånga effekten på det stora hela och därmed kunna sätta av en reserv. Det är alltså viktigt med en andel flyttar som är likt den verkliga andelen flyttar.

Det finns flera olika metoder för att ta fram flyttantaganden. I den här uppsatsen kommer vi att undersöka om GLM, GLMM samt random forest är lämpliga metoder. Som motivation till varför dessa metoder har valts ut är att det går att hitta drivare samt samspel mellan drivare som påverkar annullationer och därmed få antaganden på en mer detaljerad nivå.

Kapitel 2

Teori

Det finns flera olika metoder för att prediktera framtida annullationsbeteende. I den här uppsatsen kommer vi att undersöka metoderna GLM, GLMM samt random forests. Det finns ett flertal artiklar där som undersöker om GLM är en lämplig metod för att förutsäga annullationer. Exempelvis har (Kim 2005), (Cox and Lin 2006) och (Cerchiara et al. 2008) undersökt metoden. GLMM och random forests verkar inte vara vanliga sätt att prediktera annullationsbeteende i dagsläget.

2.1 GLM

Genom linjär regression modelleras ett samband mellan en variabel vi önskar förklara, responsvariabeln, och kovariater, även kallade förklaringsvariabler. En utökning av linjär regression är GLM (Generalized linear models). Till exempel kan GLM användas vid räknedata eller, som i vårt fall, om responsvariabeln är binär. Att responsvariabeln är binär innebär att den antar värdena 0 eller 1.

För att beskriva metoden kan vi utgå från linjär regression. Då antas ett linjärt samband mellan väntevärdet för en responsvariabel Y och förklarande variabler X enligt ekvation (2.1).

$$E[Y_i] = \mu_i = \sum_{j=1}^{r} x_{ij}\beta_j \tag{2.1}$$

Vid GLM modelleras en transformation av väntevärdet, istället för väntevärdet själv. Vi får då ekvationen (2.2).

$$g(\mu_i) = \eta_i = \sum_{j=1}^{r} x_{ij}\beta_j \tag{2.2}$$

$g(\mu_i)$ är länkfunktionen, som är en av dom tre komponenterna i GLM. De andra två är responsvariabeln och den systematiska komponenten (Madsen and Thyregod 2010). Vi fortsätter med att gå igenom dom olika komponenterna.

Responsvariabel

Responsvariabeln kallas även den slumpmässiga komponenten och är den variabel som vi vill modellera. Responsvariabeln kommer från en fördelning som tillhör exponentiella spridningsfamiljen (EDM). Att en fördelning tillhör EDM innebär att dess täthetsfunktion kan skrivas på formen enligt ekvation (2.3).

$$f_{Y_i}(y_i; \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi/w_i} + c(y_i, \phi, w_i)\right\} \tag{2.3}$$

θ_i är en transformation av μ_i, ϕ är spridningsmått, w_i är vikter och $b(\theta_i)$ är den kumulanta funktionen. Den kumulanta funktionen går att härleda från den kumulantgenererande funktionen enligt ekvation (2.4).

$$E[\boldsymbol{Y}] = \boldsymbol{\Psi}'(0) = b'(\boldsymbol{\theta}) \tag{2.4}$$

Normalfördelningen, poissonfördelningen och gammafördelnigen är exempel på fördelningar som tillhör EDM (Ohlsson and Johansson 2010).

Systematiska komponenten

Den systematiska komponenten är en linjär koppling mellan responsvariabeln och kovariaterna. Vi skriver den som ekvation (2.5).

$$\boldsymbol{\eta} = \boldsymbol{\beta}\boldsymbol{X} \tag{2.5}$$

I ekvationen är $\boldsymbol{\beta}$ en vektor av parametrar som skattas och $\boldsymbol{X}$ är en matris med kovariater som kallas designmatrisen (Ohlsson and Johansson 2010).

Länkfunktionen

Länkfunktionen binder samman responsvariabeln och våra kovariater, den systematiska komponenten. Länkfunktionen transformerar väntevärdet för responsvariabeln till den systematiska komponenten. Vi skriver den enligt ekvation

(2.6).

$$g(\mu_i) = \mu_i = \sum_{j=1}^{r} x_{ij}\beta_j \tag{2.6}$$

$g(.)$ är en monoton differentierbar transformation. Som exempel på länkfunktion används identitetslänken för linjär regression, vilket innebär $g(x) = x$.

I det här arbetet kommer vi att använda logit-länken som används för binära variabler vid logistisk regression. Då ges logit-länken länkfunktionen enligt (2.7).

$$logit(\mu_i) = \log(\frac{\mu_i}{1-\mu_i}) = \beta_0 + \beta_1 x_1 + ... + \beta_r x_r \tag{2.7}$$

För att få fram sannolikheten för att responsvariabeln ska anta värdet 1 ($P(Y = 1)$) kan vi genom att exponera och flytta runt termer få ekvation (2.8).

$$\mu_i = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+...+\beta_r x_r)}} \tag{2.8}$$

(Ohlsson and Johansson 2010).

Tolkning av oddskvoter för logistisk regression

Vid logistisk regression fås koefficienterna ut som logodds (LOR) av förklaringsvariabeln och definieras enligt ekvation (2.9).

$$LOR = \log(\frac{\mu}{1-\mu}) = log(\frac{p_1}{1-p_1}/\frac{p_2}{1-p_2}) \tag{2.9}$$

Vi börjar med att definiera vad ett odds är. I ekvationerna är p_i är sannolikheten för att en händelse lyckas och $1 - p_i$ är sannolikheten för att händelsen misslyckas för $i = 1, 2$. Oddset för en händelse är alltså delen $\frac{p_i}{1-p_i}$. Vi förtydligar med ett exempel: om sannolikheten att lyckas är 0,7 (p=0,7) beräknas oddset som 0,7/(1-0,7)=2,33. Detta innebär att oddset för att lyckas är 2,33 mer troligt än att misslyckas.

Logodds är svårtolkade och mer vanligt är att istället ta exponenten av logoddsen för att istället få ut oddskvoten (OR). Oddskvoten anger oddsen för att en händelse lyckas mot att händelsen misslyckas och definieras enligt ekvation (2.10).

$$OR = \frac{p_1}{1-p_1}/\frac{p_2}{1-p_2} \tag{2.10}$$

Oddskvoten beskriver hur oddsen förändras när en av variablerna förändras. Detta kan vara antingen en numerisk variabel eller en kategorisk variabel. Vi illustrerar med exempel. Vi antar för enkelhetens skull att vi har en ekvation med en konstantterm, β_0 och en förklaringsvariabel, x. Vi har då ekvationen enligt (2.11).

$$logit(\mu) = \beta_0 + \beta_1 x \tag{2.11}$$

Om x är en kategorisk variabel som kan anta värdena 0 eller 1 har vi att oddskvoten kommer att beräknas enligt (2.12).

$$OR = \frac{P(Y=1|x=1)}{P(Y=0|x=1)} / \frac{P(Y=1|x=0)}{P(Y=0|x=0)} = \frac{e^{\beta_0+\beta_1*1}}{e^{\beta_0+\beta_1*0}} = e^{\beta_1} \tag{2.12}$$

På samma sätt kommer vi kunna tolka en förändring om x istället är en numerisk variabel. Vi ser hur oddsen förändras när variabeln förändras med en enhet enligt ekvation (2.13).

$$OR = \frac{P(Y=1|x=2)}{P(Y=0|x=2)} / \frac{P(Y=1|x=1)}{P(Y=0|x=1)} = \frac{e^{\beta_0+\beta_1*2}}{e^{\beta_0+\beta_1*1}} = e^{\beta_1} \tag{2.13}$$

Om täljare och nämnare är lika stora kommer oddskvoten att anta värdet 1. Då innebär det att det inte blir någon skillnad när vi förändrar förklaringsvariabeln (Agresti 2002).

2.2 GLMM

Ett av dom grundläggande antaganden i GLM är att observationerna i data är oberoende. Detta antagande är inte alltid uppfyllt, t.ex. vid longitudinella studier där samma individ observeras vid upprepade tillfällen eller när vi har ett kluster som är gemensamt för observationer i data. Ett exempel på ett sådant kluster är vilken läkare som utför en behandling i en studie. Det är rimligt att anta att det finns individuella skillnader mellan läkare och att dessa kanske utför behandlingen något olika. För att hantera beroendet i vår data kan vi använda GLMM (generalized linear mixed models). I GLMM tas hänsyn till att det finns individuella skillnader mellan subjekt i data genom att slumpmässiga effekter för subjekten införs (Antonio and Beirlant 2005).

I GLM har vi fixa effekter som gäller för alla observationer i vår data, våra förklaringsvariabler. I GLMM inför vi även slumpmässiga effekter. Dom slumpmässiga effekterna tillämpas bara för en delmängd av observationerna i data. Om

observationerna tillhör samma subjekt kommer dom ha samma slumpmässiga effekt, däremot kommer olika subjekt ha olika slumpmässiga effekter. Vi menar här med subjekt den variabel som klassificerar våra observationer i kluster. T.ex. om vi undersöker ett läkemedels effekt på populationen och vi behandlar doktorn som behandlar patienten som subjekt kommer alla observationer från den doktorn ha samma slumpmässiga effekt. Dom slumpmässiga effekterna observerar vi inte och vi hanterar dom som om dom varierar slumpmässigt mellan subjekten (Agresti 2002).

För att beskriva hur vi anpassar GLM för att hantera dom slumpmässiga effekterna backar vi några steg och utgår från en linjär regression enligt ekvation (2.14).

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.14}$$

I ekvation (2.14) är $\boldsymbol{\beta}$ en vektor innehållande parametrarna för dom p fixa effekterna. Skattningarna för dessa är gemensamma för alla subjekten. Vi utökar sedan ekvation med en variabel som fångar den slumpmässiga effekten för dom i olika subjekten, vilket innebär termen $\boldsymbol{Zb}$ i ekvationen (2.15).

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Zb} + \boldsymbol{\epsilon} \tag{2.15}$$

I ekvation (2.15) är $\boldsymbol{b}$ är en vektor med dom q slumpmässiga effekterna och antas vara normalfördelad, $\boldsymbol{b} \sim N(0, \boldsymbol{\Gamma})$. Vi har alltså två typer av slumpmässiga variabler i GLMM - dom vi observerar, $\boldsymbol{Y}$, och dom vi inte observerar, $\boldsymbol{b}$. Vi har att $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Sigma})$ och $\boldsymbol{b}$ är oberoende (Antonio and Beirlant 2005). Kovariansmatriserna $\boldsymbol{\Sigma}$ och $\boldsymbol{\Gamma}$ beror på okända parametrar $\boldsymbol{\theta}$ som även dessa behöver skattas. $\boldsymbol{Z}$ är en designmatris bestående av värdena 0 och 1 för att indikera om observationen tillhör ett subjekt eller inte. Vi specificerar dimensionerna för ekvation (2.15) nedan.

$$\underbrace{\boldsymbol{Y}}_{n \times 1} = \underbrace{\boldsymbol{X}}_{n \times p} \underbrace{\boldsymbol{\beta}}_{p \times 1} + \underbrace{\boldsymbol{Z}}_{n \times q} \underbrace{\boldsymbol{b}}_{q \times 1} + \underbrace{\epsilon}_{n \times 1}$$

(Madsen and Thyregod 2010).

För att gå från linjär regression med slumpmässiga effekter till GLMM transformerar vi slutligen vår responsvariabel vilket leder till ekvation (2.16)

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Zb} \tag{2.16}$$

(Antonio and Beirlant 2005).

2.2.1 Likelihoodfunktionen

I GLM skattar vi våra koefficienter för våra förklaringsvariabler med hjälp av maximum-likelihood. Detta är dock inte helt enkelt vid GLMM, vi gör därför detta i två steg.

Det första steget innebär att vi, betingat på våra slumpmässiga effekter, antar att våra observationer följer en GLM. Då är $\boldsymbol{Zb}$ i ekvation (2.16) en känd offset och vi antar att våra observationer inom samma subjekt är oberoende av varandra (Agresti 2002). Vi betecknar denna som $f_{\boldsymbol{y}|\boldsymbol{b}}(\boldsymbol{y};\boldsymbol{b},\boldsymbol{\beta})$.

Det andra steget innebär att vi antar att våra slumpmässiga effekter är oberoende och kommer från en normalfördelning (Agresti 2002). Den slumpmässiga effekten för individ i ges av vektorn $\boldsymbol{b}_i$. Vi antar att $\boldsymbol{b}_i$ är oberoende och likafördelade med täthetsfunktionen $f_{\boldsymbol{b}}(\boldsymbol{b};\boldsymbol{\theta})$. Där $\boldsymbol{\theta}$ är dom okända parametrarna i täthetsfunktionen. Vi antar att våra slumpmässiga effekter är normalfördelade med väntevärde 0 och med en kovarianmatris $\boldsymbol{\Gamma}$ som bestäms av $\boldsymbol{\theta}$.

Genom att kombinera dom två stegen får vi den simultana likelihoodfunktionen enligt (2.17).

$$L(\boldsymbol{\beta},\boldsymbol{\theta};\boldsymbol{b},\boldsymbol{y}) = f_{\boldsymbol{y}|\boldsymbol{b}}(\boldsymbol{y};\boldsymbol{b},\boldsymbol{\beta})f_{\boldsymbol{b}}(\boldsymbol{b};\boldsymbol{\theta}) \tag{2.17}$$

I fallet då vi har en identitetslänk, det vill säga en normalfördelning, kan integralen beräknas. För andra länkfunktioner är så inte fallet och likelihooden måste approximeras (Antonio and Beirlant 2005). Det finns flera olika tekniker för att approximera integralen. Vi kommer gå igenom teknikerna för en förenklad approximationsteknik och Laplace. Båda använder metoden PIRLS. Skillnaden mellan metoderna är hur $\beta-$värdena, koefficienterna för våra fixa effekter skattas.

Laplace

Vi vill alltså approximera likelihood-funktionen och mer specifikt mål skatta $\boldsymbol{\theta}$ i vår vår simultana likelihood $L(\boldsymbol{\beta},\boldsymbol{\theta};\boldsymbol{b},\boldsymbol{y})$. För att göra detta behöver vi ta fram vår marginella likelihood genom att integrerar över $\boldsymbol{b}$, våra oobserverade slumpmässiga effekter. Vi gör detta i ekvation (2.18).

$$L_M(\boldsymbol{\beta},\boldsymbol{\theta};\boldsymbol{y}) = \int_{\mathbb{R}^q} L(\boldsymbol{\beta},\boldsymbol{\theta};\boldsymbol{b},\boldsymbol{y})\ d\boldsymbol{b} \tag{2.18}$$

I ekvationen är $L(\boldsymbol{\beta}, \boldsymbol{\theta}; \boldsymbol{b}, \boldsymbol{y})$ vår simultana likelihood för både dom stokastiska variabler vi observerar ($\boldsymbol{y}$) och dom vi inte observerar ($\boldsymbol{b}$). $\boldsymbol{\theta}$ betecknar våra parametrar som vi vill skatta.

Eftersom integralen är svår att lösa görs en approximering av log-likelihoodfunktionen, där log-likelihoodfuntionen är $l(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{y}) = \log(L(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{y}))$. Vi gör detta med en andra ordningens Taylorutveckling kring maxvärdet med avseende på dom ej observerade slumpmässiga variablerna $\mathbf{b}$. Detta resulterar i ekvation (2.19).

$$l(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{y}) \approx l(\boldsymbol{\beta}, \boldsymbol{\theta}, \tilde{\boldsymbol{b}}, \boldsymbol{y}) - \frac{1}{2}(\boldsymbol{b} - \tilde{\boldsymbol{b}})^T \boldsymbol{H}(\tilde{\boldsymbol{b}})(\boldsymbol{b} - \tilde{\boldsymbol{b}}) \tag{2.19}$$

Där $\boldsymbol{H}(\boldsymbol{\beta}, \tilde{\boldsymbol{b}}) = -l''(\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{y})|_{\tilde{\boldsymbol{b}}=\boldsymbol{b}}$ är den negativa hessianen av log-likelihoodfunktionen utvärderad i punkten $\tilde{\boldsymbol{b}}$. I Taylorutvecklingen försvinner första ordningens derivata eftersom funktionen utvecklas kring maximum. Vi får då att Laplace approximationen för den marginella log-likelihooden enligt ekvation (2.20)

$$\begin{aligned}
l_{M,LA}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y}) &= \log \int_{\mathbb{R}^q} e^{l(\boldsymbol{\beta}, \boldsymbol{\theta}, \tilde{\boldsymbol{b}}, \boldsymbol{y}) - \frac{1}{2}(\boldsymbol{b} - \tilde{\boldsymbol{b}})^T \boldsymbol{H}(\tilde{\boldsymbol{b}})(\boldsymbol{b} - \tilde{\boldsymbol{b}})} \mathrm{d}\boldsymbol{b} \\
&= l(\boldsymbol{\beta}, \boldsymbol{\theta}, \tilde{\boldsymbol{b}}, \boldsymbol{y}) + \log \int_{\mathbb{R}^q} e^{-\frac{1}{2}(\boldsymbol{b} - \tilde{\boldsymbol{b}})^T \boldsymbol{H}(\tilde{\boldsymbol{b}})(\boldsymbol{b} - \tilde{\boldsymbol{b}})} \mathrm{d}\boldsymbol{b} \\
&= l(\boldsymbol{\beta}, \boldsymbol{\theta}, \tilde{\boldsymbol{b}}, \boldsymbol{y}) + \log |\frac{2\pi}{\boldsymbol{H}(\tilde{\boldsymbol{b}})}|^{\frac{1}{2}} \int_{\mathbb{R}^q} \frac{e^{-\frac{1}{2}(\boldsymbol{b} - \tilde{\boldsymbol{b}})^T \boldsymbol{H}(\tilde{\boldsymbol{b}})(\boldsymbol{b} - \tilde{\boldsymbol{b}})}}{(2\pi)^{\frac{q}{2}} |\boldsymbol{H}^{-1}(\tilde{\boldsymbol{b}})|^{\frac{1}{2}}} \mathrm{d}\boldsymbol{b} \\
&= l(\boldsymbol{\beta}, \boldsymbol{\theta}, \tilde{\boldsymbol{b}}, \boldsymbol{y}) + \log |\frac{2\pi}{\boldsymbol{H}(\tilde{\boldsymbol{b}})}|^{\frac{1}{2}} \\
&= l(\boldsymbol{\beta}, \boldsymbol{\theta}, \tilde{\boldsymbol{b}}, \boldsymbol{y}) - \frac{1}{2} \log |\frac{\boldsymbol{H}(\tilde{\boldsymbol{b}})}{2\pi}|
\end{aligned} \tag{2.20}$$

(Madsen and Thyregod 2010).

Förenklad approximationsteknik (PIRLS)

Det finns en förenklad approximationsteknik där den slumpmässiga effekten inte integreras. Istället optimeras koefficienterna för dom fixa och dom slumpmässiga effekterna med PIRLS (Penalized Iteratively Reweighted Least Squares). Detta leder till snabbare uträkningar, men inte lika exakta skattningar. Ju fler observationer per subjekt desto mer lika blir skattningarna approximationen med Laplace och skillnaderna blir försumbara (Wu et al. 2019).

För att approximera våra skattningar använder vi en metod som bygger på

viktad minstakvadratmetoden, men där vi inför en straff-term och vi upprepar proceduren, med en för varje iteration, uppdaterad viktmatris tills vi uppfyller ett stoppkriterium (Kouwayè et al. 2013). Minstakvadratskattningen är ett sätt för att skatta en okänd parameter, $\boldsymbol{\theta}$. Den kräver inte att vi har en känd fördelning, vilket Maximum-likelihood kräver. Se Appendix A för detaljer kring hur minstakvadratskattningen beräknas. I stora drag kommer vi att utföra följande procedur:

1. Sätt startvärde för $\boldsymbol{b}$; $\boldsymbol{b}^{(0)} = 0$
2. Vid varje iteration uppdaterar vi värdena för vektorn med väntevärden, $\hat{\boldsymbol{\mu}}^{(r)}$, vektorn med derivator, $\boldsymbol{G}^{(r)}$, och den diagonala viktmatrisen, $\boldsymbol{W}^{(r)}$.
3. Vi använder våra uppdaterade värden för att beräkna $\boldsymbol{z}$ enligt $\boldsymbol{z}^{(r)} \approx \hat{\boldsymbol{\eta}}^{(\boldsymbol{r})} + \boldsymbol{G}^{(r)}(\boldsymbol{Y} - \hat{\boldsymbol{\mu}}^{(\boldsymbol{r})})$
4. Vi beräknar en skattning för våra slumpmässiga effekter genom att lösa ut $\hat{\boldsymbol{b}}^{r+1}$ ur $(\boldsymbol{Z}^T\boldsymbol{W}^{(r)}\boldsymbol{Z})\hat{\boldsymbol{b}}^{r+1} = \boldsymbol{Z}^T\boldsymbol{W}^{(r)}\boldsymbol{z}^{(r)}$
5. Vi upprepar steg 2 till 4 med våra uppdaterade parametrar tills ett stoppkriterium är uppnått.

Vi går nu in i detalj på hur vi beräknar våra skattningar. Vi betecknar vilken iteration vi befinner oss med r. $\hat{\boldsymbol{b}}^{(r)}$ innebär alltså skattningen för $\boldsymbol{b}$ (dom slumpmässiga effekterna som antas vara normalfördelade) för iteration r. Till exempel innebär $\hat{\boldsymbol{b}}^{(1)}$ skattningen för dom slumpmässiga effekterna vid den första iterationen.

Steg 1

Vi börjar med att sätta startvärdet för att beräkna värdena för iteration 1 ($r = 1$). Vi använder då $\boldsymbol{b}^{(0)} = 0$. Vi utför därefter steg 2-4 och använder den uppdaterade vektorn $b^{(1)}$ som input till iteration 2. Vi fortsätter på samma sätt med värdena som räknas fram i iteration 2 till iteration 3 o.s.v. tills vi når vårt stopkriterium (se steg 5 nedan) (Schelldorfer et al. 2014).

Steg 2

För varje iteration r uppdaterar vi värdena i vektorerna $\hat{\boldsymbol{\mu}}^{(r)}$ och $\boldsymbol{G}^{(r)}$ samt matrisen $\boldsymbol{W}^{(r)}$.

Vår vektor med väntevärden, $\hat{\boldsymbol{\mu}}^{(r)}$, uppdaterar vi enligt ekvation (2.21).

$$\hat{\boldsymbol{\mu}}^{(r)} = g^{-1}(\hat{\boldsymbol{\eta}}^{(r)}) \tag{2.21}$$

Vi har då definierat sambandet mellan koefficienterna för iteration r som ekvation (2.22) där vi då behandlar våra $\boldsymbol{\beta}$ som fixa

$$\hat{\boldsymbol{\eta}}^{(r)} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}\hat{\boldsymbol{b}}^{(\boldsymbol{r})} \tag{2.22}$$

(Kouwayè et al. 2013).

Eftersom vi använder logistisk regression har vi en logit-länk vilket innebär att vi får väntevärdet enligt ekvation (2.23).

$$\hat{\boldsymbol{\mu}}^{(r)} = g^{-1}(\hat{\boldsymbol{\eta}}^{(r)}) = \frac{1}{1 + e^{-\hat{\boldsymbol{\eta}}^{(r)}}} \tag{2.23}$$

(Bates 2011).

Vi uppdaterar vektorn med derivator, $\boldsymbol{G}^{(r)}$, enligt ekvation (2.24)

$$\boldsymbol{G}^{(r)} = \frac{d\boldsymbol{\eta}^{(r)}}{d\boldsymbol{\mu}^{(r)}} \tag{2.24}$$

(Kouwayè et al. 2013). $\boldsymbol{G}^{(r)}$ är alltså derivatan för länkfunktionen utvärderad i punkten $\boldsymbol{\mu}^{(r)}$ (McCullagh and Nelder 1999).

Eftersom vi använder logistisk regression innebär det att vi får derivator enligt ekvation (2.25)

$$\frac{d\boldsymbol{\eta}^{(r)}}{d\boldsymbol{\mu}^{(r)}} = \frac{d}{d\boldsymbol{\mu}^{(r)}} \log(\frac{\hat{\boldsymbol{\mu}}^{(r)}}{1 - \hat{\boldsymbol{\mu}}^{(r)}}) = \frac{1}{\hat{\boldsymbol{\mu}}^{(r)}} + \frac{1}{1 - \hat{\boldsymbol{\mu}}^{(r)}} = \frac{1}{\hat{\boldsymbol{\mu}}^{(r)}(1 - \hat{\boldsymbol{\mu}}^{(r)})} \tag{2.25}$$

(Bates 2011).

Slutligen uppdaterar vi vår diagonala viktmatris, $\boldsymbol{W}^{(r)}$, enligt ekvation (2.26). I ekvationen är $V^{(r)}$ den betingade variansen utvärderad i punkten $\hat{\boldsymbol{\mu}}^{(r)}$

$$\boldsymbol{W}^{(r)} = \frac{1}{(\boldsymbol{G}^{(r)})^2 V^{(r)}} = \hat{\boldsymbol{\mu}}^{(r)}(1 - \hat{\boldsymbol{\mu}}^{(r)}) \tag{2.26}$$

(McCullagh and Nelder 1999).

Steg 3

Vi kommer använda viktad minstakvadratskattning, men istället för att använda $\boldsymbol{y}$ som beroende variabel kommer vi att använda $\boldsymbol{z}$. $\boldsymbol{z}$ är en linjäriserad form av länkfunktionen tillämpad med $\boldsymbol{y}$. Vi använder då första ordningens Taylorutveckling i punkten $\boldsymbol{\mu}$ som ges av ekvation (2.27).

$$g(\boldsymbol{y}) \approx g(\boldsymbol{\mu}) + (\boldsymbol{y} - \boldsymbol{\mu})g'(\boldsymbol{\mu}) \tag{2.27}$$

Vilket innebär att vi i vårt fall för iteration r använder ekvation (2.28)

$$\boldsymbol{z}^{(r)} \approx \hat{\boldsymbol{\eta}}^{(\boldsymbol{r})} + \boldsymbol{G}^{(r)}(\boldsymbol{y} - \hat{\boldsymbol{\mu}}^{(\boldsymbol{r})}) \tag{2.28}$$

(McCullagh and Nelder 1999).

Steg 4

Det uppdaterade parametervärdet för $\hat{\boldsymbol{b}}^{(r+1)}$ är lösningen till den viktade minstakvadratskattningen som ges av ekvation (2.29).

$$(\boldsymbol{Z}^T\boldsymbol{W}^{(r)}\boldsymbol{Z})\hat{\boldsymbol{b}}^{r+1} = \boldsymbol{Z}^T\boldsymbol{W}^{(r)}\boldsymbol{z}^{(r)} \tag{2.29}$$

Dock så behöver vi ta ta hänsyn till variationen för våra slumpmässiga effekter vilket gör att vi inför en straffterm, $\boldsymbol{\Gamma}^{-1}$. Vi får då ekvation (2.30).

$$(\boldsymbol{Z}^T\boldsymbol{W}^{(r)}\boldsymbol{Z} + \boldsymbol{\Gamma}^{-1})\hat{\boldsymbol{b}}^{r+1} = \boldsymbol{Z}^T\boldsymbol{W}^{(r)}\boldsymbol{z}^{(r)} \tag{2.30}$$

För att förenkla våra uträkningar använder vi Choleskys faktorisering. Choleskys faktorisering innebär att vi kan faktorisera en symmetrisk, positivt definit matris A som $A = LL^T$. Där L är en undertriangulär matris med positiva diagonalelement. Vi använder Choleskys faktorisering och låter: $\boldsymbol{\Gamma}^{-1} = \Delta^{-1}\Delta$. Vi inför $\hat{\boldsymbol{b}}^* = \Delta\hat{\boldsymbol{b}}$ och $\boldsymbol{Z}^* = \boldsymbol{Z}\Delta^{-1}$. Detta gör att vi kan skriva om ekvation (2.30) till ekvation (2.31)

$$(\boldsymbol{Z}^{*T}\boldsymbol{W}^{(r)}\boldsymbol{Z}^* + \boldsymbol{I})\boldsymbol{b}^{*(r+1)} = \boldsymbol{Z}^{*T}\boldsymbol{W}^{(r)}\boldsymbol{z}^{(r)} \tag{2.31}$$

Genom att flytta om termerna får vi slutligen ekvation (2.32).

$$\hat{\boldsymbol{b}}^{*(r+1)} = (\boldsymbol{Z}^{*T}\boldsymbol{W}^{(r)}\boldsymbol{Z}^* + \boldsymbol{I})^{-1}(\boldsymbol{Z}^{*T}\boldsymbol{W}^{(r)}\boldsymbol{z}^{(r)}) \tag{2.32}$$

Steg 5

Värdena $\hat{\boldsymbol{b}}^{*(1)}, \hat{\boldsymbol{b}}^{*(2)}, .., \hat{\boldsymbol{b}}^{*(n)}$ konvergerar till $\hat{\boldsymbol{b}}^{*}(\boldsymbol{\beta}, \boldsymbol{\theta})$ när ett gränsvärde, a, är uppfyllt enligt ekvation (2.33).

$$\frac{||\hat{\boldsymbol{\eta}}^{(r+1)} - \hat{\boldsymbol{\eta}}^{(r)}||}{||\hat{\boldsymbol{\eta}}^{(r)}||} \leq a \tag{2.33}$$

Skattning av β-värden

När vi sätter inställningen $nagq = 0$ i R innebär det att förutom att dom slumpmässiga effekterna skattas genom PIRLS, görs även dom fixa effekterna det, det vill säga våra β-värden. Dessa skattas då genom ekvation (2.34).

$$\hat{\boldsymbol{\beta}} = \arg \max_{\beta} [d(\boldsymbol{Y}|\boldsymbol{\beta}, \hat{\boldsymbol{b}}) - \hat{\boldsymbol{b}}^{*T} \hat{\boldsymbol{b}}^{*} + \log(\det(\hat{\boldsymbol{\Gamma}}))] \tag{2.34}$$

I ekvationen är $d(\boldsymbol{Y}|\boldsymbol{\beta}, \hat{\boldsymbol{b}})$ modellens deviance och $\boldsymbol{\Gamma}^{(r)}$ är varians-kovariansmatrisen för $\boldsymbol{b}$. Den uppdateras för varje iteration enligt ekvation (2.35)

$$Var(\hat{\boldsymbol{b}}^{(r)}) \approx \hat{\boldsymbol{\Gamma}}^{(r)} = (\boldsymbol{Z}^{*T} \boldsymbol{W}^{(r)} \boldsymbol{Z}^{*} + \boldsymbol{I}) \tag{2.35}$$

(Kouwayè et al. 2013).

2.2.2 Tolkning av oddskvoter vid GLMM

När vi använder GLMM kommer vi precis som när vi använder GLM få oddskvoter. Vid GLM jämförs oddskvoterna mot referensgruppen som är det värde som antas för konstanttermen, β_0, men vid GLMM kommer dom olika subjekten ha olika värden för konstanttermen. Vi kommer alltså få ut oddskvoter för våra olika kovariater precis som vanligt, men hur dom tolkas beror på varje subjekt eftersom vi har olika värden för vår konstantterm för dessa.

Oftast är resultaten vi får ut från en modell svårtolkade och vi vill transformera dessa till oddskvoter, t.ex. är det lättare att tolka oddskvoter än logodds när vi har använt logistisk regression. En svårighet uppstår då eftersom transformationen av responsvariabeln oftast är icke-linjär vilket innebär att resultaten inte behöver bli additiva utan kan bli multiplikativa. Detta innebär att vi kan få oväntade resultat som är mer svårtolkade (Agresti 2002).

2.3 Mått för modelljämförelse & variabelselektion

För både GLM och GLMM kan vi kan bedöma både parametrar och modeller mot varandra. Vi vill även att modellen ska kunna prediktera framtida flyttar och därför behöver även detta testas.

VIF

Vi har ett antagande om oberoende mellan variablerna. Om detta inte är uppfyllt har vi multikollinaritet. Om vi har multikollinaritet mellan variabler kan det få konsekvenser som att variabler inte blir statistiskt signifikanta. För att mäta detta kan vi använda måttet variansinflationsfaktor (VIF). Vi definierar R_j^2 som den hur mycket variabeln x_j förklaras av de övriga $j-1$ variablerna. VIF definieras då enligt ekvation (2.36).

$$VIF = \frac{1}{1 - R_j^2} \tag{2.36}$$

Om VIF antar värdet 1 innebär det att det inte finns någon korrelation mellan variablerna. Ju högre värde på VIF desto mer korrelerade är variablerna. Vanliga gränser för när en variabel bör plockas bort brukar vara 5 eller 10 (Sundberg 2014).

Om en av variablerna har mer än en frihetsgrad kommer GVIF (Generaliserad variansinflationsfaktor) beräknas istället för VIF. VIF justeras då med antal frihetsgrader (DF) enligt ekvation (2.37)

$$GVIF = VIF^{\frac{1}{2DF}} \tag{2.37}$$

(Nilsson and Fox 2020).

LRT & Deviance

Vi kommer att vilja testa om det innebär en förbättring av modellen när vi inför en ny variabel. För att testa om den nya variabeln bidrar till modellen kan vi använda Likelihood Ratio Test (LRT). Vid LRT testas om en modell där variabler har tagits bort presterar lika bra som modellen med variablerna. Detta testas genom att en teststatistiska räknas ut genom att jämföra log-likelihooden för dom olika modellerna och definieras enligt (2.38).

$$LRT = 2[l(y) - l(\hat{\mu})] \tag{2.38}$$

I ekvation (2.38) är $l(y)$ log-likelihooden för modellen med variabeln vi önskar testa om den ger ett bidrag till modellen och $l(\hat{\mu})$ log-likelihooden för modellen utan variabeln (Ohlsson and Johansson 2010).

Teststatistikan LRT är χ^2-fördelad och där antalet frihetsgrader är lika med antalet parametrar som exkluderas i modellen vi jämför mot. Detta innebär att om vi testar en modell där vi har lagt till en variabel mot en modell där variabeln inte är inkluderad är antalet frihetsgrader 1 (Agresti 2002).

LRT kallas även scaled deviance. Deviance (D) fås genom ekvation (2.39).

$$D = \phi LRT \tag{2.39}$$

I ekvationen är ϕ en skalparameter som ingår i beräkningen av LRT. Genom att vi multiplicerar med ϕ kommer deviance inte vara beroende av denna (Ohlsson and Johansson 2010).

AIC & BIC

Vi vill skapa en modell som fångar verkligheten så bra som möjligt men som inte är alltför komplex och som blir överanpassad till data. Ett mått för att uppnå detta är Akaikes informations kriterie (AIC). AIC bedömer en modell efter hur nära dom predikterade värdena kommer dom faktiska enligt ekvation (2.40).

$$AIC = -2(\text{Max } l(y) - r) \tag{2.40}$$

Där $l(y)$ är log-likelihoodfunktionen och r är antalet parametrar. Den modell som har lägst AIC är den modell som bäst förklarar verkligheten. Som vi ser i ekvation (2.40) kommer en modell med många parametrar ha ett större AIC än en modell med färre parametrar.

Nära besläktad med AIC är Bayesian Information Criterion (BIC). Skillnaden mellan AIC och BIC är att BIC även tar hänsyn till antalet observationer. Detta definieras enligt ekvation (2.41).

$$BIC = -2 \log L + \log(n) r \tag{2.41}$$

Även för BIC föredras en modell med lägre värde jämfört en med ett högre värde (Agresti 2002).

2.4 Beslutsträd

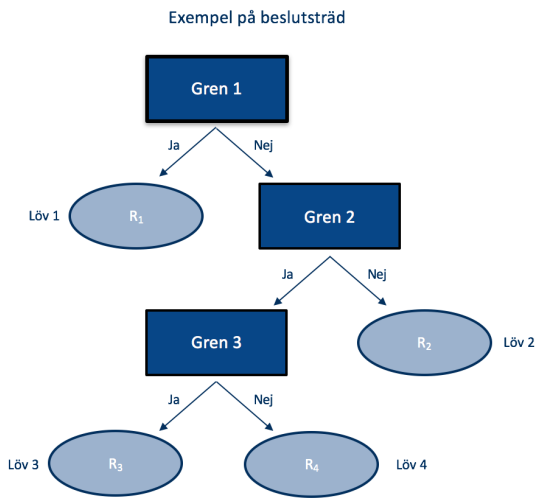
Det finns två typer av beslutsträd; regressionsträd, som används när responsvariabel är numerisk, och klassifikationsträd, som används när responsvariabeln är kategorisk. Vi kommer i det här arbetet fokusera på klassifikationsträd då vår responsvariabel i det här arbetet kommer vara av den typen, lapse ($Y = 1$) eller ej lapse ($Y = 0$). Då teorin skiljer sig åt mellan de olika typerna kommer vi i teoriavsnittet att fokusera på hur klassifikationsträd hanteras.

Fördelar med beslutsträd är att det är enkelt att både förklara och att åskådliggöra för icke-experter. Nackdelar är att dom oftast inte är lika bra på att prediktera som regressioner samt att dom är icke-robusta, en lite ändring i data kan ha stor påverkan. För att förbättra prediktionsförmågan kan tekniker som bagging och random forest användas. Vi kommer att beskriva dessa tekniker i avsnitt 2.4.3 och 2.4.5, men vi ska först gå igenom den allmänna teorin kring beslutsträd.

All teori under avsnitten om beslutsträd är hämtat från (James et al. 2013).

2.4.1 Konstruktion av beslutsträd

Vid förklaringen av hur beslutsträd fungerar utgår vi från bild 2.1. Ett beslutsträd är uppbyggt av grenar och löv. Grenarna är beslutsregler där observationen antingen kommer att gå åt det ena eller åt det andra hållet beroende på om beslutsregeln är uppfylld eller ej. Grenarna kallas även splits. Löven är längst ned på beslutsträdet när det inte längre finns några grenar. Det är dessa som är utfallen för responsvariabeln Y. De olika löven i beslutsträdet betecknas $R_1, R_2, \ldots, R_M$. I exemplet är $M = 4$, vilket innebär fyra löv.



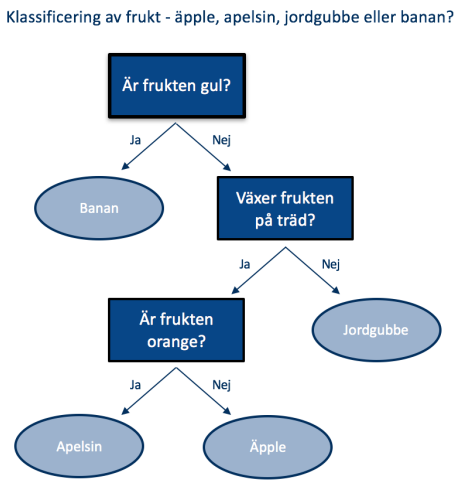
Figur 2.1: *Exempel på beslutsträd*

Modellen för klassifikationsträd skrivs som ekvation (2.42).

$$f(X) = \sum_{m=1}^{M} c_m * 1(X \in R_M) \tag{2.42}$$

c_m är utfallet som klassas som det som majoriteten i lövet klassas som.

Ett exempel på hur ett beslutträd kan vara utformat kan vi se i bild 2.2 där frukt klassificeras. Vi har en kategorisk responsvariabel som kan anta värdena äpple, banan, jordgubbe eller äpple. När vi börjar uppdelningen i träd tillhör alla observationer samma grupp. Vid det första ”splitet” delas gruppen upp i två delar. Då kommer gula frukter att delas upp till vänster medan frukter som inte är gula kommer att sorteras till höger. Gruppen till vänster kommer då till ett löv där de klassificeras som banan medan gruppen till höger kommer delas upp ytterligare. Nästa beslutsregel delar upp frukterna i om dom växer på träd vilket kommer innebära en ytterligare uppdelning. Om frukten inte växer på träd kommer frukten att klassificeras som jordgubbe. Om frukten växer på träd kommer den till en sista beslutsregeln. Då delas upp dom upp efter om dom är orange eller någon annan färg. Orange frukter kommer att klassificeras som apelsiner medan alla övriga kommer att klassificeras som äpplen.



Figur 2.2: *Exempel på klassificering av frukt mha beslutsträd.*

I klassifikationen kan det förekomma missar, en banan kan t.ex. vara grön och hamnar då felaktigt i kategorin äpple. Detta för oss vidare till mått på hur bra ett split är.

2.4.2 Mått på hur bra ett split är

Vi vill klassa en observation genom att den delas upp i grupper, grenar, tills den slutligen når ett löv. Vi vill att grenarna ska vara placerade på ett så bra sätt som möjligt och vi vill att dessa ska klassa rätt. Hur bra ett split, en grenuppdelning, är kan mätas med olika mått. Vi kommer här att gå igenom måttet Gini index.

Vi börjar med att definiera $\hat{p_{mk}}$ som anger proportionen av träningsdata i den m:e regionen som tillhör klass k. Detta innebär att om alla observationer tillhör klass k kommer $\hat{p_{mk}} = 1$. Gini index definieras enligt ekvation (2.43) och anger total varians över alla k klasser. Ett litet värde är bättre, då det innebär mindre variation inom ett löv, det vill säga att dom flesta observationerna tillhör samma klass.

$$G = \sum_{k=1}^{K} \hat{p_{mk}}(1 - \hat{p_{mk}}) \tag{2.43}$$

2.4.3 Bagging

Beslutsträd har hög varians. För att minska den kan vi använda tekniken bagging. Bagging kallas även Bootstrap aggregation och bygger på tekniken bootstrap. Bagging utgår ifrån teorin att ta medelvärdet av n oberoende observationer $Z_1, Z_2, .., Z_n$ som alla har varians σ^2 ger variansen σ^2/n. På så sätt minskar variansen och därmed blir prediktionen mer exakt.

Vid bagging tas många träningsstickprov från populationen. Vi får då B bootstrapade träningssätt. För varje stickprov b bygger vi beslutsträd som ger oss prediktionen $\hat{f}^{*b}$. Vi får alltså prediktionerna $\hat{f}^{*1}, \hat{f}^{*2}, .., \hat{f}^{*B}$ för vart och en av de B träningsstickproven. För varje träningsstickprov klassas utfallet efter det mest förekommande utfallet, det vill säga majoriteten.

2.4.4 Out-of-bag felskattning

Dom observationerna som inte väljs ut vid bagging för att skapa modellen för ett beslutsträd kallas out-of-bag (OOB) observationer. För att utvärdera hur stor felskattningen är för vår modell vi fått fram genom bagging kan vi använda dessa observationer och prediktera utfallet med hjälp av dom beslutsträd som inte använts för den observationen.

Vi predikterar utfall för observation i som tillhör OOB för vart och en av de olika beslutsträden där i är OOB. Vi använder precis som tidigare majoriteten vid klassifikation. Vi får då ett utfall för observation i. Vi kan på så sätt få en prediktion för samtliga observationer i OOB. Från detta kan vi sedan beräkna classification error.

2.4.5 Random forests

Vid random Forests använder vi en variant av tekniken bagging. Vi bygger att antal beslutsträd på bootstrapad data (för att förbättra prediktionsförmågan), men vid byggandet av beslutsträdet väljs en delmängd av förklaringsvariablerna ut som möjliga att använda vid varje split.

Beslutsträdet skapas genom att vi vid varje split slumpmässigt väljer ut m från p möjliga förklaringsvariabler. Vanligtvis är $m \approx \sqrt{p}$. Detta innebär att flera möjliga förklaringsvariabler kommer att väljas bort vid varje steg. Detta görs

eftersom vi vill undvika korrelation mellan de olika beslutsträden. Om vi skulle ha en förklaringsvariabel som har en stor påverkan på vår responsvariabel kommer den ingå i alla beslutsträd och träden skulle vara snarlikt konstruerade. Genom att slumpmässigt välja ut förklaringsvariabler kommer den förklaringsvariabel som har stor påverkan inte ingå i alla beslutsträd och vi kommer att få en större variation av hur träden är utformade. Vi fortsätter sedan att skapa beslutsträd på det här sättet med vår bootstrapade data tills vi når vårt stoppkriterium.

2.5 Obalanserat data

När vi vill bygga en modell som klassificerar om en händelse inträffar eller ej kan det uppstå problem om data är obalanserat. Att data är obalanserat innebär att det innehåller en väldigt liten andel av den klassen vi önskar prediktera i vår modell och väldigt stor andel av den andra klassen. Detta är t.ex. ett vanligt problem vid klassifikation av kreditbedrägerier. Data innehåller då oftast få observationer av kreditbedrägerier och många fall som inte är kreditbedrägerier. Om vi då skulle bygga en modell för att prediktera kreditbedrägerier förekommer det då en stor risk att modellen som ska klassificera framtida kreditbedrägerier kan se ut att ha en bra prediktionsförmåga, medan den i själva verket bara är bra på att prediktera den överrepresenterade gruppen, gruppen som inte är kreditbedrägerier. Vi kommer i det här arbetet att benämna den underrepresenterade gruppen som minoritetsgruppen och den överrepresenterade gruppen som majoritetsgruppen.

För att komma runt problematiken kan vi balansera data. Det finns flera olika metoder för att balansera data och dessa metoder delas in tre olika kategorier: sampling, ensemble och distansbaserade. Vilken av dom som är bäst avgörs från fall till fall och beror på vilken typ av data och hur den är fördelad. Vi kommer i det här arbetet att fokusera på den första gruppen, sampling, som innehåller två olika tekniker; over och under sampling (Dal Pozzolo et al. 2013).

2.5.1 Over sampling

Over sampling innebär att observationer slumpmässigt läggs till den grupp vi vill prediktera, minoritetsgruppen. Det finns även en variant på over sampling som heter SMOTE. Där genereras nya observationer genom interpolation från befintliga observationer (Dal Pozzolo et al. 2013). I vårt exempel med kreditbedrägerier skulle vi vid over sampling replikera observationer i data för fall där det skett kredit-

bedrägerier och på så sätt få en högre andel kreditbedrägerier av totala observationer.

En fördel med metoden over sampling är att ingen information förloras då alla observationer från båda grupperna behålls. Däremot kan problem uppstå när antalet observationer i data ökar. Då kan modellkörningar ta betydligt längre tid och det finns en risk att arbetsminnet tar slut (Rahman and Davis 2013).

2.5.2 Under sampling

Metoden under sampling innebär att observationer tas bort slumpmässigt från majoritetsgruppen tills data är balanserat. Idén bakom metoden är att det oftast är realistiskt att anta att många av observationerna från majoritetsgruppen är överflödiga och att det därför går att ta bort observationer från den här gruppen utan allt för stora konsekvenser (Dal Pozzolo et al. 2015). I vårt exempel med kreditbedrägerier skulle vi vid under sampling ta bort observationer i data för fall där det inte skett kreditbedrägerier och på så sätt få en högre andel kreditbedrägerier av totala observationer.

Metoden under sampling är till fördel vid stor data då tekniken leder till kortare körtider och mindre risk för att få slut på arbetsminne. Dock är det en nackdel med den här metoden är att information förloras eftersom vi tar bort observationer från data (Rahman and Davis 2013).

2.5.3 Hur sannolikheter påverkas vid under sampling

När vi balanserar data och bygger en modell på vår nya data för att sedan prediktera en klass i testdata (se avsnitt *Uppdelning av data* i kapitel 2.6) kommer vi efter att vi har balanserat data ha två olika fördelningar, en fördelning i data som vi använder för att bygga modellen (balanserad data) och en fördelning för vår testdata (obalanserad data). Vi kommer att ha en betydligt högre andel av minoritetsgruppen i vår data som vi använder för att bygga modellen än vad vi har i vår testdata. Detta innebär att det blir det en snedvridning i våra prediktioner och dessa får betydligt högre sannolikheter vilket gör att vi kommer att få en större andel som blir klassificerade som minoritetsklassen. Detta kan vi rätta upp genom att använda två olika metoder, en metod som används för random forest och en annan som används för logistisk regression.

Vid random forest

Vid random forest måste vi justera sannolikheten för händelsen att en observation tillhör minoritetsklassen. Vi börjar med att införa några definitioner. Vi har vårt ursprungliga, obalanserade, data som vi betecknar som $(\mathcal{X}, \mathcal{Y})$ och vår balanserade data som vi betecknar med (X, Y). (X, Y) är alltså en delmängd av $(\mathcal{X}, \mathcal{Y})$. Vi använder variabeln s för att beteckna om en observation tillhör (X, Y) enligt (2.44).

$$s = \begin{cases} 1, & \text{om en observation tillhör } (X, Y) \\ 0, & \text{annars} \end{cases} \tag{2.44}$$

Vi använder y för att beteckna om en observation tillhör minoritetsklassen eller inte enligt (2.45).

$$y = \begin{cases} 1, & \text{om en observation tillhör minoritetsklassen} \\ 0, & \text{om en observation tillhör majoritetsklassen} \end{cases} \tag{2.45}$$

Vi antar ett oberoende mellan s och x givet y. Detta eftersom vi slumpmässigt väljer ut observationer från majoritetsklassen från $(\mathcal{X}, \mathcal{Y})$.

Vi låter p_s beteckna sannolikheten för att i det balanserade data få en observation från minoritetsklassen, det vill säga $p_s = p(y = 1|x, s = 1)$. Genom att använda Bayes sats tillsammans med att $p(s|y, x) = p(s|y)$ kan vi skriva p_s enligt ekvation (2.46).

$$p_s = p(y = 1|x, s = 1) = \frac{p(s = 1|y = 1)p(y = 1|x)}{p(s = 1|y = 1)p(y = 1|x) + p(s = 1|y = 0)p(y = 0|x)} \tag{2.46}$$

Genom att utnyttja att $p(s = 1|y = 1) = 1$ (eftersom samtliga observationer från minoritetsklassen är kvar i vårt balanserade data) får vi ekvation (2.47).

$$p_s = p(y = 1|x, s = 1) = \frac{p(y = 1|x)}{p(y = 1|x) + p(s = 1|y = 0)p(y = 0|x)} \tag{2.47}$$

Vi inför beteckningen $\beta = p(s = 1|y = 0)$, som är sannolikheten för att en observation tillhör majoritetsklassen i det balanserade data. Vi har sambandet enligt ekvation (2.48). Där är N^+ antal 1:or och N^- antal 0:or i vårt obalanserade data.

$$\beta = \frac{p(y = 1)}{p(y = 0)} \approx \frac{N^+}{N^-} \tag{2.48}$$

När vi balanserar vår data kommer vi ha samma antal 1:or, det vill säga: $N_s^+ = N^+$, men eftersom vi tar bort observationer från majoritetsklassen får vi istället N_s^- observationer som är 0:or och vi har sambandet i ekvation (2.49).

$$N_s^- = \beta N^- \tag{2.49}$$

Vilket innebär att vi kan beräkna värdet för β när vi har balanserat data med ekvation (2.50).

$$\beta = \frac{N_s^-}{N^-} \tag{2.50}$$

Vi inför även beteckningen p för sannolikheten att en observation tillhör minoritetsklassen i data som är obalanserat, det vill säga $p = p(y = 1|x)$, vilket ger som följd att $1 - p = p(y = 0|x)$. Nu kan vi skriva sambandet mellan p_s och p som ekvation (2.51).

$$p_s = \frac{p}{p + \beta(1 - p)} \tag{2.51}$$

Vi kan nu få fram sannolikheten som vi är ute efter, p, genom att använda ekvation (2.52)

$$p = \frac{p_s}{p_s + \frac{1-p_s}{\beta}} \tag{2.52}$$

I ekvationen är p_s alltså sannolikheten att få en observation från minoritetsklassen i vårt balanserade data (som vi får från vår modell) och β är antalet 0:or i vårt balanserade data dividerat med antal 0:or i vårt obalanserade data (Dal Pozzolo et al. 2015).

Om vi utgår från vårt exempel med kreditbedrägerier och antar att vi har totalt 10 000 observationer varav 1 000 av våra observationer är kreditbedrägerier och 9 000 är fall då det inte skett kreditbedrägerier. Då har vi att $\beta = 1000/9000$. Om vi använder metoden under sampling för att ta bort observationer tills vi når jämnvikt kommer vi då ha 1 000 fall av kreditbedrägerier och 1 000 fall av icke-kreditbedrägerier. Vi kommer då att ha:

$$1000 = \beta * 9000 \tag{2.53}$$

Vilket leder till att vi får $\beta = 1/9$. Vi antar vidare att ett kontrakt i vårt testdata har fått sannolikheten 0,2 för att vara ett kreditbedrägeri. Då kan vi använda ekvation (2.52) och får då sannolikheten för p enligt ekvation (2.54).

$$p = \frac{p_s}{p_s + \frac{1-p_s}{\beta}} = \frac{0,2}{0,2 + \frac{1-0,2}{1/9}} \approx 0,027 \tag{2.54}$$

Vid logistisk regression

När vi använder logistisk regression på balanserad data korrigerar vi inte sannolikheten för händelsen, utan vi behöver vi bara korrigera vår konstantterm. Detta eftersom våra lutningskoefficienter är statistiskt konsekventa skattningar av dom verkliga värdena (King and Zeng 2001b). Vi använder en metod som kallas *prior correction* där vi justerar vår konstantterm enligt ekvation (2.55).

$$\beta_0 = \hat{\beta}_0 - \ln\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\overline{y}}{1-\overline{y}}\right)\right] \tag{2.55}$$

I ekvationen är $\hat{\beta}_0$ vår skattning av konstanttermen för modellen med balanserad data, τ är observerad andel 1:or i vår obalanserade data och $\overline{y}$ är observerade 1:or i vår balanserade data (King and Zeng 2001a).

2.6 Bedömning av prediktionsförmåga

Uppdelning av data

För att bedöma modellens prediktionsförmåga kommer data att delas in i två delar; en del för att ta fram en lämplig modell och del för att testa modellens prediktionsförmåga. Detta är en metod som bland annat användes av (Cerchiara et al. 2008). Då användes 70% av data för att ta fram modellen och 30% används för att testa prediktionsförmågan.

Klassifikationstabeller & ROC-kurvor

Vi kan jämföra det faktiska utfallet med det predikterade i en klassifikationstabell. Vi kommer för varje observation i vår testdata få fram en sannolikhet. Om sannolikheten överstiger ett gränsvärde kommer vi att klassa observationen som 1. Understiger sannolikheten gränsvärdet kommer vi att klassa den som 0. Ett vanligt gränsvärde är 0,5 (Agresti 2002). Vi kommer då kunna dela in våra observationer i fyra olika grupper:

- *Correct rejections* är fallet om både det observerade och det predikterade antar värdet 0.

- *Misses* är fallet om vi observerar värdet 1 men det predikterade antar värdet 0.
- *False* är fallet om vi predikterar 1 men det observerade värdet är 0.
- *Success* är fallet om vi både har predikterat 1 och det observerade antar värdet 1.

Grupperna och hur dessa klassificeras ses i tabell 2.1.

		Observerade	
		0	**1**
Predikterade	**0**	*Correct rejections*	*Misses*
	1	*False*	*Success*

Tabell 2.1: *Klassifikationstabell*

Utifrån dessa kan vi ta fram måtten Sensitivity, antalet 1:or som vi rätt klassificerar som 1:or:

$$\text{Sensitivity} = \frac{Success}{Success + Misses}$$

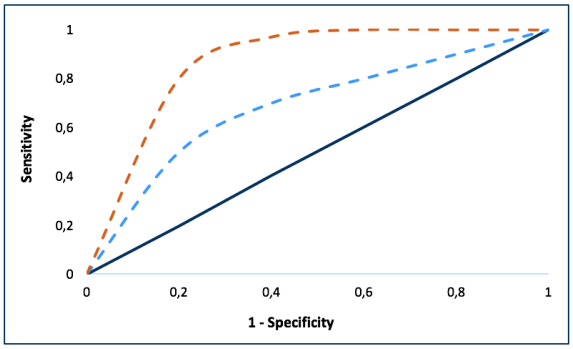
och Specificity, antalet 0:or som vi rätt klassificerar som 0:or:

$$\text{Specificity} = \frac{Correct\ rejections}{Correct\ rejections + False}$$

(Milhaud 2013).

En ROC-kurva (Receiver Operating Characteristic) är en kurva för att bedöma en modells prediktionsförmåga. På x-axeln har vi 1-specificity, som är dom möjliga gränsvärdena. Dessa kallas även False positive rate. På y-axeln ser vi vilket värde vi får på sensitivity, även kallad True Positive rate, för respektive gränsvärde (Agresti 2002). Ett exempel på hur ROC-kuvor kan se ut ses i figur 2.3. Där visar den svarta linjen när utfallet är densamma som slumpen. Vi har två streckade kurvor i vårt exempel. När vi skapar kurvorna stegar vi fram genom olika gränsvärden från 0 till 1. För varje gränsvärde observerar vi hur många 1:or som vi rätt klassificerar som 1:or - vår Sensitivity, som blir vårt y-värde i grafen. I punkten (0,0) kommer Sensitivity att vara 0, vilket innebär att inga av våra verkliga 1:or kommer att klassas som 1:or, däremot kommer Specificity vara 1 vilket innebär att alla 0:or rätt kommer att klassificeras som 0:or. I punkten (1,1) kommer vi klassa alla 1:or som 1:or, men vi kommer inte klassa några 0:or korrekt som 0:or. I vårt exempel ser vi att den orangea kurvan får en högre Sensitivity, alltså en större andel rätt klassificerade 1:or,

för ett lägre gränsvärde. Gränsvärdet visar andelen 0:or som vi felaktigt klassificerar som 1:or. Kurvan visar med andra ord hur många rätt klassificerade 1:or vi får till priset av fel klassificerade 0:or (James et al. 2013). I figuren har vi alltså att den orangea streckade kurvan ger ett bättre utfall än den blåa streckade kurvan.



Figur 2.3: *Exempel på ROC-Kurvor*

För att utvärdera ROC-kurvor mot varandra kan måttet AUC (Area Under the Curve) användas. Ju högre värde, närmare 1, på AUC desto bättre. Ett värde på 0,5 eller lägre innebär att klassifikationen är lika bra eller sämre än slumpen (James et al. 2013). I figur 2.3 kommer den orangea kurvan ha en större AUC än den blåa kurvan eftersom ytan under den kurvan är större än ytan under den blå streckade linjen (James et al. 2013).

Kapitel 3

Data

Data som har använts i det här arbetet kommer från flera olika datakällor som har kombinerats. Då databaserna som utgör grunden för data inte är skapade för att användas i det här syftet finns det brister i data.

På grund utav sekretesskäl kommer vi inte skriva ut alla variablerna som ingår i analysen. Variablerna kommer istället att betecknas med löpnummer. Variablerna samt vilken typ av variabel som den är återfinns i tabell B.1 i Appendix B.

Regelverk för flyttar

Det finns ett regelverk för vilka kontrakt och vilket belopp som får flytta. Vi har förenklat urvalet av kontrakt till privattecknade pensionsförsäkringar, kollektivavtalad tjänstepension och tjänstepensionsförsäkringar med teckningsdatum fr.o.m. 2006-01-01. Vår data innehåller både försäkringar under uppskov och utbetalning. I samtliga fall tas bara kontrakt som får flytta med.

Vid flytt av ett kontrakt tas en fast och rörlig avgift ut. Den rörliga avgiften beror på kontraktets duration. Ju tidigare flytten sker desto högre avgift. Exempel på hur dom rörliga flyttavgifterna kan vara utformade ses i tabell 3.1.

Försäkringsår	**Flyttavgift**
1-3	4,0 %
4-6	3,0 %
7-9	2,0 %
10-	1,0 %

Tabell 3.1: *Exempel på rörliga flyttavgifter per år från tecknandet.*

Longitudinell data

I det här arbetet används data för var och ett av åren 2010 till 2019. Samma kontrakt kan finnas med alla åren, vilket innebär att data inte är oberoende. Om ett kontrakt flyttar år x innebär det att det inte har flyttat under åren $x-1$, $x-2$ etc. för alla åren innan år x. Att data är utformat på det här sättet, samma kontrakt under flera tidsperioder kallas longitudinell data eller paneldata.

Kapitel 4

Modellering

Vi kommer i det här arbetet använda tre olika metoder för att prediktera flyttar. I samtliga metoder kommer vi att betrakta händelsen flytt som en binär variabel. Vi har då definitionen enligt ekvation (4.1).

$$Y = \begin{cases} 1, & \text{om kontraktet flyttat under perioden} \\ 0, & \text{annars} \end{cases} \tag{4.1}$$

Vi har delat upp vår data i två delar, en del för att bygga modeller på och en del för att testa och utvärdera utfallet av våra modeller. Vi får genom våra olika modeller en sannolikhet för flytt för varje observation i vår testdata. Utifrån dessa sannolikheter gör vi 10 000 simuleringar. Då får vi 10 000 utfall där kontrakten i vår testdata antingen har flyttat eller inte flyttat. Utifrån dessa tar vi fram histogram, medelvärde och medianen och jämför med verkligt utfall för andel flyttade kontrakt samt andel av flyttat belopp. Vi utför detta förfarande för samtliga modeller som vi skapar genom metoderna GLM, GLMM och random forest.

Utöver klassifikationstabeller använder vi ROC-kurvor för att undersöka hur bra våra prediktioner blir. Vi använder paketet *ROCR* för att göra ROC-kurvor i R. Paketet hjälper oss att rita upp vilken Sensitivty vi får för varje gränsvärde.

4.1 GLM

Den första metoden vi använder oss utav är GLM på obalanserad data. Vi börjar vår analys med att undersöka boxplottar och undersöka variablerna var och en för sig.

Vi tar inte med variabler där vi har för många saknade värden eller som är en mer detaljerad uppdelning av en annan variabel och där vi inte har tillräckligt många observationer i den mer detaljerade uppdelningen, men har det i den grövre. Detta innebär att vi tar bort *Variabel 3*.

Det finns flera olika metoder för att välja ut vilka variabler som ska ingå i en logistisk regression, som t.ex. backward elimination och forward elimination. Då vi har väldigt många observationer och förklaringsvariabler i vår data lämpar sig inte backward elimination, vi får felmeddelanden om att arbetsminnet tar slut. Vi börjar därför med att undersöka dom olika förklaringsvariablerna var för sig. Utifrån dom variabler som är signifikanta tar vi in dessa i en modell och undersöker hur mycket dom bidrar till modellen genom att undersöka AIC, BIC och Deviance. Dom variabler som inte bidrar med mycket och inte är signifikanta på 5% signifikansnivå tas bort. Vi undersöker även kontinuerligt om GVIF antar ett värde större än 5 för någon förklaringsvariabel. Om så är fallet tas denna bort. Vi testar modeller med och utan samspel.

För vår slutliga modell vill vi beräkna flyttsannolikheten för varje observation i vår testdata. Vi använder funktionen *predict* i R med specifikationen $type = 'response'$ för att beräkna sannolikheten enligt ekvation (2.7).

4.2 Balansering av data

I snitt flyttar ca 1% av dom flyttbara kontrakten (ca 4 000 kontrakt) för varje år. Detta är en liten andel vilket gör att data blir obalanserat. För att balansera data följer vi stegen i figur 4.1.

Figur 4.1: *Stegen vi gör vid balansering av data.*

Vi utgår från vår data och balanserar denna genom metoden under sampling. Vi väljer just den metoden eftersom vi har ett stort dataunderlag och redan nu har vi problem med att RAM-minnet tar slut vid vissa beräkningar. Vi använder metoden på två olika sätt i den här rapporten - ett sätt för när vi använder GLM och ett annat för när vi använder GLMM. Detta eftersom vi vill ta hänsyn till beroende när

vi använder metoden GLMM. På vårt nya data bygger vi modeller. Vi kommer att bygga modellerna med logistisk regression genom GLM och GLMM, och random forest. Från dessa modeller tar vi fram sannolikheten för flytt för varje observation. Denna sannolikhet är beräknad på balanserad data och är därför betydligt högre än den verkliga sannolikheten. Vi måste därför justera tillbaka sannolikheten till en obalanserad sannolikhet.

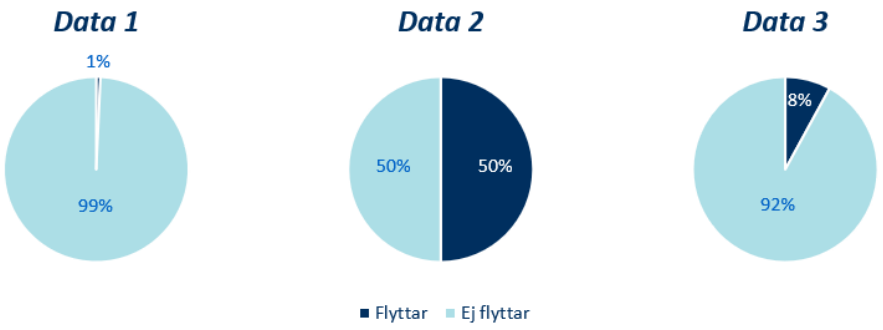
För att göra den första balanseringen av data med metoden under sampling använder vi *downSample* från paketet *caret* i R. Då kommer observationer från majoritetsgruppen, det vill säga kontrakt som inte har flyttat, slumpvis tas bort tills vi kommer till en jämnvikt där 50 % av observationerna är kontrakt som har flyttat och 50 % av observationerna är kontrakt som inte har flyttat. Vi kommer i rapporten härmed benämna det här data som *data 2*.

Den andra balanseringen av data gör vi för att kunna använda GLMM. När vi modellerar med GLMM vill vi inte ta bort data helt slumpmässigt, då vi har ett beroende i data eftersom vi följer kontrakten under flera år. Om vi slumpmässigt tar bort observationer kan det innebära att ett kontrakt kommer att finnas kvar i data år 1, är borttaget år 2 och sedan dyker upp i data igen år 3. Detta kanske blir ett mindre problem vid användning av GLM, men blir ett större problem med GLMM då vi vill ta hänsyn till att det finns ett beroende mellan kontrakten. För att komma till rätta med det här problemet väljer vi ut alla kontraktsnummer som har flyttat samt alla kontraktsnummer som inte har flyttat. Vi tar sedan slumpmässigt bort kontraktsnummer som ej har flyttat tills vi har en jämnvikt. Vi kommer då inte ha perfekt jämnvikt, då kontrakten kan förekomma olika antal år i data. Vi kommer i rapporten härmed benämna den här data som *data 3*.

Dom olika data som används i modelleringen visas i tabell 4.1 och fördelningen för antal flyttar för dom olika data ses i figur 4.2.

Namn	*Förklaring*
Data 1	Orginaldata
Data 2	Balanserad data med under sampling
Data 3	Balanserad data med hänsyn till beroende

Tabell 4.1: *Tabell över dom olika data som används vid modelleringen.*



Figur 4.2: *Andel flyttar för dom olika data.*

4.3 Logistisk regression med balanserad data

Vi fortsätter vår analys med att återigen använda metoden GLM, men nu på balanserad data. Vi har två olika balanserade data; *data 2* och *data 3*. Vi utgår för båda dessa från en modell med samtliga förklaringsvariabler utan samspel. Vi eliminerar variabler, en i taget, efter signifikansnivå. Vi tar i varje steg bort variabeln med högst p-värde. Vi upprepar detta förfarande tills vi når en modell där samtliga variabler är signifikanta på 5%-signifikansnivå.

Vi fortsätter därefter vår variabelselektion med att undersöka om det finns ett beroende mellan variabler. Detta gör vi genom att ta fram GVIF per förklaringsvariabel för våra modeller. Vi använder samma metod för våra två data. Vi ser VIF för modellen som använder *data 3* i tabell 4.2. Vi ser i tabellen att variablerna *Variabel 6* och *Variabel 7* når över gränsvärdet 5. Vi skapar två nya modeller, en där *Variabel 6* elimineras och en där *Variabel 7* elimineras. Vi väljer att fortsätta med den modell som har lägst AIC, vilket blir modellen där *Variabel 7* elimineras. Vi utför ett analysen på ett analogt sätt för *data 2*.

Variabel	*GVIF*
Variabel 4	1
Variabel 6	25
Variabel 7	24
Variabel 12	1
Variabel 13	2
Variabel 18	2
Variabel 19	1
Variabel 21	1
Variabel 30	2

Tabell 4.2: *Tabell över VIF för data 3.*

Efter att vi har modeller där alla variabler är signifikanta på 5%-signifikansnivå och där vi inte har en GVIF som överstiger 5 provar vi att lägga till samspel mellan variabler. Vi utgår från vår modell och lägger till alla möjliga samspel ett i taget. Vi observerar värdena på AIC och LRT samt att p-värdet för samspelet är signifikant. Detta resulterar i att vi för båda modellerna väljer att inkludera samspel mellan *Variabel 19* och *Variabel 18*.

När vi har vår slutliga modell tar vi fram sannolikheten för flytt för varje observation. Då vi har balanserat data behöver vi justera vår flyttsannolikhet. Närmare bestämt behöver vi justera konstanttermen eftersom vi har använt oss av logistisk regression. För att rätta till sannolikheterna använder vi formel (2.55) för att justera vår konstantterm som vi använder i ekvation (2.7). Detta gör vi både för modellen som använder *data 2* och *data 3*. Vi får samma värde på τ, andel 1:or i populationen, men olika värden på $\overline{y}$, andel 1:or i balanserade data, för dom två olika data.

4.4 GLMM

Vi fortsätter vår analys med att i vår logistiska regression ta hänsyn till att vi faktiskt har ett beroende i data. Data som används till analysen är longitudinell data, vilket innebär att vid en studie över tid kan ett subjekt kan förekomma flera gånger. Detta bryter mot ett av antagandena för GLM. För att hantera detta använder vi GLMM istället för GLM. Våra subjekt är försäkringsnummer och dessa kommer då få en slumpmässig effekt och varje försäkringsnummer får därmed en egen konstantterm. För att utföra GLMM använder vi *glmer* från paket *lme4* i R.

Vi börjar med att undersöka om det ger en effekt av att ha med försäkringsnummer som en slumpmässig effekt. Vi jämför då AIC för en modell med bara konstantterm mot en modell med som bara har konstantterm och försäkringsnummer som slumpmässig effekt. Resultaten ses i tabell 4.3.

Modell	*AIC*
Bara konstantterm	240 675
Konstantterm och försäkringsnummer	240 668

Tabell 4.3: *Tabell över AIC för modell med och utan försäkringsnummer som random effect.*

Vi ser att AIC minskar när vi inför försäkringsnummer som slumpmässig effekt vilket tyder på att den faktiskt har en påverkan. Dock ser vi att det är en väldigt liten skillnad. Även LRT visar att försäkringsnummer har en effekt då p-värdet på 0,00281 understiger 5%-signifikansnivå. Samma test görs på data där beroendet inte tas hänsyn till vid under sampling. Då visar det istället att AIC är något högre för modellen som använder GLMM mot modellen som använder GLM.

Det går att använda flera olika tekniker för parameterskattningar för GLMM. Vissa av dessa är väldigt tidskrävande och minneskrävande. Den enda metoden som fungerar på vårt data är när *number of adaptive Gauss-Hermite quadrature points* sätts till 0. Denna metod ger inte lika exakta resultat men gör att våra uträkningar går snabbare.

Vi utgår från balanserad data där vi har tagit hänsyn till att kontrakten har ett beroende, det vill säga det andra balanserade data som vi skapade, *data 3*. Vi använder metoden forward selection och jämför värden på AIC och BIC samt p-värden. Vi vill att AIC och BIC ska sjunka när den nya variabeln läggs till och att variabeln ska vara signifikant på 5 %-signifikansnivå.

När vi kommit fram till en modell där alla variabler är signifikanta undersöker vi att det inte finns ett beroende mellan variabler. Vi ser i tabell 4.4 att *Variabel 8* och *Variabel 29* har värden som överstiger 5. Vi provar därför att ta bort en av dom i taget för att se om GVIF-värdena då kommer att understiga 5. Då detta inte är fallet tar vi båda förklaringsvariablerna.

Variabel	*GVIF*
Variabel 8	48
Variabel 4	1
Variabel 12	1
Variabel 21	1
Variabel 18	4
Variabel 13	3
Variabel 30	2
Variabel 29	27

Tabell 4.4: *Tabell över VIF för GLMM på data 3.*

Vi fortsätter vår analys med att undersöka om vi kan lägga till samspel mellan variabler. Vi undersöker återigen om variabler ska läggas till genom att jämföra värden på AIC, BIC och om variabeln är signifikant på 5% signifikansnivå. Efter vår analys lägger vi till samspel mellan variablerna *Variabel 12* och *Variabel 18*.

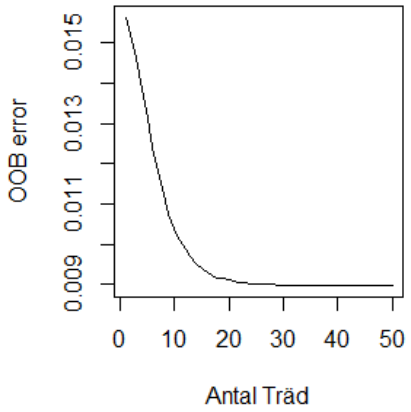
När vi har vår slutliga modell tar vi fram sannolikheten för flytt för varje observation. Då vi har balanserat data behöver vi justera vår flyttsannolikhet. Eftersom vi har använt oss av logistisk regression är det konstantterm som vi behöver justera. För att rätta till sannolikheterna använder vi *coef()* för att få fram konstanttermen för varje subjekt justerat för dess slumpmässiga effekt. Vi använder sedan formel (2.55) precis som i fallet för GLM, med skillnaden att varje subjekt nu har en egen konstantterm, istället för en gemensam för alla subjekt.

4.5 Random forest

Slutligen analyserar vi flyttar med random forest. Vi utför analysen på samtliga tre data. Innan vi skapar modellen tar vi tar bort försäkringsnummer från data. Detta gör vi eftersom vi anser att det inte är en lämplig förklaringsvariabel. Även om den hade varit det innehåller variabeln försäkringsnummer för många nivåer för att den ska kunna hanteras av modellen. Vi utför random forest med *randomForest* från paketet med samma namn i R.

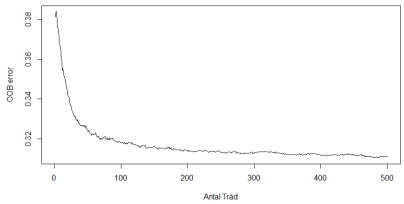
När vi sätter upp random forest i R får vi göra några val av inställningar för algoritmen. Vi måste välja hur många variabler som algoritmen kommer att välja mellan vid skapande av varje gren. Vi sätter så som brukligt antalet $m = \sqrt{p} = 5$. Detta innebär att vid varje gren kommer 5 olika variabler slumpmässigt väljas fram

som möjliga variabler. Detta innebär att alla träd som vi skapar kommer vara olika då olika variabler har varit möjliga alternativ vid skapandet av varje gren i dom olika träden. Vi måste även välja hur många träd vi vill skapa i modelleringen. Vi sätta antalet träd till ett relativt högt tal (vi behöver inte oroa oss för överanpassning enligt (James et al. 2013)). Detta gör vi för att vara säkra på våra skattningar, men vi vill inte att beräkningarna ska ta allt för lång tid. För data som inte är balanserat, *data 1*, som är ett stort data, kan vi inte använda många träd utan vi får nöja oss med 50. Detta verkar dock vara ett tillräckligt antal, vilket vi kan se i figur 4.3. Vi ser i figuren att felet för OOB konvergerar innan 50.

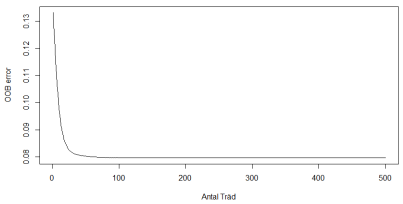


Figur 4.3: *Kurva över hur felet för OOB minskar med antal träd för data 1.*

För *data 2* och *data 3* använder vi standardvärdet 500 träd i modelleringen. Vi undersöker även att detta är ett tillräckligt antal genom att okulärbesiktiga figurerna 4.4 och 4.5. Vi ser i figurerna att felet för OOB konvergerar innan 500 träd och att vi därför har ett tillräckligt stort antal träd för både *data 2* och *data 3*.



Figur 4.4: *Kurva över hur felet för OOB minskar med antal träd för data 2.*



Figur 4.5: *Kurva över hur felet för OOB minskar med antal träd för data 3.*

Vi använder funktionen *predict* med specifikationen $type = \ 'prob'$ i R för att prediktera flyttar för vårt testdata. Vi får genom funktionen en flyttsannolikhet för varje kontrakt i testdata. För *data 2* och *data 3* behöver vi sedan konvertera flyttsannolikheten till en obalanserad flyttsannolikhet. Detta görs genom ekvation (2.51).

4.6 Simulering av data

För att utvärdera våra modeller utför vi även en simulering av data som vi sedan använder för att bygga olika modeller på.

Vi börjar med att rensa data. Kontrakt som försvinner av andra anledningar en flytt tas bort. Vi får då ett data som utgår från år 2010. Vi utgår därefter från vår modell där vi använder GLM. Vi får genom den fram sannolikhet för flytt per kontrakt. Vi slumpar flyttar utifrån våra beräknade flyttsannolikheter genom att dra från en binomialfördelning med dom beräknade flyttsannolikheter (*rbinom* i R). Vi låter dessa kontrakt försvinna från vår data så att det inte förekommer under tidsperioder efter det "flyttat". Vid nästa år tillkomma nya kontrakt. Vi utför förfarandet på samma sätt: utifrån sannolikheterna som vi fått från vår modell flyttar kontrakt slumpmässigt och tas bort från data. Vi upprepar detta förfarande till för samtliga år fram till och med år 2019. Genom att simulera data på det här sättet har vi rensat bort effekter av att kontrakt kanske flyttar på grund av andra anledningar än dom parameterspecifika, som erbjudanden från andra bolag, regelverk etc.

Vi delar även in det här data i en del för att bygga modellen (år 2010-2018) och en del för att testa modellen (år 2019). Utifrån vårt nya data bygger vi sedan modeller med metoderna GLM, GLMM och random forest. Från resultaten från dom nya modellerna predikterar vi sedan sannolikhet för flytt på vårt data. Vi simulerar sedan detta 10 000 gånger på vår testdata och tar medelvärden av resultaten. Vi utför även samma test på vårt data som vi byggt modellen med. Där får vi, på grund utav storleken på data, nöja oss med 1 000 simuleringar.

Kapitel 5

Resultat

5.1 GLM

Vi börjar vår analys av resultaten med att jämföra GLM på *data 1* mot GLM på *data 2*. Modellerna med skattade parametrar presenteras i tabellerna C.1 och C.2 i Appendix C.

Sannolikheter

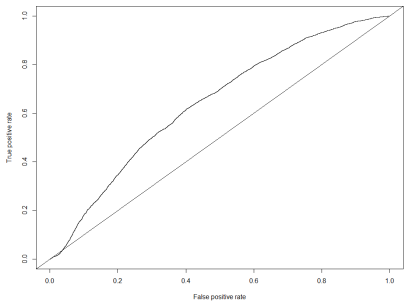
Vi ser i tabell 5.1 sannolikheten för flytt när vi använder våra modeller för GLM på *data 1* och GLM på *data 2* för att förutsäga flyttar på vårt testdata. Vi ser i tabellen att vi har något högre sannolikheter när vi använder modellen som gjordes på vår balanserade data.

Modell	*Min*	*Median*	*Medelvärde*	*Max*
GLM på *Data 1*	0,0004	0,0063	0,0075	0,0399
GLM på *Data 2*	0,0000	0,0063	0,0086	0,0545

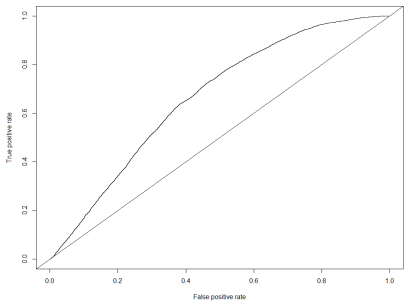
Tabell 5.1: *Tabell över summering av sannolikheter för flytt för metoden GLM.*

ROC-kurvor & AUC

ROC-kurvorna i figurerna 5.1 och 5.2 visar att vi har en något större yta under kurvan när vi utför GLM på *data 2* mot när vi utför GLM på obalanserad data, *data 1*.



Figur 5.1: *ROC-kurva för GLM på Data 1*



Figur 5.2: *ROC-kurva för GLM på Data 2*

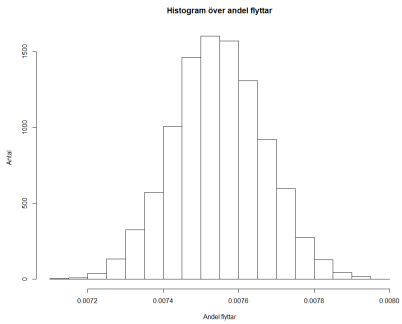
Detta bekräftas även i tabell 5.2 där vi ser att vi har gått från ett värde på AUC på 0,64 till 0,67 när vi balanserar data.

Modell	*AUC*
GLM på *Data 1*	0,64
GLM på *Data 2*	0,67

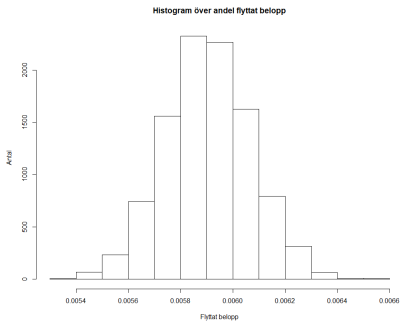
Tabell 5.2: *Tabell över AUC för dom olika modellerna.*

Simulering av 10 000 utfall

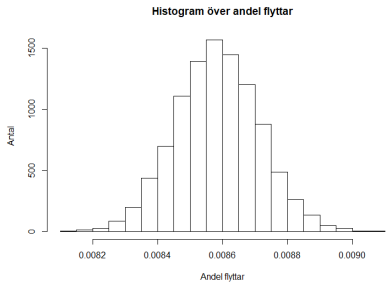
När vi utför 10 000 simuleringar baserade på sannolikheterna från dom olika modellerna får vi ett medelvärde om 0,75% för antal flyttar och 0,59 % för flyttat belopp för GLM på *Data 1* och ett medelvärde om 0,86% för antal flyttar och 0,73% för flyttat belopp för GLM på *Data 2*. Vi ser i figurer 5.3 och 5.4 att vi underskattar både antalet flyttar och andelen flyttbart belopp. Vi når inte upp till det verkliga utfallet, som uppgår till 1,05% respektive 0,77%, i någon av dom 10 000 simuleringarna. Det gör vi däremot i några av våra simuleringar när vi gör 10 000 simuleringar för flyttat belopp för modellen som bygger på *data 2*, vilket vi ser i figur 5.6 där den blåa linjen representerar faktiskt flyttat belopp.



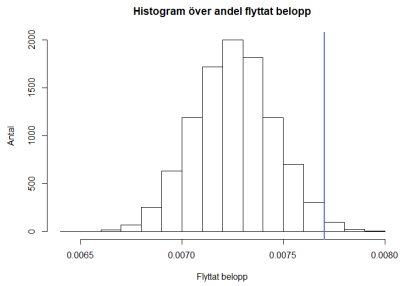
Figur 5.3: *10 000 simuleringar av flyttat andel antal utifrån resultaten från GLM på Data 1.*



Figur 5.4: *10 000 simuleringar av andel flyttat belopp utifrån resultaten från GLM på Data 1.*



Figur 5.5: *10 000 simuleringar av flyttat andel antal utifrån resultaten från GLM på Data 2.*



Figur 5.6: *10 000 simuleringar av andel flyttat belopp utifrån resultaten från GLM på Data 2.*

5.2 GLM & GLMM på *Data 3*

Vi fortsätter vår analys av resultaten genom att jämföra GLM mot GLMM. För båda metoderna använder vi *data 3*. Modellen med skattade parametrar som gjordes med GLMM presenteras i tabell C.4 och modell för GLM presenteras i tabell C.3 i Appendix C.

Sannolikheter

Tabellen 5.3 visar att både genom att använda GLM och GLMM får vi låga sannolikheter för flyttar när vi använder modellerna för att prediktera flyttar på vår

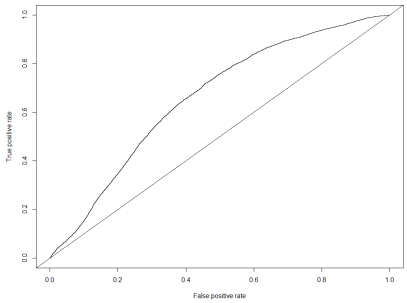
testdata. Vi ser att vi får ett högre maxvärde för sannolikheten genom att använda GLM.

Modell	*Min*	*Median*	*Medelvärde*	*Max*
GLM på *Data 3*	0,0013	0,0072	0,0081	0,0178
GLMM på *Data 3*	0,0017	0,0085	0,0083	0,0141

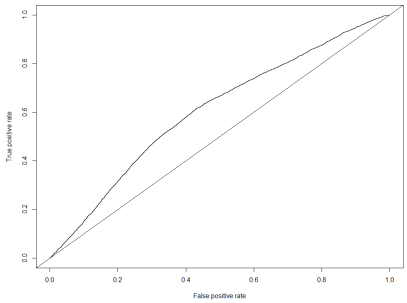
Tabell 5.3: *Tabell över summering av sannolikheter för flytt.*

ROC-kurvor & AUC

ROC-kurvorna i figurerna 5.7 och 5.8 visar att vi får en större AUC när vi använder GLM jämfört mot GLMM.



Figur 5.7: *ROC-kurva för GLM på Data 3.*



Figur 5.8: *ROC-kurva för GLMM på Data 3.*

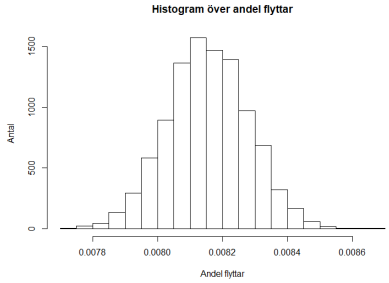
Detta bekräftas även i tabell 5.4 där vi ser att GLM ger en AUC på 0,66 medan GLMM ger en AUC på 0,61.

Modell	*AUC*
GLM på *Data 3*	0,66
GLMM på *Data 3*	0,61

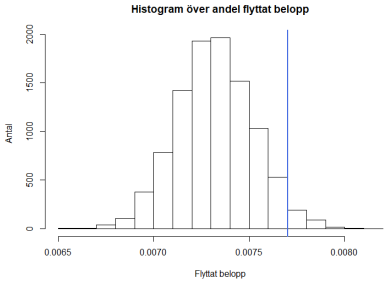
Tabell 5.4: *Tabell över AUC för dom olika modellerna.*

Simulering av 10 000 utfall

Vi ser resultatet av simulering av 10 000 utfall ifrån sannolikheter från modellen som använder GLM med *Data 3* i figurerna 5.9 och 5.10. För simuleringarna för antalet flyttar underskattar vi antalet i samtliga simuleringar. Vi får 0,82% som medelvärde, medan verkligt utfall är 1,05%. Vi underskattar även beloppet i majoriteten av våra simuleringar, men som vi kan se i grafen i figur 5.10 träffar vi rätt i några. Medelvärdet för våra simuleringar är 0,73% och den verkliga andelen av beloppet, 0,76%.

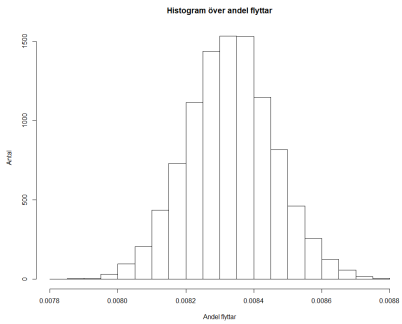


Figur 5.9: *10 000 simuleringar av flyttat andel antal utifrån resultaten från GLM på Data 3.*

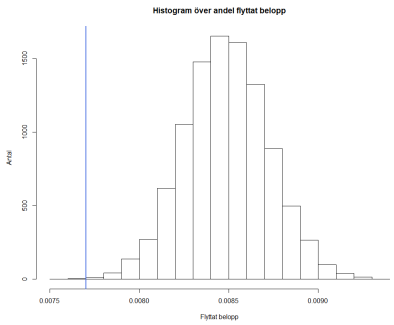


Figur 5.10: *10 000 simuleringar av andel flyttat belopp utifrån resultaten från GLM på Data 3.*

Vi ser i Figurerna 5.9 och 5.11 resultaten av 10 000 simuleringar för antal flyttar för modellen som använder GLMM med *Data 3*. Även för denna kommer vi inte upp i den verkliga andelen för antalet flyttar. Vi får 0,83% som medelvärde, medan verkligt utfall är 1,05%. Däremot överskattar vi beloppet för antal flyttar i majoriteten av simuleringarna som vi kan se i grafen i figur 5.12. Den blå linjen representerar verkligt utfall. Medelvärdet för våra simuleringar är 0,85% medan verkligt utfall är 0,76%.



Figur 5.11: *10 000 simuleringar av flyttat andel antal utifrån resultaten från GLMM på Data 3.*



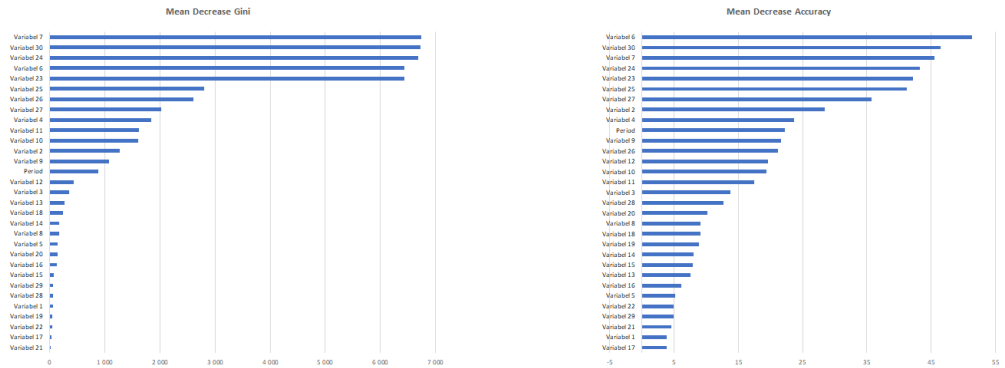
Figur 5.12: *10 000 simuleringar av andel flyttat belopp utifrån resultaten från GLMM på Data 3.*

5.3 Random forest

Till sist analyserar vi resultaten från våra modeller med random forest. Vi börjar med att analysera hur stor betydelse varje variabel har i modelleringen.

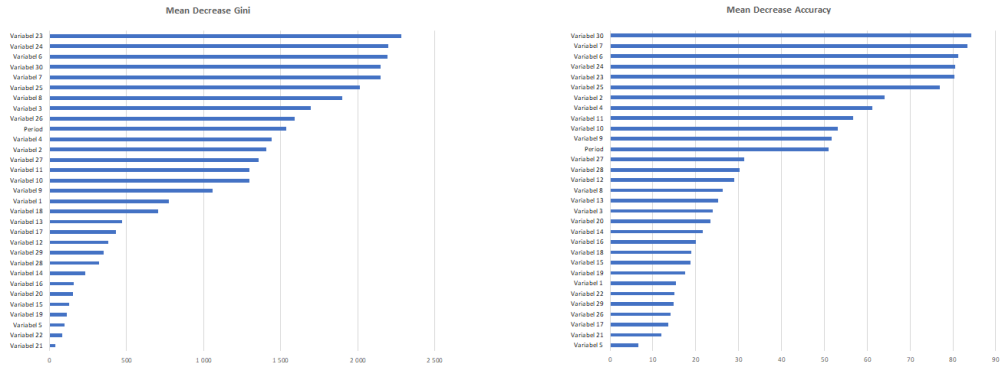
Variabels påverkan

Hur stor betydelse varje variabel har i modelleringen för *data 1* ses i bild 5.13. Vi kan avläsa från grafen till vänster i figuren att variablerna *Variabel 7*, *Variabel 30*, *Variabel 24*, *Variabel 6* och *Variabel 23* har högst värden. Detta innebär att dessa har en stor betydelse och att när vi placerar grenar för dessa minskar Gini indexet som mest. Vi ser att samma variabler även dyker upp i grafen till höger som är ett annat mått som visar hur mycket exaktheten minskar när en variabel utelämnas. Detta innebär att exaktheten minskar som mest när *Variabel 6* utelämnas.



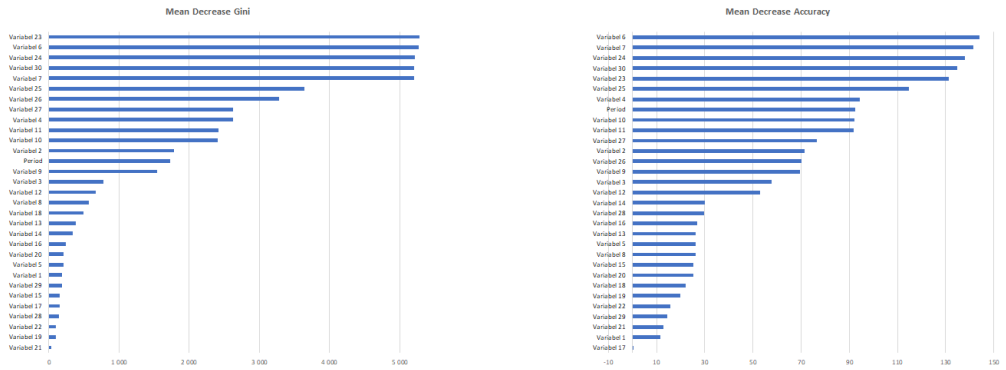
Figur 5.13: *Variablers påverkan för Data 1.*

På samma ser vi hur stor betydelse varje variabel har i modelleringen för *data 2* i Figur 5.14. Vi kan avläsa från grafen till vänster i figuren att variablerna *Variabel 23*, *Variabel 24*, *Variabel 6* och *Variabel 30* har högst värden. I den högra grafen i figuren ser vi att samma variabler har en stor betydelse för exaktheten, men nu i en annan ordning.



Figur 5.14: *Variablers påverkan för Data 2.*

Slutligen ser vi på samma sätt hur stor betydelse varje variabel har i modelleringen för *data 3* i Figur 5.15. Vi kan avläsa från grafen till vänster i figuren att variablerna *Variabel 23*, *Variabel 6*, *Variabel 24*, *Variabel 30* och *Variabel 7* har högst värden. Precis som tidigare ser vi att samma variabler dyker upp i den högra grafen, men i en annan ordning.



Figur 5.15: *Variablers påverkan för Data 3.*

Klassifikationstabell för OOB-observationer

Vid skapande av beslutsträden i random forest görs en klassifikationstabell på våra OOB-observationer. Vi ser klassifikationstabellen för modellen som använder *data 1* i tabell 5.5. Vi ser i tabellen att vi predikterar rätt för om kontraktet flyttar i mer än hälften av utfallen (64%), vilket är bättre än vi hade slumpat utfallen. Vi predikterar rätt i 99 % av fallen för dom kontrakt som inte har flyttat.

		Observerade	
		0	**1**
Predikterade	**0**	3 889 609 (99%)	28 (36%)
	1	35 187 (1%)	49 (64%)

Tabell 5.5: *Klassifikationstabell Random Forest på OOB för Data 1.*

Vi ser klassifikationstabellen för våra OOB-observationer för modellen som använder *data 2* i tabell 5.6. Vi har en bättre prediktion av flyttar i det här fallet, 68% av dom som har flyttat har vi predikterat som flyttar. Vi ser dock att vi har ett sämre utfall för kontrakten som inte har flyttat. Då predikterar vi att 27% av dessa kontrakt kommer att flytta.

		Observerade	
		0	**1**
Predikterade	**0**	22 776 (73%)	12 460 (32%)
	1	8 376 (27%)	26 860 (68%)

Tabell 5.6: *Klassifikationstabell Random Forest på OOB för Data 2.*

Vi ser klassifikationstabellen för våra OOB-observationer för modellen som använder *data 3* i tabell 5.7. I tabellen ser vi att vi lyckas prediktera verkliga flyttar i 59% av fallen. När vi predikterar kontrakt som inte har flyttat vi bättre än för *data 2* då vi predikterar rätt i 91% av fallen.

		Observerade	
		0	**1**
Predikterade	**0**	362 094 (91%)	227 (41%)
	1	35 460 (9%)	324 (59%)

Tabell 5.7: *Klassifikationstabell Random Forest på OOB för Data 3.*

Sannolikheter

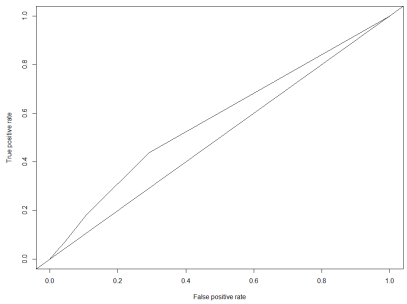
Vi ser sannolikheter för flytt för dom olika modellerna i tabell 5.8. Vi ser att vi får en låg sannolikhet för flytt för modellen som använder *data 1* och högre sannolikheter för modellerna som använder *data 2* och *data 3*.

Modell	*Min*	*Median*	*Medelvärde*	*Max*
Random forest på *Data 1*	0,0000	0,0000	0,0095	0,48000
Random forest på *Data 2*	0,0000	0,0065	0,0086	0,2265
Random forest på *Data 3*	0,0000	0,0096	0,0118	0,1497

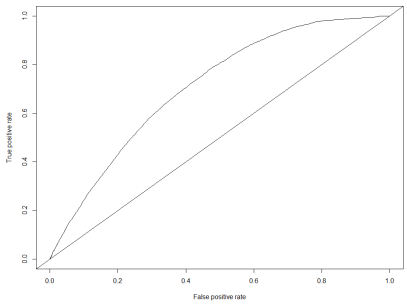
Tabell 5.8: *Tabell över sannolikheter för flytt.*

ROC-kurvor & AUC

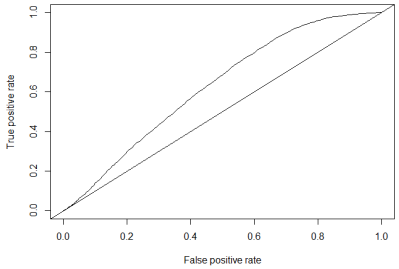
I Figurerna 5.16, 5.17 och 5.18 ser vi ROC-kurvor för vår testdata för dom olika modellerna. Vi ser att vi kurvan för *data 2* som visas i Figur 5.17 ger störst area under kurvan följt av kurvan för *data 3* som ses i Figur 5.18. Detta bekräftas av när vi beräknar AUC för dom tre modellerna som återfinns i tabell 5.9.



Figur 5.16: *ROC-kurva för Random forest på Data 1.*



Figur 5.17: *ROC-kurva för Random forest på Data 2.*



Figur 5.18: *ROC-kurva för Random forest på Data 3.*

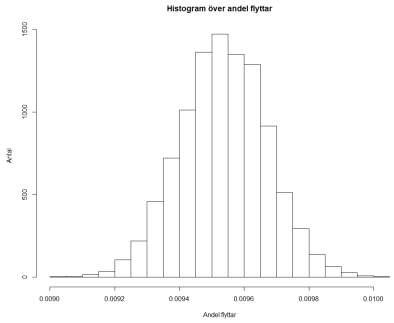
Modell	*AUC*
Random forest på *Data 1*	0,58
Random forest på *Data 2*	0,71
Random forest på *Data 3*	0,63

Tabell 5.9: *Tabell över AUC för modellerna.*

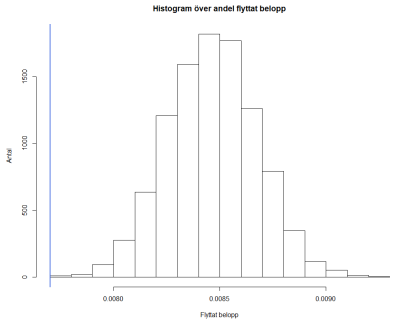
Simulering av 10 000 utfall

Vid simulering av 10 000 utfall ifrån sannolikheter från modellen får vi utfallen för andel av antalet kontrakt som flyttar och andel flyttbart belopp för *data 1* i figurerna 5.19 och 5.20. För andelen av antalet kontrakt underskattar vi den verkliga andelen av antalet. Medelvärde för dom 10 000 simuleringarna är 0,95% medan verkligt utfall är 1,05%. Vi ser i figur 5.19 att vi inte kommer upp till den verkliga utfallet i någon av dom 10 000 simuleringarna. När vi istället simulerar flyttat belopp överskattar vi

beloppet vilket vi ser i figur 5.20. Medelvärdet för våra simuleringar är 0,85% medan verkligt utfall är 0,77%. Vi ser dock att vi i några simuleringar kommer fram till rätt belopp, som representeras av linjen i grafen.

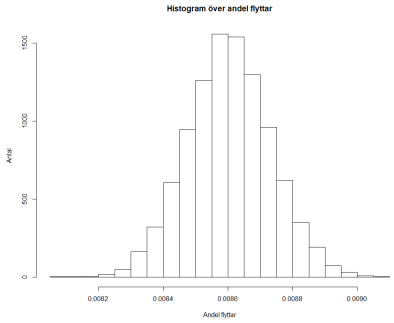


Figur 5.19: *10 000 simuleringar av flyttat andel antal utifrån resultaten från Random forest på Data 1.*

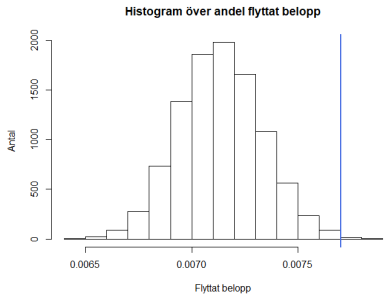


Figur 5.20: *10 000 simuleringar av andel flyttat belopp utifrån resultaten från Random forest på Data 1.*

Vi ser utfallen av simulering av 10 000 ifrån sannolikheter från modellen för modellen som använder *data 2* i Figurerna 5.21 och 5.22. För både andelen av antalet kontrakt och andelen av flyttbart belopp underskattar vi dom verkliga utfallen. För antal flyttar når vi aldrig upp till rätt antal. Medelvärde för dom 10 000 simuleringarna är 0,86% medan verkligt utfall är 1,05%. För flyttat belopp når vi däremot upp till verkligt flyttat belopp, vilket vi kan se i grafen i figur 5.22. Vi ser dock att vi ligger något lägre i dom flesta av simuleringarna. Medelvärde för dom 10 000 simuleringarna är 0,71% medan verkligt utfall är 0,77%.



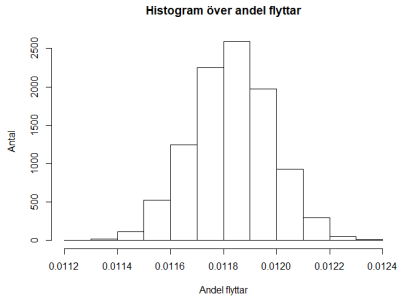
Figur 5.21: *10 000 simuleringar av flyttat andel antal utifrån resultaten från randomforest på Data 2.*



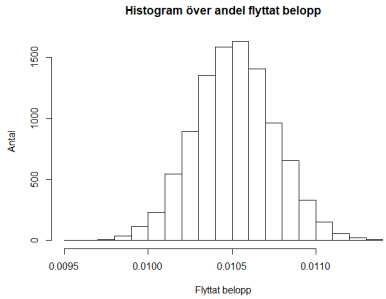
Figur 5.22: *10 000 simuleringar av andel flyttat belopp utifrån resultaten från randomforest på Data 2.*

För modellen där vi använt *data 3* kan vi se i Figurerna 5.23 och 5.24 att vi överskattar både andelen av antalet kontrakt och andelen av flyttbart belopp.

Medelvärde för dom 10 000 simuleringarna för antal kontrakt är 1,18% medan verkligt utfall är 1,05%. Medelvärde för dom 10 000 simuleringarna för belopp är 1,05% medan verkligt utfall är 0,77%. Vi ser i figurerna att både för antal och belopp överskattar vi i samtliga simuleringar.



Figur 5.23: *10 000 simuleringar av flyttat andel antal utifrån resultaten från randomforest på Data 3.*



Figur 5.24: *10 000 simuleringar av andel flyttat belopp utifrån resultaten från randomforest på Data 3.*

5.4 Jämförelser mellan modeller

Vi ser en sammanställning av medelvärde och standardavvikelser för 10 000 simuleringar från samtliga modeller på dom olika data i tabell 5.10.

Modell	Antal flyttar		Flyttat belopp	
	$\overline{x}$	σ	$\overline{x}$	σ
Verkligt (testdata)	1,05%	-	0,77%	-
GLM på *Data 1*	0,7547%	0,0119%	0,5904%	0,0164%
GLM på *Data 2*	0,8585%	0,0128%	0,7257%	0,0196%
GLM på *Data 3*	0,8155%	0,0125%	0,7321%	0,0199%
GLMM på *Data 3*	0,8332%	0,0127%	0,8486%	0,0234%
Random forest på *Data 1*	0,9537%	0,0133%	0,8464%	0,0209%
Random forest på *Data 2*	0,8604%	0,0129%	0,7148%	0,0195%
Random forest på *Data 3*	1,1830%	0,0150%	1,0515%	0,0241%

Tabell 5.10: *Tabell över medelvärden från 10 000 simuleringar för dom olika modellerna.*

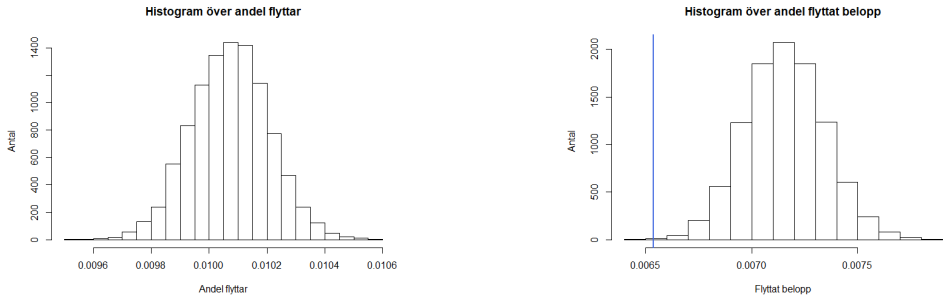
Vi kan avläsa i tabellen att vi inte har någon träffsäker skattning av andelen av antalet. Vi underskattar antalet för samtliga modeller, förutom random forest på

data 3 där vi överskattar antalet. För beloppet kommer vi närmare det verkliga utfallet för samtliga metoder, men det är i fyra modeller som vi träffar det verkliga beloppet i några av simuleringarna; random forest på *data 2*, GLM *data 2*, GLM *data 3* och GLMM *data 3*. Får alltså en bättre modell när vi balanserar data.

Vi ser att vi med alla tre metoder får ett högre medelvärde av skattningarna för både antal och belopp när vi använder *data 3*.

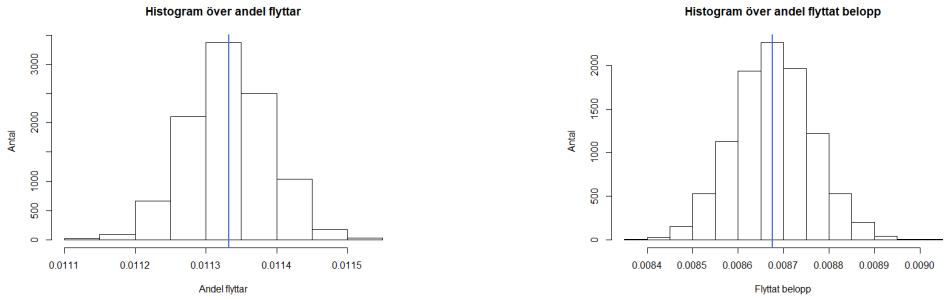
5.5 Resultat från simulering

I figur 5.25 ser vi resultaten av 10 000 simuleringar för antal flyttar och flyttbart belopp för modellen som använder GLM och där vi applicerar resultaten från denna på vår testdata. Både för antal flyttar (1,01% mot 0,89%) och för flyttbart belopp (0,72% mot 0,65%) får vi ett högre medelvärde än det faktiska utfallet.



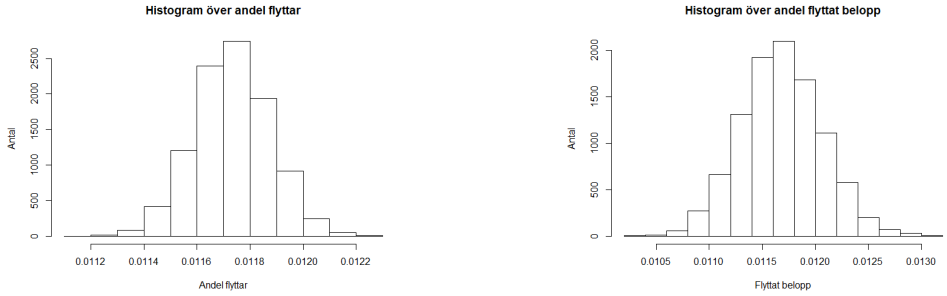
Figur 5.25: *Resultat från simuleringar av flyttat andel antal och belopp på testdata utifrån resultaten av GLM på simulerad data.*

I figur 5.26 ser vi resultaten av 1 000 simuleringar för antal flyttar och flyttbart belopp för modellen som använder GLM och där vi applicerar resultaten från denna på vår modellbyggedata. Vi ser i figuren att vi träffar det verkliga utfallet väldigt bra för både antal flyttar (1,13% mot 1,13%) och för flyttbart belopp (0,87% mot 0,87%).



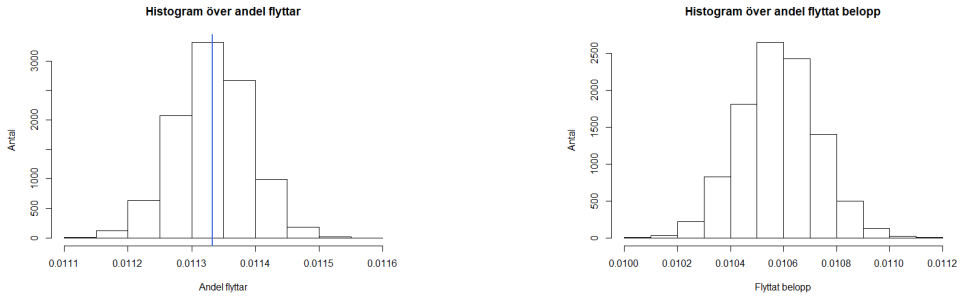
Figur 5.26: *Resultat från simuleringar av flyttat andel antal och belopp på modellbyggedata utifrån resultaten av GLM på simulerad data.*

I figur 5.27 ser vi resultaten av 10 000 simuleringar för antal flyttar och flyttbart belopp för modellen som använder GLMM och där vi applicerar resultaten från denna på vår testdata. Både för antal flyttar (1,17% mot 0,89%) och för flyttbart belopp (1,16% mot 0,65%) får vi ett högre medelvärde än det faktiska utfallet.



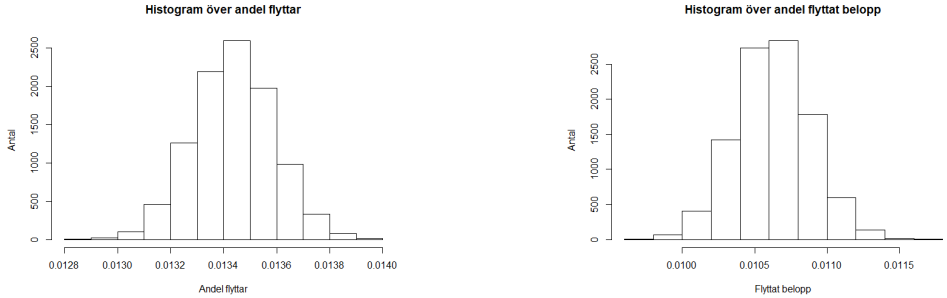
Figur 5.27: *Resultat från simuleringar av flyttat andel antal och belopp på testdata utifrån resultaten av GLMM på simulerad data.*

I figur 5.28 ser vi resultaten av 1 000 simuleringar för antal flyttar och flyttbart belopp för modellen som använder GLMM och där vi applicerar resultaten från denna på vår modellbyggedata. Vi ser i figuren att vi kommer nära det verkliga antalet (1,34% mot 1,13%), men att vi får ett högre medelvärde än det faktiska utfallet för flyttbart belopp (1,06% mot 0,87%) .



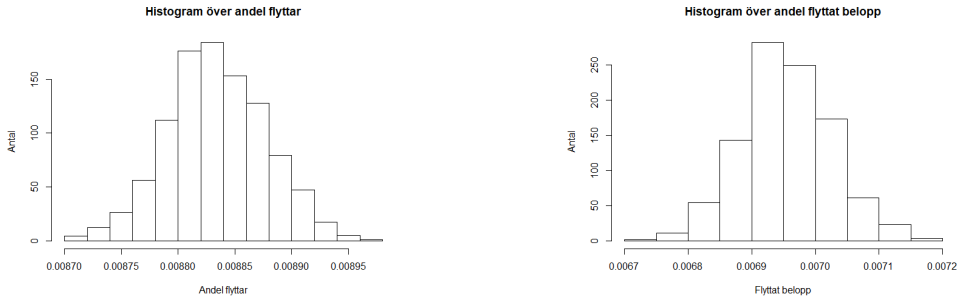
Figur 5.28: *Resultat från simuleringar av flyttat andel antal och belopp på modellbyggedata utifrån resultaten av GLMM på simulerad data.*

I figur 5.29 ser vi resultaten av 10 000 simuleringar för antal flyttar och flyttbart belopp för modellen som använder Random forest och där vi applicerar resultaten från denna på vår testdata. Både för antal flyttar (1,34% mot 0,89%) och för flyttbart belopp (1,06% mot 0,65%) får vi ett högre medelvärde än det faktiska utfallet.



Figur 5.29: *Resultat från simuleringar av flyttat andel antal och belopp på testdata utifrån resultaten av Random forest på simulerad data.*

I figur 5.30 ser vi resultaten av 1 000 simuleringar för antal flyttar och flyttbart belopp för modellen som använder Random forest och där vi applicerar resultaten från denna på vår modellbyggedata. Både för antal flyttar (0,88% mot 1,13%) och för flyttbart belopp (0,70% mot 0,87%) får vi ett lägre medelvärde än det faktiska utfallet.



Figur 5.30: *Resultat från simuleringar av flyttat andel antal och belopp på modellbyggedata utifrån resultaten av Random forest på simulerad data.*

Modell	Modellbyggedata		Testdata	
	Andel av antal flyttar	**Andel av flyttbart belopp**	**Andel av antal flyttar**	**Andel av flyttbart belopp**
Verkligt utfall	1,1333	0,8676	0,8883	0,6537
GLM	1,1332	0,8678	1,0073	0,7156
GLMM	1,3437	1,0627	1,1732	1,168027
Random forest	0,8834	0,6954	1,3437	1,0627

Tabell 5.11: *Tabell över medelvärden av resultaten från simulering.*

Vi kan avläsa i tabell 5.11 att modellerna i samtliga fall presterar bättre för modellbyggedata mot data för år 2019.

Kapitel 6

Slutsatser & diskussion

Vi har i det här arbetet undersökt tre olika metoder för att förutsäga flyttar och använt både originaldata och balanserad data. Vi har analyserat resultaten både efter hur bra modellerna har varit på att förutsäga andelen belopp och andelen flyttar. Vi lyckas inte pricka rätt i antalet för några av våra modeller, men vi lyckas bättre med beloppet som kan anses vara viktigare att prediktera rätt. För beloppet kommer vi närmare det verkliga utfallet för samtliga metoder, men det är i fyra modeller som vi träffar det verkliga beloppet i några av simuleringarna; random forest på *data 2*, GLM *data 2*, GLM *data 3* och GLMM *data 3*. Vi får alltså en bättre modell när vi balanserar data. I det här arbetet har data balanserats med metoden under sampling. En stor nackdel med den här metoden är att vi tappar information när vi tar bort observationer.

Den modell som ger bäst utfall är GLM på *data 2*. Då lyckas vi komma närmast andelen flyttbart belopp. Som vi ser i grafen i figur 5.10 träffar vi verkligt belopp i svansen till höger av våra simuleringar, d.v.s. vi underskattar generellt nivån av flyttbart belopp som flyttar. Det innebär att om vi använde den här nivån skulle vi underskatta hur mycket som faktiskt flyttar, vilket inte är önskvärt. Då kanske det vore ett bättre alternativ att använda en modell som överskattar beloppet.

Vi får högst AUC random forest på *data 2* (0,71), men predikterar en lägre andel flyttbart belopp än när vi använder GLM på samma data. En nackdel med random forest är att den har svårt att anpassa sig efter nya nivåer på data. Om vi skulle t.ex. få en ny flyttavgift så skulle inte den kunna anpassa sig på samma sätt som en regressionsmodell. För att använda den här metoden skulle mer analyser behöva göras och vi skulle vilja säkerställa att modellen är stabil över tiden.

Det är svårt att avgöra om GLMM ger ett bättre resultat än GLM då vi använde en förenklad metod för GLMM. I både GLM och GLMM träffar vi rätt i beloppet i några av våra simuleringar. För GLMM överskattar vi dock beloppet i majoriteten av simuleringarna medan vi underskattar beloppet i majoriteten av simuleringarna när vi använder GLM. Vi får en högre AUC och medelvärdet av våra simuleringar kommer närmare det verkliga utfallet för flyttat belopp när vi använder GLM vilket innebär, att med den metoden vi använde för GLMM, är GLM ett bättre alternativ för att prediktera flyttar.

Analysen med GLMM har stora förbättringsmöjligheter. Det är främst två hinder i det här arbetet. Det första är att vi inte fritt kunde anpassa en modell genom att ha ett högre värde på *nagq* på grund av minnesproblematiken. Med nuvarande metod blir resultaten bara approximativa. Ett annat problem är att vi kan ha haft för få observationspunkter per försäkringsnummer. Det hade kunnat vara en lämpligare indelning av att använda person/företag som subjekt och då få med alla försäkringar som subjektet äger. Då skulle det kanske vara möjligt att se mönster som att t.ex en försäkring sägs upp vid det tillfället då det inte längre finns en avgift. Ytterligare ett annat alternativ hade kunnat vara att använda mäklarbolag som subjekt då dessa skulle kunna ha olika beteenden.

När vi simulerade data fick vi bäst resultat med GLM, vilket inte är överraskande - vi använder ju simulerad data utifrån en modell med GLM. Då GLMM är närbesläktat med GLM är det inte förvånande att den ger näst bäst resultat. För modellerna när vi använder GLM och GLMM presterade modellerna betydligt bättre på modellbyggedata än på testdata. Då testdata är data från ett år som inte ingår i modellbyggedatat kan det vara så att det är något speciellt som inträffar det året - ett nytt regelverk, förändring av produkter, eller andra anledningar. Vi har inte gjort någon extra analys av anledningen till detta.

Det finns flera förbättringsområden för det här arbetet. Dels går det att göra modeller med bättre data då kvaliteten på data inte är optimal. Även andra typer av variabler hade kunnat vara intressanta att undersöka. Flera artiklar har undersökt makroekonomiska variabler som arbetslöshet, inflation etc. Dessa variabler kanske hade gett en ännu bättre modell än att bara använda kontraktsspecifika variabler. Det finns troligen även ett samband mellan andra bolags erbjudanden och försäkringstagarnas beslut om att flytta. Sådana variabler är dock svåra att fånga i en modell och det går inte att förutsäga hur andra bolag ska agera i framtiden.

Ett stort problem i det här arbetet har varit att RAM-minnet tar slut vid

vissa beräkningar i R. Detta har gjort att alla metoder inte har kunnat testas eller behövt begränsas. Speciellt har detta varit ett problem när *Data 1* har använts och när vi anpassade modellen med GLMM.

Ett annat problem i det här arbetet är att det är svårt att utvärdera modellerna med bara ett år att prediktera på. Att använda fler år att testa på hade varit ett alternativ, men det kvarstår fortfarande problem med av olika skäl kan testdata skilja sig åt mot åren som modellen byggdes på. Omvärlden förändras ständigt med nya regelverk och förändrade beteenden hos både kunder och konkurrenter etc. Då hade ett bättre alternativt kunna vara att dela upp data i 30% testdata och 70% modellbyggedata.

Vi skulle även behöva göra en analys på vilka kontrakt som den klassificerar som flyttar, även om det är sekundärt att vi i våra klassifikationstabeller inte träffar precis rätt, finns det ett intresse av att pricka in rätt kontrakt som flyttar om det handlar om stora flyttbara belopp. Om vi lyckas felklassificera alla kontrakt med större flyttbara belopp så kan detta innebära att vi får helt fel andelar. Det skulle kunna vara en slump som gör att vi för vår testdata lyckades få någorlunda rätt andel, men för ett annat år skulle resultaten kunna se annorlunda ut.

Slutligen hade det varit intressant att prova andra metoder, som exempelvis överlevnadsanalys, men även andra tillvägagångssätt inom dom metoder som har undersökts i detta arbete. T.ex. finns det finns andra metoder för att balansera data som kanske hade gett ett bättre resultat eller att hitta en metod för att komma runt problem med att arbetsminnet för datorn tar slut.

Appendix A

Minstakvadratskattning

Vi definierar $x_1, ..., x_n$ som ett slumpmässigt stickprov från en slumpvariabel, X. X har väntevärdet $m(\theta)$, där m är en känd funktion. Minstakvadratskattningen av θ är det värde, som vi betecknar θ^*, som minimerar $Q(\theta)$ i ekvation (A.1).

$$Q(\theta) := \sum_{i=1}^{m} (x_i - m(\theta))^2 \tag{A.1}$$

Om observationerna kommer från olika fördelningar behöver vi modifiera (A.1) till ekvation (A.2). Då tar vi hänsyn till att x_i är en observation av X_i som har väntevärde $m_i(\theta)$ och med standardavvikelse σ.

$$Q(\theta) := \sum_{i=1}^{m} (x_i - m_i(\theta))^2 \tag{A.2}$$

I ekvation A.2 är standardavvikelsen samma för alla i. När vi har olika standardavvikelser vill vi vikta minstakvadratskattningen. Detta eftersom en observation med ett stor standardavvikelse inte ger samma precision som en med liten standardavvikelse. Vi inför därför $w_i = \lambda/\sigma_i^2$ där λ är en godtycklig konstant som är oberoende av θ. Vi får då en viktad minstakvadratskattning som ges av ekvation (A.3)

$$Q(\theta) := \sum_{i=1}^{m} w_i (x_i - m_i(\theta))^2 \tag{A.3}$$

(Alm and Britton 2008).

Uttryckt i matrisform, när vi vill använda den viktade minstakvadratskatt-

ningen för att skatta β, får vi ekvation (A.4).

$$(\boldsymbol{Y} - \boldsymbol{X\beta})^T \boldsymbol{W} (\boldsymbol{Y} - \boldsymbol{X\beta}) \tag{A.4}$$

Vi kan då skriva om enligt normalekvationerna enligt ekvation (A.5) och sedan lösa ut β.

$$(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{Y} \tag{A.5}$$

På samma sätt kan vi skriva när vi är intresserade av att skatta b istället för β. Vi får då normalekvationen (A.6).

$$(\boldsymbol{Z}^T \boldsymbol{W} \boldsymbol{Z})\boldsymbol{b} = \boldsymbol{Z}^T \boldsymbol{W} \boldsymbol{z} \tag{A.6}$$

(Doran et al. 2007).

Appendix B

Variabler

Variabelnamn	*Typ av variabel*
Period	Datum
Försäkringsnummer	Kategorisk
Skatteklass	Kategorisk
Flyttat	Binär
Variabel 1	Kategorisk
Variabel 2	Numerisk
Variabel 3	Kategorisk
Variabel 4	Numerisk
Variabel 5	Numerisk
Variabel 6	Numerisk
Variabel 7	Numerisk
Variabel 8	Kategorisk
Variabel 9	Numerisk
Variabel 10	Numerisk
Variabel 11	Numerisk
Variabel 12	Binär
Variabel 13	Kategorisk
Variabel 14	Kategorisk
Variabel 15	Kategorisk
Variabel 16	Kategorisk
Variabel 17	Kategorisk
Variabel 18	Numerisk
Variabel 19	Binär
Variabel 20	Kategorisk
Variabel 21	Binär
Variabel 22	Binär
Variabel 23	Numerisk
Variabel 24	Numerisk
Variabel 25	Numerisk
Variabel 26	Numerisk
Variabel 27	Numerisk
Variabel 28	Kategorisk
Variabel 29	Binär
Variabel 30	Numerisk

Tabell B.1: *Tabell över variabler. Namn och typ.*

Appendix C

Parameterskattningar

Parameterskattningar GLM

Variabelnamn	*Logodds*	*Oddskvot*	*p-värde*
(Intercept)	-5.11	0.01	0.0000000000000002 ***
Variabel 17 a	-0.31	0.73	0.0000000000000002 ***
Variabel 17 b	0.87	2.39	0.0000000000000002 ***
Variabel 19	0.06	1.07	0.00301 **
Variabel 12	0.23	1.26	0.0000000000000002 ***
Variabel 18	-16.91	0.00	0.0000000000000002 ***
Variabel 6	0	1.00	0.0000000000000002 ***
Variabel 13 a	-0.33	0.72	0.0000000000000002 ***
Variabel 13 b	-0.6	0.55	0.0000000000000002 ***
Variabel 13 b	-0.58	0.56	0.0000000000000002 ***
Variabel 30	0.12	1.13	0.0000000000000002 ***
Variabel 19* Variabel 13 a	-1.21	0.30	0.0000000000000002 ***
Variabel 19 * Variabel 13 b	-0.47	0.63	0.0000000000000002 ***

Tabell C.1: *Skattningar för resultat från GLM på Data 1*

Variabelnamn	*Logodds*	*Oddskvot*	*p-värde*
(Intercept)	-22	0	0.0000000000000002 ***
Variabel 1 a	0.49	1.63	0.000066967598 ***
Variabel 1 b	-0.77	0.46	0.000000000504 ***
Variabel 1 c	-3.88	0.02	0.0000000000000002 ***
Variabel 1 d	-1.23	0.29	0.0000000000000002 ***
Variabel 4	0.01	1.01	0.0000000000000002 ***
Variabel 19	0.74	2.09	0.0000000000000002 ***
Variabel 12	0.24	1.27	0.0000000000000002 ***
Variabel 21	-0.92	0.4	0.0000000000000002 ***
Variabel 18	-9.46	0	0.0000000000000002 ***
Variabel 9	-0.03	0.97	0.0000000000000002 ***
Variabel 13 a	-0.41	0.67	0.0000000000000002 ***
Variabel 13 b	-0.63	0.54	0.0000000000000002 ***
Variabel 30	0.05	1.05	0.0000000000000002 ***
Variabel 18 * Variabel 19	-158.11	0	0.0000000000000002 ***

Tabell C.2: *Skattningar för resultat från GLM på Data 2*

Variabelnamn	*Logodds*	*Oddskvot*	*p-värde*
(Intercept)	-11.09	0	0.0000000000000002 ***
Variabel 4	0	1	0.000000000132 ***
Variabel 19	0.22	1.25	0.0000000000000002 ***
Variabel 12	0.1	1.11	0.0000000000000002 ***
Variabel 21	-0.61	0.54	0.0000000000000002 ***
Variabel 18	-11.35	0	0.0000000000000002 ***
Variabel 6	0	1	0.0000000000000002 ***
Variabel 13 a	-0.21	0.81	0.0000000000000002 ***
Variabel 13 b	-0.24	0.79	0.0000000000000002 ***
Variabel 30	0.1	1.1	0.0000000000000002 ***
Variabel 18 * Variabel 19	-85.53	0	0.0000000000000002 ***

Tabell C.3: *Skattningar för resultat från GLM på Data 3.*

Parameterskattningar GLMM

Variabelnamn	*Logodds*	*Oddskvot*	*p-värde*
(Intercept)	-14.32	0.00	0.0000000000000002 ***
Variabel 4	0.01	1.01	0.0000000000000002 ***
Variabel 12	0.05	1.05	0.00116 **
Variabel 22	-0.65	0.52	0.0000000000000002 ***
Variabel 18	-12.28	0.00	0.0000000000000002 ***
Variabel 13 a	-0.58	0.56	0.0000000000000002 ***
Variabel 13 b	-0.42	0.66	0.0000000000000002 ***
Variabel 30	0.06	1.07	0.0000000000000002 ***
Variabel 12 * Variabel 18	4.25	70.32	0.0000000000785 ***

Tabell C.4: *Skattningar för resultat från GLMM på Data 3.*

Referenser

A. Agresti. *Categorical Data Analysis*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2002.

S. E. Alm and T. Britton. *Stokastik*. Liber, Stockholm, 2008.

K. Antonio and J. Beirlant. Actuarial statistics with generalized linear mixed models. *Insurance: Mathematics and Economics*, 40:58–76, 04 2005.

D. Bates. Mixed models in r using the lme4 package part 5: Generalized linear mixed models, 2011. `https://lme4.r-forge.r-project.org/slides/2011-01-11-Madison/5GLMM.pdf`, URL-datum: 2022-05-09.

R. R. Cerchiara, E. Matthew, and A. Gambrini. Generalized Linear Models in Life Insurance: Decrements and Risk Factor analysis Under Solvency II, 2008. `http://www.actuaries.org/AFIR/Colloquia/Rome2/Cerchiara_Edwards_Gambini.pdf`, URL-datum: 2019-09-30.

S. H. Cox and Y. Lin. Annuity lapse rate modeling: Tobit or not tobit? *Society of actuaries*, 1, 10 2006.

A. Dal Pozzolo, S. Caelen Oliver, Waterschoot, and G. Bontempi. *Racing for Unbalanced Methods Selection*. Springer, Berlin, Heidelberg, 2013.

A. Dal Pozzolo, O. Caelen, R. Johnson, and G. Bontempi. Calibrating probability with undersampling for unbalanced classification. *2015 IEEE Symposium Series on Computational Intelligence*, page 159–166, 12 2015.

H. Doran, D. Bates, P. Bliese, and M. Dowling. Estimating the multilevel rasch model: With the lme4 package. *Journal of Statistical Software, Articles*, 20(2): 1–18, 2007.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning : with applications in R*. Springer, New York, 2013.

C. Kim. Modeling surrender and lapse rates with economic variables. *North American Actuarial Journal*, 9, 10 2005.

G. King and L. Zeng. Logistic regression in rare events data. *Political Analysis*, 9:137–163, Spring 2001a.

G. King and L. Zeng. Explaining rare events in international relations. *International Organization*, 55:693–715, Summer 2001b.

konsumenternas.se. Flytta din pension, 2015. `https://www.konsumenternas.se/pension/pensionens-olika-delar/om-flyttratt`, URL-datum: 2019-10-26.

B. Kouwayè, N. Fonton, F. Rossi, G. Cottrell, and M. Hounkonnou. Variables selection by the lasso method : application to malaria data of tori-bossito (benin). 01 2013.

H. Madsen and P. Thyregod. *Introduction to General and Generalized Linear Models*. CRC Press, Boca Raton, FL, 2010.

P. McCullagh and J. Nelder. *Generalized Linear Models*. Chapman & Hall/CRC, Boca Raton, FL, 1999.

X. Milhaud. Exogenous and endogenous risk factors management to predict surrender behaviours. *ASTIN Bulletin*, 43, 09 2013.

H. Nilsson and J. Fox. Package 'car', 2020. `https://cran.r-project.org/web/packages/car/car.pdf`, URL-datum: 2020-09-20.

E. Ohlsson and B. Johansson. *Non-Life Insurance Pricing with Generalized Linear Models*. Springer-Verlag Berlin Heidelberg, 2010.

M. Rahman and D. N. Davis. Addressing the class imbalance problem in medical datasets. *International Journal of Machine Learning and Computing*, 3:224, 04 2013.

J. Schelldorfer, L. Meier, and P. Bühlmann. Glmmlasso: An algorithm for high-dimensional generalized linear mixed models using l-penalization. *Journal of Computational and Graphical Statistics*, 23(2), 2014.

R. Sundberg. *Lineära statistiska modeller*. Matematisk statistik, 2014.

Q. Wu, D. Debeer, J. Buchholz, J. Hartig, and R. Janssen. Predictors of individual performance changes related to item positions in pisa assessments. *Large-scale Assessments in Education 7*, 5, 2019.